Информатика и её применения

Том 15 Выпуск 3 Год 2021

СОДЕРЖАНИЕ

Технический редактор *Л. Кокушкина* Художественный редактор *М. Седакова*
Сдано в набор 31.08.21. Подписано в печать 17.09.21. Формат 60 x 84 / 8
Бумага офсетная. Печать цифровая. Усл.-печ. л. 13,72. Уч.-изд. л. 12,5. Тираж 100 экз.
Заказ № 1142
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
Отпечатано в НИПКЦ «Восход-А» с готовых файлов
Москва 123103, наб. Новикова-Прибоя, д. 3, корп. 2

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 2–8

УДАЛЕННЫЙ МОНИТОРИНГ РАБОЧИХ ПРОЦЕССОВ*

А. А. Грушо[1], Н. А. Грушо[2], М. И. Забежайло[3], Е. Е. Тимонина[4]

Аннотация: Рассмотрена схема контроля рабочих процессов в распределенной информационной системе (РИС), экономная с точки зрения объема информации, передаваемой удаленному системному администратору (СА) или офицеру безопасности. Предложенная схема допускает автоматизацию контроля, основана на реальном опыте СА и позволяет реализовать логику определения, классификации и приближенной локализации аномалий. Системный администратор получает информацию о функционировании РИС по каналам связи. Предполагается, что источником сообщений для СА служат сенсоры. Сенсоры являются сущностями, способными распознать информацию, поступившую на вход связанного с сенсором преобразования информации, т. е. если преобразование получает на вход информацию, то сенсор распознает факт и время передачи входной информации первому преобразованию того блока, в котором он находится, в рамках реализующейся информационной технологии (ИТ). Схема основана на вычислении и анализе моментов времени, когда сенсор «видит» передачу данных на вход преобразования в конкретном экземпляре ИТ. Особенностями подхода к мониторингу рабочих процессов являются оценки систематических задержек рабочих процессов и анализ остановов с использованием «параллельных» сенсоров. Построенная схема позволяет динамически детализировать контроль с целью уточнения приближенного места аномалии.

Ключевые слова: информационная безопасность; удаленный мониторинг информационной системы; оценка данных мониторинга по информационным характеристикам

DOI: 10.14357/19922264210301

1 Введение

В работе рассматривается проблема удаленного мониторинга рабочих процессов в больших РИС. Поиск аномалий в Big Data становится актуальной проблемой, обзор работ в этом направлении приведен в [1]. Удаленный поиск аномалий в РИС рассматривался в ряде работ [2–4].

Анализ данных мониторинга обычно ведется СА и/или офицером безопасности. Далее будем предполагать, что весь анализ ведется одним СА. Основная задача СА — бесперебойная работа информационной системы. Системный администратор должен знать, что каждый из элементов РИС функционирует правильно. В своей работе по выявлению, устранению и предотвращению сбоев в компьютерных системах СА сталкивается со сложным взаимодействием множества программно-аппаратных элементов. Поэтому СА опирается на некоторую вспомогательную вычислительную систему, содержащую вспомогательные данные и инструменты обработки данных мониторинга (вычислительный центр (ВЦ) СА). Таким образом, СА должен решать следующие задачи:

- определение работоспособности РИС, т. е. проведение анализа данных о состоянии всей инфраструктуры от бизнес-процессов и квалификации пользователей до решения о правильной работе этой инфраструктуры;
- выявление причин нарушения работоспособности для компенсации потерь из-за сбоев, ошибок и конфликтов с требованиями информационной безопасности (далее — сбоев или аномалий), а также для восстановления работоспособности и предотвращения повторных случаев.

Элементами РИС будем называть оборудование, программное обеспечение, кабельные сети, сетевое оборудование, механизмы и средства обеспечения информационной безопасности. Произвольное множество элементов РИС будем называть объектом. Суммарный вычислительный процесс, объединенный единой целью и формой исходных данных, будем называть информационной технологией [5].

В статье рассмотрены архитектура и возможности реализации системы удаленного мониторинга

*Работа частично поддержана РФФИ (проект 18-29-03081).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, grusho@yandex.ru
[2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, info@itake.ru
[3]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, m.zabezhailo@yandex.ru
[4]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, eltimon@yandex.ru

(СУМ) и удаленного анализа данных мониторинга с целью получения достоверной информации о рабочих процессах в РИС, определения причин нарушения работоспособности и информационной безопасности РИС [6]. Архитектура системы мониторинга и методы решения указанных выше задач основаны на отслеживании времени происходящих в РИС событий.

2 Модель информационной технологии

Для моделирования ИТ в РИС используем аппарат теории графов [7]. Будем моделировать ИТ в виде DAG (Directed Acyclic Graph) с учетом того, что каждая вершина графа представляется в виде тройки: входные данные –преобразование – выходные данные [8]. Тогда циклы графового представления ИТ исчезают из-за различий входных и выходных данных. Такое представление позволяет агрегировать в одну вершину связные подграфы DAG, называть их не преобразованиями, а блоками, агрегировать дуги между блоками и опять получать новый DAG, который представляет ИТ с новыми вершинами и дугами. Указанная процедура обратима. Если блоки заменять агрегированными ранее подграфами, а агрегированные дуги детализировать и соединять с соответствующими вершинами подграфов, то можно детализировать агрегированный DAG вплоть до восстановления исходного DAG [8]. Далее, следуя обозначениям работы [8], прописными латинскими буквами $A, B, \ldots$ будем обозначать данные (объекты), служащие входными или выходными данными преобразований информации в ИТ, преобразования и агрегированные преобразования, т. е. преобразования в блоках будем обозначать строчными латинскими буквами $a, b, \ldots$, а блоки вместе с данными будем обозначать строчными греческими буквами $\alpha, \beta, \ldots$

Множество построенных представлений ИТ в виде множества DAG образует частичный порядок [9]. Максимальной вершиной корневого дерева, представляющего это частично упорядоченное множество DAG, служит DAG, состоящий из одного блока. Построенное семейство DAG является частью метаданных ИТ [9].

3 Модель удаленного администратора

Системный администратор получает информацию о функционировании РИС по каналам связи. В работе предполагается, что источником сообщений для СА служат сенсоры. Сенсоры являются сущностями, способными распознать информацию, поступившую на вход связанного с сенсором преобразования информации, т. е. если преобразование a получает на вход информацию A, то сенсор распознает факт и время передачи входной информации первому преобразованию a того блока, в котором он находится, в рамках реализующейся ИТ.

В работе не рассматривается вопрос об оптимальном расположении сенсоров. Теория размещения сенсоров развита в работе [4], однако практика ее применения требует дополнительных исследований.

Предположим, что работающий сенсор, связанный с первым преобразованием из блока α, является единственным для этого блока, и будем обозначать его $s(\alpha)$. Тогда DAG(IT), описывающий ИТ, порождает новый граф DAG(C), где C — множество работающих сенсоров в РИС. Вершинами DAG(C) служат все сенсоры, привязанные к вершинам DAG(IT). Дуга от $s(\alpha)$ идет к вершине $s(\beta)$ тогда и только тогда, когда в DAG(IT) существует ориентированный путь от блока α к блоку β, не содержащий других сенсоров.

Этот граф корректно определен для тех представлений DAG(IT), в которых сенсоры не попадают внутрь блоков. Если сенсор лежит внутри блока, то построенная выше его идентификация, связанная с возможностью вычисления момента времени поступления входных данных в преобразование блока, не определена. Поэтому сенсоры внутри блоков не работают и не передают СА информацию.

Кроме того, граф сенсоров для каждого допустимого DAG(IT) также является DAG. В самом деле, если сенсорный граф DAG(C) имеет ориентированный цикл, то из условия, что с каждым блоком может быть связано не более одного работающего сенсора, следует, что в исходном DAG тоже должен быть ориентированный цикл. Это противоречит определению DAG(IT).

Но даже для корректно определенных графов сенсоров необходимо отметить, что они приобретают у СА различные наименования вершин, так как идентификаторы блоков меняются в зависимости от агрегации DAG.

Пусть DAG_1 и DAG_2 описывают одну ИТ при фиксированной системе работающих сенсоров C, из которых ни один не попадает внутрь блока. Функционал сенсора не зависит от функционала ИТ, но зависит от имени блока, с которым он связан. Считаем, что сенсор сам вычисляет регистрируемое время поступления на вход блока

информации, а также без взаимодействия с функционалом блока формирует и отправляет сообщение СА. Тогда можно утверждать, что все корректно построенные графы сенсоров изоморфны друг другу. В самом деле, имя сенсора будет меняться у СА в зависимости от агрегации или детализации DAG(IT).

Рассмотрим один шаг агрегации DAG(IT), в котором a является первым преобразованием блока α. С этим преобразованием a связан сенсор $s(a)$, по которому построен граф с вершиной $s(\alpha)$. Агрегация, при которой работающий $s(a)$ попадает внутрь какого-либо блока, недопустима. Тогда при допустимой агрегации $s(a)$ остается вне агрегированного блока со стороны поступающих в a данных. Поскольку существует ориентированный путь из a в b, с которым связан $s(b)$ в блоке β, то в агрегированном графе, в котором α переходит в α^*, по-прежнему существует ориентированный путь из агрегированной вершины α^* в β. Этот путь не может иметь длину 0, так как тогда $s(b)$ окажется внутри блока α^*. Таким образом, если агрегация не нарушает правило непокрытия сенсора, то оно допустимо для исходного DAG. А значит, исходный сенсор $s(a)$ приобретает у СА новое имя $s(\alpha^*)$ и по-прежнему существует путь до следующего сенсора $s(\beta)$.

В графе сенсоров, построенном по новым именам вершин, по-прежнему существует дуга, соединяющая $s(\alpha^*)$ с сенсором $s(\beta)$. Следовательно, если прежний DAG может при выполнении указанных ограничений перейти в новый DAG, то новый граф сенсоров отличается от прежнего только переименованием вершин, т. е. новый граф изоморфен прежнему. Поэтому можем считать, что каждый сенсор имеет уникальное имя, не связанное с переименованием вершин DAG и это имя известно каждому сенсору, что не мешает СА осуществлять привязку к соответствующему блоку в любом допустимом DAG(IT).

Справедливо и обратное утверждение. Пусть фиксирован DAG_1(IT) и множество C связанных с ним работающих сенсоров. Тогда существует единственный DAG(IT), который не может быть далее агрегирован у СА при данной системе работающих сенсоров.

В самом деле, каждый ориентированный путь (без работающих сенсоров) между блоками с сенсорами из C можно стянуть к новому блоку, расположенному вместо блока, из которого выходит рассматриваемый путь. Тогда получим граф, изоморфный графу DAG(C). Таким образом, СА не может получить больше информации о рабочих процессах ИТ, чем в построенной максимальной модели DAG(IT), и более детализированные DAG ему не нужны.

Далее будет показано, как в этом случае СА может вычислить блок с аномалией.

Предположим, что в блоке с аномалией есть неработающие сенсоры. Тогда СА, рассматривая отдельно аномальный блок, может включить в работу не работавшие ранее сенсоры и, рассмотрев более детализированный DAG(IT), уточнить локализацию аномалии. Этим объясняется целесообразность рассмотрения множества различных DAG и существования только части работающих в определенное время сенсоров. При нормальной работе ИТ максимальный DAG для заданного множества работающих сенсоров C представляется более оптимальным, чем детализированный DAG с множеством сенсоров C^*, для $|C^*| > |C|$.

4 Алгоритмы контроля распределенных информационных систем

Рассмотрим язык общения сенсоров с ВЦ СА, который далее будем часто отождествлять с СА. Язык общения сенсора с ВЦ СА для данной ИТ имеет следующие элементы. Каждая реализация ИТ приобретает уникальный идентификатор ID(IT), который сопровождает передачу данных между блоками ИТ. При передаче входных данных первому преобразованию a блока α сенсор $s(a)$ фиксирует время такой передачи и формирует сообщение для СА, содержащее ID(IT), время и имя передающего сенсора.

По предположению, существует канал связи для передачи этого сообщения СА. На стороне СА вместе с формированием ID(IT) для запуска ИТ фиксируется стартовое время, которое далее сравнивается со всеми моментами времени во входящих сообщениях. Вместе с метаданными СА получает возможность отслеживать процесс реализации каждой ИТ и времена достижения очередных блоков преобразований соответствующей ИТ. Если в ИТ участвуют пользователи, которые запускают следующие шаги ИТ, то вместе с этими стартами передаются идентификаторы экземпляров ИТ, которые пользователи получили от уже реализованной части ИТ.

Предположим, что последовательности времен предыдущих экземпляров каждой ИТ, закончившихся успешно, хранятся в памяти ВЦ СА. Тогда появляется возможность для каждой ИТ оценить статистические характеристики хранящихся времен. Однако далее будет показано, что традиционное использование методов математической статистики здесь не работает.

Предположим, что для каждого сенсора в ИТ можно определить в ВЦ СА две константы $C_1(s)$ и $C_2(s)$. Пусть $\Delta(s)$ — разница между временем в сообщении сенсора s и стартовым временем. Если $\Delta(s) \le C_1(s)$, то процесс реализации ИТ проходит нормально, что будем обозначать символом «1». Если $C_1(s) < \Delta(s) \le C_2(s)$, то процесс реализации ИТ проходит с задержкой, что будем обозначать символом «0». Если же $C_2(s) < \Delta(s)$, то это значит, что процесс реализации ИТ остановился и по дальнейшим путям от s не рассматривается, т. е. произошел останов, что будем обозначать символом «-1».

Рассмотрим логику анализа полученных временны́х характеристик одной ИТ. Если сообщение от сенсора s соответствует 1, то на любом из следующих ближайших сенсоров может появиться 1. Если сообщение от сенсора s соответствует 0, то на любом из следующих ближайших сенсоров не может появиться 1, а может появиться 0 или -1, если к СА перестанут поступать сообщения. Переходы 0 в 0 могут означать, что задержка от одного блока передается по цепочке другому блоку.

Однако может так случиться, что задержка породит далее останов, т. е. 0 перейдет в -1. Переходы 0 в 1 и -1 в 0 невозможны. Отсюда возникает возможность отслеживать блок, в котором произошла аномалия, а именно: переходы времен от соседних сенсоров 1 в 0 и 0 в -1 однозначно определяют блок в максимальном DAG(IT) для данного C, в котором произошла аномалия.

Отметим, что переход 1 в 0 может произойти по трем причинам:

(1) образовалась очередь к первому преобразованию блока;

(2) в блоке произошел неявный сбой;

(3) в блоке существует хоть один процесс, порождающий повышенную нагрузку на компьютерную систему.

Сенсор контролирует передачу данных только в первое преобразование блока α. В ходе следующих преобразований могут поступать извне новые данные, а выходные данные промежуточных преобразований передаваться преобразованиям, не входящим в данный блок, т. е. при 1, которая соответствует сообщению сенсора на входе рассматриваемого блока α, на вход одного сенсора после данного блока может попасть 1, но на другой сенсор после данного блока — 0.

Приведем пример распространения 1 и 0 в рассматриваемом случае:

$$\begin{array}{c} 1 \\ \uparrow \\ 1 \to \alpha \to 0\,. \end{array}$$

В этом примере в блоке α в некотором преобразовании произошла аномалия, но до нее процесс выполнялся нормально.

Теперь рассмотрим, почему данная схема контроля работает плохо и что нужно делать, чтобы она работала хорошо.

Рассмотрим блок α и сенсор $s(\alpha)$. Время работы блока α оценивается как разность временны́х меток следующего сенсора $s(\beta)$ и сенсора $s(\alpha)$. Пусть в идеале это время равно $T(\alpha)$. Это время подвержено случайным искажениям. Поэтому точное время $T(\alpha)$ будет смещено в ту или иную сторону за счет случайных искажений. Запишем реальное значение времени обработки в виде $T(\alpha) + \xi_1$, где ξ_1 — случайная величина, изменяющая время работы за счет случайных искажений.

Но если к блоку α образовалась очередь, то время обработки второго набора в очереди будет $2T(\alpha)$ и время реализации для этого набора будет иметь вид $2T(\alpha) + \xi_i + \xi_2$, где $\xi_1 + \xi_2$ — сумма случайных величин, изменяющих время работы за счет случайных искажений.

Если из k реализаций данного блока доля тех заданий, которые попадают в очередь вторыми, равна λk, а остальные обрабатываются без очереди, то среднее время увеличится и составит

$$\Delta T = T(\alpha)(1 + \lambda) + \frac{1}{k}\sum \xi_i\,.$$

Поскольку случайные величины ξ_i имеют разные знаки, то их среднее арифметическое с ростом k стремится по вероятности к 0. Отсюда следует, что одноразовые задержки трудно выявляются, но систематические задержки, связанные с аномалиями в обработке, могут выявляться достаточно устойчиво. В этом случае надо запоминать два параметра, а именно: стандартное время одной обработки (минимальное время $T(\alpha)$) и среднее время по последней серии длины k наблюдений времен реализаций работы блока α.

Другие типы задержек также целесообразно выявлять с помощью нескольких наблюдений, так как только в этом случае они представляют реальную аномалию.

В случае останова стратегия вычисления «-1» становится другой. Для повышения достоверности останова желательно сравнивать времена задержки сразу у нескольких сенсоров, непосредственно связанных с $s(\alpha)$. Если для одного сенсора $s(b)$ наблюдается $C_2(s(b)) < \Delta(s(b))$, то одновременное превышение $C_2(s)$ еще для одного или нескольких сенсоров, не следующих за $s(b)$, значительно усиливает информацию об останове в блоке α (против очереди).

5 Случай нескольких одновременно выполняемых информационных технологий

Рассмотрим случай одновременного выполнения в РИС нескольких ИТ. Пусть множество работающих сенсоров C одно и то же для всех ИТ. Однако для каждой ИТ имеется свое множество DAG, и блоки в них отличаются, но не отличаются преобразования, с которыми связаны сенсоры. В сообщениях от сенсоров присутствуют идентификаторы ИТ, с которыми связаны эти сообщения. Поэтому множество приходящих сообщений однозначно разделяется по ИТ, которые функционируют в данное время. Стартовые времена в различных ИТ могут отличаться, но для каждой технологии время развития процесса выполнения ИТ вычисляется однозначно. Если для двух технологий $ИТ_1$ и $ИТ_2$ приходит информация от одного сенсора, то это значит, что по крайней мере первые преобразования у соответствующих блоков $ИТ_1$ и $ИТ_2$ совпадают, т. е. аномалия в одном блоке с этим преобразованием и отсутствие аномалии в другом блоке с этим преобразованием показывают только, что преобразование работает правильно.

При нескольких обращениях к этому преобразованию может образоваться очередь. Если в $ИТ_1$ и $ИТ_2$ результат 1 на всех ближайших следующих сенсорах $DAG(IT_1)$ и результат 0 на всех ближайших следующих сенсорах $DAG(IT_2)$, то, по-видимому, это результат очереди.

Приведенные примеры показывают, что существуют дедуктивные и правдоподобные рассуждения и выводы, в значительной степени сокращающие полный перебор при анализе данных мониторинга СА.

6 Заключение

В работе рассмотрена схема контроля рабочих процессов в РИС, экономная с точки зрения объема информации, передаваемой удаленному СА. Предложенная схема допускает автоматизацию контроля, основана на реальном опыте СА и позволяет реализовать логику определения, классификации и приближенной локализации аномалий.

Схема основана на вычислении и анализе моментов времени, когда сенсор «видит» передачу данных на вход преобразования в конкретном экземпляре ИТ. Особенности подхода к мониторингу рабочих процессов заключаются в оценке систематических задержек рабочих процессов и анализе остановов с использованием «параллельных» сенсоров. Построенная схема позволяет динамически детализировать контроль с целью уточнения приближенного места аномалии.

Данный подход использует вспомогательную информацию в виде математических моделей ИТ, представленных в форме множеств ориентированных ациклических графов и базы данных накапливаемых значений времен обработки информации различными преобразованиями и блоками.

Нет оснований полагать, что описанный подход — единственный инструмент удаленного СА для контроля рабочих процессов в РИС. Задача ставилась как поиск простых решений для удаленного контроля, которые могут быть автоматизированы.

Литература

1. *Thudumu S., Branch P., Jin J., Singh J. J.* A comprehensive survey of anomaly detection techniques for high dimensional big data // J. Big Data, 2020. Vol. 7. Art. No. 42. 30 p. doi: 10.1186/s40537-020-00320-x.
2. *Dereszynski E. W., Dietterich T. G.* Probabilistic models for anomaly detection in remote Sensor data streams // arXiv.org, 20 Jun 2012. arXiv:1206.5250 [cs.AI]. P. 75–82.
3. *Шкодырев В. П., Ягафаров К. И., Баштовенко В. А., Ильина Е. Э.* Обзор методов обнаружения аномалий в потоках данных // CEUR Workshop Procee., 2017. Vol. 1864. Art. 8. 7 p.
4. *Hooi B.* Anomaly detection in graphs and time series: Algorithms and applications: PhD Thesis. — Pittsburgh, PA, USA: Carnegie Mellon University, 2019. 222 p. http://reports-archive.adm.cs.cmu.edu/anon/ml2019/CMU-ML-19-100.pdf.
5. *Самуйлов К. Е., Чукарин А. В., Яркина Н. В.* Бизнес-процессы и информационные технологии в управлении телекоммуникационными компаниями. — М.: Альпина Паблишерс, 2009. 442 с.
6. *Grusho A., Grusho N., Zabezhailo M., Timonina E., Senchilo V.* Metadata for root cause analysis // Communications ECMS, 2021. Vol. 35. Iss. 1. P. 267–271. doi: 10.7148/2021-0267.
7. *Brandón Á., Solé M., Huélamo A., Solans D., Pérez M. S., Muntés-Mulero V.* Graph-based root cause analysis for service-oriented and microservice architectures // J. Syst. Software, 2020. Vol. 159. Art. No. 110432. 17 p. doi: 10.1016/j.jss.2019. 110432.
8. *Grusho A., Grusho N., Zabezhailo M., Timonina E.* Generation of metadata for network control // Distributed computer and communication networks / Eds. V. M. Vishnevskiy, K. E. Samouylov, D. V. Kozyrev. — Lecture notes in computer science ser. — Springer, Cham, 2020. Vol. 12563. P. 723–735. doi: 10.1007/978-3-030-66471-8_55.
9. *Грушо Н. А., Грушо А. А., Забежайло М. И., Тимонина Е. Е.* Методы нахождения причин сбоев в информационных технологиях с помощью метаданных // Информатика и её применения, 2020. Т. 14. Вып. 2. С. 33–39. doi: 10.14357/19922264200205.

Поступила в редакцию 01.07.2021

REMOTE MONITORING OF WORKFLOWS

A. A. Grusho, N. A. Grusho, M. I. Zabezhailo, and E. E. Timonina

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

Abstract: The paper discusses the workflows control scheme in the distributed information system which is economical in terms of the amount of information provided to the remote system administrator or security officer. The proposed scheme allows automation of control, is based on real experience of the system administrator, and allows implementing logic of determination, classification, and approximating localization of anomalies. The system administrator receives information about the operation of the distributed information system through the communication channels. In operation, it is assumed that the sources of messages for the system administrator are sensors. Sensors are entities capable of recognizing information received at the input of sensor-related transformation of information, that is, if the transformation receives information at the input, then the sensor recognizes the fact and time of transmitting the input information to the first transformation of the block in which it is located within the framework of the information technology being implemented. The scheme is based on the calculation and analysis of the moments when the sensor "sees" the transfer of data to the transformation input in a particular instance of information technology. The characteristics of the approach to workflows monitoring are estimations of systematic process delays and analysis of outages using "parallel" sensors. The constructed scheme allows one to dynamically detail the control to clarify the approximate location of the anomaly.

Keywords: information security; remote monitoring of information system; evaluation of monitoring data by information characteristics

DOI: 10.14357/19922264210301

Acknowledgments

The paper was partially supported by the Russian Foundation for Basic Research (project 18-29-03081).

References

1. Srikanth, T., P. Branch, J. Jin, and J. Singh. 2020. A comprehensive survey of anomaly detection techniques for high dimensional big data. *J. Big Data* 7:42. 30 p. doi: 10.1186/s40537-020-00320-x.
2. Dereszynski, E. W., and T. G. Dietterich. 2012. Probabilistic models for anomaly detection in remote sensor data streams. *arXiv.org*. 8 p. Available at: https://arxiv.org/ftp/arxiv/papers/1206/1206.5250.pdf (accessed July 9, 2021).
3. Shkodyrev, V., K. Yagafarov, V. Bashtovenko, and E. Ilyina. 2017. Obzor metodov obnaruzheniya anomaliy v potokakh dannykh [The overview of anomaly detection methods in data streams]. *CEUR Workshop Procee*. 1864:8. 7 p.
4. Hooi, B. 2019. *Anomaly detection in graphs and time series: Algorithms and applications*. Pittsburgh, PA: Carnegie Mellon University. PhD Thesis. 222 p. Available at: http://reports-archive.adm.cs.cmu.edu/anon/ml2019/CMU-ML-19-100.pdf (accessed August 7, 2021).
5. Samuylov, K. E., A. V. Chukarin, and N. V. Yarkina. 2009. *Biznes-protsessy i informatsionnye tekhnologii v upravlenii telekommunikatsionnymi kompaniyami* [Business processes and information technologies in management of the telecommunication companies]. Moscow: Alpina Publs. 442 p.
6. Grusho, A., N. Grusho, M. Zabezhailo, E. Timonina, and V. Senchilo. 2021. Metadata for root cause analysis. *Communications ECMS* 35(1):267–271. doi: 10.7148/2021-0267.
7. Brandón, Á., M. Solé, A. Huélamo, D. Solans, M. S. Pérez, and V. Muntés-Mulero. 2020. Graph-based root cause analysis for service-oriented and microservice architectures. *J. Syst. Software* 159:110432. 17 p. doi: 10.1016/j.jss.2019.110432.
8. Grusho, A., N. Grusho, M. Zabezhailo, and E. Timonina. 2020. Generation of metadata for network control. *Distributed computer and communication networks*. Eds. V. M. Vishnevskiy, K. E. Samouylov, and D. V. Kozyrev. Lecture notes in computer science ser. Springer, Cham. 12563:723–735. doi: 10.1007/978-3-030-66471-8_55.
9. Grusho, N. A., A. A. Grusho, M. I. Zabezhailo, and E. E. Timonina. 2020. Metody nakhozhdeniya prichin sboev v informatsionnykh tekhnologiyakh s pomoshch'yu metadannykh [Methods of finding the causes of information technology failures by means of meta data]. *Informatika i ee Primeneniya — Inform. Appl.* 14(2):33–39. doi: 10.14357/19922264200205.

Received July 1, 2021

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

Grusho Nikolai A. (b. 1982) — Candidate of Science (PhD) in physics and mathematics, senior 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; info@itake.ru

Zabezhailo Michael I. (b. 1956) — Doctor of Science in physics and mathematics, principal scientist, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, 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, 2021. Т. 15. Вып. 3. С. 9–15

ФИЛЬТРАЦИЯ СОСТОЯНИЙ МАРКОВСКИХ СКАЧКООБРАЗНЫХ ПРОЦЕССОВ ПО КОМПЛЕКСНЫМ НАБЛЮДЕНИЯМ II: ЧИСЛЕННЫЙ АЛГОРИТМ*

А. В. Борисов[1], Д. Х. Казанчян[2]

Аннотация: Работа представляет заключительную часть исследований, начатых в статье «Фильтрация состояний марковских скачкообразных процессов по комплексным наблюдениям I: точное решение задачи» (Информатика и её применения, 2021. Т. 15. Вып. 2. С. 12–19). Предложен алгоритм численной реализации решения задачи фильтрации состояний марковского скачкообразного процесса (МСП) по совокупности наблюдаемых считающих процессов и диффузии с мультипликативными шумами. Исходную задачу оценивания предлагается приблизить последовательностью соответствующих задач фильтрации по наблюдениям, дискретизованным по времени. В работе выведен рекуррентный вид дискретизованной оценки, определен показатель ее точности на одном шаге и получена зависимость этой точности от характеристик применяемой схемы численного интегрирования.

Ключевые слова: марковский скачкообразный процесс; оптимальная фильтрация; мультипликативные шумы в наблюдениях; дискретизованные наблюдения; точность аппроксимации

DOI: 10.14357/19922264210302

1 Введение

В первой части цикла [1] представлено решение задачи фильтрации состояний *марковского скачкообразного процесса* с непрерывным временем и конечным множеством состояний по наблюдениям совокупности диффузионных и считающих процессов. Интенсивности шумов диффузионных наблюдений и скачков считающих наблюдений являются функциями оцениваемого состояния. Шумы такого вида называются *мультипликативными*.

Для корректного использования математического аппарата стохастического анализа решаемая задача фильтрации была сформулирована в форме нахождения *условного математического ожидания* (УМО) относительно не исходного потока σ-алгебр наблюдений, а его модификации, непрерывной справа. Было предложено эквивалентное с информационной точки зрения преобразование диффузионных наблюдений в совокупность диффузионного процесса с единичной интенсивностью, считающих процессов, фиксирующих часть скачков оцениваемого состояния, и косвенных наблюдений состояния, выполненных в дискретные неслучайные моменты времени. Оценка оптимальной фильтрации была получена в виде решения непрерывно-дискретной стохастической дифференциальной системы с преобразованными наблюдениями в правой части. В [1] также доказано утверждение о том, что при выполнении некоторых достаточно необременительных условий идентифицируемости оптимальная оценка фильтрации *почти наверное* (п. н.) совпадает с оцениваемым состоянием МСП, т. е. имеется возможность точного восстановления состояния по имеющимся косвенным зашумленным наблюдениям.

Теоретическое решение задачи фильтрации и ее свойства выглядят многообещающими для использования в решении практических задач оценивания и управления по неполной информации. Однако применение этих результатов сопряжено с рядом трудностей. Во-первых, на практике «сглаживание» справа потока σ-алгебр наблюдений означает доступность будущих наблюдений на некотором небольшом отрезке времени либо, что эквивалентно, решение задачи фильтрации с неко-

*Работа выполнена при частичной поддержке РФФИ (проект 19-07-00187 А) и в соответствии с программой Московского центра фундаментальной и прикладной математики.

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук; Московской авиационный институт; Центр фундаментальной и прикладной математики Московского государственного университета имени М. В. Ломоносова, ABorisov@frccsc.ru

[2]Факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова, Drastamat94@gmail.com

торым малым запаздыванием. Во-вторых, преобразование наблюдений, использованное в теоретическом решении задачи, не позволяет реализовать его с помощью конструктивного численного алгоритма с некоторой гарантированной точностью. Дело в том, что данное формально корректное преобразование диффузионных наблюдений является результатом двойного предельного перехода, причем сходимость в этих переходах подразумевается по вероятности. Результатом первого перехода является квадратичная характеристика исходных диффузионных наблюдений, результатом второго — ее производная справа. Сложности реализации этих преобразований средствами цифровой вычислительной техники вполне очевидны.

Целью данной части цикла ставится разработка численных алгоритмов реализации решения поставленной задачи фильтрации. По своей сути, это обобщение алгоритмов, предложенных в цикле [2–4] на случай дополнительно доступных считающих наблюдений.

Статья организована следующим образом.

Раздел 2 содержит формальную постановку задачи фильтрации состояний однородного МСП по диффузионным и считающим наблюдениям, дискретизованным по времени с постоянным шагом.

В разд. 3 дано теоретическое решение этой задачи. Оно представляет собой рекуррентную процедуру, на каждом шаге которой применяется обобщенная формула Байеса. В ее числителе и знаменателе присутствуют интегралы — сдвиг-масштабные смеси произведений «гауссиана – пуассоновская вероятность», в которых в качестве смешивающих выступают распределения времени пребывания МСП в каждом из возможных состояний на отрезке временно́й дискретизации. Упомянутые интегралы не могут быть вычислены аналитически, и их предлагается вычислять приближенно с помощью интегральных сумм. В разделе представлена локальная характеристика получаемых аппроксимаций, определяющая точность приближения на одном шаге. Доказано утверждение, связывающее точность схемы численного интегрирования и данную характеристику точности.

Анализ представленных результатов, заключительные замечания и перспективные направления дальнейших исследований даны в разд. 4.

2 Постановка задачи фильтрации

На полном вероятностном пространстве с фильтрацией $(\Omega, \mathcal{F}, \mathcal{P}, \{\mathcal{F}_t\}_{t\geqslant 0})$ рассматривается стохастическая динамическая система наблюдения

$$X_t = X_0 + \int_0^t \Lambda^\top X_s\, ds + \mu_t^X; \tag{1}$$

$$\mathcal{Y}_r = \int_{t_{r-1}}^{t_r} f X_s\, ds + \int_{t_{r-1}}^{t_r} \sum_{n=1}^{N} X_s^n g_n^{1/2}\, dW_s,$$
$$r \in \mathbb{N},\ \ t_r = r\Delta; \tag{2}$$

$$\mathcal{Z}_r = \int_{t_{r-1}}^{t_r} h X_s\, ds + \left(\mu_{t_r}^Z - \mu_{t_{r-1}}^Z\right),$$
$$r \in \mathbb{N},\ \ t_r = r\Delta, \tag{3}$$

где

$X_t \in \mathbb{S}^N$ — ненаблюдаемое состояние системы — однородный МСП с множеством состояний $\mathbb{S}^N \triangleq \{e_1, \ldots, e_N\}$ ($\mathbb{S}^N$ — множество единичных векторов пространства $\mathbb{R}^N$), матрицей интенсивностей переходов Λ и начальным распределением π; $\mu_t^X \in \mathbb{R}^N$ — $\mathcal{F}_t$-согласованный мартингал;

$\mathcal{Y}_r \in \mathbb{R}^M$ — дискретизованные по времени с шагом Δ непрерывные косвенные наблюдения, зашумленные $\mathcal{F}_t$-согласованным стандартным винеровским процессом $W_t \in \mathbb{R}^M$; $(M \times N)$-мерная матрица f характеризует план наблюдений, а набор $(M \times M)$-мерных симметричных матриц $\{g_n\}_{n=\overline{1,N}}$ определяет интенсивности шумов в зависимости от текущего состояния X_t;

$\mathcal{Z}_r \in \mathbb{R}^K$ — дискретизованные по времени с шагом Δ косвенные наблюдения, компоненты которых являются считающими процессами: элементы $(K \times N)$-мерной матрицы h определяют интенсивность скачков отдельных компонент в зависимости от текущего состояния X_t; $\mu_t^Z \in \in \mathbb{R}^K$ — $\mathcal{F}_t$-согласованный мартингал.

Обозначим через $\mathfrak{O}_r \triangleq \{\mathcal{Y}_q, \mathcal{Z}_q :\ q = \overline{1,r}\}$ последовательность σ-алгебр всех наблюдений, полученных на отрезке времени $[0, t_r]$, $r \in \mathbb{N}$, $\mathfrak{O}_0 \triangleq \{\varnothing, \Omega\}$.

Задача оптимальной фильтрации состояния МСП X_t по дискретизованным наблюдениям заключается в построении УМО $\widehat{\mathcal{X}}_r \triangleq \mathsf{E}\{X_{t_r}|\mathfrak{O}_r\}$, $r \in \mathbb{N}$.

Ниже представлены ограничения на исследуемую систему наблюдения (1)–(3), необходимые для корректного решения поставленной задачи фильтрации.

A. Исследуемый вероятностный базис с фильтрацией является пространством Винера–Пуассона [5, 6].

Б. Шумы в $\mathcal{Y}$ равномерно невырожденны [7], т. е. $\min_{1\leqslant n\leqslant N} g_n > 0$.

В. Компоненты мартингала μ^Z в наблюдениях $\mathcal{Z}$ (3) ортогональны друг другу, т. е. $\langle \mu^Z\rangle_t = \int_0^t \operatorname{diag}(hX_s)\,ds$. Мартингалы в уравнении состояния μ^X и в наблюдениях μ^Z также ортогональны $\langle \mu^X, \mu^Z\rangle_t \equiv 0$.

Рассматриваемая в данной части цикла система наблюдения (1)–(3) является автономной в отличие от более общего случая, рассмотренного в [1]. Для решения задачи фильтрации по дискретизованным наблюдениям это существенно. В автономном случае оценка фильтрации будет зависеть только от распределения времени пребывания МСП в каждом из возможных состояний на интервале дискретизации. В неавтономном случае оценка определяется вероятностным распределением отрезка траекторий в целом. Предположение об однородности может быть ослаблено: коэффициенты системы наблюдения должны быть постоянными на интервалах дискретизации.

Рассмотрим непрерывный вариант наблюдений, соответствующий (2), (3):

$$Y_t = \int_0^t fX_s\,ds + \int_0^t \sum_{n=1}^N X_s^n g_n^{1/2}\,dW_s;$$

$$Z_t = \int_0^t hX_s\,ds + \mu_t^Z,$$

и соответствующий поток σ-алгебр $\{\mathcal{O}_t\}_{t\geqslant 0}$: $\mathcal{O}_t \triangleq \sigma\{Y_s, Z_s : 0 \leqslant s \leqslant t\}$. Зафиксируем произвольный момент времени $t > 0$ и построим вложенную последовательность двоичных разбиений отрезка $[0,t]$ с уменьшающимся шагом $\Delta_n \triangleq t/2^n$ и соответствующие наблюдения $(\mathcal{Y}_r^n, \mathcal{Z}_r^n)$, подобные (2), (3), с шагом Δ_n. Далее построим последовательность σ-алгебр $\{\mathfrak{O}_{2^n}^n\}_{n\in\mathbb{N}}$: $\mathfrak{O}_{2^n}^n \triangleq \sigma\{\mathcal{Y}_r^n, \mathcal{Z}_r^n : 1 \leqslant r \leqslant 2^n\}$. В силу того что векторный процесс (Y_t, Z_t) обладает свойством сепарабельности, имеет место сходимость σ-алгебр $\mathcal{O}_t = \lim_{n\to\infty} \mathfrak{O}_{2^n}^n$, и по теореме Леви [7] $\mathsf{E}\{X_t|\mathfrak{O}_{2^n}^n\} \to \mathsf{E}\{X_t|\mathcal{O}_t\}$ $\mathcal{P}$-п. н. при $n \to \infty$. Этот факт означает, что решение задачи оптимальной фильтрации состояния МСП по наблюдениям, дискретизованным по времени, помимо самостоятельной практической значимости может рассматриваться как один из способов приближенного решения задачи оптимальной фильтрации по совокупности диффузионных и считающих наблюдений. При этом строить версию потока σ-алгебр наблюдений $\mathcal{O}_{t+}$, непрерывных справа, не требуется.

3 Оценка фильтрации по дискретизованным наблюдениям и ее численная реализация

Ниже в статье будут использоваться следующие обозначения.

1. $\|\alpha\|_Q^2 \triangleq \alpha^\top Q\alpha$.
2. I — единичная матрица подходящей размерности; $\mathbf{0}$ — нулевая матрица подходящей размерности; $\mathbf{1}$ — вектор-строка подходящей размерности, составленная из единиц.
3. $\mathbf{I}_{\mathcal{A}}(x)$ — характеристическая функция множества $\mathcal{A}$.
4. $\tau_r \triangleq \int_{t_{r-1}}^{t_r} X_s\,ds$ — вектор, компоненты которого равны случайному времени пребывания процесса X в каждом из возможных состояний на отрезке $[t_{r-1}, t_r]$.
5. $\mathcal{D} \triangleq \{u = \operatorname{col}(u^1,\dots,u^N) :\ u^n \geqslant 0,\ \sum_{n=1}^N u^n = \Delta\}$ — $(N-1)$-мерный симплекс в пространстве $\mathbb{R}^N$ — носитель распределения вектора τ_r.
6. $\Pi \triangleq \{\pi = \operatorname{col}(\pi^1,\dots,\pi^N) :\ \pi^n \geqslant 0,\ \sum_{n=1}^N \pi^n = 1\}$ — «вероятностный симплекс», множество возможных начальных распределений π.
7. $\mathbf{p}^{k\ell}(du)$ — распределение вектора $X_{t_r}^\ell \tau_r$ при условии $X_{t_{r-1}} = e_k$, т. е. для любого $\mathcal{G} \in \mathcal{B}(\mathbb{R}^N)$ верно равенство
$$\mathsf{E}\left\{\mathbf{I}_{\mathcal{G}}(\tau_r) X_{t_r}^\ell | X_{t_{r-1}} = e_k\right\} = \int_{\mathcal{G}} \mathbf{p}^{k\ell}(du).$$
8. $\mathcal{N}(y,m,Q) \triangleq (2\pi)^{-M/2}\det{}^{-1/2} Q e^{-\|y-m\|_{Q^{-1}}^2/2}$ — M-мерная плотность гауссовского распределения с математическим ожиданием m и невырожденной ковариационной матрицей Q.
9. $\mathcal{P}(z,\alpha) \triangleq e^{-\sum_{k=1}^K \alpha^k} \prod_{k=1}^K (\alpha^k)^{z^k}/(z^k)!$ ($z = \operatorname{col}(z^1,\dots,z^K) \in \mathbb{Z}_+^K$; $\alpha = \operatorname{col}(\alpha^1,\dots,\alpha^K) \in \mathbb{R}_+^K$) — распределение K-мерного случайного вектора с независимыми компонентами, имеющими пуассоновское распределение с параметрами α^k, $k = \overline{1,K}$.
10. Функции
$$\theta^{kj} = \theta^{kj}(y,z) \triangleq$$
$$\triangleq \int_{\mathcal{D}} \mathcal{N}\left(y, fu, \sum_{p=1}^N u^p g_p\right) \mathcal{P}(z, hu)\mathbf{p}^{kj}(du); \quad (4)$$
$$\theta = \theta(y,z) \triangleq \|\theta^{kj}(y,z)\|_{k,j=\overline{1,N}};$$
$$\theta_r \triangleq \theta(\mathcal{Y}_r, \mathcal{Z}_r).$$

Теорема 1. *В условиях A–B оптимальная оценка фильтрации по дискретизованным наблюдениям определяется следующей рекуррентной схемой*:

$$\widehat{\mathcal{X}}_0 = \pi\,;$$

$$\widehat{\mathcal{X}}_r^j = \sum_{k=1}^N \widehat{\mathcal{X}}_{r-1}^k \int\limits_{\mathcal{D}} \mathcal{N}\left(\mathcal{Y}_r, fu, \sum_{p=1}^N u^p g_p\right) \times$$

$$\times\, \mathcal{P}\left(\mathcal{Z}_r, hu\right) \mathbf{p}^{kj}(du) \Bigg/$$

$$\left(\sum_{i,\ell=1}^N \widehat{\mathcal{X}}_{r-1}^i \int\limits_{\mathcal{D}} \mathcal{N}\left(\mathcal{Y}_r, fv, \sum_{q=1}^N v^q g_q\right) \times\right.$$

$$\left. \times\, \mathcal{P}\left(\mathcal{Z}_r, hv\right) \mathbf{p}^{i\ell}(dv)\right), \quad j = \overline{1,N}\,. \quad (5)$$

Доказательство теоремы 1 дано в приложении.

Используя введенные выше определения, оценку (5) можно записать в форме матричной рекурсии:

$$\widehat{\mathcal{X}}_r = \frac{1}{\mathbf{1}\theta_r^\top \widehat{\mathcal{X}}_{r-1}}\, \theta_r^\top \widehat{\mathcal{X}}_{r-1}, \;\; r \in \mathbb{N}, \;\; \widehat{\mathcal{X}}_0 = \pi\,.$$

Интегралы в $\theta(y,z)$ не могут быть вычислены аналитически, поэтому они заменяются некоторыми аппроксимациями $\psi(y,z)$, а значит, доступными являются не оценки $\widehat{\mathcal{X}}_r$, а их аппроксимации, определяемые рекурсивно:

$$\overline{\mathcal{X}}_r = \frac{1}{\mathbf{1}\psi_r^\top \overline{\mathcal{X}}_{r-1}}\, \psi_r^\top \overline{\mathcal{X}}_{r-1}, \;\; r \in \mathbb{N}\,, \;\; \widehat{\mathcal{X}}_0 = \pi\,.$$

Везде ниже предполагается, что элементы матрицы $\psi(y,z)$ неотрицательны.

Определим локальную характеристику близости оценок $\{\overline{\mathcal{X}}_r\}$ и $\{\widehat{\mathcal{X}}_r\}$ как

$$\sigma \triangleq \sup_{\pi\in\Pi} \mathsf{E}\left\{\|\widehat{\mathcal{X}}_1 - \overline{\mathcal{X}}_1\|_1\right\}.$$

Теорема 2. *Если для аппроксимации $\psi(y,z)$ выполняется неравенство*

$$\sum_{z\in\mathbb{Z}_+^K} \int\limits_{\mathbb{R}^M} \|[\psi(y,z) - \theta(y,z)]^\top\|_1\, dy \leqslant \varepsilon\,, \qquad (6)$$

то локальная характеристика точности $\sigma \leqslant 2\varepsilon$.

Доказательство теоремы 2 дано в приложении.

Утверждение о локальной точности не является асимптотическим, т. е. справедливо для любого соотношения шага дискретизации по времени Δ и точности численного интегрирования ε.

4 Заключение

Вторая часть цикла посвящена вычислительному аспекту решения задачи оптимальной фильтрации состояний МСП по имеющемуся комплексу считающих наблюдений и наблюдаемой диффузии с мультипликативными шумами. Разница с результатами [2–4] заключается именно во включении в наблюдения помимо диффузионных еще и считающих компонент.

Исходную задачу оценивания МСП в непрерывном времени предлагается заменить соответствующей задачей фильтрации состояния МСП по дискретизованным наблюдениям. При этом точность данного приближения гарантируется, во-первых, тем, что в дискретной задаче ищется оптимальная оценка, и, во-вторых, фактом сходимости дискретных аппроксимаций к исходному непрерывному пределу. Вычисление оптимальной оценки фильтрации по дискретизованным наблюдениям (см. теорему 1), однако, аналитически невозможно и требует привлечения численных методов. Теорема 2 определяет влияние ошибки той или иной схемы численного интегрирования на точность полученного приближения оценки фильтрации по дискретизованным наблюдениям. Отличие полученного результата от представленного в [2] заключается в отказе от разложения интегралов по абстрактной мере $\mathbf{p}^{kj}$ (4) в ряд конечномерных интегралов по мере Лебега. Это связано с тем, что существуют схемы численного интегрирования, не требующие такого представления, например метод квадратур Гаусса [8]. Обычно для этих схем известны оценки их точности, обеспечиваемой при вычислении интегралов по мере Лебега. Поэтому для корректного применения той или иной схемы необходимо предварительно получить характеристики точности схемы при вычислении интегралов типа (4). Помимо этого, для использования алгоритмов фильтрации состояний МСП по дискретизованным наблюдениям необходимо изучение асимптотического поведения глобальных показателей точности аппроксимации решения задачи фильтрации. Оно заключаются в том, что при убывающем шаге дискретизации Δ исследуется точность аппроксимации оценки в некоторый фиксированный момент времени t. Решения этих задач для различных схем численного интегрирования могут рассматриваться как направления дальнейших исследований.

Приложение

Доказательство теоремы 1. Воспользуемся подходом из [9]. Применим метод математической индукции. При $r = 0$

$$\widehat{\mathcal{X}}_0 = \mathsf{E}\left\{X_0|\mathcal{D}_0\right\} = \mathsf{E}\left\{x_0\right\} = \pi\,.$$

Пусть для некоторого $r \in \mathbb{N}$ известно условное распределение $\widehat{\mathcal{X}}_{r-1} = \mathsf{E}\{X_{t_{r-1}}|\mathfrak{O}_{r-1}\}$. Определим УМО $\widehat{\mathcal{X}}_r$ на следующем шаге. Для этого необходимо найти совместное распределение $(X_{t_r}, \mathcal{Y}_r, \mathcal{Z}_r)$ относительно $\mathfrak{O}_{r-1}$. Условие A гарантирует, что векторы $\mathcal{Y}_r$ и $\mathcal{Z}_r$ являются условно независимыми относительно σ-алгебры $\mathcal{F}_{t_r}^X \vee \mathfrak{O}_{r-1}$: $\mathcal{Y}_r$ имеет гауссовское распределение с параметрами

$$\mathsf{E}\left\{\mathcal{Y}_r|\mathcal{F}_{t_r}^X\right\} = f\tau_r,\ \ \mathrm{cov}\left(\mathcal{Y}_r, \mathcal{Y}_r|\mathcal{F}_{t_r}^X\right) = \sum_{n=1}^{N} \tau_r^n g_n,$$

а $\mathcal{Z}_r$ состоит из независимых компонент, имеющих пуассоновское распределение с векторным параметром $h\tau_r$.

В силу марковского свойства $\{(X_{t_r}, \mathcal{Y}_r, \mathcal{Z}_r)\}_{r\in\mathbb{Z}_+}$, формулы полной вероятности и теоремы Фубини для любых $\mathcal{A} \in \mathcal{B}(\mathbb{R}^M)$ и $z \in \mathbb{Z}_+^K$ верна следующая цепочка равенств:

$$\begin{aligned}
&\mathsf{E}\left\{X_{t_r}\mathbf{I}_{\mathcal{A}}(\mathcal{Y}_r)\mathbf{I}_{\{z\}}(\mathcal{Z}_r)\,\middle|\mathfrak{O}_{r-1}\right\} = \\
&= \mathsf{E}\left\{\mathsf{E}\left\{X_{t_r}\mathbf{I}_{\mathcal{A}}(Y_r)\mathbf{I}_{\{z\}}(\mathcal{Z}_r)\,\middle|\mathcal{F}_{t_r}^X \vee \mathfrak{O}_{r-1}\right\}\middle|\mathfrak{O}_{r-1}\right\} = \\
&= \mathsf{E}\left\{X_{t_r}\int\limits_{\mathcal{A}} \mathcal{N}\left(y, f\tau_r, \sum_{p=1}^{N}\tau_r^p g_p\right) dy\mathcal{P}(z, h\tau_r)\middle|\mathfrak{O}_{r-1}\right\} = \\
&= \mathsf{E}\left\{\mathsf{E}\left\{X_{t_r}\int\limits_{\mathcal{A}} \mathcal{N}\left(y, f\tau_r, \sum_{p=1}^{N}\tau_r^p g_p\right) dy \times\right.\right. \\
&\qquad \left.\left.\times\,\mathcal{P}(z, h\tau_r)\middle|X_{t_{r-1}} \vee \mathfrak{O}_{r-1}\right\}\middle|\mathfrak{O}_{r-1}\right\} = \\
&= \mathsf{E}\left\{\sum_{\ell=1}^{N} e_\ell \sum_{k=1}^{N} X_{t_{r-1}}^k \int\limits_{\mathcal{D}}\int\limits_{\mathcal{A}} \mathcal{N}\left(y, fu, \sum_{p=1}^{N} u^p g_p\right) dy \times\right. \\
&\qquad \left.\times\,\mathcal{P}(z, hu)\mathbf{p}^{k\ell}(du)\middle|\mathfrak{O}_{r-1}\right\} = \\
&= \sum_{\ell=1}^{N} e_\ell \int\limits_{\mathcal{A}}\left[\sum_{k=1}^{N}\widehat{\mathcal{X}}_{r-1}^k \int\limits_{\mathcal{D}} \mathcal{N}\left(y, fu, \sum_{p=1}^{N} u^p g_p\right) \times\right. \\
&\qquad \left.\times\,\mathcal{P}(z, hu)\mathbf{p}^{k\ell}(du)\right] dy\,,
\end{aligned}$$

из чего следует, что интегранд в квадратных скобках в последнем выражении определяет искомое совместное распределение $(X_{t_r}, \mathcal{Y}_r, \mathcal{Z}_r)$ относительно $\mathfrak{O}_{r-1}$. Тогда условное распределение $\widehat{\mathcal{X}}_r$ покомпонентно определяется с помощью обобщенного варианта формулы Байеса:

$$\begin{aligned}
\widehat{\mathcal{X}}_r^j &= \sum_{k=1}^{N}\widehat{\mathcal{X}}_{r-1}^k \int\limits_{\mathcal{D}} \mathcal{N}\left(\mathcal{Y}_r, fu, \sum_{p=1}^{N} u^p g_p\right) \times \\
&\times\mathcal{P}(\mathcal{Z}_r, hu)\mathbf{p}^{kj}(du)\Bigg/\left(\sum_{i,\ell=1}^{N}\widehat{\mathcal{X}}_{r-1}^i \int\limits_{\mathcal{D}}\mathcal{N}\left(\mathcal{Y}_r, fv, \sum_{q=1}^{N} v^n g_q\right)\times\right. \\
&\qquad \left.\times\,\mathcal{P}(\mathcal{Z}_r, hv)\mathbf{p}^{i\ell}(dv)\right),\ \ j = \overline{1, N}\,.
\end{aligned}$$

Теорема 1 доказана.

Доказательство теоремы 2. Для доказательства теоремы потребуется следующая вспомогательная

Лемма 1. *Если $\phi \triangleq \phi(y, z)$ — функция $\mathbb{R}^M \times \mathbb{Z}_+^K \to \mathbb{R}_+$ такая, что*

$$\sum_{z\in\mathbb{Z}_+^K}\int\limits_{\mathbb{R}^M} |\phi(y, z)|\, dy < \infty,$$

и

$$\Phi \triangleq \frac{\phi(\mathcal{Y}_1, \mathcal{Z}_1)}{\mathbf{1}\theta_1^\top\pi}\,,$$

то

$$\mathsf{E}\{\Phi\} = \sum_{z\in\mathbb{Z}_+^K}\int\limits_{\mathbb{R}^M}\phi(y, z)\, dy\,.$$

Доказательство леммы 1. Рассмотрим суммируемую функцию $\phi = \phi(y, z) : \mathbb{R}^M \times \mathbb{Z}_+^K \to \mathbb{R}$ и $\mathfrak{O}_1$-измеримую случайную величину

$$\begin{aligned}
\Phi &\triangleq \frac{\phi(\mathcal{Y}_1, \mathcal{Z}_1)}{\mathbf{1}\theta_1^\top\pi} = \\
&= \frac{\phi(\mathcal{Y}_1, \mathcal{Z}_1)}{\sum_{i,j=1}^{N}\int_{\mathcal{D}}\mathcal{N}(\mathcal{Y}_1, fu, \sum_{p=1}^{N}u^p g_p)\mathcal{P}(\mathcal{Z}_1, hu)\mathbf{p}^{ij}(du)\pi_i}\,.
\end{aligned}$$

Найдем $\mathsf{E}\{\Phi\}$:

$$\begin{aligned}
\mathsf{E}\{\Phi\} &= \sum_{z\in\mathbb{Z}_+^K}\int\limits_{\mathbb{R}^M}\int\limits_{\mathcal{D}}\phi(y, z)\sum_{k,\ell=1}^{N}\mathcal{N}\Big(y, fv, \sum\nolimits_{q=1}^{N} v^q g_q\Big) \times \\
&\times\,\mathcal{P}(z, hv)\mathbf{p}^{k\ell}(dv)\pi_k\Bigg/\left(\sum_{i,j=1}^{N}\int\limits_{\mathcal{D}}\mathcal{N}\left(y, fu, \sum_{p=1}^{N}u^p g_p\right)\times\right. \\
&\qquad \left.\times\,\mathcal{P}(z, hu)\mathbf{p}^{ij}(du)\pi_i\right) dy = \\
&= \sum_{z\in\mathbb{Z}_+^K}\int\limits_{\mathbb{R}^M}\phi(y, z)\int\limits_{\mathcal{D}}\mathcal{N}\left(y, fv, \sum_{q=1}^{N}v^q g_q\right)\mathcal{P}(z, hv)\times \\
&\times\sum_{k\ell=1}^{N}\mathbf{p}^{k\ell}(dv)\pi_k\Bigg/\left(\int\limits_{\mathcal{D}}\mathcal{N}\left(y, fu, \sum_{p=1}^{N}u^p g_p\right)\times\right. \\
&\left.\times\,\mathcal{P}(z, hu)\sum_{i,j=1}^{N}\mathbf{p}^{ij}(du)\pi_i\right) dy = \sum_{z\in\mathbb{Z}_+^K}\int\limits_{\mathbb{R}^M}\phi(y, z)\, dy\,.
\end{aligned}$$

Лемма 1 доказана.

Вернемся к доказательству теоремы 2. Обозначим $\gamma = \gamma(y, z) \triangleq \psi(y, z) - \theta(y, z)$, $\gamma_1 \triangleq \gamma(\mathcal{Y}_1, \mathcal{Z}_1)$. Заметим, что $(\gamma^\top\pi\mathbf{1} - \mathbf{1}\gamma^\top\pi I)\gamma^\top\pi = 0$, а также $\|1/(\mathbf{1}\psi(y, z)^\top\pi)\psi(y, z)^\top\pi\|_1 \equiv 1$. Используя эти равенства, можно показать, что верна следующая цепочка неравенств:

$$\|\overline{\mathcal{X}}_1 - \widehat{\mathcal{X}}_1\|_1 = \frac{1}{\mathbf{1}\theta_1^\top \pi \mathbf{1}\psi_1^\top \pi}\|\mathbf{1}\theta_1^\top \pi\psi_1^\top \pi - \mathbf{1}\psi_1^\top \pi\theta_1^\top \pi\|_1 =$$
$$= \frac{1}{\mathbf{1}\theta_1^\top \pi \mathbf{1}\psi_1^\top \pi}\|\mathbf{1}\theta_1^\top \pi\gamma_1^\top \pi - \mathbf{1}\gamma_1^\top \pi\theta_1^\top \pi\|_1 =$$
$$= \frac{1}{\mathbf{1}\theta_1^\top \pi \mathbf{1}\psi_1^\top \pi}\|(\gamma_1^\top \pi\mathbf{1} - \mathbf{1}\gamma_1^\top \pi I)\theta_1^\top \pi\|_1 =$$
$$= \frac{1}{\mathbf{1}\theta_1^\top \pi \mathbf{1}\psi_1^\top \pi}\|(\gamma_1^\top \pi\mathbf{1} - \mathbf{1}\gamma_1^\top \pi I)(\theta_1^\top \pi + \gamma_1^\top \pi)\|_1 =$$
$$= \frac{1}{\mathbf{1}\theta_1^\top \pi}\|(\gamma_1^\top \pi\mathbf{1} - \mathbf{1}\gamma_1^\top \pi I)\frac{1}{\mathbf{1}\psi_1^\top \pi}\psi_1^\top \pi\|_1 \leqslant$$
$$\leqslant \frac{1}{\mathbf{1}\theta_1^\top \pi}\|\gamma_1^\top \pi\mathbf{1} - \mathbf{1}\gamma_1^\top \pi I\|_1 \leqslant \frac{1}{\mathbf{1}\theta_1^\top \pi}(\|\gamma_1^\top\|_1 + |\mathbf{1}\gamma_1^\top \pi|) \leqslant$$
$$\leqslant 2\frac{1}{\mathbf{1}\theta_1^\top \pi}\max_{j=\overline{1,N}}\sum_i |\gamma^{ji}(\mathcal{Y}_1, \mathcal{Z}_1)| = \frac{2}{\mathbf{1}\theta_1^\top \pi}\|\gamma_1^\top\|_1.$$

Применим неравенство (6) и результат леммы 1 к левой и правой части данной цепочки неравенств:

$$\mathsf{E}\left\{\|\overline{\mathcal{X}}_1 - \widehat{\mathcal{X}}_1\|_1\right\} \leqslant 2\sum_{z\in\mathbb{Z}_+^K}\int_{\mathbb{R}^M}\|\gamma^\top(y,z)\|_1\,dy \leqslant 2\varepsilon\,.$$

Тогда $\sigma = \sup_{\pi\in\Pi}\mathsf{E}\left\{\|\overline{\mathcal{X}}_1 - \widehat{\mathcal{X}}_1\|_1\right\} \leqslant 2\varepsilon$. Теорема 2 доказана.

Литература

1. *Борисов А., Казанчян Д.* Фильтрация состояний марковских скачкообразных процессов по комплексным наблюдениям I: точное решение задачи // Информатика и её применения, 2021. Т. 15. Вып. 2. С. 12–19.
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: 10.14357/19922264190411.
3. *Борисов А.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям II: случай аддитивных шумов // Информатика и её применения, 2020. Т. 14. Вып. 1. С. 17–23. doi: 10.14357/19922264200103.
4. *Борисов А.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям III: случай мультипликативных шумов // Информатика и её применения, 2020. Т. 14. Вып. 2. С. 10–18. doi: 10.14357/19922264200103.
5. *Ishikawa Y., Kunita H.* Malliavin calculus on the Wiener–Poisson space and its application to canonical SDE with jumps // Stoch. Proc. Appl., 2006. Vol. 116. P. 1743–1769. doi: 10.1016/j.spa.2006.04.013.
6. *Borisov A., Miller G., Stefanovich A.* Controllable Markov jump processes. I. Optimum filtering based on complex observations // J. Computer Systems Sciences International, 2018. Vol. 57. No. 6. P. 890–906. doi: 10.1134/S1064230718060035.
7. *Liptser R., Shiryaev A.* Statistics of random processes I: General theory. — Berlin/Heidelberg: Springer, 2001. 427 p.
8. *Isaacson E., Keller H.* Analysis of numerical methods. — New York, NY, USA: Dover Publications, 1994. 541 p.
9. *Bertsekas D., Shreve S.* Stochastic optimal control: The discrete-time case. — Boston, MA, USA: Athena Scientific, 1978. 330 p.

Поступила в редакцию 05.03.2021

FILTERING OF MARKOV JUMP PROCESSES GIVEN COMPOSITE OBSERVATIONS II: NUMERICAL ALGORITHM

A. V. Borisov[1,2,3] and D. Kh. Kazanchyan[4]

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

[2]Moscow Aviation Institute (National Research University), 4 Volokolamskoe Shosse, Moscow 125080, Russian Federation

[3]Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation

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

Abstract: The note represents the second, final part of the series initiated by the article Borisov, A., and D. Kazanchyan. 2021. Filtering of Markov jump processes given composite observations I: Exact solution. *Informatika i ee primeneniya — Inform. Appl.* 15(2):12–19. The authors propose a new numerical algorithm of the optimal state estimation for the Markov jump processes given observable both the counting processes and the diffusion ones with the multiplicative noises. The authors approximate the initial continuous-time estimation problem by a sequence of the corresponding filtering problems given the time-discretized observations. The paper contains the explicit recursive form of the discretized estimate and introduces its one-step precision characteristic along with dependence of the characteristics on the utilized numerical estimation scheme.

Keywords: Markov jump process; optimal filtering; multiplicative observation noises; time-discretized observations; approximation precision

DOI: 10.14357/19922264210302

Acknowledgments

The work was supported in part by the Russian Foundation for Basic Research (project 19-07-00187 A). The research was conducted in accordance with the program of the Moscow Center for Fundamental and Applied Mathematics.

References

1. Borisov, A., and D. Kazanchyan. 2021. Fil'tratsiya sostoyaniy markovskikh skachkoobraznykh protsessov po kompleksnym nablyudeniyam I: tochnoe reshenie zadachi [Filtering of Markov jump processes given composite observations I: Exact solution]. *Informatika i ee Primeneniya — Inform. Appl.* 15(2):12–19.
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. Borisov, A. 2020. Chislennye skhemy fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam II: sluchay additivnykh shumov [Numerical schemes of Markov jump process filtering given discretized observations II: Additive noises case]. *Informatika i ee primeneniya — Inform. Appl.* 14(1):17–23. doi: 10.14357/19922264200103.
4. Borisov, A. 2020. Chislennye skhe my fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam III: sluchay mul'tiplikativnykh shumov [Numerical schemes of Markov jump process filtering given discretized observations III: Multiplicative noises case]. *Informatika i ee primeneniya — Inform. Appl.* 14(2):10–18. doi: 10.14357/19922264200103.
5. Ishikawa, Y., and H. Kunita. 2006. Malliavin calculus on the Wiener–Poisson space and its application to canonical SDE with jumps. *Stoch. Proc. Appl.* 116:1743–1769. doi: 10.1016/j.spa.2006.04.013.
6. Borisov, A., G. Miller, and A. Stefanovich. 2018. Controllable Markov jump processes. I. Optimum filtering based on complex observations. *J. Computer Systems Sciences International*. 57(6):890–906. doi: 10.1134/S1064230718060035.
7. Liptser, R., and A. Shiryaev. 2001. *Statistics of random processes I: General theory.* Berlin/Heidelberg: Springer. 427 p.
8. Isaacson, E., and H. Keller. 1994. *Analysis of numerical methods.* New York, NY: Dover Publications. 541 p.
9. Bertsekas, D., and S. Shreve. 1978. *Stochastic optimal control: The discrete-time case.* Boston, MA: Athena Scientific. 330 p.

Received March 5, 2021

Contributors

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; professor, Moscow Aviation Institute (National Research University), 4 Volokolamskoe Shosse, Moscow 125080, Russian Federation; senior scientist, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; aborisov@frccsc.ru

Kazanchyan Drastamat Kh. (b. 1994) — PhD student, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; drastamat94@gmail.com

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 16–23

АЛГОРИТМЫ СЖАТИЯ ДАННЫХ МАССИВОВ СИЛОВЫХ КРИВЫХ II: КОДИРОВАНИЕ КОМПОНЕНТ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ

Д. В. Сушко[1]

Аннотация: Представлена вторая часть исследования задачи обратимого (без потерь) сжатия данных массивов силовых кривых — трехмерных массивов, элементы которых суть 16-битные целые числа. Предложены алгоритмы сжатия массивов силовых кривых, основанные на универсальном арифметическом кодировании компонент, получаемых в результате применения одномерного дискретного вейвлет-преобразования (ДВП) по системе вейвлетов (5–3) к строкам массивов. Преобразование реализуется в рамках лифтинг-схемы и является обратимым. Для построения эффективных алгоритмов использован метод повторного применения вейвлет-преобразования и два уже апробированных в первой части работы метода универсального кодирования (разложение на вычислимые состояния, выбор веса при построении кодовых вероятностей). Для предложенных алгоритмов на пяти тестовых массивах построены оценки скорости кодирования. Результаты показывают, что каждый из упомянутых выше методов позволяет уменьшить скорость кодирования, а комбинация всех трех методов дает наиболее эффективный алгоритм. Скорость кодирования тестовых массивов этим алгоритмом составляет 3,8160, 3,4050, 3,3678, 4,1309 и 4,0996 бит/пиксель, а выигрыш по сравнению с алгоритмом обратимого сжатия стандарта JPEG 2000 составляет 6%–9%.

Ключевые слова: атомно-силовой микроскоп; массив силовых кривых; обратимое сжатие; арифметическое кодирование; универсальное кодирование

DOI: 10.14357/19922264210303

1 Введение

Рассмотрение задачи обратимого сжатия данных массивов силовых кривых было начато в работе [1], где были определены потенциальные возможности некоторых алгоритмов сжатия. В число таких алгоритмов вошли стандартные алгоритмы (DEFLATE, JPEG 2000) и ряд простых алгоритмов арифметического кодирования. В работе [2] были построены и исследованы более сложные алгоритмы, основанные на универсальном арифметическом кодировании ошибок предсказания. Данная работа является продолжением работы [2] и имеет целью построить и исследовать алгоритмы, основанные на универсальном арифметическом кодировании компонент ДВП.

В качестве основной величины, характеризующей эффективность алгоритма сжатия, используется *скорость кодирования* R, определяемая как отношение длины кодового слова L (в битах), порождаемого алгоритмом для описания массива данных, к полному числу элементов (пикселей) N этого массива; единица измерения скорости кодирования — бит/пиксель (бт/п). Коэффициент сжатия равен отношению длины элемента массива в битах к скорости кодирования. Скорость кодирования рассматриваемых алгоритмов оценивается на массивах, полученных при сканировании мягких биологических образцов в режиме измерения силовых карт на микроскопе MultiMode V (Veeco, США). Те же экспериментальные данные были использованы в работах [1, 2], что позволяет непосредственно сравнивать результаты. Вычисления проводятся программами, написанными на языке Python.

2 Предварительные сведения

Данный раздел содержит краткое изложение фактов, положений и утверждений работы [2], используемых в настоящей работе.

Массив силовых кривых представляет собой трехмерный массив размерами $(I, J, 2K)$:

$$\mathbf{V} = \{V(i,j,k)\},\ \ i = 0,1,\ldots,I-1,\\ j = 0,1,\ldots,J-1,\ \ k = 0,1,\ldots,2K-1\,,$$

элементы которого суть 16-битные целые числа, т. е. целые числа в диапазоне $[-2^{15}, 2^{15}-1]$. Ин-

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, dsushko@ipiran.ru

дексы (i,j) нумеруют узлы равномерной решетки в поле наблюдения — прямоугольной области в горизонтальной плоскости OXY, совпадающей с плоскостью подложки образца. Строки $F^{\mathrm{A}}_{(i,j)}(k) = V(i,j,k)$ и $F^{\mathrm{R}}_{(i,j)}(k) = V(i,j,K+k)$, i,j фиксированы, $k = 0,1,\ldots,K-1$, представляют собой силовые кривые подвода и отвода в узле (i,j). Значения $F^{\mathrm{A,R}}_{(i,j)}(k)$ пропорциональны силе, действующей на сканирующий зонд в точке пространства, находящейся на расстоянии $k\Delta z$ по вертикали (вдоль оси OZ) от поверхности образца в узле (i,j), при подводе зонда к образцу и отводе зонда от образца; Δz — шаг по вертикали. Для некоторых узлов (i,j) требуемое число значений (K) кривой подвода $F^{\mathrm{A}}_{(i,j)}(k)$ не может быть измерено. В таком случае осуществляется дополнение соответствующих строк до требуемой длины минимально возможным значением -2^{15}, записываемым в конец строки.

Тестовыми экспериментальными данными являются пять массивов силовых кривых (I–V), полученных при сканировании образцов, представляющих собой абсорбированные из раствора на твердую подложку вирусы. Первый образец (I) — это риновирус 2 на подложке из слюды, второй и третий образцы (II и III) — вирус мягкой мозаики ячменя на подложке из слюды, четвертый и пятый образцы (IV и V) — вирус табачной мозаики на подложке из стекла. Массивы силовых кривых имеют следующие размеры: $I = J = 64$, $K = 4096$ (массив I); $I = J = 128$, $K = 1024$ (массивы II–V). Полное число элементов всех массивов равно 2^{25}. Характерный вид силовых кривых показан на рисунке в работе [2].

Традиционный метод решения задач обратимого сжатия цифровых данных заключается в применении к исходному массиву данных некоторых обратимых преобразований, обеспечивающих декорреляцию его отсчетов и/или уменьшение диапазона их значений, и последующем арифметическом кодировании [3] полученного таким образом массива как последовательности независимых отсчетов.

При арифметическом кодировании конечной числовой последовательности $\mathbf{x} = \{x_n\}$, $n = 0,1,\ldots,N-1$, принимающей значения в априори известном диапазоне $\mathfrak{A}$, множество *условных кодовых распределений вероятностей* (или просто *кодовых распределений*) $\{q_n(a) = q_n(a|x_{n-1},\ldots,x_0), a \in \mathfrak{A}\}$ используется для того, чтобы приписать последовательности $\mathbf{x}$ кодовую вероятность $Q(\mathbf{x})$ и кодовое слово (результат сжатия). Построение кодовых распределений, обеспечивающих получение возможно более коротких кодовых слов при неизвестной статистике, — задача универсального кодирования [4]. Восстановление исходной последовательности по кодовому слову осуществляется в процессе декодирования без задержки. В момент восстановления очередного значения x_n декодеру уже известны все предыдущие значения $\{x_0,\ldots,x_{n-1}\}$, и кодовые распределения могут быть построены декодером так же, как они были ранее построены кодером в процессе кодирования. Это позволяет декодеру восстановить значение x_n.

Кодовые распределения строятся по формуле:

$$q_n\left(a|x_{n-1},\ldots,x_0\right) = \begin{cases} \dfrac{1+w\theta_n(a)}{a_{\max}(\mathbf{x}) - a_{\min}(\mathbf{x}) + 1 + wn} \\ \qquad\qquad \text{при } a \in [a_{\min}(\mathbf{x}), a_{\max}(\mathbf{x})]\,; \\ 0 \qquad \text{в противном случае}\,, \end{cases} \tag{1}$$

где $\theta_n(a)$ — число элементов, принимающих значение a, на начальном участке последовательности $\mathbf{x}$ до $(n-1)$-го члена включительно; $a_{\min}(\mathbf{x})$ и $a_{\max}(\mathbf{x})$ — нижняя и верхняя границы диапазона значений последовательности $\mathbf{x}$; параметр $w = 1,2,\ldots$ — вес. Границы $a_{\min}(\mathbf{x})$ и $a_{\max}(\mathbf{x})$ могут быть вычислены кодером и должны быть переданы декодеру помимо кодового слова, что, вообще говоря, слегка увеличивает скорость кодирования. Однако для массивов силовых кривых, которые содержат 2^{25} элементов, этим можно пренебречь.

Скорость арифметического кодирования оценивается по формуле:

$$R(\mathbf{x}) = \frac{1}{N}\sum_{n=0}^{N-1} -\log_2 q_n(x_n)\,. \tag{2}$$

Величины $\theta(x)/N$, $x \in \mathfrak{A}$, где $\theta(x)$ — число элементов, принимающих значение x, в последовательности $\mathbf{x}$, образуют *частотное* (или *эмпирическое*) распределение вероятностей значений последовательности. Величина

$$H(\mathbf{x}) = \sum_{x\in\mathfrak{A}} \frac{\theta(x)}{N}\left[-\log_2\frac{\theta(x)}{N}\right] \tag{3}$$

(используется соглашение о том, что $0\log_2 0 = 0$) называется *квазиэнтропией* последовательности. Единица измерения квазиэнтропии — бт/п. Квазиэнтропия зависит только от самой последовательности и представляет собой нижнюю границу скорости арифметического кодирования (см., например, [4]). Разность $R - H \geq 0$ — *избыточность* арифметического кодирования. Величина избыточности характеризует качество решения одной из

задач универсального кодирования — задачи построения кодовых распределений.

Если последовательность $\mathbf{x}$ разложена на две подпоследовательности $\mathbf{x}_0$ и $\mathbf{x}_1$ с числом элементов N_0 и N_1, то квазиэнтропия пары подпоследовательностей и общая скорость их независимого арифметического кодирования равны

$$H(\{\mathbf{x}_0, \mathbf{x}_1\}) = \frac{N_0}{N} H(\mathbf{x}_0) + \frac{N_1}{N} H(\mathbf{x}_1),$$
$$R(\{\mathbf{x}_0, \mathbf{x}_1\}) = \frac{N_0}{N} R(\mathbf{x}_0) + \frac{N_1}{N} R(\mathbf{x}_1). \quad (4)$$

Арифметическое кодирование трехмерного массива данных требует их одномерного упорядочения. В работе принято так называемое строчное упорядочение, при котором трехмерный массив $\mathbf{X}$ размерами (I, J, K) превращается в последовательность $\mathbf{x}$ размером $N = IJK$ так, что $x(n) = X(i, j, k)$, $n = JKi + Kj + k$.

Полученные в [1] результаты показывают, что эффективность алгоритмов сжатия возрастает при использовании двух предварительных преобразований. Первое преобразование заключается в разделении массива силовых кривых $\mathbf{V}$ на массивы кривых подвода $\mathbf{V}^{\mathrm{A}}$ и отвода $\mathbf{V}^{\mathrm{R}}$: $\mathbf{V} \to \{\mathbf{V}^{\mathrm{A}}, \mathbf{V}^{\mathrm{R}}\}$,

$$V^{\mathrm{A}}(i, j, k) = V(i, j, k), \quad V^{\mathrm{R}}(i, j, k) = V(i, j, K + k),$$
$$k = 0, 1, \ldots, K - 1\,, \quad (5)$$

и последующую обработку полученных массивов по отдельности. Второе преобразование применяется к полученному массиву кривых подвода $\mathbf{V}^{\mathrm{A}}$ и заключается в сужении диапазона значений этого массива, непомерно широкого из-за наличия значения -2^{15}, используемого для дополнения «неполных» строк: $\mathbf{V}^{\mathrm{A}} \to \bar{\mathbf{V}}^{\mathrm{A}}$,

$$\bar{V}^{\mathrm{A}} = \begin{cases} V^{\mathrm{A}}_{d\min} - 1, & V^{\mathrm{A}} = -2^{15}; \\ V^{\mathrm{A}}, & V^{\mathrm{A}} > -2^{15}, \end{cases} \quad (6)$$

где $V^{\mathrm{A}}_{d\min}$ — динамический минимум значений массива $\mathbf{V}^{\mathrm{A}}$ (т. е. минимум значений без учета значения -2^{15}). Чтобы обратить преобразование сужения диапазона (6), декодеру нужна информация о том, встречалось ли значение -2^{15} в исходном массиве кривых подвода. Для передачи декодеру этой известной кодеру информации требуется один бит, и соответствующее увеличение скорости кодирования пренебрежимо мало.

Одним из преобразований, обеспечивающих декорреляцию отсчетов данных, является переход к ошибкам предсказания следующего вида: $\mathbf{X} \to \mathbf{D}$,

$$D(i, j, k) =$$
$$= \begin{cases} X(0,0,0), & i = j = k = 0; \\ X(i,0,0) - X(i-1,0,0), & i = 1, \ldots, I-1, \\ & j = k = 0; \\ X(i,j,0) - X(i,j-1,0), & i = 1, \ldots, I-1, \\ & j = 1, \ldots, J-1, \ k = 0; \\ X(i,j,k) - X(i,j,k-1), & i = 1, \ldots, I-1, \\ & j = 1, \ldots, J-1, \ k = 1, \ldots, K-1. \end{cases} \quad (7)$$

3 Алгоритмы кодирования компонент вейвлет-преобразования

Обратимым ДВП конечной целочисленной последовательности является ДВП по системе вейвлетов (5–3) [5] при его реализации с помощью лифтинг-схемы [6]. Преобразование входит в стандарт сжатия JPEG 2000. Вид как прямого, так и обратного преобразования представлен, например, в [7].

Результат применения ДВП к данной последовательности $\mathbf{x}$ длины N (N — четное) представляет собой разложение этой последовательности на две последовательности (компоненты) длины $N/2$ каждая: $\mathbf{x} \to \{\mathbf{x}^0, \mathbf{x}^1\}$. Соответствующие формулы приведены в [1]. Компонента $\mathbf{x}^0$ — приближение последовательности $\mathbf{x}$ вдвое меньшего разрешения; компонента $\mathbf{x}^1$ — детальная составляющая. Применение ДВП для сжатия данных обусловлено тем, что корреляция элементов детальной составляющей существенно меньше, чем корреляция элементов исходного сигнала, а ее значения распределены более неравномерно, чем исходные.

Обратимые ДВП многомерных массивов строятся на основе описанного выше ДВП последовательностей. Рассмотрим, что представляет собой одномерное ДВП трехмерного массива по строкам (по третьему измерению), используемое в работе в качестве декорреляционного преобразования. Пусть $\mathbf{X} = \{X(i, j, k)\}$, $i = 0, 1, \ldots, I - 1$, $j = 0, 1, \ldots, J - 1$, $k = 0, 1, \ldots, K - 1$ (K — четное) — трехмерный массив. При фиксированных значениях i и j элементы массива образуют последовательность $\mathbf{x}_{(i,j)} = \{x_{(i,j)}(k)\} \doteq \{X(i, j, k)\}$. Применим ДВП ко всем таким последовательностям:

$$\mathbf{x}_{(i,j)} \to \{\mathbf{x}^0{}_{(i,j)}, \mathbf{x}^1{}_{(i,j)}\};$$
$$\mathbf{x}^{0,1}{}_{(i,j)} = \{x^{0,1}{}_{(i,j)}(k)\}, \quad k = 0, 1, \ldots, \frac{K}{2} - 1\,.$$

Трехмерные массивы $\mathbf{X}^{0,1} = \{X^{0,1}(i, j, k)\} \doteq \doteq \{x^{0,1}{}_{(i,j)}(k)\}$, $i = 0, 1, \ldots, I - 1$, $j = 0, 1, \ldots, J - 1$,

$k = 0, 1, \ldots, K/2-1$ — искомый результат одномерного ДВП: $\mathbf{X} \to \{\mathbf{X}^0, \mathbf{X}^1\}$, массив $\mathbf{X}^0$ — приближение, массив $\mathbf{X}^1$ — детальная составляющая.

Из результатов [1] следует, что самый эффективный способ применения ДВП для сжатия массива силовых кривых $\mathbf{V}$ состоит в применении одномерного ДВП к строкам (т. е. по третьему измерению) массивов $\bar{\mathbf{V}}^{\mathrm{A}}$ и $\mathbf{V}^{\mathrm{R}}$, полученных в результате преобразований (5) и (6). Все рассматриваемые в работе алгоритмы используют указанные преобразования.

Начнем с рассмотрения алгоритма A[1|1|0], который обеспечил наименьшую скорость кодирования среди всех исследованных в [1] алгоритмов. Алгоритм предполагает разложение массивов $\bar{\mathbf{V}}^{\mathrm{A}}$ и $\mathbf{V}^{\mathrm{R}}$ посредством ДВП на компоненты $\bar{\mathbf{V}}^{\mathrm{A}} \to \{\mathbf{W}^{\mathrm{A}\,0}, \mathbf{W}^{\mathrm{A}\,1}\}$ и $\mathbf{V}^{\mathrm{R}} \to \{\mathbf{W}^{\mathrm{R}\,0}, \mathbf{W}^{\mathrm{R}\,1}\}$, переход к ошибкам предсказания (7) для приближений $\mathbf{W}^{\mathrm{A}\,0} \to \mathbf{D}^{\mathrm{A}\,0}$ и $\mathbf{W}^{\mathrm{R}\,0} \to \mathbf{D}^{\mathrm{R}\,0}$ и арифметическое кодирование полученных компонент $\mathbf{D}^{\mathrm{A}\,0}$, $\mathbf{W}^{\mathrm{A}\,1}$, $\mathbf{D}^{\mathrm{R}\,0}$ и $\mathbf{W}^{\mathrm{R}\,1}$ с использованием построенных по формуле (1) с весом $w = 2$ кодовых распределений. Размер кодируемых компонент одинаков, поэтому квазиэнтропия $H[1|1]$ и скорость кодирования $R[1|1|0]$ совокупности этих четырех компонент равны средним значениям квазиэнтропии (3) и скорости кодирования (2) компонент по отдельности. Значения квазиэнтропии $H[1|1]$ для массивов I–V представлены в строке 1 табл. 1, а значения скорости кодирования $R[1|1|0]$ — в строке 7. Значения в таблице приводятся в единицах бт/п с точностью до четырех знаков после десятичной запятой. Избыточность кодирования составляет 0,0013–0,0028 бт/п в зависимости от массива.

Рассмотрим метод сжатия, основанный на двукратном применении ДВП. Обозначим через A[2|1|0] алгоритм, отличающийся от предыдущего повторным применением ДВП к приближениям, полученным после первого ДВП:

$$\mathbf{W}^{\mathrm{A}\,0} \to \{\mathbf{W}^{\mathrm{A}\,00}, \mathbf{W}^{\mathrm{A}\,01}\};\ \ \mathbf{W}^{\mathrm{R}\,0} \to \{\mathbf{W}^{\mathrm{R}\,00}, \mathbf{W}^{\mathrm{R}\,01}\}.$$

Переход к ошибкам предсказания (7) осуществляется для приближений второго ДВП $\mathbf{W}^{\mathrm{A}\,00} \to \mathbf{D}^{\mathrm{A}\,00}$ и $\mathbf{W}^{\mathrm{R}\,00} \to \mathbf{D}^{\mathrm{R}\,00}$, после чего выполняется независимое арифметическое кодирование полученных компонент $\mathbf{D}^{\mathrm{A}\,00}$, $\mathbf{W}^{\mathrm{A}\,01}$, $\mathbf{W}^{\mathrm{A}\,1}$, $\mathbf{D}^{\mathrm{R}\,00}$, $\mathbf{W}^{\mathrm{R}\,01}$ и $\mathbf{W}^{\mathrm{R}\,1}$ с использованием построенных по формуле (1) с весом $w = 2$ кодовых распределений. Поскольку каждое применение ДВП уменьшает размер массива вдвое, квазиэнтропия $H[2|1]$ совокупности указанных шести компонент равна

$$H[2|1] = \\ = \frac{H(\mathbf{D}^{\mathrm{A}\,00}) + H(\mathbf{W}^{\mathrm{A}\,01}) + H(\mathbf{D}^{\mathrm{R}\,00}) + H(\mathbf{W}^{\mathrm{R}\,01})}{8} + \\ + \frac{H(\mathbf{W}^{\mathrm{A}\,1}) + H(\mathbf{W}^{\mathrm{R}\,1})}{4}$$

и такая же формула (с заменой H на R) имеет место для скорости кодирования $R[2|1|0]$ совокупности компонент. Квазиэнтропия и скорость кодирования каждой отдельной компоненты вычисляются по формулам (3) и (2).

Значения квазиэнтропии $H[2|1]$ для массивов I–V представлены в строке 2 табл. 1, а значения скорости кодирования $R[2|1|0]$ — в строке 10. Среднее по массивам уменьшение квазиэнтропии $H[2|1]$

Таблица 1 Квазиэнтропия и скорость кодирования

№	Величина	I	II	III	IV	V
1	$H[1\|1]$	3,8701	3,4364	3,4069	4,2198	4,1724
2	$H[2\|1]$	3,8464	3,4131	3,3784	4,1912	4,1501
3	$H[1\|2,\mathrm{opt}]$	3,8519	3,4125	3,3766	4,1783	4,1339
4	$H[1\|2,\mathrm{fix}]$	3,8535	3,4127	3,3768	4,1793	4,1344
5	$H[2\|2,\mathrm{opt}]$	3,8144	3,4026	3,3649	4,1271	4,0963
6	$H[2\|2,\mathrm{fix}]$	3,8149	3,4031	3,3653	4,1289	4,0976
7	$R[1\|1\|0]$	3,8714	3,4392	3,4096	4,2217	4,1742
8	$R[1\|1\|\mathrm{opt}]$	3,8706	3,4374	3,4081	4,2207	4,1733
9	$R[1\|1\|\mathrm{fix}]$	3,8706	3,4374	3,4081	4,2207	4,1733
10	$R[2\|1\|0]$	3,8481	3,4164	3,3824	4,1938	4,1528
11	$R[2\|1\|\mathrm{opt}]$	3,8471	3,4145	3,3804	4,1925	4,1515
12	$R[2\|1\|\mathrm{fix}]$	3,8471	3,4145	3,3804	4,1925	4,1515
13	$R[1\|2,\mathrm{fix}\|0]$	3,8556	3,4159	3,3804	4,1820	4,1373
14	$R[1\|2,\mathrm{fix}\|\mathrm{opt}]$	3,8541	3,4138	3,3783	4,1805	4,1357
15	$R[1\|2,\mathrm{fix}\|\mathrm{fix}]$	3,8542	3,4139	3,3783	4,1805	4,1357
16	$R[2\|2,\mathrm{fix}\|0]$	3,8176	3,4074	3,3709	4,1330	4,1021
17	$R[2\|2,\mathrm{fix}\|\mathrm{opt}]$	3,8159	3,4049	3,3677	4,1309	4,0996
18	$R[2\|2,\mathrm{fix}\|\mathrm{fix}]$	3,8160	3,4050	3,3678	4,1309	4,0996

Таблица 2 Оптимальные и фиксированные пороги

Компонента	I	II	III	IV	V	I–V	Компонента	I	II	III	IV	V	I–V
$\mathbf{D}^{\mathrm{A}\,0}$	8	9	9	10	10	9	$\mathbf{D}^{\mathrm{R}\,0}$	8	9	9	10	10	9
$\mathbf{W}^{\mathrm{A}\,1}$	3	2	2	3	3	3	$\mathbf{W}^{\mathrm{R}\,1}$	3	2	4	3	3	3
$\mathbf{D}^{\mathrm{A}\,00}$	40	42	19	17	17	18	$\mathbf{D}^{\mathrm{R}\,00}$	10	18	18	16	16	17
$\mathbf{W}^{\mathrm{A}\,01}$	5	5	5	6	6	5	$\mathbf{W}^{\mathrm{R}\,01}$	5	4	4	6	6	5

по сравнению с $H[1|1]$ составляет 0,0253 бт/п, минимальное — 0,0223 бт/п (массив V), максимальное — 0,0287 бт/п (массив IV). Такого же порядка уменьшение скорости кодирования $R[2|1|0]$ по сравнению с $R[1|1|0]$: среднее по массивам — 0,0245 бт/п, минимальное — 0,0214 бт/п (массив V), максимальное — 0,0279 бт/п (массив IV). Избыточность кодирования алгоритма A[2|1|0] составляет 0,0017–0,0040 бт/п в зависимости от массива и в 1,2–1,5 раза больше избыточности кодирования алгоритма A[1|1|0].

Применим метод сжатия, основанный на статистической модели *источника с вычислимой последовательностью состояний*, по схеме, уже использованной в [2]. Выберем некоторое натуральное число t — порог. Разложим подлежащую кодированию последовательность $\mathbf{x}$ (в роли которой может выступать любая из рассмотренных выше одномерно упорядоченных компонент) на две подпоследовательности $\mathbf{x}_0$ и $\mathbf{x}_1$, называемые элементами нулевого и первого состояний. К нулевому состоянию отнесем те элементы x_n, для которых $n \geq 1$ и $|x_{n-1}| < t$, прочие элементы отнесем к первому состоянию. Элементы $\mathbf{x}_0$ и $\mathbf{x}_1$ состояний будем кодировать/декодировать независимо. Это возможно, поскольку в момент обработки текущего элемента значение предыдущего элемента известно как кодеру, так и декодеру (декодирование осуществляется без задержки), поэтому известно, к какому состоянию принадлежит текущий элемент. Совокупная квазиэнтропия пары состояний и совместная скорость их независимого арифметического кодирования даются формулами (4) и зависят от выбора порога.

При любом значении порога t квазиэнтропия пары состояний $H(\{\mathbf{x}_0, \mathbf{x}_1\})$ не превышает квазиэнтропии всей последовательности $H(\mathbf{x})$ (см., например, [4]). Оптимизационная задача нахождения порога, при котором квазиэнтропия пары состояний принимает минимальное значение, может быть решена для конкретных данных путем перебора. Результаты численного решения этой задачи для всех подлежащих кодированию компонент каждого из массивов I–V приведены в колонках табл. 2, обозначенных I,. . .,V.

Применим метод разложения на состояния к фигурирующим в алгоритме A[1|1|0] компонентам $\mathbf{D}^{\mathrm{A}\,0}$, $\mathbf{W}^{\mathrm{A}\,1}$, $\mathbf{D}^{\mathrm{R}\,0}$ и $\mathbf{W}^{\mathrm{R}\,1}$, используя оптимальные пороги. Квазиэнтропия H[1|2,opt] совокупности полученных компонент $\{\mathbf{D}^{\mathrm{A}\,0}{}_0, \mathbf{D}^{\mathrm{A}\,0}{}_1, \dots\}$ для массивов I–V представлена в строке 3 табл. 1. Среднее по массивам уменьшение квазиэнтропии по сравнению с $H[1|1]$ составляет 0,0305 бт/п, минимальное — 0,0182 бт/п (массив I), максимальное — 0,0415 бт/п (массив IV).

Нахождение оптимальных порогов требует значительного времени счета и не может быть реализовано на этапе кодирования в режиме реального времени. Поэтому для построения состояний в алгоритме, предназначенном для практического применения, следует использовать заранее выбранные фиксированные значения порогов, общие для всей совокупности сжимаемых массивов. Такие общие значения порогов для подлежащих кодированию компонент массивов I–V представлены в колонках табл. 2, обозначенных I–V.

В строке 4 табл. 1 представлены значения совокупной квазиэнтропии $H[1|2,\mathrm{fix}]$, отвечающие построенным с фиксированными значениями порогов состояниям для компонент $\mathbf{D}^{\mathrm{A}\,0}$, $\mathbf{W}^{\mathrm{A}\,1}$, $\mathbf{D}^{\mathrm{R}\,0}$ и $\mathbf{W}^{\mathrm{R}\,1}$ массивов I–V. Увеличение квазиэнтропии по сравнению с оптимальным значением $H[1|2,\mathrm{opt}]$ составляет 0,0002–0,0016 бт/п в зависимости от массива. Таким образом, использование общих фиксированных порогов не приводит к значимому увеличению квазиэнтропии по сравнению с оптимальными значениями.

В строке 13 табл. 1 представлены скорости кодирования $R[1|2,\mathrm{fix}|0]$ массивов I–V алгоритмом A[1|2,fix|0], который независимо кодирует элементы состояний компонент $\mathbf{D}^{\mathrm{A}\,0}$, $\mathbf{W}^{\mathrm{A}\,1}$, $\mathbf{D}^{\mathrm{R}\,0}$ и $\mathbf{W}^{\mathrm{R}\,1}$, построенных с фиксированными порогами, и использует формулу (1) с весом $w = 2$ для построения кодовых распределений. Для всех массивов наблюдается уменьшение скорости кодирования по сравнению с алгоритмом A[1|1|0], которое лишь немного меньше, чем уменьшение квазиэнтропии $H[1|2,\mathrm{fix}]$ по сравнению с $H[1|1]$. Среднее по массивам уменьшение составляет 0,0290 бт/п, минимальное — 0,0158 бт/п (массив I), максимальное —

Таблица 3 Оптимальные и фиксированные веса

Последовательность	I	II	III	IV	V	I–V	Последовательность	I	II	III	IV	V	I–V
$\mathbf{D}^{\mathrm{A}0}$	24	18	11	9	10	10	$\mathbf{D}^{\mathrm{R}0}$	50	235	59	51	30	50
$\mathbf{W}^{\mathrm{A}1}$	20	13	17	14	16	15	$\mathbf{W}^{\mathrm{R}1}$	7	8	8	22	7	8
$\mathbf{D}^{\mathrm{A}00}$	22	16	11	10	10	10	$\mathbf{D}^{\mathrm{R}00}$	44	159	43	32	22	35
$\mathbf{W}^{\mathrm{A}01}$	16	10	13	11	12	12	$\mathbf{W}^{\mathrm{R}01}$	5	10	8	11	9	9
$\mathbf{D}^{\mathrm{A}0}{}_0$	99	13	12	12	13	20	$\mathbf{D}^{\mathrm{R}0}{}_0$	27	37	18	29	42	20
$\mathbf{D}^{\mathrm{A}0}{}_1$	19	14	9	8	9	10	$\mathbf{D}^{\mathrm{R}0}{}_1$	54	246	62	49	23	50
$\mathbf{W}^{\mathrm{A}1}{}_0$	22	13	17	18	21	18	$\mathbf{W}^{\mathrm{R}1}{}_0$	7	8	8	17	6	8
$\mathbf{W}^{\mathrm{A}1}{}_1$	38	174	258	20	25	25	$\mathbf{W}^{\mathrm{R}1}{}_1$	7	6	5	21	7	7
$\mathbf{D}^{\mathrm{A}00}{}_0$	40	19	28	20	26	25	$\mathbf{D}^{\mathrm{R}00}{}_0$	24	29	12	15	33	15
$\mathbf{D}^{\mathrm{A}00}{}_1$	16	12	9	8	8	10	$\mathbf{D}^{\mathrm{R}00}{}_1$	47	153	41	27	19	40
$\mathbf{W}^{\mathrm{A}01}{}_0$	24	12	15	14	16	15	$\mathbf{W}^{\mathrm{R}01}{}_0$	5	7	8	8	9	7
$\mathbf{W}^{\mathrm{A}01}{}_1$	17	10	13	12	13	13	$\mathbf{W}^{\mathrm{R}01}{}_1$	5	9	6	11	9	8

0,0397 бт/п (массив IV). Избыточность кодирования A[1|2,fix|0] составляет 0,0022–0,0036 бт/п и в 1,2–1,7 раза больше избыточности алгоритма A[1|1|0].

Применим теперь метод разложения на состояния к компонентам, получаемым в результате двукратного применения ДВП, т. е. к компонентам $\mathbf{D}^{\mathrm{A}00}$, $\mathbf{W}^{\mathrm{A}01}$, $\mathbf{W}^{\mathrm{A}1}$, $\mathbf{D}^{\mathrm{R}00}$, $\mathbf{W}^{\mathrm{R}01}$ и $\mathbf{W}^{\mathrm{R}1}$, фигурирующим в алгоритме A[2|1|0]. В строках 5 и 6 табл. 1 представлены значения совокупной квазиэнтропии H[2|2,opt] и H[2|2,fix], отвечающие построенным с оптимальными и фиксированными значениями порогов состояниям компонент массивов I–V. Среднее по массивам уменьшение квазиэнтропии H[2|2,opt] по сравнению с H[2|1] составляет 0,0348 бт/п, минимальное — 0,0104 бт/п (массив II), максимальное — 0,0640 бт/п (массив IV). Разность значений H[2|2,fix] и H[2|2,opt] невелика, она составляет 0,0004–0,0017 бт/п в зависимости от массива. Можно снова констатировать, что замена оптимальных порогов на фиксированные не приводит к значимому увеличению квазиэнтропии.

В строке 16 табл. 1 представлены скорости кодирования R[2|2,fix|0] массивов I–V алгоритмом A[2|2,fix|0], который независимо кодирует элементы состояний компонент $\mathbf{D}^{\mathrm{A}00}$, $\mathbf{W}^{\mathrm{A}01}$, $\mathbf{W}^{\mathrm{A}1}$, $\mathbf{D}^{\mathrm{R}00}$, $\mathbf{W}^{\mathrm{R}01}$ и $\mathbf{W}^{\mathrm{R}1}$, построенных с фиксированными порогами, и использует формулу (1) с весом $w = 2$ для построения кодовых распределений. Наблюдается уменьшение скорости кодирования по сравнению с алгоритмом A[2|1|0], которое лишь немного меньше, чем уменьшение квазиэнтропии H[2|2,fix] по сравнению с H[2|1]. Среднее по массивам уменьшение составляет 0,0325 бт/п, минимальное — 0,0090 бт/п (массив II), максимальное — 0,0608 бт/п (массив IV). Избыточность кодирования A[2|2,fix|0] составляет 0,0027–0,0056 бт/п и в 1,3–1,7 раза больше избыточности кодирования алгоритма A[2|1|0].

Приведенные результаты показывают, что с увеличением числа независимо кодируемых последовательностей (компонент и их состояний) избыточность кодирования несколько возрастает. Поэтому, как и в [2], рассмотрим метод уменьшения избыточности кодирования, основанный на выборе веса w в формуле (1) для кодовых распределений. Оптимизационная задача нахождения веса w, при котором для конкретной последовательности данных скорость кодирования (2) минимальна, может быть решена путем перебора. Результаты численного решения этой задачи для всевозможных независимо кодируемых последовательностей (компонент и их состояний) каждого из массивов I–V приведены в колонках табл. 3, обозначенных I,. . . ,V; состояния построены с фиксированными значениями порогов (см. соответствующие колонки табл. 2).

Вычисление оптимальных весов требует значительного времени. Поэтому в алгоритме, предназначенном для практического применения, следует использовать фиксированные значения весов, общие для всей совокупности сжимаемых массивов. Выбранные для совокупности массивов I–V значения весов представлены в колонках табл. 3, обозначенных I–V.

В строках 7–9 табл. 1 представлены скорости кодирования R[1|1|0] массивов I–V описанным выше базовым алгоритмом A[1|1|0] и скорости кодирования R[1|1|opt] и R[1|1|fix] этих массивов производными от него алгоритмами A[1|1|opt] и A[1|1|fix] соответственно. Производные алгоритмы отличаются от базового алгоритма использованием оптимальных и фиксированных весов w в (1) при построении кодовых распределений. Аналогично содержание строк 10–12, 13–15 и 16–18 таблицы.

Анализ представленных результатов позволяет сделать следующие выводы. Во-первых, использование в алгоритме фиксированных весов позволяет уменьшить избыточность кодирования в 1,9–3,0 раза по сравнению с базовым алгоритмом. Во-вторых, замена в алгоритме фиксированных весов на оптимальные практически не уменьшает избыточность кодирования: во всех случаях уменьшение избыточности при такой замене не превысило 0,0001 бт/п.

Данные табл. 1 показывают, что наиболее эффективным среди практически применимых алгоритмов является алгоритм A[2|2,fix|fix], который состоит в двукратном применении ДВП, разложении полученных компонент на состояния с использованием фиксированных порогов и независимом кодировании элементов построенных состояний с использованием фиксированных весов в (1) при построении кодовых распределений.

4 Заключение

Предложен ряд алгоритмов обратимого сжатия массивов силовых кривых, основанных на универсальном арифметическом кодировании компонент ДВП, и построены оценки скорости кодирования этих алгоритмов. Полученные результаты показывают, что повторное применение ДВП, разложение компонент ДВП на независимо кодируемые вычислимые состояния и выбор подходящих весов в формуле для кодовых распределений позволяют уменьшить скорость кодирования, т. е. увеличить степень сжатия.

Наиболее эффективным среди предложенных в настоящей работе и в работе [2] алгоритмов, которые могут быть применены на практике, является алгоритм A[2|2,fix|fix]. Скорость кодирования R[2|2,fix|fix] этого алгоритма (строка 18 табл. 1) заметно меньше скорости кодирования алгоритма обратимого сжатия стандарта JPEG 2000 [1]. Для массивов I,. . .,V выигрыш составляет 0,2863, 0,3185, 0,2936, 0,2908 и 0,2366 бт/п соответственно.

Литература

1. *Стефанович А. И., Сушко Д. В.* О сжатии данных массивов силовых кривых // Информационные процессы, 2020. Т. 20. № 3. С. 284–296.
2. *Сушко Д. В.* Алгоритмы сжатия данных массивов силовых кривых I: кодирование ошибок предсказания // Информатика и её применения, 2021. Т. 15. Вып. 2. С. 82–88. doi: 10.14357/19922264210212.
3. *Witten I. H., Neal R. M., Cleary J. G.* Arithmetic coding for data compression // Commun. ACM, 1987. Vol. 30. No. 6. P. 520–540. doi: 10.1145/214762.214771.
4. *Штарьков Ю. М.* Универсальное кодирование. Теория и алгоритмы. — М.: Физматлит, 2013. 288 с.
5. *Le Gall D., Tabatabai A.* Sub-band coding of digital images using symmetric short kernel filters and arithmetic coding techniques // IEEE Conference (International) on Acoustics, Speech, and Signal Processing. — Piscataway, NJ, USA: IEEE, 1988. Vol. 2. P. 761–764. doi: 10.1109/ICASSP.1988.196696.
6. *Sweldens W.* The lifting scheme: A custom-design construction of biorthogonal wavelets // Appl. Comput. Harmon. A., 1996. Vol. 3. No. 2. P. 186–200. doi: 10.1006/acha.1996.0015.
7. *Taubman D. S., Marcellin M. W.* JPEG2000: Image compression fundamentals, standards, and practice. — New York, NY, USA: Springer Science + Business Media, 2002. 773 p. doi: 10.1007/978-1-4615-0799-4.

Поступила в редакцию 30.12.2020

COMPRESSION ALGORITHMS FOR FORCE VOLUME DATA II: CODING OF WAVELET TRANSFORM COMPONENTS

D. V. Sushko

Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The author presents the second part of the investigation of the problem of reversible (lossless) compression of force volume data which are the three-dimensional arrays with 16-bit integer elements. The author proposes reversible compression algorithms of force volume data based on the universal arithmetic coding of components obtained as the result of applying the one-dimensional discrete wavelet transform using (5–3) wavelet system to the rows of the arrays. The transform is realized in the frame of lifting scheme and it is reversible. To construct effective algorithms, the author uses the method of wavelet transform reapplication and two methods of universal coding previously tested in the first part of the investigation (decomposition into calculable states and choice of the weight while constructing the code probabilities). The author constructs bit rate estimations for the proposed

algorithms for five test arrays. The results show that each of the methods mentioned above decreases the bit rate and the combination of all three methods leads to the most efficient algorithm. The bit rates of this algorithm for the test arrays are 3.8160, 3.4050, 3.3678, 4.1309, and 4.0996 bit/pixel, benefit in comparison with the standard JPEG 2000 reversible compression algorithm is 6%–9%.

Keywords: atomic force microscope; force volume data; reversible compression; arithmetic coding; universal coding

DOI: 10.14357/19922264210303

References

1. Stefanovich, A. I., and D. V. Sushko. 2020. O szhatii dannykh massivov silovykh krivykh [On data compression of force volumes]. *Informatsionnye protsessy* [Information Processes] 20(3):284–296.
2. Sushko, D. V. 2021. Algoritmy szhatiya dannykh massivov silovykh krivykh I: kodirovanie oshibok predskazaniya [Compression algorithms for force volume data I: Prediction errors coding]. *Informatika i ee Primeneniya — Inform. Appl.* 15(2):81–87. doi: 10.14357/19922264210212.
3. Witten, I. H., R. M. Neal, and J. G. Cleary. 1987. Arithmetic coding for data compression. *Commun. ACM* 30(6):520–540. doi: 10.1145/214762.214771.
4. Shtar'kov, Yu. M. 2013. *Universal'noe kodirovanie. Teoriya i algoritmy* [Universal coding. Theory and algorithms]. Moscow: Fizmatlit. 288 p.
5. Le Gall, D., and A. Tabatabai. 1988. Sub-band coding of digital images using symmetric short kernel filters and arithmetic coding techniques. *IEEE Conference (International) on Acoustics, Speech, and Signal Processing Proceedings*. Piscataway, NJ: IEEE. 761–764. doi: 10.1109/ICASSP.1988.196696.
6. Sweldens, W. 1996. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. A.* 3(2):186–200. doi: 10.1006/acha.1996.0015.
7. Taubman, D. S., and M. W. Marcellin. 2002. *JPEG2000: Image compression fundamentals, standards, and practice*. New York, NY: Springer Science + Business Media. 773 p. doi: 10.1007/978-1-4615-0799-4.

Received December 30, 2020

Contributor

Sushko Dmitry V. (b. 1962) — Candidate of Science (PhD) in physics and mathematics, 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; dsushko@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 24–28

МАКСИМАЛЬНЫЕ МЕЖУЗЛОВЫЕ ПОТОКИ ПРИ ПРЕДЕЛЬНОЙ ЗАГРУЗКЕ МНОГОПОЛЬЗОВАТЕЛЬСКОЙ СЕТИ

Ю. Е. Малашенко[1]

Аннотация: Рассматривается метод поиска угловых точек на гранях выпуклого многогранного множества допустимых межузловых потоков, передаваемых между всеми узлами сети одновременно. Базовыми считаются точки пересечения осей координат с внешней границей множества. В качестве опорной выбирается реперная точка, в которой сумма межузловых потоков является максимально возможной среди всех допустимых распределений. На основе полученных данных формируется система опорных векторов с компонентами, равными межузловым потокам, при одновременной передаче которых достигается предельная загрузка сети. Для целей анализа допустимые распределения межузловых потоков предлагается записывать как выпуклую комбинацию опорных векторов. Полученное агрегированное представление может быть использовано при разработке нормативных показателей стационарных режимов функционирования при передаче информационных потоков, превышающих функциональные возможности сети. В качестве примеров рассматриваются оценки равнодолевого распределения максимально возможных потоков.

Ключевые слова: многопользовательская сеть; предельная загрузка сети; множество межузловых потоков

DOI: 10.14357/19922264210304

1 Введение

Для многопользовательской сети связи предлагается развитие метода агрегированного описания множества достижимых межузловых потоков [1, 2]. В настоящей работе строятся гарантированные оценки потоков, которые могут одновременно передаваться между всеми парами узлов-корреспондентов. Базовые векторы межузловых потоков соответствуют угловым точкам на границе множества достижимости [3]. Соответствующие значения компонент векторов находятся как решение последовательности однопродуктовых задач поиска максимального потока для выделенной пары узлов при фиксированных нулевых значениях для всех остальных [4]. Компоненты исходных векторов определяются при поиске совместных распределений межузловых потоков, при которых достигается предельная загрузка сети. Выпуклая комбинация полученных опорных векторов задает набор гарантированных нижних оценок допустимых значений межузловых потоков [5].

2 Математическая модель

В качестве математической модели для описания многопользовательской сетевой системы используется следующая формальная запись условий и ограничений, которые должны выполняться при одновременной передаче потоков различных видов между всеми парами узлов-корреспондентов.

Сеть $G(\mathbf{d})$ задается множествами $\langle V, R, U, P\rangle$:

– узлов (вершин) сети

$$V = \{v_1, v_2, \ldots, v_n, \ldots, v_N\};$$

– неориентированных ребер

$$R = \{r_1, r_2, \ldots, r_k, \ldots, r_E\}.$$

Ребро r_k *соединяет* концевые вершины v_{n_k} и v_{j_k}. Ребру r_k ставятся в соответствие две ориентированные дуги $\{u_k, u_{k+E}\}$ из множества ориентированных дуг

$$Um = \{u_1, u_2, \ldots, u_k, \ldots, u_{2E}\}.$$

Дуги $\{u_k, u_{k+E}\}$ определяют прямое и обратное направление передачи потока по ребру r_k между концевыми вершинами $\{v_{n_k}, v_{j_k}\}$.

Обозначения:

$S(v_n)$ — множество номеров исходящих дуг, по которым поток покидает узел v_n;

$T(v_n)$ — множество номеров входящих дуг, по которым поток поступает в узел v_n.

Состав множеств $S(v_n)$ и $T(v_n)$ однозначно определяется в ходе выполнения следующей процедуры. Пусть некоторое ребро $r_k \in R$ соединяет вершины с номерами n и j такими, что $n < j$,

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, mala-yur@yandex.ru

тогда ориентированная дуга u_k, направленная из вершины v_n в v_j, считается *исходящей* из вершины v_n, и ее номер k заносится в множество $S(v_n)$, а дуга u_{k+E}, направленная из вершины v_j в v_n, — *входящая* для v_n, и ее номер $(k+E)$ помещается в список $T(v_n)$. Соответственно дуга u_k является *входящей* для v_j, и ее номер k попадает в $T(v_j)$, а дуга u_{k+E} — *исходящей*, и номер $(k+E)$ вносится в список исходящих дуг $S(v_j)$.

В многопользовательской сети $G(\mathbf{d})$ рассматривается $M = N(N-1)$ независимых, невзаимозаменяемых и равноправных потоков различных видов, которые передаются между узлами-корреспондентами из множества $P = \{p_1, p_2, \ldots, p_M\}$.

По определению, каждой паре узлов-корреспондентов p_m соответствуют:

– вершина-источник с номером s_m, через которую входной поток m-го вида z_m поступает в сеть;

– вершина-приемник с номером t_m, из которой поток m-го вида z_m покидает сеть.

В множестве P выделяется подмножество $P(R^+)$ пар узлов-корреспондентов, расположенных в концевых вершинах ребра r_k, $k = \overline{1,E}$.

Следуя вышеизложенной процедуре нумерации направленных дуг, вводятся следующие обозначения: пусть ребро r_k соединяет вершины с номерами n и j такими, что $n < j$. Для соответствующей пары узлов-корреспондентов p_k, расположенных в узлах $\{v_n, v_j\}$, узел v_n считается источником, а узел v_j — приемником потока z_k k-го вида, который передается из узла с номером n в узел с номером j для пары p_k. Для пары $p_{k+} \Longleftrightarrow \{v_j, v_n\}$ узел v_j считается источником и устанавливается соответствие $s_k \Longleftrightarrow j$, а узел v_n — приемником и $t_k \Longleftrightarrow n$.

Пары p_m из подмножества $P(R^+)$ называются *смежными* узлами-корреспондентами. Все остальные пары узлов-корреспондентов относятся к множеству $P(R^-)$:

$$P = P(R^+) \cup P(R^-);$$
$$P(R^+) \cap P(R^-) = \varnothing\,.$$

Обозначим:

z_m — величина *межузлового* потока m-го вида, который поступает в сеть из узла с номером s_m и покидает сеть из узла с номером t_m;

x_{mi} — величина потока m-го вида, который передается по ребру r_i, $m = \overline{1,M}$, $i = \overline{1,E}$.

Во всех узлах $v_n \in V$, $n = \overline{1,N}$, для всех видов потоков должны выполняться условия сохранения потоков:

$$\sum_{i \in S(v_n)} x_{mi} - \sum_{i \in T(v_n)} x_{mi} = \begin{cases} z_m, & \text{если } v = v_{S_m}; \\ -z_m, & \text{если } v = v_{t_m}; \\ 0 & \text{в остальных случаях,} \end{cases}$$
$$n = \overline{1,N}\,,\ m = \overline{1,M}\,,\ x_{mi} \geq 0\,,\ z_m \geq 0\,. \quad (1)$$

Величина z_m равна входному потоку m-го вида, который пропускается от источника к приемнику пары p_m при распределении потоков x_{mi} по дугам сети.

Каждому ребру $r_k \in R$ приписывается неотрицательное число d_k, определяющее суммарный предельно допустимый поток, который можно передать по ребру r_k в обоих направлениях. В исходной сети компоненты вектора пропускных способностей $\mathbf{d} = (d_1, d_2, \ldots, d_k, \ldots, d_E)$ — наперед заданные положительные числа $d_k > 0$. Вектором $\mathbf{d}$ определяются следующие ограничения на сумму дуговых потоков всех видов, передаваемых по ребру r_k:

$$\sum_{m=1}^{M} (x_{mk} + x_{m(k+E)}) \leq d_k,$$
$$x_{mk} \geq 0\,,\ \ x_{m(k+E)} \geq 0\,,\ \ k = \overline{1,E}\,. \quad (2)$$

В рамках данной модели пропускная способность ребер сети — вектор $\mathbf{d}$ — трактуется как «*ресурсное ограничение*», а сумма дуговых потоков рассматривается как показатель использования «*ресурсов*» сети при передаче межузловых потоков различных видов.

Для всех z_m и x_{mi}, удовлетворяющих условиям (1) и (2), вычисляются суммарные потоки:

$$y_m = \sum_{i=1}^{2E} x_{mi},\ \ m = \overline{1,M}\,. \quad (3)$$

Суммарный реберный поток y_m характеризует «*нагрузку*» на сеть при передаче межузлового потока величины z_m из узла-источника s_m в узел-приемник t_m. Величина y_m показывает, какой суммарный *ресурс* — пропускная способность сети — требуется для передачи межузлового потока z_m, а отношение $w_m = y_m/z_m$, $m = \overline{1,M}$, показывает, какие *ресурсы* необходимы для передачи единичного потока m-го вида между узлами s_m и t_m.

Ограничения (1)–(3) задают подмножество допустимых значений компонент вектора межузловых потоков $\mathbf{z} = (z_1, z_2, \ldots, z_m, \ldots, z_M)$:

$$Z(\mathbf{d}) = \{\mathbf{z} \geq 0 \mid (\mathbf{z}, \mathbf{x}, \mathbf{y}) \text{ удовлетворяют (1)–(3)}\},$$

а все допустимые распределения ресурсов принадлежат подмножеству

$$Y(\mathbf{d}) = \{\mathbf{y} \geq 0 \mid (\mathbf{z}, \mathbf{x}, \mathbf{y}) \text{ удовлетворяют (1)–(3)}\}.$$

3 Монопольные режимы передачи потока

В рамках данного модельного описания, по определению, монопольным режимом называется способ управления, при котором все ресурсы сети используются для передачи потока одной выделенной пары узлов-корреспондентов $p_a \in P$, а для всех остальных потоки полагаются равными нулю.

Предельно допустимый поток, который можно передать между фиксированной парой узлов-корреспондентов p_a в монопольном режиме, является решением стандартной, в данном случае однопродуктовой, задачи о максимальном потоке.

Задача 1. Найти:

$$z_a^0 = \max_{\langle z,x\rangle \in Z(d)} z_a$$

при условии $z_i = 0$, $i = \overline{1,M}$, $i \neq a$.

При решении задачи 1 для пары p_a вычисляются: межузловой поток z_a^0; дуговые потоки $\{x_{ak}^0; x_{a(k+E)}^0\}$, $k = \overline{1,E}$; суммарное значение реберного потока $y_a^0 = \sum_{i=1}^{2E} x_{ai}^0$; *удельные затраты ресурсов (пропускной способности)* $w(a) = y_a^0/z_a^0$, требуемые для передачи одной единицы потока z_a^0.

Поток величины z_a^0 называется МРМ-потоком и является *максимальным потоком*, передаваемым в *монопольном режиме* для пары узлов-корреспондентов p_a.

Задача 1 решается последовательно для всех $p_m \in P$, вычисляются значения z_m^0 для $m = \overline{1,M}$, на основе которых формируется вектор предельных межузловых потоков

$$\mathbf{z}^0(0) = \langle z_1^0, z_2^0, \ldots, z_M^0 \rangle$$

с компонентами, равными МРМ-потокам.

Вектор $\mathbf{z}_m^0 = \langle 0, 0, \ldots, z_m^0, \ldots, 0, 0 \rangle$ определяет координаты угловой точки множества достижимости $Z(\mathbf{d})$, которая лежит на пересечении границы $Z(\mathbf{d})$ с соответствующей координатной осью z_m.

Множество векторов

$$\mathbf{Z}^{(0)} = \langle \mathbf{z}_1^0, \mathbf{z}_2^0, \ldots, \mathbf{z}_M^\circ \rangle$$

определяет угловые точки *базового* МРМ-сечения.

Любой вектор $\mathbf{z}(\cdot)$, который можно представить как выпуклую комбинацию элементов $\mathbf{Z}^{(0)}$:

$$\mathbf{z}(\cdot) = \sum_{m=1}^{M} \gamma_m(\cdot)\mathbf{z}_m^0, \quad \sum_{m=1}^{M} \gamma_m(\cdot) = 1; \quad \gamma_m(\cdot) \geq 0,$$

задает координаты точки на поверхности базового МРМ-сечения.

В задаче 2 рассматривается луч $\{0, \beta, \mathbf{z}^0(0)\}$ с направляющим вектором $\mathbf{z}^0(0)$ и вычисляется значение $\beta^*(0)$ в точке пересечения луча с базовым МРМ-сечением.

Задача 2. Найти $\beta^*(0)$ и $\gamma_m^*(0)$ такие, что

$$\beta^*(0)\mathbf{z}^0(0) = \sum_{m=1}^{M} \gamma_m^*(0)\mathbf{z}_m^0, \quad \sum_{m=1}^{M} \gamma_m^*(0) = 1\,.$$

Представление искомого вектора в виде выпуклой комбинации угловых точек $\mathbf{z}_m^0(0)$ записывается покомпонентно:

$$\beta^*(0)z_m^0 = \gamma_m^*(0)z_m^0, \quad m = \overline{1,M}\,; \quad \sum_{m=1}^{M} \gamma_m^*(0) = 1\,.$$

Таким образом, решение

$$\beta^*(0) = \gamma_m^*(0) = \frac{1}{M}\,, \quad m = \overline{1,M}\,,$$

определяет искомый вектор $\mathbf{z}^* = (1/M)\mathbf{z}^0(0)$ с компонентами

$$z_m^* = \frac{1}{M}\, z_m^0, \quad m = \overline{1,M}\,,$$

равными величинам потоков, которые можно одновременно передать между всеми корреспондентами $p_m \in P$. Вектор $\mathbf{z}^*$ принадлежит базовому МРМ-сечению и задает допустимое *равнодолевое* распределение потоков, *пропорциональных* базовым значениям МРМ-потоков.

4 Предельная загрузка сети

В множестве $Z(\mathbf{d})$ существует допустимый вектор $\mathbf{z}(d)$ с компонентами:

$$z_m(d) = 0, \quad p_m \in P(R^-);$$
$$z_k(d) = d_k, \quad p_k \in P(R^+).$$

Сумма межузловых потоков

$$\sigma(d) = \sum_{k=1}^{E} z_k(d) = \sum_{k=1}^{E} d_k = \mathrm{D}^*$$

является максимально возможной среди всех допустимых векторов $\mathbf{z} \in Z(\mathbf{d})$.

Для каждой пары $p_k \in P(R^+)$ межузловому потоку $z_k(d)$ соответствует реберный поток $x_k^0(d)$, а $y_k^0(d) = x_k^0(d)$. Суммарная *нагрузка*

$$\sum_{k=1}^{2E} y_k^0(d) = \sum_{k=1}^{2E} x_k^0(d) = \sum_{k=1}^{E} z_k(d) = \sum_{k=1}^{E} d_k = \mathrm{D}^*.$$

Вектор $\mathbf{z}(d)$ задает распределение PLD-потоков (от *англ.* Peak-Load-Distribution), поскольку при одновременной передаче всех межузловых потоков $z_k(d)$, $p_k \in P(R^+)$ достигается *предельно допустимая загрузка* всех ребер сети.

Далее рассматривается процедура формирования PLD-векторов на базе вектора $\mathbf{z}(d)$ и векторов — решений задачи 1 для $p_m \in P$.

PLD-процедура.

1. Пусть для некоторой исходной пары $p \in P(R^-)$ в результате решения задачи 1 найден МРМ-поток z^0 и дуговые потоки $\langle (x^0_{ak} + x^0_{a(k+E)})\rangle$, $k = \overline{1,E}$. Значению z^0 ставится в соответствие вектор

$$\mathbf{z}^1(a) = \langle z^1_1(a), z^1_2(a), z^1_3(a), \ldots, z^1_M(a)\rangle$$

с компонентами:

$z^1_m(a) = 0$ для $m \neq a$, $p_m \in P(R^-)$, $y^1_m(a) = 0$;

$z^1_m(a) = z^0$ для $m = a$, $p_m \in P(R^-)$, $y^1_a(a) = \sum_{k=1}^{E}(x^0_{ak} + x^0_{a(k+E)})$;

$z^1_k(a) = d_k - (x^0_{ak} + x^0_{a(k+E)})$, $p_a \in P(R^+)$, $y^1_a(a) = d_k - (x^0_{ak} + x^0_{a(k+E)})$.

2. Пусть исходная пара $p_b \in P(R^+)$ и z^0_b — МРМ-поток для пары $p_b \in P(R^+)$. Компоненты соответствующего вектора $\mathbf{z}^1(b)$:

$z^1_m(b) = 0$ для $p_m \in P(R^-)$, $y^1_m(b) = 0$;

$z^1_k(b) = d_k - (x^o_{bk} + x^0_{b(k+E)})$ для $p_k \in P(R^+)$, $k \neq b$, $y^1_k(b) = d_k - (x^0_{bk} + x^0_{b(k+E)})$;

$z^1_k(b) = z^0_b$ при $k = b$, $p_k \in P(R^+)$, $y^1_b(b) = \sum_{k=1}^{E}(x^0_{bk} + x^0_{b(k+E)})$.

По построению для любого вектора $\mathbf{z}^1(j), p_j \in P$ суммарная (общая) *загрузка*:

$$\sigma(j) = \sum_{m=1}^{M} y^1_m(j) =$$

$$= \sum_{k=1}^{E}(x^0_{jk} + x^0_{j(k+E)}) + \sum_{k=1}^{E}[d_k - (x^0_{jk} + x^0_{j(k+E)})] =$$

$$= \sum_{k=1}^{E} d_k + \sum_{k=1}^{E}(x^0_{jk} + x^0_{j(k+E)}) - \sum_{k=1}^{E}(x^0_{jk} + x^0_{j(k+E)}) =$$

$$= \mathrm{D}^*.$$

Множество векторов $\mathbf{Z}^1 = \{\mathbf{z}^1(1), \mathbf{z}^1(2), \mathbf{z}^1(3), \ldots, \mathbf{z}^1(m), \ldots, \mathbf{z}^1(M)\}$ определяет угловые точки PLD-сечения. На основе вектора $\mathbf{z}(d)$ и угловых векторов из множеств $\mathbf{Z}^0$ и $\mathbf{Z}^1$ формируется опорный внутренний каркас: $\mathrm{SiF}(0) = \{\mathbf{z}(d), \mathbf{Z}^0, \mathbf{Z}^1\}$ (от *англ.* Support-internal-Frame – опорный внутренний каркас). Любая выпуклая комбинация векторов из множества $\mathrm{SiF}(0)$ задает допустимое распределение потоков, которые могут одновременно передаваться между всеми парами узлов-корреспондентов.

Для примера рассмотрим луч $\langle 0, \Theta\mathbf{z}^0(0)\rangle$ и определим значение Θ^* в точке пересечения с выпуклой оболочкой $\mathrm{SiF}(0)$.

Задача 3. Найти $\Theta^* = \max\limits_{\Theta,\gamma} \Theta$ при условиях:

$$\Theta\mathbf{z}^0(0) = \sum_{m=1}^{M} \gamma_m(0)\mathbf{z}^0_m + \sum_{m=1}^{M} \gamma_m(1)\mathbf{z}^1(m) + \lambda\mathbf{z}(d);$$

$$\Theta \geq 0\,;\ \gamma_m(0) \geq 0\,;\ \gamma_m(1) \geq 0\,;\ \lambda \geq 0\,;$$

$$\sum_{m=1}^{M} \gamma_m(0) + \sum_{m=1}^{M} \gamma_m(1) + \lambda = 1\,.$$

Решение задачи 3 позволяет получить гарантированные нижние оценки межузловых потоков $z^*_m = \Theta^* z^0_m$, $m = \overline{1,M}$, при *равнодолевом* распределении МРМ-потоков.

5 Заключение

В рамках модели предложенная алгоритмическая схема рассматривается как способ определения исходных нормативных параметров и ограничений при исследовании функциональных возможностей сети. Полученные оценки можно использовать в условиях неопределенности о текущих состояниях передающих подсистем и нестационарных входных потоках, превышающих функциональные возможности сети. Для целей анализа вектор входных потоков, умноженных на нормирующий коэффициент, записывается как выпуклая комбинация опорных векторов. Расчетное значение коэффициента меньше единицы будет означать, что пропускные возможности системы не позволяют одновременно передать все запрашиваемые потоки и следует ввести ограничения на передачу в соответствии с нормативными рекомендациями. При формировании опорных векторов многократно решается задача о максимальном потоке и минимальном разрезе [4]. Результирующие вычислительные затраты оцениваются полиномиальной функцией от общего числа узлов сети.

Литература

1. *Малашенко Ю. Е., Назарова И. А., Новикова Н. М.* Экспресс-анализ и агрегированное представление множества достижимых потоков многопродуктовой сетевой системы // Изв. РАН. ТиСУ, 2019. № 6. С. 63–72.
2. *Малашенко Ю. Е., Назарова И. А.* Аппроксимация множества достижимых потоков многопользовательской

сети // Информатика и её применения, 2020. Т. 14. Вып. 3. С. 81–85.
3. *Лотов А. В., Поспелова И. И.* Многокритериальные задачи принятия решений. — М.: Макс Пресс, 2008. 197 с.
4. *Йенсен П., Барнес Д.* Потоковое программирование / Пер. с англ. — М.: Радио и связь, 1984. 392 с. (*Jensen P. A., Barnes J. W.* Network flow programming. — New York, NY, USA: Wiley, 1980. 408 p.)
5. *Данциг Дж.* Линейное программирование, его применения и обобщения / Пер. с англ. — М.: Прогресс, 1966. 589 с. (*Dantzig G.* Linear programming and extensions. — Princeton, NJ, USA: Princeton University Press, 1963. 600 p.)

Поступила в редакцию 04.06.2021

MAXIMUM INTERNODE FLOWS AT PEAK LOAD OF A MULTIUSER NETWORK

Yu. E. Malashenko

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

Abstract: The paper discusses a method of angular points searching on the edges of the convex polyhedral set of permissible internode flows transmitted between all network nodes simultaneously. The points of coordinate axes intersection with the outer boundary of the set are considered as basic. A point, in which the sum of internode flows is the maximum possible among all permissible distributions, is selected as that reference point. Based on the data obtained, a system of support vectors is generated with components equal to internode flows, with simultaneous transmitting of which the peak network load is achieved. For the purpose of the analysis, the permissible distributions of internode flows are proposed to record as a convex combination of support vectors. The resulting aggregated representation can be used in the development of regulatory indicators of stationary operation modes when transferring information flows exceeding the network functionality. As examples, estimates of the equal distribution of maximum possible flows are considered.

Keywords: multiuser network; network peak load; internode flows set

DOI: 10.14357/19922264210304

References

1. Malashenko, Yu. E., I. A. Nazarova, and N. M. Novikova. 2019. Express analysis and aggregated representation of the set of reachable flows for a multicommodity network system. *J. Comput. Sys. Sc. Int.* 58(6):889–897.
2. Malashenko, Yu. E., and I. A. Nazarova. 2020. Approksimatsiya mnozhestva dostizhimykh potokov mnogopol'zovatel'skoy seti [Approximation of the multiuser network feasible flows set]. *Informatika i ee Primeneniya — Inform. Appl.* 14(3):81–85.
3. Lotov, A. V., and I. I. Pospelova. 2008. *Mnogokriterial'nye zadachi prinyatiya resheniy* [Multicriteria decision making tasks]. Moscow: Maks Press. 197 p.
4. Jensen, P. A., and J. W. Barnes. 1980. *Network flow programming*. New York, NY: Wiley. 408 p.
5. Dantzig, G. 1963. *Linear programming and extensions*. Princeton, NJ: Princeton University Press. 600 p.

Received June 4, 2021

Contributor

Malashenko Yuri E. (b. 1946) — Doctor of Science in physics and mathematics, principal scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; mala-yur@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 29–40

ЭКСПЕРТНАЯ СИСТЕМА ДЛЯ МОНИТОРИНГА И ПРОГНОЗИРОВАНИЯ ПРОЦЕССОВ РАСПРЕДЕЛЕНИЯ РЕСУРСОВ

А. В. Босов[1], Д. В. Жуков[2]

Аннотация: Представлен проект экспертной системы (ЭС) мониторингового типа, предназначенной для поддержки принятия решений при управлении процессами распределения (потребления и воспроизводства) ресурсов. Результаты анализа процесса потребления представляются в виде сценариев интегрального оценивания его состояния. Сценарии готовятся экспертом, носят ситуативный характер и в реальных расчетах применяются как для оценки текущего состояния, так и для прогнозирования развития ситуаций. Традиционная интерпретация процесса потребления основана на понятии ресурсов и субъектов потребления, использующих ресурсы в соответствии с нормами потребления, а также простым описанием воспроизводства ресурсов объектами производства. Предполагается, что компоненты информационной модели имеют географическую и темпоральную привязку, в частности привязаны ко времени данные о запасах ресурсов. Основную интеллектуальную нагрузку несет методика подготовки сценариев интегрального оценивания состояния. Эта методика основана на идеологии экспертного оценивания типовых ситуаций и формировании сценариев расчета простыми методами машинного обучения. В последних используются широко распространенные подходы к оптимизации — минимизация средних квадратов и модулей. Представленный проект ЭС носит инструментальный характер, может применяться в различных приложениях. Процесс подготовки и оценки результативности подготовленных экспертом сценариев интегрального оценивания иллюстрируется числовым и графическим материалом.

Ключевые слова: экспертная система; ресурсы и потребление; машинное обучение; метод наименьших квадратов; метод наименьших модулей

DOI: 10.14357/19922264210305

1 Введение

В работе рассматривается прикладное решение задачи анализа распространенного во многих практических областях процесса потребления-воспроизводства ресурсов. Используется универсальная терминология — традиционное понимание термина *ресурс* с акцентом на его экономическую интерпретацию [1], т. е. ресурс рассматривается как средство для реализации некоторых потребностей [2]. Типизация ресурсов не важна, существенно только то, что ресурсы можно количественно измерять или оценивать и результат этого измерения носит динамический характер, т. е. *объемы ресурсов* могут возрастать и уменьшаться, обеспечивая показателю числовую динамику во времени. Типичным источником для этой динамики может выступать, например, доктрина экономического потребления и производства [3]. Исходя из такой интерпретации, ресурсы считаются *объектами потребления*, предназначенными для использования *субъектами потребления*. Типовой субъект — человек, возможно, наделенный дополнительными свойствами, например пациент, потребитель, спасатель и т. п. Компенсируют потребление *объекты производства*, обеспечивающие увеличение объема для тех ресурсов, которые являются *воспроизводимыми*. Эта модель носит довольно универсальный характер и позволяет описывать потребление в различных прикладных интерпретациях, не ограничиваясь распределением создаваемой в процессе производства продукции. Ресурсами могут быть как материальные сущности, потребляемые ежедневно людьми, так и факторы производства (экономические и/или производственные), природные ресурсы типа полезных ископаемых. Также можно описывать административные ресурсы, услуги, задействуемые при решении организационных вопросов, муниципальном или государственном управлении, или еще менее традиционные информационные и вычислительные ресурсы, обеспечивающие работу инфотелекоммуникационных систем.

Решаемая задача традиционна еще и в том смысле, что наряду с потребляемыми ресурсами, субъектами потребления и объектами производства предполагается наличие средств мониторинга, способных получать информацию о характеристиках процесса потребления-воспроизводства. Глав-

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, ABosov@frccsc.ru
[2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, DZhukov@frccsc.ru

ное же предложение в рассматриваемой задаче — это наличие эксперта, заинтересованного не просто в мониторинге описанного процесса, а в формировании оценки состояния процесса и прогнозирования его развития.

Задачи такого рода наиболее активно рассматривались примерно в 1980-х гг., и решение их чаще всего связывалось с *экспертными системами* [4, 5]. Это программные системы, основанные на созданном людьми знании, которое собирается в хорошо изученных областях от экспертов (как правило, путем их заинтересованного интервьюирования) и выражается в виде правил (например, конструкции «если, то»), которые реализуются в программной форме. Можно считать, что такие системы породили то научное направление, которое принято называть «искусственным интеллектом» [5]. Точнее, ЭС — это образец технологий искусственного интеллекта, характерных для первой его волны, относящейся к концу прошлого века [6, 7]. Далее состоялась вторая волна технологий искусственного интеллекта, основная черта которой — развитие машинного обучения [8], прежде всего для искусственных нейронных сетей.

Сейчас принято говорить о начальной стадии третьей волны, которая фокусируется на объяснительных и общих технологиях искусственного интеллекта. Но пока эта волна охватывает лишь научные исследования и только начинают формироваться фундаментальные результаты, прикладные задачи успешно решаются уже имеющимися и подтвердившими практическую полезность и востребованность методами первой-второй волн.

Стимул реализации представленного проекта состоял в том, чтобы в отношении процесса потребления ресурсов реализовать мониторинговую ЭС в традиционном смысле, наполнив ее адаптированными к прикладной задаче средствами машинного обучения. При этом важной была идея использовать эксперта не только в качестве конечного пользователя программной реализации, но и в качестве источника знаний для обучения системы интегральному оцениванию состояния процесса потребления ресурсов и прогнозированию его развития. В этом смысле предлагаемая ЭС отличается от традиционной, в которой база знаний формируется полностью на этапе создания, а сами знания вносятся в систему специально обученным инженером по знаниям. В описываемой системе корректировать базу знаний эксперт может самостоятельно в процессе работы, в том числе и модифицируя алгоритм анализа и прогнозирования с помощью машинного обучения [9].

Обсуждение ЭС построено в основном на терминологии [10]. Выделяются самые важные элементы: база знаний, содержащая интеллектуальное описание процесса потребления, и пользователь-эксперт. Кроме того, реализуются два режима функционирования системы — *режим ввода знаний*, в котором эксперт (что важно, без помощи инженера по знаниям) наполняет базу знаний, и *режим консультации*, в котором эксперт выступает пользователем, ведя диалог с системой и получая от нее рекомендации.

В рамках общей классификации ЭС представляемую систему следует отнести к системам мониторинга и прогнозирования.

Завершая вводную часть, подчеркнем, что широкий интерес к таким мониторинговым ЭС имеет вполне современный характер, и, не претендуя на объемный обзор, приведем некоторые примеры.

Значимой и очень популярной прикладной областью являются системы экологического мониторинга. Здесь можно найти как задачи сбора комплексных наблюдений за состоянием окружающей среды, за происходящими в ней процессами, явлениями, так и задачи оценки и прогноза изменения состояния [11, 12]. Есть множество современных примеров практического применения мониторинговых ЭС в промышленности, например при производстве синтетических волокон [13], ремонте авиационных двигателей, в машиностроении [14, 15], при мониторинге морской обстановки [16].

Широкую область применения представляют вопросы прогнозирования и реагирования в чрезвычайных ситуациях [17], а также задачи телемедицины [18]. К таким же актуальным прикладным областям может применяться и обсуждаемая система мониторинга процесса потребления ресурсов. В качестве типового, но далеко не единственного приложения можно указать на систему жизнеобеспечения населения в чрезвычайных ситуациях [19].

Поскольку процесс создания ЭС хорошо изучен и описан, в разд. 2 кратко обсуждается информационная модель и функциональные процессы. Основное внимание уделено алгоритмическому обеспечению процесса обучения, которому посвящен разд. 3. Численные и графические иллюстрации к предложенной методике представлены в разд. 4.

2 Модель и функционирование экспертной системы мониторинга потребления ресурсов

В качестве образца приложения, которому должна соответствовать информационная модель

процесса потребления, использовалась существующая система управления потоками ресурсов жизнеобеспечения населения, принятая в РФ. Это позволило представить и обосновать основные информационные элементы, а также провести эксперименты, имитировать работу эксперта, использующего оценки состояния процесса и прогноз как источник принятия управленческих решений в прикладной сфере.

Итак, к основным элементам информационной модели процесса потребления-воспроизводства ресурсов отнесены ресурсы, нормы потребления, субъекты потребления, объекты производства и маркеры геопривязки.

А. *Ресурс*, интерпретируемый как источник удовлетворения некоторых потребностей субъектов потребления, определяется набором свойств, позволяющих гибко формировать описание этой формальной сущности в конкретных приложениях. Так, с ресурсами связан классификатор, позволяющий объединять их в группы, строить иерархии. Ряд свойств позволяет их типизировать, особо отметим свойство воспроизводимости (для таких ресурсов к процессу потребления добавляется процесс производства). Также отметим свойство ресурса «быть услугой», в частности потребляться без уменьшения запаса ресурса.

Б. *Норма потребления*. Это основное свойство ресурса, во многом определяемое его типизацией. Нормы потребления могут задаваться в типовых единицах, например тонн нефтепродуктов в сутки, т. е. характеризовать объем ресурса, периодически извлекаемый из имеющегося запаса. Другой пример — число больничных коек на человека — ресурс, не обеспечиваемый производством, с постоянным условно неуменьшаемым запасом, но ограниченный ситуацией с численностью пациентов. Нормы предусматриваются и для услуг, например такие, как охват транспорта средствами мониторинга положения, населения — системами оповещения, пожароопасных природных территорий — средствами контроля. Вариантов здесь неограниченно много, свой набор задает каждая предметная область.

В. *Субъекты потребления*. Примеры субъектов прямо следуют из норм потребления. Если это нефтепродукты, то субъект — завод или регион, если больничные койки, то — население, пациенты и т. д. Для модели важна только связка: есть ресурс, он потребляется субъектом согласно периодичности и объемам, заданным нормой потребления.

Г. *Объекты производства*. Принципиально здесь имеются два вида сущностей — это производственные мощности и хранилища. Производственные мощности описывают процесс, «обратный» потреблению: запас воспроизводимого ресурса в соответствии с заданным объемом производства (аналог нормы потребления) увеличивается согласно заданной периодичности. Хранилища в традиционном смысле производством не являются, этот объект описывает значительные объемы ресурса, имеющиеся в резерве. Как у традиционного производства объем хранения может изменяться, в его отношении может приниматься управляющее воздействие «использовать резерв», т. е. для такого объекта характерным примером служит государственный резерв, другой актуальный пример — законсервированная больница.

Д. *Геопривязка*. Моделирование процесса потребления-воспроизводства ресурсов имеет смысл, если одновременно требуется выполнить анализ и получить прогноз в виде многих оценок — как с учетом динамики, так и статически. Это возможно только в том случае, когда процесс потребления многомерный, т. е. одинаковые ресурсы потребляются во многих местах. Такое представление обеспечивается в форме географической привязки, связывающей все субъекты потребления и объекты производства с их географическим расположением. Так, если речь идет о глобальном процессе, то геопривязка — это перечень стран, если ограничиться государственным масштабом РФ, тогда это субъекты РФ, если городским масштабом, то районы города/субъекта.

Назначение информационной модели, точнее данных, собираемых в рамках мониторинга процесса потребления-воспроизводства, — *интегральное оценивание текущего состояния* этого процесса и *прогнозирование его развития*. Предполагается, что в рамках этапа мониторинга, с которого начинается функционирование ЭС, выполняется сбор информации, т. е. в рамках выбранного набора ресурсов, субъектов и объектов осуществляется сбор и накопление данных. Отметим, что предусматривается вариант, когда система работает с максимально скудным набором данных, ограниченным только текущими сведениями, но для «хорошей» работы лучше статистику накопить, чтобы эксперт хотя бы примерно мог оценить динамику изменений ситуации с ресурсами в целом и в каждой географической области.

Для выполнения анализа в системе предусмотрены два заранее подготовленных алгоритма. Пер-

вый — это *интегральная оценка состояния процесса потребления* в выбранной (выбранных) географической области. Этот алгоритм, получая команду эксперта, выполняет расчет, формируя статическую оценку состояния по имеющимся данным. Оценка — это число от 0 до 100. Для каждой заданной территории эксперт получит свое число, его удобно цветом отобразить средствами геоинформационной системы (ГИС), если таковая имеется.

Второй алгоритм *моделирует развитие процесса*. В простейшем случае расчета на заданный временной диапазон имеющиеся запасы уменьшаются за счет потребляемых и пополняются за счет воспроизводимых ресурсов. Типовой временной диапазон — это месяц с суточным шагом. На каждом шаге вычисляется оценка состояния по тому же статическому алгоритму. При наличии накопленной статистики для некоторых показателей можно включить в расчет прогноз их динамики. Например, если есть динамика по заболевшим/выздоровевшим, то можно вычислить простой линейный прогноз, что будет явно лучше, чем использовать текущие значения, считая их постоянными и на перспективу. Такими техническими, вспомогательными прогнозами можно снабдить многие из перечисленных выше показателей в отношении численности населения, объемов производства, услуг и т. д. При любом варианте использования итог второго алгоритма — это последовательность интегральных оценок состояния, привязанных ко времени и к территории.

Метод интегрального оценивания состояния процесса потребления является основным интеллектуальным средством обсуждаемой ЭС, а его идея опирается на два положения.

1. Оценивание состоит в расчете интегральной характеристики состояния ситуации в заданный момент времени для заданной территории по имеющимся (в том числе полученным в результате прогнозирования) данным. Главное в этом определении — это понятие *ситуации*, описываемой ограниченным перечнем типов ресурсов. Ситуация может быть как самой общей, т. е. включать все без исключения ресурсы, заданные используемым классификатором ресурсов, так и произвольную их комбинацию вплоть до пары каких-то ресурсов. Например, выполнялись эксперименты с моделью ситуации, включающей ресурсы: средства индивидуальной защиты населения, аппаратура ИВЛ (искусственной вентиляции легких), число мест в эпидемиологических профильных отделениях больниц. Для объемов этих ресурсов основным значимым показателем, характеризующим субъект потребления, очевидно, является численность пациентов, а сам пример, как нетрудно догадаться, вызван вниманием к текущей ситуации с пандемией COVID-19.
2. Для формирования правила оценивания применяются *экспертные оценки*, предоставляемые пользователем системы, т. е. пользователь по данной конкретной ситуации имеет возможность задать некоторое множество собственных оценок-образцов, а система использует их для формирования универсального алгоритма оценивания состояния. Именно в этом месте, с одной стороны, формируется база знаний ЭС, с другой — задействуется средство машинного обучения, обобщающее заданные экспертом знания на любой вариант рассматриваемой ситуации (формальному алгоритму посвящен следующий раздел).

Важное свойство обсуждаемой ЭС — отсутствие требования к эксперту во владении особыми навыками в информационных технологиях, чтобы заполнять базу знаний. Напомним, что в традиционных ЭС этим занимается специально подготовленный инженер по знаниям. В обсуждаемой системе описание ситуации, оно называется *шаблоном*, легко создает обычный пользователь. Для этого ему дается простой интерфейс к текущим данным, так что просто выбирается любая из описанных территорий и произвольная комбинация из имеющихся в классификаторе ресурсов. Такой шаблон понятен эксперту, потому что основан на реальных данных, его легко создать, а в дальнейшем легко интерпретировать. Примерная иллюстрация действий эксперта, создающего шаблон, представлена на рис. 1. Созданный шаблон именуется и сохраняется в базе знаний вместе с набором параметров — текущих данных территории, использованной в качестве образца при создании шаблона. Но при этом геопривязка устраняется и шаблон далее можно использовать много раз уже как абстрактное описание ситуации.

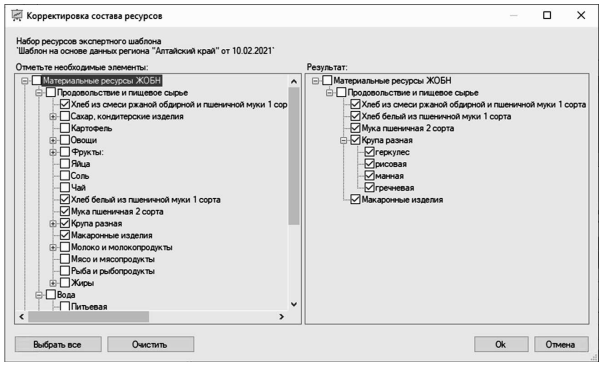


Рис. 1 Вариант интерфейса для задания шаблона ситуации

3 Алгоритм интегрального оценивания

При формировании алгоритма оценивания ситуации пользователь выполняет три шага.

Первый — самый важный — добавляет в базу знаний экспертные оценки. С этой целью удобнее всего использовать данные шаблона, изменять их в ручном режиме, описывая ситуацию конкретными данными, используя наборы значений, которые представляются важными и понятными эксперту, и задавая для каждого набора значений свою интегральную оценку. Таких оценок можно заранее создать произвольно много. Кроме того, в системе предусмотрен набор простых базовых оценок (подробнее о них сказано далее). Формализовано при этом и понятие «хватает» (задается диапазон достаточности, когда запас ресурса удовлетворяет потребности без пополнения). Промежуточные значения определяются в базовом варианте простым линейным сглаживанием. Своими оценками эксперт смещает эту базовую оценку, изменяя значимость конкретного ресурса (как в сторону пересмотра состояния запаса, так и в сторону изменения влияния диапазона достаточности).

Наполнив базу знаний такими оценками, эксперт переходит ко второму этапу — создает *сценарий*, выбирая набор оценок для расчета. Этот шаг нужен для традиционной классификации расчетов такими терминами, как пессимистический/оптимистический.

Последний шаг — формирование алгоритма — система выполняет сама, преобразуя базовый прогноз, максимально приближая его поверхность к точкам — оценкам, заданным экспертом. Расчет, прежде чем сохраниться для дальнейшего использования, может быть протестирован пользователем (этот этап проиллюстрирован численным примером в следующем разделе).

Формальное описание алгоритма вычисления интегральной оценки начинается с определения *нормированной оценки каждого ресурса*, включенного в шаблон ситуации. Пусть R — объем некоторого ресурса, P — установленная для него норма потребления, C — численность субъектов потребления данного ресурса. Зададим *интервал* Δ *автономного жизнеобеспечения* (типичное значение — неделя), на протяжении которого имеющегося в резерве объема ресурса должно быть достаточно (без воспроизведения, если ресурс возобновляемый). Пусть также для ресурса определен *атрибут возобновляемый / постоянный / услуга* (типичный возобновляемый ресурс — это изделия производства; постоянный — это места в объектах коллективного пользования; услуга — это сервисы, предоставляемые субъектам потребления). Нормированная оценка x для возобновляемого ресурса:

$$x = \begin{cases} 100\,, & \text{если } R > \Delta PC\,; \\ \dfrac{R}{\Delta PC}\,100\,, & \text{если } PC < R \le \Delta PC\,; \\ 0\,, & \text{если } R \le PC\,. \end{cases} \qquad (1)$$

Эта оценка равна 100, если имеющегося запаса ресурса хватает на период Δ, 0, если отсутствует суточный запас, и относительную по отношению к периоду Δ величину в промежуточном случае. Для постоянного ресурса соотношение примерно такое же, но без интервала Δ:

$$x = \begin{cases} 100\,, & \text{если } R > PC\,; \\ \dfrac{R}{PC}\,100\,, & \text{если } 0 < R \le PC\,; \\ 0\,, & \text{если } R = 0\,. \end{cases} \qquad (2)$$

Это оценка достаточности ресурса, например численности комплектов спасения, врачей, наборов медикаментов, больничных коек, т. е. применяемых ситуативно. При отсутствии достаточного объема ресурса дается относительная оценка имеющегося объема.

Ресурс типа услуги описывает доступные сервисы, предоставляемые субъектам, и может характеризоваться либо показателем охвата (обеспеченность), либо пропускной способностью. Его оценка имеет вид:

$$x = \begin{cases} 100\,, & \text{если } R > P\,; \\ \dfrac{R}{P} \cdot 100 & \text{для типа «обеспеченность»}\,; \\ \dfrac{P}{R} \cdot 100 & \text{для типа «пропускная способность»}\,; \\ 0\,, & \text{если } R = 0\,. \end{cases} \qquad (3)$$

Это оценка объема охваченных услугой по отношению к нормативу.

Применяя (1)–(3) к данным мониторинга ресурса, изменяя численность субъектов C, интервал Δ и объем P, для каждого ресурса можно получить оценку, принимающую значения от 0 до 100 и имеющую понятный эксперту смысл. Это дает возможность обратной интерпретации данных при включении экспертных оценок в сценарий: изменяя абстрактные нормированные оценки ресурсов, эксперт изменяет данные по численности субъектов и запасам.

Теперь переходим к интегральным оценкам. Пусть в шаблон включены p ресурсов с объемами запаса $R_1, \ldots, R_p$, для которых будут вычисляться (задаваться экспертом) нормированные оценки

состояния $x_1, \ldots, x_p$. Цель — сформировать сценарий, включив в него выбранный шаблон и N оценок, задающих экспертное представление о ситуациях и соответствующих им интегральных оценках. Для этого через X^i обозначим p-мерный вектор, составленный из нормированных оценок состояния ресурсов: $X^i = \left(x_1^i, \ldots, x_p^i\right)^p$, а отвечающие им интегральные оценки (заданные по умолчанию, сформированные ЭС или сформированные экспертом) оформим в виде равенства

$$I_i = f\left(X^i\right), \quad i = 1, \ldots, N. \tag{4}$$

Далее требуется указать еще один важный атрибут ресурса — *витальный/невитальный*, указывающий, является ли наличие ресурса «жизненно необходимым» или без него можно обойтись. В каждый сценарий при создании ЭС всегда добавляет такие оценки:

$I_1 = f(100, \ldots, 100) = 100$, т. е. если все ресурсы в состоянии 100, то интегральная оценка 100;

$I_i = f(100, \ldots, 0, \ldots, 100) = 0$, т. е. если состояние i-го витального ресурса имеет оценку 0 (для каждого ресурса в отдельности), то интегральная оценка 0 (число витальных ресурсов в дальнейшем обозначается K);

$I_i = f(100, \ldots, 0, \ldots, 100) = 100^{(p-1)/p}\alpha^{1/p}$, т. е. если состояние i-го невитального ресурса имеет оценку 0 (для каждого ресурса в отдельности), то интегральная оценка есть среднее геометрическое оценок, в котором множитель, соответствующий этому ресурсу, равен заданной величине α, по умолчанию $\alpha = 50$ и меняется экспертом, как и параметр автономности Δ (число невитальных ресурсов обозначается M, так что $K + M = p$).

Это минимальный набор знаний, который формируется ЭС автоматически при инициализации сценария и дает начальные условия для подбора функции $f(X)$ в (4), т. е. для проведения обучающего расчета.

Будем считать, что векторы оценок X^i в (4) пронумерованы так, что первыми идут перечисленные оценки описанного набора знаний: $I_1 = 100$, $I_i = 0$, $i = 2, \ldots, K+1$, $I_i = 100^{(p-1)/p}\alpha^{1/p}$, $i = K + 2, \ldots, K + M + 1$. Это первые $p + 1$ участвующих в обучающем расчете оценок I_i (по одной для каждого ресурса $R_1, \ldots, R_p$ в состоянии 0 и одна для общего состояния 100). Объясняется выбор оценок минимального набора тем, что они представляют граничные оценки умолчательной функции $\Phi(X) = \varphi(x_1)\varphi(x_2)\cdots\varphi(x_p)$, где

$$\varphi(x_i) = \begin{cases} x_i^{1/p}, & \text{если } i\text{-й ресурс витальный;} \\ \left(\dfrac{100-\alpha}{100}\, x_i + \alpha\right)^{1/p}, & \\ & \text{если } i\text{-й ресурс невитальный.} \end{cases}$$

Таким образом, предлагаемый ЭС по умолчанию алгоритм интегрального оценивания представляет собой вычисление $f(X)$ в форме среднего геометрического $\Phi(X)$ имеющихся нормированных оценок витальных ресурсов и пропорционально α сдвинутых от нуля оценок невитальных ресурсов. Результирующий алгоритм экспертного оценивания, представляемый в (4) функцией $f(X)$, должен учитывать включенные экспертом в (4) оценки X^i, $i = p + 2, \ldots, N$. Цель этого алгоритма состоит в том, чтобы, учитывая знания минимального набора, скорректировать $\Phi(X)$ так, чтобы уменьшить отклонения $f(X^i)$ от экспертных оценок I_i. Для этого представим функцию $f(X)$ в виде:

$$f(X) = a^{\mathrm{T}} h(X), \quad h(X) = (\Phi(X), x_1, \ldots, x_p)^{\mathrm{T}}. \tag{5}$$

Такая $f(X)$ — это линейная регрессия аргумента $h(X)$, определяемая коэффициентами $a^{\mathrm{T}} = (a_1, \ldots, a_{p+1})$ и регрессорами $\Phi(X), x_1, \ldots, x_p$, т. е. значения умолчательного алгоритма $\Phi(X)$ будут корректироваться линейной комбинацией $x_1, \ldots, x_p$.

Первый предлагаемый вариант вычисления параметров a в (5) основан на *методе наименьших квадратов* (МНК), т. е. минимизируемой характеристикой близости $f(X^i)$ от экспертных оценок I_i служит $\sum_{i=1}^{N}(f(X^i) - I_i)^2$. Варианты расчета для этого критерия хорошо известны, как и свойства оптимального решения. Это решение будем обозначать $\hat{a}$ и использовать наиболее удобный в практической реализации рекуррентный алгоритм расчета, а именно: определим соотношения для параметров $\hat{a}^i$, $i = 1, \ldots, N$, вычисляемых по экспертным данным $X^1, \ldots, X^i$, так что $\hat{a} = \hat{a}^N$, следующим образом [20]. Во-первых, заметим, что выражение $\hat{a}^i$ для $i = p + 1$ обеспечивается выбором регрессоров и значений $X^1, \ldots, X^{p+1}$: $\hat{a}^{p+1} = (1, 0, \ldots, 0)^{\mathrm{T}}$, так как при таком выборе $\hat{a}^{p+1}$ функция $f(X) = \Phi(X)$ проходит через все точки $X^1, \ldots, X^{p+1}$. Матрицу регрессоров для этого шага итерации $i = p + 1$ обозначим $H(X) = (h^{\mathrm{T}}(X^1), \ldots, h^{\mathrm{T}}(X^{p+1}))^{\mathrm{T}}$, так что первые $p + 1$ оценки I_i могут быть представлены в виде:

$$(I_1, \ldots, I_{p+1})^{\mathrm{T}} = H(X)\hat{a}^{p+1}.$$

Далее вычисляется матрица

$$B^{p+1} = \left(H^{\mathrm{T}}(X)H(X)\right)^{-1}.$$

Здесь существование обратной матрицы обеспечивается выбором регрессоров и значений

$X^1, \ldots, X^{p+1}$ ($H(X)$ имеет полный столбцовый ранг, поэтому $H^{\mathrm{T}}(X)H(X)$ невырожденна). С указанными $\hat{a}^{p+1}$ и B^{p+1} дальнейшие расчеты для $i = p+2, \ldots, N$ выглядят так [20]:

$$\hat{a}^{i+1} = \hat{a}^i + \frac{B^i h\left(X^{i+1}\right)}{1 + h^{\mathrm{T}}(X^{i+1}) B^i h\left(X^{i+1}\right)} \times \times \left(I_{i+1} - h^{\mathrm{T}}\left(X^{i+1}\right)\hat{a}^i\right);$$

$$B^{i+1} = B^i - \frac{B^i h\left(X^{i+1}\right) h^{\mathrm{T}}\left(X^{i+1}\right) B^{i\mathrm{T}}}{1 + h^{\mathrm{T}}\left(X^{i+1}\right) B^i h\left(X^{i+1}\right)}.$$

Окончательный расчет, т. е. искомая оценка $\hat{a}^N$ — это наилучшая линейная несмещенная оценка Гаусса–Маркова. Именно эта величина вместе с описанием (и всеми использованными данными) и представляет собой результат расчета — методику, которая сохраняется для будущего использования экспертом в рамках данного сценария. Завершение обучающего расчета обозначается пользователем прямым указанием сохранить методику. Предварять это действие может апробация рассчитанной методики. Для этого в интерфейс обучающего расчета включается возможность тестирования, т. е. применения методики на тренировочных данных. Иллюстрируется это примером в следующем разделе.

Второй вариант вычисления параметров a в (5) предлагается на основании *метода наименьших модулей* (МНМ) [21], т. е. минимизируется $\sum_{i=1}^{N} \left|f(X^i) - I_i\right|$. Стимулом для применения этого метода служат его известные робастные свойства, которые в рассматриваемом случае могут оказаться полезными в случае, если эксперт назначит отдельные оценки ситуациям, чрезмерно несогласующиеся с умолчательной методикой $\Phi(X)$ и другими оценками. Решение МНМ будем обозначать $\tilde{a}$, наиболее практичная реализации расчета — это также рекуррентный алгоритм [22]. Простоты МНК в этом варианте уже нет, рекурсия — это последовательное приближение к оптимальному решению, а основывается расчет на взвешенном МНК, а именно: решение $\tilde{a}$ по МНМ представим в виде предела последовательных оценок $\tilde{a}^t$, $t = 0, 1, \ldots$, вычисляемых рекуррентно взвешенным МНК. Для этого обозначим матрицы регрессоров и весов

$$\tilde{H}(X) = \left(h^{\mathrm{T}}\left(X^1\right), \ldots, h^{\mathrm{T}}\left(X^N\right)\right)^{\mathrm{T}};$$

$$B^t = \mathrm{diag}\left(\left|I_1 - \left(\tilde{a}^t\right)^{\mathrm{T}} h\left(X^1\right)\right|, \ldots \right. \left. \ldots, \left|I_N - \left(\tilde{a}^{t-1}\right)^{\mathrm{T}} h\left(X^N\right)\right|\right),$$

с помощью которых представим оценку на итерации t в виде:

$$\tilde{a}^t = \left(\tilde{H}^{\mathrm{T}}(X)\left(B^t\right)^{-1}\tilde{H}(X)\right)^{-1} \times \times \tilde{H}^{\mathrm{T}}(X)\left(B^t\right)^{-1}(I_1, \ldots, I_N)^{\mathrm{T}}. \quad (6)$$

Прекращая итерирование в (6) по достижении заданной точности, получим приближенное решение $\tilde{a}$.

Завершают обсуждение алгоритмов обучения в ЭС краткие положения о порядке проведения реального расчета. В рамках реального расчета пользователь ЭС может применить любую из подготовленных методик к текущим данным, поступающим в систему в рамках мониторинга состояния, а также выполнить прогноз развития ситуации на заданную перспективу, например на один месяц. В простейшем случае в прогнозной части определяется исходный начальный запас $R_i(0)$, $i = 1, \ldots, p$, для каждого ресурса из выбранного сценария, по каждому из возобновляемых ресурсов формируется динамика его изменения

$$R_i(n) = R_i(n-1) - P_i C + W_i,$$

где P_i — установленная для него норма потребления, W_i — объем воспроизводства ресурса за сутки; для невозобновляемых — динамика отсутствует:

$$R_i(n) = R_i(n-1).$$

В более наполненной ретроспективными данными мониторинга ЭС вместо приведенных тривиальных выражений прогноза могут использоваться более сложные варианты прогнозирования. И при любом варианте на каждый день в заданном диапазоне прогнозирования выполняется расчет интегральной оценки по выбранному сценарию, тем самым формируется обобщенный сценарий развития ситуации. Выполненный расчет прогноза стандартно представляется пользователю в табличной форме, а при наличии функциональных возможностей системы — визуализируется средствами ГИС.

4 Численный пример

Учитывая, что презентационные возможности ограничены изображением трехмерной поверхности, рассматриваемый пример реализации представленных методов интегрального оценивания состояния процесса распределения-воспроизводства ресурсов включает всего два ресурса, $p = 2$, а также исходно заданное значение $a = 1/2$. Можно рассматривать ограничение размерности как проектирование шаблона с $p > 2$ для выполнения удобного анализа экспертом создаваемого сценария «покоординатно».

Для запасов ресурсов R_1 и R_2 рассмотрим два случая:

(1) оба ресурса витальные;

(2) R_1 — витальный, R_2 — невитальный.

В иллюстративных расчетах приводятся примеры экспертных оценок и анализируются результаты обучения ЭС интегральному оцениванию обоими методами: МНК и МНМ.

Случай 1. Базовая функция $\Phi(x_1, x_2) = (x_1, x_2)^{1/2}$, обученный алгоритм представляет функция $f(X) = a_1(x_1x_2)^{1/2} + a_2x_1 + a_3x_2$.

В расчетах МНК $a_i = \hat{a}_i$, в расчетах МНМ $a_i = \tilde{a}_i$. Экспертные оценки и результаты расчетов представлены в табл. 1.

Таблица 1 Данные и результаты расчета для $\Phi(x_1, x_2) = (x_1\,x_2)^{1/2}$

x_1	x_2	$\Phi(x_1, x_2)$	I	$\lvert\Phi(x_1, x_2) - I\rvert$	$f(x_1, x_2)$, $a_i = \hat{a}_i$	$\lvert f(x_1, x_2) - I\rvert$, $a_i = \hat{a}_i$	$f(x_1, x_2)$, $a_i = \tilde{a}_i$	$\lvert f(x_1, x_2) - I\rvert$, $a_i = \tilde{a}_i$
0	100	0,0	0	0,0	16,8	16,8	41,3	41,3
100	0	0,0	0	0,0	−6,9	6,9	0,0	0,0
0	0	0,0	0	0,0	0,0	0,0	0,0	0,0
10	**100**	**31,6**	**60**	28,3	48,5	11,5	60,0	0,0
20	**100**	**44,7**	**70**	25,3	61,3	8,7	67,7	2,3
30	**100**	**54,8**	**80**	25,2	70,9	9,0	73,7	6,3
100	100	100	100	0,0	112,5	12,5	100,4	0,4
			$\sum \lvert\Phi(x_1, x_2) - I\rvert = 78{,}9$; $\sum (\Phi(x_1, x_2) - I)^2 = 2080{,}7$		$\sum \lvert\Phi(x_1, x_2) - I\rvert = 65{,}4$; $\sum (\Phi(x_1, x_2) - I)^2 = 775{,}0$; $\hat{a}_i = 1{,}02$; $\hat{a}_2 = -0{,}07$; $\hat{a}_3 = 0{,}17$		$\sum \lvert\Phi(x_1, x_2) - I\rvert = 50{,}3$; $\sum (\Phi(x_1, x_2) - I)^2 = 6778{,}3$; $\hat{a}_1 = 0{,}59$; $\hat{a}_2 = 0{,}0$; $\hat{a}_3 = 0{,}41$	

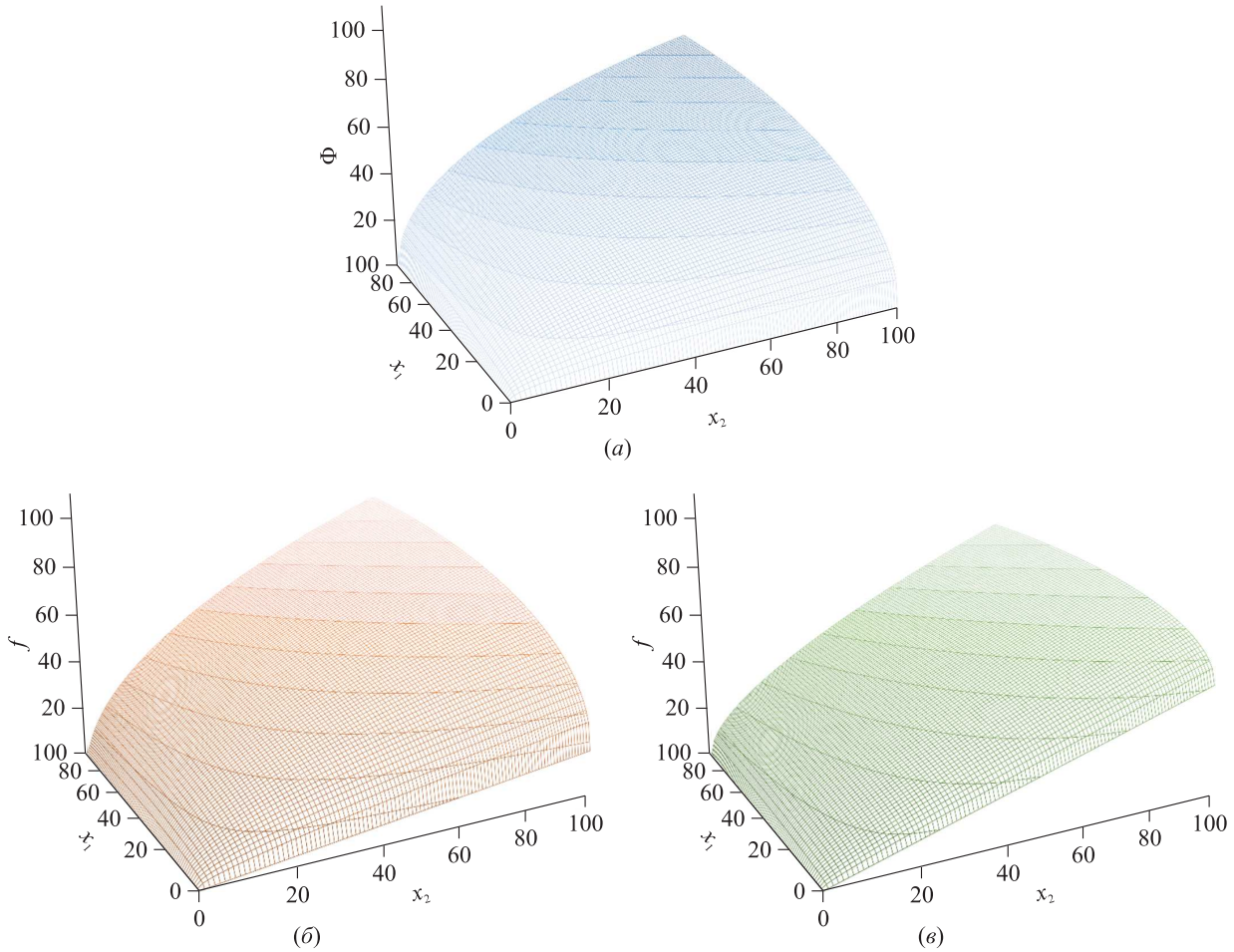


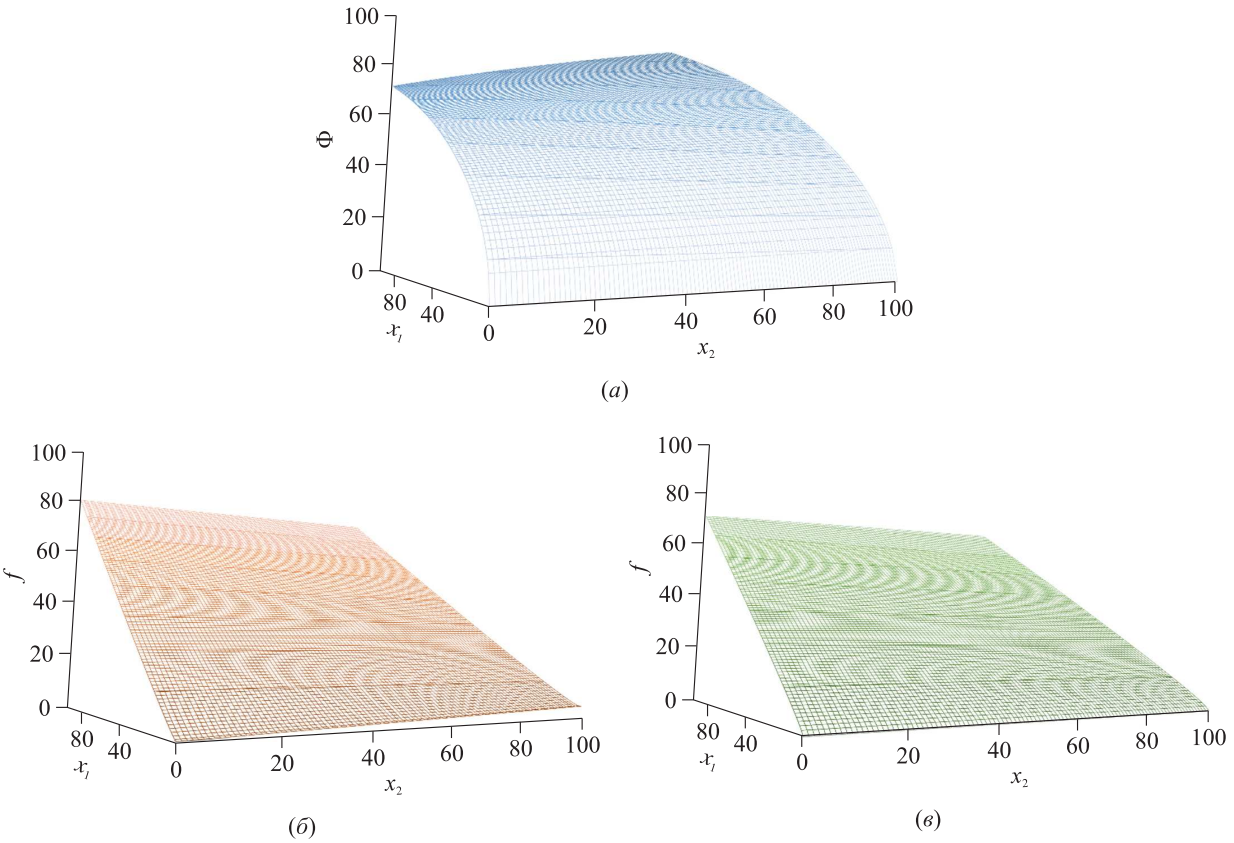
Рис. 2 Поверхности $\Phi(x_1, x_2) = (x_1x_2)^{1/2}$ (*а*), $f(x_1, x_2) = \hat{a}_1(x_1x_2)^{1/2} + \hat{a}_2x_1 + \hat{a}_3x_2$ (*б*) и $f(x_1, x_2) = \tilde{a}_1(x_1x_2)^{1/2} + \tilde{a}_2x_1 + \tilde{a}_3x_2$ (*в*)

Дополняющие базовую методику $\Phi(x_1, x_2)$ три экспертные оценки выделены, их смысл заключается в том, чтобы значительно поднять интегральную оценку в случаях, когда запас R_1 значительно меньше недельного. Для этого интегральные оценки I, отвечающие малым значениям нормированной оценки x_1, сделаны значительно больше, чем базовые $\Phi(x_1, x_2)$ при $x_2 = 100$. Потребоваться такое могло, например, чтобы учесть быструю воспроизводимость первого ресурса.

Визуализируют результаты поверхности, представленные на рис. 2. Заметим, что и табличные данные, и изображения результирующих поверхностей показывают более предпочтительный результат, полученный МНМ. Поверхность $f(x_1, x_2) = \tilde{a}_1(x_1x_2)^{1/2} + \tilde{a}_2x_1 + \tilde{a}_3x_2$ явно лучше отражает и заложенный изначально смысл интегральной оценки, и коррективы, внесенные экспертными оценками. Отдельного внимания заслуживает то, что даже нормировка интегральной оценки сохра-

Таблица 2 Данные и результаты расчета для $\Phi(x_1, x_2) = (x_1\,(x_2/2 + 50))^{1/2}$

x_1	x_2	$\Phi(x_1, x_2)$	I	$\lvert\Phi(x_1, x_2) - I\rvert$	$f(x_1, x_2)$, $a_i = \hat{a}_i$	$\lvert f(x_1, x_2) - I\rvert$, $a_i = \hat{a}_i$	$f(x_1, x_2)$, $a_i = \tilde{a}_i$	$\lvert f(x_1, x_2) - I\rvert$, $a_i = \tilde{a}_i$
0	100	0,0	0,0	0,0	2,9	2,9	0,0	0,0
100	0	70,7	70,7	0,0	79,2	8,5	70,7	0,0
0	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
60	**100**	**77,5**	**40**	37,5	47,7	7,7	45,1	5,1
70	**100**	**83,7**	**50**	33,7	55,7	5,7	52,1	2,1
80	**100**	**89,4**	**60**	29,4	63,7	3,7	59,1	0,9
100	100	100	100	0,0	79,9	20,1	73,0	27,0
			$\sum \lvert\Phi(x_1, x_2) - I\rvert = 100{,}6$; $\sum (\Phi(x_1, x_2) - I)^2 = 3403{,}5$		$\sum \lvert\Phi(x_1, x_2) - I\rvert = 48{,}8$; $\sum (\Phi(x_1, x_2) - I)^2 = 593{,}0$; $\hat{a}_i = -0{,}08$; $\hat{a}_2 = 0{,}85$; $\hat{a}_3 = 0{,}03$		$\sum \lvert\Phi(x_1, x_2) - I\rvert = 35{,}2$; $\sum (\Phi(x_1, x_2) - I)^2 = 762{,}7$; $\hat{a}_1 = 0{,}08$; $\hat{a}_2 = 0{,}65$; $\hat{a}_3 = 0{,}0$



(*а*) (*б*) (*в*)

Рис. 3 Поверхности $\Phi(x_1, x_2) = (x_1(x_2/2 + 50))^{1/2}$ (*а*), $f(x_1x_2) = \hat{a}_1(x_1(x_2/2 + 50))^{1/2} + \hat{a}_2x_1 + \hat{a}_3x_2$ (*б*) и $f(x_1x_2) = \tilde{a}_1(x_1(x_2/2 + 50))^{1/2} + \tilde{a}_2x_1 + \tilde{a}_3x_2$ (*в*)

нена без дополнительных усилий. Расчет МНК дает худший результат по той причине, что критерий заставляет поверхность точнее отрабатывать оценку $I(100, 100) = 100$, которую внесенные экспертные оценки сделали аномальной. Соответственно, МНМ с этой аномалией справился, а МНК — нет, вплоть до того что рассчитанная поверхность $f(x_1, x_2) = \hat{a}_1(x_1x_2)^{1/2} + \hat{a}_2x_1 + \hat{a}_3x_2$ допускает и отрицательные значения, и значения, большие 100.

Случай 2. Базовая функция $\Phi(x_1, x_2) = (x_1(x_2/2 + 50))^{1/2}$, обученный алгоритм представляет функция $f(X) = a_1(x_1(x_2/2 + 50))^{1/2} + a_2x_1 + a_3x_2$.

Экспертные оценки и результаты расчетов представлены в табл. 2.

Выделенные дополняющие базовую методику $\Phi(x_1, x_2)$ три экспертные оценки введены, чтобы значительно поднять интегральную оценку в случаях, когда запас R_1 значительно меньше недельного. Для этого интегральные оценки I, отвечающие малым значениям нормированной оценки x_1, сделаны значительно больше, чем базовые $\Phi(x_1, x_2)$ при $x_2 = 100$. Потребоваться такое могло, например, чтобы учесть быструю воспроизводимость первого ресурса.

Визуализируют результаты поверхности, представленные на рис. 3. Альтернативой случаю 1 приведенные здесь данные показывают более предпочтительный результат, полученный МНК, так как здесь нет явной аномалии.

5 Заключение

Представленная в статье ЭС — это, конечно, инициативный проект, а не конечный продукт, который заинтересует потенциального пользователя. Проект целенаправленно создавался как инструментальное решение. В этом его принципиальное отличие от традиционных ЭС, основной недостаток которых — узкая специализация. Числовой материал, рисунки, которые включены в текст, — это результаты экспериментов, выполненных с модельными материалами, например с данными Роспотребнадзора и Минздрава по ситуации с пандемией COVID-19. Так что опыт, подтверждающий адаптируемость представленной ЭС к конкретным задачам, есть, и он положительный.

Литература

1. *McConnell C. R., Brue S. L., Flynn S. M.* Economics: Principles, problems, and policies. — 19th ed. — New York, NY, USA: McGraw-Hill/Irwin, 2011. 896 p.
2. *Black J., Nigar H., Gareth M. A.* Dictionary of economics. — 5th ed. — Oxford, U.K.: Oxford University Press, 2017. 584 p.
3. *Angus D.* Understanding consumption. — Oxford, U.K.: Oxford University Press, 1992. 256 p.
4. *Jackson P.* Introduction to expert systems. — Reading, MA, USA: Addison-Wesley, 1999. 542 p.
5. *Giarratano J. C., Riley G. D.* Expert systems: Principles and programming. — 4th ed. — Course Technology, 2004. 288 p.
6. *Haenlein M., Kaplan A. A.* Brief history of artificial intelligence: On the past, present, and future of artificial intelligence // Calif. Manage. Rev., 2019. Vol. 61. No. 4. P. 5–14.
7. *Russell S., Norvig P.* Artificial intelligence: A modern approach. — New York, NY, USA: Pearson, 2020. 1136 p.
8. The National Artificial Intelligence Research and Development Strategic Plan. — National Science and Technology Council, Networking and Information Research and Development Subcommittee, 2016. https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf.
9. *Mitchell T. M.* Machine learning. — New York, NY, USA: McGraw Hill, 1997. 432 p.
10. *Гаврилова Т. А., Хорошевский В. Ф.* Базы знаний интеллектуальных систем. — СПб.: Питер, 2000. 384 с.
11. *Кузенкова Г. В.* Введение в экологический мониторинг. — Н. Новгород: НФ УРАО, 2002. 72 с.
12. *Дмитриков В. П., Дудников И. А.* Экспертные системы как составная часть экологического мониторинга // Успехи современного естествознания, 2014. № 9-2. С. 170–172.
13. *Буров А. Н., Калабин А. Л., Козлов А. В., Пакшвер Э. А.* Экспертная система мониторинга технологического процесса // Программные продукты и системы, 2015. № 2. С. 39–43.
14. *Жернаков С. В.* Активная экспертная система мониторинга и управления ремонтом авиационных газотурбинных двигателей // Автоматизация и современные технологии, 2001. № 6. С. 16–21.
15. *Самойлова Е. М.* Построение экспертной системы поддержки принятия решения как интеллектуальной составляющей системы мониторинга технологического процесса // Вестник Пермского национального исследовательского политехнического университета. Машиностроение, материаловедение, 2016. Т. 18. № 2. С. 128–142.
16. *Смагин А. А., Медведев Д. М., Мельниченко А. С., Липатова С. В., Рудковский Ю. А.* Разработка программного комплекса экспертной системы морского мониторинга // Автоматизация процессов управления, 2008. Т. 12. № 2. С. 56–68.
17. *Арефьева Е. В., Бабусенко М. С., Барышев Е. М. и др.* Проблемы защиты населения и территорий в чрезвычайных ситуациях в условиях современных вызовов и угроз: Справочное пособие. — М.: ВНИИ ГОЧС, 2017. 452 с.

18. *Yanase J., Triantaphyllou E.* A systematic survey of computer-aided diagnosis in medicine: Past and present developments // Expert Syst. Appl., 2019. Vol. 138. Art. 112821. 25 p.
19. ГОСТ Р 22.3.01-94 Безопасность в чрезвычайных ситуациях. Жизнеобеспечение населения в чрезвычайных ситуациях. Общие требования. — М.: Изд-во стандартов, 1994. 11 с.
20. *Albert A.* Regression, and the Moore–Penrose pseudoinverse. — New York, NY, USA: Academic Press, 1972. 179 p.
21. *Мудров В. И., Кушко В. Л.* Метод наименьших модулей. — 2-е изд. — М.: URSS, 2013. 57 с.
22. *Schlossmacher E. J.* An iterative technique for absolute deviations curve fitting // J. Am. Stat. Assoc., 1973. Vol. 68. No. 344. P. 857–859.

Поступила в редакцию 04.06.2021

EXPERT SYSTEM FOR MONITORING AND FORECASTING OF RESOURCE ALLOCATION PROCESSES

A. V. Bosov and D. V. Zhukov

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

Abstract: The paper presents a project of an expert monitoring system designed to support decision-making in managing the processes of distribution (consumption and reproduction) of resources. The results of the analysis of the consumption process are presented in the form of scenarios for the integral assessment of its state. Scenarios are prepared by an expert, are situational in nature and are used in real calculations both to assess the current state and to predict the development of situations. The traditional interpretation of the consumption process is based on the concept of resources and subjects of consumption who consume resources in accordance with consumption rates, as well as a simple description of the reproduction of resources by production objects. It is assumed that the components of the information model are geographically and temporally referenced, in particular, data on resource reserves are linked to time. The main intellectual load is carried by the methodology for preparing scenarios for the integral assessment of the state. This technique is based on the ideology of expert evaluation of typical situations and the formation of calculation scenarios using simple machine learning methods. The latter use widespread approaches to optimization — minimization of mean squares and modules. The presented project of the expert system is instrumental in nature and can be used in various applications. The process of preparing and evaluating the effectiveness of the scenarios of integral assessment prepared by the expert is illustrated by numerical and graphic material.

Keywords: expert system; resources and consumption; machine learning; least squares method; least modules method

DOI: 10.14357/19922264210305

References

1. McConnell, C. R., S. L. Brue, and S. M. Flynn. 2011. *Economics: Principles, problems, and policies*. 19th ed. New York, NY: McGraw-Hill/Irwin. 896 p.
2. Black, J., H. Nigar, and M. A. Gareth. 2017. *Dictionary of economics*. 5th ed. Oxford, U.K.: Oxford University Press. 584 p.
3. Angus, D. 1992. *Understanding consumption*. Oxford, U.K.: Oxford University Press. 256 p.
4. Jackson, P. 1999. *Introduction to expert systems*. Reading, MA: Addison-Wesley. 542 p.
5. Giarratano, J. C., and G. D. Riley. 2004. *Expert systems: Principles and programming*. 4th ed. Course Technology. 288 p.
6. Haenlein, M., and A. A. Kaplan. 2019. Brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *Calif. Manage. Rev.* 61(4):5–14.
7. Russell, S., and P. Norvig. 2020. *Artificial intelligence: A modern approach*. New York, NY: Pearson. 1136 p.
8. National Science and Technology Council, Networking and Information Research and Development Subcommittee. 2016. The National Artificial Intelligence Research and Development Strategic Plan. Available at: https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf (accessed June 23, 2021).
9. Mitchell, T. M. 1997. *Machine learning*. New York, NY: McGraw Hill. 432 p.
10. Gavrilova, T. A., and V. F. Khoroshevskiy. 2000. *Bazy znaniy intellektual'nykh sistem* [Knowledge bases of intelligent systems]. St. Petersburg: Piter. 384 p.
11. Kuzenkova, G. V. 2002. *Vvedenie v ekologicheskiy monitoring* [Introduction to environmental monitoring]. N. Novgorod: NF URAO. 72 p.

12. Dmitrikov, V. P., and I. A. Dudnikov. 2014. Ekspertnye sistemy kak sostavnaya chast' ekologicheskogo monitoringa [Expert systems as component part of ecological monitoring]. *Uspekhi sovremennogo estestvoznaniya* [Successes of Modern Natural Science] 9-2:170–172.
13. Burov, A. N., A. L. Kalabin, A. V. Kozlov, and E. A. Pakshver. 2015. Ekspertnaya sistema monitoringa tekhnologicheskogo protsessa [Expert system for monitoring of technological processes]. *Programmnye produkty i sistemy* [Software & Systems] 2:39–43.
14. Zhernakov, S. V. 2001. Aktivnaya ekspertnaya sistema monitoringa i upravleniya remontom aviatsionnykh gazoturbinnykh dvigateley [An active expert system for monitoring and managing the repair of aircraft gas turbine engines]. *Avtomatizatsiya i sovremennye tekhnologii* [Automation and Modern Technologies] 6:16–21.
15. Samoylova, E. M. 2016. Postroenie ekspertnoy sistemy podderzhki prinyatiya resheniya kak intellektual'noy sostavlyayushchey sistemy monitoringa tekhnologicheskogo protsessa [Construction of an expert decision support system as an intellectual component of a technological process monitoring system]. *Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Mashinostroenie, materialovedenie* [Bulletin of the Perm National Research Polytechnic University. Mechanical Engineering, Materials Science] 18(2):128–142.
16. Smagin, A. A., D. M. Medvedev, A. S. Mel'nichenko, Lipatova, S. V., and Yu. A. Rudkovskiy. 2008. Razrabotka programmnogo kompleksa ekspertnoy sistemy morskogo monitoringa [Development of the software complex for the expert system of marine monitoring]. *Avtomatizatsiya protsessov upravleniya* [Automation of Control Processes] 12(2):56–68.
17. Aref'eva, E. V., M. S. Babusenko, E. M. Baryshev, *et al.* 2017. *Problemy zashchity naseleniya i territoriy v chrezvychaynykh situatsiyakh v usloviyakh sovremennykh vyzovov i ugroz: spravochnoye posobiye* [Problems of protection of the population and territories in emergency situations in the face of modern challenges and threats: Reference book]. Moscow: VNII GOChS. 452 p.
18. Yanase, J., and E. Triantaphyllo. 2019. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Syst. Appl.* 138:112821. 25 p.
19. GOST R 22.3.01-94. 1994. Bezopasnost' v chrezvychaynykh situatsiyakh. Zhizneobespechenie naseleniya v chrezvychaynykh situatsiyakh. Obshchie trebovaniya [Safety in emergency situations. Life support of the population in emergency situations. General requirements]. Moscow: Izd-vo Standartov. 11 p.
20. Albert, A. 1972. *Regression and the Moore–Penrose pseudoinverse*. New York, NY: Academic Press. 179 p.
21. Mudrov, V. I., and V. L. Kushko. 2013. *Metod naimen'shikh moduley* [Method of least modules]. 2nd ed. Moscow: URSS. 57 p.
22. Schlossmacher, E. J. 1973. An iterative technique for absolute deviations curve fitting. *J. Am. Stat. Assoc.* 68(344):857–859.

Received June 4, 2021

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

Zhukov Denis V. (b. 1979) — 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; DZhukov@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 41–50

ДИСПЕТЧЕРИЗАЦИЯ В СИСТЕМЕ С ПАРАЛЛЕЛЬНЫМ ОБСЛУЖИВАНИЕМ С ПОМОЩЬЮ РАСПРЕДЕЛЕННОГО ГРАДИЕНТНОГО УПРАВЛЕНИЯ МАРКОВСКОЙ ЦЕПЬЮ*

М. Г. Коновалов[1], Р. В. Разумчик[2]

Аннотация: Рассматривается система распределения заданий, состоящая из диспетчера и нескольких параллельно работающих однопроцессорных серверов, каждый из которых имеет очередь неограниченной емкости. Задания без внутренней структуры поступают в систему по рекуррентному потоку и имеют случайный размер. В момент поступления очередное задание в соответствии с некоторой стратегией направляется диспетчером, имеющим полную информацию о текущем состоянии системы, в один из серверов. Задания обслуживаются из очереди по одному в соответствии с дисциплиной FIFO (first in, first out), после чего навсегда покидают систему. Скорости процессоров фиксированы и, вообще говоря, различны. Ставится задача нахождения стратегии, минимизирующей стационарное среднее время пребывания задания в системе. Излагается способ решения подобного класса задач, основанный на использовании метода Монте-Карло в сочетании с оригинальным адаптивным алгоритмом управления марковскими цепями с континуальным множеством состояний. Его главная составляющая — динамическая дискретизация множества состояний и использование в каждой ячейке разбиения «локального» квазиградиентного алгоритма. На простых численных примерах показано, что построенная диспетчеризация успешно конкурирует с лучшими из известных решений, а сам метод рассчитан на более широкую область применения.

Ключевые слова: гетерогенная система с параллельным обслуживанием; управление марковской цепью с несчетным множеством состояний; оптимизация времени пребывания

DOI: 10.14357/19922264210306

1 Введение

В статье рассматривается модель системы распределения заданий, с которой часто приходится сталкиваться при решении задач планирования ресурсов и которая, несмотря на свою простоту, остается предметом активных научных исследований [1]. Система состоит из нескольких узлов, работающих параллельно и независимо. Каждый узел представляет собой обслуживающий прибор (процессор) с неограниченной очередью для хранения поступающих к нему заданий, которые выполняются по одному в соответствии с дисциплиной FIFO[3]. В систему поступает рекуррентный поток однородных заданий. В момент поступления очередного задания становится известным его размер, вообще говоря случайный, и диспетчером в автоматическом или ручном режиме принимается решение о направлении на обслуживание в один из узлов. Для выбора решения может быть использована любая информация о текущем и прошлом состоянии системы[4]. После окончания обслуживания задание покидает систему. Каждое задание обслуживается в узле некоторое время, которое зависит от размера задания и от (известной диспетчеру) скорости процессора в данном узле. Задача заключается в нахождении такой стратегии диспетчера, которая обеспечивала бы минимальное среднее время пребывания задания в системе.

Ключевым моментом для решения поставленной задачи является то, что динамика системы описывается как марковский процесс принятия решений, причем диспетчеризация заданий является стратегией управления, а время пребывания заданий играет роль одношаговых доходов. Несмотря на простоту и «прозрачность» аналитического описания, за исключением частных случаев (см., например, [2]), удовлетворительного и универсального прикладного решения задачи пока не найдено. Причина кроется в том, что сочетания естественных предположений о входящем потоке и длительностях обслуживания приводят к трудностям (главная

*Работа выполнена при поддержке РФФИ (проект 20-07-00804).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, mkonovalov@ipiran.ru

[2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, rrazumchik@ipiran.ru

[3]Для излагаемого решения это предположение, вообще говоря, несущественно и в узлах может использоваться любая консервативная дисциплина обслуживания.

[4]Предлагаемый подход позволяет ослабить и это предположение, например на случай, когда размер задания заранее не известен.

из которых, пожалуй, — несчетность множества состояний), преодолеть которые, следуя какой-то единой схеме, не удается. Естественным решением для диспетчеризации в рассматриваемой общей системе выглядело бы использование простого в реализации правила[1], оптимальность которого известна в частных случаях. Однако, как показывают численные эксперименты, вместе с условиями видоизменяется и структура оптимальной стратегии. Например, для пороговой стратегии[2] (см. [2, 4]), в зависимости от вида задающих процесс функций распределения в пространстве состояний (которое в данном случае является трехмерным), для каждого узла может существовать более чем одна связная область, в которой оптимальным решением является выбор этого узла (см. [5, Fig. 4]).

Вернемся к исходной постановке. С принципиальной точки зрения для решения задачи достаточна методология динамического программирования. Искомую стратегию можно приближенно найти из соответствующих уравнений (см., например, [6, уравнения (8.46)]). Один из методов их решения — дискретизация множества состояний. Однако обычно с повышением требований к точности решения слишком быстро возрастает сложность решающих алгоритмов [7–9]. Другой метод — аппроксимировать целевую функцию некоторой простой известной (суррогатной) функцией, зависящей от одного или нескольких параметров (см., например, [10–12]). Однако нестабильность и потенциальная расходимость — основные недостатки реализующих такой метод[3] алгоритмов [13]. Известны также и другие, специальные подходы (см., например, [14, 15]), позволяющие «сглаживать» некоторые недостатки разработанных методов, при этом зачастую требуется накладывать такие дополнительные ограничения (например, возможность факторизации многомерных переходных плотностей), которые сужают область их применения.

Предлагаемое в этой статье общее решение сформулированной в начале раздела задачи не опирается на методологию динамического программирования, а получается за счет сочетания метода Монте-Карло и адаптивных алгоритмов управления частично наблюдаемыми марковскими цепями[4] [18, 19]. В его основе лежит динамическая дискретизация множества состояний с переменным количеством, размером и составом элементов разбиения. В каждой подобласти управление (выбор сервера) происходит с помощью отдельного вероятностного распределения. Коррекция значений вероятностей осуществляется (независимо) с помощью градиентного алгоритма, оценивающего производную целевой функции по результатам наблюдений за имитируемой траекторией. Не приходится рассчитывать, что получающееся таким образом решение окажется в каждом конкретном случае оптимальным. Однако, как показывают численные эксперименты, оно, во-первых, успешно конкурирует с другими известными решениями, а во-вторых, в отличие от последних, не требует специальных ограничений ни на входящий поток, ни на процесс обслуживания.

В разд. 2 приводится формальная постановка задачи. Затем в разд. 3 излагаются соображения, на которых основан алгоритм оптимизации. Здесь же описывается и идея, лежащая в основе его численной реализации. Несколько иллюстрирующих основные положения статьи примеров приведены в разд. 4. В последнем разделе резюмируются основные результаты и затрагиваются некоторые открытые вопросы.

2 Процесс диспетчеризации в системе с параллельным обслуживанием как управляемая марковская цепь

Система состоит из K независимых серверов, каждый из которых может выполнять единовременно не более одного задания. Производительности серверов постоянны, однако, вообще говоря, различны. Задания поступают в систему в виде рекуррентного потока, причем интервал между поступлениями определяется функцией распределения F. Задания имеют случайный размер с функцией распределения G. Каждое поступившее задание мгновенно направляется на один из серверов, где либо сразу же начинает выполняться без прерывания, если сервер свободен, либо ставится в очередь, емкость которой предполагается бесконечной. Переходы заданий между очередями исключены. Об-

[1] См., например, [2, 3].

[2] Отметим, что такие популярные правила, как SITA (size interval task assignment), JSQ (join the shortest queue) (и ее разновидности JSQ(d), HJSQ(d)) и т. п., вообще могут быть далеки от оптимальных. Этот эффект особенно заметен, когда распределения времен обслуживания имеют «тяжелый хвост». Подобный «перекос» характерен, по-видимому, для большинства статических и динамических (неадаптивных) правил принятия решений.

[3] По другим причинам, но это же замечание относится и к предложенному в статье методу.

[4] Новизна же заключается не в идее использования метода статистических испытаний в сочетании с адаптивными алгоритмами (она хорошо известна из литературы [16, 17]), а в способе реализации — путем распределенного управления цепью Маркова с континуальным множеством состояний с помощью переменного набора адаптивных алгоритмов.

служивание очередей происходит согласно дисциплине FIFO.

Рассмотрим функционирование системы в дискретные моменты поступления заданий, которые перенумеруем как $n = 0, 1, 2, \ldots$ Пусть $\chi_n^{(k)}$ означает индикатор события

$$A_n^{(k)} = \{\text{задание, поступившее в момент } n, \text{ направлено на сервер } k\}$$

(очевидно, $\sum_{k=1}^{K} \chi_n^{(k)} = 1$).

Обозначим через $x_n^{(i)}$ объем работы на сервере i в момент n, когда задание поступило, но еще не было распределено на какой-нибудь сервер. Этот объем складывается за счет задания, которое, возможно, выполняется, а также за счет заданий, возможно, находящихся в очереди. Для определенности можно считать, что в начальный момент система не содержит заданий, т. е. $x_0^{(k)} = 0$. Через $\tilde{x}_n^{(k)}$ обозначим ту же величину, но уже после выбора сервера. Очевидно, что

$$\tilde{x}_n^{(k)} = x_n^{(k)} + \chi_n^{(k)} \xi_n \,,$$

где ξ_n — размер задания, поступившего в момент n.

Далее, к моменту $n+1$ следующего поступления задания имеем очевидное соотношение:

$$x_{n+1}^{(k)} = \max\left\{0, \tilde{x}_n^{(k)} - r^{(k)}\tau_n\right\} \,,$$

где $r^{(k)}\tau_n$ — работа, выполненная сервером k, имеющим производительность $r^{(k)}$, за время τ_n, прошедшее между поступлениями заданий с номерами n и $n+1$. (Величины τ_n и ξ_n независимы между собой и от других компонент процесса и имеют распределения соответственно F и G.)

Таким образом, динамика процесса выполнения заданий описывается рекуррентными уравнениями Линдли:

$$x_{n+1}^{(k)} = \max\left\{0, x_n^{(k)} + \chi_n^{(k)}\xi_n - r^{(k)}\tau_n\right\} \,, \quad k = 1, \ldots, K \,. \quad (1)$$

Вектор $\chi_n = \chi_n^{(k)}$ играет роль управления, примененного в момент n, поскольку наступление событий $A_n^{(k)}$ определяется внешним, управляющим действием диспетчера. Будем считать, что это действие — выбор сервера в момент n — реализуется с помощью вероятностного распределения на множестве $\{1, \ldots, K\}$:

$$\sigma_n = \left(\sigma_n^{(k)}, k = 1, \ldots, K\right) ,$$

так что событие $A_n^{(k)}$ наступает с вероятностью $\sigma_n^{(k)}$.

Определим вектор

$$x_n = \left(x_n^{(1)}, \ldots, x_n^{(k)}, \xi_n\right) ,$$

который назовем состоянием системы в момент n. Легко понять, что значение вектора x_n полностью определяется значениями векторов x_{n-1} и χ_n, т. е. предыдущим состоянием и выбранным управлением. Это означает, что последовательность x_n, определенная системой (1), представляет собой управляемую марковскую цепь, а последовательность $\sigma_0, \sigma_1, \sigma_2, \ldots$ является стратегией управления. Компоненты стратегии могут в общем случае быть устроены по-разному и зависеть от сколь угодно далекой предыстории, однако особую роль играют так называемые однородные марковские стратегии. Для таких стратегий все «правила» σ_n устроены одинаково, причем таким образом, что вероятность выбора сервера зависит только от текущего состояния: $\sigma_n = \sigma(x_n)$.

Для управляемой марковской цепи x_n естественным образом определяется «одношаговый доход» θ_n, который задан в терминах времени пребывания задания в системе. С учетом заданной дисциплины обслуживания очереди время пребывания задания, поступившего в систему в момент n и направленного на сервер k, есть

$$\theta_n = x_n^k + \frac{\xi_n}{r^{(k)}} \,.$$

Рассматривая процесс на бесконечном интервале времени, предположим, что существует непустое подмножество Σ однородных марковских стратегий таких, что предел

$$W(\sigma) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} M_\sigma \theta_n$$

конечен для всех $\sigma \in \Sigma$ (M_σ — математическое ожидание относительно стратегии σ). Заметим, что для рассматриваемой системы вопрос о существовании конечного предела решается, как правило, достаточно легко и связан с условием, чтобы нагрузка на серверы была меньше единицы.

Величина $W(\sigma)$ является, очевидно, критерием качества стратегии σ с точки зрения среднего времени выполнения заданий, поэтому цель управления формулируется как отыскание минимума функции $W(\sigma)$ на множестве Σ. Заметим еще, что ограничение области поиска множеством Σ не сужает возможности минимизации. Из общей теории известно (см., например, [6, с. 253]), что стратегии, учитывающие зависимость от более глубокой предыстории, не улучшают значение $W(\sigma)$ по сравнению с однородными марковскими стратегиями.

3 Идея построения алгоритма оптимизации

Пусть X — некоторое компактное подмножество $\mathbb{R}^m$. Управляемая марковская цепь с дискретным временем $n = 0, 1, 2, \ldots$ и с множеством состояний X задана конечным семейством переходных плотностей $p_k(x, y)$, $x, y \in X$, $k = 1, \ldots, K$.

Параметр k играет роль управления и его значение подлежит выбору на каждом шаге n.

Рассмотрим множество $\Sigma_{\text{o}} = \{\sigma\}$ однородных по времени стратегий управления, имеющих вид:

$$\sigma : \; X \to \mathcal{K}\,,$$

где $\mathcal{K}$ — совокупность всех вероятностных распределений на множестве $\{1, \ldots, K\}$. Применение стратегии $\sigma \in \Sigma_{\text{o}}$ означает, что управление k выбирается каждый раз с фиксированной вероятностью $\sigma_k = \sigma_k(x)$, зависящей от текущего состояния цепи x. Эта вероятность является одной из компонент данной стратегии, т. е.

$$\sigma = \sigma(x) = \{\sigma_k(x), x \in X, k = 1, \ldots, K\}\,.$$

Пусть выбрана стратегия $\sigma \in \Sigma_{\text{o}}$. Тогда плотность переходной вероятности выражается как

$$p(x, y) = \sum_{k=1}^{K} \sigma_k(x) p_k(x, y) \tag{2}$$

и, естественно, зависит от σ, т. е. $p(x, y) = p_\sigma(x, y)$. Плотность переходной вероятности за n шагов имеет вид

$$p^{(n)}(x, y) = \int\limits_X p^{(n-1)}(x, z) p(z, y)\, dz\,, \quad n = 1, 2, \ldots$$

и также зависит от выбранной стратегии. Обозначим через $\mathsf{P} = \mathsf{P}(x, A)$ и $\mathsf{P}^n = \mathsf{P}^n(x, A)$ переходные вероятности, порожденные соответственно плотностями p и $p^{(n)}$:

$$\left.\begin{aligned} \mathsf{P}(x, A) &= \int\limits_A p(x, y)\, dy\,; \\ \mathsf{P}^n(x, A) &= \int\limits_A p^{(n)}(x, y)\, dy\,, \quad x \in X\,, \end{aligned}\right\} \tag{3}$$

где $A \subset X$ — борелевское множество. Разобьем множество состояний цепи на непересекающиеся подмножества:

$$X = \bigcup_{i=1}^{N} X_i\,, \quad X_i \cap X_j = \varnothing\,, \quad i \neq j\,, \tag{4}$$

и обозначим через 1_i индикатор множества X_i. Будем далее рассматривать подмножество однородных стратегий $\Sigma \subset \Sigma_\sigma$ с компонентами вида

$$\sigma_k(x) = \sum_{i=1}^{N} \sigma_{ik} 1_i(x)\,,$$

где параметр σ_{ik} означает вероятность выбора управления k при условии, что цепь находится в состоянии $x \in X_i$. Ясно, что $\sigma_{ik} \geq 0$ и $\sum_{k=1}^{K} \sigma_{ik} = 1$. Для дальнейшего удобно исключить компоненты σ_{i1} по формуле:

$$\sigma_{i1} = 1 - \sum_{k=2}^{K} \eta_{ik}\,,$$

где $\eta_{ik} = \sigma_{ik}$, $k > 1$. Выражение (2) для плотности переходной вероятности теперь запишется в виде:

$$p(x, y) = p_1(x, y) + \\ + \sum_{i=1}^{N} \sum_{k=2}^{K} \eta_{ik} 1_i(x) \left(p_k(x, y) - p_1(x, y)\right). \tag{5}$$

Стратегии из Σ, так же как и соответствующие им переходные вероятности, оказываются зависящими от матрицы $\eta = \{\eta_{ik}\}$ размера $N \times (K - 1)$, элементы которой принимают значения из множества

$$H = \left\{ \eta_{ik} : \; \sum_{k=2}^{K} \eta_{ik} < 1\,; \; \eta_{ik} \geq 0\,; \right. \\ \left. i = 1, \ldots, N\,; k = 2, \ldots, K \right\}. \tag{6}$$

Предположим, что для любой стратегии из множества Σ существует предельное (стационарное) распределение π цепи Σ, $\pi(A) = \lim_{n\to\infty} \mathsf{P}^n(x, A)$, не зависящее от $x \in \mathbb{R}^m$ и удовлетворяющее уравнению

$$\pi(I - \mathsf{P}) = 0\,, \tag{7}$$

где I — единичный оператор, а $(\pi\mathsf{P})(A) = \int_X \pi(dx)\mathsf{P}(x, A)$.

Ясно, что мера π, так же как и переходная вероятность P, зависит от матрицы η. Зафиксируем индексы α и β и будем обозначать штрихом операцию дифференцирования по переменной $\eta_{\alpha\beta}$. Тогда результат дифференцирования равенства (7) запишется в виде:

$$\pi'(I - \mathsf{P}) - \pi\mathsf{P}' = 0\,.$$

Можно показать, что из этой формулы следует соотношение

$$\pi' = \pi\mathsf{P}' \left(I + \mathsf{P} + \mathsf{P}^2 + \cdots\right), \tag{8}$$

причем (формальное) дифференцирование P с использованием (3) и (5) дает

$$(\mathsf{P}(x,A))' = 1_\alpha(x)\int\limits_A (p_\beta(x,y) - p_1(x,y))\,dy\,. \quad (9)$$

Пусть на множестве состояний цепи задана конечнозначная функция $g:\ X \to \mathbb{R}$, которая интерпретируется как «доход». Предельный средний доход, получаемый от управления марковской цепью с помощью стратегии $\sigma \in \Sigma$, определяется как

$$W = W(\eta) = \int\limits_X g(x)\pi(dx)\,.$$

Формулы (8) и (9) позволяют формально выписать выражение для частной производной W по переменной $\eta_{\alpha\beta}$:

$$\begin{aligned} W' &= \int\limits_X g(x)\pi'(dx) = \\ &= \sum_{n=0}^{\infty} \int\limits_{x\in X} g(x) \int\limits_{u\in X_\alpha} \pi(du) \int\limits_{v\in X} (p_\beta(u,v) - p_1(u,v)) \times \\ &\qquad \times \mathsf{P}^n(v,dx)\,dv = \sum_{n=0}^{\infty} \left(G^{(n)}_{\alpha\beta} - G^{(n)}_{\alpha 1}\right), \end{aligned} \quad (10)$$

где выражение

$$G^{(n)}_{\alpha\beta} = \int\limits_{x\in X} g(x) \int\limits_{u\in X_\alpha} \pi(du) \int\limits_{v\in X} p_\beta(u,v)\mathsf{P}^n(v,dx)\,dv$$

имеет смысл предельного среднего дохода на шаге $n+1$ после попадания в множество X_α и выбора управления β.

Схема максимизации функции $W(\eta)$ может быть основана на коррекции матрицы η в процессе эволюции марковской цепи методом проекции (квази)градиента:

$$\eta\,(n_{t+1}) = \Pi_H\left[\eta(n_t) + a_t\widehat{\nabla W}\,(\eta(n_t))\right], \qquad t = 0,1,2,\ldots$$

В этой формуле:

- n_t — возрастающая подпоследовательность натуральных чисел (моменты изменения матрицы η; в промежутках между этими моментами матрица η не изменяется);
- $\eta(n_t)$ — матрица η в момент n_t;
- $\widehat{\nabla W}(\eta(n_t))$ — оценка градиента функции W с помощью формулы (10) по наблюдениям до момента n_t;
- a_t — убывающая к нулю числовая последовательность;
- $\Pi_H[\cdot]$ — оператор проектирования на определенное формулой (6) выпуклое множество H.

Имеются две принципиальные трудности в реализации этой схемы, и обе они связаны с оценкой градиента:

(1) формула (10) для частной производной по компоненте матрицы η содержит бесконечное число слагаемых;

(2) каждое слагаемое в (10) представляет собой усреднение по предельному распределению, соответствующему фиксированному значению η.

Оба обстоятельства «стимулируют» увеличение промежутков между моментами изменения матрицы η для накопления статистики, что снижает скорость коррекции. Для преодоления указанных трудностей ниже используются «оценки с забыванием», которые ранее были использованы для решения задачи управления конечными и счетными цепями [19]. Далее следует краткое описание одной из возможных реализаций оптимизационного алгоритма.

Пусть $\mathcal{X} = \{X_1,\ldots,X_N\}$ — разбиение (4) пространства состояний и пусть X_i — произвольный элемент разбиения. Элементу X_i сопоставляется вектор управляющих вероятностей $\eta_i = (\eta_{i2},\ldots,\eta_{iK})$, а также вектор оценок этих вероятностей $G_i = (G_{i2},\ldots,G_{iK})$. Оба вектора трансформируются в процессе управления. Вектор η_i в определенные моменты времени изменяется следующим образом:

$$\eta_i := \prod_{H_i}^{\varepsilon} [\eta_i + aG_i]\ ,$$

где $\prod_{H_i}^{\varepsilon}$ — оператор проектирования на множество

$$H_i = \left\{\eta_{ik} : \sum_{k=2}^{K} \eta_{ik} < 1\,,\ \eta_{ik} \ge \varepsilon,\ k = 2,\ldots,K\right\}, \qquad a > 0\,,\ \varepsilon > 0\,.$$

Оценки корректируются следующим образом. Пусть в некоторый момент τ цепь в l-й раз с начала процесса попадает в состояние $x \in X_i$, и пусть при этом (с вероятностью η_{ik}) применяется управление k. Начиная с момента $\tau+1$, в течение $d > 0$ тактов накапливается суммарный одношаговый доход g_{ld}, после чего происходит изменение одной из оценок:

$$G_{ik} := \frac{bG_{ik} + g_{ld}}{1 + b + b^2 + \cdots + b^l}\,,$$

где $b \lesssim 1$. Параметры алгоритма a, b, d и ε, в принципе, также могут корректироваться в процессе управления.

Выбор моментов изменения векторов η_i может осуществляться разными способами. В алгоритме, который применялся в описанных ниже экспериментах, эти изменения осуществлялись синхронно с моментами коррекции вектора оценок G_i.

Прежде чем применять предложенною схему оптимизации предельного среднего дохода, необходимо задать разбиение пространства состояний управляемой марковской цепи. Можно предложить разнообразные способы дискретизации (см., например, [9]), самый простой из которых в идейном плане — это равномерная статическая прямоугольная сетка. Метод динамического, пространственно неоднородного разбиения[1], который кратко будет описан ниже, более сложный, однако, судя по вычислительным экспериментам, оказывается тем самым «правильным углом атаки» для рассматриваемого типа задач.

Для пояснения способа разбиения ограничимся рассмотрением случая одномерного пространства состояний, в качестве которого возьмем отрезок числовой прямой $X = [a, b]$. Пусть $\mathcal{X} = \{X_1, \ldots, X_N\}$ — разбиение (10). В начальный момент $N = 1$ и $X_1 = X$. В некоторый момент происходит разбиение отрезка X_1 на отрезок $X_{10} = [a, c]$ и полуинтервал $X_{11} = (c, b]$, причем точка c выбирается случайно на отрезке $[a, b]$. Новое разбиение пространства состояний имеет вид:

$$\mathcal{X} = \{X_1 = X_{10}, X_2 = X_{11}\}.$$

Далее если на некотором этапе имеется разбиение $\mathcal{X}$, то любой его элемент, скажем X_i, может «разделиться» на два «соседних» непересекающихся подмножества: X_{i0} и X_{i1}. Число элементов в $\mathcal{X}$ при этом увеличится на 1. Параллельно с процессом дробления идет процесс «склеивания», в котором два соседних элемента разбиения (типа только что отмеченных X_{i0} и X_{i1}) заменяются на исходный элемент (X_i).

Динамическую структуру разбиения удобно отображать с помощью направленного дерева с переменным множеством вершин и дуг. С каждой вершиной v связан элемент разбиения X_v. Корневой вершине соответствует отрезок X. Из каждой вершины v либо выходят ровно две дуги (к вершинам, соответствующим разбиению элемента X_v), либо не выходит ни одной дуги (оконечные вершины). В каждую вершину дерева, кроме корневой вершины, входит ровно одна дуга. В оконечных вершинах расположены элементы разбиения, актуального на текущий момент. В промежуточных вершинах хранятся элементы разбиения, которые были разделены на предыдущих этапах. Также с каждой вершиной связан вектор управляющих вероятностей, однако активными (т. е. используемыми для выбора управления) на текущий момент являются только вероятности, относящиеся к оконечным вершинам.

Выбор моментов «деления» и «склейки», а также их интенсивность могут определяться следующими эвристическими соображениями:

- те элементы разбиения, для которых накоплена бóльшая статистика для оценок градиента, целесообразно видоизменять чаще других;
- элементы разбиения, у которых вероятности мало отделены от нуля или от единицы, имеет смысл дробить реже;
- имеет смысл чаще склеивать те пары соседних элементов разбиения, у которых векторы вероятностей «близки»[2] между собой.

Общая интенсивность процессов изменения разбиения может корректироваться в зависимости от наблюдаемой в конкретных случаях скорости оптимизации.

4 Примеры

Остановимся на нескольких простых примерах, дающих общее представление о возможностях предложенного алгоритма. Рассмотрим простейшую систему из $K = 2$ серверов и сравним значения целевого функционала при следующих стратегиях[3]: рандомизированная (далее — RND), циклическая (далее — RR), программная (далее — SG), миопическая (далее — Myopic), Static Deep[4] глубиной 2 из [4, разд. 3.1] и стратегия из разд. 3 (далее — RL). Отметим, что правила RND, SG и Static Deep являются параметрическими; в табл. 1 и 2 оптимальные значения параметров указаны в квадратных скобках.

Начнем с полностью марковского случая. В табл. 1 и 2 приведены значения целевого функционала W, когда сервер 1 и сервер 2 имеют соответственно одинаковую и различную производи-

[1] Идея которого, конечно же, не нова (см., например, [20, 21]).

[2] Влияние метрики на качество результатов не изучалось. В вычислительных экспериментах, речь о которых пойдет ниже, использовалась метрика L1.

[3] За подробным описанием стратегий можно обратиться к [1] или [4, разд. 5].

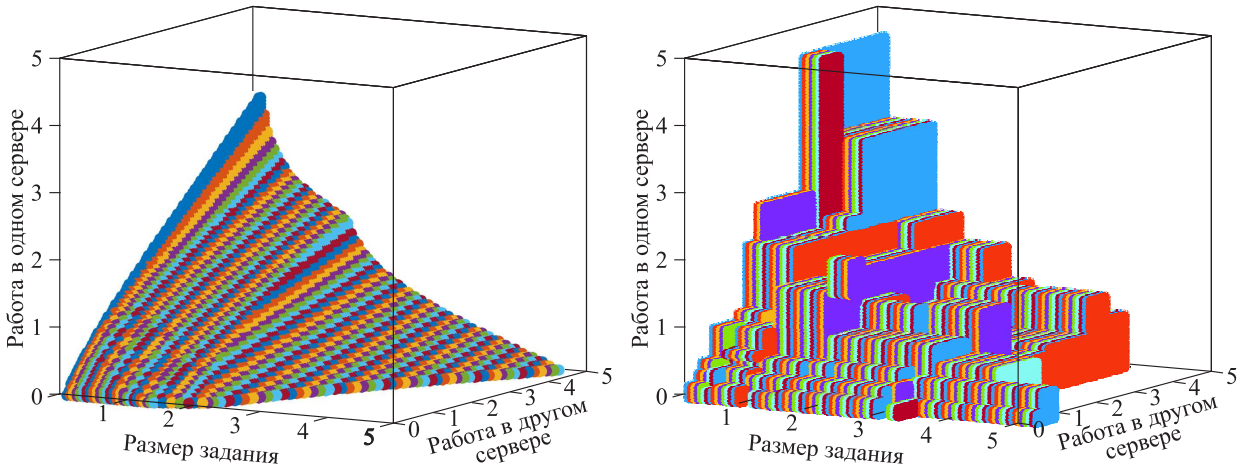
[4] Судя по публикациям, для рассматриваемой системы эта стратегия является наилучшей в том смысле, что, варьируя ее глубину, можно, по-видимому, получить близкое к неизвестному оптимальному значение целевой функции.

Таблица 1 Стационарное среднее время пребывания задания в полностью марковской системе из двух серверов единичной производительности

Стратегия	Загрузка		
	0,4	0,6	0,8
RND	1,667 [0,5]	2,5 [0,5]	5 [0,5]
RR	1,380	1,984	3,843
Myopic	1,191	1,563	2,778
Static Deep	1,191 [0,5]	1,529 [0,37]	2,518 [0,44]
RL	1,191	1,508	2,464

Таблица 2 Стационарное среднее время пребывания задания в полностью марковской системе из двух серверов с производительностью 1 и 2

Стратегия	Загрузка		
	0,4	0,6	0,8
RND	1,032 [0,21]	1,587 [0,27]	3,215 [0,31]
SG	0,915 [0,75]	1,314 [0,70]	2,539 [0,68]
Myopic	0,764	1,030	1,847
StaticDeep	0,725 [0,49]	0,980 [0,43]	1,687 [0,38]
RL	0,724	0,974	1,653



(*а*) (*б*)

Рис. 1 Структура стратегии Static Deep (*а*) и новой стратегии RL (*б*) из табл. 1 при загрузке 0,8 в области наибольшего сгущения. Отмеченные цветом области соответствуют выбору сервера 1

тельность, а средний размер заданий равен единице ($\mathsf{M}\xi = 1$).

Как видно из табл. 1 и 2, новый алгоритм, будучи динамическим и принципиально немарковским, учитывающим всю предысторию функционирования системы, оптимизирует целевой функционал существенно лучше, чем статические алгоритмы RND, RR и SG. Кроме того, он успешно конкурирует с наилучшей из известных стратегий (Static Deep). В связи с последним обстоятельством примечательно, что новый алгоритм добивается такого результата «схожим образом» (из рис. 1 видно, что структуры стратегий в области наибольшего сгущения имеют определенное визуальное сходство).

В полностью немарковских системах обычно не выполняются условия, необходимые для применения известных специальных алгоритмов. Однако новый алгоритм свободен от подобных ограничений и, кроме того, оптимизирует целевой функционал W не хуже классических стратегий (применимых при любых условиях, гарантирующих существование W). Пусть, например, серверы

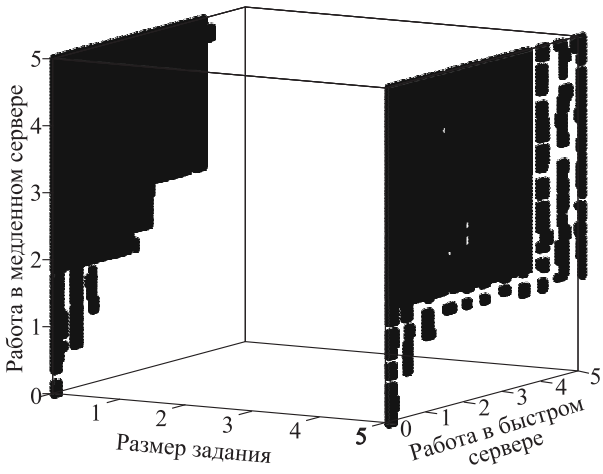


Рис. 2 Структура новой стратегии в полностью немарковской системе в области наибольшего сгущения. Отмеченные черным области соответствуют выбору быстрого сервера

имеют производительность 1,2 и 1 и через каждые 1,5 единицы времени в систему поступает задание, распределение размера которого — бимодальное в точках 2 и 4 с равными вероятностями. В такой системе, загруженной более чем на 90%, наилучшая из классических стратегий — Myopic — дает 3,338, а новая стратегия — 3,316 (см. структуру стратегии на рис. 2). Примечательно, что, хотя пример формально находится вне области применимости стратегии Static Deep, она дает результат 3,373 и имеет похожую структуру.

5 Заключение

В статье предложен подход к диспетчеризации в системе с параллельными серверами, основанный на идее градиентного подхода к марковскому процессу принятия решений. Главная особенность рассматриваемой математической задачи — несчетность множества состояний марковской цепи. Решение этой задачи основывается на двух принципиальных моментах:

(1) на аналитическом выражении для градиента целевой функции, в качестве которой выступает предельное среднее время пребывания задания в системе;

(2) на использовании динамического, постоянно меняющегося разбиения пространства состояний.

Диспетчеризация имеет распределенный децентрализованный характер: каждой области разбиения соответствует локальный алгоритм, корректирующий вероятность выбора сервера, когда состояние системы находится в данной области. Коррекция осуществляется на основе оценок частных производных целевой функции, строящихся по предыстории процесса. Таким образом, алгоритм в целом представляет собой немарковскую стратегию, которая осуществляет поиск в подмножестве «дискретных» однородных марковских стратегий, аппроксимирующих оптимальную диспетчеризацию. Здесь заключается принципиальное отличие новой конструкции от существующих диспетчеризаций. Так, стратегии Myopic и Static Deep представляют собой однородные (вырожденные) марковские стратегии, которые предписывают выбор сервера в данном состоянии системы однозначным образом, рассчитываемым априори с помощью соответствующих формул.

Способность новой диспетчеризации корректировать вероятности выбора сервера в процессе функционирования относит ее к разряду адаптивных алгоритмов [19]. Основным преимуществом такого подхода является возможность применения и получения выигрыша практически без оглядки на то, с какой системой приходится иметь дело: добавление коррелированных потоков, большее число серверов, наличие группового поступления заданий и т. п. лишь увеличат «размерность задачи», но не изменят ничего по существу. Однако это преимущество дается не бесплатно: отказ от дополнительных предположений повышает как вычислительную, так и идейную сложность алгоритма. Вообще, надо отметить достаточную сложность предложенной конструкции, особенно по сравнению с такими алгоритмами, как RND, RR или Myopic, что, однако, в современных условиях не является принципиальным препятствием для практического применения.

Относительно перспективы дальнейших исследований следует указать на два направления. Необходимо провести формальное доказательство сходимости алгоритма (численные эксперименты такую сходимость, безусловно, подтверждают). Необходимо также продвинуться, аналитически и экспериментально, в решении задачи оптимальной настройки параметров алгоритма, в частности параметров оценки градиента целевой функции. Численные примеры свидетельствуют о существенной зависимости скорости и качества работы алгоритма от такой настройки.

Литература

1. *Hyytiä E., Righter R., Virtamo J., Viitasaari L.* On value functions for FCFS queues with batch arrivals and general cost structures // Perform. Evaluation, 2020. Vol. 138. Art. 102083. 29 p.
2. *Hyytiä E.* Optimal routing of fixed size jobs to two parallel servers // INFOR, 2013. Vol. 51. No. 4. P. 215–224.
3. *Коновалов М. Г., Разумчик Р. В.* Обзор моделей и алгоритмов размещения заданий в системах с параллельным обслуживанием // Информатика и её применения, 2015. Т. 9. Вып. 4. С. 56–67.
4. *Konovalov M., Razumchik R.* Iterative algorithm for threshold calculation in the problem of routing fixed size jobs to two parallel servers // J. Telecommun. Inf. Technol., 2015. Vol. 3. P. 32–38.
5. *Hyytiä E.* Lookahead actions in dispatching to parallel queues // Perform. Evaluation, 2013. Vol. 70. No. 10. P. 859–872.
6. *Уолрэнд Дж.* Введение в теорию сетей массового обслуживания. — М.: Мир, 1993. 336 с. (*Walrand J.* An introduction to queueing networks. — Englewood Cliffs, NJ, USA: Prentice Hall, 1989. 384 p.)
7. *Chow C.-S.* Multigrid algorithms and complexity results for discrete-time stochastic control and related fixed-point problems: PhD Thesis. — Campbridge, MA, USA: MIT, 1989. 162 p.

8. *Rust J.* Using randomization to break the curse of dimensionality // Econometrica, 1997. Vol. 65. No. 3. P. 487–516.
9. *Guihenneuc-Jouyaux C., Robert C. P.* Discretization of continuous Markov chains and Markov chain Monte Carlo convergence assessment // J. Am. Stat. Assoc., 1998. Vol. 93. No. 443. P. 1055–1067.
10. *Puterman M. L., Brumelle S.* On the convergence of policy iteration in stationary dynamic programming // Math. Oper. Res., 1979. Vol. 4. No. 1. P. 60–69.
11. *Antos A., Munos R., Szepesvari C.* Fitted Q-iteration in continuous action-space MDPs // 20th Conference (International) on Neural Information Processing Systems Proceedings. — Red Hook, NY, USA: Curran Associates Inc., 2007. P. 9–16.
12. *Zhang S., Wan Yi, Sutton R., Whiteson S.* Average-reward off-policy policy evaluation with function approximation // arXiv.org, 2021. arXiv:2101.02808 [cs.LG].
13. *Bertsekas D. P.* A counter-example to temporal differences learning // Neural Comput., 1994. Vol. 7. P. 270–279.
14. *Cooper W., Henderson S., Lewis M.* Convergence of simulation-based policy iteration // Probab. Eng. Inform. Sc., 2003. Vol. 17. P. 213–234.
15. *Kveton B., Hauskrecht M.* Heuristic refinements of approximate linear programming for factored continuous-state Markov decision processes // 14th Conference (International) on Automated Planning and Scheduling Proceedings. — Menlo Park, CA, USA: AAAI Press, 2004. P. 306–314.
16. *Cao X.* Stochastic learning and optimization — a sensitivity-based approach // Annu. Rev. Control, 2009. Vol. 33. No. 1. P. 11–24.
17. *Samúelsson S. G., Hyytiä E.* Applying reinforcement learning to basic routing problem // Queueing theory and network applications / Eds. Y. Takahashi, T. Phung-Duc, S. Wittevrongel, W. Yue. — Lecture notes in computer science ser. — Springer, 2018. Vol. 10932. P. 238–249.
18. *Срагович В. Г.* Адаптивное управление. — М.: Наука, 1981. 381 с.
19. *Коновалов М. Г.* Методы адаптивной обработки информации и их приложения. — М.: ИПИ РАН, 2007. 212 с.
20. *Uther W. T. B., Veloso M. M.* Tree based discretization for continuous state space reinforcement learning // 15th National/10th Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence Proceedings. — Menlo Park, CA, USA: AAAI Press, 1998. P. 769–774.
21. *Ortner R.* Adaptive aggregation for reinforcement learning in average reward Markov decision processes // Ann. Oper. Res., 2013. Vol. 208. P. 321–336.

Поступила в редакцию 13.07.2021

ROUTING JOBS TO HETEROGENEOUS PARALLEL QUEUES USING DISTRIBUTED POLICY GRANDIENT ALGORITHM

M. G. Konovalov and R. V. Razumchik

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

Abstract: The problem of dispatching to heterogeneous servers, operating independently in parallel, is considered. Each server has a single processor and a dedicated FIFO (first in, first out) queue of infinite capacity. Homogeneous jobs (without preceding constraints) arrive one by one to the dispatcher which immediately makes a routing decision. Both jobs interarrival times and their sizes are assumed to be independent and identically distributed random variables with general distributions. Upon making a decision, full information about the current system's state, including the arriving job size, is available to the dispatcher. The problem is to minimize the long-run system's mean response time. A new sample-path-based policy gradient algorithm is proposed which allows one to construct such a policy. Its main ingredients are the dynamically changing discretization of the continuous state space and individual policy gradient algorithms acting in each cell. Simple numerical examples are given which demonstrate that the new algorithm can outperform best known solutions and is applicable in quite general cases.

Keywords: heterogeneous parallel queues; Markov chains with continuous state space; sojourn time optimization

DOI: 10.14357/19922264210306

Acknowledgments

The reported study was funded by RFBR, project No. 20-07-00804.

References

1. Hyytiä, E., R. Righter, J. Virtamo, and L. Viitasaari. 2020. On value functions for FCFS queues with batch arrivals and general cost structures. *Perform. Evaluation* 138:102083. 29 p.
2. Hyytiä, E. 2013. Optimal routing of fixed size jobs to two parallel servers. *INFOR* 51(4):215–224.
3. Konovalov, M. G., and R. V. Razumchik. 2015. Obzor modeley i algoritmov razmeshcheniya zadaniy v sistemakh s parallel'nym obsluzhivaniem [Methods and algorithms for job scheduling in systems with parallel service: A survey]. *Informatika i ee Primeneniya — Inform. Appl.* 9(4):56–67.
4. Konovalov, M., and R. Razumchik. 2015. Iterative algorithm for threshold calculation in the problem of routing fixed size jobs to two parallel servers. *J. Telecommun. Inf. Technol.* 3:32–38.
5. Hyytiä, E. 2013. Lookahead actions in dispatching to parallel queues. *Perform. Evaluation* 70(10):859–872.
6. Walrand, J. 1989. *An introduction to queueing networks*. Englewood Cliffs, NJ: Prentice Hall. 384 p.
7. Chow, C.-S. 1989. Multigrid algorithms and complexity results for discrete-time stochastic control and related fixed-point problems. Campbridge, MA: MIT. PhD Thesis.162 p.
8. Rust, J. 1997. Using randomization to break the curse of dimensionality. *Econometrica* 65(3):487–516.
9. Guihenneuc-Jouyaux, C., and C. P. Robert. 1998. Discretization of continuous Markov chains and Markov chain Monte Carlo convergence assessment. *J. Am. Stat. Assoc.* 93(443):1055–1067.
10. Puterman, M. L., and S. Brumelle. 1979. On the convergence of policy iteration in stationary dynamic programming. *Math. Oper. Res.* 4(1):60–69.
11. Antos, A., R. Munos, and C. Szepesvari. 2007. Fitted Q-iteration in continuous action-space MDPs. *20th Conference (International) on Neural Information Processing Systems Proceedings*. Red Hook, NY: Curran Associates Inc. 9–16.
12. Zhang, S., Y. Wan, R. Sutton, and S. Whiteson. 2021. Average-reward off-policy policy evaluation with function approximation. Available at: https://arxiv.org/abs/2101.02808 (accessed July 7, 2021).
13. Bertsekas, D. P. 1994. A counter-example to temporal differences learning. *Neural Comput.* 7:270–279.
14. Cooper, W., S. Henderson, and M. Lewis. 2003. Convergence of simulation-based policy iteration. *Probab. Eng. Inform. Sc.* 17:213–234.
15. Kveton, B., and M. Hauskrecht. 2004. Heuristic refinements of approximate linear programming for factored continuous-state Markov decision processes. *14th Conference (International) on International Conference on Automated Planning and Scheduling Proceedings*. Menlo Park, CA: AAAI Press. 306–314.
16. Cao, X. 2009. Stochastic learning and optimization — a sensitivity-based approach. *Annu. Rev. Control* 33(1):11–24.
17. Samúelsson, S. G., and E. Hyytiä. 2018. Applying reinforcement learning to basic routing problem. *Queueing theory and network applications*. Eds. Y. Takahashi, T. Phung-Duc, S. Wittevrongel, and W. Yue. Lecture notes in computer science ser. Springer. 10932:238–249.
18. Sragovich, V. G. 1981. *Adaptivnoe upravlenie* [Adaptive control]. Moscow: Nauka. 381 p.
19. Konovalov, M. G. 2007. *Metody adaptivnoy obrabotki informatsii i ikh prilozheniya* [Methods of adaptive information processing and their applications]. Moscow: IPI RAN. 212 p.
20. Uther, W. T. B., and M. M. Veloso. 1998. Tree based discretization for continuous state space reinforcement learning. *15th National/10th Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence Proceedings*. Menlo Park, CA: AAAI Press. 769–774.
21. Ortner, R. 2013. Adaptive aggregation for reinforcement learning in average reward Markov decision processes. *Ann. Oper. Res.* 208:321–336.

Received July 13, 2021

Contributors

Konovalov Mikhail G. (b. 1950) — 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; mkonovalov@ipiran.ru

Razumchik Rostislav V. (b. 1984) — 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 119333, Russian Federation; rrazumchik@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 51–56

ПОРОГОВЫЕ ФУНКЦИИ В МЕТОДАХ ПОДАВЛЕНИЯ ШУМА, ОСНОВАННЫХ НА ВЕЙВЛЕТ-РАЗЛОЖЕНИИ СИГНАЛА*

О. В. Шестаков[1]

Аннотация: При передаче по каналам связи сигналы, как правило, загрязняются шумом. Методы подавления шума, основанные на пороговой обработке коэффициентов вейвлет-разложения, стали популярными благодаря своей простоте, скорости и возможности адаптации к нестационарным сигналам. Анализ погрешностей этих методов является важной практической задачей, поскольку дает возможность оценивать качество как самих методов, так и используемого для обработки оборудования. Самые популярные виды пороговой обработки — жесткая и мягкая пороговая обработка, однако каждая из них имеет свои недостатки. В попытке избавиться от этих недостатков в последние годы были предложены различные альтернативные виды пороговой обработки. В работе рассматривается модель сигнала, загрязненного аддитивным гауссовским шумом, и обсуждается общая постановка задачи пороговой обработки с пороговой функцией, принадлежащей некоторому классу. Описывается алгоритм вычисления порога, минимизирующего несмещенную оценку риска. Также приводятся условия, при которых эта оценка риска является асимптотически нормальной и сильно состоятельной.

Ключевые слова: вейвлеты; пороговая обработка; адаптивный порог; несмещенная оценка риска

DOI: 10.14357/19922264210307

1 Введение

Одна из основных задач анализа и обработки сигналов — подавление шума. Традиционные методы используют преобразование Фурье для разделения сигнала на высокочастотные и низкочастотные компоненты с последующим удалением высокочастотных компонент, что приводит к подавлению большей части шума. Однако при этом также удаляется полезная информация, которая содержится в высокочастотной компоненте, и зачастую сигнал оказывается сильно сглаженным. Вейвлет-разложение обладает свойством частотно-временно́й локализации и тем самым позволяет анализировать эволюцию частотного спектра во времени. При этом подавление шума, как правило, осуществляется с помощью методов пороговой обработки вейвлет-коэффициентов, которые обнуляют коэффициенты, не превышающие заданного порога. Наиболее часто применяется жесткая или мягкая пороговая обработка, однако каждая из них имеет свои недостатки. Жесткая пороговая обработка использует разрывную пороговую функцию, что приводит к появлению артефактов и невозможности построения несмещенной оценки риска, а при мягкой пороговой обработке в функции сигнала появляется дополнительное смещение. В ряде работ предложены некоторые альтернативные функции пороговой обработки, справляющиеся с указанными недостатками [1–9]. При этом значение порога обычно предлагается вычислять, решая задачу минимизации оценки среднеквадратичного риска, построенной по методу Стейна [10]. В данной работе рассматривается постановка задачи пороговой обработки, в которой пороговая функция имеет достаточно общий вид. Обсуждается метод вычисления адаптивного порога, минимизирующего оценку риска, а также асимптотические свойства этой оценки, такие как сильная состоятельность и асимптотическая нормальность.

2 Вейвлет-разложение

Для функции сигнала $f \in L^2(\mathbf{R})$ вейвлет-разложение представляет собой ряд вида

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

где функции $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ описывают сдвиг и растяжение (сжатие) выбранной вейвлет-

*Работа выполнена при частичной финансовой поддержке РФФИ (проект 19-07-00352); статья опубликована при финансовой поддержке Минобрнауки РФ в рамках реализации программы Московского центра фундаментальной и прикладной математики по соглашению № 075-15-2019-1621.

[1]Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики; Федеральный исследовательский центр «Информатика и управление» Российской академии наук; Московский центр фундаментальной и прикладной математики, oshestakov@cs.msu.ru

функции ψ. Индекс j в этом разложении называется масштабом, а индекс k — сдвигом. Чтобы последовательность $\{\psi_{j,k}\}$ образовывала ортонормированный базис, функция ψ должна удовлетворять определенным требованиям, однако ее можно выбрать таким образом, чтобы она обладала некоторыми полезными свойствами, например была дифференцируемой M раз и имела M нулевых моментов [11].

Пусть функция f определена на некотором конечном отрезке $[a,b]$ и равномерно регулярна по Липшицу с некоторым показателем γ ($0 < \gamma \leq M$). Также предположим, что ψ удовлетворяет условию:

$$\int\limits_{-\infty}^{\infty} |t^{\gamma}\psi(t)|\,dt < \infty\,.$$

В этом случае справедливо неравенство [11]:

$$|\langle f, \psi_{j,k}\rangle| \leq \frac{C_f}{2^{j(\gamma+1/2)}}\,, \qquad (1)$$

где C_f — некоторая положительная константа, зависящая от вида функции f. Неравенство (1) показывает, что в пространстве вейвлет-коэффициентов регулярные по Липшицу функции имеют разреженное (или экономное) представление.

При передаче по цифровому каналу функции сигнала заданы в дискретных отсчетах. Предположим, что число этих отсчетов составляет $N = 2^J$ при некотором $J > 0$. Дискретное вейвлет-преобразование представляет собой умножение вектора значений функции f на ортогональную матрицу W, определяемую вейвлет-функцией ψ. При этом получается набор дискретных вейвлет-коэффициентов $\mu_{j,k}$, для которых справедливо соотношение $\mu_{j,k} \approx \sqrt{N}\langle f, \psi_{j,k}\rangle$ [11]. Это приближение становится точнее с ростом N.

3 Подавление шума с помощью пороговой обработки

На практике сигналы, как правило, загрязнены шумом. В большинстве случаев можно считать, что это аддитивный белый гауссовский шум, т. е. загрязненный сигнал описывается следующей моделью:

$$X_i = f_i + z_i,\ \ i = 1, \ldots, N\,,$$

где f_i — чистые значения функции сигнала, а z_i — независимые случайные величины, имеющие нормальное распределение с нулевым средним и дисперсией σ^2. Поскольку матрица W ортогональна, дискретные вейвлет-коэффициенты имеют вид:

$$Y_{j,k} = \mu_{j,k} + \epsilon_{j,k}, \quad j = 0, \ldots, J-1,\ k = 0, \ldots, 2^j - 1\,,$$

где $\epsilon_{j,k}$ также независимы и нормально распределены с нулевым средним и дисперсией σ^2.

Для подавления шума к вейвлет-коэффициентам применяется функция пороговой обработки, смысл которой заключается в обнулении достаточно маленьких коэффициентов, которые считаются шумом. Наиболее распространены функция жесткой пороговой обработки

$$\rho_H(y,T) = \begin{cases} y & \text{при } |y| > T\,; \\ 0 & \text{при } |y| \leqslant T \end{cases}$$

и функция мягкой пороговой обработки

$$\rho_S(y,T) = \begin{cases} y - T & \text{при } y > T\,; \\ y + T & \text{при } y < -T\,; \\ 0 & \text{при } |y| \leqslant T \end{cases}$$

с некоторым порогом T. При этом каждая из них имеет свои недостатки. Функция ρ_H разрывна, что приводит к отсутствию устойчивости и появлению дополнительных артефактов, а функция ρ_S приводит к появлению дополнительного смещения в оценке функции сигнала. В ряде работ предложены некоторые альтернативные функции пороговой обработки $\rho(y,T)$, которые, по сути, являются компромиссом между жесткой и мягкой пороговой обработкой [1–9]. Эти функции непрерывны, как и функция $\rho_S(y,T)$, но при этом $\rho(y,T) \to y$ при $|y| \to \infty$, т. е. для больших абсолютных значений коэффициентов они похожи на функцию жесткой пороговой обработки. Некоторые из этих функций зависят от дополнительных параметров.

В связи с появлением разнообразных видов функций пороговой обработки имеет смысл рассмотреть некоторый общий класс таких функций. Пусть функция $h(y,T)$ обладает следующими свойствами:

(1) $h(-y,T) = -h(y,T)$ (нечетность);

(2) $0 \leqslant h(y,T) \leqslant T + c$ при $y \geqslant 0$ для некоторой константы $c > 0$ (ограниченность);

(3) $h(T,T) = T$ (непрерывность пороговой функции).

Определим функцию пороговой обработки

$$\rho_h(y,T) = \begin{cases} y - h(y,T) & \text{при } |y| > T\,; \\ 0 & \text{при } |y| \leqslant T\,. \end{cases}$$

Такой вид имеет, например, функция мягкой пороговой обработки, функция гибридной пороговой обработки [2, 8], функция пороговой обработки на основе гиперболического тангенса [7], сигмоидная функция [9] и некоторые другие.

4 Значение порога и статистические свойства оценки риска

Погрешность (или риск) пороговой обработки определяется следующим образом:

$$R_J(T) = \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \mathsf{E}\left(\rho_h(Y_{j,k}, T) - \mu_{j,k}\right)^2. \quad (2)$$

Методы выбора порогового значения T, как правило, ориентированы на минимизацию риска (2). Заметим, что в выражении (2) присутствуют неизвестные величины $\mu_{j,k}$ и вычислить значение $R_J(T)$ на практике нельзя. Поэтому при вычислении T используется минимаксный подход в предположении о принадлежности функции сигнала какому-либо классу [12–14]. Отметим также, что популярен так называемый универсальный порог $T_U = \sigma\sqrt{2\ln N}$. Этот порог является в определенном смысле максимальным (в работах [15, 16] показано, что можно не рассматривать $T > T_U$).

Другой подход к вычислению порогового значения основан на минимизации статистической оценки риска (2), построенной по методу Стейна [10]. Если $h(y,T)$ дифференцируема по переменной y, то справедливо соотношение:

$$\mathsf{E}\left(\rho_h(Y_{j,k}, T) - \mu_{j,k}\right)^2 = \sigma^2 + \mathsf{E}g^2(Y_{j,k}) - 2\sigma^2\mathsf{E}g'(Y_{j,k}),$$

где

$$g\left(Y_{j,k}\right) = Y_{j,k} - \rho_h\left(Y_{j,k}, T\right).$$

Таким образом, в качестве несмещенной оценки риска можно использовать следующую величину:

$$\widehat{R}_J(T) = \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} F\left(Y_{j,k}, T\right), \quad (3)$$

где

$$F(y,T) = \begin{cases} y^2 - \sigma^2 & \text{при } |y| \leqslant T; \\ h^2(y,T) + \sigma^2 - 2\sigma^2 h'_y(y,T) & \text{при } |y| > T. \end{cases}$$

Порог T_S минимизирует оценку (3):

$$\widehat{R}_J(T_S) = \min_{T\in[0,T_U]} \widehat{R}_J(T).$$

Этот порог имитирует теоретический «идеальный» порог T_{Min}, минимизирующий теоретический риск (2):

$$R_J(T_{\mathrm{Min}}) = \min_{T\in[0,T_U]} R_J(T).$$

Если дополнительно предположить, что при любом фиксированном y функция $h(y,T)$ возрастает по T, а $h'_y(y,T)$ убывает по T (этим условиям удовлетворяют, например, функция гибридной пороговой обработки и функция пороговой обработки на основе гиперболического тангенса), то алгоритм вычисления значения T_S в общем случае аналогичен алгоритму вычисления этого порога при мягкой пороговой обработке [11].

Упорядочим $Y_{j,k}$ по убыванию абсолютных значений и обозначим через $Y_{(i)}$ i-ю координату полученного упорядоченного вектора. Пусть l — такой индекс, что

$$\left|Y_{(l)}\right| \leqslant T < \left|Y_{(l-1)}\right|.$$

Тогда

$$\begin{aligned}
\widehat{R}_J(T) &= \sum_{i=1}^{N-1} F\left(Y_{(i)}, T\right) = \\
&= \sum_{i=1}^{l-1} F\left(Y_{(i)}, T\right) + \sum_{i=l}^{N-1} F\left(Y_{(i)}, T\right) = \\
&= \sum_{i=1}^{l-1} \left(h^2(Y_{(i)}, T) + \sigma^2 - 2\sigma^2 h'_y(Y_{(i)}, T)\right) + \\
&\qquad + \sum_{i=l}^{N-1} \left(Y_{(i)}^2 - \sigma^2\right) = \\
&= \sum_{i=1}^{l-1} \left(h^2(Y_{(i)}, T) - 2\sigma^2 h'_y\left(Y_{(i)}, T\right)\right) + \\
&\qquad + \sum_{i=l}^{N-1} Y_{(i)}^2 + (2l - 1 - N)\sigma^2. \quad (4)
\end{aligned}$$

При введенных ограничениях на функцию $h(y,T)$ выражение (4) при фиксированном l возрастает по T. Следовательно, минимум достигается при $T = |Y_{(l)}|$. Порог T_S, минимизирующий $\widehat{R}_J(T)$, ищется сравнением $N-1$ значений выражения (4) в точках $T_m = |Y_{(m)}|$, $1 \leqslant m \leqslant N-1$. То значение $|Y_{(m)}|$, на котором достигается минимум, принимается за T_S.

Для регулярных по Липшицу функций сигнала f можно установить асимптотические свойства оценки риска. Если дополнительно предположить, что существует положительная константа c, такая что $\left|h'_y(y,T)\right| \leqslant cT^2$ при всех y (что, как правило, выполнено), то будут справедливы следующие утверждения, доказательство которых полностью аналогично доказательству соответствующих утверждений в работах [17, 18].

Теорема 1. *Пусть f задана на отрезке $[a,b]$ и равномерно регулярна по Липшицу с показателем $\gamma > 1/2$. Тогда равномерно по $x \in \mathbf{R}$*

$$\mathsf{P}\left(\frac{\widehat{R}_J\left(T_S\right)-R_J\left(T_{\mathrm{Min}}\right)}{\sigma^2\sqrt{2^{J+1}}}<x\right)\to\Phi(x)\ \ \textit{при}\ J\to\infty\,,$$

где $\Phi(x)$ *— функция распределения стандартного нормального закона.*

Теорема 2. *Пусть* f *задана на отрезке* $[a,b]$ *и равномерно регулярна по Липшицу с показателем* $\gamma>1/2$. *Тогда при любом* $\lambda>0$

$$\frac{\widehat{R}_J\left(T_S\right)-R_J\left(T_{\mathrm{Min}}\right)}{2^{J(1/2+\lambda)}}\to 0\ \text{ п. в. при } J\to\infty\,.$$

Данные теоремы обосновывают использование величины $\widehat{R}_J(T_S)$ для оценивания погрешности и позволяют строить асимптотические доверительные интервалы для теоретического риска $R_J(T_{\mathrm{Min}})$.

Литература

1. *Gao H.-Y.* Wavelet shrinkage denoising using the non-negative garrote // J. Comput. Graph. Stat., 1998. Vol. 7. No. 4. P. 469–488.
2. *Chmelka L., Kozumplik J.* Wavelet-based wiener filter for electrocardiogram signal denoising // Comput. Cardiol., 2005. Vol. 32. P. 771–774.
3. *Poornachandra S., Kumaravel N., Saravanan T. K., Somaskandan R.* WaveShrink using modified hyper-shrinkage function // 27th Annual Conference (International) of the IEEE Engineering in Medicine and Biology Society Proceedings. — Piscataway, NJ, USA: IEEE, 2005. P. 30–32.
4. *Lin Y., Cai J.* A new threshold function for signal denoising based on wavelet transform // Conference (International) on Measuring Technology and Mechatronics Automation Proceedings. — Piscataway, NJ, USA: IEEE, 2010. P. 200–203.
5. *Huang H.-C., Lee T. C. M.* Stabilized thresholding with generalized sure for image denoising // 17th Conference (International) on Image Processing Proceedings. — Piscataway, NJ, USA: IEEE, 2010. P. 1881–1884.
6. *Zhao R.-M., Cui H.-M.* Improved threshold denoising method based on wavelet transform // 7th Conference (International) on Modelling, Identification and Control Proceedings. — Piscataway, NJ, USA: IEEE, 2015. Art. 7409352. 4 p. doi: 10.1109/ICMIC.2015.7409352.
7. *He C., Xing J., Li J., Yang Q., Wang R.* A new wavelet thresholding function based on hyperbolic tangent function // Math. Probl. Eng., 2015. Vol. 2015. Art. 528656. 10 p.
8. *Priya K. D., Rao G. S., Rao P. S.* Comparative analysis of wavelet thresholding techniques with wavelet-wiener filter on ECG signal // Procedia Comput. Sci., 2016. Vol. 87. P. 178–183.
9. *He H., Tan Y.* A novel adaptive wavelet thresholding with identical correlation shrinkage function for ECG noise removal // Chinese J. Electron., 2018. Vol. 27. No. 3. P. 507–513.
10. *Stein C.* Estimation of the mean of a multivariate normal distribution // Ann. Stat., 1981. Vol. 9. No. 6. P. 1135–1151.
11. *Mallat S.* A wavelet tour of signal processing. — New York, NY, USA: Academic Press, 1999. 857 p.
12. *Donoho D., Johnstone I. M.* Ideal spatial adaptation via wavelet shrinkage // Biometrika, 1994. Vol. 81. No. 3. P. 425–455.
13. *Donoho D., Johnstone I. M.* Minimax estimation via wavelet shrinkage // Ann. Stat., 1998. Vol. 26. No. 3. P. 879–921.
14. *Jansen M.* Noise reduction by wavelet thresholding. — Lecture notes in statistics. — New York, NY: Springer Verlag, 2001. Vol. 161. 217 p.
15. *Donoho D., Johnstone I. M.* Adapting to unknown smoothness via wavelet shrinkage // J. Am. Stat. Assoc., 1995. Vol. 90. P. 1200–1224.
16. *Marron J. S., Adak S., Johnstone I. M., Neumann M. H., Patil P.* Exact risk analysis of wavelet regression // J. Comput. Graph. Stat., 1998. Vol. 7. P. 278–309.
17. *Шестаков О. В.* Асимптотическая нормальность оценки риска пороговой обработки вейвлет-коэффициентов при выборе адаптивного порога // Докл. Акад. наук, 2012. Т. 445. № 5. С. 513–515.
18. *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 24.07.2021

THRESHOLDING FUNCTIONS IN THE NOISE SUPPRESSION METHODS BASED ON THE WAVELET EXPANSION OF THE SIGNAL

O. V. Shestakov[1,2,3]

[1]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[2]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[3]Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation

Abstract: When transmitted over communication channels, signals are usually contaminated with noise. Noise suppression methods based on thresholding of wavelet expansion coefficients have become popular due to their simplicity, speed, and ability to adapt to nonstationary signals. The analysis of the errors of these methods is an important practical task, since it makes it possible to assess the quality of both the methods themselves and the equipment used for processing. The most popular types of thresholding are hard and soft thresholding but each has its own drawbacks. In an attempt to address these shortcomings, various alternative thresholding methods have been proposed in recent years. The paper considers a model of a signal contaminated with additive Gaussian noise and discusses the general formulation of the thresholding problem with a thresholding function belonging to a certain class. An algorithm for calculating the threshold that minimizes the unbiased risk estimate is described. Conditions are also given under which this risk estimate is asymptotically normal and strongly consistent.

Keywords: wavelets; thresholding; adaptive threshold; unbiased risk estimate

DOI: 10.14357/19922264210307

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (project No. 19-07-00352). The paper was published with the financial support of the Ministry of Education and Science of the Russian Federation as a part of the Program of the Moscow Center for Fundamental and Applied Mathematics under agreement No. 075-15-2019-1621.

References

1. Gao, H.-Y. 1998. Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Stat.* 7(4):469–488.
2. Chmelka, L., and J. Kozumplik. 2005. Wavelet-based wiener filter for electrocardiogram signal denoising. *Comput. Cardiol.* 32:771–774.
3. Poornachandra, S., N. Kumaravel, T. K. Saravanan, and R. Somaskandan. 2005. WaveShrink using modified hyper-shrinkage function. *27th Annual Conference (International) of the IEEE Engineering in Medicine and Biology Society Proceedings*. Piscataway, NJ: IEEE. 30–32.
4. Lin, Y., and J. Cai. 2010. A new threshold function for signal denoising based on wavelet transform. *Conference (International) on Measuring Technology and Mechatronics Automation Proceedings*. Piscataway, NJ: IEEE. 200–203.
5. Huang, H.-C., and T. C. M. Lee. 2010. Stabilized thresholding with generalized sure for image denoising. *17th Conference (International) on Image Processing Proceedings*. Piscataway, NJ: IEEE. 1881–1884.
6. Zhao, R.-M., and H.-M. Cui. 2015. Improved threshold denoising method based on wavelet transform. *7th Conference (International) on Modelling, Identification and Control Proceedings*. Piscataway, NJ: IEEE. Art. ID: 7409352. 4 p. doi: 10.1109/ICMIC.2015.7409352.
7. He, C., J. Xing, J. Li, Q. Yang, and R. Wang. 2015. A new wavelet thresholding function based on hyperbolic tangent function. *Math. Probl. Eng.* 2015:528656. 10 p.
8. Priya, K. D., G. S. Rao, and P. S. Rao. 2016. Comparative analysis of wavelet thresholding techniques with wavelet-wiener filter on ECG signal. *Procedia Comput. Sci.* 87:178–183.
9. He, H., and Y. Tan. 2018. A novel adaptive wavelet thresholding with identical correlation shrinkage function for ECG noise removal. *Chinese J. Electron.* 27(3):507–513.
10. Stein, C. 1981. Estimation of the mean of a multivariate normal distribution. *Ann. Stat.* 9(6):1135–1151.
11. Mallat, S. 1999. *A wavelet tour of signal processing*. New York, NY: Academic Press. 857 p.
12. Donoho, D., and I. M. Johnstone. 1994. Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81(3):425–455.

13. Donoho, D., and I. M. Johnstone. 1998. Minimax estimation via wavelet shrinkage. *Ann. Stat.* 26(3):879–921.

14. Jansen, M. 2001. *Noise reduction by wavelet thresholding*. Lecture notes in statistics ser. New York, NY: Springer Verlag. Vol. 161. 217 p.

15. Donoho, D., and I. M. Johnstone. 1995. Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* 90:1200–1224.

16. Marron, J. S., S. Adak, I. M. Johnstone, M. H. Neumann, and P. Patil. 1998. Exact risk analysis of wavelet regression. *J. Comput. Graph. Stat.* 7:278–309.

17. Shestakov, O. V. 2012. Asymptotic normality of adaptive wavelet thresholding risk estimation. *Dokl. Math.* 86(1):556–558.

18. Shestakov, O. V. 2016. On the strong consistency of the adaptive risk estimator for wavelet thresholding. *J. Math. Sci.* 214(1):115–118.

Received July 24, 2021

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 Leninskie 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; leading scientist, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; oshestakov@cs.msu.su

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 57–62

МЕТОД ОЦЕНИВАНИЯ ПАРАМЕТРОВ ИЗГИБА, ФОРМЫ И МАСШТАБА ГАММА-ЭКСПОНЕНЦИАЛЬНОГО РАСПРЕДЕЛЕНИЯ*

А. А. Кудрявцев[1], О. В. Шестаков[2], С. Я. Шоргин[3]

Аннотация: Рассматривается модифицированный метод моментов для оценивания трех из пяти параметров гамма-экспоненциального распределения. Предлагается оценивать параметры распределения, основываясь на его логарифмических моментах. Приводится явный вид оценок параметров изгиба, формы и масштаба при фиксированных параметрах концентрации гамма-экспоненциального распределения; обосновывается сильная состоятельность полученных оценок. Также обсуждается метод отсева лишних решений системы уравнений для логарифмических моментов; приводится ряд численных примеров, иллюстрирующих получение оценок по модельным выборкам. Поскольку анализируемое распределение тесно связано с обобщенным гамма-распределением и обобщенным бета-распределением второго рода, результаты работы могут найти широкое применение в прикладных задачах, использующих для моделирования непрерывные распределения с неограниченным неотрицательным носителем.

Ключевые слова: оценивание параметров; гамма-экспоненциальное распределение; смешанные распределения; обобщенное гамма-распределение; метод моментов; состоятельная оценка

DOI: 10.14357/19922264210308

1 Введение

Большинство современных математических моделей, описываемых при помощи вероятностных распределений с неограниченным неотрицательным носителем, оперируют частными случаями обобщенного гамма-распределения [1, 2] и обобщенного бета-распределения второго рода [3]. В работе [4] было предложено распределение, позволяющее распространить полезные свойства популярных модельных распределений на более широкий круг прикладных задач.

Определение. Будем говорить, что случайная величина ζ имеет гамма-экспоненциальное распределение $\mathrm{GE}\,(r, \nu, s, t, \delta)$ с параметрами изгиба $0 \le r < 1$, формы $\nu \ne 0$, концентрации $s, t > 0$ и масштаба $\delta > 0$, если ее плотность при $x > 0$ задается соотношением

$$g_E(x) = \frac{|\nu| x^{t\nu-1}}{\delta^{t\nu}\Gamma(s)\Gamma(t)}\,\mathsf{Ge}_{r,\,tr+s}\left(-\left(\frac{x}{\delta}\right)^{\nu}\right), \qquad (1)$$

где $E = (r, \nu, s, t, \delta)$, а $\mathsf{Ge}_{\alpha,\,\beta}(x)$ — гамма-экспоненциальная функция [5]:

$$\mathsf{Ge}_{\alpha,\,\beta}(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}\,\Gamma(\alpha k + \beta), \quad x \in \mathbb{R}\,,\ \ 0 \le \alpha < 1\,,\ \ \beta > 0\,.$$

Потенциал использования распределения (1) в приложениях обусловлен не только гибкостью пятипараметрического распределения при описании всевозможных процессов, но также полезным свойством, связывающим гамма-экспоненциальное распределение с обобщенным гамма-распределением $\mathrm{GG}\,(v, q, \theta)$, имеющим плотность

$$f(x) = \frac{|v| x^{vq-1} e^{-(x/\theta)^v}}{\theta^{vq}\Gamma(q)}, \quad v \ne 0\,,\ \ q > 0\,,\ \ \theta > 0\,,\ \ x > 0\,. \qquad (2)$$

Справедливо следующее утверждение [4].

Лемма 1. 1. *Пусть независимые случайные величины λ и μ имеют соответственно распределения* $\mathrm{GG}\,(v, q, \theta)$ *и* $\mathrm{GG}\,(u, p, \alpha)$, $uv > 0$. *Тогда распределение λ совпадает с* $\mathrm{GE}\,(0, v, \cdot, q, \theta)$*; распределение λ/μ при $|u| > |v|$ совпадает с* $\mathrm{GE}\,(v/u, v, p, q, \theta/\alpha)$*; распределение λ/μ при $|v| > |u|$ совпадает с* $\mathrm{GE}\,(u/v, -u, q, p, \theta/\alpha)$.

*Работа выполнена при частичной финансовой поддержке РФФИ (проект 20–07–00655); статья опубликована при финансовой поддержке Минобрнауки РФ в рамках реализации программы Московского центра фундаментальной и прикладной математики по соглашению No. 075-15-2019-1621.

[1]Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики; Московский центр фундаментальной и прикладной математики, nubigena@mail.ru

[2]Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики; Федеральный исследовательский центр «Информатика и управление» Российской академии наук; Московский центр фундаментальной и прикладной математики, oshestakov@cs.msu.ru

[3]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, sshorgin@ipiran.ru

2. ***При*** $0 < r < 1$ ***плотность*** $g_E(x)$, $E = (r, \nu, s, t, \delta)$, ***совпадает с плотностью отношения независимых случайных величин, имеющих обобщенные гамма-распределения*** $\mathrm{GG}(\nu, t, \delta)$ ***и*** $\mathrm{GG}(\nu/r, s, 1)$.

Лемма 1 позволяет не только использовать гамма-экспоненциальное распределение для описания параметров моделей, распределения которых представимы в виде масштабных смесей обобщенных гамма-законов (например, индексов баланса независимых факторов [6]), но также дает возможность оценивать неизвестные параметры распределения (1) при помощи вероятностных характеристик распределения (2).

Заметим, что ввиду представления плотности (1) в терминах специальной гамма-экспоненциальной функции применение метода максимального правдоподобия для оценивания неизвестных параметров представляется затруднительным. Кроме того, поскольку моменты распределения (1) имеют вид [4]

$$\mathsf{E}\zeta^k = \frac{\delta^k}{\Gamma(t)\Gamma(s)}\,\Gamma\left(t+\frac{k}{\nu}\right)\Gamma\left(s-\frac{rk}{\nu}\right), \quad t+\frac{k}{\nu}>0\,,\ \ s-\frac{rk}{\nu}>0\,,$$

определены не для всех значений параметров и представимы через немонотонную гамма-функцию, применение прямого метода моментов также не представляется разумным.

В работах [7, 8] для оценивания параметров было предложено использовать модифицированный метод, основанный на логарифмических моментах гамма-экспоненциального распределения; был найден аналитический вид оценок пар параметров изгиба-масштаба и формы-масштаба при фиксированных остальных трех параметрах распределения (1) и исследованы некоторые статистические свойства полученных оценок, такие как сильная состоятельность и асимптотическая нормальность.

В данной работе метод, предложенный в [7], рассматривается для отыскания оценок неизвестных параметров изгиба, формы и масштаба при фиксированных параметрах концентрации гамма-экспоненциального распределения (1).

2 Логарифмические моменты гамма-экспоненциального распределения

Введем обозначение для дигамма-функции:

$$\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}\,.$$

Найдем первые четыре логарифмических момента случайной величины ζ, имеющей гамма-экспоненциальное распределение $\mathrm{GE}(r, \nu, s, t, \delta)$. Поскольку, ввиду справедливости леммы 1, преобразование Меллина случайной величины ζ имеет вид [9]

$$\mathcal{M}_\zeta(z) = \frac{\delta^z}{\Gamma(t)\Gamma(s)}\,\Gamma\left(t+\frac{z}{\nu}\right)\Gamma\left(s-\frac{rz}{\nu}\right), \quad t+\frac{\mathsf{Re}(z)}{\nu}>0,\ \ s-\frac{r\mathsf{Re}(z)}{\nu}>0\,,$$

характеристическая функция логарифма ζ вычисляется по формуле

$$\mathsf{E}e^{iy\ln\zeta} = \frac{\delta^{iy}}{\Gamma(t)\Gamma(s)}\,\Gamma\left(t+\frac{iy}{\nu}\right)\Gamma\left(s-\frac{iry}{\nu}\right), \quad y\in\mathbb{R}\,. \tag{3}$$

Нужное число раз продифференцировав соотношение (3), получаем

$$\mathsf{E}\ln\zeta = \frac{\nu\ln\delta+\psi(t)-r\psi(s)}{\nu}\,;$$

$$\mathsf{E}\ln^2\zeta = \frac{[\nu\ln\delta+\psi(t)-r\psi(s)]^2}{\nu^2} + \frac{\psi'(t)+r^2\psi'(s)}{\nu^2}\,;$$

$$\mathsf{E}\ln^3\zeta = \frac{[\nu\ln\delta+\psi(t)-r\psi(s)]^3}{\nu^3} + \frac{3(\psi'(t)+r^2\psi'(s))\,[\nu\ln\delta+\psi(t)-r\psi(s)]}{\nu^3} + \frac{\psi''(t)-r^3\psi''(s)}{\nu^3}\,;$$

$$\mathsf{E}\ln^4\zeta = \frac{[\nu\ln\delta+\psi(t)-r\psi(s)]^4}{\nu^4} + \frac{6(\psi'(t)+r^2\psi'(s))[\nu\ln\delta+\psi(t)-r\psi(s)]^2}{\nu^4} + \frac{4(\psi''(t)-r^3\psi''(s))[\nu\ln\delta+\psi(t)-r\psi(s)]}{\nu^4} + \frac{3(\psi'(t)+r^2\psi'(s))^2+\psi'''(t)+r^4\psi'''(s)}{\nu^4}\,.$$

3 Оценивание параметров изгиба, формы и масштаба при фиксированных параметрах концентрации

Зафиксируем параметры концентрации s и t. Оценим при помощи модифицированного метода моментов параметры изгиба r, формы ν и масштаба δ распределения $\mathrm{GE}(r, \nu, s, t, \delta)$.

Введем обозначение для выборочных логарифмических моментов ζ:

$$L_k(X) = \frac{1}{n}\sum_{i=1}^{n} \ln^k X_i,$$

где $X = (X_1, \ldots, X_n)$ — выборка из распределения $\zeta \sim \mathrm{GE}\,(r, \nu, s, t, \delta)$.

Составим систему из четырех уравнений с тремя неизвестными r, ν и δ (s и t фиксированы):

$$\mathsf{E} \ln^k \zeta = L_k(X), \quad k = 1, 2, 3, 4\,.$$

Рассмотрим статистики

$$\phi_k = \left(\psi^{(k)}(s)\right)^{-1} \psi^{(k)}(t);$$

$$K(X) = (\psi'(s))^{-1} \left(L_2(X) - L_1^2(X)\right);$$

$$M(X) = (\psi''(s))^{-1} \left(L_3(X) - 3L_1(X)\left(L_2(X) - L_1^2(X)\right) - L_1^3(X)\right);$$

$$N(X) = (\psi'''(s))^{-1} \Big(L_4(X) - 6L_1^2(X)(L_2(X) - L_1^2(X)) - 4L_1(X)\left(L_3(X) - 3L_1(X)\left(L_2(X) - L_1^2(X)\right) - L_1^3(X)\right) - 3\left(L_2(X) - L_1^2(X)\right)^2 - L_1^4(X)\Big)\,.$$

Очевидно, что исходная система уравнений эквивалентна следующей:

$$\ln\delta + \frac{\psi(t)}{\nu} - \frac{r\psi(s)}{\nu} = L_1(X); \tag{4}$$

$$\frac{\phi_1}{\nu^2} + \frac{r^2}{\nu^2} = K(X); \tag{5}$$

$$\frac{\phi_2}{\nu^3} - \frac{r^3}{\nu^3} = M(X); \tag{6}$$

$$\frac{\phi_3}{\nu^4} + \frac{r^4}{\nu^4} = N(X). \tag{7}$$

Из уравнений (5) и (7) получаем

$$\left(K^2(X) - N(X)\right)\nu^4 - 2K(X)\phi_1\nu^2 + \phi_1^2 + \phi_3 = 0\,,$$

откуда следует, что возможные оценки квадрата параметра ν имеют вид:

$$\hat\nu_\pm^2(X) = \frac{K(X)\phi_1 \pm \sqrt{N(X)(\phi_1^2 + \phi_3) - K^2(X)\phi_3}}{K^2(X) - N(X)}\,. \tag{8}$$

Следовательно, возможные оценки квадрата параметра r имеют вид:

$$\hat r_\pm^2(X) = \frac{1}{K^2(X) - N(X)}\left(N(X)\phi_1 \pm \pm K(X)\sqrt{N(X)(\phi_1^2 + \phi_3) - K^2(X)\phi_3}\right). \tag{9}$$

Заметим, что полученные решения не определяют однозначно оценки параметров r и ν, причем если знак r всегда определен, то ν может быть как положительной, так и отрицательной величиной. Кроме того, численные эксперименты показывают, что для фиксированной реализации выборки выражения (8) и (9) могут давать непротиворечивые значения оценок при любом знаке перед радикалом. По этой причине при обработке реальных данных следует альтернативными методами определить знак параметра ν, а затем воспользоваться алгоритмом отсева лишнего решения системы.

Алгоритм выбора «правильного» решения $(\hat r(X), \hat\nu(X))$ системы заключается в следующем. На первом этапе для заданной реализации выборки вычисляются значения правых частей (9). В случае если значение $\hat r_\pm^2(X)$ при некотором знаке перед радикалом не принадлежит интервалу $[0, 1)$, соответствующая пара решений отсеивается и в качестве решения $(\hat r(X), \hat\nu(X))$ выбирается пара с противоположным знаком перед радикалом. На втором этапе, если значения (9) при любом знаке перед радикалом непротиворечивы, следует дополнительно воспользоваться уравнением (6) для определения решения $(\hat r(X), \hat\nu(X))$ системы. Поскольку оценки $\hat r(X)$ и $\hat\nu(X)$ параметров изгиба и формы наряду со статистикой $M(X)$ представляют собой непрерывные функции выборочных логарифмических моментов, имеет место асимптотика при $n \to \infty$:

$$\Delta \equiv \left|\frac{\phi_2}{\hat\nu^3(X)} - \frac{\hat r^3(X)}{\hat\nu^3(X)} - M(X)\right| \longrightarrow 0\,,$$

почти наверное.

Данное соотношение при фиксированном объеме выборки n дает возможность определять «правильное» решение системы исходя из критерия

$$\Delta = \min\{\Delta_+, \Delta_-\}\,,$$

где

$$\Delta_\pm = \left|\frac{\phi_2}{\hat\nu_\pm^3(X)} - \frac{\hat r_\pm^3(X)}{\hat\nu_\pm^3(X)} - M(X)\right|. \tag{10}$$

Оценка параметра масштаба δ находится из уравнения (4) подстановкой найденного при помощи алгоритма решения $(\hat r(X), \hat\nu(X))$.

Другими словами, оценки $\hat r(X)$, $\hat\nu(X)$ и $\hat\delta(X)$ неизвестных параметров $0 \le r < 1$, $\nu > 0$ и $\delta > 0$ следует искать по формулам:

$$\hat{r}(X) = \\ = \frac{1}{\sqrt{K^2(X) - N(X)}} \left(N(X)\phi_1 - \operatorname{sgn}(\Delta_+ - \Delta_-) \times \right. \\ \left. \times K(X) \sqrt{N(X)(\phi_1^2 + \phi_3) - K^2(X)\phi_3} \right)^{1/2}; \quad (11)$$

$$\hat{\nu}(X) = \\ = \frac{1}{\sqrt{K^2(X) - N(X)}} \left(K(X)\phi_1 - \operatorname{sgn}(\Delta_+ - \Delta_-) \times \right. \\ \left. \times \sqrt{N(X)(\phi_1^2 + \phi_3) - K^2(X)\phi_3} \right)^{1/2}; \quad (12)$$

$$\hat{\delta}(X) = \exp\left\{ L_1(X) - \frac{\psi(t)}{\hat{\nu}(X)} + \frac{\psi(s)\hat{r}(X)}{\hat{\nu}(X)} \right\}. \quad (13)$$

Из усиленного закона больших чисел непосредственно вытекает следующее утверждение.

Теорема 1. *Определенные в* (11)–(13) *оценки* $\hat{r}(X)$, $\hat{\nu}(X)$ *и* $\hat{\delta}(X)$ *неизвестных параметров* $0 \le r < 1$, $\nu > 0$ *и* $\delta > 0$ *при фиксированных параметрах* s *и* t *распределения* (1) *обладают свойством сильной состоятельности*.

Замечание. В случае если известно, что $\nu < 0$, в формуле (12) следует поставить знак «минус» перед внешним радикалом.

4 Результаты численного моделирования

В данном разделе приводится ряд численных результатов, иллюстрирующих метод выбора «правильных» оценок параметров изгиба r, формы ν и масштаба δ при фиксированных параметрах концентрации s и t гамма-экспоненциального распределения (1).

В табл. 1–3 приводятся значения оценок $\hat{r}_\pm(X)$ и $\hat{\nu}_\pm(X)$, получаемые из соотношений (8) и (9). Предполагается априори известным, что $\nu > 0$. Индекс в обозначениях оценок соответствует рассматриваемому знаку перед радикалом в правых частях (8) и (9); тот же смысл имеет индекс $\Delta_\pm$ из (10). Оценки для δ получаются аналогично из соотношения (13).

При получении численных результатов использовались реализации выборок объема n из модельного гамма-экспоненциального распределения с параметрами $E = (r, \nu, s, t, \delta)$.

В табл. 1 приводятся значения оценок параметров r, ν и δ, полученные по реализации выборки из модельного распределения с набором параметров $E = (0{,}4; 1{,}7; 1{,}8; 1{,}4; 1)$. Оценки, соответствующие положительному знаку перед радикалом, автоматически отсеиваются ввиду отрицательных значений правой части (9).

Таблица 1 Оценки параметров модельного распределения GE (0,4; 1,7; 1,8; 1,4; 1)

n	$\hat{r}_+(X)$	$\hat{\nu}_+(X)$	$\hat{\delta}_+(X)$	Δ_+	$\hat{r}_-(X)$	$\hat{\nu}_-(X)$	$\hat{\delta}_-(X)$	Δ_-
10^3	—	—	—	—	0,2916	1,6848	0,9836	0,0500
10^4	—	—	—	—	0,3224	1,6740	0,9874	0,0192
10^5	—	—	—	—	0,3785	1,6861	0,9968	0,0096
10^6	—	—	—	—	0,3948	1,7003	0,9985	0,0016

Таблица 2 Оценки параметров модельного распределения GE (0,7; 0,5; 5,2; 3,7; 1)

n	$\hat{r}_+(X)$	$\hat{\nu}_+(X)$	$\hat{\delta}_+(X)$	Δ_+	$\hat{r}_-(X)$	$\hat{\nu}_-(X)$	$\hat{\delta}_-(X)$	Δ_-
10^3	—	—	—	—	—	—	—	—
10^4	2,9737	1,1629	22,580	29,793	0,8375	0,5329	1,4975	4,2301
10^5	3,3620	1,2817	27,648	30,929	0,7857	0,5173	1,3011	2,0116
10^6	4,0023	1,4974	34,098	32,457	0,7278	0,5055	1,0937	0,4892

Таблица 3 Оценки параметров модельного распределения GE (0,5; 0,9; 1,2; 2,7; 1)

n	$\hat{r}_+(X)$	$\hat{\nu}_+(X)$	$\hat{\delta}_+(X)$	Δ_+	$\hat{r}_-(X)$	$\hat{\nu}_-(X)$	$\hat{\delta}_-(X)$	Δ_-
10^3	0,6925	1,0964	1,1326	0,1240	0,0828	0,7207	0,9004	0,3813
10^4	0,6245	0,9839	1,0561	0,0779	0,1631	0,7032	0,8588	0,4085
10^5	0,4567	0,8683	0,9792	0,0508	0,3188	0,7812	0,9146	0,2042
10^6	0,5119	0,9087	1,0062	0,0143	0,2678	0,7551	0,8942	0,2518

В табл. 2 приводятся значения оценок параметров r, ν и δ, полученные по реализации выборки из модельного распределения с набором параметров $E = (0{,}7; 0{,}5; 5{,}2; 3{,}7; 1)$. Можно заметить, что для конкретной реализации выборки объема $n = 10^3$ оценки параметров не определены при любом знаке перед радикалом в (8) и (9). Это объясняется отрицательностью самого подкоренного выражения, вызванной тем, что данного объема выборки может быть недостаточно для использования рассматриваемых оценок. При бóльших объемах выборки следует выбирать тройку решений с отрицательным знаком перед радикалом, поскольку в противном случае оценка параметра r выходит за область определения параметра.

В табл. 3 приводятся значения оценок параметров r, ν и δ, полученные по реализации выборки из модельного распределения с набором параметров $E = (0{,}5; 0{,}9; 1{,}2; 2{,}7; 1)$. В данной таблице обе тройки оценок имеют непротиворечивые значения. Следует отсеять решение $(\hat{r}_-(X), \hat{\nu}_-(X), \hat{\delta}_-(X))$, поскольку $\Delta_- > \Delta_+$.

5 Заключение

В статье приведены явные виды оценок параметров изгиба, формы и масштаба при фиксированных параметрах концентрации гамма-экспоненциального распределения и алгоритм отсева лишних решений системы уравнений, основанной на логарифмических моментах. Поиск оценок параметров концентрации аналогичными методами сопряжен с принципиальными трудностями, связанными с обращением полигамма-функций.

Литература

1. *Крицкий С. Н., Менкель М. Ф.* О приемах исследования случайных колебаний речного стока // Труды НИУ ГУГМС. Сер. IV, 1946. Вып. 29. С. 3–32.
2. *Крицкий С. Н., Менкель М. Ф.* Выбор кривых распределения вероятностей для расчетов речного стока // Известия АН СССР. Отд. техн. наук, 1948. № 6. С. 15–21.
3. *McDonald J. B.* Some generalized functions for the size distribution of income // Econometrica, 1984. Vol. 52. No. 3. P. 647–665.
4. *Кудрявцев А. А.* О представлении гамма-экспоненциального и обобщенного отрицательного биномиального распределений // Информатика и её применения, 2019. Т. 13. Вып. 4. С. 78–82.
5. *Кудрявцев А. А., Титова А. И.* Гамма-экспоненциальная функция в байесовских моделях массового обслуживания // Информатика и её применения, 2017. Т. 11. Вып. 4. С. 104–108.
6. *Кудрявцев А. А.* Байесовские модели баланса // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 18–27.
7. *Кудрявцев А. А., Шестаков О. В.* Метод логарифмических моментов для оценивания параметров гамма-экспоненциального распределения // Информатика и её применения, 2020. Т. 14. Вып. 3. С. 49–54.
8. *Kudryavtsev A. A., Shestakov O. V.* Asymptotically normal estimators for the parameters of the gamma-exponential distribution // Mathematics, 2021. Vol. 9. Iss. 3. Art. 273. 13 p. doi: 10.3390/math9030273.
9. *Арутюнов Е. Н., Кудрявцев А. А., Недоливко Ю. Н.* Вероятностные характеристики индекса баланса факторов, имеющих обобщенные гамма-распределения // Информатика и её применения, 2021. Т. 15. Вып. 1. С. 65–71.

Поступила в редакцию 03.07.2021

A METHOD FOR ESTIMATING BENT, SHAPE AND SCALE PARAMETERS OF THE GAMMA-EXPONENTIAL DISTRIBUTION

A. A. Kudryavtsev[1,2], O. V. Shestakov[1,2,3], and S. Ya. Shorgin[3]

[1] Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[2] Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[3] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The article discusses a modified method of moments for estimating three of five parameters of the gamma-exponential distribution. It is proposed to estimate the distribution parameters based on its logarithmic moments. An explicit form of estimates of the bent, shape, and scale parameters is given for fixed concentration parameters of the gamma-exponential distribution; the strong consistency of the obtained estimates is justified. The article also discusses the method of eliminating unnecessary solutions to the system of equations for logarithmic

article moments; a number of numerical examples are presented to illustrate the derivation of estimates from model samples. Since the analyzed distribution is closely related to the generalized gamma distribution and the generalized beta distribution of the second kind, the results of this work can be widely used in applied problems using continuous distributions with an unbounded nonnegative support for modeling.

Keywords: parameter estimation; gamma-exponential distribution; mixed distributions; generalized gamma distribution; method of moments; consistent estimate

DOI: 10.14357/19922264210308

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (project 20-07-00655). The paper is published with the financial support of the Ministry of Education and Science of the Russian Federation as a part of the program of the Moscow Center for Fundamental and Applied Mathematics under the agreement No. 075-15-2019-1621.

References

1. Kritsky, S. N., and M. F. Menkel. 1946. O priemakh issledovaniya sluchaynykh kolebaniy rechnogo stoka [Methods of investigation of random fluctuations of river flow]. *Trudy NIU GUGMS Ser. IV* [Proceedings of GUGMS Research Institutions. Ser. IV] 29:3–32.
2. Kritsky, S. N., and M. F. Menkel. 1948. Vybor krivykh raspredeleniya veroyatnostey dlya raschetov rechnogo stoka [Selection of probability distribution curves for river flow calculations]. *Izvestiya AN SSSR. Otd. tekhn. nauk* [Herald of the USSR Academy of Sciences. Technical Sciences] 6:15–21.
3. McDonald, J. B. 1984. Some generalized functions for the size distribution of income. *Econometrica* 52(3):647–665.
4. Kudryavtsev, A. A. 2019. O predstavlenii gamma-eksponentsial'nogo i obobshchennogo otritsatel'nogo binomial'nogo raspredeleniy [On the representation of gamma-exponential and generalized negative binomial distributions]. *Informatika i ee Primeneniya — Inform. Appl.* 13(4):78–82.
5. Kudryavtsev, A. A., and A. I. Titova. 2017. Gamma-eksponentsial'naya funktsiya v bayesovskikh modelyakh massovogo obsluzhivaniya [Gamma-exponential function in Bayesian queuing models]. *Informatika i ee Primeneniya — Inform. Appl.* 11(4):104–108.
6. Kudryavtsev, A. A. 2018. Bayesovskie modeli balansa [Bayesian balance models]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):18–27.
7. Kudryavtsev, A. A., and O. V. Shestakov. 2020. Metod logarifmicheskikh momentov dlya otsenivaniya parametrov gamma-eksponentsial'nogo raspredeleniya [Method of logarithmic moments for estimating the gamma-exponential distribution parameters]. *Informatika i ee Primeneniya — Inform. Appl.* 14(3):49–54.
8. Kudryavtsev, A. A., and O. V. Shestakov. 2021. Asymptotically normal estimators for the parameters of the gamma-exponential distribution. *Mathematics* 9(3):273. 13 p. doi: 10.3390/math9030273.
9. Arutyunov, E. N., A. A. Kudryavtsev, and Iu. N. Nedolivko. 2021. Veroyatnostnye kharakteristiki indeksa balansa faktorov, imeyushchikh obobshchennye gamma-raspredeleniya [Probabilistic characteristics of balance index of factors with generalized gamma distribution]. *Informatika i ee Primeneniya — Inform. Appl.* 15(1):65–71.

Received July 3, 2021

Contributors

Kudryavtsev Alexey A. (b. 1978) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; nubigena@mail.ru

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 Leninskie 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; leading scientist, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; oshestakov@cs.msu.su

Shorgin Sergey Ya. (b. 1952) — 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 119333, Russian Federation; sshorgin@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 63–74

МЕТОД ПОВЫШЕНИЯ ТОЧНОСТИ НЕЙРОСЕТЕВЫХ ПРОГНОЗОВ С ИСПОЛЬЗОВАНИЕМ СМЕШАННЫХ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ И ЕГО РЕАЛИЗАЦИЯ В ВИДЕ ЦИФРОВОГО СЕРВИСА*

А. К. Горшенин[1], В. Ю. Кузьмин[2]

Аннотация: Описан метод комбинированного использования классических вероятностно-статистических моделей и нейронных сетей, ориентированный на повышение точности прогнозирования. При этом моменты математических моделей используются для нетривиального расширения признакового пространства при обучении нейронных сетей. На примере анализа нескольких ансамблей экспериментальных данных стелларатора Л-2М продемонстрирована эффективность предложенного подхода, особенно при использовании моментов моделей, полученных для приращений исходных наблюдаемых данных. Для реализации методов статистического анализа и предложенных алгоритмов машинного обучения создан цифровой сервис, архитектура и основные возможности которого также обсуждаются в рамках данной статьи.

Ключевые слова: нейронные сети; конечные нормальные смеси; вероятностные модели; прогнозирование; цифровой сервис; высокопроизводительные вычисления; турбулентная плазма; стелларатор

DOI: 10.14357/19922264210309

1 Введение

Создание современных методов анализа данных в финансовой сфере, медицине или метеорологии зачастую связано с использованием различных математических моделей, в том числе вероятностно-статистических, основанных на выборках случайного объема [1]. В частности, могут быть упомянуты исследования распределений размеров частиц реголита [2] или различных характеристик осадков [3]. Возникающие при этом смешанные вероятностные модели позволяют учитывать существенно неоднородный характер анализируемых данных. При этом их характеристики (например, моменты распределений) могут быть использованы в различных методах машинного обучения как дополнительные данные-признаки [4]. Таким образом, в рамках этого подхода при обучении возможно использование большего набора данных без непосредственного увеличения объема выборок исходных наблюдений.

Впервые данная идея была предложена авторами в работе [5] в рамках использования статистических подходов в области анализа процессов в турбулентной плазме [6]. Действительно, методы машинного обучения и нейронные сети в исследованиях турбулентной плазмы позволяют добиваться заметных результатов как в вопросах моделирования наблюдаемых явлений [7–10], так и в задачах анализа и прогнозирования нестабильностей и разрушительных для стеллараторов и токамаков эффектов [11, 12].

В данной статье акцент делается на анализе экспериментальных данных стелларатора Л-2М [13]. Созданные методы на стыке классического вероятностно-статистического моделирования и машинного обучения реализованы в виде инструментов цифрового сервиса. Он ориентирован на предоставление единого интерфейса для доступа как к различным статистическим методам, так и к алгоритмам нейросетевого прогнозирования, развиваемым коллективом авторов для анализа данных в широком спектре научных областей.

2 Описание метода и конфигурации используемых нейронных сетей

Как было упомянуто выше, смешанные вероятностные модели являются качественными аппрок-

*Работа выполнялась с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва). Работа выполнена при частичной поддержке Стипендии Президента Российской Федерации молодым ученым (СП-3956.2021.5).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, agorshenin@frccsc.ru

[2]Факультет космических исследований Московского государственного университета имени М. В. Ломоносова, shadesilent@yandex.ru

симациями для описания распределений процессов в различных областях. Поэтому в качестве дополнительных данных (признаков), которые могут использоваться в методах машинного обучения, было предложено использование первых четырех моментов соответствующего распределения, получаемых на итерационных шагах процедуры, называемой методом скользящего разделения смесей (CPC-метод) [1].

Каждому вектору $\mathbf{X}$, который подается в качестве входного в нейронную сеть, ставится в соответствие набор величин $(\mathbb{E}_X^{(n)}, \mathbb{D}_X^{(n)}, \gamma_X^{(n)}, \kappa_X^{(n)})$, определяемых следующими формулами (см. статью [14]):

– математическое ожидание:

$$\mathbb{E}_X^{(n)} = \sum_{i=1}^{k(n)} p_i(n) a_i(n); \tag{1}$$

– дисперсия:

$$\mathbb{D}_X^{(n)} = \sum_{i=1}^{k(n)} p_i(n) \left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 + \sum_{i=1}^{k(n)} p_i(n) \sigma_i^2(n); \tag{2}$$

– коэффициент асимметрии:

$$\gamma_X^{(n)} = \left[\sum_{i=1}^{k(n)} p_i(n) \left(a_i^3(n) + 3 a_i(n) \sigma_i^2(n) \right) - \left(\sum_{i=1}^{k(n)} p_i(n) a_i(n) \right) \times \left(3 \sum_{i=1}^{k(n)} p_i(n) \left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 + 3 \sum_{i=1}^{k(n)} p_i(n) \sigma_i^2(n) - \left(\sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 \right) \right] \times \left[\sum_{i=1}^{k(n)} p_i(n) \left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 + \sum_{i=1}^{k(n)} p_i(n) \sigma_i^2(n) \right]^{-3/2}; \tag{3}$$

– коэффициент эксцесса:

$$\kappa_X^{(n)} = \left[\sum_{i=1}^{k(n)} p_i(n) \left(a_i^4(n) + 6 a_i^2 \sigma_i^2(n) + 3 \sigma_i^4(n) \right) - 3 \left(\sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^4 - 4 \left(\sum_{i=1}^{k(n)} p_i(n) a_i(n) \right) \times \left(\sum_{i=1}^{k(n)} p_i(n) \left(a_i^3(n) + 3 a_i(n) \sigma_i^2(n) \right) \right) + 6 \left(\sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 \left(\sum_{i=1}^{k(n)} p_i(n) \left(a_i^2(n) + \sigma_i^2(n) \right) \right) \right] \times \left[\sum_{i=1}^{k(n)} p_i(n) \left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 + \sum_{i=1}^{k(n)} p_i(n) \sigma_i^2(n) \right]^{-2} - 3. \tag{4}$$

Здесь предполагается, что все элементы данного вектора имеют распределение типа конечной нормальной смеси $\sum_{i=1}^{k(n)} p_i \Phi\left((x - a_i(n)) \sigma_i^{-1}(n) \right)$, $x \in \mathbb{R}$, со стандартными ограничениями на параметры: $a_i(n) \in \mathbb{R}$, $\sigma_i(n) \in \mathbb{R}$ и $\sigma_i(n) > 0$, $\sum_{i=1}^{k(n)} p_i(n) = 1$, $p_i(n) \geqslant 0$ для всех $i = \overline{1, k(n)}$.

Аргумент n у каждой из данных величин показывает зависимость от номера шага CPC-метода, т. е. данные моменты определяются не для всего ряда, а только для его части — окна. При этом предполагается, что указанные моменты определяются по наблюдениям, которые отстоят от первого элемента $\mathbf{X}$ (согласно его расположению в анализируемом ряду) на величину L скользящего окна CPC-метода. Это позволяет увеличить объем данных для обучения без необходимости расширения экспериментальных выборок. Кроме того, указанные моменты не содержат информацию о том, как ведет себя ряд после данного наблюдения, и поэтому могут быть корректно использованы при построении прогнозов. Помимо таких модельных моментов могут быть использованы их выборочные аналоги. Подобная конфигурация для расширения признакового пространства также будет рассматриваться при сравнении результатов.

Выходными данными нейронной сети являются M последовательных предсказанных наблюдений, т. е. прогноз, соответствующий истинным наблюдениям $D_{i+L}, \ldots, D_{i+L+M-1}$.

Для выбора наилучшей архитектуры конструируемых нейронных сетей необходима оптимизация набора гиперпараметров. В данном исследовании гиперпараметры разделены на две категории:

(1) фиксированные: общие для всех обучаемых моделей;
(2) изменяемые: в пространстве которых производится поиск оптимальной архитектуры.

К первым из них относятся следующие гиперпараметры:

- каждая нейронная сеть является многослойным перцептроном;
- для нейронов, за исключением выходного слоя, используется функция активации `ReLU`;
- для выходного слоя используется функция линейной активации;
- в качестве функции потерь выбрана стандартная метрика — среднеквадратичная ошибка (`RMSE`, Root Mean Square Error);
- скорость обучения при выходе на плато функции потерь снижается.

В качестве переменных гиперпараметров выбраны:

- оптимизаторы: рассмотрены четыре оптимизатора из семейства `Adam` (`Adam`, `AdaMax`, `AdaDelta` и `NAdam`) [15];
- коэффициент дропаута [16]: возможные значения — 0, 0,1, 0,25, 0,5;
- количество слоев и нейронов в каждом слое:
  - один скрытый слой с 50/100/150/200 нейронами;
  - два скрытых слоя с 100/200 нейронами в каждом;
  - три скрытых слоя с 50/100/200 нейронами в каждом;
- варианты обогащения исходного набора данных, т. е. нетривиального расширения признакового пространства.

Для оптимизации гиперпараметров использовался поиск в гиперкубе [17], т. е. для каждой комбинации гиперпараметров была создана и обучена нейронная сеть. Всего проанализированы 576 комбинаций. Обучение модели завершалось по истечению 500 эпох или в случае неуменьшения функции потерь на протяжении 35 эпох. Для определения величин моментов (1)–(4) на шагах CPC-метода использованы `EM`-алгоритмы, по сей день остающиеся одними из наиболее востребованных методов оценивания параметров различных распределений [18–21], с разнообразными настройками.

3 Описание сервиса и пример анализа экспериментальных данных

Для автоматизации анализа данных создан гетерогенный вычислительный сервис, реализующий комбинацию статистических методов, в частности алгоритмов скользящего разделения смесей, и инструментов машинного обучения. Отдельные вопросы, связанные с архитектурой и интерфейсом подобного сервиса, а также с реализацией в его рамках CPC-алгоритма, были представлены авторами в статьях [14, 22–25]. В текущей версии сервиса к имеющимся возможностям добавлена реализация быстрого решения задач классификации и предсказания разнообразных данных, например финансовых индексов или экспериментальных рядов турбулентной плазмы.

Для пользователя сервис предоставляет следующий функционал:

- возможность загрузки и предварительного анализа данных;
- инструменты для CPC-анализа данных [25] и получения моментных характеристик смешанных моделей;
- обучение нейронных сетей по заданной архитектуре, включая опции нетривиального расширения признакового пространства (см. разд. 2);
- подбор гиперпараметров нейронных сетей для заданного набора архитектур с использованием различных подходов [26];
- визуализация процесса обучения;
- сравнительный анализ точности обученных моделей;
- визуализация полученных результатов;
- сохранение обученных моделей для их дальнейшего использования.

При анализе временны́х рядов нестационарной структуры большое значение имеет подбор корректной архитектуры и оптимизация гиперпараметров. Таким образом, логично разделить процесс обработки данных на несколько этапов.

Первый этап — загрузка, предварительная обработка и визуализация данных. Каждая загруженная выборка является отдельной сущностью, которую можно использовать как для создания моделей на последующих шагах, так и для построения новых выборок: например, приведение временно́го ряда к разностному временно́му ряду либо добавление дополнительных характеристик. Кроме то-

го, предусмотрены средства для визуализации ряда и/или его части, основных трендов и набора выборочных статистических характеристик. Предварительная обработка данных осуществляется с помощью EM-алгоритма. Доступны графики эволюции компонент смесей во времени, сравнительный анализ полученных статистических свойств выборки в зависимости от настройки размера и шага окон.

В дальнейшем загруженная выборка и порожденные ею выборки могут быть использованы в качестве исходных данных для задач машинного обучения. Пользователь сервиса выбирает варианты архитектур, диапазон изменения доступных гиперпараметров и набор статистических методов для обогащения данных. Процесс обучения в силу высокой вычислительной сложности происходит асинхронно.

Финальная стадия обработки данных — сравнение точности полученных моделей, визуализация решения задачи прогнозирования и/или классификации и сохранение обученной модели для дальнейшего использования. Примеры графиков, используемых для сравнения различных методов обогащения, приведены ниже. Так, на рис. 1 приведено сравнение точности прогнозов для одного из моментов эксперимента с турбулентной плазмой. На графиках изображены исходные данные и нейросетевые прогнозы на 1 и 30 шагов на основе только исходных наблюдений (слева вверху) и с расширением признакового пространства за счет:

- выборочных моментов (справа вверху);
- модельных моментов для исходных данных (слева внизу);
- смешанных нормальных моментов для приращений (справа внизу).

Здесь и далее на горизонтальной оси нанесена временна́я шкала эксперимента (от начала его активной стадии), а на вертикальной — значение физической величины, наблюдаемой в эксперименте.

На рис. 2 продемонстрированы возможности сервиса в области сравнения величины ошибок, получаемых для различных конфигурациях нейронных сетей, с использованием классической метрики RMSE (на примере различных вариантов расширения признакового пространства, описанных выше).

Структурно сервис состоит из пяти частей:

(1) Frontend — обеспечивает создание пользовательского интерфейса, отображение графиков, первичную обработку загруженных данных;

(2) Backend — отвечает за взаимодействие Frontend и других компонент сервиса, валидацию данных, преобразование данных внутри сервиса;

(3) сервер баз данных — служит персистентным хранилищем информации;

(4) балансировщик нагрузки на рабочие серверы;

(5) множество рабочих серверов — реализует алгоритмы статистического анализа и методы машинного обучения с высокой вычислительной сложностью.

В такой структуре сервиса динамическое масштабирование достигается путем изменения числа рабочих серверов и балансировки нагрузки на соединении Backend и рабочих серверов. Балансировка на уровне Backend и сервера баз данных возможна без концептуального изменения архитектуры сервиса. При этом рабочие серверы должны иметь доступ к ресурсам графических карт (`GPU`). Пул таких серверов может быть создан на основе множества провайдеров облачных сервисов. В данной работе для непосредственного обучения нейронных сетей использовалась инфраструктура ЦКП «Информатика» ФИЦ ИУ РАН. Таким образом, данный сервис может предоставлять интерфейс доступа к удаленному высокопроизводительному вычислительному оборудованию.

Непосредственно для создания сервиса были использованы:

- vue.js — для создания пользовательского интерфейса;
- PHP-фреймворк Yii2 — для реализации Backend;
- mySQL/PostgreSQL как база данных;
- Python как основа для рабочих серверов и связка Tensorflow + NVIDIA CUDA для осуществления доступа к GPU-ресурсам.

С использованием реализованных в сервисе инструментов проанализированы пять ансамблей физических экспериментов, полученных в разных режимах функционирования стелларатора Л-2М. Данные каждого эксперимента представляют собой временной ряд колебаний плотности плазмы, причем содержательную часть составляют около 60 000 наблюдений. Значения параметров L (размер окна) и M (максимальная длительность прогноза) выбраны как $L = 200$ и $M = 30$. Перед началом обработки данных состояние генератора случайных чисел фиксируется для обеспечения воспроизводимости результатов. Исходные данные были разделены на обучающие и тестовые выборки в соотношении 70% к 30%. Расчеты проводились с использованием гибридного высокопроизводительного вычислительного кластера на базе архитектуры `Power9` с тактовой частотой 2,0 ГГц (20 ядер)

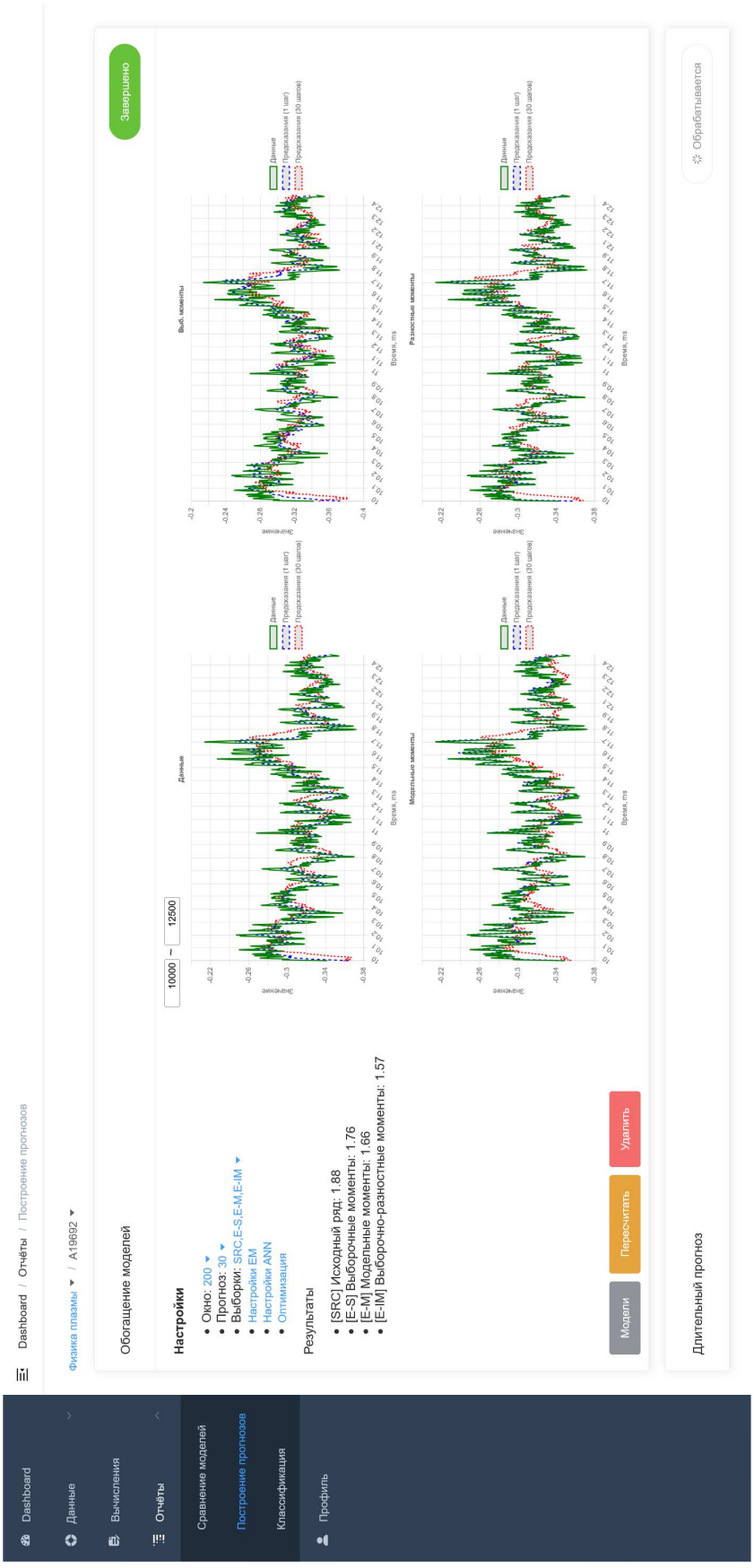


Рис. 1 Пример прогноза для ряда A19692 (10,00–12,5 мс) в интерфейсе сервиса

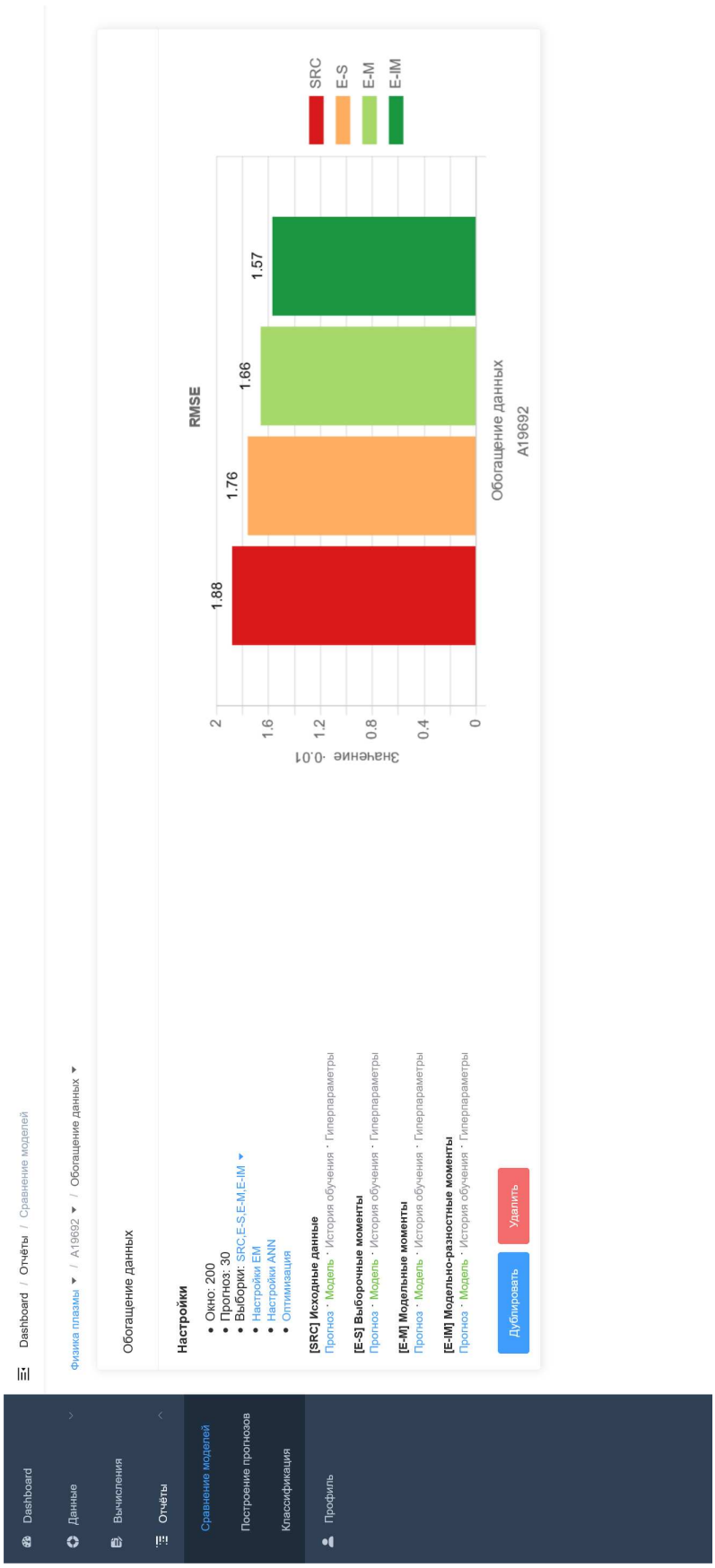


Рис. 2 Сравнение величин ошибок прогноза для ряда A19692 в метрике RMSE для различных конфигураций в интерфейсе сервиса

и 4 видеокартами NVIDIA Volta V100 (общий объем памяти 16 ГБ).

Оптимальные конфигурации гиперпараметров для обогащенных и исходных данных для всех пяти ансамблей близки. Нейронные сети с одним скрытым слоем, состоящим из большого числа нейронов, приводят к более точным прогнозам, чем глубокие нейронные сети, в которых скрытые слои содержат схожее суммарное число нейронов. Выбор оптимизатора зависит от наблюдаемых наборов данных, в среднем точность и скорость обучения оказались немного выше при использовании оптимизатора Adam.

Во всех построенных нейронных сетях не наблюдалось эффекта переобучения. Ненулевой коэффициент для дропаута отрицательно сказывался на функции потерь и скорости обучения, наилучшие результаты были достигнуты без его применения. В большинстве случаев обучение модели заканчивалось до истечения максимально допустимого порога в 500 эпох.

Выбор между обогащенными и необогащенными данными сильно повлиял на точность, скорость обучения и значение функции потерь. У обогащенных наборов данных время обучения модели было на 20%–30% медленнее, при этом не наблюдалось серьезных различий в скорости обучения между различным образом обогащенными наборами данных.

На рис. 3 и 4 продемонстрированы полученные результаты прогнозирования еще для нескольких ансамблей (помимо рассмотренного на рис. 1) в некоторых случайно выбранных диапазонах. Для всех проанализированных экспериментов прогнозы на 1 и 30 наблюдений следуют общим трендам в данных. При этом обучение на обогащенных наборах позволяет создать модели, которые лучше адаптируются к быстрым изменениям данных. В качестве примера можно привести ансамбль A20229 (см. рис. 3). Кроме того, модельное и модельно-разностное обогащения приводят к лучшему предсказанию пиковых значений (локальных минимумов и максимумов) по сравнению с анализом исходных данных или данными с простым обогащением (см., например, рис. 1, интервал 11,40–11,60 мс).

Таблица демонстрирует различие в функции потерь (RMSE) для трех экспериментов, полученных при использовании данных с различными вариантами обогащения.

Для всех рядов получено повышение точности прогнозирования (относительно значений метрики RMSE) при расширении признакового пространства. При этом минимальная ошибка получена для исследуемых ансамблей в случае модельно-разностного расширения признакового пространства (см. правый столбец в таблице. Наибольшая разница наблюдается со случаем, когда не проведено расширение признакового пространства — в такой ситуации преимущество может доходить до 80,71% (ансамбль A20264). По сравнению с выборочными моментами наибольший прирост точности для моментно-разностной конфигурации составил 30,31% (ряд A20229), а по сравнению с CPC-моментами для сходных наблюдений — 18,91% (также для ансамбля A20229).

Необходимо отметить, что прогнозирование данных стелларатора представляет существенный прикладной интерес в таких задачах обработки турбулентной плазмы, как:

- верификация рядов, полученных в рамках проведения эксперимента с едиными начальными условиями;
- восстановление сигналов при временном прекращении работы регистрационного оборудования;
- анализ профилей плотности токов увлечения и поглощения электронно-циклотронного нагрева (в частности, для термоядерного реактора ITER [12, 27]).

4 Заключение

В статье представлен подход к проблеме предсказания поведения турбулентной плазмы с использованием машинного обучения. Продемонстрирована эффективность применения статистических моделей на основе конечных смесей нормальных распределения для приращений данных в качестве источника наблюдений для нетривиального расширения признакового пространства.

Предлагаемый подход может быть использован и для верификации смешанной модели, построенной для исходного ряда. Действительно, в случае значительного повышения точности прогноза данных при использовании именно характеристик аппроксимирующей смешанной модели это может служить индикатором корректности ее выбора в качестве математического описания изучаемого явления.

В качестве направления дальнейших исследований можно указать усложнение архитектуры используемых нейронных сетей, например за счет использования рекуррентных LSTM-слоев [28] или ансамблей нейронных сетей. Все они будут в дальнейшем интегрированы в созданный цифровой сервис. Однако даже для относительно простой конфигурации на примере многослойного перцептрона

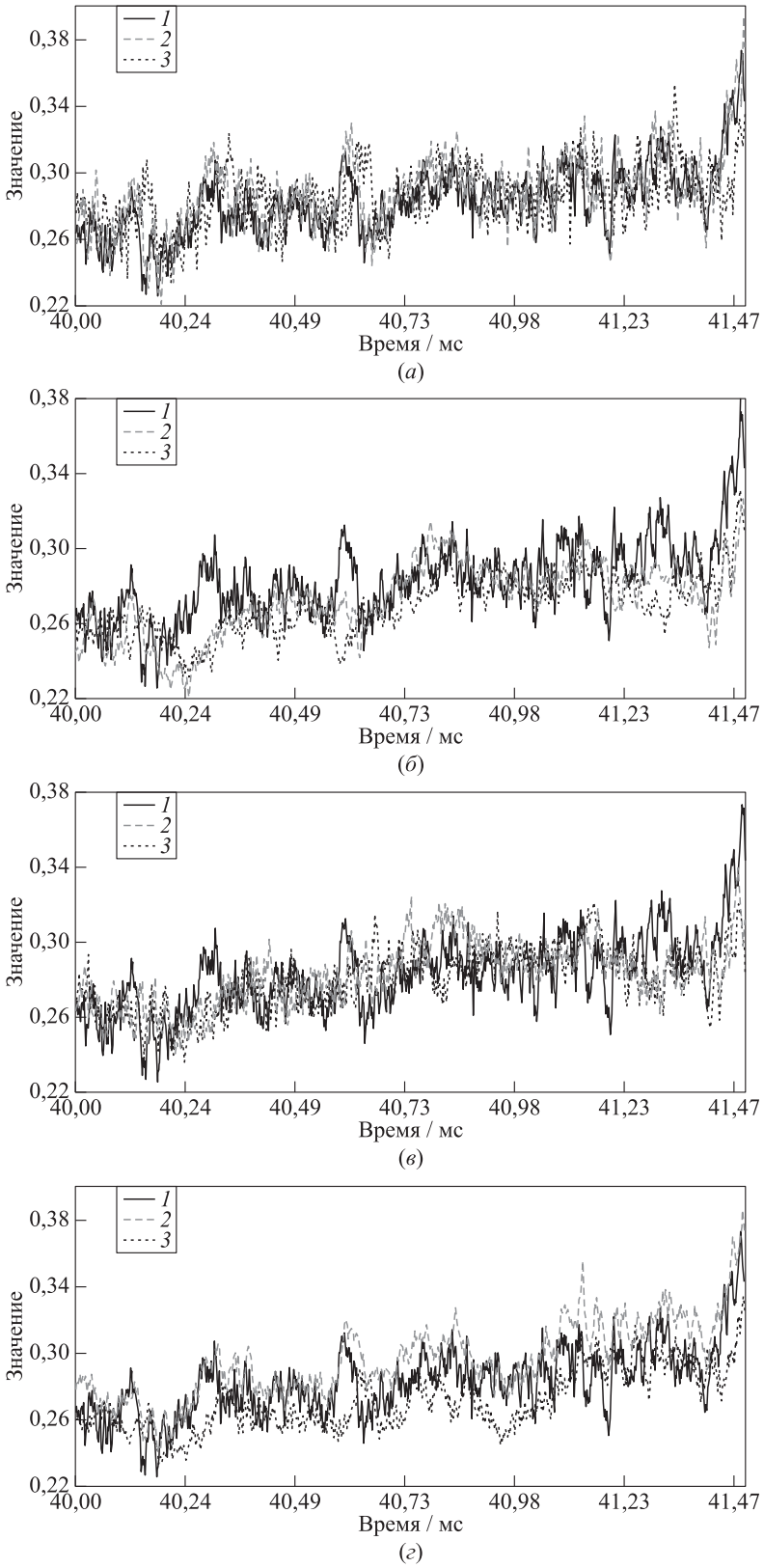


Рис. 3 Пример прогноза для ряда A20229 (40,00–41,50 мс): (*а*) исходные данные; (*б*) данные с обогащением выборочными моментами; (*в*) данные с модельным обогащением; (*г*) данные с модельно-разностным обогащением; *1* — исходные данные; *2* — прогноз на 1 шаг; *3* — прогноз на 30 шагов

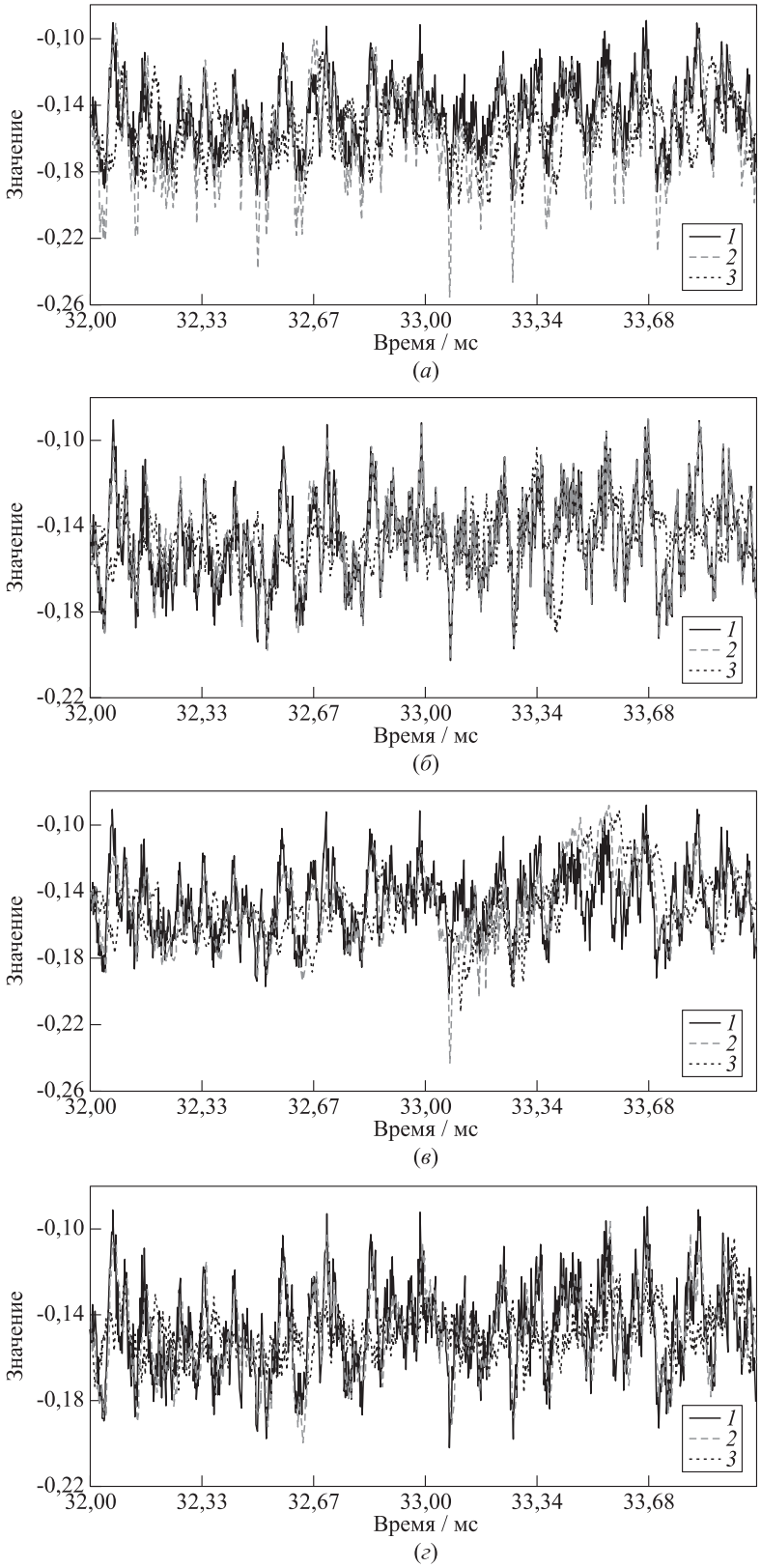


Рис. 4 Пример прогноза для ряда A20264 (32,00–34,00 мс): (*а*) исходные данные; (*б*) данные с обогащением выборочными моментами; (*в*) данные с модельным обогащением; (*г*) данные с модельно-разностным обогащением; *1* — исходные данные; *2* — прогноз на 1 шаг; *3* — прогноз на 30 шагов

Сравнение точности прогнозирования в метрике RMSE для различных конфигураций

Ансамбль	Исходные данные	Выборочные моменты	Модельное обогащение	Модельно-разностное обогащение
A19692ch1	0,0842	0,0817	0,0815	**0,0809**
A19692ch2	0,035	0,0348	0,0345	**0,0341**
A19692	0,0188	0,0176	0,0166	**0,0157**
A20229	0,0631	0,0503	0,0459	**0,0386**
A20264	0,0562	0,0325	0,0323	**0,0311**

продемонстрирована эффективность предлагаемого комбинированного подхода на основе использования математической модели и ее характеристик при решении задач машинного обучения.

Литература

1. *Королев В. Ю.* Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. — М.: Изд-во Моск. ун-та, 2011. 512 с.
2. *Gorshenin A. K., Korolev V. Yu., Zeifman A. I.* Modeling particle size distribution in lunar regolith via a central limit theorem for random sums // Mathematics, 2020. Vol. 8. Iss. 9. Art. 1409.
3. *Korolev V. Yu., Gorshenin A. K.* Probability models and statistical tests for extreme precipitation based on generalized negative binomial distributions // Mathematics, 2020. Vol. 8. Iss. 4. Art. 604.
4. *Chandrashekar G., Sahin F.* A survey on feature selection methods // Comput. Electr. Eng., 2014. Vol. 40. Iss. 1. P. 16–28.
5. *Gorshenin A., Kuzmin V.* On feature expansion with finite normal mixture models in machine learning // Advances in intelligent systems, computer science and digital economics II / Eds. Zh. Hu, S. Petoukhov, M. He. — Advances in intelligent systems and computing ser. — Springer, 2021. Vol. 1402. P. 82–90.
6. *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. 075006. 7 p.
7. *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. 060702. 4 p.
8. *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 Computing Applications, 2018. Vol. 29. Iss. 6. P. 83–109.
9. *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. 30LT02. 9 p.
10. *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. 106018. 17 p.
11. *Parsons M. S.* Interpretation of machine-learning-based disruption models for plasma control // Plasma Phys. Contr. F., 2017. Vol. 59. Iss. 8. Art. 085001. 5 p.
12. *Kates-Harbeck J., Svyatkovskiy A., Tang W.* Predicting disruptive instabilities in controlled fusion plasmas through deep learning // Nature, 2019. Vol. 568. Iss. 7753. P. 526–531.
13. *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.
14. *Горшенин А. К.* Концепция онлайн-комплекса для стохастического моделирования реальных процессов // Информатика и её применения, 2016. Т. 10. Вып. 1. С. 72–81. doi: 10.14357/19922264160107.
15. *Buduma N.* Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. — Sebastopol, CA, USA: O'Reilly Media, 2017. 298 p.
16. *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.
17. *Gorshenin A. K., Kuzmin V. Yu.* Improved architecture and configurations 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. *Gorshenin A. K.* On implementation of EM-type algorithms in the stochastic models for a matrix computing on GPU // AIP Conf. Proc., 2015. Vol. 1648. Art. 250008. 4 p.
19. *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.
20. *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.
21. *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.
22. *Gorshenin A., Kuzmin V.* Online system for the construction of structural models of information flows // 7th Congress

(International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings. — Piscataway, NJ, USA: IEEE, 2015. P. 216–219.

23. *Gorshenin A., Kuzmin V.* On an interface of the online system for a stochastic analysis of the varied information flows // AIP Conf. Proc., 2016. Vol. 1738. Art. 220009. 4 p.
24. *Горшенин А. К.* О некоторых математических и программных методах построения структурных моделей информационных потоков // Информатика и её применения, 2017. Т. 11. Вып. 1. С. 58–68. doi: 10.14357/19922264170105.
25. *Gorshenin A. K., Kuzmin V. Yu.* Research support system for stochastic data processing // Pattern Recognition Image Analysis, 2017. Vol. 27. No. 3. P. 518–524.
26. *Bergstra J., Bengio Y.* Random search for hyper-parameter optimization // J. Mach. Learn. Res., 2012. Vol. 13. P. 281–305.
27. *Aymar R., Barabaschi P., Shimomura Y.* The ITER design // Plasma Phys. Contr. F., 2002. Vol. 44. Iss. 5. P. 519–565.
28. *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.

Поступила в редакцию 17.07.2021

METHOD FOR IMPROVING ACCURACY OF NEURAL NETWORK FORECASTS BASED ON PROBABILITY MIXTURE MODELS AND ITS IMPLEMENTATION AS A DIGITAL SERVICE

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

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

[2]Faculty of Space Research, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation

Abstract: A method aimed at improving the forecasting accuracy is presented. It uses a combination of classical probabilistic-statistical models and neural networks. Moments of mathematical models are used as a nontrivial expansion of the feature space. The efficiency of the proposed approach is demonstrated by the analysis of several experimental data ensembles of the L-2M stellarator. Error decrease is especially noticeable when using the moments of the statistical models based on the increments of the initial observed data. To implement the methods of statistical analysis and the proposed machine learning algorithms, a digital service has been created. Its architecture and capabilities are also outlined.

Keywords: neural networks; finite normal mixtures; probability models; forecasting; digital service; high-performance computing; turbulence plasma; stellarator

DOI: 10.14357/19922264210309

Acknowledgments

The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow). The research was partially supported by the RF Presidential scholarship program (project No. 3956.2021.5).

References

1. 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.
2. Gorshenin, A. K., V. Yu. Korolev, and A. I. Zeifman. 2020. Modeling particle size distribution in lunar regolith via a central limit theorem for random sums. *Mathematics* 8(9):1409. 24 p.
3. Korolev, V. Yu., and A. K. Gorshenin. 2020. Probability models and statistical tests for extreme precipitation based on generalized negative binomial distributions. *Mathematics* 8(4):604.
4. Chandrashekar, G., and F. Sahin. 2014. A survey on feature selection methods. *Comput. Electr. Eng.* 40(1):16–28.
5. Gorshenin, A., and V. Kuzmin. 2021. On feature expansion with finite normal mixture models in machine learning. *Advances in intelligent systems, computer science and digital economics II.* Eds. Zh. Hu, S. Petoukhov, and M. He. Advances in intelligent systems and computing ser. Springer. 1402:82–90.
6. 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 dur-

ing transient process under electron cyclotron resonance heating of the L-2M stellarator plasma. *Plasma Phys. Contr. F.* 61(7):075006. 7 p.

7. 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. 4 p.
8. 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 Computing Applications* 29(6):83–109.
9. 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. 9 p.
10. 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. 17 p.
11. Parsons, M. S. 2017. Interpretation of machine-learning-based disruption models for plasma control. *Plasma Phys. Contr. F.* 59(8):085001. 5 p.
12. Kates-Harbeck, J., A. Svyatkovskiy, and W. Tang. 2019. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature* 568(7753):526–531.
13. 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.
14. Gorshenin, A. K. 2016. Kontseptsiya onlayn-kompleksa dlya stokhasticheskogo modelirovaniya real'nykh protsessov [Concept of online service for stochastic modeling of real processes]. *Informatika i ee Primeneniya — Inform. Appl.* 10(1):72–81. doi: 10.14357/19922264160107.
15. Buduma, N. 2017. *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms*. Sebastopol, CA: O'Reilly Media. 298 p.
16. 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.
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):68–77.
18. Gorshenin, A. K. 2015. On implementation of EM-type algorithms in the stochastic models for a matrix computing on GPU. *AIP Conf. Proc.* 1648:250008. 4 p.
19. 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.
20. 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.
21. 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.
22. Gorshenin, A., and V. Kuzmin. 2015. Online system for the construction of structural models of information flows. *7th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings*. Piscataway, NJ: IEEE. 216–219.
23. Gorshenin, A., and V. Kuzmin. 2016. On an interface of the online system for a stochastic analysis of the varied information flows. *AIP Conf. Proc.* 1738:220009. 4 p.
24. Gorshenin, A. K. 2017. O nekotorykh matematicheskikh i programmnykh metodakh postroeniya strukturnykh modeley informatsionnykh potokov [On some mathematical and programming methods for construction of structural models of information flows]. *Informatika i ee Primeneniya — Inform. Appl.* 11(1):58–68. doi: 10.14357/19922264170105.
25. Gorshenin, A. K., and V. Yu. Kuzmin. 2017. Research support system for stochastic data processing. *Pattern Recognition Image Analysis* 27(3):518–524.
26. Bergstra, J., and Y. Bengio. 2012. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13:281–305.
27. Aymar, R., P. Barabaschi, and Y. Shimomura. 2002. The ITER design. *Plasma Phys. Contr. F.* 44(5):519–565.
28. 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.

Received July 17, 2021

Contributors

Gorshenin Andrey K. (b. 1986) — Doctor of Science in physics and mathematics, associate professor, head of department, leading scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; associate professor, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1,Moscow 119991, Russian Federation; agorshenin@frccsc.ru

Kuzmin Victor Yu. (b. 1986) — programmer, Faculty of Space Research, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; shadesilent@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 75–82

МЕТОД ВИЗУАЛИЗАЦИИ СТИМУЛЯЦИИ КОНФЛИКТОВ В ГИБРИДНЫХ ИНТЕЛЛЕКТУАЛЬНЫХ МНОГОАГЕНТНЫХ СИСТЕМАХ

С. Б. Румовская[1], И. А. Кириков[2]

Аннотация: Рассмотрение сложных задач (проблем) в малых коллективах специалистов различных направлений позволяет получить более качественное интегрированное решение. В таких коллективах неизбежно возникают конфликты — деструктивные (личные) и конструктивные (инструментальные). Моделирование работы таких коллективов и стимулирование конструктивных конфликтов в них позволит повысить качество решения и выработать метод решения, релевантный проблеме. Визуализация — мощный инструмент для извлечения, анализа и понимания информации. Работа посвящена разработке метода визуализации процессов стимулирования конфликтов в рамках гибридных интеллектуальных многоагентных систем (ГиИМАС), моделирующих агентами рассуждения отдельных специалистов и отображающих в памяти компьютера макроуровневые процессы, возникающие в результате взаимодействия специалистов при решении проблемы «за круглым столом».

Ключевые слова: коллектив экспертов; конфликт агентов; визуализация стимуляции конфликта

DOI: 10.14357/19922264210310

1 Введение

В практической деятельности коллективное решение проблем все более предпочтительнее перед индивидуальным ввиду их гетерогенности, растущего разнообразия знаний, необходимых для их решения, и растущих объемов обрабатываемой информации. В коллективах неизбежны конфликты различного характера: деструктивные конфликты, связанные преимущественно с отношениями, личными вопросами и проблемами, и конструктивные — инструментальные конфликты относительно проблемы или процесса ее решения. Фасилитатор (член коллектива, отвечающий за организацию эффективной совместной работы) осуществляет управление конфликтами, и первые им предотвращаются, а вторые — стимулируются (интенсифицируются) и разрешаются. Инструментальные конфликты подталкивают к дискуссиям и дебатам, помогающим коллективам повысить производительность за счет лучшего понимания различных точек зрения и альтернативных решений, а также помогают исключить эффект группсинка, когда принимается решение большинства, несмотря на то что есть члены коллектива, не согласные с этой точкой зрения, но умалчивающие свое мнение. Поэтому актуально исследование (в том числе с помощью визуализации процессов стимулирования) и моделирование интенсификации конфликтов в рамках ГиИМАС, которые моделируют рассуждения отдельных специалистов (агентами), а также отображают в памяти компьютера макроуровневые процессы, возникающие в результате взаимодействия специалистов при решении проблемы «за круглым столом».

В [1] предложена модель ГиИМАС с проблемно- и процессно-ориентированными конфликтами, затем в [2] описан метод идентификации конфликтов между агентами в рамках предложенной в [1] модели, а в [3] описан один из методов управления конфликтами (МУК) — стимуляция конструктивных инструментальных конфликтов (проблемно- и процессно-ориентированных), что повышает релевантность ГиИМАС работе реальных малых коллективов. На основе [1–3] в [4] был разработан подход к визуализации конфликтов, повышающий прозрачность работы ГиИМАС для пользователя. Цель настоящей работы — разработка метода визуализации интенсификации (стимуляции) конфликтов на базе предложенных методов их идентификации [2] и стимуляции [3] в рамках представленной в [1] модели ГиИМАС.

[1]Калининградский филиал Федерального исследовательского центра «Информатика и управление» Российской академии наук, sophiyabr@gmail.com

[2]Калининградский филиал Федерального исследовательского центра «Информатика и управление» Российской академии наук, baltbipiran@mail.ru

2 Походы к визуализации динамики конфликтов

В литературе превалируют работы с визуализацией динамики конфликтов в общем случае [5–8], представления которых можно свести к варианту, предложенному в [9] (рис. 1). На рис. 1 предконфликт (фазы начальная и подъема) включает:

(1) возникновение конфликтной ситуации — возникновение противоречия между субъектами (их целями, действиями, интересами и т. п.) и осознание объективной конфликтной ситуации — восприятие реальности как проблемной, в сочетании с необходимостью принять какие-либо действия для ее разрешения;

(2) попытки решить проблему неконфликтными способами — участники стремятся решить возникшую проблему, спорную ситуацию неконфликтными способами (убеждение, просьба, разъяснение и т. п.);

(3) предконфликтная ситуация восприятия угрозы своей безопасности одним из оппонентов — «пусковой механизм» конфликта, ведет к переходу развития конфликта в острый период.

Конфликтное взаимодействие (острый период) включает:

(1) инцидент — первое столкновение сторон, попытка с помощью этого решить проблему в свою пользу. Если взаимные конфликтные действия продолжаются, то это перерастает в эскалацию — резкое обострение противоречия (процесс углубления противоречий; расширение числа участников и т. д.);

(2) сбалансированное противодействие, которое характеризуется снижением интенсивности борьбы при продолжающемся противостоянии сторон;

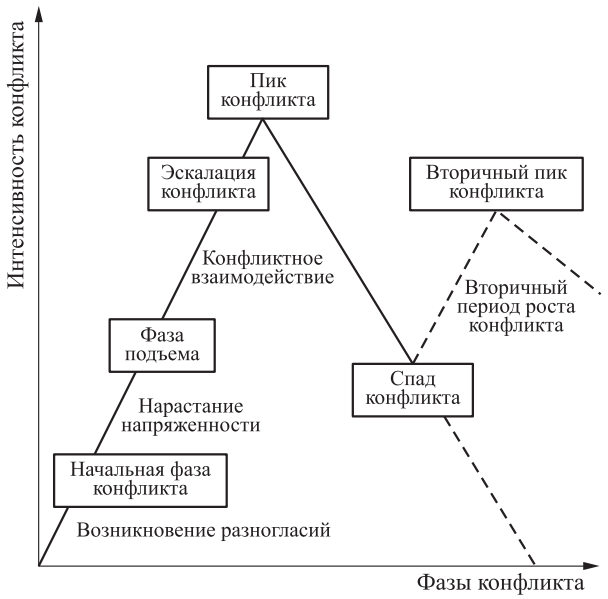


Рис. 1 Динамика конфликта

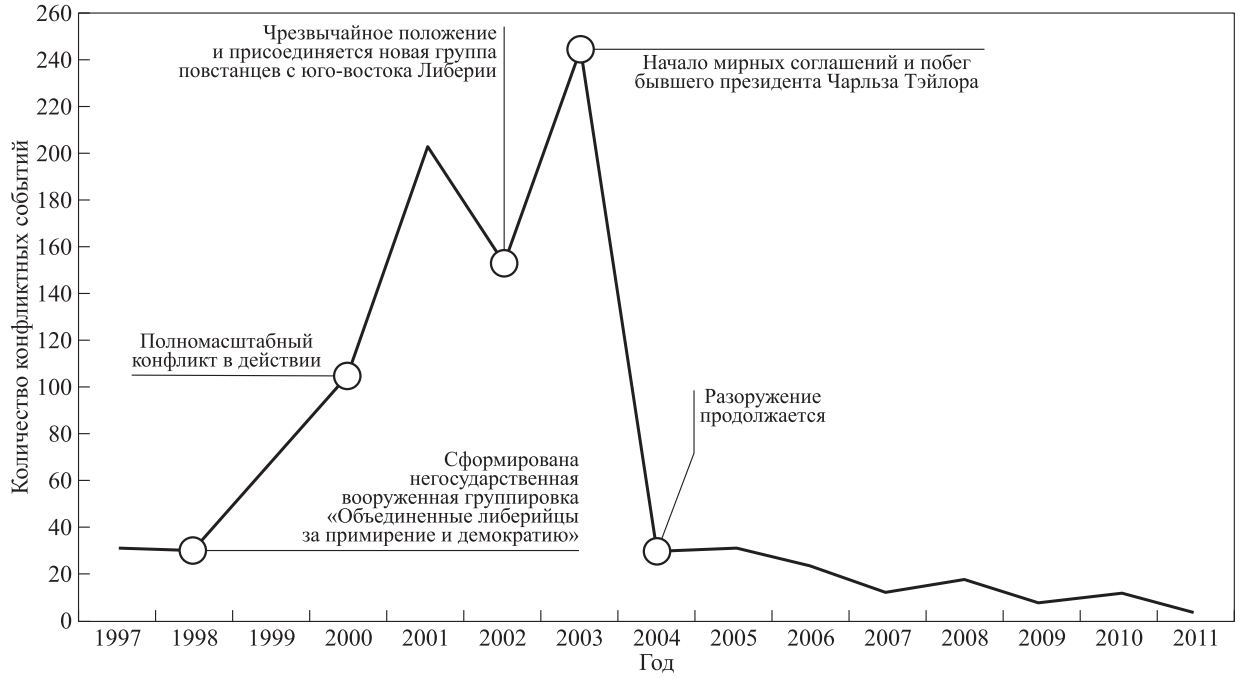


Рис. 2 Хронология конфликтных событий в Либерии (1977–2011 гг.)

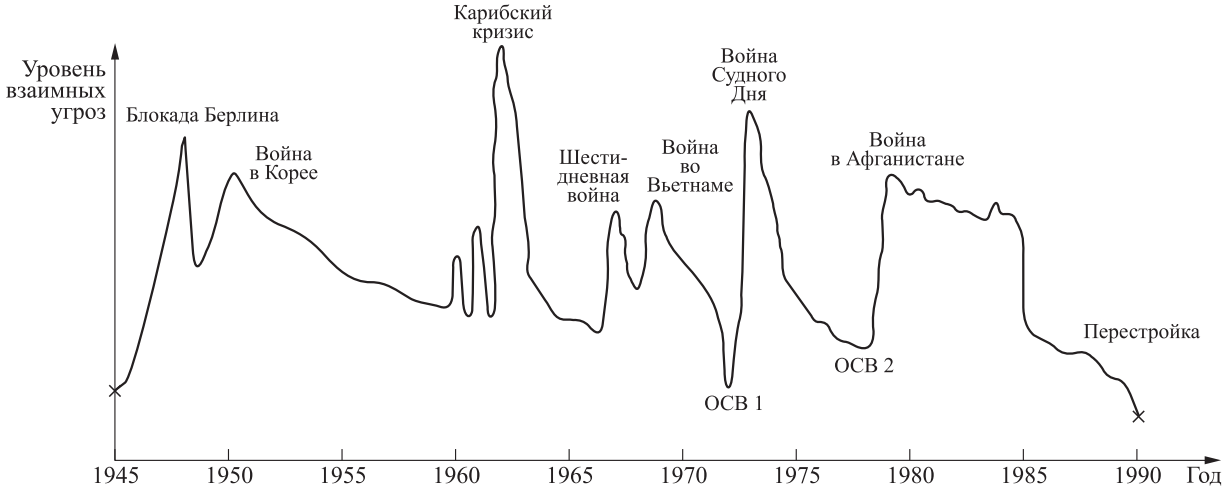


Рис. 3 Динамика макроконфликта США–СССР (1945–1990 гг.): ОСВ — переговоры об ограничении стратегических вооружений

(3) завершение конфликта — переход к поиску выхода из конфликтной ситуации (остается возможность перерастания в другой конфликт). Послеконфликтный период (спад конфликта) — частичная и/или полная нормализация отношений. Возможно вторичное зарождение конфликта.

Работ, содержащих визуализацию конкретных конфликтов, найдено мало и только для военных макроконфликтов. Например, развитие конфликта в Либерии [10] визуализировано с помощью графического представления распределения числа конфликтных событий на протяжении всех лет противостояния (рис. 2). На графике выделены переломные моменты.

В [11] история 35-летнего большого американо-советского конфликта визуализируется посредством выделения переломных моментов и построения графика зависимости числа взаимных угроз от года их возникновения (рис. 3) — выделены попытки изменить статус-кво, отражающие баланс антагонистических сил (события описываются в контексте эскалации и деэскалации).

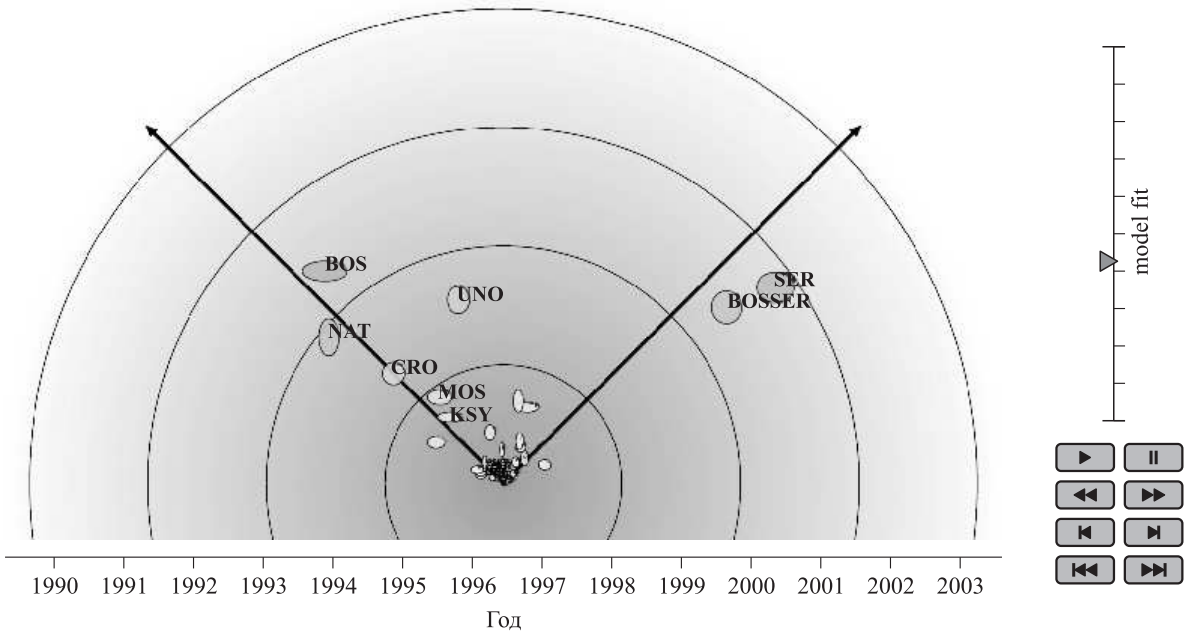


Рис. 4 Биполярная визуализация Балканского конфликта (1989–2003 гг.)

В [12] предлагается метод визуализации конфликтных сетей (рис. 4). Метод обеспечивает графический обзор конфликта и позволяет плавно анимировать динамику конфликта по годам. Акторы (доминирующие участники) — члены двух групп в зависимости от близости к левой или правой оси, а их вовлеченность пропорциональна расстоянию от источника. Форма актора кодирует соотношение между активностью (высотой) и пассивностью (шириной). Однако никакой количественной информации визуализация не содержит, и для ее получения требуются дополнительные действия.

Таким образом, найдены только работы, отображающие динамику деструктивных макроконфликтов без деталей взаимодействия участников, и ни одной работы с визуализацией динамики конфликта в малых группах специалистов, решающих проблему (в том числе стимуляции конструктивного конфликта). Ввиду этого для агентов в ГиИМАС необходима разработка нового метода, позволяющего построить более детальную визуализацию стимуляции конфликта, более явно отображающую (нежели в [12]) вовлеченность в конфликт — тип конфликта, напряженность между участниками и их изменение.

3 Метод визуализации стимуляции конфликта между агентами

В [1] было введено понятие процесса управления конфликтами

$$\mathrm{cnfm} = \langle \mathbf{CNF}, \mathrm{cnfcl}, \mathrm{cmkb}, \mathrm{act}_{\mathrm{cnfm}}, \mathrm{ACT}_{\mathrm{agcr}} \rangle, \quad (1)$$

где **CNF** — матрица, описывающая конфликт кортежем из (2); cnfcl — классификатор конфликтов агентов, идентифицирующий их характер и оценивающий напряженность [2]; cmkb — база знаний об эффективности МУК; $\mathrm{act}_{\mathrm{cnfm}}$ — функция агента-фасилитатора (АФ) «управление конфликтом»; $\mathrm{ACT}_{\mathrm{agcr}}$ — множество допустимых действий агентов по разрешению противоречий.

Конфликт между агентами (элемент матрицы **CNF** из (1)):

$$\mathrm{cnf}_{ij\,\mathrm{cnft}} = \\ = \langle \mathrm{id}_i, \mathrm{id}_j, \mathrm{cnfin}, \mathrm{cnft}, \mathrm{ACT}_{\mathrm{agcr}\,i}, \mathrm{ACT}_{\mathrm{agcr}\,j} \rangle, \quad (2)$$

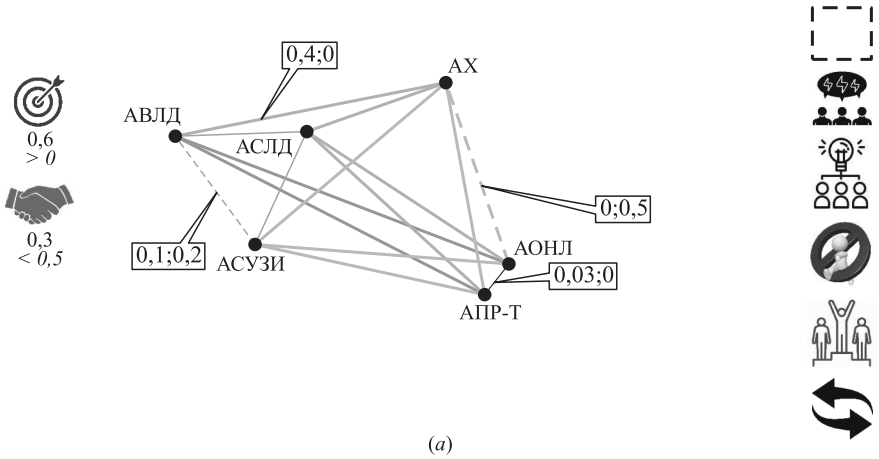
где id_i и id_j — идентификаторы агентов — субъектов конфликта; $\mathrm{cnfin} \in [0,1]$ — напряженность конфликта; cnft — «тип конфликта» (проблемно- и/или процессно-ориентированный); $\mathrm{ACT}_{\mathrm{agcr}\,i}, \mathrm{ACT}_{\mathrm{agcr}\,i} \subseteq \mathrm{ACT}_{\mathrm{agcr}}$ — множества допустимых действий агентов ag_i и ag_j по разрешению противоречий.

Функция $\mathrm{act}^{\mathrm{cnfm}}$ из (1) была задана в [3]. На базе матрицы **CNF** (1) в [4] был разработан метод визуализации конфликта (МВК). Модифицируем последовательность шагов, описывающих функцию «управления конфликтом» (МПШФУК) $\mathrm{act}^{\mathrm{cnfm}}$ АФ [3], дополняя ее запуском МВК [4], а также визуализацией среднего арифметического показателей взаимозависимости целей агентов $\mathrm{gd}^{\mathrm{himas}}$ и общего показателя напряженности конфликта в ГиИМАС $\mathrm{cnf}^{\mathrm{himas}}$ с их пороговыми значениями.

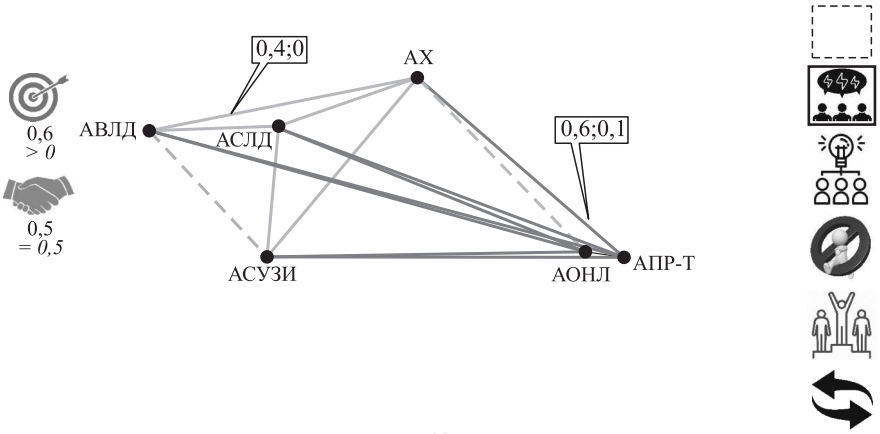
Пример результата работы МПШФУК — визуализация стимуляции конфликта между агентами — представлен на рис. 5. Толщина и цвет (от светло-серого до темно-серого) линии указывают на величину среднего квадратического напряженностей конфликтов между агентами. У каждой вершины — подписи, идентификаторы агентов. Если $\eta > 0$, то для агентов с напряженностью конфликта ниже η ребра на графе не отображаются. При наведении указателя мыши на ребро отображаются напряженности конфликтов. С учетом матриц [4] **CP** (элементы cp_{ij} описывают величину напряженности проблемно-ориентированного конфликта между агентами) и **CPR** (элементы cpr_{ij} описывают величину напряженности процессно-ориентированного конфликта), если превалирует проблемно-ориентированный конфликт ($\mathrm{cp}_{ij} \geq \mathrm{cpr}_{ij}$), то ребро между двумя вершинами (агентами) прорисовывается сплошной линией, а если процессно-ориентированный ($\mathrm{cp}_{ij} < \mathrm{cpr}_{ij}$) — штриховой.

Последовательность шагов МПШФУК следующая.

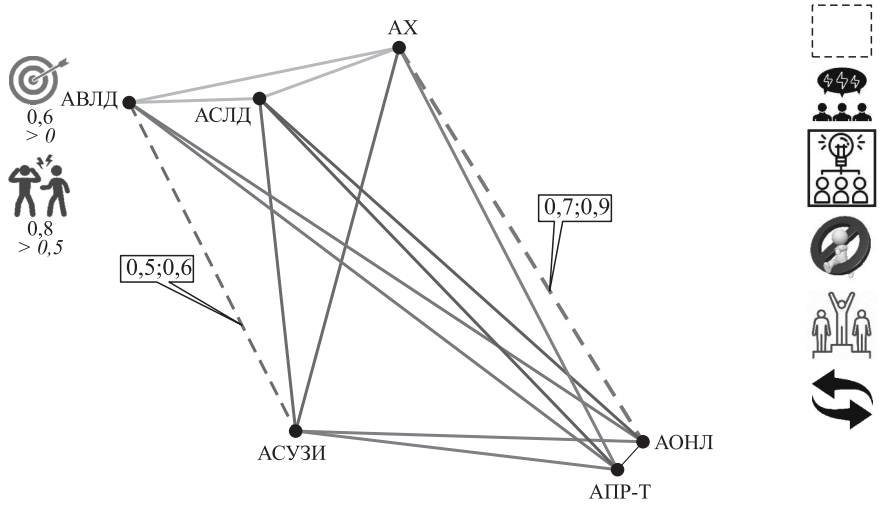
1. Инициализировать переменную «стадия»: stg = = “начало”.
2. Вычислить среднее арифметическое показателей взаимозависимости целей агентов по формуле [3]:
$$\mathrm{gd}^{\mathrm{himas}} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{2\mathrm{gd}_{ij}(n-2)!}{n!}.$$
3. Запустить функцию идентификации конфликтов $\mathrm{act}^{\mathrm{cnft}}$ из (1), чтобы с помощью классификатора конфликтов **cnfcl** из (1) на основе решений, предложенных агентами-специалистами (АС), сформировать матрицу конфликтов **CNF** из (1) между парами агентов, после чего вычислить общий показатель напряженности конфликта в ГиИМАС:



(*а*)



(*б*)



(*в*)

Рис. 5 Визуализация стимуляции конфликта между агентами: (*а*) начало, предконфликт (начальная фаза); (*б*) стимуляция, инцидент (фаза подъема); (*в*) стимуляция, эскалация (пик конфликта), конец стимуляции

$$\mathrm{cnf}^{\mathrm{himas}} = \\ = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{\mathrm{cnft} \in \mathrm{CNFT}} \frac{2\mathrm{cnf}_{ij\,\mathrm{cnft}}(n-2)!}{|\mathrm{CNFT}|n!}.$$

4. Задать пороговое η минимальное значение напряженности визуализируемого конфликта при необходимости (по умолчанию $\eta = 0$), запустить МВК и отобразить на экране $\mathrm{gd}^{\mathrm{himas}}$ и $\mathrm{cnf}^{\mathrm{himas}}$ с их порогами, используя их значения, цвет и символы (для $\mathrm{gd}^{\mathrm{himas}}$: 🎯 темно-серого цвета, если выше нуля, и 🎯 светло-серого цвета, если ниже или равно нулю; для $\mathrm{cnf}^{\mathrm{himas}}$: 🤝 темно-серого цвета, если меньше $\mathrm{cnf}_h^{\mathrm{himas}}$, 🤝 светло-серого цвета, если равно $\mathrm{cnf}_h^{\mathrm{himas}}$, и 👫 , если больше $\mathrm{cnf}_h^{\mathrm{himas}}$).
5. **Если** $\mathrm{gd}^{\mathrm{himas}} > 0$, $\mathrm{cnf}^{\mathrm{himas}} < \mathrm{cnf}_h^{\mathrm{himas}}$ и $\mathrm{stg} \in$ $\in$ {“начало”, “стимуляция”}, где $\mathrm{cnf}_h^{\mathrm{himas}}$ — определяемое в ходе тестирования системы значение, по умолчанию равное 0,5, **то** установить stg = “стимуляция”, выполнить функцию «стимуляция конфликтов» $\mathrm{act}^{\mathrm{cnfs}}$ и подсветить рамкой символ выбранного при выполнении функции $\mathrm{act}^{\mathrm{cnfs}}$ метода коллективных рассуждений АС из набора доступных АФ metcolag = {« », «мозговой штурм с наводящими вопросами», «пул мозговой записи», «оспаривание ограничений», «конкуренция за вознаграждение», «инвертирования целей АС»} (metcolag биективно отображается на smetcolag = {⬚, 💬, 👥, 🎯, 🏛, 🔄} и по умолчанию перед запуском функции «управление конфликтом» устанавливаются metcolag = « », а smetcolag = ⬚, **иначе** сохранить визуализацию стимуляции конфликта, установить stg = “разрешение” и выполнить функцию «разрешение конфликтов» $\mathrm{act}^{\mathrm{cnfr}}$.
6. Если stg = “конец” после выполнения функций «стимуляция конфликтов» $\mathrm{act}^{\mathrm{cnfs}}$ или «разрешение конфликтов» $\mathrm{act}^{\mathrm{cnfr}}$, то отправить агенту, принимающему решение (АПР), сообщение о необходимости завершения работы ГиИМАС и закончить работу.
7. Ожидать поступления сообщений от АС или АПР.
8. Если получено сообщение от АС, содержащее решение проблемы или ее части, и metcolag $\neq$ «инвертирования целей АС», то перейти к п. 3, иначе к п. 2.
9. Если получено сообщение от АПР о завершении работы, то закончить работу, иначе сообщить агенту-отправителю об ошибке и перейти к п. 7.

Таким образом, данный метод визуализации стимуляции конфликта между агентами в ГиИМАС интегрирует в себе пиктографическую, визуальную (с помощью графов) и численную информации, что в совокупности дает пользователю исчерпывающее представление о процессах стимуляции конфликта агентов.

4 Заключение

Стимуляция конструктивных конфликтов в малых коллективах специалистов приводит к более качественному решению проблемы. Моделирование и визуализация процессов стимуляции агентов делает работу ГиИМАС релевантной реальному процессу поиска разрешения проблемной ситуации коллективом. В работе рассмотрены подходы к визуализации стимуляции конфликтов (исключительно деструктивных военных макроконфликтов) и предложен новый метод ее визуализации, базирующийся на представленном в [4] методе визуализации конфликта между агентами. Предложенный метод предоставляет возможность отследить изменение напряженности между агентами при каждом запуске функции «стимуляция конфликта», а также отследить, какой метод коллективных рассуждений агентов-специалистов внес наибольший вклад в процесс стимуляции конструктивного конфликта в коллективе.

Литература

1. *Листопад С. В., Кириков И. А.* Моделирование конфликтов агентов в гибридных интеллектуальных многоагентных системах // Системы и средства информатики, 2019. Т. 29. № 3. С. 139–148. doi: 10.14357/08696527190312.
2. *Листопад С. В., Кириков И. А.* Метод идентификации конфликтов агентов в гибридных интеллектуальных многоагентных системах // Системы и средства информатики, 2020. Т. 30. № 1. С. 56–65. doi: 10.14357/08696527200105.
3. *Листопад С. В., Кириков И. А.* Стимуляция конфликтов агентов в гибридных интеллектуальных многоагентных системах // Системы и средства информатики, 2021. Т. 31. № 2. С. 47–58. doi: 10.14357/08696527210205.
4. *Румовская С. Б., Кириков И. А.* Метод визуального представления конфликтов в гибридных интеллектуальных многоагентных системах // Информатика и её применения, 2020. Т. 14. Вып. 4. С. 77–82. doi: 10.14357/19922264200411.
5. *Rummel R. J.* Understanding conflict and war. — London / New York: John Wiley and Sons, 1975. Vol. 1. 342 p.

6. *Cupach W. R., Canary D. J.* Competence in interpersonal conflict. — New York, NY, USA: McGraw-Hill, 1997. 347 p.
7. *Анцупов А. Я., Шипилов А. И.* Конфликтология. — 5-е изд. — М.: ЮНИТИ, 2013. 512 с.
8. *Гришина Н. В.* Психология конфликта. — 3-е изд. — СПб.: Питер, 2016. 576 с.
9. *Плещиц С. Г., Мармышева Л. Н., Дергаль П. П. и др.* Конфликтология. — СПб.: СПбГУЭФ, 2012. 207 с.
10. *Herbert S.* Conflict analysis: Topic guide. — Birmingham, U.K.: GSDRC, University of Birmingham, 2017. 34 p. https://gsdrc.org/wp-content/uploads/2017/05/ConflictAnalysis.pdf.
11. *Jeong H.-W.* Understanding conflict and conflict analysis. — Los Angeles, CA, USA: SAGE, 2008. 264 p. https://is.muni.cz/el/1423/podzim2015/MVZ208/um/59326109/Ho-Won_Jeong_Understanding_Conflict_and_Conflict_Analysis_2008.pdf.
12. *Brandes U., Lerner J.* Visualization of conflict networks // Building and using datasets on armed conflicts / Ed. M. Kauffmann. — NATO science for peace and security ser. E: Human and societal dynamics. — IOS Press, 2008. Vol. 36. P. 169–188.

Поступила в редакцию 23.06.2021

VISUAL REPRESENTATION METHOD FOR THE CONFLICT STIMULATION IN HYBRID INTELLIGENT MULTIAGENT SYSTEMS

S. B. Rumovskaya and I. A. Kirikov

Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation

Abstract: Consideration of complex tasks (problems) in small teams of specialists in various fields makes it possible to get a better integrated solution. In such collectives, both destructive (personal) and constructive (instrumental) conflicts inevitably arise. Modeling the work of such teams and stimulating constructive conflicts in them will improve the quality of the solution and develop a solution method that is relevant to the problem. Visualization is a powerful tool for extracting, analyzing, and understanding of information. The work is devoted to the development of the method for visualizing the processes of stimulating conflicts within hybrid intelligent multiagent systems that simulate the reasoning of individual specialists by agents and display in the computer memory macrolevel processes arising as a result of the interaction of specialists when solving a problem "at a round table."

Keywords: collective of experts; conflict; visualization of the conflict stimulation

DOI: 10.14357/19922264210310

References

1. Listopad, S. V., and I. A. Kirikov. 2019. Modelirovanie konfliktov agentov v gibridnykh intellektual'nykh mnogoagentnykh sistemakh [Modeling of agent conflicts in hybrid intelligent multiagent systems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(3): 139–148. doi: 10.14357/08696527190312.
2. Listopad, S. V., and I. A. Kirikov. 2020. Metod identifikatsii konfliktov agentov v gibridnykh intellektual'nykh mnogoagentnykh sistemakh [Agent conflict identification method in hybrid intelligent multiagent systems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 30(1):56–65. doi: 10.14357/08696527200105.
3. Listopad, S. V., and I. A. Kirikov. 2021. Stimulyatsiya konfliktov agentov v gibridnykh intellektual'nykh mnogoagentnykh sistemakh [Stimulation of agent conflicts in hybrid intelligent multiagent systems]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 31(2):47–58. doi: 10.14357/08696527210205.
4. Rumovskaya, S. B., and I. A. Kirikov. 2020. Metod vizual'nogo predstavleniya konfliktov v gibridnykh intellektual'nykh mnogoagentnykh sistemakh [Conflict visual representation method in collective decision-making within hybrid intelligent multiagent systems]. *Informatika i ee Primeneniya — Inform. Appl.* 14(4):77–82. doi: 10.14357/19922264200411.
5. Rummel, R. J. 1975. *Understanding conflict and war*. London / New York: John Wiley and Sons. Vol. 1. 342 p.
6. Cupach, W. R., and D. J. Canary. 1997. *Competence in interpersonal conflict*. New York, NY: McGraw-Hill. 347 p.
7. Antsupov, A. Ya., and A. I. Shipilov. 2013. *Konfliktologiya* [Conflictology]. 5th ed. Moscow: YuNITI. 512 p.
8. Grishina, N. V. 2016. *Psikhologiya konflikta* [The psychology of conflict]. St. Petersburg: Piter. 576 p.
9. Pleshchits, S. G., L. N. Marmysheva, P. P. Dergal', *et al.* 2012. *Konfliktologiya* [Conflictology]. St. Petersburg: SPbGUEF. 207 p.
10. Herbert, S. 2017. *Conflict analysis: Topic guide*. Birmingham: GSDRC, University of Birmingham. 34 p. Available at: https://gsdrc.org/wp-content/uploads/2017/05/ConflictAnalysis.pdf (accessed June 30, 2021).

11. Jeong, H.-W. 2008. *Understanding conflict and conflict analysis*. Los Angeles, CA: SAGE. 264 p. Available at: https://is.muni.cz/el/1423/podzim2015/MVZ208/um/59326109/Ho-Won_Jeong_Understanding_Conflict_and_Conflict_Analysis_2008.pdf (accessed June 30, 2021).

12. Brandes, U., and J. Lerner. 2008. Visualization of conflict networks. *Building and using datasets on armed conflicts*. Ed. M. Kauffmann. NATO science for peace and security ser. E: Human and societal dynamics. IOS Press. 36:169–188.

Received June 23, 2021

Contributors

Rumovskaya Sophiya B. (b. 1985) — Candidate of Science (PhD) in technology, scientist, Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation; sophiyabr@gmail.com

Kirikov Igor A. (b. 1955) — Candidate of Science (PhD) in technology, director, Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation; baltbipiran@mail.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 83–90

ФОРМЫ ПРЕДСТАВЛЕНИЯ НОВОГО ЗНАНИЯ, ИЗВЛЕЧЕННОГО ИЗ ТЕКСТОВ*

И. М. Зацман[1]

Аннотация: Рассматривается модель целенаправленного извлечения нового знания из текстов коллективом экспертов, а также формы его представления в лингвистических типологиях и словарях терминов баз медицинских знаний. Использование модели иллюстрируется двумя примерами: извлечение нового знания о значениях немецких модальных глаголов из параллельных текстов и о значениях терминов из медицинских документов. Процесс извлечения нового знания основан на лингвистическом аннотировании текстов экспертами. Основная цель аннотирования состоит в пополнении типологий рубриками (баз знаний (БЗ) — терминами), которые удовлетворяют заданному критерию новизны, согласованы в коллективе экспертов и представляют извлеченное знание. Рассматриваемая модель включает этап согласования понимания экспертами как извлеченного знания, так и форм его представления. В рассматриваемом примере базы знаний используются три формы: новый термин, измененная дефиниция уже существующего термина (без увеличения числа его значений) и расширенная дефиниция уже существующего термина (с увеличением числа его значений).

Ключевые слова: извлечение знания из текстов; модель ИТО; концепт; рубрика; термин; контекстное значение; немецкие модальные глаголы; типология; база медицинских знаний

DOI: 10.14357/19922264210311

1 Введение

В статье рассматривается модель извлечения нового знания из текстов в процессе их лингвистического аннотирования экспертами [1, 2], а также формы представления извлеченного нового знания. Применение модели и используемые в ней формы иллюстрируются двумя примерами:

(1) извлечение новых значений немецких модальных глаголов из параллельных текстов и их представление в лингвистических типологиях (проект по гранту № 20-012-00166) [3];

(2) извлечение из медицинских текстов нового знания и его представление в словаре терминов базы знаний об исследуемой болезни (проект по гранту № 21-57-53018) [4].

В проекте создания базы медицинских знаний эксперты могут выбрать одну из трех форм представления нового извлеченного знания или их сочетание:

(1) новый термин;

(2) измененная дефиниция уже существующего термина (без увеличения числа его значений);

(3) расширенная дефиниция уже существующего термина (с увеличением числа его значений).

Модель извлечения нового знания была разработана в рамках проекта по гранту РФФИ № 18-07-00192 «Метод и информационная технология для целенаправленного формирования новых лингвистических типологий» [5–8]. В этой модели первоначально использовалась одна форма представления нового знания — дефиниция *новой рубрики* лингвистической типологии.

Цель проекта состояла в создании модели целенаправленного извлечения нового знания в интересах пополнения типологии описаниями новых значений исследуемых языковых единиц и проектировании информационной технологии на ее основе.

Для демонстрации реализуемости технологии использовалась только одна форма представления нового знания, извлеченного из фрагментов параллельных текстов, — дефиниция рубрики, которая добавлялась в типологию при обнаружении в текстах нового значения исследуемой языковой единицы [9, 10]. При этом объем смыслового содержания концепта добавляемой рубрики должен был быть не меньше объема обнаруженного в текстах нового

*Исследование выполнено при финансовой поддержке РФФИ (проект 20-012-00166) с использованием ЦКП «Информатика» ФИЦ ИУ РАН, а также при финансовой поддержке РФФИ и Государственного фонда естественных наук Китая в рамках проекта 21-57-53018.

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, izatsman@yandex.ru

значения. В следующем разделе будет дано описание различия между концептами рубрик типологии и значениями языковой единицы (которые будем называть *контекстными значениями*), обнаруженными экспертами в текстах.

Отметим, что разработанная модель позволяет не ограничиваться только одной формой его представления, а использовать несколько форм. При этом рассматриваемый спектр форм не является неотъемлемой компонентой только этой модели. Эти же формы можно использовать и в традиционных моделях генерации нового знания, известных в экономической науке [11–17].

Разработанная модель дает возможность не только использовать несколько форм представления нового знания, но и перераспределять его между рубриками типологии или, соответственно, уже существующими терминами словаря базы знаний, меняя их дефиниции. Необходимость в перераспределении была выявлена в рамках проекта по гранту РФФИ № 20-012-00166 «Исследование семантики модальных конструкций по данным немецко-русского параллельного корпуса», который выполняется в Институте проблем информатики ФИЦ ИУ РАН.

Цель этого проекта состоит в поиске новых контекстных значений немецких модальных глаголов, добавлении соответствующих рубрик в их типологию, а также в уточнении дефиниций уже существующих рубрик. При обнаружении новых контекстных значений и была выявлена необходимость не только создавать новые рубрики типологии, но и перераспределять новое содержание между уже существующими рубриками, меняя их дефиниции [18, 19].

Цель статьи состоит в описании форм представления нового знания, извлеченного из текстов, с помощью дефиниций рубрик типологий (классификационных схем) и терминов словарей баз знаний. Использование нескольких форм представления нового знания рассматривается в контексте разработанной ранее модели целенаправленного извлечения нового знания [5–10].

Раздел 2 посвящен определению понятий «концепт» рубрики (термина) и «контекстное значение», а также описанию различий между ними на примере результатов проекта по исследованию немецких модальных глаголов.

В разд. 3 дается описание варианта ранее разработанной модели с тремя формами представления знания, адаптированной для медицинского проекта [4].

2 Контекстные значения языковых единиц и концепты

При описании форм представления нового знания используются два ключевых понятия: «концепт» рубрики типологии (классификационной схемы) или словарного термина и «контекстное значение» языковой единицы. Начнем с определения понятия «контекст». Согласно энциклопедическому словарю, контекст — это «фрагмент *текста*, включающий избранную для анализа единицу, необходимый и достаточный для определения значения этой единицы, являющегося непротиворечивым по отношению к общему смыслу текста» [20][1]. На основе этой словарной дефиниции определим понятие «*контекстное значение*» языковой единицы (например, модальный глагол или медицинский термин) как ее смысловое содержание, определенное в границах контекста и являющееся непротиворечивым по отношению к общему смыслу текста. Понятие «*концепт*» рубрики типологии (классификационной схемы) или словарного термина определяется как смысловое содержание, извлекаемое из ее (его) дефиниции, определенное в ее границах и являющееся: (1) непротиворечивым по отношению к общему смыслу типологии (словаря) и (2) охватывающим те контекстные значения, которым эта рубрика присваивается.

Понятия «концепт», «контекстное значение» и различие между ними проиллюстрируем на примере результатов, полученных в рамках проекта по аннотированию предложений с немецкими модальными глаголами [8–10], в котором используется их типология со структурой, представленной на рис. 1.

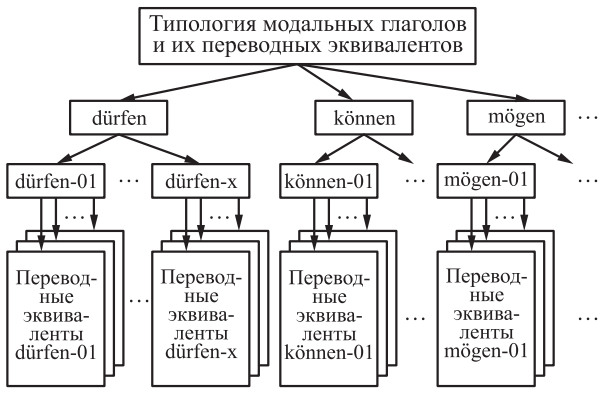


Рис. 1 Структура типологии значений немецких модальных глаголов и их переводных эквивалентов

[1]Возможны ситуации, когда разные эксперты могут увидеть и описать разные значения языковой единицы в одном и том же контексте (см. рис. 2).

Таблица 1 Рубрики значений трех немецких модальных глаголов

Рубрики значений глагола dürfen	Число немецких предложений	Рубрики значений глагола können	Число немецких предложений	Рубрики значений глагола mögen	Число немецких предложений
dürfen-01	304	können-01	2244	mögen-01	157
dürfen-02	40	können-02	138	mögen-02	4
dürfen-03	18	können-03	103	mögen-03	33
dürfen-04	18	können-04	11	mögen-04	21
dürfen-x	3	können-05	13	mögen-05	35
		können-06	88	mögen-06	74
		können-07	79	mögen-07	13
		können-08	18	mögen-x	0
		können-09	2		
		können-x	3		
Число проаннотированных предложений	383		2699		337

На рис. 1 показаны три уровня типологии. Верхний (постоянный) уровень включает немецкие модальные глаголы dürfen, können, mögen, müssen, sollen и wollen (показаны первые три глагола). Средний уровень содержит рубрики значений немецких модальных глаголов (показаны две рубрики из пяти для dürfen и по одной для können и mögen). В типологии у каждого глагола есть специальная рубрика с неопределенным концептом и без дефиниции, имеющая символ «x» на конце. Она проставляется в тех случаях, когда контекстное значение глагола в аннотируемом предложении, по мнению экспертов, *не принадлежит ни одному концепту* рубрик текущего варианта типологии.

Рубрики трех модальных глаголов и число проаннотированных немецких предложений с каждой из них, включая специальные рубрики с символом «x», указаны в табл. 1[1]. Например, рубрике dürfen-01[2] соответствует 304 немецких предложения, проаннотированных экспертами с применением надкорпусной базы данных [2, 3], а рубрике dürfen-x — 3 немецких предложения. Поэтому рубрика dürfen-01 и была проставлена в аннотациях[3] 304 немецких предложений, а рубрика dürfen-x — в аннотациях 3 из 383 немецких предложений с dürfen.

Число рубрик в типологии, соответствующих каждому глаголу, может увеличиваться в процессе семантического анализа и аннотирования предложений параллельных текстов. Новые рубрики добавляются в типологию, если эксперты находят в предложениях новые контекстные значения глаголов, которые не представлены рубриками текущей версии типологии. До очередного пополнения типологии таким предложениям временно присваиваются рубрики с «x». Например, до начала проекта по модальным глаголам в типологии было только две рубрики, охватывающих контекстные значения глагола dürfen в проаннотированных предложениях (см. рисунок на с. 95 в [7]), а по состоянию на 20 июня 2021 г. их число увеличилось до четырех, не считая рубрики с символом «x» (см. табл. 1).

Нижний уровень типологии включает для каждой рубрики все те переводные эквиваленты на русском языке, которые встретились экспертам в процессе аннотирования. Переводные эквиваленты используются экспертами при описании *перевода* каждого контекстного значения модального глагола в процессе аннотирования. Восемь эквивалентов, встретившихся экспертам 8 и более раз в переводах на русский язык 304 немецких предложений с глаголом dürfen в значении dürfen-01, приведены в табл. 2 (по состоянию на 20.06.21).

Всего эксперты описали в процессе аннотирования 72 эквивалента для значения dürfen-01. Для 64 из 72 эквивалентов (с частотностью менее 8) в табл. 2 отведена одна строка. Сорок три из 64 эквивалентов встретились только один раз (в табл. 1 не показаны). В 46 из 304 эквивалентов модальность не передана. Таблица 3 содержит три примера предложений

[1] Все численные данные в таблицах, полученные с использованием ЦКП «Информатика» ФИЦ ИУ РАН, указаны по состоянию надкорпусной базы данных немецких модальных глаголов [3] на 20 июня 2021 г.

[2] Дефиниция концепта этой рубрики по состоянию на 20 июня 2021 г. определена экспертами как «*Наличие разрешения делать что-л. (часто в неполных синтаксических конструкциях): мочь; можно; под отрицанием: нельзя, не разрешается, запрещено*».

[3] Примеры аннотаций немецких предложений с модальным глаголом sollen приведены в табл. 2 и 3 работы [5] с указанием рубрик этого глагола (описание концептов рубрик дано в [21]).

Таблица 2 Русские переводные эквиваленты, соответствующие значению dürfen-01

	Переводной эквивалент	Частотность эквивалентов
	[*эквиваленты, не передающие модальность в переводе*]	46
1	мочь	34
2	должен	26
3	нельзя	25
4	можно (*см. первый пример в табл. 3*)	18
5	иметь право	14
6	разрешить	9
7	модальность передана в переводе формой повелительного наклонения (*см. второй пример в табл. 3*)	8
8	вправе (*см. третий пример в табл. 3*)	8
	[*остальные 64 переводных эквивалента*]	116
	Число немецких предложений с dürfen в значении dürfen-01	304

Таблица 3 Три предложения с глаголом dürfen в первом значении и их переводы

Оригинальный текст	Перевод
Darf ich dir einen Kuss geben? [Thomas Mann. Buddenbrooks (1896–1900)]	Можно, я тебя поцелую? [Перев. — Н. Ман (1953)]
Du darfst nicht wieder fort, darfst mich nicht wieder allein lassen. [Theodor Fontane. Effi Briest (1894–1895)]	Не уезжай больше, не оставляй меня одну! [Перев. — Ю. Светланов, Г. Эгерман (1960)]
Das Volk, hieß es in der Urteilsbegründung, müsse von den finanziell und gesellschaftlich bessergestellten Kreisen einen sittlich einwandfreien Lebenswandel nicht nur fordern dürfen, sondern auch vorgelebt sehen können. [Friedrich Dürrenmatt. Justiz (1985)]	Народ, как гласило обоснование приговора, вправе не только требовать, чтобы круги, благополучные с финансовой точки зрения и вышестоящие — с общественной, вели нравственно безупречный образ жизни, но и вправе своими глазами наблюдать его. [Перев. — С. Фридлянд (1988)]

с dürfen в первом значении dürfen-01 (с переводными эквивалентами «можно», императивная конструкция и «вправе»).

В проекте по исследованию значений немецких модальных глаголов используются две формы представления нового знания, извлеченного экспертами из параллельных текстов:

(1) измененные дефиниции уже существующих рубрик глагола (без увеличения их числа);

(2) новые рубрики и их дефиниции, которые добавляются в типологию на ее среднем уровне. Операции изменения дефиниций в надкорпусной базе данных описаны в [19].

Форма, аналогичная первой из трех перечисленных в разд. 1, в этом проекте не используется, так как эксперты до начала исследования фиксируют число модальных глаголов, исследуемых ими в немецком языке в рамках проекта. Эта форма (добавляемый новый медицинский термин) вместе с двумя другими формами будет описана в следующем разделе.

3 Модель с тремя формами представления нового знания

Разработанная ранее модель [5–10] получила название «information-technology-oriented — ITO» (модель ИТО) [22]. Ее адаптированный вариант используется в настоящее время при проектировании информационной технологии извлечения новых контекстных значений из текстов документов, относящихся к исследуемой болезни, и пополнения словаря БЗ о ней терминами, представляющими эти значения [4]. В рамках этой технологии, сочетающей программно реализованные и экспертные стадии семантического анализа текстов, извлеченные значения вместе с их контекстами распределяются по трем категориям (см. стрелки 1, 2 и 3 на рис. 2):

(1) контекст(ы) извлеченного значения указывает (указывают) на необходимость *обновления описания* одного из концептов в дефиниции термина, уже представленного в словаре БЗ;

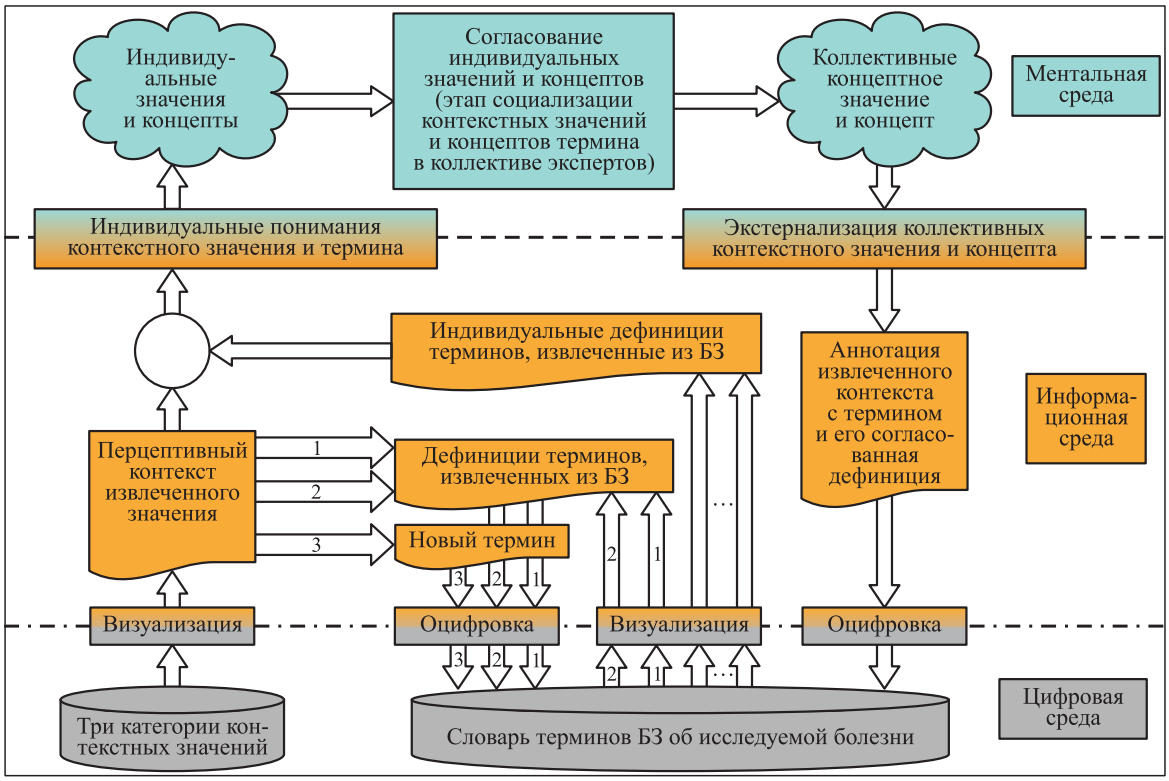


Рис. 2 Модель ИТО [22], адаптированная для обработки медицинских документов

(2) контекст(ы) говорит (говорят) о необходимости включения *описания нового концепта* в дефиницию термина, уже существующего в словаре БЗ;

(3) извлеченное значение и его контексты указывают на необходимость *формирования дефиниции нового термина* и добавления его в словарь БЗ.

В БЗ дефиниция одного термина может включать описание нескольких его концептов, т. е. термины могут быть многозначными (для модальных глаголов каждому концепту рубрики соответствовала своя дефиниция). Сопоставительный анализ структурных компонентов дескрипторов терминов проектируемой БЗ и терминов тезауруса Национального института рака (National Cancer Institute — NCI) [23, 24], а также описание информационной технологии извлечения новых контекстных значений из текстов документов даны в [4, 22].

Модель ИТО, описанная в работе [22], была адаптирована (см. рис. 2) в медицинском проекте в интересах разработки финальной из пяти стадий технологии извлечения и представления новых контекстных значений в БЗ:

(1) формирование массива документов, описывающих исследуемое заболевание, извлекаемых из баз данных, например PubMed Central [25];

(2) автоматическое обнаружение в этом массиве слов (словосочетаний) и формирование массива их контекстов, которые могут говорить о *потенциально новых контекстных значениях* (согласно заданному критерию их новизны [22]);

(3) автоматизированное распределение потенциально новых контекстных значений по трем перечисленным выше категориям;

(4) анализ каждой категории и уточнение ее состава коллективом экспертов;

(5) изменение дефиниций уже существующих терминов (стрелки 1 и 2 на рис. 2) или пополнение словаря БЗ новыми терминами (стрелки 3) экспертами, которые согласовывают свое понимание извлеченных контекстных значений и концептов.

На первых двух стадиях решается также задача сокращения массива контекстов за счет уменьшения поискового шума (на первой стадии при формировании массива документов, на второй — массива контекстов), а на третьей — задача уменьшения числа ошибок распределения методом машинного обучения. От степени решения этих двух задач зависит объем последующей работы экспертов.

4 Заключение

Иногда бывает необходимо аннотировать нетекстовые данные в процессе терминологического описания в БЗ некоторых заболеваний. В частности, аннотации случаев инфицирования коронавирусами могут включать рубрики, представляющие смысл нетекстовых данных. Такими данными могут быть изображения, отражающие симптомы заболевания (например, участки уплотнений в легких по типу «матового стекла»), а также биоинформационные последовательности геномов штаммов коронавирусов. Для таких случаев необходимо будет обобщить процедуру лингвистического аннотирования и применять ее обобщенный вариант при аннотировании сочетания текстовых и нетекстовых данных. Создание обобщенной процедуры может стать основой извлечения нового знания из сочетаний его разнородных потенциальных источников (текстов, изображений, биоинформационных последовательностей).

Литература

1. Handbook of linguistic annotation / Eds. N. Ide, J. Pustejovsky. — Dordrecht, The Netherlands: Springer Science + Business Media, 2017. 1468 p.
2. *Гончаров А. А., Инькова О. Ю., Кружков М. Г.* Методология аннотирования в надкорпусных базах данных // Системы и средства информатики, 2019. Т. 29. № 2. С. 148–160.
3. *Добровольский Д. О., Зализняк Анна А.* Немецкие конструкции с модальными глаголами и их русские соответствия: проект надкорпусной базы данных // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам Междунар. конф. «Диалог». — М.: РГГУ, 2018. С. 172–184.
4. *Зацман И. М.* Проблемно-ориентированная актуализация словарных статей двуязычных словарей и медицинской терминологии: сопоставительный анализ // Информатика и её применения, 2021. Т. 15. Вып. 1. С. 94–101.
5. *Зацман И. М.* Стадии целенаправленного извлечения знаний, имплицированных в параллельных текстах // Системы и средства информатики, 2018. Т. 28. № 3. С. 175–188.
6. *Zatsman I.* Finding and filling lacunas in knowledge systems // 20th European Conference on Knowledge Management Proceedings. — Reading: Academic Publishing International Ltd., 2019. Vol. 2. P. 1143–1151.
7. *Зацман И. М.* Целенаправленное развитие систем лингвистических знаний: выявление и заполнение лакун // Информатика и её применения, 2019. Т. 13. Вып. 1. С. 91–98.
8. *Зацман И. М.* Проблемно-ориентированная верификация полноты темпоральных онтологий и заполнение понятийных лакун // Информатика и её применения, 2020. Т. 14. Вып. 3. С. 119–128.
9. *Zatsman I.* Finding and filling lacunas in linguistic typologies // 15th Forum (International) on Knowledge Asset Dynamics Proceedings. — Matera: Institute of Knowledge Asset Management, 2020. P. 780–793.
10. *Zatsman I.* Three-dimensional encoding of emerging meanings in AI-systems // 21st European Conference on Knowledge Management Proceedings. — Reading: Academic Publishing International Ltd., 2020. P. 878–887.
11. *Nonaka I.* The knowledge-creating company // Harvard Bus. Rev., 1991. Vol. 69. No. 6. P. 96–104.
12. *Nonaka I.* A dynamic theory of organizational knowledge creation // Organ. Sci., 1994. Vol. 5. No. 1. P. 14–37.
13. *Нонака И., Такеучи Х.* Компания — создатель знания / Пер. с англ. — М.: Олимп-бизнес, 2003. 384 с. (*Nonaka I., Takeuchi H.* The knowledge-creating company. — Oxford, NY, USA: Oxford University Press, 1995. 284 p.).
14. *Wierzbicki A. P., Nakamori Y.* Basic dimensions of creative space // Creative space: Models of creative processes for knowledge civilization age / Eds. A. P. Wierzbicki, Y. Nakamori. — Berlin: Springer Verlag, 2006. P. 59–90.
15. *Nissen M. E.* Harnessing knowledge dynamics: Principled organizational knowing & learning. — London: IRM Press, 2006. 278 p.
16. *Wierzbicki A. P., Nakamori Y.* Knowledge sciences: Some new developments // Z. Betriebswirt., 2007. Vol. 77. No. 3. P. 271–295.
17. *Nakamori Y.* Knowledge and systems science — enabling systemic knowledge synthesis. — London – New York: CRC Press, 2013. 234 p.
18. *Гончаров А. А., Зацман И. М., Кружков М. Г.* Эволюция классификаций в надкорпусных базах данных // Информатика и её применения, 2020. Т. 14. Вып. 4. С. 108–116.
19. *Гончаров А. А., Зацман И. М., Кружков М. Г.* Представление новых лексикографических знаний в динамических классификационных системах // Информатика и её применения, 2021. Т. 15. Вып. 1. С. 86–93.
20. Лингвистический энциклопедический словарь / Под ред. В. Н. Ярцевой. — М.: Советская энциклопедия, 1990. 685 с.
21. *Добровольский Д. О., Зализняк Анна А.* О семантике немецкого глагола sollen // ВАПросы языкознания. — М.: Буки-Веди, 2020. С. 459–464.
22. *Zatsman I., Khakimova A.* New knowledge discovery for creating terminological profiles of diseases // 22nd European Conference on Knowledge Management Proceedings. — Reading: Academic Publishing International Ltd., 2021. P. 837–846.
23. Overview of National Cancer Institute Thesaurus (NCIt). https://wiki.nci.nih.gov/pages/viewpage.action?pageId=7472532.
24. National Cancer Institute Thesaurus. https://ncit.nci.nih.gov/ncitbrowser.
25. PubMed Central Overview. https://www.ncbi.nlm.nih.gov/pmc/about/intro.

Поступила в редакцию 14.07.2021

FORMS REPRESENTING NEW KNOWLEDGE DISCOVERED IN TEXTS

I. M. Zatsman

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

Abstract: The model of goal-oriented discovery of new knowledge by a team of experts in texts and the forms of its representation in linguistic typologies and term dictionaries of medical knowledge bases are considered. The usage of the model is illustrated by two examples: the discovery of new knowledge about the meanings of German modal verbs in parallel texts and about the meanings of terms in medical documents. The process of discovering new knowledge is based on linguistic annotation of texts performed by experts. The main goal of annotation is to enrich typologies with headings or, accordingly, augment knowledge bases with terms that meet the specified criterion of novelty, are agreed on within the team of experts, and represent the discovered knowledge. The model includes the coordination of experts' understanding of both the discovered knowledge and the forms of its representation. In the example of the term dictionary of a medical knowledge base, three forms are used: a new term, a changed definition of an existing term without increasing the number of its meanings, and an enlarged definition of an existing term with its new meanings.

Keywords: discovery of knowledge in texts; ITO model; concept; heading; term; contextual meaning; German modal verbs; typology; medical knowledge base

DOI: 10.14357/19922264210311

Acknowledgments

The reported study was funded by RFBR, project number 20-012-00166, and by RFBR and NSFC, project number 21-57-53018. The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow).

References

1. Ide, N., and J. Pustejovsky, eds. 2017. *Handbook of linguistic annotation*. Dordrecht, The Netherlands: Springer Science + Business Media. 1468 p.
2. Goncharov, A. A., O. Yu. Inkova, and M. G. Kruzhkov. 2019. Metodologiya annotirovaniya v nadkorpusnykh bazakh dannykh [Annotation methodology of supracorpora databases]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(2):148–160.
3. Dobrovol'skij, D. O., and Anna A. Zalizniak. 2018. Nemetskie konstruktsii s modal'nymi glagolami i ikh russkie sootvetstviya: proekt nadkorpusnoy bazy dannykh [German constructions with modal verbs and their Russian correlates: A supracorpora database project]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po mat-lam Mezhdunar. konf. "Dialog"* [Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference (International) "Dialogue"]. Moscow: RSHI. 17(24):172–184.
4. Zatsman, I. 2021. Problemno-orientirovannaya aktualizatsiya slovarnykh statey dvuyazychnykh slovarey i meditsinskoy terminologii: sopostavitel'nyy analiz [Problem-oriented updating of dictionary entries of bilingual dictionaries and medical terminology: Comparative analysis]. *Informatika i ee Primeneniya — Inform. Appl.* 15(1):94–101.
5. Zatsman, I. 2018. Stadii tselenapravlennogo izvlecheniya znaniy, implitsirovannykh v parallel'nykh tekstakh [Stages of goal-oriented discovery of knowledge implied in parallel texts]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(3):175–188.
6. Zatsman, I. 2019. Finding and filling lacunas in knowledge systems. *20th European Conference on Knowledge Management Proceedings*. Reading: Academic Publishing International Ltd. 2:1143–1151.
7. Zatsman, I. 2019. Tselenapravlennoe razvitie sistem lingvisticheskikh znaniy: vyyavlenie i zapolnenie lakun [Goal-oriented development of linguistic knowledge systems: Identifying and filling of lacunae]. *Informatika i ee Primeneniya — Inform. Appl.* 13(1):91–98.
8. Zatsman, I. 2020. Problemno-orientirovannaya verifikatsiya polnoty temporal'nykh ontologiy i zapolnenie ponyatiynykh lakun [Problem-oriented verifying the completeness of temporal ontologies and filling conceptual lacunas]. *Informatika i ee Primeneniya — Inform. Appl.* 14(3):119–128.
9. Zatsman, I. 2020. Finding and filling lacunas in linguistic typologies. *15th Forum (International) on Knowledge Asset Dynamics Proceedings*. Matera: Institute of Knowledge Asset Management. 780–793.

10. Zatsman, I. 2020. Three-dimensional encoding of emerging meanings in AI-systems. *21st European Conference on Knowledge Management Proceedings*. Reading: Academic Publishing International Ltd. 878–887.
11. Nonaka, I. 1991. The knowledge-creating company. *Harvard Bus. Rev.* 69(6):96–104.
12. Nonaka, I. 1994. A dynamic theory of organizational knowledge creation. *Organ. Sci.* 5(1):14–37.
13. Nonaka, I., and H. Takeuchi. 1995. *The knowledge-creating company*. Oxford, NY: Oxford University Press. 284 p.
14. Wierzbicki, A. P., and Y. Nakamori. 2006. Basic dimensions of creative space. *Creative space: Models of creative processes for knowledge civilization age*. Eds. A. P. Wierzbicki and Y. Nakamori. Berlin: Springer Verlag. 59–90.
15. Nissen, M. E. 2006. *Harnessing knowledge dynamics: Principled organizational knowing & learning*. London: IRM Press. 278 p.
16. Wierzbicki, A. P., and Y. Nakamori. 2007. Knowledge sciences: Some new developments. *Z. Betriebswirt.* 77(3):271–295.
17. Nakamori, Y. 2013. *Knowledge and systems science — enabling systemic knowledge synthesis*. London – New York: CRC Press. 234 p.
18. Goncharov, A. A., I. M. Zatsman, and M. G. Kruzhkov. 2020. Evolyutsiya klassifikatsiy v nadkorpusnykh bazakh dannykh [Evolution of classifications in supracorpora databases]. *Informatika i ee Primeneniya — Inform. Appl.* 14(4):108–116.
19. Goncharov, A. A., I. M. Zatsman, and M. G. Kruzhkov. 2021. Predstavlenie novykh leksikograficheskikh znaniy v dinamicheskikh klassifikatsionnykh sistemakh [Representation of new lexicographical knowledge in dynamic classification systems]. *Informatika i ee Primeneniya — Inform. Appl.* 15(1):86–93.
20. Yartseva, V. N., ed. 1990. *Lingvisticheskiy entsiklopedicheskiy slovar'* [Linguistic encyclopedic dictionary]. Moscow: Soviet Encyclopedia. 685 p.
21. Dobrovol'skij, D. O., and Anna A. Zalizniak. 2020. O semantike nemetskogo glagola sollen [About the semantics of the German verb sollen]. *VAProsy yazykoznaniya* [Topics in the study of language]. Moscow: Buki-Vedi. 459–464.
22. Zatsman, I., and A. Khakimova. 2021. New knowledge discovery for creating terminological profiles of diseases. *22nd European Conference on Knowledge Management Proceedings*. Reading: Academic Publishing International Ltd. 837–846.
23. Overview of NCI Thesaurus. Available at: https://wiki.nci.nih.gov/pages/viewpage.action?pageId=7472532 (accessed July 20, 2021).
24. NCI Thesaurus. Available at: https://ncit.nci.nih.gov/ncitbrowser (accessed July 20, 2021).
25. National library of medicine. Available at: https://pubmed.ncbi.nlm.nih.gov (accessed July 20, 2021).

Received July 14, 2021

Contributor

Zatsman Igor M. (b. 1952) — Doctor of Science in technology, Head of Department, 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; izatsman@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 91–97

СИСТЕМА МАССОВОГО ОБСЛУЖИВАНИЯ С УПРАВЛЯЕМЫМ ПО СИГНАЛАМ ПЕРЕРАСПРЕДЕЛЕНИЕМ ПРИБОРОВ ДЛЯ АНАЛИЗА НАРЕЗКИ РЕСУРСОВ СЕТИ 5G*

И. А. Кочеткова[1], А. С. Власкина[2], Н. Н. Ву[3], В. С. Шоргин[4]

Аннотация: Нарезка радиоресурсов — одна из ключевых технологий беспроводных сетей пятого поколения с постоянно растущим числом пользователей и предоставляемых услуг. Особенность данной технологии заключается в возможности организации логически изолированных сегментов радиоресурсов под конкретные требования оператора и его пользователей, причем перераспределение объемов ресурса между сегментами возможно, но только в моменты обращения к ним контроллера. В работе построена математическая модель для двух классов нетерпеливого эластичного трафика с минимальным порогом скорости в виде системы массового обслуживания с очередями и сигналами, управляющими перераспределением ресурса. Приведены формулы для расчета параметров эффективности перераспределения ресурса — коэффициента соответствия начальному распределению ресурса, коэффициента успеха перераспределения ресурса и коэффициента использования ресурса.

Ключевые слова: 5G; нарезка ресурсов; система массового обслуживания с сигналами; перераспределение ресурса; эластичный трафик

DOI: 10.14357/19922264210312

1 Введение

В настоящее время наблюдается постоянный рост числа устройств и пользователей различных услуг мобильной сети. Стремясь удовлетворить потребности пользователей по качеству обслуживания, операторы мобильной связи разрабатывают новые принципы построения сети.

Основу современных сетей пятого поколения 5G составляет концепция нарезки виртуальных сетевых ресурсов (*англ.* network slicing). При этом подразумевается логическое разделение радиоресурсов на сегменты в зависимости от установленных требований [1]. Внедрение такой технологии привело к задаче проектирования модели сети, учитывающей различные характеристики предоставляемых услуг.

Для мультисервисных беспроводных сетей, реализующих технологию нарезки радиоресурсов, авторами [2] предлагается одна из возможных моделей схемы распределения ресурсов в виде системы массового обслуживания с очередями и повторными вызовами.

Политика обслуживания пользователей может быть сформулирована как комбинация задачи выпуклого программирования для распределения ресурсов и управления доступом по приоритету [3]. Пока число пользователей в каждом сегменте остается в пределах установленного порога, все пользователи обрабатываются одинаково. Но если число пользователей в каком-то из сегментов превысит пороговое значение, пользователям сегмента сначала будет выделен меньший объем ресурсов, а затем, если пропускной способности окажется недостаточно, будет присвоен более низкий приоритет.

Более подробно алгоритмы оптимизации сегментирования на основе различных стратегий разделения ресурсов описаны в работах [4, 5]. При этом оценка верхней и нижней границ выделенного диапазона радиоресурсов [6], необходимого для поддержания требуемых гарантий по скорости обслуживания сегмента, позволяет избежать простоя радиоресурсов и перегрузки отдельных сегментов.

*Исследование выполнено при финансовой поддержке РФФИ (проект 20-37-70079). Публикация создана при поддержке Программы стратегического академического лидерства РУДН.

[1]Российский университет дружбы народов; Федеральный исследовательский центр «Информатика и управление» Российской академии наук, gudkova-ia@rudn.ru

[2]Российский университет дружбы народов, vlaskina-as@rudn.ru

[3]Российский университет дружбы народов, 1032168360@rudn.ru

[4]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, vshorgin@ipiran.ru

Исследование алгоритмов адаптивного управления ресурсами, используемых для управления качеством видеопотока с камер видеонаблюдения на удаленные серверы через действующую беспроводную сеть с видео- и веб-трафиком, представлено в работе [7].

Данная работа служит продолжением исследований [8–10], где были разработаны математические и имитационные модели систем массового обслуживания с эластичным трафиком и конечной очередью. Предложены три показателя, которые можно использовать для оценки эффективности динамического перераспределения радиоресурсов. Будем исследовать систему с динамическим перераспределением радиоресурсов, управляемую по внешнему сигналу.

2 Нарезка ресурсов

Рассмотрим задачу оценки эффективности перераспределения ресурса между классами трафика исходя из трех критериев:

(1) насколько сильно происходит отклонение от начального распределения ресурса;

(2) насколько часты случаи, когда при поступлении сигнала перераспределения ресурса не происходит;

(3) насколько много простаивает свободного ресурса при ожидающих в это время сессиях трафика.

В системе K классов трафика, между которыми динамически перераспределяется ресурс объема V. Обозначим $V_1(t_0), \ldots, V_K(t_0)$, $\sum_{k=1}^{K} V_k(t_0) = V$, начальное распределение ресурса, а $V_1(t), \ldots, V_K(t)$, $\sum_{k=1}^{K} V_k(t) = V$, распределение в некоторый момент времени t. Пусть $t_1, t_2, \ldots, t_i, \ldots$ — моменты времени поступления сигналов, управляющих перераспределением ресурса между классами трафика. Тогда параметрами эффективности перераспределения ресурса, соответствующими критериям 1–3, будут коэффициент α соответствия начальному распределению ресурса, коэффициент β успеха перераспределения ресурса и коэф-

Таблица 1 События в системе

№ п/п	Условие	Результат
1. Поступление запроса на передачу эластичного трафика k-класса		
а	k-ресурса для обеспечения минимального порога скорости передачи эластичного трафика с учетом новой сессии достаточно	Запрос будет принят, и передача трафика будет начата на k-ресурсе, а скорость передачи всех сессий на нем будет пропорционально снижена
б	k-ресурса для обеспечения минимального порога скорости передачи эластичного трафика с учетом новой сессии недостаточно, но в k-очереди есть место для ожидания начала обслуживания	Запрос будет принят и отправлен на ожидание начала обслуживания в k-очередь
в	k-ресурса для обеспечения минимального порога скорости передачи эластичного трафика с учетом новой сессии недостаточно и в k-очереди места для ожидания начала обслуживания отсутствуют	Запрос будет заблокирован
2. Завершение передачи эластичного трафика k-класса		
а	Ожидает начала обслуживания в k-очереди хотя бы одна сессия эластичного трафика	Сессия будет завершена и передача трафика из k-очереди будет начата, а скорость передачи всех сессий останется прежней
б	Сессии эластичного трафика, ожидающие начала обслуживания в k-очереди, отсутствуют	Сессия будет завершена, а скорость передачи всех оставшихся сессий на k-ресурсе будет пропорционально повышена
3. Наступление порога для времени ожидания начала обслуживания сессии эластичного трафика k-класса		
а	—	Сессия не будет обслужена
4. Поступление сигнала		
а	Условия для перераспределения ресурса в соответствии с заданной политикой управления выполнены	Ресурс будет перераспределен в соответствии с заданной политикой
б	Условия для перераспределения ресурса в соответствии с заданной политикой управления не выполнены	Перераспределения ресурса не будет

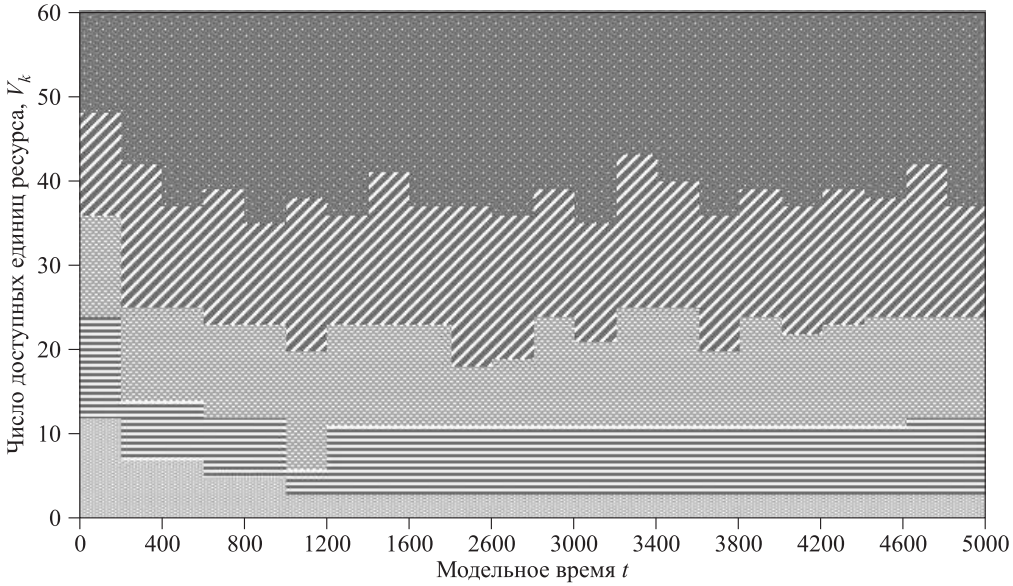


Рис. 1 Перераспределение ресурса

фициент γ использования ресурса, которые определяются по формулам:

$$\alpha = \frac{1}{K}\sum_{k=1}^{K}\alpha_k\,,$$

$$\alpha_k = \lim_{T\to\infty}\sum_{i=1}^{I(T)}\frac{t_i - t_{i-1}}{TY}\,\frac{\min\{V_k(t_i), V_k(t_0)\}}{V_k(t_0)}\,;$$

$$\beta = \lim_{T\to\infty}\sum_{i=1}^{I(T)}\frac{\max\limits_{k=1,\ldots,K} 1\{V_k(t_i)\neq V_k(t_i-0)\}}{I(T)}\,;$$

$$\gamma = \frac{1}{K}\sum_{k=1}^{K}\gamma_k\,,\quad \gamma_k = \lim_{T\to\infty}\sum_{j=1}^{J(T)}\frac{\tau^e_{kj}-\tau^b_{kj}}{T}\,,$$

где моменты τ^b_{kj} и τ^e_{kj} обозначают начало и завершение j-периода занятости для k-класса трафика.

Будем далее рассматривать нетерпеливый эластичный трафик с минимальным порогом скорости и возможностью ожидания начала обслуживания сессии в очереди.

В табл. 1 приведено тезисное описание функционирования системы для разных вариантов происходящих событий. Отметим, что перераспределение ресурса происходит при последовательном выполнении двух условий — поступления сигнала и выполнения условий заданной политики управления, решение которой зависит от состояния, в котором находится система. На рис. 1 проиллюстрировано распределение ресурса во времени для политики из работы авторов [9].

3 Система массового обслуживания

Перейдем к построению математической модели для двух классов эластичного трафика в виде системы массового обслуживания с сигналами. Предположим, что входящие потоки запросов на передачу эластичного трафика пуассоновские с интенсивностями λ_1/λ_2, а объем трафика распределен по экспоненциальному закону с параметрами μ_1/μ_2. С учетом порога b для скорости передачи трафика максимальное число обслуживаемых сессий составляет $N = \lfloor V/b\rfloor$, а число мест в очереди для ожидающих начала обслуживания сессий — R_1/R_2. Пусть пороги для времени ожидания начала обслуживания сессий распределены по экспоненциальному закону с параметрами $\varepsilon_1/\varepsilon_2$, а поток сигналов, управляющих перераспределением ресурса, является пуассоновским с интенсивностью δ.

Опишем функционирование системы при помощи случайного процесса $X(t)$ с состояниями вида $x = (m_1, m_2, n_1, n_2, r_1, r_2)$, где m_k — порог для максимального числа обслуживаемых сессий k-класса; n_k — число обслуживаемых сессий k-класса; r_k — число ожидающих начала обслуживания сессий k-класса, $k = 1, 2$. Тогда пространство состояний $X(t)$ будет иметь вид:

$$\begin{aligned}\mathsf{X} = \{&(m_1, N-m_1, n_1, n_2, 0, 0):\\ &0\le m_1\le N\,,\, n_1\ge 0\,,\, n_2\ge 0\,,\, n_1+n_2\le N\}\cup\\ &\cup\{(m_1, N-m_1, m_1, N-m_1, r_1, r_2):\\ &0\le m_1\le N\,,\, 0<r_1\le R_1\,,\, 0<r_2\le R_2\}\,.\end{aligned}$$

Таблица 2 Интенсивности переходов

№ п/п	Интенсивность события	Условие на x	Состояние x'
1а-1	λ_1	$n_1+1\le m_1$	$(m_1, N-m_1, n_1+1, n_2, 0, 0)$
1б-1	λ_1	$n_1+1>m_1, r_1+1\le R_1$	$(m_1, N-m_1, n_1, n_2, r_1+1, r_2)$
1а-2	λ_2	$n_2+1\le m_2$	$(m_1, N-m_1, n_1, n_2+1, 0, 0)$
1б-2	λ_2	$n_2+1>m_2, r_2+1\le R_2$	$(m_1, N-m_1, n_1, n_2, r_1, r_2+1)$
2а-1	$\frac{m_1}{N}V\mu_1$	$r_1>0$	$(m_1, N-m_1, n_1, n_2, r_1-1, r_2)$
2б-1	$\frac{m_1}{N}V\mu_1$	$r_1=0, n_1>0$	$(m_1, N-m_1, n_1-1, n_2, 0, r_2)$
2а-2	$\frac{m_2}{N}V\mu_2$	$r_2>0$	$(N-m_2, m_2, n_1, n_2, r_1, r_2-1)$
2б-2	$\frac{m_2}{N}V\mu_2$	$r_2=0, n_2>0$	$(N-m_2, m_2, n_1, n_2-1, r_1, 0)$
3а-1	$r_1\varepsilon_1$	$r_1>0$	$(m_1, N-m_1, m_1, n_2, r_1-1, r_2)$
3а-2	$r_2\varepsilon_2$	$r_2>0$	$(N-m_2, m_2, n_1, m_2, r_1, r_2-1)$
4а-1	δ	$n_1=m_1, r_1>0,$ $n_2<m_2, r_2=0,$ $r_1\le m_2-n_2$	$(m_1+r_1, m_2-r_1, m_1+r_1, n_2, 0, 0)$
4б-1	δ	$n_1=m_1, r_1>0,$ $n_2<m_2, r_2=0,$ $r_1>m_2-n_2$	$(m_1+m_2-n_2, n_2, m_1+m_2-n_2, n_2, r_1-m_2+n_2, 0)$
4а-2	δ	$n_2=m_2, r_2>0,$ $n_1<m_1, r_1=0,$ $r_2\le m_1-n_1$	$(m_1-r_2, m_2+r_2, n_1, m_2+r_2, 0, 0)$
4б-2	δ	$n_2=m_2, r_2>0,$ $n_1<m_1, r_1=0,$ $r_2>m_1-n_1$	$(n_1, m_2+m_1-n_1, n_1, m_2+m_1-n_1, 0, r_2-m_1+n_1)$

В табл. 2 перечислены возможные интенсивности переходов между состоянием x и другими состояниями системы, нумерация строк соответствует нумерации в табл. 1, а политика управления перераспределением ресурсов определена в строках 4. Исходя из описанных правил записывается матрица интенсивностей переходов и могут быть найдены стационарные вероятности $\pi(x)$, $x\in\mathsf{X}$.

4 Пример численного анализа

Зная стационарные вероятности $\pi(x)$, $x\in\mathsf{X}$, показатели эффективности перераспределения ресурса — коэффициент α соответствия начальному распределению ресурса, коэффициент β успеха перераспределения ресурса и коэффициент γ использования ресурса — можно найти по формулам:

$$\alpha=\frac{\alpha_1+\alpha_2}{2};$$

$$\alpha_k=\sum_{x\in\mathsf{X}}\frac{\min\{m_k,\tilde m_k\}}{\tilde m_k}\,\pi\left(m_1,m_2,n_1,n_2,r_1,r_2\right);$$

$$\beta=\sum_{x\in\mathsf{B}}\pi\left(m_1,m_2,n_1,n_2,r_1,r_2\right),$$

$$\begin{aligned}\mathsf{B}=\{x\in\mathsf{X}:\ &n_1=m_1,\ r_1>0,\ n_2<m_2,\ r_2=0,\ r_1\le m_2-n_2\}\cup\\ \cup\{x\in\mathsf{X}:\ &n_1=m_1,\ r_1>0,\ n_2<m_2,\ r_2=0,\\ &r_1>m_2-n_2\}\cup\{x\in\mathsf{X}:\ n_2=m_2,\ r_2>0,\\ &n_1<m_1,\ r_1=0,\ r_2\le m_1-n_1\}\cup\{x\in\mathsf{X}:\ n_2=m_2,\\ &r_2>0,\ n_1<m_1,\ r_1=0,\ r_2>m_1-n_1\},\end{aligned}$$

$$\gamma=\frac{\gamma_1+\gamma_2}{2},$$

$$\gamma_k=\sum_{x\in\mathsf{X}:\ m_k>0}\frac{n_k}{m_k}\,\pi\left(m_1,m_2,n_1,n_2,r_1,r_2\right).$$

Проиллюстрируем зависимость коэффициентов эффективности перераспределения ресурса от среднего значения $1/\delta$ интервала между поступлениями сигналов. На рис. 2 показаны коэффициенты α, β и γ для следующих исходных данных: $\lambda_1=1$, $\lambda_2=5$, $\mu_1=\mu_2=0{,}01$, $\varepsilon_1=\varepsilon_2=0$, $N_1=N_2=2$, $R_1=R_2=1$.

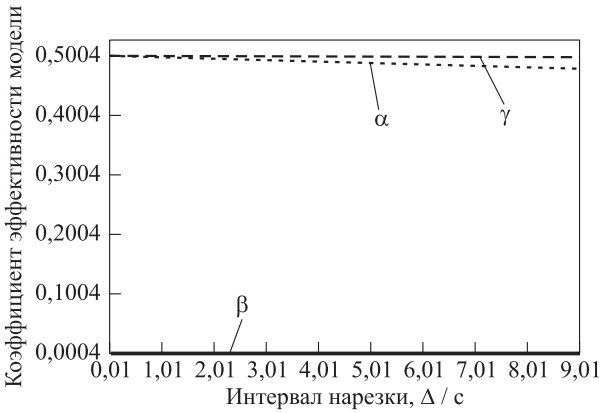


Рис. 2 Коэффициенты эффективности перераспределения ресурса

5 Заключение

Разработана математическая модель для двух классов эластичного трафика в виде системы массового обслуживания с очередью и минимальным порогом скорости с перераспределением ресурса, управляемым сигналами.

Предложены показатели эффективности перераспределения ресурса: коэффициент соответствия начальному распределению ресурса, коэффициент успеха перераспределения ресурса и коэффициент использования ресурса.

В дальнейшем предполагается рассмотреть другие модели, в частности с управлением перераспределением ресурсов с использованием методов машинного обучения.

Авторы благодарят студентку кафедры прикладной информатики и теории вероятностей РУДН Софию Бурцеву за помощь в проведении численного анализа.

Литература

1. *Muhizi S., Ateya A. A., Muthanna A., Kirichek R., Koucheryavy A.* A novel slice-oriented network model // Distributed computer and communication networks / Eds. V. Vishnevskiy, D. Kozyrev. — Communications in computer and information science ser. — Springer, 2018. Vol. 919. P. 421–431. doi: 10.1007/978-3-319-99447-5_36.
2. *Markova E., Adou Y., Ivanova D., Golskaia A., Samouylov K.* Queue with retrial group for modeling best effort traffic with minimum bit rate guarantee transmission under network slicing // Distributed computer and communication networks. — Lecture notes in computer science ser. — Springer, 2019. Vol. 11965. P. 432–442. doi: 10.1007/978-3-030-36614-8_33.
3. *Yarkina N., Gaidamaka Y., Correia L. M., Samouylov K.* An analytical model for 5G network resource sharing with flexible SLA-oriented slice isolation // Mathematics, 2020. Vol. 8. No. 7. Art. 1177. doi: 10.3390/math8071177.
4. *De Cola T., Bisio I.* QoS optimisation of eMBB services in converged 5G-satellite networks // IEEE T. Veh. Technol., 2020. Vol. 69. No. 10. P. 12098–12110. doi: 10.1109/TVT.2020.3011963.
5. *Ageev K., Sopin E., Samouylov K.* Resource sharing model with minimum allocation for the performance analysis of network slicing // Information technologies and mathematical modelling. Queueing theory and applications / Eds. A. Dudin, A. Nazarov, A. Moiseev. — Communications in computer and information science ser. — Springer, 2021. Vol. 1391. P. 378–389. doi: 10.1007/978-3-030-72247-0_28.
6. *Koucheryavy Y., Lisovskaya E., Moltchanov D., Kovalchukov R., Samuylov A.* Quantifying the millimeter wave new radio base stations density for network slicing with prescribed SLAs // Comput. Commun., 2021. Vol. 174. P. 13–27. doi: 10.1016/j.comcom.2021.04.010.
7. *Zhirnov N. S., Lyakhov A. I., Khorov E. M.* Mathematical model of a network slicing approach for video and Web traffic // J. Commun. Technol. El., 2019. Vol. 64. No. 8. P. 890–899. doi: 10.1134/ S1064226919080278.
8. *Vlaskina A., Polyakov N., Gudkova I.* Modeling and performance analysis of elastic traffic with minimum rate guarantee transmission under network slicing // Internet of things, smart spaces, and next generation networks and systems / Eds. O. Galinina, S. Andreev, S. Balandin, Y. Koucheryavy. — Lecture notes in computer science ser. — Springer, 2019. Vol. 11660. P. 621–634. doi: 10.1007/978-3-030-30859-9_54.
9. *Kochetkova I., Vlaskina A., Burtseva S., Savich V., Hosek J.* Analyzing the effectiveness of dynamic network slicing procedure in 5G network by queuing and simulation models // Internet of things, smart spaces, and next generation networks and systems / Eds. O. Galinina, S. Andreev, S. Balandin, Y. Koucheryavy. — Lecture notes in computer science ser. — Springer, 2020. Vol. 12525. P. 71–85. doi: 10.1007/978-3-030-65726-0_7.
10. *Власкина А. С., Поляков Н. А., Гудкова И. А., Гайдамака Ю. В.* Анализ вероятностно-временных характеристик обслуживания эластичного трафика с минимальной скоростью в сегменте беспроводной сети с нарезкой радиоресурсов // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика, 2020. Т. 20. № 3. С. 378–387. doi: 10.18500/1816-9791-2020-20-3-378-387.

Поступила в редакцию 29.07.2021

QUEUING SYSTEM WITH SIGNALS FOR DYNAMIC RESOURCE ALLOCATION FOR ANALYZING NETWORK SLICING IN 5G NETWORKS

I. A. Kochetkova[1,2], A. S. Vlaskina[1], N. N. Vu[1], and V. S. Shorgin[2]

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

[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119133, Russian Federation

Abstract: Network slicing is one of the key technologies of fifth generation wireless networks with an ever-increasing number of users and services. A feature of this technology is the ability to organize logically isolated segments of radio resources for the specific requirements of the operator and its users, and the redistribution of the resource volumes between the segments is possible, but only at the moments of the controller accessing them. In this work, a mathematical model is built for two classes of impatient elastic traffic with a minimum rate guarantee transmission by queuing system with queues and signals that control resource redistribution. Formulas are given for calculating the parameters of the efficiency of resource redistribution — the coefficient of compliance with the initial distribution of the resource, the success rate of resource redistribution, and the coefficient of resource utilization.

Keywords: 5G; network slicing; queuing system with signals; dynamic resource allocation; elastic traffic

DOI: 10.14357/19922264210312

Acknowledgments

This paper has been supported by the RUDN University Strategic Academic Leadership Program. The reported study was funded by RFBR, project No. 20-37-70079.

References

1. Muhizi, S., A. A. Ateya, A. Muthanna, R. Kirichek, and A. Koucheryavy. 2018. A novel slice-oriented network model. *Distributed computer and communication networks*. Eds. V. Vishnevskiy and D. Kozyrev. Communications in computer and information science ser. Springer. 919:421–431. doi: 10.1007/978-3-319-99447-5_36.
2. Markova, E., Y. Adou, D. Ivanova, A. Golskaia, and K. Samouylov. 2019. Queue with retrial group for modeling best effort traffic with minimum bit rate guarantee transmission under network slicing. *Distributed computer and communication networks*. Lecture notes in computer science ser. Springer. 11965:432–442. doi: 10.1007/978-3-030-36614-8_33.
3. Yarkina, N., Y. Gaidamaka, L. M. Correia, and K. Samouylov. 2020. An analytical model for 5G network resource sharing with flexible SLA-oriented slice isolation. *Mathematics* 8(7):1177. 19 p. doi: 10.3390/math8071177.
4. De Cola, T., and I. Bisio. 2020. QoS optimization of eMBB services in converged 5G-satellite networks. *IEEE T. Veh. Technol.* 69(10):12098–12110. doi: 10.1109/TVT.2020.3011963.
5. Ageev, K., E. Sopin, and K. Samouylov. 2021. Resource sharing model with minimum allocation for the performance analysis of network slicing. *Information technologies and mathematical modelling. Queueing theory and applications*. Eds. A. Dudin, A. Nazarov, and A. Moiseev. Communications in computer and information science ser. Springer. 1391:378–389. doi: 10.1007/978-3-030-72247-0_28.
6. Koucheryavy, Y., E. Lisovskaya, D. Moltchanov, R. Kovalchukov, and A. Samuylov. 2021. Quantifying the millimeter wave new radio base stations density for network slicing with prescribed SLAs. *Comput. Commun.* 174:13–27. doi: 10.1016/j.comcom.2021.04.010.
7. Zhirnov, N. S., A. I. Lyakhov, and E. M. Khorov. 2019. Mathematical model of a network slicing approach for video and web traffic. *J. Commun. Technol. El.* 64(8):890–899. doi: 10.1134/S1064226919080278.
8. Vlaskina, A., N. Polyakov, and I. Gudkova. 2019. Modeling and performance analysis of elastic traffic with minimum rate guarantee transmission under network slicing. *Internet of things, smart spaces, and next generation networks and systems*. Eds. O. Galinina, S. Andreev, S. Balandin, and Y. Koucheryavy. Lecture notes in computer science ser. Springer. 11660:621–634. doi: 10.1007/978-3-030-30859-9_54.
9. Kochetkova, I., A. Vlaskina, S. Burtseva, V. Savich, and J. Hosek. 2020. Analyzing the effectiveness of dynamic network slicing procedure in 5G network by queuing and simulation models. *Internet of things, smart spaces, and*

next generation networks and systems. Eds. O. Galinina, S. Andreev, S. Balandin, and Y. Koucheryavy. Lecture notes in computer science ser. Springer. 12525:71–85. doi: 10.1007/978-3-030-65726-0_7.

10. Vlaskina, A. S., N. A. Polyakov, I. A. Gudkova, and Yu. V. Gaidamaka. 2020. Analiz veroyatnostno-vremennykh kharakteristik obsluzhivaniya elastichnogo trafika s minimal'noy skorost'yu v segmente besprovodnoy seti s narezkoy radioresursov [Performance analysis of elastic traffic with minimum bit rate guarantee transmission in wireless network under network slicing]. *Izvestiya Saratovskogo universiteta. Novaya seriya. Seriya: Matematika. Mekhanika. Informatika* [Izvestiya of Saratov University. Mathematics. Mechanics. Informatics.]. 20(3):378–387. doi: 10.18500/1816-9791-2020-20-3-378-387.

Received July 29, 2021

Contributors

Kochetkova Irina A. (b. 1985) — Candidate of Science (PhD) in physics and mathematics, associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, 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 119133, Russian Federation; gudkova-ia@rudn.ru

Vlaskina Anastasia S. (b. 1995) —PhD student, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; vlaskina-as@rudn.ru

Vu Nhat N. (b. 1997) — student, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; 1032168360@rudn.ru

Shorgin Vsevolod S. (b. 1978) — 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 119133, Russian Federation; vshorgin@ipiran.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 3. С. 98–111

АНАЛИЗ СТРАТЕГИИ РАЗГРУЗКИ БАЗОВЫХ СТАНЦИЙ 5G NR С ПОМОЩЬЮ ТЕХНОЛОГИИ NR-U*

А. В. Дараселия[1], Э. С. Сопин[2], Д. А. Молчанов[3], К. Е. Самуйлов[4]

Аннотация: Ожидается, что базовые станции (БС) сетей пятого поколения (*англ.* Fifth Generation, 5G) Новое радио (*англ.* New Radio, NR) миллиметрового диапазона частот будут развернуты в областях с чрезвычайно высокими и резко колеблющимися требованиями к трафику, где это приводит к частым нарушениям качества обслуживания с точки зрения предоставляемой скорости на интерфейсе доступа, особенно в часы пик. В качестве одной из мер борьбы с перегрузками 3GPP (3rd Generation Partnership Project) рассматривает технологию NR-U (NR-Unlicensed), позволяющую использовать на БС помимо лицензируемого спектра также и нелицензируемый спектр частот, например 60 ГГц. В этом случае сессия, которая не может быть обслужена в лицензируемом спектре вследствие нехватки ресурсов, может быть перенаправлена в нелицензируемый спектр, где происходит конкуренция за ресурсы с абонентами технологии WiGig. Цель данного исследования – оценить параметры качества обслуживания в области, характеризующейся определенной плотностью пользователей NR и WiGig, где пользователи NR могут использовать технологию NR-U, если выполняются их требования к скорости. В качестве исследуемой метрики используется вероятность потери сессии NR и достижимая скорость передачи в нелицензируемом спектре частот. Проведенное численное исследование показало, что на исследуемые характеристики помимо плотности пользователей NR и WiGig оказывают влияние размер окна конкурентного доступа, плотность блокаторов и минимальная необходимая скорость. Представленные численные результаты позволяют сделать вывод о том, что рассматриваемый подход может значительно увеличить достижимую скорость пользовательских сессий, однако для этого требуется достаточно плотное развертывание технологии NR.

Ключевые слова: NR-U; New Radio; WiGig; QoS; теория массового обслуживания; ресурсная система массового обслуживания; марковский процесс

DOI: 10.14357/19922264210313

1 Введение

Технология 5G Новое радио, стандартизованная в рамках 3GPP Release 15 и Release 16, обещает резкое повышение скорости доступа на последней миле [1], особенно за счет использования частот миллиметрового диапазона (*англ.* millimeter wave, mmWave) [2]. Ожидается, что технология NR, работающая как часть гетерогенной технологии 5G, станет решающим шагом на пути к удовлетворению возросших требований пользователей к интерфейсу доступа.

На первом этапе проникновения на рынок БС mmWave NR будут развернуты в местах с высокой концентрацией пользовательского трафика, например в торговых центрах, на концертах. В этих условиях можно ожидать резких колебаний потребности в трафике. Так как технология NR, как ожидается, обеспечит определенный уровень качества обслуживания с точки зрения достигнутой скорости на интерфейсе доступа, эти колебания потенциально могут привести к ухудшению скорости сверх допустимого предела.

В данной статье проводится анализ подхода, обеспечивающего жизнеспособный вариант для сглаживания колебаний трафика. Рассматривается совместная реализация технологий NR и WiGig с использованием спектра 60 ГГц на одной физической БС. Так как обе технологии работают в диапазоне mmWave, они характеризуются изначально согласованными скоростями на интерфейсе доступа, а также, как ожидается, будут широко поддерживаться современными и будущими пользовательскими устройствами (ПУ). Ожидается, что такие системы, основанные на методе агрегации несущих, стандартизованном 3GPP, станут частью

*Публикация подготовлена при поддержке РНФ (проект 20-71-00124).

[1]Российский университет дружбы народов, avdaraseliya@rudn.ru

[2]Российский университет дружбы народов; Федеральный исследовательский центр «Информатика и управление» Российской академии наук, sopin-es@rudn.ru

[3]Российский университет дружбы народов; Университет Тампере, Финляндия, dmitri.moltchanov@tuni.fi

[4]Российский университет дружбы народов; Федеральный исследовательский центр «Информатика и управление» Российской академии наук, samuylov-ke@rudn.ru

будущих сетей 5G. Однако протоколы управления доступом в нелицензируемом диапазоне частот, использующие процедуру случайного доступа, не всегда могут гарантировать требуемый уровень скорости передачи, что ставит задачу расчета необходимой плотности таких NR-нелицензированных БС для поддержки заданной плотности ПУ NR-U и ПУ WiGig, работающих только с использованием технологии WiGig.

В этой статье исследуется процесс совместного обслуживания сессий с использованием базовых станций NR-U, объединяющих технологии NR и WiGig. С помощью инструментов теории массового обслуживания и марковских процессов разработана модель, позволяющая вычислять скорости, предоставляемые ПУ NR-U и ПУ WiGig. Эта промежуточная метрика дает возможность определить вероятность потери сессии ПУ NR-U, а также оценить требуемую плотность базовых станций NR-U для заданной плотности ПУ NR-U и WiGig.

2 Модель системы

2.1 Модель развертывания

Рассмотрим систему с технологиями NR и WiGig, физически расположенную на одной БС NR-U. Технология NR использует лицензируемый диапазон 28 ГГц с шириной канала $B_N = 400$ МГц [3]. Предполагается, что технология WiGig работает в диапазоне 60 ГГц, используя канал шириной $B_W = 2{,}16$ ГГц [4]. Предположим, что БС NR-U развернуты в соответствии с точечным процессом Пуассона в $\mathcal{R}^2$ с плотностью λ_A БС на квадратный метр (рис. 1). Высота БС NR-U — h_B.

В рассматриваемом развертывании есть два типа устройств: ПУ NR-U и ПУ WiGig. Устройства первого типа способны работать как в диапазонах NR, так и в диапазонах WiGig, а ПУ WiGig используют только технологию WiGig. Расположение ПУ NR-U и WiGig определяется также в соответствии с точечным процессом Пуассона с плотностями $\chi_{B,N}$ и $\chi_{B,W}$ на 1 м^2 соответственно. Высота всех пользовательских устройств составляет h_U.

Совместная реализация NR и WiGig достигается за счет использования технологии агрегации несущих 3GPP [5]. Никакого дополнительного взаимодействия или сигнализации между технологиями NR и WiGig на стороне БС NR-U не предполагается, поскольку вся логика реализована на ПУ NR-U. Нелицензируемый диапазон использует БС NR-U для обслуживания сессий ПУ NR-U, выгруженных из лицензируемой части NR. В частности, если гарантии скорости для ПУ не могут быть предоставлены в NR-части базовой станции NR-U, ПУ пробует использовать нелицензируемую технологию WiGig. Если достигнутой скорости недостаточно, ПУ NR-U покидает систему.

2.2 Механизм сосуществования

Все ПУ, использующие нелицензируемый диапазон, применяют подход с прослушиванием канала перед передачей (*англ.* Listen-Before-Talk, LBT). Предполагается, что технология NR-U использует механизм CoLBT (Channel observation-based LBT)

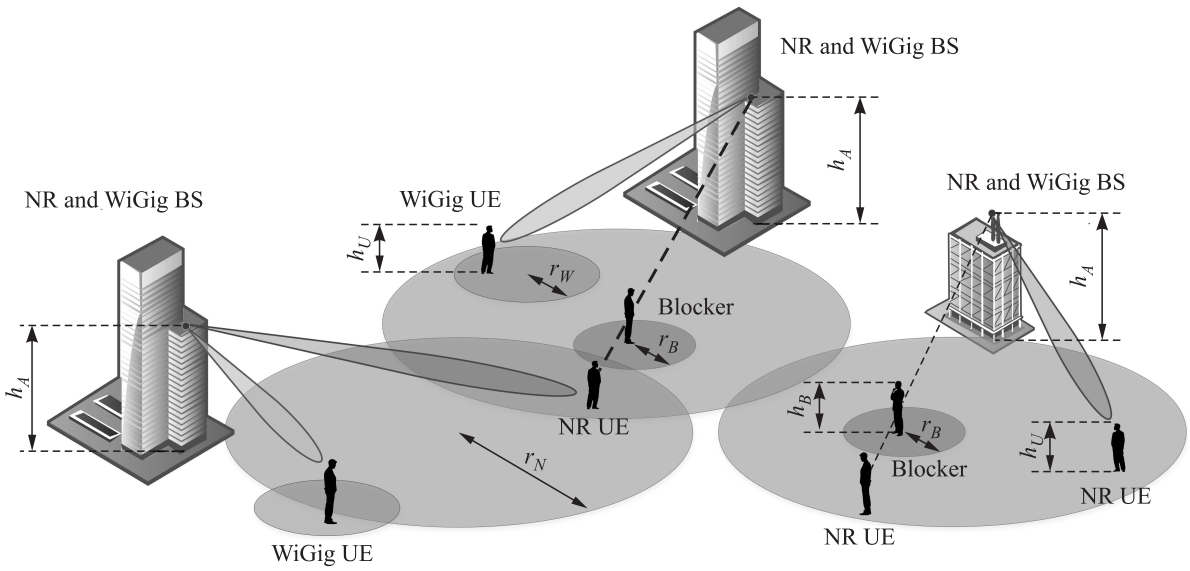


Рис. 1 Иллюстрация рассматриваемого сценария развертывания (UE — user equipment)

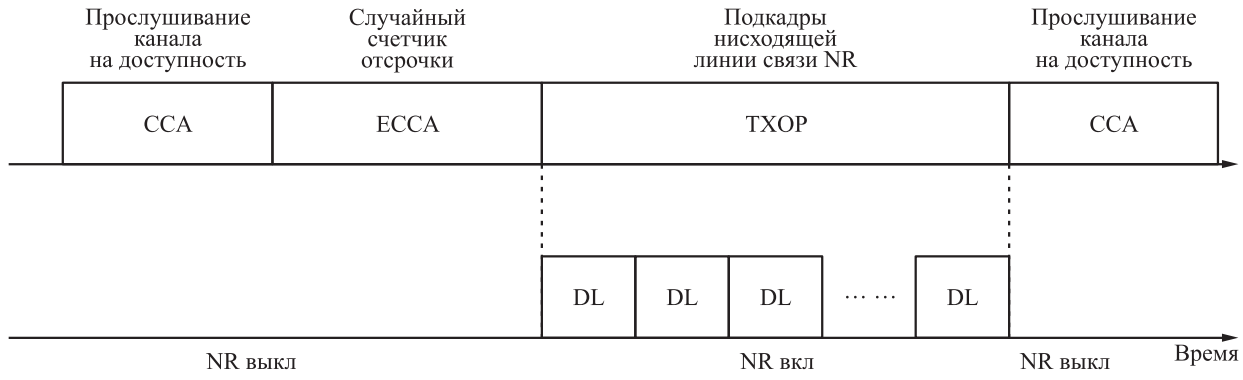


Рис. 2 Иллюстрация рассматриваемого механизма случайного доступа

на основе окна конкурентного доступа (*англ.* contention window, CW) и концепции счетчика отсрочки (рис. 2), который аналогичен рекомендованному 3GPP для технологии LAA (Licensed Assisted Access) [6]. Изначально размер CW равен 32, а максимальное число повторных передач равно T.

Пользовательские устройства, у которых есть пакет, готовый к передаче, генерируют случайное целое число тактов отсрочки на отрезке $[1, \mathrm{CW}]$ в соответствии с равномерным распределением. Значение счетчика отсрочки уменьшается на единицу в каждом такте, где канал считается свободным. Если канал занят, счетчик приостанавливается и продолжается прослушивание канала. Когда значение счетчика отсрочки достигает единицы, ПУ передает свой пакет. Возможны три исхода:

(1) успешная передача;
(2) неуспешная передача из-за коллизии (из-за конфликта с другой передачей пользовательского устройства NR-U или WiGig);
(3) неуспешная передача из-за блокировки пути прямой видимости (*англ.* Line of Sight, LoS).

В случаях неуспешной передачи максимальное значение счетчика отсрочки удваивается, а при успешной — сбрасывается до исходного значения.

На процедуру разрешения конфликтов как для БС NR-U, так и для точек доступа WiGig влияет направленность используемых антенных решеток, т. е. коллизия может произойти только тогда, когда ПУ, расположенные в одном секторе WiGig, пытаются передавать в одно и то же время. С точки зрения моделирования это означает, что эффективное число ПУ ограничено направленностью антенной решетки WiGig.

2.3 Модели распространения, блокировки и антенны

1. Модель блокировки. В модели учитывается перекрытие пешеходами путей распространения. Предполагается, что пешеходы движутся в $\mathcal{R}^2$ в соответствии с моделью подвижности со случайным направлением [7] со скоростью v метров в секунду и экспоненциально распределенной длиной пробега со средним значением τ метров. Пешеходы моделируются как цилиндры высотой h_B и радиусом r_B.

Обозначим через $p_b(r)$ вероятность того, что ПУ, расположенное на расстоянии r от БС NR-U, заблокировано. Обобщая результаты из [7, 8], получаем следующее выражение для этой вероятности:

$$p_b(r) = 1 - e^{-2\lambda_B r_B (r(h_B - h_U)/(h_W - h_U) + r_B)},$$

где λ_B — плотность пешеходов.

2. Модель распространения. Поскольку обе рассматриваемые технологии работают в миллиметровом диапазоне, используются аналогичные модели распространения. Отношение сигнал–шум плюс помехи (*англ.* signal to interference & noise ratio, SINR) на приемнике, расположенном на расстоянии e, составляет

$$S(y) = \frac{P_{N,U} G_{N,A} G_{N,U}}{(N_0 W + M_I) L(y)},$$

где $P_{N,U}$ — мощность передачи пользовательского устройства; $G_{N,A}$ и $G_{N,U}$ — коэффициенты усиления антенной решетки на БС и ПУ соответственно; N_0 — спектральная плотность мощности шума; W — рабочая полоса пропускания; M_I — граница интерференции; $L(y)$ — потери распространения. Для заданной плотности развертывания БС NR-U можно оценить значение интерференции, используя модели на основе стохастической геометрии [9–11].

Согласно [12], потери распространения, измеряемые в децибелах, определяются как

$$L_{\mathrm{dB}}(y) = \begin{cases} 32{,}4 + 21{,}0\log(y) + 20\log f_{M,c} \\ \text{в незаблокированном состоянии;} \\ 32{,}4 + 3{,}19\log(y) + 20\log f_{M,c} \\ \text{в заблокированном состоянии,} \end{cases} \quad (1)$$

где $f_{M,c}$ — рабочая частота, ГГц; x — расстояние между БС и ПУ.

Потери распространения в (1) можно представить в линейном масштабе, используя модель вида $A_i y^{-\zeta_i}$, где A_i и ζ_i — коэффициенты распространения. Вводя коэффициенты (A_1, ζ_1) и (A_2, ζ_2) которые соответствуют незаблокированным и заблокированным состояниям, получаем

$$A = A_1 = A_2 = 10^{2\log_{10} f_{M,c}+3{,}24}, \qquad \zeta_1 = 2{,}1\,, \quad \zeta_2 = 3{,}19\,.$$

Тогда значение SINR на ПУ можно записать как

$$S(y) = \\ = \frac{P_{N,U}G_{N,A}G_{N,U}}{(N_0W + M_I)A}\left[y^{-\zeta_1}\left[1 - p_b(y)\right] + y^{-\zeta_2}p_b(y)\right],$$

где $p_b(y)$ — вероятность блокировки на расстоянии y.

Вводя коэффициент

$$C = \frac{P_{N,U}G_{N,A}G_{N,U}}{(N_0W + M_I)A}\,, \quad (2)$$

модель распространения можно представить в виде:

$$S(y) = Cy^{-\zeta_1}[1 - p_b(y)] + C_2y^{-\zeta_2}p_b(y). \quad (3)$$

3. Модель антенны. Аналогично [9] предполагается, что диаграмма направленности антенной решетки представляет собой коническую зону с углом α, совпадающим с HPBW (half-power beamwidth) антенной решетки. Согласно [13], HPBW антенной решетки α пропорциональна числу элементов:

$$\alpha = 2|\theta_m - \theta_{3\mathrm{dB}}|.$$

Здесь $\theta_{3\mathrm{dB}}$ — угол, при котором значение излучаемой мощности на 3 дБ ниже максимума; θ_m — положение максимума массива:

$$\theta_m = \arccos\left(-\frac{\beta}{\pi}\right),$$

где β — ориентация массива, точнее азимутальный угол, представляющий физическую ориентацию массива, т. е. $\theta_m = \pi/2$ для $\beta = 0$.

Среднее усиление антенны по HPBW можно найти, как в [13]:

$$G = \frac{1}{\theta^+_{3\mathrm{dB}} - \theta^-_{3\mathrm{dB}}}\int\limits_{\theta^-_{3\mathrm{dB}}}^{\theta^+_{3\mathrm{dB}}} \frac{\sin(N\pi\cos(\theta)/2)}{\sin(\pi\cos(\theta)/2)}\,d\theta, \quad (4)$$

где $\theta^{\pm}_{3\mathrm{dB}} = \arccos[-\beta \pm 2{,}782/(N\pi)]$; N — количество антенных элементов.

2.4 Схемы трафика, распределения ресурсов и разгрузки

Предположим, что ПУ NR-U и WiGig генерируют эластичные требования к трафику. Согласно этой модели (также известной как модель полного буфера в 3GPP) ПУ всегда имеют данные для передачи. Однако ПУ NR-U характеризуются некоторыми требованиями к минимальной скорости $R_{\min}$, которые предоставляются как часть соглашения QoS между оператором сети и пользователями. Обратим внимание, что в зависимости от расстояния между БС NR-U и ПУ одна и та же минимальная скорость $R_{\min}$ требует разного объема ресурса.

Пользовательские устройства WiGig используют только нелицензируемый диапазон. Предполагается, что выбор точки доступа для подключения происходит на основе мощности приема опорного сигнала. Таким образом, ПУ связаны с ближайшей БС WiGig. Для ПУ NR-U все несколько иначе. Изначально ПУ NR-U пытается установить соединение с ближайшей БС NR-U и использовать технологию NR. Однако если текущая скорость, предоставляемая другим ПУ NR-U, в данное время использующим технологию NR на этой БС, упадет ниже требуемой минимальной скорости $R_{\min}$, сессия будет перенаправлена на связанную технологию WiGig. Если скорость, предоставляемая на интерфейсе WiGig, недостаточна для удовлетворения требований к минимальной скорости $R_{\min}$, сессия NR-U сбрасывается.

Также отметим, что в WiGig могут быть выгружены только те сессии, которые ближе к БС NR-U, чем r_W, где r_W — покрытие точки доступа WiGig.

3 Распределение объема запрашиваемых ресурсов

Для параметризации процесса обслуживания в NR-части базовой станции NR-U нужны следующие параметры:

(1) распределение объема ресурса, необходимого для одной сессии пользовательского устройства NR-U;

(2) доля сессий, которые могут быть выгружены в WiGig-часть базовой станции NR-U.

Для определения искомых параметров требуется определить эффективные радиусы покрытия NR и WiGig частей БС NR-U, r_N и r_W. Поскольку

оба радиуса получаются аналогично, ниже рассматривается только r_N. В области БС NR-U эффективный радиус покрытия r_N равен минимуму от расстояния между БС NR-U $r_{N,V}$ и максимального покрытия NR-части базовой станции NR-U $r_{N,S}$, т. е. $r_N = \min(r_{N,S}, r_{N,V})$. Ниже получим эти компоненты.

Радиус $r_{N,S}$ определяется как максимальное расстояние между ПУ NR-U и БС NR-U, так что ПУ NR-U в условиях блокировки LoS не находится в условиях простоя. Согласно используемой модели распространения, отношение сигнал/шум на максимальном двумерном расстоянии $r_{N,S}$ определяется выражением:

$$S = C_2 \left(r_{N,S}^2 + (h_B - h_U)^2\right)^{-\zeta/2} = S_{\text{th}},$$

где S_{th} является SNR (*англ.* signal to noise ratio), соответствующим самой низкой возможной схеме модуляции и кодирования (*англ.* modulation and coding scheme, MCS) NR. Решая это уравнение относительно $r_{N,S}$, получаем:

$$r_{N,S} = \sqrt{\left(\frac{S_{\text{th}} M_{S,B}}{C_2}\right)^{2/\zeta} - (h_B - h_U)^2}.$$

Здесь $M_{S,B}$ — граница теневого замирания, определяемая как

$$M_{S,B} = \sqrt{2}\sigma_{S,B}\text{erfc}^{-1}(2p_C)\,,$$

где $\text{erfc}^{-1}(\cdot)$ — обратная дополнительная функция ошибок; p_C — вероятность покрытия границы соты; $\sigma_{S,B}$ — стандартное отклонение распределения теневого замирания для состояния блокировки LoS, которое представлено в [12]. Отметим, что $r_{N,S}$ зависит от C_2 из (2), который, в свою очередь, зависит от угла сектора α согласно (4). Отметим, что обычно $r_{W,S} < r_{N,S}$ из-за различий в несущих частотах и допустимой излучаемой мощности.

Радиус $r_{N,V}$, равный половине расстояния между соседними БС NR-U, определяется аппроксимацией круга ячейки Вороного, вызванной местоположением БС NR-U в $\mathcal{R}^2$. Поскольку фактическая площадь ячейки Вороного неизвестна [14], используем компьютерное моделирование, чтобы получить $r_{N,V}$. Радиус покрытия WiGig-части базовой станции NR-U r_W получается аналогично.

Как только радиусы r_N и r_S получены, можно приступить к определению объема ресурсов, запрашиваемых ПУ NR-U. Для этого потребуется получить функцию распределения (ФР) SNR. Заметим, что в рассматриваемой модели случайность SNR обусловлена двумя факторами: расположением ПУ относительно БС и затуханием [12]. Для учета затухания, имеющего логнормальное распределение по линейной шкале, сначала преобразуем его в шкалу децибел, где оно описывается нормальным распределением.

Выведем теперь ФР трехмерного расстояния между ПУ и БС NR, предполагая, что ПУ равномерно распределены в плоскости покрытия. Расстояние на плоскости распределено согласно плотности распределения $w_R(x) = 2x/r_N^2$ [15]. Теперь трехмерное расстояние может быть получено с помощью расстояния на плоскости через $\phi_D(r) = \sqrt{(h_B - h_U)^2 + r^2}$ при помощи метода преобразования случайных величин [16]. В частности, если случайная величина Y с плотностью распределения $w(y)$ является функцией $Y = \phi(X)$ от другой случайной величины X с плотностью распределения вероятностей $f(x)$, то

$$w(y) = \sum_{\forall i} f(\psi_i(y)) \left|\frac{d\psi_i'(y)}{dy}\right|, \qquad (5)$$

где $x = \psi_i(y) = \phi^{-1}(x)$ — обратные функции.

Подставляя $w_r(x)$ и $\phi_r(x)$ в (5), получаем

$$W_d(x) = \frac{2h_B h_U - h_B^2 + h_U^2 + x^2}{d_N^2}$$

для $h_B - h_U < x < \sqrt{r_N^2 + h_B^2 - 2h_B h_U + h_U^2}$.

Обратную функцию от отношения сигнал/шум в децибелах без затухания можно найти с помощью того же метода, где SNR в децибелах является функцией трехмерного расстояния:

$$\phi_{\text{SNR,dB}}(d) = 10\log_{10}\left(Ad^{-\zeta}\right).$$

Здесь d — трехмерное расстояние между ПУ и БС NR; ζ является показателем потерь распространения; A определяет все усиления и потери, кроме потерь при распространении и флуктуаций из-за затухания. Подставляя $\phi_{\text{SNR,dB}}(x)$ и $W_d(x)$ в (5), получаем ФР SNR:

$$W_{S^{\text{dB}}}(x) = 1 - \frac{10^{-x/(5\zeta)} A^{2/\zeta} - (h_B - h_U)^2}{r_N^2}$$

при $x > 10\log_{10}[A(r_N^2 + (h_B - h_U)^2)^{-\zeta/2}]$.

Учитывая, что теневое затухание характеризуется логнормальным распределением по линейной шкале, приводящим к нормальному распределению по шкале децибел, случайная величина SNR может быть записана в виде суммы $S_{\text{SF}} = S^{\text{dB}} + \mathcal{N}(0,\sigma)$, где σ — стандартное отклонение теневого затухания. Наконец, определяем ФР SNR как свертку $W_S(y)$ и нормального распределения с нулевым средним и стандартным отклонением σ, т. е.

$$W_{S_{\text{SF}}}(y) = \int_{-\infty}^{\infty} W_S\,(y+u)\frac{e^{-u^2/(2\sigma^2)}}{\sqrt{2\pi}\,\sigma}\,du\,.$$

К сожалению, последние не могут быть оценены в замкнутой форме с использованием техники преобразования случайных величин, но могут быть представлены в терминах функции Лапласа:

$$W_{S_{\rm SF}}(x) = \frac{1}{2r_N^2}\Big[A^{2/\zeta}10^{-x/(5\zeta)}e^{\sigma^2\log^2(10)/(50\zeta^2)}\times$$
$$\times\Big[\mathrm{erf}\Big(\big(50\zeta\log A - 25\zeta^2\log B + \sigma^2\log^2(10) -$$
$$-5\zeta x\log(10)\big)\Big/\Big(5\sqrt{2}\zeta\sigma\log(10)\Big)\Big) -$$
$$-\mathrm{erf}\Big(\big(50\zeta(\log A - \zeta\log(h_B-h_U)) + \sigma_S^2\log^2(10) -$$
$$-5\zeta x\log(10)\big)\Big/\Big(5\sqrt{2}\zeta\sigma\log(10)\Big)\Big)\Big] +$$
$$+\big(r_N^2 + (h_B-h_U)^2\big)\times$$
$$\times\mathrm{erf}\left(\frac{-10\log A + 5\zeta\log B + x\log(10)}{\sqrt{2}\sigma\log(10)}\right) - (h_B-h_U)^2\times$$
$$\times\,\mathrm{erf}\Big(\Big(\sqrt{2}\,(-10\log A + 10\zeta\log(h_B-h_U) +$$
$$+x\log(10))\Big)\Big/\big(\sigma\log(100)\big)\Big) + r_N^2\Big], \quad (6)$$

где $B = r_N^2 + (h_B - h_U)^2$; $\mathrm{erf}(\cdot)$ — функция Лапласа. Включая потери, вызванные блокировкой L_B в A и подставляя $\sigma_{S,B}$ и $\sigma_{S,\mathrm{nB}}$ в (6), можем получить две ФР SNR $W_{S_{\mathrm{nB}}}$ и W_{S_B} для незаблокированного и заблокированного состояний.

Взвешивая эти две ФР с помощью вероятностей того, что ПУ находится в заблокированном/незаблокированном состоянии, получаем окончательное выражение для ФР SNR:

$$W_S(x) = p_b W_{S_B}(x) + (1-p_b)\,W_{S_{\mathrm{nB}}}(x).$$

Функция распределения требования сессии к ресурсам получается путем сопоставления ФР SNR с граничными значениями SNR для каждой кодово-модуляционной схемы NR [17] при заданной частоте ошибок блока (*англ.* block error rate, BLER). Отметим, что из-за различий в r_N и r_W, а также в системных параметрах технологий NR и WiGig эти скорости могут различаться даже для одного и того же ПУ NR-U. Учитывая, что ПУ NR-U распределены по точечному процессу Пуассона в $\mathcal{R}^2$ для NR-части базовой станции NR-U, а также принимая во внимание значения спектральной эффективности кодово-модуляционных схем, среднюю спектральную эффективность можно получить по формуле:

$$E[S_e] = \int_0^{r_N} \frac{2x}{r_N}\,\log_2(1+S(x))\,dx\,,$$

где $S(x)$ определяется в (3).

4 Анализ стратегии разгрузки

Для анализа стратегии разгрузки разработана модель, состоящая из двух основных компонентов:

(1) ресурсная система массового обслуживания, описывающая обслуживание сессий в лицензируемом диапазоне;

(2) цепь Маркова для анализа эффективности случайного доступа к разделяемой среде WiGig.

Как показано на рис. 3, вероятность потери на NR, вызванная нехваткой ресурсов, является вероятностью выгрузки сессий на WiGig, и только при потере еще и на WiGig сессия теряется окончательно. Вероятность потери на WiGig определяется как вероятность того, что достижимая скорость меньше минимально необходимой $R_{\min}$.

Из-за различия между радиусами покрытия NR- и WiGig-частей базовой станции NR-U r_N и r_W (т. е. $r_N > r_W$), вводятся два типа сессий. Сессии первого типа, расстояние до БС которых удовлетворяет неравенству $r_W < x < r_N$, могут обслуживаться только частью NR базовой станции NR-U. Если

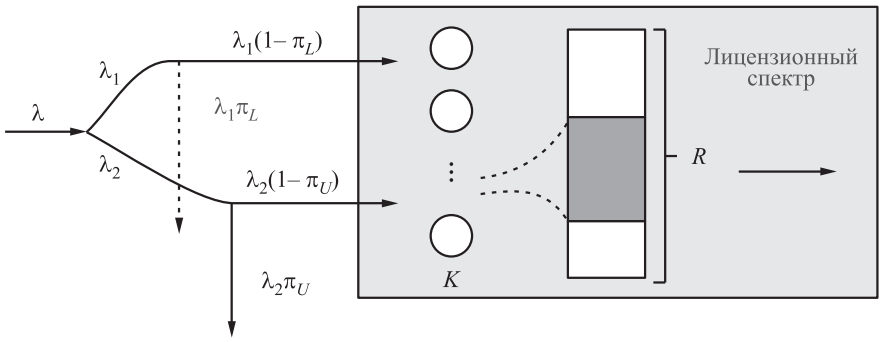


Рис. 3 Схема ресурсной системы массового обслуживания

ресурсов для обслуживания таких сессий недостаточно, то они теряются. Сессии второго типа ($x \leq r_W$) могут быть выгружены в WiGig при нехватке ресурсов на NR.

Пусть $\lambda = \chi_{B,N} r_N^2 \pi p_N \lambda_S$ — интенсивность поступления всех сессий NR-U, где $\chi_{B,N}$ — плотность пользовательских устройств NR на квадратный метр, определенная в подразд. 2.1; $r_N^2\pi$ — зона покрытия базовой станции для технологии NR, представленной в подразд. 2.4; p_N — вероятность того, что пользователь передает в определенный момент времени; λ_S — интенсивность генерации сессии от каждого ПУ NR. Интенсивность λ является суммой интенсивностей λ_1 и λ_2 поступления заявок первого (из кольца $r_W < x < r_W$) и второго (из окружности $0 < x < r_W$) типов.

Интенсивность поступления в лицензируемую полосу равна интенсивности поступления всех сессий, т. е. $\lambda_L = \lambda$. Интенсивность поступления λ_U на нелицензируемый спектр равна интенсивности потерь сессий второго типа на NR, т. е. $\lambda_U = \lambda_2 \pi_U$, где π_U — вероятность потерь сессий второго типа в лицензируемой части. Кроме того, обозначим π_L вероятность потерь сессий первого типа, которая приводит к окончательной ее потере.

Основной метрикой служит требуемая плотность БС NR-U, которые должны быть развернуты в области для поддержания минимальной скорости сессии NR-U $R_{\min}$ для заданных плотностей ПУ NR-U и WiGig, $\chi_{B,N}$ и $\chi_{B,W}$ соответственно.

В следующем подразделе выводятся формулы для расчета этой метрики, а также для вероятности потери сессии ПУ NR-U.

4.1 Процесс обслуживания в лицензируемом диапазоне

Для моделирования процесса обслуживания сессии в лицензируемом диапазоне используются ресурсные системы массового обслуживания [18–21]. С этой целью рассмотрим систему массового обслуживания с $K < \infty$ сессиями и дискретным объемом $R < \infty$ единиц ресурса, где K обозначает максимальное число ПУ NR-U в системе, т. е. максимальное количество заявок, которые могут одновременно обслуживаться в NR-части базовой станции NR-U. В систему поступают два пуассоновских потока сессий двух типов с интенсивностями λ_1 и λ_2. Времена обслуживания распределены экспоненциально с параметром μ.

Процесс обслуживания каждой сессии требует случайного объема ресурса $0 \leq r \leq R$. Распределение требований к ресурсу для рассматриваемых типов заявок обозначается $\{p_{l,j}\}_{j\geq 0}$, $l = 1, 2$, где $p_{l,j}$ — вероятность того, что сессия типа l требует j единиц ресурса. Согласно [18], ресурсная система массового обслуживания с двумя потоками может быть проанализирована как система с одним агрегированным потоком, распределение требований к ресурсам в котором определяется по формуле:

$$p_{j,L} = \frac{\rho_1}{\rho} p_{1,j} + \frac{\rho_2}{\rho} p_{2,j},$$

где предлагаемая нагрузка определяется как $\rho = \rho_1 + \rho_2$, $\rho_i = \lambda_i/\mu$, $i = 1, 2$.

Система работает следующим образом. Поступающая сессия принимается в систему, если на момент прибытия в системе достаточно доступного ресурса для удовлетворения требований этой сессии. Поступающая сессия сбрасывается, если в момент поступления объем ресурса, требуемого для нее, превышает объем доступного ресурса. В этом случае сессия первого типа будет потеряна, а сессия второго типа — перенаправлена на нелицензируемый спектр. Когда время обслуживания сессии истекает, она покидает систему, освобождая все занятые ресурсы. Поведение системы можно описать случайным процессом $X(t) = (\xi(t), \gamma(t))$. Здесь $\xi(t)$ — число сессий в системе, а $\gamma(t) = (\gamma_1(t), \gamma_2(t), \ldots, \gamma_{\xi(t)}(t))$, где $\gamma_i(t)$ — объем ресурса, выделенного i-й обслуживаемой сессии в момент времени t.

Обозначим через $\mathsf{P}_k(r)$ стационарную вероятность того, что в системе обслуживаются k сессий, которые занимают суммарно r ресурсов, т. е.

$$\mathsf{P}_k(r) = \lim_{t\to\infty} \mathsf{P}\left\{\xi(t) = k, \sum_{i=1}^{\xi(t)} \gamma_i(t) = r\right\}, \quad 0 \leq r \leq R.$$

Согласно [22], стационарное распределение задается выражениями

$$\mathsf{P}_k(r) = \mathsf{P}_0 \frac{\rho^k}{k!} p_{r,L}^{(k)},\ k = 1, 2, \ldots, K,\ r = 1, 2, \ldots, R,$$

где

$$\mathsf{P}_0 = \left(1 + \sum_{k=1}^{K} \frac{\rho^k}{k!} \sum_{r=0}^{R} p_{r,L}^{(k)}\right)^{-1};$$

$\{p_{r,L}^{(k)}\}_{r\geq 0}$ — k-кратная свертка распределения $\{p_{r,L}\}_{r\geq 0}$, вычисляемая рекуррентно:

$$p_{r,L}^{(k)} = \sum_{j=0}^{r} p_{r-j,L}^{(k-1)} p_{j,L}.$$

Вероятность потери π_U сессии второго типа определяется выражением:

$$\pi_U = 1 - \mathsf{P}_0 \sum_{k=0}^{K-1} \frac{\rho^k}{k!} \sum_{r=0}^{R} p_{r,L}^{(k+1)}. \qquad (7)$$

Для больших значений K и R расчет по формуле (7) является вычислительно трудоемким. В этом случае можно применять сверточный вычислительный алгоритм, разработанный в [19]. В соответствии с ним вероятность потерь может быть рассчитана по формуле:

$$\pi_U = 1 - G^{-1}(K,R)\sum_{i=0}^{R} p_{2,i}G(K-1,R-i),$$

где значения $G(n,r)$ задаются как

$$G(n,r) = \sum_{i=0}^{n}\frac{\rho^i}{i!}\sum_{j=0}^{r}p_{j,L}^{(i)}$$

и вычисляются рекуррентно.

Аналогично, вероятность потерь сессий первого типа вычисляется по формуле:

$$\pi_L = 1 - G^{-1}(K,R)\sum_{i=0}^{R} p_{1,i}G(K-1,R-i).$$

4.2 Процесс обслуживания в нелицензируемом спектре

Здесь охарактеризуем распределение требований к ресурсу и интенсивность поступления сессий в WiGig-часть базовой станции NR-U.

Распределение требований к ресурсу сброшенных на лицензируемой полосе сессий рассчитывается по формуле:

$$p_{j,U} = \frac{p_{2,j}}{\pi_U}\left(\sum_{r=0}^{R}\mathsf{P}_K(r) + \sum_{k=0}^{K-1}\sum_{r=R-j+1}^{R}\mathsf{P}_k(r)\right), \quad 0 \le j \le R. \quad (8)$$

Распределение (8) можно также получить при помощи функций $G(k,r)$, используя их определение и правила преобразования сумм:

$$p_{j,U} = \frac{1}{\pi_U}p_{2,j}\frac{G(K,R) - G(K-1,R-j)}{G(K,R)}.$$

Сессии ПУ NR-U, выгруженные на WiGig-полосу, конкурируют за ресурсы передачи с ПУ WiGig. Перейдем к выводу вероятности успешной передачи для обоих типов ПУ, которая используется далее для определения скорости, полученной ПУ NR в нелицензируемом диапазоне.

Пусть p_c — вероятность коллизии, а p_b — вероятность того, что путь прямой видимости заблокирован. Тогда вероятность успешной передачи может быть выражена как

$$\theta = (1-p_c)(1-p_b). \quad (9)$$

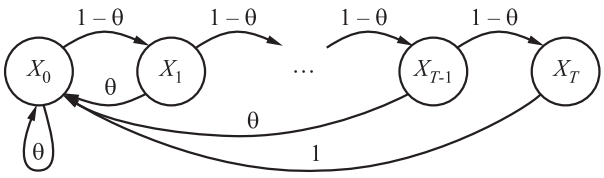


Рис. 4 Диаграмма переходов состояний марковской модели

Процедуру случайного доступа к среде передачи, представленной в подразд. 2.2, можно описать цепью Маркова $\{X_n,\ n \ge 0\}$, где X_n обозначает число неудачных попыток передать данные с момента последней успешной попытки и принимает значения от 0 до T. Граф переходных вероятностей цепи Маркова, представленный на рис. 4, позволяет получить систему уравнений для ее стационарного распределения:

$$\left.\begin{aligned} &q_0 = q_0\theta + q_1\theta + \cdots + q_{T-1}\theta + q_T;\\ &q_i = q_{i-1}(1-\theta),\ \ i = 1,\ldots,T-1;\\ &q_T = q_{T-1}(1-\theta). \end{aligned}\right\} \quad (10)$$

Решая систему (10), получаем формулу для вычисления стационарных вероятностей q_i в следующем виде:

$$q_i = \frac{\theta}{1-(1-\theta)^{T+1}}(1-\theta)^i,\ \ i = 0,\ldots,T. \quad (11)$$

Пусть теперь π_N и π_W — вероятности того, что ПУ NR-U и WiGig соответственно делают попытку передачи в произвольно выбранный такт времени. Тогда, если существуют конкурирующие сессии n_N NR-U и n_W WiGig, вероятность коллизии равна

$$p_c = (1-p_b)\left(1-(1-\pi_N)^{n_N}(1-\pi_W)^{n_W}\right). \quad (12)$$

Поскольку ПУ пытаются передавать только в одном временно́м такте каждого состояния цепи Маркова $\{X_n\}$, вероятность передачи π_N может быть вычислена как обратное значение средней длительности (в тактах) пребывания в одном состояния цепи Маркова:

$$\pi_N = \left[\sum_{i=0}^{T} q_j b_j\right]^{-1}, \quad (13)$$

где среднее число тактов b_j в состоянии j

$$b_j = \sum_{i=1}^{2^jW}\frac{1}{2^jW}\,i = \frac{2^jW+1}{2},\ \ j = 0,1,\ldots,T. \quad (14)$$

Подставив (11) и (14) в (13), вероятность передачи π_N можно записать в следующем виде:

$$\pi_N = \left[\frac{\theta W\left(1-2^{T+1}(1-\theta)^{T+1}\right)}{2\left(1-(1-\theta)^{T+1}\right)(2\theta-1)}+\frac{1}{2}\right]^{-1}. \quad (15)$$

Отметим, что формула для π_W имеет точно такой же вид.

Также заметим, что вероятности θ, p_c, π_N и π_W фактически являются функциями от числа сессий n_N и n_W, конкурирующих за возможность передачи. Таким образом, решая нелинейную систему (9), (12) и (15) для каждой пары значений n_N и n_W, полученные значения можно использовать для вычисления успешной передачи Π_N:

$$\Pi_N = \sum_{i=1}^{\infty}\frac{(\rho_N^\star)^i}{i!}e^{-\rho_N^\star}\sum_{j=0}^{\infty}\frac{(\rho_W^\star)^j}{j!}e^{-\rho_W^\star}\pi_N(i,j)\theta(i,j)\,,$$

где $\rho_N^\star = \lambda_U/\mu$; $\rho_W^\star = \lambda_W/\mu_W$.

Вероятность успешной передачи для ПУ WiGig рассчитывается аналогично. Поскольку скорость, достигаемая сессией со спектральной эффективностью m_j, определяется выражением $R_{U,j}^N = \Pi_N B_U m_j$, то средняя скорость, достигаемая NR-U UE в нелицензируемом диапазоне, составляет

$$E[R_U^N] = \sum_{j=0}^{R} p_{j,U}\Pi_N B_U m_j\,.$$

Скорость, достигаемая ПУ WiGig, получается аналогичным образом. Чтобы определить возможную вероятность потери сессии NR-U, определим Q_U как вероятность потери сессии NR-U, т. е. вероятность того, что минимальная скорость $R_{\min}$ не будет достигнута в нелицензируемой полосе частот. Таким образом,

$$Q_U = \mathsf{P}\{R < R_{\min}\} = \sum_{R_{U,j}^N < R_{\min}} p_{j,U}.$$

Наконец, вероятность потери сессии Q в системе в целом равна

$$Q = \frac{\lambda_1 + \lambda_2(1-\pi_U)}{\lambda}\pi_L + \frac{\lambda_2\pi_U}{\lambda}Q_U.$$

5 Численный пример

В данном разделе приводится численный пример расчета искомых характеристик. Исходные данные, использованные для получения графиков, представлены в таблице.

Графики на рис. 5 и 6 представляют вероятность коллизии как функцию от числа ПУ для нескольких значений плотности блокаторов λ_B (см. рис. 5) и начального значения окна конкурентного доступа W (см. рис. 6). Анализируя данные, представленные на рис. 5 для $W = 16$, отметим, что при увеличении числа ПУ вероятность коллизии повышается для всех рассматриваемых значений λ_B. Однако блокаторы оказывают сильное влияние на значения вероятности коллизии. В частности, исследуемые характеристики принимают наименьшие значения при $\lambda_B = 0{,}1$. При увеличении λ_B до 0,3 и далее до 0,5 наблюдается рост вероятности коллизии. Это происходит вследствие уменьшения вероятности

Параметры системы по умолчанию

Параметр	**Значение**
Рабочие частоты NR/WiGig, $f_{M,c}$	28/60 ГГц
Ширина полосы пропускания NR/WiGig, B_N, B_W	400 МГц / 2,16 ГГц
Высота БС NR-U, h_A	10 м
Радиус блокатора, r_B	0,2 м
Высота блокатора, h_B	1,7 м
Высота ПУ, h_U	1,5 м
Мощность передачи NR/WiGig, $P_{N,U}$	33/23 дБ·м
Тепловой шум, N_0	-174 дБ·м/Гц
Запас помехоустойчивости, M_I	3 дБ
Порог мощности приема сигнала, S_{th}	-9 дБ
Интенсивность блокаторов, λ_B	0,3
Антенна точки доступа/ПУ WiGig	$16\times4\,/\,8\times4$
Антенна БС/ПУ NR-U	$64\times4\,/\,8\times4$
Вероятность активной сессии ПУ NR-U, p_N	0,1
Вероятность активной сессии ПУ WiGig, p_W	0,1
Размер начального окна конкурентного доступа, W	16
Попытки повторной передачи с удвоением CW, T	10
Минимальная требуемая скорость в нелицензируемой полосе частот, $R_{\min}$	50 Мбит/с

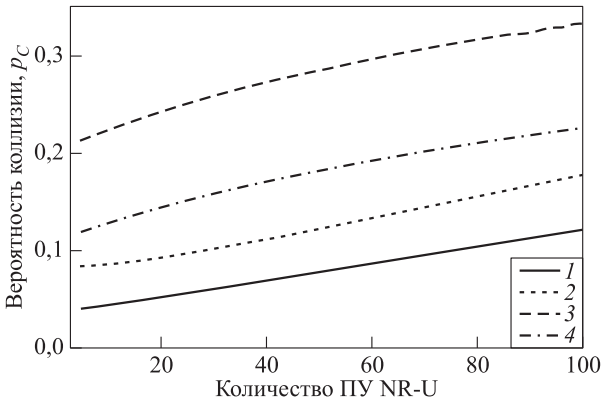


Рис. 5 Вероятность коллизии в нелицензируемом спектре, $W = 16$: *1* — $\lambda_B = 0{,}1$; *2* — 0,3; *3* — 0,5; *4* — $\lambda_B = 0{,}7$

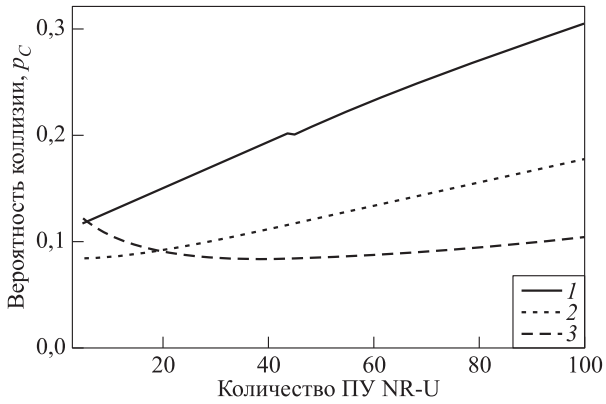


Рис. 6 Вероятность коллизии в нелицензируемом спектре, $\lambda_B = 0{,}3$: *1* — $W = 8$; *2* — 16; *3* — $W = 32$

успешной передачи и последующего накопления устройств, имеющих пакет, готовый к передаче. Однако дальнейшее увеличение плотности блокаторов приводит к снижению вероятности коллизии вследствие роста окон передачи на ПУ.

Начальный размер окна конкурентного доступа W также оказывает влияние на вероятности коллизии (см. рис. 6 для $\lambda_B = 0{,}3$). Анализируя представленные данные, следует отметить, что для малых значений начального размера окна конкурентного доступа наблюдается практически линейная зависимость вероятности коллизии. При больших значениях W зависимость становится более сложной, что объясняется совместным влиянием вероятности блокировки передачи и вероятности успешной передачи.

Перейдем теперь к рассмотрению системных метрик. На рис. 7 и 8 представлена вероятность успешной передачи как функция от плотности БС для разных значений плотности блокаторов λ_B (см. рис. 7) и размера окна конкурентного доступа W (см. рис. 8).

Анализируя представленные данные, можно отметить, что вероятность успешной передачи повышается для всех представленных кривых. Здесь основную роль играет уменьшение радиуса обслуживания соты и, как следствие, уменьшение нагрузки, что приводит к указанному положительному эффекту. Отметим также, что увеличение плотности блокаторов приводит к значительному снижению вероятности успешной передачи. Идентичный эффект наблюдается также и для размера окна W.

Рассмотрим среднюю скорость передачи в нелицензируемом спектре $E[R_U^N]$, а также вероятность потери сессии Q, представленные на рис. 9 и 10 как функции от плотности БС для разных значений нагрузки и размера окна конкурентного до-

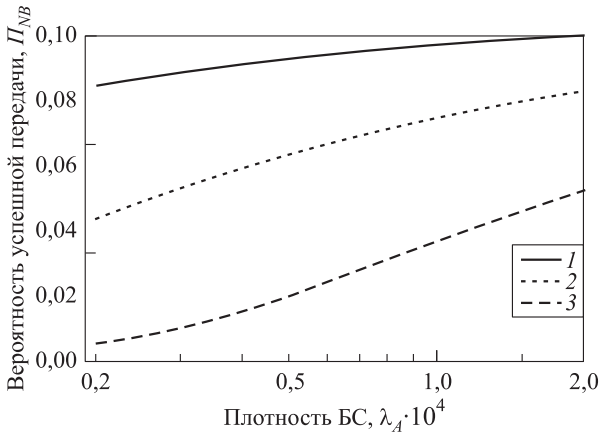


Рис. 7 Вероятность успешной передачи как функция от плотности БС: *1* — $\lambda_B = 0{,}1$; *2* — 0,3; *3* — $\lambda_B = 0{,}5$

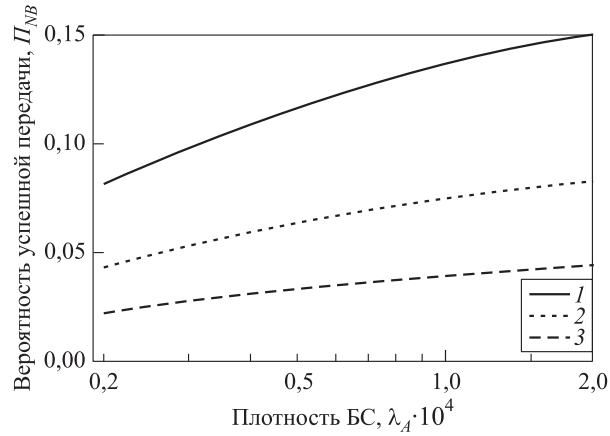


Рис. 8 Вероятность успешной передачи как функция от плотности БС: *1* — $W = 8$; *2* — 16; *3* — $W = 32$

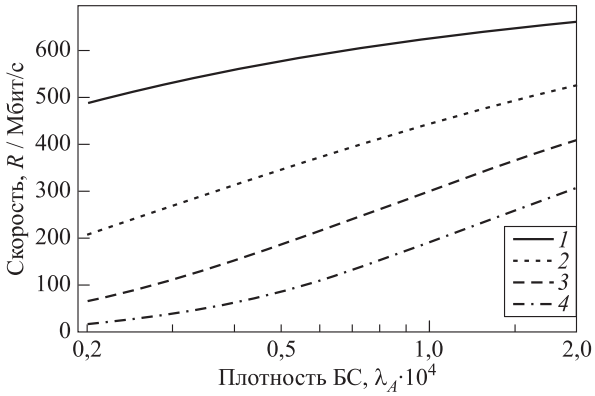


Рис. 9 Средняя скорость при $\lambda = 0{,}1$: *1* — $\lambda_B = 0{,}1$; *2* — 0,3; *3* — 0,5; *4* — $\lambda_B = 0{,}7$

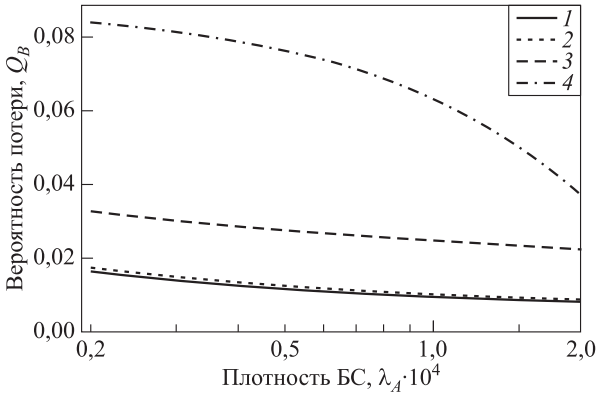


Рис. 10 Вероятность потери сессии при $\lambda_B = 0{,}3$: *1* — $\lambda = 10^{-3}$; *2* — 10^{-2}; *3* — 10^{-1}; *4* — $\lambda = 10^{0}$

ступа. Анализируя данные, представленные на рис. 9 для $W = 16$ и $\lambda_B = 0{,}3$, следует отметить, что повышение нагрузки приводит в увеличению вероятности потери сессии. Кроме того, вероятность потери сессии незначительно изменяется с ростом размера окна конкурентного доступа, хотя и скорость сессии в нелицензируемом спектре растет, как видно из рис. 10. Это связано с тем, что в рассматриваемой системе для большей части диапазона плотностей БС радиус покрытия WiGig значительно меньше, чем радиус покрытия технологии NR.

6 Заключение

В этой статье был проанализирован механизм выгрузки сессий в сетях 5G NR на основе технологии NR-U.

Используя вероятность потери сессии NR-U в качестве исследуемой метрики, авторы предложили метод оценки плотности развертывания БС NR-U, необходимой для поддержки заданной плотности ПУ NR с предписанными гарантиями QoS с точки зрения пропускной способности.

Показано, что на исследуемые характеристики помимо плотности ПУ NR и WiGig оказывает значительное влияние размер окна конкурентного доступа, плотность блокаторов и минимальная необходимая скорость. Представленные численные результаты позволяют сделать вывод о том, что рассматриваемый подход действительно дает возможность значительно увеличить достижимую скорость пользовательских сессий. Однако для этого требуется достаточно плотное развертывание технологии NR. Эти соображения необходимо учитывать при планировании развертывания сетей NR-U.

Литература

1. *Parkvall S., Dahlman E., Furuskar A., Frenne M.* NR: The New 5G Radio access technology // IEEE Communications Standards Magazine, 2017. Vol. 1. Iss. 4. P. 24–30.
2. *Wang T., Li G., Huang B., Miao Q., Fang J., Li P., Tan H., Li W., Ding J., Li J., Wang Y.* Spectrum analysis and regulations for 5G // 5G mobile communications. — Springer, 2017. P. 27–50.
3. 3GPP TR 38.101-1 v.16.7.0 — NR. User Equipment (UE) radio transmission and reception. Part 1: Range 1 Standalone (Release 16). https://www.3gpp.org/ftp/Specs/archive/38_series/38.101-1/38101-1-g70.zip.
4. IEEE Std. 802.11ad-2012. IEEE Standard for Information technology. Telecommunications and information exchange between systems. Local and metropolitan area networks. Specific requirements. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Amendment 3: Enhancements for very high throughput in the 60 GHz band. https://ieeexplore.ieee.org/document/6392842.
5. 3GPP TR 38.808 v.10.1.0: Evolved Universal Terrestrial Radio Access (E-UTRA). Carrier Aggregation. Base Station (BS) radio transmission and reception (Release 10). 2013. https://www.3gpp.org/ftp/Specs/archive/36_series/36.808/36808-a10.zip.
6. *Ali R., Shahin N., Musaddiq A., Kim B. S., Kim S. W.* Fair and efficient channel observation-based listen-before talk (CoLBT) for LAA-WiFi coexistence in unlicensed LTE // 10th Conference (International) on Ubiquitous and Future Networks. — Piscataway, NJ, USA: IEEE, 2018. P. 154–158.
7. *Nain P., Towsley D., Liu B., Liu Z.* Properties of random direction models // 24th Annual Joint Conference of the IEEE Computer and Communications Societies Proceedings. — Piscataway, NJ, USA: IEEE, 2005. P. 1897–1907.
8. *Gapeyenko M., Samuylov A., Gerasimenko M., Moltchanov D., Singh S., Aryafar E., Yeh S., Himayat N., Andreev S., Koucheryavy Y.* Analysis of human-body blockage in urban millimeter-wave cellular communications // IEEE

Conference (International) on Communications. — Piscataway, NJ, USA: IEEE, 2016. Art. 7511572. 7 p. doi: 10.1109/ICC.2016.7511572.

9. *Petrov V., Komarov M., Moltchanov D., Jornet J. M., Koucheryavy Y.* Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas // IEEE T. Wirel. Commun., 2017. Vol. 16. Iss. 3. P. 1791–1808.
10. *Kovalchukov R., Moltchanov D., Samuylov A., Ometov A., Andreev S., Koucheryavy Y., Samouylov K.* Evaluating SIR in 3D millimeter-wave deployments: Direct modeling and feasible approximations // IEEE T. Wirel. Commun., 2018. Vol. 18. Iss. 2. P. 879–896.
11. *Kovalchukov R., Moltchanov D., Samuylov A., Ometov A., Andreev S., Koucheryavy Y., Samouylov K.* Analyzing effects of directionality and random heights in drone-based mmWave communication // IEEE T. Veh. Technol., 2018. Vol. 67. Iss. 10. P. 10064–10069.
12. 3GPP TR 38.901. Study on channel model for frequencies from 0.5 to 100 GHz (Release 14). 2017. https://www.3gpp.org/ftp/Specs/archive/38_series/38.901/38901-e30.zip.
13. *Constantine A. B.* Antenna theory: Analysis and design. — 3rd ed. – Hoboken, NJ, USA: John Wiley & Sons, 2005. 1072 p.
14. *Tanemura M.* Statistical distributions of poisson voronoi cells in two and three dimensions // Forma-Tokyo, 2003. Vol. 18. Iss. 4. P. 221–247.
15. *Moltchanov D.* Distance distributions in random networks // Ad Hoc Networks, 2012. Vol. 10. Iss. 6. P. 1146–1166.
16. *Ross S. M.* Introduction to probability models. — 10th ed. — Academic Press, 2014. 784 p.
17. 3GPP TR 38.211 v.15.8.0. 5G. NR. Physical channels and modulation. 2018. https://www.3gpp.org/ftp/Specs/archive/38_series/38.211/38211-f80.zip.
18. *Samouylov K., Sopin E., Vikhrova O.* Analyzing blocking probability in LTE wireless network via queuing system with finite amount of resources // Information technologies and mathematical modelling — queueing theory and applications. — Communications in computer and information science ser. — Springer, 2015. Vol. 564. P. 393–403.
19. *Sopin E., Ageev K., Markova E., Vikhrova O., Gaidamaka Yu.* Performance analysis of M2M traffic in LTE network using queuing systems with random resource requirements // Autom. Control Comp. S., 2018. Vol. 52. Iss. 5. P. 345–353.
20. *Begishev V., Moltchanov D., Sopin E., Samuylov A., Andreev S., Koucheryavy Y., Samouylov K.* Quantifying the impact of guard capacity on session continuity in 3GPP New Radio systems // IEEE T. Veh. Technol., 2019. Vol. 68. Iss. 12. P. 12345–12359.
21. *Бегишев В. О., Сопин Э. С., Молчанов Д. А., Самуйлов А. К., Гайдамака Ю. В., Самуйлов К. Е.* Оценка эффективности механизма резервирования полосы пропускания для технологии mmWave в сетях связи пятого поколения // Информационно-управляющие системы, 2019. № 5. С. 51–63.
22. *Наумов В. А., Самуйлов К. Е., Самуйлов А. К.* О суммарном объеме ресурсов, занимаемых обслуживаемыми заявками // Автоматика и телемеханика, 2016. № 8. С. 125–132.

Поступила в редакцию 25.07.2021

ANALYSIS OF 5G NR BASE STATIONS OFFLOADING BY MEANS OF NR-U TECHNOLOGY

A. V. Daraseliya[1], E. S. Sopin[1,2], D. A. Moltchanov[1,3], and K. E. Samouylov[1,2]

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

[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation

[3] Tampere University, 7 Korkeakoulunkatu, Tampere 33720, Finland

Abstract: Fifth generation millimeter wave New Radio (NR) base stations (BS) are expected to be deployed in areas with extremely high and drastically fluctuating traffic demands resulting in frequent QoS (Quality of Service) violations in terms of provided rate at the access interface, especially, during busy hour conditions. As one of the measures to combat congestion, 3GPP (3rd Generation Partnership Project) considers the NR-U (NR-Unlicensed) technology, which allows to use the unlicensed frequency spectrum, for example, 60 GHz, on the BS in addition to the licensed spectrum. In this case, a session that cannot be served in the licensed spectrum due to lack of resources can be redirected to the unlicensed spectrum where competition for resources with WiGig technology subscribers takes place. The aim of this study is to evaluate the QoS (quality of service) parameters in an area characterized by a certain density of NR and WiGig users, where NR users can use NR-U technology if their rate requirements are met. The studied metric is the NR session loss probability and the achievable transmission rate in the unlicensed frequency spectrum. The performed numerical study shows that in addition to the density of NR and WiGig UE, the studied characteristics are influenced by the size of the contention window, the density of blockers, and

the minimum required rate. The presented numerical results allow one to conclude that the approach under consideration can significantly increase the attainable data rate of user sessions but this requires dense deployments of NR technology.

Keywords: NR-U; New Radio; WiGig; QoS; queuing theory; resource queuing system; Markov process

DOI: 10.14357/19922264210313

Acknowledgments

The publication was funded by the Russian Science Foundation, project No. 20-71-00124.

References

1. Parkvall, S., E. Dahlman, A. Furuskar, and M. Frenne. 2017. NR: The New 5G Radio access technology. *IEEE Communications Standards Magazine* 1(4):24–30.
2. Wang, T., G. Li, B. Huang, Q. Miao, J. Fang, P. Li, H. Tan, W. Li, J. Ding, J. Li, and Y. Wang. 2017. Spectrum analysis and regulations for 5G. *5G mobile communications*. Springer. 27–50.
3. TR 38.101-1. 2021. NR. User Equipment (UE) radio transmission and reception. Part 1: Range 1 Standalone (Release 16). Available at: https://www.3gpp.org/ftp/Specs/archive/38_series/38.101-1/38101-1-g70.zip (accessed August 12, 2021).
4. IEEE Std 802.11ad-2012. 2012. IEEE Standard for information technology. Telecommunications and information exchange between systems. Local and metropolitan area networks. Specific requirements. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Amendment 3: Enhancements for very high throughput in the 60 GHz band. Available at: https://ieeexplore.ieee.org/document/6392842 (accessed August 12, 2021).
5. 3GPP TR 38.808. 2013. Evolved Universal Terrestrial Radio Access (E-UTRA). Carrier Aggregation. Base Station (BS) radio transmission and reception (Release 10). Available at: https://www.3gpp.org/ftp/Specs/archive/36_series/36.808/36808-a10.zip (accessed August 12, 2021).
6. Ali, R., N. Shahin, A. Musaddiq, B. S. Kim, and S. W. Kim. 2018. Fair and efficient channel observation-based listen-before talk (CoLBT) for LAA-WiFi coexistence in unlicensed LTE. *10th Conference (International) on Ubiquitous and Future Networks Proceedings*. Piscataway, NJ: IEEE. 154–158.
7. Nain, P., D. Towsley, B. Liu, and Z. Liu. 2005. Properties of random direction models. *24th Annual Joint Conference of the IEEE Computer and Communications Societies Proceedings*. Piscataway, NJ: IEEE. 1897–1907.
8. Gapeyenko, M., A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, E. Aryafar, S. Yeh, N. Himayat, S. Andreev, and Y. Koucheryavy. 2016. Analysis of human-body blockage in urban millimeter-wave cellular communications. *IEEE Conference (International) on Communications*. Piscataway, NJ: IEEE. Art. 7511572. 7 p. doi: 10.1109/ICC.2016.7511572.
9. Petrov, V., M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy. 2017. Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas. *IEEE T. Wirel. Commun.* 16(3):1791–1808.
10. Kovalchukov, R., D. Moltchanov, A. Samuylov, A. Ometov, S. Andreev, Y. Koucheryavy, and K. Samouylov. 2018. Evaluating SIR in 3D millimeter-wave deployments: Direct modeling and feasible approximations. *IEEE T. Wirel. Commun.* 18(2):879–896.
11. Kovalchukov, R., D. Moltchanov, A. Samuylov, A. Ometov, S. Andreev, Y. Koucheryavy, and K. Samouylov. 2018. Analyzing effects of directionality and random heights in drone-based mmWave communication. *IEEE T. Veh. Technol.* 67(10):10064–10069.
12. 3GPP TR 38.901. 2017. Study on channel model for frequencies from 0.5 to 100 GHz (Release 14). Available at: https://www.3gpp.org/ftp/Specs/archive/38_series/38.901/38901-e30.zip (accessed August 12, 2021).
13. Constantine, A. B. 2005. *Antenna theory: Analysis and design*. 3rd ed. Hoboken, NJ: John Wiley & Sons. 1072 p.
14. Tanemura, M. 2003. Statistical distributions of poisson voronoi cells in two and three dimensions. *Forma-Tokyo* 18(4):221–247.
15. Moltchanov, D. 2012. Distance distributions in random networks. *Ad Hoc Networks* 10(6):1146–1166.
16. Ross, S. M. 2014. *Introduction to probability models*. 10th ed. Academic Press. 784 p.
17. 3GPP TR 38.211 v.15.8.0. 2018. 5G. NR. Physical channels and modulation. Available at: https://www.etsi.org/deliver/etsi_ts/138200_138299/138211/15.02.00_60/ts_138211v150200p.pdf (accessed August 12, 2021).
18. Samouylov, K., E. Sopin, and O. Vikhrova. 2015. Analyzing blocking probability in LTE wireless network via queuing system with finite amount of resources. *Information technologies and mathematical modelling — queueing theory and applications*. Communications in computer and information science ser. Springer. 564:393–403.
19. Sopin, E., K. Ageev, E. Markova, O. Vikhrova, and Yu. Gaidamaka. 2018. Performance analysis of M2M traffic in LTE network using queuing systems with random resource requirements. *Autom. Control Comp. S.* 52(5):345–353.

20. Begishev, V., D. Moltchanov, E. Sopin, A. Samuylov, S. Andreev, Y. Koucheryavy, and K. Samouylov. 2019. Quantifying the impact of guard capacity on session continuity in 3GPP New Radio systems. *IEEE T. Veh. Technol.* 68(12):12345–12359.
21. Begishev, V., E. Sopin, D. Molchanov, A. Samuylov, Yu. Gaydamaka, and K. Samuylov. 2019. Otsenka effektivnosti mekhanizma rezervirovaniya polosy propuskaniya dlya tekhnologii mmWave v setyakh svyazi pyatogo pokoleniya [Performance evaluation of the bandwidth reservation mechanism for mmWave technology in 5G networks]. *Informatsionno-upravlyayushchie sistemy* [Information and Control Systems] 5:51–63.
22. Naumov, V. A., K. E. Samuilov, and A. K. Samuilov. 2016. On the total amount of resources occupied by serviced customers. *Automat. Rem. Contr.* 77(8):1419–1427.

Received July 25, 2021

Contributors

Daraseliya Anastasia V. (b. 1994) — PhD student, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; avdaraseliya@sci.pfu.edu.ru

Sopin Eduard S. (b. 1987) — Candidate of Science (PhD) in physics and mathematics, associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, 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 119133, Russian Federation; sopin-es@rudn.ru

Moltchanov Dmitry A. (b. 1978) — Doctor of Science in technology, lecturer, Tampere University, 7 Korkeakoulunkatu, Tampere 33720, Finland; associate professor, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, Russian Federation; dmitri.moltchanov@tuni.fi

Samouylov Konstantin E. (b. 1955) — Doctor of Science in technology, professor, Head of Department, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya Str., Moscow 117198, 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 119133, Russian Federation; samuylov-ke@rudn.university

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; профессор Московского авиационного института; старший научный сотрудник Московского центра фундаментальной и прикладной математики Московского государственного университета имени М. В. Ломоносова

Босов Алексей Вячеславович (р. 1969) — доктор технических наук, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Власкина Анастасия Сергеевна (р. 1995) — аспирант кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов

Ву Ньят Нам (р. 1997) — студент кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов

Горшенин Андрей Константинович (р. 1986) — доктор физико-математических наук, доцент, руководитель отдела и ведущий сотрудник Федерального исследовательского цента «Информатика и управление» Российской академии наук; доцент факультета вычислительной математики и кибернетики Московского государственного университета имени 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

Грушо Николай Александрович (р. 1982) — кандидат физико-математических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Дараселия Анастасия Валерьевна (р. 1994) — аспирант кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов

Жуков Денис Владимирович (р. 1979) — главный специалист Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

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

Зацман Игорь Моисеевич (р. 1952) — доктор технических наук, заведующий отделом Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Казанчян Драстамат Хачатурович (р. 1994) — аспирант кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова

Кириков Игорь Александрович (р. 1955) — кандидат технических наук, директор Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Коновалов Михаил Григорьевич (р. 1950) — доктор технических наук, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Кочеткова Ирина Андреевна (р. 1985) — кандидат физико-математических наук, доцент Российского университета дружбы народов; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Кудрявцев Алексей Андреевич (р. 1978) — кандидат физико-математических наук, доцент кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; старший научный сотрудник Московского центра фундаментальной и прикладной математики Московского государственного университета имени М. В. Ломоносова

Кузьмин Виктор Юрьевич (р. 1986) — программист факультета космических исследований Московского государственного университета имени М. В. Ломоносова

Малашенко Юрий Евгеньевич (р. 1946) — доктор физико-математических наук, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Молчанов Дмитрий Александрович (р. 1978) — доктор технических наук, лектор Университета Тампере (Финляндия, Тампере); доцент Российского университета дружбы народов

Разумчик Ростислав Валерьевич (р. 1984) — кандидат физико-математических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Румовская София Борисовна (р. 1985) — кандидат технических наук, научный сотрудник Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Самуйлов Константин Евгеньевич (р. 1955) — доктор технических наук, профессор, заведующий кафедрой прикладной информатики и теории вероятностей Российского университета дружбы народов; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Сопин Эдуард Сергеевич (р. 1987) — кандидат физико-математических наук, доцент Российского университета дружбы народов; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Сушко Дмитрий Викторович (р. 1962) — кандидат физико-математических наук, старший 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

Шестаков Олег Владимирович (р. 1976) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; ведущий научный сотрудник Московского центра фундаментальной и прикладной математики Московского государственного университета имени М. В. Ломоносова

Шоргин Всеволод Сергеевич (р. 1978) — кандидат технических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Шоргин Сергей Яковлевич (р. 1952) — доктор физико-математических наук, профессор, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

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/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. *Russ. 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. — New York, NY, USA: Wiley, 1974. 521 p.)
2. В англоязычной части:
   Timoshenko, S. P., D. H. Young, and W. Weaver. 1974. *Vibration problems in engineering*. 4th ed. New York: Wiley. 521 p.

Описание неопубликованного документа:

Latypov, A. R., M. M. Khasanov, and V. A. Baikov. 2004 (unpubl.). Geologiya i dobycha (NGT GiD) [Geology and production (NGT GiD)]. Certificate on official registration of the computer program No. 2004611198.

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]. Moscow. D.Sc. Diss. 272 p.

Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. Moscow: IPI RAN. PhD Thesis. 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: iiep@frccsc.ru (Стригина Светлана Николаевна)
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. *Russ. 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]. Moscow: IPI RAN. PhD Thesis. 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: iiep@frccsc.ru (to Svetlana Strigina)
http://www.ipiran.ru/english/journal.asp