Информатика и её применения

Том 14 Выпуск 1 Год 2020

СОДЕРЖАНИЕ

Информатика и её применения

Том 14 Выпуск 1 Год 2020

СОДЕРЖАНИЕ

Технический редактор *Л. Кокушкина* Художественный редактор *М. Седакова*
Сдано в набор 10.03.20. Подписано в печать 29.03.20. Формат 60 x 84 / 8
Бумага офсетная. Печать цифровая. Усл.-печ. л. 16,51. Уч.-изд. л. 14,5. Тираж 100 экз.
Заказ № 1101
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
Отпечатано в НИПКЦ «Восход-А» с готовых файлов
Москва 109052, ул. Смирновская, д. 25, стр. 3

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2020. Т. 14. Вып. 1. С. 3–9

АСИМПТОТИЧЕСКАЯ РЕГУЛЯРНОСТЬ ВЕЙВЛЕТ-МЕТОДОВ ОБРАЩЕНИЯ ЛИНЕЙНЫХ ОДНОРОДНЫХ ОПЕРАТОРОВ ПО НАБЛЮДЕНИЯМ, РЕГИСТРИРУЕМЫМ В СЛУЧАЙНЫЕ МОМЕНТЫ ВРЕМЕНИ*

О. В. Шестаков[1]

Аннотация: При решении обратных статистических задач часто приходится обращать некоторый линейный однородный оператор, и обычно бывает необходимо использовать методы регуляризации, поскольку наблюдаемые данные, как правило, зашумлены. Популярными методами подавления шума являются процедуры пороговой обработки коэффициентов разложения наблюдаемой функции по специальному базису. Преимущества данных методов заключаются в их вычислительной эффективности и возможности адаптации как к виду оператора, так и к локальным особенностям оцениваемой функции. Анализ погрешностей этих методов представляет собой важную практическую задачу, поскольку позволяет оценить качество как самих методов, так и используемого оборудования. Иногда природа данных такова, что регистрация наблюдений проводится в случайные моменты времени. Если точки отсчетов образуют вариационный ряд, построенный по выборке из равномерного распределения на отрезке регистрации данных, то использование обычных процедур пороговой обработки оказывается адекватным. В данной работе проводится анализ оценки среднеквадратичного риска при обращении линейных однородных операторов и показывается, что при определенных условиях данная оценка является сильно состоятельной и асимптотически нормальной.

Ключевые слова: пороговая обработка; линейный однородный оператор; случайные отсчеты; оценка среднеквадратичного риска

DOI: 10.14357/19922264200101

1 Введение

Во многих прикладных задачах используются математические модели, в которых предполагается, что данные наблюдаются не напрямую, а после некоторого линейного преобразования, и если в наблюдаемых данных содержится шум, то необходимо применять методы регуляризации. Нелинейные методы подавления шума с помощью вейвлет-разложения и процедур пороговой обработки приобрели значительную популярность. Эти методы хорошо изучены, и предложены способы нахождения оптимальных параметров для различных классов функций, описывающих наблюдаемые данные [1–4]. Также изучены статистические свойства оценки среднеквадратичного риска. Показано, что при определенных условиях она оказывается сильно состоятельной и асимптотически нормальной [5–7].

В некоторых ситуациях нет возможности (или она сильно затруднена) регистрировать наблюдения через равные промежутки времени [8]. Иногда природа поступающих данных такова, что регистрация отсчетов производится в случайные моменты времени. В работе [9] показано, что если точки отсчетов образуют вариационный ряд, построенный по выборке из равномерного распределения на отрезке регистрации данных, то при использовании обычной пороговой обработки вейвлет-коэффициентов порядок среднеквадратичного риска остается с точностью до логарифмического множителя равным оптимальному порядку в классе функций, регулярных по Липшицу.

В данной работе рассматривается статистическая оценка среднеквадратичного риска пороговой обработки коэффициентов при обращении линейных однородных операторов и показывается, что статистические свойства этой оценки также не меняются при переходе от фиксированной равномерной сетки отсчетов к случайной. Оценка остается сильно состоятельной и асимптотически нормальной, т. е. сохраняет асимптотическую регулярность.

*Работа выполнена при финансовой поддержке Российского научного фонда (проект 18-11-00155).

[1]Московский государственный университет им. М. В. Ломоносова, кафедра математической статистики факультета вычислительной математики и кибернетики; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, oshestakov@cs.msu.ru

2 Обращение линейных однородных операторов

Линейным однородным оператором называется такое линейное преобразование K искомой функции f, что

$$K\left[f(a(x-x_0))\right] = a^{-\beta}(Kf)\left[a(x-x_0)\right]$$

для любого x_0 и любого $a > 0$. Параметр β называется показателем однородности. Примерами линейных однородных операторов служат оператор интегрирования, преобразование Гильберта и преобразование Абеля. Математические модели с такими операторами используются при решении задач вычислительной томографии, физики плазмы, оптики и др.

Рассмотрим методы обращения оператора K, основанные на свойствах вейвлет-разложений [1–3]. Преимущество этих методов заключается в адаптации не только к свойствам оператора K, но и к свойствам самой искомой функции f.

Вейвлет-разложение функции $f \in L^2(\mathbb{R})$ имеет вид:

$$f = \sum_{j,k\in Z} \langle f, \psi_{j,k}\rangle \psi_{j,k}, \tag{1}$$

где $\psi_{j.k}(x) = 2^{j/2}\psi(2^j x - k)$, а $\psi(x)$ — некоторая материнская вейвлет-функция (семейство $\{\psi_{j,k}\}_{j,k\in Z}$ образует ортонормированный базис в $L^2(\mathbb{R})$). Индекс j в (1) называется масштабом, а индекс k — сдвигом. В дальнейшем будут рассматриваться функции f на отрезке $[0,1]$, равномерно регулярные по Липшицу с некоторым показателем $\gamma > \beta$ и константой Липшица $L > 0$. Для таких функций известно [10], что если вейвлет-функция M раз непрерывно дифференцируема ($M \geqslant \gamma$), имеет M нулевых моментов и достаточно быстро убывает на бесконечности, т. е. существует такая константа $C_A > 0$, что

$$\int_{-\infty}^{\infty} (1 + |t|^\gamma)\,|\psi(t)|\,dt \leqslant C_A,$$

то найдется такая константа $A > 0$, что

$$|\langle f, \psi_{j,k}\rangle| \leqslant \frac{A}{2^{j(\gamma+1/2)}}. \tag{2}$$

Поскольку оператор K линеен и однороден, существуют такие функции $\xi_{j,k}$, что $\langle Kf, \xi_{j,k}\rangle = \langle f, \psi_{j,k}\rangle$ [1]. Функции $\xi_{j,k}$ называются вейглетами. По своим свойствам они похожи на вейвлеты и также представляют собой сдвиги и растяжения некоторой материнской функции ξ.

Далее пусть $\xi_{j,k} = \lambda_{j,k} u_{j,k}$, где $\lambda_{j,k} = \|(K^*)^{-1}\psi_{j,k}\|$. Можно показать, что $\lambda_{j,k} = 2^{\beta j}\lambda_{0,0}$. При этом функция f представляется в виде ряда

$$f = \sum_{j,k\in Z} \lambda_{j,k}\langle Kf, u_{j,k}\rangle \psi_{j,k}. \tag{3}$$

Как видно, в (3) коэффициенты разложения выражаются через Kf, а не через f. Эта формула лежит в основе метода обращения K, который называется вейвлет-вейглет-разложением.

Аналогично по базису вейвлет-фукций можно разложить Kf:

$$Kf = \sum_{j,k\in Z} \langle Kf, \psi_{j,k}\rangle \psi_{j,k}.$$

Функции $\psi_{j,k}$ не обязаны совпадать с функциями в разложении (1), но для удобства будем обозначать их так же. Если функции Kf и ψ удовлетворяют перечисленным выше условиям, то найдется такая константа $C_K > 0$, что

$$|\langle Kf, \psi_{j,k}\rangle| \leqslant \frac{C_K}{2^{j(\gamma+1/2)}}. \tag{4}$$

Далее через $\mathrm{Lip}(\gamma)$ будем обозначать класс регулярных по Липшицу функций, коэффициенты разложения которых удовлетворяют (2) или (4) в зависимости от используемого метода обращения.

Пусть теперь $\lambda_{j,k} = \left\|K^{-1}\psi_{j,k}\right\|$, тогда $\lambda_{j,k} = 2^{\beta j}\lambda_{0,0}$, а функция f представляется в виде ряда [3]:

$$f = \sum_{j,k\in Z} \lambda_{j,k}\langle Kf, \psi_{j,k}\rangle u_{j,k}, \tag{5}$$

где $u_{j,k} = K^{-1}\psi_{j,k}/\lambda_{j,k}$. Функции $u_{j,k}$ не совпадают с функциями в разложении (3), однако по аналогии также называются вейглетами. Формула (5) лежит в основе еще одного метода обращения, который называется вейглет-вейвлет-разложением.

Последовательности $\{u_{j,k}\}$ в обоих разложениях не образуют ортонормированную систему, однако если выполнены некоторые условия гладкости, то они образуют устойчивые базисы [2, 5].

3 Пороговая обработка коэффициентов

Пусть функция $Kf(x)$ задана на отрезке $[0,1]$. Предположим, что отсчеты $Kf(x)$ регистрируются в случайные моменты времени и содержат аддитивный шум, т. е. рассмотрим следующую модель данных:

$$Y_i = Kf(x_i) + z_i, \qquad i = 1,\ldots,N \quad \left(N = 2^J\right),$$

где x_i независимы и равномерно распределены на $[0,1]$, а z_i — не зависящие от x_i и между собой «шумовые» коэффициенты, относительно которых предполагается, что они имеют нормальное распределение с нулевым средним и дисперсией σ^2.

Пусть $0 \leqslant x_{(1)} < \cdots < x_{(N)} \leqslant 1$ — вариационный ряд, построенный по выборке x_i, $i = 1, \ldots, N$. Тогда, перенумеровав Y_i и z_i, получаем модель

$$Y_i = Kf\left(x_{(i)}\right) + \varepsilon_i, \;\; i = 1, \ldots, N\,, \tag{6}$$

где ε_i имеют такую же структуру, как z_i. Наблюдения состоят из пар $(x_{(1)}, Y_1), \ldots, (x_{(N)}, Y_N)$, в которых расстояния между отсчетами в общем случае не равны. При этом $\mathsf{E}x_{(i)} = i/(N+1)$. Наряду с (6) рассмотрим выборку с равными расстояниями между отсчетами

$$\left(\frac{1}{N+1}, Z_1\right), \ldots, \left(\frac{N}{N+1}, Z_N\right). \tag{7}$$

где

$$Z_i = Kf\left(\frac{i}{N+1}\right) + \varepsilon_i, \;\; i = 1, \ldots, N\,.$$

Применяя к выборке (7) дискретное вейглет- или вейвлет-преобразование [6, 7], можно перейти к моделям дискретных коэффициентов.

Для метода вейвлет-вейглет-разложения имеем

$$Z^W_{j,k} = \mu^W_{j,k} + w_{j,k}, \tag{8}$$

где $\mu^W_{j,k} \approx 2^{J/2}\langle Kf, u_{j,k}\rangle$, а шумовые коэффициенты $w_{j,k}$ имеют нормальное распределение с нулевым средним и не являются независимыми. Дисперсии σ_1^2 коэффициентов $w_{j,k}$ зависят от вида оператора и выбранного вейвлет-базиса, но не зависят от j и k [4].

Модель вейглет-вейвлет-коэффициентов имеет вид:

$$Z^V_{j,k} = \mu^V_{j,k} + v_{j,k}, \tag{9}$$

где $\mu^V_{j,k} \approx 2^{J/2}\langle Kf, \psi_{j,k}\rangle$, а шумовые коэффициенты $v_{j,k}$ независимы и имеют нормальное распределение с нулевым средним и дисперсией $\sigma_2^2 = \sigma^2$.

Популярным методом подавления шума является пороговая обработка эмпирических коэффициентов. К коэффициентам в моделях (8) или (9) применяется функция жесткой пороговой обработки $\rho_H(x,T) = y\mathbf{1}(|x| > T)$ или мягкой пороговой обработки $\rho_S(x,T) = \mathbf{sgn}(x)\,(|x| - T)_+$ с порогом T. Смысл пороговой обработки заключается в удалении достаточно маленьких коэффициентов, которые считаются шумом.

Далее для сокращения записи будем обозначать через $W_{j,k}$ «зашумленные» коэффициенты моделей (8) и (9), а через $\mu_{j,k}$ — «чистые» коэффициенты этих моделей. Через $\widehat{W}_{j,k}$ будем обозначать оценки $\mu_{j,k}$, полученные с помощью пороговой обработки. Также дисперсии шумовых коэффициентов σ_1^2 и σ_2^2 будем обозначать одним символом σ^2 (хотя эти дисперсии, вообще говоря, различны).

Если применить дискретное вейглет- или вейвлет-преобразование к выборке (6), то получится набор эмпирических коэффициентов

$$V_{j,k} = \nu_{j,k} + \xi_{j,k},\; j = 0, \ldots, J-1,\; k = 0, \ldots, 2^j - 1,$$

где $\xi_{j,k}$ равны $w_{j,k}$ или $v_{j,k}$ в зависимости от того, используется модель (8) или (9). Здесь $\nu_{j,k}$ — коэффициенты дискретного преобразования «чистой» выборки $Kf\left(x_{(1)}\right), \ldots, Kf\left(x_{(N)}\right)$. В общем случае $V_{j,k}$ не равны $W_{j,k}$ и $\nu_{j,k}$ не равны $\mu_{j,k}$. Однако к $V_{j,k}$ можно применить ту же процедуру, что и к коэффициентам $W_{j,k}$, и получить оценки $\widehat{V}_{j,k}$. В следующих разделах обсуждаются свойства таких оценок.

4 Среднеквадратичный риск пороговой обработки

Среднеквадратичный риск пороговой обработки для выборки со случайными точками отсчетов определим как

$$R_\nu(f,T) = \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \lambda^2_{j,k}\mathsf{E}(\widehat{V}_{j,k} - \mu_{j,k})^2.$$

Также определим среднеквадратичный риск для выборки с равными расстояниями между отсчетами:

$$R_\mu(f,T) = \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \lambda^2_{j,k}\mathsf{E}(\widehat{W}_{j,k} - \mu_{j,k})^2.$$

Выбор величины порога — одна из основных задач при пороговой обработке. Для класса $\mathrm{Lip}(\gamma)$ близким к оптимальному является порог

$$T_\gamma = \sigma\sqrt{\frac{4\gamma}{2\gamma+1}(1+2\beta)\ln 2^J}\,.$$

Используя результаты работ [1, 4], можно убедиться, что справедливо следующее утверждение о порядке $R_\mu(f, T_\gamma)$.

Теорема 1. *Пусть $Kf \in \mathrm{Lip}(\gamma)$ на отрезке $[0,1]$ с $\gamma > \beta$ и вейвлет-функция удовлетворяет перечисленным выше условиям. Тогда при выборе порога T_γ справедливо*

$$R_\mu(f, T_\gamma) \leqslant C \cdot 2^{J(2\beta+1)/(2\gamma+1)} J^{(2\gamma+2\beta+2)/(2\gamma+1)},$$

где C — некоторая положительная константа.

Также, повторяя рассуждения работы [9], можно показать, что при $\gamma > \max(\beta, 1/2)$ аналогичное утверждение справедливо для $R_\nu(f, T_\gamma)$. Таким образом, замена равноотстоящих точек отсчетов на случайные не ухудшает оценку порядка среднеквадратичного риска.

5 Свойства статистической оценки среднеквадратичного риска

В практических ситуациях вычислить значение среднеквадратичного риска нельзя, поскольку оно зависит от ненаблюдаемых «чистых» коэффициентов. Однако можно построить его оценку, используя только наблюдаемые данные. Эта оценка определяется выражением [10]:

$$\widehat{R}_\nu(f, T) = \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \lambda_{j,k}^2 F\left[V_{j,k}, T\right], \qquad (10)$$

где

$$F\left[V_{j,k}, T\right] = \begin{cases} \left(V_{j,k}^2 - \sigma^2\right) \mathbf{1}(|V_{j,k}| \leqslant T) + \sigma^2 \mathbf{1}(|V_{j,k}| > T) \\ \qquad \text{в случае жесткой пороговой обработки;} \\ \left(V_{j,k}^2 - \sigma^2\right) \mathbf{1}(|V_{j,k}| \leqslant T) + \\ \qquad + \left(\sigma^2 + T^2\right) \mathbf{1}(|V_{j,k}| > T) \\ \qquad \text{в случае мягкой пороговой обработки.} \end{cases}$$

Оценка (10) дает возможность получить представление о погрешности, с которой оценивается функция f. Докажем утверждение о ее асимптотической нормальности.

Теорема 2. *Пусть $Kf \in \mathrm{Lip}(\gamma)$ на отрезке $[0, 1]$ с $\gamma > 1/2 + \beta$ и вейвлет-функция удовлетворяет перечисленным выше условиям. Тогда при жесткой и мягкой пороговых обработках*

$$\mathsf{P}\left(\frac{\widehat{R}_\nu\left(f, T_\gamma\right) - R_\nu\left(f, T_\gamma\right)}{D_J} < x\right) \to \Phi(x) \quad \text{при } J \to \infty,$$

где $\Phi(x)$ — функция распределения стандартного нормального закона. При использовании метода вейглет-вейвлет-разложения

$$D_J = \sigma^2 \lambda_{0,0}^2 \sqrt{2(2^{4\beta+1} - 1)^{-1}}\, 2^{J(1/2+2\beta)},$$

а при использовании метода вейвлет-вейглет-разложения

$$D_J = C_\beta 2^{J(1/2+2\beta)},$$

где константа C_β имеет более сложную структуру [6] *и зависит от используемого базиса и вида оператора K.*

Доказательство. Докажем теорему для случая жесткой пороговой обработки. В случае мягкой пороговой обработки доказательство аналогично.

Наряду с $\widehat{R}_\nu(f, T_\gamma)$ рассмотрим

$$\widehat{R}_\mu(f, T_\gamma) = \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \lambda_{j,k}^2 F[W_{j,k}, T_\gamma]$$

и запишем разность $\widehat{R}_\nu(f, T_\gamma) - R_\nu(f, T_\gamma)$ в виде

$$\widehat{R}_\nu\left(f, T_\gamma\right) - R_\nu\left(f, T_\gamma\right) = \widehat{R}_\mu\left(f, T_\gamma\right) - R_\mu\left(f, T_\gamma\right) + \widetilde{R},$$

где

$$\widetilde{R} = \widehat{R}_\nu\left(f, T_\gamma\right) - \widehat{R}_\mu\left(f, T_\gamma\right) - \left(R_\nu(f, T_\gamma) - R_\mu\left(f, T_\gamma\right)\right).$$

В [6, 11] показано, что при $\gamma > 1/2 + \beta$

$$\mathsf{P}\left(\frac{\widehat{R}_\mu(f, T_\gamma) - R_\mu\left(f, T_\gamma\right)}{D_J} < x\right) \to \Phi(x) \quad \text{при } J \to \infty.$$

Следовательно, для доказательства теоремы достаточно показать, что

$$\frac{\widetilde{R}}{2^{J(1/2+2\beta)}} \xrightarrow{\mathsf{P}} 0 \text{ при } J \to \infty.$$

Если $\gamma > \max(\beta, 1/2)$, то в силу теоремы 1 и аналогичного утверждения для $R_\nu(f, T)$

$$\frac{R_\nu(f, T_\gamma) - R_\mu(f, T_\gamma)}{2^{J(1/2+2\beta)}} \to 0 \text{ при } J \to \infty.$$

Далее пусть

$$j_0 \approx \frac{J}{2\gamma + 1} + \frac{1}{2\gamma + 1} \log_2 J.$$

Представим $\widehat{R}_\nu(f, T_\gamma) - \widehat{R}_\mu(f, T_\gamma)$ в виде

$$\widehat{R}_\nu(f, T_\gamma) - \widehat{R}_\mu(f, T_\gamma) = S_1 + S_2,$$

где

$$S_1 = \sum_{j=0}^{j_0-1} \sum_{k=0}^{2^j-1} \lambda_{j,k}^2 \left(F\left[V_{j,k}, T_\gamma\right] - F\left[W_{j,k}, T_\gamma\right]\right);$$

$$S_2 = \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \lambda_{j,k}^2 \left(F\left[V_{j,k}, T_\gamma\right] - F\left[W_{j,k}, T_\gamma\right]\right).$$

Поскольку как в случае жесткой, так и в случае мягкой пороговой обработки

$$\left.\begin{aligned} |F[V_{j,k},T_\gamma]| &\leq T_\gamma^2+\sigma^2,\\ |F[W_{j,k},T_\gamma]| &\leq T_\gamma^2+\sigma^2 \ \text{ п.в.,}\end{aligned}\right\} \qquad (11)$$

то для $\gamma > \max(\beta, 1/2)$

$$\frac{S_1}{2^{J(1/2+2\beta)}} \xrightarrow{\mathsf{P}} 0 \text{ при } J\to\infty.$$

Далее

$$\begin{aligned} S_2 &= \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2\left(F\left[V_{j,k},T_\gamma\right]-F[W_{j,k},T_\gamma]\right) = \\ &= \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2(V_{j,k}^2-W_{j,k}^2)+ \\ &+\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2(W_{j,k}^2-2\sigma^2)\times \\ &\times \mathbf{1}(|V_{j,k}|\leqslant T_\gamma,|W_{j,k}|>T_\gamma)+ \\ &+\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2(2\sigma^2-V_{j,k}^2)\mathbf{1}(|V_{j,k}|>T_\gamma,|W_{j,k}|\leqslant T_\gamma)+ \\ &+\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2(W_{j,k}^2-V_{j,k}^2)\times \\ &\times \mathbf{1}(|V_{j,k}|>T_\gamma,|W_{j,k}|>T_\gamma). \end{aligned} \qquad (12)$$

Рассмотрим сумму

$$\begin{aligned} &\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2(V_{j,k}^2-W_{j,k}^2) = \\ &= \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2(\nu_{j,k}^2-\mu_{j,k}^2)+ \\ &+2\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\xi_{j,k}\lambda_{j,k}^2(\nu_{j,k}-\mu_{j,k}). \end{aligned}$$

Учитывая устойчивость вейглет-базиса и результаты работ [9, 12], можно показать, что условное распределение этой суммы при фиксированных x_i нормально с математическим ожиданием $\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^J-1}\lambda_{j,k}^2(\nu_{j,k}^2-\mu_{j,k}^2)$ и дисперсией, не превосходящей $C_\lambda\sigma^2\sum_{j=j_0}^{J-1}\sum_{k=0}^{2J-1}\lambda_{j,k}^4(\nu_{j,k}-\mu_{j,k})^2$, где константа C_λ зависит от выбранного базиса и вида оператора K.

Так как $Kf\in \mathrm{Lip}(\gamma)$, то, повторяя рассуждения работы [12], можно показать, что

$$\begin{aligned} &\mathsf{E}_x\left|\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2\left(\nu_{j,k}^2-\mu_{j,k}^2\right)\right| \leqslant \\ &\leqslant C_1\cdot 2^{J(1+2\beta-\min(2,\gamma))}+C_2\cdot 2^{J(1+(2\beta-2\gamma)/(2\gamma+1))}, \end{aligned}$$

где C_1 и C_2 — некоторые положительные константы.

Также, используя оценку из работы [9], получаем, что

$$\begin{aligned} &\mathsf{E}_x\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^4\left(\nu_{j,k}-\mu_{j,k}\right)^2 \leqslant \\ &\leqslant C_3\cdot 2^{J(1+4\beta-\min(1,\gamma))}, \end{aligned} \qquad (13)$$

где C_3 — некоторая положительная константа. Следовательно, если $\gamma > \max(\beta, 1/2)$, то, применяя неравенство Маркова, получаем, что

$$\frac{1}{2^{J(1/2+2\beta)}}\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^2\left(\nu_{j,k}^2-\mu_{j,k}^2\right)\xrightarrow{\mathsf{P}}0\,,$$

$$\frac{1}{2^{J(1+4\beta)}}\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1}\lambda_{j,k}^4(\nu_{j,k}-\mu_{j,k})^2\xrightarrow{\mathsf{P}}0$$

при $J\to\infty$. Таким образом,

$$\frac{\sum_{j=j_0}^{J-1}\sum_{k=0}^{2^J-1}\lambda_{j,k}^2\left(V_{j,k}^2-W_{j,k}^2\right)}{2^{J(1/2+2\beta)}}\xrightarrow{\mathsf{P}}0 \text{ при } J\to\infty\,.$$

В оставшихся суммах в (12) содержатся индикаторы, в которых либо $|V_{j,k}|>T_\gamma$, либо $|W_{j,k}|>T_\gamma$, причем в силу (2) и (4) для всех слагаемых $|\mu_{jk}|\leq C_4J^{-1/2}$ с некоторой константой $C_4>0$. Повторяя рассуждения из работы [13] с использованием (13), можно показать, что эти суммы при делении на $2^{J(1/2+2\beta)}$ также сходятся к нулю по вероятности. Теорема доказана.

Помимо асимптотической нормальности оценка (10) также обладает свойством сильной состоятельности.

Теорема 3. *Пусть выполнены условия теоремы* 2. *Тогда при жесткой и мягкой пороговых обработках для любого* $\alpha>1/2$

$$\frac{\widehat{R}_\nu(f,T_\gamma)-R_\nu(f,T_\gamma)}{2^{(\alpha+2\beta)J}}\to 0 \text{ п.в. при } J\to\infty.$$

Поскольку выполнено (11) и при фиксированных x_i слагаемые в (10) условно независимы (или слабо зависимы), доказательство этой теоремы аналогично доказательству соответствующего утверждения из работы [14].

Литература

1. *Donoho D.* Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition // Appl. Comp. Harm. Anal., 1995. Vol. 2. P. 101–126.
2. *Lee N.* Wavelet-vaguelette decompositions and homogenous equations. — West Lafayette, IN, USA: Purdue University, 1997. PhD Thesis. 103 p.
3. *Abramovich F., Silverman B. W.* Wavelet decomposition approaches to statistical inverse problems // Biometrika, 1998. Vol. 85. No. 1. P. 115–129.
4. *Johnstone I. M.* Wavelet shrinkage for correlated data and inverse problems adaptivity results // Statist. Sinica, 1999. Vol. 9. P. 51–83.
5. *Кудрявцев А. А., Шестаков О. В.* Асимптотика оценки риска при вейглет-вейвлет разложении наблюдаемого сигнала // T-Comm — Телекоммуникации и транспорт, 2011. № 2. С. 54–57.
6. *Ерошенко А. А., Шестаков О. В.* Асимптотическая нормальность оценки риска при вейвлет-вейглет-разложении функции сигнала в модели с коррелированным шумом // Вестн. Моск. ун-та. Сер. 15: Вычисл. матем. и киберн., 2014. № 3. С. 23–30.
7. *Ерошенко А. А., Кудрявцев А. А., Шестаков О. В.* Предельное распределение оценки риска метода вейглет-вейвлет-разложения сигнала в модели с коррелированным шумом // Вестн. Моск. ун-та. Сер. 15: Вычисл. матем. и киберн., 2015. № 1. С. 12–18.
8. *Cai T., Brown L.* Wavelet shrinkage for nonequispaced samples // Ann. Stat., 1998. Vol. 26. No. 5. P. 1783–1799.
9. *Cai T., Brown L.* Wavelet estimation for samples with random uniform design // Stat. Probabil. Lett., 1999. Vol. 42. P. 313–321.
10. *Mallat S.* A wavelet tour of signal processing. — New York, NY, USA: Academic Press, 1999. 857 p.
11. *Шестаков О. В.* Вероятностно-статистические методы анализа и обработки сигналов на основе вейвлет-алгоритмов. — М.: АРГАМАК-МЕДИА, 2016. 200 с.
12. *Шестаков О. В.* Свойства вейвлет-оценок сигналов, регистрируемых в случайные моменты времени // Информатика и её применения, 2019. Т. 13. Вып. 2. С. 30–35.
13. *Шестаков О. В.* Аппроксимация распределения оценки риска пороговой обработки вейвлет-коэффициентов нормальным распределением при использовании выборочной дисперсии // Информатика и её применения, 2010. Т. 4. Вып. 4. С. 73–81.
14. *Shestakov O. V.* On the strong consistency of the adaptive risk estimator for wavelet thresholding // J. Math. Sci., 2016. Vol. 214. No. 1. P. 115–118.

Поступила в редакцию 16.12.19

ASYMPTOTIC REGULARITY OF THE WAVELET METHODS OF INVERTING LINEAR HOMOGENEOUS OPERATORS FROM OBSERVATIONS RECORDED AT RANDOM TIMES

O. V. Shestakov[1,2]

[1]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: When solving inverse statistical problems, it is often necessary to invert some linear homogeneous operator and it is usually necessary to use regularization methods, since the observed data are noisy. Popular methods for noise suppression are the procedures of thresholding the expansion coefficients of the observed function. The advantages of these methods are their computational efficiency and the ability to adapt to both the type of operator and the local features of the estimated function. An analysis of the errors of these methods is an important practical task, since it allows one to evaluate the quality of both the methods themselves and the equipment used. Sometimes, the nature of the data is such that observations are recorded at random times. If the observation points form a variational series constructed from a sample of a uniform distribution on the data recording interval, then the use of conventional threshold processing procedures is adequate. The present author analyzes the estimate of the mean square risk in the problem of inversion of linear homogeneous operators and demonstrates that under certain conditions, this estimate is strongly consistent and asymptotically normal.

Keywords: threshold processing; linear homogeneous operator; random observation points; mean square risk estimate

DOI: 10.14357/19922264200101

Acknowledgments

This research is supported by the Russian Science Foundation (project No. 18-11-00155).

References

1. Donoho, D. 1995. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comp. Harm. Anal.* 2:101–126.
2. Lee, N. 1997. Wavelet-vaguelette decompositions and homogenous equations. West Lafayette, IN: Purdue University. PhD Thesis. 103 p.
3. Abramovich, F., and B. W. Silverman. 1998. Wavelet decomposition approaches to statistical inverse problems. *Biometrika* 85(1):115–129.
4. Johnstone, I. M. 1999. Wavelet shrinkage for correlated data and inverse problems adaptivity results. *Statist. Sinica* 9:51–83.
5. Kudryavtsev, A. A., and O. V. Shestakov. 2011. Asimptotika otsenki riska pri veyglet-veyvlet razlozhenii nablyudaemogo signala [The average risk assessment of the wavelet decomposition of the signal]. *T-Comm — Telekommunikatsii i transport* [T-Comm — Telecommunications and Transport] 2:54–57.
6. Eroshenko, A. A., and O. V. Shestakov. 2014. Asymptotic normality of estimating risk upon the wavelet-vaguelette decomposition of a signal function in a model with correlated noise. *Mosc. Univ. Comput. Math. Cybern.* 38(3):110–117.
7. Eroshenko, A. A., A. A. Kudryavtsev, and O. V. Shestakov. 2015. Limit distribution of a risk estimate using the vaguelette-wavelet decomposition of signals in a model with correlated noise. *Mosc. Univ. Comput. Math. Cybern.* 39(1):6–13.
8. Cai, T., and L. Brown. 1998. Wavelet shrinkage for nonequispaced samples. *Ann. Stat.* 26(5):1783–1799.
9. Cai, T., and L. Brown. 1999. Wavelet estimation for samples with random uniform design. *Stat. Probabil. Lett.* 42:313–321.
10. Mallat, S. 1999. *A wavelet tour of signal processing.* New York, NY: Academic Press. 857 p.
11. Shestakov, O. V. 2016. *Veroyatnostno-statisticheskie metody analiza i obrabotki signalov na osnove veyvlet-algoritmov* [Probabilistic-statistical methods of signal analysis and processing based on wavelet algorithms]. Moscow: Argamak-Media Publs. 200 p.
12. Shestakov, O. V. 2019. Svoystva veyvlet-otsenok signalov, registriruemykh v sluchaynye momenty vremeni [Properties of wavelet estimates of signals recorded at random time points]. *Informatika i ee Primeneniya — Inform. Appl.* 13(2):30–35.
13. Shestakov, O. V. 2010. Approksimatsiya raspredeleniya otsenki riska porogovoy obrabotki veyvlet-koeffitsientov normal'nym raspredeleniem pri ispol'zovanii vyborochnoy dispersii [Normal approximation for distribution of risk estimate for wavelet coefficients thresholding when using sample variance]. *Informatika i ee Primeneniya — Inform. Appl.* 4(4):73–81.
14. Shestakov, O. V. 2016. On the strong consistency of the adaptive risk estimator for wavelet thresholding. *J. Math. Sci.* 214(1):115–118.

Received December 16, 2019

Contributor

Shestakov Oleg V. (b. 1976) — Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; oshestakov@cs.msu.su

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2020. Т. 14. Вып. 1. С. 10–16

АНАЛИЗ КОНФИГУРАЦИЙ LSTM-СЕТЕЙ ДЛЯ ПОСТРОЕНИЯ СРЕДНЕСРОЧНЫХ ВЕКТОРНЫХ ПРОГНОЗОВ*

А. К. Горшенин[1], В. Ю. Кузьмин[2]

Аннотация: Проанализированы 36 конфигураций архитектур LSTM-сетей (Long Short-Term Memory, долгая краткосрочная память) для построения прогнозов длительностью до 70 шагов по данным, размер которых составляет 300–500 элементов. Для вероятностной аппроксимации наблюдений применена модель на основе конечных смесей нормальных распределений, поэтому в качестве исходных данных для прогнозирования использованы математическое ожидание, дисперсия, коэффициенты асимметрии и эксцесса этих смесей. Определены оптимальные конфигурации нейронных сетей и продемонстрирована практическая возможность построения качественных среднесрочных прогнозов при ограниченном времени обучения. Полученные результаты важны для развития вероятностно-статистического подхода к описанию эволюции турбулентных процессов в магнитоактивной высокотемпературной плазме.

Ключевые слова: LSTM; прогнозирование; глубокое обучение; высокопроизводительные вычисления; CUDA

DOI: 10.14357/19922264200102

1 Введение

Традиционно при анализе турбулентного состояния плазмы исследователи пытаются установить связь между скоростями роста неустойчивых режимов, условиями их возбуждения и спектрами флуктуаций, полученными с помощью гирокинетического моделирования или в реальных экспериментах. При этом основное внимание уделяется стационарным режимам, необходимым для работы в устойчивом состоянии будущего управляемого термоядерного реактора, а нелинейной стадией развития турбулентности, ее насыщения, образования вихрей и их хаотизации обычно пренебрегают. Поэтому исследования, представленные в статье [1], ориентированы на развитие статистического подхода к описанию эволюции турбулентных процессов в магнитоактивной высокотемпературной плазме.

В качестве экспериментальных данных были использованы ансамбли диагностик, которые учитывают флуктуации плотности плазмы даже в центральных областях плазменного столба (подробно физические эксперименты описаны в статье [2]).

Для решения указанной задачи был использован математический аппарат на основе конечных нормальных смесей, методов их скользящего разделения [3, 4] и оценивания параметров с помощью алгоритмов EM-типа (expectation-maximization) [5–10]. С учетом эволюции во времени характеристик предложенных моделей естественным выглядит вопрос о возможности их прогнозирования.

Методы машинного обучения и нейронные сети в исследованиях турбулентной плазмы используются, возможно, не слишком часто, однако позволяют добиваться достаточно заметных результатов как в вопросах моделирования наблюдаемых явлений [11–14], так и в задачах анализа и прогнозирования нестабильностей и разрушительных для стеллараторов и токамаков эффектов [15, 16]. Ранее авторами были предложены несколько архитектур нейронных сетей [17, 18], в том числе и рекуррентных, для эффективного краткосрочного прогнозирования (т. е. от одного до трех наблюдений вперед), а также разработаны инструменты совместного (векторного) предсказания значений с помощью сетей прямого распространения [19] математического ожидания, дисперсии, коэффициентов асимметрии и эксцесса конечных нормальных смесей в задачах обработки указанных экспериментальных данных.

В данной статье рассматриваются вопросы среднесрочного прогнозирования, т. е. предсказания

*Работа выполнена при частичной поддержке РФФИ (проекты 19-07-00352 и 18-29-03100) и Стипендии Президента Российской Федерации молодым ученым и аспирантам (СП-538.2018.5). Для ускорения обучения был использован гибридный высокопроизводительный вычислительный комплекс ЦКП «Информатика» ФИЦ ИУ РАН: http://ckp.frccsc.ru.

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова, agorshenin@frccsc.ru

[2]ООО «Вай2Гео», shadesilent@yandex.ru

значений на 10–70 шагов по входных векторам, размер которых составляет 300–500 элементов, с помощью LSTM-сетей [20] и их обучения и валидации на современном высокопроизводительном вычислительном оборудовании.

2 Конфигурации LSTM-сетей для среднесрочных прогнозов

В этом разделе рассмотрим набор базовых архитектур и выбранные настройки гиперпараметров нейронных сетей для построения среднесрочных прогнозов моментов конечных смешанных вероятностных распределений. Как показано в статье [19], векторное прогнозирование (т. е. совместное предсказание значений сразу для всех четырех рассматриваемых моментов) ведет к уменьшению времени обучения по сравнению с последовательной обработкой отдельных рядов, при этом общая точность прогнозов для каждого ряда только увеличивается.

Для построения среднесрочных векторных прогнозов были использованы LSTM-архитектуры — разновидность рекуррентных нейронных сетей, успешно зарекомендовавшая себя при решении задач обработки и прогнозирования различных временны́х рядов. Выбраны три базовые конфигурации архитектур.

I. Входной слой — скрытый слой из 100 нейронов — выходной слой.

II. Входной слой — два скрытых слоя из 150 и 100 нейронов — выходной слой.

III. Входной слой — три скрытых слоя из 200, 150 и 100 нейронов — выходной слой.

Выбор такого количества нейронов связан с тем, что по результатам предварительного анализа различных возможных конфигураций было установлено, что постепенное уменьшение числа нейронов в скрытых слоях приводит к повышению скорости обучения при сравнимой точности результатов.

На вход каждой нейронной сети подается $4N$ наблюдений, где N — ширина окна, на основе которого делается предсказание, на выходе — $4M$ наблюдений, где M — выбранная длина прогноза. В данной работе рассмотрены следующие наборы значений:

$$N = \{300, 400, 500\};\ \ M = \{10, 30, 50, 70\}.$$

Обучение проводится на протяжении 500 эпох, при этом возможна досрочная остановка при отсутствии значимого убывания функции потерь в течение 35 эпох подряд. В качестве метода оптимизации выбран `Adam`, так как аналогично случаю построения краткосрочных прогнозов [18, 19] использование других функций (`NAdam`, `AdaDelta`, `SGD`, `AdaMax` [21]) не приводит к улучшению результатов. Использована функция активации рациональная сигмоида, определяемая выражением $x(1+|x|)^{-1}$. Эффект переобучения при таком выборе гиперпараметров для тестовых данных не наблюдается, применение дропаут-слоев [22] в описанных конфигурациях приводит к ухудшению результатов, поэтому от их использования в итоговых вариантах было решено отказаться.

3 Алгоритм среднесрочного прогнозирования нестационарных данных и программные реализации

В данном разделе опишем алгоритм среднесрочного прогнозирования нестационарных данных, которыми и выступают анализируемые моменты смесей. Ниже он представлен в виде псевдокода.

Для сравнения скорости построения прогнозов данный алгоритм тестировался на двух различных вычислительных системах. Первая из них — персональный компьютер (ПК) с процессором `i7-8750H` и видеокартой `NVIDIA GeForce RTX 2070`; вторая — гибридный высокопроизводительный вычислительный кластер (ГВВК) с двумя процессорами `Power9` с тактовой частотой 2,0 ГГц (20 ядер) и 4 видеокартами `NVIDIA Volta V100` (общий объем памяти 16 ГБ). Обучение нейронных сетей реализовано на языке программирования `Python` (версия 3.6 с использованием библиотек `Keras-GPU 2.1` и `Tensorflow-GPU 1.2`). Для инициализации системы использована технология контейнеризации [23], что позволило существенно сократить время развертывания, накладные расходы на получение и установку совместимых версий пакетов и абстрагировать приложение от хоста. В текущей версии время развертывания приложения на ГВВК составило около 3–5 с. Было установлено, что использование ГВВК повышает скорость обучения в 12–27 раз для всех рассмотренных архитектур. Например, характерная продолжительность одной эпохи обучения для архитектуры I при предсказании на 30 шагов вперед по известной выборке в 300 элементов на ГВВК составила 0,93 с, а на ПК — 24,5 с. Поэтому все рассмотренные в следующем разделе конфигурации тестировались именно на ресурсах ГВВК.

Алгоритм среднесрочного прогнозирования нестационарных данных

```
data = [exp, var, kurt, skew]; // Четыре момента смесей
Цикл i = 1 : length(data)
  NORMALIZEARRAY(data[i]); // Нормализация каждого момента

  // Формирование списка окон с последовательным сдвигом на 1 шаг
windows = [];
Цикл i = 1 : length(data) − windowLength
  window = data[i, i + windowLength];
  windows.push(window);

[train, test] = DIVIDETRAINTESTDATA(windows, 0.7);
nn = SEQUENTIAL(); // Создание объекта нейронной сети
Цикл idx = 1 : layerNumber
  NN.ADDLAYER(LSTM); // LSTM-слои

NN.ADDLAYER(Flatten); // Слой выравнивания
NN.ADDLAYER(Linear[4 ∗ predictionLength]);
   // Обучение модели с использованием Adam и RMSE
NN.COMPILE(optimizer = adam, loss = rmse, metrics = [accuracy, rmse, mae]);
NN.FIT(train, validation = test, callbacks = [EarlyStopping, ReduceLROnPlateau], shuffle = False)
```

4 Выбор оптимальных конфигураций нейросетей для векторных прогнозов различной длины

Для анализа результатов прогнозирования использованы классические метрики — среднеквадратичная ошибка (RMSE, Root Mean Square Error) и средняя абсолютная ошибка (MAE, Mean Absolute Error). Исходные данные преобразуются таким образом, чтобы обрабатываемые наблюдения принадлежали отрезку $[0, 1]$. Соотношение между обучающими и тестовыми наборами составляет 70% к 30%.

В таблице приведены данные по ошибкам и скорости обсчета данных на ГВВК для 36 различных конфигураций нейронных сетей. Наибольший прирост точности получается в результате перехода от архитектуры типа I с одним скрытым слоем к архитектуре типа II с двумя скрытым слоями. В среднем архитектура II дает на 18% меньшую ошибку RMSE и на 20% меньшую ошибку MAE. Однако подобное увеличение точности ведет к повышению длительности обучения в среднем на 42%.

Важную роль играет соотношение между размером прогноза и известным окном. Эмпирически (см. также таблицу) установлено, что если данное отношение меньше 0,1, то эффект от использования архитектуры типа II менее заметен. Так, переход от одного скрытого слоя к двум скрытым слоям при предсказании 70 наблюдений по 300 предшествующим значениям (упомянутое соотношение составляет около 0,23) ведет к уменьшению RMSE на 37% и MAE на 44%, а аналогичный переход при предсказании 10 наблюдений по окну в 500 (т. е. 0,02) — всего лишь к уменьшению на 7,2% и 6% соответственно. В среднем для всех архитектур время обучения увеличивается на 44%, RMSE уменьшается на 23%, MAE — на 28%.

Добавление еще одного скрытого слоя (т. е. переход к архитектуре типа III) увеличивает время обучения, при этом точность остается сопоставимой с точностью архитектуры II. В среднем использование третьего слоя уменьшает RMSE на 3% и MAE на 2%, при этом время обучения возрастает на 14%. Также стоит отметить, что при этом в ряде случаев (см., например, предсказание на 30 шагов для окна в 400 элементов, предсказание на 50 шагов для окна в 500 элементов) ошибки могут несколько возрасти — обучение моделей с большим числом скрытых слоев требует менее строгого ограничения по числу эпох. Дальнейшее наращивание числа слоев — даже при условии изменения критериев останова обучения — не дает значимого прироста точности.

Была рассмотрена модификация архитектуры II, в которой во втором скрытом слое вместо 100 нейронов использованы 150. На тестовых данных было установлено, что при увеличении времени обучения на 5%–10% увеличение точности составляет (в терминах рассматриваемых метрик) около 1%. Поэтому использование такой конфигурации не представляется оптимальным.

На рисунке представлены примеры графиков рассчитанных моментов и сделанных прогнозов

Сравнение точности прогнозирования в метриках RMSE и MAE, а также времени обучения для различных конфигураций

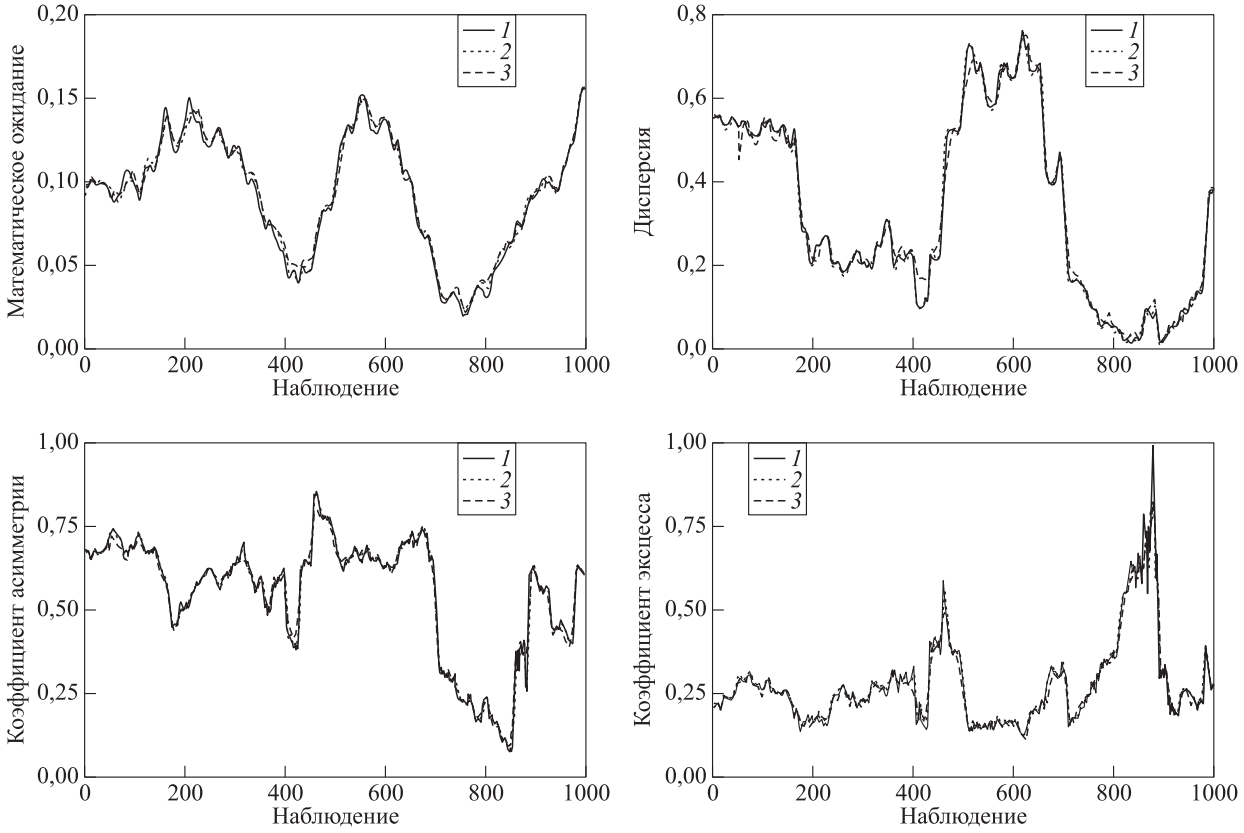
Размер окна	Длительность прогноза	Конфигурация архитектуры	RMSE	MAE	Время обучения, с
300	10	I	0,029	0,020	231,61
		II	0,025	0,017	391,87
		III	0,025	0,017	445,22
	30	I	0,033	0,023	275,05
		II	0,028	0,019	401,55
		III	0,026	0,018	456,62
	50	I	0,034	0,023	291,20
		II	0,029	0,02	415,51
		III	0,026	0,019	473,71
	70	I	0,044	0,034	290,25
		II	0,028	0,019	420,83
		III	0,025	0,017	486,46
400	10	I	0,028	0,018	628,06
		II	0,023	0,015	529,25
		III	0,023	0,015	534,85
	30	I	0,030	0,020	350,74
		II	0,025	0,017	519,58
		III	0,026	0,02	509,92
	50	I	0,031	0,022	356,43
		II	0,024	0,016	528,15
		III	0,022	0,014	615,32
	70	I	0,028	0,024	361,19
		II	0,024	0,018	534,51
		III	0,025	0,017	629,95
500	10	I	0,026	0,017	457,90
		II	0,024	0,016	667,49
		III	0,022	0,014	804,82
	30	I	0,029	0,02	463,36
		II	0,023	0,016	683,53
		III	0,022	0,015	824,50
	50	I	0,026	0,018	481,52
		II	0,022	0,014	696,62
		III	0,023	0,016	842,29
	70	I	0,027	0,019	495,84
		II	0,024	0,017	714,66
		III	0,024	0,017	840,07

на 1 и 50 шагов для окна в 500 наблюдений для архитектуры II.

При рассмотрении наблюдения с номером n строится среднесрочный прогноз по окну, составленному из наблюдений с номерами от $n - 500 - l$ до $n - l$, где l соответствует длине прогноза (в данном случае — 1 или 50). Спрогнозированные ряды хорошо приближают исходные (даже с учетом их явной нестационарности), при этом существенные изменения локальных трендов, выбросы, ведут к естественному увеличению ошибки (например, см. наблюдения 820–900 для коэффициента эксцесса). Этот эффект может быть частично скомпенсирован менее строгими ограничениями на обучение модели. Аналогично среднесрочный прогноз не всегда способен предсказать точную величину изменения поведения ряда, наиболее сильно это проявляется на наблюдениях 400–420, 430–460 для дисперсии. При этом предсказания среднесрочных трендов (рост или падение ряда на протяжении десятков наблюдений) достаточно точны, например номер наблюдения с максимальной величиной у краткосрочного/среднесрочного прогноза и исходных данных на наблюдениях 620–640 для коэффициента асимметрии совпадает.

5 Заключение

В работе рассмотрено несколько типов архитектур LSTM-сетей, которые при различных конфигурациях позволяют с достаточной точностью (в терминах метрик RMSE и MAE) и за умеренное время строить среднесрочные прогнозы существенно нестационарных рядов. При этом удается доста-



Сравнение значений исходных данных (*1*), прогноза на 1 шаг (*2*) и на 50 шагов (*3*) для четырех моментов

точно успешно учитывать среднесрочные тренды, присутствующие в данных. Установлено, что точность прогнозирования выбросов и быстрой смены направлений локальных трендов может быть повышена за счет увеличения числа эпох обучения, например до 5000, однако это ведет к существенному замедлению построения нейросетевых моделей и практически не требуется для точного прогнозирования большинства участков анализируемых рядов. Таким образом, создана методология обработки подобных данных, сформулированы практические рекомендации по развертыванию реализующих ее программных решений на высокопроизводительном вычислительном оборудовании. Полученные результаты важны для развития вероятностно-статистического подхода к описанию эволюции турбулентных процессов в магнитоактивной высокотемпературной плазме.

Литература

1. *Batanov G. M., Borzosekov V. D., Gorshenin A. K., Kharchev N. K., Korolev V. Yu., Sarskyan K. A.* Evolution of statistical properties of microturbulence during transient process under electron cyclotron resonance heating of the L-2M stellarator plasma // Plasma Phys. Contr. F., 2019. Vol. 61. Iss. 7. Art. No. 075006.
2. *Batanov G. M., Berezhetskii M. S., Borzosekov V. D., et al.* Reaction of turbulence at the edge and in the center of the plasma column to pulsed impurity injection caused by the sputtering of the wall coating in L-2M stellarator // Plasma Phys. Rep., 2017. Vol. 43. Iss. 8. P. 818–823.
3. *Королев В. Ю.* Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. — М.: Изд-во Моск. ун-та, 2011. 512 с.
4. *Gorshenin A., Korolev V., Kuzmin V., Zeifman A.* Coordinate-wise versions of the grid method for the analysis of intensities of non-stationary information flows by moving separation of mixtures of gamma-distribution // 27th European Conference on Modelling and Simulation Proceedings. — Dudweiler, Germany: Digitaldruck Pirrot GmbH, 2013. P. 565–568.
5. *Lee S. X., Leemaqz K. L., McLachlan G. J.* A block EM algorithm for multivariate skew normal and skew t-mixture models // IEEE T. Neur. Net. Lear., 2018. Vol. 29. Iss. 11. P. 5581–5591.
6. *Cai T. T., Ma J., Zhang L.* CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality // Ann. Stat., 2019. Vol. 47. Iss. 3 P. 1234–1267.
7. *Liu C., Li H.-C., Fu K., Zhang F., Datcu M., Emery W. J.* A robust EM clustering algorithm for Gaussian mixture models // Pattern Recogn., 2019. Vol. 87. P. 269–284.
8. *Wu D., Ma J.* An effective EM algorithm for mixtures of Gaussian processes via the MCMC sampling and approximation // Neurocomputing, 2019. Vol. 331. P. 366–374.

9. *Ben Hassen H., Masmoudi K., Masmoudi A.* Model selection in biological networks using a graphical EM algorithm // Neurocomputing, 2019. Vol. 349. P. 271–280.
10. *Zeller C. B., Cabral C. R. B., Lachos V. H., Benites L.* Finite mixture of regression models for censored data based on scale mixtures of normal distributions // Adv. Data Anal. Classi., 2019. Vol. 13. Iss. 1. P. 89–116.
11. *Meneghini O., Luna C. J., Smith S. P., Lao L. L.* Modeling of transport phenomena in tokamak plasmas with neural networks // Phys. Plasmas, 2014. Vol. 21. Iss. 6. Art. No. 060702.
12. *Raja M. A. Z., Shah F. H., Tariq M., Ahmad I., Ahmad S. U.* Design of artificial neural network models optimized with sequential quadratic programming to study the dynamics of nonlinear Troesch's problem arising in plasma physics // Neural Comput. Appl., 2018. Vol. 29. Iss. 6. P. 83–109.
13. *Mesbah A., Graves D. B.* Machine learning for modeling, diagnostics, and control of non-equilibrium plasmas // J. Phys. D Appl. Phys., 2019. Vol. 52. Iss. 30. Art. No. 30LT02.
14. *Narita E., Honda M., Nakata M., Yoshida M., Hayashi N., Takenaga H.* Neural-network-based semi-empirical turbulent particle transport modelling founded on gyrokinetic analyses of JT-60U plasmas // Nucl. Fusion, 2019. Vol. 59. Iss. 10. Art. No. 106018.
15. *Parsons M. S.* Interpretation of machine-learning-based disruption models for plasma control // Plasma Phys. Contr. F., 2017. Vol. 59. Iss. 8. Art. No. 085001.
16. *Kates-Harbeck J., Svyatkovskiy A., Tang W.* Predicting disruptive instabilities in controlled fusion plasmas through deep learning // Nature, 2019. Vol. 568. Iss. 7753. P. 526.
17. *Gorshenin A. K., Kuzmin V. Yu.* Improved architecture of feedforward neural networks to increase accuracy of predictions for moments of finite normal mixtures // Pattern Recognition Image Analysis, 2019. Vol. 29. No. 1. P. 79–88.
18. *Горшенин А. К., Кузьмин В. Ю.* Применение рекуррентных нейронных сетей для прогнозирования моментов конечных нормальных смесей // Информатика и её применения, 2019. Т. 13. Вып. 3. С. 114–121.
19. *Gorshenin A., Kuzmin V.* A machine learning approach to the vector prediction of moments of finite normal mixtures // Adv. Intell. Syst., 2020. Vol. 1127. P. 307–314.
20. *Greff K., Srivastava R. K., Koutnik J., Steunebrink B. R., Schmidhuber J.* LSTM: A search Space Odyssey // IEEE T. Neur. Net. Lear., 2017. Vol. 28. Iss. 10. P. 2222–2232.
21. *Buduma N.* Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. — Sebastopol, CA, USA: O'Reilly Media, 2017. 298 p.
22. *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: A simple way to prevent neural networks from overfitting // J. Mach. Learn. Res., 2014. Vol. 15. P. 1929–1958.
23. *Peinl R., Holzschuher F., Pfitzer F.* Docker cluster management for the cloud — survey results and own solution // J. Grid Comput., 2016. Vol. 14. Iss. 2. P. 265–282.

Поступила в редакцию 15.01.20

ANALYSIS OF CONFIGURATIONS OF LSTM NETWORKS FOR MEDIUM-TERM VECTOR FORECASTING

A. K. Gorshenin[1,2] and V. Yu. Kuzmin[3]

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[2]Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991, Russian Federation
[3]"Wi2Geo LLC," 3-1 Mira Ave., Moscow 129090, Russian Federation

Abstract: The paper analyzes 36 configurations of LSTM (long short-term memory) architectures for forecasting with a duration up to 70 steps based on data whose size is 300–500 elements. For probabilistic approximation of observations, a model based on finite normal mixtures is used; therefore, the mathematical expectation, variance, skewness, and kurtosis of these mixtures are used as initial data for forecasting. The optimal configurations of neural networks were determined and the practical possibility of constructing high-quality medium-term forecasts with a limited training time was demonstrated. The results obtained are important for the development of a probabilistic-statistical approach to the description of the evolution of turbulent processes in a magnetically active high-temperature plasma.

Keywords: LSTM; forecasting; deep learning; high-performance computing; CUDA

DOI: 10.14357/19922264200102

Acknowledgments

The research is partially supported by the Russian Foundation for Basic Research (projects 19-07-00352 and 18-29-03100) and the RF Presidential scholarship program (project No. 538.2018.5). The calculations were performed using Hybrid high-performance computing cluster of FRC CSC RAS (http://ckp.frccsc.ru/).

References

1. Batanov, G. M., V. D. Borzosekov, A. K. Gorshenin, N. K. Kharchev, V. Yu. Korolev, and K. A. Sarskyan. 2019. Evolution of statistical properties of microturbulence during transient process under electron cyclotron resonance heating of the L-2M stellarator plasma. *Plasma Phys. Contr. F.* 61(7):075006.
2. Batanov, G. M., M. S. Berezhetskii, V. D. Borzosekov, *et al.* 2017. Reaction of turbulence at the edge and in the center of the plasma column to pulsed impurity injection caused by the sputtering of the wall coating in L-2M stellarator. *Plasma Phys. Rep.* 43(8):818–823.
3. Korolev, V. Yu. 2011. *Veroyatnostno-statisticheskie metody dekompozitsii volatil'nosti khaoticheskikh protsessov* [Probabilistic and statistical methods of decomposition of volatility of chaotic processes]. Moscow: Moscow University Publishing House. 512 p.
4. Gorshenin, A., V. Korolev, V. Kuzmin, and A. Zeifman. 2013. Coordinate-wise versions of the grid method for the analysis of intensities of non-stationary information flows by moving separation of mixtures of gamma-distribution. *27th European Conference on Modelling and Simulation Proceedings*. Dudweiler, Germany: Digitaldruck Pirrot GmbH. 565–568.
5. Lee, S. X., K. L. Leemaqz, and G. J. McLachlan. 2018. A block EM algorithm for multivariate skew normal and skew t-mixture models. *IEEE T. Neur. Net. Lear.* 29(11):5581–5591.
6. Cai, T. T., J. Ma, and L. Zhang. 2019. CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *Ann. Stat.* 47(3):1234–1267.
7. Liu, C., H.-C. Li, K. Fu, F. Zhang, M. Datcu, and W. J. Emery. 2019. A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recogn.* 87:269–284.
8. Wu, D., and J. Ma. 2019. An effective EM algorithm for mixtures of Gaussian processes via the MCMC sampling and approximation. *Neurocomputing* 331:366–374.
9. Ben Hassen, H., K. Masmoudi, and A. Masmoudi. 2019. Model selection in biological networks using a graphical EM algorithm. *Neurocomputing* 349:271–280.
10. Zeller, C. B., C. R. B. Cabral, V. H. Lachos, and L. Benites. 2019. Finite mixture of regression models for censored data based on scale mixtures of normal distributions. *Adv. Data Anal. Classi.* 13(1):89–116.
11. Meneghini, O., C. J. Luna, S. P. Smith, and L. L. Lao. 2014. Modeling of transport phenomena in tokamak plasmas with neural networks. *Phys. Plasmas* 21(6):060702.
12. Raja, M. A. Z., F. H. Shah, M. Tariq, I. Ahmad, and S. U. Ahmad. 2018. Design of artificial neural network models optimized with sequential quadratic programming to study the dynamics of nonlinear Troesch's problem arising in plasma physics. *Neural Comput. Appl.* 29(6):83–109.
13. Mesbah, A., and D. B. Graves. 2019. Machine learning for modeling, diagnostics, and control of non-equilibrium plasmas. *J. Phys. D Appl. Phys.* 52(30):30LT02.
14. Narita, E., M. Honda, M. Nakata, M. Yoshida, N. Hayashi, and H. Takenaga. 2019. Neural-network-based semi-empirical turbulent particle transport modelling founded on gyrokinetic analyses of JT-60U plasmas. *Nucl. Fusion* 59(10):106018.
15. Parsons, M. S. 2017. Interpretation of machine-learning-based disruption models for plasma control. *Plasma Phys. Contr. F.* 59(8):085001.
16. Kates-Harbeck, J., A. Svyatkovskiy, and W. Tang. 2019. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature* 568(7753):526.
17. Gorshenin, A. K., and V. Yu. Kuzmin. 2019. Improved architecture of feedforward neural networks to increase accuracy of predictions for moments of finite normal mixtures. *Pattern Recognition Image Analysis* 29(1):79–88.
18. Gorshenin, A. K., and V. Yu. Kuzmin. 2019. Primenenie rekurrentnykh neyronnykh setey dlya prognozirovaniya momentov konechnykh normal'nykh smesey [Application of recurrent neural networks to forecasting the moments of finite normal mixtures]. *Informatika i ee Primeneniya — Inform. Appl.* 13(3):114–121.
19. Gorshenin, A., and V. Kuzmin. 2020. A machine learning approach to the vector prediction of moments of finite normal mixtures. *Adv. Intell. Syst.* 1127:307–314.
20. Greff, K., R. K. Srivastava, J. Koutnik, B. R Steunebrink, and J. Schmidhuber. 2017. LSTM: A search Space Odyssey. *IEEE T. Neur. Net. Lear.* 28(10):2222–2232.
21. Buduma, N. 2017. *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms*. Sebastopol, CA: O'Reilly Media. 298 p.
22. Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15:1929–1958.
23. Peinl, R., F. Holzschuher, and F. Pfitzer. 2016. Docker cluster management for the cloud — survey results and own solution. *J. Grid Comput.* 14(2):265–282.

Received January 15, 2020

Contributors

Gorshenin Andrey K. (b. 1986) — Candidate of Science (PhD) in physics and mathematics, associate professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilova Str., Moscow 119333, Russian Federation; leading scientist, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991, Russian Federation; agorshenin@frccsc.ru

Kuzmin Victor Yu. (b. 1986) — Head of Development, "Wi2Geo LLC," 3-1 Mira Ave., Moscow 129090, Russian Federation; shadesilent@yandex.ru

ЧИСЛЕННЫЕ СХЕМЫ ФИЛЬТРАЦИИ МАРКОВСКИХ СКАЧКООБРАЗНЫХ ПРОЦЕССОВ ПО ДИСКРЕТИЗОВАННЫМ НАБЛЮДЕНИЯМ II: СЛУЧАЙ АДДИТИВНЫХ ШУМОВ*

А. В. Борисов[1]

Аннотация: Статья продолжает цикл исследований, начатых в работе «Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям I: характеристики точности». На основании полученных оценок точности приближенного решения задачи фильтрации состояний однородных марковских скачкообразных процессов (МСП) по косвенным непрерывным зашумленным наблюдениям предложены различные алгоритмы ее численной реализации и проведен их сравнительный анализ. При этом класс систем наблюдения ограничен системами с *аддитивными* винеровскими шумами: интенсивность шумов в наблюдениях является неслучайной постоянной. Для построения аппроксимаций использовались численные схемы «левых» и «средних» прямоугольников порядка точности 2 и 3 соответственно, а также квадратуры Гаусса порядка 5. В итоге были получены численные схемы порядка точности $1/2$, 1 и 2.

Ключевые слова: марковский скачкообразный процесс; оптимальная фильтрация; аддитивные шумы в наблюдениях; стохастическое дифференциальное уравнение; аналитическая и численная аппроксимация

DOI: 10.14357/19922264200103

1 Введение

Данная статья продолжает исследования, начатые в [1, 2]. Они посвящены численной реализации решения задачи фильтрации состояний однородных *марковских скачкообразных процессов* по непрерывным косвенным наблюдениям в присутствии винеровских шумов [3, 4]. Оценка оптимальной фильтрации является решением системы нелинейных *стохастических дифференциальных уравнений* (СДУ) типа Кушнера–Стратоновича, не имеющим аналитического представления. Искомая оценка оптимальной фильтрации является условным распределением состояния МСП, имеющим неотрицательные компоненты и удовлетворяющим условию нормировки. Традиционные методы решения систем СДУ, например метод Эйлера–Маруямы [5], не гарантируют сохранение этих свойств для своих реализаций и поэтому бесполезны для решения систем Кушнера–Стратоновича. Те численные методы, которые обеспечивают своим оценкам выполнение свойств неотрицательности компонентов и нормировки, названы в [1, 2] *устойчивыми*. Для их построения предлагается аппроксимировать исходную задачу фильтрации по непрерывным наблюдениям задачей оптимальной фильтрации по дискретизованным наблюдениям. В [1] эта задача решена: соответствующая рекуррентная формула представляет собой вариант абстрактной формулы Байеса. Она содержит бесконечные суммы, в которых слагаемые — интегралы. Предложена замена бесконечного ряда его конечным отрезком длины s, соответствующая учету в оценке возможности того, что состояние на временно́м интервале дискретизации наблюдений совершит не более s переходов; полученное таким образом приближение названо в работе *аналитической аппроксимацией порядка* s. В [1] доказано утверждение, определяющее потерю точности при переходе от оптимальной оценки к ее аналитической аппроксимации в зависимости от параметров системы наблюдения и выбранного порядка s.

Аналитические аппроксимации также не могут быть вычислены точно, так как содержат интегралы, не имеющие явного аналитического представления. В [2] предложено заменить их конечными интегральными суммами. Полученные аппроксимации названы *численными*. Они уже могут быть реализованы средствами вычислительной техники. Доказано утверждение, определяющее потерю точности при переходе от аналитической аппроксимации к ее численной реализации в зависимости от точности используемой схемы интегрирования. Результаты [1, 2] также позволяют оценить общую

*Работа выполнена при частичной поддержке РФФИ (проект 19-07-00187 А).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, aborisov@frccsc.ru

потерю точности при переходе от оценки оптимальной фильтрации по дискретизованным наблюдениям к ее численной аппроксимации, заданной схемой численного интегрирования.

Цель данной статьи — разработка и сравнительный анализ различных численных аппроксимаций оценок оптимальной фильтрации для класса систем с аддитивными шумами в наблюдениях. Это значит, что интенсивность шумов является неслучайной. Несмотря на то что теоретическое решение данной задачи в непрерывном времени известно давно и определяется классическим фильтром Вонэма [3], создание алгоритмического обеспечения для его эффективной устойчивой численной реализации до сих пор является актуальной задачей [6–8].

Статья организована следующим образом. Раздел 2 содержит постановку задачи оптимальной фильтрации по дискретизованным наблюдениям и сведения из [1, 2], необходимые для ее численного решения. В разд. 3 рассматривается подкласс систем наблюдения, включающий в себя лишь системы с аддитивными шумами в наблюдениях. Для них выбраны аналитические аппроксимации порядка $s = 1$ и 2, для реализации которых использованы схемы численного интегрирования «левых» и «средних» прямоугольников (для $s = 1$), а также комбинация квадратур Гаусса (для $s = 2$). Полученные в результате численные аппроксимации имеют порядок $1/2$, 1 и 2. Заключительные замечания представлены в разд. 4.

2 Постановка задачи фильтрации и необходимые сведения об аналитических и численных аппроксимациях

На триплете с фильтрацией $(\Omega^X \times \Omega^W, \mathcal{F}^X \times \mathcal{F}^W, \mathcal{P}^X \times \mathcal{P}^W, \{\mathcal{F}_t^X \times \mathcal{F}_t^W\}_{t\geqslant 0})$ рассматривается система наблюдения

$$X_t = X_0 + \int_0^t \Lambda^\top X_s\, ds + \mu_t\,;$$

$$Y_r = \int_{(r-1)h}^{rh} f X_s\, ds + \int_{(r-1)h}^{rh} \sum_{n=1}^{N} X_s^n g_n^{1/2}\, dW_s,\quad r \in \mathbb{N},$$

где

– $X_t \triangleq \operatorname{col}(X_t^1, \ldots, X_t^N) \in \mathbb{S}^N$ — ненаблюдаемое состояние системы — однородный МСП с конечным множеством состояний $\mathbb{S}^N \triangleq \{e_1, \ldots, e_N\}$ ($\mathbb{S}^N$ — множество единичных векторов евклидова пространства $\mathbb{R}^N$), матрицей интенсивностей переходов Λ и начальным распределением π;

– $\mu_t \triangleq \operatorname{col}(\mu_t^1, \ldots, \mu_t^N) \in \mathbb{R}^N$ — $\mathcal{F}_t^X$-согласованный мартингал;

– $\{Y_r\}_{r\in\mathbb{N}}:\ Y_r \triangleq \operatorname{col}(Y_r^1, \ldots, Y_r^M) \in \mathbb{R}^M$ — последовательность дискретизованных наблюдений, доступных в известные равноотстоящие моменты времени $\{rh\}_{r\in\mathbb{N}}$;

– $W_t \triangleq \operatorname{col}(W_t^1, \ldots, W_t^M) \in \mathbb{R}^M$ — $\mathcal{F}_t^W$-согласованный стандартный винеровский процесс;

– f — $(M \times N)$-мерная матрица;

– $\{g_n\}_{n=\overline{1,N}}$ — симметрические положительно определенные матрицы.

Задача оптимальной фильтрации состояния X по дискретизованным наблюдениям Y заключается в нахождении *условного математического ожидания* (УМО)

$$\widehat{X}_r \triangleq \mathbf{E}\left\{X_{t_r} | \mathcal{O}_r\right\},$$

где $\mathcal{O}_r \triangleq \sigma\{Y_\ell :\ 1 \leqslant \ell \leqslant r\}$ — σ-алгебра, порожденная наблюдениями, полученными до момента времени rh включительно; $\mathcal{O}_0 \triangleq \{\varnothing,\ \Omega\}$.

Пусть $N_r^X(\omega)$ — число скачков процесса X, произошедших на отрезке $[(r-1)h, rh]$; $\tau_r \triangleq \int_0^t X_s\, ds$ — случайный вектор времени пребывания процесса X в различных состояниях на отрезке $[(r-1)h, rh]$, а $\rho^{n,j,m}(\cdot)$ — распределение вектора $\tau_r X_{t_r}^j \mathbf{I}_{\{m\}}(N_r^X)$ при условии $X_{t_{r-1}} = e_k$. Это означает, что для любого $\mathcal{G} \in \mathcal{B}(\mathbb{R}^M)$ верно равенство

$$\mathbf{E}\left\{\mathbf{I}_{\mathcal{G}}(\tau_r) X_{t_r}^j \mathbf{I}_{\{m\}}\left(N_r^X\right) | X_{t_{r-1}} = e_k\right\} = \int_{\mathcal{G}} \rho^{k,j,m}(du).$$

Аналитическая аппроксимация $\overline{X}_r$ порядка s определяется рекурсивной схемой

$$\overline{X}_r = (\mathbf{1}\xi_r^\top \overline{X}_{r-1})^{-1} \xi_r^\top \overline{X}_{r-1},\ \ r \geqslant 1,\ \ \overline{X}_0 = \pi, \quad (1)$$

где $\mathbf{1} = \operatorname{row}(1, \ldots, 1)$ — вектор-строка подходящей размерности, а $\xi_q \triangleq \|\xi^{ij}(Y_q)\|_{i,j=\overline{1,N}}$ — $(N \times N)$-мерные случайные матрицы — функции наблюдений Y_q:

$$\xi^{ij}(y) \triangleq \sum_{m=0}^{s} \int_{\mathcal{D}} \mathcal{N}\left(y, fu, \sum_{p=1}^{N} u^p g_p\right) \rho^{i,j,m}(du). \quad (2)$$

В последней формуле $\mathcal{N}(y, m, K) \triangleq (2\pi)^{-M/2} \det{}^{-1/2} K \exp\left\{-(1/2)\|y - m)\|_{K^{-1}}^2\right\}$ — M-мерная плотность гауссовского распределения с математическим ожиданием m и невырожденной ковариационной матрицей K.

Из построения оценки $\overline{X}_r$ (1) следует, что она обладает свойством устойчивости.

Обычно значения интегралов $\xi^{ij}(y)$ приближенно вычисляются в виде интегральных сумм:

$$\xi^{ij}(y) \approx \psi^{ij}(y) \triangleq \sum_{\ell=1}^{L} \mathcal{N}\left(y, fw_\ell, \sum_{p=1}^{N} w_\ell^p g_p\right) \varrho_\ell^{ij};$$
$$\psi(y) \triangleq \|\psi^{ij}(y)\|_{i,j=\overline{1,N}},$$

определяемых набором пар $\{(w_\ell, \varrho_\ell^{ij})\}_{\ell=\overline{1,L}}$. Здесь $\varrho_\ell^{ij} \geqslant 0$, $\ell = \overline{1,L}$, — веса, $\sum_{\ell=1}^{L} \varrho_\ell^{ij} \leqslant 1$, а $w_\ell \triangleq \operatorname{col}(w_\ell^1, \ldots, w_\ell^N) \in \mathcal{D}$ — точки и $\psi_q \triangleq \triangleq \|\psi^{ij}(Y_q)\|_{ij=\overline{1,N}}$.

По построению ψ_q^{ij} являются положительными случайными величинами, поэтому *численная аппроксимация* $\widetilde{X}_r$ оценки $\overline{X}_r$, вычисляемая рекурсивно

$$\widetilde{X}_r \triangleq (\mathbf{1}\psi_r^\top \widetilde{X}_{r-1})^{-1}\psi_r^\top \widetilde{X}_{r-1}, \quad r \geqslant 1, \quad \widetilde{X}_0 = \pi, \quad (3)$$

также обладает свойством устойчивости.

Если для схемы численного интегрирования выполнено условие

$$\int_{\mathbb{R}^M} |\psi^{kj}(y) - \xi^{kj}(y)|dy < \delta,$$

то глобальный показатель точности ограничен сверху

$$\sup_{\pi\in\Pi} \mathbf{E}\left\{\|\widehat{X}_r - \widetilde{X}_r\|_1\right\} \leqslant$$
$$\leqslant 4\left[1 - \left(1 - \frac{(\overline{\lambda}h)^{s+1}}{(s+1)!}\right)^r\right] + 2r\delta, \quad (4)$$

где $\overline{\lambda} \triangleq \max_k |\lambda_{kk}|$. Для фиксированного момента времени T с уменьшением шага дискретизации $h \to 0$ неравенство (4) принимает асимптотический вид:

$$\sup_{\pi\in\Pi} \mathbf{E}\left\{\|\widetilde{X}_{T/h} - \widehat{X}_{T/h}\|_1\right\} \leqslant$$
$$\leqslant 2T\left(2\overline{\lambda}\frac{(\overline{\lambda}h)^s}{(s+1)!} + \frac{\delta}{h}\right). \quad (5)$$

Для эффективного выбора схемы численного интегрирования необходимо подбирать δ так, чтобы $\overline{\lambda}\delta/(\overline{\lambda}h)^{s+1} \to C \geqslant 0$ при $h \to 0$.

3 Приближенное решение задачи фильтрации по наблюдениям с аддитивными шумами

Ниже исследуются аппроксимации порядка $s = 1$ и 2. Для них с помощью обобщенной формулы полной вероятности легко получить вид интегралов (2), используемых в дальнейшем изложении:

$$\int_{\mathcal{D}} \mathcal{N}\left(Y_r, fu, \sum_{p=1}^{N} u^p g_p\right) \rho^{k,j,0}(du) =$$
$$= \delta_{kj} e^{\lambda_{kk}h} \mathcal{N}\left(Y_r, hf^k, hg_k\right); \quad (6)$$

$$\int_{\mathcal{D}} \mathcal{N}\left(Y_r, fu, \sum_{p=1}^{N} u^p g_p\right) \rho^{k,j,1}(du) =$$
$$= (1-\delta_{kj})\,\lambda_{kj} e^{\lambda_{jj}h} \int_0^h e^{(\lambda_{kk}-\lambda_{jj})u} \times$$
$$\times \mathcal{N}\left(Y_r, uf^k + (h-u)f^j, ug_k + (h-u)g_j\right) du; \quad (7)$$

$$\int_{\mathcal{D}} \mathcal{N}\left(Y_r, fu, \sum_{p=1}^{N} u^p g_p\right) \rho^{k,j,2}(du) =$$
$$= \sum_{\substack{i:i\neq k,\\ i\neq j}} \lambda_{ki}\lambda_{ij} e^{\lambda_{jj}h} \int_0^h \int_0^{h-u^k} e^{(\lambda_{kk}-\lambda_{ii})u^k + (\lambda_{ii}-\lambda_{jj})u^i} \times$$
$$\times \mathcal{N}\left(Y_r, u^k f^k + u^i f^i + (h-u^k-u^i)f^j,\right.$$
$$\left. u^k g_k + u^i g_i + (h-u^k-u^i)g_j\right) du^i du^k, \quad (8)$$

где f^j — j-й столбец матрицы f. За исключением (6) остальные интегралы не имеют явного аналитического представления, и для них будут рассмотрены различные варианты численной реализации.

3.1 Порядок $s = 1$, схема «левых» прямоугольников

Для системы наблюдения с чисто аддитивными шумами ($g_n \equiv g$) исследуем точность аппроксимации первого порядка ($s = 1$) при использовании численного интегрирования методом «левых» прямоугольников. В этом случае

$$\xi^{kj}(y) = \delta_{kj} e^{\lambda_{jj}h} \mathcal{N}(y, hf^j, hg) +$$
$$+ (1-\delta_{kj})\lambda_{kj} e^{\lambda_{jj}h} \int_0^h e^{(\lambda_{kk}-\lambda_{jj})u} \times$$
$$\times \mathcal{N}\left(y, uf^k + (h-u)f^j, hg\right) du; \quad (9)$$

$$\psi^{kj}(y) =$$
$$= e^{\lambda_{jj}h} \mathcal{N}(y, hf^j, hg)\left[\delta_{kj} + (1-\delta_{kj})\lambda_{kj}h\right]. \quad (10)$$

Ошибка численного интегрирования [9] определяется следующим образом:

$$\gamma^{kj}(y) = (1-\delta_{kj})\,\frac{\lambda_{kj}h^2}{2}\,e^{\lambda_{jj}h}\frac{d}{du}\big[e^{(\lambda_{kk}-\lambda_{jj})u}\times$$
$$\times\,\mathcal{N}\left(y, uf^k + (h-u)f^j, hg\right)\big]\Big|_{u=z} =$$
$$= (1-\delta_{kj})\,\frac{\lambda_{kj}h^2}{2}e^{\lambda_{jj}h}e^{(\lambda_{kk}-\lambda_{jj})z}\times$$
$$\times\,\mathcal{N}\left(y, zf^k + (h-z)f^j, hg\right)\zeta_0(y,z),$$

где $z = z(y) \in [0,h]$ — некоторый параметр, зависящий от y, и

$$\zeta_0(y,z) \triangleq \lambda_{kk} - \lambda_{jj} + \langle f^j, f^k - f^j\rangle_{g^{-1}} -$$
$$- \frac{z}{h}\|f^k - f^j\|^2_{g^{-1}} + \frac{1}{h}\langle y, f^k - f^j\rangle_{g^{-1}}. \quad (11)$$

Непосредственно интегрировать абсолютную величину γ^{kj} проблематично, так как

$$\int\limits_{\mathbb{R}^M} |\gamma^{kj}(y)|\,dy = \int\limits_{\mathbb{R}^M} |\gamma^{kj}(y, z^{kj}(y))|\,dy\,,$$

а зависимость $z^{kj}(y)$ в общем случае неизвестна. Поэтому предварительно оценим $|\gamma^{kj}|$ сверху. Прежде всего, можно непосредственно проверить истинность следующего неравенства:

$$\|y - z^{kj}f^k - (h-z^{kj})f^j\|^2_{(hg)^{-1}} \geqslant$$
$$\geqslant \|y\|^2_{(2hg)^{-1}} - \|z^{kj}f^k + (h-z^{kj})f^j\|^2_{(hg)^{-1}}. \quad (12)$$

Отсюда следует, что

$$h^2\frac{\lambda_{kj}}{2}\,e^{\lambda_{jj}h+(\lambda_{kk}-\lambda_{jj})z}\mathcal{N}(y, z^{kj}f^k + (h-z^{kj})f^j, hg) =$$
$$= h^2\frac{\lambda_{kj}}{2}\,e^{\lambda_{jj}h+(\lambda_{kk}-\lambda_{jj})z}(2\pi)^{-M/2}|hg|^{-1/2}\times$$
$$\times\exp\left(-\frac{1}{2}\|y - z^{kj}f^k - (h-z^{kj})f^j\|^2_{(hg)^{-1}}\right)\leqslant$$
$$\leqslant h^2\frac{\lambda_{kj}}{2}\,e^{\lambda_{jj}h+(\lambda_{kk}-\lambda_{jj})z}(2\pi)^{-M/2}|hg|^{-1/2}\times$$
$$\times\exp\left(-\frac{1}{2}\|y\|^2_{(2hg)^{-1}}\right)\times$$
$$\times\exp\left(\frac{1}{2}\|z^{kj}f^k + (h-z^{kj})f^j\|^2_{(hg)^{-1}}\right)\leqslant$$
$$\leqslant h^2C\mathcal{N}(y,0,2hg), \quad (13)$$

где

$$C \triangleq \max_{\substack{k,j:\\ k\neq j}} \lambda_{kj}e^{(h/2)\max_k\|f^k\|^2_{g^{-1}}}.$$

Тогда для интеграла от абсолютной величины ошибки верна следующая оценка сверху:

$$\int\limits_{\mathbb{R}^M}|\gamma^{kj}(y)|\,dy \leqslant Ch^2\int\limits_{\mathbb{R}^M}\mathcal{N}(y,0,2hg)|\zeta_0(y,z)|\,dy \leqslant$$
$$\leqslant Ch^2\int\limits_{\mathbb{R}^M}\left|\lambda_{kk} - \lambda_{jj} + \langle f^j, f^k - f^j\rangle_{g^{-1}} -\right.$$
$$\left. - \frac{z}{h}\|f^k - f^j\|^2_{g^{-1}}\right|\mathcal{N}(y,0,2hg)\,dy +$$
$$+ Ch^2\int\limits_{\mathbb{R}^M}\left|\frac{1}{h}\langle y, f^k - f^j\rangle_{g^{-1}}\right|\mathcal{N}(y,0,2hg)\,dy =$$
$$= Ch^2\int\limits_{\mathbb{R}^M}\left|\lambda_{kk} - \lambda_{jj} + \langle f^j, f^k - f^j\rangle_{g^{-1}} -\right.$$
$$\left. - \frac{z}{h}\|f^k - f^j\|^2_{g^{-1}}\right|\mathcal{N}(y,0,2hg)\,dy +$$
$$+ \sqrt{2}Ch^{3/2}\int\limits_{\mathbb{R}^M}\left|\frac{1}{h}\left\langle y, g^{-1/2}\left(f^k - f^j\right)\right\rangle_I\right|\times$$
$$\times\,\mathcal{N}(y,0,I)dy = K_1h^2 + K_2h^{3/2} \quad (14)$$

для некоторых неотрицательных констант K_1 и K_2. Окончательно получаем, что

$$\int\limits_{\mathbb{R}^M}|\gamma^{kj}(y)|\,dy = O\left(h^{3/2}\right),$$

и в этом случае согласно (5)

$$\sup_{\pi\in\Pi}\mathbf{E}\left\{\|\widetilde{X}_{T/h} - \widehat{X}_{T/h}\|_1\right\} \leqslant CTh^{1/2}$$

для некоторой константы $C > 0$ и достаточно малого шага h. Данный результат вполне согласуется с известным фактом, согласно которому порядок глобальной точности аппроксимации методом Эйлера–Маруямы решения СДУ, описывающего фильтр Вонэма, равен $1/2$. Однако, в отличие от последнего, алгоритм (3), (10) обеспечивает своим реализациям устойчивость.

3.2 Порядок $s = 1$, схема «средних» прямоугольников

Несмотря на широкое применение метода «левых» прямоугольников, его выбор при фиксированном порядке аналитической аппроксимации $s = 1$ не является оптимальным: выбранная схема численного интегрирования снижает общий порядок точности до $1/2$.

Вновь рассмотрим аналитическую аппроксимацию (9) порядка $s = 1$ и для ее приближения вместо (10) используем схему «средних» прямоугольников:

$$\psi^{kj}(y) = \delta_{kj} e^{\lambda_{jj} h} \mathcal{N}(y, hf^j, hg) +$$
$$+(1-\delta_{kj})\lambda_{kj} h e^{(\lambda_{jj}+\lambda_{kk})h/2} \mathcal{N}\left(y, \frac{h}{2}\left(f^j + f^k\right), hg\right).$$

При этом ошибка численного интегрирования определяется формулой:

$$\gamma^{kj}(y) = (1-\delta_{kj})\,\frac{\lambda_{kj} h^3}{24}\, e^{\lambda_{jj} h} \times$$
$$\times \frac{d^2}{du^2}\left[e^{(\lambda_{kk}-\lambda_{jj})u} \mathcal{N}(y, uf^k + (h-u)f^j, hg)\right]\Big|_{u=z} =$$
$$= (1-\delta_{kj})\,\frac{\lambda_{kj} h^2}{2}\, e^{\lambda_{jj} h} e^{(\lambda_{kk}-\lambda_{jj})z} \times$$
$$\times \mathcal{N}\left(y, zf^k + (h-z)f^j, hg\right)\left[\zeta_0^2(y,z) - \zeta_1\right],$$

где вновь $z = z(y) \in [0,h]$ — некоторый параметр, зависящий от y, а

$$\zeta_1 \triangleq \frac{\partial}{\partial z}\zeta_0(y,z) = \frac{1}{h}\|f^j - f^k\|^2_{g^{-1}}. \qquad (15)$$

Используя неравенства (12) и (13) и действуя аналогично выводу (14), можно показать, что для схемы «средних» прямоугольников верна оценка

$$\int\limits_{\mathbb{R}^M} |\gamma^{kj}(y)|\, dy = O(h^2),$$

и в этом случае согласно (5)

$$\sup_{\pi\in\Pi} \mathbf{E}\left\{\|\widetilde{X}_{T/h} - \widehat{X}_{T/h}\|_1\right\} \leqslant CTh$$

для некоторой константы $C > 0$ и достаточно малого шага h. Таким образом, заменой одной схемы численного интегрирования другой, без увеличения вычислительных затрат, возможно повысить общий порядок точности до первого.

Необходимо отметить, что дальнейшая фиксация порядка $s = 1$ и использование более точных методов численного интегрирования не приведет к значительному уточнению оценок, так как в суммарной погрешности основную роль будет играть ошибка аналитической аппроксимации, а не численного интегрирования. Для увеличения общей точности следует увеличить порядок аналитической аппроксимации до второго.

3.3 Порядок $s = 2$, квадратуры Гаусса

Формулы (6)–(8) для $s = 2$ позволяют получить вид функций ξ^{kj}:

$$\xi^{kj}(y) = \delta_{kj} e^{\lambda_{jj} h} \mathcal{N}(y, hf^j, hg) +$$
$$+ (1-\delta_{kj})\,\lambda_{kj} e^{\lambda_{jj} h} \int\limits_0^h e^{(\lambda_{kk}-\lambda_{jj})u} \times$$
$$\times \mathcal{N}\left(y, uf^k + (h-u)f^j, hg\right) du +$$
$$+ \sum_{\substack{i:i\neq k,\\ i\neq j}} \lambda_{ki}\lambda_{ij} e^{\lambda_{jj} h} \int\limits_0^h \int\limits_0^{h-u} e^{(\lambda_{kk}-\lambda_{ii})u + (\lambda_{ii}-\lambda_{jj})v} \times$$
$$\times \mathcal{N}\left(y, uf^k + vf^i + (h-u-v)f^j, hg\right) dvdu\,. \qquad (16)$$

В случае наблюдений с аддитивными шумами для эффективного использования повышения порядка аналитической аппроксимации s удается подобрать схемы вычисления интегралов без дополнительного дробления шага h. Для вычисления одномерного интеграла в (16) будем использовать двухточечную квадратуру Гаусса, для повторного интеграла — трехточечную:

$$\int\limits_0^h e^{(\lambda_{kk}-\lambda_{jj})u} \mathcal{N}(y, uf^k + (h-u)f^j, hg)\, du =$$
$$= \frac{h}{2}\left[e^{(\lambda_{kk}-\lambda_{jj})(\sqrt{3}-1)h/(2\sqrt{3})} \times\right.$$
$$\times \mathcal{N}\left(y, 2hf^j + \frac{h}{2\sqrt{3}}\left(f^k - f^j\right), hg\right) +$$
$$+ e^{(\lambda_{kk}-\lambda_{jj})(\sqrt{3}+1)h/(2\sqrt{3})} \times$$
$$\left.\times \mathcal{N}\left(y, 2hf^k + \frac{h}{2\sqrt{3}}\left(f^j - f^k\right), hg\right)\right] + e_1(y);$$

$$\int\limits_0^h \int\limits_0^{h-u} e^{(\lambda_{kk}-\lambda_{ii})u + (\lambda_{ii}-\lambda_{jj})v} \times$$
$$\times \mathcal{N}\left(y, uf^k + vf^i + (h-u-v)f^j, hg\right) dvdu =$$
$$= \frac{h^2}{6}\left[e^{(\lambda_{kk}-\lambda_{ii})h/6 + (\lambda_{ii}-\lambda_{jj})h/6} \times\right.$$
$$\times \mathcal{N}\left(y, \frac{h}{6}f^k + \frac{h}{6}f^i + \frac{2h}{3}f^j, hg\right) +$$
$$+ e^{(\lambda_{kk}-\lambda_{ii})2h/3 + (\lambda_{ii}-\lambda_{jj})h/6} \times$$
$$\times \mathcal{N}\left(y, \frac{h}{6}f^k + \frac{2h}{3}f^i + \frac{h}{6}f^j, hg\right) +$$
$$+ e^{(\lambda_{kk}-\lambda_{ii})h/6 + (\lambda_{ii}-\lambda_{jj})2h/3} \times$$
$$\left.\times \mathcal{N}\left(y, \frac{2h}{3}f^k + \frac{h}{6}f^i + \frac{h}{6}f^j, hg\right)\right] + e_2(y),$$

где $e_1(y)$ и $e_2(y)$ — ошибки интегрирования. Согласно [9], абсолютные значения ошибок ограничены следующим образом:

$$|e_1(y)| \leqslant h^5 K_3 \max_{u\in[0,h]} \left| \frac{\partial^4}{\partial u^4} \Big[e^{(\lambda_{kk}-\lambda_{jj})u} \times \right.$$
$$\left. \times \mathcal{N}(y, uf^k + (h-u)f^j, hg) \Big] \right|; \quad (17)$$

$$|e_2(y)| \leqslant$$
$$\leqslant h^5 K_4 \max_{\substack{(u,v)\in\mathcal{D},\\ k=1,2,3}} \left| \frac{\partial^3}{\partial u^k \partial v^{3-k}} \Big[e^{(\lambda_{kk}-\lambda_{ii})u + (\lambda_{ii}-\lambda_{jj})v} \times \right.$$
$$\left. \times \mathcal{N}\left(y, uf^k + vf^i + (h-u-v)f^j, hg\right) \Big] \right|,$$

где K_3 и K_4 — некоторые положительные константы.

Производная в (17) имеет вид:

$$\frac{\partial^4}{\partial u^4} \left[e^{(\lambda_{kk}-\lambda_{jj})u} \mathcal{N}(y, uf^k + (h-u)f^j, hg) \right] =$$
$$= e^{(\lambda_{kk}-\lambda_{jj})u} \mathcal{N}\left(y, uf^k + (h-u)f^j, hg\right) \times$$
$$\times \left(\zeta_0^4(y,z) + 6\zeta_0^2(y,z)\zeta_1 + 3\zeta_1^2\right),$$

где ζ_0 и ζ_1 определены формулами (11) и (15). Строя оценки сверху интеграла от абсолютного значения $e_1(y)$ подобно (14), можно получить неравенство

$$\int\limits_{\mathbb{R}^M} |e_1(y)|\, dy \leqslant K_5 h^3,$$

и аналогичная оценка для $|e_2(y)|$ имеет вид:

$$\int\limits_{\mathbb{R}^M} |e_2(y)|\, dy \leqslant K_6 h^3$$

для некоторых неотрицательных констант K_5 и K_6. В этом случае согласно (5)

$$\sup_{\pi\in\Pi} \mathbf{E} \left\{ \|\widetilde{X}_{T/h} - \widehat{X}_{T/h}\|_1 \right\} \leqslant CTh^2$$

для некоторой константы $C > 0$ и достаточно малого шага h. Таким образом, путем незначительного увеличения вычислительных затрат возможно повысить общий порядок точности аппроксимации до второго.

4 Заключение

На основании результатов [1, 2] в статье разработан ряд численных алгоритмов решения задачи фильтрации состояний однородных МСП по дискретизованным косвенным наблюдениям в присутствии аддитивных винеровских шумов. Эти алгоритмы имеют одинаковую структуру: оценки считаются рекуррентно в виде дроби, в которой числитель и знаменатель являются суммами некоторых интегралов. Отличие алгоритмов заключается лишь в числе слагаемых в этих суммах ($s = 1, 2$) и схеме, реализующей численное интегрирование. Примечательно, что в случае аддитивных шумов в наблюдениях интегралы таковы, что для их численного интегрирования с выбранной точностью не требуется дополнительного дробления интервала отрезка интегрирования. В итоге получены аппроксимации с точностью порядка $1/2$, 1 и 2.

Литература

1. *Борисов А.* Фильтрация состояний марковских скачкообразных процессов по дискретизованным наблюдениям // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 115–121. doi: 10.14357/19922264180316.
2. *Борисов А.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям I: характеристики точности // Информатика и её применения, 2019. Т. 13. Вып. 4. С. 68–75. doi: 10.14357/1992226419041.
3. *Wonham W.* Some applications of stochastic differential equations to optimal nonlinear filtering // SIAM J. Control Optim., 1964. Vol. 2. No. 3. P. 347–369. doi: 10.1137/0302028.
4. *Борисов А.* Фильтрация Вонэма по наблюдениям с мультипликативными шумами // Автоматика и телемеханика, 2018. № 1. С. 52–65.
5. *Kloeden P., Platen E.* Numerical solution of stochastic differential equations. — Berlin: Springer, 1992. 636 p.
6. *Yin G., Zhang Q., Liu Y.* Discrete-time approximation of Wonham filters // J. Control Theory Applications, 2004. No. 2. P. 1–10.
7. *Platen E., Rendek R.* Quasi-exact approximation of hidden Markov chain filters // Commun. Stoch. Anal., 2010. Vol. 4. Iss. 1. P. 129–142.
8. *Bäuerle N., Gilitschenski I., Hanebeck U.* Exact and approximate hidden Markov chain filters based on discrete observations // Statistics Risk Modeling, 2016. Vol. 32. Iss. 3-4. P. 159–176.
9. *Isaacson E., Keller H.* Analysis of numerical methods. — New York, NY, USA: Dover Publications, 1994. 541 p.

Поступила в редакцию 11.10.19

NUMERICAL SCHEMES OF MARKOV JUMP PROCESS FILTERING GIVEN DISCRETIZED OBSERVATIONS II: ADDITIVE NOISE CASE

A. V. Borisov

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The note is a sequel of investigations initialized in the article Borisov, A. 2019. Numerical schemes of Markov jump process filtering given discretized observations I: Accuracy characteristics. *Inform. Appl.* 13(4):68–75. The basis is the accuracy characteristics of the approximated solution of the filtering problem for the state of homogeneous Markov jump processes given the continuous indirect noisy observations. The paper presents a number of the algorithms of their numerical realization together with the comparative analysis. The class of observation systems under investigation is bounded by ones with additive observation noises. This presumes that the observation noise intensity is a nonrandom constant. To construct the approximation, the authors use the left and midpoint rectangle rule of the accuracy order 2 and 3, respectively, and the Gaussian quadrature of the order 5. Finally, the presented numerical schemes have the accuracy of the order 1/2, 1, and 2.

Keywords: Markov jump process; optimal filtering; additive and multiplicative observation noises; stochastic differential equation; analytical and numerical approximation

DOI: 10.14357/19922264200103

Acknowledgments

The work was supported in part by the Russian Foundation for Basic Research (Project No. 19-07-00187 A).

References

1. Borisov, A. 2018. Fil'tratsiya sostoyaniy markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam [Filtering of Markov jump processes by discretized observations]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):115–121. doi: 10.14357/19922264180316.
2. Borisov, A. 2019. Chislennye skhemy fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam I: kharakteristiki tochnosti [Numerical schemes of Markov jump process filtering given discretized observations I: Accuracy characteristics]. *Informatika i ee Primeneniya — Inform. Appl.* 13(4):68–75. doi: 10.14357/19922264190411.
3. Wonham, W. 1964. Some applications of stochastic differential equations to optimal nonlinear filtering. *SIAM J. Control Optim.* 2:347–369.
4. Borisov, A. 2018. Wonham filtering by observations with multiplicative noises. *Automat. Rem. Contr.* 79(1):39–50. doi: 10.1134/S0005117918010046.
5. Kloeden, P., and E. Platen. 1992. *Numerical solution of stochastic differential equations.* Berlin: Springer. 636 p.
6. Yin, G., Q. Zhang, and Y. Liu. 2004. Discrete-time approximation of Wonham filters. *J. Control Theory Applications* 2:1–10.
7. Platen, E., and R. Rendek. 2010. Quasi-exact approximation of hidden Markov chain filters. *Commun. Stoch. Anal.* 4(1):129–142.
8. Bäauerle, N., I. Gilitschenski, and U. Hanebeck. 2016. Exact and approximate hidden Markov chain filters based on discrete observations. *Statistics Risk Modeling* 32(3-4):159–176.
9. Isaacson, E., and H. Keller. 1994. *Analysis of numerical methods.* New York, NY: Dover Publications. 541 p.

Received October 11, 2019

Contributor

Borisov Andrey V. (b. 1965) — Doctor of Science in physics and mathematics, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; aborisov@frccsc.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2020. Т. 14. Вып. 1. С. 24–30

УПРАВЛЕНИЕ ВЫХОДОМ СТОХАСТИЧЕСКОЙ ДИФФЕРЕНЦИАЛЬНОЙ СИСТЕМЫ ПО КВАДРАТИЧНОМУ КРИТЕРИЮ. IV. АЛЬТЕРНАТИВНОЕ ЧИСЛЕННОЕ РЕШЕНИЕ*

А. В. Босов[1], А. И. Стефанович[2]

Аннотация: В исследовании задачи оптимального управления для диффузионного процесса Ито и линейного управляемого выхода с квадратичным критерием качества подводится промежуточный итог: для приближенного вычисления оптимального решения предлагается альтернативный классическому численному интегрированию метод на основе компьютерного моделирования. Метод позволяет применять статистическое оценивание для определения коэффициентов $\beta_t(y)$ и $\gamma_t(y)$ полученной ранее функции Беллмана $V_t(y,z) = \alpha_t z^2 + \beta_t(y)z + \gamma_t(y)$, определяющей оптимальное решение в исходной задаче оптимального стохастического управления. Реализуется метод на основании свойств линейных уравнений в частных производных параболического типа, описывающих $\beta_t(y)$ и $\gamma_t(y)$, — их эквивалентного описания в форме стохастических дифференциальных уравнений и теоретико-вероятностного представления решения, известного как уравнение А. Н. Колмогорова, или эквивалентной интегральной форме, известной как формула Фейнмана–Каца. Стохастические уравнения, соотношения для оптимального управления и ряд вспомогательных параметров объединяются в одну дифференциальную систему, для которой формулируется алгоритм имитационного моделирования решения, обеспечивающий необходимые выборки для статистического оценивания коэффициентов $\beta_t(y)$ и $\gamma_t(y)$. Поставленный ранее численный эксперимент дополняется расчетами, выполненными представленным альтернативным методом, и сравнительным анализом результатов.

Ключевые слова: стохастическое дифференциальное уравнение; оптимальное управление; функция Беллмана; линейные уравнения параболического типа; уравнение А. Н. Колмогорова; формула Фейнмана–Каца; имитационное компьютерное моделирование; метод Монте-Карло

DOI: 10.14357/19922264200104

1 Введение

В работах [1–3] представлено детальное исследование задачи управления линейным выходом стохастической дифференциальной системы по квадратичному критерию качества: решены уравнения динамического программирования и получено оптимальное управление, применены традиционные сеточные методы для приближенного вычисления коэффициентов найденной функции Беллмана, предложено альтернативное описание этих коэффициентов, базирующееся на уравнении А. Н. Колмогорова [4] или эквивалентной ему интегральной формуле Фейнмана–Каца [5]. Основание результатам обеспечило то обстоятельство, что функцию Беллмана в рассматриваемой задаче управления удалось привести к виду

$$V_t(y,z) = \alpha_t z^2 + \beta_t(y)z + \gamma_t(y)\,, \qquad (1)$$

в котором коэффициенты $\beta_t(y)$ и $\gamma_t(y)$ описываются линейными уравнениями в частных производных второго порядка параболического типа, коэффициент α_t описывается уравнением Риккати и легко вычисляется любым методом решения обыкновенных дифференциальных уравнений.

Формула Фейнмана–Каца или теоретико-вероятностное представление терминального условия решения задачи Коши для уравнения А. Н. Колмогорова в данной работе используются как основание для представления численного метода, альтернативного классическим сеточным методам решения параболических уравнений, а именно: дополнив имеющееся описание исходной стохастической дифференциальной системы уравнениями для оптимального управления, для коэффициента α_t и для нескольких вспомогательных переменных, приближенное вычисление коэффициентов $\beta_t(y)$

*Работа выполнена при частичной поддержке РФФИ (проект 19-07-00187-А).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ABosov@frccsc.ru

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, AStefanovich@frccsc.ru

и $\gamma_t(y)$ удается реализовать в форме статистического оценивания параметров случайных функций — решений общей дифференциальной системы. Таким образом, приближенное решение обеспечивается методом Монте-Карло, примененным к смоделированным решениям стохастических дифференциальных уравнений.

Структура статьи такова. Привлекаемые исходные положения из [1–3] кратко приведены в разд. 2, описание общей дифференциальной системы дано в разд. 3, процедура имитационного моделирования детально расписана в разд. 4. Анализу качества предложенной численной процедуры посвящен разд. 5, где продолжено исследование того же численного эксперимента, что был представлен в [2], а именно: проделанные в [2] расчеты дополнены аналогичными расчетами, в которых вместо сеточных методов для расчета $\beta_t(y)$ используется имитационное моделирование. Сравнительный анализ результатов представлен как в форме сравнения поверхностей $\beta_t(y)$, вычисленных разными способами, так и динамикой целевого функционала в исходной задаче управления.

2 Краткие сведения о задаче управления выходом

Рассматривается задача оптимизации квадратичного целевого функционала

$$J\left(U_0^T\right) = E\left\{\int_0^T \left(S_t\left(s_t y_t - g_t z_t - h_t u_t\right)^2 + G_t z_t^2 + H_t u_t^2\right)dt + S_T\left(s_T y_T - g_T z_T\right)^2 + G_T z_T^2\right\}, \quad U_0^T = \{u_t, 0 \le t \le T\}, \quad (2)$$

определяемого состоянием y_t стохастической дифференциальной системы

$$dy_t = A_t(y_t)\,dt + \Sigma_t(y_t)\,dv_t, \quad y_0 = Y, \qquad (3)$$

и выходом z_t, связанным с y_t линейно:

$$dz_t = a_t y_t\,dt + b_t z_t\,dt + c_t u_t\,dt + \sigma_t\,dw_t, \quad z_0 = Z,$$

где v_t и w_t — независимые стандартные винеровские процессы; Y и Z — случайные величины с конечным вторым моментом; u_t — управление. Ограничения на параметры модели, обеспечивающие корректность постановки задачи и существование решения, а также решение уравнений динамического программирования приведены в [1]. В частности, показано, что оптимальное управление

$$u_t^* = -\frac{1}{2}\left(S_t h_t^2 + H_t\right)^{-1}\left(c_t\left(2\alpha_t z_t + \beta_t(y_t)\right) + 2S_t\left(s_t y_t - g_t z_t\right)h_t\right), \quad (4)$$

доставляющее минимум функционалу $J(U_0^T)$, обеспечивается представлением функции Беллмана в виде (1), в частности имеет место равенство:

$$\min_{U_0^T} J\left(U_0^T\right) = J\left(\left(U^*\right)_0^T\right) = E\left\{V_0(Y,Z)\right\} = E\left\{\alpha_0 Z^2 + \beta_0(Y)Z + \gamma_0(Y)\right\}. \quad (5)$$

Коэффициент α_t описывается уравнением Риккати (приведено ниже), коэффициенты $\beta_t(y)$ и $\gamma_t(y)$ задаются следующими дифференциальными уравнениями в частных производных:

$$\frac{\partial\beta_t(y)}{\partial t} + A_t(y)\frac{\partial\beta_t(y)}{\partial y} + \frac{1}{2}\Sigma_t^2(y)\frac{\partial^2\beta_t(y)}{\partial^2 y} + M_t y + N_t\beta_t(y) = 0, \quad \beta_T(y) = -2S_T s_T g_T y;$$

$$\frac{\partial\gamma_t(y)}{\partial t} + A_t(y)\frac{\partial\gamma_t(y)}{\partial y} + \frac{1}{2}\Sigma_t^2(y)\frac{\partial^2\gamma_t(y)}{\partial y^2} + L_t(y) = 0, \quad \gamma_T(y) = S_T s_T^2 y^2, \quad (6)$$

где

$$M_t = 2\left(\alpha_t\left(a_t + \left(S_t h_t^2 + H_t\right)^{-1} c_t S_t h_t s_t\right) - \left(S_t - \left(S_t h_t^2 + H_t\right)^{-1} S_t^2 h_t^2\right) s_t g_t\right);$$

$$N_t = b_t - \left(S_t h_t^2 + H_t\right)^{-1} c_t S_t h_t g_t - \left(S_t h_t^2 + H_t\right)^{-1} c_t^2 \alpha_t;$$

$$L_t(y) = \beta_t(y)\left(a_t + \left(S_t h_t^2 + H_t\right)^{-1} c_t S_t h_t s_t\right) y + \left(S_t - \left(S_t h_t^2 + H_t\right)^{-1} S_t^2 h_t^2\right) s_t^2 y^2 - \frac{1}{4}\left(S_t h_t^2 + H_t\right)^{-1} c_t^2 \beta_t^2(y).$$

Исследование уравнений (4) и (5), выполненное в [3], позволило записать систему стохастических дифференциальных уравнений А. Н. Колмогорова, решение которой связано теоретико-вероятностным соотношением с искомыми коэффициентами $\beta_t(y)$ и $\gamma_t(y)$. Соответствующее свойство в окончательном виде представлено следующими соотношениями:

$$\beta_t(y) = E\left\{-2S_T s_T g_T \exp\left\{\int\limits_t^T N_\tau\, d\tau\right\} y_T + \right.$$
$$\left. + \int\limits_t^T \exp\left\{\int\limits_t^\tau N_s\, ds\right\} M_\tau y_\tau\, d\tau \,|\mathcal{F}_t^y\right\}; \quad (7)$$

$$\gamma_t(y) = E\left\{S_T s_T^2 y_T^2 + \int\limits_t^T L_\tau(y_\tau)\, d\tau \,|\mathcal{F}_t^y\right\}, \quad (8)$$

где $\mathcal{F}_t^y$ — σ-алгебра, порожденная значениями y_τ до момента t включительно. Равенство (7) представляет собой частный случай формулы Фейнмана–Каца [5].

3 Общая дифференциальная система

Соотношения (7)–(8) в полной мере выражают теоретико-вероятностную связь между решениями линейных уравнений в частных производных второго порядка параболического типа, т. е. уравнений для искомых коэффициентов $\beta_t(y), \gamma_t(y)$ и стохастическим дифференциальным уравнением (3) для состояния оптимизируемой динамической системы. Но форма их записи не слишком удобна для численной реализации, поэтому искомое решение ниже представлено в эквивалентной дифференциальной форме:

$$dy_t = A_t(y_t)\, dt + \Sigma_t(y_t)\, dv_t\,, \;\; y_0 = Y\,; \quad (9)$$

$$\frac{\partial \alpha_t}{\partial t} + 2\alpha_t\left(b_t - \left(S_t h_t^2 + H_t\right)^{-1} c_t S_t h_t g_t\right) +$$
$$+ \left(S_t - \left(S_t h_t^2 + H_t\right)^{-1} S_t^2 h_t^2\right) g_t^2 + G_t -$$
$$- \left(S_t h_t^2 + H_t\right)^{-1} c_t^2 \alpha_t^2 = 0\,, \;\; \alpha_T = S_T g_t^2 + G_T\,; \quad (10)$$

$$\left.\begin{aligned} dy_t^{(1)} &= M_t I_t^{-1} y_t\, dt\,, \;\; y_0^{(1)} = 0\,; \\ dy_t^{(2)} &= L_t(y_t)\, dt\,, \;\; y_0^{(2)} = 0\,; \\ di_t &= N_t\, dt\,, \;\; i_0 = 0\,, \; I_t = \exp\{-i_t\}\,. \end{aligned}\right\} \quad (11)$$

Здесь уравнение (9) такое же, как и уравнение (3), но формально оно должно пониматься в смысле другого вероятностного пространства, так как описывает не состояние исходной стохастической системы, а некоторый инструментальный процесс, построенный для уравнения А. Н. Колмогорова. То обстоятельство, что использованы одинаковые обозначения y_t, объясняется желанием подчеркнуть стохастическую эквивалентность (9) и (3) и не вводить без необходимости дополнительные обозначения. Уравнение (10) — уравнение Риккати для коэффициента α_t. Соотношениями (11) задаются вспомогательные переменные, используемые для более компактной записи результата.

Предложенная дифференциальная система (9)–(11) позволяет записать следующие выражения для коэффициентов $\beta_t(y)$ и $\gamma_t(y)$:

$$\left.\begin{aligned} \beta_t(y) &= I_t E\left\{-2S_T s_T g_T I_T^{-1} y_T(t,y) + \right. \\ &\qquad \left. + y_T^{(1)}(t,y) - y_t^{(1)}(t,y)\right\}; \\ \gamma_t(y) &= E\left\{S_T s_T^2 y_T^2(t,y) + y_T^{(2)}(t,y) - \right. \\ &\qquad \left. - y_t^{(2)}(t,y)\right\}, \end{aligned}\right\} \quad (12)$$

где через $y_T(t,y)$ обозначена терминальная точка решения уравнения (9) при условии $y_t = y$, через $y_t^{(1)}(t,y)$ и $y_t^{(2)}(t,y)$ обозначены решения (11), удовлетворяющие тому же условию $y_t = y$, в том числе и терминальные значения $y_T^{(1)}(t,y)$ и $y_T^{(2)}(t,y)$.

Заметим, что совокупность выписанных уравнений очевидным образом декомпозируется на три самостоятельные части. Во-первых, это обыкновенные дифференциальные уравнения для α_t и i_t, которые решаются первыми любым численным методом. Вторая часть — это стохастическое дифференциальное уравнение (6), решение которого требует соответствующего подхода. Третья часть — это уравнения для вспомогательных переменных, которые представляют собой среднеквадратические интегралы от решения (9).

4 Приближенное решение методом имитационного моделирования

Представление уравнений для искомых коэффициентов с помощью системы (9)–(11) преследует очевидную цель — использовать для ее решения метод Монте-Карло. Для этого под решением уравнения (9) будем понимать смоделированный пучок $\{(y_t)_l\}_{l=1}^L$, содержащий L траекторий процесса y_t. Тогда вместо точных расчетов по формулам (10) можно получить приближенные решения, используя статистические оценки $\overline{E}\{X\} = (1/L)\sum_{l=1}^L X_l$, построенные по выборке $\{X_l\}_{l=1}^L$, порождаемой пучком $\{(y_t)_l\}_{l=1}^L$, вместо математических ожиданий $E\{X\}$.

Для формального описания алгоритма введем обозначения:

– разбиение (для простоты равномерное) во временно́й области $0 \le t \le T$: $\{t_n\}_{n=0}^N$, $t_0 = 0 < t_1 < \cdots < t_{N-1} < t_N = T$, $\delta = t_n - t_{n-1} = T/N \ll 1$;

– разбиение (равномерное и одинаковое для всех t) в области значений y_t: $\{y_m\}_{m=0}^M$, $-\infty < y_0 < y_1 < \cdots < y_{M-1} < y_M < +\infty$, $\varepsilon = y_m - y_{m-1} = y_M/M \ll 1$;

– оператор «ближайшей точки» $y_{m^*} = \mathrm{near}\,(y)$, где y_{m^*} — та точка разбиения $\{y_m\}_{m=0}^M$, для которой $m^* = \arg\min\limits_{0\le m\le M} |y - y_m|$.

При росте N будет уменьшаться δ. При росте M дополнительно потребуем, чтобы расширялся интервал $[y_0, y_M]$ и уменьшалось ε так, чтобы при $M \to +\infty$ выполнялось $y_0 \to -\infty$, $y_M \to +\infty$, $\varepsilon \to 0$. Ближайшая точка в рассматриваемом скалярном случае, естественно, определяется элементарно — сравнением двух расстояний от заданной точки до границ интервала разбиения, в который она попала. Однако сохраним этот оператор как заготовку для обобщения на многомерный случай, что добавит содержательности этому определению. Вопросы сходимости выходят за рамки данного исследования, точнее в рассмотрении этих вопросов не видится какого-либо принципиального содержания, поскольку относятся они не столько к предлагаемому алгоритму, сколько к используемым в нем классическим методам. Итак, алгоритм.

Шаг 1. Используя разбиение $\{t_n\}_{n=0}^N$ решить численно обыкновенные дифференциальные уравнения (10) и (11) для α_t и i_t. Упрощая обозначения, для полученных приближенных решений будем использовать те же, что и для точных: α_{t_n} и i_{t_n}, $n = 0, \ldots, N$. Аналогично поступим далее с обозначениями для приближенных решений стохастических уравнений.

Шаг 2. Используя разбиения $\{t_n\}_{n=0}^N$ и $\{y_m\}_{m=0}^M$, смоделировать приближенно пучок L траекторий $\{(y_{t_n})_l, (y_{t_n}^{(1)})_l, (y_{t_n}^{(2)})_l\}_{l=1}^L$ процессов $y_t, y_t^{(1)}, y_t^{(2)}$ для всех $t \in \{t_n\}_{n=0}^N$, так чтобы $(y_{t_n})_l \in \{y_m\}_{m=0}^M$, т. е. значения всех смоделированных приближенно траекторий y_t проходили через узлы выбранного разбиения. Для этого выполнить следующие шаги.

Шаг 2.1. Записать приближенные уравнения для $y_t, y_t^{(1)}, y_t^{(2)}$, заменив в точных уравнениях (9), (11) коэффициенты α_t и i_t их приближенными значениями, полученными на шаге 1.

Шаг 2.2. Записать разностный аналог стохастической дифференциальной системы (9), используя любую (предпочтительно явную) схему численного интегрирования [6], например Эйлера, Мильштейна, Тейлора, Рунге–Кутты.

Шаг 2.3. Используя на шаге n, т. е. для $t = t_n$, решение $\{(y_{t_{n-1}})_l, (y_{t_{n-1}}^{(1)})_l, (y_{t_{n-1}}^{(2)})_l\}_{l=1}^L$, полученное на шаге $n-1$ и удовлетворяющее привязке к разбиению $\{y_m\}_{m=0}^M$, смоделировать (согласно выбранной схеме численного решения стохастического уравнения) промежуточное решение для $\{(y_{t_n})_l, (y_{t_n}^{(1)})_l, (y_{t_n}^{(2)})_l\}_{l=1}^L$ и заменить точку $(y_{t_n})_l$ промежуточного решения на «ближайшую точку»: $\mathrm{near}((y_{t_n})_l)$. Отличие для $t = 0$ состоит в том, что вместо разностной схемы моделируется выборка для случайной величины Y, задающей начальное условие в (3). Шаги 2.1–2.3 повторяются вплоть до $t = T$ или $n = N$.

Шаг 3. Вычислить приближенные значения $\{\beta_{t_n}(y_m), \gamma_{t_n}(y_m)\}_{m=0,\ldots,M}^{n=0,\ldots,N}$, для чего выполнить следующие шаги.

Шаг 3.1. Для каждого $t = t_n$ и каждого $y = y_m$ из имеющегося пучка траекторий выбрать те, что отвечают условию $y_t = y$, и сформировать выборки $\{(y_T(t,y))_l, (y_T^{(1)}(t,y))_l, (y_T^{(2)}(t,y))_l\}_{l=1}^{L_{n,m}}$ некоторой длины $L_{n,m}$.

Шаг 3.2. Вычислить приближенные решения, используя в соответствующих соотношениях в (12) вместо математического ожидания $E\{\cdot\}$ аппроксимацию $\overline{E}\{\cdot\}$, суммируя в ней именно тот пучок, что выбран на предыдущем шаге для $t = t_n$ и $y = y_m$ Повторять шаг 3.1 для следующих значений t, y.

Шаг 4. Дополнив имеющиеся смоделированные данные выборкой для начального значения $z_0 = Z$ (с использованием уже имеющейся выборки для Y и вычисленных на предыдущем шаге $\beta_0(Y_l)$ и $\gamma_0(Y_l)$), вычислить, используя (5), оценку Монте-Карло для $J((U^*)_0^T)$.

Шаг 5. Смоделировать приближенно пучок L траекторий $\{(z_{t_n})_l, (u_{t_n}^*)_l\}_{l=1}^L$ процессов z_t, u_t^* для всех $t \in \{t_n\}_{n=0}^N$.

Выбор на шаге 1 метода приближенного интегрирования не имеет принципиального значения для алгоритма в целом. Это может быть и явный метод Эйлера, и метод Рунге–Кутты 4-го порядка, как и любой неявный метод численного интегрирования обыкновенного дифференциального уравнения. От этого метода требуется быть устойчивым в имеющейся постановке и обеспечивать сходимость приближенного решения к точному при расширении разбиения $\{t_n\}_{n=0}^N$. Выбор на шаге 2.2 конкретного метода для решения уравнения Ито [6] также не принципиален, естественно, при условии,

что будет получено «хорошее» приближение искомого сильного решения. Соответственно, условия и свойства сходимости этого метода принимаются такими, как есть. Следует обратить внимание, что на шаге 2 привязка к разбиению выполняется только для переменной, соответствующей y_t, что, как уже несколько раз подчеркивалось, вытекает из вспомогательного характера $y_t^{(1)}, y_t^{(2)}$. Точность, обеспечиваемая шагом 3, конечно, во многом будет определяться как качеством выбора разбиения фазовой переменной, так и гладкостью всех элементов системы. Если система окажется «удачной», то исходно смоделированного пучка траекторий окажется достаточно для выполнения всех осреднений на всем интервале управления. Если нет, то нехватку, возможно, компенсирует то обстоятельство, что вероятность траекторий, проходящих через «неудачные» точки разбиения $\{y_m\}_{m=0}^{M}$, мала, а значит, мал их вклад в целевую функцию (2). В любом случае следует исходить из того, что, наращивая объем моделируемой информации, всегда можно добиться любой точности оценивания, оправдывая это действием закона больших чисел [7]. Наконец, отметим, что последние два шага выполняются в том случае, когда требуется провести какой-либо анализ оптимального управления (определить достижимое качество управления, сравнить с другими управлениями, визуально охарактеризовать траектории выхода и/или управления и т. п.).

5 Численный пример

Для применения предложенного алгоритма были продолжены эксперименты с модельным примером, детально проанализированным в [2], а именно: использована простая модель для показателя RTT (Round-Trip Time) сетевого протокола TCP (Transmission Control Protocol), предложенная в [8] и конкретизированная следующим уравнением:

$$dy_t = (1 - 0{,}1y_t)\,dt + 0{,}5\sqrt{y_t}\,dv_t\,,$$
$$y_0 = Y \sim N(15{,}9)\,. \quad (13)$$

Здесь $N(M, D)$ — нормальное распределение со средним M и дисперсией D.

Равномерное разбиение временно́й области $0 \le t \le T$ для $T = 5$ выполнено для разбиения в области $[-10, 40]$ значений y_t — для $\varepsilon = 0{,}01$, $M = 5000$. Итоговый расчет в отношении $\beta_t(y)$ в [2] проведен для неявной сеточной схемы и граничного условия в задаче Дирихле: $\beta_t(y) = 0$, $y = -10; 40$. Соответствующее приближенное решение обозначается $\beta_t^0(y)$, порождаемое им управление — u_t^0.

Альтернативное численное решение $\beta_t^*(y)$ (соответствующее управление u_t^*), реализованное представленным выше алгоритмом, использует ту же сеточную структуру и выборку из $L = N\,\mathrm{M} \cdot 1000 = 25 \cdot 10^9$ траекторий y_t, аппроксимирующих (13) согласно явному методу Эйлера.

Результаты иллюстрируются, во-первых, рис. 1, на котором представлена поверхность $\beta_t^0(y)$, и рис. 2, на котором изображена поверхность $\beta_t^*(y)$, а также табл. 1, в которой демонстрируются отклонения $|\beta_t^0(y) - \beta_t^*(y)|$, и табл. 2, в которой показана динамика целевых функционалов $J((U^0)_0^T)$ и $J((U^*)_0^T)$.

В отношении показанных на рис. 1 и 2 поверхностей можно сделать вывод, что они аппроксимируют одну и ту же функцию, причем первая аппроксимация обеспечивает некоторую визуальную гладкость, вторая отличается некоторым «дрожанием», характерным для оценок Монте-Карло.

В табл. 1 каждое отклонение дополнено в скобках относительным отклонением по отношению к поверхности $\beta_t^*(y)$. Как видно, совпадение поверхностей с точностью 1%–4% имеет место в большинстве точек внутри интервала $[-10, 40]$. Исключение составляют точки левой границы $y =$

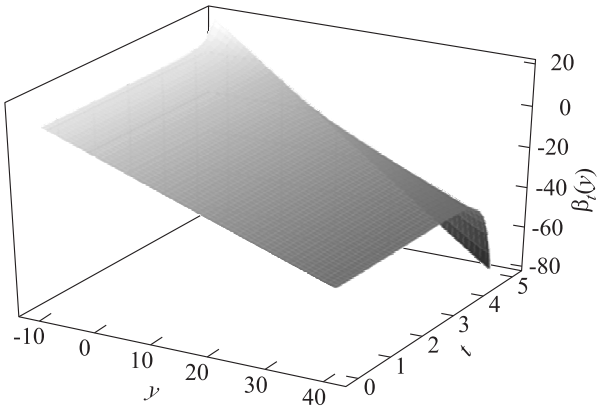


Рис. 1 Поверхность $\beta_t^0(y)$

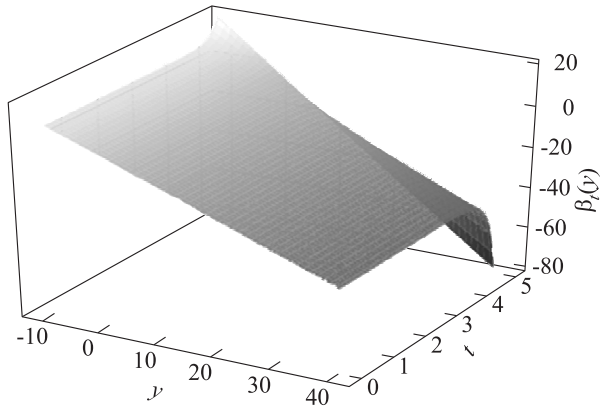


Рис. 2 Поверхность $\beta_t^*(y)$

Таблица 1 Отклонения $|\beta_t^0(y) - \beta_t^*(y)|$

t	y					
	$-10{,}0$	0,0	10,0	20,0	30,0	40,0
0	0,57 (6,2%)	0,00 (0,7%)	0,16 (1,5%)	0,17 (0,8%)	0,25 (0,8%)	1,41 (3,5%)
1	1,07 (11,7%)	0,00 (0,3%)	0,06 (0,5%)	0,13 (0,6%)	0,57 (1,8%)	1,33 (3,3%)
2	0,47 (5,1%)	0,01 (1,3%)	0,30 (2,8%)	0,39 (1,9%)	0,14 (0,4%)	0,56 (1,4%)
3	0,64 (7,0%)	0,04 (7,5%)	0,25 (2,4%)	0,75 (3,6%)	0,16 (0,5%)	0,98 (2,4%)
4	0,43 (4,6%)	0,07 (12,4%)	0,18 (1,6%)	0,27 (1,2%)	0,31 (1,0%)	2,51 (6,1%)
5	0,00 (0,0%)	0,00 (0,0%)	0,00 (0,0%)	0,00 (0,0%)	0,00 (0,0%)	0,00 (0,0%)

Таблица 2 Динамика целевых функционалов $J((U^0)_0^T)$ и $J((U^*)_0^T)$

t	$J((U^0)_0^t)$	$J((U^*)_0^t)$	$\lvert J((U^0)_0^t) - J((U^*)_0^t)\rvert$
1	134,17	132,10	2,08 (1,5%)
2	266,11	258,87	7,23 (2,7%)
3	394,32	381,55	12,77 (3,2%)
4	517,93	497,49	20,44 (3,9%)
5	742,21	704,48	37,73 (5,1%)

$= -10{,}0$, относительно которых можно предполагать неудачу в выборе граничного условия. Также несколько неудачных значений присутствуют в точке $y = 0{,}0$, где сечение поверхности $\beta_t^0(y)$ принимает значения, близкие к нулю. Заметим, что, хотя поверхность $\beta_t^0(y)$ интерпретируется как некоторый эталон, на самом деле неизвестно, какое именно решение ближе к $\beta_t(y)$. Косвенно оценить это помогают результаты применения соответствующих управлений $(U^0)_0^T = \{u_t^0, 0 \le t \le 5\}$ и $(U^*)_0^T = \{u_t^*, 0 \le t \le 5\}$, иллюстрируемые в динамике табл. 2.

Приведенные результаты не только подтвердили высокую точность расчета, обеспечиваемую предложенным алгоритмом, но и показали превосходство, обеспечиваемое управлением u_t^*, что косвенно свидетельствует о более высокой точности аппроксимации $\beta_t(y)$, обеспечиваемой представленным алгоритмом.

6 Заключение

В статье подведен промежуточный итог исследованию задачи оптимизации линейного выхода нелинейной дифференциальной системы по квадратичному критерию, выполненному в [1–3]. В рассматриваемой задаче оптимизации имеется оптимальное решение, полученное методом динамического программирования, приближенное решение на основе сеточных методов решения дифференциальных уравнений и альтернативное приближенное решение, базирующееся на теоретико-вероятностной связи решений параболических уравнений в частных производных и стохастических дифференциальных систем. Завершающим в полном объеме исследование вопросом могло бы стать изучение рассматриваемой постановки для случая неполной информации о состоянии системы y_t, т. е. предположения, что состояние системы доступно посредством косвенных наблюдений, обеспечиваемых выходом z_t. Данный вопрос предполагается рассмотреть в последующих публикациях.

Литература

1. *Босов А. В., Стефанович А. И.* Управление выходом стохастической дифференциальной системы по квадратичному критерию. I. Оптимальное решение методом динамического программирования // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 99–106.
2. *Босов А. В., Стефанович А. И.* Управление выходом стохастической дифференциальной системы по квадратичному критерию. II. Численное решение уравнений динамического программирования // Информатика и её применения, 2019. Т. 13. Вып. 1. С. 9–15.
3. *Босов А. В., Стефанович А. И.* Управление выходом стохастической дифференциальной системы по квадратичному критерию. III. Анализ свойств оптимального управления // Информатика и её применения, 2019. Т. 13. Вып. 3. С. 41–49.
4. *Гихман И. И., Скороход А. В.* Теория случайных процессов. — М.: Наука, 1975. Т. III. 496 с.
5. *Øksendal B.* Stochastic differential equations. An introduction with applications. — New York, NY, USA: Springer-Verlag, 2003. 379 p.
6. *Kloden P. E., Platen E.* Numerical solution of stochastic differential equations. — Berlin–Heidelberg: Springer-Verlag, 1992. 636 p.
7. *Ширяев А. Н.* Вероятность. — 2-е изд. — М.: Наука, 1989. 640 с.
8. *Bohacek S., Rozovskii B.* A diffusion model of roundtrip time // Comput. Stat. Data An., 2004. Vol. 45. Iss. 1. P. 25–50.

Поступила в редакцию 28.08.19

STOCHASTIC DIFFERENTIAL SYSTEM OUTPUT CONTROL BY THE QUADRATIC CRITERION. IV. ALTERNATIVE NUMERICAL DECISION

A. V. Bosov and A. I. Stefanovich

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: In the study of the optimal control problem for the Ito diffusion process and the controlled linear output with a quadratic quality criterion, an intermediate result is resumed: for approximate calculation of the optimal solution, an alternative to classical numerical integration method based on computer simulation is proposed. The method allows applying statistical estimation to determine the coefficients $\beta_t(y)$ and $\gamma_t(y)$ of the previously obtained Bellman function $V_t(y,z) = \alpha_t z^2 + \beta_t(y)z + \gamma_t(y)$, determining the optimal solution in the original problem of optimal stochastic control. The method is implemented on the basis of the properties of linear parabolic partial differential equations describing $\beta_t(y)$ and $\gamma_t(y)$ — their equivalent description in the form of stochastic differential equations and a theoretical-probability representation of the solution, known as A. N. Kolmogorov equation, or an equivalent integral form known as the Feynman–Katz formula. Stochastic equations, relations for optimal control and for auxiliary parameters are combined into one differential system, for which an algorithm for simulating a solution is stated. The algorithm provides the necessary samples for statistical estimation of the coefficients $\beta_t(y)$ and $\gamma_t(y)$. The previously performed numerical experiment is supplemented by calculations presented by an alternative method and a comparative analysis of the results.

Keywords: stochastic differential equation; optimal control; Bellman function; linear differential equations of parabolic type; Kolmogorov equation; Feynman–Katz formula; computer simulations; Monte-Carlo method

DOI: 10.14357/19922264200104

Acknowledgments

This work was partially supported by the Russian Foundation for Basic Research (grant 19-07-00187-A).

References

1. Bosov, A. V., and A. I. Stefanovich. 2018. Upravlenie vykhodom stokhasticheskoy differentsial'noy sistemy po kvadratichnomu kriteriyu. I. Optimal'noe reshenie metodom dinamicheskogo programmirovaniya [Stochastic differential system output control by the quadratic criterion. I. Dynamic programming optimal solution]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):99–106.
2. Bosov, A. V., and A. I. Stefanovich. 2019. Upravlenie vykhodom stokhasticheskoy differentsial'noy sistemy po kvadratichnomu kriteriyu. II. Chislennoe reshenie uravneniy dinamicheskogo programmirovaniya [Stochastic differential system output control by the quadratic criterion. II. Dynamic programming equations numerical solution]. *Informatika i ee Primeneniya — Inform. Appl.* 13(1):9–15.
3. Bosov, A. V., and A. I. Stefanovich. 2019. Upravlenie vykhodom stokhasticheskoy differentsial'noy sistemy po kvadratichnomu kriteriyu. III. Analiz svoystv optimal'nogo upravleniya [Stochastic differential system output control by the quadratic criterion. III. Optimal control properties analysis]. *Informatika i ee Primeneniya — Inform. Appl.* 13(3):41–49.
4. Gihman, I.I., and A. V. Skorohod. 2012. *The theory of stochastic processes*. New York, NY: Springer-Verlag. Vol. III. 388 p.
5. Øksendal, B. 2003. *Stochastic differential equations. An introduction with applications*. New York, NY: Springer-Verlag. 379 p.
6. Kloden, P. E., and E. Platen. 1992. *Numerical solution of stochastic differential equations*. Berlin–Heidelberg: Springer-Verlag. 636 p.
7. Shiryaev, A. N. 1996. *Probability*. New York, NY: Springer-Verlag. 624 p.
8. Bohacek, S., and B. Rozovskii. 2004. A diffusion model of roundtrip time. *Comput. Stat. Data An.* 45(1):25–50.

Received August 28, 2019

Contributors

Bosov Alexey V. (b. 1969) — Doctor of Science in technology, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; AVBosov@ipiran.ru

Stefanovich Alexey I. (b. 1983) — principal specialist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; AStefanovich@frccsc.ru

ВЫРАВНИВАНИЕ ДЕКАРТОВЫХ ПРОИЗВЕДЕНИЙ УПОРЯДОЧЕННЫХ МНОЖЕСТВ*

А. В. Гончаров[1], В. В. Стрижов[2]

Аннотация: Работа посвящена исследованию метрических методов анализа объектов сложной структуры. Предлагается обобщить метод динамического выравнивания двух временных рядов на случай объектов, определенных на двух и более осях времени. В дискретном представлении такие объекты являются матрицами. Метод динамического выравнивания временных рядов обобщается как метод динамического выравнивания матриц. Предложена функция расстояния, устойчивая к монотонным нелинейным деформациям декартова произведения двух и более временных шкал. Определен выравнивающий путь между объектами. В дальнейшем объектом называется матрица, в которой строки и столбцы соответствуют осям времени. Исследованы свойства предложенной функции расстояния. Для иллюстрации метода решаются задачи метрической классификации объектов на модельных данных и данных из набора MNIST.

Ключевые слова: функция расстояния; динамическое выравнивание; расстояние между матрицами; нелинейные деформации времени; пространственно-временные ряды

DOI: 10.14357/19922264200105

1 Введение

Временны́е ряды представляют собой набор измерений, упорядоченных по оси времени. Анализ временны́х рядов производится при решении задач, связанных с классификацией активности человека по измерениям акселерометра телефона, поиском паттернов в EEG-сигналах (электроэнцефалограмма), кластеризации набора ECoG (электрокортикограмма) данных и во многих других задачах [1]. Рассматриваются объекты, для которых время между измерениями фиксированно. В данной работе для построения адекватной функции расстояния между объектами требуется учесть нелинейные деформации относительно оси времени: глобальные и локальные сдвиги, растяжения и сжатия [2].

В [3] приводятся различные методы решения задач анализа временны́х рядов: классификации, детектирования паттернов, кластеризации и др. В [4] описание временных рядов строится с помощью анализа параметров моделей, в [5] используется их признаковое описание, в [6] анализируется их форма. Комбинации этих подходов описаны в [3].

Метрические методы находят схожие объекты в наборе. Используются функции расстояния над временны́ми рядами: расстояние Хаусдорфа [7], MODH [8], расстояние, основанное на HMM (hiden Markov model) [9], евклидово расстояние в исходном пространстве или в пространстве сниженной размерности [6], LCSS (longest common subsequence) [10]. Показано [11], что в случае локальных или глобальных деформаций времени при решении задач, требующих анализа исходной формы временно́го ряда, метод динамического выравнивания оси времени DTW (Dynamic Time Warping) превосходит другие функции расстояния [12] по качеству итогового решения задачи, так как при наличии смещений двух объектов относительно друг друга требуется выравнивать их оптимальным образом для вычисления расстояния между ними.

В данной работе предлагается перейти от рассмотрения объекта $\mathbf{s}(t)$, временно́го ряда, к более общему случаю $\mathbf{s}(\mathbf{t})$, в котором компоненты вектора $\mathbf{t}$ — оси времени. Из-за существенного роста вычислительной сложности при увеличении числа осей времени предлагается рассмотреть объекты $\mathbf{s}(t_1, t_2)$, определенные на двух осях времени. Оси времени считаются независимыми. В случае единственной дискретной и ограниченной сверху шкалы времени объект представим вектором фиксированной размерности. Аналогично объект настоящего исследования представим матрицей.

*Работа выполнена при частичной финансовой поддержке РФФИ (проекты 19-07-1155 и 19-07-00885). Настоящая статья содержит результаты проекта «Статистические методы машинного обучения», выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по договору МГУ им. М. В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 11.12.2018 № 13/1251/2018.

[1]Московский физико-технический институт, alex.goncharov@phystech.edu

[2]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук; Московский физико-технический институт, strijov@ccas.ru

Вводятся ограничения на зависимости осей времени в декартовом произведении для таких объектов. Определена гипотеза порождения данных: объекты одного класса эквивалентности получены при помощи допустимых преобразований, а именно: локальных деформаций (растяжений и сжатий) каждой из осей времени по отдельности. В дискретном случае преобразование представимо дуплицированием строк и столбцов матриц. В число допустимых преобразований попадают и глобальные деформации: сдвиги по осям времени, представимые добавлением и удалением крайних строк и столбцов исходных матриц. Для каждой из осей времени выполняются свойства времени: монотонность и непрерывность. Похожими на описанные свойствами обладает, например, частотный спектр сигнала, где одна ось определяет время, а другая — частоту, величину, обратную времени.

Между двумя объектами, матрицами, в случае допустимых преобразований требуется определить инвариантную к преобразованиям осей времени функцию расстояния, которая сможет выделить классы эквивалентности множества преобразованных объектов. Работа посвящена определению такой функции расстояния, как обобщения метода динамического выравнивания временных рядов DTW для матриц.

Цель данной работы — построение метода, основанного на динамическом выравнивании осей времени для матриц. Метод динамического выравнивания временны́х рядов [13] определен только для объектов с одной осью времени, что делает его неприменимым для описанного случая. Однако концепции, используемые на каждой стадии вычисления оптимального выравнивания, обобщены на рассматриваемый случай. Работа исследует свойства предложенного метода и сравнивает результаты применения метода к задачам классификации изображений [14] с результатами функции расстояния L_2.

Для иллюстрации и анализа результатов решается задача метрической классификации объектов (матриц низкой размерности). Используются наборы данных: модельные данные, которые согласуются с выдвинутой гипотезой порождения данных для временны́х рядов, подмножество набора MNIST сниженной размерности и частотный спектр сигнала.

2 Постановка задачи построения функции расстояния

Рассмотрим задачу построения функции расстояния между объектами. Функция расстояния инвариантна к допустимым преобразованиям осей времени: глобальным и локальным линейным и нелинейным деформациям временно́й шкалы. Ниже приведены две постановки задачи, с помощью которых определены свойства предложенной функции расстояния, оценено ее качество и проведено сравнение нескольких функций расстояния: предложенной и L_2.

Первая постановка задачи использует общее свойство функций расстояния: объединение схожих объектов и разделение непохожих объектов. Вводится определение свойства инвариантности функции расстояния к допустимым преобразованиям осей времени. Вторая постановка задачи уточняет первую и заключается в проведении метрической классификации методом ближайшего соседа.

Постановка задачи выбора функции расстояния между двумя объектами. На двух временны́х осях заданы объекты вида $\mathbf{A}(t_1, t_2) \in \mathbb{R}^{n\times n}$. Функция $G_w(\mathbf{A}) : \mathbb{R}^{n\times n} \to \mathbb{R}^{\hat{n}\times\hat{n}}$ задает допустимые преобразования исходного объекта $\mathbf{A}$: глобальные сдвиги, локальные линейные и нелинейные деформации, а именно: растяжения и сжатия оси времени, сдвиги значений по оси времени. Скалярный параметр $w \in \mathbb{R}^+$ функции G фиксирует набор этих преобразований.

Допустимым элементарным преобразованием матрицы $\mathbf{A}$ назовем дуплицирование случайных строк и столбцов исходной матрицы, добавление или удаление крайних строк и столбцов. Допустимым преобразованием примем некоторую последовательность допустимых элементарных преобразований матрицы $\mathbf{A}$ и обозначим как $G_w(\mathbf{A})$.

Будем называть объект $\mathbf{B} \in \mathbb{R}^{\hat{n}\times\hat{n}}$ полученным из объекта $\mathbf{A}$ при помощи допустимых преобразований $G_{\hat{w}}$, если существует $\hat{w} \in \mathbb{R}^+ : \mathbf{B} = G_{\hat{w}}(\mathbf{A})$.

Функцию расстояния между двумя объектами $\rho : \mathbb{R}^{n\times n} \times \mathbb{R}^{\hat{n}\times\hat{n}} \to \mathbb{R}^+$ оценим на выборке $\mathfrak{D} = \{\mathbf{A}_i\}_{i=1}^m$ объектов вида $\mathbf{A}_i \in \mathbb{R}^{n\times n}$.

Для каждого объекта выборки $\mathbf{A}_i$ и объекта $\mathbf{B}_j$ его класса эквивалентности $\{\mathbf{B}_j\}_i = \{\mathbf{B} \in \mathfrak{D} | \exists w_i, w_j : G_{w_i}(\mathbf{A}_i) = G_{w_j}(\mathbf{B}_j)\}$ заданы допустимые трансформации с параметрами w_i и w_j, такие что $G_{w_i}(\mathbf{A}_i) = G_{w_j}(\mathbf{B}_j)$. Для каждого объекта выборки $\mathbf{A}_i$ и объекта $\mathbf{C}_j$ из других классов эквивалентности $\{\mathbf{C}_k\}_i = \{\mathbf{C} \in \mathfrak{D} | \nexists w_i, w_k : G_{w_i}(\mathbf{A}_i) = G_{w_k}(\mathbf{C})\}$ не существует таких $w_i, w_k : G_{w_i}(\mathbf{A}_i) = G_{w_k}(\mathbf{C}_k)$.

Решается задача поиска функции расстояния ρ, значение которой на паре объектов одного класса эквивалентности меньше, чем на любой паре объектов из разных: для любых $i, j, k \in \{1, \ldots, m\}$ $\rho(\mathbf{A}_i, \mathbf{B}_j) < \rho(\mathbf{A}, \mathbf{C}_k)$. Функцию расстояния, обладающую таким свойством, назовем инвариантной на классах эквивалентности.

Критерием качества для функции расстояния ρ на выборке $\mathfrak{D}$ примем долю объектов, для которых указанное неравенство выполняется:

$$S_\rho(\mathfrak{D}) = \frac{1}{m}\sum_{i=1}^{m}\prod_{\{\mathbf{B}_j\}_i}\prod_{\{\mathbf{C}_k\}_i}\left[\rho(\mathbf{A}_i,\mathbf{B}_j) < \rho(\mathbf{A}_i,\mathbf{C}_k)\right].$$

Постановка задачи выбора функции расстояния ρ сводится к задаче максимизации критерия качества.

Прикладное использование функции расстояния. Задана выборка $\mathfrak{D} = \{(\mathbf{A}_i, y_i)\}_{i=1}^{m}$, состоящая из пар объект–ответ. Объектами служат объекты сложной структуры: $\mathbf{A}_i \in \mathbb{R}^{n\times n}$, а ответами выступают метки класса — $y_i \in Y = \{1,\ldots,E\}$, где $E \ll m$. Выборка разделена на обучение $\mathfrak{D}_l = \{(\mathbf{A}_i, y_i)\}_{i=1}^{m_1}$ и контроль $\mathfrak{D}_t = \{(\mathbf{A}_i, y_i)\}_{m_1}^{m_1+m_2}$.

Модель классификации f принадлежит множеству моделей метрической классификации 1NN, которые классифицируемому объекту ставят в соответствие метку класса ближайшего объекта из обучающей выборки по заданной функции расстояния ρ:

$$\hat{y} = f(\mathbf{B}|\rho) = y\,\underset{i=1,\ldots,m_1}{\arg\min}\,\rho\left(B, A_i\right).$$

Критерий качества S модели f для задачи классификации — доля правильно проставленного класса на контрольной выборке:

$$S(f|\rho) = \frac{1}{m_2}\sum_{i=m_1}^{m_1+m_2}\left[f(\mathbf{A}_i|\rho) = y_i\right].$$

Требуется выбрать функцию расстояния ρ для модели классификации $f : \mathbb{R}^{n\times n} \to Y$, максимизирующую критерий качества S на контрольной выборке:

$$f = \underset{\rho\in\{\mathrm{mDTW}, L_2\}}{\arg\max}\left(S(f|\rho)\right).$$

3 Вычисление матричного расстояния mDTW

Предлагается использовать функцию расстояния DTW, модифицированную для случая выравнивания двойной шкалы времени.

Определение 1. Даны два объекта $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$. Тензор невязок $\mathbf{\Omega}^{n\times n\times n\times n}$ — такой тензор, что его элемент $\mathbf{\Omega}(i,j,k,l)$ равен квадрату разности между элементами $\mathbf{A}(i,j)$ и $\mathbf{B}(k,l)$:

$$\mathbf{\Omega}(i,j,k,l) = (\mathbf{A}(i,j) - \mathbf{B}(k,l))^2.$$

Определение 2. Путем $\boldsymbol{\pi}$ между двумя объектами $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ назовем множество индексов тензора $\mathbf{\Omega}$:

$$\boldsymbol{\pi} = \{(i,j,k,l)\}, \quad i,j,k,l \in \{1,\ldots,n\},$$

удовлетворяющее следующим условиям:

Частичный порядок. Для элементов пути $\boldsymbol{\pi}$ с фиксированными значениями i,k задан порядок: выравнивающий путь для фиксированных строк двух матриц упорядочен — $\{(i,j_r,k,l_r))\}_{r=1}^{R} \subset \boldsymbol{\pi}$ мощностью R. Аналогично для фиксированных столбцов с индексами j,l.

Граничные условия. Пусть $(i,j,k,l) \in \boldsymbol{\pi}$, тогда $(1,j,1,l) \in \boldsymbol{\pi}$ и $(i,1,k,1) \in \boldsymbol{\pi}$. Путь $\boldsymbol{\pi}$ содержит элементы тензора $\mathbf{\Omega}$: $(1,1,1,1) \in \boldsymbol{\pi}$ и $(n,n,n,n) \in \boldsymbol{\pi}$.

Непрерывность по направлению. Для упорядоченного подмножества пути $\{(i,j_r,k,l_r)\}_{r=1}^{R} \subset \boldsymbol{\pi}$ выполняется условие непрерывности:

$$j_r - j_{r-1} \le 1, \quad l_r - l_{r-1} \le 1, \quad r = 2,\ldots,R.$$

На шаге пути $\boldsymbol{\pi}$ по фиксированному направлению времени i,k встречаются только соседние элементы матрицы (включая соседние по диагонали). Аналогично для фиксированных j,l.

Монотонность по направлению. Для упорядоченного подмножества пути $\{(i,j_r,k,l_r)\}_{r=1}^{R} \subset \boldsymbol{\pi}$ выполняется хотя бы одно из условий монотонности функции выравнивания времени:

$$j_r - j_{r-1} \ge 1, \quad l_r - l_{r-1} \ge 1, \quad r = 2,\ldots,R.$$

Свойства пути между матрицами обобщают свойства пути между двумя временными рядами.

Определение 3. Стоимость $\mathrm{Cost}\,(\mathbf{A},\mathbf{B},\boldsymbol{\pi})$ пути $\boldsymbol{\pi}$ между объектами $\mathbf{A}, \mathbf{B}$:

$$\mathrm{Cost}\,(\mathbf{A},\mathbf{B},\boldsymbol{\pi}) = \sum_{(i,j,k,l)\in\boldsymbol{\pi}}\mathbf{\Omega}(i,j,k,l).$$

Определение 4. Выравнивающий путь $\hat{\boldsymbol{\pi}}$ между объектами $\mathbf{A}, \mathbf{B}$ — путь наименьшей стоимости среди всех возможных путей между объектами:

$$\hat{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}}{\arg\min}\,\mathrm{Cost}\,(\mathbf{A},\mathbf{B},\boldsymbol{\pi}).$$

Функция расстояния $\rho(\mathbf{A},\mathbf{B}) = \mathrm{mDTW}\,(\mathbf{A},\mathbf{B})$ между объектами $\mathbf{A}$ и $\mathbf{B}$ рассчитывается как стоимость выравнивающего пути $\hat{\boldsymbol{\pi}}$:

$$\mathrm{mDTW}(\mathbf{A},\mathbf{B}) = \mathrm{Cost}\,(\mathbf{A},\mathbf{B},\hat{\boldsymbol{\pi}}). \qquad (1)$$

Алгоритм вычисления значения расстояния (4). Построение алгоритма вычисления значения функции расстояния между матрицами основан на алгоритме расчета функции расстояния между временны́ми рядами. В случае выравнивания одной

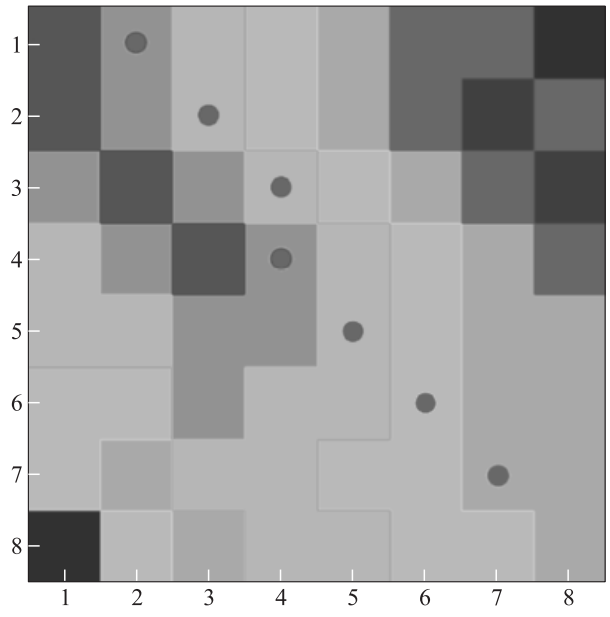


Рис. 1 Матрица стоимости оптимального выравнивания, по обеим осям отложены временные отсчеты

временно́й шкалы итоговая матрица расстояний $\boldsymbol{D}$ (рис. 1) в каждом элементе $\boldsymbol{D}(i,j)$ содержит расстояние между подрядом первого временно́го ряда и подрядом второго временно́го ряда. Рассмотрим алгоритм динамического выравнивания двух временны́х рядов $\mathbf{s} \in R^n$ и $\mathbf{c} \in R^m$ на рис. 2.

Элемент $\boldsymbol{D}(i,j)$ матрицы $\boldsymbol{D}$ соответствует стоимости выравнивающего пути между подпоследовательностями исходных временны́х рядов: $\mathbf{s}(1:i) = \mathbf{s}(t),\ t = 1,\ldots,i$, и $\mathbf{c}(1:j) = \mathbf{c}(t),\ t = 1,\ldots,j$. Алгоритм построения наилучшего выравнивания времени подразумевает, что выравнивающий путь между этими подпоследовательностями получен одним из трех способов — если стоимость выравнивающего пути между подпоследовательностями $\mathbf{s}(1:\overline{i})$ и $\mathbf{c}(1:\overline{j})$ минимальна для $\overline{i,j}$ из множества

$$\overline{i,j} \in \{\{i-1,j\},\{i,j-1\},\{i-1,j-1\}\},$$

тогда выравнивающий путь между $\mathbf{s}(1:i)$ и $\mathbf{c}(1:j)$ получен добавлением пары (i,j) к выбранному выравнивающему пути с минимальной стоимостью из трех.

Предложенный алгоритм переносит эти рассуждения на случай выравнивания двух матриц $\mathbf{A}$ и $\mathbf{B}$. Элемент $\boldsymbol{D}(i,j,k,l)$ четырехиндексного тензора расстояний $\boldsymbol{D}$ соответствует стоимости выравнивающего пути между $\mathbf{A}(1:i,1:j) = \mathbf{A}(t_1,t_2)$, $t_1 = 1,\ldots,i$, $t_2 = 1,\ldots,j$, и $\mathbf{B}(1:k,1:l) = \mathbf{B}(t_1,t_2)$, $t_1 = 1,\ldots,k$, $t_2 = 1,\ldots,l$. Выравнивающий путь между этими подматрицами получен одним из семи способов — если стоимость выравнивающего пути между подматрицами $\mathbf{A}(1:\overline{i},1:\overline{j})$ и $\mathbf{B}(1:\overline{k},1:\overline{l})$ минимальна для $\overline{i,j,k,l}$ из множества

$$\begin{aligned}\overline{i,j,k,l} \in \{&\{i-1,j,k,l\},\{i,j-1,k,l\},\\ &\{i,j,k-1,l\},\{i,j,k,l-1\},\{i-1,j,k-1,l\},\\ &\{i,j-1,k,l-1\},\{i-1,j-1,k-1,l-1\}\},\end{aligned}$$

то к выравнивающему пути между этими подматрицами добавляется элемент пути (i,j,k,l) и по-

DTW(**s**,**c**):
  $\boldsymbol{D}$(1:n+1,1:m+1) = inf;
  $\boldsymbol{D}$(1,1) = 0;
  for $i = 2{:}\ n+1$
    for $j = 2 : m+1$
      $d = (\mathbf{s}(i-1) - \mathbf{c}(j-1))^2$;
      $\boldsymbol{D}(i,j) = d + \min([\boldsymbol{D}(i-1,j), \boldsymbol{D}(i,j-1), \boldsymbol{D}(i-1,j-1)])$;
return sqrt($\boldsymbol{D}(n+1,m+1)$)

Рис. 2 Алгоритм вычисления DTW для временных рядов

Correction $(\overline{i,j,k,l}, \boldsymbol{\pi}(\overline{i,j,k,l}))$:
  if $\overline{i,j,k,l} \in \{(i-1,j,k,l);(i,j,k-1,l);(i-1,j,k-1,l)\}$:
    $\widehat{\pi} = \{(\overline{i},r,\overline{k},f) \in \boldsymbol{\pi}(\overline{i,j,k,l}) | r,f \in \mathbb{N}\}$
  elif $\overline{i,j,k,l} \in \{(i,j-1,k,l);(i,j,k,l-1);(i,j-1,k,l-1)\}$:
    $\widehat{\pi} = \{(r,\overline{j},f,\overline{l}) \in \boldsymbol{\pi}(\overline{i,j,k,l}) | r,f \in \mathbb{N}\}$
  elif $\overline{i,j,k,l} = i-1, j-1, k-1, l-1$:
    $\widehat{\pi} = \{(\overline{i},r,\overline{k},f) \in \boldsymbol{\pi}(\overline{i,j,k,l}) | r,f \in \mathbb{N}\} \cup$
    $\cup\{(r,\overline{j},f,\overline{l}) \in \boldsymbol{\pi}(\overline{i,j,k,l}) | r,f \in \mathbb{N}\}$
  $\boldsymbol{d\pi} = \{\text{element} \in \widehat{\pi} : \text{произведены замены индексов } \overline{i} = i,\ \overline{j} = j,\ \overline{k} = k,\ \overline{l} = l\}$
return $\boldsymbol{d\pi}$

Рис. 3 Алгоритм вычисления поправки $\boldsymbol{d\pi}$ пути $\boldsymbol{\pi}$

mDTW $(\mathbf{A}, \mathbf{B})$:
  $\mathbf{D}(1:n+1, 1:n+1, 1:n+1, 1:n+1) = inf$;
  $\mathbf{D}(1,1,1,1) = 0$;
  $\boldsymbol{\pi}(1,1,1,1) = ((1,1),(1,1))$
  $for\ i,j,k,l \in \mathbb{N}^{2:n+1} \times \mathbb{N}^{2:n+1} \times \mathbb{N}^{2:n+1} \times \mathbb{N}^{2:n+1}$:
    $\overline{i,j,k,l} = \arg\min($ [$\mathbf{D}$(i-1, j, k, l), $\mathbf{D}$(i, j-1, k, l), $\mathbf{D}$(i, j, k-1, l), $\mathbf{D}$(i, j, k, l-1),
      $\mathbf{D}(i-1,j,k-1,l), \mathbf{D}(i,j-1,k,l-1), \mathbf{D}(i-1,j-1,k-1,l-1)]);$
  $\boldsymbol{d\pi} = \text{Correction}\,(\overline{i,j,k,l}, \boldsymbol{\pi}(\overline{i,j,k,l}))$
  $\boldsymbol{\pi}(i,j,k,l) = \boldsymbol{d\pi} \cup \{(\overline{i,j,k,l})\}$
  $\text{cost} = (\mathbf{A}(i,j) - \mathbf{B}(k,l))^2 + \sum_{(r,f,t,g)\in \boldsymbol{d\pi}} (\mathbf{A}(r,f) - \mathbf{B}(t,g))^2$;
  $\mathbf{D}(i,j,k,l) = \text{cost} + \mathbf{D}(\overline{i,j,k,l})$
return sqrt($\mathbf{D}(n+1,n+1,n+1,n+1)$)

Рис. 4 Алгоритм вычисления расстояния между матрицами

правка $\boldsymbol{d\pi}$ пути $\boldsymbol{\pi}$, алгоритм вычисления которой приведен ниже.

Обозначим выравнивающий путь между $\mathbf{A}(1:i, 1:j)$ и $\mathbf{B}(1:k, 1:l)$ как $\boldsymbol{\pi}(i,j,k,l)$, тогда поправка $\boldsymbol{d\pi}$ пути $\boldsymbol{\pi}(i,j,k,l)$ при фиксированных $\overline{i,j,k,l}$ вычисляется приведенным на рис. 3 образом.

Алгоритм динамического выравнивания двух матриц и вычисления расстояния mDTW между ними с учетом приведенного выше алгоритма примет вид, представленный на рис. 4.

Следует отметить, что алгоритм [15] имеет высокую сложность вычисления — $O(n^4)$. Предполагается ускорение метода с использованием ограничения Sakoe-Chiba band, что сократит вычислительную сложность алгоритма до $O(n^2k^2)$, где k — параметр ограничения.

4 Вычислительный эксперимент

Вычислительный эксперимент проведен на модельных данных с допустимыми преобразованиями и на реальных данных: объектах коллекции MNIST с допустимыми преобразованиями и на спектрограммах зашумленных сигналов.

Решается задача метрической классификации методом ближайшего соседа. В таблице приведены значения критерия качества функции расстояния $S_\rho(\mathfrak{D})$ и критерия качества метрической классификации $S(f|p)$ при использовании двух функций расстояния: предложенной в работе mDTW и L_2.

Модельные данные — это нулевые матрицы со случайными ненулевыми строками, столбцами, подпрямоугольниками с наложенным шумом. К ним применены допустимые преобразования, согласованные с гипотезой наличия локальных и глобальных искажений. На рис. 5 показан пример оптимального выравнивания двух объектов. Линиями показаны элементы пути $\boldsymbol{\pi}$.

Подготовлена подвыборка набора данных MNIST. Она состоит из 100 объектов классов 0 и 1 сниженной размерности с допустимыми преобразованиями. На рис. 6 показан пример оптимального выравнивания объектов.

Аналогичный эксперимент проведен для решения задачи метрической классификации спектров различных сигналов, пример которых приведен на рис. 7. На рисунке показаны примеры Фурье-спектров этих сигналов. Спектр получен путем применения быстрого преобразования Фурье к исходному сигналу для различных окон с фиксированным размером и сдвигом. Исходные временны́е ряды обладали свойством периодичности, период выбирался случайным образом.

Тестирование проведено на разного рода данных: исходных модельных данных без наложения

Снижение расстояний при выполнении преобразований для различных наборов данных

Данные	Метод			
	L_2		MatrixDTW	
	$S(f\|p)$	$S_\rho(\mathfrak{D})$	$S(f\|p)$	$S_\rho(\mathfrak{D})$
Модельные данные без преобразований	92%	78%	100%	85%
Модельные данные с преобразованиями	86%	65%	100%	82%
Модельные данные с преобразованиями и шумом	69%	61%	92%	78%
MNIST без преобразований	95%	—	95%	—
MNIST с преобразованиями	53%	—	92%	—
Спектр сигнала	83%	—	96%	—

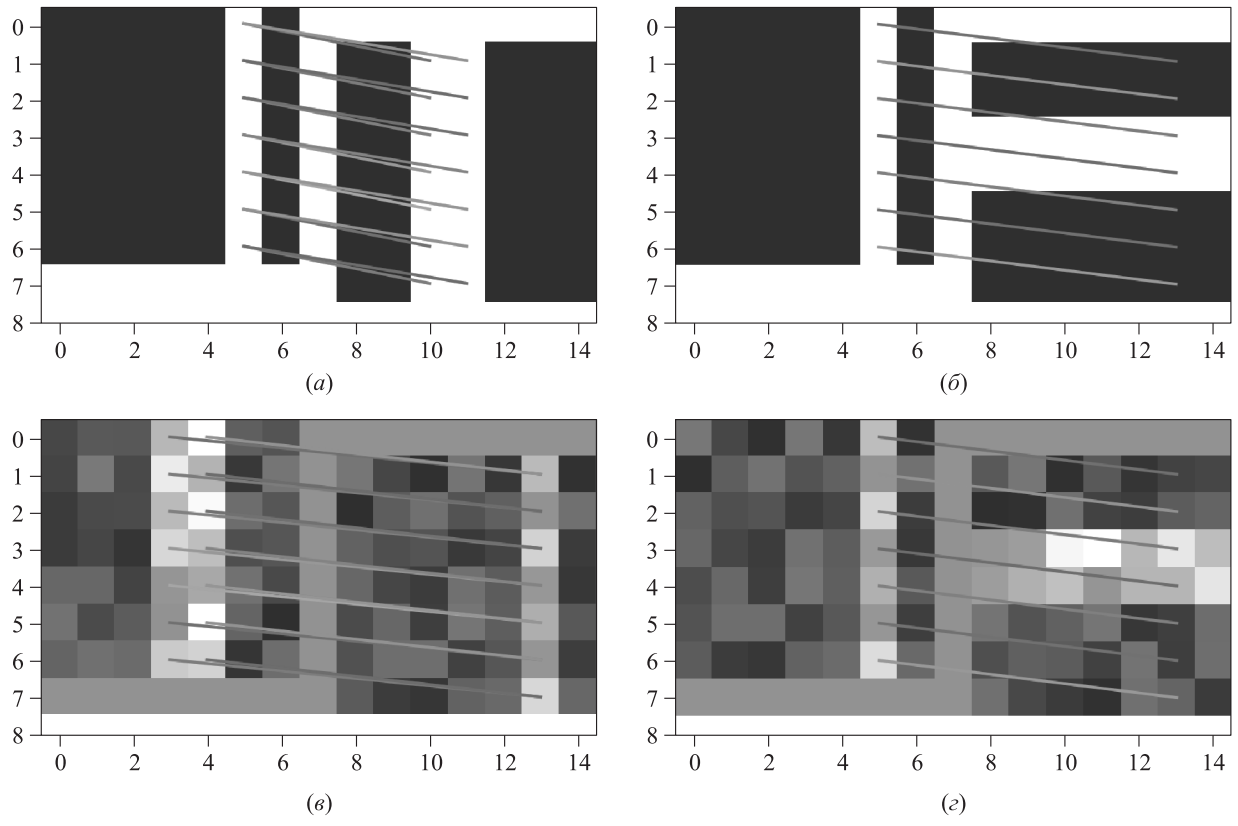


Рис. 5 Выравнивание модельных данных: (*а*) один класс без шума; (*б*) разные классы без шума; (*в*) один класс с шумом; (*г*) разные классы с шумом

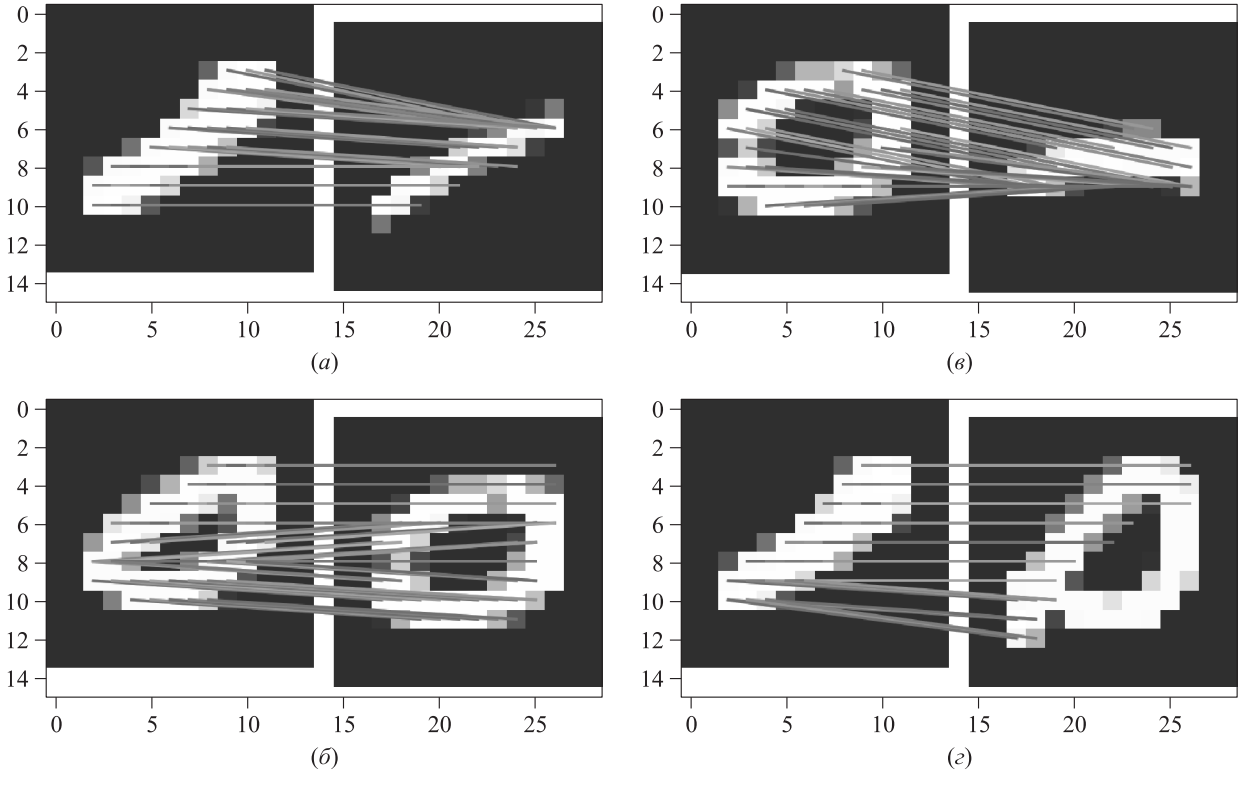


Рис. 6 Выравнивание данных MNIST: левый столбец — один класс; правый столбец — разные классы; (*а*) mDTW = = 720,1; (*б*) 948,6; (*в*) 2017,0; (*г*) mDTW = 2071,4

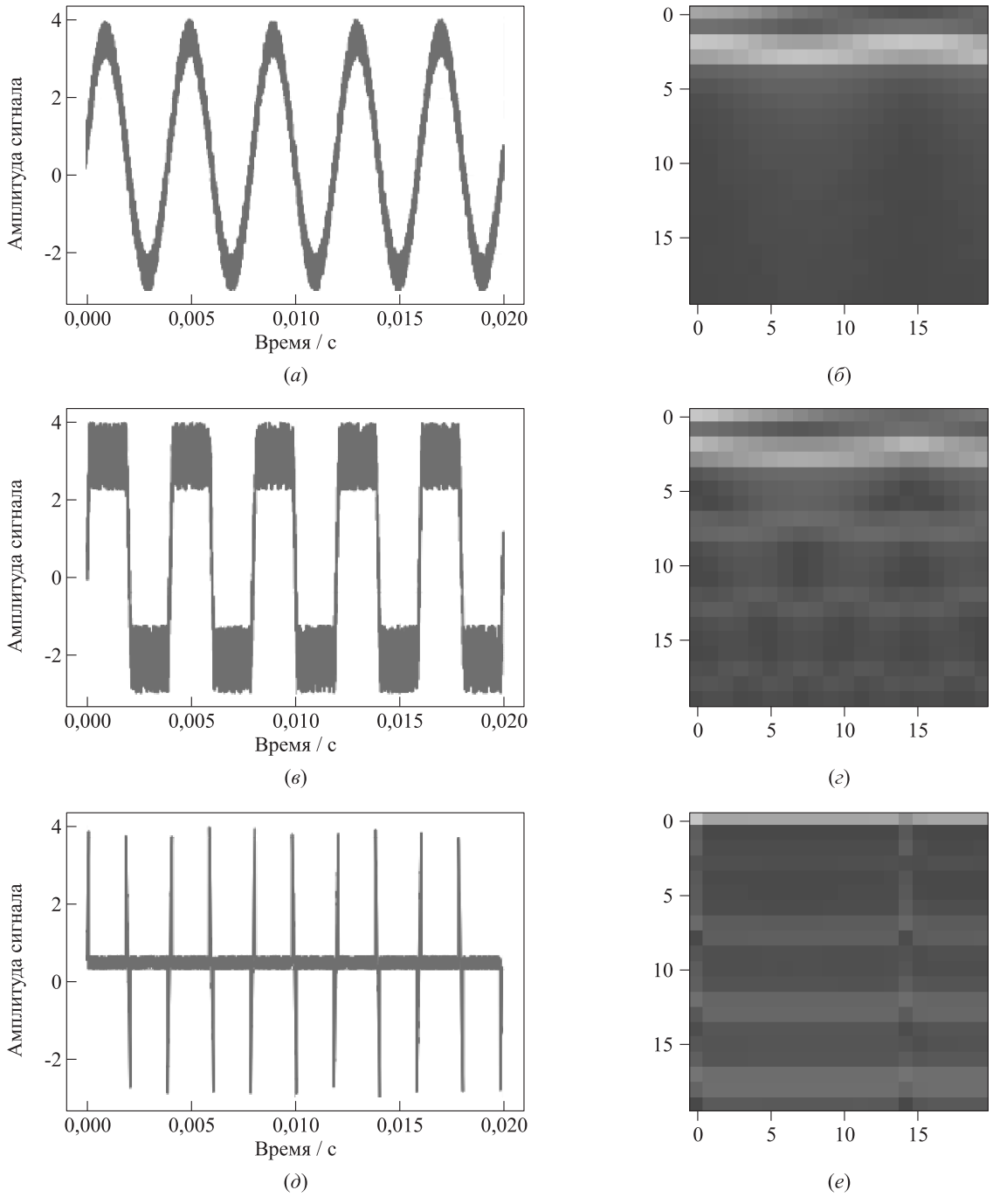


Рис. 7 Данные спектров сигнала: (*а*) класс 1; (*б*) спектр класса 1; (*в*) класс 2; (*г*) спектр класса 2; (*д*) класс 3; (*е*) спектр класса 3

допустимых преобразований, с ними, а также на модельных данных с наложенным поверх объектов случайным шумом.

В каждом из проведенных экспериментов была продемонстрирована устойчивость предложенного подхода к допустимым преобразованиям. Наилучшее значение критерия качества задачи классификации было достигнуто при использовании предложенной функции расстояния.

5 Заключение

В работе предложено обобщение метода динамического выравнивания временны́х рядов для

случая объектов, определенных на двух осях времени. Существует теоретическое обобщение предлагаемых методов на случай конечного множества осей времени. Вычислительный эксперимент позволил проанализировать свойства подхода: устойчивость к допустимым преобразованиям и разделяющая способность функции расстояния как на реальных, так и на модельных данных. Качество решения задачи метрической классификации выше решения, основанного на евклидовом расстоянии. Вычислительная сложность метода высокая, что ограничивает его применимость на объектах высокой размерности.

Литература

1. *Hill N. J., Lal T. N., Schroder M., Hinterberger T., Wilhelm B., Nijboer F., Mochty U., Widman G., Elger C., Scholkopf B., Kubler A., Birbaumer N.* Classifying EEG and ECoG signals without subject training for fast BCI implementation: Comparison of nonparalyzed and completely paralyzed subjects // IEEE T. Neur. Sys. Reh., 2006. Vol. 14. Iss. 2. P. 183–186.
2. *Sakoe H., Chiba S.* A dynamic programming approach to continuous speech recognition // 7th Congress (International) on Acoustics Proceedings, 1971. Vol. 3. P. 65–69.
3. *Aghabozorgi S., Ali S. S., Wah T. Y.* Time-series clustering — a decade review // Inform. Syst., 2015. Vol. 53. P. 16–38.
4. *Warrenliao T.* Clustering of time series data — a survey // Pattern Recogn., 2005. Vol. 38. Iss. 11. P. 1857–1874.
5. *Hautamaki V., Nykanen P., Franti P.* Time-series clustering by approximate prototypes // 19th Conference (International) on Pattern Recognition Proceedings, 2008. No. D. P. 1–4.
6. *Faloutsos C., Ranganathan M., Manolopoulos Y.* Fast subsequence matching in time-series databases // SIGMOD Rec., 1994. Vol. 23. Iss. 2. P. 419–429.
7. *Basalto N., Bellotti R., Carlo F. D., Facchi P., Pascazio S.* Hausdorff clustering of financial time series // Physica A, 2007. Vol. 379. Iss. 2. P. 635–644.
8. *Gorelick L., Blank M., Shechtman E., Irani M., Basri R.* Actions as space-time shapes // IEEE T. Pattern Anal., 2007. Vol. 29. Iss. 12. P. 2247–2253.
9. *Smyth P.* Clustering sequences with hidden Markov models // Adv. Neural In., 1997. Vol. 9. P. 648–654.
10. *Banerjee A., Ghosh J.* Clickstream clustering using weighted longest common subsequences // Workshop on Web Mining, SIAM Conference on Data Mining Proceedings, 2001. P. 33–40.
11. *Aach J., Church G.M.* Aligning gene expression time series with time warping algorithms // Bioinformatics, 2001. Vol. 17. Iss. 6. P. 495–508.
12. *Yi B. K., Faloutsos C.* Fast time sequence indexing for arbitrary $\mathcal{L}_p$ norms // 26th Conference (International) on Very Large Data Bases Proceedings, 2000. P. 385–394.
13. *Goncharov A. V., Strijov V. V.* Analysis of dissimilarity set between time series // Computational Mathematics Modeling, 2018. Vol. 29. Iss. 3. P. 359–366.
14. *Alon J., Athitsos V., Sclaroff S.* Online and offline character recognition using alignment to prototypes // 8th Conference (International) on Document Analysis and Recognition, 2005. Vol. 2. P. 839–843.
15. *Гончаров А. В.* Выравнивания декартовых произведений упорядоченных множеств mDTW. Программная реализация алгоритма, 2019. https://github.com/Intelligent-Systems-Phystech/PhDThesis/tree/master/Goncharov2019/MatrixDTW/code.

Поступила в редакцию 24.04.19

ALIGNMENT OF ORDERED SET CARTESIAN PRODUCT

A. V. Goncharov[1] and V. V. Strijov[1,2]

[1] Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation

[2]A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The work is devoted to the study of metric methods for analyzing objects with complex structure. It proposes to generalize the dynamic time warping method of two time series for the case of objects defined on two or more time axes. Such objects are matrices in the discrete representation. The DTW (Dynamic Time Warping) method of time series is generalized as a method of matrices dynamic alignment. The paper proposes a distance function resistant to monotonic nonlinear deformations of the Cartesian product of two time scales. The alignment path between objects is defined. An object is called a matrix in which the rows and columns correspond to the axes of time. The properties of the proposed distance function are investigated. To illustrate the method, the problems of metric classification of objects are solved on model data and data from the MNIST dataset.

Keywords: distance function; dynamic alignment; distance between matrices; nonlinear time warping; space–time series

DOI: 10.14357/19922264200105

Acknowledgments

This work was supported by the Russian Foundation for Basic Research (projects 19-07-1155 and 19-07-00885). The paper contains results of the project Statistical methods of machine learning, which is carried out within the framework of the Program "Center of Big Data Storage and Analysis" of the National Technology Initiative Competence Center. It is supported by the Ministry of Science and Higher Education of the Russian Federation according to the agreement between the M. V. Lomonosov Moscow State University and the Foundation of project support of the National Technology Initiative from 11.12.2018, No. 13/1251/2018.

References

1. Hill, N. J., T. N. Lal, M. Schroder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. Elger, B. Scholkopf, A. Kubler, and N. Birbaumer. 2006. Classifying EEG and ECoG signals without subject training for fast BCI implementation: Comparison of nonparalyzed and completely paralyzed subjects. *IEEE T. Neur. Sys. Reh.* 14(2):183–186.
2. Sakoe, H., and S. Chiba. 1971. A dynamic programming approach to continuous speech recognition. *7th Congress (International) on Acoustics Proceedings*. 3:65–69.
3. Aghabozorgi, S., S. S. Ali, and T. Y. Wah. 2015. Time-series clustering — a decade review. *Inform. Syst.* 53:16–38.
4. Warrenliao, T. 2005. Clustering of time series data — a survey. *Pattern Recogn.* 38(11):1857–1874.
5. Hautamaki, V., P. Nykanen, and P. Franti. 2008. Time-series clustering by approximate prototypes. *19th Conference (International) on Pattern Recognition Proceedings*. D:1–4.
6. Faloutsos, C., M. Ranganathan, and Y. Manolopoulos. 1994. Fast subsequence matching in time-series databases. *SIGMOD Rec.* 23(2):419–429.
7. Basalto, N., R. Bellotti, F. D. Carlo, P. Facchi, and S. Pascazio. 2007. Hausdorff clustering of financial time series. *Physica A* 379(2):635–644.
8. Gorelick, L., M. Blank, E. Shechtman, M. Irani, and R. Basri. 2007. Actions as space-time shapes. *IEEE T. Pattern Anal.* 29(12):2247–2253.
9. Smyth, P. 1997. Clustering sequences with hidden Markov models. *Adv. Neural In.* 9:648–654.
10. Banerjee, A., and J. Ghosh. 2001. Clickstream clustering using weighted longest common subsequences. *Workshop on Web Mining, SIAM Conference on Data Mining Proceedings*. 33–40.
11. Aach, J., and G. M. Church. 2001. Aligning gene expression time series with time warping algorithms. *Bioinformatics* 17(6):495–508.
12. Yi, B. K., and C. Faloutsos. 2000. Fast time sequence indexing for arbitrary $\mathcal{L}_p$ norms. *26th Conference (International) on Very Large Data Bases Proceedings*. 385–394.
13. Goncharov, A. V., and V. V. Strijov. 2018. Analysis of dissimilarity set between time series. *Computational Mathematics Modeling* 29(3):359–366.
14. Alon, J., V. Athitsos, and S. Sclaroff. 2005. Online and offline character recognition using alignment to prototypes. *8th Conference (International) on Document Analysis and Recognition*. 2:839–843.
15. Goncharov, A. V. Alignment of Ordered Set Cartesian Product mDTW. Software implementation of the algorithm. Available at: https://github.com/Intelligent-Systems-Phystech/PhDThesis/tree/master/Goncharov 2019/MatrixDTW/code (accessed December 27, 2019).

Received April 24, 2019

Contributors

Goncharov Alexey V. (b. 1995) — PhD student, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141701, Russian Federation; alex.goncharov@phystech.edu

Strijov Vadim V. (b. 1967) — Doctor of Science in physics and mathematics, leading scientist, A. A. Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; professor, Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141701, Russian Federation; strijov@ccas.ru

НЕЙРОФИЗИОЛОГИЯ КАК ПРЕДМЕТНАЯ ОБЛАСТЬ ДЛЯ РЕШЕНИЯ ЗАДАЧ С ИНТЕНСИВНЫМ ИСПОЛЬЗОВАНИЕМ ДАННЫХ*

Д. О. Брюхов[1], С. А. Ступников[2], Д. Ю. Ковалёв[3], И. А. Шанин[4]

Аннотация: Цель данного обзора — анализ нейрофизиологии как области с интенсивным использованием данных. В настоящее время происходит заметный рост числа исследований в области изучения человеческого мозга. Появляются крупные международные проекты, поддерживающие исследования, направленные на улучшение понимания работы человеческого мозга, а также на обнаружение и поиск способов лечения основных заболеваний, связанных с человеческим мозгом. Объем данных, генерируемых в типичной лаборатории, проводящей исследования в области нейрофизиологии, растет в геометрической прогрессии. При этом данные представляются с использованием большого числа разнообразных форматов. Это приводит к необходимости создания инфраструктур и баз данных, а также веб-сайтов, предоставляющих единый доступ к данным и обеспечивающим обмен этими данными между исследователями по всему миру. Для анализа собранных данных применяются методы и средства из области нейроинформатики — науки на стыке нейрофизиологии и информатики. Для решения нейрофизиологических задач применяются различные методы информатики, такие как статистический анализ и машинное обучение, в частности нейронные сети.

Ключевые слова: нейрофизиология; нейроинформатика; интенсивное использование данных; анализ данных

DOI: 10.14357/19922264200106

1 Введение

Нейрофизиология — один из ярких примеров научной области с интенсивным использованием данных. Она представляет собой комбинацию различных областей знаний: анатомии, физиологии, генетики, биохимии, психологии — и стала передовой областью в исследовании и моделировании работы человеческого мозга.

В настоящее время во всем мире растет интерес к научному пониманию работы человеческого мозга, выражающийся в количестве исследований в области нейрофизиологии. Появляется новое, более качественное оборудование, позволяющее получать более точные данные различного вида, в частности данные магнитно-резонансной томографии (МРТ), электроэнцефалографии (ЭЭГ), магнитоэнцефалографии (МЭГ) и др. Новое оборудование позволяет за несколько дней собирать больше данных, чем всего десять лет назад собиралось за целый год.

С увеличением объема данных встает проблема совместного использования этих данных для решения разнообразных задач в области нейрофизиологии. Стали появляться как региональные консорциумы и проекты, поддерживающие исследователей для решения различных задач в области нейрофизиологии (например, американская инициатива BRAIN (The Brain Research through Advancing Innovative Neurotechnologies® Initiative), европейские проекты HBP (Human Brain Project) и BNCI (Brain-Neural-Computer-Interaction) из программы Horizon 2020), так и консорциумы, объединяющие исследователей во всем мире для решения конкретных задач, такие как проект исследования коннектома человека HCP (Human Connectome Project), инициатива по нейровизуализации болезни Альцгеймера ADNI

* Работа выполнена при частичной финансовой поддержке РФФИ (проект 18-29-22096).

[1] Институт проблем информатики Федерального исследовательского центра «Информатика и управлени» Российской академии наук, dbriukhov@ipiran.ru

[2] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sstupnikov@ipiran.ru

[3] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, dm.kovalev@gmail.com

[4] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ivan.shanin@gmail.com

(Alzheimer's Disease Neuroimaging Initiative), инициатива по развитию маркеров для болезни Паркинсона PPMI (Parkinson's Progression Markers Initiative).

С ростом объема данных растет и разнообразие этих данных. К сожалению, в области нейрофизиологии в настоящее время нет единых стандартов для представления данных. Это относится практически ко всем видам данных, в частности к нейроизображениям и биомедицинским сигналам. Разнообразие форматов данных вызвано большим числом видов медицинского оборудования, а также средств визуализации и анализа получаемых данных.

С ростом объема и разнообразия данных продолжает усиливаться стремление разместить их в доступных репозиториях (базах данных, веб-сайтах). Такие репозитории могут содержать петабайты нейрофизиологических данных и позволяют обмениваться ими исследователям по всему миру. Некоторые базы содержат данные для решения конкретного класса задач, другие — широкий набор различных данных. При добавлении новых данных в эти базы данные обычно проходят процесс рецензирования.

Базы данных могут содержать данные как в каком-то определенном формате, так и поддерживать несколько разных форматов, принятых в сообществе. Базы данных и веб-сайты предоставляют единый интерфейс доступа к зарегистрированным в них данным. Некоторые сайты предоставляют также программные средства для визуализации содержащихся в них данных.

За последние годы созданы десятки программных средств для сбора, обработки, анализа и визуализации данных в области нейрофизиологии, основанные на методах и средствах из области информатики и применяющих методы моделирования из области нейрофизиологии. Таким образом, формируется нейроинформатика как междисциплинарная область сотрудничества исследователей-нейрофизиологов с исследователями-информатиками.

В рамках статьи предоставлена информация о текущем состоянии дел в области нейрофизиологии (при этом основное внимание уделяется моделированию когнитивных функций на основе нейрофизиологических данных): основные мировые стратегические инициативы и проекты (разд. 2), крупные базы данных, содержащие данные исследований (разд. 3), основные форматы представления нейроизображений и биомедицинских сигналов (разд. 4), программные средства для обработки и анализа нейроизображений (разд. 5).

2 Крупные международные консорциумы и проекты в области нейрофизиологии

С увеличением объема данных и числа исследований в области нейрофизиологии встает задача компьютерной поддержки этих исследований и совместного использования полученных данных. Появляются как региональные консорциумы и проекты, поддерживающие исследователей для решения различных задач в области нейрофизиологии, так и консорциумы, объединяющие исследователей во всем мире для решения конкретных задач или для лечения различных заболеваний, связанных с мозгом.

Инициатива исследования мозга с помощью продвинутых инновационных технологий (BRAIN) [1] была объявлена в США в 2013 г. и представляет собой 10-летнюю программу, направленную на революцию в понимании работы человеческого мозга.

Начатый в 2013 г. проект *Human Brain Project* [2] — это десятилетний проект поддержки исследований человеческого мозга, курируемый Европейским Союзом (ЕС). Цель проекта — создание современной исследовательской инфраструктуры, которая позволит исследователям расширять знания в понимании работы человеческого мозга.

Стартовавший в 2010 г. проект *Human Connectome Project* [3] является попыткой картирования нервных путей, лежащих в основе функционирования человеческого мозга. Цель проекта — сбор и обмен данными о структурной и функциональной связанности человеческого мозга (коннектома) в макромасштабе (в сантиметровом и миллиметровом масштабе).

Проект *BNCI Horizon 2020* [4] в рамках 7-й рамочной программы ЕС направлен на поддержку и координацию усилий в области интерфейсов мозг–компьютер (BCI, Brain–Computer Interface) и нейроинтерфейсов мозг–компьютер (BNCI). Основная цель этого проекта — разработка дорожной карты для области BCI с особым упором на промышленные приложения BCI и конечных пользователей. Этот проект объединяет 12 европейских университетов.

Потребность в использовании данных различных дисциплин для исследования процессов и способов лечения основных заболеваний была признана несколько лет назад [5]. Также была осознана необходимость сотрудничества между центрами и дисциплинами для интеграции и совместного использования разнообразных данных [6] путем организации междисциплинарных консорциумов.

Примером такого консорциума может служить *инициатива по нейровизуализации болезни Альцгеймера* (ADNI) [7], объединяющая исследователей с данными исследований для улучшения профилактики и лечения болезни Альцгеймера. Основные цели инициативы: выявление болезни на ранней стадии и определение способа отслеживания болезни с помощью биомаркеров, применение методов ранней диагностики (когда вмешательство может быть наиболее эффективным), предоставление данных исследований для ученых всего мира.

Другими примерами междисциплинарных консорциумов, использующих обработку нейроизображений, являются инициативы, направленные на лечения таких заболеваний, как болезнь Паркинсона (PPMI), психиатрические расстройства. Поддерживаются базы данных для сбора нейровизуальных, генетических и феноменальных данных об аутизме (National Database of Autism Research) и повреждениях головного мозга (Federal Interacgency Traumatic Brain Injury Research).

3 Инфраструктуры и базы данных в области нейрофизиологии

С целью дальнейшего использования данных, полученных исследователями со всего мира, создаются и поддерживаются инфраструктуры доступа к данным и отдельные базы данных, объединяющие данные от различных исследовательских групп и предоставляющие единый интерфейс доступа к этим данным. Инфраструктуры предоставляют единую среду для доступа к различным данным и использования различных программных средств для обработки этих данных. Ниже рассмотрены основные современные инфраструктуры и базы данных.

Проект *1000 функциональных коннектомов* (1000 Functional Connectomes project) [8] предоставляет доступ к фМРТ-изображениям со всего мира. Проект содержит данные о более 1200 наборах фМРТ-изображений состояния покоя, собранных с 33 разных сайтов. Проект содержит как сырые, так и предобработанные данные, представленные в формате BIDS (brain imaging data structure).

OpenNEURO [9] — бесплатная и открытая платформа для обмена данными МРТ, МЭГ и ЭЭГ. Она является развитием проекта по созданию базы данных *OpenfMRI*, законченного в 2010 г. Первоначально база данных включала только наборы данных, содержащих фМРТ-действия (task based fMRI). В настоящее время она открыта для любых видов МРТ-нейроизображений. Все изображения, хранящиеся в базе данных, представлены в формате BIDS.

База данных ConnectomeDB [10] была разработана в рамках проекта HCP [3] и содержит данные о структурной и функциональной связанности человеческого мозга (коннектома). База данных в настоящее время включает в себя несколько видов данных МРТ, ЭЭГ и МЭГ. Изображения, хранящиеся в ConnectomeDB, представлены в формате NIFTI. Для обработки данных проекта был создан *Connectome Workbench* — свободно предоставляемый инструмент для визуализации и анализа данных, полученных в рамках проекта HCP.

XNAT [11] — это открытая информационная платформа для работы с нейроизображениями, разработанная исследовательской группой по нейроинформатике в Вашингтонском университете. Она облегчает общие задачи управления, обеспечения производительности и качества обработки нейроизображений и связанных данных. *XNAT Central* является общедоступным хранилищем медицинских изображений, основанным на открытой информационной платформе обработки изображений XNAT. В отличие от большинства других хранилищ, таких как ConnectomeDB и Open fMRI, XNAT Central не модерируется для контроля содержимого и не предназначен для поддержки решения каких-либо конкретных научных задач и подходов. Все изображения, хранящиеся в XNAT Central, представлены в формате DICOM.

NITRC [12] — это бесплатный веб-ресурс, который предлагает информацию о постоянно расширяющемся наборе программного обеспечения и данных для нейроинформатики. Он состоит из трех компонентов: реестра ресурсов (NITRC-R), репозитория изображений (NITRC-IR) и вычислительной среды (NITRC-CE). *Вычислительная среда NITRC-CE* представляет собой виртуальную облачную платформу, содержащую предустановленный набор программных средств для работы с нейроизображениями. *Репозиторий изображений NITRC* включает в себя изображения в форматах DICOM и NIFTI.

База данных, разработанная в рамках проекта BNCI Horizon 2020 [4], является общедоступной коллекцией наборов данных в области BCI. Цель создания базы данных — повышение научной прозрачности и эффективности. База данных способствует также валидации опубликованных методов и способствует разработке новых алгоритмов. Данные могут храниться в различных форматах ЭЭГ-данных.

4 Форматы данных в нейрофизиологии

В области нейрофизиологии в настоящее время нет единых стандартов для хранения данных [13]. Это относится как к нейроизображениям, так и к биомедицинским сигналам. Многообразие форматов представления данных вызвано разнообразием как медицинского оборудования, так и средств визуализации и анализа получаемых данных.

4.1 Форматы магнитно-резонансной томографии

Данные нейрофизиологических изображений должны содержать не только сами изображения, но и дополнительную информацию (метаданные), обеспечивающую интероперабельность и повторное использование этих данных. Изображения без связанных с ними метаданных практически бесполезны. К метаданным относятся информация об изображении (размер пикселя, ширина и высота изображения, число изображений), информация об оборудовании, информация об объекте наблюдения, информация о положении объекта наблюдения относительно оборудования.

Наиболее распространенные форматы представления нейроизображений — DICOM (digital imaging and communications in medicine) [14], используемый в большинстве медицинских сканеров, и ANALYZE 7.5 [15], разработанный в клинике Mayo в рамках создания пакета программ Analyze для хранения, визуализации и обработки многомерных биомедицинских изображений. Среди других форматов данных в нейрофизиологии можно отметить NIFTI [16] и BIDS [17].

Форматы определяют, как изображения и метаданные хранятся в файле. Названия конкретных метаданных в каждом формате свои, и их число варьируется от сотни (ANALYZE, NIFTI) до нескольких тысяч (DICOM). В ряде форматов (ANALYZE, NIFTI, BIDS) определяется неизменяемый список используемых метаданных, и любой файл должен содержать значения всех этих метаданных, даже если они неизвестны. Другие форматы (DICOM) используют гибкий набор метаданных, когда конкретные метаданные присутствуют в файле только в том случае, если они определены.

Нейрофизиологические изображения представляются в виде трехмерного массива вокселей, описывающего положение вокселей в трехмерном пространстве. Также может добавляться четвертое измерение — время. Каждый формат определяет свой способ представления этого массива в файле в виде одномерной последовательности вокселей. Форматы отличаются способом задания ориентации изображения относительно сканера: неявная фиксированная ориентация (ANALYZE), кватернионы (NIFTI) и направляющие косинусы (DICOM, NIFTI). Для определения ориентации объекта наблюдения относительно сканера используются в основном два подхода: нейрологический (NIFTI) и радиологический (DICOM).

4.2 Форматы электроэнцефалографии

В области хранения биомедицинских сигналов существует множество разнообразных форматов [18]. Форматы определяют, как метаданные (заголовки) и данные хранятся в файле. Заголовки файлов (метаданные) обычно хранятся в бинарном виде (EDF (European data format) [19], GDF (general data format) [20]), но в некоторых форматах они хранятся в текстовом виде или в виде XML (OpenXDF [21]).

Некоторые биомедицинские данные могут содержать различные виды биомаркеров, для этого форматы (EDF, GDF, OpenXDF) должны поддерживать частоту дискретизации и коэффициенты масштабирования. Первоначально форматы поддерживали хранение 8-битных данных, затем 16-битных (EDF), а все последние форматы поддерживают и типы данных более 16 бит (GDF, OpenXDF). При хранении данных важно знать физическую единицу записанного сигнала, т. е. представляют ли значения выборки милливольт (мВ) или микровольт (мкВ). Большинство форматов поддерживают все физические единицы, представленные в стандарте ISO11073:10101. Но некоторые старые форматы (EDF, GDF 1.0) отводят на это только 8 байт, чего недостаточно для хранения всех единиц.

Биомедицинские сигналы зачастую содержат артефакты. Часть форматов (GDF, OpenXDF) позволяют задавать диапазон изменения значения единиц, что позволяет автоматически находить некорректные данные. Для анализа больших баз данных и архивов важно иметь доступную информацию (поддерживаются в форматах OpenXDF, GDF 2.1) о демографии пациентов, записывающем оборудовании, исследователе и т. д.

5 Программные средства работы с нейрофизиологическими данными

Программные средства работы с нейроизображениями помогают исследователям в изучении мозга человека. Они позволяют визуализировать

данные в виде трехмерных изображений, применять различные методы анализа данных.

Средства визуализации позволяют визуализировать как 2D-, так и 3D- и 4D-нейроизображения (3D Slicer [22], Mango [23]). Кроме визуализации они также позволяют выполнять операции над изображениями, такие как ручная сегментация и создание трехмерной модели поверхности (3D Slicer), создание и редактирование областей интереса в изображениях, рендеринг поверхности, наложение изображений (Mango).

Средства анализа нейроизображений позволяют применять различные методы информатики для анализа нейроизображений. Написанное на MATLAB программное обеспечение *CONN* [24] предназначено для вычисления, визуализации и анализа функциональных связностей в фМРТ. Предоставляются также функции обнаружения и очистки артефактов, динамического анализа связности и анализа на основе поверхности и объема. Система *SPM* [25] предназначена для статистического параметрического картирования, используемого для определения различий в зарегистрированной активности мозга с использованием пространственно расширенных статистических процессов.

Набор инструментов для работы с ЭЭГ *NBT* [26] обеспечивает расчет и интеграцию нейрофизиологических биомаркеров. NBT предлагает конвейер, включающий различные этапы обработки данных: от хранения данных до применения статистических методов, вычисление отклонения артефактов, визуализацию сигналов, вычисление биомаркеров и статистическое тестирование. Программные средства *EEGLAB* [27], *FieldTrip* [28] и *BioSig* [29], реализованные в MATLAB, предназначены для обработки биомедицинских сигналов, таких как ЭЭГ, МЭГ и других электрофизиологических сигналов. EEGLAB реализует метод независимых компонент, частотно-временной анализ, вычисление отклонения артефактов и несколько режимов визуализации данных. FieldTrip предлагает методы предварительной обработки и расширенного анализа, такие как частотно-временной анализ, восстановление источников с использованием диполей, распределенных источников и непараметрическое статистическое тестирование. BioSig предоставляет средства визуализации данных и средства для сбора данных, обработки артефактов, контроля качества, извлечения характеристик, классификации, моделирования данных.

Библиотеки обработки нейроизображений на языке Python помогают разрабатывать собственные программы для работы с нейроизображениями.

Библиотека *NiPy* [30] — библиотека, состоящая из нескольких частей, которые позволяют пользователю выполнять как простые операции с изображениями fMRI (например, чтение и запись), так и сложные алгоритмы анализа нейроизображений. *Nibabel* предоставляет прикладной программный интерфейс для чтения и записи различных форматов файлов нейроизображений, таких как ANALYZE, NIFTI, MINC, MGH. *Niwidgets* предоставляет средства визуализации нейроизображений. Nitime предоставляет средства для анализа временны́х рядов в области нейровизуализации. *Nilearn* предоставляет средства для статистического исследования данных нейровизуализации на основе метода независимых компонент CanICA.

MNE-Python [31] — это программный пакет с открытым исходным кодом, предназначенный для анализа данных МЭГ и ЭЭГ. Он предоставляет современные алгоритмы, которые охватывают несколько методов предварительной обработки данных, локализации источников, статистического анализа, методы машинного обучения.

6 Заключение

Нейрофизиологию можно рассматривать как область с интенсивным использованием данных, где данные играют ключевую роль в исследованиях в понимании работы головного мозга и в обнаружении и лечении заболеваний, связанных с головным мозгом. По всему миру создаются проекты, поддерживающие исследования в этой области. Рост числа исследований и появление нового оборудования ведут к лавинообразному увеличению объема данных. Эти данные могут представляться в различных форматах. Требуются новые средства для хранения данных больших объемов (которые могут достигать нескольких петабайт), средства для интеграции данных, представленных в разных форматах, средства анализа такого объема данных. Отдельные компьютеры больше не подходят для анализа данных в области нейрофизиологии, а потому необходимо развивать новые инфраструктуры, позволяющие хранить и обрабатывать такие объемы данных. Актуальные задачи в области нейрофизиологии требуют применения современных методов анализа данных, включая статистический анализ и машинное обучение, реализованных в распределенных вычислительных инфраструктурах.

Литература

1. BRAIN Initiative. https://braininitiative.nih.gov.
2. Human Brain Project home page. https://www.humanbrainproject. eu.

3. *Elam J. S., Van Essen D.* Human Connectome project // Encyclopedia of computational neuroscience / Eds. D. Jaeger, R. Jung. — New York, NY, USA: Springer, 2013. 4 p.
4. *Brunner C., Blankertz B., Cincotti F., et al.* BNCI Horizon 2020 — towards a roadmap for brain/neural computer interaction // 8th Conference (International) on Universal Access in Human–Computer Interaction Proceedings. — Lecture notes in computer science ser. — Springer, 2014. Vol. 8513. P. 475–486.
5. *Jiang T., Liu Y., Shi F., Shu N., Liu B., Jiang J., Zhou Y.* Multimodal magnetic resonance imaging for brain disorders: Advances and perspectives // Brain Imaging Behav., 2008. Vol. 2. Iss. 4. P. 249–257.
6. *Van Horn J. D., Toga A. W.* Multisite neuroimaging trials // Curr. Opin. Neurol., 2009. Vol. 22. Iss. 4. P. 370–378.
7. *Jack C. R., Bernstein M. A., Fox N. C., et al.* The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods // J. Magn. Reson. Imaging, 2008. Vol. 27. Iss. 4. P. 685–691.
8. *Biswal B. B., Mennes M., Zuo X. N., et al.* Toward discovery science of human brain function // P. Natl. Acad. Sci. USA, 2010. Vol. 107. Iss. 10. P. 4734–4739.
9. *Poldrack R. A., Barch D.,M., Mitchell J., et al.* Toward open sharing of task-based fMRI data: The OpenfMRI project // Front. Neuroinform., 2013. Vol. 7. Art. No. 12. P. 1–12.
10. *Hodge M. R., Horton W., Brown T., et al.* ConnectomeDB-sharing human brain connectivity data // NeuroImage, 2016. Vol. 124. P. 1102–1107.
11. *Marcus D., Olsen T. R., Ramaratnam M., Buckner R. L.* The extensible neuroimaging archive toolkit (XNAT): An informatics platform for managing, exploring, and sharing neuroimaging data // Neuroinformatics, 2007. Vol. 5. P. 11–34.
12. NITRC home page. https://www.nitrc.org.
13. *Neu S. C., Crawford K. L., Toga A. W.* Practical management of heterogeneous neuroimaging metadata by global neuroimaging data repositories // Front. Neuroinform., 2012. Vol. 6. Art. No. 8. P. 1–9.
14. Digital Imaging Communication in Medicine (DICOM): NEMA Standards Publication PS 3. — Washington, DC, USA: National Electrical Manufacturers Association, 1999.
15. ANALYZE 7.5 file format. http://eeg.sourceforge.net/ANALYZE75.pdf.
16. NIFTI home page. http://nifti.nimh.nih.gov.
17. The Brain Imaging Data Structure (BIDS) specification. https://bids.neuroimaging.io/bids_spec.pdf.
18. *Schlögl A.* An overview on data formats for biomedical signals // World Congress on Medical Physics and Biomedical Engineering. — Berlin–Heidelberg: Springer, 2009. P. 1557–1560.
19. *Kemp B., Värri A., Rosa A. C., Nielsen K. D., Gade J.* A simple format for exchange of digitized polygraphic recordings // Electroen. Clin. Neuro., 1992. Vol. 82. Iss. 5. P. 391–393.
20. *Schlögl A.* GDF — a general data format for biomedical signals // arXiv.org, 11 Aug 2006 (v. 1), 26 Mar 2013 (v. 10). arxiv:cs/0608052.
21. *Smith J., Johnson J., Schubert J., Widell R.* A new format for polysomnography data // Sleep, 2005. Vol. 28. Iss. 11. P. 1473–1473.
22. *Fedorov A., Beichel R., Kalpathy-Cramer J., et al.* 3D slicer as an image computing platform for the quantitative imaging network // Magn. Reson. Imaging, 2012. Vol. 30. Iss. 9. P. 1323–1241.
23. *Sadigh-Eteghad S., Majdi A., Farhoudi M., Talebi M., Mahmoudi J.* Different patterns of brain activation in normal aging and Alzheimer's disease from cognitional sight: Meta analysis using activation likelihood estimation // J. Neurol. Sci., 2014. Vol. 343. Iss. 1-2. P. 159–166.
24. *Whitfield-Gabrieli S., Nieto-Castanon A.* Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks // Brain Connectivity, 2012. Vol. 2. Iss. 3. P. 125–141.
25. *Friston K. J., Ashburner J. T., Kiebel S. J., Nichols T. E., Penny W. D.* Statistical parametric mapping: The analysis of functional brain images: The analysis of functional brain images. — Academic Press, 2011. 688 p.
26. *Poil S.* Neurophysiological Biomarkers of cognitive decline: From criticality to toolbox. — Amsterdam: VU University, 2013. 218 p.
27. *Delorme A., Makeig S.* EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics // J. Neurosci. Meth., 2004. Vol. 134. P. 9–21.
28. *Oostenveld R., Fries P., Maris E., Schoffelen J. M.* FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data // Comput. Intell. Neurosc., 2011. Vol. 2011. Art. ID: 156869. P. 1–9.
29. *Vidaurre C., Sander T. H., Schlögl A.* BioSig: The free and open source software library for biomedical signal processing // Comput. Intell. Neurosc., 2011. Vol. 2011. Art. ID: 935364. P. 1–12.
30. *Brett M., Taylor J., Burns C., et al.* NIPY: An open library and development framework for FMRI data analysis // NeuroImage, 2009. Vol. 47. Suppl. 1. P. S196.
31. *Gramfort A., Luessi M., Larson E., et al.* MEG and EEG data analysis with MNE-Python // Front. Neurosci., 2013. Vol. 7. Art. No. 267. P. 1–13.

Поступила в редакцию 14.11.19

NEUROPHYSIOLOGY AS A SUBJECT DOMAIN FOR DATA INTENSIVE PROBLEM SOLVING

D. O. Briukhov, S. A. Stupnikov, D. Yu. Kovalev, and I. A. Shanin

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The goal of this survey is to analyze neurophysiology as a data intensive domain. Nowadays, the number of researches on the human brain is increasing. International projects and researches are aimed at improvement of the understanding of the human brain function. The amount of data obtained in typical laboratories in the field of neurophysiology is growing exponentially. The data are represented using a large number of various formats. This requires creation of infrastructures, databases, and websites that provide unified access to data and support the exchange of data between researchers all over the world. Specific methods and tools forming the field of neuroinformatics (that is, an intersection of neurophysiology and computer science) are used to analyze collected data and to solve neurophysiological problems. These methods include, in particular, statistical analysis, machine learning, and neural networks.

Keywords: neurophysiology; neurophysiological resources; neuroinformatics; data intensive research; analysis of neurophysiological data

DOI: 10.14357/19922264200106

Acknowledgments

This research was partially supported by the Russian Foundation for Basic Research (project 18-29-22096).

References

1. BRAIN Initiative Home Page. Available at: https://braininitiative.nih.gov/ (accessed November 12, 2019)
2. Human Brain Project Home Page. Available at: https://www.humanbrainproject.eu (accessed November 12, 2019).
3. Elam, J. S., and D. Van Essen. 2013. Human Connectome project. *Encyclopedia of computational neuroscience*. Eds. D. Jaeger and R. Jung. New York, NY: Springer. 4 p.
4. Brunner, C., B. Blankertz, F. Cincotti, *et al.* 2014. BNCI Horizon 2020 — towards a roadmap for brain/neural computer interaction. *8th Conference (International) on Universal Access in Human–Computer Interaction Proceedings*. Lecture notes in computer science ser. Springer. 8513: 475–486.
5. Jiang, T., Y. Liu, F. Shi, N. Shu, B. Liu, J. Jiang, and Y. Zhou. 2008. Multimodal magnetic resonance imaging for brain disorders: advances and perspectives. *Brain Imaging Behav.* 2(4):249–257.
6. Van Horn, J. D., and A. W. Toga. 2009. Multisite neuroimaging trials. *Curr. Opin. Neurol.* 22(4):370–378.
7. Jack, C. R., M. A. Bernstein, N. C. Fox, *et al.* 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* 27(4):685–691.
8. Biswal, B. B., M. Mennes, X. N. Zuo, *et al.* 2010. Toward discovery science of human brain function. *P. Natl. Acad. Sci. USA* 107(10):4734–4739.
9. Poldrack, R. A., D. M. Barch, J. Mitchell, *et al.* 2013. Toward open sharing of task-based fMRI data: The OpenfMRI project. *Front. Neuroinform.* 7:12.
10. Hodge, M. R., W. Horton, T. Brown, *et al.* 2016. ConnectomeDB-sharing human brain connectivity data. *NeuroImage* 124:1102–1107.
11. Marcus, D., T. R. Olsen, M. Ramaratnam, and R. L. Buckner. 2007. The extensible neuroimaging archive toolkit (XNAT): An informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5:11–34.
12. NITRC Home Page. Available at: https://www.nitrc.org/ (accessed November 12, 2019).
13. Neu, S. C., K. L. Crawford, and A. W. Toga. 2012. Practical management of heterogeneous neuroimaging metadata by global neuroimaging data repositories. *Front. Neuroinform.* 6:8.
14. Digital Imaging Communication in Medicine (DICOM). 1999. NEMA Standards Publication PS 3. Washington, DC: National Electrical Manufacturers Association.
15. ANALYZE 7.5 file format. Available at: http://eeg.sourceforge.net/ANALYZE75.pdf (accessed November 12, 2019).
16. NIFTI home page. Available at: http://nifti.nimh.nih.gov (accessed November 12, 2019).
17. The Brain Imaging Data Structure (BIDS) specification. Available at: https://bids.neuroimaging.io/bids_spec.pdf (accessed November 12, 2019).

18. Schlögl, A. 2009. An overview on data formats for biomedical signals. *World Congress on Medical Physics and Biomedical Engineering*. Berlin–Heidelberg: Springer. 1557–1560.
19. Kemp, B., A. Värri, A. C. Rosa, K. D. Nielsen, and J. Gade. 1992. A simple format for exchange of digitized polygraphic recordings. *Electroen. Clin. Neuro.* 82(5):391–393.
20. Schlögl, A. GDF — a general data format for biomedical signals Version 2.51. Available at: https://arxiv.org/abs/cs/0608052 (accessed November 12, 2019).
21. Smith, J., J. Johnson, J. Schubert, and R. Widell. 2005. A new file format for polysomnography data. *Sleep* 28(11):1473–1473.
22. Fedorov, A., R. Beichel, J. Kalpathy-Cramer, *et al.* 2012. 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* 30(9):1323–1241.
23. Sadigh-Eteghad, S., A. Majdi, M. Farhoudi, M. Talebi, and J. Mahmoudi. 2014. Different patterns of brain activation in normal aging and Alzheimer's disease from cognitional sight: Meta analysis using activation likelihood estimation. *J. Neurol. Sci.* 343(1-2):159–166.
24. Whitfield-Gabrieli, S., and A. Nieto-Castanon. 2012. Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity* 2(3):125–141.
25. Friston, K. J., J. T. Ashburner, S. J. Kiebel, T. E. Nichols, and W. D. Penny. 2011. *Statistical parametric mapping: The analysis of functional brain images*. Academic Press. 688 p.
26. Poil, S. 2013. *Neurophysiological Biomarkers of cognitive decline: From criticality to toolbox*. Amsterdam: VU University. 218 p.
27. Delorme, A., and S. Makeig. 2004. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics. *J. Neurosci. Meth.* 134:9–21.
28. Oostenveld, R., P. Fries, E. Maris, and J. M. Schoffelen. 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosc.* 2011:156869.
29. Vidaurre, C., T. H. Sander, and A. Schlögl. 2011. BioSig: The free and open source software library for biomedical signal processing. *Comput. Intell. Neurosc.* 2011:935364.
30. Brett, M., J. Taylor, C. Burns, *et al.* 2009. NIPY: An open library and development framework for FMRI data analysis. *NeuroImage* 47:S196.
31. Gramfort, A., M. Luessi, and E. Larson. 2013. EEG data analysis with MNE-Python. *Front. Neurosci.* 7:267.

Received November 14, 2019

Contributors

Briukhov Dmitry O. (b. 1971) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; dbriukhov@ipiran.ru

Stupnikov Sergey A. (b. 1978) — Candidate of Science (PhD) in technology, lead scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sstupnikov@ipiran.ru

Kovalev Dmitry Yu. (b. 1988) — junior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; dkovalev@ipiran.ru

Shanin Ivan A. (b. 1991) — junior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; v08shanin@gmail.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2020. Т. 14. Вып. 1. С. 48–55

РИСК-НЕЙТРАЛЬНАЯ ДИНАМИКА ДЛЯ МОДЕЛИ ARIMA-GARCH С ОШИБКАМИ, РАСПРЕДЕЛЕННЫМИ ПО ЗАКОНУ S_U ДЖОНСОНА

А. Р. Данилишин[1], Д. Ю. Голембиовский[2]

Аннотация: Риск-нейтральный мир выступает одной из фундаментальных моделей финансовой математики, на которой основывается определение справедливой стоимости производных финансовых инструментов. В статье рассматривается построение риск-нейтральной динамики для случайного процесса ARIMA-GARCH (Autoregressive Integrated Moving Average, Generalized AutoRegressive Conditional Heteroskedasticity — интегрированная модель авторегрессии и скользящего среднего, обобщенная авторегрессионная условная гетероскедастичность) с ошибками, распределенными по закону S_U Джонсона. Для нахождения коэффициентов модели, соответствующих риск-нейтральной динамике, в большинстве преобразований (примерами таких преобразований являются преобразование Эшера, расширенный принцип Гирсанова) необходимо существование производящей функции моментов. Для таких распределений, как распределение Стьюдента и S_U Джонсона, данная функция неизвестна. В статье формируется производящая функция моментов для распределения S_U Джонсона и доказывается, что, используя модификацию расширенного принципа Гирсанова, можно получить риск-нейтральную меру относительно выбранного распределения.

Ключевые слова: ARIMA; GARCH; риск-нейтральная мера; расширенный принцип Гирсанова; распределение S_U Джонсона; ценообразование опционов

DOI: 10.14357/19922264200107

1 Введение

Основой для нахождения справедливой стоимости производных финансовых инструментов является моделирование поведения цен базовых активов [1]. Существует множество моделей, однако в случае дискретного времени широкую популярность получила ARIMA-GARCH — модель, позволяющая описывать переменную волатильность случайного процесса [2]. В статье рассматривается модель ARIMA-GARCH с ошибками, распределенными по закону S_U Джонсона. Данное распределение получается нелинейным преобразованием нормального распределения и, в силу свойств данного преобразования, характеризуется асимметричностью и наличием «тяжелых хвостов», что позволяет достаточно хорошо приближать реальные цены базовых активов [3, 4].

На финансовых рынках одним из главных принципов ценообразования финансового инструмента выступает безарбитражность, т. е. отсутствие возможности получения безрисковой прибыли при нулевых затратах [5]. Свойству безарбитражности рынка отвечает существование риск-нейтральной вероятностной меры. Полный рынок характеризуется наличием единственной риск-нейтральной меры, неполный же рынок имеет множество подобных мер.

Главным ограничением многих принципов получения риск-нейтральных мер (преобразование Эшера, расширенный принцип Гирсанова) является наличие производящей функции моментов исследуемого распределения. Производящая функция моментов распределения S_U Джонсона [6, 7] не описана в литературе. Цель статьи — получение данной производящей функции моментов, обобщение расширенного принципа Гирсанова и нахождение риск-нейтральной меры для рассматриваемого случайного процесса. Подробное описание расширенного принципа Гирсанова можно найти в статье [8], а его применение к различным типам моделей GARCH — в работе [9].

Статья построена следующим образом. В разд. 2 введены основные положения модели, описаны принципы построения риск-нейтральной меры на основе расширенного принципа Гирсанова, опре-

[1]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, danilishin-artem@mail.ru

[2]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики; Московский финансово-промышленный университет «Синергия», golemb@cs.msu.su

делены понятия ARIMA, GARCH и распределения S_U Джонсона, приведен вывод производящей функции моментов [10] для распределения S_U Джонсона. В разд. 3 применяется расширенный принцип Гирсанова для распределения S_U Джонсона. В разд. 4 приводится модификация расширенного принципа Гирсанова и получение риск-нейтральных коэффициентов модели ARIMA-GARCH [11]. В заключении статьи приводятся выводы исследования.

2 Модель динамики цены базового актива

Пусть случайный процесс (цена базового актива опциона) $S = (S_t)_{t=0}^T$ задан на вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P})$ с фильтрацией $(\mathcal{F}_t)_{t=0}^T$, где $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_T = \mathcal{F}$. Рассмотрим случайный процесс логарифмических приращений цены базового актива

$$Y_t = \ln\left(\frac{S_t}{S_{t-1}}\right), \; Y_0 = 0, \; t = 1, 2, \ldots, T.$$

Для отсутствия арбитражных возможностей процесс эволюции цены базового актива должен быть согласован с мартингальной мерой $\mathbb{Q}$. Данная мера должна быть эквивалентна исходной (физической) мере $\mathbb{P}$ ($\mathbb{P} \approx \mathbb{Q}$, $\forall\, B \in \mathcal{F}$: $\mathbb{Q}(B) = 0 \Leftrightarrow$ $\Leftrightarrow \mathbb{P}(B) = 0$). Мера $\mathbb{Q}$ является мартингальной, если справедливо равенство

$$\mathbb{E}^{\mathbb{Q}}\left[\tilde{S}_t | \mathcal{F}_{t-1}\right] = \tilde{S}_{t-1}.$$

Здесь $\mathbb{E}^{\mathbb{Q}}[*|*]$ — условное математическое ожидание относительно меры $\mathbb{Q}$; $\tilde{S}_t$ — дисконтированные цены базового актива относительно безрисковой ставки r ($\tilde{S}_t = e^{-rt} S_t$). Условие мартингальности можно переписать в следующем виде:

$$\mathbb{E}^{\mathbb{Q}}\left[e^{Y_t} | \mathcal{F}_{t-1}\right] = e^r.$$

Расширенный принцип Гирсанова описывает динамику дисконтированных цен базового актива следующим образом [8, 9]:

$$\tilde{S}_t = \tilde{S}_0 A_t M_t, \; A_t = \prod_{k=1}^{t} \mathbb{E}^{\mathbb{P}}\left[\frac{\tilde{S}_k}{\tilde{S}_{k-1}} | \mathcal{F}_{t-1}\right], \qquad (1)$$

где процесс M_t является мартингалом [12]. Поделив левую и правую части выражения (1) на $\tilde{S}_{t-1}$ и сделав соответствующие преобразования, получим:

$$\tilde{S}_t = \tilde{S}_{t-1} e^{-r + \ln\left(\mathbb{E}^{\mathbb{P}}\left[e^{Y_t} | \mathcal{F}_{t-1}\right]\right)} W_t = \tilde{S}_{t-1} e^{v_t} W_t,$$

$$W_t = \frac{M_t}{M_{t-1}}.$$

В расширенном принципе Гирсанова утверждается, что процесс

$$Z_t = \prod_{k=1}^{t} \frac{g^{\mathbb{P}}_{W_k}\left(\tilde{S}_k / \tilde{S}_{k-1}\right) e^{v_k}}{g^{\mathbb{P}}_{W_k}\left(e^{-v_k} \tilde{S}_k / \tilde{S}_{k-1}\right)},$$

порождаемый производной Радона–Никодима $d\mathbb{Q}/d\mathbb{P}$ [13, 14], где $g^{\mathbb{P}}_{W_t}$ — условная плотность распределения W_t, обеспечивает риск-нейтральную динамику для $\tilde{S}_t$ в новой мере $\mathbb{Q}$, относительно старой $\mathbb{P}$ [8]. Данное утверждение можно записать следующим образом:

$$\mathcal{L}^{\mathbb{Q}}\left(\tilde{S}_t | \mathcal{F}_{t-1}\right) = \mathcal{L}^{\mathbb{P}}\left(M_t | \mathcal{F}_{t-1}\right).$$

Для процесса Y_t с плотностью распределения $f^{\mathbb{P}}_{Y_t}$ выражение для Z_t примет следующий вид [9]:

$$Z_t = \prod_{k=1}^{t} \frac{f^{\mathbb{P}}_{Y_k}\left(Y_k - r + \ln(M_{Y_k|\mathcal{F}_{t-1}}(1))\right)}{f^{\mathbb{P}}_{Y_t}(Y_k)}.$$

Переход от физической меры к риск-нейтральной осуществляется с помощью производящей функции моментов [8]:

$$M^{\mathbb{Q}}_{Y_t}(c) = e^{-c(-r + \ln(M_{Y_t}(1)))} M^{\mathbb{P}}_{Y_t}(c). \qquad (2)$$

Модель ARIMA (p, d, q)-GARCH (P, Q) является комбинацией моделей ARIMA (p, d, q) — интегрированной модели авторегрессии и скользящего среднего — и GARCH (P, Q) — обобщенной авторегрессионной условной гетероскедастичности [7, 9]:

$$\Delta^d y_t = m_t + \delta_t \varepsilon_t, \quad \varepsilon_t|_{\mathcal{F}_{t-1}} \sim \text{JSU}(\xi, \lambda, \gamma, \delta);$$

$$m_t = \mathbb{E}\left[\Delta^d y_t | \mathcal{F}_{t-1}\right] = \phi_0 + \phi_1 \Delta^d y_{t-1} + \cdots$$

$$\cdots + \phi_p \Delta^d y_{t-p} + \theta_1 \delta_{t-1} \varepsilon_{t-1} + \cdots + \theta_q \delta_{t-q} \varepsilon_{t-q};$$

$$\delta_t^2 = \text{Var}\left[\Delta^d y_t | \mathcal{F}_{t-1}\right] = \alpha_0 + \alpha_1 \delta_{t-1}^2 + \cdots$$

$$\cdots + \alpha_P \delta_{t-P}^2 + \beta_1 \delta_{t-1}^2 \varepsilon_{t-1}^2 + \cdots + \beta_Q \delta_{t-Q}^2 \varepsilon_{t-Q}^2.$$

Будем рассматривать модель ARIMA (p, d, q)-GARCH (P, Q) с ошибками ε_t, имеющими распределение S_U Джонсона. Распределение S_U Джонсона — JSU $(\xi, \lambda, \gamma, \delta)$ — представляет собой четырехпараметрическое вероятностное распределение, которое образуется в результате нелинейного преобразования нормально распределенной случайной величины $X \sim N(0, 1)$:

$$\varepsilon_t = \xi + \lambda \sinh\left(\frac{X_t - \gamma}{\delta}\right) = g(X_t),$$

где $-\infty < \xi < \infty$ — параметр сдвига местоположения; $0 < \lambda < \infty$ — параметр масштабирования; $-\infty < \gamma < \infty$ — параметр асимметрии; $0 < \delta < \infty$ — показатель эксцесса. Функция плотности распределения имеет следующий вид:

$$f_{\varepsilon_t}(\varepsilon) = \frac{\delta}{\lambda\sqrt{2\pi}} \times \\ \times \frac{1}{\sqrt{1+((\varepsilon-\xi)/\lambda)^2}}\, e^{-(\gamma+\delta\sinh^{-1}((\varepsilon-\xi)/\lambda))^2/2}. \quad (3)$$

Математическое ожидание и дисперсия ε_t должны равняться 0 и 1 соответственно:

$$\mathbb{E}[\varepsilon_t] = \xi - \lambda e^{1/(2\delta^2)}\sinh\left(\frac{\gamma}{\delta}\right) = 0\,;$$

$$\mathrm{Var}[\varepsilon_t] = \frac{\lambda^2}{2}\left(e^{1/\delta^2}-1\right)\left(e^{1/\delta^2}\cosh\left(\frac{2\gamma}{\delta}\right)+1\right) = 1\,.$$

Введем соответствующие замены переменных:

$$\tilde{\xi} = \tilde{\lambda} e^{1/(2\delta^2)}\sinh\left(\frac{\gamma}{\delta}\right);$$

$$\tilde{\lambda} = \sqrt{2}\left(\left(e^{1/\delta^2}-1\right)\left(e^{1/\delta^2}\cosh\left(\frac{2\gamma}{\delta}\right)+1\right)\right)^{-1/2},$$

тогда ошибки будут иметь распределение с уже новыми параметрами:

$$\varepsilon_t|_{\mathcal{F}_{t-1}} \sim \mathrm{JSU}\left(\tilde{\xi},\tilde{\lambda},\gamma,\delta\right).$$

Утверждение 1. *Производящая функция моментов для распределения S_U Джонсона имеет следующий вид*:

$$M_Y(c) = e^{\xi c}\sum_{n=0}^{\infty}\left(\frac{c\lambda}{2}\right)^n\frac{1}{n!}\times \\ \times\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{(n-2j)^2/(2\delta^2)+\gamma(n-2j)/\delta}.$$

Доказательство.

$$\frac{Y-\xi}{\lambda} = \sinh\frac{X-\gamma}{\delta} = \\ = \frac{1}{2}\left(e^{(X-\gamma)/\delta} - e^{-(X-\gamma)/\delta}\right) \Rightarrow \mathbb{E}\left[\left(\frac{Y-\xi}{\lambda}\right)^n\right] = \\ = \frac{1}{2^n}\int\limits_{-\infty}^{\infty}\left(e^{(x-\gamma)/\delta} - e^{-(x-\gamma)/\delta}\right)^n f_{\mathrm{norm}}(x)\,dx = \\ = \frac{1}{2^n}\int\limits_{-\infty}^{\infty}\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{-(x-\gamma)(n-2j)/\delta} f_{\mathrm{norm}}(x)\,dx = \\ = \frac{1}{2^n}\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{\gamma(n-2j)/\delta}\times \\ \times\int\limits_{-\infty}^{\infty} e^{-x(n-2j)/\delta} f_{\mathrm{norm}}(x)\,dx = \\ = \frac{1}{2^n}\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{(n-2j)^2/(2\delta^2)+\gamma(n-2j)/\delta}.$$

Свойство

$$M_{(Y-\xi)/\lambda}(c) = e^{-\xi c/\lambda}M_y\left(\frac{c}{\lambda}\right)$$

производящей функции моментов завершает доказательство.

3 Риск-нейтральная динамика на основе расширенного принципа Гирсанова

Зная распределение ошибок $\varepsilon_t|_{\mathcal{F}_{t-1}} \sim \mathrm{JSU}(\tilde{\xi},\tilde{\lambda},\gamma,\delta)$ (3), можно найти распределение случайного процесса Y_t:

$$f_{Y_t}(y_t) = \frac{\delta}{\tilde{\lambda}\delta_t\sqrt{2\pi}}\,\frac{1}{\sqrt{1+((y_t-(m_t+\tilde{\xi}\delta_t))/(\delta_t\tilde{\lambda}))^2}}\times \\ \times e^{-\left(\gamma+\delta\sinh^{-1}\left((y_t-(m_t+\tilde{\xi}\delta_t))/(\delta_t\tilde{\lambda})\right)\right)^2/2}. \quad (4)$$

Тогда, сравнивая плотности (3) и (4), получаем, что $Y_t|_{\mathcal{F}_{t-1}}$ имеет распределение $\mathrm{JSU}(m_t+\tilde{\xi}\delta_t, \tilde{\lambda}\delta_t, \gamma, \delta)$. Производящая функция моментов для случайного процесса Y_t будет иметь следующий вид:

$$M_{Y_t}^{\mathbb{P}}(c) = e^{(m_t+\delta_t\tilde{\xi})c}\sum_{n=0}^{\infty}\left(\frac{c\tilde{\lambda}\delta_t}{2}\right)^n\frac{1}{n!}A_n\,, \quad (5)$$

где

$$A_n = \sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{(n-2j)^2/(2\delta^2)+\gamma(n-2j)/\delta}.$$

Пользуясь выражением (2), найдем производящую функцию моментов относительно риск-нейтральной меры $\mathbb{Q}$:

$$M_{Y_t}^{\mathbb{Q}}(c) = e^{-c(-r+\ln M_{Y_t}^{\mathbb{P}}(1))}M_{Y_t}^{\mathbb{P}}(c) = \\ = e^{c(m_t+\delta_t\tilde{\xi}+r-\ln M_{Y_t}^{\mathbb{P}}(1))}\sum_{n=0}^{\infty}\left(\frac{c\tilde{\lambda}\delta_t}{2}\right)^n\frac{1}{n!}A_n = \\ = e^{c\left(r-\ln\left(\sum_{n=0}^{\infty}\left(\tilde{\lambda}\delta_t/2\right)^n A_n/n!\right)\right)}\sum_{n=0}^{\infty}\left(\frac{c\tilde{\lambda}\delta_t}{2}\right)^n\frac{1}{n!}A_n. \quad (6)$$

Сравнивая выражения (5) и (6), приходим к выводу, что в новой мере $\mathbb{Q}$ процесс $Y_t|_{\mathcal{F}_{t-1}}$ имеет распределение JSU $\left(r-\ln\left(\sum_{n=0}^{\infty}(\tilde{\lambda}\delta_t/2)^n A_n/n!\right),\tilde{\lambda}\delta_t,\gamma,\delta\right)$. Модель ARIMA-GARCH примет следующий вид:

$$Y_t = r - \ln\left(\sum_{n=0}^{\infty}\left(\frac{\tilde{\lambda}\delta_t}{2}\right)^n \times \right.$$
$$\left. \times \frac{1}{n!}\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{(n-2j)^2/(2\delta^2)+\gamma(n-2j)/\delta}\right) -$$
$$- \tilde{\lambda}\delta_t e^{1/(2\delta^2)}\sinh\left(\frac{\gamma}{\delta}\right) + \delta_t\varepsilon_t\,,$$
$$\varepsilon_{t|\mathcal{F}_{t-1}} \sim \text{JSU}\left(\tilde{\xi},\tilde{\lambda},\gamma,\delta\right).$$

Исследуем вопрос о существовании условного математического ожидания в новой метрике. Заметим, что для существования первого момента необходимо, чтобы сходился следующий ряд:

$$\sum_{n=0}^{\infty}\left(\frac{\tilde{\lambda}\delta_t}{2}\right)^n\frac{1}{n!}\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{(n-2j)^2/(2\delta^2)+\gamma(n-2j)/\delta}. \quad (7)$$

Ряд (7) является степенным рядом и имеет вид $\sum_{n=0}^{\infty}a_n X^n$, где

$$X = \frac{\tilde{\lambda}\delta_t}{2};$$
$$a_n = \frac{1}{n!}\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{(n-2j)^2/(2\delta^2)+\gamma(n-2j)/\delta}.$$

Найдем радиус сходимости данного степенного ряда:

$$R = \lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right| = \lim_{n\to\infty}(n+1)\times$$
$$\times\left|\frac{\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{\frac{(n-2j)^2}{2\delta^2}+\frac{\gamma(n-2j)}{\delta}}}{\sum_{j=0}^{n+1}(-1)^{n+1-j}C_{n+1}^j e^{\frac{(n+1-2j)^2}{2\delta^2}+\frac{\gamma(n+1-2j)}{\delta}}}\right| =$$
$$= \lim_{n\to\infty}(n+1)\frac{e^{n^2/(2\delta^2)+\gamma n/\delta}}{e^{(n+1)^2/(2\delta^2)+\gamma(n+1)/\delta}}\times$$
$$\times\left|\frac{\sum_{j=0}^{n}(-1)^{n-j}C_n^j e^{-\frac{2jn}{\delta^2}+\frac{2j^2}{\delta^2}\delta^2-\frac{2j\gamma}{\delta}}}{\sum_{j=0}^{n+1}(-1)^{n+1-j}C_{n+1}^j e^{-\frac{2j(n+1)}{\delta^2}+\frac{2j^2}{\delta^2}-\frac{2j\gamma}{\delta}}}\right| =$$
$$= \lim_{n\to\infty}(n+1)\frac{e^{n^2/(2\delta^2)+\gamma n/\delta}}{e^{(n+1)^2/(2\delta^2)+\gamma(n+1)/\delta}} =$$
$$= \lim_{n\to\infty}\frac{n+1}{e^{n/\delta^2+1/(2\delta^2)+\gamma/\delta}} = 0\,.$$

Здесь использовано то обстоятельство, что

$$0 \le \lim_{n\to\infty} C_n^j e^{-2jn/\delta^2+2j^2/\delta^2-2j\gamma/\delta} \le$$
$$\le \lim_{n\to\infty}\frac{n^j}{j!}e^{-2jn/\delta^2+2j^2/\delta^2-2j\gamma/\delta} = 0\,.$$

Таким образом, ряд сходится только в нуле, а ввиду того, что $\tilde{\lambda}\neq 0$ и $\delta_t \not\equiv 0$, условное математическое ожидание в новой метрике будет отсутствовать.

4 Модификация расширенного принципа Гирсанова

В расширенном принципе Гирсанова рассматриваются логарифмы отношения цен базового актива $Y_t = \ln(S_t/S_{t-1})$. Предлагается рассматривать относительные цены $\tilde{Y}_t = S_t/S_{t-1} - 1$. При таком предположении динамика базового актива будет иметь следующий вид:

$$\tilde{S}_t = \tilde{S}_{t-1}(1+\mu_t)W_t\,,$$

где

$$\mu_t = \frac{\mathbb{E}^{\mathbb{P}}\left[\tilde{Y}_t+1|\mathcal{F}_{t-1}\right]}{(1+r/n)^n} - 1\,;$$

n — количество начислений риск-нейтральной ставки за год.

Теорема 1. *Процесс*

$$Z_t = \prod_{k=1}^{t}\frac{g^{\mathbb{P}}_{W_k}\left(\tilde{S}_k/\tilde{S}_{k-1}\right)(1+\mu_k)}{g^{\mathbb{P}}_{W_k}((1+\mu_k)^{-1}\tilde{S}_k/\tilde{S}_{k-1})}$$

обеспечивает риск-нейтральную динамику для $\tilde{S}_t$ в новой мере $\mathbb{Q}$ относительно старой $\mathbb{P}$:

$$\mathcal{L}^{\mathbb{Q}}\left(\tilde{S}_t|\mathcal{F}_{t-1}\right) = \mathcal{L}^{\mathbb{P}}\left(M_t|\mathcal{F}_{t-1}\right).$$

Доказательство.

$$Z_t = Z_{t-1}\frac{g^{\mathbb{P}}_{W_k}(\tilde{S}_t/\tilde{S}_{t-1})(1+\mu_t)}{g^{\mathbb{P}}_{W_k}((1+\mu_t)^{-1}\tilde{S}_t/\tilde{S}_{t-1})} =$$
$$= Z_{t-1}\frac{g^{\mathbb{P}}_{W_k}((1+\mu_t)W_t)(1+\mu_t)}{g^{\mathbb{P}}_{W_k}(W_t)};$$

$$\mathbb{E}^{\mathbb{P}}\left[Z_t|\mathcal{F}_{t-1}\right] =$$
$$= Z_{t-1}\int_{-\infty}^{\infty} g^{\mathbb{P}}_{W_k}\left((1+\mu_t)\,w_t\right)(1+\mu_t)\,dw_t = Z_{t-1}\,.$$

Необходимо показать, что закон распределения дисконтированной цены базового актива $\tilde{S}_t$ для меры $\mathbb{Q}$ совпадает с законом распределения случайного процесса M_t для меры $\mathbb{P}$. Обозначим условную

плотность распределения случайного процесса M_t для меры $\mathbb{P}$ как $\rho_t(M_t)$:

$$\rho_t(M_t) = P\left(M_t < a\right)'_{a=M_t} = \\ = P\left(\frac{M_t}{M_{t-1}} < \frac{a}{M_{t-1}}\right)'_{a=M_t} = \\ = P\left(W_t < \frac{a}{M_{t-1}}\right)'_{a=M_t} = \frac{g^{\mathbb{P}}_{W_k}(M_t/M_{t-1})}{M_{t-1}}.$$

Введем обозначение

$$\tilde{W}_t = (1+\mu_t)\, W_t,$$

тогда структура уравнений, описывающих динамики M_t и $\tilde{S}_t$, будет совпадать:

$$M_t = M_{t-1} W_t\,; \;\; \tilde{S}_t = \tilde{S}_{t-1}\tilde{W}_t\,.$$

Далее обозначим условную плотность случайного процесса $\tilde{S}_t$ по метрике $\mathbb{Q}$ как $\tilde{\rho}_t(\tilde{S}_t)$:

$$\tilde{\rho}_t(\tilde{S}_t) = Q\left(\tilde{S}_t < a\right)'_{a=\tilde{S}_t} = \\ = Q\left(\frac{\tilde{S}_t}{\tilde{S}_{t-1}} < \frac{a}{\tilde{S}_{t-1}}\right)'_{a=\tilde{S}_t} = Q\left(\tilde{W}_t < \frac{a}{\tilde{S}_{t-1}}\right)'_{a=\tilde{S}_t} = \\ = \frac{\tilde{g}^{\mathbb{Q}}_{\tilde{W}_t}(\tilde{S}_t/\tilde{S}_{t-1})}{\tilde{S}_{t-1}},$$

где $\tilde{g}^{\mathbb{Q}}_{\tilde{W}_t}$ — условная плотность распределения случайного процесса $\tilde{W}_t$ по мере $\mathbb{Q}$.

Осталось показать, что закон распределения случайного процесса W_t совпадает с законом распределения случайного процесса $\tilde{W}_t$. Определим функцию распределения $\tilde{G}^{\mathbb{Q}}_{\tilde{W}_t}$ для $\tilde{W}_t$:

$$\tilde{G}^{\mathbb{Q}}_{\tilde{W}_t}(a) = \frac{\mathbb{E}^{\mathbb{P}}\left[Z_t I_{\{\tilde{W}_t<a\}}|\mathcal{F}_{t-1}\right]}{\mathbb{E}^{\mathbb{P}}\left[Z_t|\mathcal{F}_{t-1}\right]} = \\ = \int\limits_{-\infty}^{\infty} g^{\mathbb{P}}_{W_t}\left((1+\mu_t)\, w_t\right)(1+\mu_t) I_{\{\tilde{W}_t<a\}} dw_t = \\ = \int\limits_{-\infty}^{\infty} g^{\mathbb{P}}_{W_t}(\tilde{w}_t) I_{\{\tilde{W}_t<a\}} d\tilde{w}_t = \\ = \int\limits_{-\infty}^{\infty} g^{\mathbb{P}}_{W_t}(\tilde{w}_t)\, d\tilde{w}_{t\,\tilde{W}_t<a} \Rightarrow \tilde{g}^{\mathbb{Q}}_{\tilde{W}_t} = g^{\mathbb{P}}_{W_t}.$$

Теорема 2.

$$\prod_{k=1}^{t} \frac{g^{\mathbb{P}}_{W_k}(\tilde{S}_k/\tilde{S}_{k-1})(1+\mu_k)}{g^{\mathbb{P}}_{W_k}((1+\mu_k)^{-1}\tilde{S}_k/\tilde{S}_{k-1})} = \\ = \prod_{k=1}^{t} \frac{f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{Y}_k(1+\mu_k)+\mu_k)(1+\mu_k)}{f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{Y}_k)}.$$

Доказательство.

$$g^{\mathbb{P}}_{W_k}\left(\frac{\tilde{S}_k}{\tilde{S}_{k-1}}\right) = P\left(W_k < a\right)'_{a=\tilde{S}_k/\tilde{S}_{k-1}} = \\ P\left(\frac{S_k}{S_{k-1}} - 1 < a(1+\mu_k)\left(1+\frac{r}{n}\right)^n - 1\right)'_{a=\tilde{S}_k/\tilde{S}_{k-1}} = \\ = f^{\mathbb{P}}_{\tilde{Y}_k}\left(\frac{\tilde{S}_k}{\tilde{S}_{k-1}}(1+\mu_k)\left(1+\frac{r}{n}\right)^n - 1\right) \times \\ \times (1+\mu_k)\left(1+\frac{r}{n}\right)^n = \\ = f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{Y}_k(1+\mu_k)+\mu_k)(1+\mu_k)\left(1+\frac{r}{n}\right)^n;$$

$$g^{\mathbb{P}}_{W_k}\left((1+\mu_k)^{-1}\frac{\tilde{S}_k}{\tilde{S}_{k-1}}\right) = \\ = P\left(W_k < a\right)'_{a=(1+\mu_k)^{-1}\tilde{S}_k/\tilde{S}_{k-1}} = P\left(\frac{S_k}{S_{k-1}} - 1 < \right. \\ \left. < a(1+\mu_k)\left(1+\frac{r}{n}\right)^n - 1\right)'_{a=(1+\mu_k)^{-1}\tilde{S}_k/\tilde{S}_{k-1}} = \\ = f^{\mathbb{P}}_{\tilde{Y}_k}\left((1+\mu_k)^{-1}\frac{\tilde{S}_k}{\tilde{S}_{k-1}}(1+\mu_k)\left(1+\frac{r}{n}\right)^n - 1\right) \times \\ \times (1+\mu_k)\left(1+\frac{r}{n}\right)^n = f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{Y}_k)(1+\mu_k)(1+r);$$

$$\frac{g^{\mathbb{P}}_{W_k}(\tilde{S}_k/\tilde{S}_{k-1})(1+\mu_k)}{g^{\mathbb{P}}_{W_k}((1+\mu_k)^{-1}\tilde{S}_k/\tilde{S}_{k-1})} = \\ = \frac{f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{Y}_k(1+\mu_k)+\mu_k)(1+\mu_k)(1+r/n)^n(1+\mu_k)}{f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{Y}_k)(1+\mu_k)(1+r/n)^n} = \\ = \frac{f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{Y}_k(1+\mu_k)+\mu_k)(1+\mu_k)}{f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{Y}_k)}.$$

Воспользовавшись теоремами 1 и 2, найдем производящую функцию моментов в новой метрике.

Утверждение 2.

$$M^{\mathbb{Q}}_{\tilde{Y}_t}(c) = e^{-\mu_t c/(1+\mu_t)} M^{\mathbb{P}}_{\tilde{Y}_t}\left(\frac{c}{1+\mu_t}\right). \qquad (8)$$

Доказательство.

$$M^{\mathbb{Q}}_{\tilde{Y}_t}(c) = \mathbb{E}^{\mathbb{Q}}\left[e^{\tilde{Y}_t c}|\mathcal{F}_{t-1}\right] = \mathbb{E}^{\mathbb{P}} \times \\ \times\left[e^{\tilde{Y}_t c}\frac{f^{\mathbb{P}}_{\tilde{Y}_t} c(\tilde{Y}_t(1+\mu_t)+\mu_t)(1+\mu_t)}{f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{Y}_t)} Z_{t-1}\middle|\mathcal{F}_{t-1}\right] = \\ = \mathbb{E}^{\mathbb{P}}\left[e^{\tilde{Y}_t c}\frac{f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{Y}_t(1+\mu_t)+\mu_t)(1+\mu_t)}{f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{Y}_t)}\middle|\mathcal{F}_{t-1}\right] \times \\ \times \mathbb{E}^{\mathbb{P}}\left[Z_{t-1}|\mathcal{F}_{t-1}\right];$$

$$\mathbb{E}^{\mathbb{P}}\left[Z_{t-1}|\mathcal{F}_{t-1}\right] = \mathbb{E}^{\mathbb{P}}\left[Z_{t-1}\right] =$$
$$= \int\limits_{-\infty}^{\infty}\cdots\int\limits_{-\infty}^{\infty}\prod_{k=1}^{t-1}\frac{f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{y}_k(1+\mu_k)+\mu_k)(1+\mu_k)}{f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{y}_k)}\times$$
$$\times f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{y}_k)\,d\tilde{y}_k =$$
$$= \int\limits_{-\infty}^{\infty}\cdots\int\limits_{-\infty}^{\infty}\prod_{k=1}^{t-1} f^{\mathbb{P}}_{\tilde{Y}_k}(\tilde{y}_k(1+\mu_k)+\mu_k)(1+\mu_k)\,d\tilde{y}_k\,.$$

Введем замену переменной:

$$\tilde{y}_k(1+\mu_k)+\mu_k = u_k \Rightarrow d\tilde{y}_k = \frac{du_k}{1+\mu_k}\,.$$

Тогда

$$\int\limits_{-\infty}^{\infty}\cdots\int\limits_{-\infty}^{\infty}\prod_{k=1}^{t-1} f^{\mathbb{P}}_{\tilde{Y}_k}\left(\tilde{y}_k(1+\mu_k)+\mu_k\right)(1+\mu_k)\,d\tilde{y}_k =$$
$$= \int\limits_{-\infty}^{\infty}\cdots\int\limits_{-\infty}^{\infty}\prod_{k=1}^{t-1} f^{\mathbb{P}}_{\tilde{Y}_k}(u_k)\,du_k = 1\,;$$

$$\mathbb{E}^{\mathbb{P}}\left[e^{\tilde{Y}_t c}\frac{f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{Y}_t(1+\mu_t)+\mu_t)(1+\mu_t)}{f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{Y}_t)}\middle|\,\mathcal{F}_{t-1}\right] =$$
$$= \int\limits_{-\infty}^{\infty} e^{\tilde{y}_t c}\frac{f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{y}_t(1+\mu_t)+\mu_t)(1+\mu_t)}{f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{y}_t)}\,f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{y}_t)\,d\tilde{y}_t =$$
$$= \int\limits_{-\infty}^{\infty} e^{\tilde{y}_t c} f^{\mathbb{P}}_{\tilde{Y}_t}(\tilde{y}_t(1+\mu_t)+\mu_t)(1+\mu_t)\,d\tilde{y}_t =$$
$$= \int\limits_{-\infty}^{\infty} e^{c(u_t-\mu_t)/(1+\mu_t)} f^{\mathbb{P}}_{\tilde{Y}_t}(u_t)\,du_t =$$
$$= e^{-\mu_t c/(1+\mu_t)}\int\limits_{-\infty}^{\infty} e^{u_t c/(1+\mu_t)} f^{\mathbb{P}}_{\tilde{Y}_t}(u_t)\,du_t =$$
$$= e^{-\mu_t c/(1+\mu_t)} M^{\mathbb{P}}_{\tilde{Y}_t}\left(\frac{c}{1+\mu_t}\right).$$

Применим формулу (8) к производящей функции моментов:

$$M^{\mathbb{Q}}_{\tilde{Y}_t}(c) = e^{-\mu_t c/(1+\mu_t)}\times$$
$$\times e^{c(m_t+\delta_t\tilde{\xi})/(1+\mu_t)}\sum_{n=0}^{\infty}\left(\frac{c\tilde{\lambda}\delta_t}{2(1+\mu_t)}\right)^n\frac{1}{n!}A_n =$$
$$= e^{c(m_t+\delta_t\tilde{\xi}-\mu_t)/(1+\mu_t)}\sum_{n=0}^{\infty}\left(\frac{c\tilde{\lambda}\delta_t}{2(1+\mu_t)}\right)^n\frac{1}{n!}A_n\,.$$

В итоге получаем, что процесс $\tilde{Y}_t|_{\mathcal{F}_{t-1}}$ относительно метрики $\mathbb{Q}$ имеет распределение $\mathrm{JSU}\left((m_t+\delta_t\xi-\right.$ $\left.-\mu_t)/(1+\mu_t), \tilde{\lambda}\delta_t/(1+\mu_t), \gamma, \delta\right)$, риск-нейтральный ARIMA-GARCH-процесс динамики базового актива $\tilde{Y}_t = S_t/S_{t-1}-1$ описывается уравнением:

$$\tilde{Y}_t = \left(1+\frac{r}{n}\right)^n - 1 + \delta_t\frac{(1+r/n)^n}{1+m_t}\,\varepsilon_t\,,$$
$$\varepsilon_{t|\mathcal{F}_{t-1}} \sim \mathrm{JSU}\left(\tilde{\xi}, \tilde{\lambda}, \gamma, \delta\right).$$

5 Заключение

В статье рассмотрена задача получения риск-нейтрального преобразования динамики доходности базового актива производных финансовых инструментов. Для аппроксимации динамики была использована модель ARIMA-GARCH с ошибками, распределенными по закону S_U Джонсона. Распределение S_U Джонсона на данный момент часто используется при моделировании временны́х рядов базовых активов для разного рода производных финансовых инструментов [11]. Главное преимущество данного распределения заключается в возможности моделировать временны́е ряды с «тяжелыми хвостами». Однако для данного распределения неизвестна производящая функция моментов, что делало невозможным получение соответствующей риск-нейтральной динамики методами, которые используют данную функцию.

Первым результатом статьи является получение производящей функции моментов в виде степенного ряда, что дает возможность, используя расширенный принцип Гирсанова, получить коэффициенты, обеспечивающие риск-нейтральную динамику для модели ARIMA-GARCH. Однако условное математическое ожидание данного случайного процесса не существует, так как соответствующий ряд не сходится.

Вторым результатом работы является переход от логарифма отношения цен базового актива непосредственно к доходности. Это дает возможность не вычислять значение производящей функции моментов в конкретной точке, а, используя ее общий вид, получать распределение случайного процесса в новой (риск-нейтральной) метрике. Полученное риск-нейтральное распределение используется для построения окончательного вида риск-нейтрального ARIMA-GARCH-процесса динамики стоимости базового актива.

Литература

1. *Hull J.* Options, futures, and other derivatives. — 10th ed. — Pearson, 2018. 896 p.

2. *Patton A.* Quantitative finance. — London: University of London Press Publisher, 2015. 65 p.
3. *Akgiray V.* Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts // J. Bus., 1989. Vol. 62. Iss. 1. P. 55–80. doi: 10.1086/296451.
4. *Teräsvirta T.* An introduction to univariate GARCH models // Handbook of financial time series / Eds. T. G. Andersen, R. A. Davis, J.-P. Kreiss, Th. V. Mikosch. — Berlin–Heidelberg: Springer, 2009. Vol. 10. P. 17–42. doi: 10.1007/978-3-540-71297-8_1.
5. *Follmer H., Schied A.* Stochastic finance: An introduction in discrete time. — Berlin: Walter de Gruyter, 2002. 422 p.
6. *Bollerslev T.* A conditionally heteroskedastic time series model for speculative prices and rates of return // Rev. Econ. Stat., 1987. Vol. 69. Iss. 3. P. 542–547. doi: 10.2307/1925546.
7. *Simonato J. G.* GARCH processes with skewed and leptokurtic innovations: Revisiting the Johnson S_U case. May 16, 2012. https://ssrn.com/abstract=2060994.
8. *Elliott R. J., Madan D. B.* A discrete time equivalent martingale measure // Math. Financ., 1998. Vol. 8. Iss. 2. P. 127–152. doi: 10.1111/1467-9965.00048.
9. *Yi Xi.* Comparison of option pricing between ARMA-GARCH and GARCH-M models. — London, Ontario, Canada: University of Western Ontario, 2013. MoS Thesis. 73 p.
10. *Enrique R., Escobar L.* Using moment generating functions to derive mixture distributions // Am. Stat., 2006. Vol. 60. Iss. 1. P. 75–80. doi: 10.1198/000313006X90819.
11. *Simonato J. G., Stentoft L.* Which pricing approach for options under GARCH with non-normal innovations? July 2015. https://www.degroote.mcmaster.ca/files/2015/11/SimonatoStentoft.pdf.
12. *Williams D.* Probability with martingales. — Cambridge: Cambridge University Press, 1991. 251 p.
13. *Cameron R. H., Martin W. T.* Transformation of Wiener integrals under a general class of linear transformations translations // T. Am. Math. Soc., 1945. Vol. 58. P. 184–219. doi: 10.1090/S0002-9947-1945-0013240-1.
14. *Bell D.* Transformations of measures on an infinite-dimensional vector space // Seminar on stochastic processes, 1990 / Eds. E. Çinlar, P. J. Fitzsimmons, R. J. Williams. — Progress in probability book ser. — Birkhäuser Boston, 1991. Vol. 24. P. 15–25. doi: 10.1007/978-1-4684-0562-0_3.

Поступила в редакцию 23.06.19

RISK-NEUTRAL DYNAMICS FOR THE ARIMA-GARCH RANDOM PROCESS WITH ERRORS DISTRIBUTED ACCORDING TO THE JOHNSON'S S_U LAW

A. R. Danilishin[1] and D. Yu. Golembiovsky[1,2]

[1]Department of Operations Research, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, Moscow 119991, GSP-1, Russian Federation
[2]Department of Banking, Sinergy University, 80-G Leningradskiy Prospect, Moscow 125190, Russian Federation

Abstract: Risk-neutral world is one of the fundamental principles of financial mathematics, for definition of a fair value of derivative financial instruments. The article deals with the construction of risk-neutral dynamics for the ARIMA-GARCH (Autoregressive Integrated Moving Average, Generalized AutoRegressive Conditional Heteroskedasticity) random process with errors distributed according to the Johnson's S_U law. Methods for finding risk-neutral coefficients require the existence of a generating function of moments (examples of such transformations are the Escher transformation, the extended Girsanov principle). A generating function of moments is not known for Student and Johnson's S_U distributions. The authors form a generating function of moments for the Johnson's S_U distribution and prove that a modification of the extended Girsanov principle may obtain a risk-neutral measure with respect to the chosen distribution.

Keywords: ARIMA; GARCH; risk-neutral measure; Girsanov extended principle; Johnson's S_U; option pricing

DOI: 10.14357/19922264200107

References

1. Hull, J. 2018. *Options, futures, and other derivatives*. 10th ed. Pearson. 896 p.
2. Patton, A. 2015. *Quantitative finance*. London: University of London Press Publisher. 65 p.
3. Akgiray, V. 1989. Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts. *J. Bus.* 62(1):55–80. doi: 10.1086/296451.
4. Teräsvirta, T. 2009. An introduction to univariate GARCH models. *Handbook of financial time series*. Eds. T. G. Andersen, R. A. Davis, J.-P. Kreiss, and

Th. V. Mikosch. Berlin–Heidelberg: Springer. 10:17–42. doi: 10.1007/978-3-540-71297-8_1.

5. Follmer, H., and A. Schied. 2002. *Stochastic finance: An introduction in discrete time*. Berlin: Walter de Gruyter. 422 p.
6. Bollerslev, T. 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. *Rev. Econ. Stat.* 69(3):542–547. doi: 10.2307/1925546.
7. Simonato, J. G. 2012. GARCH processes with skewed and leptokurtic innovations: Revisiting the Johnson S_U case. Available at: https://ssrn.com/abstract=2060994 (accessed May 18, 2012).
8. Elliott, R. J., and D. B. Madan. 1998. A Discrete time equivalent martingale measure. *Math. Financ.* 8(2):127–152. doi: 10.1111/1467-9965.00048.
9. Yi, X. 2013. Comparison of option pricing between ARMA-GARCH and GARCH-M models. London, Ontario, Canada: University of Western Ontario. MoS Thesis. 73 p.
10. Enrique, R., and L. Escobar. 2006. Using moment generating functions to derive mixture distributions. *Am. Stat.* 60(1):75–80. doi: 10.1198/000313006X90819.
11. Simonato, J. G., and L. Stentoft. 2015. Which pricing approach for options under GARCH with non-normal innovations? Available at: https://www.degroote.mcmaster.ca/files/2015/11/SimonatoStentoft.pdf (accessed November 2015).
12. Williams, D. 1991. *Probability with martingales*. Cambridge: Cambridge University Press. 251 p.
13. Cameron, R. H., and W. T. Martin. 1945. Transformation of Wiener integrals under a general class of linear transformations translations. *T. Am. Math. Soc.* 58:184–219. doi: 10.1090/S0002-9947-1945-0013240-1.
14. Bell, D. 1991. Transformations of measures on an infinite-dimensional vector space. *Seminar on stochastic processes, 1990*. Eds. E. Çinlar, P. J. Fitzsimmons, and R. J. Williams. Progress in probability book ser. Birkhäuser Boston. 24:15–25. doi: 10.1007/978-1-4684-0562-0_3.

Received June 23, 2019

Contributors

Danilishin Artem R. (b. 1992) — PhD student, Department of Operations Research, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; danilishin-artem@mail.ru

Golembiovsky Dmitry Y. (b. 1960) — Doctor of Science in technology, professor, Department of Operation Research, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; professor, Department of Banking, Sinergy University, 80-G Leningradskiy Prospect, Moscow 125190, Russian Federation; golemb@cs.msu.su

ПОВЫШЕНИЕ ТОЧНОСТИ РЕШЕНИЯ ОБРАТНЫХ ЗАДАЧ ЗА СЧЕТ УТОЧНЕНИЯ ГРАНИЧНЫХ УСЛОВИЙ*

С. М. Серебрянский[1], А. Н. Тырсин[2]

Аннотация: Рассматриваются вопросы устойчивости решения обратных задач относительно точного задания граничных условий. В практических приложениях, как правило, теоретический вид функциональной зависимости граничных условий не определен или неизвестен, а также присутствуют случайные погрешности измерений. Исследования показали, что это приводит к существенному снижению точности решения обратной задачи. С целью повышения точности решения обратных задач предложено уточнять функциональный вид граничных условий с помощью распознавания вида математической модели зависимости с последующей аппроксимацией этой функцией поведения физической величины на границе. Восстановление вида зависимости выполнено методами распознавания зависимостей на основе структурных разностных схем и распознавания на основе обратного отображения. Приведены модельные примеры реализации в условиях присутствия аддитивных случайных погрешностей измерений и неизвестного вида зависимости граничных условий.

Ключевые слова: обратная задача; распознавание; функциональная зависимость; модель; разностная схема; обратная функция; выборка; дисперсия; аппроксимация

DOI: 10.14357/19922264200108

1 Введение

Исследование так называемых обратных задач, когда исходя из некоторых характеристик физического поля необходимо восстановить характеристики самой среды, порождающей это поле, можно описать операторным уравнением 1-го рода [1]:

$$Au = f\,.$$

Трудности, возникающие при исследовании таких уравнений, связаны, главным образом, с неограниченностью обратного оператора A и отсутствием непрерывной зависимости решения от правой части (неустойчивость или некорректность задачи).

Обычные методы, используемые для приближенного решения корректных задач, оказываются, как правило, непригодными. Для эффективного решения неустойчивых задач к настоящему времени созданы специальные регулярные методы, основанные на замене исходной некорректной (неустойчивой) задачи задачей или последовательностью задач, корректных в обычном смысле [2].

При решении таких задач могут возникать возможные ошибки из-за неточного задания вида функциональных зависимостей, описывающих граничные условия, а также наличия случайных погрешностей измерений.

Рассматривается обратная задача для уравнения теплопроводности с ненулевыми граничными условиями [3]. Граничные условия заданы приближенно с некоторой погрешностью, поэтому определение вида функциональной зависимости в них имеет большой практический смысл. Как показывает практика, достаточно некоторой ошибки в начальных данных для нарушения устойчивости решения задачи.

Цель исследования — попытка преодоления этого недостатка за счет восстановления граничных условий задачи с помощью методов идентификации [4–10].

2 Постановка задачи

Рассмотрим задачу

$$\frac{\partial u_1(x,t)}{\partial t} = \frac{\partial^2 u_1(x,t)}{\partial x^2}\,,\ 0 \le x \le x_2\,,\ t \ge 0\,; \quad (1)$$

$$\frac{\partial u_2(x,t)}{\partial t} = \kappa\frac{\partial^2 u_2(x,t)}{\partial x^2}\,,\ x_2 < x \le 1\,,\ t \ge 0\,; \quad (2)$$

$$u_1(x,0) = 0\,,\ 0 \le x \le x_2\,; \quad (3)$$

$$u_2(x,0) = 0\,,\ x_2 \le x \le 1\,; \quad (4)$$

*Работа выполнена при финансовой поддержке РФФИ (проект 20-41-660008 р_а).

[1]Троицкий филиал Челябинского государственного университета, tf_chelgu@mail.ru

[2]Научно-инженерный центр «Надежность и ресурс больших систем и машин» УрО РАН; Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, at2001@yandex.ru

$$u_1(0,t) = f_\delta(t)\,,\ t \geq 0\,; \tag{5}$$

$$u_1(x_1,t) = \psi_\delta(t)\,,\ 0 < x_1 < x_2\,,\ t \geq 0\,; \tag{6}$$

$$u_1(x_2,t) = u_2(x_2,t)\,,\ t \geq 0\,; \tag{7}$$

$$\lambda_1 \frac{\partial u_1(x_2,t)}{\partial x} = \lambda_2 \frac{\partial u_2(x_2,t)}{\partial x}\,,\ t \geq 0\,. \tag{8}$$

Здесь x — пространственная координата; t — время; $u_i(x,t)$ — значение температуры в точке x в момент времени t в i-м слое; $\kappa = \kappa_1/\kappa_2$, где $\kappa_1 = \lambda_1/(\rho_1 c_1)$, $\kappa_2 = \lambda_2/(\rho_2 c_2)$, c_i — теплоемкость в i-м слое, ρ_i — плотность материала в i-м слое, λ_i — коэффициент теплопроводности i-го слоя; f_δ — значение температуры в точке 0, рассчитанное в результате эксперимента, при этом если бы было известно точное значение f_0 температурного распределения в этой точке, то выполнялось бы условие $\|f_\delta(t) - f_0(t)\| \leq \leq \delta$, где δ — погрешность измерений; ψ_δ — значение температуры в промежуточной точке x_1, рассчитанное в результате эксперимента, при этом если бы было известно точное значение ψ_0 температурного распределения в этой точке, то выполнялось бы условие $\|\psi_\delta(t) - \psi_0(t)\| \leq \delta$, где δ — погрешность измерений.

Физическая картина смоделированных процессов восстановления функции температуры представлена на рис. 1.

Требуется, используя исходную информацию f_δ, ψ_δ и δ задачи (1)–(8), рассчитать значения $u_{2\delta}(1,t)$, наиболее близкие по норме значениям $u_{20}(1,t)$.

Задача (1)–(8) считается некорректно поставленной и требует специального решения [2, 11–14]. Ее решение подробно рассматривалось в [3]. В результате была получена точная по порядку оценка

$$\sqrt{\|u_{2\varepsilon}(1,t) - u_{20}(1,t)\|^2 + \left\| v_{2\varepsilon}(1,t) - \frac{\partial u_{20}(1,t)}{\partial x} \right\|^2} \leq \leq \sqrt{2}\, l_2 \ln^{-2} \frac{1}{\varepsilon}\,,$$

где

$$\varepsilon = \frac{\lambda_1}{\lambda_2}\,\delta \left[\sqrt{2}\, r^{2/3} \left[\frac{1 + e^{-(x_1/\sqrt{2}}}{1 - e^{-\sqrt{2}\,x_1}} \right] + \frac{4}{\sqrt{2}\,x_1} \right],$$

а $r = const$ — радиус на множестве M_r, являющемся классом корректности.

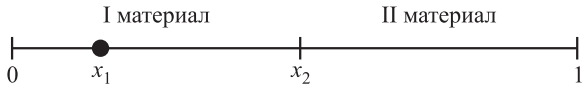


Рис. 1 Восстановление функции температуры в композиционных материалах

3 Численная реализация

В [15] приведено численное решение задачи (1)–(8), которое выглядит следующим образом:

$$u\left(1,\overline{t}\right) \approx \psi_0(\overline{t}) + \left(\frac{\psi_0(\overline{t}) - f_0(\overline{t})}{x_1} \right)(1 - x_1) + \\ + \frac{1}{2}\,\psi_0'(\overline{t})\,(1 - x_1)^2\,\kappa\,. \tag{9}$$

Если вместо точных значений $f_0(\overline{t})$ и $\psi_0(\overline{t})$ даны приближения, то использование (9), очевидно, приведет к неверному результату.

Для решения этой проблемы в [15] подробно рассмотрена методика, построенная на выполнении условий невязки и нахождении параметра регуляризации α, согласованного с погрешностью задания входных данных.

В качестве входных данных для численной реализации задачи (1)–(8) были использованы: шаг по координате $\Delta x = 0{,}1$; длина отрезка по координате $lx = 1$; длина отрезка по времени $lt = 10$; параметр регуляризации $\alpha = 0{,}05$; погрешности исходных данных $\delta = 0{,}01$, $0{,}03$, $0{,}05$ и $0{,}1$; исходная функция $u(1,t) = te^{-t}$.

Экспериментально полученные результаты представлены на рис. 2.

Аппроксимация проводилась с помощью схожих по свойствам моделей: $\overline{u}(1,t) = 1{,}1te^{-t}$; $\overline{u}(1,t) = t^\alpha e^{-t}$ и $\overline{u}(1,t) = t \cdot 3^{-\alpha t}$.

Результаты вычислений для модели $u(1,t) = \\ = te^{-t}$ с различными погрешностями представлены в табл. 1. Погрешность δ распределена равномерно с нулевым математическим ожиданием. Выделим в качестве характеристики точности среднеквадратичную ошибку между точным и приближенным решением, вычисляемую по формуле:

$$S_n^1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (u_i - \overline{u}_i)^2}\,.$$

Наличие ошибок неизбежно приводит к понижению устойчивости решения задачи, что наглядно представлено в табл. 1. Необходима более точная идентификация граничных условий задачи.

Для идентификации вида зависимостей в граничных условиях (5)–(6) были использованы следующие методы:

(1) метод распознавания зависимостей на основе структурных разностных схем (РС) [9];

(2) метод распознавания зависимостей на основе обратного отображения [10] (расширение области применения метода распознавания на основе структурных РС).

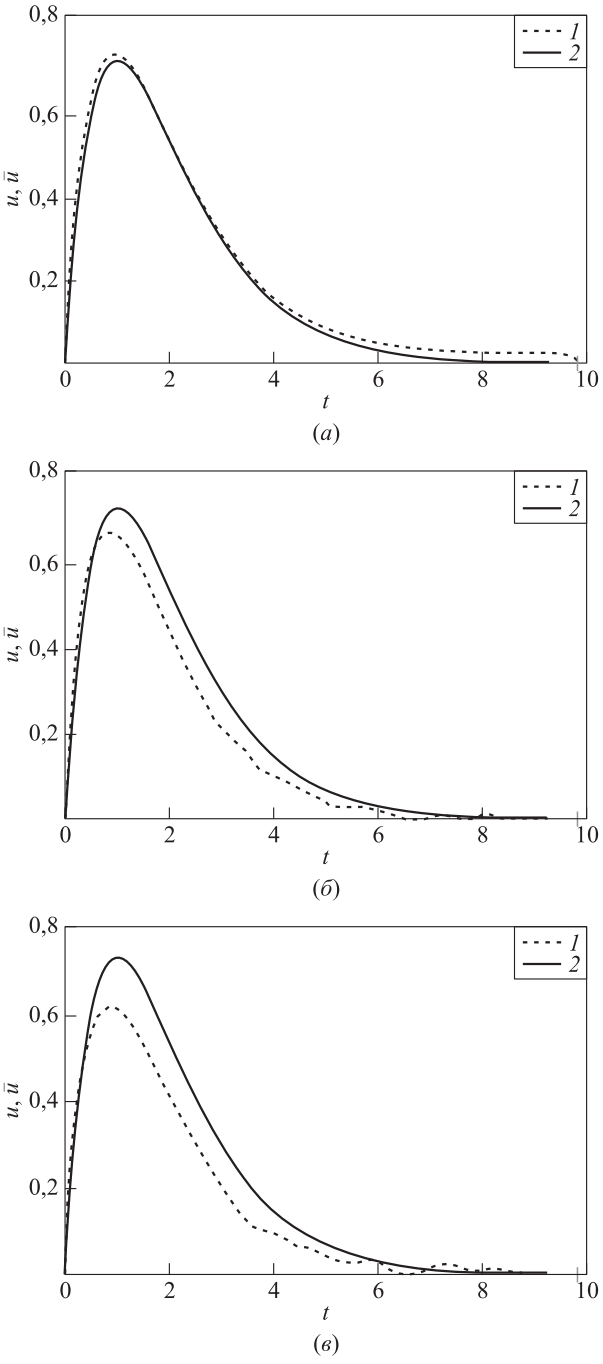


Рис. 2 Аппроксимация моделями $\overline{u}(1,t) = 1{,}1te^{-t}$ (погрешность $\delta = 0{,}01$) (*а*), $\overline{u}(1,t) = t^{\alpha}e^{-t}$ (*б*) и $\overline{u}(1,t) = t \cdot 3^{-\alpha t}$ (*в*): *1* — приближенное решение; *2* — точное решение

Результаты апробации на модельных данных, выполненной методом статистических испытаний [16, 17], показали достаточно высокую достоверность распознавания. В результате распознавания функциональных зависимостей $f_\delta(t)$ и $\psi_\delta(t)$ в граничных условиях (5)–(6) можно получить более полную картину распространения тепла в материалах и тем самым повысить точность решения задачи (1)–(8).

Таблица 1 Среднеквадратичная ошибка при различных погрешностях

Погрешность δ	Среднеквадратичная ошибка S_n^1
0,01	0,04
0,03	0,07
0,05	0,12
0,10	0,37

4 Примеры повышения точности численного решения обратной задачи за счет распознавания зависимостей в граничных условиях

Для распознавания зависимостей на основе структурных РС [10, 18, 19] рассмотрим в табл. 2 некоторый класс моделей с соответствующими им разностными схемами, которые могут быть взяты в качестве граничных условий задачи (1)–(8).

Пример 1. Пусть задана последовательность

$$y_k = f(k) = A + Bt_k = A + B\Delta k\,,\ n_0 \le k \le n_1\,,$$

где $A = 5$, $B = 3$, $\sigma_\varepsilon = 0{,}02$, $\Delta = 1$, $n_0 = 6$ и $n_1 = 100$.

Идентифицируем каждую из зависимостей

$$\widehat{f}_j(k) = f_j(k) + \varepsilon_k$$

по экспериментальным данным, т. е. вычислим оценки ее параметров $\widehat{\mathbf{b}}^{\,j}$, найденные методом наименьших квадратов, которые являются состоятельными и несмещенными (ε_k — случайные ошибки, имеющие нормальное распределение $N(0, \sigma_\varepsilon^2)$).

Используя определение обратного отображения, для каждой $\widehat{f}_j(k)$ построим последовательность $\{\widehat{u}_k^{\,j}\}$, элементы которой $\widehat{u}_k^{\,j} = \widehat{f}_j^{\,-1}(y_k)$. Отсюда, построив для последовательности $\mathbf{u} = \{u_k\}$, элементы которой $u_k = f^{-1}[f(k)] = k$, РС 2-го порядка, получим:

$$u_k = a_1u_{k-1} + a_2u_{k-2}\,,\ \ \mathbf{a}^\circ = (a_1, a_2) = (2, -1)\,,$$

т. е. отображение функции f в точку $\mathbf{a}^\circ = (2, -1)$. Таким образом, $\mathbf{u} \overset{F_p}{\to} \mathbf{a}^\circ$.

Таблица 2 Класс моделей с соответствующими им РС

№ модели	Зависимость	Структурная РС
1	$y_k = A + B\sin(\omega t_k + \varphi)$	$y_k = a_1\dfrac{y_{k-1}(y_{k-1}+y_{k-3})}{y_{k-2}} + a_2 y_{k-2},\ a_1 = 1,\ a_2 = -1$
2	$y_k = A + B\ln(Ct_k)$	$y_k = a_1 b_k^{(2)} y_{k-1} + a_2 c_k^{(2)} y_{k-2}\,,\ a_1 = 1,\ a_2 = -1,\ C > 0\,,$ $b_k^{(2)} = \dfrac{\ln(k/(k-2))}{\ln((k-1)/(k-2))}\,,\ c_k^{(2)} = \dfrac{\ln(k/(k-1))}{\ln((k-1)/(k-2))}$
3	$y_k = \dfrac{At_k^2 + Bt_k}{t_k^2 + C}$	$v_k = a_1 v_{k-1} + a_2 v_{k-2}\,,\ a_1 = 2\,,\ a_2 = -1\,,$ $v_k = 2\Delta\dfrac{y_k[(k-1)(2k-2)y_{k-2} - (k-2)(2k-1)y_{k-1}]}{(k-1)(k-2)y_k - 2k(k-2)y_{k-1} + k(k-1)y_{k-2}}$
4	$y_k = A\exp\left\{-Be^{-\alpha t_k}\right\}$	$v_k = a_1 v_{k-1} + a_2\dfrac{(v_{k-1} - v_{k-2})^2}{v_{k-2} - v_{k-3}},\ a_1 = 1,\ a_2 = 1,\ v_k = \ln y_k$
5	$y_k = A + Be^{\alpha t_k}$	$v_k = a_1 v_{k-1} + a_2 v_{k-2},\ a_1 = 2,\ a_2 = -1,$ $v_k = \ln[r(y_k - y_{k-1})],\ r = \operatorname{sign}(B\alpha)$
6	$y_k = A + Bt_k$	$y_k = a_1 y_{k-1} + a_2 y_{k-2},\ a_1 = 2,\ a_2 = -1$

Таблица 3 Результаты идентификации модели $y_k = A + Bt_k$

№ модели	Расстояние до ОДЗ
1	0,0355
2	0,0238
3	0,0238
4	0,0373
5	0,0238
6	0,0076

Таблица 4 Набор моделей и их обратные преобразования

Модель $u = F_j(t)$	Обратное преобразование $t = F_j^{-1}(u)$
$u = te^{-t}$	$-W(-t)$, $W(\cdot)$ — функция Ламберта
$u = t^\beta e^{-\alpha t}$	$-\dfrac{1{,}2W(-0{,}83333\alpha^{5/6}t)}{t}$
$u = ta^{-\alpha t}$	$-\dfrac{W(\alpha a\ln(t))}{\alpha\ln(t)}$
$u = 1 - e^{-(t/\lambda)^k}$	$\lambda\sqrt[t]{-\ln(1-k)}$

Если $\widehat{f}_j \neq f$, то

$$u_k^j = \widehat{f}_j^{\,-1}(y_k) \xrightarrow{F_p} \left(a_1^j, a_2^j\right) = \mathbf{a}^j \neq \mathbf{a}^\circ\,.$$

Поставим в соответствие каждой функции $\widehat{f}_j$ функционал вида

$$d(\mathbf{y}, f_j) = \rho_R\left(F\left[f_l^{-1}(\mathbf{y})\right], F\left[f^{-1}(\mathbf{y})\right]\right) = \sqrt{(\mathbf{a}^j, \mathbf{a}^\circ)}\,.$$

Выражение представляет собой евклидово расстояние. В итоге получим последовательность евклидовых расстояний для выбора наилучшей модели по критерию минимума расстояния до области допустимых значений (ОДЗ) вектора оценок коэффициентов АР-моделей (авторегрессионных моделей). Результаты идентификации моделей приведены в табл. 3.

Из табл. 3 видно, что расстояние до ОДЗ в случае модели $y_k = A + Bt_k$ существенно меньше, чем для остальных моделей [20]. Поэтому в данном примере, безусловно, лучшей оказалась фактическая модель $y_k = A + Bt_k$.

Для распознавания зависимостей на основе обратного отображения [15] рассмотрим в табл. 4 некоторый набор моделей заданного конечного множества моделей и их обратные преобразования, которые могут быть взяты в качестве граничных условий задачи (1)–(8).

Таблица 5 Среднеквадратичная ошибка с аддитивным шумом

Аддитивный шум δ	Среднеквадратичная ошибка S_n^2
0,01	0,0015
0,03	0,0027
0,05	0,0035
0,10	0,0080

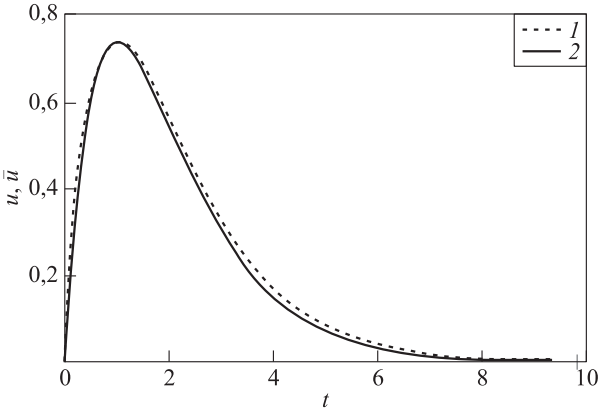


Рис. 3 Аппроксимация моделью $\overline{u}(1,t) = t^{1,05}e^{-0,95t}$: *1* — приближенное решение; *2* — точное решение

Таблица 6 Сравнение среднеквадратичных ошибок при наличии аддитивного шума

Аддитивный шум, δ	Среднеквадратичная ошибка, S_n^1	Среднеквадратичная ошибка, S_n^2	Отношение среднеквадратичных ошибок, S_n^1/S_n^2
0,01	0,04	0,0015	26,66
0,03	0,07	0,0027	25,92
0,05	0,12	0,0035	34,28
0,10	0,37	0,0080	46,25

В процессе применения предлагаемого метода распознается функциональная зависимость с различными параметрами и производится ее сглаживание.

Пример 2. Пусть задана модель $u = t^\beta e^{-\alpha t}$, где $\alpha = \beta = 1$.

Результаты вычислений [15] для рассматриваемой модели с аддитивным шумом представлены в табл. 5.

Подбор оптимальных параметров аппроксимируемой модели выполнялся по сетке в пределах: $\alpha = [0{,}7; 1{,}3]$, $\beta = [0{,}7; 1{,}3]$ с шагом 0,05.

На рис. 3 представлена графическая интерпретация проведенных вычислений.

В табл. 6 показано сравнение среднеквадратичных ошибок S_n^1 и S_n^2, полученных регулярным методом с параметром регуляризации α, результаты которого представлены в табл. 1, и методом распознавания на основе обратного отображения, которое показывает, насколько точнее происходит распознавание модели.

5 Заключение

В статье были рассмотрены вопросы, связанные с устойчивостью решения обратных задач относительно правильного задания граничных условий.

Показано, что наличие случайных погрешностей измерений и неправильное задание функциональной зависимости граничных условий приводит к значительному снижению точности решения обратной задачи.

Для решения этой проблемы предложено предварительно идентифицировать вид функциональной зависимости граничных условий и затем аппроксимировать результаты измерений на границе данной моделью. Для идентификации вида функциональной зависимости можно использовать предложенные методы распознавания. Результаты апробации на модельных данных, выполненные методом статистических испытаний, показали значительное повышение точности решения обратной задачи за счет идентификации вида функциональной зависимости граничных условий.

Литература

1. *Иванов В. К., Мельникова И. В., Филинков А. И.* Дифференциально-операторные уравнения и некорректные задачи. — М.: Физматлит, 1995. 176 с.
2. *Иванов В. К., Васин В. В., Танана В. П.* Теория линейных некорректных задач и её приложения. — М.: Наука, 1978. 206 с.
3. *Серебрянский С. М.* Об оценках погрешности методов приближенного решения одной обратной задачи // Сибирский ж. индустриальной математики, 2010. № 2(42). С. 135–148.
4. *Налимов В. В.* Теория эксперимента. — М.: Наука, Физматлит, 1971. 208 с.
5. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: исследование зависимостей. — М.: Финансы и статистика, 1985. 487 с.
6. *Клейнер Г. Б., Смоляк С. А.* Эконометрические зависимости: принципы и методы построения. — М.: Наука, 2000. 104 с.
7. *Тырсин А. Н.* Идентификация зависимостей на основе моделей авторегрессии // Автометрия, 2005. Т. 41. № 1. С. 43–49.
8. *Орлов А. И.* Прикладная статистика. — 2-е изд., перераб. и доп. — М.: Экзамен, 2007. 671 с.
9. *Тырсин А. Н., Серебрянский С. М.* Распознавание зависимостей во временных рядах на основе структурных разностных схем // Автометрия, 2015. Т. 51. № 2. С. 54–60.
10. *Тырсин А. Н., Серебрянский С. М.* Распознавание типа зависимости на основе обратного отображения // Информатика и её применения, 2016. Т. 10. Вып. 2. С. 58–64.
11. *Тихонов А. Н.* Об устойчивости обратных задач // Докл. Акад. наук СССР, 1943. Т. 39. № 5. С. 195–198.
12. *Танана В. П.* Методы решения операторных уравнений. — М.: Наука, 1981. 160 с.
13. *Алифанов О. М.* Обратные задачи теплообмена. — М.: Машиностроение, 1988. 279 с.
14. *Танана В. П., Япарова Н. М.* Об оптимальном по порядку методе решения условно-корректных задач // Сибирский ж. вычислительной математики, 2006. Т. 9. № 4. С. 353–368.

15. *Серебрянский С. М., Тырсин А. Н.* Повышение точности решения обратных задач при ошибках в начальных данных // Вестник Бурятского государственного университета. Математика, информатика, 2018. Вып. 4. С. 58–71.
16. *Ермаков С. Е.* Метод Монте-Карло и смежные вопросы. — 2-е изд. — М.: Наука, Физматлит, 1975. 472 с.
17. *Михайлов Г. А., Войтишек А. В.* Численное статистическое моделирование. Методы Монте-Карло. — М.: Академия, 2006. 368 с.
18. *Семенычев В. К.* Идентификация экономической динамики на основе моделей авторегрессии. — Самара: СНЦ РАН, 2004. 243 с.
19. *Зотеев В. Е.* Параметрическая идентификация диссипативных механических систем на основе разностных уравнений. — М.: Машиностроение, 2009. 344 с.
20. *Серебрянский С. М.* Распознание зависимостей на основе структурных разностных схем: Свидетельство о регистрации электронного ресурса № 19869. — М.: ИНИПИ РАО, ОФЭРНиО; зарегистр. 10.01.14.

Поступила в редакцию 21.04.19

IMPROVEMENT OF THE ACCURACY OF SOLUTION OF TASKS FOR THE ACCOUNT OF THE CONSTRUCTION OF BOUNDARY CONDITIONS

S. M. Serebryanskii[1] and A. N. Tyrsin[2]

[1]Troitsk Branch of Chelyabinsk State University, 9 S. Rasin Str., Troitsk 457100, Russian Federation
[2]Science and Engineering Center "Reliability and Resource of Large Systems and Machines," Ural Branch of the Russian Academy of Sciences; 54a Studencheskaya Str., Yekaterinburg 620049, Russian Federation

Abstract: The problems of stability of the solution of inverse problems with respect to the exact setting of boundary conditions are considered. In practical applications, as a rule, the theoretical form of the functional dependence of the boundary conditions is a form that is not defined or not known, and there are also random measurement errors. Studies have shown that this leads to a significant reduction in the accuracy of solving the inverse problem. In order to increase the accuracy of solving inverse problems, it was proposed to refine the functional form of the boundary conditions by recognizing the form of the mathematical model of dependence with the subsequent approximation by this function of the behavior of a physical quantity at the boundary. Dependency recovery was performed using dependency recognition methods based on structural difference schemes and inverse mapping recognition. Model examples of implementation in the presence of additive random measurement errors and an unknown type of dependence of the boundary conditions are given.

Keywords: inverse problem; recognition; functional dependence; model; difference schemes; inverse function; sampling; variance; approximation

DOI: 10.14357/19922264200108

Acknowledgments

This research was partially supported by the Russian Foundation for Basic Research (project 20-41-660008 r_a).

References

1. Ivanov, V. K., I. V. Melnikova, and A. I. Filinkov. 1995. *Differentsial'no-operatornye uravneniya i nekorrektnye zadachi* [Differential-operator equations and ill-posed problems]. Moscow: Fizmatlit. 176 p.
2. Ivanov, V. K., V. V. Vasin, and V. P. Tanana. 1978. *Teoriya lineynykh nekorrektnykh zadach i ee prilozheniya* [Theory of linear ill-posed problems and its applications]. Moscow: Nauka. 206 p.
3. Serebryanskii, S. M. 2010. Ob otsenkakh pogreshnosti metodov priblizhennogo resheniya odnoy obratnoy zadachi [On estimating the errors of approximate solution methods for an inverse problem]. *Sibirskiy zh. industrial'noy matematiki* [Siberian J. Iindustrial Mathematics] 2(42):135–148.
4. Nalimov, V. V. 1971. *Teoriya eksperimenta* [Theory of experiment]. Moscow: Nauka, Fizmatlit. 208 p.
5. Aivazyan, S. A., I. S. Enyukov, and L. D. Meshalkin. 1985. *Prikladnaya statistika: issledovanie zavisimostey* [Applied statistics: The study of dependencies]. Moscow: Finansy i statistika. 487 p.
6. Kleiner, G. B., and S. A. Smolyak. 2000. *Ekonometricheskie zavisimosti: printsipy i metody postroeniya* [Econometric dependences: Principles and methods of construction]. Moscow: Nauka. 104 p.

7. Tyrsin, A. N. 2005. Identifikatsiya zavisimostey na osnove modeley avtoregressii [Identification of dependencies based on autoregression models]. *Optoelectronics Instrumentation Data Processing* 41(1):43–49.
8. Orlov, A. I. 2007. *Prikladnaya statistika* [Applied statistics]. 2nd ed. Moscow: Ekzamen. 671 p.
9. Tyrsin, A. N., and S. M. Serebryanskii. 2015. Dependence identification in a time series on the basis of structural difference schemes. *Optoelectronics Instrumentation Data Processing* 51(2):149–154.
10. Tyrsin, A. N., and S. M. Serebryanskii. 2016. Raspoznavanie tipa zavisimosti na osnove obratnogo otobrazheniya [Recognition of dependences on the basis of inverse mapping]. *Informatika i ee Primeneniya — Inform. Appl.* 10(2):58–64.
11. Tikhonov, A. N. 1943. Ob ustoychivosti obratnykh zadach [On the stability of inverse problems]. *Dokl. Akad. nauk SSSR* 39(5):195–198.
12. Tanana, V. P. 1981. *Metody resheniya operatornykh uravneniy* [Methods for solving operator equations]. Moscow: Nauka. 160 p.
13. Alifanov, O. M. 1988. *Obratnye zadachi teploobmena* [Inverse heat transfer problems]. Moscow: Mechanical Engineering. 279 p.
14. Tanana, V. P., and N. M. Yaparova. 2006. Ob optimal'nom po poryadku metode resheniya uslovno-korrektnykh zadach [On the optimal order method for solving ill-posed problems]. *Sibirskiy zh. vychislitel'noy matematiki* [Siberian J. Numerical Mathematics] 9(4):353–368.
15. Serebryansky, S. M., and A. N. Tyrsin. 2018. Povyshenie tochnosti resheniya obratnykh zadach pri oshibkakh v nachal'nykh dannykh [Improving the accuracy of solving inverse problems with inherent errors]. *Vestnik Buryatskogo gosudarstvennogo universiteta. Matematika, informatika* [BSU Bull. Mathematics Informatics] 4:58–71.
16. Ermakov, S. E. 1975. *Metod Monte-Karlo i smezhnye voprosy* [Monte-Carlo method and related questions]. 2nd ed. Moscow: Nauka, Fizmatlit. 472 p.
17. Mikhailov, G. A., and A. V. Voytishek. 2006. *Chislennoe statisticheskoe modelirovanie. Metody Monte-Karlo* [Numerical statistical modeling. Monte-Carlo methods]. Moscow: Akademiya. 368 p.
18. Semenychev, V. K. 2004. *Identifikatsiya ekonomicheskoy dinamiki na osnove modeley avtoregressii* [Identification of economic dynamics on the basis of autoregressive models]. Samara: SNTs RAN. 243 p.
19. Zoteev, V. E. 2009. *Parametricheskaya identifikatsiya dissipativnykh mekhanicheskikh sistem na osnove raznostnykh uravneniy* [Parametric identification of dissipative mechanical systems on the basis of difference equations]. Moscow: Mashinostroenie. 344 p.
20. Serebryanskii, S. M. 2014. Raspoznanie zavisimostey na osnove strukturnykh raznostnykh skhem [Recognition of dependences on the basis of structural difference schemes]. Certificate of registration of electronic resource No. 19869.

Received April 21, 2019

Contributors

Serebryanskii Sergey M. (b. 1983) — senior lecturer, Troitsk branch of Chelyabinsk State University, 9 S. Razina Str., Troitsk 457100, Chelyabinsk Region, Russian Federation; tf_chelgu@mail.ru

Tyrsin Alexander N. (b. 1961) — Doctor of Science in technology, leading scientist, Science and Engineering Center "Reliability and Resource of Large Systems and Machines," Ural Branch of the Russian Academy of Sciences, 54a Studencheskaya Str., Ekaterinburg 620049, Russian Federation; at2001@yandex.ru

О МЕТОДАХ ПОВЫШЕНИЯ ТОЧНОСТИ МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ НА НЕСБАЛАНСИРОВАННЫХ ДАННЫХ*

Л. А. Севастьянов[1], Е. Ю. Щетинин[2]

Аннотация: Проведены исследования методов преодоления разбалансированности классов в данных с целью повышения качества классификации с точностью, более высокой, чем при непосредственном использовании алгоритмов классификации к несбалансированным данным. Для повышения точности классификации в работе предложена схема, состоящая в использовании комбинации алгоритмов классификации и методов отбора признаков RFE (Recursive Feature Elimination), Random Forest и Boruta с предварительным использованием балансирования классов методами случайного семплирования, SMOTE (Synthetic Minority Oversamplimg TEchnique) и ADASYN (ADAptive SYNthetic sampling). На примере данных о заболеваниях кожи проведены компьютерные эксперименты, показавшие, что применение алгоритмов семплирования для устранения дисбаланса классов, а также отбора наиболее информативных признаков значительно повышает точность результатов классификации. Наиболее эффективным по точности классификации оказался алгоритм случайного леса при семплировании данных с использованием алгоритма ADASYN.

Ключевые слова: классификация; несбалансированные данные; семплирование; случайный лес; ADASYN; SMOTE

DOI: 10.14357/19922264200109

1 Введение

Задачи классификации относятся к наиболее популярным в анализе данных [1]. В качестве методов, используемых для установления принадлежности объекта к тому или иному классу, чаще всего используют машинное обучение с учителем. Основная идея этого подхода — индуктивный вывод функции на основе размеченных данных для обучения. Это означает, что успешность применения алгоритма машинного обучения с учителем во многом зависит от той выборки объектов, на основе которых он «обучается». Большинство подобных алгоритмов требуют от исследователя включения сопоставимого числа примеров для каждого из классов, однако зачастую сделать сбалансированные наборы данных не представляется возможным в связи с рядом факторов. Нередко возникают ситуации, когда в наборе данных доля примеров некоторого класса незначительна (этот класс будем называть миноритарным, а другой, преобладающий над первым, — мажоритарным). Ключевые из них — специфика целевой области (балансировка данных может понизить показатель их репрезентативности) и разная цена ошибок первого и второго рода при классификации. Такие тенденции хорошо заметны, например, в кредитном скоринге, медицине, маркетинге [2, 3].

Вследствие этого возникает проблема обучения модели на несбалансированных данных (таковыми являются данные, в распределении которых наблюдается асимметрия, а показатели моды и среднего значения не равны): в соответствии с базовыми предположениями, заключенными в большинстве алгоритмов, целью обучения ставится максимизация доли правильных решений по отношению ко всем принятым решениям, а данные для обучения и генеральная совокупность подчиняются одному и тому же распределению. Однако учет данных предположений и несбалансированности выборки приводит к тому, что модель оказывается неспособна классифицировать данные лучше, чем тривиальная модель, полностью игнорирующая менее представленный класс и маркирующая все объекты для классификации как принадлежащие к мажоритарному классу.

С другой стороны, возможно построение слишком сложной модели, включающей большое множество правил, которое при этом будет охватывать малое число объектов. Такой классификатор мо-

*Работа выполнена при поддержке РФФИ (проект 18-07-00567).

[1]Российский университет дружбы народов, leonid.sevast@gmail.com

[2]Финансовый университет при Правительстве РФ, riviera-molto@mail.ru

жет оказаться неэффективным, что приведет модель к переобучению и некорректным оценкам прогноза. Следует отметить, что могут отличаться и последствия ошибочной классификации, причем неверная классификация примеров миноритарного класса, как правило, обходится в разы дороже, чем ошибочная классификация объекта из мажоритарного класса. Правильный выбор признаков может оказаться более значимой задачей, чем уменьшение времени обработки данных или повышения точности классификации. К примеру, в медицине нахождение минимального набора признаков, оптимального для задачи классификации, может стать необходимым условием для постановки диагноза. Таким образом, чтобы избежать подобного явления и достичь хорошего результата, необходимо исследовать методы работы с несбалансированными данными.

В настоящей работе проведены исследования методов преодоления разбалансированности классов с целью повышения качества классификации с точностью, более высокой, чем при непосредственном применении алгоритмов классификации к несбалансированным классам. Для повышения точности классификации в работе предложена схема, состоящая в использовании комбинации алгоритмов классификации и методов отбора признаков RFE, Random Forest и Boruta с предварительным использованием балансирования классов методами случайного семплирования, SMOTE и ADASYN.

2 Алгоритмы балансирования классов

Один из подходов к решению указанной проблемы — применение различных стратегий семплинга, которые можно разделить на две группы: случайные и специальные [2]. В первом случае удаляют некоторое число примеров мажоритарного класса (undersampling), во втором — увеличивают число примеров миноритарного класса (oversampling).

2.1 Удаление примеров мажоритарного класса. Алгоритм случайного семплирования мажоритарного класса (random undersampling)

Сначала рассчитывается K — число мажоритарных примеров, которые необходимо удалить для достижения требуемого уровня соотношения различных классов. Затем случайным образом выбираются K мажоритарных примеров и удаляются. В случае исследуемых данных естественными представляются методы по увеличению миноритарного класса. Перейдем к рассмотрению таких стратегий.

2.2 Увеличение миноритарного класса. Дублирование примеров миноритарного класса (oversampling). Случайная наивная выборка

Самый простой способ увеличить число примеров миноритарного класса — случайным образом выбрать наблюдения из него и добавить их в общий набор данных, пока не будет достигнут баланс между классами большинства и меньшинства. В зависимости от того, какое соотношение классов необходимо, выбирается число случайных записей для дублирования. Одна из проблем со случайной наивной выборкой заключается в том, что она просто дублирует уже существующие данные. К достоинствам такого подхода относятся его простота, легкость реализации и предоставляемая им возможность изменить баланс в любую нужную сторону. Про недостатки нужно говорить отдельно в соответствии с тем, какая стратегия семплинга используется: несмотря на то что обе из них изменяют общий размер данных с целью поиска баланса, их применение влечет разные последствия.

В случае undersampling удаление данных может привести к потере классом важной информации и, как следствие, к понижению показателя его презентативности. В свою очередь, применение oversampling может привести к переобучению [2].

Такой подход к восстановлению баланса не всегда эффективен, поэтому был предложен специальный метод увеличения числа примеров миноритарного класса — алгоритм SMOTE [4].

Алгоритм SMOTE основан на идее генерации некоторого числа искусственных примеров, которые были бы «похожи» на имеющиеся в миноритарном классе, но при этом не дублировали их. Для создания новой записи находят разность $d = X_b - X_a$, где X_a и X_b — векторы признаков «соседних» примеров a и b из миноритарного класса. Их находят, используя алгоритм ближайших соседей (KNN, k-nearest neighbors). В данном случае необходимо и достаточно для примера b получить набор из k соседей, из которого в дальнейшем будет выбрана запись b. Остальные шаги алгоритма KNN не требуются. Далее из d путем умножения каждого его элемента на случайное число в интервале $(0, 1)$ получают $\tilde{d}$. Вектор признаков нового примера вычисляется путем сложения X_a и $\tilde{d}$. Алгоритм SMOTE позволяет задавать число записей, которое необходимо искусственно сгенерировать. Степень

сходства примеров a и b можно регулировать путем изменения значения k (числа ближайших соседей).

Алгоритм SMOTE решает многие проблемы, которые присущи методу случайной выборки, и действительно увеличивает изначальный набор данных таким образом, что модель обучается гораздо эффективнее [5]. Тем не менее данный алгоритм имеет и свои недостатки, главный из которых — игнорирование мажоритарного класса. Это может проявиться в том, что при сильно разреженном распределении объектов миноритарного класса относительно мажоритарного наборы данных «смешаются», т. е. расположатся в таком виде, что отделить объекты одного класса от другого будет очень трудно. Примером данного явления может служить случай, когда между объектом и его соседом, на основе которых генерируется новый экземпляр, находится объект другого класса. В результате синтетически созданный объект будет находиться ближе к противоположному классу, чем к классу своих родителей. Кроме того, число сгенерированных с помощью SMOTE экземпляров задается заранее; следовательно, уменьшается возможность изменения баланса и гибкость метода.

Важно отметить существенное ограничение SMOTE. Поскольку он работает путем интерполяции между редкими примерами, то может генерировать примеры только внутри тела доступных примеров — никогда снаружи. Формально SMOTE может только заполнить выпуклую оболочку существующих примеров меньшинства, но не создавать для них новые внешние области.

Основное преимущество SMOTE по сравнению с традиционной случайной наивной чрезмерной выборкой заключается в том, что при создании синтетических наблюдений вместо повторного использования существующих наблюдений данный классификатор с меньшей вероятностью будет переобучен. В то же время всегда необходимо убедиться, что наблюдения, созданные SMOTE, реалистичны.

2.3 Адаптивный синтетический семплинг и его обобщения

В основе данного метода лежат алгоритмы синтетического семплинга, основные из которых — Borderline-SMOTE и ADASYN [6, 7]. Borderline-SMOTE накладывает ограничения на выбор объектов миноритарного класса, на основе которых генерируются новые экземпляры. Происходит это следующим образом: для каждого объекта миноритарного класса определяется набор k ближайших соседей, затем производится подсчет, сколько экземпляров из этого набора принадлежат к мажоритарному классу (это число принимается за m). После этого отбираются те объекты миноритарного класса, для которых верно неравенство $k/2 \leq m < k$. Полученный набор представляет собой экземпляры миноритарного класса, находящиеся на границе распределения, и именно у них вероятность оказаться некорректно классифицированными выше, чем у прочих. Следует отметить, почему неравенство, определяющее отбор объектов, исключает случаи, при которых все k соседей принадлежат к мажоритарному классу: это связано с тем, что подобные экземпляры расположены в зоне «смешивания» двух классов и на их основе могут быть сгенерированы лишь искажающие процесс обучения модели объекты. В связи с этим они объявляются шумом (*англ.* noise) и игнорируются алгоритмом.

Алгоритм ADASYN же, в свою очередь, основывается на систематическом методе, позволяющем адаптивно генерировать разные объемы данных в соответствии с их распределениями [7]. Входные данные для алгоритма — обучающий набор данных: D_r с m выборками с $\{x_i, y_i\}$, $i = \overline{1, m}$, где x_i — n-мерный вектор в пространстве признаков, а y_i — соответствующий класс.

Пусть m_r и m_x — число образцов классов меньшинства и большинства соответственно, такие что $m_r \ll m_x$ и $m_r + m_x = m$. Псевдокод алгоритма имеет следующий вид.

1. Вычислить пропорцию классов $d = m_r/m_x$.
2. Если $d < d_x$ (где d_x — заданный порог для максимально допустимого дисбаланса классов), то:
   (а) найти число синтетически создаваемых образцов минорного класса $G = (m_x - m_r)\beta$, где β — параметр, используемый для определения желаемого уровня баланса ($\beta = 1$ означает полный баланс классов);
   (б) для каждого $x_i \in$ minority class найти K ближайших соседей, используя евклидово расстояние, и вычислить $r_i = \Delta_i/K$;
   (в) нормализовать $r_x = r_i/\sum_i r_i$ так, чтобы r_x стал плотностью распределения;
   (г) вычислить $g_i = r_x G$ синтетической выборки, сформированной для каждого образа из класса меньшинства, где G — общее число примеров синтетических данных;
   (д) для каждого примера данных из класса меньшинства x_i создать примеры синтетических данных g_i в соответствии со следующими шагами:

– *в цикле от* 1 *до* i:

(i) случайным образом выбрать один пример данных меньшинства, x_u из K ближайших соседей для данных x_i;

(ii) создать пример синтетических данных:

$$g_i = x_i + (x_u - x_i)\lambda,$$

где $(x_u - x_i)$ — n-мерный вектор евклидова пространства; λ — случайное число: $\lambda \in [0, 1]$.

Основное различие между SMOTE и ADASYN заключается в способах создания синтетических выборочных образцов для класса меньшинства. В ADASYN используется функция плотности r_x, определяющая число синтетических образцов, которые будут созданы для конкретной точки, тогда как в SMOTE существует единый вес для всех точек меньшинства.

3 Исследуемые данные: описание и характеристики

В настоящей работе для тестирования и сравнительного анализа описанных выше методов устранения дисбаланса классов был использован набор данных о заболеваниях кожи. Диагностика эритематозно-плоскоклеточных заболеваний — серьезная проблема в дерматологии, а современные принципы диагностики и лечения опираются на наиболее раннее обнаружение заболевания. Все они имеют общие клинические особенности с очень небольшими различиями. Еще одна трудность для диагностики заключается в том, что заболевание может проявлять признаки другого заболевания на начальной стадии и иметь характерные признаки на последующих стадиях.

Исследуемые данные были созданы компанией Nielsen в 1998 г. и содержат 366 наблюдений, формирующих 6 классов, которые могут быть охарактеризованы 34 признаками [8]. Классами являются:

– псориаз (класс 1) — 112 случаев;
– себорейный дерматит (класс 2) — 72 случая;
– плоский лишай (класс 3) — 61 случай;
– розовый лишай (класс 4) — 49 случаев;
– хронический дерматит (класс 5) — 52 случая;
– красный волосяной лишай (класс 6) — 20 случаев.

Полное описание данных приведено в [9].

4 Компьютерные эксперименты

Исследования данных проводились по следующему алгоритму.

1. Предварительная обработка данных: заполнение пропусков в данных и использование кодирования признаков.
2. Балансирование классов с помощью описанных выше алгоритмов семплинга.
3. Отбор признаков по их важности.
4. Классификация с использованием логистической регрессии и метода опорных векторов (SVM — Support Vector Machine).
5. Оценка качества классификации.

В настоящей работе отбор признаков по их важности и информативности был осуществлен следующими методами:

(а) рекурсивное исключение признаков RFE [5];

(б) деревья решений RF [10];

(в) Boruta [11].

Алгоритм Random Forest представляет собой ансамбль многочисленных алгоритмов классификации (деревьев решений). Каждый из этих классификаторов строится на случайном подмножестве объектов и случайном подмножестве признаков. Пусть обучающая выборка состоит из N примеров, размерность пространства признаков равна M и задан дополнительный параметр m. Все деревья строятся независимо друг от друга по следующей процедуре.

1. Сгенерируем случайную подвыборку с повторением размером n из обучающей выборки.
2. Построим решающее дерево, классифицирующее примеры данной подвыборки, причем в ходе создания очередного узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех M признаков, а лишь из m случайно выбранных.
3. Дерево строится до полного исчерпания подвыборки и не подвергается процедуре отсечения ветвей.

Классификация объектов проводится путем голосования: каждое дерево ансамбля относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев. Для применения RF в задаче оценки важности признаков необходимо обучить алгоритм на выборке и для каждого примера обучающей выборки посчитать out-of-bag-ошибку [10].

Пусть X_n — бутстрэпированная выборка дерева b_n. Бутстрэппинг представляет собой выбор l объектов из выборки с возвращением, в результате чего некоторые объекты выбираются несколько раз, а некоторые — ни разу. Помещение нескольких копий одного объекта в бутстрэпированную выборку соответствует выставлению веса при данном объекте, соответствующее ему слагаемое несколько раз войдет в функционал, и поэтому штраф за ошибку на нем будет больше. Пусть $L(y, z)$ — функция потерь, y_i — ответ на i-м объекте обучающей выборки, тогда out-of-bag-ошибка вычисляется по следующей формуле:

$$\text{OOB} = \sum_{i=1}^{l} L\left(y_i, \frac{\sum_{n=1}^{N}[x_i \notin X_n^l]b_n(x_i)}{\sum_{n=1}^{N}[x_i \notin X_n^l]}\right).$$

Затем для каждого объекта такая ошибка усредняется по всему случайному лесу. Чтобы оценить важность признака, его значения перемешиваются для всех объектов обучающей выборки и out-of-bag-ошибка считается снова. Важность признака оценивается путем усреднения по всем деревьям разности показателей out-of-bag-ошибок до и после перемешивания значений. При этом значения таких ошибок нормализуются на стандартное отклонение.

Boruta — эвристический алгоритм отбора значимых признаков, основанный на использовании Random Forest [11]. На каждой итерации удаляются признаки, у которых Z-мера меньше максимальной Z-меры среди добавленных признаков. Чтобы получить Z-меру признака, необходимо посчитать важность признака, полученную с помощью встроенного алгоритма в Random Forest, и поделить ее на стандартное отклонение важности признака. Добавленные признаки получаются следующим образом: копируются признаки, имеющиеся в выборке, а затем каждый новый признак заполняется путем перетасовки его значений. В целях получения статистически значимых результатов эта процедура повторяется несколько раз, переменные генерируются независимо на каждой итерации. Запишем пошагово алгоритм Boruta.

1. Добавить в данные копии всех признаков. В дальнейшем копии будем называть скрытыми признаками.

2. Случайным образом перемешать каждый скрытый признак.

3. Запустить Random Forest и получить Z-меру всех признаков.

4. Найти максимальную Z-меру из всех Z-мер для скрытых признаков.

5. Удалить признаки, у которых Z-мера меньше, чем найденная на предыдущем шаге.

6. Повторять все шаги до тех пор, пока Z-мера всех признаков не станет больше, чем максимальная Z-мера скрытых признаков.

Таблица 1 Результаты классификации методом опорных векторов

Семплинг для несбалансированных классов	Методы выбора признаков	Число выбранных признаков	Точность	F1-мера	Полнота
Несбалансированные данные	Все признаки	630	0,9324	0,9337	0,9324
	RFE	65	0,9595	0,9598	0,9595
	Случайный лес	32	0,9595	0,9590	0,9595
	Boruta	207	0,9324	0,9330	0,9324
Случайная выборка	Все признаки	630	0,9324	0,9337	0,9324
	RFE	44	0,9359	0,9468	0,9459
	Случайный лес	44	0,9465	0,9466	0,9465
	Boruta	284	0,9595	0,9598	0,9595
SMOTE	Все признаки	630	0,9324	0,9337	0,9324
	RFE	68	0,9595	0,9730	0,9730
	Случайный лес	42	0,9595	0,9072	0,9054
	Boruta	257	0,9459	0,9337	0,9324
ADASYN	Все признаки	630	0,9324	0,9337	0,9324
	RFE	44	0,9459	0,9459	0,9459
	Случайный лес	40	**0,9845**	0,9330	0,9324
	Boruta	276	0,9459	0,9602	0,9595

Таблица 2 Результаты классификации с использованием логистической регрессии

Семплинг для несбалансированных классов	Методы выбора признаков	Число выбранных признаков	Точность	F1-мера	Полнота
Несбалансированные данные	Все признаки	630	0,9330	0,9234	0,9231
	RFE	19	0,9459	0,9459	0,9459
	Случайный лес	32	0,9595	0,9590	0,9595
	Boruta	200	0,9595	0,9590	0,9595
Случайная выборка	Все признаки	630	0,9630	0,9634	0,9630
	RFE	48	0,9665	0,9866	0,9865
	Случайный лес	44	0,9730	0,9730	0,9730
	Boruta	290	0,9730	0,9765	0,9730
SMOTE	Все признаки	630	0,9730	0,9734	0,9730
	RFE	20	0,9459	0,9459	0,9459
	Случайный лес	41	0,9324	0,9330	0,9324
	Boruta	264	0,9595	0,9590	0,9595
ADASYN	Все признаки	630	0,9530	0,9534	0,9530
	RFE	67	0,9595	0,9602	0,9595
	Случайный лес	42	**0,9895**	**0,9859**	**0,9893**
	Boruta	245	0,9595	0,9590	0,9595

5 Результаты исследования и их обсуждение

Для решения задачи многоклассовой классификации на несбалансированных данных были выбраны алгоритмы машинного обучения: логистическая регрессия и метод опорных векторов с линейным ядром (Linear SVM). Все вычисления были программно реализованы на языке PYTHON, их результаты, данные, а также коды программ размещены в репозитории авторов данной статьи [9]. Для сравнения результатов классификации были использованы три метрики: точность (accuracy), полнота (precision) и F1-мера. Результаты проведенных исследований представлены в табл. 1.

В ее первом столбце перечислены применявшиеся методы семплирования. Во втором столбце приведены применявшиеся методы отбора признаков, в третьей — число отобранных при этом признаков. В остальных столбцах приведены значения метрик качества, полученные в результате применения к преобразованным данным алгоритма опорных векторов (SVM).

Аналогично построена табл. 2, содержащая результаты классификации с применением логистической регрессии.

Из анализа полученных результатов, приведенных в табл. 1 и 2, можно видеть, что во всех случаях применение методов семплирования позволило получить более высокую точность классификации, чем на несбалансированных данных. В рамках применения описанной в работе схемы наилучшая точность классификации была достигнута в результате применения алгоритма балансирования классов ADASYN и затем отбора признаков алгоритмом случайного леса. Для сравнения, в работах других исследователей, проводивших подобные исследования, например [5, 10], точность классификации достигала лишь 93%.

6 Заключение

В настоящей работе предложена схема повышения точности классификации на несбалансированных данных с использованием алгоритмов балансирования классов и отбора признаков, таких как RFE, Boruta, Random Forest и др. Результаты проведенных в работе вычислительных экспериментов показали эффективность ее применения для решения поставленной задачи. В частности, алгоритм ADASYN, по сравнению с другими алгоритмами, повысил точность классификации до 98%. В заключение стоит отметить, что рассмотренная в работе проблема по-прежнему актуальна, а существующие методы могут быть улучшены.

Литература

1. *Паттерсон Дж., Гибсон А.* Глубокое обучение с точки зрения практика / Пер. с англ. А. А. Слинкина. — М.: ДМК Пресс, 2018. 417 с. (*Patterson J., Gibson A.* Deep learning: A practitioner's approach. — O'Reilly Media, 2017. 532 p.)

2. *Japkowicz N., Stephen S.* The class imbalance problem: A systematic study // Intell. Data Anal., 2002. Vol. 6. Iss. 5. P. 429–449. doi: 10.3233/IDA-2002-6504.

3. *He H., Garcia A.* Learning from imbalanced data // IEEE T. Knowl. Data En., 2009. Vol. 21. Iss. 9. P. 1263–1284. doi: 10.1109/TKDE.2008.239.

4. *Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P.* SMOTE: Synthetic minority over-sampling technique // J. Artif. Intell. Res., 2002. Vol. 16. P. 321–357. doi: 10.1613/jair.953.

5. *Lin X., Yang F., Zhou L.* A support vector machine recursive feature elimination feature selection method based on artificial contrast variables and mutual information // J. Chromatogr. B, 2012. Vol. 10. P. 149–155. doi: 10.1016/j.jchromb.2012.05.020.

6. *Han H., Wen-Yuan W., Bing-Huan M.* Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning // Advances in intelligent computing / Eds. De-Shuang Huang, Xiao-Ping Zhang, Guang-Bin Huang. — Lecture notes in computer science book ser. — Springer, 2005. Vol. 3644. P. 878–887. doi: 10.1007/11538059_91.

7. *He H., Bai Ya., Garcia A., Li Sh.* ADASYN: Adaptive synthetic sampling approach for imbalanced learning // IEEE Joint Conference (International) on Neural Networks (IEEE World Congress on Computational Intelligence). — IEEE, 2008. P. 1322–1328.

8. *Murphy P. M., Aha D. W.* UCI repository of machine learning databases. — Irvine, CA, USA: University of California, Department of Information and Computer Science, 1998. https://www.ics.uci.edu/mlearn/MLRepository.html.

9. Dermatology-article. https://github.com/riviera2015/Dermatology-article.

10. *Tuv E., Borisov A., Runger G., Torkkola K.* Feature selection with ensembles, artificial variables, and redundancy elimination // J. Mach. Learn. Res., 2009. Vol. 10. P. 1341–1366.

11. *Kursa M., Rudnicki W.* Feature selection with the Boruta package // J. Stat. Softw., 2010. Vol. 36. Iss. 11. P. 1–13. doi: 10.18637/jss.v036.i11.

Поступила в редакцию 29.11.19

ON METHODS FOR IMPROVING THE ACCURACY OF MULTICLASS CLASSIFICATION ON IMBALANCED DATA

L. A. Sevastianov[1] and E. Yu. Shchetinin[2]

[1]Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation

[2]Financial University under the Government of the Russian Federation, 49 Leningradsky Prospekt, Moscow 125993, Russian Federation

Abstract: This paper studies methods to overcome the imbalance of classes in order to improve the quality of classification with accuracy higher than the direct use of classification algorithms to unbalanced data. The scheme to improve the accuracy of classification is proposed, consisting in the use of a combination of classification algorithms and methods of selection of features such as RFE (Recursive Feature Elimination), Random Forest, and Boruta with the preliminary use of balancing classes by random sampling methods, SMOTE (Synthetic Minority Oversamplimg TEchnique) and ADASYN (ADAptive SYNthetic sampling). By the example of data on skin diseases, computer experiments were conducted which showed that the use of sampling algorithms to eliminate the imbalance of classes as well as the selection of the most informative features significantly increases the accuracy of the classification results. The most effective classification accuracy was the Random Forest algorithm for sampling data using the ADASYN algorithm.

Keywords: imbalanced data; classification; sampling; random forest; ADASYN; SMOTE

DOI: 10.14357/19922264200109

Acknowledgments

The paper was prepared with the support of the Russian Foundation for Basic Research (project 18-07-00567).

References

1. Patterson, J., and A. Gibson. 2017. *Deep learning: A practitioner's approach*. O'Reilly Media. 532 p.

2. Japkowicz, N., and S. Stephen. 2002. The class imbalance problem: A systematic study. *Intell. Data Anal.* 6(5):429–449. doi: 10.3233/IDA-2002-6504.

3. He, H., and A. Garcia. 2009. Learning from imbalanced data. *IEEE T. Knowl. Data En.* 21(9):1263–1284. doi: 10.1109/TKDE.2008.239.

4. Chawla, N. V., K. W. Bowyer, L. O.Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16:321–357.

5. Lin, X., F. Yang, and L. Zhou. 2012. A support vector machine recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *J. Chromatogr. B* 10:149–155. doi: 10.1016/j.jchromb.2012.05.020.
6. Han, H., W. Wen-Yuan, and M. Bing-Huan. 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *Advances in intelligent computing*. Eds. De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang. Lecture notes in computer science book ser. Springer. 3644:878–887. http://dx.doi.org/10.1007/11538059_91.
7. He, H., Ya. Bai, A. Garcia, and Sh. Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE Joint Conference (International) on Neural Networks (IEEE World Congress on Computational Intelligence)*. China. 1322–1328.
8. Murphy, P. M., and D. W. Aha. 1998. UCI repository of machine learning databases. Irvine, CA: University of California-Irvine, Department of Information and Computer Science. Available at: https://www.ics.uci.edu/mlearn/MLRepository.html (accessed December 27, 2019).
9. Dermatology-article. Available at: https://github.com/riviera2015/Dermatology-article (accessed December 27, 2019).
10. Tuv, E., A. Borisov, G. Runger, and K. Torkkola. 2009. Feature selection with ensembles, artificial variables, and redundancy elimination. *J. Mach. Learn. Res.* 10:1341–1366.
11. Kursa, M., and W. Rudnicki. 2010. Feature selection with the Boruta package. *J. Stat. Softw.* 36(11):1–13. doi: 10.18637/jss.v036.i11.

Received November 29, 2019

Contributors

Sevastianov Leonid A. (b. 1949) — Doctor of Science in physics and mathematics, professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; leonid.sevast@gmail.com

Shchetinin Eugene Yu. (b. 1962) — Doctor of Science in physics and mathematics, professor, Department of Data Analysis, Decision-Making and Financial Technology, Financial University under the Government of the Russian Federation, 49 Leningradsky Prosp., Moscow 125993, Russian Federation; riviera-molto@mail.ru

МОДЕЛИРОВАНИЕ ПРОЦЕССА МОНИТОРИНГА СИСТЕМ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ НА ОСНОВЕ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ*

Г. А. Попов[1], С. Ж. Симаворян[2], А. Р. Симонян[3], Е. И. Улитина[4]

Аннотация: Рассматривается задача моделирования процесса мониторинга в системах информационной безопасности по выявлению необнаруженных злоумышленных атак на основе использования методов теории массового обслуживания. Процесс мониторинга сводится к анализу потока заявок на обслуживание системой обработки данных как потока потенциально возможных злоумышленных действий. При выявлении вызова мониторинг немедленно прекращается и начинается обслуживание выявленного вызова. В рамках указанной модели получены функциональные соотношения для следующих двух наиболее важных характеристик: вероятности состояний системы и вероятности числа невыявленных вызовов в моменты окончания обслуживания. Нахождение указанных характеристик позволит более эффективно организовать процесс выявления злоумышленных атак на систему обработки данных при данной схеме обработки выявленных вызовов.

Ключевые слова: защита информации; информационная безопасность; система массового обслуживания; вероятность

DOI: 10.14357/19922264200110

1 Введение

Одной из актуальных проблем процесса обеспечения информационной безопасности в системах обработки данных является проблема выявления успешных злоумышленных атак, оказавшихся незамеченными для систем обнаружения вторжений, т. е. атак, которые система обнаружения вторжений не обнаружила на этапе поступления запроса на обработку и пропустила как незлоумышленное действие. В этом случае система обработки данных либо даже не имеет понятия о том, что была подвергнута успешной злоумышленной атаке, либо узнает об этом слишком поздно, когда уже необходимо предпринять меры по локализации и ликвидации последствий несанкционированного вторжения. Примером может служить появление новых вирусов, шпионских программ, когда существующие средства защиты еще не выработали соответствующих механизмов противодействия и в течение некоторого периода новый вирус или программа абсолютно безнаказанно действует в компьютерных системах.

В настоящее время большое внимание уделяется разработкам интеллектуальных систем защиты информации на основе нейронных сетей для решения задач, связанных с обнаружением атак, и механизмов искусственных иммунных структур, которые успешно решают задачу противодействия выявленным угрозам на этапе проникновения их в системы обработки данных [1, 2]. Статей по разработке эффективных методов и механизмов противодействия невыявленным угрозам почти нет [1].

Можно перечислить ряд методов, позволяющих в определенных ограниченных рамках выявлять и/или нейтрализовать невыявленные злоумышленные атаки [1]. В частности, периодическое полное обновление программного обеспечения, периодический мониторинг систем обработки данных с целью выявления каких-либо нетиповых отклонений, использование интеллектуальных методов выявления атак. В данной работе рассматривается процедура мониторинга систем обработки данных с целью выявления возможных следов злоумышленных атак.

Для формализованного изучения процесса мониторинга как составной части системы информационной безопасности в работе предлагается использовать аппарат теории систем массового обслуживания (СМО) [3, 4]. Использованы методы

*Исследование выполнено при финансовой поддержке РФФИ (проект 19-01-00383).

[1]Астраханский государственный технический университет, popov@astu.org

[2]Сочинский государственный университет, simsim58@mail.ru

[3]Сочинский государственный университет, oppm@mail.ru

[4]Сочинский государственный университет, ulitinaelena@mail.ru

анализа соответствующих СМО, аналогичные приведенным в [5].

2 Описание модели системы массового обслуживания с мониторингом

Рассматривается СМО, в которую поступает простейший поток вызовов с интенсивностью a. Поступающий вызов направляется в очередь. Процесс обслуживания состоит из двух этапов. На первом этапе проводится мониторинг системы с целью выявления хотя бы одного вызова, который необходимо обслужить. Будем различать два вида мониторинга. Первый выполняется перед каждым обслуживанием: он заключается в проверке всех стандартных атрибутов с целью выявления потребности в обслуживании (например, противодействии атаке). Решение о том, какой из вызовов обслуживается следующим, принимается, исходя из результатов мониторинга. Назовем этот мониторинг мониторингом очереди. Потребность в обслуживании вызова, находящегося в системе, при мониторинге очереди возникает с вероятностью $\alpha > 0$. Обслуженный вызов покидает систему. Функции распределения (ФР) времени мониторинга и времени обслуживания равны $B_1(t)$ и $B_2(t)$ соответственно. В случае если вызов не был выбран для обслуживания в процессе мониторинга, он остается в очереди. Предположим, что после цикла мониторинга не был выбран ни один вызов для обслуживания. Тогда обслуживающий прибор начинает выполнять общий мониторинг, предполагающий более системный, общий и глубокий анализ и контроль состояния системы. Этот мониторинг назовем общим мониторингом. Он выполняется периодически и непрерывно, после окончания одного цикла мониторинга сразу же начинается другой. Если в процессе данного мониторинга возникает потребность в обслуживании, то мониторинг сразу же прерывается и начинается обслуживание вызова. Вероятность обнаружения потребности в обслуживании (например, скрытой или явной атаки в системах безопасности) равна γ, а ФР длительности одного цикла мониторинга равна $B_3(t)$. Пусть $B_1(+0) = 0$, $B_3(+0) = 0$, т. е. мгновенный мониторинг исключается.

3 Основной результат

Ниже используется следующая лемма.

Лемма. *Пусть случайные величины (СВ)* $\{\xi_i, i = \overline{-n, M}\}$ *($n \geq 0$) независимы и равномерно распределены на промежутке* $[0, A]$, *где* M *— СВ, имеющая пуассоновское распределение с параметром* λ. *Каждая СВ реализуется с вероятностью* γ *и «окрашивается» в красный цвет с вероятностью* z_1 *для* $i \leq 0$ *и вероятностью* z_2 *для* $i \geq 1$; $\zeta = \min \xi_i$, *где минимум берется по всем реализованным СВ. Тогда вероятность* $\Phi(z, x, A) = \Phi(z, x)$ *события «все имеющиеся СВ окрашены в красный цвет и* $\zeta < x$*» равна*:

$$\Phi(z,x) = \\ = \begin{cases} 0 & \text{при } x < 0\,; \\ z_1^n e^{-\lambda A(1-z_2)} \left(1 - \left(1 - \gamma\dfrac{x}{A}\right)^n e^{-\lambda z_2 \gamma x}\right) & \\ & \text{при } x \in [0, A)\,; \\ z_1^n & \text{при } x \geq A\,. \end{cases}$$

Доказательство. Введем СВ η_i ($i \geq -n$):

$$\eta_i = \begin{cases} \xi_i & \text{с вероятностью } \gamma\,; \\ +\infty & \text{с вероятностью } 1-\gamma\,. \end{cases}$$

Тогда $\zeta = \min\{\eta_i | -n \leq i \leq M\}$. Пусть $B(N)$ есть событие «СВ $\{\xi_i,\ i = \overline{-n, N}\}$ окрашены в красный цвет». Для любого фиксированного $N \geq 0$ и любого $x \in [0, A]$ имеем

$$\mathsf{P}\left\{B(N), \min\left(\eta_i | i \in [-n; N]\right) < x\right\} = \\ = \mathsf{P}\{B(N)\} - \mathsf{P}\left\{B(N), \min\left(\eta_i | i \in [-n; N]\right) \geq x\right\} = \\ = z_1^n z_2^N \left(1 - \prod_{i=-n}^{N} P\left(\eta_i \geq x\right)\right) = \\ = z_1^n z_2^N \left(1 - \prod_{i=-n}^{N} \left(1 - \gamma + \gamma\left(1 - \frac{x}{A}\right)^{N+n}\right)\right) = \\ = z_1^n z_2^N - z_1^n z_2^N \left(1 - \gamma\frac{x}{A}\right)^{N+n},$$

откуда

$$\Phi(z,x) = \sum_{N \geq 0} \mathsf{P}(M = N) \times \\ \times\, \mathsf{P}\left\{B(N), \min\left(\eta_i | -n \leq i \leq N\right) < x\right\} = \\ = \sum_{N=0}^{\infty} \frac{(\lambda A)^N}{N!} e^{-\lambda A} z_1^n z_2^N \left[1 - \left(1 - \gamma\frac{x}{A}\right)^{N+n}\right] = \\ = z_1^n e^{-\lambda A(1-z_2)} \left(1 - \left(1 - \frac{\gamma x}{A}\right)^n e^{1-\gamma x/A}\right),$$

что влечет утверждение леммы.

На основе леммы доказывается следующая теорема.

Теорема. *Справедливо соотношение*

$$\Phi'_x(z,x) = \\ = \begin{cases} 0\,, & \text{если } x \notin (0,A)\,; \\ z_1^n e^{-\lambda A(1-z_2)-\lambda z_2 \gamma x}\overline{\Phi}(z,x)\,, & \text{если } x \in (0,A)\,, \end{cases}$$

где

$$\overline{\Phi}(z,x) = \left(\frac{n}{A}\left(1-\gamma\frac{x}{A}\right)^{n-1} + \left(1-\gamma\frac{x}{A}\right)^{n}\lambda z_2\gamma\right).$$

Введем следующие обозначения. Пусть $q(m,n,t)$ ($n \geq 0$; $0 \leq m \leq n$; $t \geq 0$) есть вероятность того, что в очереди в момент t находится n вызовов, из которых m поступили в систему во время обслуживания других вызовов; $q(z,w,t) = \sum_{n\geq 0}\sum_{m=0}^{n} q(m,n,t)z^m w^n$ ($0 \leq z \leq 1$); $\beta_i(s) = \int_0^\infty e^{-st}\,dB_i(t)$ — преобразование Лапласа–Стилтьеса (ПЛС) ФР $B_i(t)$ ($i = \overline{1,3}$); $\beta(s) = \beta_1(s)\beta_2(s)$; $p(m,n,t)$ — вероятность следующего события: в момент t заканчивается обслуживание вызова, в системе имеется n вызовов, из которых m пришли при обслуживании других вызовов, и нет выявленных для обслуживания вызовов ($m \leq n$).

Выведем соотношение для $q(z,w,t)$. Заметим, что функции $q(z,w,t)$ можно дать следующую вероятностную интерпретацию. Предположим, что каждый вызов, поступивший во время обслуживания другого вызова, с вероятностью w окрашивается в розовый цвет, а с вероятностью $(1-w)$ не окрашивается; кроме того, каждый вызов, поступивший во время обслуживания другого вызова, окрашивается с вероятностью z в красный цвет, а с вероятностью $(1-z)$ — в синий. Тогда $q(z,w,t)$ есть вероятность того, что в очереди в момент t все вызовы окрашены в красный цвет (если очередь не пуста) и нет синих вызовов, пришедших в систему во время обслуживания других вызовов. Потоки как красных, так и синих вызовов получаются из поступающего простейшего потока на основе процедуры его просеивания с вероятностями w и $(1-w)$ соответственно и, следовательно, также являются пуассоновскими с параметрами aw и $a(1-w)$. Аналогичные утверждения справедливы для потоков красных и неокрашенных вызовов, а также для комбинаций цветов.

Назовем вызов плохим, если за время его обслуживания пришли синие или неокрашенные вызовы. Вероятность того, что данный вызов не является ни синим, ни окрашенным, равна $1-zw$. Поэтому вероятность того, что данный вызов является хорошим (т. е. за время его обслуживания не пришло ни одного синего или неокрашенного вызова), равна $\beta(a-azw) \stackrel{\text{def}}{=} \tau$.

Составим соотношения для потока хороших вызовов, поступивших в систему за время от 0 до t. Поток плохих вызовов является просеянным пуассоновским потоком с вероятностью просеивания $1-\tau$. Следовательно, вероятность того, что за время t в систему не поступило ни одного плохого вызова, равна $\exp(-a(1-\tau)t)$. Отметим, что при этом синие вызовы могли поступить в систему в промежутках, когда система была свободна от обслуживания и занята общим мониторингом; это вызовы, с которых начинается период занятости, и вызовы, которые оказались необнаруженными обслуживающим прибором (с вероятностью $1-\alpha$ или $1-\gamma$ в зависимости от этапа работы обслуживающего устройства). Описанное событие имеет место в следующих случаях.

1. За время t вообще не было синих и неокрашенных вызовов (вероятность равна $e^{-a(1-zw)t}$) и в очереди в момент t нет синих и неокрашенных вызовов (вероятность равна $q(z,w,t)$). Таким образом, вероятность указанного случая равна $e^{-a(1-zw)t}q(z,w,t)$.

2. Первый синий вызов поступил в систему, когда она не обслуживала вызовов, а занималась общим мониторингом, а именно: в некоторый момент времени u система оказалась в состоянии, описываемом вероятностью $p(n,m,u)$; при этом вызовы, пришедшие в систему во время обслуживания других вызовов, были красными (вероятность z^m). Просмотрев все вызовы, обслуживающий прибор не обнаружил ни одного вызова, нуждающегося в обслуживании (вероятность $(1-\alpha)^n$), и длительность промежутка поиска вызовов для обслуживания (т. е. длительность мониторинга очереди) равна v_0 (вероятность $dB_1(v_0)$); за этот промежуток не пришли неокрашенные и плохие вызовы, а также вызовы, которые будут выявлены до конца N-го этапа общего мониторинга (т. е. при $(N+1)$-й попытке выявления, включая данный этап). Процедуру выявления можно рассматривать как процедуру просеивания с вероятностью просеивания α на данном этапе и вероятностью γ на последующих этапах. Вероятность того, что вызов не будет выявлен на этапах с данного по N-й, равна $\overline{\alpha}\,\overline{\gamma}^N$, где $\overline{\alpha} = 1-\alpha$, $\overline{\gamma} = 1-\gamma$. Поэтому вероятность того, что хотя бы один из уже просеянных по критериям окраски (с вероятностью $1-\tau w$) вызовов будет выявлен, равна $(1-\tau w)(1-\overline{\alpha}\,\overline{\gamma}^N)$. А вероятность того, что за время v_0 подобных вызовов не придет, равна $e^{-a(1-\tau w)(1-\overline{\alpha}\,\overline{\gamma}^N)v_0}$. Далее начался цикл из N этапов общей профилактики

($N \geq 0$), длительность i-го этапа равна v_i (вероятность $dB_3(v_i)$) ($1 \leq i \leq N$). Ни один из первоначально имевшихся n вызовов за все N этапов не был выявлен с целью обслуживания (вероятность $(1-\gamma)^{nN}$); не был выявлен также и ни один из вызовов, пришедших в систему во время общего мониторинга. Так как при общем мониторинге процесс выявления вызовов с целью обслуживания можно рассматривать как процедуру просеивания с вероятностью γ, то поток выявленных вызовов, пришедших на i-м этапе, во время обслуживания которых не было синих и неокрашенных вызовов, является просеянным пуассоновским с вероятностью просеивания $\gamma^{N-i+1}\tau w$, а вероятность того, что ни один из пришедших на i-м этапе вызовов не будет выявлен до конца N-го этапа и является хорошим, равна $e^{-a(1-\gamma^{N-i+1}\tau w)v_i}$. В силу свойств пуассоновского потока промежутки времени между последовательными моментами выявления, а также остаточные времена до выявления вызовов имеют показательное распределение с параметром $a\gamma^{N-i+1}\tau w$. Наконец, на $(N+1)$-м этапе, длительность которого равна v_{N+1} (вероятность $dB_3(v_{N+1})$), один из пришедших вызовов был выявлен через время v ($v \leq v_{N+1}$) после начала этого этапа, и ни один из пришедших за это время вызовов не был неокрашенным или плохим; вероятность этого события, в силу леммы 1, равна $\Phi'_v(\tau w, v, v_{N+1})\,dv$. А затем за оставшийся промежуток длиной $(t-u-\sum_{i=0}^{N} v_i - v)$ не пришли вызовы, за время обслуживания которых поступили в систему синие и неокрашенные вызовы (вероятность $e^{-a(1-\tau w)(t-u-\sum_{i=0}^{N} v_i - v)}$).

Просуммировав по всем значениям n, m, N, u, v, v_i ($0 \leq i \leq N+1$), получим следующее выражение для вероятности события, описываемого в п. 2:

$$\sum_{n>0}\sum_{m=0}^{n}\sum_{N\geq 0}\int\limits_{D}\cdots\int\limits_{v=0}^{v_{N+1}} p(m,n,u)z^m w^n \times$$
$$\times (1-\alpha)^n(1-\gamma)^{nN}(1-\gamma)^{nN}e^{-a(1-\alpha\gamma^N\tau w)v_0} \times$$
$$\times \prod_{i=1}^{N} e^{-a(1-\gamma^{N-i+1}\tau w)\sum_{j=i}^{N} v_j}\tau^n e^{-av_{N+1}(1-\tau w)} \times$$
$$\times d_v\left(1-\left(1-\gamma\frac{v}{v_{N+1}}\right)^n e^{-a\tau w\gamma v}\right) \times$$
$$\times e^{-a(1-\tau w)(t-u-\sum_{i=0}^{N} v_i - v)}dudB_1(v_0)\prod_{i=1}^{N+1} dB_3(v_i)\,,$$

где область интегрирования

$$D = \{(u; v_0; v_1; \ldots; v_{N+1}) : u + v_0 + v_1 + \cdots$$
$$\cdots + v_{N+1} < t,\ u \geq 0\,,\ v_i \geq 0\ (0 \leq i \leq N+1)\}.$$

Таким образом, получаем уравнение

$$e^{-a(1-\tau)t} = e^{-a(1-zw)t}q(z,t) \times$$
$$\times \sum_{n>0}\sum_{m=0}^{n}\sum_{N\geq 0}\int\limits_{D}\cdots\int\limits_{v=0}^{v_{N+1}} p(m,n,u)z^m w^n \times$$
$$\times (1-\alpha)^n(1-\gamma)^{nN}e^{-a(1-\tau w)v_0} \times$$
$$\times e^{-a\sum_{i=1}^{N}(1-\gamma^{N-i+1}\tau)\sum_{j=1}^{N} v_j e^{-av_{N+1}(1-\tau w)}} \times$$
$$\times d_v\left(1-\left(1-\gamma\frac{v}{v_{N+1}}\right)^n e^{-a\tau w\gamma v}\right) \times$$
$$\times e^{-a(1-\tau w)(t-u-\sum_{i=0}^{N} v_i - v)}dudB_1(v_0)\prod_{i=1}^{N+1} dB_3(v_i)\,.$$

Отсюда, полагая

$$p(w,z,t) = \sum_{n\geq 0}\sum_{m=0}^{n} p(m,n,t)z^m w^n,$$

выводим ($\theta = v/v_{N+1}$):

$$q(z,t) = e^{a[(1-zw)-(1-\tau)]t}\Bigg(1 +$$
$$+\sum_{N\geq 0}\int\limits_{D}\cdots\int\limits_{v=0}^{v_{N+1}}\!\!d_\theta\left(p\left((1-\alpha)(1-\gamma)^N(1-\gamma\theta), z, u\right)\times\right.$$
$$\times e^{-a\tau w\gamma v_{N+1}\theta}\Big)\, e^{-a(1-\tau w)v_0}e^{-av_{N+1}(1-\tau w)} \times$$
$$\times e^{-a(1-\tau w)(t-u-\sum_{i=0}^{N} v_i - v)} \times$$
$$\times e^{-a\sum_{j=1}^{N} v_j\left(j-\tau w(\gamma^{N-j+1}-\gamma^n)/(1-\gamma)\right)}dudB_1(v_0)\times$$
$$\left.\times \prod_{i=1}^{N+1} dB_3(v_i)\right).$$

При достаточно малых $z > 0$ выражение $\beta(a - -az) - zw > 0$. Поэтому при $t \to \infty$ величина $e^{a[\beta(a-az))-zw]t} \to \infty$ при указанных значениях z. Но так как величина $q(z,t)$ при $t \to \infty$ ограничена, то выражение в правой части в скобках должно стремиться к нулю, и после алгебраических преобразований получаем:

$$\sum_{N\geq 0}\int\limits_{\theta=0}^{\infty} d_\theta(p\left((1-\alpha)(1-\gamma)^N(1-\gamma\theta), z, a\tau(1-w)\right)\times$$
$$\times \beta_1\left(a(1-\tau w)\right)\times$$
$$\times \beta_3\left(a\left(1+\theta\right)+(\tau w(\gamma\theta-1)-\tau\theta)\,v_{N+1}\right)\times$$
$$\times \prod_{j=1}^{N}\beta_3\left(a\left(j-\frac{\gamma^{N-j+1}-\gamma^n}{1-\gamma}\tau w\right)-a(1-\tau w)\right) =$$
$$= -1\,. \quad (1)$$

Равенство (1) требует дальнейшего анализа. Оно может быть использовано в процессе компьютерного моделирования процесса мониторинга.

Исследуем теперь характеристику, которая описывает число невыявленных атак. Пусть $q_n(z,\overline{Z})$ есть вероятность того, что после окончания n-го по порядку обслуживания вызовов в системе нет выявленных z-синих вызовов, ожидающих обслуживания, все вызовы, не выявленные за i попыток, являются i-красными ($i \geq 1$); здесь $\overline{Z} = (z_1, z_2, \ldots, z_i, \ldots)$. Запишем рекуррентное соотношение, связывающее $q_{n+1}(z,\overline{Z})$ с $q_n(z,\overline{Z})$.

В момент окончания $(n+1)$-го по порядку обслуживания возможны следующие ситуации.

1. В момент окончания n-го по порядку обслуживания в очереди имеются выявленные вызовы (вероятность $q_n(z,\overline{z}) - q_n(0,\overline{z})$), с вероятностью α_i каждый вызов, не выявленный в предыдущих $(i-1)$-й попытках, будет выявлен и перейдет в основную очередь, а с дополнительной вероятностью $\overline{\alpha}_i = 1 - \alpha_i$ не будет выявлен. В последнем случае он с вероятностью z_i получит дополнительную окраску i-красного цвета и перейдет в очередь из вызовов, не выявленных при i попытках. Распишем указанное событие в терминах исходных вероятностей: $\sum_{N\geq 1}\sum_{N_i,\, i\geq 1} Q_n(N; N_i,\ i \geq 1) z^N \prod_{i\geq 1} Z_i^{N_i}$, где $Z_i = \prod_{j=1}^{i} z_j$. С вероятностью α_{i+1} каждый вызов переходит в основную очередь и окрашивается в красный цвет с вероятностью z (при этом старая окраска убирается (делится на Z_i) — в момент $(n+1)$-го окончания его прошлая окраска не представляет интереса), а с вероятностью $\overline{\alpha_{i+1}}$ вызов переходит в $(i+1)$-очередь и получает дополнительную $(i+1)$-окраску с вероятностью z_{i+1}. Таким образом, необходимо заменить Z_i на z с вероятностью α_{i+1} и заменить Z_i на Z_{i+1} с вероятностью $\overline{\alpha_{i+1}}$, т. е. Z_i заменяется на $\alpha_{i+1} z + (1-\alpha_{i+1}) Z_{i+1}$. Далее начинается обслуживание одного из основных вызовов (по предположению, в случае 1 эта очередь не пуста); при этом необходимо убрать окраску этого вызова (т. е. разделить на z). За время обслуживания этого вызова не должно поступить ни одного синего вызова; при этом каждый поступающий вызов с вероятностью α_1 становится выявленным и ставится в основную очередь, а с вероятностью $\overline{\alpha_1} = 1 - \alpha_1$ не выявляется и ставится в 1-очередь. Это равносильно тому, что поступающий поток просеивается на три потока: выявленных 0-красных вызовов (параметр потока $\alpha_1 z$), не выявленных 1-красных вызовов параметр потока $(1-\alpha_1) z_1$ и на поток всех остальных вызовов, и требуется, чтобы за время обслуживания не поступали вызовы третьего потока. Вероятность этого события равна $\beta(a - a(\alpha_1 z + (1-\alpha_1) z_1))$. Таким образом, получаем, что вероятность, описываемая в случае 1, имеет вид: $z^{-1}\left(q_n\left(z, Az + (1-A) * \overline{Z}_{+1}\right) - q_n\left(0, Az + (1-A) * \overline{Z}_{+1}\right)\right) \beta\left(a - a\left(\alpha_1 z + (1-\alpha_1 z_1)\right)\right)$. Здесь «*» означает покомпонентное умножение векторов, а индекс «+1» указывает на увеличение на 1 индексов всех компонентов вектора.
2. В момент окончания n-го по порядку обслуживания в очереди нет выявленных вызовов (вероятность $q_n(0,\overline{Z})$). Затем началось проведение общего мониторинга системы, и в течение N циклов ($N \geq 1$) не было выявлено ни одного вызова из имевшихся (вероятность $q_n(0, \{\prod_{j=1}^{N}(1-A_j)\}_{j\geq 1} * \overline{Z}_{+N})$), ни одного из вызовов, пришедших на i-м цикле мониторинга (вероятность $\beta_3(a - a\prod_{j=i}^{N}(1-\alpha_j) z_i)$ для всех $i = \overline{1,N}$. Наконец, на $(N+1)$-м цикле один из вызовов был выявлен. Так как поток красных вызовов, не выявленных на этапах с i-го по N-й, может быть получен на основе процедуры просеивания с вероятностью $\prod_{j=i}^{N}(1-\alpha_j) z_i$, то вероятность того, что за время t не будет выявлен ни один из вызовов, пришедших в систему на i-м цикле, равна

$$q_n\left(0,\overline{Z}\right) = q_n\left(0, \left\{\prod_{j=1}^{N}(1-A_j)\right\}_{j\geq 1} * \overline{Z}_{+N}\right) \times \beta_3\left(a - a\prod_{j=i}^{N}(1-\alpha_j) z_i\right).$$

Пусть $\overline{M_0}$ — вектор числа невыявленных вызовов, прошедших различные стадии выявления, в момент n-го окончания обслуживания (когда нет выявленных вызовов); M_j — число вызовов, поступивших во время j-го цикла общего мониторинга (и они не были выявлены) ($1 \leq j \leq N$); M_{N+1} — число вызовов, поступивших во время $(N+1)$-го цикла общего мониторинга до момента τ, когда был выявлен первый из всех перечисленных вызовов (т. е. до момента τ они не были выявлены). Тогда вероятность того, что хотя бы один из этих вызовов будет выявлен на $(N+1)$-м цикле и не будет выявлен на предыдущих, а все вызовы — соответствующего типа красности, равна

$$\sum_{l\geq 1} \alpha_{l+N+1} \times \prod_{k=0}^{N} \left(1-\alpha_{l+k}\right)^{M_{0,l}} M_{0,l} \left(1-\alpha_{l+N+1}\right)^{M_{0,l}-1} \times$$

$$\times z_{l+N+1}^{M_{0,l}-1}\prod_{j=1}^{N+1}\left[(1-\alpha_{N-j+1})^{M_j}z_{N-j+1}^{M_j}\right]\times$$
$$\times\prod_{j\geq 1,j\neq l}\left\{\prod_{k=0}^{N+1}(1-\alpha_{j+k})\,z_{j+N+1}\right\}^{M_{0,j}}+$$
$$+\sum_{l=1}^{N+2}\alpha_l M_l\,(1-\alpha_{l+N+1})^{M_l-1}z_{l+N+1}^{M_l-1}\times$$
$$\times\prod_{j\geq 1}(1-\alpha_{N-j+1})^{M_{0,j}}z_{n-j+1}^{M_j}\times$$
$$\times\prod_{j=1,i\neq l}^{N+2}(1-\alpha_{N-j+1})^{M_j}z_{N-j+1}^{M_j}.$$

Считаем, что момент τ выявления вызова является СВ, равномерно распределенной на интервале τ_{N+1} длительности $(N+1)$-го цикла общего мониторинга.

Заметим, что распределение вектора M_0 задается производящей функцией $q_n(0,\overline{Z})$; M_j распределено по закону Пуассона с параметром a на промежутке τ_j длительности j-го цикла. Просуммировав последнее выражение по всем возможным состояниям векторов M_0 и M_j $(1\leq j\leq N)$, получаем:

$$\sum_{l\geq 1}\alpha_{l+N+1}\prod_{k=0}^{N}(1-\alpha_{l+k})\times$$
$$\times q'_{n,l}\left(0,\left\{\prod_{k=0}^{N+1}(1-\alpha_{l+k})\,z_{l+N+1}\right\}_{l\geq 1}\right)\times$$
$$\times b_3\left(a-a((1-\alpha_1)\,z_1\right)\times$$
$$\times\prod_{j=1}^{N}\beta_3\left(a-a\prod_{k=1}^{N-j+2}(1-\alpha_k)\,z_{N-j+1}\right)+$$
$$+\sum_{l=1}^{N+1}\alpha_{N-l+2}\beta'_3\left(\prod_{k=1}^{N-l+1}(1-\alpha_k)\,z_{N-j+1}\right)\times$$
$$\times\prod_{k=1}^{N-l+1}(1-\alpha_k)\,q_n\left(0,\left\{\prod_{k=0}^{N+1}(1-\alpha_{l+k})\,z_{l+N+1}\right\}\right)\times$$
$$\times b_3\left(a-a\,(1-\alpha_1)\,z_1\right)\times$$
$$\times\prod_{j=1,i\neq l}^{N}\beta_3\left(a-a\prod_{k=1}^{N-j+2}(1-\alpha_k)\,z_{N-j+1}\right)\times$$
$$\times\prod_{j=1,j\neq l}^{N}\beta_3\left(a-a\prod_{k=1}^{N-j+2}(1-\alpha_k)\,z_{N-j+1}\right)+$$
$$+\alpha_1\prod_{j=1}^{N}\beta_3\left(\prod_{k=1}^{N-j+2}(1-\alpha_k)\,z_{N-j+1}\right)\times$$
$$\times b'_3\left(a-a\,(1-\alpha_1)\,z_1\right),$$

где

$$q_{n,l'}\left(z,\overline{Z}\right)=\frac{\partial}{\partial z_l}\,q_n\left(z,\overline{Z}\right);$$
$$b_3(s)=\int_0^{\infty}\frac{1-e^{-s\tau_{N+1}}}{s\tau_{N+1}}\,dB_3\left(\tau_{N+1}\right).$$

Дополнительно надо умножить приведенную вероятность на вероятность z того, что выявленный вызов — красный, и на вероятность $\beta(a-a(\alpha_1 z+$ $+(1-\alpha_1)z_1))$ того, что за время его обслуживания не поступят в систему синие вызовы и невыявленные 1-синие вызовы. Предполагается для простоты, что по окончании обслуживания очередной мониторинг не производится.

При этом возможны следующие три схемы обработки выявленного вызова.

1. При выявлении вызова мониторинг немедленно прекращается и начинается обслуживание выявленного вызова.
2. При выявлении вызова мониторинг доводится до конца и только затем начинается обслуживание выявленных вызовов.
3. При выявлении вызова на первом цикле мониторинга прерывание обслуживания не происходит; на втором и последующих циклах мониторинга происходит его прерывание и начинается обслуживание вызова.

В статье рассматривается только первая схема. Остальные схемы предполагается рассмотреть в последующих работах авторов. Тогда надо потребовать, чтобы за время начавшегося обслуживания красного вызова не пришли выявленные синие вызовы и 1-красные вызовы, не выявленные по окончании обслуживания этого вызова (вероятность $\beta(a-a(\alpha_1 z+(1-\alpha_1)z_1))$). Получаем:

$$q_{n+1}\left(z,\overline{Z}\right)=\beta\left(a-a\left(\alpha_1 z+(1-\alpha_1)\,z_1\right)\right)\times$$
$$\times\left\{z^{-1}\left(q_n\left(z,Az+(1-A)*\overline{Z}_{+1}\right)-\right.\right.$$
$$-q_n\left(0,Az+(1-A)*\overline{Z}_{+1}\right))+$$
$$+\left[\sum_{l\geq 1}\alpha_{l+N+1}\prod_{k=0}^{N}(1-\alpha_{l+k})\times\right.$$
$$\times q'_{n,l}\left(0,\left\{\prod_{k=0}^{N+1}(1-\alpha_{l+k})\,z_{l+N+1}\right\}_{l\geq 1}\right)\times$$
$$\times\prod_{j=1}^{N}\beta_3\left(a-a\prod_{k=1}^{N-j+2}(1-\alpha_k)\,z_{N-j+1}\right)\times$$

$$\times b_3\left(a - a\left(1-\alpha_1\right)z_1\right) + \sum_{l=1}^{N+1}\alpha_{N-l+2}\times$$
$$\times \prod_{k=1}^{N-l+1}\left(1-\alpha_k\right)\beta_3'\left(\prod_{k=1}^{N-l+1}\left(1-\alpha_k\right)z_{N-j+1}\right)\times$$
$$\times q_n\left(0,\left\{\prod_{k=0}^{N+1}\left(1-\alpha_{l+k}\right)z_{l+N+1}\right\}\right)\times$$
$$\times \prod_{j=1,j\neq l}^{N}\beta_3\left(a - a\prod_{k=1}^{N-j+2}\left(1-\alpha_k\right)z_{N-j+1}\right)\times$$
$$\times b_3\left(a - a\left(1-\alpha_1\right)z_1\right) +$$
$$+\alpha_1\prod_{j=1}^{N}\beta_3\left(\prod_{k=1}^{N-j+2}\left(1-\alpha_k\right)z_{N-j+1}\right)\times$$
$$\times b_3'\left(a - a\left(1-\alpha_1\right)z_1\right)\Bigg]\Bigg\}.$$

Помножим данное соотношение на w^n и просуммируем по $n \geq 1$. Получим:

$$\frac{q(w,z,\overline{Z}) - q_0(z,\overline{Z})}{w} =$$
$$= \beta\left(a - a\left(\alpha_1 z + \left(1-\alpha_1\right)z_1\right)\right)\times$$
$$\times\left\{\frac{q(w,z,Az+(1-A)*\overline{Z}_{+1})}{z} -\right.$$
$$\left. - \frac{q(w,z,Az+(1-A)*\overline{Z}_{+1})}{z} +\right.$$
$$+\left[\sum_{l\geq 0} q_l'\left(w,0,\left\{\prod_{j=0}^{N+1}\left(1-\alpha_{l+j}\right)z_{l+N+1}\right\}_{l\geq 1}\times\right.\right.$$
$$\left.\left.\times\,\alpha_{l+N+1}\right)\prod_{k=0}^{N}\left(1-\alpha_{l+k}\right)\times\right.$$
$$\times\prod_{j=1}^{N}\beta_3\left(a - a\prod_{k=1}^{N-j+2}\left(1-\alpha_k\right)z_{N-j+1}\right)\times$$
$$\times b_3\left(a - a\left(1-\alpha_1\right)z_1\right) +$$
$$+ q\left(w,0,\left\{\prod_{j=0}^{N+1}\left(1-\alpha_{l+j}\right)z_{l+N+1}\right\}\right)\times$$
$$\times\left(\sum_{l=1}^{N+1}\alpha_{N-l+2}\beta_3'\left(\prod_{k=1}^{N-l+1}\left(1-\alpha_k\right)z_{N-j+1}\right)\times\right.$$
$$\left.\times\prod_{k=1}^{N-l+1}\left(1-\alpha_k\right)\right)b_3\left(a - a\left(1-\alpha_1\right)z_1\right)\times$$
$$\times\prod_{j=1,j\neq l}^{N}\beta_3\left(a - a\prod_{k=1}^{N-j+2}\left(1-\alpha_k\right)z_{N-j+1}\right) +$$
$$+\alpha_1\prod_{j=1}^{N}\beta_3\left(\prod_{k=1}^{N-j+2}\left(1-\alpha_k\right)z_{N-j+1}\right)\times$$
$$\left.\times\, b_3'\left(a - a\left(1-\alpha_1\right)z_1\right)\right]\Bigg\}, \quad (2)$$

где $q_0(z,\overline{Z})$ есть производящая функция числа вызовов разных типов в системе в начальный момент времени. В частности, если вызовов в начальный момент нет, то $q_0(z,\overline{Z}) = 1$. Заметим, что $q_0(z,\overline{Z}) = q(0,z,\overline{Z})$.

Соотношение (2) является достаточно сложным интегродифференциальным уравнением для нахождения функции $q(w,z,\overline{Z})$, описывающей распределение во времени числа невыявленных атак. В настоящее время нет эффективных методов решения указанного уравнения. Однако соотношение (2) позволяет разработать рекуррентные вычислительные процедуры для нахождения $q(w,z,\overline{Z})$, которые предполагается привести в последующих работах.

4 Заключение

В работе рассмотрена задача построения модели мониторинга систем обработки данных по показателям информационной безопасности на основе использования методов теории массового обслуживания. Получены соотношения для следующих двух наиболее важных характеристик указанных систем: вероятности состояний системы и вероятности числа невыявленных вызовов в моменты окончания обслуживания. Здесь под состоянием понимается число вызовов, ожидающих обслуживания (т. е. атак, ожидающих нейтрализации), включая невыявленные атаки. Нахождение указанных характеристик позволит более эффективно организовать процесс выявления злоумышленных атак на систему обработки данных.

Литература

1. *Грушо А. А., Грушо Н. А., Тимонина Е. Е.* Методы защиты информации от атак с помощью скрытых каналов и враждебных программно-аппаратных агентов в распределенных системах // Вестник РГГУ. Сер.: Документоведение и архивоведение. Информатика. Защита информации и информационная безопасность, 2009. № 10. С. 33–45.
2. *Kopyrin A. S., Simavoryan S. Zh., Simonyan A. R., Ulitina E. I.* The methodology of risk analysis in assessing information security threats // Modeling Artificial Intelligence, 2017. No. 4-2. P. 78–85. doi: 10.13187/mai.2017.2.78.

3. *Бажаев Н. А., Давыдов А. Е., Кривцова И. Е., Лебедев И. С., Салахутдинова К. И.* Подход к анализу состояния информационной безопасности беспроводной сети // Прикладная информатика, 2016. Т. 11. № 6(66). С. 121–128.
4. *Коляденко Ю. Ю., Лукинов И. Г.* Модель распределенных атак в программно-конфигурируемых сетях связи // Вестник ЮУрГУ. Сер.: Компьютерные технологии, управление, радиоэлектроника, 2017. Т. 17. № 3. С. 34–43. doi: 10.14529/ctcr170304.
5. *Гнеденко Б. В., Даниелян Э. А., Димитров Б. Н., Климов Г. П., Матвеев В. Ф.* Приоритетные системы массового обслуживания. — М.: МГУ, 1973. 448 с.

Поступила в редакцию 03.08.19

MODELING OF MONITORING OF INFORMATION SECURITY PROCESS ON THE BASIS OF QUEUING SYSTEMS

G. A. Popov[1], S. Zh. Simavoryan[2], A. R. Simonyan[2], and E. I. Ulitina[2]

[1]Astrakhan State Technical University, 16 Tatischeva Str., Astrakhan 414056, Russian Federation
[2]Sochi State University, 94 Plastunskaya Str., Sochi 354003, Russian Federation

Abstract: The paper is devoted to the mathematical modeling of monitoring process by the information security systems, aimed at detection of hidden malicious attacks. The modeling is based on the queueing theory formalism. The monitoring process is reduced to the analysis of the customer flow arriving at the queueing system, in which each customer is regarded as carrying potential malicious attacks. Functional relations between the system state probability distribution and the distribution of the number of undetected malicious attacks on service completion epochs are obtained. These characteristics may allow one to improve the efficiency of malicious attacks detection process in the data processing systems.

Keywords: protection of information; information security; queuing system; probability

DOI: 10.14357/19922264200110

Acknowledgments

The reported study was funded by the Russian Foundation for Basic Researh, project No. 19-01-00383.

References

1. Grusho, A. A., N. A. Grusho, and E. E. Timonina. 2009. Metody zashchity informatsii ot atak s pomoshch'yu skrytykh kanalov i vrazhdebnykh programmno-apparatnykh agentov v raspredelennykh sistemakh [Methods of information protection aganst covert channels attacks and malicious software/hardware agents in distributed systems]. *Vestnik RGGU. Ser. Dokumentovedenie i arkhivovedenie. Informatika. Zashchita informatsii i informatsionnaya bezopasnost'* [RGGU Bulletin. Informatics. Information security. Mathematician ser.] 10:33–45.
2. Kopyrin, A. S., S. Zh. Simavoryan, A. R. Simonyan, and E. I. Ulitina. 2017. The methodology of risk analysis in assessing information security threats. *Modeling Artificial Intelligence* 4-2:78–85.
3. Bazhayev, N. A., A. E. Davydov, I. E. Krivtsova, I. S. Lebedev, and K. I. Salakhutdinova. 2016. Podkhod k analizu sostoyaniya informatsionnoy bezopasnosti besprovodnoy seti [Wireless security information analysis approach]. *Prikladnaya informatika* [J. Applied Informatics] 11(6(66)):121–128.
4. Kolyadenko Yu. Yu., and I. G. Lukinov. 2017. Model' raspredelennykhatak v programmno-konfiguriruemykh setyakh svyazi [Model of distributed attacks in program-configurable communication networks]. *Vestnik YUUrGU. Ser. Komp'yuternye tekhnologii, upravleniye, radioelektronika* [Bulletin of SUSU. Computer technologies, automatic control, radioelectronics ser.] 17(3):34–43.
5. Gnedenko, B. V., E. A. Danielyan, B. N. Dimitrov, G. P. Klimov, and V. F. Matveev. 1973. *Prioritetnyye sistemy massovogo obsluzhivaniya* [Priority queues]. Moscow: MSU. 448p.

Received August 3, 2019

Contributors

Popov Georgy A. (b. 1950) — Doctor of Science in technology, professor, Head of Department, Astrakhan State Technical University, 16 Tatischeva Str., Astrakhan 414056, Russian Federation; popov@astu.org

Simavoryan Simon Zh. (b. 1958) — Candidate of Science (PhD) in technology, associate professor, Sochi State University, 94 Plastunskaya Str., Sochi 354003, Russian Federation; simsim58@mail.ru

Simonyan Arsen R. (b. 1960) — Candidate of Science (PhD) in physics and mathematics, associate professor, Sochi State University, 94 Plastunskaya Str., Sochi 354003, Russian Federation; oppm@mail.ru

Ulitina Elena I. (b. 1978) — Candidate of Science (PhD) in physics and mathematics, associate professor, Sochi State University, 94 Plastunskaya Str., Sochi 354003, Russian Federation; ulitinaelena@mail.ru

О КАУЗАЛЬНОЙ РЕПРЕЗЕНТАТИВНОСТИ ОБУЧАЮЩИХ ВЫБОРОК ПРЕЦЕДЕНТОВ В ЗАДАЧАХ ДИАГНОСТИЧЕСКОГО ТИПА*

А. А. Грушо[1], М. И. Забежайло[2], Е. Е. Тимонина[3]

Аннотация: Работа посвящена некоторым особенностям анализа причинности в задачах интеллектуального анализа данных. Обсуждаются возможности использования так называемых открытых логических теорий в задачах диагностики (классификации) для описания пополняемых наборов эмпирических данных. В задачах этого типа необходимо установить (спрогнозировать, диагностировать и др.) наличие или отсутствие целевого свойства у нового прецедента, заданного описанием на том же языке представления гетерогенных данных, которым описаны примеры, обладающие целевым свойством, и контрпримеры, не обладающие целевым свойством. Представлен вариант построения открытых теорий, описывающих коллекции прецедентов средствами специальных логических выражений — характеристических функций (ХФ). Характеристические функции позволяют избавиться от гетерогенности в описаниях прецедентов. Предложена процедурная конструкция формирования ХФ обучающей выборки прецедентов. Исследованы свойства ХФ и некоторые условия их существования.

Ключевые слова: диагностика; каузальный анализ; интеллектуальный анализ данных; открытые логические теории

DOI: 10.14357/19922264200111

1 Введение

Работа посвящена некоторым особенностям анализа причинности в задачах диагностического типа. К этому классу традиционно относят такие задачи, в которых исходные данные представлены описаниями двух типов прецедентов — тех, которые обладают заданным целевым свойством (будем называть их *примерами* наличия анализируемого эффекта), и тех, которые, будучи «похожими» на примеры, тем не менее таковым свойством не обладают (будем называть их *контрпримерами*). В задачах этого типа необходимо установить (спрогнозировать, диагностировать и др.) наличие или отсутствие целевого свойства у нового (предлагаемого для формирования «диагноза») прецедента, заданного описанием на том же языке представления данных/знаний, которым описаны примеры и контрпримеры.

По-видимому, исторически первые примеры подобных проблем сформировала медицина, однако сегодня предлагаемый перечень содержит и такие не менее востребованные предметные области, как:

- техническая диагностика [1] (рассматриваемая в широком диапазоне от диагностики отказов и поддержания работоспособности сложного оборудования — транспортного, энергетического, нефтегазового и т. п. — до обеспечения устойчивой работы крупных центров обработки данных, компьютерных сетей, телекоммуникационной инфраструктуры и др.);
- финансовый мониторинг и противодействие мошенничеству в финансовой сфере [2];
- информационная безопасность (идентификация компьютерных атак и организация противодействия целенаправленным вредоносным воздействиям в сфере информационно-коммуникационных технологий) [3, 4];
- экологический мониторинг (в частности, выявление потенциально опасных химических соединений в окружающей человека среде) [5, 6];
- социологический анализ (типология социума и анализ рациональных оснований принятия решений различными социальными группами) [7] и др.

*Работа частично поддержана РФФИ (проект 18-29-03081).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, grusho@yandex.ru

[2]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, m.zabezhailo@yandex.ru

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, eltimon@yandex.ru

Соответствующие подходы, математические модели, методы и алгоритмы формируют в этих областях своего рода сквозную технологию компьютерного анализа данных и поддержки принятия управленческих решений.

Наиболее распространенными в рассматриваемой области анализа данных и поддержки принятия решений оказались так называемые *интерполяционно-экстраполяционные* математические техники решения задач диагностического типа: имеющаяся «обучающая» выборка описаний прецедентов (примеров и контрпримеров) интерполируется той или иной системой эмпирических зависимостей (ЭЗ) (регрессионных, логических и т. п.) так, чтобы «диагностика» нового прецедента могла быть сведена к проверке корректной экстраполяции каких-либо из уже найденных ЭЗ на описание этого нового прецедента.

Критически значимым аспектом, характерным для задач этого типа, оказывается проблема анализа причинности (идентификации каузальных оснований) формируемых диагностических заключений. С управленческой точки зрения существенно иметь возможности выделять именно причинные факторы влияния, «вынуждающие» появление исследуемого целевого эффекта. Это позволяет, оказывая влияние именно на причины, целенаправленно воздействовать на ситуацию и объект управления, чтобы не допустить возникновения негативных последствий.

При изучении эффектов причинности (каузальных «влияний») путем анализа эмпирического материала важно принять во внимание ряд фундаментальных обстоятельств, к наиболее критически значимым из которых следует отнести:

- отсутствие ясных, корректных и точных представлений о «структуре» искомых причинных «влияний», в частности об исчерпывающем перечне факторов «влияния» и полном комплексе взаимосвязей между ними;
- отсутствие достаточно надежных оснований считать используемый язык представления анализируемых эмпирических данных *каузально полным*, т. е. адекватно представляющим как все факторы причинного влияния, так и все взаимосвязи между ними [8].

В подобной ситуации говорить об анализе собственно причин в задачах рассматриваемого типа не совсем корректно. Более естественным представляется ограничиться лишь эмпирическими «причинами» — описаниями каузальных влияний, которые могут быть выделены (найдены, восстановлены из данных и т. п.) в имеющейся «обучающей» выборке прецедентов и задействованном для описания накапливаемого эмпирического материала языке представления данных.

Далее в этой работе, упоминая о причинности и каузальных влияниях, будем иметь в виду именно такое, учитывающее специфику анализа эмпирических данных, уточнение представлений о причинности.

Принимая во внимание особую значимость каузальных факторов «влияния», характеризующих в каждом конкретном случае наличие изучаемого целевого эффекта, и дополнительно отмечая критически важную роль идентификации таких факторов для организации целенаправленного противодействия негативным последствиям их влияния, более подробно рассмотрим возможности их выявления с помощью подходящих средств компьютерного анализа данных. Результативной здесь оказывается следующая очевидная эвристика: так как необходимые факторы влияния должны отражать «каузальность» во всех описаниях прецедентов наличия анализируемого целевого эффекта (примеров), то, как следствие, значимые для возникновения исследуемого целевого эффекта комбинации таких факторов должны отображаться сходствами описаний примеров [9, 10]. При этом естественно потребовать, чтобы выделяемые таким образом значимые комбинации факторов влияния не «встречались» ни в одном из описаний контрпримеров. Случаи, когда «причина» имеется, а эффекта — нет, должны быть использованы для удаления артефактов из формируемого на примерах множества значимых комбинаций факторов целенаправленного каузального влияния.

Располагая «непротиворечивыми», характерными только для примеров значимыми комбинациями факторов влияния, можно строить соответствующие логические условия. Будем называть их характеристическими функциями текущей базы фактов (БФ). Характеристические функции охватывают имеющееся на текущий момент множество прецедентов (примеров и контрпримеров), каждое из которых истинно на всех примерах и ложно на всех контрпримерах из текущей БФ. Непустота множества ХФ, порождаемых на текущей выборке прецедентов БФ, может рассматриваться как характеристика ее каузальной репрезентативности, т. е. корректной «разделимости» примеров и контрпримеров логическими условиями каузального характера.

Учитывая, что «природа» диагностических задач требует возможностей оперировать расширяющимися за счет новых описаний прецедентов коллекциями эмпирических данных, естественно рассматривать соответствующую «динамику» накапливания и изменений анализируемого эмпири-

ческого материала, т. е. динамику изменений каждого конкретного множества соответствующих ХФ. Особый интерес вызывают такие подсемейства ХФ, которые «наследуются», т. е. сохраняют корректность описания конкретной БФ при ее конкретном расширении. С математической точки зрения интерес представляет проблема эффективного выделения именно таких подсемейств ХФ. В данном случае речь идет о проблеме эффективного поиска ЭЗ, описывающих природу диагностируемого целевого эффекта.

2 Математическая модель формирования факторов каузального влияния

Пусть заданы два множества:

$$\boldsymbol{U} = \{a_1, a_2, \ldots, a_n\};$$
$$\boldsymbol{\Omega} = \{O_1, O_2, \ldots, O_m\} \subseteq 2^U \backslash \phi,$$

первое из которых будем называть исходным алфавитом, или множеством образующих элементов для описания анализируемых прецедентов из Ω, а второе — множеством собственно описаний прецедентов, построенных над универсумом U.

Примерами будем называть такие прецеденты $O_1^+, O_2^+, \ldots, O_{m^+}^+$ из множества $\boldsymbol{\Omega}$, которые обладают исследуемым целевым свойством P. Обозначим это подмножество множества $\boldsymbol{\Omega}$ через $\boldsymbol{\Omega}^+$. *Контрпримерами* будем называть такие прецеденты $O_1^-, O_2^-, \ldots, O_{m^-}^-$ из множества $\boldsymbol{\Omega}$, которые не обладают этим целевым свойством P. Обозначим это подмножество множества $\boldsymbol{\Omega}$ через $\boldsymbol{\Omega}^-$. Таким образом, множество $\boldsymbol{\Omega}$ будет представлять описание текущего состояния БФ.

Для операции $\otimes$ сходства описаний прецедентов из БФ используем операцию $\cap$ пересечения множеств образующих, формирующих описание каждого прецедента из $\boldsymbol{\Omega}$, т. е. результатом ее применения будет множество попарно совпадающих значений признаков из множества образующих $\boldsymbol{U}$. Несложно убедиться, что такое определение операции сходства корректно, т. е. по непустой операции это отношение рефлексивно, симметрично и ассоциативно.

Пусть $\boldsymbol{\Omega} = \{O_1, O_2, \ldots, O_m\}$. Построим по $\boldsymbol{\Omega}$ и $\otimes$ множество $\mathrm{Dom}\,(\boldsymbol{\Omega})$ следующим образом:

(а) $\boldsymbol{\Omega} = \{O_1, O_2, \ldots, O_m\} \subset \mathrm{Dom}\,(\boldsymbol{\Omega})$;

(б) $\{[A \in \mathrm{Dom}\,(\boldsymbol{\Omega})] \& [B \in \mathrm{Dom}\,(\boldsymbol{\Omega})] \& [(A \otimes B) \neq \neq \phi]\} \rightarrow [(A \otimes B) \in \mathrm{Dom}\,(\boldsymbol{\Omega})]$;

(в) других элементов в множестве $\mathrm{Dom}\,(\boldsymbol{\Omega})$нет.

Аналогичным образом определяются множества $\mathrm{Dom}\,(\boldsymbol{\Omega}^+)$ и $\mathrm{Dom}\,(\boldsymbol{\Omega}^-)$.

Фиксируя каждый конкретный, непустой результат $V = V_0$ вычисления операции $\otimes$ сходства на элементах множества $\mathrm{Dom}\,(\boldsymbol{\Omega})$, где $A \otimes B = V_0$, можно выделять соответствующие классы сходства $\mathbf{E}_{V_0}^{\otimes}$:

$$\mathbf{E}_{V_0}^{\otimes} = \{\langle O_{i_1}, O_{i_2}\rangle | O_{i_1} \otimes O_{i_2} \otimes V_0 = V_0\}\ .$$

Таким образом, фиксируя конкретное значение параметра сходства V_0, можно определить множество всех прецедентов $\mathbf{E}_{V_0}$, которые содержат множество признаков V_0. Тогда можно разбить исходное множество прецедентов $\boldsymbol{\Omega}$ на два непересекающихся подмножества $\mathbf{E}_{V_0}$ и $\boldsymbol{\Omega}\backslash\mathbf{E}_{V_0}$: $\mathbf{E}_{V_0} \cap \boldsymbol{\Omega}\backslash\mathbf{E}_{V_0} = \phi$. При этом надо рассматривать каждый раз при таком разделении множества $\boldsymbol{\Omega}$ на два непересекающихся подмножества с помощью соответствующего V_0 все содержащие общую часть V_0 прецеденты.

Для каждого сформированного таким образом класса сходства примеров должно дополнительно выполняться логическое условие, называемое *запретом на контрпримеры* (ЗКП) [9]. В случае использования ЗКП запрещается вложимость примеров одного «знака» в какой-либо из сформированных классов сходства противоположного «знака».

Условие ЗКП для примеров можно представить в следующем виде:

$$\forall\, V \left\{ \left[\left(V \in \mathrm{Dom}\ \left(\boldsymbol{\Omega}^+\right)\right) \&(V \neq \phi)\right] \& \right.$$
$$\& \left[\exists\, O_r^+, \exists\, O_s^+ \left(O_r^+ \otimes O_s^+ \otimes V = V\right)\right] \rightarrow$$
$$\left. \rightarrow \neg \left(\exists\, O_p^- \left[\left(O_p^- \in \left(\boldsymbol{\Omega}^-\right)\right) \& \left(O_p^- \otimes V = V\right)\right]\right)\right\}\ .$$

При этом условие ЗКП для контрпримеров может быть формализовано симметричным образом заменой $\boldsymbol{\Omega}^+$ на $\boldsymbol{\Omega}^-$, примеров O_r^+ и O_s^+ — на контрпримеры O_r^- и O_s^-, а вместе с ними контрпримера O_p^- на пример O_p^+.

Исходя из того принципа, что для появления свойства P должна быть причина, при выполнении условия ЗКП можно ожидать, что причина содержится в примерах $\boldsymbol{\Omega}^+$. Однако неизвестны характеристики причины, а именно: является ли причина единственной и какова структура причины. Из упомянутого выше принципа сходства можно искать причины свойства P в порождающих элементах V классов сходства. Но в силу указанной неопределенности о всех порождающих классы сходства элементах необходимо говорить как о возможных факторах влияния на прогноз появления свойства P, или факторах эмпирической причинности свойства P. При этом сами порождающие элементы

классов сходства можно называть ЭЗ. Аналогично можно определять ЭЗ на множестве $\mathbf{\Omega}^-$.

Итак, чтобы ожидать, что новый предложенный для прогноза свойств прецедент O_0 также имеет свойство P, в обсуждаемой схеме рассуждений следует убедиться, что данный O_0 «попадает» хотя бы в один из классов сходства, сформированных на исходном множестве примеров $\mathbf{\Omega}^+$.

Легко видеть, что покрытие всего исходного множества прецедентов $\mathbf{\Omega}^+$ классами сходства прецедентов-примеров, порождаемыми на его примерах с непременным выполнением условия ЗКП, позволяет разделить с помощью выделяемых комбинаций значений факторов «причинности» примеры и контрпримеры в текущем множестве $\mathbf{\Omega}$ (текущей БФ).

Ясно, что один класс сходства $\mathbf{E}_V$ может не покрывать множество $\mathbf{\Omega}^+$. Однако несколько классов сходства $\mathbf{E}_{V_1}, \ldots, \mathbf{E}_{V_k}$ могут покрыть все множество $\mathbf{\Omega}^+$. Такое покрытие, если оно существует, может не быть единственным. Из условия ЗКП для любого покрытия множества $\mathbf{\Omega}^+$ классами сходства $\mathbf{E}_{V_1}, \ldots, \mathbf{E}_{V_k}$ порождающие элементы $V_1, \ldots, V_k$ не могут встречаться в прецедентах из множества $\mathbf{\Omega}^-$.

Для покрытия $\mathbf{E}_{V_1}, \ldots, \mathbf{E}_{V_k}$ определим бинарные переменные следующим образом:

$$x_{ij} = \begin{cases} 1, & \text{если характеристика } a_i \in V_j\,; \\ 0 & \text{в противном случае.} \end{cases}$$

Тогда с множеством V_j взаимно однозначно связана конъюнкция $\bigwedge\limits_{a_i \in V_j} x_{ij}$.

Определение. Характеристической функцией покрытия $\mathbf{\Omega}^+$ классами сходства $\mathbf{E}_{V_1}, \ldots, \mathbf{E}_{V_k}$ называется двоичная функция $\bigvee\limits_{j=1}^{k} \bigwedge\limits_{a_i \in V_j} x_{ij}$.

Если возможно построить s покрытий классами сходства множества $\mathbf{\Omega}^+$, то существует s ХФ.

Утверждение. Каждая построенная в соответствии с определением функция ХФ принимает на факте φ из текущей БФ значение 1 (истина) тогда и только тогда, когда данный факт характеризуется наличием анализируемого целевого свойства P, и 0 (ложь) тогда и только тогда, когда данный факт характеризуется отсутствием анализируемого целевого свойства P.

Доказательство. *Необходимость.* Если φ не обладает свойством P, то $\varphi \notin \mathbf{\Omega}^+$. Тогда при любом покрытии $\mathbf{\Omega}^+$ в силу ЗКП ни один из порождающих элементов из покрытия не может входить в φ. Следовательно, ни одна конъюнкция ХФ не может равняться 1, а тогда значение ХФ на φ равно 0.

Если φ обладает свойством P, то $\varphi \in \mathbf{\Omega}^+$ и, следовательно, входит в один из классов сходства покрытия $\mathbf{\Omega}^+$. Тогда порождающий элемент этого класса сходства принадлежит φ и соответствующая конъюнкция равняется 1.

Достаточность. Если $\text{ХФ}(\varphi) = 1$, то в определяющем ХФ покрытии существует конъюнкция, принимающая значение 1. Тогда φ принадлежит соответствующему классу сходства в $\mathbf{\Omega}^+$. Если $\text{ХФ}(\varphi) = 0$, то в определяющем ХФ покрытии не существует ни одной конъюнкции, принимающей значение 1. Тогда φ не принадлежит ни одному классу сходства в $\mathbf{\Omega}^+$. Тогда из покрытия следует, что $\varphi \in \mathbf{\Omega}^-$.

3 Примеры

Несложно убедиться, что в общем случае ХФ, формируемая на текущей БФ, не является единственной.

Пример 1. Пусть $\boldsymbol{U} = \{a_1, a_2, a_3, a_4\}$ и $\mathbf{\Omega} = \{O_1, O_2, \ldots, O_7\}$, где

$$\begin{aligned} O_1 &= \{a_1, a_4\}\ ; \\ O_2 &= \{a_1, a_2\}\ ; \\ O_3 &= \{a_2, a_4\}\ ;, \\ O_4 &= \{a_1, a_3\}\ ; \\ O_5 &= \{a_1, a_2, a_3, a_4\}\ ; \\ O_6 &= \{a_2, a_3\}\ ; \\ O_7 &= \{a_3, a_4\}\ , \end{aligned}$$

при этом множество примеров $\mathbf{\Omega}^+ = \{O_1^+, O_2^+, \ldots, O_7^+\}$, а множество контрпримеров — $\mathbf{\Omega}^- = \phi$.

Несложно убедиться, что дизъюнкции

$$\begin{aligned} \text{ХФ}_1(\mathbf{\Omega}) &= a_1 \vee a_2 \vee a_3\,(\{O_1, O_2, O_4, O_5\} \cup \\ &\quad \cup \{O_2, O_3, O_5, O_6\} \cup \{O_4, O_5, O_6, O_7\})\ ; \\ \text{ХФ}_2(\mathbf{\Omega}) &= a_1 \vee a_3 \vee a_4\,(\{O_1, O_2, O_4, O_5\} \cup \\ &\quad \cup \{O_4, O_5, O_6, O_7\} \cup \{O_1, O_3, O_5, O_7\})\ ; \\ \text{ХФ}_3(\mathbf{\Omega}) &= a_2 \vee a_3 \vee a_4\,(\{O_2, O_3, O_5, O_6\} \cup \\ &\quad \cup \{O_4, O_5, O_6, O_7\} \cup \{O_1, O_3, O_5, O_7\}) \end{aligned}$$

соответствуют трем различным покрытиям исходного множества примеров $\mathbf{\Omega}^+$ классами сходства примеров из $\mathbf{\Omega}$.

Множество всех ХФ, построенных на заданной БФ, может оказаться пустым.

Пример 2. Пусть $\boldsymbol{U} = \{a_1, a_2, \ldots, a_n, x, y, z\}$ и $\mathbf{\Omega} = \{O_1, O_2, O_3\}$, где

$$\begin{aligned} O_1 &= \{a_1, a_2, \ldots, a_n, x\}\ ; \\ O_2 &= \{a_1, a_2, \ldots, a_n, y\}\ ; \\ O_3 &= \{a_1, a_2, \ldots, a_n, z\}\ , \end{aligned}$$

при этом множество примеров $\mathbf{\Omega}^+ = \{O_1^+, O_2^+\}$, а множество контрпримеров $\mathbf{\Omega}^- = \{O_3^-\}$.

Тогда единственная ЭЗ, формируемая на примерах $\{O_1^+, O_2^+\}$, порождается множеством признаков $\{a_1, a_2, \ldots, a_n\}$, и при этом условие ЗКП на контрпримере O_3^- не выполняется. Таким образом, множество ХФ, формируемых на БФ, оказывается пустым.

4 Заключение

1. В работе исследован подход к построению методов описания причинности появления свойства P в БФ. Открытость множества БФ не позволяет однозначно определять причину свойства P, если не определена структура этой причины и единственность причины. Поэтому в исследовании речь идет о выявлении фактов влияния на появление ЭП. Пополнение БФ позволяет лишь делать предположения о появлении свойства P в новых фактах.
2. Для удобства анализа факторов влияния на появление ЭП определены ХФ, которые позволяют формализовать рассматриваемый подход в условиях выполнения ЗКП.
3. В терминах ХФ можно автоматизировать поиск свойства P в новых прецедентах.
4. Развитие предложенного метода предполагает его практическое использование в областях, перечисленных во введении, а также дальнейшее исследование причинно-следственных связей в условиях сложных причин и сложной структуры свойства P.

Литература

1. *Грушо А. А., Забежайло М. И., Грушо Н. А., Тимонина Е. Е.* Поиск эмпирических причин сбоев и ошибок в компьютерных системах и сетях с использованием метаданных // Системы и средства информатики, 2019. Т. 29. № 4. С. 28–38. doi: 10.14357/08696527190403.
2. *Грушо А. А., Забежайло М. И., Грушо Н. А., Тимонина Е. Е.* Архитектурные решения в задаче выявления мошенничества при анализе информационных потоков в цифровой экономике // Информатика и её применения, 2019. Т. 13. Вып. 2. С. 21–27. doi: 10.14357/19922264190204.
3. *Грушо А. А., Забежайло М. И., Грушо Н. А., Тимонина Е. Е.* Интеллектуальный анализ данных в обеспечении информационной безопасности // Проблемы информационной безопасности. Компьютерные системы, 2016. № 3. С. 55–60.
4. *Grusho A. A.* Data mining and information security // Computing network security / Eds. J. Rak, J. Bay, I. Kotenko, *et al.* — Lecture notes in computer science ser. — Springer, 2017. Vol. 10446. P. 28–33.
5. *Забежайло М. И.* Комбинаторные средства формализации эмпирической индукции: Дис. . . . докт. физ.-мат. наук. — М., 2015. 440 с.
6. *Забежайло М. И., Трунин Ю. Ю.* К проблеме доказательности медицинского диагноза: интеллектуальный анализ данных о пациентах в выборках ограниченного размера // Научно-техническая информация. Сер. 2, 2019. № 12. С. 12–18.
7. *Михеенкова М. А. и др.* ДСМ-метод в социологии: анализ данных и прогнозирование // Автоматическое порождение гипотез в интеллектуальных системах / Под общ. ред. В. К. Финна. — М.: ЛИБРОКОМ, 2009. С. 409–492.
8. *Грушо А. А., Грушо Н. А., Забежайло М. И., Смирнов Д. В., Тимонина Е. Е.* Параметризация в прикладных задачах поиска эмпирических причин // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 62–66.
9. *Финн В. К.* Индуктивные методы Д. С. Милля в системах искусственного интеллекта. Часть I // Искусственный интеллект и принятие решений, 2010. № 3. С. 3–21.
10. *Финн В. К.* Индуктивные методы Д. С. Милля в системах искусственного интеллекта. Часть II // Искусственный интеллект и принятие решений, 2010. № 4. С. 14–40.

Поступила в редакцию 12.01.20

ON CAUSAL REPRESENTATIVENESS OF TRAINING SAMPLES OF PRECEDENTS IN DIAGNOSTIC TYPE TASKS

A. A. Grusho[1], M. I. Zabezhailo[2], and E. E. Timonina[1]

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation

[2]Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The work focuses on some features of causality analysis in data mining tasks. The possibilities of using so-called open logic theories in diagnostic (classification) tasks to describe replenished sets of empirical data are discussed. In tasks of this type, it is necessary to establish (predict, diagnose, etc.) the presence or absence of a target property in a new precedent given by a description in the same presentation language of heterogeneous data, which describes examples having a target property and counter-examples not having a target property. The variant of construction of open theories describing collections of precedents by means of special logical expressions — characteristic functions — is presented. Characteristic functions allow to get rid of heterogeneity in descriptions of precedents. The procedural design of formation of characteristic functions of a training sample of precedents is proposed. The properties of characteristic functions and some conditions of their existence are studied.

Keywords: diagnostics; causal analysis; intelligent data analysis; open logic theories

DOI: 10.14357/19922264200111

Acknowledgments

The paper was partially supported by the Russian Foundation for Basic Research (project 18-29-03081).

References

1. Grusho, A. A., N. A. Grusho, M. I. Zabezhailo, and E. E. Timonina. 2019. Poisk empiricheskikh prichin sboev i oshibok v komp'yuternykh sistemakh i setyakh s ispol'zovaniem metadannykh [Search of empirical causes of failures and errors in computer systems and networks using metadata]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(4):28–38.
2. Grusho, A. A., N. A. Grusho, M. I. Zabezhailo, and E. E. Timonina. 2019. Arkhitekturnye resheniya v zadache vyyavleniya moshennichestva pri analize informatsionnykh potokov v tsifrovoy ekonomike [Architectural decisions in the problem of identification of fraud in the analysis of information flows in digital economy]. *Informatika i ee Primeneniya — Inform. Appl.* 13(2):21–27.
3. Grusho, A. A., N. A. Grusho, M. I. Zabezhailo, and E. E. Timonina. 2016. Intelligent data analysis in information security. *Autom. Control Comp. S.* 50(8):722–725.
4. Grusho, A. 2017. Data mining and information security. *Computer network security*. Eds. J. Rak, J. Bay, I. Kotenko, *et al.* Lecture notes in computer science ser. Springer. 10446:28–33.
5. Zabezhailo, M. I. 2015. Kombinatornye sredstva formalizatsii empiricheskoy induktsii [Combinatorial means of formalizing empirical induction]. Moscow. D.Sc. Diss. 440 p.
6. Zabezhailo, M. I., and Y. Y. Trunin. 2019. On the problem of medical diagnostic evidence: Intelligent analysis of empirical data on patients in samples of limited size. *Autom. Doc. Math. Linguist.* 53:322–328. doi: 10.3103/S0005105519060086.
7. Mikheenkova, M. A., *et al.* 2009. DSM-metod v sotsiologii: analiz dannykh i prognozirovanie [DSM method in sociology: Data analysis and forecasting]. *Avtomaticheskoe porozhdenie gipotez v intellektual'nykh sistemakh* [Automatic hypotheses generation in intelligent systems]. Ed. V. K. Finn. Moscow: KD LIBROKOM. 409–492.
8. Grusho, A. A., N. A. Grusho, M. I. Zabezhailo, D. V. Smirnov, and E. E. Timonina. 2018. Parametrizatsiya v prikladnykh zadachakh poiska empiricheskikh prichin [Parametrization in applied problems of search of the empirical reasons]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):62–66.
9. Finn, V. K. 2011. J. S. Mill's inductive methods in artificial intelligence systems. Part I. *Scientific Technical Information Processing* 38(6):385–302.
10. Finn, V. K. 2012. J. S. Mill's inductive methods in artificial intelligence systems. Part II. *Scientific Technical Information Processing* 39(5):241–260.

Received January 12, 2020

Contributors

Grusho Alexander A. (b. 1946) — Doctor of Science in physics and mathematics, professor, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation; grusho@yandex.ru

Zabezhailo Michael I. (b. 1956) — Doctor of Science in physics and mathematics, principal scientist, Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; m.zabezhailo@yandex.ru

Timonina Elena E. (b. 1952) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation; eltimon@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2020. Т. 14. Вып. 1. С. 87–93

ПРОИЗВОДИТЕЛЬНОСТЬ ОГРАНИЧЕННОГО КОНВЕЙЕРА

А. А. Хусаинов[1]

Аннотация: Работа посвящена изучению производительности ограниченного конвейера — вычислительного конвейера, число активных ступеней которого в каждый момент времени ограничено сверху некоторым значением. Рассмотрены ограниченные конвейеры с заданными суммой и максимумом задержек ступеней. Ступени могут иметь разные задержки. Основная задача — построение аналитической модели для расчета времени обработки заданного объема данных с помощью этого ограниченного конвейера. Решение упрощается, если ограничение рассматривать как структурный конфликт конвейера. Эта аналитическая модель построена для случая, когда работа ограниченного конвейера обладает свойством непрерывности обработки каждого входного элемента. Для таких конвейеров в работе доказана гипотеза о том, что минимальное число процессоров, при котором достигается наибольшая производительность, равно наименьшему целому числу, не меньшему отношения суммы задержек ступеней к их максимальной задержке. Установлено, что если не требовать свойства непрерывности, то эта гипотеза неверна. Построенная модель может быть применена для синхронизации работы ступеней ограниченного конвейера со свойством непрерывности. Если не требовать свойства непрерывности, то получаем асинхронный ограниченный конвейер, синхронизация работы ступеней которого осуществляется на основе готовности данных. Разработано программное обеспечение, позволяющее вычислять время обработки данных с помощью асинхронного ограниченного конвейера.

Ключевые слова: вычислительный конвейер; моноид трасс; нормальная форма Фоаты; производительность конвейера; структурный конфликт

DOI: 10.14357/19922264200112

1 Введение

Вычислительный конвейер, состоящий из p ступеней, называется *ограниченным* некоторым числом q, если в каждый момент времени могут одновременно выполняться не более чем q ступеней. В данной работе найдена формула для расчета времени обработки n входных элементов с помощью ограниченного конвейера, обладающего свойством непрерывности обработки для каждого входного элемента конвейера. С помощью этой формулы для ограниченного конвейера со свойством непрерывности подтверждена выдвинутая в [1] на основании экспериментов гипотеза о том, что минимальное число процессоров, при котором достигается наибольшая производительность ограниченного конвейера, будет равно наименьшему целому q, удовлетворяющему неравенству $q \geq \sigma/\mu$, где σ — сумма задержек ступеней конвейера, а μ — задержка самой медленной ступени. Приведен пример, показывающий, что в общем случае эта гипотеза неверна.

Проведенные исследования тесно связаны с конфликтами, возникающими при работе конвейера. Под конфликтами подразумеваются состояния, приводящие к замедлению работы конвейера. Теория конфликтов применяется при разработке конвейерных процессоров [2], сигнальных процессоров [3], сопроцессоров [4]. Существуют три типа конфликтов [5]: структурные конфликты, конфликты по данным и конфликты управления. Структурный конфликт — оборудование не может поддержать комбинацию инструкций, которые необходимо выполнить одновременно в некоторый момент времени. Ограниченный конвейер можно рассматривать как конвейер со структурным конфликтом. Аналитические модели для расчета производительности конвейеров с конфликтами построены в [2, 6, 7]. Эти модели предназначены для однородных конвейеров — конвейеров, ступени которых имеют одинаковые задержки. Заметим, что в работах [8, 9] изучались неоднородные конвейеры и был предложен метод динамического отображения конвейера (dynamic pipeline mapping) для улучшения производительности, решались задачи, где число процессоров превышает число ступеней, но проблемы, связанные с расчетом производительности ограниченных конвейеров, не были решены.

Автором в работе [10] была построена аналитическая модель для неоднородного конвейера с единственным конфликтом, вызывающим рестарт. В предлагаемой работе строится аналогичная модель для конвейера с одним конфликтом, соответствующего ограниченному конвейеру со свойством непрерывности.

[1] Комсомольский-на-Амуре государственный университет, husainov51@yandex.ru

Для оценки времени ускорения работы программы с помощью q процессоров можно использовать закон Амдала [11]. Для конвейеров существуют некоторые варианты этого закона, описанные в [12, п. 1.4.1.3]. Естественно предположить, что для ограниченного конвейера со свойством непрерывности имеет место

$$T_q(n) \approx \sigma + \frac{(n-1)\sigma}{q}\,.$$

Будет доказано, что это равенство верно с точностью до суммы задержек ступеней σ.

В разд. 2 рассмотрен однородный ограниченный конвейер. Для построения аналитической модели для него достаточно рассмотреть таблицу занятости процесса обработки n входных элементов. В разд. 3 построена и доказана формула для расчета времени обработки данных с помощью неоднородного ограниченного конвейера со свойством непрерывности. В разд. 4 описано программное обеспечение для расчета производительности асинхронных ограниченных конвейеров, синхронизация работы ступеней которых осуществляется на основе готовности данных, передаваемых между ступенями. В конце разд. 4 приведен пример, показывающий, что в общем случае гипотеза о минимальном числе процессоров конвейера неверна.

2 Однородный ограниченный конвейер

Обозначим ступени вычислительного конвейера через $a_1, \ldots, a_p$. Ступень, выполняющаяся в некоторый момент времени на некотором процессоре (функциональном устройстве) конвейера, называется *активной* в этот момент. Задержкой ступени называется время обработки ступенью одного входного элемента конвейера. Это время включает в себя логические операции и операции обмена данными с другими ступенями через входные и выходные каналы. Будем предполагать, что процессоры конвейера имеют одинаковую тактовую частоту, и измерять время в тактах.

Под *таблицей занятости* [13] конвейера будем подразумевать матрицу, строки которой соответствуют ступеням конвейера и имеют номера $1 \leq \leq i \leq p$, а столбцы — тактам времени $1, 2, 3, \ldots$ Коэффициенты этой матрицы a_{ij} равны $k \geq 1$ тогда и только тогда, когда i-я ступень обрабатывает k-й входной элемент в течение такта j. В этом случае в клетку (i, j) ставится число k. Если i-я ступень в момент j не активна, то $a_{ij} = 0$ и соответствующая клетка в таблице остается пустой.

Конвейер, состоящий из p ступеней, называется *ограниченным числом* q, если в каждый момент времени число активных ступеней не больше q.

Ограниченный конвейер имеет следующие свойства:

- в каждый момент времени активна по крайней мере одна ступень;
- в каждый момент времени активно не более q ступеней;
- перед обработкой входного элемента конвейера для каждого $i > 1$ ступень a_i ожидает окончания обработки этого входного элемента с помощью ступени a_{i-1};
- для каждого $i \geq 1$ ступень a_i ожидает окончания своего предыдущего выполнения.

Конвейер называется имеющим свойство *непрерывности* (работы), если для всякого входного элемента разность между временем конца обработки и временем начала обработки этого элемента равна сумме задержек ступеней конвейера. В частности, свойством непрерывности обладает однородный конвейер — конвейер, все ступени которого имеют одинаковые задержки, равные некоторому числу h.

Обозначим через p число его ступеней. Если нет конфликтов, то время обработки n элементов равно $(p+n-1)h$. Пусть однородный конвейер ограничен числом q, $1 \leq q \leq p$. На вход конвейера поступает n элементов входных данных. Таблица 1 показывает занятость конвейера при $p = 4$, $q = 3$ и $n = 5$. При попытке запустить больше чем q параллельно работающих ступеней возникает структурный конфликт, в результате которого каждая ступень будет ожидать освобождения одного из процессоров и время работы этой ступени увеличится на $(p-q)h$.

Таблица 1 Однородный ограниченный конвейер

	01	02	03	04	05	06	07	08	09	10
1	1	2	3		4	5				
2		1	2	3		4	5			
3			1	2	3		4	5		
4				1	2	3		4	5	

Учитывая случай $p < q$, приходим к следующему утверждению.

Предложение 1 [14, Prop. 1]. *Время обработки n элементов с помощью q процессоров для однородного конвейера из p ступеней с задержкой h равно*

$$T_q(n) = \left(p + n - 1 + (p-q)^+ \left[\frac{n-1}{q}\right]\right) h\,. \qquad (1)$$

Здесь $[x]$ обозначает целую часть вещественного числа x, а $(x)^+$ означает число, равное x, если $x \geq 0$, и равное 0 в случае $x \leq 0$.

Эта формула была применена в [14] для расчета оптимальной глубины однородного ограниченного конвейера.

3 Неоднородный ограниченный конвейер

В данном разделе всюду, где не оговорено противное, неоднородный ограниченный конвейер будет обладать свойством непрерывности.

Ступени неоднородного конвейера могут иметь разные задержки $\tau_1, \ldots, \tau_p$. Время обработки n элементов равно

$$T_p(n) = \sigma + (n-1)\mu\,,$$

где $\sigma = \sum_{i=1}^{p} \tau_i$ — сумма, а $\mu = \max\{\tau_i | 1 \leq i \leq p\}$ — максимум задержек ступеней конвейера. Правую часть этой формулы можно получить из времени обработки для однородного конвейера $(p+n-1)h$, подставляя вместо p отношение σ/μ, а вместо h — максимальную задержку ступеней μ. Это приводит к предположению о том, что аналогичным образом из формулы (1) может быть получена формула для времени обработки с помощью ограниченного неоднородного конвейера. Из этой формулы придется удалить слагаемые, для которых $\sigma/\mu - q < 0$. Сформулируем и докажем полученное утверждение.

Теорема 1. *Время обработки n элементов с помощью неоднородного конвейера со свойством непрерывности, ограниченного числом q и состоящего из p ступеней, равно*

$$T_q(n) = \sigma + (n-1)\mu + (\sigma - q\mu)^+ \left[\frac{n-1}{q}\right]. \quad (2)$$

Доказательство. Обычный конвейер обрабатывает n элементов за время $T_p(n) = \sigma + (n-1)\mu$. В случае, когда активны $q < p$ ступеней, к этому времени добавляется время ожидания свободных процессоров. Это время ожидания называется штрафным. Рассмотрим таблицу занятости при обработке n элементов. Первые q элементов будут обрабатываться без лишних торможений. Они будут обработаны за время $T(q) = \sigma + (q-1)\mu$. При попытке обработать $(q+1)$-й элемент возникает (структурный) конфликт, связанный с тем, что число одновременно работающих ступеней не должно быть больше чем q. Этот конфликт будет разрешен после окончания обработки первого элемента, ибо в этом случае появится свободный процессор. В силу свойства непрерывности отсюда вытекает, что обработку $(q+1)$-го элемента можно начать в момент времени σ. Время обработки $q+1$ элементов будет равно 2σ. Поскольку для обычного конвейера время обработки $q+1$ элементов равно $\sigma + q\mu$, то штрафное время будет равно $2\sigma - (\sigma + q\mu) = \sigma - q\mu$. Эти конфликты возникают при обработке элементов с номерами $q+1, 2q+1, \ldots, mq+1$, где m — наибольшее целое, для которого $mq+1 \leq n$. Ясно, что $m = [(n-1)/q]$. При обработке остальных элементов конфликты не возникают. Отсюда вытекает, что если $\sigma - q\mu \geq 0$, то $T_q(n) = \sigma + (n-1)\mu + m(\sigma - q\mu)$. Если же $\sigma - q\mu \leq 0$, то в момент времени $q\mu$ начала обработки $(q+1)$-го элемента первый элемент будет обработан и закончит занимать один из процессоров. В этом случае штрафное время будет равно нулю и $T_q(n) = \sigma + (n-1)\mu$. Теорема 1 доказана.

Из доказанной формулы (2) вытекает, что при $n-1 \geq q$ время обработки n элементов будет минимальным тогда и только тогда, когда имеет место неравенство $\sigma - q\mu \leq 0$. Это приводит к следующему утверждению.

Следствие 1. Минимальное число процессоров, при котором достигается наибольшая производительность ограниченного конвейера со свойством непрерывности, равно наименьшему целому q, удовлетворяющему неравенству $q \geq \sigma/\mu$. В этом случае время обработки равно $\sigma + (n-1)\mu$.

Отсюда вытекает, что предположение, выдвинутое в работе [1], верно для ограниченных конвейеров со свойством непрерывности.

Обозначим через $T_q^A(n)$ следующую оценку для производительности конвейера:

$$T_q^A(n) = \begin{cases} \left(1 + \dfrac{n-1}{q}\right)\sigma\,, & \text{если } q \leq \dfrac{\sigma}{\mu}\,; \\ \sigma + (n-1)\mu\,, & \text{если } q \geq \dfrac{\sigma}{\mu}\,. \end{cases}$$

Следствие 2. Имеют место неравенства: $0 \leq T_q^A(n) - T_q(n) < \sigma$.

Доказательство. Пусть $\sigma \geq q\mu$. Воспользуемся тем, что $n-1-q\,[(n-1)/q]$ равно остатку $(n-1) \bmod q$ от деления числа $n-1$ на q. Из теоремы 1 вытекает, что

$$T_q(n) = \sigma + (n-1)\mu + (\sigma - q\mu)\left[\frac{n-1}{q}\right] = \sigma\left(1 + \left[\frac{n-1}{q}\right]\right) + \mu\left((n-1) \bmod q\right).$$

Следовательно,

$$T_q^A(n) - T_q(n) = (\sigma - q\mu)\frac{(n-1)\bmod q}{q} < \sigma\,.$$

Пусть $S_q(n) = T_1(n)/T_q(n)$ — ускорение вычисления с помощью ограниченного конвейера. Обозначим $S_q = \lim_{n\to\infty} S_q(n)$. Рассматривая случаи $q < \sigma/\mu$ и $q \geq \sigma/\mu$, получаем

Следствие 3. Для конвейера со свойством непрерывности, ограниченного числом $q > 1$, имеет место равенство $S_q = \min(q, \sigma/\mu)$.

4 Асинхронный ограниченный конвейер

Попытаемся сравнить производительность ограниченного конвейера со свойством непрерывности с производительностью асинхронного ограниченного конвейера, синхронизация работы ступеней которого осуществляется на основе готовности данных, передаваемых между ступенями. Но не ясно, как вычислять производительность асинхронного конвейера. Возможный ответ дает компьютерная программа, которой посвящен данный раздел. Эта программа для введенных пользователем числовых значений задержек ступеней асинхронного ограниченного конвейера и объема данных вычисляет значения времени обработки данных в зависимости от числа процессоров и выводит эти значения в виде графиков. Она создает также и сохраняет в файл таблицы занятости конвейера при различных числах процессоров.

Программа основана на методе, предложенном Дикертом [15] и использующем теорию трасс — слов, состоящих из букв алфавита, на котором задано антирефлексивное симметричное бинарное отношение независимости.

Опишем этот метод. Рассмотрим множество операций (машинных команд) $A = \{a_0, \ldots, a_{m-1}\}$ и отношение *независимости* $I \subseteq A^2$, состоящее из пар операций (a_i, a_j), которые могут выполняться одновременно. Каждое слово можно интерпретировать как процесс, состоящий из команд, принадлежащих этому слову. Если команды могут выполняться одновременно, то их можно переставлять между собой. Два слова, составленные из букв алфавита A, определим как *эквивалентные*, если одно из них можно получить из другого с помощью последовательности перестановок рядом стоящих независимых букв. *Трассой* называется класс эквивалентности множества A^* всех слов по этому отношению эквивалентности. Для произвольного слова $w \in A^*$ обозначим через $[w]$ его класс эквивалентности. Определим *композицию* по формуле $[w_1][w_2] = [w_1 w_2]$. Операция композиции превращает множество классов эквивалентности в моноид, который называется *моноидом трасс*. Идея алгоритма вычисления времени работы параллельного процесса описана в [15]. Пусть время выполнения каждой команды, принадлежащей алфавиту A, равно одному такту. Каждая трасса может быть представлена в виде последовательности максимальных блоков (ярусов) параллельно выполняющихся команд. Это представление называется *нормальной формой Фоаты*. Число блоков называется *высотой* нормальной формы, и оно равно времени выполнения трассы.

Аналогично нормальной форме Фоаты введем *нормальную форму относительно числа* $q \geq 1$ как состоящую из последовательности блоков, длины которых не превышают число q. Для этой цели модифицируем алгоритм приведения к нормальной форме Фоаты и за определение возьмем результат этого алгоритма.

Опишем алгоритм. Пусть на входе задано некоторое непустое слово $w \in A^*$. Считываем из него первый символ и нормальную форму полагаем равной одному блоку, состоящему из этого символа. Далее в цикле считываем очередной символ x и для него выполняем 3 действия:

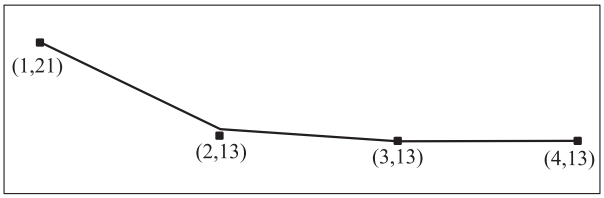
(1) ищем блок с наименьшим номером $k \geq 1$, такой что все элементы блоков, имеющих номера $\geq k$, независимы от x. Если таких k нет, то добавляем новый блок, содержащий единственный элемент x, и переходим к следующему символу;

(2) цикл: пока k-й блок имеет q элементов, увеличиваем k на 1;

(3) если k остался не больше номера последнего блока, то добавляем x в k-й блок. Иначе добавляем новый блок, состоящий из элемента x.

Число слов, из которых состоит нормальная форма относительно q, называется ее *высотой относительно* q. Если каждая буква обозначает операцию, время выполнения которой равно единице измерения, и независимые операции могут выполняться параллельно, то время выполнения операций слова w будет равно высоте нормальной формы этого слова относительно q.

Для того чтобы находить время выполнения процесса обработки асинхронным ограниченным конвейером, зададим алфавит $A = \{a_0, \ldots, a_{m-1}\}$, состоящий из $m = 3p$ букв. Слово w, соответствующее обработке n элементов конвейером, задается следующим образом. Всякая ступень с номером i разбивается на полутакты и записывается как слово $a_{3i}a_{3i+1}^{2\tau_i-2}a_{3i+2}$. Здесь a_{3i} — начальный полутакт,



Сравнение асинхронной и непрерывной обработки

а a_{3i+2} — последний полутакт ступени. Между ними находятся полутакты, выполняющиеся последовательно, и их можно обозначить одинаковой буквой a_{3i+1}. Слово w будет равно

$$\left(a_0 a_1^{2\tau_0-2} a_2 a_3 a_4^{2\tau_1-2} a_5 \cdots a_{3(p-1)} a_{3p-2}^{2\tau_{p-1}-2} a_{3p-1}\right)^n .$$

Отношение независимости состоит из пар полутактов, которые могут выполняться параллельно:

$$I = A^2 \setminus \Big(\{a_0, a_1, a_2\}^2 \cup \{a_3, a_4, a_5\}^2 \cup \cdots \\ \cdots \cup \{a_{3p-3}, a_{3p-2}, s_{3p-1}\}^2 \cup \\ \cup \{a_2, a_3\}^2 \cup \{a_5, a_6\}^2 \cup \cdots \cup \{a_{3p-4}, a_{3p-3}\}^2 \Big) .$$

Нормальная форма относительно q дает последовательность параллельно работающих блоков (ярусов) операций. Поскольку время выполнения операции равно полутакту, то время обработки n элементов с помощью ограниченного конвейера будет равно половине высоты этой нормальной формы.

Рассмотрим пример работы программы. На рисунке крупными точками показан график зависимости времени обработки трех входных элементов с помощью асинхронного конвейера, ограниченного числом процессоров X. Задержки ступеней равны $\tau_0 = 1$, $\tau_1 = 3$, $\tau_2 = 1$, $\tau_3 = 2$. График, полученный по формуле (2), изображен в виде ломаной линии. По формуле (2) время $T_2(3) = 14$. Крупная точка $(2, 13)$ показывает, что для асинхронного конвейера это время равно 13. Приведенный пример показывает также, что для асинхронного конвейера следствие 1 неверно, поскольку наименьшее целое q, удовлетворяющее $q \geq \sigma/\mu$, равно 3, а минимальное число процессоров, при котором достигается наибольшая производительность, равно 2.

Таблица 2 представляет собой таблицу занятости для этого примера.

5 Заключение

Получена формула для расчета производительности ограниченного конвейера (теорема 1). Вытекающее из нее следствие 1 говорит о том, что для некоторых конвейеров число функциональных устройств можно уменьшить и это не приведет к снижению производительности. Это можно применять для совершенствования архитектуры процессора. Другим возможным продолжением данной работы может стать развитие метода, подсказавшего формулировку теоремы 1. Он позволяет обобщать некоторые утверждения об однородных конвейерах на неоднородные.

В последнее время конвейеры широко применяются при разработке многопоточных приложений для облачных вычислений. Многопоточный конвейер, работающий на компьютере с многоядерным процессором, ограничен числом процессорных ядер. В этом случае работа каждого процессорного ядра сопровождается переключением контекста нити, что приводит к замедлению работы конвейера. Аналогичная проблема возникает при реализации ограниченных конвейеров для многоядерных сигнальных процессоров и вычислительных систем с массовым параллелизмом — появляются дополнительные накладные расходы, связанные с общим управлением ядрами в рамках одного процессора. Полученная в работе формула для расчета производительности не учитывает такого замедления. Построение аналитической модели, учитывающей эти накладные расходы, было бы одним из перспективных продолжений данной работы.

Литература

1. *Хусаинов А. А., Чернов А. М., Маевская Е. Д., Романченко А. А.* Модели для расчета времени работы вычисли-

Таблица 2 Работа асинхронного ограниченного конвейера

	01	02	03	04	05	06	07	08	09	10	11	12	13	14
1	1	2	3											
2		1	1	1	2	2	2	3	3	3				
3					1			2			3			
4						1	1		2	2		3	3	

тельных конвейеров // Актуальные проблемы науки: Мат-лы XXIII Междунар. научно-практич. конф. — М.: Спутник+, 2016. С. 83–91.

2. *Emma P. G., Davidson E. S.* Characterization of branch and data dependencies in programs for evaluating pipeline performance // IEEE T. Comput., 1987. Vol. 7. P. 859–875.
3. *Cheah H. Y., Fahmy S. A., Kapre N.* On data forwarding in deeply pipelined soft processors // ACM/SIGDA Symposium (International) on Field-Programmable Gate Arrays Proceedings. — New York, NY, USA: ACM, 2015. P. 181–189.
4. *Merchant F., Chattopadhyay A., Raha S., Nandy S. K., Narayan R.* Accelerating BLAS and LAPACK via efficient floating point architecture design // Parallel Process. Lett., 2017. Vol. 27. No. 03n04. P. 1–7.
5. *Паттерсон Д., Хеннесси Дж.* Архитектура компьютера и проектирование компьютерных систем / Пер. с англ. — СПб.: Питер, 2012. 784 с. (*Patterson D. A., Hennessy J. L.* Computer organization and design. — 4th ed. — Amsterdam: Elsevier, 2012. 703 p.)
6. *Hartstein A., Puzak T. R.* The optimum pipeline depth for a microprocessor // ACM Comp. Ar., 2002. Vol. 30. Iss. 2. P. 7–13.
7. *Yao J., Miwa S., Shimada H.* Optimal pipeline depth with pipeline stage unification adoption // ACM Comp. Ar., 2007. Vol. 35. Iss. 5. P. 3–9.
8. *Moreno A., César E., Guevara A., Sorribes J., Margalef T.* Load balancing in homogeneous pipeline based applications // Parallel Comput., 2012. Vol. 38. Iss. 3. P. 125–139.
9. *Moreno A., Sikora A., César E., Sorribes J., Margalef T.* HeDPM: Load balancing of linear pipeline applications on heterogeneous systems // J. Supercomput., 2017. Vol. 73. Iss. 9. P. 3738–3760.
10. *Хусаинов А. А., Титова Е. А.* Оптимальная глубина вычислительного конвейера при заданном объеме данных // Вычислительные технологии, 2018. Т. 23. № 1. С. 96–104.
11. *Amdahl G. M.* Validity of the single processor approach to achieving large scale computing capabilities // AFIPS Spring Joint Computer Conference Proceedings. — New York, NY, USA: ACM, 1967. P. 483–485.
12. *Shen J. P., Lipasti M. H.* Model processor design: Fundamental of superscalar processors. — New York, NY, USA: McGraw-Hill, 2005. 643 p.
13. *Коуги П. М.* Архитектура конвейерных ЭВМ / Пер. с англ. — М.: Радио и связь, 1985. 360 с. (*Kogge P. M.* The architecture of pipelined computers. — Washington, D.C., USA: McGraw-Hill, 1981. 335 p.)
14. *Husainov A. A.* Optimum depth of the bounded pipeline. — New York, NY, USA: Cornell University, 2018. Preprint. 11 p. http://arxiv.org/abs/cs.DC/1807.11022v1.
15. *Diekert V.* Combinatorics on traces. — Lecture notes in computer science ser. — Berlin: Springer-Verlag, 1990. Vol. 454. 169 p.

Поступила в редакцию 30.08.19

PERFORMANCE OF THE BOUNDED PIPELINE

A. A. Khusainov

Komsomolsk-na-Amure State University, 27 Lenina Prosp., Komsomolsk-on-Amur, Khabarovsk Region 681013, Russian Federation

Abstract: The paper is devoted to studying the performance of a bounded pipeline that is a computational pipeline, the number of active stages of which is bounded at any time by a fixed number. The bounded pipelines with the given sum and the maximum of delays of stages are considered. The stages can have different delays. The main problem is to build an analytical model for calculating the processing time of a given amount of data using this bounded pipeline. The solution is simplified if the constraint is treated as a structural pipeline hazard. This analytical model is constructed for the case when the operation of a bounded pipeline has the property of continuity of processing for each input element. For such pipelines, the conjecture is proved in the paper that the minimum number of processors at which the greatest productivity is achieved is equal to the smallest integer not less than the ratio of the sum of stage delays to the maximum delay. It is established that if the property of continuity is not required, then this conjecture is not true. The constructed model can be used to synchronize the operation of the stages of a bounded pipeline with the continuity property. If we do not require the property of continuity, then we get an asynchronous bounded pipeline, the synchronization of the work for the stages is carried out on the basis of the data readiness. The software is developed, which is based on the theory of trace monoids and allows one to calculate the processing time with an asynchronous bounded pipeline.

Keywords: computational pipeline; trace monoid; Foata normal form; pipeline performance; structural hazard

DOI: 10.14357/19922264200112

References

1. Khusainov, A. A., A. M. Chernov, E. D. Mayevskaya, and A. A. Romanchenko. 2016. Modeli dlya rascheta vremeni raboty vychislitel'nykh konveyerov [Models for calculating the operating time of computational pipelines]. *Aktual'nyye problemy nauki: Mat-ly XXIII Mezhdunar. nauchno-praktich. konf.* [23rd Conference (International) on Actual Problems of Science Proceedings]. 83–91.
2. Emma, P. G., and E. S. Davidson. 1987. Characterization of branch and data dependencies in programs for evaluating pipeline performance. *IEEE T. Comput.* 7:859–875.
3. Cheah, H. Y., S. A. Fahmy, and N. Kapre. 2015. On data forwarding in deeply pipelined soft processors. *ACM/SIGDA International Symposium on Field-Programmable Gate Arrays Proceedings*. New York, NY: ACM. 181–189.
4. Merchant, F., A. Chattopadhyay, S. Raha, S. K. Nandy, and R. Narayan. 2017. Accelerating BLAS and LAPACK via efficient floating point architecture design. *Parallel Process. Lett.* 27(03n04):1–7.
5. Patterson, D. A., and J. L. Hennessy. 2012. Computer organization and design: The hardware/software interface. 4th ed. Amsterdam: Elsevier. 703 p.
6. Hartstein, A., and T. R. Puzak. 2002. The optimum pipeline depth for a microprocessor. *ACM Comp. Ar.*. 30(2):7–13.
7. Yao, J., S. Miwa, and H. Shimada. 2007. Optimal pipeline depth with pipeline stage unification adoption. *ACM Comput. Ar.* 35(5):3–9.
8. Moreno, A., E. César, A. Guevara, J. Sorribes, and T. Margalef. 2012. Load balancing in homogeneous pipelinebased applications. *Parallel Comput.* 38(3):125–139.
9. Moreno, A., A. Sikora, E. César, J. Sorribes, and T. Margalef. 2017. HeDPM: Load balancing of linear pipeline applications on heterogeneous systems. *J. Supercomput.* 73(9):3738–3760.
10. Husainov, A. A., and E. A. Titova. 2018. Optimal'naya glubina vychislitel'nogo konveyera pri zadannom ob"eme dannykh [Optimal depth of the computational pipeline for a given amount of input data]. *Vychislitel'nyye tekhnologii* [Computational Technologies] 23(1):96–104.
11. Amdahl, G. M. 1967. Validity of the single processor approach to achieving large scale computing capabilities. *AFIPS Spring Joint Computer Conference Proceedings*. New York, NY: ACM. 483–485.
12. Shen, J. P., and M. H. Lipasti. 2005. *Model processor design: Fundamental of superscalar processors*. New York, NY: McGraw-Hill. 643 p.
13. Kogge, P. M. 1981. *The architecture of pipelined computers*. Washington, D.C.: McGraw-Hill. 335 p.
14. Husainov, A. A. 2018. Optimum depth of the bounded pipeline. arXiv 1807.11022 v1[cs.DC]. Available at: http://arxiv.org/abs/cs.DC/1807.11022v1 (accessed December 27, 2019).
15. Diekert, V. 1990. *Combinatorics on traces*. Lecture notes in computer science ser. Berlin: Springer-Verlag. Vol. 454. 169 p.

Received August 30, 2019

Contributor

Khusainov Akhmet A. (b. 1951) — Doctor of Science in physics and mathematics, professor, Komsomolsk-na-Amure State University, 27 Lenina Prosp., Komsomolsk-on-Amur, Khabarovsk Region 681013, Russian Federation; husainov51@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2020. Т. 14. Вып. 1. С. 94–100

МЕТОД ЗАДАНИЯ КОНЕЧНЫХ НЕКОММУТАТИВНЫХ АССОЦИАТИВНЫХ АЛГЕБР ПРОИЗВОЛЬНОЙ ЧЕТНОЙ РАЗМЕРНОСТИ ДЛЯ ПОСТРОЕНИЯ ПОСТКВАНТОВЫХ КРИПТОСХЕМ

А. А. Костина[1], А. Ю. Мирин[2], Д. Н. Молдовян[3], Р. Ш. Фахрутдинов[4]

Аннотация: Представлен новый унифицированный метод задания конечных некоммутативных ассоциативных алгебр (КНАА) произвольной четной размерности m и описаны исследуемые свойства алгебр для случаев $m = 4$ и 6 при задании алгебр над конечным простым полем $GF(p)$ с большим размером простого числа p. Получены формулы, описывающие множество p^2 (p^4) глобальных левосторонних единиц, содержащихся в 4-мерной (6-мерной) алгебре. В исследованных алгебрах имеет место только локальная обратимость. Для каждой из алгебр выведены формулы для вычисления единственного локального двустороннего элемента, связанного с фиксированным локально обратимым вектором. Новая форма скрытой задачи дискретного логарифмирования (СЗДЛ) предложена в качестве постквантового криптографического примитива и использована для разработки постквантовой схемы электронной цифровой подписи (ЭЦП).

Ключевые слова: конечная некоммутативная алгебра; ассоциативная алгебра; вычислительно трудная задача; дискретный логарифм; цифровая подпись; постквантовая криптография

DOI: 10.14357/19922264200113

1 Введение

Актуальным текущим вызовом в области криптографии стала разработка криптосхем с открытым ключом, пригодных для принятия на их основе постквантовых криптографических стандартов [1, 2] взамен текущих стандартов, основанных на вычислительной трудности задачи дискретного логарифмирования (ЗДЛ), которая решается на пока еще гипотетическом квантовом компьютере за полиномиальное время [3]. Перспективным направлением решения указанной проблемы представляется использование вычислительной трудности СЗДЛ, задаваемой в КНАА [4]. Для реализации потенциала СЗДЛ как постквантового криптографического примитива важное значение имеют задачи поиска и исследования новых КНАА как носителей СЗДЛ и новых форм последней, для которых СЗДЛ не будет сводиться к ЗДЛ в конечном поле [5].

В настоящей работе предлагается общий способ задания КНАА произвольной четной размерности $m > 2$ и исследуются свойства 4- и 6-мерных КНАА, построенных в соответствии с предложенным общим способом. Характерным свойством рассмотренных алгебр является наличие большого множества глобальных левосторонних единиц, и для реализации на их основе постквантовых криптосхем предлагается новая форма СЗДЛ, использующая указанную особенность примененных в качестве ее алгебраического носителя КНАА.

2 Задача дискретного логарифмирования в скрытой группе

Известно автоморфное отображение некоммутативной группы Г, задаваемое следующей формулой:

$$\varphi_V(G) = V^{-1} \circ G \circ V\,,$$

где V — фиксированный элемент группы Г; G — элемент, пробегающий все значения в группе Г. Элементы G и $Y = V^{-1} \circ G \circ V$ называются элементами, сопряженными через элемент V. Для фиксированного значения G операция автоморфного отображения и операция возведения в сте-

[1]Санкт-Петербургский институт информатики и автоматизации Российской академии наук, anna-kostina1805@mail.ru
[2]Санкт-Петербургский институт информатики и автоматизации Российской академии наук, mirin@cobra.ru
[3]Санкт-Петербургский институт информатики и автоматизации Российской академии наук, mdn.spectr@mail.ru
[4]Санкт-Петербургский институт информатики и автоматизации Российской академии наук, fahr@cobra.ru

пень являются перестановочными (взаимно коммутативными), т. е. имеет место соотношение:

$$\left(V^{-1}\circ G\circ V\right)^x = V^{-1}\circ (G^x)\circ V\,.$$

Это свойство может быть использовано для формулирования следующей вычислительно трудной задачи, пригодной для использования в качестве примитива криптосистем с открытым ключом.

Пусть задан некоторый элемент G. Из некоторой коммутативной подгруппы $\Gamma_{\text{комм}}\subset\Gamma$ выбираются элемент X и произвольное число x и вычисляется элемент $Y = X^{-1}\circ(G^x)\circ X$. По заданным Y и G требуется вычислить X и x. Поскольку вычислить эти два неизвестных элемента по отдельности нельзя, то эта задача в общем случае не сводится к задаче дискретного логарифмирования в циклической подгруппе. Нахождение неизвестных X и x по значениям Y и G представляет собой самостоятельную трудную вычислительную задачу, отличную от задачи дискретного логарифмирования. При известном значении X можно вычислить $Y' = X\circ Y\circ X^{-1}$ или $G' = X^{-1}\circ G\circ X$, после чего число x можно найти из уравнения $Y' = G^x$ или из уравнения $Y = G'^x$ соответственно, т. е. решая задачу дискретного логарифмирования. Однако значение X остается неизвестным, поэтому ЗДЛ в явном виде не стоит. Криптосхемы на основе СЗДЛ, заданной в этой форме, описаны в работе [4].

3 Конечные некоммутативные ассоциативные алгебры

3.1 Алгебры как векторные пространства с дополнительной операцией умножения векторов

Рассмотрим m-мерное векторное пространство, элементами которого выступают всевозможные векторы вида

$$A = (a_0, a_1, \ldots, a_1) = \\ = \left(a_0\mathbf{e}_0 + a_1\mathbf{e}_1 + \cdots + a_{m-1}\mathbf{e}_{m-1}\right),$$

где $a_i\in GF(p)$, p — простое число; $\mathbf{e}_i$ — формальные базисные векторы. Дополнительно к стандартным операциям в векторном пространстве — операции сложения векторов и операции умножения вектора на скаляр — определим операцию умножения ($\circ$) векторов $A = \sum_{i=0}^{m-1} a_i\mathbf{e}_i$ и $B = \sum_{j=0}^{m-1} b_j\mathbf{e}_j$ в соответствии со следующей формулой:

$$A\circ B = \sum_{i=0}^{m-1}\sum_{j=0}^{m-1} a_ib_j\mathbf{e}_i\circ\mathbf{e}_j\,, \qquad (1)$$

где каждое из всевозможных произведений пар базисных векторов заменяется на однокомпонентный вектор в соответствии с некоторым правилом, задаваемым в виде таблицы умножения базисных векторов (ТУБВ). Векторное пространство с определенной таким образом операцией умножения векторов называется m-мерной алгеброй. Для построения криптосхем на основе СЗДЛ интерес представляют КНАА.

3.2 Общий способ задания конечных некоммутативных ассоциативных алгебр произвольной четной размерности $m\geq 4$

В качестве общего способа задания конечной ассоциативной алгебры четной размерности $m>1$ предлагается задать значение произведения $\mathbf{e}_i\circ\mathbf{e}_j$ в формуле (1) в соответствии со следующим выражением:

$$\mathbf{e}_i\circ\mathbf{e}_j = \begin{cases}\mathbf{e}_j\,, & \text{если } i \bmod 2 = 0\,;\\ \mathbf{e}_{m-1-j}\,, & \text{если } i\bmod 2 = 1\,.\end{cases} \qquad (2)$$

Докажем, что правило (2) задает ассоциативное умножение векторов. Рассмотрим произведение векторов A, B и $C = \sum_{k=0}^{m-1} c_k\mathbf{e}_k$, осуществляемое в соответствии со следующими двумя вариантами:

$$(A\circ B)\circ C = \\ = \left(\sum_{i=0}^{m-1}\sum_{j=0}^{m-1} a_ib_j\mathbf{e}_i\circ\mathbf{e}_j\right)\circ\sum_{k=0}^{m-1} c_k\mathbf{e}_k = \\ = \sum_{i=0}^{m-1}\sum_{j=0}^{m-1}\sum_{k=0}^{m-1} a_ib_jc_k\left(\mathbf{e}_i\circ\mathbf{e}_j\right)\circ\mathbf{e}_k\,; \qquad (3)$$

$$A\circ(B\circ C) = \\ = \left(\sum_{i=0}^{m-1} a_i\mathbf{e}_i\right)\circ\left(\sum_{j=0}^{m-1}\sum_{k=0}^{m-1} b_jc_k\mathbf{e}_i\circ\mathbf{e}_j\right) = \\ = \sum_{i=0}^{m-1}\sum_{j=0}^{m-1}\sum_{k=0}^{m-1} a_ib_jc_k\mathbf{e}_i\circ\left(\mathbf{e}_j\circ\mathbf{e}_k\right)\,. \qquad (4)$$

Равенство правых частей выражений (3) и (4) имеет место, если равенство

$$\left(\mathbf{e}_i\circ\mathbf{e}_j\right)\circ\mathbf{e}_k = \mathbf{e}_i\circ\left(\mathbf{e}_j\circ\mathbf{e}_k\right) \qquad (5)$$

справедливо для всех возможных троек значений (i,j,k). Справедливость равенства (5) можно легко показать, рассматривая следующие четыре случая:

Случай 1: i и j — четные значения:

$$\left.\begin{matrix}(\mathbf{e}_i \circ \mathbf{e}_j) \circ \mathbf{e}_k = \mathbf{e}_j \circ \mathbf{e}_k = \mathbf{e}_k \\ \mathbf{e}_i \circ (\mathbf{e}_j \circ \mathbf{e}_k) = \mathbf{e}_i \circ \mathbf{e}_k = \mathbf{e}_k\end{matrix}\right\} \Rightarrow$$
$$\Rightarrow (\mathbf{e}_i \circ \mathbf{e}_j) \circ \mathbf{e}_k = \mathbf{e}_i \circ (\mathbf{e}_j \circ \mathbf{e}_k)\ .$$

Случай 2: i — четное значение; j — нечетное значение:

$$\left.\begin{matrix}(\mathbf{e}_i \circ \mathbf{e}_j) \circ \mathbf{e}_k = \mathbf{e}_j \circ \mathbf{e}_k = \mathbf{e}_{m-1-k} \\ \mathbf{e}_i \circ (\mathbf{e}_j \circ \mathbf{e}_k) = \mathbf{e}_i \circ \mathbf{e}_{m-1-k} = \mathbf{e}_{m-1-k}\end{matrix}\right\} \Rightarrow$$
$$\Rightarrow (\mathbf{e}_i \circ \mathbf{e}_j) \circ \mathbf{e}_k = \mathbf{e}_i \circ (\mathbf{e}_j \circ \mathbf{e}_k)\ .$$

Случай 3: i — нечетное значение; j — четное значение:

$$\left.\begin{matrix}(\mathbf{e}_i \circ \mathbf{e}_j) \circ \mathbf{e}_k = \mathbf{e}_{m-1-j} \circ \mathbf{e}_k = \mathbf{e}_{m-1-k} \\ \mathbf{e}_i \circ (\mathbf{e}_j \circ \mathbf{e}_k) = \mathbf{e}_i \circ \mathbf{e}_k = \mathbf{e}_{m-1-k}\end{matrix}\right\} \Rightarrow$$
$$\Rightarrow (\mathbf{e}_i \circ \mathbf{e}_j) \circ \mathbf{e}_k = \mathbf{e}_i \circ (\mathbf{e}_j \circ \mathbf{e}_k)\ .$$

Случай 4: i и j — нечетные значения:

$$\left.\begin{matrix}(\mathbf{e}_i \circ \mathbf{e}_j) \circ \mathbf{e}_k = \mathbf{e}_{m-1-j} \circ \mathbf{e}_k = \mathbf{e}_k \\ \mathbf{e}_i \circ (\mathbf{e}_j \circ \mathbf{e}_k) = \mathbf{e}_i \circ \mathbf{e}_{m-1-k} = \\ = \mathbf{e}_{m-1-(m-1-k)} = \mathbf{e}_k\end{matrix}\right\} \Rightarrow$$
$$\Rightarrow (\mathbf{e}_i \circ \mathbf{e}_j) \circ \mathbf{e}_k = \mathbf{e}_i \circ (\mathbf{e}_j \circ \mathbf{e}_k)\ .$$

Таким образом, операция умножения векторов, задаваемая правилом (2), ассоциативна, а при $m \geq 4$ также и некоммутативна.

3.3 Четырехмерная алгебра

Рассмотрим случай 4-мерной КНАА, для которой найдено распределение структурного коэффициента $\mu \in GF(p)$, представленное в табл. 1.

Нахождение левосторонних единиц 4-мерной алгебры, задаваемой табл. 1, связано с решением следующего векторного уравнения:

$$X \circ A = A\,, \tag{6}$$

где $A = (a_0, a_1, a_2, a_3)$ — некоторый заданный вектор, для которого требуется найти левостороннюю единицу; $X = (x_0, x_1, x_2, x_3)$ — неизвестный вектор. С учетом табл. 1 данное векторное уравнение сводится к решению следующей системы линейных уравнений с четырьмя неизвестными:

$$\left.\begin{matrix}\mu x_0 a_0 + x_1 a_3 + x_2 a_0 + \mu x_3 a_3 = a_0\,; \\ \mu x_0 a_1 + x_1 a_2 + x_2 a_1 + \mu x_3 a_2 = a_1\,; \\ \mu x_0 a_2 + x_1 a_1 + x_2 a_2 + \mu x_3 a_1 = a_2\,; \\ \mu x_0 a_3 + x_1 a_0 + x_2 a_3 + \mu x_3 a_0 = a_3\,.\end{matrix}\right\} \tag{7}$$

Таблица 1 Предлагаемая ТУБВ для случая $m = 4$

$\circ$	$\mathbf{e}_0$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$
$\mathbf{e}_0$	$\mu\mathbf{e}_0$	$\mu\mathbf{e}_1$	$\mu\mathbf{e}_2$	$\mu\mathbf{e}_3$
$\mathbf{e}_1$	$\mathbf{e}_3$	$\mathbf{e}_2$	$\mathbf{e}_1$	$\mathbf{e}_0$
$\mathbf{e}_2$	$\mathbf{e}_0$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$
$\mathbf{e}_3$	$\mu\mathbf{e}_3$	$\mu\mathbf{e}_2$	$\mu\mathbf{e}_1$	$\mu\mathbf{e}_0$

Эта система распадается на следующие две независимые системы из двух уравнений:

$$\left.\begin{matrix}\left.\begin{matrix}(\mu x_0 + x_2)a_0 + (x_1 + \mu x_3)a_3 = a_0\,; \\ (x_1 + \mu x_3)a_0 + (\mu x_0 + x_2)a_3 = a_3\,;\end{matrix}\right\} \\ \left.\begin{matrix}(\mu x_0 + x_2)a_1 + (x_1 + \mu x_3)a_2 = a_1\,; \\ (x_1 + \mu x_3)a_1 + (\mu x_0 + x_2)a_2 = a_2\,.\end{matrix}\right\}\end{matrix}\right. \tag{8}$$

Выполнив в (8) замену переменных по формулам $z_1 = \mu x_0 + x_2$ и $z_1 = z_2 + \mu x_3$, легко показать, что решения системы (7) совпадают с решениями следующей системы из двух уравнений с четырьмя неизвестными x_0, x_1, x_2 и x_3:

$$\left.\begin{matrix}\mu x_0 + x_2 = 1\,; \\ x_1 + \mu x_3 = 0\,.\end{matrix}\right\}$$

Поскольку решения системы (7) не зависят от координат вектора A, это означает, что найденные решения описывают глобальные левосторонние единицы, т. е. левосторонние единицы, действующие на все элементы рассматриваемой 4-мерной КНАА. Множество всех p^2 глобальных левосторонних единиц описывается следующей формулой:

$$L = (l_0, l_1, l_2, l_3) = \left(x_0, x_1, 1 - \mu x_0, -\mu^{-1} x_1\right)\,,$$

где $x_0, x_1 = 0, 1, \ldots, p-1$.

Для нахождения правосторонних единиц вектора $A = (a_0, a_1, a_2, a_3)$ рассмотрим следующее векторное уравнение:

$$A \circ X = A\,, \tag{9}$$

которое сводится к системе линейных уравнений вида

$$\left.\begin{matrix}\mu a_0 x_0 + a_1 x_3 + a_2 x_0 + \mu a_3 x_3 = a_0\,; \\ \mu a_0 x_1 + a_1 x_2 + a_2 x_1 + \mu a_3 x_2 = a_1\,; \\ \mu a_0 x_2 + a_1 x_1 + a_2 x_2 + \mu a_3 x_1 = a_2\,; \\ \mu a_0 x_3 + a_1 x_0 + a_2 x_3 + \mu a_3 x_0 = a_3\,.\end{matrix}\right\} \tag{10}$$

Если координаты вектора A удовлетворяют неравенству

$$\Delta = (\mu a_0 + a_2)^2 - (a_1 + \mu a_3)^2 \neq 0\,,$$

то система (10) имеет единственное решение:

$$x_0 = r_0 = \frac{a_0(\mu a_0 + a_2) - a_3(a_1 + \mu a_3)}{\Delta};$$
$$x_1 = r_1 = \frac{\mu(a_0 a_1 - a_2 a_3)}{\Delta};$$
$$x_2 = r_2 = \frac{a_2(\mu a_0 + a_2) - a_1(a_1 + \mu a_3)}{\Delta};$$
$$x_3 = r_3 = \frac{a_2 a_3 - a_0 a_1}{\Delta},$$

которое определяет существование единственной локальной правосторонней единицы $R_A = (r_0, r_1, r_2, r_3)$, соответствующей вектору A. Вектор R_A действует как правая единица на некотором подмножестве элементов рассматриваемой алгебры, поэтому она называется локальной.

3.4 Шестимерная алгебра

Для случая 6-мерной КНАА найдены распределения независимых структурных коэффициентов $\mu, \lambda \in GF(p)$, представленные в табл. 2.

Таблица 2 Предлагаемая ТУБВ для случая $m = 6$

$\circ$	$\mathbf{e}_0$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{e}_4$	$\mathbf{e}_5$
$\mathbf{e}_0$	$\mu\mathbf{e}_0$	$\mu\mathbf{e}_1$	$\mu\mathbf{e}_2$	$\mu\mathbf{e}_3$	$\mu\mathbf{e}_4$	$\mu\mathbf{e}_5$
$\mathbf{e}_1$	$\mathbf{e}_5$	$\mathbf{e}_4$	$\mathbf{e}_3$	$\mathbf{e}_2$	$\mathbf{e}_1$	$\mathbf{e}_0$
$\mathbf{e}_2$	$\lambda\mathbf{e}_0$	$\lambda\mathbf{e}_1$	$\lambda\mathbf{e}_2$	$\lambda\mathbf{e}_3$	$\lambda\mathbf{e}_4$	$\lambda\mathbf{e}_5$
$\mathbf{e}_3$	$\lambda\mathbf{e}_5$	$\lambda\mathbf{e}_4$	$\lambda\mathbf{e}_3$	$\lambda\mathbf{e}_2$	$\lambda\mathbf{e}_1$	$\lambda\mathbf{e}_0$
$\mathbf{e}_4$	$\mathbf{e}_0$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{e}_4$	$\mathbf{e}_5$
$\mathbf{e}_5$	$\mu\mathbf{e}_5$	$\mu\mathbf{e}_4$	$\mu\mathbf{e}_3$	$\mu\mathbf{e}_2$	$\mu\mathbf{e}_1$	$\mu\mathbf{e}_0$

Нахождение левосторонних единиц 6-мерной КНАА, задаваемой табл. 2, по векторному уравнению (6), в котором $A = (a_0, a_1, a_2, a_3, a_4, a_5)$ и $X = (x_0, x_1, x_2, x_3, x_4, x_5)$, приводит к решению следующей системы из шести линейных уравнений с неизвестными координатами вектора X:

$$\left.\begin{aligned}
&\mu x_0 a_0 + x_1 a_5 + \lambda x_2 a_0 + \lambda x_3 a_5 + \\
&\qquad + x_4 a_0 + \mu x_5 a_5 = a_0;\\
&\mu x_0 a_1 + x_1 a_4 + \lambda x_2 a_1 + \lambda x_3 a_4 + \\
&\qquad + x_4 a_1 + \mu x_5 a_4 = a_1;\\
&\mu x_0 a_2 + x_1 a_3 + \lambda x_2 a_2 + \lambda x_3 a_3 + \\
&\qquad + x_4 a_2 + \mu x_5 a_3 = a_2;\\
&\mu x_0 a_3 + x_1 a_2 + \lambda x_2 a_3 + \lambda x_3 a_2 + \\
&\qquad + x_4 a_3 + \mu x_5 a_2 = a_3;\\
&\mu x_0 a_4 + x_1 a_1 + \lambda x_2 a_4 + \lambda x_3 a_1 + \\
&\qquad + x_4 a_4 + \mu x_5 a_1 = a_4;\\
&\mu x_0 a_5 + x_1 a_0 + \lambda x_2 a_5 + \lambda x_3 a_0 + \\
&\qquad + x_4 a_5 + \mu x_5 a_0 = a_5.
\end{aligned}\right\} \quad (11)$$

Выделим в этой системе следующие три системы из двух уравнений:

$$\left.\begin{aligned}
&(\mu x_0 + \lambda x_2 + x_4)a_0 + \\
&\qquad + (x_1 + \lambda x_3 + \mu x_5)a_5 = a_0;\\
&(x_1 + \lambda x_3 + \mu x_5)a_0 + \\
&\qquad + (\mu x_0 + \lambda x_2 + x_4)a_5 = a_5;
\end{aligned}\right\}$$
$$\left.\begin{aligned}
&(\mu x_0 + \lambda x_2 + x_4)a_1 + \\
&\qquad + (x_1 + \lambda x_3 + \mu x_5)a_4 = a_1;\\
&(x_1 + \lambda x_3 + \mu x_5)a_1 + \\
&\qquad + (\mu x_0 + \lambda x_2 + x_4)a_4 = a_4;
\end{aligned}\right\}$$
$$\left.\begin{aligned}
&(\mu x_0 + \lambda x_2 + x_4)a_2 + \\
&\qquad + (x_1 + \lambda x_3 + \mu x_5)a_3 = a_2;\\
&(x_1 + \lambda x_3 + \mu x_5)a_2 + \\
&\qquad + (\mu x_0 + \lambda x_2 + x_4)a_3 = a_3.
\end{aligned}\right\} \quad (12)$$

Легко видеть, что решение системы (12) можно найти, выполнив замену переменных по формулам $z_1 = \mu x_0 + \lambda x_2 + x_4$ и $z_2 = x_1 + \lambda x_3 + \mu x_5$. После такой замены переменных каждая из трех подсистем системы (12) включает два уравнения с одинаковыми двумя неизвестными z_1 и z_2 и приобретает вид:

$$\left.\begin{aligned}
&z_1 a_0 + z_2 a_5 = a_0;\\
&z_1 a_5 + z_2 a_0 = a_5;
\end{aligned}\right\}$$
$$\left.\begin{aligned}
&z_1 a_1 + z_2 a_4 = a_1;\\
&z_1 a_4 + z_2 a_2 = a_4;
\end{aligned}\right\} \quad (13)$$
$$\left.\begin{aligned}
&z_1 a_2 + z_2 a_3 = a_2;\\
&z_1 a_3 + z_2 a_2 = a_3.
\end{aligned}\right\}$$

Система (13) имеет единственное решение в виде пары значений $z_1 = 1$ и $z_2 = 0$ для всех возможных значений вектора A, кроме случая одновременного выполнения условий $a_0 = a_5$, $a_1 = a_4$ и $a_2 = a_3$. В последнем случае имеется множество дополнительных решений в виде пар значений $z_1 \in \in GF(p)$ и $z_2 = 1 - z_1$. Этот особый случай выпадает из множества значений векторов, используемых при построении криптосхем на основе рассматриваемой 6-мерной конечной алгебры.

Выполнение обратной замены переменных показывает, что исходная система (11) имеет решения, совпадающие с решениями следующей системы из двух линейных уравнений с шестью неизвестными x_0, x_1, x_2, x_3, x_4 и x_5:

$$\left.\begin{aligned}
&\mu x_0 + \lambda x_2 + x_4 = 1;\\
&x_1 + \lambda x_3 + \mu x_5 = 0.
\end{aligned}\right\} \quad (14)$$

Решения системы (14) не зависят от координат вектора A, т. е. они описывают следующее множество p^4 глобальных левосторонних единиц:

$$L = (l_0, l_1, l_2, l_3, l_4, l_5) = \\ = \left(x_0, x_1, x_2, x_3, 1 - \mu x_0 - \lambda x_2, -\mu^{-1}(x_1 + \lambda x_3)\right), \quad (15)$$

где $x_0, x_1, x_2, x_3 = 0, 1, \ldots, p-1$.

Правосторонние единицы, соответствующие вектору $A = (a_0, a_1, a_2, a_3, a_4, a_5)$, удовлетворяют векторному уравнению (9), рассмотрение которого приводит к системе уравнений, представимой в виде трех независимых систем из двух линейных уравнений с двумя неизвестными:

$$\left.\begin{cases} k_1 x_0 + k_2 x_5 = a_0\,; \\ k_2 x_0 + k_1 x_5 = a_5\,; \end{cases} \right\} \quad \left.\begin{cases} k_1 x_1 + k_2 x_4 = a_1\,; \\ k_2 x_1 + k_1 x_4 = a_4\,; \end{cases} \right\} \quad \left.\begin{cases} k_1 x_2 + k_2 x_3 = a_2\,; \\ k_2 x_2 + k_1 x_3 = a_3\,, \end{cases} \right\} \tag{16}$$

где введены обозначения $k_1 = \mu a_0 + \lambda a_2 + a_4$ и $k_2 = a_1 + \lambda a_3 + \mu a_5$. В каждой из трех независимых систем из двух уравнений главный определитель равен одному и тому же значению:

$$\Delta = (\mu a_0 + \lambda a_2 + a_4)^2 - (a_1 + \lambda a_3 + \mu a_5)^2\ .$$

При выполнении условия $\Delta = k_1^2 - k_2^2 \neq 0$ система (16) имеет единственное решение:

$$\left.\begin{aligned} x_0 = r_0 &= \frac{a_0 k_1 - a_5 k_2}{\Delta}\,; \\ x_1 = r_1 &= \frac{a_1 k_1 - a_4 k_2}{\Delta}\,; \\ x_2 = r_2 &= \frac{a_2 k_1 - a_3 k_2}{\Delta}\,; \\ x_3 = r_3 &= \frac{a_3 k_1 - a_2 k_2}{\Delta}\,; \\ x_4 = r_4 &= \frac{a_4 k_1 - a_1 k_2}{\Delta}\,; \\ x_5 = r_5 &= \frac{a_5 k_1 - a_0 k_2}{\Delta}\,, \end{aligned} \right\} \tag{17}$$

которое определяет существование единственной локальной правосторонней единицы $R_A = (r_0, r_1, r_2, r_3, r_4, r_5)$, соответствующей вектору A и всевозможным степеням последнего.

Подставляя значения $x_0 = r_0$, $x_1 = r_1$, $x_2 = r_2$, $x_3 = r_3$ из формул (17) в формулу (15), описывающую множество глобальных левосторонних единиц, можно получить $x_4 = 1 - \mu r_0 - \lambda r_2 = r_4$ и $x_5 = -\mu^{-1}(r_1 + \lambda r_3) = r_5$ Последнее означает, что локальная правосторонняя единица R_A содержится в множестве глобальных левосторонних единиц (15), т. е. она является локальной двухсторонней единицей E_A вектора A.

Легко показать, что в бесконечной последовательности $A, A^2, A^3, \ldots, A^i, \ldots$ отсутствует нулевой вектор и при некоторой минимальной натуральной степени d имеет место $A^d = A$. Следовательно, $A^d = A \Rightarrow A^{d-1} \circ A = A \circ A^{d-1}$, т. е. вектор $E_A = A^\omega$, где $\omega = d - 1$, есть локальная двухсторонняя единица вектора A (значение ω будем называть локальным порядком локально обратимого вектора A). Множество $\{A, A^2, A^3, \ldots, A^i, \ldots, A^\omega\}$ представляет собой циклическую мультипликативную группу с единицей E_A.

4 Задание новой формы скрытой задачи дискретного логарифмирования и схема цифровой подписи на ее основе

Для построения алгоритмов ЭЦП предлагается новая форма СЗДЛ, которая отличается использованием открытого ключа в виде трех элементов КНАА, принадлежащих разным циклическим группам, и описывается следующим образом.

1. В качестве характеристики поля берем простое число p достаточно большой разрядности (например, 384 бит).
2. Выбираем три случайных локально обратимых вектора A, B и N, локальный порядок которых содержит достаточно большой простой делитель.
3. Выбираем две случайные глобальные левосторонние единицы L_1 и L_2.
4. Вычисляем вектор A' из уравнения
$$A \circ A' = L_1\,. \tag{18}$$
5. Вычисляем вектор B' из уравнения
$$B \circ B' = L_2\,. \tag{19}$$
6. Вычисляем векторы T и L_3 из уравнения
$$A \circ T \circ B' = L_3\,. \tag{20}$$
7. Выбираем случайное натуральное число $x < \omega$, где ω — значение порядка вектора N.
8. Вычисляем векторы Y и U по формулам:
$$Y = A' \circ N^x \circ A\,;\ \ U = B' \circ N \circ B\,.$$

В силу локальной обратимости векторов A и B уравнения (18) и (19) имеют единственное решение (см. решение системы (16)). Уравнение (20) решается в два этапа. Сначала вычисляются векторы T' и L_3 как неизвестные в уравнении $T' \circ B' = L_3$ (решение является единственным), а затем находится вектор T из уравнения $A \circ T = T'$, которое имеет единственное решение.

Открытым ключом служит тройка векторов Y, U и T. Число x и все другие векторы, использованные для вычисления открытого ключа, остаются секретными. Владелец открытого ключа должен хранить в качестве своего личного секретного ключа число x и два вектора A' и B. Остальные секретные элементы могут быть уничтожены после завершения процедуры вычисления открытого ключа. Предлагаемая форма СЗДЛ состоит в вычислении значения x по открытому ключу. Схема ЭЦП на ее основе описывается следующим образом.

Алгоритм генерации электронной цифровой подписи

1. Выбрать случайное число $k < \omega$ и вычислить вектор $V = A' \circ N^k \circ B$.
2. Вычислить значение $e = F_h(M, V)$, где F_h — некоторая специфицированная хеш-функция; M — электронный документ, который должен быть подписан.
3. Вычислить число $s = k + ex \bmod q$.

Пара чисел (e, s) представляет собой ЭЦП к документу M.

Алгоритм проверки подлинности электронной цифровой подписи

1. По значениям e и s вычислить вектор $V' = Y^e \circ T \circ U^s$.
2. Используя хеш-функцию F_h, вычислить значение $e' = F_h(M, V')$.
3. Если $e' = e$, то подпись признается подлинной, иначе подпись отвергается как ложная.

Доказательство корректности работы схемы ЭЦП:

$$\begin{aligned} V' = Y^e \circ T \circ U^s = \\ = (A' \circ N^x \circ A)^e \circ T \circ (B' \circ N \circ B)^s = \\ = A' \circ N^{xe} \circ (A \circ T \circ B') \circ N^s \circ B = \\ = A' \circ N^{xe} \circ L_3 \circ N^{k-xe} \circ B = \\ = A' \circ N^{xe+k-xe} \circ B = A' \circ N^k \circ B = V \Rightarrow \\ \Rightarrow e' = F_h(M, V') = F_h(M, V) = e\,. \end{aligned}$$

Предложенная в данном разделе схема ЭЦП расширяет ранее известные типы криптосхем с открытым ключом [4], основанные на вычислительной трудности СЗДЛ. Вопрос о сверхполиномиальной сложности решения предложенной формы СЗДЛ на квантовом компьютере требует выполнения специальных математических исследований с привлечением теории конечных алгебр и связан с изучением возможности и трудоемкости сведения СЗДЛ к обычной ЗДЛ. Это представляет собой самостоятельную исследовательскую задачу.

5 Заключение

Разработан общий способ задания m-мерных конечных некоммутативных ассоциативных алгебр для произвольного четного значения размерности $m \geq 4$, свойства которых позволили предложить новую форму СЗДЛ и постквантовую схему цифровой подписи.

Литература

1. Announcing request for nominations for public-key post-quantum cryptographic algoritms. Federal Register. Department of Commerce. Vol. 81. No. 244. P. 92787–92788. https://www.gpo.gov/fdsys/pkg/FR-2016-12-20/pdf/2016-30615.pdf.
2. Post-quantum cryptography / Eds. T. Lange, R. Steinwandt. — Security and cryptology ser. — Springer, 2018. Vol. 10786. 542 p.
3. *Shor P. W.* Polynomial-time algorithms for prime factorization and discrete logarithms on quantum computer // SIAM J. Comput., 1997. Vol. 26. P. 1484–1509.
4. *Moldovyan D. N.* Non-commutative finite groups as primitive of public-key cryptoschemes // Quasigroups Related Systems, 2010. Vol. 18. P. 165–176.
5. *Kuzmin A. S., Markov V. T., Mikhalev A. A., Mikhalev A. V., Nechaev A. A.* Cryptographic algorithms on groups and algebras // J. Math. Sci., 2017. Vol. 223. Iss. 5. P. 629–641.

Поступила в редакцию 27.06.19

METHOD FOR DEFINING FINITE NONCOMMUTATIVE ASSOCIATIVE ALGEBRAS OF ARBITRARY EVEN DIMENSION FOR DEVELOPMENT OF THE POSTQUANTUM CRYPTOSCHEMES

A. A. Kostina, A. Yu. Mirin, D. N. Moldovyan, and R. Sh. Fahrutdinov

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 39, 14th Line V.O., St. Petersburg 199178, Russian Federation

Abstract: The paper introduces a new unified method for defining finite noncommutative associative algebras of arbitrary even dimension m and describes the investigated properties of the algebras for the cases $m = 4$ and 6, when the algebras are defined over the ground field $GF(p)$ with a large size of the prime number p. Formulas describing the set of p^2 (p^4) global left-sided units contained in the 4-dimensional (6-dimensional) algebra are derived. Only local invertibility takes place in the algebras investigated. Formulas for computing the unique local two-sided unit related to the fixed locally invertible vector are derived for each of the algebras. A new form of the hidden discrete logarithm problem is proposed as postquantum cryptographic primitive. The latter was used to develop the postquantum digital signature scheme.

Keywords: finite noncommutative algebra; associative algebra; computationally difficult problem; discrete logarithm; digital signature; postquantum cryptography

DOI: 10.14357/19922264200113

References

1. Depatment of Commerce. 2016. Announcing request for nominations for public-key post-quantum cryptographic algorithms. Federal Register 81(244):92787–92788. Available at: https://www.gpo.gov/fdsys/pkg/FR-2016-12-20/pdf/2016-30615.pdf (accessed March 2, 2020).
2. Lange, T., and R. Steinwandt, eds. 2018. *Post-quantum cryptography*. Security and cryptology ser. Springer. Vol. 10786. 542 p.
3. Shor, P. W. 1997. Polynomial-time algorithms for prime factorization and discrete logarithms on quantum computer. *SIAM J. Comput.* 26:1484–1509.
4. Moldovyan, D. N. 2010. Non-commutative finite groups as primitive of public-key cryptoschemes. *Quasigroups Related Systems* 18:165–176.
5. Kuzmin, A. S., V. T. Markov, A. A. Mikhalev, A. V. Mikhalev, and A. A. Nechaev. 2017. Cryptographic algorithms on groups and algebras. *J. Math. Sci.* 223(5):629–641.

Received June 27, 2019

Contributors

Kostina Anna A. (b. 1983) — scientist, Laboratory of Cybersecurity and Postquantum Cryptosystems, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 39, 14th Line V.O., St. Petersburg 199178, Russian Federation; anna-kostina1805@mail.ru

Mirin Anatoly Yu. (b. 1979) — Candidate of Science (PhD), senior scientist, Laboratory of Cybersecurity and Postquantum Cryptosystems, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 39, 14th Line V.O., St. Petersburg 199178, Russian Federation; mirin@cobra.ru

Moldovyan Dmitriy N. (b. 1986) — Candidate of Science (PhD), scientist, Laboratory of Cybersecurity and Postquantum Cryptosystems, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 39, 14th Line V.O., St. Petersburg 199178, Russian Federation; mdn.spectr@mail.ru

Fahrutdinov Roman Sh. (b. 1972) — Candidate of Science (PhD), Head of Laboratory of Cybersecurity and Postquantum Cryptography, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 39, 14th Line V.O., St. Petersburg 199178, Russian Federation; fahr@cobra.ru

МЕТОД НАВИГАЦИИ И СОСТАВЛЕНИЯ КАРТЫ В ТРЕХМЕРНОМ ПРОСТРАНСТВЕ НА ОСНОВЕ КОМБИНИРОВАННОГО РЕШЕНИЯ ВАРИАЦИОННОЙ ПОДЗАДАЧИ ТОЧКА–ТОЧКА ICP ДЛЯ АФФИННЫХ ПРЕОБРАЗОВАНИЙ*

А. В. Вохминцев[1], А. В. Мельников[2], С. А. Пачганов[3]

Аннотация: Одновременная навигация и картографирование относятся к проблеме, в которой данные кадра используются в качестве единственного источника внешней информации для того, чтобы установить положение движущейся камеры в пространстве и в то же время построить карту зоны исследования. На сегодняшний день эта проблема считается решенной для построения двумерных карт небольших статических сцен с использованием датчиков дальности. Однако для динамичных, сложных и крупномасштабных сцен построение точной трехмерной карты окружающего пространства стало активной областью научных исследований. Для решения поставленной проблемы в работе предложено решение задачи точка–точка для аффинных преобразований и разработан быстрый итерационный алгоритм регистрации кадров в трехмерном пространстве. Производительность и вычислительная сложность предлагаемого метода реконструкции трехмерных сцен представлены и обсуждены на примере эталонных данных. Результаты могут быть применены в задачах навигации мобильного робота в реальном масштабе времени.

Ключевые слова: задача регистрации данных; локализация; методы одновременной навигации и составления карты; аффинное преобразование; двумерные дескрипторы; итеративный алгоритм ближайших точек

DOI: 10.14357/19922264200114

1 Введение

Разработка динамической системы для надежного решения проблемы одновременной навигации мобильного робота и составления карты окружающей его среды (Simultaneous Localization And Mapping, SLAM) в реальном масштабе времени стала одной из ключевых задач в современной робототехнике и машинном зрении, так как на ее решении основано создание автономных интеллектуальных робототехнических комплексов и систем [1–3]. Для построения качественных трехмерных моделей необходимо совместное использование полученных с разных датчиков данных, таких как изображения, положение используемого датчика и карты глубины [4, 5]. В большинстве случаев для построения трехмерных моделей по картам глубины используется итеративный алгоритм ближайших точек (Iterative Closest Point, ICP) [6]. Главный этап алгоритма регистрации ICP связан с поиском соответствующего геометрического преобразования (ортогонального или аффинного), которое наилучшим образом совмещает два облака точек в разных RGB-D-кадрах для выбранной метрики (вариационная подзадача алгоритма). Точность реконструкции трехмерной сцены существенно зависит от выбора метрики для оценки геометрического преобразования и метода решения вариационной задачи [7].

Результат применения итерационных методов для решения задачи минимизации выбранного функционала зависит от правильности выбора начального приближения параметров геометрического преобразования: итерационный процесс может сходиться медленно, сходиться к локальному оптимуму или вообще не сходиться. Использование решений вариационной задачи в замкнутой форме позволяет избежать этих проблем [8, 9]. Выбор класса геометрических преобразований также оказывает значительное влияние на результат реконструкции трехмерной сцены [10].

Для класса ортогональных преобразований решение задачи точка–точка в замкнутой форме получено с помощью кватернионов [8] или с помощью

*Работа выполнена при поддержке РФФИ (проект 18-37-20032) и Российского научного фонда (проект 15-19-10010).

[1]Челябинский государственный университет; Югорский государственный университет, vav@csu.ru

[2]Югорский государственный университет, melnikovav@uriit.ru

[3]Югорский государственный университет, pachganovsa@uriit.ru

ортогональных матриц [9]. На основе метода Хорна сформулирован алгоритм ICP в варианте точка–точка [11]. Известно, что метрика точка–плоскость превосходит метрику точка–точка по точности и скорости сходимости.

Использование аффинных преобразований позволяет решать задачу регистрации для нежестких объектов [12]. Отметим, что если истинное геометрическое преобразование, связывающее два облака точек, ортогонально, то применение вариационной задачи точка–плоскость для аффинных преобразований даст правильный результат только в том случае, если соответствие между точками двух облаков близко к идеальному. На практике соответствие между точками двух облаков в большинстве случаев неидеально: например, все точки первого облака могут соответствовать одной точке второго облака или небольшому локальному подмножеству точек второго облака [13]. Решение описанной выше условной вариационной задачи для метрики точка–точка известно как метод Хорна [8]. Для метрики точка–плоскость [14] существуют решения в замкнутой форме для аффинных преобразований.

2 Метод SLAM и постановка задачи

Решение задачи одновременной навигации и картографирования состоит из следующих этапов:

- сопоставление и регистрация последовательностей изображений в RGB-D-кадрах;
- пространственное совмещение трехмерных облаков точек в RGB-D-кадрах;
- обнаружение замыканий цикла;
- построение трехмерной карты доступной окружающей среды;
- определение позиции робота в относительной системе координат в каждый момент времени.

В данной работе предложено новое решение вариационной задачи точка–точка в замкнутой форме на основе комбинации данных о глубине и цвете в кадре, которое направлено на решение второго этапа задачи SLAM. Основные недостатки итерационных методов регистрации данных связаны с ограничением области сходимости и большой вычислительной сложностью. Кроме того, результат решения вариационной задачи зависит от правильности выбора начального приближения. Для преодоления данного недостатка в работе предлагается использовать визуально связанные характеристики RGB-D-кадра (особые точки), которые позволяют совмещать кадры без требования начальной инициализации. Хорн предложил решение условной вариационной задачи для метрики точка–точка в замкнутой форме для ортогональных преобразований. В данной работе получено решение в замкнутой форме для аффинных преобразований, что, во-первых, создает математическую основу для решения задачи регистрации неригидных объектов на сцене; во-вторых, позволяет находить точное решение вариационной задачи для вырожденных случаев, например, когда все точки трехмерного облака точек находятся в одной плоскости. При решении задачи SLAM в динамическом пространстве такая необходимость возникает при идентификации и отслеживании основных структурных элементов сцены: например, для замкнутых пространств (помещений) такими элементами могут выступать потолок, стены, пол.

3 Сопоставление визуальных признаков на RGB-D-кадрах

Для обработки визуальных характеристик сцены используется алгоритм сопоставления изображений на основе рекурсивного вычисления гистограмм направленных градиентов (ГНГ) по нескольким круглым скользящим окнам и пирамидальному разложению изображения [15, 16]. Для работы с особыми признаками используется следующая схема.

1. Вычисление ГНГ на изображениях.
2. Сопоставление между особыми точками для выбранных подмножеств.
3. Отбрасывание некоторых пар особых точек для поиска [17].
4. Решение вариационной задачи регистрации данных для визуально связанных характеристик изображения.

Рассмотрим более подробно пп. 2 и 4. В работе используется корреляционный оператор, при помощи которого осуществляется процедура сопоставления данных из различных RGB-D-кадров. Введем формулу для определения нормализованной центрированной ГНГ эталонного изображения:

$$\overline{\mathrm{HOG}_i^R}(\alpha) = \frac{\mathrm{HOG}_i^R(\alpha) - \mathrm{Mean}^R}{\sqrt{\mathrm{Var}^R}},$$

где Mean^R — среднее значение ГНГ; Var^R — дисперсия ГНГ. Тогда для каждого i-го медианного

фильтра (МФ) в позиции k можно определить корреляционную функцию:

$$C_i^k(\alpha) = \\ = \mathrm{IFT}\left[\frac{\mathrm{HS}_i^k(\omega)\mathrm{HR}_i^*(\omega)}{\sqrt{Q\sum_{q=0}^{Q-1}\left(\mathrm{HOG}_i^k(q)\right)^2 - \left(\mathrm{HS}_i^k(0)\right)^2}}\right],$$

где $\mathrm{HS}_i^k(\omega)$ — преобразование Фурье ГНГ внутри i-го МФ входной сцены; $\mathrm{HR}_i(\omega)$ — преобразование Фурье $\overline{\mathrm{HOG}_i^R}(\alpha)$; $*$ для i-го преобразования Фурье обозначает комплексное сопряжение. Для определения подобия двух ГНГ применяется корреляционный пик $P_i^k = \max_\alpha\left\{C_i^k(\alpha)\right\}$.

Решение вариационной задачи для визуально связанных характеристик изображения может быть представлено в виде:

$$J(\mathrm{RV}) = \frac{1}{|A_f|}\sum_{i\in A_f}^{n} w_i \left\|M\left(\mathrm{RV}\,F_x^i\right) - M\left(F_y^i\right)\right\|^2 .$$

Здесь RV — матрица аффинного преобразования для визуально связанных характеристик сцены; w_i — весовые характеристики данных; F_x^i и F_y^i — визуально связанные характеристики сцены в исходном и целевом кадре соответственно:

$$F_x^i = \left(x_{1f}^i, x_{2f}^i, x_{3f}^i\right)^{\mathrm{T}} ; \;\; F_y^i = \left(y_{1f}^i, y_{2f}^i, y_{3f}^i\right)^{\mathrm{T}} .$$

Функция M осуществляет преобразование координат точек F_x^i и $F_y^i \in \mathbb{R}^3$ из трехмерной системы координат относительно камеры в систему координат камеры $C^i = \left(C_x^i, C_y^i, D^i\right) \in \mathbb{R}^3$, где C_x^i и C_y^i — соответствующие координаты точек в пиксельном пространстве; D^i — значение глубины в пиксельном пространстве:

$$C_x^i = \frac{f}{x_{3f}^i}\,x_{1f}^i + O_x ;$$
$$C_y^i = \frac{f}{x_{3f}^i}\,x_{2f}^i + O_y ;$$
$$D^i = \sqrt{{x_{1f}^i}^2 + {x_{2f}^i}^2 + {x_{3f}^i}^2} .$$

Здесь O_x и O_y — координаты центра изображения в пиксельном пространстве; f — фокус камеры. Аналогичным образом могут быть определены координаты точек в системе координат камеры для точек F_y^i.

4 Решение задачи точка–точка метода ICP для аффинных преобразований

Задача регистрации трехмерных данных (данных кадра о глубине) на основе алгоритма ICP состоит из следующих шагов.

1. Формирование разреженных подмножеств точек из двух плотных трехмерных облаков точек.
2. Определение соответствующих точек в каждом из разреженных подмножеств.
3. Определение весовых коэффициентов для каждой полученной пары.
4. Отбрасывание некоторых пар в облаках точек (RANSAC-метод или аналог).
5. Выбор метрики ошибки для пар точек.
6. Решение вариационной задачи на основе минимизации функции ошибки.

Обозначим через $X = \{x_1, \ldots, x_n\}$ множество точек в исходном RGB-кадре и через $Y = \\ = \{y_1, \ldots, y_m\}$ множество точек в целевом RGB-D-кадре в $\mathbb{R}^3$. Предположим, что отношения между точками в кадрах X и Y такие, что для каждой точки в x_i можно вычислить соответствующую точку в y_i. Тогда алгоритм ICP можно рассмотреть как геометрическое преобразование из X в Y следующего вида:

$$Rx_i + T ,$$

где R — матрица поворота; T — вектор параллельного переноса; $i = 1, \ldots, n$. Тогда аффинное преобразование в $\mathbb{R}^3$ можно представить в виде функции от двенадцати переменных и решить вариационную задачу алгоритма регистрации кадров для произвольного преобразования.

Пусть $J(R,T)$ — функция вида:

$$J(R,T) = \sum_{i=1}^{n} \left\|Rx_i + T - y_i\right\|^2 ,$$

где

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ x_{3i} \end{pmatrix} \; ; \;\; y_i = \begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} .$$

Тогда вариационная задача ICP может быть определена как $\arg\min\limits_{R,T} J(R,T)$, где

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \; ; \;\; T = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} .$$

Можно заметить, что

$$J(R,T) = \\ = \sum_{i=1}^{n} \left(r_{11}x_{1i} + r_{12}x_{2i} + r_{13}x_{3i} + t_1 - y_{1i}\right)^2 + \\ + \left(r_{21}x_{1i} + r_{22}x_{2i} + r_{23}x_{3i} + t_2 - y_{2i}\right)^2 + \\ + \left(r_{31}x_{1i} + r_{32}x_{2i} + r_{33}x_{3i} + t_3 - y_{3i}\right)^2 . \quad (1)$$

Применим к множеству точек X преобразование переноса следующего вида:

$$\left(x_{1i}^t = x_{1i} - \frac{1}{n}\sum_{j=1}^{n} x_{1j} \, , x_{2i}^t = x_{2i} - \frac{1}{n}\sum_{j=1}^{n} x_{2j} \, , \right. \\ \left. x_{3i}^t = x_{3i} - \frac{1}{n}\sum_{j=1}^{n} x_{3j} \right) .$$

Далее выполним аналогичное преобразование для облака точек Y и подставим новые координаты в выражение (1). Очевидно, что функционал $J(R,T)$ не зависит от элементов вектора параллельного переноса. Решение вариационной задачи ICP может быть найдено методом Хорна в общем случае. Распространим метод Хорна на случай аффинных преобразований. Решение задачи для вырожденных случаев представлено в работе [17]. Положим, что

$$\sum_{i=1}^{n} x_{1i}^2 \neq 0 \, ; \ \sum_{i=1}^{n} x_{2i}^2 \neq 0 \, ; \ \sum\nolimits_{i=1}^{n} x_{3i}^2 \neq 0 \, .$$

Тогда можно решить вариационную задачу относительно матрицы R:

$$\frac{\partial J(R)}{\partial r_{1k}} = \\ = \sum_{i=1}^{n} 2\left(r_{11}x_{1i} + r_{12}x_{2i} + r_{13}x_{3i} - y_{1i}\right)x_{ki} = 0 \, , \\ k = 1,2,3 \, ; \\ \sum_{i=1}^{n} \left(r_{1m}x_{mi} + r_{1n}x_{ni} - y_{1i}\right)x_{ki} + r_{1k}\sum_{i=1}^{n} x_{ki}^2 = 0 \, , \\ k,m,n = 1,2,3 \, ; \ m,n \neq k \, . \quad (2)$$

Из выражения (2) можем получить значения параметров матрицы поворота:

$$r_{1k} = -\frac{\sum\nolimits_{i=1}^{n}(r_{1m}x_{mi} + r_{1n}x_{ni} - y_{1i})x_{ki}}{\sum\nolimits_{i=1}^{n} x_{ki}^2} \, . \quad (3)$$

Учитывая выражение (3), можем представить $J(R)$ в следующем виде:

$$J(R) = \sum_{i=1}^{n} \left(r_{1m}x_{mi} - \right. \\ - \frac{\sum\nolimits_{j=1}^{n}(r_{1m}x_{mj} + r_{1n}x_{nj} - y_{1j})x_{kj}}{\sum\nolimits_{j=1}^{n} x_{kj}^2} x_{ki} + r_{1n}x_{ni} - \\ \left. - y_{1i} \right)^2 + \left(r_{21}x_{1i} + r_{22}x_{2i} + r_{23}x_{3i} - y_{2i}\right)^2 + \\ + \left(r_{31}x_{1i} + r_{32}x_{2i} + r_{33}x_{3i} - y_{3i}\right)^2 . \quad (4)$$

Рассмотрим более подробно первое слагаемое в выражении (4) и раскроем скобки под знаком суммы. Если подставить полученное выражение в функционал $J(R)$, то имеем

$$J(R) = \sum_{i=1}^{n} \left(r_{1m}\left(x_{mi} - x_{ki}\frac{\sum\nolimits_{j=1}^{n} x_{mj}x_{kj}}{\sum\nolimits_{j=1}^{n} x_{kj}^2} \right) + \right. \\ + r_{1n}\left(x_{ni} - x_{ki}\frac{\sum\nolimits_{j=1}^{n} x_{nj}x_{kj}}{\sum\nolimits_{j=1}^{n} x_{kj}^2} \right) - \\ \left. - \left(y_{1i} - x_{ki}\frac{\sum\nolimits_{j=1}^{n} y_{1j}x_{kj}}{\sum\nolimits_{j=1}^{n} x_{kj}^2} \right) \right)^2 + \\ + \left(r_{21}x_{1i} + r_{22}x_{2i} + r_{23}x_{3i} - y_{2i}\right)^2 + \\ + \left(r_{31}x_{1i} + r_{32}x_{2i} + r_{33}x_{3i} - y_{3i}\right)^2 . \quad (5)$$

Введем следующие обозначения для упрощения вида выражения (5):

$$G_{mi} = x_{mi} - x_{ki}\frac{\sum\nolimits_{j=1}^{n} x_{mj}x_{kj}}{\sum\nolimits_{j=1}^{n} x_{kj}^2} \, ; \\ G_{pi} = x_{ni} - x_{ki}\frac{\sum\nolimits_{j=1}^{n} x_{nj}x_{kj}}{\sum\nolimits_{j=1}^{n} x_{kj}^2} \, ; \\ G_{ki} = y_{1i} - x_{ki}\frac{\sum\nolimits_{j=1}^{n} y_{1j}x_{kj}}{\sum\nolimits_{j=1}^{n} x_{kj}^2} \, .$$

Тогда с учетом обозначений можем переписать выражение (5) как

$$J(R) = \sum_{i=1}^{n} \left(r_{1m}G_{mi} + r_{1n}G_{pi} - G_{ki}\right)^2 + \\ + \left(r_{21}x_{1i} + r_{22}x_{2i} + r_{23}x_{3i} - y_{2i}\right)^2 + \\ + \left(r_{31}x_{1i} + r_{32}x_{2i} + r_{33}x_{3i} - y_{3i}\right)^2 . \quad (6)$$

Определим частную производную $J(R)$ относительно r_{1m}:

$$\frac{\partial J(R)}{\partial r_{1m}} = 2\sum_{i=1}^{n}\left(r_{1m}G_{mi} + r_{1n}G_{pi} - G_{ki}\right)G_{mi} = 0.$$

Тогда найдем r_{1m}:

$$r_{1m} = -\frac{r_{1n}\sum_{i=1}^{n} G_{mi}G_{pi} - \sum_{i=1}^{n} G_{mi}G_{ki}}{\sum_{i=1}^{n} G_{mi}^2}.$$

Подставим полученное решение в функционал $J(R)$:

$$\begin{aligned} J(R) = \\ = \sum_{i=1}^{n}\left(-\frac{r_{1n}\sum_{j=1}^{n} G_{mj}G_{pj} - \sum_{j=1}^{n} G_{mj}G_{kj}}{\sum_{j=1}^{n} G_{mj}^2}G_{mi} + \right. \\ \left. + r_{1n}G_{pi} - G_{ki}\right)^2 + \left(r_{21}x_{1i} + r_{22}x_{2i} + r_{23}x_{3i} - \right. \\ \left. - y_{2i}\right)^2 + \left(r_{31}x_{1i} + r_{32}x_{2i} + r_{33}x_{3i} - y_{3i}\right)^2. \quad (7) \end{aligned}$$

Произведем некоторые преобразования в первом слагаемом выражения (7):

$$\begin{aligned} J(R) = \sum_{i=1}^{n}\left(\left(-r_{1n}G_{mi}\sum_{j=1}^{n} G_{mj}G_{pj} - \right.\right. \\ \left.\left. - G_{mi}\sum_{j=1}^{n} G_{mj}G_{kj}\right)\Big/\sum_{j=1}^{n} G_{mj}^2 + r_{1n}G_{pi} - G_{ki}\right)^2 + \\ + \left(r_{21}x_{1i} + r_{22}x_{2i} + r_{23}x_{3i} - y_{2i}\right)^2 + \\ + \left(r_{31}x_{1i} + r_{32}x_{2i} + r_{33}x_{3i} - y_{3i}\right)^2. \end{aligned}$$

Затем сделаем группировку слагаемых следующего вида:

$$\begin{aligned} J(R) = \sum_{i=1}^{n}\left(r_{1n}\left(G_{pi} - \frac{G_{mi}\sum_{j=1}^{n} G_{mj}G_{pj}}{\sum_{j=1}^{n} G_{mj}^2}\right) - \right. \\ \left. - \left(G_{ki} - \frac{G_{mi}\sum_{j=1}^{n} G_{mj}G_{kj}}{\sum_{j=1}^{n} G_{mj}^2}\right)\right)^2 + \\ + \left(r_{21}x_{1i} + r_{22}x_{2i} + r_{23}x_{3i} - y_{2i}\right)^2 + \\ + \left(r_{31}x_{1i} + r_{32}x_{2i} + r_{33}x_{3i} - y_{3i}\right)^2. \quad (8) \end{aligned}$$

Введем следующие обозначения:

$$\left.\begin{aligned} \Omega_1 = G_{pi} - \frac{G_{mi}\sum_{j=1}^{n} G_{mj}G_{pj}}{\sum_{j=1}^{n} G_{mj}^2}; \\ \Omega_2 = G_{ki} - \frac{G_{mi}\sum_{j=1}^{n} G_{mj}G_{kj}}{\sum_{j=1}^{n} G_{mj}^2}. \end{aligned}\right\} \quad (9)$$

С учетом (8) и (9) выражение (6) может быть представлено в виде:

$$\begin{aligned} J(R) = \sum_{i=1}^{n}\left(r_{1n}\Omega_1 - \Omega_2\right)^2 + \\ + \left(r_{21}x_{1i} + r_{22}x_{2i} + r_{23}x_{3i} - y_{2i}\right)^2 + \\ + \left(r_{31}x_{1i} + r_{32}x_{2i} + r_{33}x_{3i} - y_{3i}\right)^2. \end{aligned}$$

Теперь определим частную производную $J(R)$ относительно r_{1n}:

$$\frac{\partial J(R)}{\partial r_{1n}} = 2\sum_{j=1}^{n}\left(r_{1n}\Omega_1 - \Omega_2\right)\Omega_1 = 0.$$

Тогда

$$r_{1n} = \frac{\sum_{k=1}^{n}\Omega_1\Omega_2}{\sum_{k=1}^{n}\Omega_1^2}.$$

При $\sum_{k=1}^{n}\Omega_1^2 \neq 0$ можем определить параметры матрицы поворота. Тогда первая строка матрицы будет выглядеть следующим образом:

$$\begin{aligned} r_{1m} = -\frac{r_{1n}\sum_{i=1}^{n} G_{mi}G_{pi} - \sum_{i=1}^{n} G_{mi}G_{ki}}{\sum_{i=1}^{n} G_{mi}^2}; \\ r_{1n} = \frac{\sum_{k=1}^{n}\Omega_1\Omega_2}{\sum_{k=1}^{n}\Omega_1^2}; \\ r_{1k} = -\frac{\sum_{i=1}^{n}\left(r_{1m}x_{mi} + r_{1n}x_{ni} - y_{1i}\right)x_{ki}}{\sum_{i=1}^{n} x_{ki}^2}. \end{aligned}$$

Вторая строка параметров матрицы поворота:

$$\begin{aligned} r_{2m} = -\frac{r_{2n}\sum_{i=1}^{n} G_{mi}G_{pi} - \sum_{i=1}^{n} G_{mi}G_{ki}}{\sum_{i=1}^{n} G_{mi}^2}; \\ r_{2n} = \frac{\sum_{k=1}^{n}\Omega_1\Omega_2}{\sum_{k=1}^{n}\Omega_1^2}; \\ r_{2k} = -\frac{\sum_{i=1}^{n}\left(r_{2m}x_{mi} + r_{2n}x_{ni} - y_{2i}\right)x_{ki}}{\sum_{i=1}^{n} x_{ki}^2}. \end{aligned}$$

Третья строка параметров матрицы поворота:

$$r_{3m} = -\frac{r_{3n}\sum_{i=1}^{n} G_{mi}G_{pi} - \sum_{i=1}^{n} G_{mi}G_{ki}}{\sum_{i=1}^{n} G_{mi}^2};$$

$$r_{3n} = \frac{\sum_{k=1}^{n} \Omega_1\Omega_2}{\sum_{k=1}^{n} \Omega_1^2};$$

$$r_{3k} = -\frac{\sum_{i=1}^{n} \left(r_{3m}x_{mi} + r_{3n}x_{ni} - y_{3i}\right) x_{ki}}{\sum_{i=1}^{n} x_{ki}^2}.$$

Определим элементы вектора параллельного переноса T через элементы матрицы поворота:

$$t_k = \frac{1}{n}\sum_{i=1}^{n}\left(y_{ki} - \left(r_{k1}x_{1i} + r_{k2}x_{2i} + r_{k3}x_{3i}\right)\right) = 0\,,$$

$$k = 1, 2, 3\,.$$

5 Комбинированное решение задачи точка–точка для аффинных преобразований в трехмерном пространстве

Вариационную задачу точка–точка для визуально связанных характеристик сцены и данных глубины можно представить в виде: $\arg\min\limits_{\mathrm{RV},\mathrm{RD}} J\left(\mathrm{RV},\mathrm{RD}\right)$, где

$$J\left(\mathrm{RV},\mathrm{RD}\right) =$$
$$= \alpha\frac{1}{W}\frac{1}{|A_f|}\sum_{i\in A_f}^{m} w_i \left\|M\left(\mathrm{RV}\,F_x^i\right) - M\left(F_y^i\right)\right\|^2 +$$
$$+ (1-\alpha)\frac{1}{W}\frac{1}{|A_d|}\sum_{j\in A_d}^{n} w_j \left\|\mathrm{RD}\,x_j + T - y_j\right\|^2,$$

где RV и RD — матрицы аффинного преобразования для визуально связанных характеристик сцены и для данных о глубине сцены соответственно; α и W — набор параметров для нормировки данных, подбираемый эмпирическим путем; w_i и w_j — весовые характеристики данных, связанные с семантической маркировкой пространства; A_f — подмножество, которое содержит связи между особыми точками в двух кадрах; A_d — содержит связи между соответствующими точками x_j и y_j в трехмерных облаках точек в двух кадрах данных; F_x^i и F_y^i — визуально связанные характеристики сцены.

В общем случае $\mathrm{RV} \neq \mathrm{RD}$, в данной работе находится совместное решение вариационной задачи для частного случая, когда $\mathrm{RV} = \mathrm{RD} = \mathrm{RT}$. Пусть T_{km} — начальная оценка для ICP с использованием кинематической модели движения камеры; $k_{\max}$ и ε — пороги алгоритма ICP по числу шагов и по величине ошибки E соответственно; RT^* — лучшее преобразование на i-м шаге метода; δ — порог для точек инлайнеров. Общая схема комбинированного метода регистрации может быть представлена в следующем виде.

Шаг 1. Вычисление ГНГ на изображениях.

Шаг 2. Сопоставление между особыми точками F_x^i и F_y^i для выбранных подмножеств. Решение вариационной задачи регистрации данных для визуально связанных характеристик изображения. Получим RT^* и A_f.

Шаг 3. Проверка: если $|A_f| < \delta$, то $\mathrm{RT}^* = T_{km}$ и $A_f = \emptyset$. Положим $i = 1$.

Шаг 4. Определение соответствующих точек A_d в исходном и целевом облаке точек с использованием метода ближайших соседей. Определение весовых коэффициентов для каждой полученной пары A_d.

Шаг 5. Решение комбинированной вариационной задачи относительно RT^*, A_f и A_d.

Шаг 6. Проверка: $(E(\mathrm{RT}_i^*) - E(\mathrm{RT}_{i+1}^*) \leq \varepsilon)$ или (Номер итерации $> \mathrm{k}_{\max}$). Если условие неверно, то возврат на шаг 4 и $i = i + 1$, иначе получено искомое преобразование RT^*.

В качестве продолжения направления исследований представляет интерес разработка метода решения комбинированной вариационной задачи точка–плоскость в замкнутой форме для аффинных и ортогональных преобразований. Обозначим:

$$\eta_1 = \alpha\frac{1}{W}\frac{1}{|A_f|},\ \eta_2 = (1-\alpha)\frac{1}{W}\frac{1}{|A_d|}.$$

Пусть n^i есть унитарная нормаль к касательной плоскости $T(y^i)$ к поверхности $S(Y)$ в точке y^i; RD^M — комбинированная матрица аффинных преобразований, содержащая компоненты параллельного переноса и поворота. Тогда

$$J\left(\mathrm{RD}^M\right) = \sum_{i=1}^{m}\eta_1 w_i\left(M\left(\mathrm{RD}^M F_x^i\right) - M\left(F_y^i\right)\right)^2 +$$
$$+ \sum_{j=1}^{n}\eta_2 w_j\left(\left\langle \mathrm{RD}^M x'^j - y'^j, n^j\right\rangle\right)^2,$$

где

$$x'^j = \left(x_1'^j, x_2'^j, x_3'^j, 1\right)^{\mathrm{T}};\ y'^j = \left(y_1'^j, y_2'^j, y_3'^j, 1\right)^{\mathrm{T}}.$$

Предполагается, что метод будет основан на проецировании элемента многообразия матриц линейных преобразований на подмногообразие ортогональных матриц.

6 Компьютерное моделирование

В данном разделе представлены результаты компьютерного моделирования. Для проведения компьютерного моделирования были выбраны четыре набора данных из базы данных NYU Depth Dataset V2 [18, 19], содержащих фрагменты крупномасштабных сцен помещений: Classrooms, Living Rooms (1/4), Offices (1/2) и Offices (2/2). Компьютерное моделирование проводилось на персональном компьютере Intel Core i7 с многоядерным графическим процессором. Для экспериментальных исследований использовалась камера Kinect 2.0 в качестве RGB-D-сенсора. Каждый из четырех наборов данных из тестовой базы содержит файлы дампа, изображение в формате ppm (рис. 1, *а*), данные о глубине в формате pgm (рис. 1, *б*). Данные в RGB-D-кадрах получены из разных положений камеры на сцене. Например, набор данных Classrooms содержит 688 ключевых кадров размером 640×480.

Для повышения точности решения вариационной задачи точка–точка в предложенном методе используются данные о визуально связанных характеристиках изображений, а значит решение вариационной задачи в целом зависит от точности сопоставления данных на изображениях в RGB-D-кадре. Проведем сравнительный анализ известных дескрипторных методов (SIFT (scale-invariant feature transform), SURF (speeded-up robust features) и ORB (oriented fast and rotated brief)) и предложенного метода сопоставления особых точек (HOGs, histogram oriented gradients) [15, 16].

Будем оценивать точность методов по числу правильных сопоставлений при разных значениях угла поворота и изменения масштаба. Были получены следующие результаты зависимости точности реконструкции трехмерной комбинированной карты и производительности от выбранного типа двумерного дескриптора (см. таблицу) [20]. В таблице приведены средние значения точности для разного типа аффинных преобразований (угол поворота и масштаб) для различных наборов данных из тестовой базы данных и различных двумерных дескрипторов.

В работе проведен сравнительный анализ скорости сходимости предложенного метода регистрации данных и известных методов регистрации на основе итеративного алгоритма ближайших точек в зависимости от выбора метрики ошибки пар точек: точка–точка [6], точка–точка с экстраполяцией [8], точка–плоскость [14] в терминах среднеквадратичной ошибки. Для проведения компьютерного моделирования было выбрано два

(*а*)

(*б*)

Рис. 1 Тестовый набор данных Classrooms (классная комната) из NYU Depth Dataset: (*а*) визуальные данные RGB-D-кадра; (*б*) данные глубины RGB-D-кадра

Зависимость точности/производительности метода реконструкции трехмерной комбинированной плотной карты от типа двумерного дескриптора

Метод	Набор данных Classrooms	Набор данных Living Rooms (1/4)	Набор данных Offices (1/2)	Набор данных Offices (2/2)
	Точность, %			
SIFT	93	93	94	95
SURF	57	61	63	59
ORB	78	82	75	75
HOGs	93	96	97	96
	Производительность, с			
SIFT	17,2	26,21	12,82	18,9
SURF	0,23	0,7	0,4	0,44
ORB	3,34	3,42	5,15	3,33
HOGs	0,79	1,11	0,68	0,81

набора данных из базы данных NYU Depth Dataset: в контролируемых условиях (Offices (1/2), сцена 1) и в неконтролируемых условиях (Offices (1/2), сцена 2).

В результате серии тестов установлены зависимости скорости сходимости методов от выбора ошибки метрики и условий проведения экспериментов (рис. 2). Было установлено, что в контролируемых условиях предложенный метод регистрации данных имеет сходимость, близкую к обеспечиваемой методом, который использует метрику точка–плоскость (рис. 2, *а*). В неконтролируемых условиях (при неравномерном освещении) предложенный метод показывает лучшую сходимость, чем указанные выше методы регистрации данных (рис. 2, *б*). Дополнительно было установлено, что точность метода реконструкции зависит от числа особых точек в RGB-D-кадре нелинейным образом — в виде функции с одним ярко выраженным пиком для всех типов дескрипторов [21, 22].

Как известно, дескриптор HOG инвариантен к неравномерному освещению и фотометрическим преобразованиям, которые проявляются на больших сценах [15]. Был проведен сравнительный анализ зависимости скорости сходимости методов регистрации от выбора типа дескриптора и условий проведения экспериментов (рис. 3). Было установлено, что в контролируемых условиях тип используемого дескриптора имеет ограниченное влияние на сходимость метода регистрации данных (рис. 3, *а*). В неконтролируемых условиях использование дескриптора HOGs позволяет получить лучшую сходимость в сравнении с другими дескрипторными методами: предложенный метод регистрации сходится уже после 11-й итерации, тогда как при использовании дескриптора ORB метод сходится только после 16-й итерации (рис. 3, *б*).

На рис. 4 показаны результаты применения предложенного метода регистрации для последовательности ключевых RGB-D-кадров из базы дан-

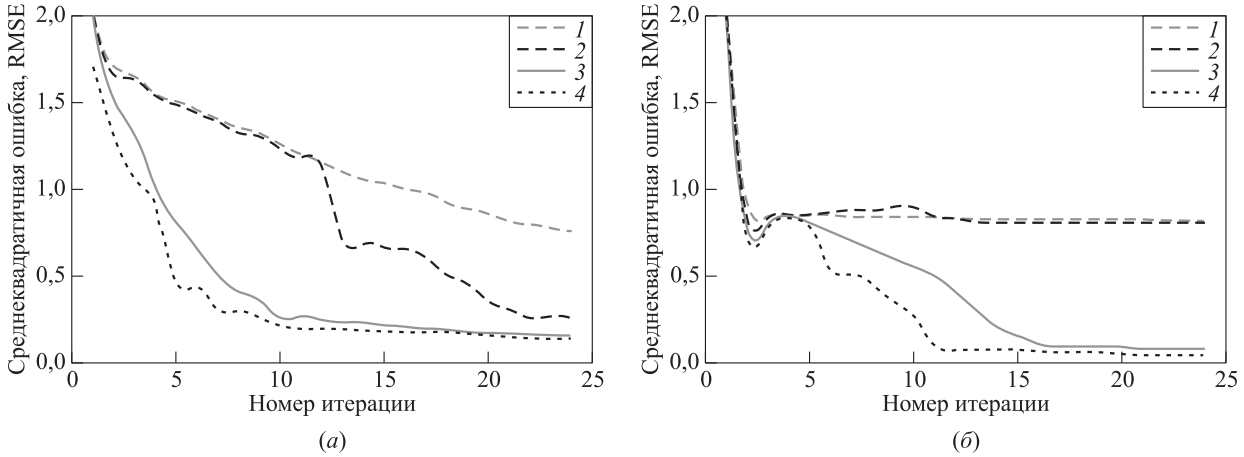


Рис. 2 Сравнение скорости сходимости в зависимости от ошибки метрики и условий наблюдения: (*а*) изменение скорости сходимости в контролируемых условиях; (*б*) изменение скорости сходимости в неконтролируемых условиях; *1* — точка–точка; *2* — точка–точка с экстраполяцией; *3* — точка–плоскость; *4* — комбинированный подход

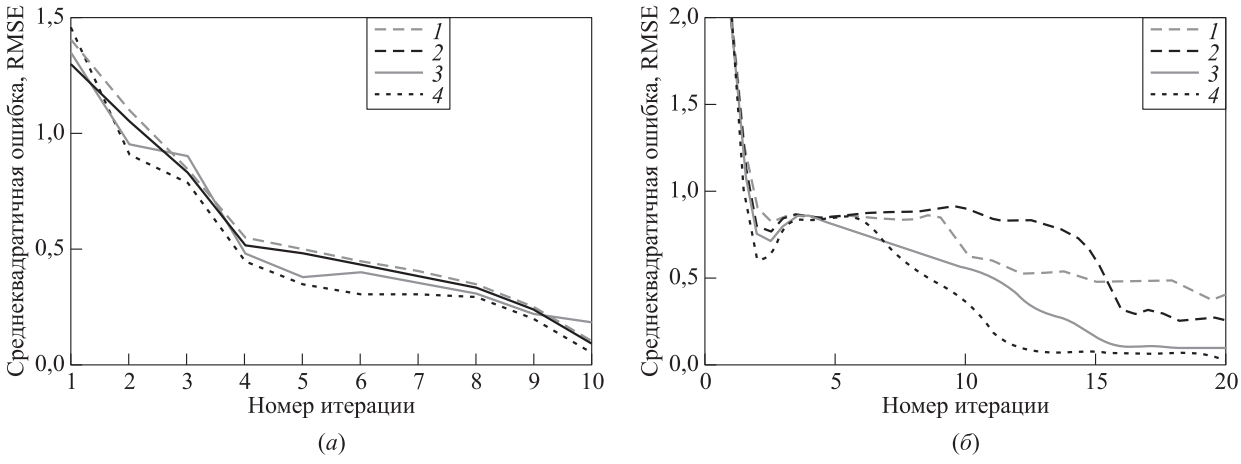


Рис. 3 Сравнение скорости сходимости в зависимости от типа дескриптора и условий наблюдения: (*а*) изменение скорости сходимости в контролируемых условиях; (*б*) изменение скорости сходимости в неконтролируемых условиях; *1* — SIFT; *2* — SURF; *3* — ORB; *4* — HOGs

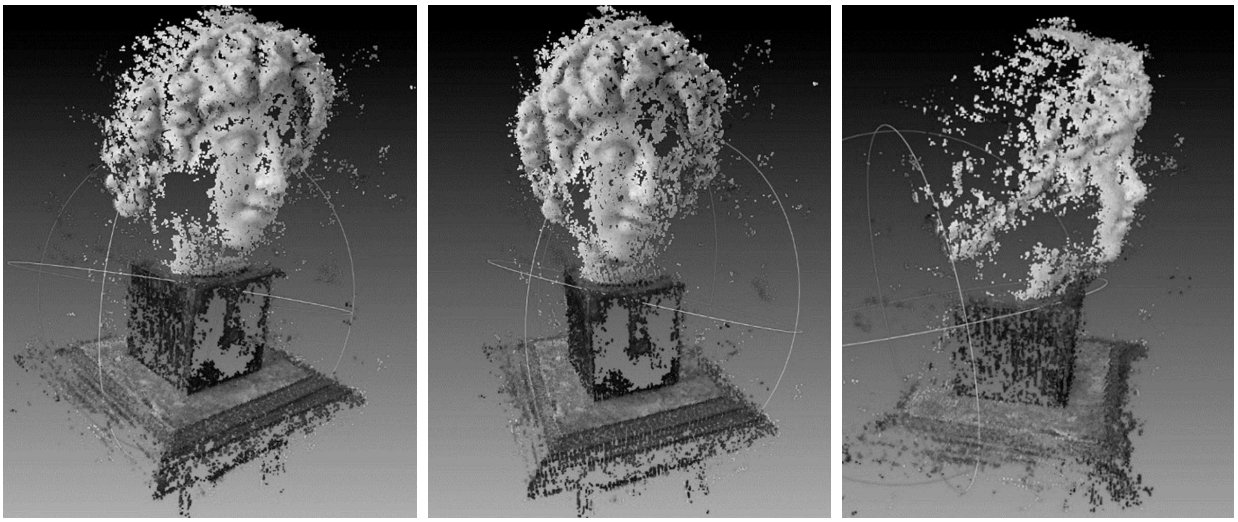


Рис. 4 Реконструкция трехмерного объекта на основе предложенного метода регистрации данных с разных точек обзора

ных NYU Depth Dataset: исходное облако точек визуально совмещено с целевым облаком точек на примере статуи головы человека. Недостаток классического метода ICP [6] — большая вычислительная сложность. Проекционные методы могут сократить вычислительную сложность метода регистрации ICP с $O(N_S \log(N_T))$ для метода ICP с k-D-деревом до $O(N_S)$ для метода ICP с ограничением в виде сферы или треугольника.

Для ускорения процедуры сопоставления данных в работе используется пирамидальный подход, основанный на низкочастотной фильтрации и последующей дискретизации результирующего изображения/облака точек. Вычислительная сложность предлагаемого метода регистрации может быть оценена следующим образом: $n_1 + n_2 O_1/F^1 + \cdots$ $\cdots + n_k O_1/F^{k-1}$, где k — число шагов дискретизации; n_i — число итераций в методе регистрации, выполненных на шаге k; O_1 — вычислительная сложность первого шага алгоритма; F — параметр, определяющий разбиение RGB-D-кадра.

7 Выводы

В работе предложен комбинированной метод решения вариационной задачи точка–точка в замкнутой форме для аффинных преобразований, который создает основу для распространения метода Хорна на случай с неригидными объектами на сцене. Было проведено сравнение предлагаемого

метода регистрации данных с методом Хорна для метрики точка–точка с экстраполяцией и без экстраполяции, а также с методом регистрации для метрики точка–плоскость. В результате компьютерного моделирования было установлено, что применение визуально связанных характеристик для решения вариационной задачи алгоритма ICP позволяет улучшить сходимость метода в неконтролируемых условиях. Использование визуально связанных характеристик изображений позволяет решить проблему зависимости результата решения вариационной задачи от правильности выбора начальных значений. Двумерный дескриптор HOGs обладает лучшими характеристиками по сравнению с известными дескрипторами при малых поворотах в области сцены. Предложенный метод используется для регистрации облаков точек с произвольным пространственным разрешением и масштабом относительно друг друга, дает точные оценки для сложных крупномасштабных сцен.

Литература

1. *Vidal-Calleja T. A., Berger C., Sola J., Lacroix S.* Large scale multiple robot visual mapping with heterogeneous landmarks in semi-structured terrain // J. Robotics Autonomous Systems, 2011. Vol. 59. Iss. 9. P. 654–674. doi: 10.1016/j.robot.2011.05.008.
2. *Vokmintsev A., Timchenko M., Yakovlev K.* Simultaneous localization and mapping in unknown environment using dynamic matching of images and registration of point clouds // 2nd Conference (International) on Industrial Engineering, Applications and Manufacturing. — IEEE, 2017. Art. ID 7910967. 6 p. doi: 10.1109/ICIEAM.2016.7910967.
3. *Bokovoy A., Yakovlev K.* Sparse 3D point-cloud map upsampling and noise removal as a vSLAM post-processing step: Experimental evaluation // Interactive collaborative robotics / Eds. A. Ronzhin, G. Rigoll, R. Meshcheryakov. — Lecture notes in computer science ser. — Springer, 2018. Vol. 11097. P. 23–33.
4. *Tam G., Cheng Z.-Q., Lai Y.-K., Langbein F., Liu Y., Marshall D., Martin R., Rosin P.* Registration of 3D point clouds and meshes: A survey from rigid to nonrigid // IEEE T. Vis. Comput. Gr., 2013. Vol. 19. Iss. 7. P. 1199–1217. doi: 10.1109/tvcg.2012.310.
5. *Picos K., Diaz-Ramirez V. H., Kober V., Montemayor A. S., Pantrigo J. J.* Accurate three-dimensional pose recognition from monocular images using template matched filtering // Opt. Eng., 2016. Vol. 55. Iss. 6. Art. ID 063102. doi: 10.1117/1.oe.55.6.063102.
6. *Besl P., McKay N.* A method for registration of 3-D shapes // IEEE T. Pattern Anal., 1992. Vol. 14. Iss. 2. P. 239–256. doi: 10.1109/34.121791.
7. *Cheng S., Marras I., Zafeiriou S., Pantic M.* Statistical non-rigid ICP algorithm and its application to 3D face alignment // Image Vision Comput., 2017. Vol. 58. P. 3–12. doi: 10.1016/j.imavis.2016.10.007.
8. *Horn B.* Closed-form solution of absolute orientation using unit quaternions // J. Opt. Soc. Am. A, 1987. Vol. 4. Iss. 4. P. 629–642. doi: 10.1364/josaa.4.000629.
9. *Horn B., Hilden H., Negahdaripour S.* Closed-form solution of absolute orientation using orthonormal matrices // J. Opt. Soc. Am. A, 1988. Vol. 5. Iss. 7. P. 1127–1135. doi: 10.1364/josaa.5.001127.
10. *Khoshelham K.* Closed-form solutions for estimating a rigid motion from plane correspondences extracted from point clouds // ISPRS J. Photogramm., 2016. Vol. 114. P. 78–91. doi: 10.1016/j.isprsjprs.2016.01.010.
11. *Du S., Liu J., Zhang C., Zhu J., Li K.* Probability iterative closest point algorithm for m-D point set registration with noise // Neurocomputing, 2015. Vol. 157. Iss. 1. P. 187–198. doi: 10.1016/j.neucom.2015.01.019.
12. *Cheng S., Marras I., Zafeiriou S.* Active nonrigid ICP algorithm // 11th IEEE Conference (International) and Workshops on Automatic Face and Gesture Recognition Proceedings. — IEEE, 2015. Art. ID 7163161. 8 p. doi: 10.1109/FG.2015.7163161.
13. *Echeagaray-Patron B. A., Kober V., Karnaukhov V., Kuznetsov V.* A method of face recognition using 3D facial surfaces // J. Commun. Technol. El., 2017. Vol. 62. Iss. 6. P. 648–652. doi: 10.1134/s1064226917060067.
14. *Low K. L.* Linear least-squares optimization for point-to-plane ICP surface registration. — Chapel Hill, NC, USA: University of North Carolina at Chapel Hill, Department of Computer Science, 2004. Technical Report TTR04-004. https://www.comp.nus.edu.sg/~lowkl/publications/lowk_point-to-plane_icp_techrep.pdf.
15. *Вохминцев А. В., Соченков И. В., Кузнецов В. В., Тихоньких Д. В.* Распознавание лиц на основе алгоритма сопоставления изображений с рекурсивным вычислением гистограмм направленных градиентов // Докл. Акад. наук, 2016. Т. 466. № 3. С. 261. doi: 10.7868/S0869565216030087.
16. *Diaz-Escobar J., Kober V.* A robust HOG-based descriptor for pattern recognition // Proc. SPIE, 2016. Vol. 9971. Art. ID 99712A. doi: 10.1117/12.2237963.
17. *Vokhmintcev A., Yakovlev K.* A real-time algorithm for mobile robot mapping based on rotation-invariant descriptors and ICP // Comm. Comp. Inf. Sc., 2016. Vol. 661. P. 357–369.
18. *Silberman N., Kohli P., Hoiem D., Fergus R.* NYU Depth Dataset V2. https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.
19. *Silberman N., Hoiem D., Kohli P., Fergus R.* Indoor segmentation and support inference from RGBD images // Computer vision / Eds. A. W. Fitzgibbon, S. Lazebnik, P. Perona, *et al.* — Lecture notes in computer science ser. — 2012. Vol. 7576. P. 746–760.

20. *Vokhmintcev A., Botova T., Sochenkov I., Sochenkova A., Makovetskii A.* Robot mapping algorithm based on Kalman filtering and symbolic tags // Proc. SPIE, 2017. Vol. 10396. Art. ID 103962I. doi: 10.1117/12.2273562.
21. *Vokhmintcev A., Timchenko M., Melnikov A., Kozko A., Makovetskii A.* Robot path planning algorithm based on symbolic tags in dynamic environment // Proc. SPIE, 2017. Vol. 10396. Art. ID 103962E. doi: 10.1117/12.2273279.
22. *Sochenkov I., Vokhmintsev A.* Visual duplicates image search for a non-cooperative person recognition at a distance // Procedia Engineer., 2015. Vol. 129. P. 440–445. doi: 10.1016/j.proeng.2015.12.147.

Поступила в редакцию 25.02.19

SIMULTANEOUS LOCALIZATION AND MAPPING METHOD IN THREE-DIMENSIONAL SPACE BASED ON THE COMBINED SOLUTION OF THE POINT–POINT VARIATION PROBLEM ICP FOR AN AFFINE TRANSFORMATION

A. V. Vokhmintcev[1,2], A. V. Melnikov[2], and S. A. Pachganov[2]

[1]Chelyabinsk State University, 129 Br. Kashirinyh Str., Chelyabinsk 454001, Russian Federation
[2]Ugra State University, 16 Chekhov Str., Khanty-Mansiysk 628012, Russian Federation

Abstract: Simultaneous localization and mapping is a problem in which frame data are used as the only source of external information to define the position of a moving camera in space and at the same time, to reconstruct a map of the study area. Nowadays, this problem is considered solved for the construction of two-dimensional maps for small static scenes using range sensors such as lasers or sonar. However, for dynamic, complex, and large-scale scenes, the construction of an accurate three-dimensional map of the surrounding space is an active area of research. To solve this problem, the authors propose a solution of the point–point problem for an affine transformation and develop a fast iterative algorithm for point clouds registering in three-dimensional space. The performance and computational complexity of the proposed method are presented and discussed by an example of reference data. The results can be applied for navigation tasks of a mobile robot in real-time.

Keywords: registration problem; localization; simultaneous localization and mapping; affine transformation; two-dimensional descriptors; iterative closest point

DOI: 10.14357/19922264200114

Acknowledgments

This work was partially supported by the Russian Foundation for Basic Research (grant 18-37-20032) and by the Russian Science Foundation (project No. 15-19-10010).

References

1. Vidal-Calleja, T. A., C. Berger, J. Sola, and S. Lacroix. 2011. Large scale multiple robot visual mapping with heterogeneous landmarks in semi- structured terrain. *J. Robotics Autonomous Systems* 59(9):654–674. doi: 10.1016/j.robot.2011.05.008.
2. Vokmintsev, A., M. Timchenko, and K. Yakovlev. 2017. Simultaneous localization and mapping in unknown environment using dynamic matching of images and registration of point clouds. *2nd Conference (International) on Industrial Engineering, Applications and Manufacturing*. IEEE. Art. ID 7910967. 6 p. doi: 10.1109/ICIEAM.2016.7910967.
3. Bokovoy, A., and K. Yakovlev. 2018. Sparse 3D point-cloud map upsampling and noise removal as a vSLAM post-processing step: Experimental evaluation. *Interactive collaborative robotics*. Eds. A. Ronzhin, G. Rigoll, and R. Meshcheryakov. Lecture notes in computer science ser. Springer. 11097:23–33.
4. Tam, G., Z.-Q. Cheng, Y.-K. Lai, F. Langbein, Y. Liu, D. Marshall, R. Martin, and P. Rosin. 2013. Registration of 3D point clouds and meshes: A survey from rigid to nonrigid. *IEEE T. Vis. Comput. Gr.* 19(7):1199–1217. doi: 10.1109/tvcg.2012.310.
5. Picos, K., V. H. Diaz-Ramirez, V. Kober, A. S. Montemayor, and J. J. Pantrigo. 2016. Accurate three-dimensional pose recognition from monocular images

using template matched filtering. *Opt. Eng.* 55(6):063102. doi: 10.1117/1.oe.55.6.063102.

6. Besl, P., and N. McKay. 1992. A method for registration of 3-D shapes. *IEEE T. Pattern Anal.* 14(2):239–256. doi: 10.1109/34.121791.
7. Cheng, S., I. Marras, S. Zafeiriou, and M. Pantic. 2017. Statistical non-rigid ICP algorithm and its application to 3D face alignment. *IEEE Image Vision Comput.* 58:3–12. doi: 10.1016/j.imavis.2016.10.007.
8. Horn, B. 1987. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A* 4(4):629–642. doi: 10.1364/josaa.4.000629.
9. Horn, B., H. Hilden, and S. Negahdaripour. 1988. Closed-form solution of absolute orientation using orthonormal matrices. *J. Opt. Soc. Am. A* 5(7):1127–1135. doi: 10.1364/JOSAA.5.001127.
10. Khoshelham, K. 2016. Closed-form solutions for estimating a rigid motion from plane correspondences extracted from point clouds. *J. ISPRS Photogramm.* 114:78–91. doi: 10.1016/j.isprsjprs.2016.01.010.
11. Du, S., J. Liu, C. Zhang, J. Zhu, and K. Li. 2015. Probability iterative closest point algorithm for m-D point set registration with noise. *Neurocomputing* 157(1):187–198. doi: 10.1016/j.neucom.2015.01.019.
12. Cheng, S., I. Marras, and S. Zafeiriou. 2015. Active nonrigid ICP algorithm. *IEEE 11th Conference (International) and Workshops on Automatic Face and Gesture Recognition Proceedings*. Art. ID 7163161. 8 p. doi: 10.1109/FG.2015.7163161.
13. Echeagaray-Patron, B. A., V. Kober, V. Karnaukhov, and V. Kuznetsov. 2017. A method of face recognition using 3D facial surfaces. *J. Commun. Technol. El.* 62(6):648–652. doi: 10.1134/s1064226917060067.
14. Low, K. L. 2004. Linear least-squares optimization for point-to-plane ICP surface registration. Chapel Hill, NC: University of North Carolina at Chapel Hill, Department of Computer Science. Technical Report TTR04-004. Available at: https://www.comp.nus.edu.sg/~lowkl/publications/lowk_point-to-plane_icp_techrep.pdf (accessed December 17, 2019).
15. Vokhmintcev, A. V., I. V. Sochenkov, V. V. Kuznetsov, and D. V. Tikhonkikh. 2016. Face recognition based on a matching algorithm with recursive calculation of oriented gradient histograms. *Doklady Mathematics* 93(1):37–41. doi: 10.1134/s1064562416010178.
16. Diaz-Escobar, J., and V. Kober. 2016. A robust HOG-based descriptor for pattern recognition. *Proc. SPIE* 9971:99712A. doi: 10.1117/12.2237963.
17. Vokhmintcev, A., and K. Yakovlev. 2016. A real-time algorithm for mobile robot mapping based on rotation-invariant descriptors and ICP. *Comm. Comp. Inf. Sc.* 661:357–369.
18. Silberman, N., P. Kohli, D. Hoiem, and R. Fergus. NYU depth dataset V2. Available at: https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html (accessed December 17, 2019).
19. Silberman, N., D. Hoiem, P. Kohli, and R. Fergus. 2012. Indoor segmentation and support inference from RGBD Images. *Computer vision*. Eds. A. W. Fitzgibbon, S. Lazebnik, P. Perona, *et al.* Lecture notes in computer science ser. 7576:746–760.
20. Vokhmintcev, A., T. Botova, I. Sochenkov, A. Sochenkova, and A. Makovetskii. 2017. Robot mapping algorithm based on Kalman filtering and symbolic tags. *Proc. SPIE* 10396:103962I. doi: 10.1117/12.2273562.
21. Vokhmintcev, A., M. Timchenko, A. Melnikov, A. Kozko, and A. Makovetskii. 2017. Robot path planning algorithm based on symbolic tags in dynamic environment. *Proc. SPIE* 10396:103962E. doi: 10.1117/12.2273279.
22. Sochenkov, I., and A. Vokhmintsev. 2015. Visual duplicates image search for a non-cooperative person recognition at a distance. *Procedia Engineer.* 129:440–445. doi: 10.1016/j.proeng.2015.12.147.

Received February 25, 2019

Contributors

Vokhmintcev Alexander V. (b. 1978) — Candidate of Science (PhD) in technology, associate professor; head of laboratory, Chelyabinsk State University, 129 Br. Kashirinyh Str., Chelyabinsk 454001, Russian Federation; associate professor, Ugra State University, 16 Chekhov Str., Khanty-Mansiysk, 628012, Russian Federation; vav@csu.ru

Melnikov Andrey V. (b. 1956) — Doctor of Science in technology, professor, Ugra State University, 16 Chekhov Str., Khanty-Mansiysk 628012, Russian Federation; melnikovav@uriit.ru

Pachganov Stepan A. (b. 1994) — PhD student, Ugra State University, 16 Chekhov Str., Khanty-Mansiysk 628012, Russian Federation; pachganovsa@uriit.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2020. Т. 14. Вып. 1. С. 113–120

АНАЛИТИЧЕСКАЯ ТЕКСТОЛОГИЯ В СИСТЕМАХ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ*

Е. Б. Козеренко[1], М. Ю. Михеев[2], Н. В. Сомин[3], Л. И. Эрлих[4], К. И. Кузнецов[5]

Аннотация: Представлено новое направление исследований на пересечении лингвистики, информатики и филологии с привлечением логико-статистических методов анализа неструктурированных данных в виде естественно-языковых текстов с целью решения целого ряда задач извлечения эксплицитных и имплицитных знаний из текстов с использованием семантически-ориентированного лингвистического процессора (СОЛП), формирования лексико-статистических представлений текстов, построения аналитических заключений, определения идиостиля автора и текстуального сходства произведений на основе анализа служебных слов и других микротекстовых элементов; выявления эмоциональной окрашенности текстов, построения полного профиля авторского текста на основе суперпозиции методов. Рассматривается пример текстологического анализа «Синей книги» из «Петербургского дневника» З. Н. Гиппиус.

Ключевые слова: обработка естественного языка; статистические методы; когнитивные технологии; лексико-семантический анализ; извлечение знаний из текстов; аналитические системы

DOI: 10.14357/19922264200115

1 Введение

В статье рассматриваются актуальные задачи и методы обработки естественно-языковых текстов с целью решения целого ряда задач извлечения эксплицитных и имплицитных знаний из текстов [1] с использованием семантически-ориентированного лингвистического процессора [2], формирования лексико-статистических представлений текстов [3–8], построения аналитических заключений [9, 10], определения идиостиля автора и текстуального сходства произведений на основе анализа служебных слов и других микротекстовых элементов [11]; выявления эмоциональной окрашенности текстов, построения полного профиля авторского текста на основе суперпозиции методов. Настоящая работа ставит своей целью описать методы аналитической текстологии и их возможные применения для обработки неструктурированных данных [12–15].

Способы представления информации, знаний многообразны. Огромный объем данных представлен в виде текстов естественного языка, что делает задачу извлечения и структурирования информации из текстов весьма важной. Это относится к различным предметным областям. Для оперирования данными на компьютере необходимо выделить из текста объекты, их атрибуты, связи между объектами, процессы, в которых эти объекты задействованы, другую важную информацию, которая бы позволяла не только описать ситуацию, но и строить выводы, характерные для конкретной предметной области, прогнозировать развитие ситуации [14–18].

Решение проблемы связано с анализом больших массивов текстов русского языка на основе технологической цепочки, организованной как последовательное применение инструментов построения лексико-статистического представления исследуемого текста, извлечения именованных сущностей и связей, определения идиостиля автора на основе служебных слов и микротекстовых элементов и определения близости текстов различных авторов.

*Работа выполнена при частичной поддержке РФФИ (проект 18-012-00220-а).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, kozerenko@mail.ru

[2]Научно-исследовательский вычислительный центр Московского государственного университета им. М. В. Ломоносова, mihej57@yandex.ru

[3]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, chri-soc@yandex.ru

[4]Научно-исследовательский вычислительный центр Московского государственного университета им. М. В. Ломоносова, levehr@yandex.ru

[5]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, k.smith@mail.ru

2 Формирование лексико-статистических представлений текстов

Для получения лексико-частотных характеристик текстов была разработана программная система ISV (Intelligent Statistical Verbalizer), реализующая три основные функции анализа текстов (разработчик программного обеспечения — Н. В. Сомин):

(1) лексический анализ, определяющий лексические особенности лексем;

(2) морфологический анализ русских и английских слов с выдачей канонической формы слова;

(3) каунтинг, т. е. подсчет числа вхождений слов с предварительным вычислением их канонической формы.

Необходимость разработки вербалайзера обусловлена прежде всего расширением и усложнением работ, связанных с аналитической обработкой текстов на естественном языке. В очень многих задачах предварительный морфологический анализ текста позволяет получить очень важную информацию о лексемах, которая в дальнейшем может использоваться для синтаксического анализа, семантического анализа и других фаз аналитической обработки текстов.

Предложим пример. Так, в комплексе задач, связанных с обнаружением плагиата и выяснением авторства, очень важен статистический анализ текстов. Однако прямой подсчет вхождений лексем оказывается малоэффективным, поскольку статистика числа вхождений «расплывается» по различным словоформам одного и того же слова и не дает объективной картины. Гораздо более точные результаты можно получить, применяя подсчет числа вхождений не для словоформ, а канонических форм слова.

Помимо этого, поскольку очень часто верхние строчки файла вхождений занимают предлоги, союзы, местоимения и другие вспомогательные слова, то в процессе анализа их следует отделять в особую группу (для последующей обработки микротекстовых элементов), для чего необходим достаточно глубокий морфологический анализ.

Кроме того, с помощью морфологического анализа возможен подсчет статистики по частям речи: только для существительных, только для глаголов и т. д., что позволит сделать анализ более дифференцированным, а значит выявить более тонкие закономерности.

Таким образом, одним из требований, предъявляемых к вербалайзеру, должна быть возможность комбинирования всех его возможностей с выбором необходимых для данной задачи.

Другое существенное требование — обеспечение высокой эффективности вычислений. Анализ больших текстов позволяет строить, например, терминологические портреты многих предметных областей, а также выявлять закономерности связей между текстами.

Чтобы успешно обрабатывать тексты больших объемов, необходима очень высокая скорость поиска и осуществления морфологического анализа. Это требует включения в вербалайзер техники обработки больших данных (big data), а также разработки особой структуры информационных массивов вербалайзера.

Данным блоком прежде всего решается задача структуризации текста. От правильного распознавания структуры текста в значительной степени зависит корректность всего анализа. При этом задача структуризации распадается на цепочку локальных задач: выделение из входного потока лексем; выделение предложений; выделение абзацев; унификация текста; исправление опечаток и грамматических ошибок; определение лексических признаков слов.

Отметим, что блок морфологического анализа выдает информацию о 97 различных морфологических признаках русского и английского языков, которых достаточно для анализа самых сложных языковых нюансов. В общем случае морфологический блок выдает несколько вариантов морфологического анализа (так называемая морфологическая омонимия). Это свойство морфологического анализа обеспечивается особой структурой словарей, а именно: в Словаре основ (СО) может быть несколько записей с одинаковой основой (но с разными классами окончаний), а на один и тот же класс окончаний может ссылаться несколько слов с разными основами. Возможны случаи пустой основы (пример: «хорошо»—«лучше») и пустого класса окончаний (для неизменяемых слов). Кроме основы и вариантов окончаний в Словаре классов окончаний (СКО) хранятся морфологические признаки, соответствующие определенному классу окончаний в целом (постоянная морфологическая информация) и каждому окончанию парадигмы в отдельности (переменная морфологическая информация). Алгоритм предполагает, что в общем случае могут быть найдены несколько вариантов морфологического разбора. Этот факт хорошо известен лингвистам как морфологическая омонимия. Например, слово «стекло» имеет по крайней мере два варианта разбора: как существительное (вставить *стекло*) и как глагол (что-то *стекло* с крыши).

3 Концептуальный анализ текстов на основе семантически-ориентированного лингвистического процессора

Задача автоматического анализа текстовой информации, представленной в интернете, актуальна во всем мире. Для решения полного спектра задач обработки естественного языка создан СОЛП [2]. Центральным компонентом СОЛП является инструментальный пакет (SDK-модуль) PullEnti (Puller of Entities). Этот процессор в рамках соревнований, проводившихся на конференции «Диалог-2016», занял два первых места при анализе текстов в рамках решения задач извлечения именованных сущностей. Разработчик PullEnti — Константин Игоревич Кузнецов. В системе PullEnti используются динамически подключаемые компоненты (плагины), что позволяет без перекомпилирования запускать различные функциональные возможности. Именно таким образом активируется блок семантического анализа.

В процессе анализа выделяются семантические единицы (токены), которые представляют собой типизированные фразы, такие как текстовые, числовые и т. д. Например, в результате анализа фразы «в 1917 году» будут выделены три токена: «в» — текстовый, «году» — текстовый, «1917» — числовой. Такие токены можно назвать простыми. Кроме того, выделяются метатокены — сложные токены, которые объединяют несколько простых токенов, например существительные с определителями, скобки, кавычки и т. п.

Первый этап концептуального анализа текстовых сообщений — выделение параметров команд. Этот этап проводится с помощью инструментального пакета PullEnti, предназначенного для решения задачи выделения именованных сущностей, их свойств и связей из неструктурированных русскоязычных текстов в рамках информационных систем, разрабатываемых на .NET Framework 2.0 и выше. PullEnti состоит из общей и специализированной частей. Общая часть обеспечивает реализацию общих алгоритмов морфологического и синтаксического анализа, а также поддержку модели данных. Специализированная часть состоит из отдельных сборок (анализаторов), реализующих выделение именованных сущностей определенных типов (персоны, даты, локации, организации и др.). Предполагается также выявление ассоциативных связей между выделенными сущностями в определенной предметной области. При этом для расчета силы ассоциативной связи между именованными сущностями используется косинусная мера между контекстными векторами (компонентами вектора именованных сущностей служат частоты их совместной встречаемости в одном и том же контексте). Такие векторы образуют семантическое контекстное пространство. Для вычисления косинусной меры между контекстными векторами используется следующая формула:

$$\frac{x \cdot y}{|x| \cdot |y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}.$$

В зависимости от того, какие контексты считаются идентичными, различают типы контекстов. Классификация и даже перечисление типов контекстов — проблема, которая в силу своей новизны требует особого рассмотрения [3, 16].

4 Стейплинг: построение идиостиля автора и определение близости текстов

Метод стейплинга (от английского staple — скрепа) основан на результатах исследований М. Ю. Михеева и Л. И. Эрлиха, в которых продемонстрирована высокая степень информативности слов закрытых классов (служебных слов — союзов, предлогов, частиц, — дискурсивных слов, вводных оборотов, устойчивых наречных конструкций и т. п.) для решения задач определения авторства текста по частотам служебных слов, построения идиостилистического профиля автора, установления близости текстов, определения текстовых заимствований [11]. Такие элементы языка получили название языковой, или текстовой, скрепы (staple) и служат потенциальными статистическими маркерами (СМ) автора.

Стейплинг-подход к текстологическим исследованиям является новым и менее известен, чем составление авторских словарей ключевых и опорных слов. Автора и его текст могут исчерпывающим образом характеризовать не только ключевые слова, но и слова наименее значимые, при помощи списка из сотни наиболее частотных подобных скреп можно будет находить реального автора текста, определять близость исследуемого текста к стилистике текстов базы сравнения. Для составления 100-словного, а затем 300-словного списка русских служебных слов и выражений, т. е. наиболее частых в языке союзов, предлогов, частиц, дискурсивных слов,

вводных оборотов, устойчивых наречных конструкций, фразеологизмов, стандартных средств пара- и гипотаксиса, а также наиболее употребительных сочетаний из них, проведены исследования в рамках двух этапов проекта РФФИ «Создание алгоритма идентификации авторского идиостиля на основании частотности употребления служебных слов» (руководитель М. Ю. Михеев). В настоящее время ведутся работы по расширению, уточнению списка скреп, рассмотрению их контекстных характеристик, выявлению дистрибутивных признаков этих элементов текста.

Для проведения достоверных исследований необходима база авторских текстов. За исходную базу взят Национальный корпус русского языка (НКРЯ), а из него — тексты семи наиболее известных русских писателей-прозаиков XIX–XX вв.: Гоголя, Тургенева, Достоевского, Толстого, Бунина, Горького и Набокова. В этих подкорпусах подсчитано число употреблений всех единиц 100-словного списка и вычислены их частоты (в миллионных долях — ipm, или миллипромилле, т. е. 1/10000-й части процента).

Как показали системные исследования авторских текстов, высокой информативностью обладают относительные частоты маркеров: более наглядным представляется сравнивать не сами значения частот данного СМ в разных текстах разных авторов, а их отношения к среднему уровню, т. е. выраженное в процентах отношение частоты какой-то конкретной скрепы у конкретного автора к средней частоте этой же скрепы по НКРЯ. Метод стейплинга эффективно применяется для создания идиостилистического профиля автора. Истоки метода обнаруживаются в первых исследованиях идиостиля и связаны с именами Ю. Н. Тынянова, Ю. Н. Караулова и В. В. Виноградова. В частности, В. В. Виноградов ввел термин «языковая личность». Идиостиль — это система содержательных и формальных лингвистических характеристик, присущих произведениям определенного автора, которая делает уникальным воплощенный в этих произведениях авторский способ языкового выражения. Идиостиль близок к понятию идиолект.

В задаче распознавания авторства более важным оказывается не то, ***о чем*** говорит автор, а то, ***как*** он это говорит. В лингвостатистике существует 100-словный список Сводеша, задающий, как известно, лексику, наименее подверженную изменениям в данном языке, по которой можно рассчитать скорость синонимических замен базового лексического фонда. Сам список предполагается примерно одинаковым для любых языков и служит как бы «лингвистическими часами» — по изменениям в нем можно определить время распада языка.

Методика выявления идиостилистического профиля автора: выявляется набор наиболее частых в его текстах, характерных именно для него служебных единиц языка, элементов 100-словного списка. Для краткости далее все элементы такого списка будут именоваться просто текстовыми, или языковыми скрепами. Они же выступают потенциальными СМ стиля писателя, обособляя один идиостиль от другого.

5 Полный профиль авторского текста (на примере «Синей книги» из «Петербургского дневника» З. Н. Гиппиус)

Авторский текст — непосредственный вход в интеллектуальный и эмоционально-аффективный мир человека, данный нам в «ощущении через восприятие текста», актуализированные мысль и чувство пишущего. В дополнение к текстам, представленным в НКРЯ был проведен ряд исследований текстов «Петербургского дневника» З. Н. Гиппиус [9, 10]. Несмотря на большое влияние поэта, критика и философа З. Н. Гиппиус на современников, ее произведения [19, 20] относительно мало изучены отечественной лингвистикой в связи с их недоступностью в советский период. На примере анализа текста «Синей книги» рассмотрим последовательное применение методов аналитической текстологии.

«Синяя книга» — первая часть «Петербургского дневника» З. Н. Гиппиус [18, с. 51–241; 19] — содержит описание хронологии событий и настроений с августа 1914 г. по ноябрь 1917 г., впечатления и мысли автора. Ярко и образно представлены события предреволюционного периода и непосредственно двух всплесков революции — в феврале и октябре 1917 г. Построение лексико-статистического образа текста с помощью системы ISV позволило получить наиболее частотные ключевые слова и скрепы. Далее были выделены семантически значимые объекты, именованные сущности и связи с использованием СОЛП.

Рассмотрим полученные результаты.

Наиболее частотные значимые (опорные) слова — «герои повествования»: война, революция.

Базовые семантические классы (типы именованных сущностей): даты, события, люди, стихии, атмосфера, общий строй жизни, настроение автора, впечатления и оценки автора.

Основная тема предреволюционного периода звучит рефреном: тяжесть войны и ожидание надвигающейся катастрофы: «Война — в статике», «Греция замерла», «никому нет никуда выхода. И не предвидится»; «Какая-то ЧРЕВАТОСТЬ в воздухе; ведь нельзя же только — ЖДАТЬ!»; «В атмосфере глубокий и зловещий ШТИЛЬ»; «Оцепенели»; «Спокойствие. . . отчаянья».

Ключевые узлы развития событий по датам:

Зима 1916 года; Лето 1916 года; 1917 2 февраля. Четверг; 11 февраля. Суббота; 22 февраля. Среда; 23 февраля. Четверг: начало движения; «26 февраля. Воскресенье»; 27 февраля. Понедельник 12 ч. дня; 2 ч. дня; 3 ч. дня; 4 1/2 часа; 5 часов; 5 1/2.

Персоны: Чхенкели, Вильсон, Керенский, Милюков, Гришка, Пуришкевич, Протопопов, Родзянко, Брусилов, Рузский, ген. Алексеев, Коновалов, Дмитрюков, Чхеидзе, Шульгин, Шидловский, Милюков, Караулов, Львов и Ржевский.

Локации: Санкт-Петербург, Выборгская сторона, Таврический сад.

Организации: Дума, Городская Дума, Комитет «для водворения порядка и для сношения с учреждениями и лицами».

Оценочные суждения автора: «При этом плохо везде. Истощение и неустройство»; «У нас особенно худо. Нынешняя зима впятеро тяжелее и дороже прошлогодней. Рядом — постыдная роскошь наживателей»; «Грозная, страшная сказка»; «. . .столько знакомых, милых лиц, молодых и старых»; «Но все лица, и незнакомые, — милые, радостные, верящие какие-то. . . Незабвенное утро, алые крылья и марсельеза в снежной, золотом отливающей, белости. . .»

Скрепы в «Синей книге» образуют отчетливое тематическое гнездо «неопределенности»: какая-то, кто-то где-то, будто бы, кем-то и другие: никому нет никуда {выхода}; какая-то {ЧРЕВАТОСТЬ}; никто не {сомневается}; никто не {знает}; никто не {думает}; кто-то где-то {обмолвился}; {опираются} на какие-то {слова}; {случилось большое} «Ничего»; {было —} Ничего; как-то {внезапно}; никто, {конечно, в точности} ничего не {знает}; кое-где {остановили трамваи (и разбили)}; будто бы {убили}; будто бы {пошли}; {все} «будто бы»; где оно и кто; и что бы ни было {дальше}; {верящие} какие-то; {не простится —} кем-то, чем-то; кто-нибудь. Какие-нибудь {третьи}; какой-то {подлый слой}; {министерством} якобы {«доверия»}; {Дума} будто бы {решила}; {она}, кажется, {там сидит}; {Солдаты}, кажется, {были выпивши}; в какой {зале} — не {знаю}; В какой {они связи с Комитетом} — не {выясняется}.

6 Заключение

Лексико-статистический анализ текстов естественных языков предназначен для установления статистических закономерностей встречаемости наименований понятий, служебных слов, оборотов. Полученные в результате такого анализа закономерности позволяют не только автоматически распознавать именованные сущности, но и использовать их для установления системы взаимосвязей понятий при формировании предварительных словарей парадигматических и ассоциативных связей.

Привлечение ряда методов текстологического анализа и выстраивание их в последовательную технологическую цепочку: лексико-статистический анализ, извлечение именованных сущностей и формирование классификации ключевых и опорных слов (основы для построения онтологии), выявление и частотный анализ служебных слов и других скреп позволяет более эффективно решать целый ряд задач аналитической обработки неструктурированных данных, каковыми являются естественно-языковые тексты.

Литература

1. *Kuznetsov I. P., Kozerenko E. B., Matskevich A. G.* Intelligent extraction of knowledge structures from natural language texts // IEEE/WIC/ACM Joint Conferences (International) on Web Intelligence and Intelligent Agent Technology — Workshops WI-IAT Proceedings. — Lyon, France: IEEE Computer Society, 2011. P. 269–272.
2. *Kozerenko E. B., Kuznetsov K. I., Morozova Yu. I., Romanov D. A.* Semantic proximity establishment in the tasks of knowledge extraction and named entities recognition // 19th Conference (International) on Artificial Intelligence, WORLDCOMP'17 Proceedings. — Las Vegas, NV, USA: CSREA Press, 2017. P. 339–344.
3. *Dempster A. P., Laird N. M., Rubin D. B.* Maximum likelihood from incomplete data via the EM algorithm // J. Roy. Stat. Soc. B, 1977. Vol. 39. No. 1. P. 1–22.
4. *Rapp R.* Word sense discovery based on sense descriptor dissimilarity // 9th Machine Translation Summit Proceedings. — New Orleans, LA, USA, 2003. P. 315–322.
5. *Lenci A.* Distributional semantics in linguistic and cognitive research // Riv. Linguist., 2008. Vol. 1. P. 1–30.
6. *Turney P.* A uniform approach to analogies, synonyms, antonyms and associations // 22nd Conference (International) on Computational Linguistic Proceedings. — Manchester, 2008. P. 905–912.
7. *Baroni M., Lenci A.* Distributional memory: A general framework for corpus-based semantics // Comput. Linguist., 2010. Vol. 36. Iss. 4. P. 673–721.
8. *Schumann A.* Towards the automated enrichment of multilingual terminology databases with knowledge-rich contexts // Computational Linguistics and Intellectual Technologies: Conference (International) "Dialogue 2012"

Proceedings. — Moscow: RGGU, 2012. Vol. 1. Iss. 11. P. 559–567.

9. *Козеренко Е. Б.* «Наших дедов мечта невозможная» — Учредительное собрание в Черных тетрадях Зинаиды Гиппиус // Маргиналии-2015: границы культуры и текста: Тезисы докл. Междунар. конф. / Под ред. Е. Б. Козеренко, А. Г. Кравецкого, М. Ю. Михеева. — Полоцк, 2015. http://uni-persona.srcc.msu.ru/site/conf/marginalii-2015/thesis.htm.
10. *Козеренко Е. Б.* Февраль 17-го в «Синей книге» З. Н. Гиппиус: опыт текстологического анализа // Маргиналии-2017: границы культуры и текста: Тезисы докл. Междунар. конф. / Под ред. А. Г. Кравецкого, М. Ю. Михеева.— Торжок, 2017. http://uni-persona.srcc.msu.ru/site/conf/marginalii-2017/thesis.htm.
11. *Михеев М. Ю., Эрлих Л. И.* Идиостилевой профиль и определение авторства текста по частотам служебных слов // НТИ. Сер. 2: Информ. процессы и системы, 2018. № 2. С. 25–34.
12. *Charnine M. M., Kuznetsov I. P., Kozerenko E. B.* Semantic navigator for Internet search // Conference (International) on Machine Learning Proceeding. — Las Vegas, NV, USA: CSREA Press, 2005. P. 60–68.
13. *Кузнецов И. П., Сомин Н. В.* Выявление имплицитной информации из текстов на естественном языке: проблемы и методы // Информатика и её применения, 2012. Т. 6. Вып. 1. С. 48–57.
14. *Kuznetsov I. P., Kozerenko E. B., Charnine M. M.* Technological peculiarity of knowledge extraction for logical-analytical systems // WORLDCOMP'12: ICAI'12 Proceedings. — Las Vegas, NV, USA: CSREA Press, 2012. Vol. II. P. 762–768.
15. *Шарнин М. М., Кузнецов И. П.* Особенности семантического поиска информационных объектов на основе технологии баз знаний // Информатика и её применения, 2012. Т. 6. Вып. 2. С. 47–56.
16. *Lund K., Burgess C.* Producing high-dimensional semantic spaces from lexical co-occurrence // Behav. Res. Meth. Ins. C., 1996. Vol. 28. No. 2. P. 203–208.
17. *McCarthy D., Koeling R., Weeds J., Carroll J.* Finding predominant senses in untagged text // 42nd Annual Meeting of the Association for Computational Linguistics Proceedings. — Barcelona, Spain: ACL, 2004. P. 280–287.
18. *Baroni M., Zamparelli R.* Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space // Conference on Empirical Methods in Natural Language Processing Proceedings. — Stroudsburg, PA, USA: ACL, 2010. P. 1183–1193.
19. *Гиппиус З. Н.* Дневники. — М.: Захаров, 2017. 528 с.
20. Синяя книга // Интернет-ресурс произведений З. Н. Гиппиус. https://gippius.com/doc/memory/sinyaya-kniga.html.

Поступила в редакцию 15.01.20

ANALYTICAL TEXTOLOGY IN INTELLIGENT PROCESSING SYSTEMS FOR UNSTRUCTURED DATA

E. B. Kozerenko[1], M. Y. Mikheev[2], N. V. Somin[1], L. I. Ehrlich[2], and K. I. Kuznetsov[1]

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation

[2]Research Computing Center Lomonosov Moscow State University, 1, bld. 4 Leninskie Gory, Moscow, GSP-1, 119991, Russian Federation

Abstract: The paper presents a new field of research at the intersection of linguistics, computer science, and philology involving logical and statistical methods of analyzing unstructured data in the form of natural language texts in order to solve a number of the tasks of extracting explicit and implicit knowledge from texts using a semantics-oriented linguistic processor, forming lexical statistical representations of texts, building analytical conclusions, discovery of the author's idiostyle and textual similarity of literary works based on the analysis of service words and other microtext elements; identifying the sentiment of texts, building a full profile of the author's text based on the superposition of methods. The example of the textological analysis of the "Blue Book" of the "Petersburg Diary" by Zinaida Hippius is considered.

Keywords: natural language processing; statistical methods; cognitive technology; lexical semantic analysis; knowledge extraction from texts; analytical systems

DOI: 10.14357/19922264200115

Acknowledgments

The paper was partially supported by the Russian Foundation for Basic Research (project 18-012-00220-a).

References

1. Kuznetsov, I. P., E. B. Kozerenko, and A. G. Matskevich. 2011. Intelligent extraction of knowledge structures from natural language texts. *IEEE/WIC/ACM Joint Conferences (International) on Web Intelligence and Intelligent Agent Technology Proceedings — Workshops WI-IAT Proceedings*. Lyon, France: IEEE Computer Society. 269–272.
2. Kozerenko, E. B., K. I. Kuznetsov, Yu. I. Morozova, and D. A. Romanov. 2017. Semantic proximity establishment in the tasks of knowledge extraction and named entities recognition. *Conference (International) on Artificial Intelligence, WORLDCOMP'17 Proceedings*. Las Vegas, NV: CSREA. 339–344.
3. Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* 39(1):1–22.
4. Rapp, R. 2003. Word sense discovery based on sense descriptor dissimilarity. *9th Machine Translation Summit Proceedings*. New Orleans, LA. 315–322.
5. Lenci, A. 2008. Distributional semantics in linguistic and cognitive research. *Riv. Linguist.* 1:1–30.
6. Turney, P. 2008. A uniform approach to analogies, synonyms, antonyms and associations. *22nd Conference (International) on Computational Linguistic Proceedings*. Manchester. 905–912.
7. Baroni, M., and A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Comput. Linguist.* 36(4):673–721.
8. Schumann, A. 2012. Towards the automated enrichment of multilingual terminology databases with knowledge-rich contexts. *Computational Linguistics and Intellectual Technologies: Conference (International) "Dialogue 2012" Proceedings*. Moscow. 1(11):559–567.
9. Kozerenko, E. B. 2015 «Nashikh dedov mechta nevozmozhnaya» — Uchreditel'noe sobranie v Chernykh Tetradyakh Zinaidy Gippius ["The impossible dream of our grandfathers" — Constituent assembly in the Black Notebooks of Z. N. Hippius]. Eds. E. B. Kozerenko, A. G. Kravetsky, and M. Y. Mikheev. *Conference (International) "Marginalia-2015: Borders of Culture and Text" Proceedings*. Polotsk. Available at: http://uni-persona.srcc.msu.ru/site/conf/marginalii-2015/thesis.htm (accessed March 10, 2020).
10. Kozerenko, E. B. 2017. Fevral' 17-go v «Siney knige» Z. N. Gippius: opyt tekstologicheskogo analiza [February of 17th in the "Blue book" of Z. N. Hippius: The case of the textological analysis]. Eds. A. G. Kravetsky and M. Y. Mikheev. *Conference (International) "Marginalia-2017: Borders of Culture and Text" Proceedings*. Torzhok. Available at: http://uni-persona.srcc.msu.ru/site/conf/marginalii-2017/thesis.htm (accessed March 10, 2020).
11. Mikheev, M. Yu., and L. I. Ehrlich. 2018. Idiostilevoy profil' i opredelenie avtorstva teksta po chastotam sluzhebnykh slov [Individual style profile and text authorship detection based on the service words frequencies]. *Nauchno-technicheskaya informatsia. Ser. 2. Informatsionnye protsessy i sistemy* [Scientific Technical Information. Ser. 2. Information Processes and Systems] 2:25–34.
12. Charnine, M. M., I. P. Kuznetsov, and E. B. Kozerenko. 2005. Semantic navigator for Internet search. *Conference (International) on Machine Learning Proceeding*. Las Vegas, NV: CSREA Press. 60–68.
13. Kuznetsov, I. P., and N. V. Somin. 2012. Vyyavlenie implitsitnoy informatsii iz tekstov na estestvennom yazyke: problemy i metody [Revealing implicit information from texts in natural language: Problems and methods]. *Informatika i ee Primeneniya — Inform. Appl.* 6(1):48–57.
14. Kuznetsov, I. P., E. B. Kozerenko, and M. M. Charnine. 2012. Technological peculiarity of knowledge extraction for logical-analytical systems. *WORLDCOMP'12: ICAI'12 Proceedings*. Las Vegas, NV: CSREA Press. II:762–768.
15. Charnine, M. M., and I. P. Kuznetsov. 2012. Osobennosti semanticheskogo poiska informatsionnykh ob"ektov na osnove tekhnologii baz znaniy [The peculiarities of the semantic search of information objects founded on the knowledge bases technology]. *Informatika i ee Primeneniya — Inform. Appl.* 6(2):47–56.
16. Lund, K., and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Meth. Ins. C.* 28(2):203–208.
17. McCarthy, D., R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant senses in untagged text. *42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona, Spain: ACL. 280–287.
18. Baroni, M., and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. *Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: ACL. 1183–1193.
19. Hippius, Z. N. 2017. *Dnevniki* [Diaries]. Moscow: Zakharov. 528 p.
20. The Internet resource of Z. N. Hippius works, "Sinyaya kniga." Available at: https://gippius.com/doc/memory/sinyaya-kniga.html (accessed January 27, 2020).

Received January 15, 2020

Contributors

Kozerenko Elena B. (b. 1959) — Candidate of Science (PhD) in linguistics, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; kozerenko@mail.ru

Mikheev Michael Yu. (b. 1957) — Doctor of Science in linguistics, leading scientist, Research Computing Center Lomonosov Moscow State University, 1, bld. 4 Leninskie Gory, Moscow, GSP-1, 119991, Russian Federation; mihej57@yandex.ru

Somin Nikolai V. (b. 1947) — Candidate of Science (PhD) in physics and mathematics, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; chri-soc@yandex.ru

Ehrlich Lev I. (b. 1948) — leading engineer, Research Computing Center Lomonosov Moscow State University, 1, bld. 4 Leninskie Gory, Moscow, GSP-1, 119991, Russian Federation; levehr@yandex.ru

Kuznetsov Konstantin I. (b. 1968) — leading engineer, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; k.smith@mail.ru

ИНКАПСУЛЯЦИЯ СЕМАНТИЧЕСКИХ ПРЕДСТАВЛЕНИЙ В ЭЛЕМЕНТЫ ГРАММАТИКИ

Ш. Б. Шихиев[1], Ф. Ш. Шихиев[2]

Аннотация: Предлагается новый математический аппарат представления естественного языка (ЕЯ) для компьютерной лингвистики — морфология, синтаксис и семантика описаны как предметы дискретной математики, образующие иерархию и целостную информационную систему. Предлагаемая конструктивная теория языка представляет собой новый подход к изучению языка путем разделения полномочий синтаксиса и семантики; построения автономных моделей синтаксиса и семантики; формирования языка как отображения элементов двух множеств: синтаксиса и семантики.

Ключевые слова: естественный язык; граф; синтаксис; семантика; лексика; словоформа; морфологический признак; лексическая группа; словарь; предложение; алгоритм

DOI: 10.14357/19922264200116

1 Введение

А. В ЕЯ переплетены два автономных явления: дискретное (грамматика) и аналоговое (семантика). Грамматика (морфология и синтаксис) может быть описана на языке математики. Морфологические формы слов позволяют их различать и приписывать им и их сочетаниям различные значения, которые фиксируются в семантическом словаре. В данной статье показано, как это можно сделать.

Модель морфологии заключает в себе правила построения словоформ; словоформа представлена *словом* в исходной форме и ее *морфологическими параметрами* из десятичных цифр. Словоформы с одинаковыми *морфологическими параметрами* (*формами*) образуют *лексическую группу*.

Модель синтаксиса ЕЯ строится исходя из следующих предположений: в ЕЯ имеются такие *элементарные предложения*, что из n членов предложения всегда можно составить $(n-1)$ *словосочетаний* (*синтаксически связанных пар словоформ*), таких что члены предложения образуют *связный граф*; следовательно, объединение *элементарных предложений* образует *синтаксический граф* $\mathrm{Sint} = (X, Y)$ из множеств X (*словоформ*) и Y (*словосочетаний*). Критерий связности двух словоформ определяется через их *морфологические формы* (*параметры*). Нетрудно проверить [1], что миллионы *словосочетаний*, имеющих место в грамматике русского языка, распределены по трем сотням *синтаксических отношений* — прямых произведений лексических групп, что упрощает представление *графа* Sint в памяти компьютера.

Б. Предлагаемая сетевая модель грамматики Sint открывает следующие возможности в изучении ЕЯ и его реализации на компьютере:

1. Корневые деревья из модели Sint порождают *элементарные предложения-деревья*. Обход *предложения-дерева* сопоставляет ему *предложение-последовательность*; корректность обратной задачи, известной как *синтаксический анализ предложения*, зависит от того, в какой степени соблюдены принципы *фрагментарности сегментов* между этими предложениями, т. е. являются ли *сегменты* (вершины ветвей) в *предложении-дереве фрагментами* в соответствующем *предложении-последовательности* [2].
2. Из элементарных предложений строятся более сложные предложения по правилам синтаксиса. Из правил построения модели Sint следует возможность построения такой модель грамматики, что любое предложение русского языка (например, предложения, встречающиеся в литературе на русском языке) порождается элементами синтаксиса Sint; следовательно, имеется формальное определение *синтаксически правильно построенного предложения* и соответствующий алгоритм распознавания таких предложений [2].
3. *Модель морфологии* (правила преобразования слов) представлена в *морфологическом словаре*; программа, реализующая эти правила, образует *компьютерную модель морфологии*. Приемы реализации *модели синтаксиса* (алгоритмов анализа и синтеза предложений) демонстрируются

[1]Дагестанский государственный университет, sh_sh_b51@mail.ru
[2]Дагестанский государственный университет, fuad@mail.ru

в [3]. Значения предложений в синтаксисе Sint представлены в *семантическом словаре* в виде классов для реализации *семантической модели семантики* посредством объектных технологий программирования.

В. Обращаясь к истории вопроса, можно напомнить следующее. В данной работе реализована идея Ф. де Соссюра и Л. Ельмслева, согласно которой строится автономное и конструктивное описание синтаксиса, порождающего «планы выражений», далее в каждый «план выражения» инкапсулируется «план содержания» — элемент семантики, а *язык* становится отображением (биекцией) элементов двух множеств: синтаксиса и семантики.

Представляется, что только разделение полномочий синтаксиса и семантики открывает путь к формализации ЕЯ, иначе придется согласиться с утверждениями типа: «Никто не может сформулировать все правила английской грамматики. . .» [4].

В учебниках по «Общему синтаксису» [5] и в работах И. А. Мельчука [6] осторожно указывалось на древовидность структуры предложения и графы использовались только для демонстрации сетевой структуры предложения. Нужно было решиться и выделить класс *элементарных предложений*, имеющих структуру *корневого дерева*, а далее заметить, что все другие формы предложения (с однородными членами, сложные предложения и т. д.) собраны из элементарных предложений.

Многие правила построения синтаксически правильных сочетаний можно представить в виде правил контекстно-свободной грамматики (например, выражения, образованные из согласованных и несогласованных определений). Этот факт был использован для представления *правил синтаксиса* ЕЯ посредством *подстановок* и *деревьев разбора*. Предложение «Сколько чувствительности контекста требуется, чтобы предоставлять разумные структурные описания?» (How much context-sensitivity is required to provide reasonable structural descriptions?) из [7] указывает на то, что автор «Грамматики сложения деревьев» (Tree adjoining grammars) далек от мысли раздельного исследования синтаксиса и семантики.

2 Морфология

Морфология есть структура, заданная тройкой (A, L_0, F_0), где A — *алфавит*; L_0 — *исходная лексика*, представляющая собой конечное множество *исходных слов* над алфавитом A; F_0 — конечный набор *исходных морфологических признаков*, представляющих собой двухразрядные десятичные числа.

Исходные признаки разбиты на непересекающиеся подмножества; элементы каждого подмножества образуют линейный массив, который называется *категорией* (*признаков*). Категорий в морфологии русского языка меньше десяти, и они именованы кодами: 10 (род), 20 (число), 30 (падеж), 40 (степень) и т. д., а исходные признаки в категориях кодированы следующим образом: $10 = (11, 12, 13)$, $20 = (21, 22)$, $30 = (31, 32, 33, 34, 35, 36)$, $40 = (41, 42, 43)$ и т. д., или в более привычной записи: $10 =$ (м. род, ср. род, ж. род), $20 =$ (ед. число, мн. число), $30 =$ (И., Р., Д., В., Т., П.), $40 =$ (полная форма ИП, краткая форма ИП, сравнительная степень ИП) и т. д.

Исходная лексика также разбита на непересекающиеся подмножества — *исходные части речи*; к каждой *исходной части речи* D_0 прикреплен свой набор категорий Ψ и множество *признаков* Ω; *признак* представляет собой строку $f =$ «$\alpha_1\alpha_2\cdots\alpha_k$» из *исходных признаков*, принадлежащих различным категориям из Ψ. Один из признаков f_0 называется *начальным признаком*. Слова из D_0 обладают *начальным признаком*; *исходное слово* s_0 представлено строкой вида «$s_0 : f_0$»; форма s слова s_0 с признаком f будет представлена в виде «$s_0 : f$»: например, **дом** : 2133 = **дому**.

Части речи именованы кодами: 01 — имя существительное (ИС), 02 — имя прилагательное (ИП), 07 — глагол и т. д. Категориями для ИС являются 20 и 30, а признаками будут $\Omega = \{2131, 2132, 2133, 2134, 2135, 2136, 2231, 2232, 2233, 2234, 2235, 2236\}$. Для ИП из категорий 10, 20, 30 и 40 составлено 29 признаков: 41112131, $41112132, 41112133, 41112134, 41112135, \ldots, 43$.

Множество $s_0 : \Omega = \{s_0 : f | f \in \Omega\}$ называется *морфологической группой* слова s_0, а ее элементы называются формами слова s_0, или просто *словоформами*. Группа $s_0 : \Omega$ для ИС s_0 состоит из 12 слов, а для ИП — из 29 словоформ; нетрудно заметить, что элементы множества $s_0 : \Omega$ различны.

Объединение множеств $s_0 : \Omega$ по всем s_0 из D_0 обозначается через D и называется *частью речи*. Морфология теперь может быть представлена четверкой $(A, L, \Psi 1, \Omega 1)$, где лексика L — объединение всех *частей речи*; $\Psi 1$ — множество категорий; $\Omega 1$ — множество признаков.

Числовые признаки при исходной форме слова обозначают формы слова, на которых будет построен синтаксис (и предложения) языка. Для преобразования словоформ ЕЯ в слова с числовыми признаками и обратно нужны соответствующие *морфологические правила*; их, как известно, можно найти в словообразовательных словарях.

Рассмотрим слово $s_0 : \alpha_1\alpha_2\cdots\alpha_k$. По определению исходный признак α_k принадлежит не-

которой категории F_k. Морфология ЕЯ обладает таким свойством, что признаками морфологии выступают все строки $\alpha_1\alpha_2\cdots\alpha_k$, где α_k пробегает элементы массива F_k; они образуют множество $\alpha_1\alpha_2\cdots\alpha_{k-1}F_k$ — *парадигму* слова s_0 по категории F_k. Формы слова сгруппированы по *парадигмам*. В электронном словаре вместо элементов *парадигм* будут записаны алгоритмы (морфологические правила), порождающие их элементы.

Например, в строке

«2130дом,дома,дому,дом,домом,доме»

за *парадигмой* 2130 слова **дом** перечислены его элементы. Есть возможность вместо словоформ записать их постфиксы следующим образом:

дом0115:213000,а,у,,ом,е;

223000а,ов,ам,а,ами,ах. (1)

В первой части «дом0115» статьи (1) за словом «дом» указаны его *грамматические атрибуты* (01 — код ИС, 15 — мужской род и неодушевленное); во второй части перечислены два *морфологических правила* (разделенных точкой с запятой), по которым строятся элементы *парадигмы* 2130 и 2230. К коду *парадигмы* приписано двухразрядное число — длина изменяемой части словоформ из этой парадигмы. Для экономии памяти имеет смысл хранить отдельным списком постфиксы элементов *парадигмы*, а в статьях словаря указать их порядковые номера.

Построение словарных статей представляет собой рутинную работу, которую также можно запрограммировать.

Пусть s — обычная форма слова s_0 с признаком $\alpha_1\alpha_2\cdots\alpha_k$, т. е. $s = s_0 : \alpha_1\alpha_2\cdots\alpha_k$. В языковом явлении морфология решает две задачи: *синтеза* — перехода от $s_0 : \alpha_1\alpha_2\cdots\alpha_k$ к s — и *анализа* — перехода от s к $s_0 : \alpha_1\alpha_2\cdots\alpha_k$.

Чтобы осуществить *анализ* словоформы, потребуется осуществить синтез всех элементов множества $s_0 : \Omega$. Анализ словоформы усложняется тем, что по форме слова s практически невозможно точно угадать его исходную форму s_0; а словарная статья начинается с исходной формы слова s. Анализ словоформы — трудоемкая процедура, для повышения ее эффективности требуется использовать различные приемы поиска из дискретного анализа.

3 Синтаксис

Пусть $D^1, D^2, \ldots, D^q$ — коды частей речи (как изменяемых, так и наречия, которые образуют одну *лексическую группу*) в морфологии $\mu = (A, L, \Psi 1, \Omega 1)$, для них определены множества признаков $\Omega^1, \Omega^2, \ldots, \Omega^q$ соответственно.

Если $a \in \Omega^i$, то через $a : D^i$, или $D^i a$, обозначается множество слов из D^i, обладающих признаком a, и называется *лексической группой* по признаку a. Например, $2135(01) = 012135 = \{$**домом**, **точкой**, **едой**, $\ldots\}$.

Через $\Omega^i(D^i)$, или $[D^i]$, обозначается множество, состоящее из *лексических групп* $D^i a$ по всем признакам a из Ω^i. Например, множество [01] состоит из 12 *лексических групп*: $012131, 012132, 012133, 012134, \ldots, 012236$.

Через Λ_μ обозначено объединение всех $[D^1], [D^2], \ldots, [D^q]$. Построение синтаксиса начинается с выбора нескольких пар *лексических групп* из множества Λ_μ:

$$X_1 \text{ и } Y_1\,;\ X_2 \text{ и } Y_2\,;\ X_3 \text{ и } Y_3\,;\ \ldots\,;\ X_k \text{ и } Y_k\,. \qquad (2)$$

Через R обозначено объединение произведений:

$$R = X_1 * Y_1 \cup X_2 * Y_2 \cup \cdots \cup X_k * Y_k\,. \qquad (3)$$

Произведение *лексических групп* $X_i * Y_i$ ($i = 1, \ldots, k$) называется *синтаксическим отношением* (СО), их элементы — *синтаксически связанными словоформами* или *словосочетаниями*; в *словосочетании* (x, y) слово x — главный член, y — зависимый член сочетания. Например, $012131 * 012132 = \{$(**дом**, **моды**), (**запах**, **дыма**), (**небо**, **сна**), $\ldots\}$.

Орграф (L, R) задает *синтаксис* Sint. Орграф $(L1, R1)$, где $L1$ состоит из *лексических групп* (2), а $R1$ — из пар (X_i, Y_i), где $i = 1, \ldots, k$, также задает *синтаксис* Sint в упакованном виде; пусть $\mathrm{Sint}1 = (L1, R1)$. Через $\mathrm{Sint}\,(\Lambda_\mu)$ обозначается некоторый синтаксис (грамматика), заданный на Λ_μ.

Синтаксические отношения (3) задаются парами морфологических признаков

$$f_1 \text{ и } g_1\,;\ f_2 \text{ и } g_2\,;\ \ldots\,;\ f_k \text{ и } g_k\,, \qquad (4)$$

где признаки f_i и g_i определяют *лексические группы* X_i и Y_i ($i = 1, \ldots, k$). Следовательно, необъятного размера синтаксис Sint задается небольшим набором (около двухсот пар) признаков (4) из $\Omega 1$.

Например, если D^1 — ИС, D^2 — ИП, а $f_1 = 012131$ и $g_1 = 012132$, $f_2 = 012131$ и $g_2 = 022131$, $f_3 = 012132$ и $g_3 = 022132$, то в графе $(A, L, \Psi 1, \Omega 1)$ будут связаны дугой (v, w) только те вершины v и w, которые принадлежат СО: $012131 * 012132$, $012131 * 022131, 012132 * 022132$.

В каждом из трех множеств содержатся сотни тысяч элементов. Элементами СО $012131 * 012132$ ($012131 * 022131, 012132 * 022132$) являются *несогласованные* (*согласованные*) определения.

Корневое дерево в графе Sint называется *выражением-деревом*. Среди лексических групп имеются

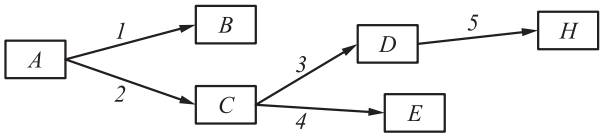


Рис. 1 Корневое дерево в Sint1

две группы: *группа сказуемых* GV и *группа подлежащих* GS. Если корень *выражения-дерева* принадлежит GV и связан дугой с вершиной из GS, то такое *выражение-дерево* называется *предложением-деревом*. Обход *выражения-дерева* называется *выражением-последовательностью* (или просто выражением), обход *предложения-дерева* называется *предложением-последовательностью* (просто *предложением*).

Если в грамматике Sint (Λ_μ):

(1) множество L — лексика русского языка, (3) — функции из морфологии русского языка;

(2) элементы $X_i * Y_i$ $(i = 1, \ldots, k)$ — словосочетания (пары связанных словоформ), допустимые в синтаксисе русского языка;

(3) связанные пары словоформ в выражениях русского языка образуют корневое дерево,

то Sint (Λ_μ) должен иметь много общего с синтаксисом русского языка; поэтому Sint (Λ_μ) будем называть *моделью* грамматики русского языка. Нетрудно показать существование такой грамматики Sint (Λ_μ), что предложения, встречающиеся в литературе на русском языке, будут предложениями в грамматике Sint (Λ_μ). (Но в грамматике Sint (Λ_μ) будут предложения «сомнительного» значения.)

Грамматика Sint будет использована для программного построения и распознавания *выражений* в синтаксисе Sint. Sint — открытая система, в ней могут появляться новые слова и синтаксически связанные пары слов.

В частности, известная задача *синтаксического анализа предложения* одинаково формулируется и решается как в ЕЯ, так и в грамматике Sint: *синтаксически правильно построенное* в грамматике Sint *предложение* будет *синтаксически правильно построенным предложением* и в грамматике русского языка. Последовательность словоформ образует *синтаксически правильно построенное* в грамматике Sint *предложение*, если в графе Sint найдется *предложение-дерево*, порожденное этим набором словоформ. А алгоритмы поиска корневого дерева, порожденного заданным множеством вершин, хорошо известны. Строгая формулировка задачи *синтаксического анализа предложения* и наличие алгоритма ее решения в грамматике ЕЯ уже дорогого стоит.

Важным понятием в грамматике Sint является СФ — *синтаксическая форма*. Показанное на рис. 1 дерево имеет скобочные формы представления: через вершины — $A(B, C(D(H), E))$ — и дуги — $0(1, 2(3(5), 4))$.

Вершинами дерева $A(B, C(D(H), E))$ являются *лексические группы*, поэтому оно порождает предложения-деревья $a(b, c(d(h), e))$, где строчной буквой обозначена словоформа из группы, обозначенной этой же буквой в верхнем регистре. Предложения из Sint, порожденные *корневым деревом* из Sint1, имеют одну и ту же синтаксическую форму, поэтому такие деревья называются СФ. Более того, СФ представляет собой правило, порождающее класс предложений определенной формы.

Предположим, что дуга i $(i = 1, \ldots, 5)$ дерева с рис. 1 (СФ1) представлена произведением пары признаков $f_i * g_i$. Синтаксические связи между вершинами дерева требуют, чтобы $f_3 = f_4 = g_2$, $f_5 = g_3$; любые шесть слов с указанными признаками могут оказаться вершинами предложения, порожденного СФ1.

Присваивая различным вершинам СФ1 различные словоформы, можно строить различные предложения. Обозначив через $g(t)$ словоформу $t : g$ (форму g слова t), можно выписать форму порождаемых СФ1 предложений:

$$f_1(s_1)\,(g_1(t_1), g_2(t_2)\,(g_3(t_3)(g_5(t_5)), g_4(t_4)))\,. \qquad (5)$$

Выражение (5) удобно представить в виде двух изоморфных деревьев:

$$f_1\,(g_1, g_2\,(g_3(g_5), g_4)) : s_1\,(t_1, t_2\,(t_3(t_5), t_4))\,. \qquad (6)$$

Нетрудно заметить, что СФ1 представлена выражением $f_1(g_1, g_2(g_3(g_5), g_4))$. Примером СФ служит выражение

$$01112131(02112131, 01112132(02112132))\,,$$

которое порождает выражения типа «**белый дом старого охотника**».

В форме (6) будут храниться выражения в *семантическом словаре*. Исследования текстов показывают, что число различных СФ, порождающих простые предложения на русском языке, не превышает сотни; три десятка СФ позволяют носителю русского языка вполне красноречиво выражать свои мысли; а у каждого автора текстов имеются характерные для него СФ, которыми он пользуется для выражения своих мыслей.

4 Семантика

При всей своей содержательности формальная грамматика без семантики не образует языка. *Семантика* строится на элементах синтаксиса; эле-

менты синтаксиса (слова и их сочетания) должны *выражать знания*. Знание, выраженное элементом синтаксиса, называется его *значением* или *семантикой*.

Знание есть специфическая форма *ощущения* сознанием активного состояния определенной области памяти человека; свидетелем существования *знания* является человек, *ощущающий* его; *знание* о слове (как последовательности букв) — назовем его *именем* слова — также хранится в памяти; к *знанию* о слове прикреплено иное *знание*, называемое его *значением*; человек способен воспроизводить слово; воспринятое человеком *слово* активизирует его *значение*; слово и его значение способны активизировать друг друга — в этом сущность ЕЯ. *Значение* и *имя* слова состоят в таких же отношениях, как информация в ячейке оперативной памяти и адрес этой ячейки, который хранится в другой ячейке.

Элементы синтаксиса и семантики — проявления в физиологии человека. Отношения между двумя явлениями: *ощущением-словом* и *ощущением-знанием*, видимо, имел в виду Ф. де Соссюр, говоря о языке как об отображении друг в друга «двух сущностей»: *элементов грамматики* («план выражения») и *элементов семантики* («план значения»).

Чтобы цифровая техника, способная оперировать элементами грамматики, стала имитатором языка, требуется в ней (в технике) найти нечто, представляющее собой знание. Например, функции «плана значений» могли бы сыграть «нейронные сети» (электронные схемы), если бы ОП состояла из них; но в современных компьютерах организация памяти такова, что слово на экране монитора и в памяти компьютера — сущности одной и той же природы.

Собеседники, пользуясь только элементами синтаксиса, обмениваются знаниями. Такое общение доступно и двум компьютерам, если они будут наделены одной и той же *моделью мира* (сетью знаний) и идентичными правилами трансформации знаний в предложения и наоборот.

Знание в *модели мира* может быть сохранено в памяти машины и в форме элементов синтаксиса. Остается придумать технологию хранения, подобную той, которая наблюдается в языковой способности человека, а не в статьях *толкового словаря*.

Далее излагается один из возможных вариантов построения *семантического словаря*, позволяющего хранить большие объемы *знаний* в форме, удобной для программной обработки.

В статье W словаря хранятся *элементарные знания* о понятии W в виде корневого дерева $T(W)$, в котором отображены *семантические отношения* между W и другими понятиями. При дереве $T(W)$ имеется СФ для преобразования $T(W)$ в *элементарное* синтаксически правильное выражение.

Синтаксическую форму, привязанную к дереву $T(W)$, обозначим через $\mathrm{SF}\,(T(W))$. Семантический словарь состоит из пар $\langle T, \mathrm{SF}\,(T)\rangle$, которые были описаны в (6) и являются предложениями синтаксиса Sint. Таким образом, элементы синтаксиса используются для представления знания.

На примере *статьи*, посвященной понятию *аист* (рис. 2), рассмотрим структуру самой статьи и оценим, какие возможности кроются в таком словаре для формирования языка.

По структуре статья состоит из двух частей: заголовка (1) и тела (2)–(11). В заголовке статьи за именем *понятия* следует его *полный код* — 040613, в котором 04 и 0406 — вложенные друг в друга коды двух предков (*животное* и *птица*) понятия *аист*. За косой чертой указаны имя родительского понятия (*птица*) и код части речи (01) понятия *аист*.

Тело статьи состоит из пяти разделов: *форма*, *свойство*, *элемент*, *событие*, *метод*. В разделе «*форма*» перечислены *семантические формы* слова *аист*; в разделе «*свойство*» — качества данного понятия (*цвет*, *вес*, *форма* и т. д.); в разделе «*элемент*» — составные части, образующие описываемое понятие.

аист040613/01птица((1)
форма: (01аист; 02аистовый, 02аистиный; 07); (2)
свойство: (3)
(СФ1: цвет(белый); (4)
СФ2: вес(до 7 кг); (5)
СФ4: местонахождение(деревня, поле); (6)
СФ3: местожительство(гнездо(СФ6: кровля, СФ: дерево)); (7)
элемент: (СФ5: клюв(СФ1: длинный)); (8)
событие: (СФ6: сидеть, летать); (9)
метод: (СФ7: клевать(трава)) (10)
) (11)

Рис. 2 Статья семантического словаря, описывающая понятие *аист*

В разделе «событие» перечислены действия и состояния, которым может подвергаться описываемое понятие. В разделе «*метод*» указаны действия, которые может совершать «*аист*» над другими понятиями.

В каждом разделе фиксированы знания, представленные в виде *семантического дерева* в скобочной форме; корнем для всех деревьев служит имя описываемого понятия (в теле статьи оно не дублируется). Деревьям предписаны СФ для преобразования их в выражения синтаксиса Sint.

Например, запись «СФ1: цвет(белый)» может обозначать

«01112131(01112132(02112132)):
аист(цвета(белого))»,

т. е. выражение «аист цвета белого».

Далее для построения сложного предложения из *элементарных* можно использовать богатый опыт *синтаксической* и *логической семантик*.

Литература

1. Грамматика русского языка / Под ред. В. В. Виноградова, Е. С. Истриной, С. Г. Бархударова. — В 2 т. — М.: Изд-во Академии наук СССР, 1960. 720 с.
2. *Шихиев Ф. Ш.* Формализация и сетевая формулировка задачи синтаксического анализа: Дис. . . . канд. физ.-мат. наук. — СПб.: СпбГУ, 2006. 171 с.
3. *Мирзабеков Я. М., Шихиев Ш. Б.* Формальная грамматика русского языка в примерах // Прикладная дискретная математика, 2018. № 40. С. 114–126.
4. *Слобин Д., Грин Дж.* Психолингвистика. — М.: Прогресс, 1976. 336 с.
5. *Тестелец Я. Г.* Введение в общий синтаксис. — М.: РГГУ, 2001. 798 с.
6. *Мельчук И. А.* Опыт теории лингвистических моделей Смысл–Текст. — М.: Языки русской культуры, 1999. 346 с.
7. Natural language parsing / Eds. D. Dowty, L. Karttunen, A. Zwicky. — Cambridge: Cambridge University Press, 1985. 413 p.

Поступила в редакцию 25.12.18

INCAPSULATION OF SEMANTIC REPRESENTATIONS INTO ELEMENTS OF A GRAMMAR

Sh. B. Shihiev and F. Sh. Shihiev

Department of Discrete Mathematics and Computer Science, Dagestan State University, 43-a Gadzhiyev Str., Makhachkala 367000, Republic of Dagestan, Russian Federation

Abstract: The article proposes a new mathematical apparatus of natural language representation for computer linguistics: morphology, syntax, and semantics are described as the objects of discrete mathematics forming a hierarchy and an integral information system. The proposed constructive language theory is a new approach to language learning by separating the domains of syntax and semantics, constructing the autonomous models of syntax and semantics, language formation as the mapping of elements of two sets: syntax and semantics.

Keywords: natural language; graph; syntax; semantics; lexicon; word form; morphological feature; lexical group; dictionary; sentence; algorithm

DOI: 10.14357/19922264200116

References

1. Vinogradov, V. V., E. S. Istrina, and S. G. Barkhudarova, eds. 1960. *Grammatika russkogo yazyka* [Russian language grammar]. Moscow: USSR Acad. Sci. Publs. 720 p.
2. Shihiev, F. Sh. 2006. Formalizatsiya i setevaya formulirovka zadachi sintaksicheskogo analiza [Formalization and network interpretation of a parsing task]. St. Petersburg: St. Petersburg State University. PhD Diss. 171 p.
3. Mirzabekov, Ya. M., and Sh. B. Shihiev. 2018. Formal'naya grammatika russkogo yazyka v primerakh [Formal grammar of Russian language in examples]. *Prikladnaya diskretnaya matematika* [Applied Discrete Mathematics] 40:114–126.
4. Slobin, D., and G. Green. 1976. *Psikholingvistika* [Psycholinguistics]. Moscow: Progress. 336 p.
5. Testelets, Ya. G. 2001. *Vvedenie v obshchiy sintaksis* [Introduction to general syntax]. Moscow: RGGU. 798 p.
6. Mel'chuk, I. A. 1999. *Opyt teorii lingvisticheskikh modeley Smysl–Tekst* [Experience in the theory of linguistic models Sense–Text]. Moscow: Yazyki russkoy kul'tury. 346 p.
7. Dowty, D., L. Karttunen, and A. Zwicky, eds. 1985. *Natural language parsing*. Cambridge: Cambridge University Press. 413 p.

Received December 25, 2018

Contributors

Shihiev Shukur B. (b. 1951) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Discrete Mathematics and Computer Science, Dagestan State University, 43-a Gadzhiyev Str., Makhachkala 367000, Republic of Dagestan, Russian Federation; sh_sh_b51@mail.ru

Shihiev Fuad B. (b. 1980) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Discrete Mathematics and Computer Science, Dagestan State University, 43-a Gadzhiyev Str., Makhachkala 367000, Republic of Dagestan, Russian Federation; fuad@mail.ru

INFORMATICS AND APPLICATIONS 2020. Vol. 14. Iss. 1. P. 128–135

INFORMATION FUSION OF DOCUMENTS

S. K. Dulin[1], N. G. Dulina[2], and P. V. Ermakov[3]

Abstract: The paper considers the problems associated with the creation of an expert base of documents that require prompt processing of incoming information and, as a consequence, restructuring of the knowledge base. The authors propose procedures that reduce the search of the optimal consistent state of interrelated documents. An approach to assessing the relationship of text documents and informational messages as poorly structured objects was developed. The practical implementation of this approach is described.

Keywords: information fusion; controlled data and knowledge consistency; knowledge base restructuring

DOI: 10.14357/19922264200117

1 Introduction

Combining information of various origins for integrative analysis and processing has been called "Information Fusion"[1], implying that the synthesized data carrying information combine type properties of source data and possess more information than merely conjunction of information sources considered separately. The main difficulty of the synthesis problem is that information sources contain heterogeneous data represented by various formats and structures and employed in different types of platforms.

The main factors of data heterogeneity and their sources are: various types of data, diversity in data origin, various models of database representation, various data presentation formats, differentiating in the organization of data storage systems, differences in the degree of reliability and accuracy of data, and variety of a degree and form of data structure.

The process of information fusion is a multilevel process that includes five basic stages [2–4]:

- zero stage — the stage of combining sensor signals, designed to obtain data indicating semantically clear and interpretable attributes of objects and participating in the applications of the research being performed;
- the first stage is aimed at processing data of the zero stage in order to make a decision on the classes of the objects in question and the states of these objects;
- the second stage of Information Fusion, designed to assess the situation, including the zero and the first stages. It is used to assess the situational interaction of objects considered as a whole;
- the third stage — the stage of evaluation of the interaction "Impact Assessment," designed to perform an antagonistic assessment, based on the prediction of the situation;
- the fourth stage — the stage of feedbacks, evaluating the possibility of using feedbacks in the system in question; and
- the fifth stage — the final stage, the level of man–machine interaction, performing correctional actions of the operator for the sake of the system control.

Research in the field of Information Fusion mainly focuses on the synthesis of data represented by digital images and arrays of data and documents [1, 2, 5].

Current trends in the development of corporative informational systems show that, along with traditional informational resources, the results of intelligent activity of experts and analysts become very important for the successful operation of large and middle-sized companies. A unified informational environment of the company incorporates these formalized results in an accumulated form such that all executives can jointly use this resource in the context of their assignments. The role played by the knowledge accumulated in such a way in the enterprise-wide systems allows us to consider this knowledge as very valuable and a notably important resource for a company, which, together with the traditional resources, such as financial, material, human, etc., characterizes the reliability of the company. The totality of this knowledge, presented mainly in text form, is the intelligent assets of the company, and the competitiveness of the company and its adaptability to changing the business environment depends on how efficiently this resource is used.

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation, skdulin@mail.ru

[2]A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation, ngdulina@mail.ru

[3]TeleRetail GmbH, 30 Markenstraße, Düsseldorf 40227, Germany; petcazay@gmail.com

An intelligent asset is a specific resource that requires specialized knowledge management systems. These systems enable the search, accumulation, and processing of knowledge by experts in solving various analytical problems. This tendency in knowledge engineering appeared relatively recently, but interest in the development and usage of such systems is permanently growing. This is largely due to the significant results achieved by some companies that have successfully implemented knowledge management systems into their manufacturing activity.

Complex technological solutions designed to support various stages of composition and usage of corporative data and knowledge have been embodied in the knowledge management systems. At each of these stages, individual problems are solved, with the most important of them being associated with tasks related to searching, processing documents, and extracting knowledge from them.

Text processing tasks are solved in practically all fields of human activity, and the analysis of the current environment is an integral part of practically each corporative management system securing a timely and adequate reaction to changes in the business environment. Actually, operativeness is the basic characteristics of monitoring problems, which distinguishes them from the problems related to prediction, planning, etc., because the main goal of the monitoring is the timely reaction of corresponding management subsystems of the general technological scheme of company functioning to changes of internal or external factors.

In the general case, the purpose of text processing tasks is to accumulate necessary information from different sources, process it analytically, and, on this basis, generate corresponding decisions. The character of text processing tasks is permanent in the sense that the environment and the parameters of the company operation are subject to permanent changes, which requires regular (or periodic) sampling of ever changing information.

Text processing tasks can conventionally be divided into two classes: internal monitoring and external monitoring.

Internal monitoring is associated mainly with the monitoring of internal operation parameters, e. g., regular monitoring of the operation of complex installations, cargo moving, etc. Possible examples are control systems for energy plants, freight management, etc. The typical feature of these problems is a relatively constant set of parameters used to estimate the state of the process (production, physical parameters of an installation, etc.).

In contrast to the internal monitoring, the external monitoring is mainly related to the estimation of the state of the environment and external conditions of the company operation. As an example, an analysis of consumer demand carried out by a commodity-producing company falls into this category. The typical feature of these problems is that, first, the parameters to be estimated are poorly formalized and, second, the set of these parameters is variable. The latter factor requires the restructuring of the analyst knowledge according to the changed conditions. All this makes us consider the "restructurability" of the expert knowledge base as one of the characteristic features of the problems of external monitoring.

In the problems of external monitoring, special requirements must be imposed on the sources of information used by experts for the localization of required knowledge and data. The development of informational technologies during recent years has strongly suggested that the Internet is gradually becoming the most important source of information in solving analytical problems in practically all areas of human activity. Coming up to printed and electronic mass media, Internet is often ranked first in operativeness, which makes the Internet the most valuable information source in monitoring problems. It is for this reason that, in this work, special attention is paid to the solution of monitoring problems associated with search and processing of text information in Internet.

2 Approach to Provision of Knowledge Consistency

In previous works (see [2, 6, 7]), the authors put forward a procedure providing the consistency of the knowledge base dynamically formed by an expert, which is based on the analysis of structural interrelations between separate components of the knowledge base with subsequent restructuring of it aimed at reducing existing inconsistency. In so doing, the basic criterion of structural consistency was a concept of polyconsonance of power n [4].

Consider a knowledge base formed on the basis of search and analysis of Internet information. In solving the monitoring problems associated with the formation of such a knowledge base, the application of this procedure faces certain difficulties resulting from poor formalization and an obscure or ambiguous structure of the data (text or multimedia documents). Besides, for the monitoring problems considered here, a large number of informational messages directed to the expert for analytical processing and replenishment of the knowledge base are characteristic. As a result, the amount of resources (especially, time) required for the restructuring of a dynamically changing knowledge base is increased significantly, which is, perhaps, the main obstacle to the successful practical implementation of any procedure of the above type.

One of the major disadvantages of the algorithm proposed in [2] is that it is oriented to problems of the search type; that is why, the authors made special efforts to reduce the search and thus increase the algorithm efficiency in its practical implementation. The results presented below are aimed at the solution of the latter problem.

Consider a set of mutually related objects $O = \{o_i\}$ with a similarity function f [3] satisfying the condition

$$0 \le f(o_i, o_j) \le 1 .$$

Numbers α and β will denote the lower and upper similarity thresholds, respectively, satisfying the condition

$$0 \le \alpha \le \beta \le 1 .$$

Now, let us introduce the concepts of a negative, positive, and indifferent link between two arbitrary elements o_i and o_j of the set O. The link is called "negative" if its value does not exceed the lower similarity threshold: $0 \le f(o_i, o_j) \le \alpha$; it is called "positive" if the value of the similarity function is not less than the upper similarity threshold: $\beta \le f(o_i, o_j) \le 1$; and, if $\alpha < f < \beta$, it is called "indifferent" (zero).

Consider a partition of the given set into a number of nonempty subsets $K_1, \ldots, K_n$.

A link between two arbitrary elements o_i and o_j of the entire set O is called "bad" if one of the following conditions is satisfied:

(1) the elements o_i and o_j belong to the same subset K_x, and the link between them is negative; or

(2) the elements o_i and o_j belong to different subsets K_1 and K_2, and the link between them is positive.

Using this definition, let us to each object o_k from the set considered assign the number v_k of its bad links for a given partition into subsets. Now, let us construct a vector V consisting of these values (this vector has a dimension equal to the number of objects in the set) and call it the nodewise difference vector (NDV) [2]. The sum of the elements of this vector is denoted by S_{NDV}.

Clearly, different partitions of the original set correspond to different NDVs and different values of S_{NDV}. According to the algorithm considered, the main problem is to find a partition of the given set O such that the sum S_{NDV} takes its minimal value; i. e., the total number of bad links tends to zero.

The algorithm [2] developed by the authors consists in successive transformations of the set of informational objects on the basis of the condition

$$S_{\mathrm{NDV}} > \frac{n(N-n)}{2}$$

where S_{NDV} is the sum of nodewise differences for the given set of n elements belonging to a pair of consonant subsets of the total cardinality N. If this condition is fulfilled, then the restructuring of the considered set results in a decrease of the total sum S_{NDV}.

Theorem 1. *Let K_1 and K_2 be two subsets of a given set of mutually related objects O:*

$$K_1 = \{o_i\} \ , \ i = 1, \ldots, n_1 ;$$
$$K_2 = \{o_j\} \ , \ j = 1, \ldots, n_2 .$$

A set containing m elements from these two subsets satisfies the condition of the algorithm if, and only if, the set consisting of all remaining elements of these two subsets satisfies the same condition.

Proof. First, let us prove the necessity. Let the set of objects $\{o_k\}$, $k = 1, \ldots, m$, satisfy the condition of the algorithm:

$$\sum v_k > \frac{m(n_1 + n_2 - m)}{2}$$

where v_k are the NDV values for the element with the number k. This formula can be transformed to the form:

$$\sum v_k > \frac{(n_1 + n_2 - m)\left((n_1 + n_2) - (n_1 + n_2 - m)\right)}{2}$$

which means that the set of $n_1 + n_2 - m$ vectors not belonging to the original set also satisfies the condition of the algorithm.

The sufficiency of the condition is proved similarly. The theorem is proved.

Corollary. In order to find a set of objects from two given subsets that satisfies the condition of the algorithm, it is sufficient to check the fulfillment of this condition only for the subsets consisting of $(n_1+n_2)/2$ objects. In other words, only subsets with cardinalities not exceeding half of the sum of the cardinalities of the original subsets K_1 and K_2 should be checked.

Proof. Indeed, if some set consisting of more than $(n_1 + n_2)/2$ elements satisfies the condition, then the complement to it also satisfies this condition, with the cardinality of the complement being not greater than $(n_1 + n_2)/2$.

Theorem 2. *Let K_1 and K_2 be two subsets of a given set of mutually related objects O:*

$$K_1 = \{o_i\} \ , \ i = 1, \ldots, n_1 ;$$
$$K_2 = \{o_j\} \ , \ i = 1, \ldots, n_2 .$$

Let a set $\{o_k\}$ of $m < (n_1 + n_2)/2$ elements belonging to these two subsets satisfy the condition of the algorithm. If a zero NDV element corresponds to some element o_x

from this set, then the set of the vectors corresponding $O^* = \{o_1, \ldots, o_{x-1}, o_{x+1}, \ldots, o_m\}$ *also satisfies the condition of the algorithm.*

Proof. According to the assumption of the theorem, the sum S^*_{NDV} for the set $O^* = \{o_1, \ldots, o_{x-1}, o_{x+1}, \ldots, o_m\}$ is equal to the sum S_{NDV} of the original set of the elements from the two subsets K_1 and K_2:

$$S^*_{\text{NDV}} = S_{\text{NDV}}\,.$$

Denote by N the total cardinality of the considered subsets: $N = n_1 + n_2$. Then,

$$(m-1)(N-(m-1)) = m(N-m) + (2m - N - 1)\,.$$

According to the assumption of the theorem, $m \le N/2$; hence, $2m - N - 1 < 0$. To complete the proof, let us write the following inequality:

$$S^*_{\text{NDV}} = S_{\text{NDV}} = \sum v_k > \frac{m(N-m)}{2} > \frac{(m-1)(N-(m-1))}{2}$$

which means that the set $\{o_1, \ldots, o_{x-1}, o_{x+1}, \ldots, o_m\}$ satisfies the condition of the algorithm.

Obviously enough, it follows from this theorem that, in the practical implementation of the proposed algorithm, it is sufficient to search for a set of elements for the next iteration among those with nonzero NDV values.

3 Thematic Role of Similarity

The most significant factor affecting the operation of the algorithm considered is the similarity function on the basis of which interrelations between different elements of a given set are determined. As far as the support of monitoring problems is considered, with the texts (in particular, news) and the Internet being the elements and the main information source, respectively, the construction of the similarity function becomes a fairly difficult problem. Perhaps, one of the solutions to this problem could be the use of various methods of linguistic analysis to determine the degree of "likeness" of two different documents, although these methods are not free from some shortcomings associated with the hardship of their implementation, adjustment, etc. To determine the similarity function in practical applications, the authors have put forward another approach. One of the advantages of this new approach is the simplicity of implementation and the "notional transparency."

The basis of this approach schematically shown in Fig. 1 is the determination of vocabulary groups [6], which denote the sets of keywords defined by the expert. The expert assorts the keywords according to some criterion, e. g., "thematic meaning:"

$$G_k = \{w_i\}\,, \quad i = 1, \ldots, n_k.$$

Consider an arbitrary element o_j from a given set O. This object is a text document; so, it can be represented as an aggregate of lexical units, i. e., words. For o_j, let us define its coefficient of correspondence with the dictionary group G_i as the ratio $S(G_i)_j$ of the number of keywords specified in this dictionary group and available

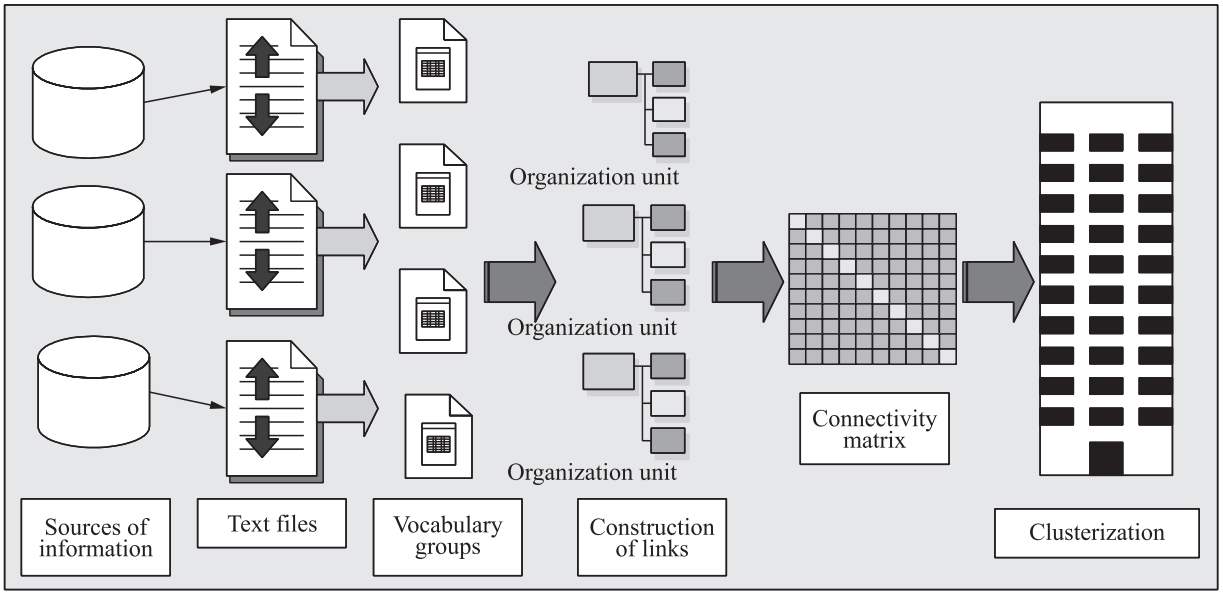


Figure 1 Determination of vocabulary groups

in the text of the information object itself, to the total number of keywords from all dictionary groups, $S(G)_j$ found in this text. Then, one can define the factor of correspondence of the object o_j to the vocabulary group G_i as

$$L_j^i = \frac{S(G_i)_j}{S(G)_j}.$$

On the basis of these coefficients, let us define the degree of thematic coupling between two arbitrary informational objects as follows:

(A) $f(o_k, o_l) = 1$ if $S(G)_k = 0$ and $S(G)_l = 0$;

(B) $f(o_k, o_l) = 0$ if $S(G)_k \neq S(G)_l$ and $S(G)_k S(G)_l = 0$; and

(C) $f(o_k, o_l) = \max\left(\min\left(L_k^i, L_l^i\right)\right)$, $i = 1, \ldots, n$, for $S(G)_k S(G)_l \neq 0$ where n is the number of the vocabulary groups.

Note that the similarity function defined above takes the values on the interval from 0 to 1 but lacks associativity, because $0 \leq f(o_i, o_j) \leq 1$. In the works devoted to the theoretical grounds of the considered algorithm of structural transformations of a set of objects, the associativity of the similarity function has not been used; therefore, the fact that the function introduced above is not associative does not require any changes in the proposed algorithm. Moreover, the lack of associativity here has an additional meaning, which makes it possible to treat the function introduced above as a *thematic* similarity function.

Indeed, if, in the considered text, there are keywords from different vocabulary groups, then all the coefficients L_j^i for this element will be less than one. Hence, the value of the similarity function f will also be less than one, and the more the number of the vocabulary groups, the less this value. In practice, this could mean that the considered document is of a review nature and, most probably, has no distinct "thematic meaning."

4 Consistency Controlling Module iiProcessor

The authors' technique for providing structural consistency of the knowledge base in solving monitoring problems has been implemented in a specialized system called an iiProcessor. This system is designed to compose expert knowledge bases for social, political, and international sciences. The knowledge bases are constructed from the information supplied by various mass media through their Internet servers. The main purpose of the system is to accumulate informational messages (news) related to the themes of user's interest from various Internet sources, to integrate the information into a unified knowledge base, to create links between different elements of the knowledge base, and to make subsequent restructuring of the knowledge base on the basis of these links, with the result of this restructuring being the representation of the body of the information accumulated as a logical system of classes. The latter system can be treated as an informational model of the problem examined by the expert (for example, the social and political situation in a particular region of the world). A general scheme of operation of the system is shown in Fig. 2.

As a source of information, this system uses the CNN Internet site (http://cnn.com). Several times a day, this

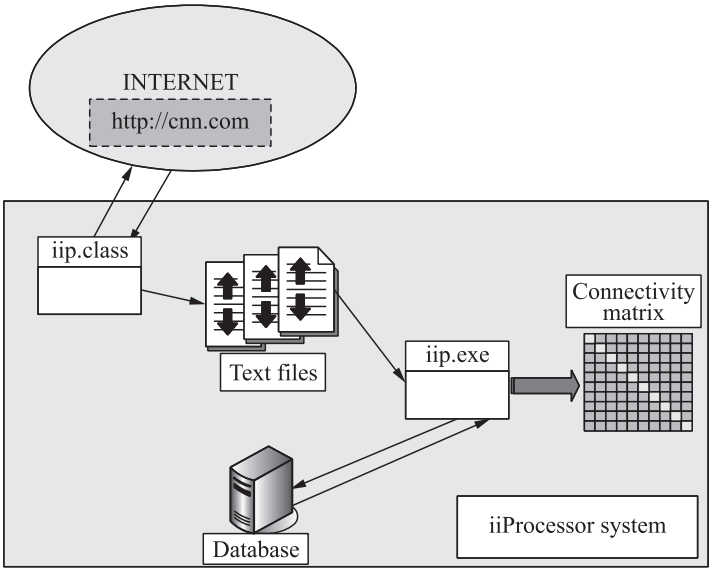


Figure 2 A general scheme of operation of iiProcessor system

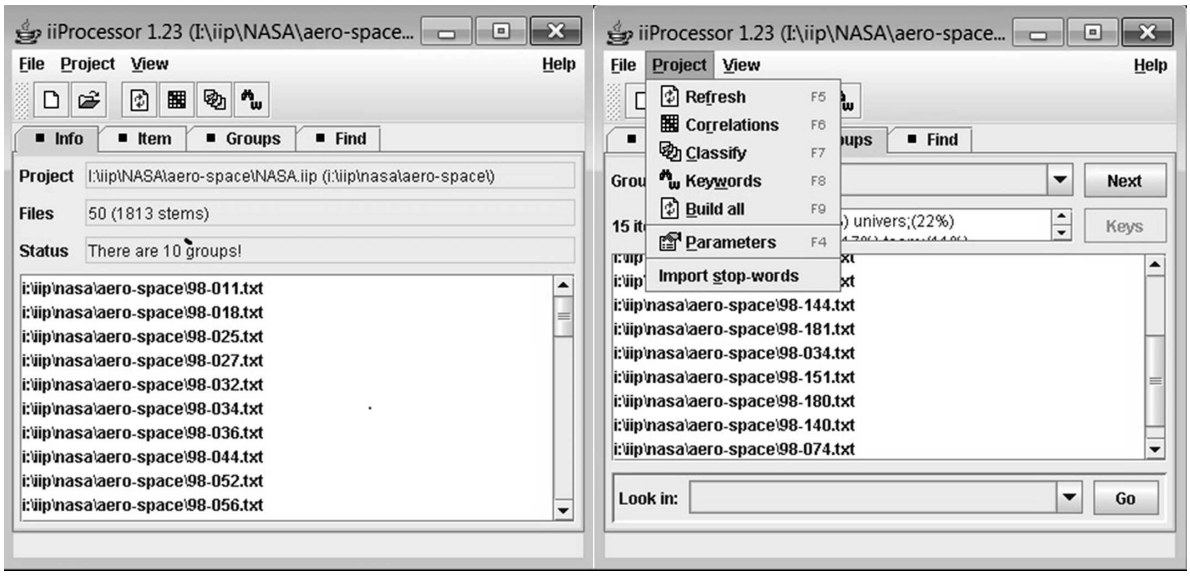


Figure 3 Example of use of vocabulary group technique to establish links between different documents

site publishes information covering many aspects of social and political life in many countries. In most cases, the informational messages are weakly-structured text documents. In order to establish links between different documents, the vocabulary group technique described above is used (Fig. 3). If various informational messages contain common keywords belonging to different vocabulary groups, this technique estimates the "likeness" of the messages. The similarity function classifies these links as positive or negative, which makes it possible to construct a connectivity matrix on the set of the informational messages received by the user (see Fig. 3).

The mode of "Keywords" allows one to get 10 of the most significant key words for a given document with an indication of their weighting factors (Fig. 4).

The mode of interrelations ("Correlations") will allow to get several documents that have the greatest interrelations with selected document. This mode works only if the loaded document belongs to the current project of the iiProcessor system, in which the relationship was evaluated (Fig. 5).

The choice of the CNN server as a source of information is explained by the fact that this server is one of the most informationally abundant servers providing real-time information. Of course, the choice of the sources of information is strongly determined by the character of the problem considered. In this sense, the CNN server is not universal. In view of the above considerations, the Restructor system is implemented as a complex of two program modules. The rsn.exe module is the basic one. An auxiliary iip.class module executes a real-time search for new information in a specified information source in the Internet. With such an architecture, this

Figure 4 "Keywords" mode

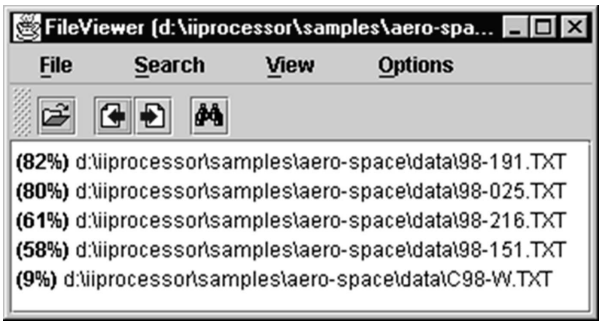


Figure 5 "Correlation" mode

system can be adopted to operation with any informational servers in the Internet (and beyond) by replacing only the auxiliary module, without changing its kernel

where the major mathematical results of the authors' approach are implemented.

5 Concluding Remarks

The implementation of the results of Theorems 1 and 2 in the inference engine made it possible to considerably reduce the time expenses of the built-in algorithm for restructuring the database. The use of the connectivity matrix as the major visualization means for the informational objects improved the clearness of the representation of the information model of the problem considered by an expert. The system has been tested in analyzing the events related to NASA's aerospace research.

References

1. Dasarathy, B. 2001. Information fusion — what, where, why, when, and how? *Inform. Fusion* 2(2):75–76.
2. Dulin, S. K. 1995. The approach to structural consistency of situations' models in an active knowledge base. *Workshop of 10th IEEE Symposium (International) on Intelligent Control Proceedings*. Monterey, CA: AdRem, Inc. 253–258.
3. Duckham, M., and M. Worboys. 2007. Automated geographic information fusion and ontology alignment. *Spatial data on the Web*. Eds. A. Belussi, B. Catania, E. Clementini, and E. Ferrari. Berlin: Springer. Ch. 6:109–132.
4. Pravia, M. 2008. Generation of a fundamental data set for hard/soft information fusion. *11th Conference (International) on Information Fusion Proceedings*. Cologne: International Society of Information Fusion. 134–145.
5. Landauer, T. K., K. Kireyev, and C. Panaccione. 2011. Word maturity: A new metric for word knowledge. *Sci. Stud. Read.* 15(1):92–108.
6. Dulina, N., and O. Kozhunova. 2010. Information monitoring system: A problem of linguistic resources consistency and verification. *Problems of Cybernetics and Informatics: 3rd Conference (International) Proceedings*. Baku. 56–58.
7. Dulin, S. K., and N. G. Dulina. 2018. Ispol'zovanie diseminatsionnykh algoritmov dlya formirovaniya nestrukturirovannoy tekstovoy informatsii v baze geodannykh [Using dissemination algorithms for the formation of unstructured textual information in the geodatabase]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(2):42–59.

Received February 26, 2019

Contributors

Dulin Sergey K. (b. 1950) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; principal scientist, Research & Design Institute for Information Technology, Signalling and Telecommunications on Railway Transport (JSC NIIAS), 27-1 Nizhegorodskaya Str., Moscow 109029, Russian Federation; skdulin@mail.ru

Dulina Natalia G. (b. 1947) — Candidate of Science (PhD) in technology, leading programmer, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; ngdulina@mail.ru

Ermakov Petr V. (b. 1985) — Senior Software Developer, TeleRetail GmbH, 30 Markenstraße, Düsseldorf 40227, Germany; petcazay@gmail.com

ИНФОРМАЦИОННЫЙ СИНТЕЗ ДОКУМЕНТОВ

С. К. Дулин[1], Н. Г. Дулина[2], П. В. Ермаков[3]

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, skdulin@mail.ru
[2]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, ngdulina@mail.ru
[3]TeleRetail GmbH, Düsseldorf, Germany

Аннотация: Рассматриваются проблемы, связанные с созданием экспертной базы документов, требующей оперативной обработки поступающей информации и, как следствие, реструктуризации базы знаний. Предложены процедуры, уменьшающие время поиска оптимального согласованного состояния взаимосвязанных документов. Был разработан подход к оценке взаимосвязи текстовых документов и информационных сообщений как плохо структурированных объектов. Описана практическая реализация этого подхода.

Ключевые слова: информационный синтез; контролируемая согласованность данных и знаний; реструктуризация базы знаний

DOI: 10.14357/19922264200117

Литература

1. *Dasarathy B.* Information fusion — what, where, why, when, and how? // Inform. Fusion, 2001. Vol. 2. Iss. 2. P. 75–76.
2. *Dulin S. K.* The approach to structural consistency of situations' models in an active knowledge base // Workshop of 10th IEEE Symposium (International) on Intelligent Control Proceedings. — Monterey, CA, USA: AdRem, Inc., 1995. P. 253–258.
3. *Duckham M., Worboys M.* Automated geographic information fusion and ontology alignment // Spatial data on the Web / Eds. A. Belussi, B. Catania, E. Clementini, E. Ferrari. — Berlin: Springer, 2007. Ch. 6. P. 109–132.
4. *Pravia M.* Generation of a fundamental data set for hard/soft information fusion // 11th Conference (International) on Information Fusion. — Cologne: International Society of Information Fusion, 2008. P. 134–145.
5. *Landauer T. K., Kireyev K., Panaccione C.* Word maturity: A new metric for word knowledge // Sci. Stud. Read., 2011. Vol. 15. Iss. 1. P. 92–108.
6. *Dulina N., Kozhunova O.* Information monitoring system: A problem of linguistic resources consistency and verification // Problems of Cybernetics and Informatics: 3rd Conference (International) Proceedings. — Baku, 2010. P. 56–58.
7. *Дулин С. К., Дулина Н. Г.* Использование диссеминационных алгоритмов для формирования неструктурированной текстовой информации в базе геоданных // Системы и средства информатики, 2018. Т. 28. № 2. С. 42–59.

Поступила в редакцию 26.02.2019

ОБ АВТОРАХ

Борисов Андрей Владимирович (р. 1965) — доктор физико-математических наук, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Босов Алексей Вячеславович (р. 1969) — доктор технических наук, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Брюхов Дмитрий Олегович (р. 1971) — кандидат технических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Вохминцев Александр Владиславович (р. 1978) — кандидат технических наук, доцент, заведующий научно-исследовательской лабораторией Челябинского государственного университета; доцент Югорского государственного университета

Голембиовский Дмитрий Юрьевич (р. 1960) — доктор технических наук, профессор, профессор кафедры исследования операций факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; профессор кафедры банковского дела университета «Синергия»

Гончаров Алексей Владимирович (р. 1995) — аспирант Московского физико-технического института

Горшенин Андрей Константинович (р. 1986) — кандидат физико-математических наук, доцент, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; ведущий научный сотрудник факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Грушо Александр Александрович (р. 1946) — доктор физико-математических наук, профессор, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Данилишин Артём Ростиславович (р. 1992) — аспирант кафедры исследования операций факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Дулин Сергей Константинович (р. 1950) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; главный научный сотрудник Научно-исследовательского и проектно-конструкторского института информатизации, автоматизации и связи на железнодорожном транспорте (ОАО «НИИАС»)

Дулина Наталья Георгиевна (р. 1947) — кандидат технических наук, ведущий программист Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ермаков Петр Вячеславович (р. 1985) — ведущий программист TeleRetail GmbH, Düsseldorf, Germany

Забежайло Михаил Иванович (р. 1956) — доктор физико-математических наук, доцент, главный научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ковалёв Дмитрий Юрьевич (р. 1988) — младший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Козеренко Елена Борисовна (р. 1959) — кандидат филологических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Костина Анна Александровна (р. 1983) — научный сотрудник лаборатории кибербезопасности и постквантовых криптосистем Санкт-Петербургского института информатики и автоматизации Российской академии наук

Кузнецов Константин Игоревич (р. 1968) — ведущий инженер Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Кузьмин Виктор Юрьевич (р. 1986) — руководитель Департамента разработки ООО «Вей2Гео»

Мельников Андрей Витальевич (р. 1956) — доктор технических наук, профессор Югорского государственного университета

Мирин Анатолий Юрьевич (р. 1979) — кандидат технических наук, старший научный сотрудник лаборатории кибербезопасности и постквантовых криптосистем Санкт-Петербургского института информатики и автоматизации Российской академии наук

Михеев Михаил Юрьевич (р. 1957) — доктор филологических наук, ведущий научный сотрудник Научно-исследовательского вычислительного центра МГУ им. М. В. Ломоносова

Молдовян Дмитрий Николаевич (р. 1986) — кандидат технических наук, научный сотрудник лаборатории кибербезопасности и постквантовых криптосистем Санкт-Петербургского института информатики и автоматизации Российской академии наук

Пачганов Степан Александрович (р. 1994) — аспирант Югорского государственного университета

Попов Георгий Александрович (р. 1950) — доктор технических наук, профессор, заведующий кафедрой Астраханского государственного технического университета

Севастьянов Леонид Антонович (р. 1949) — доктор физико-математических наук, профессор кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов

Серебрянский Сергей Михайлович (р. 1983) — старший преподаватель Троицкого филиала Челябинского государственного университета

Симаворян Симон Жоржевич (р. 1958) — кандидат технических наук, доцент Сочинского государственного университета

Симонян Арсен Рафикович (р. 1960) — кандидат физико-математических наук, доцент Сочинского государственного университета

Сомин Николай Владимирович (р. 1947) — кандидат физико-математических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Стефанович Алексей Игоревич (р. 1983) — главный специалист Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Стрижов Вадим Викторович (р. 1967) — доктор физико-математических наук, ведущий научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук; профессор Московского физико-технического института

Ступников Сергей Александрович (р. 1978) — кандидат технических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Тимонина Елена Евгеньевна (р. 1952) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Тырсин Александр Николаевич (р. 1961) — доктор технических наук, ведущий научный сотрудник Научно-инженерного центра «Надежность и ресурс больших систем и машин» Уральского отделения Российской академии наук

Улитина Елена Ивановна (р. 1978) — кандидат физико-математических наук, доцент Сочинского государственного университета

Фахрутдинов Роман Шафкатович (р. 1972) — кандидат технических наук, заведующий лабораторией кибербезопасности и постквантовых криптосистем Санкт-Петербургского института информатики и автоматизации Российской академии наук

Хусаинов Ахмет Аксанович (р. 1951) — доктор физико-математических наук, профессор Комсомольского-на-Амуре государственного университета

Шанин Иван Андреевич (р. 1991) — младший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Шестаков Олег Владимирович (р. 1976) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Шихиев Фуад Шукурович (р. 1980) — кандидат физико-математических наук, доцент Дагестанского государственного университета

Шихиев Шукур Бабаевич (р. 1951) — кандидат физико-математических наук, доцент Дагестанского государственного университета

Щетинин Евгений Юрьевич (р. 1962) — доктор физико-математических наук, профессор Департамента анализа данных, принятия решений и финансовых технологий Финансового университета при Правительстве РФ

Эрлих Лев Исаакович (р. 1948) — ведущий инженер Научно-исследовательского вычислительного центра МГУ им. М. В. Ломоносова

Правила подготовки рукописей для публикации в журнале «Информатика и её применения»

Журнал «Информатика и её применения» публикует теоретические, обзорные и дискуссионные статьи, посвященные научным исследованиям и разработкам в области информатики и ее приложений.

Журнал издается на русском языке. По специальному решению редколлегии отдельные статьи могут печататься на английском языке.

Тематика журнала охватывает следующие направления:

- теоретические основы информатики;
- математические методы исследования сложных систем и процессов;
- информационные системы и сети;
- информационные технологии;
- архитектура и программное обеспечение вычислительных комплексов и сетей.

1. В журнале печатаются статьи, содержащие результаты, ранее не опубликованные и не предназначенные к одновременной публикации в других изданиях.

   Публикация предоставленной автором(ами) рукописи не должна нарушать положений глав 69, 70 раздела VII части IV Гражданского кодекса, которые определяют права на результаты интеллектуальной деятельности и средства индивидуализации, в том числе авторские права, в РФ.

   Ответственность за нарушение авторских прав, в случае предъявления претензий к редакции журнала, несут авторы статей.

   Направляя рукопись в редакцию, авторы сохраняют свои права на данную рукопись и при этом передают учредителям и редколлегии журнала неисключительные права на издание статьи на русском языке (или на языке статьи, если он отличен от русского) и на перевод ее на английский язык, а также на ее распространение в России и за рубежом. Каждый автор должен представить в редакцию подписанный с его стороны «Лицензионный договор о передаче неисключительных прав на использование произведения», текст которого размещен по адресу http://www.ipiran.ru/publications/licence.doc. Этот договор может быть представлен в бумажном (в 2-х экз.) или в электронном виде (отсканированная копия заполненного и подписанного документа).

   Редколлегия вправе запросить у авторов экспертное заключение о возможности публикации представленной статьи в открытой печати.

2. К статье прилагаются данные автора (авторов) (см. п. 8). При наличии нескольких авторов указывается фамилия автора, ответственного за переписку с редакцией.

3. Редакция журнала осуществляет экспертизу присланных статей в соответствии с принятой в журнале процедурой рецензирования.

   Возвращение рукописи на доработку не означает ее принятия к печати.

   Доработанный вариант с ответом на замечания рецензента необходимо прислать в редакцию.

4. Решение редколлегии о публикации статьи или ее отклонении сообщается авторам.

   Редколлегия может также направить авторам текст рецензии на их статью. Дискуссия по поводу отклоненных статей не ведется.

5. Редактура статей высылается авторам для просмотра. Замечания к редактуре должны быть присланы авторами в кратчайшие сроки.

6. Рукопись предоставляется в электронном виде в форматах MS WORD (.doc или .docx) или LaTeX (.tex), дополнительно — в формате .pdf, на дискете, лазерном диске или электронной почтой. Предоставление бумажной рукописи необязательно.

7. При подготовке рукописи в MS Word рекомендуется использовать следующие настройки.

   Параметры страницы: формат — A4; ориентация — книжная; поля (см): внутри — 2,5, снаружи — 1,5, сверху — 2, снизу — 2, от края до нижнего колонтитула — 1,3.

   Основной текст: стиль — «Обычный», шрифт — Times New Roman, размер — 14 пунктов, абзацный отступ — 0,5 см, 1,5 интервала, выравнивание — по ширине.

Рекомендуемый объем рукописи — не свыше 10 страниц указанного формата. При превышении указанного объема редколлегия вправе потребовать от автора сокращения объема рукописи.

Сокращения слов, помимо стандартных, не допускаются. Допускается минимальное количество аббревиатур.

Все страницы рукописи нумеруются.

Шаблоны примеров оформления представлены в Интернете: http://www.ipiran.ru/journal/template.doc

8. Статья должна содержать следующую информацию на ***русском и английском языках***:
   - название статьи;
   - Ф.И.О. авторов, на английском можно только имя и фамилию;
   - место работы, с указанием почтового адреса организации и электронного адреса каждого автора;
   - сведения об авторах, в соответствии с форматом, образцы которого представлены на страницах: http://www.ipiran.ru/journal/issues/2013_07_01_rus/authors.asp и http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;
   - аннотация (не менее 100 слов на каждом из языков). Аннотация — это краткое резюме работы, которое может публиковаться отдельно. Она является основным источником информации в информационных системах и базах данных. Английская аннотация должна быть оригинальной, может не быть дословным переводом русского текста и должна быть написана хорошим английским языком. В аннотации не должно быть ссылок на литературу и, по возможности, формул;
   - ключевые слова — желательно из принятых в мировой научно-технической литературе тематических тезаурусов. Предложения не могут быть ключевыми словами;
   - источники финансирования работы (ссылки на гранты, проекты, поддерживающие организации и т. п.).

9. Требования к спискам литературы.

   Ссылки на литературу в тексте статьи нумеруются (в квадратных скобках) и располагаются в каждом из списков литературы в порядке первых упоминаний.

   Списки литературы представляются в двух вариантах:

   (1) **Список литературы к русскоязычной части**. Русские и английские работы — на языке и в алфавите оригинала;

   (2) **References**. Русские работы и работы на других языках — в латинской транслитерации с переводом на английский язык; английские работы и работы на других языках — на языке оригинала.

   Необходимо для составления списка “References” пользоваться размещенной на сайте http://www.translit.net/ru/bgn/ бесплатной программой транслитерации русского текста в латиницу.

   Список литературы “References” приводится полностью отдельным блоком, повторяя все позиции из списка литературы к русскоязычной части, независимо от того, имеются или нет в нем иностранные источники. Если в списке литературы к русскоязычной части есть ссылки на иностранные публикации, набранные латиницей, они полностью повторяются в списке “References”.

   Ниже приведены примеры ссылок на различные виды публикаций в списке “References”.

   Описание статьи из журнала:

   Zagurenko, A. G., V. A. Korotovskikh, A. A. Kolesnikov, A. V. Timonov, and D. V. Kardymon. 2008. Tekhniko-ekonomicheskaya optimizatsiya dizayna gidrorazryva plasta [Technical and economic optimization of the design of hydraulic fracturing]. *Neftyanoe hozyaystvo* [*Oil Industry*] 11:54–57.

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Rus. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Описание статьи из электронного журнала:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Описание статьи из продолжающегося издания (сборника трудов):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental’noe issledovanie prochnosti soedineniy “stal’–kompozit” [Experimental study of the strength of joints “steel–composite”]. *Trudy MGTU “Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem”* [*Bauman MSTU “Mathematical Modeling of Complex Technical Systems” Proceedings*]. 593:125–130.

Описание материалов конференций:

Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

Описание книги (монографии, сборники):

Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

Latyshev, V. N. 2009. *Tribologiya rezaniya. Kn. 1: Friktsionnye protsessy pri rezanii metallov* [*Tribology of cutting. Vol. 1: Frictional processes in metal cutting*]. Ivanovo: Ivanovskii State Univ. 108 p.

Описание переводной книги (в списке литературы к русскоязычной части необходимо указать: / Пер. с англ. — после названия книги, а в конце ссылки указать оригинал книги в круглых скобках):

1. В русскоязычной части:
   Тимошенко С. П., Янг Д. Х., Уивер У. Колебания в инженерном деле / Пер. с англ. — М.: Машиностроение, 1985. 472 с. (*Timoshenko S. P., Young D. H., Weaver W.* Vibration problems in engineering. — 4th ed. — N.Y.: Wiley, 1974. 521 p.)
2. В англоязычной части:
   Timoshenko, S. P., D. H. Young, and W. Weaver. 1974. *Vibration problems in engineering*. 4th ed. N.Y.: Wiley. 521 p.

Описание неопубликованного документа:

Latypov, A. R., M. M. Khasanov, and V. A. Baikov. 2004. Geology and production (NGT GiD). Certificate on official registration of the computer program No. 2004611198. (In Russian, unpubl.)

Описание интернет-ресурса:

Pravila tsitirovaniya istochnikov [Rules for the citing of sources]. Available at: http://www.scribd.com/doc/1034528/ (accessed February 7, 2011).

Описание диссертации или автореферата диссертации:

Semenov, V. I. 2003. Matematicheskoe modelirovanie plazmy v sisteme kompaktnyy tor [Mathematical modeling of the plasma in the compact torus]. D.Sc. Diss. Moscow. 272 p.

Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. PhD Thesis. Moscow: IPI RAN. 23 p.

Описание ГОСТа:

GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. Moscow: Standardinform Publs. 10 p.

Описание патента:

Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

10. Присланные в редакцию материалы авторам не возвращаются.
11. При отправке файлов по электронной почте просим придерживаться следующих правил:
    - указывать в поле subject (тема) название журнала и фамилию автора;
    - использовать attach (присоединение);
    - в состав электронной версии статьи должны входить: файл, содержащий текст статьи, и файл(ы), содержащий(е) иллюстрации.
12. Журнал «Информатика и её применения» является некоммерческим изданием. Плата за публикацию не взимается, гонорар авторам не выплачивается.

Адрес редакции журнала «Информатика и её применения»:
Москва 119333, ул. Вавилова, д. 44, корп. 2, ФИЦ ИУ РАН
Тел.: +7 (499) 135-86-92 Факс: +7 (495) 930-45-05
e-mail: rust@ipiran.ru (Сейфуль-Мулюков Рустем Бадриевич)
http://www.ipiran.ru/journal/issues/

Requirements for manuscripts submitted to Journal "Informatics and Applications"

Journal "Informatics and Applications" (Inform. Appl.) publishes theoretical, review, and discussion articles on the research and development in the field of informatics and its applications.

The journal is published in Russian. By a special decision of the editorial board, some articles can be published in English.

The topics covered include the following areas:

- theoretical fundamentals of informatics;
- mathematical methods for studying complex systems and processes;
- information systems and networks;
- information technologies; and
- architecture and software of computational complexes and networks.

1. The Journal publishes original articles which have not been published before and are not intended for simultaneous publication in other editions. An article submitted to the Journal must not violate the Copyright law. Sending the manuscript to the Editorial Board, the authors retain all rights of the owners of the manuscript and transfer the nonexclusive rights to publish the article in Russian (or the language of the article, if not Russian) and its distribution in Russia and abroad to the Founders and the Editorial Board. Authors should submit a letter to the Editorial Board in the following form:

   Agreement on the transfer of rights to publish:

   "*We, the undersigned authors of the manuscript ". . . ", pass to the Founder and the Editorial Board of the Journal "Informatics and Applications" the nonexclusive right to publish the manuscript of the article in Russian (or in English) in both print and electronic versions of the Journal. We affirm that this publication does not violate the Copyright of other persons or organizations.*

   Author(s) signature(s): (name(s), address(es), date).

   This agreement should be submitted in paper form or in the form of a scanned copy (signed by the authors).

2. A submitted article should be attached with **the data on the author(s)** (see item 8). If there are several authors, the contact person should be indicated who is responsible for correspondence with the Editorial Board and other authors about revisions and final approval of the proofs.
3. The Editorial Board of the Journal examines the article according to the established reviewing procedure. If the authors receive their article for correction after reviewing, it does not mean that the article is approved for publication. The corrected article should be sent to the Editorial Board for the subsequent review and approval.
4. The decision on the article publication or its rejection is communicated to the authors. The Editorial Board may also send the reviews on the submitted articles to the authors. Any discussion upon the rejected articles is not possible.
5. The edited articles will be sent to the authors for proofread. The comments of the authors to the edited text of the article should be sent to the Editorial Board as soon as possible.
6. The manuscript of the article should be presented electronically in the MS WORD (.doc or .docx) or LaTeX (.tex) formats, and additionally in the .pdf format. All documents may be sent by e-mail or provided on a CD or diskette. A hard copy submission is not necessary.
7. The recommended typesetting instructions for manuscript.

   Pages parameters: format A4, portrait orientation, document margins (cm): left — 2.5, right — 1.5, above — 2.0, below — 2.0, footer 1.3.

   Text: font —Times New Roman, font size — 14, paragraph indent — 0.5, line spacing — 1.5, justified alignment.

   The recommended manuscript size: not more than 10 pages of the specified format. If the specified size exceeded, the editorial board is entitled to require the author to reduce the manuscript.

   Use only standard abbreviations. Avoid abbreviations in the title and abstract. The full term for which an abbreviation stands should precede its first use in the text unless it is a standard unit of measurement.

   All pages of the manuscript should be numbered.

   The templates for the manuscript typesetting are presented on site: http://www.ipiran.ru/journal/template.doc.
8. The articles should enclose data both in **Russian and English**:

   - title;
   - author's name and surname;
   - affiliation — organization, its address with ZIP code, city, country, and official e-mail address;
   - data on authors according to the format: (see site)
     http://www.ipiran.ru/journal/issues/2013_07_01/authors.asp and
     http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;

- abstract (not less than 100 words) both in Russian and in English. Abstract is a short summary of the article that can be published separately. The abstract is the main source of information on the article and it could be included in leading information systems and data bases. The abstract in English has to be an original text and should not be an exact translation of the Russian one. Good English is required. In abstracts, avoid references and formulae;
- indexing is performed on the basis of keywords. The use of keywords from the internationally accepted thematic Thesauri is recommended.

  Important! Keywords must not be sentences;
- Acknowledgments.

9. References. Russian references have to be presented both in English translation and Latin transliteration (refer http://www.translit.net/ru/bgn/).

   Please take into account the following examples of Russian references appearance:

   Article in journal:

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Rus. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Journal article in electronic format:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Article from the continuing publication (collection of works, proceedings):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental'noe issledovanie prochnosti soedineniy "stal'–kompozit" [Experimental study of the strength of joints "steel–composite"]. *Trudy MGTU "Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem"* [*Bauman MSTU "Mathematical Modeling of Complex Technical Systems" Proceedings*]. 593:125–130.

   Conference proceedings:

   Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

   Books and other monographs:

   Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

   Dissertation and Thesis:

   Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. PhD Thesis. Moscow: IPI RAN. 23 p.

   State standards and patents:

   GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. M.: Standardinform Publs. 10 p.

   Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

   References in Latin transcription are presented in the original language.

   References in the text are numbered according to the order of their first appearance; the number is placed in square brackets. All items from the reference list should be cited.

10. Manuscripts and additional materials are not returned to Authors by the Editorial Board.
11. Submissions of files by e-mail must include:
    - the journal title and author's name in the "Subject" field;
    - an article and additional materials have to be attached using the "attach" function;
    - an electronic version of the article should contain the file with the text and a separate file with figures.
12. "Informatics and Applications" journal is not a profit publication. There are no charges for the authors as well as there are no royalties.

Editorial Board address:
FRC CSC RAS, 44, block 2, Vavilov Str., Moscow 119333, Russia
Ph.: +7 (499) 135 86 92, Fax: +7 (495) 930 45 05
e-mail: rust@ipiran.ru (to Prof. Rustem Seyful-Mulyukov)
http://www.ipiran.ru/english/journal.asp