Информатика и её применения

Том 4 Выпуск 2 Год 2010

Тематический выпуск

Вероятностно-статистические методы и задачи информатики и информационных технологий

СОДЕРЖАНИЕ

Выпускающий редактор *Т. Торжкова;* Технический редактор *Л. Кокушкина;* Художественный редактор *М. Седакова*
Сдано в набор 01.04.10. Подписано в печать 14.04.10. Формат 60 x 84 / 8
Бумага офсетная. Печать офсетная. Усл.-печ. л. 12,25. Уч.-изд. л. 10. Тираж 200 экз.
Заказ №
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
torus@torus-press.ru; http://www.torus-press.ru
Отпечатано в ППП «Типография «Наука» с готовых диапозитивов, Москва 121099, Шубинский пер., д. 6.

Предисловие

Вниманию читателей журнала «Информатика и её применения» предлагается очередной тематический выпуск «Вероятностно-статистические методы и задачи информатики и информационных технологий». Предыдущие тематические выпуски журнала по данному направлению вышли в 2008 г. (том 2, вып. 2) и в 2009 г. (том 3, вып. 3).

Статьи, собранные в данном журнале, посвящены разработке новых вероятностно-статистических методов, ориентированных на решение конкретных задач информатики и информационных технологий, а также — в ряде случаев — и других прикладных задач. Проблематика, охватываемая публикуемыми работами, развивается в рамках научного сотрудничества между Институтом проблем информатики Российской академии наук (ИПИ РАН) и Факультетом вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова в ходе работ над совместными научными проектами (в том числе в рамках функционирования научно-образовательного центра ИПИ РАН – ВМК МГУ «Вероятностно-статистические методы анализа рисков»). Многие авторы статей данного выпуска журнала являются активными участниками традиционного международного семинара по проблемам устойчивости стохастических моделей, руководимого В. М. Золотарёвым и В. Ю. Королевым; регулярные сессии этого семинара проводятся под эгидой МГУ и ИПИ РАН.

Наряду с представителями ИПИ РАН и МГУ, в число авторов входят ученые из ВЦ РАН им. А. А. Дородницына, Института прикладных математических исследований Карельского НЦ РАН, Московского физико-технического института, МГТУ «Станкин», Тамбовского государственного университета им. Г. Р. Державина, Российского отделения Института микроэлементов ЮНЕСКО, Helsinki Institute for Information Technology (Хельсинки, Финляндия).

В данном выпуске традиционно присутствует тематика, весьма активно разрабатываемая в течение многих лет специалистами ИПИ РАН и МГУ, — методы моделирования и управления для информационно-телекоммуникационных и вычислительных систем. В статье М. Г. Коновалова рассматривается проблема анализа и оптимизации распределения потоков заданий и ценообразования в системах коллективного использования распределенных вычислительных ресурсов. Предложен подход к построению математических моделей систем вычислительных ресурсов, основанный на укрупненном описании потоков заданий в виде динамических балансовых соотношений. В статье А. С. Лукьяненко и Е. В. Морозова исследован класс сетевых протоколов контроля несущей среды, где окно передачи сообщения является произвольной возрастающей функцией числа конфликтов сообщения, посланного с данной станции.

Ряд статей выпуска посвящен разработке и применению стохастических методов и информационных технологий для решения различных прикладных задач. В работе М. П. Кривенко рассматривается задача построения эмпирического байесовского классификатора, обеспечивающего распознавание текста, когда отдельные знаки имеют различные размеры. Представлен комбинированный метод построения оценки элементов байесовского классификатора, включающий непараметрическую ядерную оценку и параметрическую оценку с помощью плотности нормального распределения. В статье К. В. Рудакова и И. Ю. Торшина разработан формализм для применения алгебраического подхода к проблеме распознавания вторичной структуры белка. Этот формализм позволил сформулировать математическое описание принятой у биологов гипотезы о локальном характере зависимости вторичной структуры от первичной и получить конструктивные критерии разрешимости задачи. Работа А. В. Маркина и О. В. Шестакова посвящена решению задачи реконструкции изображения по радоновскому образу с помощью вейвлет-вейглет разложения. Статьи О. В. Крючина и С. Ю. Степанова представляют результаты в области нейронных сетей. В работе Е. Б. Козеренко рассматриваются задачи создания лингвистических фильтров в статистических моделях машинного перевода и совершенствования механизмов выравнивания параллельных текстов для повышения точности и адекватности переводов.

Две статьи посвящены развитию перспективных теоретических вероятностно-статистических методов, которые могут найти широкое применение в различных задачах информатики и информационных технологий. В работе В. Е. Бенинга и Р. А. Королева рассмотрены вопросы анализа статистических критериев проверки гипотез о параметрах обобщенного распределения Лапласа, которое применяется при математическом моделировании многих процессов в телекоммуникационных системах, в экономике, финансовом деле, технике и других областях. Работа И. Г. Шевцовой посвящена уточнению оценок абсолютной константы в оценке точности нормальной аппроксимации (такая аппроксимация весьма важна при математическом моделировании и анализе многих реальных систем, в том числе информационных и телекоммуникационных).

Редакционная коллегия журнала выражает надежду, что данный тематический выпуск будет интересен специалистам в области теории вероятностей и математической статистики и их применения к решению задач информатики и информационных технологий.

Заместитель главного редактора журнала «Информатика и её применения»,
директор ИПИ РАН, академик *И. А. Соколов*

Редактор-составитель тематического выпуска, профессор кафедры математической статистики
факультета вычислительной математики и кибернетики МГУ им. М. В. Ломоносова,
ведущий научный сотрудник ИПИ РАН, доктор физико-математических наук *В. Ю. Королев*

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 3–12

О ПЛАНИРОВАНИИ ПОТОКОВ В СИСТЕМАХ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ*

М. Г. Коновалов[1]

Аннотация: Рассмотрена проблема анализа и оптимизации распределения потоков заданий и ценообразования в системах коллективного использования распределенных вычислительных ресурсов. Проведен обзор литературных источников. Предложен подход к построению математических моделей систем вычислительных ресурсов, основанный на укрупненном описании потоков заданий в виде балансовых соотношений и использовании функций качества. Участники системы представляются как субъекты, обладающие собственными стратегиями поведения и преследующие индивидуальные цели, сформулированные в терминах качества обслуживания и стоимости. В качестве варианта стратегии участников рассматривается распределенный децентрализованный алгоритм градиентного типа. Приведен численный пример и обсуждены перспективы развития и использования модели.

Ключевые слова: системы вычислительных ресурсов; распределение потоков; качество обслуживания; коллективное поведение

1 Введение

В современном мире к ряду наиболее существенных для человечества ресурсов, таких как вода, нефть и т. п., добавился еще один, хотя и искусственный, но жизненно важный ресурс, который можно назвать вычислительным (или, более широко, информационно-вычислительным). Его запасы, в отличие от естественных ресурсов, увеличиваются. «Можно констатировать, что экспоненциальный характер роста вычислительных мощностей. . . и систем хранения информации сохранится еще на многие годы. . . » [1].

Несмотря на то, что количество компьютеров увеличивается, а их мощность продолжает расти, эффективность использования вычислительной техники, по общераспространенному мнению, отстает от этого процесса, оставаясь невысокой. Имеет место ситуация, при которой сопряженное с огромными материальными и интеллектуальными затратами наращивание вычислительных ресурсов (ВР) не дает должной отдачи, что, учитывая подверженность данного специфического ресурса моральному старению, делает проблему особенно острой.

Вполне естественная идея увеличить действенность имеющегося и обновляющегося парка компьютеров за счет обеспечения более широкого, стандартизованного и облегченного доступа нашла свою реализацию в виде систем коллективного использования распределенных ВР, в первую очередь гридов [2]. Данная статья касается одного из общих аспектов, связанных с разработкой и эксплуатацией систем коллективного использования ВР, и посвящена проблематике выбора принципов и алгоритмов распределения заданий между имеющимися ресурсами.

Всякая система коллективного использования ВР предполагает наличие двух типов составляющих, участвующих в ее работе: потребителей ресурсов и собственно ресурсов, т. е. вычислительной техники. С узкотехнической точки зрения потребители ассоциируются с источниками заданий. Последние могут и должны быть выполнены на том или ином ресурсе, они имеют определенную трудоемкость, сопряжены с передачей определенного объема информации, а также обладают рядом других показателей, включая требования к качеству обслуживания. Точно так же ресурс в узком смысле является той или иной разновидностью компьютера (персональный компьютер, кластер, суперкомпьютер и т. д.) или хранилища данных и обладает определенными техническими параметрами: производительностью, емкостью памяти, программным обеспечением и пр. Таким образом, проблема планирования потоков заданий в узком смысле может пониматься как составление расписания, предписывающего место и очередность выполнения того или иного задания и увязывающего рабочие характеристики исполняемых заданий и используемых ресурсов.

В то же время в современных больших вычислительных системах, за счет их масштаба, разнородно-

* Работа выполнена при поддержке РФФИ, гранты 09-07-12032-офи_м, 08-07-00152-а.

[1] Институт проблем информатики Российской академии наук, mkonovalov@ipiran.ru

сти оборудования, различий в административном подчинении и других особенностей, при планировании появляются качественно новые моменты, не учитываемые в классической теории расписаний и связанные с взаимозависимостью отдельных участников системы.

К примеру, приходится принимать во внимание, что потребители не только порождают потоки заданий, но и преследуют при их выполнении довольно сложные цели, порожденные соображениями финансового, или приоритетного, или секретного и т. д. характера. К тому же потребители часто объединяются в группы или сообщества, оставаясь при этом в определенной мере самостоятельными, и к тому же разделенными географически.

Аналогично за ресурсом как техническим устройством обычно стоит еще и владелец ресурса со своими собственными интересами. Приходится констатировать, что отдельные составные части распределенной системы ВР обладают собственным поведением, целенаправленность которого, вообще говоря, не совпадает с приоритетами системы в целом. Проблема планирования в такой системе в широком смысле не может быть сведена к составлению расписания и созданию соответствующего связующего программного обеспечения. Необходимо концептуальное понимание принципов и наличие методов, позволяющих организовать поведение участников системы и учитывающих как собственные интересы участников, так и цели, стоящие перед всей системой.

Термин *система вычислительных ресурсов* не относится к числу точных или даже вызывающих однозначные ассоциации. По-видимому, безусловно общими для всех систем ВР являются следующие признаки:

(1) наличие собственно ВР как технических устройств (процессоров и структур, составленных из процессоров, носителей информации);

(2) наличие источников заданий, выполняемых с помощью этих устройств;

(3) объединение поименованных в первых двух пунктах элементов в систему, позволяющее различным источникам заданий обращаться к разным ресурсам, обмениваться информацией между элементами, осуществлять более или менее согласованную политику функционирования составных частей.

Перечисленным признакам удовлетворяют самые различные и по масштабу, и по целям системы ВР. Основным «примером» для данной работы можно считать системы грид с их, как правило, глобальным масштабом, гетерогенным составом и автономностью поведения подсистем. Однако было бы неверно ограничивать статью только областью проблематики грида. Например, в крупных и не очень крупных компаниях повсеместно приходят к мысли о необходимости организации более эффективного менеджмента в области использования ВР. «Территориально-распределенное предприятие — более 1700 км магистральных трубопроводов, которые протянулись от Полярного круга до юга Тюменской области, 17 компрессорных станций, один из крупнейших в России завод стабилизации газового конденсата. . . Сбой в любом звене может привести к достаточно серьезным последствиям. Поэтому года три назад. . . была осознана необходимость создания системы управления информационно-вычислительными ресурсами». (Это цитата 2010 г. из публикации о компании ООО «Сургутгазпром» [3].)

Данная работа направлена на изучение таких систем ВР, в которых составные части обладают определенной самостоятельностью в выборе критериев деятельности и стратегий управления потоками заданий и распределения ресурсов. Другая особенность постановки задачи заключается в том, что отдельные субъекты, составляющие систему, многофункциональны. Они могут одновременно являться как источниками заданий, так и представлять запас вычислительных ресурсов или выполнять посреднические функции.

2 Краткий обзор литературы

За последние годы опубликовано много работ по планированию в системах ВР, прежде всего, в системах грид. Это отражено в обзорах, посвященных данной тематике [4, 5]. Публикации самого последнего времени, которые упоминаются ниже, представляют большое разнообразие постановок задач, методов и приложений.

Проблема управления ресурсами в системах ВР может трактоваться как имеющая две основные составляющие. Первая — это нахождение алгоритмов, которые бы эффективно обслуживали конкретные задания, направляя их на конкретные ресурсы. Эта задача в целом находится в области классических оптимизационных задач (за рубежом для ее обозначения повсеместно используется термин scheduling), хотя и осложнена большой размерностью и обилием специфических особенностей. Вторая — связана с уже упоминавшейся автономностью субъектов, обладающих собственными целями и стратегиями поведения. Реализация стратегии, оптимизирующей глобальную для всей системы целевую функцию (даже если такая

стратегия будет найдена), натолкнется на сопротивление отдельных элементов системы, если не будут учтены их локальные интересы.

Работы первого направления продолжают широко использовать классические статические оптимизационные постановки задач с «интегральным» описанием потоков заданий [6] и применением традиционных потоковых алгоритмов [7]. По-прежнему популярна *задача о рюкзаке*, которая применительно к проблеме планирования ресурсов дополняется использованием функций полезности и метрик качества обслуживания [8], стоимостными соображениями [9], а также эвристическими методами [10].

Во многих работах, однако, рассматривается «штучная» обработка заданий. В этих случаях описание моделей осуществляется в терминах случайных процессов, а для оптимизации часто употребляется марковский процесс принятия решений [11, 12]. В [5] предложена модель, в которой для оперативного управления коллективным доступом к распределенным вычислительным ресурсам используются алгоритмы, основанные на адаптивном варианте теории управления марковскими цепями.

Поскольку точные методы решения классических оптимизационных задач часто неэффективны, то широкую популярность приобрели эвристические алгоритмы [13–15]. Не являясь сугубой принадлежностью планирования потоков заданий, они фигурируют за рубежом под экзотическими названиями (artificial fish swarm algorithm, genetic algorithm, simulated annealing и пр.)

Некоторые работы используют менее распространенные подходы.

Так, в [16] предлагается механизм планирования заданий в гриде на основе предварительного резервирования в режиме онлайн. В [17] описан алгоритм диспетчеризации заданий также в режиме реального времени, но на основе балансировки нагрузки путем прогнозирования производительности серверов. Алгоритмы балансировки нагрузки разработаны [18, 19], а статистическое прогнозирование явилось исходной посылкой для создания стратегии планирования в [20]. В [21] для распределения заданий в вычислительном гриде используются нечеткие множества и нейронные сети. В [22] та же задача решается с применением техники так называемых *сложных сетей* (complex network). В [23] обсуждаются специфические проблемы планирования, возникающие при обслуживании заданий разного типа: ориентированных на вычисления и связанных преимущественно с передачей данных. В [24] описана необычная постановка задачи распределения мобильных ресурсов.

Отмеченное выше второе направление исследований, связанное с автономностью поведения участников системы ВР, представлено значительно беднее. Работы, которые можно было бы отнести к этому направлению, часто отражают в большей степени гетерогенность составных блоков системы, нежели их самостоятельность.

В [25] предлагается модель размещения заданий, параллельно выполняемых в автономных доменах. Другой подход к организации параллельных вычислений в гетерогенной среде основан на привлечении так называемых систем с агентами [26].

В [27] рассмотрена технология организации планирования в гриде, составные части которого основаны на разных стандартах, что затрудняет коллективное использование ресурсов. Рассмотренное в [27] понятие *федерации гридов* (grid-federation) широко разрабатывается авторами статьи [28], в которой дана схема кооперированного использования ресурсов в гриде, основанная на концепции соглашений об уровне обслуживания (service level agreement).

Стратегии *согласований* (negotiations) рассмотрены в [29].

Большие надежды в организации менеджмента в системах ВР возлагаются на экономический подход, побуждающий, как принято считать, разнородные и конкурирующие элементы действовать, соблюдая интересы системы в целом. Эта часть публикаций заслуживает отдельного обзора (в некоторой степени он проведен в [4]). Здесь упомянем в качестве примера сравнительный анализ менеджмента в гриде, основанный на различных механизмах рыночной экономики [30], а также две модели планирования ресурсов, основанные на оценке их стоимости [31, 32].

Проведенный беглый обзор показывает, прежде всего, что проблема управления потоками заданий в системах вычислительных ресурсов находится в стадии интенсивного изучения. В настоящее время нельзя говорить о том, что для ее решения сформирован в какой-то степени окончательный круг теоретических положений и практических методов. Существующие разработки отличаются большим многообразием используемых подходов и средств.

В данной работе сделана попытка дать простое математическое описание распределенной системы ВР, которое, не отражая частных деталей, годилось бы для изучения общих вопросов управления потоками заданий в широком классе таких систем. Избранный способ моделирования продолжает подход, начатый в [33].

3 Моделирование процесса распределения потоков в системе вычислительных ресурсов

3.1 Балансовое соотношение для потоков заданий

Рассмотрим модель распределенной системы ВР, составные элементы которой будем называть субъектами (или иногда, для разнообразия, участниками). Всего система содержит N субъектов, пронумерованных от 1 до N. Каждый субъект способен осуществлять, вообще говоря, троякую функцию:

- являться источником потоков заданий;
- выполнять задания на имеющихся у него ресурсах;
- являться транзитным и коммутационным пунктом для перемещения заданий между субъектами.

Таким образом, в разрабатываемой модели каждый субъект может одновременно выступать в роли потребителя ресурсов, владельца ресурса и посредника.

Субъекты взаимодействуют через коммуникационную сеть, которая не выделяется как самостоятельный элемент системы, но ее наличие будет подтверждено косвенно при дальнейшем описании.

Условимся измерять объемы заданий (нагрузку) в некоторых условных единицах, физический смысл которых в данном случае не играет особой роли. (Можно, например, представлять себе, что единица объема задания измеряется временем его выполнения на стандартном процессоре.) Поток заданий, порождаемый субъектом i, представляет собой случайный процесс $Z_i(t)$, заданный, как и все функционирование системы, в дискретном времени, $t = 0, 1, \ldots$ Этот процесс определяет объем нагрузки, поступающей в систему извне на каждом такте времени через посредство субъекта i.

Помимо экзогенного потока заданий $Z_i(t)$, субъект получает входные потоки от других участников. Объем заданий, который накапливается у субъекта i в момент t, обозначается через $X_i(t)$. Этим объемом заданий (будем называть его для краткости очередью в момент t) субъект обязан распорядиться в момент t, имея следующие возможности:

- выполнение заданий на собственном ресурсе;
- отправка заданий другим субъектам;
- оставление заданий в собственной очереди;
- уничтожение заданий (потери).

Чтобы описать реализацию указанных возможностей, определим вектор

$$\alpha_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{iN}) \,,$$

компоненты которого удовлетворяют условиям

$$\sum_{i=0}^{N} \alpha_{ij} = 1 \,, \quad 0 \le \alpha_{ij} \le 1 \,, \;\; 0 \le i, j \le N \,,$$

и означают следующее: α_{i0} — доля объема заданий, направляемая для выполнения на собственный ресурс; α_{ij}, $i \ne j$ — доля объема заданий, направляемая субъекту j; α_{ii} — доля объема заданий, оставляемая в собственной очереди (часть этих заданий может быть уничтожена).

Вектор α_i определяет способ, которым субъект i распоряжается тем объемом заданий, который у него уже имеется. В частности, параметры α_{ij}, $i \ne j$, определяют объем *заявки* на передачу части заданий на адрес субъекта j. Однако субъект j, в соответствии с обсуждавшимся выше предположением о самостоятельности поведения участников системы, имеет возможность отказаться от предложения. Для регулирования объема вновь поступающих заданий необходим дополнительный механизм. Он задается вектором

$$\beta_i = (\beta_{i1}, \ldots, \beta_{iN}) \,,$$

компоненты которого подчиняются условиям $0 \le \le \beta_{ij} \le 1$ и при $i \ne j$ означают долю от предлагаемого субъектом j объема заданий, которую субъект i согласен принять. Параметру β_{ij} придадим смысл регулирования объема заданий, сознательно удаляемых из системы субъектом i.

Согласно сказанному, в каждый момент t при передаче заданий от субъекта i к субъекту j происходит следующего рода согласование. Субъект i делает заявку на передачу заданий в объеме $\alpha_{ij} X_i(t)$. Субъект j подтверждает согласие на передачу части этого объема, и фактически передаваемый объем заданий составляет $\alpha_{ij}\beta_{ji}X_i(t)$.

Таким образом, стратегии участников системы описываются с помощью матриц α и β с компонентами α_{ij} и β_{ij}, имеющих размерности соответственно $N \times (N+1)$ и $N \times N$. О матрице α будем говорить как о стратегии *маршрутизации*, а матрицу β будем называть стратегией *согласования*.

Заметим, что стратегии поведения субъекта, описываемые с помощью матриц α и β, могут быть сколь угодно сложными, поскольку компоненты этих матриц могут зависеть в общем случае от момента времени t и даже от всей предыстории системы до этого момента.

Прежде чем описать динамику потоков в системе, необходимо еще оговорить, что происходит с теми заданиями, которые были заявлены на передачу другим субъектам, но не были согласованы. В принципе, это вопрос описания, а не существа дела, и он может быть разрешен по-разному. Договоримся в данном случае, что все несогласованные задания автоматически отправляются для выполнения на собственный ресурс.

Введем в рассмотрение матрицу λ размерностью $N \times N$ с компонентами

$$\lambda_{ij} = \alpha_{ij}\beta_{ji}\,, \quad 1 \le i, j \le N\,.$$

Объем заданий у субъекта i в момент $t+1$ складывается из экзогенного потока, а также из потоков, поступающих от всех участников системы, поэтому

$$X_i(t+1) = Z_i(t+1) + \sum_{j=1}^{N} \lambda_{ij} X_j(t)$$

или в матричной форме

$$X(t+1) = Z(t) + \lambda X(t)\,, \qquad (1)$$

где $X(t)$ и $Z(t)$ — векторы-строки с компонентами $X_i(t)$ и $Z_i(t)$.

Объем заданий, направленный для выполнения на собственный ресурс, с учетом сделанной договоренности составляет для субъекта i величину

$$Y_i(t) = \lambda_{i0} X_i(t)\,,$$

где

$$\lambda_{i0} = \sum_{j=0}^{N} \alpha_{ij} - \sum_{j \ne i} \lambda_{ij} = 1 - \alpha_{ii} - \sum_{j=1}^{N} \lambda_{ij}\,.$$

Потери заданий выражаются как

$$\alpha_{ii}(1-\beta_{ii})X_i(t) = \gamma_i X_i(t)\,. \qquad (2)$$

Величину $\gamma_i = \alpha_{ii}(1-\beta_{ii})$ — долю удаляемых заданий — назовем *коэффициентом потерь*.

В последующих рассуждениях будем предполагать, что экзогенные потоки имеют не зависящие от времени средние $z_i = \mathrm{M}Z_i(t)$.

Пусть элементы матрицы λ являются постоянными величинами и при этом для всех i выполняются соотношения $\sum\limits_{j=1}^{N} \lambda_{ij} < 1$. Тогда из (1) следует, что существуют пределы $x_i = \lim\limits_{t\to\infty} \mathrm{M}X(t)$, которые определяются соотношением

$$x = z + x\lambda \qquad (3)$$

или

$$x = z(E-\lambda)^{-1}\,,$$

где $x = (x_1, x_2, \ldots, x_N)$ и $z = (z_1, z_2, \ldots, z_N)$, а E — единичная матрица. Дальнейшие соотношения также будут относиться к предельным средним значениям переменных, участвующих в модели.

3.2 Качество обслуживания

Качество обслуживания в системе вычислительных ресурсов определяется целым рядом факторов, которые зачастую плохо поддаются количественному описанию. Тем не менее одно из самых существенных предположений, которое делается в данной статье, заключается именно в том, что количественное описание качества обслуживания имеется.

Говоря неформально о стремлении к качественному обслуживанию, можно было бы выразиться словами «быстрее, дешевле, лучше». Что такое «быстрее» и «дешевле» — вполне понятно. Например, временной фактор связан со скоростью процессоров, наличием или отсутствием очередей невыполненных заданий в буферах, задержкой при транспортировке заданий по сети и т. д. Главное в том, что временной фактор естественным образом выражается скалярно. Так же, как и стоимостный фактор. Конечно, подсчет и времени, и стоимости выполнения заданий может оказаться непростой задачей, но результатом ее решения будет количественное выражение. Сложнее может оказаться выразить числом, что значит «лучше», поскольку речь может идти о плохо формализуемых критериях (например, об использовании более или менее подходящего программного обеспечения и т. п.).

Не расшифровывая, что понимается конкретно под словом «лучше», предположим, что существует неотрицательный показатель качества субъекта в момент времени t, который будем обозначать через $q_i(t)$. При этом условимся, что значение показателя качества, равное нулю, характеризует некоторый идеальный уровень качества обслуживания, а увеличение числовой оценки качества обслуживания соответствует потере качества. Таким образом, чем больше значение показателя качества, тем хуже реальное качество обслуживания.

Следующее предположение заключается в существовании неотрицательных *функций качества* Q_i, которые будут использоваться при вычислении показателей качества субъектов $q_i(t)$ и которые характеризуют качество обслуживания на ресурсе субъекта i. Функция Q_i имеет аргументом объем заданий, поступающий на ресурс i. Характер зависимости от аргумента для каждой из функций каче-

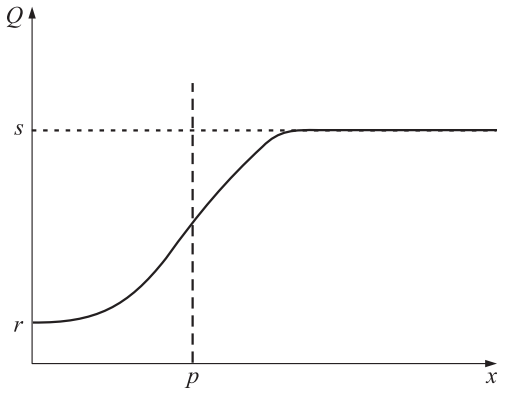

Рис. 1 Функция оценки качества Q от объема заданий x

ства предполагается примерно таким, как у функции $Q(x)$ на рис. 1. Эта функция зависит от трех параметров: p, r и s. Параметр $p > 0$ характеризует тот уровень нагрузки, достижение и превышение которого влечет существенное ухудшение качества обслуживания. Параметр p будем условно называть *емкостью*. Параметр $r \geq 0$ означает минимальную потерю качества, которая возможна при обслуживании и которая достигается при минимальной загрузке данного участника. Этот параметр будем условно называть *интенсивностью (обслуживания)*. Параметр $s > 0$ означает числовую оценку ситуации, в которой качественное обслуживание заданий практически отсутствует. Можно условно употребить выражение *(максимальный) штраф* за плохое обслуживание. Пример аналитического выражения для функции $Q(x)$:

$$Q(x) = \frac{2s \operatorname{arctg} p + \pi r + 2(s-r)\operatorname{arctg}(x-p)}{\pi + 2\operatorname{arctg} p}. \quad (4)$$

В этом примере p — точка перегиба функции $Q(x)$, $Q(0) = r$, $\lim\limits_{x\to\infty} Q(x) = s$.

При составлении соотношений для показателей качества будем исходить из следующих наводящих соображений. Располагая определенным объемом заданий, субъект обеспечивает их обслуживание, оставляя часть заданий на собственном ресурсе и отправляя остальные другим субъектам, либо уничтожая их. Эти действия приводят к разному качеству обслуживания для тех долей заданий, которые подвергаются тому или иному управлению. Например, передача заданий другому субъекту означает, что тот принимает полную ответственность за их дальнейшее обслуживание, которое будет происходить с тем уровнем качества, который обеспечивает принимающая сторона. Чтобы не усложнять дальнейшее описание, предположим, что в системе нет потерь ($\gamma_i = 0$ для всех i). Тогда распределение потоков заданий субъекта i задается набором $\lambda_{i0}, \lambda_{i1}, \ldots, \lambda_{iN}$, и для показателя качества субъекта i запишем соотношение

$$q_i = \lambda_{i0} Q_i + \sum_{n=1}^{N} \lambda_{ij} q_j \,,$$

где $Q_i = Q_i(y_i)$, $y_i = \lambda_{i0} x_i$ — функции качества, имеющие, например, вид (4), с индивидуальными для каждого субъекта параметрами (p_i, r_i, s_i). В матричной форме имеем систему

$$q = \kappa + q\lambda^{\mathrm{T}}$$

или

$$q = \kappa \left(E - \lambda^{\mathrm{T}}\right)^{-1} \,,$$

где $q = (q_1, \ldots, q_n)$; $\kappa = (\lambda_{10} Q_1, \ldots, \lambda_{N0} Q_N)$; T — знак транспонирования.

Замечание. В начале этого подраздела были выделены три составляющие качества обслуживания, которые можно определить как временну́ю, денежную и третью составляющую, учитывающую специфические трудно формализуемые критерии. Определенные выше показатели качества, в том числе функции качества, можно трактовать как *штраф за потерю качества*. Это предполагает, что упомянутая третья составляющая качества имеет не только количественное, но и, фактически, денежное выражение. Поскольку и временна́я компонента, очевидно, легко может быть переведена на тот же язык, например с помощью *штрафа за задержку*, то напрашивается естественная возможность все обсуждение качества в модели свести к стоимостным соотношениям. В данном случае, однако, такая возможность не используется, а рассматривается некий компромиссный вариант. Стоимостная составляющая выступает самостоятельно в виде арендной платы (см. следующий подраздел). Что касается фактора, связанного с задержками, то он, хотя и не входит явным образом в модель, частично отражен, поскольку большие значения функции качества при высокой нагрузке могут интерпретироваться как *штраф за задержку на ресурсе*.

3.3 Стоимостные факторы

Предполагается, что каждый субъект i имеет набор ценовых параметров (a_i, b_{ij}), где a_i — стоимость обслуживания единицы объема заданий, полученных от других субъектов; b_{ij} — тариф за передачу единицы объема заданий от субъекта i к субъекту j.

Кроме того, предположим, что задана цена c — штраф за потерю (уничтожение) единицы объема заданий — одинаковая для всей системы.

Доходы f_i субъекта i представляют собой арендную плату, равную объему поставляемых ему заданий, умноженную на коэффициент a_i. Расходы этого же субъекта складываются из платы g_i за дальнейшее обслуживание той части заданий, которая пересылается другим субъектам, и штрафа за потери h_i. Первая из указанных величин, в свою очередь, составляется из арендной платы другим субъектам и оплаты транспортировки. Согласно (2) и (3) перечисленные компоненты равняются:

$$f_i = a_i \sum_{j=1}^{N} \lambda_{ij} x_j + a_i z_i = a_i x_i \,;$$

$$g_i = \sum_{j=1}^{N} a_j \lambda_{ij} x_i + \sum_{j=1}^{N} b_{ij} \lambda_{ij} x_i \,;$$

$$h_i = c\gamma_i x_i \,.$$

Общий денежный баланс субъекта i (превышение доходов над расходами) определяется как

$$d_i = f_i - g_i - h_i = \delta_i x_i \,,$$

где $\delta_i = a_i - \sum\limits_{j=1}^{N} (a_j + b_{ij})\lambda_{ij} - c\gamma_i$.

3.4 Целевые функции и алгоритм распределения потоков

Целевые функции участников системы должны учитывать как стоимостные аспекты, так и стремление оптимизировать качество выполнения заданий.

Чтобы соединить в одном критерии эти два обычно противоречивых фактора, введем два дополнительных пороговых параметра: $\overline{d}_i$ и $\overline{q}_i$ — бюджет и требуемый уровень качества субъекта i. Субъект стремится действовать так, чтобы расходы не выходили за рамки бюджета, а качество соответствовало заданному уровню. Нарушение ограничений по одному из критериев заставляет субъект заботиться именно об этом показателе. Определим целевую функцию в этих случаях как

$$w_i = \begin{cases} -d_i \,, & \text{если } -d_i > \overline{d}_i,\ \ q_i \le \overline{q}_i \,; \\ q_i \,, & \text{если } -d_i \le \overline{d}_i \,,\ \ q_i > \overline{q}_i \,. \end{cases}$$

Такое определение соответствует стремлению потребителя:

(1) минимизировать расходы, не заботясь о качестве, если он выходит за рамки бюджета, а качество при этом находится в допустимых пределах;

(2) оптимизировать только качество обслуживания в ситуации, когда он укладывается в бюджет, а требования по качеству нарушены.

В ситуации, когда оба пороговых соотношения нарушены или, наоборот, оба выполнены, можно воспользоваться какой-либо сверкой критериев. Например, можно использовать механизм штрафов, мотивируя стремление участников улучшать качество обслуживания. Определим за ухудшение качества обслуживания штраф, размер которого прогрессивно зависит от устанавливаемых тарифов за услуги. Положим

$$w_i = -d_i - (a_i)^{1+a} q_i \,,$$

если $-d_i \le \overline{d}_i$, $q_i \le \overline{q}_i$ или $-d_i > \overline{d}_i$, $q_i > \overline{q}_i$, где $a \ge 0$ — фиксированный параметр.

Управлениями в системе будем считать определенные в п. 3.1 матрицы маршрутизации α и согласования β, которые задают распределение потоков, а также определенные в п. 3.3 стоимостные характеристики a_i. (Матрица сетевых тарифов b_{ij} считается фиксированной.) Исходное представление о децентрализованном и независимом поведении участников означает, что субъект i выбирает собственный вектор маршрутизации α_i, согласования β_i, а также тариф a_i.

Рассмотрим такой вариант поведения участников, при котором они синхронно и независимо изменяют «свои» управления согласно алгоритму проекции градиента, стремясь минимизировать определенные выше целевые функции w_i. Формальную запись алгоритма, однотипного для всех субъектов и для всех управлений, приведем для маршрутизации субъекта i. Пусть $\alpha_i^{(n)} = (\alpha_{i0}^{(n)}, \alpha_{i1}^{(n)}, \ldots, \alpha_{iN}^{(n)})$ — значение вектора α_i для n-й итерации алгоритма, $n = 0, 1, \ldots$, и пусть $\nabla w_i^{(n)}$ — градиент функции w_i в точке $\alpha_i^{(n)}$. Рекуррентное соотношение для последовательных значений вектора $\alpha_i^{(n)}$ имеет следующий вид:

$$\alpha_i^{(n+1)} = \Pi \left(\alpha_k^{(n)} - a^{(n)} \nabla w_i^{(n)} \right) \,, \qquad (5)$$

где Π означает оператор проектирования на единичный симплекс размерности $N+1$, а положительная последовательность $a^{(n)}$ удовлетворяет обычным для такого рода алгоритмов условиям $a^{(n)} \to 0$, $\sum\limits_n a^{(n)} = \infty$.

3.5 Пример

Рассмотрим систему, состоящую из трех субъектов, параметры которых указаны в табл. 1.

Функции качества для всех субъектов имеют вид (4). Требования по уровню качества для участников 1, 2, 3 равняются соответственно 20, 1,2 и 20.

Таблица 1 Параметры системы

Субъект, i	Параметры функции качества Q_i			Поток, z_i	Сетевой тариф, b_{ij}
	p_i	r_i	s_i		
1	20	10	100	100	0; 100; 1
2	120	1	20	25	100; 0; 1
3	0,01	1	1000	0	1; 1; 0

Таблица 2 Показатели системы до и после оптимизации

Субъект, i	Маршрутизация, α_i		Очередь, x_i		Качество, q_i	
	в начале	в конце	в начале	в конце	в начале	в конце
1	0,25	0 0 0,8031 0,1969	225	101	295	19,69
2	0,25	0, 0, 0, 1	150	97	271	1,22
3	0,25	0, 1, 0, 0	125	80	985	13,4

Приведенные значения параметров говорят о следующем. Субъекты 1 и 2 выполняют в системе одновременно роль источников потоков заданий и ресурсов. При этом субъект 1 создает основную нагрузку, но располагает значительно менее емким, производительным и качественным ресурсом, чем субъект 2. Субъект 3 не порождает потока заданий, но он также и не располагает сколько-нибудь значимым ресурсом. Потенциальная роль этого субъекта определяется матрицей сетевых тарифов, из которой следует, что данный участник обладает значительно более экономными возможностями общения с субъектами 1 и 2, чем те сами между собой.

Начальная маршрутизация всех субъектов (векторы α_i) равномерная, а все векторы согласования β_i имеют компоненты, равные 1. Ценовые факторы в этом примере не рассматриваются. Основные показатели, которые дает это очень неэффективное распределение потоков, содержатся в табл. 2 (округленно).

После оптимизации с помощью алгоритма (5) распределение потоков изменилось. Субъект 1 стал использовать участника 3 в качестве транзитного пункта для передачи большей части заданий на более мощный ресурс 2. На собственный ресурс направляется около 20% потока — больший объем нарушил бы требуемый уровень качества 20. Субъект 2 перешел полностью на самообслуживание и стал посылать весь поток на собственный ресурс. Характерно, что из всех компонент матрицы согласования β (в табл. 2 она не отражена) изменился только коэффициент β_{23}, который стал равным приблизительно 0,9. Это вызвано необходимостью для субъекта 2 ограничить поток, получаемый от посредника 3, и тем самым обеспечить заданный уровень качества 1,2.

4 Заключение

Рассмотрена проблема анализа и оптимизации распределения потоков заданий и ценообразования в системах коллективного использования распределенных вычислительных ресурсов. Проведенный литературный обзор показал, что, хотя существуют разнообразные подходы и методы решения проблемы, она пока далека от окончательного решения.

Предложена математическая модель системы вычислительных ресурсов, которая представляет собой интегральное описание потоков заданий в виде динамических балансовых соотношений. Необычность модели в том, что участники системы, вообще говоря, одновременно играют роль источников заданий, владельцев ресурсов и посредников в передаче потоков. Благодаря этому субъекты системы имеют одинаковое, причем математически простое, описание в плане стратегии распределения потоков заданий и выделения ресурсов. В то же время в сравнительно простую модель удалось включить целый ряд факторов, имеющих принципиальное значение для любой системы вычислительных ресурсов: планирование выбора ресурсов для заданий, степень готовности ресурса обслуживать задания, качество обслуживания, учет потерь, затраты и ценообразование, потери в системе и т. д. При этом построение модели является, в сущности, многовариантным. Перечисленные факторы в за-

висимости от потребностей той или иной системы, той или иной задачи могут быть полностью или частично отражены в модели или, наоборот, устранены из нее. В этом смысле можно говорить о том, что предложен способ моделирования системы вычислительных ресурсов.

Изложенный подход к моделированию не привязан к конкретной системе вычислительных ресурсов, поэтому его использование видится прежде всего в изучении вопросов, связанных с построением таких систем вообще. При этом надо еще раз подчеркнуть, что принципиальным соображением в работе было представление об участниках системы как о независимо действующих субъектах, преследующих индивидуальные цели. С учетом последнего замечания к числу важных вопросов, разрешение которых можно надеяться получить с помощью предложенной методологии, относятся, например, следующие.

Определение показателей, которых может достичь система ресурсов, участники которой действуют, исходя из эгоистических интересов. Определение оптимальных стратегий поведения участников.

Установление механизма ценообразования в системах ресурсов.

Определение роли и разумного числа посредников в распределении ресурсов.

Эти и другие вопросы являются направлениями дальнейших исследований.

Литература

1. Информатика: состояние, проблемы, перспективы / Под ред. И. А. Соколова. — М.: ИПИ РАН, 2009. 46 с. ISBN-978-5-902030-69-0.
2. *Демичев А. П., Ильин В. А., Крюков А. П.* Введение в грид-технологии: Препринт. — М.: НИИЯФ МГУ, 2007. 87 с.
3. http://www.systematic.ru/publikatsii/sx/art/310033/po/309844/cp/1/br/309438/discart/310033.html.
4. *Коновалов М. Г., Малашенко Ю. Е., Назарова И. А.* Модели и методы управления заданиями в системах распределенных вычислительных ресурсов: Препринт. — М.: ВЦ РАН, 2009. 110 с.
5. *Xhafa F., Abraham A.* Computational models and heuristic methods for Grid scheduling problems // Future Generation Comput. Syst., 2010. Vol. 26. P. 608–621.
6. *Cho S., Lee M., In J., Kim B., Choi E.* Policy based scheduling for resource allocation on grid / Eds. K. C. Chang *et al.* // APWeb/WAIM 2007 Ws, LNCS, 2007. Vol. 4537. P. 229–234.
7. *Топорков В. В.* Потоковые и жадные алгоритмы согласованного выделения ресурсов в распределенных системах // Известия РАН. Теория и системы управления, 2007. No. 2. P. 109–119.
8. *Vanderster D. C., Dimopoulos N. J., Sobie R. J.* Metascheduling multiple resource types using the MMKP grid // 7th IEEE/ACM Conference (International) on Grid Computing Proceedings, 2006. P. 231–237.
9. *Душин Ю. А.* Модель оценки стоимости гетерогенных ресурсов в Грид // Системы и средства информатики. Спец. вып. Математические модели в информационных технологиях. — М.: ИПИ РАН, 2006. С. 163–172.
10. *Gamst M.* Greedy and metaheuristics for the offline scheduling problem in grid computing // DTU Management Engeneering. — Technical University of Danemark, 2010. http://www.man.dtu.dk/upload/institutter/ipl/publ/publikationer%202010/rapport2.2010.pdf.
11. *Агаларов Я. М.* Динамическая стратегия распределения вычислительных ресурсов локального узла GRID // Системы и средства информатики. Вып. 17. — М.: ИПИ РАН, 2007. С. 17–29.
12. *Slegers J., Mitrani I., Thomas N.* Optimal dynamic server allocation in systems with on/off sources / Ed. K. Wolter // EPEW 2007, LNCS, 2007. Vol. 4748. P. 186–199.
13. *Farzi S.* Efficient job scheduling in grid computing with modified artificial fish swarm algorithm // Int. J. Comput. Theory Eng., 2009. Vol. 1. No. 1. http://www.ijcte.org/papers/003.pdf.
14. *Mathiyalagan P., Dhepthie U. R., Sivanandam S. N.* Grid scheduling using enhanced PSQ algorithm // Int. J. Comput. Sci. Eng., 2010. Vol. 2. No. 2. P. 140–145.
15. *Kamalam G. K., Muralibhaskaran V.* A new heuristic approach: min-mean algorithm for scheduling meta-tasks on heterogenous computing systems // Int. J. Comput. Sci. Network Security, 2010. Vol. 10. No. 1.
16. *Li B., Zhao D.* Online algorithms for single machine schedulers to support advance reservations from grid jobs / Eds. R. Perrott, B. Chapman, J. Subhlok, *et al.* // HPCC 2007, LNCS, 2007. Vol. 4782. P. 239–248.
17. *Nou R., Kounev S., Torres J.* Building online performance models of grid middleware with fine-grained load-balancing: A Globus Toolkit case study / Ed. K. Wolter // EPEW 2007, LNCS, 2007. Vol. 4748. P. 125–140.
18. *Yagoubi B., Slimani Y.* Dynamic load balancing strategy for grid computing // Trans. Eng. Comput. Technol., 2006. Vol. 13. P. 260–265.
19. *Saravanakumar E., Gomathy P.* A novel load balancing algorithm for computational grid // Int. J. Comput. Intelligence, 2010. Vol. 1. Issue 1. P. 20–26.
20. *Berten V., Gaujal B.* Brokering strategies in computational grids using stochastic prediction models // Parallel Comput., 2007. Vol. 33. P. 238–249. www.sciencedirect.com.
21. *Yu K.-M., Luo Z.-J., Chou C.-H., Chen C.-K., Zhou J.* A fuzzy neural network based scheduling algorithm for job assignment on computational grids / Eds. T. Enokido, L. Barolli, M. Takizawa // NBiS 2007, LNCS, 2007. Vol. 4658. P. 533–542.
22. *Ishii R. P., De Mello R. F., Yang L. T.* A complex network-based approach for job scheduling in grid environments /

Eds. R. Perrott, B. Chapman, J. Subhlok, *et al.* // HPCC 2007, LNCS, 2007. Vol. 4782. P. 204–215.

23. *Al-Khateeb A., Abdullah R., Rashid N. A.* Job type approach for deciding job scheduling in grid computing systems // J. Comput. Sci., 2009. Vol. 5. No. 10. P. 745–750. http://www.scipub.org/fulltext/jcs/jcs510745-750.pdf.
24. *Shah S. C., Chahdary S. H., Bashir A. K., Park M. S.* A centralized location-based job scheduling algorithm for interdependent jobs in mobile ad hoc computational grids // J. Appl. Sci., 2010. Vol. 10. No. 3. P. 174–181.
25. *Wei X., Ding Z., Xing S., Yuan Y.* VJM: A novel grid resource co-allocation model for parallel jobs // Int. J. Grid Distributed Comput., 2009. Vol. 1. No. 2.
26. *Ali G., Shaikh N. A., Shaikh Z. A.* Integration of grid and agent systems to perform parallel computations in a heterogeneous and distributed environment // Aust. J. Basic Appl. Sci., 2009. Vol. 3. No. 4. P. 3857–3863.
27. *Vazquez C., Huedo E. S. Montero R. S., Llorente I. M.* Federation of TeraGrid, EGEE and OSG infrastructures through a metascheduler. Preprint submitted to Future Generation Computer Systems, 2010. http://dsa-research.org/doku.php?id=publications:grid:utility.
28. *Ranjan R., Harwood A., Buyya R.* SLA-based cooperative superscheduling algorithms for computational grids // 8th IEEE Conference (International) on Cluster Computing (Cluster 2006) Proceedings // IEEE Computer Society Press, 2006. abs/cs/0605057.
29. *Li J., Sim K. M., Yahyapour R.* Negotiation strategies considering opportunity functions for grid scheduling / Eds. A.-M. Kermarrec, L. Bougé, T. Priol // Euro-Par, 2007, LNCS, 2007. Vol. 4641. P. 447–456.
30. *Vanmechelen K., Broeckhove J.* A comparative analysis of single-unit vickrey auctions and commodity markets for realizing grid economies with dynamic pricing / Eds. D. J. Veit, J. Altmann // GECON 2007, LNCS, 2007. Vol. 4685. P. 98–111.
31. *Krasnotcshekov V., Vakhitov A.* Adaptive scheduling and resource assessment in grid / Ed. V. Malyshkin // PaCT 2007, LNCS, 2007. Vol. 4671. P. 240–244.
32. *Агаларов Я. М.* Функция стоимости ресурсов в экономической модели грид // Информатика и её применения, 2008. Т. 2. Вып. 3. С. 27–34.
33. *Коновалов М. Г., Душин Ю. А., Малашенко Ю. Е., Шоргин С. Я.* Модель взаимодействия потребителей с удаленными вычислительными ресурсами через посредников // Системы и средства информатики. Вып. 19. — М.: Наука,1989. С. 5–33.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 13–24

НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ ЭЛЕМЕНТОВ БАЙЕСОВСКОГО КЛАССИФИКАТОРА

М. П. Кривенко[1]

Аннотация: Рассмотрена задача построения эмпирического байесовского классификатора, обеспечивающего распознавание текста в случае, когда отдельные знаки имеют различные размеры. Представлен комбинированный метод построения оценки элементов байесовского классификатора, включающий непараметрическую ядерную оценку и параметрическую оценку с помощью плотности нормального распределения. Подобная комбинированная оценка позволяет эффективно решать задачу обработки малых объемов обучающей выборки. Продуктивность предложенных идей иллюстрируется на примере распознавания реального текста.

Ключевые слова: байесовский классификатор; комбинированная оценка многомерной плотности распределения; адаптивная ядерная оценка; распознавание текста

1 Введение

Рассматривается задача восстановления (распознавания) текста по изображениям строк, содержащих знаки из алфавита некоторого языка. Используемый при написании знаков шрифт содержит начертания знаков разной ширины. Предполагается, что есть возможность формирования обучающей выборки (множества изображений и соответствующих кодов знаков). Работа является продолжением [1].

Дополнительную сложность сформулированной задаче распознавания придают следующие факторы:

- неопределенность местоположения знака в строке;
- значения ширины знаков используемого шрифта не одинаковы;
- высокая степень искажения изображения.

Рассмотрим байесовский подход, когда считаются заданными статистические свойства классов образов, а класс составляют изображения отдельного знака. В качестве модели текста рассмотрим последовательность знаков, каждый из которых появляется независимо от других с вероятностью, не зависящей от его номера в последовательности. Изображение текста формируется в строке последовательно, знак за знаком; восстановление текста осуществляется также последовательно. В силу этого основной процедурой обработки данных становится следующая: в определенных позициях изображения некоторой строки текста необходимо выявить наиболее предпочтительный знак.

Предполагается, что существует набор классов ω_j, $j = 1, 2, \ldots, M$, которые связаны с формированием в определенных позициях изображения одного из знаков $c(\omega_j)$. Вероятность появления класса ω_j равна $p(\omega_j)$, $j = 1, 2, \ldots, M$ и $\sum\limits_{j=1}^{M} p(\omega_j) = 1$. Далее рассматривается частный, но весьма распространенный вид функции потерь — единичная функция потерь. Для того чтобы приспособить байесовский подход к задаче восстановления знаков различной ширины, приходится ввести дополнительные предположения:

- распределения интенсивностей пикселов в области A_j и вне ее независимы;
- распределения интенсивностей пикселов вне области A_j не зависят от класса ω_j;
- $p(\mathbf{x}(A \backslash A_j)) \neq 0$.

Здесь A — область изображения всей строки, A_j — область изображения знака $c(\omega_j)$, $\mathbf{x}(A \backslash A_j)$ — точка выборочного пространства интенсивностей пикселов в области $A \backslash A_j$. Тогда можно показать [1], что принятие решения основывается на величинах:

$$\frac{p(\mathbf{x}(A_j)|\omega_j)p(\omega_j)}{p(\mathbf{x}(A_j))}, \quad j = 1, 2, \ldots, M. \qquad (1)$$

Для реализации байесовского классификатора необходимо при любом $j = 1, 2, \ldots, M$ знать или уметь вычислять следующие величины: $p(\omega_j)$ — априорную вероятность появления класса ω_j; условное распределение $p(\mathbf{x}(A_j)|\omega_j)$ — значение плотности распределения интенсивностей пикселов в области A_j при условии, что в ней

[1]Институт проблем информатики Российской академии наук, mkrivenko@ipiran.ru

изображается знак $c(\omega_j)$; безусловное распределение $p(\mathbf{x}(A_j))$ — значение плотности распределения интенсивностей пикселов в области A_j.

В [1] рассматривался случай, когда j-й класс образов описывался плотностью многомерного нормального распределения; там показывалось, что в этом случае распределение $p(\mathbf{x}(A_j))$ представляло собой смесь нормальных распределений, и описывались способы задания этой смеси. Основной результат указанной работы заключался в обосновании возможности применения байесовского подхода при классификации объектов, имеющих различную размерность. При этом эксперименты с реальными объектами продемонстрировали эффективность предлагаемого подхода, но вместе с тем выявили недостатки «прямолинейного» оценивания элементов байесовского классификатора с помощью параметрической модели нормального распределения.

Альтернативой традиционным параметрическим моделям служит непараметрический подход — эффективное средство при поиске новых решений при классификации многомерных данных. Здесь в первую очередь речь идет о ядерной оценке плотности распределения. Гистограммная оценка была исключена из рассмотрения сразу же, и причины здесь следующие:

- повышенная сложность задания гистограммной многомерной оценки, ее построение требует задания не только системы и размеров ячеек (классов, категорий), но и определения их ориентации в пространстве;
- выборочные свойства гистограммной оценки оказываются не наилучшими среди других оценок, причем она в большей степени ориентирована на оценивание равномерной плотности (см. разд. 5.4 [2]), которая при распознавании изображений явно отсутствует.

Применение остальных воплощений непараметрического подхода (оценка по методу ближайших соседей, ортогональные разложения) является самостоятельной задачей и не рассматривалось в данной работе.

Ядерная многомерная оценка $f^*(\mathbf{x})$ плотности распределения $f(\mathbf{x})$ обычно строится как обобщение соответствующей оценки одномерной плотности и при заданных n-мерных наблюденных значениях $\mathbf{x}_1, \ldots, \mathbf{x}_N$ имеет следующий вид:

$$f^*(\mathbf{x}) = \frac{1}{N|\mathbf{H}|} \sum_{i=1}^{N} K\left(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)\right), \quad \mathbf{x} \in \Re^n, \quad (2)$$

где $\mathbf{H}$ — несингулярная матрица размера $n \times n$ (обобщение одномерного параметра сглаживания h); $|\mathbf{H}|$ — ее определитель; ядро K — многомерная функция с ограничениями (обычно плотность некоторого распределения). Если $\mathbf{H} = h\mathbf{I}_n$, где $h > 0$, то ядро полностью определяется параметром h и (2) редуцируется к виду

$$f^*(\mathbf{x}) = \frac{1}{Nh^n} \sum_{i=1}^{N} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right), \quad \mathbf{x} \in \Re^n,$$

который и был принят за основу в данной работе.

Непараметрическое оценивание плотности сопровождается рядом общих для многомерного анализа и специфических для распознавания изображений проблем:

- существует множество реализаций метода построения оценки, с каждой из которых связана необходимость установления значений большого числа параметров;
- многомерные данные — плохой объект для визуализации, поэтому весьма трудно представить себе реальную картину происходящего;
- с ростом размерности данных существенно снижается качество получающейся оценки плотности распределения;
- «малый» объем обучающей выборки (число изображений обучающей выборки, относящихся к определенному знаку, может быть существенно меньше размерности векторного представления этого изображения) порождает существенные теоретические проблемы (матрица вторых моментов становится сингулярной).

Снижение качества оценки следует из асимптотических свойств ядерной оценки плотности. В частности, если выполняются условия регулярности для оцениваемой плотности распределения и ядра K, то при $N \to \infty$ и $h_N \to 0$ интегральная среднеквадратичная ошибка (L_2-подход) ядерной оценки стремится к нулю со скоростью $O(N^{-4/(n+4)})$ (см. обзор результатов в разд. 4.5.2 [3]), т. е. скорость снижается при возрастании размерности n.

Проблемы «проклятия размерности» можно пытаться решать с помощью приведения данных к меньшей размерности так, чтобы сохранить специфику самих данных и по возможности улучшить выборочные свойства оценок. Этим путем идут в методе проективного поиска (projection pursuit), описание которого можно найти в разд. 4.4 [4] или в разд. 4.6 [3]. В данном случае он не подходит, так как в основном ориентирован на проектирование в одно- или двухмерное пространство и не решает проблемы сингулярности. По этой причине в данной работе предлагается комбинированный метод

оценивания плотности распределения, когда для первых, наиболее «существенных» координат применяется непараметрическая ядерная оценка плотности, а для остальных — обычная параметрическая оценка с помощью нормального распределения.

2 Комбинированная оценка плотностей распределения

Рассмотрим ситуацию, когда распределение n-мерной случайной величины фактически сосредоточено в m-мерном подпространстве, $m < n$. Переход к переменной в подпространстве сниженной размерности осуществим путем перехода к первым главным компонентам, т. е. с помощью линейного преобразования $\mathbf{L}^{\mathrm{T}}$, удовлетворяющего спектральному разложению заданной ковариационной матрицы $\mathbf{C}$ следующего вида: $\mathbf{C} = \mathbf{LDL}^{\mathrm{T}}$, где $\mathbf{L}$ — ортогональная матрица; $\mathbf{D}$ — диагональная матрица с неотрицательными упорядоченными по убыванию элементами на диагонали.

Заметим сразу же, что при таком преобразовании значение плотности остается без изменений. Действительно, пусть есть случайный вектор $\mathbf{X}$, имеющий плотность распределения $\widetilde{f}(\mathbf{x})$. С помощью ортогонального преобразования $\mathbf{L}^{\mathrm{T}}$ осуществляется переход к вектору $\mathbf{Y} = \mathbf{L}^{\mathrm{T}}\mathbf{X}$, имеющему плотность распределения $f(\mathbf{y})$. Для матрицы преобразования верно следующее: $\mathbf{LL}^{\mathrm{T}} = \mathbf{I}_n$ или $\mathbf{L}^{\mathrm{T}} = \mathbf{L}^{-1}$. Тогда $\mathbf{x} = L\mathbf{y}$, а якобиан преобразования J принимает значение $J = \pm 1$. В результате получается, что плотность распределения вектора $\mathbf{Y}$ есть $f(\mathbf{y}) = |J|\widetilde{f}(\mathbf{x}) = \widetilde{f}(\mathbf{x})$.

Оценку плотности распределения $f^*(\mathbf{y})$ переменной $\mathbf{Y}$ сформируем из базовой части — оценки плотности распределения первых главных компонент, и дополнительной части — параметрической оценки плотности распределения остальных главных компонент, для которой принята аппроксимация с помощью нормального распределения. Причин для выбора такой комбинации несколько:

- эксперименты с реальными изображениями показывают, что в качестве распределения последних главных компонент приближенно можно принять нормальное распределение, для первых же главных компонент это не так, и для них не существует простых параметрических моделей;
- распространение способа оценивания базовой части на всю исходную размерность приводит к неудовлетворительному качеству получающейся оценки, а при больши́х значениях размерности порождает и вычислительные проблемы.

Таким образом, оценка $f^*(\mathbf{y})$ распадается на две составляющие, которые считаются независимыми (некоррелированность первых главных компонент подменяется независимостью совокупности первых главных компонент и всех остальных компонент), и принимает вид:

$$f^*(\mathbf{y}) = f_m^*\left(y^{(1)}, \ldots, y^{(m)}\right) \prod_{j=m+1}^{n} f_1^*\left(y^{(j)}\right),$$

где $f_m^*\left(y^{(1)}, \ldots, y^{(m)}\right)$ — оценка плотности первых m главных компонент, $f_1^*\left(y^{(j)}\right)$ — оценка плотности нормального одномерного распределения j-й главной компоненты.

Матрицы перехода $\mathbf{L}^*$ и дисперсий главных компонент $\mathbf{D}^*$ находятся из уравнения $\mathbf{C}^* = \mathbf{L}^*\mathbf{D}^*\mathbf{L}^{*\mathrm{T}}$, где оценки μ^* и $\mathbf{C}^*$ вектора среднего и ковариационной матрицы находятся обычным образом с помощью обучающей выборки. В случае, когда объемы последней предполагаются меньшими, чем размерность выборочного пространства, часть элементов $\mathbf{D}^*$ будут нулевыми. В связи с этим приходится вводить еще один параметр комбинированной оценки — критическое значение дисперсии главной компоненты d_0, $d_0 > 0$, ниже которого не может опускаться значение диагональных элементов $\mathbf{D}^*$. Этот порог является платой за малые объемы исходных данных и, как следствие, за незнание реальных характеристик распределений данных. Для определения значения d_0 допустим наличие изображений строк текста, для которых состав знаков и их расположение считаются известными. Подобную исходную информацию назовем тренировочной (уточняющей) выборкой.

3 Выбор параметров сглаживания ядерной оценки плотности распределения

Говоря о базовой части комбинированной оценки некоторой плотности распределения, будем вместо $f_m^*\left(y^{(1)}, \ldots, y^{(m)}\right)$ использовать просто запись $f_m^*\left(\widetilde{\mathbf{y}}\right)$, где $\widetilde{\mathbf{y}} \in \Re^m$. Ядерная оценка плотности имеет вид:

$$f_m^*\left(\widetilde{\mathbf{y}}\right) = \frac{1}{Nh^m} \sum_{i=1}^{N} K\left(\frac{\widetilde{\mathbf{y}} - \widetilde{\mathbf{y}}_i}{h}\right),$$

где $\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_N$ — последовательность проекций на первые m главных компонент исходных наблюденных значений $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

В настоящее время для выбора параметра сглаживания существует три основные группы методов:

– грубые (*rule-of-thumb*) методы нахождения наилучшего значения h при условии, что оцениваемая плотность распределения известна и совпадает с некоторой заданной;

– методы перепроверки (*cross-validation*), когда из выборки удаляется единственное наблюденное значение, скажем $\widetilde{\mathbf{y}}_i$, вычисляется оценка плотности в точке $\widetilde{\mathbf{y}}_i$ с помощью остальных $N-1$ выборочных значений, а h выбирается так, чтобы оптимизировать некоторый критерий, включающий все N оценок плотности;

– методы расширения (*plug-in*), в основе которых лежит выбор h для некоторой пилотной плотности распределения, а затем использование получившейся оценки для оценивания поправочного коэффициента в окончательном выражении для h. Эта процедура может быть итерационной.

В данной работе строится вариант метода перепроверки в интерпретации [5], но применительно к многомерному случаю и в форме алгоритма, который можно использовать на практике. В этом случае значение параметра сглаживания h ищется с помощью метода перепроверки как решение следующей задачи:

$$\ln L(h) \to \max\,, \tag{3}$$

где

$$\ln L(h) = \ln \prod_{i=1}^{N} f_m^*(\widetilde{\mathbf{y}}_i) = \sum_{i=1}^{N} \ln \left(\frac{1}{(N-1)h^m} \sum_{\substack{j=1 \\ j\neq i}}^{N} K\left(\frac{\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{y}}_j}{h} \right) \right).$$

В случае, когда ядро — нормальная плотность, имеем

$$\ln L(h) = \sum_{i=1}^{N} \ln \left(\frac{1}{(N-1)(2\pi)^{m/2}} \frac{1}{h^m} \times \sum_{\substack{j=1 \\ j\neq i}}^{N} \exp\left(-\frac{(\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{y}}_j)^{\mathrm{T}}(\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{y}}_j)}{2h^2} \right) \right).$$

Таким образом, необходимо при $h > 0$ найти максимум функции

$$\varphi(h) = \sum_{i=1}^{N} \ln \left(\frac{1}{h^m} \sum_{\substack{j=1 \\ j\neq i}}^{N} \exp\left(-\frac{(\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{y}}_j)^{\mathrm{T}}(\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{y}}_j)}{2h^2} \right) \right) = \sum_{i=1}^{N} \ln \left(\frac{1}{h^m} \sum_{\substack{j=1 \\ j\neq i}}^{N} \exp\left(-\frac{r_{ij}}{2h^2} \right) \right),$$

где $r_{ij} = (\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{y}}_j)^{\mathrm{T}}(\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{y}}_j)$.

Для сужения области поиска решения задачи (3) докажем

Утверждение. *Если h_0 — точка максимума функции $\varphi(h)$, то*

$$h_0 \in \left[\sqrt{\frac{\min\limits_{i\neq j} r_{ij}}{m}},\ \sqrt{\frac{\max\limits_{i,j} r_{ij}}{m}} \right].$$

Доказательство. Рассмотрим функции $\psi_{ij}(h) = (1/h^m)\exp\left(-r_{ij}/(2h^2)\right)$ и $e_{ij}(h) = \exp\left(-r_{ij}/(2h^2)\right)$ при $r_{ij} \neq 0$, $i,j = 1,\ldots,N$. Тогда имеем следующее:

$$\psi'_{ij}(h) = \frac{r_{ij} - mh^2}{h^{m+3}}\, e_{ij}(h)\,; \tag{4}$$

$$\psi''_{ij}(h) = \frac{r_{ij}^2 - (2m+3)r_{ij}h^2 + m(m+1)h^4}{h^{m+6}}\, e_{ij}(h)\,. \tag{5}$$

Из (4) следует, что уравнение $\psi'_{ij}(h) = 0$ имеет единственное допустимое решение $h_{ij} = \sqrt{r_{ij}/m}$, после подстановки которого в правую часть (5) получаем:

$$\psi''_{ij}(h_{ij}) = -\frac{2r_{ij}^2}{m}\, e_{ij}(h_{ij}) < 0\,.$$

Следовательно, точка h_{ij} — единственная точка максимума функции $\psi_{ij}(h)$ при $h > 0$.

Если $r_{ij} \neq 0$, то при $h \in (0, +\infty)$ каждая функция $\psi_{ij}(h)$ достигает единственного максимума при $h_{ij} = \sqrt{r_{ij}/m}$. Отсюда следует, что функция $\varphi(h)$ при $h < \sqrt{\min\limits_{i\neq j} r_{ij}/m}$ является возрастающей, а при $h > \sqrt{\max\limits_{i,j} r_{ij}/m}$ — убывающей, т. е. искомое решение может находиться только на отрезке

$$\left[\sqrt{\frac{\min\limits_{i\neq j} r_{ij}}{m}},\ \sqrt{\frac{\max\limits_{i,j} r_{ij}}{m}} \right], \tag{6}$$

что и требовалось доказать.

Кстати, введя обозначения $z = 1/h^2$ и $c_{ij} = r_{ij}/2$, получаем для некоторых констант c_1 и c_2:

$$\lim_{h\to 0}\psi_{ij}(h) = \lim_{z\to\infty}\psi_{ij}(z) = \lim_{z\to\infty}\frac{z^{m/2}}{\exp(c_{ij}z)} = \ldots = \lim_{z\to\infty}\begin{cases}\dfrac{c_1}{\exp(c_{ij}z)}\\ \dfrac{c_2}{\sqrt{z}\,\exp(c_{ij}z)}\end{cases} = 0\,.$$

Для поиска h_0 будем находить решение уравнения $\varphi'(h) = 0$ с помощью метода деления пополам, для чего необходимо получить выражение для производной:

$$\varphi'(h) = \sum_{i=1}^{N}\frac{\sum\limits_{\substack{j=1\\ j\neq i}}^{N}\left(r_{ij}/h^3 - m/h\right)e_{ij}(h)}{\sum\limits_{\substack{j=1\\ j\neq i}}^{N}e_{ij}(h)} = \frac{1}{h^3}\sum_{i=1}^{N}\frac{\sum\limits_{\substack{j=1\\ j\neq i}}^{N}\left(-mh^2 + r_{ij}\right)e_{ij}(h)}{\sum\limits_{\substack{j=1\\ j\neq i}}^{N}e_{ij}(h)}\,.$$

После этого отрезок (6) возможных решений уравнения $\varphi'(h) = 0$ можно несколько сузить. Введем для $i = 1,\ldots,N$ следующие обозначения: $r_{i,\min} = \min\limits_{j\neq i} r_{ij}$; $r_{i,\max} = \max\limits_{j} r_{ij}$; $\overline{r}_{\min} = (1/N)\sum\limits_i r_{i,\min}$; $\overline{r}_{\max} = (1/N)\sum\limits_i r_{i,\max}$. Тогда

$$\frac{\sum\limits_i r_{i,\min}}{h^3} - \frac{Nm}{h} \le \varphi'(h) \le \frac{\sum\limits_i r_{i,\max}}{h^3} - \frac{Nm}{h}\,.$$

Отсюда получается, что

$$h_0 \in \left[\sqrt{\frac{\overline{r}_{\min}}{m}}, \sqrt{\frac{\overline{r}_{\max}}{m}}\right]\,.$$

Кроме этого, можно записать выражение и для второй производной:

$$\varphi''(h) = -\frac{3}{h^4}\sum_{i=1}^{N}\frac{\sum\limits_{\substack{j=1\\ j\neq i}}^{N}r_{ij}e_{ij}(h)}{\sum\limits_{\substack{j=1\\ j\neq i}}^{N}e_{ij}(h)} + \frac{1}{h^6}\times$$
$$\times\sum_{i=1}^{N}\frac{\sum\limits_{\substack{j=1\\ j\neq i}}^{N}r_{ij}^2e_{ij}(h)\cdot\sum\limits_{\substack{j=1\\ j\neq i}}^{N}e_{ij}(h) - \left(\sum\limits_{\substack{j=1\\ j\neq i}}^{N}r_{ij}e_{ij}(h)\right)^2}{\left(\sum\limits_{\substack{j=1\\ j\neq i}}^{N}e_{ij}(h)\right)^2} + \frac{Nm}{h^2}\,.$$

Другим, иногда более удобным для вычислений, видом второй производной является следующий:

$$\varphi''(h) = \frac{1}{h^6}\times$$
$$\times\sum_{i=1}^{N}\left(\sum_{j\neq i}\sum_{k\neq i}\left(mh^4 - 3h^2r_{ij} - r_{ij}r_{ik} + r_{ij}^2\right)\times\right.$$
$$\left.\times\,e_{ij}(h)e_{ik}(h)\right) \Big/ \left(\sum_{j\neq i}\sum_{k\neq i}e_{ij}(h)e_{ik}(h)\right)\,.$$

Для подтверждения правильности решения (3), полученного с помощью метода деления пополам, на практике достаточно реализовать следующее:

- обеспечивать на левой границе отрезка, который делится пополам, неотрицательное значение $\varphi'(h)$, а на правой — неположительное. Или же, найдя точку, где $\varphi'(h) \approx 0$, проверять знак $\varphi''(h)$. Все это гарантирует, что полученное решение действительно служит приближением для точки максимума;
- с помощью визуального анализа поведения функции $\varphi(h)$, или $\varphi'(h)$, или $\varphi''(h)$ удостовериться, что найденная точка экстремума является единственной.

Более глубокий анализ существования и единственности точки максимума для (3) сдерживается громоздким видом $\varphi'(h)$ и $\varphi''(h)$, а также отсутствием других подходов. Следует обратить внимание, что во всех экспериментах, результаты которых использованы в данной работе, решение (3) оказывалось единственным.

Как указано в разд. 6.4 [2], ядерная оценка плотности распределения с параметром сглаживания,

полученным как решение задачи (3), является состоятельной оценкой, но задача анализа скорости сходимости и описания выборочных свойств такой оценки в реальных ситуациях ждет своего решения. В связи с этим интересна любая информация о реальном поведении оценки при увеличивающихся значениях N и m. С этой целью для нормального распределения с единичной ковариационной матрицей была проведена серия следующих экспериментов: при значениях $N = 8, 16, \ldots, 1024$ находили параметр сглаживания в соответствии с (3), а затем оценивали L_1 — расстояние между теоретической плотностью распределения и ее ядерной оценкой, а именно величину $J_N = \int |f_m(\widetilde{\mathbf{y}}) - f_m^*(\widetilde{\mathbf{y}})|\, d\widetilde{\mathbf{y}}$. Интеграл заменяли суммой, вычисляемой на отрезке $[-3, 3]$ по 7 точкам для каждой из координат в m-мерном пространстве (всего 7^m точек). Оценка $E\{J_N\}$ строилась по 100 повторениям описанного единичного эксперимента.

В случае $m = 1$ можно теоретически получить некоторые результаты, полезные для сравнения. В теореме 1 разд. 5 [2] получены точные асимптотические выражения для $E\{J_N\}$, но опереться на них при выборе параметра сглаживания h трудно. Для получения оптимального решения относительно h обратимся к верхней границе для $E\{J_N\}$. В частности, она не превышается при следующем выборе параметра сглаживания:

$$h = \left(\frac{1}{\sqrt{2\pi}} \, \frac{\sqrt{\kappa_1}}{\kappa_2} \, \frac{\nu_1}{\nu_2} \right)^{2/5} N^{-1/5} \equiv c_h N^{\gamma_h} , \quad (7)$$

где

$$\gamma_h = -\frac{1}{5} ; \qquad c_h = \left(\frac{1}{\sqrt{2\pi}} \, \frac{\sqrt{\kappa_1}}{\kappa_2} \, \frac{\nu_1}{\nu_2} \right)^{2/5} ;$$

$$\kappa_1 = \int K^2(u)\, du ; \qquad \kappa_2 = \int u^2 K(u)\, du ;$$

$$\nu_1 = \int \sqrt{f_m(u)}\, du ; \quad \nu_2 = \int |f_m''(u)|\, du .$$

При этом имеем

$$C \left(\kappa_1^2 \kappa_2\right)^{1/5} \left(\frac{1}{2}\, \nu_1^4 \nu_2 \right)^{1/5} \le N^{2/5} E\{J_N\} \le$$

$$\le 5(8\pi)^{-2/5} \left(\kappa_1^2 \kappa_2\right)^{1/5} \left(\frac{1}{2}\, \nu_1^4 \nu_2 \right)^{1/5} ,$$

где C — универсальная константа (см. теорему 2 разд. 5 [2]); далее $C \approx 1{,}03$ и $5(8\pi)^{-2/5} \approx 1{,}38$. Тогда

$$E\{J_N\} = c_J N^{\gamma_J} , \quad (8)$$

где

$$\gamma_J = -\frac{2}{5} ;$$

$$1{,}03 \left(\kappa_1^2 \kappa_2\right)^{1/5} \left(\frac{1}{2}\, \nu_1^4 \nu_2 \right)^{1/5} \le c_J \le$$

$$\le 1{,}38 \left(\kappa_1^2 \kappa_2\right)^{1/5} \left(\frac{1}{2}\, \nu_1^4 \nu_2 \right)^{1/5} .$$

В случае оценивания плотности стандартного нормального распределения и ядра — плотности опять же стандартного нормального распределения, с помощью прямых вычислений получаем:

$$\kappa_1 = \frac{1}{2\sqrt{\pi}} , \;\; \kappa_2 = 1 , \;\; \nu_1 = (8\pi)^{1/4} , \;\; \nu_2 = \sqrt{\frac{8}{\pi e}} .$$

Полученные теоретические и экспериментальные результаты сведены в табл. 1. Дадим пояснения по ее структуре и принятым обозначениям:

- знаком $*$ помечены характеристики, оценки для которых получены методом линейной регрессии;
- в строке для $m = 1$, выделенной серым цветом, даны теоретические характеристики, полученные в соответствии с (7) и (8);
- в другой строке для $m = 1$ в скобках даны 90-процентные доверительные интервалы выборочных значений.

Таблица 1 Параметры скорости сходимости характеристик ядерной оценки плотности распределения

m	c_h^*	γ_h^*	c_J^*	γ_J^*
1	$c_h = 0{,}75$	$\gamma_h = -0{,}2$	$1{,}03 \le c_J \le 1{,}38$	$\gamma_J = -0{,}4$
1	1,11 (1,08, 1,13)	−0,20 (−0,22, −0,19)	0,89 (0,86, 0,92)	−0,38 (−0,40, −0,37)
2	1,19	−0,17	1,01	−0,30
4	1,24	−0,13	1,21	−0,21
6	1,25	−0,11	1,31	−0,15

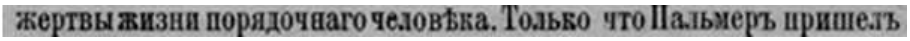

Рис. 1 Изображение одной из строк текста

Из полученных результатов можно сделать следующие выводы:

- при $m = 1$ наблюдается полное согласие эмпирических и теоретических зависимостей, правда, отказ в методе оценивания параметра сглаживания в соответствии с (3) от информации о виде распределения приводит к снижению нижней границы для c_J^* с 1,03 до 0,89;
- увеличение размерности влечет существенное снижение скорости сходимости $E\{J_N\}$ к 0, что может привести при больши́х значениях размерности и ограниченных объемах данных к проблемам в использовании ядерных оценок плотности.

4 Эксперименты

Объектом экспериментов служил отсканированный фрагмент романа Жорж Санд «Она и он», из «Собрания сочинений избранных иностранных писателей», С.-Петербург, 1897. Из-за старения бумаги и типографской краски изображение оказалось низкого качества, не достаточного для уверенного распознавания коммерческими системами (на рис. 1 дан пример изображения одной из строк). Для исследований были взяты 80 изображений строк, которые содержали как знаки, для которых строилась обучающая выборка, так и «неизвестные» знаки. Первую совокупность знаков, которая «известна» классификатору, назовем алфавитом; в нашем случае он содержит 33 строчные буквы и пробел.

Все анализируемые изображения были вручную дополнены эталонными строками, содержащими как информацию о положении знака, так и о его имени. Первые 40 строк были выделены для построения базовой обучающей выборки, общая информация о ней приведена в табл. 2. В ней приняты следующие обозначения: N_{ch} — число изображений некоторого знака в обучающей выборке, W_{ch} — ширина изображения некоторого знака, n_{ch} — число пикселов в изображении некоторого знака (высота у всех знаков одинакова и равна 14 пикселам). Заметим, что объем обучающей выборки для практически любого знака меньше числа пикселов в изображении этого знака, т. е. меньше размерности признакового пространства.

Данные о различных значениях ширины изображения для знаков обучающей выборки отдельно собраны в табл. 3.

Таблица 2 Характеристики обучающей выборки

Знак	N_{ch}	W_{ch}	n_{ch}
пробел	564	3	42
а	153	6	84
б	32	6	84
в	82	6	84
г	24	5	70
д	50	6	84
е	130	5	70
ѣ	50	7	98
ж	22	9	126
з	37	5	70
і	19	3	42
и	93	7	98
й	14	6	84
к	36	6	84
л	83	6	84
м	55	7	98
н	118	7	98
о	192	6	84
п	42	7	98
р	90	6	84
с	88	5	70
т	82	6	84
у	46	6	84
х	12	6	84
ц	7	6	84
ч	32	6	84
ш	10	10	140
щ	4	10	140
ъ	77	8	112
ы	29	8	112
ь	43	6	84
ю	12	8	112
я	60	6	84

Таблица 3 Значения ширины изображения знаков

Знак	W_{ch}
пробел, і	3
г, е, з, с	5
а, б, в, д, й, к, л, о, р, т, у, х, ц, ч, ь, я	6
ѣ , и, м, н, п	7
ъ, ы, ю	8
ж	9
ш, щ	10

Кроме этого, с помощью первых 40 строк были образованы выборки для областей со значениями ширины: 3, 5, 6, 7, 8, 9, 10 пикселов. Их длина оказалась равной 1094.

Таблица 4 Характеристики уточняющей и контрольной выборок

Тип выборки	Содержание	Номера строк	Число знаков	Число знаков из алфавита
Уточняющая	20 строк текста (исходные изображения и эталонные строки)	41–60	1196	1140
Контрольная	20 строк текста (исходные изображения и эталонные строки)	61–80	1212	1150

Таблица 5 Процедуры обработки изображений текста

Название процедуры обработки	Входные данные	Выходные данные
1. Выделение изображений строк текста (Sel_Str)	Изображение текста	Изображения строк текста
2. Формирование обучающей выборки (Create_TrainSample)	Изображения строк текста, эталонные строки текста	Изображения знаков алфавита
3. Оценивание характеристик классов (Est_ClProp)	Изображения знаков алфавита	Характеристики классов, матрицы признак–объект
4. Формирование выборок фрагментов (Create_FragObs)	Изображения строк текста	Матрицы признак–объект
5. Формирование эталонных строк текста (Create_StandStr)	Изображения строк текста	Эталонные строки текста
6. Оценивание порога черно-белого изображения (Est_BWT)	Изображения строк текста	Порог перехода к черно-белому изображению
7. Оценивание элементов байесовского классификатора (Est_BayesCl)	Матрицы признак–объект	Вероятности классов, условные и безусловные распределения
8. Подбор критических собственных значений (Est_CrEVal)	Вероятности классов, условные и безусловные распределения, эталонные строки текста	Уточненные условные и безусловные распределения
9. Распознавание строк текста (Rec_StrImage)	Изображения строк текста, порог перехода к черно-белому изображению, эталонные строки текста, вероятности классов, условные и безусловные распределения, характеристики классов	Строки текста, точность восстановления

Не все параметры алгоритмов распознавания могут быть качественно или вообще оценены по обучающей выборке. Например, обучающая выборка сама по себе не может дать оснований для выбора «достойного» критического значения d_0 дисперсий главных компонент в комбинированных оценках плотностей распределения, так как она не содержит информации обо всем многообразии данных. Поэтому надо либо применять бутстреп-методы, либо дополнять технологию распознавания этапом перепроверки (уточнения, тренировки). В связи с этим 20 имеющихся строк пришлось выделить для перепроверки процедур восстановления текста. Оставшиеся же 20 строк использовались собственно для оценки качества получаемых решений. Общие характеристики уточняющей и контрольной выборки приведены в табл. 4.

Исследование проводилось с помощью модулей обработки, кратко описанных в табл. 5.

Оценивание качества восстановления текста осуществлялось следующим образом: распознанная последовательность знаков сравнивалась с эталонной и вычислялась точность восстановления. Последняя есть доля знаков в восстановленном тексте, «совпавших» со знаками эталонного текста; при этом «совпавшими» считаются те знаки, которые находятся с точностью до параметра толерантности на одинаковых позициях и совпадают по имени. Толерантность введена по практическим соображениям, так как при измерении положения знака в изображении исходного текста и восстановлении знака обязательно имеют место ошибки; этот параметр в данной работе принимает значение 3 пиксела.

При сравнительном анализе условий обработки данных возникает проблема учета разброса выборочных значений точности восстановления, так как в случае последовательного алгоритма распо-

Таблица 6 Качество распознавания для различных значений параметров поиска знака в строке

LS	RS										
	0	1	2	3	4	5	6	7	8	9	10
0	545	821	896	888	876	857	804	711	665	645	631
1	627	906	911	895	876	862	805	709	668	648	632
2	745	925	***928***	899	880	863	805	703	662	648	631

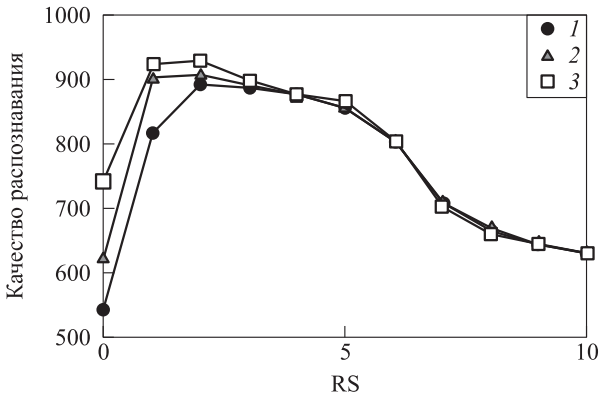

Рис. 2 Зависимость качества распознавания от RS при различных значениях LS: *1* — 0; *2* — 1; *3* — 2

знавания это бывает сделать трудно: объем исходных данных не всегда достаточен для формального статистического вывода, последовательность распознавания отдельных знаков не является последовательностью испытаний Бернулли. Поэтому при незначительных отличиях значений показателя качества будем обращать внимание на тенденции, интерпретируя результаты анализа как указание на то, что выявленная тенденция может иметь место, и поэтому представляет интерес.

Основным шагом алгоритма последовательного распознавания знаков в строке является следующий: после того как восстановлен очередной знак, от его позиции x_{t-1} в строке осуществляется продвижение на ширину w_{t-1} этого знака и поиск начала x_t очередного знака в диапазоне начальных позиций $[x_{t-1} + w_{t-1} - \mathrm{LS},\ x_{t-1} + w_{t-1} + \mathrm{RS}]$, здесь t — шаг последовательного просмотра строки. Таким образом, параметрами процесса последовательного распознавания являются следующие: RS — максимальное значение возможных сдвигов вправо и LS — максимальное значение возможных сдвигов влево. В табл. 6 для различных комбинаций указанных параметров приведены значения числа правильно распознанных знаков из уточняющей выборки. Следует обратить внимание на то, что значение $\mathrm{LS} = 3$, а это минимально возможная ширина знака (см. табл. 3), подчас приводит к тому, что алгоритм возвращается к восстановлению уже распознанного знака, т. е. возникает зацикливание. Кроме этого, обращает на себя внимание рост качества распознавания при увеличении LS, что может сопровождаться приписыванием приблизительно одним и тем же позициям строки как правильно, так и неправильно распознанных знаков (фактически нарушается принцип: при распознавании некоторого знака считается, что предшествующий знак был распознан правильно). По этим причинам значения для LS, бóльшие 2, не рассматривались. Распознавание проводилось с помощью метода, основанного на расстоянии типа хи-квадрат. В графическом виде данные табл. 6 отражены на рис. 2. Для дальнейших экспериментов были приняты следующие значения: $\mathrm{LS} = 2$ и $\mathrm{RS} = 2$, соответствующая клетка табл. 6 выделена курсивом.

Построение эффективного эмпирического байесовского классификатора связано с необходимостью выбора оценок следующих параметров:

- m и d_0 для условных распределений;
- m и d_0 для безусловных распределений;
- вероятности появления классов.

Эта задача осложняется тем, что вообще-то параметры m и d_0 могут принимать для отдельных классов и областей различные значения. Понятно, что сделать это с помощью уточняющей выборки полностью не представляется возможным, поэтому в данной работе были реализованы следующие шаги:

- перебор при $m = 0$ (это значение бралось одинаковым для условных и безусловных распределений) значений d_0 (общих для всех классов и отдельно общих для всех областей) и оценка точности распознавания; выбор среди них той пары значений d_0, которой соответствует наилучшая точность распознавания;
- перебор при найденной наилучшей паре значений d_0 различных комбинаций значений m для условного и безусловных распределений отдельно, выбор из них наилучшей комбинации с точки зрения точности распознавания.

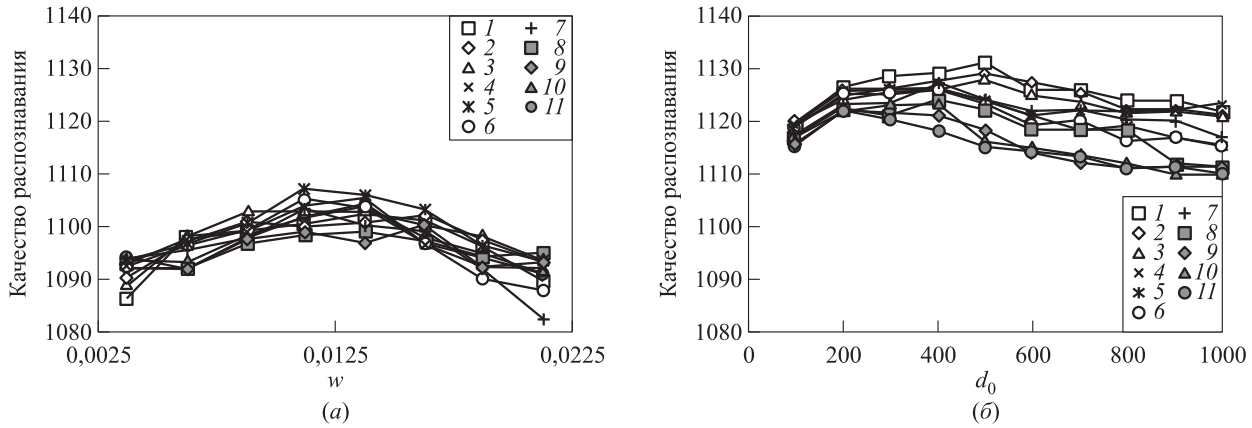

Рис. 3 Зависимость качества распознавания от d_0 для относительного (*а*) и абсолютного (*б*) способа задания при различных значениях m: *1* — 0; *2* — 1; *3* — 2; *4* — 3; *5* — 4; *6* — 5; *7* — 6; *8* — 7; *9* — 8; *10* — 9; *11* — 10

Перед реализацией описанных шагов были рассмотрены два приема задания d_0:

– относительный, когда значение d_0 определяется в терминах доли w дисперсии первой главной компоненты;

– абсолютный, когда d_0 принимает некоторое значение.

Далее в виде графиков на рис. 3 приведены результаты оценивания точности распознавания изображений строк текста тренировочной выборки для различных значений d_0 при относительном способе их задания с помощью w (рис 3, *а*) и при абсолютном способе (рис 3, *б*) . Из них можно сделать вывод о том, что абсолютный способ задания d_0 имеет преимущества.

Можно выделить три области значений d_0 — малые, средние и большие. Для малых значений d_0 характерно низкое качество распознавания, так как оно опирается на малую по объему обучающую выборку (классы уплотнены). Как следствие, появление «новых» распознаваемых образов приводит к росту числа ошибок. Для больших значений d_0 теряется индивидуальность классов (классы размыты), они начинают пересекаться, что опять же приводит к росту числа ошибок. Выбор средних значений d_0 снижает значимость отмеченных факторов (либо объективно сложившейся уплотненности, либо привнесенной размытости), что приводит к улучшению качества распознавания.

Результаты анализа точности распознавания для пар значений d_0 для условных ($d_0 = 100, 200, \ldots, 1600$) и безусловных ($d_0 = 100, 200, \ldots, 1200$) распределений говорят о том, что их выбор может стать источником повышения точности распознавания. В частности, для дальнейших экспериментов можно предложить следующую пару значений: $d_0 = 100$ для безусловных распределений и $d_0 = 300$ для условных распределений.

Для выбранной пары значений d_0 были опробованы различные комбинации значений $m = 0, 2, \ldots, 14$ для условных и безусловных распределений. Из полученных результатов видно, что для условных распределений вполне можно обойтись моделью нормального многомерного распределения. Если для распределения образов из отдельных классов (для условных распределений) это достаточно естественно, то для распределения образов из областей (безусловных распределений) логичнее использовать смесь распределений (см. [1]). Необходимость усложнения модели для безусловных распределений подтверждается и экспериментами (рост показателей качества распознавания при увеличении значения m) и должно стать предметом более тщательного исследования. По идее, качество распознавания должно достигать своего максимума для некоторого $m > 0$, что определяется наличием двух противоречивых факторов, сопровождающих увеличение размерности базовой части комбинированной оценки:

– ростом качества аппроксимации истинных распределений данных;

– снижением качества ядерной оценки плотности распределения.

В итоге анализ качества распознавания для контрольной выборки проводился при следующих значениях параметров: для условных распределений $m = 2$ и $d_0 = 300$, для безусловных распределений $m = 14$ и $d_0 = 100$. Соответствующие результаты приведены в табл. 7. При реализации классифи-

Таблица 7 Эффективность различных методов распознавания

Тип классификатора	Число распознанных знаков	Доля распознанных знаков, %	Время, с
Основанный на расстоянии типа сравнения классификаций	506	44	27
Основанный на расстоянии типа скалярного произведения	820	71	55
Основанный на расстоянии типа хи-квадрат	943	82	24
Эмпирический байесовский классификатор при равных вероятностях появления знаков из алфавита	1139	99	109
Эмпирический байесовский классификатор при оценках вероятностей появления знаков из алфавита	1138	99	106

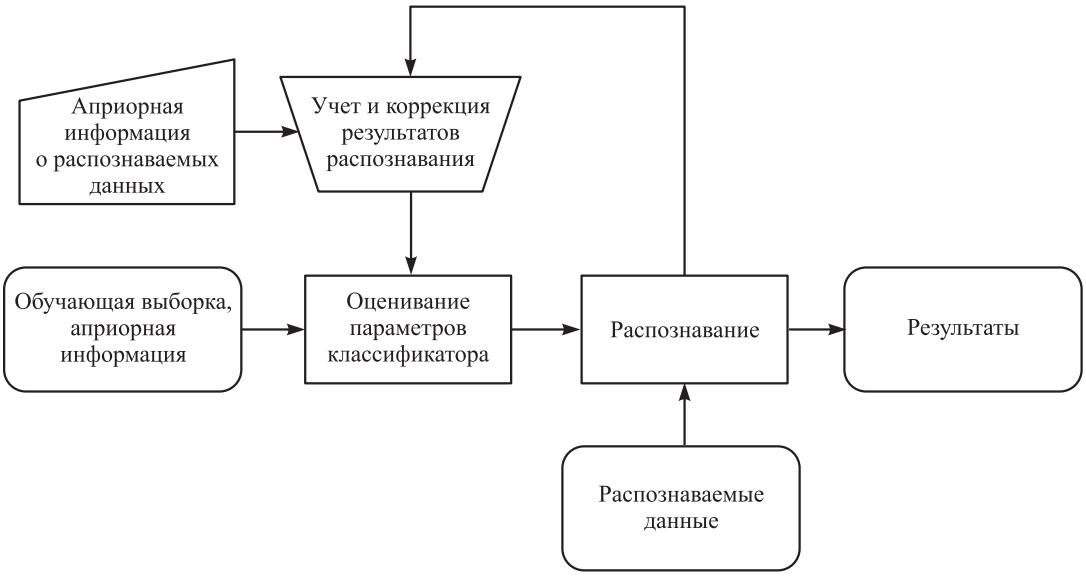

Рис. 4 Интерактивная схема распознавания

катора, основанного на расстоянии типа сравнения классификаций, для черно-белого порога было принято значение 162 (способ его нахождения дан в [6]). Полученные экспериментальные результаты говорят о том, что байесовский классификатор обладает явными преимуществами, при этом учет информации о распределении классов — вероятностях появления знаков — практически ничего не дает. Последнее, как и многие прежде полученные «нечеткие» выводы, скорее всего, объясняется большим числом распознанных знаков (1139 из 1150 знаков алфавита), что не позволяет проявиться преимуществам использования при реализации эмпирического байесовского классификатора всей имеющейся информации.

5 Заключение

Предложенная и рассмотренная в данной работе комбинированная оценка плотностей распределений, ее экспериментальный анализ показали высокую эффективность байесовского подхода при классификации объектов, имеющих различную размерность, а также работоспособность предложенных методов оценивания элементов байесовского эмпирического классификатора.

При этом продемонстрирована необходимость развития технологии распознавания с обратной связью типа той, что представлена на рис. 4.

В качестве ближайших задач можно перечислить:

- исследование эффективности использования параметрической модели смеси распределений для представления базовой части оценки плотности;
- исследование эффективности перехода от самооценки значения параметра сглаживания h типа (3) к его оцениванию с помощью тренировочной выборки;

– снижение временно́й сложности алгоритмов, реализующих байесовский классификатор, с помощью учета особенностей комбинированной оценки плотности, наличия одинаковых элементов (1) для отдельных классов и некоторой фиксированной области A_j.

Литература

1. *Кривенко М. П.* Распознавание элементов изображения, имеющих различные размеры // Системы и средства информатики. — М.: ИПИ РАН, 2007. Вып. 17. С. 30–51.
2. *Деврой Л., Дьёрфи Л.* Непараметрическое оценивание плотности: L_1-подход. — М.: Мир, 1988. 408 с.
3. *Izenman A. J.* Modern multivariate statistical techniques: Regression, classification, and manifold learning. — Springer Verlag, 2008. 731 p.
4. *Simonoff J. S.* Smoothing methods in statistics // Springer series in statistics. 2nd printing, 1998. 338 p.
5. *Duin R. P. W.* On the choice of smoothing parameters for Parzen estimators of probability density functions // IEEE Transactions on Computers, 1976. Vol. C-25. P. 1175–1179.
6. *Кривенко М. П.* Расщепление смеси вероятностных распределений на две составляющие // Информатика и её применения, 2008. Т. 2. Вып. 4. С. 48–56.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 25–35

ВОПРОСЫ РАЗРЕШИМОСТИ ЗАДАЧИ РАСПОЗНАВАНИЯ ВТОРИЧНОЙ СТРУКТУРЫ БЕЛКА*

К. В. Рудаков[1], И. Ю. Торшин[2]

Аннотация: Цель работы — разработка формализма для последующего применения алгебраического подхода к проблеме распознавания вторичной структуры белка. Проведено формальное описание задачи, рассмотрена ее разрешимость, регулярность и локальность. Введены ключевые понятия для анализа локальности, такие как окрестность, маска, система масок, монотонность и тупиковость систем масок; предложен метод построения безызбыточных систем масок. Разработанный формализм позволил сформулировать математическое описание принятой у биологов гипотезы о локальном характере зависимости вторичной структуры от первичной и получить конструктивные критерии разрешимости задачи.

Ключевые слова: алгебраический подход; вторичная структура белка; биоинформатика; окрестность; локальность; разрешимость; регулярность

1 Введение. Мотивация и постановка проблемы

Клетка — мельчайшая структурная единица организма, а белки — активные молекулярные образования, поддерживающие жизнь клетки. В современной биологии любой белок рассматривается с нескольких точек зрения:

(1) как одномерная аминокислотная последовательность (так называемая «первичная структура белка», 1D);

(2) как одномерная последовательность характерных локальных конфигураций («вторичная структура», 2D);

(3) как трехмерный объект («третичная структура», «пространственная структура», 3D);

(4) как особый механизм, выполняющий определенную роль в функционировании клетки [1].

В настоящее время основным постулатом биологии белка является утверждение о том, что первичная структура однозначно определяет вторичную и третичную структуры, а третичная структура определяет биологическую роль белка. Поэтому основной задачей теоретической биологии белка считается установление закономерностей, определяющих взаимосвязь первичной и третичной структур. Как показал проведенный ранее анализ [1], для решения этой задачи целесообразно решить промежуточную — установить взаимосвязь между первичным и вторичным уровнями структуры белка или решить задачу «распознавания вторичной структуры белка».

Следует отметить, что имеющиеся данные о первичном, вторичном и третичном уровнях описания структуры белка получены на основании существенно разных экспериментов. Первичная структура (последовательность символов в 20-буквенном алфавите) устанавливается посредством «секвенирования» (дословно «установления последовательности») — процедуры последовательной химической деградации молекулы белка. Вторичная структура (последовательность локальных конфигураций) и третичная структура (набор координат атомов) устанавливаются дифракционными методами (как правило, рентгеноструктурный анализ) или посредством исследования внутримолекулярного спин-спинового расщепления с использованием ЯМР (ядерного магнитного резонанса).

В то время как точность секвенирования определяется однозначно как совпадение–несовпадение символов и достигает 100%, трудно даже дать определение «точности» структурного эксперимента. Установленные исследователями координаты атомов молекулы белка (третичная структура) зависят не только от внутримолекулярных взаимодействий, но и от многочисленных условий проведения структурного эксперимента: выбора метода (дифракция, ЯМР), температуры, кислотности среды (pH), качества кристалла (в случае дифракционного метода), концентрации раствора белка (в случае

*Работа выполнена при поддержке РФФИ, гранты 09-07-12098, 09-07-00212-а и 09-07-00211-а.

[1]ВЦ РАН им. А. А. Дородницына, Московский физико-технический институт, rudakov@ccas.ru

[2]Российское отделение Института микроэлементов ЮНЕСКО, tiy135@yahoo.com

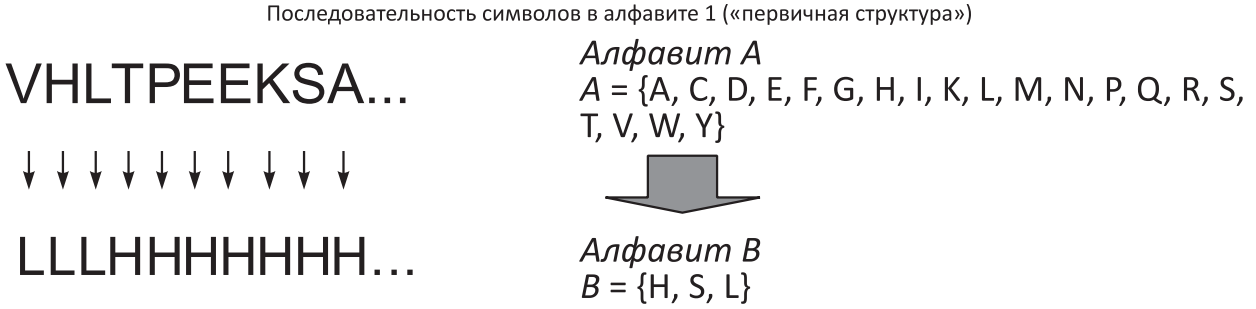

Рис. 1 Задача распознавания вторичной структуры белка

ЯМР), присутствия других молекул и т. д. Как результат, в разных экспериментах по определению структуры одного и того же белка устанавливаются отличающиеся друг от друга наборы координат атомов молекул этого белка.

Экспертный анализ третичных структур белков указал на существование ряда характерных пространственных конфигураций локальных участков молекулы белка: «спиралей», «стрэндов» и «петель». Последовательности этих пространственных конфигураций были названы «вторичной структурой белка». Вторичная структура как последовательность символов некоего алфавита (различные способы определения этого алфавита рассмотрены далее) — результат интерпретации набора координат атомов молекулы белка. Так как координаты атомов и межатомные расстояния подвержены вариациям вследствие упомянутых выше особенностей структурного эксперимента, то и вторичная структура как последовательность символов также подвержена вариациям. В то же время существующие алгоритмы расчета вторичной структуры на основе третичной, основанные на эвристиках из проблемной области, дают результаты, более чем в 90% случаев совпадающие с мнением эксперта [1–3].

В целом, распознавание вторичной структуры белка на основе его первичной структуры (аминокислотной последовательности) — одна из важнейших задач современной теоретической биологии. Актуальность задачи обусловлена значительным объемом данных по первичной структуре белка (миллионы аминокислотных последовательностей) и в сотни раз меньшим количеством экспериментальных данных по третичной и, следовательно, вторичной структуре белка. Это позволяет рассматривать накопленный материал о третичном и вторичном уровнях структуры белка как обучающую выборку для задачи распознавания вторичной структуры белка по первичной и применять алгебраический подход к проблемам распознавания [4–12]. В настоящей работе данная задача рассматривается как перевод последовательности символов из одного алфавита в другой (рис. 1).

Особого внимания заслуживает локальность рассматриваемой задачи. Применение современных физических методов для исследования структуры и свойств белковых молекул (в частности, экспериментальные исследования наносекундной динамики белка, реалистичное моделирование молекулярной динамики свертывания белка и молекулярные механизмы биосинтеза белка [1, 13, 14]), равно как и существующие эвристики для «предсказания вторичной структуры» [3], позволяют предположить локальный характер зависимости вторичной структуры от первичной.

Таким образом, противоречивость экспериментальных данных, обусловленная особенностями структурного эксперимента, неоднозначность определения алфавита для описания вторичной структуры и необходимость систематического исследования гипотезы о локальном характере взаимосвязи между первичной и вторичной структурами указывают на целесообразность разработки специализированного формализма для корректной постановки изучаемой проблемы.

Основная цель настоящей работы — разработать формализм для постановки задачи распознавания вторичной структуры белка в терминах современной теории распознавания [4–12]. Особое внимание уделяется развитию формализма для тестирования гипотезы о локальном характере зависимости вторичной структуры белка от его первичной структуры.

Приводимые далее статистические оценки были сделаны на основании анализа общедоступных экспериментальных данных по первичной, вторичной и третичной структурам (на сегодняшний день имеются данные рентгеноструктурного анализа или ЯМР для более 50 000 белков, помещенные в базу данных Protein Data Bank (PDB) [2]).

2 Введение. Об алфавитах для описания вторичной структуры белка

В рамках разрабатываемого формализма используются два алфавита: алфавит A для описания первичной структуры белка (в дальнейшем «верхнее слово») и алфавит B для описания вторичной структуры («нижнее слово»). В случае задачи распознавания вторичной структуры белка алфавит A определяется однозначно и соответствует множеству 20 типов аминокислот, образующих цепи белков, $n = |A|20$, $A = \{A, C, D, E, F, G, H, I, K, L, M, N, P, R, S, T, V, W, Y\}$. Алфавит B может быть определен существенно различающимися способами: через использование базовых алфавитов или через использование производных от них алфавитов. Ниже рассматриваются три способа определения алфавита B:

(1) базовые алфавиты;

(2) производные алфавиты на основе последовательностей литер базового алфавита;

(3) расширение базового алфавита с учетом сегментов вторичной структуры нижнего слова.

Во-первых, *базовый алфавит* типа B можно определить как трехбуквенный, $m = |B| = 3$, $B = \{H, S, L\}$, где H (от *англ.* helix) обозначает конфигурацию типа «спираль», S (strand) — конфигурацию «стрэнд» (плоский вытянутый участок белковой цепи) и L (loop) — участок произвольной структуры, т. е. не H и не S. Частота встречаемости каждого из символов при использовании 3-буквенного алфавита отражена в табл. 1, примеры конфигураций H, S, L показаны на рис. 2.

Трехбуквенный алфавит отображает принципиально различающиеся трехмерные пространственные конфигурации, которые может принимать тот или иной участок белковой цепи. На этом алфавите построено предыдущее поколение программ анализа структуры белка, которое не вполне точно отображает более тонкие геометрические различия известных конфигураций вторичной структуры. Так, при использовании более современных программ (основанных, прежде всего, на модификациях алгоритма STRIDE [13]), 3-буквенный алфавит трансформируется в 8-буквенный. Хотя основными конфигурациями по-прежнему остаются H, S и L, алфавит заметно усложняется (табл. 2). Как будет показано далее, сложность B-алфавита может сказаться на разрешимости рассматриваемой задачи распознавания.

Таблица 1 Частота встречаемости типов вторичной структуры при использовании 3-буквенного алфавита B (АК — аминокислота)

Литера	Тип вторичной структуры	Встречаемость
Спирали		
H	A-спираль, 4 АК на виток	0,36
Стрэнды		
$E(S)$	Стандартный стрэнд, минимальная длина 2 АК	0,22
Петли		
L	Петля	0,42

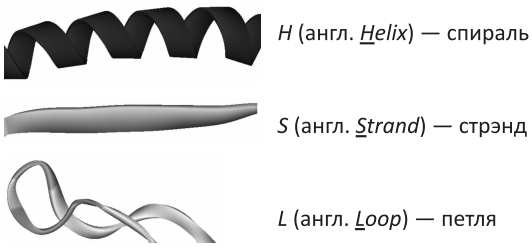

Рис. 2 Три основные формы вторичной структуры

Таблица 2 Частота встречаемости различных вариантов вторичной структуры в 8-буквенном алфавите B (АК — аминокислота)

Литера	Тип вторичной структуры	Встречаемость
Спирали		
H	A-спираль, 4 АК на виток	0,31
G	3_{10}-спираль, 3 АК на виток	0,04
I	π-спираль, 5 АК на виток	0,0002
Стрэнды		
$E(S)$	Стандартный стрэнд, минимальная длина 2 АК	0,20
Петли		
L	Петля	0,24
T	Поворот петли (turn)	0,11
S'	Изгиб петли (bend)	0,09
B	Мост к стрэнду (bridge)	0,01

Во-вторых, возможно *определение производного алфавита* B через пары букв базового B-алфавита. При этом последовательность из N символов в A-алфавите переводится в последовательность из $N - 1$ пар базового B-алфавита. Например, для трехбуквенного базового алфавита H, S, L ($m = 3$) число пар составит $m^2 = 9$ и B-алфавит будет записываться как $\{b_{ij}\}$, $B = \{HH, HS, HL, SS, SH, SL, LL, LH, LS\}$. Для примера на рис. 1 литере V в 1-й позиции верхнего слова будет соответствовать пара литер LL нижнего слова, литере T в 4-й позиции — пара литер HH и т. д. Очевидно, что B-последовательность из $N-1$

Таблица 3 Частота встречаемости пар (трехбуквенный алфавит)

Пары букв	Частота встречаемости
HH	0,33
HS	0,001
HL	0,03
SS	0,19
SH	0,003
SL	0,05
LL	0,32
LH	0,03
LS	0,05

двухбуквенных пар соответствует N символам нижнего слова. Частота встречаемости каждой из пар представляет определенный интерес (табл. 3).

Несколько важных выводов о структуре нижнего слова следует из данных табл. 3:

1. Пары HH, SS и LL встречаются наиболее часто и покрывают более 80% всех нижних слов.
2. Пары HL, LH, SL, LS встречаются на порядок реже и соответствуют границам сегментов. Тем не менее границы сегментов встречаются довольно часто (0,03–0,05) и могут распознаваться отдельным семейством алгоритмов.
3. Из пп. 1 и 2 следует преобладание во вторичной структуре *достаточно длинных однобуквенных сегментов* типа $HHHHH$, $SSSSS$, $LLLLL$. Частотные распределения длин сегментов показаны на рис. 3.

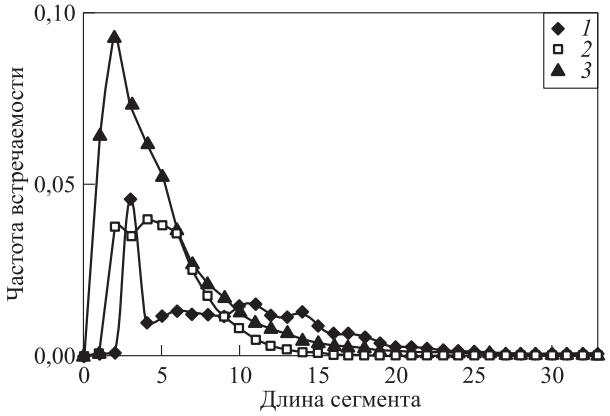

Рис. 3 Частота встречаемости длин сегментов вторичной структуры. Пик в распределении HHH при длине сегмента 3 соответствует спиралям типа G (8-буквенный алфавит): *1* — HHH (спирали); *2* — SSS (стрэнды); *3* — LLL (петли)

Таблица 4 Частота встречаемости литер расширенного трехбуквенного алфавита B

Литера	Частота встречаемости
H	0,33
S	0,19
L	0,32
H_S^b	0,0015
H_L^b	0,015
H_S^e	0,0005
H_L^e	0,015
S_H^b	0,0005
S_L^b	0,025
S_H^e	0,0015
S_L^e	0,025
L_H^b	0,015
L_S^b	0,025
L_H^e	0,015
L_S^e	0,025

4. Частоты переходов SH и HS крайне низки, т. е. границы между сегментами $SSSSS$ и $HHHHH$ — редкое явление. С физической точки зрения переходы SH и HS соответствуют резкой смене конфигурации и энергетически невыгодны вследствие Ван-дер-Ваальсова отталкивания.

В-третьих, возможно *расширение базового алфавита с учетом сегментной структуры нижних слов*. При этом базовый B-алфавит из m литер трансформируется в $(m + 2m(m-1))$-буквенный. Элементами расширенного алфавита являются элементы базового алфавита плюс $2m(m-1)$ элементов, соответствующих границам сегментов. Например, трехбуквенный базовый алфавит ($m = 3$) становится 15-буквенным расширенным B-алфавитом $B = \{H, S, L, H_S^b, H_L^b, H_S^e, H_L^e, S_H^b, S_L^b S_H^e, S_L^e, L_H^b, L_S^b, L_H^e, L_S^e\}$, частота встречаемости литер которого представлена в табл. 4.

В целом граница между сегментами $\ldots LLL$ и $HHH\ldots$ описывается четырьмя различными типами B-литер: $\ldots LLL_H^e H_L^b HH\ldots$ С физической точки зрения такой расширенный алфавит позволяет подчеркнуть различия в пространственных конфигурациях индивидуальных аминокислотных остатков вдоль белковой цепи. Например, литера L_H^e расширенного алфавита соответствует последней (*англ.* **e**nd) литере сегмента петли $\ldots LLL$, который переходит в спиральный $HHH\ldots$ сегмент. Соответственно, литера L_H^b расширенного алфавита соответствует первой (*англ.* **b**eginning) литере сегмента петли $\ldots LLL$, которому предшествует спиральный сегмент $\ldots HHH$.

Возможно использование еще более сложных B-алфавитов: трехбуквенных, четырехбуквенных и более сложных комбинаций литер базового алфавита, использование расширенных алфавитов на основе таких комбинаций и т. д. Так или иначе, *предлагаемый в настоящей работе математический аппарат применим при использовании любого способа определения B-алфавита*.

Выбор того или иного способа определения B-алфавита важен для конкретных реализаций решения данной задачи распознавания. Ниже будем говорить, что алфавит B' есть *расширение* алфавита B, если существует такая функция $f : B' \to B$, что ею определяется однозначный перевод всех слов алфавита B' в соответствующие слова алфавита B.

3 Исходные определения

Пусть заданы два алфавита: A и B, $A = \{a_1, a_2, \ldots, a_n\}$, $n > 0$, $B = \{b_1, b_2, \ldots, b_m\}$, $m > 0$. Обозначим множества слов длины k в каждом из алфавитов A^k и B^k соответственно. Тогда множество всех исходных слов в алфавите A есть $A^* = \bigcup\limits_{l=1}^{\infty} A^l$, а множество всех слов в алфавите B есть $B^* = \bigcup\limits_{l=1}^{\infty} B^l$. Решение исследуемой задачи распознавания сводится к поиску некоторой функции $F : A^* \to B^*$, причем $|F(\vec{a})| = |\vec{a}|$ ($|\vec{a}|$ — длина слова $\vec{a}$). Пусть $\mathbf{F} = \{F : A^* \to B^*;\ |F(\vec{a})| = |\vec{a}|\}$ — множество функций этого типа.

Пусть Δ — неопределенность. Введем Δ-расширенные алфавиты $\tilde{A} = A \bigcup \{\Delta\}$ и $\tilde{B} = B \bigcup \{\Delta\}$ и Δ-расширенные множества слов $\tilde{A}^* = \bigcup\limits_{l=1}^{\infty} \tilde{A}^l$ и $\tilde{B}^* = \bigcup\limits_{l=1}^{\infty} \tilde{B}^l$ соответственно.

Пусть задано конечное множество прецедентов $\mathbf{Pr} \subseteq \tilde{A}^* \times \tilde{B}^*$, $\mathbf{Pr} \neq \varnothing$, где «×» обозначает декартово произведение. Прецедент, таким образом, представляет собой пару слов $\left(\vec{a}, \vec{b}\right) \in \mathbf{Pr}$, $|\vec{a}| = |\vec{b}|$. Назовем $V = \vec{a}$ «верхним словом», а $Q = \vec{b}$ — «нижним словом» прецедента. Будем называть функцию F *корректной*, если $\underset{\mathbf{Pr}}{\forall} \left(\vec{a}, \vec{b}\right) :\ F(\vec{a}) = \vec{b}$. Тогда очевидна

Теорема 1. Функция F *существует тогда и только тогда, когда*

$$\underset{\mathbf{Pr}}{\forall} \left(\vec{a}_1, \vec{b}_1\right), \left(\vec{a}_2, \vec{b}_2\right) :\ (\vec{a}_1 = \vec{a}_2) \Rightarrow \left(\vec{b}_1 = \vec{b}_2\right). \quad (1)$$

В соответствии с теоремой 1 существование корректной функции F (т. е. разрешимость исследуемой задачи) зависит от выбора множества $\mathbf{Pr}$. Таким образом, в жесткой постановке задача о распознавании вторичной структуры разрешима при заданном множестве прецедентов $\mathbf{Pr}$, если для нее существует корректное решение F, т. е. верхнему слову каждого прецедента однозначно поставлено в соответствие нижнее слово. Все возможные $\mathbf{Pr} \in \tilde{A}^* \times \tilde{B}^*$, $\mathbf{Pr} \neq \varnothing$ подразделяются на те, на которых задача разрешима, и те, на которых задача неразрешима.

В идеале некое множество $\mathbf{Pr}$, на котором задача разрешима, должно включать все известные структуры белков (например, все данные из PDB [2]). Но в такой формулировке существование корректной F, к сожалению, опровергается наличием в базах данных о структуре белков примеров прецедентов, в которых совпадают верхние слова и не совпадают нижние. Прецеденты с одинаковыми верхними, но отличающимися нижними словами возникают в результате параллельной работы разных исследовательских групп, использования существенно различающихся физических методов установления структуры белка, разных условий структурного эксперимента, разных способов выделения и очистки белка и т. д. На рис. 4 приведены примеры таких противоречивых прецедентов.

По теореме 1 при наличии противоречивых прецедентов корректных функций F не существует. Анализ противоречивых прецедентов в реальных данных и какая-либо форма обработки этих противоречивых прецедентов (исключение, усреднение, нахождение медианы, составление профиля

```
ATVKFKYKGEEKQVDISKIKKVWRVGKMISFTYDEGGGKTGRGAVSEKDAPKELLQMLAKQ
LSSSSSSLLSSSSSSHHHSSSSSSLLLLSSSSSSLLLLLSSSSSSLLLLLLLLHHHHHHLL 1
LLSSSSSSLLSSSSSLHHHLLSSLLLLLLSSSSLLLLLLLLLLSSLLLLLLLLHHHHHHHL 2
LLSSSSSSLLSSSSSSHHHSSSSSSSLLSSSSSSLLLLLLSSSSSSSSHHHLLHHHHHHHH 3
LLSSSSSLLLLSSSSSHHHSSSSSSSLLSSSSSSLLLLLSSSSSSSLLLLLHHHHHHHHH 4
```

Рис. 4 Пример противоречивого набора экспериментальных данных и соответствующие условия экспериментов: ЯМР — ядерный магнитный резонанс (в скобках указано число модельных структур), рентген — рентгеноструктурная кристаллография белка, pH — кислотность среды, T — температура, + ДНК — структура была определена в комплексе с фрагментом дезоксирибонуклеиновой кислоты: *1* — ЯМР (35 стр.), pH = 4,5, T = 300 K; *2* — рентген, pH = 6,5, T = 123 K, + ДНК; *3* — ЯМР (50 стр.), pH = 5,0, T = 323 K, + ДНК; *4* — рентген, pH = 6,5, T = 287 K, + ДНК

вероятностей, введение весов позиций, использование данных об условиях структурного эксперимента и т. п.) представляет собой отдельное направление исследований. В дальнейшем будут рассматриваться только *множества непротиворечивых прецедентов*, для которых существует корректная функция F.

Исследуемую задачу Z, определяемую множеством прецедентов **Pr**, будем называть *разрешимой*, если для нее существует корректная функция F, т. е. выполнено условие (1) разрешимости задачи. Наряду с разрешимостью в современной теории распознавания [4–12] обычно изучается *регулярность* задач. Под регулярностью задачи понимается разрешимость самой задачи, сопровождающаяся разрешимостью задач из некоторой ее окрестности в изучаемом множестве задач.

Понятие регулярности определяется тем, как задаются окрестности задачи. Следуя идеологии научной школы академика Ю. И. Журавлева, определим окрестность задачи Z с множеством прецедентов $\mathbf{Pr} = \{\left(\vec{a}_1, \vec{b}_1\right), \left(\vec{a}_2, \vec{b}_2\right), \ldots, \left(\vec{a}_q, \vec{b}_q\right)\}$ как множество задач Z' с множеством прецедентов $\mathbf{Pr}' = \{\left(\vec{a}_1, \vec{b}'_1\right), \left(\vec{a}_2, \vec{b}'_2\right), \ldots, \left(\vec{a}_q, \vec{b}'_q\right)\}$ при произвольных $\vec{b}'_1, \vec{b}'_2, \ldots, \vec{b}'_q$. Отсюда следует, что задача Z будет регулярной на множестве прецедентов **Pr** тогда и только тогда, когда выполняется следующее *условие регулярности*:

$$\underset{\mathbf{Pr}}{\forall} \left(\vec{a}_i, \vec{b}_i\right), \left(\vec{a}_j, \vec{b}_j\right), (i \neq j) \Rightarrow (\vec{a}_i \neq \vec{a}_j) \ . \quad (2)$$

4 Локальность

Предлагаемый формализм разрабатывается с целью тестирования гипотезы о локальном характере взаимосвязи между первичным и вторичным уровнями структуры белка. В рамках формализма *локальность* означает то, что каждая литера нижнего слова определяется неким подсловом верхнего слова. Пусть дано слово $\vec{U} = \{u_1, u_2, \ldots, u_n\}$ длины n. Это может быть верхнее слово (V) или нижнее слово (W). Определим некую *ведущую позицию* i, $1 \leq i \leq n$. Дана также «маска» $\hat{m} = \{\mu_1, \mu_2, \ldots, \mu_m\}$, где $\mu_i \in Z$, $\mu_1 < \mu_2 < \ldots < \mu_m$. Будем называть μ_i *позициями маски*. Параметр m назовем *размерностью маски* $\hat{m}$ и будем обозначать как $|\hat{m}|$, а параметр $\mu_m - \mu_1 + 1$ назовем *протяженностью маски* и обозначим как $[\hat{m}]$. Определим *оператор выбора подслова* $\eta(i, \hat{m}, \vec{U})$:

$$\eta(i, \hat{m}, \vec{U}) = \begin{cases} u_{i+\mu_1} u_{i+\mu_2} \ldots u_{i+\mu_m} \ , \\ \qquad \text{если } i + \mu_1 \geq 1 \ , \ \ i + \mu_m \leq n \, ; \\ \varnothing \quad \text{в противном случае.} \end{cases}$$

Иначе говоря, оператор η выбирает определенную подпоследовательность слова $\vec{U}$ по маске $\hat{m}$, помещенной на позицию i. *Подсловом* или $(\hat{m}, i)$-*подсловом* будем называть конкретное значение оператора η на определенной позиции i некоторого слова $\vec{U}$, выбранное по маске $\hat{m}$. С точки зрения алгебраического подхода пара «маска» – «ведущая позиция» может быть рассмотрена как аналог «опорного множества», а $(\hat{m}, i)$-подслово — как аналог «представительного набора».

Пусть имеется система масок

$$M = \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_N\} \, .$$

Будем считать, что

$$\hat{m}_1 = \left(\mu_1^1, \mu_2^1, \ldots, \mu_{|m_N|}^N\right), \ldots, \quad \hat{m}_N = \left(\mu_1^N, \mu_2^N, \ldots, \mu_{|m_N|}^N\right) \ .$$

Определим одноэлементную систему масок $\hat{M}_\Sigma(M)$ как *объединенную маску* $\hat{m}$ такую, что $\hat{m} = \bigcup\limits_{k=1}^{|M|} \hat{m}_k$. Очевидно, что

$$\underset{k=1}{\overset{|M|}{\forall}} \hat{m}_k : \ (|\hat{m}_k :| \leq |\hat{m}_\Sigma|) \,\&\, \left([\hat{m}_k] \leq \left[\hat{M}_\Sigma\right]\right) \ .$$

Слова в множестве прецедентов **Pr** имеют конечную длину, поэтому для точности изложения следует описать краевые эффекты с учетом области определения оператора η. Пусть L, $R \in N \bigcup \{0\}$. Функцию F назовем (L, R)-*корректной*, если для $\forall \left(\vec{a}, \vec{b}\right) \in \mathbf{Pr}$ выполнено $F(\vec{a}) = \vec{b}'$, где $b'_1 = b'_2 = \ldots = b'_L = \Delta$, $b'_{|\vec{a}|-R+1} = \ldots = b'_{|\vec{a}|} = \Delta$ и $b'_i = b_i$ при $L < i \leq |\vec{a}| - R$. Иначе говоря, для прецедента $\left(\vec{a}, \vec{b}\right)$ F является (L, R)-корректной, если она вычисляет слово $\vec{b}$ с точностью до L, R от краев нижнего слова.

Считаем, что имеется система масок M. Можно предложить два существенно различающихся способа определения границ для описания краевых эффектов. В первом случае $L(M)$ и $R(M)$ определяются как минимальные отступы от краев верхнего слова (слева и справа соответственно), при которых применимы все маски из M. При этом $L(M) + 1 = \min(i) : \ \underset{k=1}{\overset{N}{\forall}} \left(i + \mu_1^k \geq 1\right)$, а $R(M) = |\vec{a}| - \max(i) : \ \underset{k=1}{\overset{N}{\forall}} \left(i + \mu_{|m_k|}^k \leq |\vec{a}|\right)$. Тогда $L(M) = \max\left(-\min\limits_{k=1,N} \mu_1^k, 0\right)$ и аналогично $R(M) = \max\left(\max\limits_{k=1,N} \mu_{|m_k|}^k, 0\right)$.

Во втором случае границы определяются как такие минимальные значения i (значения ведущей позиции), при которых применима, по крайней мере, одна маска из M. Обозначим левую границу (левый отступ) как $l(M)$. В этом случае $l(M)+1 = \min(i) : \overset{N}{\underset{k=1}{\exists}} \left(i+\mu_1^k = 1\right)$, и тогда $l(M) = \max\left(-\max\limits_{k=1,N}\mu_1^k, 0\right)$. Аналогично правая граница $r(M)$ определяется как $r(M) = \max\left(\min\limits_{k=1,N}\mu_{k=1,N}^k, 0\right)$.

Теорема 2. *($l(M), r(M)$)-корректная функция также ($L(M), R(M)$)-корректна.*

Доказательство. Нетрудно показать, что $l(M) \le L(M)$. Рассмотрим произвольное множество масок, первый элемент каждой маски μ_1^k, множество всех первых элементов $\{\mu_1^k\}$, $\mu_1^k \in Z$. $L(M)$ определяется через поиск минимума μ_1^k, $-\min\limits_{k=1,N}\mu_1^k$, а $l(M)$ определяется через максимум μ_1^k, $-\max\limits_{k=1,N}\mu_1^k$. Очевидно, что минимум не может превышать максимум на одном и том же множестве целых чисел, так что $\min\limits_{k=1,N}\mu_1^k \le \max\limits_{k=1,N}\mu_1^k$, $-\min\limits_{k=1,N}\mu_1^k \ge -\max\limits_{k=1,N}\mu_1^k$ и, следовательно, $l(M) \le L(M)$. Аналогично $r(M) \le R(M)$. В случае любой (L,R)-корректной функции $F(\vec{a}) = \vec{b'}$, где $b'_1 = b'_2 = \ldots = b'_L - \Delta$, $b'_{|\vec{a}|-R+1} = \ldots = b'_{|\vec{a}|} = \Delta$ и $b'_i = b_i$ для любых $\left(\vec{a}, \vec{b}\right) \in \mathbf{Pr}$. Бóльшие значения L или R соответствуют бóльшему числу ведущих позиций, в которых $b'_i = \Delta$ при $i < L$ и при $i > |\vec{a}| - R$. Поэтому, если функция F является $(l(M), r(M))$-корректной, $(L(M), R(M))$-корректность этой же функции просто соответствует увеличению числа позиций, в которых $b'_i = \Delta$, так как $l(M) \le L(M)$ и $r(M) \le R(M)$. Неопределенность (Δ) не влияет на корректность функции, так как не нарушает условия $F(\vec{a}) = \vec{b'}$. Теорема доказана.

(L,R)-корректность F важна для практических применений предлагаемого формализма. Так, $(l(M), r(M))$-корректность гарантирует корректность распознавания на концах верхнего слова, а $(L(M), R(M))$-корректность необходима для выбора безызбыточных систем масок (см. далее).

5 Условие существования локальных функций

Гипотеза о локальности исследуемой задачи распознавания формулируется ниже как гипотеза о существовании некоторой локальной функции. Пусть дана система масок M. Определим класс корректных локальных функций $f(M) \subseteq \mathbf{F}$. Функция F принадлежит $f(M)$ тогда и только тогда, когда существует функция $f : \left(\tilde{A}\right)^{|\hat{M}_\Sigma(M)|} \to \tilde{B}$ такая, что для любого $\vec{a} = (a_1, \ldots, a_n)$ выполнено $F(\vec{a}) = \vec{b} = (b_1, \ldots, b_n)$, где $b_1 = b_2 = \ldots = b_{l(M)} = \Delta$, $b_{n-r(M)+1} = \ldots = b_n = \Delta$, а при $l(M) < i \le n - r(M)$ выполнено условие

$$b_i = f\left(\eta\left(i\hat{M}_\Sigma(M), \vec{a}\right)\right). \qquad (3)$$

Отметим, что при построении корректных алгоритмов с использованием конструкций алгебраического подхода функция f будет искаться в виде:

$$f\left(\eta\left(i, \hat{M}_\Sigma(M), \vec{a}\right)\right) = \\ = g\left(h_1\left(\eta\left(i, \hat{m}_1, \vec{a}\right)\right), \ldots, h_{|M|}\left(\eta\left(i, \hat{m}_{|M|}, \vec{a}\right)\right)\right).$$

Таким образом, условие разрешимости задачи распознавания вторичной структуры белка может быть сформулировано следующим образом: дано такое множество масок M и множество прецедентов **Pr**, что существует локальная функция $F \in f(M)$ такая, что F корректна на данном множестве прецедентов **Pr**. Очевидно, что приведенные выше соображения о выборе **Pr** и о существовании противоречивых прецедентов в реальных выборках данных относятся и к локальной формулировке задачи.

При поиске решения $F \in f(M)$ естественно требовать именно $(l(M), r(M))$-корректности. *Критерием локальной разрешимости* будем называть следующее условие:

$$\underset{\mathbf{Pr}}{\forall}\left(\vec{V}_1, \vec{W}_1\right), \left(\vec{V}_2, \vec{W}_2\right) \underset{i,j \in N}{\forall}(i,j) : \eta\left(i\hat{M}_\Sigma(M), \vec{V}_1\right) = \\ = \eta\left(j, \hat{M}_\Sigma(M), \vec{V}_2\right) \Rightarrow W_1^i = W_2^j, \\ l(M) < i \le \left|\vec{V}_1\right| - r(M), \\ l(M) < j \le \left|\vec{V}_2\right| - r(M), \quad i \ne j. \qquad (4)$$

Теорема 3. *Локальная функция F, корректная на множестве прецедентов* **Pr***, существует тогда и только тогда, когда выполняется критерий локальной разрешимости.*

Доказательство. Предположим, что условие (4) не выполняется и в **Pr** существует хотя бы одна пара прецедентов, для которой при некоторых значениях i и j окрестности по $\hat{M}_\Sigma(M)$ в верхнем слове идентичны $\eta(i, \hat{M}_\Sigma, \vec{V}_1) = \eta(j, \hat{M}_\Sigma, \vec{V}_2) = \eta'$, а литеры нижнего слова различаются: $W_1^i \ne W_2^j$. Тогда в соответствии с (3) получается, что $W_1^i = f(\eta')$

и $W_2^j = f(\eta')$, т. е. происходит нарушение тождественности: $f(\eta') \neq f(\eta')$. Следовательно, критерий (4) является необходимым условием существования $F \in f(M)$. Теорема доказана.

Следствие 1. *Из теорем 2 и 3 следует критерий локальной* $(L(M), R(M))$*-разрешимости*

$$\underset{\mathbf{Pr}}{\forall}\left(\vec{V}_1, \vec{W}_1\right), \left(\vec{V}_2, \vec{W}_2\right) \underset{i,j\in N}{\forall} (i,j) : \eta\left(i, \hat{M}_\Sigma(M), \vec{V}_1\right) = \eta\left(j, \hat{M}_\Sigma(M), \vec{V}_2\right) \Rightarrow W_1^i = W_2^j\,,$$

$$L(M) < i \leq \left|\vec{V}_1\right| - R(M)\,,$$

$$L(M) < j \leq \left|\vec{V}_1\right| - R(M)\,, \quad i \neq j\,. \quad (4')$$

Следствие 2. *Расширение B-алфавита может приводить к потере разрешимости задачи* $Z(\mathrm{Pr}, M)$.

B-алфавит может быть расширен за счет использования более сложного базового алфавита (3-буквенного, 8-буквенного), использования пар литер базового алфавита, а также за счет расширения с учетом сегментной структуры нижних слов (см. разд. 1). Расширенный алфавит $B^{\text{р}}$ может быть однозначно переведен в базовый алфавит $B^{\text{б}}$ посредством функции $f_B : B^{\text{р}} \to B^{\text{б}}$. Например, в случае 3- и 8-буквенных базовых алфавитов (см. табл. 2) функция f_B определяется как $f_B : \{H, G, I \to H;\ S \to S;\ T, S'', B, L \to L\}$. Тогда, если есть разрешимость в алфавите $B^{\text{р}}$, она существует и в алфавите $B^{\text{б}} = f_B(B^{\text{р}})$, так как $\underset{\text{б}}{W_1^i} f_B(\underset{\text{р}}{W_1^i})$ и $\underset{\text{б}}{W_2^j} f_B(\underset{\text{р}}{W_2^j})$. Обратное неверно: некоторым литерам из $B^{\text{б}}$ соответствуют несколько литер $B^{\text{р}}$, так что не существует $f_B : B^{\text{б}} \to B^{\text{р}}$. Следовательно, хотя из разрешимости задачи в алфавите $B^{\text{р}}$ следует разрешимость в $B^{\text{б}}$, из разрешимости в более простом алфавите $B^{\text{б}}$ не следует разрешимость в более сложном $B^{\text{р}}$, т. е. при расширении алфавита может происходить потеря разрешимости задачи.

По теореме 3 критерий локальной разрешимости является необходимым и достаточным условием существования F. На основе условия (4) становится возможным проведение экспериментов по оценке параметров индивидуальных масок, при которых существование F возможно.

Введем *критерий локальной разрешимости* с использованием отдельных масок:

$$\underset{\mathbf{Pr}}{\forall}\left(\vec{V}_1, \vec{W}_1\right), \left(\vec{V}_2, \vec{W}_2\right) \quad \forall (i,j) \left(\underset{k=1}{\overset{|M|}{\forall}} \hat{m}_k :\ \eta\left(i, \hat{m}_k, \vec{V}_1\right) = \eta\left(j, \hat{m}_k, \vec{V}_2\right)\right) \Rightarrow W_1^i = W_2^j\,,$$

$$l(M) < i \leq \left|\vec{V}_1\right| - r(M)\,,$$

$$l(M) < j \leq \left|\vec{V}_2\right| - r(M)\,, \ i \neq j\,. \quad (4'')$$

Теорема 4. *Условия* (4) *и* (4″) *эквивалентны.*

Доказательство. Любые два подслова $\vec{v}^1 = \{v_1^1, v_2^1, \ldots, v_m^1\}$ и $\vec{v}^2 = \{v_1^2, v_2^2, \ldots, v_m^2\}$ равны, $\vec{v}^1 = \vec{v}^2$, когда совпадают литеры в каждой позиции, т. е. $\underset{i=1}{\overset{m}{\forall}} v_i^1 = v_i^2$. Поэтому выражение $\eta(i, \hat{M}_\Sigma(M), \vec{V}_1) = \eta(j, \hat{M}_\Sigma(M), \vec{V}_2)$ в (4) соответствует системе равенств S такой, что $\forall \mu \in M_\Sigma(M)$: $v_{i+\mu}^1 = v_{j+\mu}^2$, т. е.

$$S = \begin{cases} v_{i+\mu_1}^1 = v_{j+\mu_1}^2\,; \\ v_{i+\mu_2}^1 = v_{j+\mu_2}^2\,; \\ \ldots\,; \\ v_{i+\mu_m}^1 = v_{j+\mu_m}^2\,, \end{cases}$$

где $M_\Sigma(M) = \{\mu_1, \mu_2, \ldots, \mu_m\}$, $m = |M_\Sigma(M)|$

Очевидно, что каждому из $\mu_1, \mu_2, \ldots, \mu_m$ соответствует определенное равенство в S. Любой $\hat{m}_k \subset M_\Sigma(M)$ соответствует подсистема равенств s_k, которая может быть записана как $\eta(i, m_k \vec{V}_1) = \eta(j, m_k, \vec{V}_2)$. Так как $M_\Sigma(M) = \underset{k=1}{\overset{|M|}{\bigcup}} \hat{m}_k$, то $S = \underset{k=1}{\overset{|M|}{\bigcup}} s_k$, так что выражения $\eta(i, \hat{M}_\Sigma(M), \vec{V}_1) = \eta(j, \hat{M}_\Sigma(M), \vec{V}_2)$ и $\left(\underset{k=1}{\overset{|M|}{\forall}} \hat{m}_k :\ \eta(i, \hat{m}_k, \vec{V}_1) = \eta(j, \hat{m}_k, \vec{V}_2)\right)$ — эквивалентны. Последнее соответствует эквивалентности утверждений (4) и (4″). Теорема доказана.

Задачу $Z(Pr, M)$ будем называть локально разрешимой на **Pr**, если для нее существует $(l(M), r(M))$-корректная функция $F \in f(M)$. Таким образом, условия (4)–(4″) определяют наличие у задачи Z свойства *локальной разрешимости*.

Критерием, определяющим наличие у задачи свойства *локальной регулярности*, будем называть следующее условие:

$$\underset{\mathbf{Pr}}{\forall}\left(\vec{V}_1, \vec{W}_1\right), \left(\vec{V}_1, \vec{W}_2\right) \underset{i,j\in N}{\forall} (i,j), i \neq j,\ \vec{V}_1 \neq \vec{V}_2 \Rightarrow \eta\left(i, M_\Sigma(M), \vec{V}_1\right) \neq \eta\left(j, M_\Sigma(M), \vec{V}_2\right)\,,$$

$$l(M) < i \leq \left|\vec{V}_1\right| - r(M)\,,$$

$$l(M) < j \leq \left|\vec{V}_2\right| - r(M)\,. \quad (5)$$

Отметим, что из условия (5) и теоремы 2 следует

$$\underset{\mathbf{Pr}}{\forall}\left(\vec{V}_1,\vec{W}_2\right),\left(\vec{V}_2,\vec{W}_2\right)\underset{i,j\in N}{\forall}(i,j), i\neq j, \vec{V}_1\neq\vec{V}_2\Rightarrow$$
$$\Rightarrow \eta\left(i,M_\Sigma(M),\vec{V}_1\right)\neq\eta\left(j,M_\Sigma(M),\vec{V}_2\right),$$
$$L(M)<i\leq\left|\vec{V}_1\right|-R(M),$$
$$L(M)<j\leq\left|\vec{V}_2\right|-R(M),\quad i\neq j.$$

Локальная регулярность $Z(Pr,M)$ — предельный случай локальной разрешимости этой задачи. Как и условие регулярности (2), *критерий локальной регулярности* позволяет сформулировать условие разрешимости задачи Z с учетом гипотезы о локальности. В случае распознавания вторичной структуры белка регулярность задачи может быть достигнута при использовании систем масок избыточно высокой размерности и протяженности.

6 Разрешимость задачи и монотонность условия разрешимости

При выполнении условия существования локальных функций (4)–(4″), задача распознавания вторичной структуры *разрешима*, в противном случае — *неразрешима*. Наличие разрешимости задачи $Z(Pr,M)$ можно определить при заданных Pr и M. Ниже будем считать, что имеется непротиворечивое множество прецедентов **Pr**, и рассмотрим возможности варьирования множества масок M.

Варьирование M заключается в добавлении или удалении отдельных масок. В общем случае условие разрешимости (4) *немонотонно* по M, т. е. из существования разрешимости при M не следует разрешимость при произвольном M' таком, что $M\subseteq M'$. Иначе говоря, не исключена возможность нахождения маски $\hat{m}\notin M$ такой, что при включении ее в M изменятся значения $l(M)$ и $r(M)$, так что условие разрешимости (4) также нарушится вследствие невыполнения требования $(l(M),r(M))$-корректности. Поэтому целесообразно исследовать монотонность условия разрешимости (4) при условии неизменности значений $l(M)$ и $r(M)$: если для $Z(Pr,M)$ есть разрешимость, $M\subseteq M'$, $l(M)=l(M')$ и $r(M)=r(M')$, то разрешимость есть и для $Z(Pr,M')$. Условие разрешимости (4) также в общем случае не монотонно по M и при $M'\subseteq M$: из существования разрешимости задачи $Z(Pr,M)$ не следует разрешимость $Z(Pr,M')$, где M' получено при удалении масок из M. Рассмотрение монотонности свойства разрешимости задачи при $M'\subseteq M$ имеет особое значение для нахождения безызбыточных и тупиковых систем масок.

7 0-тупиковость и тупиковость систем масок

В общем случае система масок M, при которой задача разрешима, может быть избыточной в том смысле, что разрешимость сохранится при удалении некоторых масок из M. Вообще говоря, определение безызбыточной системы масок M — задача, разрешимая полным перебором подмножеств множества M. Однако в практически интересных случаях полный перебор неосуществим. Полный перебор может быть сокращен по аналогии с поиском минимальных д.н.ф. (дизъюнктивных нормальных форм) [15].

Пусть имеется система масок M, удовлетворяющая условиям (4) и (4″). Условия (4) и (5) — необходимые условия, которые ограничивают рамки рассмотрения при анализе систем масок: поиск безызбыточных систем масок должен проводиться среди множества подмножеств объединенной маски $\hat{m}=M_\Sigma(M)$ при $L,R=const$.

Условие (4) монотонно по $M_\Sigma(M)$ в следующем смысле: если (4) выполнено для $M'\subset M$ такой, что $M_\Sigma(M')\subset M_\Sigma(M)$, то (4) выполнено и для M. Если же условие (4) выполнено для M, но не выполнено для любой $M'\subset M$ такой, что $M_\Sigma(M')\subset M_\Sigma(M)$, то систему масок M назовем *0-тупиковой*. Далее рассматриваются только 0-тупиковые M. *Тупиковой* назовем такую систему масок, в которой условие (4) нарушается для любой $M'\subset M$. Отметим, что не все 0-тупиковые M являются тупиковыми.

Пусть дана 0-тупиковая система масок $M=\{\hat{m}_1,\hat{m}_2,\ldots,\hat{m}_N\}$. По аналогии с задачей упрощения д.н.ф. [15] маску $\hat{m}_{i_0}$, $i_0\in\{1,N\}$ будем называть *ядерной*, если $\hat{m}_{i_0}\notin\bigcup\limits_{j=1,N}^{j\neq i_0}\hat{m}_j$. Ядерными системами или подсистемами масок будем называть M, обладающие *свойством ядерности*:

$$\underset{i=1}{\overset{N}{\forall}}\, i\ \underset{\hat{m}_i}{\exists}\,\mu:\ j\left(\mu\notin\hat{m}_j\right). \qquad (6)$$

Теорема 5. *M — тупиковая система масок тогда и только тогда, когда M обладает свойством ядерности.*

Доказательство. Необходимость доказывается от противного. Условие (6) соответствует тому, что для каждой маски из M выполнено условие

$$\forall i \in \{1 \dots N\} \hat{m}_i \notin \bigcup_{j=1,N}^{j \neq i} \hat{m}_j \,. \qquad (6')$$

Допустим, что (6′) нарушено и $\hat{m}_i \subseteq \bigcup_{i=1,N}^{j \neq i} \hat{m}_j$. Тогда $\bigcup_{i=1,N} \hat{m}_i = \bigcup_{j=1,N}^{j \neq i} \hat{m}_j$, т. е. $M_\Sigma(M') = M_\Sigma(M)$. Поэтому условие (4) будет выполняться и для $M' = \{\hat{m}_1, \hat{m}_2, \dots, \hat{m}_{i-1}, \hat{m}_{i+1}, \dots, \hat{m}_N\}$, так что M не является тупиковой. Последнее противоречит исходной предпосылке, а следовательно, тупиковым M присуще свойство ядерности. Достаточность следует из определения свойства ядерности (6): так как любая $\hat{m}_i$ ядерной системы масок M содержит определенную уникальную позицию μ объединенной маски $M_\Sigma(M)$, удаление любой $\hat{m}_i$ из M неизбежно приведет к образованию такой M', что $M_\Sigma(M') \subset M_\Sigma(M)$, так что условие разрешимости (4) будет нарушено для M'. Последнее соответствует нарушению определения тупиковости, так что любая ядерная M является и тупиковой. Теорема доказана.

Следствие 1. *Из тупиковости следует 0-тупиковость.* В самом деле, тупиковая M обладает свойством ядерности, т. е. каждой $\hat{m}_i$ соответствует некая уникальная позиция μ в объединенной маске $M_\Sigma(M)$. Поэтому при удалении любой $\hat{m}_i$ удалится соответствующая позиция μ с образованием измененной объединенной маски $M_\Sigma(M')$ такой, что $M_\Sigma(M') \in M_\Sigma(M)$, причем M' — не 0-тупиковая.

Следствие 2. *Если в некоторой 0-тупиковой системе масок M имеется ядерная маска $\hat{m}_{i_0}$, то $\hat{m}_{i_0}$ входит во все тупиковые подсистемы M.* Если ядерная $\hat{m}_{i_o}$ представлена в 0-тупиковой M, то при удалении $\hat{m}_{i_0}$ из M произойдет потеря 0-тупиковости M (так как $M_\Sigma(M') \in M_\Sigma(M)$). Поскольку 0-тупиковость — необходимый признак тупиковости (см. следствие 1), то о тупиковости всех подсистем не может быть и речи.

Следствие 3. *Пусть в 0-тупиковой M есть несколько ядерных масок $\hat{m}_{i_1}, \hat{m}_{i_1}, \dots, \hat{m}_{i_L}$ (ядерная подсистема). Если некоторая $\hat{m} \subseteq \bigcup_{j=1.L} \hat{m}_{i_j}$, то $\hat{m}$ не входит ни в одну тупиковую M.* По теореме 5 любая маска в тупиковой M является ядерной и не может быть удалена без потери разрешимости в соответствии с условием (4). Маска $\hat{m} \subseteq \bigcup_{j=1,L} \hat{m}_{i_j}$ может быть удалена без потери разрешимости, так как ее удаление не приводит к изменению объединенной маски $M_\Sigma(M)$.

Теорема 5 и ее следствия полезны для разработки алгоритма построения безызбыточных M. На первом этапе находится 0-тупиковая система масок. Затем производится поиск еще менее избыточных систем масок на основе свойства ядерности: ядерная подсистема входит во все тупиковые системы масок, а маски, покрытые ядерной подсистемой, исключаются. Поиск ограничен снизу тупиковыми системами.

8 Заключение

Существующие методы распознавания вторичной структуры белка имеют в своей основе ряд предположений. В настоящей работе предложен математический формализм, основанный на предварительно проведенном анализе экспериментальных данных. Этот формализм в дальнейшем будет использован для анализа задачи о распознавании вторичной структуры белка с точки зрения алгебраического подхода к проблемам распознавания.

Литература

1. *Torshin I. Y.* Bioinformatics in the post-genomic era: The role of biophysics. — N.Y.: Nova Biomedical Books, 2006.
2. *Berman H. M., Henrick K., Nakamura H.* Announcing the worldwide Protein Data Bank // Nature Structural Biology, 2003. Vol. 10. No. 12. P. 980–982.
3. *Simossis V. F., Herringa J.* Integrating protein secondary structure prediction and multiple sequence alignment // Curr. Protein Pept. Sci., 2004. Vol. 5. No. 2. P. 249–266.
4. *Журавлев Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. I // Кибернетика, 1977. № 4. С. 5–17.
5. *Журавлев Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. II // Кибернетика, 1977. № 6. С. 21–27.
6. *Журавлев Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. III // Кибернетика, 1978. № 2. С. 35–43.
7. *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, 1978. Вып. 33. С. 5–68.
8. *Журавлев Ю. И., Рудаков К. В.* Об алгебраической коррекции процедур обработки (преобразования) информации // Проблемы прикладной математики и информатики. — М.: Наука, 1987. С. 187–198.

9. *Рудаков К. В.* Универсальные и локальные ограничения в проблеме коррекции эвристических алгоритмов // Кибернетика, 1987. № 2. С. 30–35.

10. *Рудаков К. В.* Полнота и универсальные ограничения в проблеме коррекции эвристических алгоритмов классификации // Кибернетика, 1987. № 3. С. 106–109.

11. *Рудаков К. В.* Симметрические и функциональные ограничения в проблеме коррекции эвристических алгоритмов классификации // Кибернетика, 1987. № 4. С. 73–77.

12. *Рудаков К. В.* О применении универсальных ограничений при исследовании алгоритмов классификации // Кибернетика, 1988. № 1. С. 1–5.

13. *Frishman D., Argos P.* Knowledge-based protein secondary structure assignment // Proteins, 1995. Vol. 23. No. 4. P. 566–579.

14. *Torshin I. Yu.* Bioinformatics in the post-genomic era: Sensing the change from molecular genetics to personalized medicine. — N.Y.: Nova Biomedical Books, 2009.

15. *Журавлев Ю. И.* Теоретико-множественные методы в алгебре логики // Проблемы кибернетики, 1962. Т. 8. № 1. С. 25–45.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 36–45

АСИМПТОТИКИ ОЦЕНКИ РИСКА ПРИ ПОРОГОВОЙ ОБРАБОТКЕ ВЕЙВЛЕТ-ВЕЙГЛЕТ КОЭФФИЦИЕНТОВ В ЗАДАЧЕ ТОМОГРАФИИ

А. В. Маркин[1], О. В. Шестаков[2]

Аннотация: Рассмотрена задача реконструкции изображения по радоновскому образу с помощью вейвлет-вейглет разложения. Исследованы свойства оценки риска пороговой обработки вейглет-коэффициентов, такие как состоятельность и асимптотическая нормальность.

Ключевые слова: вейвлеты; томография; пороговая обработка; оценка риска; предельное распределение

1 Введение

Вейвлет-преобразование является весьма популярным и удобным методом обработки нестационарных сигналов и изображений. Одна из основных задач, для которых используются вейвлеты, — удаление шума и сжатие. Эти операции производятся путем пороговой обработки вейвлет-коэффициентов. Кроме того, вейвлеты могут быть использованы для обращения линейных операторов, таких, например, как преобразование Радона. В этом случае пороговая обработка выполняет задачу регуляризации соответствующей формулы обращения.

Пусть на плоскости $(x,\, y)$ задана функция f. Определим образ (или проекции) Радона $\mathcal{R}f$ как набор интегралов от f по всевозможным прямым плоскости

$$\mathcal{R}f(s,\theta) = \int\limits_{L_{s,\theta}} f\,(x,\, y)\; dl\,, \tag{1}$$

где

$$L_{s,\theta} = \{(x,\, y) : x\cos\theta + y\sin\theta - s = 0\}\ .$$

Формула обращения преобразования (1) впервые была получена Радоном, ее можно записать в следующем виде [1]:

$$f = \frac{1}{2}\mathcal{R}^{\#}\mathcal{I}^{-1}\mathcal{R}f\,, \tag{2}$$

где $\mathcal{R}^{\#}$ — оператор обратного проецирования:

$$\left(\mathcal{R}^{\#}g\right)(x,y) = \int\limits_0^{2\pi} g(x\cos\theta + y\sin\theta,\theta)\,d\theta\,;$$

$\mathcal{I}$ — потенциал Рисса:

$$\left(\mathcal{F}_1\mathcal{I}^{\alpha}g\right)(\omega) = |\omega|^{-\alpha}\left(\mathcal{F}_1 g\right)(\omega)\,, \tag{3}$$

а $\mathcal{F}_k$ — k-мерное преобразование Фурье.

Для точного восстановления f требуется точное знание всевозможных проекций $\mathcal{R}f(s,\,\theta)$. На практике же имеют дело с конечным числом проекций, причем в проекциях присутствует шум. При этом задача томографии является некорректной, т. е. малые изменения в проекциях могут привести к восстановлению изображения, существенно отличающегося от исходного. Математически это выражается в наличии множителя $|\omega|$ в формуле (3) (и, следовательно, в (2)), который «подчеркивает» высокие частоты.

Выход видится в регуляризации (2) путем умножения $|\omega|$ на некоторый множитель, называемый частотным фильтром (или стабилизирующим множителем) [2]. Общая идея регуляризации такова: немного «испортить» проекционные данные, подавив влияние высоких частот, но при этом обеспечить реконструкцию, близкую к оригиналу. Подробнее о регуляризации формулы обращения можно прочитать в монографии [3].

2 Вейвлет-вейглет разложение

Задачу томографии можно решить и с помощью вейвлетов. Пусть $\phi(t)$ и $\psi(t)$ — одномерные отцовский и материнский вейвлеты. Определим

$$\phi_{j,k_1,k_2}(x,y) = 2^j\phi\left(2^j x - k_1\right)\phi\left(2^j y - k_2\right)\ ;$$
$$\psi^{[1]}_{j,k_1,k_2}(x,y) = 2^j\phi\left(2^j x - k_1\right)\psi\left(2^j y - k_2\right)\ ;$$

[1]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра математической статистики, artem.v.markin@mail.ru

[2]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра математической статистики, oshestakov@cs.msu.su

$$\psi^{[2]}_{j,k_1,k_2}(x,y) = 2^j\psi\left(2^jx-k_1\right)\phi\left(2^jy-k_2\right)\ ;$$
$$\psi^{[3]}_{j,k_1,k_2}(x,y) = 2^j\psi\left(2^jx-k_1\right)\psi\left(2^jy-k_2\right)\ .$$

Заметим, что параметр масштаба j контролирует сразу обе функции в произведении. Это так называемое тензорное произведение двух одномерных кратномасштабных анализов [4]. Тогда набор функций $\left\{\phi_{j_0,k_1,k_2},\, \psi^{[\lambda]}_{j,k_1,k_2},\right\}$, где $j,\ k_1,\ k_2 \in \mathbb{Z}$, $j \geq j_0$, $\lambda = \overline{1,3}$, будет ортонормированным базисом $\mathbf{L}^2(\mathbb{R}^2)$.

Донохо [5] решил задачу обращения ряда линейных операторов с помощью вейвлетов и родственных им функций специального вида, названных вейглетами (*vaguelettes*). Вейглеты для обращения оператора Радона выглядят так:

$$\xi^{[\lambda]}_{j,k_1,k_2}(s,\,\theta) = \int\limits_{-\infty}^{\infty} |\omega|\left(\mathcal{F}_2\psi^{[\lambda]}_{j,k_1,k_2}\right)\times \\ \times\,(\omega\cos\theta,\,\omega\sin\theta)\exp(i2\pi s\omega)\,d\omega\,.$$

Идея метода реконструкции заключается в том, что вейглет-коэффициенты проекций $\mathcal{R}f(s,\theta)$ равны вейвлет-коэффициентам исходной функции $f(x,y)$:

$$\left[\mathcal{R}f,\,\xi^{[\lambda]}_{j,k_1,k_2}\right] = \left\langle f,\,\psi^{[\lambda]}_{j,k_1,k_2}\right\rangle\,,$$

и поэтому

$$f = \sum_{k_1,k_2}\left[\mathcal{R}f,\,\tau_{j_0,k_1,k_2}\right]\phi_{j_0,k_1,k_2} + \\ + \sum_{j\geqslant j_0,k_1,k_2,\lambda}\left[\mathcal{R}f,\,\xi^{[\lambda]}_{j,k_1,k_2}\right]\psi^{[\lambda]}_{j,k_1,k_2}\,, \quad (4)$$

где

$$\tau_{j_0,k_1,k_2}(s,\,\theta) = \int\limits_{-\infty}^{\infty}|\omega|\left(\mathcal{F}_2\phi_{j_0,k_1,k_2}\right)\times \\ \times\,(\omega\cos\theta,\,\omega\sin\theta)\exp\left(i2\pi s\omega\right)\,d\omega\,.$$

Регуляризация вейвлет-вейглет формулы (4) производится с помощью мягкой пороговой обработки вейглет-коэффициентов (см. разд. 4).

3 Дискретизация и модель шума

Пусть функция $f(x,y)$ задана на квадрате $[0,\,1]\times$ $\times\,[0,\,1]$. Разбив стороны квадрата на $N = 2^J$ равных частей и вычислив значения f в точках отсчета, получим дискретизованную версию f. Однако на практике нередко бывает удобно нормировать длину отрезка разбиения и рассматривать вместо f ее «растянутую» версию — функцию $\bar{f}(Nx,Ny) = f(x,y)$. Тогда для вейвлет-коэффициентов функции f справедливо равенство:

$$\left\langle f,\,\psi^{[\lambda]}_{j,u_1,u_2}\right\rangle = \\ = \iint f(x,y)\,2^j\overline{\psi^{[\lambda]}\left(2^jx-k_1,\,2^jy-k_2\right)}\,dx\,dy = \\ = \left(\mathcal{W}^{[\lambda]}f\right)\left(2^{-j},k_1,k_2\right) = \\ = \frac{1}{N}\left(\mathcal{W}^{[\lambda]}\bar{f}\right)\left(N\,2^{-j},k_1,k_2\right)\ . \quad (5)$$

Заметим, что при работе с растянутой функцией растягиваются и вейвлет-функции. Коэффициенты аппроксимации, получаемые через скалярное произведение f и ϕ, не рассматриваются, так как пороговая обработка (см. разд. 4) применяется к коэффициентам деталей, которые дают функции $\psi^{[\lambda]}$. Далее везде, кроме разд. 5, предполагается, что используются именно коэффициенты растянутой версии функции f.

Задача томографии ставится следующим образом. Имеются наблюдения X, состоящие из проекций $\mathcal{R}f$ функции f и шума ϵ:

$$X = \mathcal{R}f + \epsilon\,,$$

ϵ — независимые нормальные случайные величины с нулевым средним и дисперсией σ^2. Необходимо восстановить f по X. При этом при достаточно большом N [6]

$$\left.\begin{aligned} &\mathsf{E}\left[X,\,\xi^{[\lambda]}_{j,k_1,k_2}\right] = \left[\mathcal{R}f,\,\xi^{[\lambda]}_{j,k_1,k_2}\right]\ ; \\ &\mathsf{D}\left[X,\,\xi^{[\lambda]}_{j,k_1,k_2}\right] = \sigma^2\left\|\xi^{[\lambda]}_{j,k_1,k_2}\right\|_2^2 = \sigma^2_{\lambda;j}\ ; \\ &\left\|\xi^{[\lambda]}_{j,k_1,k_2}\right\|_2^2 = 2^j\left\|\xi^{[\lambda]}_{0,0,0}\right\|_2^2\ . \end{aligned}\right\} \quad (6)$$

Как видим, дисперсия коэффициентов растет вместе с уровнем разложения. Это является следствием некорректности задачи томографии. При этом вейглеты не ортогональны, а почти ортогональны. И, стало быть, вейглет-коэффициенты не независимы, а почти независимы. Однако нередко этим фактом пренебрегают, так как исследование этой зависимости сопряжено с рядом трудностей. И потому порог выбирается исходя из предположения независимости коэффициентов. Как будет видно далее, уже только тот факт, что дисперсия растет на каждом уровне, заметно влияет на оценку риска пороговой обработки.

4 Пороговая обработка

Мягкая пороговая функция определяется следующим образом:

$$\rho(x,T)=\begin{cases}x-T & \text{при } x>T\,;\\ x+T & \text{при } x<-T\,;\\ 0 & \text{при } |x|\leq T\,.\end{cases}$$

Эта функция применяется к вейглет-коэффициентам проекций.

Допустим, что размер изображения равен $N^2 = 2^{2J}=L$, разложение идет до уровня $J-1$. В качестве порога взят порог Колашика [6, 7]:

$$T_{\lambda;j}=\sqrt{2\ln 2^{2j}}\,2^{j/2}\sigma\left\|\xi^{[\lambda]}_{0,0,0}\right\|_2\,.$$

В случае использования оценки дисперсии шума $\hat\sigma^2$ порог принимает вид

$$\hat T_{\lambda;j}=\sqrt{2\ln 2^{2j}}\,2^{j/2}\hat\sigma\left\|\xi^{[\lambda]}_{0,0,0}\right\|_2\,.$$

Идея выбора такого порога схожа с идеей выбора порога *VisuShrink* $T=\sigma\sqrt{2\ln N}$ (одномерный случай, N — размер сигнала): при таком пороге убирается почти весь шум [8, 9]. Это следует из того факта, что если $Z_1,\ldots,Z_N$ — независимые стандартные нормальные случайные величины, то

$$\mathsf{P}\left(\max_{1\leqslant i\leqslant N}|Z_i|>\sqrt{2\ln N}\right)\to 0 \text{ при } N\to\infty\,.$$

Пороговая обработка идет с уровня j_M, т. е. в формуле (4) $j_0=j_M$ (j_M определим ниже). Риск $r(f)$ такой пороговой обработки определяется следующим образом:

$$r(f)=\sum_{j=j_M}^{J-1}\sum_{\lambda=1}^{3}\sum_{k_1=0}^{2^j-1}\sum_{k_2=0}^{2^j-1}\mathsf{E}\left\{\left\langle f,\,\psi^{[\lambda]}_{j,k_1,k_2}\right\rangle-\right.$$
$$\left.-\,\rho\left(\left[X,\,\xi^{[\lambda]}_{j,k_1,k_2}\right],\,T_{\lambda;j}\right)\right\}^2\,. \quad (7)$$

Так как на практике коэффициенты $\left\langle f,\,\psi^{[\lambda]}_{j,k_1,k_2}\right\rangle$ неизвестны, то строят оценку риска. Например, на основе функции $\Phi(x,T)$ [10]:

$$\Phi(x,T_{\lambda;j})=\begin{cases}x-\sigma^2_{\lambda;j} & \text{при } x\leqslant T^2_{\lambda;j}\,;\\ \sigma^2_{\lambda;j}+T^2_{\lambda;j} & \text{при } x>T^2_{\lambda;j}\,.\end{cases}$$

Оценка риска принимает вид:

$$\tilde r(f)=\sum_{j=j_M}^{J-1}\sum_{\lambda,k_1,k_2}\Phi\left(\left|\left[X,\,\xi^{[\lambda]}_{j,k_1,k_2}\right]\right|^2,\,T_{\lambda;j}\right)\,.$$

Если вместо σ^2 используется оценка $\hat\sigma^2$, то

$$\hat r(f)=\sum_{j=j_M}^{J-1}\sum_{\lambda,k_1,k_2}\hat\Phi\left(\left|\left[X,\,\xi^{[\lambda]}_{j,k_1,k_2}\right]\right|^2,\,\hat T_{\lambda;j}\right)\,,$$

где

$$\hat\Phi(x,\hat T_{\lambda;j})=\begin{cases}x-\hat\sigma^2_{\lambda;j} & \text{при } x\leqslant \hat T^2_{\lambda;j}\,;\\ \hat\sigma^2_{\lambda;j}+\hat T^2_{\lambda;j} & \text{при } x>\hat T^2_{\lambda;j}\,.\end{cases}$$

В работах [12, 11] рассмотрены асимптотические свойства оценки риска пороговой обработки вейвлет-коэффициентов в одномерном случае при прямом наблюдении f. Показано, что $(\hat r-r)/N^a$ сходится по вероятности к нулю и по распределению к нормальному закону при соответствующих a. Величина a существенно зависит от свойств оценки $\hat\sigma^2$. Однако даже при весьма общих ограничениях на моменты $\hat\sigma^2$ порядок $a=1$ обеспечивал сходимость по вероятности к нулю. Ниже будет показано, что в задаче томографии для сходимости по вероятности к нулю недостаточно делить на число коэффициентов ($N^2=L$), т. е. некоторый аналог закона больших чисел уже не выполнен. Важнейшим фактором является то, что f наблюдается через оператор Радона $\mathcal{R}$, обратный к которому не является непрерывным (т. е. ограниченным).

5 Регулярность функции и вейвлет-коэффициенты

Известно (см., например, [10]), что если функция $f(x,y)$ является регулярной по Липшицу с параметром $0\leq\alpha\leq 1$, т. е.

$$|f(x_1,y_1)-f(x_2,y_2)|\leq$$
$$\leq C\left(|x_1-x_2|^2+|y_1-y_2|^2\right)^{\alpha/2}$$

для некоторой константы C, не зависящей от (x_1,y_1) и (x_2,y_2), то существует не зависящая от J, j, k_1 и k_2 константа A такая, что

$$\left(\mathcal{W}^{[\lambda]}f\right)\left(2^{-j},k_1,k_2\right)\leq\frac{A}{2^{j(\alpha+1)}}\,.$$

В отечественной литературе вместо регулярности по Липшицу обычно используется термин «непрерывность по Гёльдеру». С учетом (5) получаем

$$\left(\mathcal{W}^{[\lambda]}\bar f\right)\left(N\cdot 2^{-j},k_1,k_2\right)\leqslant\frac{A\cdot 2^J}{2^{j(\alpha+1)}}\,.$$

Предположение о регулярности f: будем полагать, что функция f является регулярной по Липшицу с показателем $\alpha>0$. Будем считать, что пороговая

обработка ведется с уровня $j_M \geq J/(\alpha+1)$. Заметим, что $J - j_M \to \infty$ при $J \to \infty$. Тогда при определенном выборе вейвлет-базиса [10] найдется константа C_1 такая, что для всех $j \geq j_M$ выполнено

$$\left(\mathcal{W}^{[\lambda]}\bar{f}\right)\left(N\cdot 2^{-j}, k_1, k_2\right) \leqslant C_1\,, \qquad (8)$$

причем C_1 не зависит от N. Значит, математические ожидания в (6) ограничены.

В работе используется буква C (с индексом или без индекса) для обозначения констант, причем в разных местах — вообще говоря, разных.

6 Асимптотика оценки риска при известной дисперсии шума

В работе [11] показано, что в одномерном случае при известной дисперсии шума разность риска и оценки риска при делении на $\sqrt{N}$ сходится по распределению к нормальному закону. В задаче томографии уже надо делить не на $\sqrt{L}$, а на L.

Для краткости введем обозначения:

$$Y_{\lambda;j,\mathbf{k}} = \left[X,\, \xi^{[\lambda]}_{j,k_1,k_2}\right];$$
$$\mu_{\lambda;j,\mathbf{k}} = \left\langle f,\, \psi^{[\lambda]}_{j,k_1,k_2}\right\rangle,$$

где $\mathbf{k} = (k_1,\, k_2)$. Еще раз напомним, что $\mu_{\lambda;j,\mathbf{k}}$ рассматриваются как коэффициенты растянутой версии дискретизованной функции f. С учетом предположения об ортогональности вейглетов получаем

$$Y_{\lambda;j,\mathbf{k}} \sim \mathcal{N}\left(\mu_{\lambda;j,\mathbf{k}},\, \sigma^2_{\lambda;j}\right), \qquad (9)$$

причем $Y_{\lambda;j,\mathbf{k}}$ — независимые случайные величины.

Теорема 1. *Пусть справедливы предположения о регулярности f из разд. 5. При известной дисперсии шума в задаче томографии*

$$\frac{\tilde{r}(f) - r(f)}{L\sqrt{b_2\left(\sigma^4_{1;0}+\sigma^4_{2;0}+\sigma^4_{3;0}\right)}} \Rightarrow \mathcal{N}(0,\,1)$$

при $L\to\infty$, где $b_2 = 2/(2^4-1) = 2/15$.

Доказательство. Представим разность оценки риска и самого риска в виде

$$\begin{aligned}
\tilde{r} - r = &\sum_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|\leqslant T_{\lambda;j}} + \\
&+ \sum_{\lambda,j,\mathbf{k}} \left(\sigma^2_{\lambda;j} + T^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|> T_{\lambda;j}} - \\
&- \sum_{\lambda,j,\mathbf{k}} \mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|\leqslant T_{\lambda;j}} - \\
&- \sum_{\lambda,j,\mathbf{k}} \mathsf{E}\left(\sigma^2_{\lambda;j} + T^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|> T_{\lambda;j}} = \\
&= \sum_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\right) - \\
&- \sum_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|> T_{\lambda;j}} + \\
&+ \sum_{\lambda,j,\mathbf{k}} \mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|> T_{\lambda;j}} + \\
&+ \sum_{\lambda,j,\mathbf{k}} \left(\sigma^2_{\lambda;j} + T^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|> T_{\lambda;j}} - \\
&- \sum_{\lambda,j,\mathbf{k}} \left(\sigma^2_{\lambda;j} + T^2_{\lambda;j}\right)\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}| > T_{\lambda;j}\right). \qquad (10)
\end{aligned}$$

Покажем, что при делении на L первая сумма в (10) сходится по распределению к нормальному закону, а остальные суммы — к нулю по вероятности.

Итак, рассмотрим первую сумму в (10). Имеем

$$\begin{aligned}
D^2_L = \mathsf{D}\sum_{\lambda,j,\mathbf{k}} Y^2_{\lambda;j,\mathbf{k}} &= \sum_{\lambda,j,\mathbf{k}} \mathsf{D}Y^2_{\lambda;j,\mathbf{k}} = \\
&= \sum_{\lambda}\sum_{j=j_M}^{J-1}\sum_{\mathbf{k}}\left(2\sigma^4_{\lambda;j} + 4\mu^2_{\lambda;j,\mathbf{k}}\sigma^2_{\lambda;j}\right) = \\
= \sum_{\lambda}\sum_{j=j_M}^{J-1}&\left\{2\cdot 2^{2j}\sigma^4_{\lambda;0}\cdot 2^{2j} + \sum_{\mathbf{k}} 4\mu^2_{\lambda;j,\mathbf{k}}2^j\sigma^2_{\lambda;0}\right\} \simeq \\
\simeq \sum_{\lambda}\sum_{j=j_M}^{J-1} 2\cdot 2^{4j}\sigma^4_{\lambda;0} &= \sum_{\lambda} 2\sigma^4_{\lambda;0}\frac{2^{4J}-2^{4j_M}}{2^4-1} \simeq \\
&\simeq \frac{2}{15}2^{4J}\left(\sigma^4_{1;0}+\sigma^4_{2;0}+\sigma^4_{3;0}\right). \qquad (11)
\end{aligned}$$

Знак $\simeq$ означает, что при $J\to\infty$ предел отношения левой и правой частей (11) равен единице. Если выполнено условие Линдеберга, т. е. для любого $\delta > 0$

$$\frac{1}{D^2_L}\sum_{\lambda,j,\mathbf{k}} \mathsf{E}\left\{\left(Y^2_{\lambda;j,\mathbf{k}} - \mu^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)^2 \times \right.$$
$$\left. \times\, \mathbb{1}_{|Y^2_{\lambda;j,\mathbf{k}} - \mu^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}| > \delta D_L}\right\} \to 0\,, \qquad (12)$$

то будет иметь место сходимость к нормальному распределению. Так как D_L имеет порядок L и число слагаемых в (12) имеет порядок L, то достаточно показать, что при $L\to\infty$

$$\mathsf{E}\left\{\frac{\left(Y^2_{\lambda;j,\mathbf{k}} - \mu^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)^2}{D_L} \times \right.$$
$$\left. \times\, \mathbb{1}_{\left(Y^2_{\lambda;j,\mathbf{k}} - \mu^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)^2/D_L > \delta^2 D_L}\right\} \to 0\,.$$

А последнее выполнено потому, что у случайных величин вида $\left(Y^2_{\lambda;j,\mathbf{k}} - \mu^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)^2 / D_L$ конечные математические ожидания и $D_L \to \infty$.

Теперь рассмотрим вторую сумму в (10). В силу (9) имеем

$$\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}| > T_{\lambda;j}\right) < \\ < \frac{\exp\left(-(T_{\lambda;j} - \mu_{\lambda;j,\mathbf{k}})^2/(2\sigma^2_{\lambda;j})\right)}{T_{\lambda;j}} + \\ + \frac{\exp\left(-(T_{\lambda;j} + \mu_{\lambda;j,\mathbf{k}})^2/(2\sigma^2_{\lambda;j})\right)}{T_{\lambda;j}} \leqslant \frac{C}{2^{5j/2}\sqrt{j}}$$

при $J \to \infty$ (и, следовательно, $j \to \infty$). Это можно получить из следующей цепочки равенств:

$$\exp\left(-\frac{(T_{\lambda;j} - \mu_{\lambda;j,\mathbf{k}})^2}{2\sigma^2_{\lambda;j}}\right) = \\ = \exp\left(-\frac{T^2_{\lambda;j}}{2\sigma^2_{\lambda;j}} + \frac{2T_{\lambda;j}\mu_{\lambda;j,\mathbf{k}}}{2\sigma^2_{\lambda;j}} - \frac{\mu^2_{\lambda;j,\mathbf{k}}}{2\sigma^2_{\lambda;j}}\right) = \\ = \exp\left(-\ln 2^{2j} + \frac{\sqrt{2\ln(2^{2j})}\mu_{\lambda;j,\mathbf{k}}}{2^{j/2}\sigma_{\lambda;0}} - \frac{\mu^2_{\lambda;j,\mathbf{k}}}{2\sigma^2_{\lambda;j}}\right) \simeq \\ \simeq 2^{-2j} \text{ при } j \to \infty\,,$$

так как

$$\frac{\sqrt{2\ln(2^{2j})}\mu_{\lambda;j,\mathbf{k}}}{2^{j/2}\sigma_{\lambda;0}} \to 0 \quad \text{и} \quad \frac{\mu^2_{\lambda;j,\mathbf{k}}}{2\sigma^2_{\lambda;j}} \to 0\,.$$

С помощью неравенств Чебышёва и Коши–Буняковского получаем для любого $\delta > 0$ при $J \to \infty$

$$\mathsf{P}\left(\frac{\left|\sum\limits_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right) \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > T_{\lambda;j}}\right|}{D_L} > \delta\right) \leq \\ \leq \frac{\mathsf{E}\left|\sum\limits_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right) \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > T_{\lambda;j}}\right|}{\delta D_L} \leq \\ \leq \frac{\sum\limits_{\lambda,j,\mathbf{k}} \mathsf{E}\left|Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right| \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > T_{\lambda;j}}}{\delta D_L} \leq \\ \leq \frac{\sum\limits_{\lambda,j,\mathbf{k}} \sqrt{\mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)^2 \mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}| > T_{\lambda;j}\right)}}{\delta D_L} \leq \\ \leq \frac{1}{\delta D_L} \sum_{\lambda,j,\mathbf{k}} \Big(\left(\mu^4_{\lambda;j,\mathbf{k}} + 2 \cdot 2^{2j}\sigma^4_{\lambda;0} + \right. \\ \left. + 4 \cdot 2^j \sigma^2_{\lambda;0}\mu^2_{\lambda;j,\mathbf{k}}\right) C \cdot 2^{-5j/2} j^{-1/2}\Big)^{1/2} \to 0\,.$$

Аналогично проводятся рассуждения для оставшихся сумм в (10). □

7 Свойства оценки риска при использовании оценки дисперсии шума

В работе [12] показано, что при достаточно слабых ограничениях на моменты оценки дисперсии шума для сходимости разности риска и его оценки к нулю по вероятности ее надо нормировать числом вейвлет-коэффициентов, т. е. порядок знаменателя вырастает почти на 1/2. Покажем, что в задаче томографии порядок тоже повышается почти на 1/2, но знаменатель уже будет много больше числа коэффициентов.

Введем обозначение

$$\hat{\sigma}^2_{\lambda;j} = 2^j \hat{\sigma}^2 \left\|\xi^{[\lambda]}_{0,0,0}\right\|_2^2\,.$$

Теорема 2. *Пусть справедливы предположения о регулярности f. Пусть $\hat{\sigma}^2$ — оценка дисперсии,* $\mathsf{E}\hat{\sigma}^2 - \sigma^2 = \nu_L$ *и* $\mathsf{D}\hat{\sigma}^2 = \theta_L = O(L^{-\beta})$, $\nu_L = o(1)$, $\beta > 0$. *Тогда при $L \to \infty$ выполнено*

$$\frac{\hat{r}(f) - r(f)}{L^{3/2}} \xrightarrow{\mathsf{P}} 0\,. \tag{13}$$

Доказательство. Подобно доказательству теоремы 3 в [12] запишем

$$\hat{r} - r = S_1 + S_2\,,$$

где

$$S_1 = \sum_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \hat{\sigma}^2_{\lambda;j}\right) - \\ - \sum_{\lambda,j,\mathbf{k}} \mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}} - \sigma^2_{\lambda;j}\right)\,; \tag{14}$$

$$S_2 = -\sum_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \hat{\sigma}^2_{\lambda;j}\right) \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}} + \\ + \sum_{\lambda,j,\mathbf{k}} \left(\hat{\sigma}^2_{\lambda;j} + \hat{T}^2_{\lambda;j}\right) \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}} +$$

$$+\sum_{\lambda,j,\mathbf{k}} \mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}}-\sigma^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}} - \\ -\sum_{\lambda,j,\mathbf{k}} \mathsf{E}\left(\sigma^2_{\lambda;j}+T^2_{\lambda;j}\right)\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}}. \quad (15)$$

Далее будет показано, что при делении на $L^{3/2}$ и S_1, и S_2 сходятся к нулю по вероятности.

Сначала рассмотрим S_1: по неравенству Чебышёва при любом $\delta>0$

$$\mathsf{P}\left(\frac{|S_1|}{L^{3/2}}>\delta\right)\leq \\ \leq \frac{\mathsf{E}\left(\sum\limits_{\lambda,j,\mathbf{k}}\left(Y^2_{\lambda;j,\mathbf{k}}-\hat\sigma^2_{\lambda;j}-\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}+\sigma^2_{\lambda;j}\right)\right)^2}{\delta^2L^3}= \\ =\frac{\sum\limits_{\lambda,j,\mathbf{k}}\mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}}-\hat\sigma^2_{\lambda;j}-\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}+\sigma^2_{\lambda;j}\right)^2}{\delta^2L^3}+ \\ +\frac{1}{\delta^2L^3}\sum\mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}}-\hat\sigma^2_{\lambda;j}-\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}+\sigma^2_{\lambda;j}\right)\times \\ \times\left(Y^2_{\lambda';j',\mathbf{k}'}-\hat\sigma^2_{\lambda';j'}-\mathsf{E}Y^2_{\lambda';j',\mathbf{k}'}+\sigma^2_{\lambda';j'}\right). \quad (16)$$

Во второй сумме (16) суммирование идет по индексам $(\lambda,j,\mathbf{k})\neq(\lambda',j',\mathbf{k}')$. Понятно, что первое слагаемое в (16) стремится к нулю — в сумме всего порядка L слагаемых, они имеют порядок не выше L и сумма делится на L^3 (напомним, что $L=2^{2J}$).

Рассмотрим одно из слагаемых второй суммы (16):

$$\mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}}-\hat\sigma^2_{\lambda;j}-\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}+\sigma^2_{\lambda;j}\right)\times \\ \times\left(Y^2_{\lambda';j',\mathbf{k}'}-\hat\sigma^2_{\lambda';j'}-\mathsf{E}Y^2_{\lambda';j',\mathbf{k}'}+\sigma^2_{\lambda';j'}\right)= \\ =\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}Y^2_{\lambda';j',\mathbf{k}'}-\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\hat\sigma^2_{\lambda';j'}-\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathsf{E}Y^2_{\lambda';j',\mathbf{k}'}+ \\ +\sigma^2_{\lambda';j'}\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}-\mathsf{E}\hat\sigma^2_{\lambda;j}Y^2_{\lambda';j',\mathbf{k}'}+\mathsf{E}\hat\sigma^2_{\lambda;j}\hat\sigma^2_{\lambda';j'}+ \\ +\mathsf{E}\hat\sigma^2_{\lambda;j}\mathsf{E}Y^2_{\lambda';j',\mathbf{k}'}-\sigma^2_{\lambda';j'}\mathsf{E}\hat\sigma^2_{\lambda;j}-\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathsf{E}Y^2_{\lambda';j',\mathbf{k}'}+ \\ +\mathsf{E}\hat\sigma^2_{\lambda';j'}\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}+\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathsf{E}Y^2_{\lambda';j',\mathbf{k}'}-\sigma^2_{\lambda';j'}\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}+ \\ +\sigma^2_{\lambda;j}\mathsf{E}Y^2_{\lambda';j',\mathbf{k}'}-\sigma^2_{\lambda;j}\mathsf{E}\hat\sigma^2_{\lambda';j'}-\sigma^2_{\lambda;j}\mathsf{E}Y^2_{\lambda';j',\mathbf{k}'}+ \\ +\sigma^2_{\lambda;j}\sigma^2_{\lambda';j'}=-\operatorname{cov}\left(\hat\sigma^2_{\lambda';j'},\,Y^2_{\lambda;j,\mathbf{k}}\right)- \\ -\operatorname{cov}\left(\hat\sigma^2_{\lambda;j},\,Y^2_{\lambda';j',\mathbf{k}'}\right)+\frac{\sigma^2_{\lambda;j}\sigma^2_{\lambda';j'}}{\sigma^4}\left(\nu_L^2+\theta_L\right).$$

С учетом того, что $\mathsf{D}Y^2_{\lambda;j,\mathbf{k}}$ имеет порядок 2^{2j}, а ковариацию можно оценить по неравенству Коши–Буняковского, получаем, что каждое слагаемое второй суммы (16) можно оценить как $2^{j+j'}\cdot\mathsf{o}(1)$. Всего таких слагаемых порядка L^2, а максимальное значение $2^{j+j'}$ равно $2^{J-1+J-1}=L/4$. Следовательно, после суммирования получаем, что второе слагаемое в (16) оценивается как $\mathsf{o}(1)$. Значит, $S_1/L^{3/2}$ сходится к нулю по вероятности.

Для оценки S_2 используем другую модификацию неравенства Чебышёва:

$$\mathsf{P}\left(\frac{|S_2|}{L^{3/2}}>\delta\right)\leq\frac{\mathsf{E}|S_2|}{\delta L^{3/2}}=\frac{\mathsf{E}\left[|S_2|/L^{1/2}\right]}{\delta L}.$$

Величину $\mathsf{E}|S_2|$ можно оценить сверху суммой математических ожиданий модулей сумм, входящих в S_2, а эти суммы, в свою очередь, — суммой математических ожиданий входящих в них слагаемых.

По формуле полной вероятности для некоторого $0<\gamma<1$ получаем

$$\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}\right)= \\ =\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}\,|\,\hat T_{\lambda;j}\leqslant(1-\gamma)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right)\times \\ \times\mathsf{P}\left(\hat T_{\lambda;j}\leq(1-\gamma)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right)+ \\ +\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}\,,\,\hat T_{\lambda;j}>(1-\gamma)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right).$$

В силу свойств $\hat\sigma^2$

$$\mathsf{P}\left(\hat T_{\lambda;j}\leq(1-\gamma)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right)= \\ =\mathsf{P}\left(\hat\sigma^2\leqslant(1-\gamma)^2\sigma^2\right)\leq \\ \leq\mathsf{P}\left(|\hat\sigma^2-\sigma^2-\nu_L|\geqslant(2\gamma-\gamma^2)\sigma^2+\nu_L\right)\leq \\ \leq\frac{\mathsf{D}\hat\sigma^2}{\left((2\gamma-\gamma^2)\sigma^2+\nu_L\right)^2}=\mathsf{O}\left(L^{-\beta}\right)$$

для достаточно большого L. Далее

$$\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}\,,\,\hat T_{\lambda;j}>(1-\gamma)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right)\leq \\ \leq\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>(1-\gamma)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right)= \\ =\frac{C}{2^{2j(1-\gamma)^2}\cdot 2^{j/2}\sqrt{j}}. \quad (17)$$

Теперь оцениваем математические ожидания компонентов сумм из S_2 при делении на $L^{1/2}$:

$$\mathsf{E}\left[\frac{\left|Y^2_{\lambda;j,\mathbf{k}}-\hat\sigma^2_{\lambda;j}\right|}{L^{1/2}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}}\right]\leq \\ \leq\sqrt{\mathsf{E}\left[\frac{\left(Y^2_{\lambda;j,\mathbf{k}}-\hat\sigma^2_{\lambda;j}\right)^2}{2^{2J}}\right]\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}\right)}\to 0;$$

$$\mathsf{E}\left[\frac{\left|\hat\sigma^2_{\lambda;j}+\hat T^2_{\lambda;j}\right|}{L^{1/2}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}}\right]\leq \\ \leq\left(2\ln 2^{2j}+1\right)2^{j-J}\times \\ \times\sqrt{\mathsf{E}\left(\hat\sigma^2_{\lambda;0}\right)^2\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}\right)}\to 0$$

при $j \geq j_M$ и $J \to \infty$. Остальные слагаемые оцениваются аналогично. Итак, $S_2/L^{3/2}$ тоже сходится к нулю по вероятности. □

Как и в одномерном случае (см. [11]), порядок знаменателя в (13) можно понизить, введя дополнительные ограничения на ν_L.

Теорема 3. *Пусть справедливы предположения о регулярности f. Пусть $\hat\sigma^2$ — оценка дисперсии,* $\mathsf{E}\hat\sigma^2 - \sigma^2 = \nu_L = O(L^{-\upsilon})$ *и* $\mathsf{D}\hat\sigma^2 = \theta_L = O(L^{-\beta})$, υ, $\beta > 0$. *Тогда при любом* $a > 1/2 - c$, $c = \min\{1/2, \upsilon, \beta/2\}$ *и* $L \to \infty$ *выполнено*

$$\frac{\hat r(f) - r(f)}{L^{a+1}} \xrightarrow{\mathsf{P}} 0\,.$$

Доказательство. Заметим, что $0 < c \leq 1/2$ и, стало быть, $a > 0$. Так же, как и в доказательстве теоремы 2, разобьем $\hat r - r$ на те же суммы S_1 и S_2 (см. формулы (14) и (15)), только S_1 запишем в виде

$$S_1 = \sum_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\right) - \sum_{\lambda,j,\mathbf{k}} \left(\hat\sigma^2_{\lambda;j} - \sigma^2_{\lambda;j}\right) =$$
$$= \sum_{\lambda,j,\mathbf{k}} \left(Y^2_{\lambda;j,\mathbf{k}} - \mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\right) -$$
$$- \sum_{\lambda} \sum_{j} 2^{2j}\, 2^j \left(\hat\sigma^2_{\lambda;0} - \sigma^2_{\lambda;0}\right)\,.$$

Первая сумма при делении на L сходится по распределению к нормальному закону (см. разд. 6) и, следовательно, сходится по вероятности к нулю при делении на L^{a+1}, где $a > 0$. Вторая сумма представляет собой произведение $\left(\hat\sigma^2 - \sigma^2\right)$ и множителя, имеющего порядок $2^{3J} = L^{3/2}$. Легко видеть, что

$$\frac{L^{3/2}\left(\hat\sigma^2 - \sigma^2\right)}{L^{a+1}} \xrightarrow{\mathsf{P}} 0$$

при указанных в формулировке теоремы ограничениях на a.

Покажем теперь, что S_2/L^{a+1} сходится к нулю по вероятности. Обозначим $\varkappa = a - 1/2 + c > 0$. В теореме 2 есть оценки для вероятности $\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}| > \hat T_{\lambda;j}\right)$:

$$\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}| > \hat T_{\lambda;j}\right) =$$
$$= \max\left\{\frac{C_1}{2^{2J\beta}},\, \frac{C_2}{2^{2j(1-\gamma)^2}\cdot 2^{j/2}\sqrt{j}}\right\} \quad (18)$$

для некоторого $0 < \gamma < 1$. При $J \to \infty$ имеем

$$\frac{\mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}}\right)^2 C_1/2^{2J\beta}}{L^{2a}} \leq \frac{C_3\cdot 2^{2j}\cdot 2^{-2j\beta}}{2^{2J(1-2c+2\varkappa)}} =$$
$$= \frac{C_3\cdot 2^{2j}\cdot 2^{2J\min\{1,2\upsilon,\beta\}}}{2^{2J}\cdot 2^{2J\beta}\cdot 2^{4J\varkappa}} \to 0, \quad (19)$$

$$\frac{\mathsf{E}\left(Y^2_{\lambda;j,\mathbf{k}}\right)^2 C_2\cdot 2^{-2j(1-\gamma)^2}\cdot 2^{-j/2}/\sqrt{j}}{L^{2a}} \leq$$
$$\leq \frac{C_4\cdot 2^{2j}\cdot 2^{2J\min\{1,2\upsilon,\beta\}}}{2^{2j(1-\gamma)^2+j/2}\cdot 2^{2J}\cdot 2^{4J\varkappa}\sqrt{j}} \to 0 \quad (20)$$

для достаточно малого γ. Отсюда имеем для произвольного $\delta > 0$

$$\mathsf{P}\left(\frac{\sum\limits_{\lambda,j,\mathbf{k}} Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}}}{L^{a+1}} > \delta\right) \leq$$
$$\leq \frac{\sum\limits_{\lambda,j,\mathbf{k}} \mathsf{E}\left[Y^2_{\lambda;j,\mathbf{k}}/L^a\right]\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}}}{\delta L} \to 0$$

при $J \to \infty$ в силу неравенств Чебышёва и Коши–Буняковского. Оценки для суммы с членами вида $\hat\sigma^2_{\lambda;j}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}}$ получаются аналогично. А для сумм, в которые входят $\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}}$, оценки получены в теореме 1. □

Можно сформулировать и доказать теорему сходимости по распределению к нетривиальному пределу.

Теорема 4. *Пусть справедливы предположения о регулярности f. Пусть $\hat\sigma^2$ — оценка дисперсии,* $\mathsf{E}\hat\sigma^2 - \sigma^2 = \nu_L = O(L^{-\upsilon})$ *и* $\mathsf{D}\hat\sigma^2 = \theta_L = O(L^{-\beta})$, $\upsilon > 0$, $\beta > 1/2$. *Пусть $\hat\sigma^2$ не зависит от $Y_{\lambda;j,\mathbf{k}}$ и* $\sqrt{L}\left(\hat\sigma^2 - \sigma^2\right) \Rightarrow \mathcal{N}\left(0, \Sigma^2\right)$ *при $L \to \infty$, тогда*

$$\frac{\hat r(f) - r(f)}{L\sqrt{b_2\left(\sigma^4_{1;0} + \sigma^4_{2;0} + \sigma^4_{3;0}\right)}} \Rightarrow$$
$$\Rightarrow \mathcal{N}\left(0, 1 + \frac{\left(\sigma^2_{1;0} + \sigma^2_{2;0} + \sigma^2_{3;0}\right)^2\Sigma^2}{d_2\,\sigma^4\left(\sigma^4_{1;0} + \sigma^4_{2;0} + \sigma^4_{3;0}\right)}\right),$$

где $b_2 = 2/(2^4-1) = 2/15$, $d_2 = (2(2^3-1)^2)/(2^4-1) = 98/15$.

Доказательство. В теореме 3 было существенным наличие $\varkappa > 0$, которое давало сходимость к нулю в (19) (в (20) это несущественно). Сейчас же $\varkappa = 0$, поэтому доказательство необходимо изменить.

Оценим S_2 более тонко. Имеем

$$Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}} - \mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}} =$$
$$= Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat T_{\lambda;j}} - Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}} +$$
$$+ Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}} - \mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}}\,.$$

Вопрос о двух последних слагаемых решен в теореме 1. Рассмотрим два первых:

$$\mathsf{E}\left|Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat{T}_{\lambda;j}} - Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}}\right| = \\ = \mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{T_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leqslant\hat{T}_{\lambda;j}} + \\ + \mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{\hat{T}_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leqslant T_{\lambda;j}}\,.$$

При этом

$$\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{T_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leqslant\hat{T}_{\lambda;j}} \leq \mathsf{E}\hat{T}^2_{\lambda;j}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}} \leq \\ \leq \sqrt{\frac{C\cdot j^2\cdot 2^{2j}}{2^{2j+j/2}\sqrt{j}}} \to 0, \quad J\to\infty\,,$$

и

$$\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{\hat{T}_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leq T_{\lambda;j}} \leq \\ \leq T^2_{\lambda;j}\mathsf{E}\mathbb{1}_{\hat{T}_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leq T_{\lambda;j}} \leq \\ \leq Cj\cdot 2^j\mathsf{E}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat{T}_{\lambda;j}}\,. \quad (21)$$

С учетом (18) получаем, что

$$\mathsf{E}Y^2_{\lambda;j,\mathbf{k}}\mathbb{1}_{\hat{T}_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leqslant T_{\lambda;j}} \to 0$$

при $J\to\infty$ и $\beta>1/2$. Отметим, что, в отличие от работы [11], требование на β повысилось (там требовалось только $\beta>0$). Это является следствием роста дисперсии с ростом j, которое выражается в наличии множителя 2^j в (21). Аналогично получаем соотношения для $\hat{T}_{\lambda;j}$:

$$\mathsf{E}\hat{T}^2_{\lambda;j}\mathbb{1}_{T_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leq\hat{T}_{\lambda;j}} \leq \mathsf{E}\hat{T}^2_{\lambda;j}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>T_{\lambda;j}} \leq \\ \leq \sqrt{\frac{Cj^2\cdot 2^{2j}}{2^{2j+j/2}\sqrt{j}}} \to 0\,;$$

$$\mathsf{E}\hat{T}^2_{\lambda;j}\mathbb{1}_{\hat{T}_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leq T_{\lambda;j}} \leq \\ \leq T^2_{\lambda;j}\mathsf{E}\mathbb{1}_{\hat{T}_{\lambda;j}<|Y_{\lambda;j,\mathbf{k}}|\leq T_{\lambda;j}} \to 0\,.$$

Для $\hat{\sigma}^2_{\lambda;j}$ заметим, что $\hat{\sigma}^2_{\lambda;j}\leqslant\hat{T}^2_{\lambda;j}$. После применения неравенства Чебышёва получим, что S_2/L сходится к нулю по вероятности.

В S_1 оба слагаемых сходятся по распределению к нормальному закону и при этом независимы. Поэтому их сумма тоже сходится по распределению к нормальному закону. Осталось убедиться в правильности параметров. Имеем

$$\sum_{\lambda,j,\mathbf{k}}\left(\hat{\sigma}^2_{\lambda;j}-\sigma^2_{\lambda;j}\right) = \sum_{\lambda}\sum_{j}2^{2j}\cdot 2^j\left(\hat{\sigma}^2_{\lambda;0}-\sigma^2_{\lambda;0}\right) = \\ = \left(\left\|\xi^{[1]}_{0,0,0}\right\|^2_2 + \left\|\xi^{[2]}_{0,0,0}\right\|^2_2 + \left\|\xi^{[3]}_{0,0,0}\right\|^2_2\right)\times \\ \times\frac{2^{3J}-2^{3j_M}}{2^3-1}\left(\hat{\sigma}^2-\sigma^2\right) = \\ = \frac{\sigma^2_{1;0}+\sigma^2_{2;0}+\sigma^2_{3;0}}{\sigma^2}\,\frac{2^{3J}-2^{3j_M}}{7}\left(\hat{\sigma}^2-\sigma^2\right)\,.\square$$

Замечание. Если функция f регулярная с параметром $\alpha\geq 1/4$, а $j_M\geq 4J/5$, то можно ослабить требования на $\hat{\sigma}^2$. Достаточно потребовать только состоятельность, асимптотическую нормальность и независимость от $Y_{\lambda;j,\mathbf{k}}$.

В теоремах 2 и 3 при оценке $\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>\hat{T}_{\lambda;j}\right)$ использовалось число $0<\gamma<1$. Можно заменить γ бесконечно малой последовательностью γ_L, которая не испортит порядок знаменателя в (17).

По формуле полной вероятности для любого $\delta>0$

$$\mathsf{P}\left(\sum_{\lambda,j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat{T}_{\lambda;j}}>\delta\right) = \\ = \mathsf{P}\left(\hat{T}_{\lambda;j}\leqslant(1-\gamma_L)\,\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right)\times \\ \times\mathsf{P}\left(\sum_{\lambda,j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat{T}_{\lambda;j}}>\delta\,|\,\hat{T}_{\lambda;j}\leqslant \right. \\ \left.\leq(1-\gamma_L)\,\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right) + \\ + \mathsf{P}\left(\sum_{\lambda,j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat{T}_{\lambda;j}}>\delta\,,\right. \\ \left.\hat{T}_{\lambda;j}>(1-\gamma_L)\,\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right)\,, \quad (22)$$

где $\gamma_L=1/J$. При таком γ_L получаем

$$\mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}|>(1-\gamma_L)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right) = \\ = \frac{C}{2^{2j(1-\gamma_L)^2}\cdot 2^{j/2}\sqrt{j}} \leq \frac{C_1}{2^{2J}\sqrt{j}}$$

в силу выбора j_M и того, что

$$2^{2j(1-\gamma_L)^2} = 2^{2j\left(1-2/J+1/J^2\right)} > 2^{2j-4}\,.$$

По неравенству Чебышёва

$$\mathsf{P}\left(\sum_{\lambda,j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>\hat{T}_{\lambda;j}}>\delta\,,\right. \\ \left.\hat{T}_{\lambda;j}>(1-\gamma_L)\,\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right) \leq \\ \leq\mathsf{P}\left(\sum_{\lambda,j,\mathbf{k}}\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}|>(1-\gamma_L)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}}>\delta\right)\leq$$

$$\leq \frac{\sum\limits_{\lambda,j,\mathbf{k}} \mathsf{P}\left(|Y_{\lambda;j,\mathbf{k}}| > (1-\gamma_L)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right)}{\delta} = \\ = \mathsf{O}\left(\frac{1}{\sqrt{j}}\right).$$

Используя свойство асимптотической нормальности $\hat{\sigma}^2$, можно для любого $\delta' > 0$ оценить

$$\mathsf{P}\left(\hat{T}_{\lambda;j} \leq (1-\gamma_L)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right) < \delta', \qquad (23)$$

причем отметим, что δ здесь фиксировано, а δ' можно делать произвольно малым. Имеем

$$\mathsf{P}\left(\hat{T}_{\lambda;j} \leqslant (1-\gamma_L)\sigma_{\lambda;j}\sqrt{2\ln 2^{2j}}\right) = \\ = \mathsf{P}\left(\hat{\sigma}^2 \leqslant (1-\gamma_L)^2\sigma^2\right) = \\ = \mathsf{P}\left(\left(\hat{\sigma}^2 - \sigma^2\right) \leqslant \sigma^2\left(-2\gamma_L + \gamma_L^2\right)\right) = \\ = \mathsf{P}\left(\sqrt{L}\left(\hat{\sigma}^2 - \sigma^2\right) \leqslant -\frac{\sqrt{L}\sigma^2(2J-1)}{J^2}\right).$$

Для произвольного $\delta' > 0$ найдется J_0 ($L_0 = 2^{2J_0}$) такое, что

$$F_\Sigma\left(-\frac{\sqrt{L_0}\sigma^2(2J_0-1)}{J_0^2}\right) < \frac{\delta'}{2},$$

где F_Σ — функция распределения нормального закона с нулевым средним и дисперсией Σ^2. При этом для любого $J \geq J_0$

$$\mathsf{P}\left(\sqrt{L}\left(\hat{\sigma}^2 - \sigma^2\right) \leq -\frac{\sqrt{L}\sigma^2(J-1)}{J^2}\right) \leq \\ \leq \mathsf{P}\left(\sqrt{L}\left(\hat{\sigma}^2 - \sigma^2\right) \leq -\frac{\sqrt{L_0}\sigma^2(2J_0-1)}{J_0^2}\right).$$

В силу асимптотической нормальности $\hat{\sigma}^2$ и непрерывности F_Σ для этого же δ' найдется J_1 $\left(L_1 = 2^{2J_1}\right)$ такое, что для любого $J \geq J_1$

$$\left|\mathsf{P}\left(\sqrt{L_1}\left(\hat{\sigma}^2 - \sigma^2\right) \leq x\right) - F_\Sigma(x)\right| < \frac{\delta'}{2},$$

причем J_1 не зависит от x. Возьмем $x_0 = -\sqrt{L_0}\sigma^2(2J_0-1)/J_0^2$ и $J_2 = \max\{J_0, J_1\}$. Для любого $J \geq J_2$ имеем

$$\mathsf{P}\left(\sqrt{L}\left(\hat{\sigma}^2 - \sigma^2\right) \leq x_0\right) < \delta',$$

а значит, справедливо (23).

Получаем, что сумма индикаторов в (22) сходится к нулю по вероятности:

$$\mathsf{P}\left(\sum_{\lambda,j,\mathbf{k}} \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}} > \delta\right) \to 0 \text{ при } J \to \infty.$$

Для суммы индикаторов с неслучайным порогом аналогично получаем

$$\mathsf{P}\left(\sum_{\lambda,j,\mathbf{k}} \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > T_{\lambda;j}} > \delta\right) \to 0.$$

Далее воспользуемся дискретной версией неравенства Коши–Буняковского:

$$\frac{\sum\limits_{\lambda,j,\mathbf{k}} Y_{\lambda;j,\mathbf{k}}^2 \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}}}{L} \leq \\ \leq \sqrt{\frac{\sum\limits_{\lambda,j,\mathbf{k}} Y_{\lambda;j,\mathbf{k}}^4/L}{L} \sum_{\lambda,j,\mathbf{k}} \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}}} \xrightarrow{\mathsf{P}} 0,$$

так как $\mathsf{E}\left[Y_{\lambda;j,\mathbf{k}}^4/L\right]$ ограничено,

$$\frac{\sum\limits_{\lambda,j,\mathbf{k}} \hat{T}_{\lambda;j}^2 \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}}}{L} \leq \frac{\sum\limits_{\lambda,j,\mathbf{k}} Y_{\lambda;j,\mathbf{k}}^2 \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}}}{L} \xrightarrow{\mathsf{P}} 0$$

и

$$\hat{\sigma}_{\lambda;j}^2 \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}} \leqslant \hat{T}_{\lambda;j}^2 \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}}.$$

Оценки для слагаемых с $\mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > T_{\lambda;j}}$ получены в теореме 1.

Замечание. Всюду выше в этом разделе предполагалось, что пороговая обработка и суммирование в выражении для риска (7) ведутся с уровня j_M, причем $j_M \to \infty$ при $J \to \infty$. Однако если ввести дополнительные ограничения на регулярность f, то можно вести пороговую обработку и суммирование с уровня $j_0 \nrightarrow \infty$. Если $j_M = J/(\alpha+1)$, то для коэффициентов, соответствующих $j < j_M$, неравенство (8), вообще говоря, не выполнено. Оценим вклад больших коэффициентов в оценку риска:

$$L^{-1} \sum_{j=j_0}^{j_M-1} \sum_{\lambda,\mathbf{k}} \left\{\left|Y_{\lambda;j,\mathbf{k}}^2 - \hat{\sigma}_{\lambda;j}^2\right| \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| \leqslant \hat{T}_{\lambda;j}} + \right. \\ \left. + \left(\hat{\sigma}_{\lambda;j}^2 + \hat{T}_{\lambda;j}^2\right) \mathbb{1}_{|Y_{\lambda;j,\mathbf{k}}| > \hat{T}_{\lambda;j}}\right\} \leq \\ \leq L^{-1} \sum_{j=j_0}^{j_M-1} \sum_{\lambda,\mathbf{k}} \left\{\left(\hat{\sigma}_{\lambda;j}^2 + \hat{T}_{\lambda;j}^2\right) + \left(\hat{\sigma}_{\lambda;j}^2 + \hat{T}_{\lambda;j}^2\right)\right\} \xrightarrow{\mathsf{P}} \\ \xrightarrow{\mathsf{P}} 0$$

в силу состоятельности $\hat{\sigma}^2$ и того, что

$$L^{-1}\left\{\sum_{j=j_0}^{j_M-1} j2^j\cdot 2^{2j}\right\} \le 2^{-2J}\left\{j_M\sum_{j=j_0}^{j_M-1} 2^{3j}\right\} \simeq$$
$$\simeq 2^{2J}\cdot j_M\cdot 2^{3j_M}\to 0$$

при $J\to\infty$, если $3j_M < 2J$, т. е. $\alpha > 1/2$. Слагаемые риска оцениваются аналогично. Итак, при $\alpha > 1/2$ суммирование в (7) можно начинать с произвольного j_0.

Литература

1. *Наттерер Ф.* Математические аспекты компьютерной томографии. — М.: Мир, 1990.
2. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1979.
3. *Хермен Г.* Восстановление изображений по проекциям: основы реконструктивной томографии. — М.: Наука, 1983.
4. *Добеши И.* Десять лекций по вейвлетам. — Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001.
5. *Donoho D. L.* Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition // Appl. Comput. Harmonic Anal., 1995. Vol. 2. P. 101–126.
6. *Kolaczyk E. D.* A wavelet shrinkage approach to tomographic image reconstruction // J. Amer. Statistical Association, 1996. Vol. 91. No. 435. P. 1079–1090.
7. *Kolaczyk E. D.* Wavelet methods for the inversion of certain homogeneous linear operators in the presence of noisy data. Ph.D. Thesis, 1994.
8. *Donoho D. L., Johnstone I. M.* Ideal spatial adaptation via wavelet shrinkage // Biometrika, 1994. Vol. 81. No. 3. P. 425–455.
9. *Donoho D. L., Johnstone I. M.* Adapting to unknown smoothness via wavelet shrinkage // J. Amer. Statistical Association, 1995. Vol. 90. P. 1200–1224.
10. *Mallat S.* A wavelet tour of signal processing. — Academic Press, 1999.
11. *Маркин А. В.* Предельное распределение оценки риска при пороговой обработке вейвлет-коэффициентов // Информатика и её применения, 2009. Т. 3. Вып. 4. С. 57–63.
12. *Маркин А. В., Шестаков О. В.* О состоятельности оценки риска при пороговой обработке вейвлет-коэффициентов // Вестник Московского университета. Сер. 15. Вычислительная математика и кибернетика, 2010. № 1. С. 26–33.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 46–52

АНАЛИЗ СЕТЕВОГО ПРОТОКОЛА С ОБЩЕЙ ФУНКЦИЕЙ РАСШИРЕНИЯ ОКНА ПЕРЕДАЧИ СООБЩЕНИЯ ПРИ КОНФЛИКТАХ*

А. Лукьяненко[1], Е. Морозов[2], А. Гуртов[3]

Аннотация: Исследован класс сетевых протоколов контроля несущей среды, где окно передачи сообщения является призвольной возрастающей функцией числа конфликтов сообщения, посланного с данной станции. В общепринятых предположениях, накладываемых на свойства сети, исследована функция протокола, определяемая правилом расширения окна в зависимости от числа конфликтов. Найдено выражение для функции протокола, обеспечивающей минимальное среднее время передачи сообщения. Проведен асимптотический анализ протокола при неограниченно растущем числе станций. Рассмотрены протоколы как с неограниченным, так и с ограниченным числом попыток передачи сообщения. Предложена модель распределения доступа к каналу в непрерывном времени, допускающая слоты различной длины.

Ключевые слова: передача данных; оценка производительности; моделирование протокола; доступ к каналу

1 Введение

В данной работе рассмотрен сетевой протокол, обеспечивающий отсрочку передачи данных, вызванную конфликтом в сети с коллективным доступом с контролем несущей и обнаружением/устранением конфликтов [1]. Далее для этого протокола будет использовано обозначение BP (backoff protocol). Протокол BP служит механизмом успешной передачи информации и приводит к построению легко развертываемых и недорогих локальных сетей (ЛС). Построение ЛС, использующее прямые связи всех станций друг с другом, весьма дорого, и добавление каждой новой станции ведет к стремительному увеличению затрат и сложности ЛС. Альтернативное решение, состоящее в том, что сообщение передается не напрямую, требует уверенности в том, что промежуточная станция, используемая для передачи, не уйдет из сети. В ЛС центральным элементом является *передающая среда*, которую назовем системой передачи или просто *системой*. Когда система развернута, новые узлы (рабочие станции, терминалы, принтеры, серверы) просто подключаются к ней и могут мгновенно начинать функционировать. Однако такая простота и быстрота развертывания и организации ЛС создает технические трудности. Когда некоторая станция начинает передавать сигнал, возможна ситуация, при которой другая станция, увидев систему пустой, также начинает передачу (поскольку она еще не обнаружила ранее посланный в систему сигнал). Тогда данные обеих станций перекрываются и, как следствие, разрушаются, т. е. происходит *конфликт*. Существуют методы избавления от подобных конфликтов. Например, при использовании радиосигнала или оптоволоконной среды сигналы разных станций можно передавать на различающихся частотах, однако это приведет к удорожанию технологии. К тому же число различных доступных частот обычно ограничено.

После того как станция узнает о том, что ее данные разрушены, она (как правило) инициирует повторную попытку передачи. Но если такие попытки будут предприниматься через детерминированные промежутки времени, то те же сообщения столкнутся вновь. Именно для решения этой проблемы и предназначен BP, в котором окно передачи растет вместе с числом неудачных попыток послать сообщение. Так называемый *константный* BP был впервые использован как часть протокола Aloha [2]. Позднее его модификацию (обрезанный бинарный экспоненциальный BP) успешно применили в сети Ethernet [1, 3]. Несмотря на то, что сейчас Ethernet ушел от использования BP, алгоритм расширения окна все еще активно используется в различных сетевых протоколах, в частности в беспроводных

* Работа поддерживается грантом РФФИ, 10-07-00017.

[1] Helsinki Institute for Information Technology HIIT, Aalto, Finland, firstname.secondname@hiit.fi
[2] Институт прикладных математических исследований КарНЦ РАН, emorozov@krc.karelia.ru
[3] Helsinki Institute for Information Technology HIIT, Aalto, Finland, gurtov@hiit.fi

сетях (например, IEEE 802.11 [4]) и транспортных протоколах таких, как SCTP и TFRC.

Несмотря на свою относительно долгую историю, важность и простоту, BP долгое время не поддавался удовлетворительному теоретическому анализу. В этой связи укажем на работу [5], содержащую подробную предысторию вопроса, анализ устойчивости, а также детальное исследование важных частных случаев BP. Существует множество работ, связанных с анализом BP, однако, как правило, используемые в них предпосылки являются слишком ограничительными (например, рассматривается бесконечное число станций $N = \infty$ или функция расширения окна f имеет специальную форму) [6–8]. В данной работе исследование проводится при конечном числе станций N, а затем рассматривается асимптотика при $N \to \infty$. В действительности, рассматривается целый класс протоколов описанного выше типа, где каждый протокол специфицируется выбором конкретной (возрастающей) функции f.

2 Описание протокола

Рассмотрим N станций, подключенных к передающей среде, причем предполагается, что в каждой станции постоянно имеется очередь сообщений, готовых к отправке. Считается, что все станции идентичны (в статистическом смысле) и работают независимо друг от друга. Рассмотрим работу одной такой (произвольной) станции более подробно. По алгоритму BP выбирается первое сообщение из очереди и в специальном счетчике, который называется BC, устанавливается начальное значение $i = 0$. В любой момент времени значение BC равно числу последовательных безуспешных попыток отправки, накопленных станцией к данному моменту времени. Если в некоторый момент значение BC равно i, то для осуществления следующей передачи BP строит *окно отсрочки передачи сообщения* $W(i) = [1, f(i)]$, где f — некоторая заданная монотонно возрастающая целочисленная функция, $i \geq 0$, причем $f(0) \geq 1$. (Случай произвольной функции f, требующий незначительного изменения модели, представлен в [9].) Таким образом, окно расширяется с увеличением значения BC.

Естественной единицей времени при сетевом анализе является *слот*, в течение которого может произойти одно *элементарное событие*, скажем столкновение сообщений, передача пакета и т. д. Вообще говоря, величина реального (физического) времени, соответствующего различным слотам, не является постоянной. Основные результаты данной работы получены в терминах слотов, а в разд. 5 показано, как переформулировать модель (и полученные результаты) в терминах реального времени. Таким образом, величина $f(i)$ равна числу слотов, в течение одного из которых произойдет следующая попытка отправки сообщения (при условии, что уже произошло ровно i конфликтов). Внутри окна $W(i)$ слот отправки сообщения D_i выбирается *равномерно*. Таким образом, D_i является временем задержки (в слотах) до осуществления следующей попытки отправки сообщения, если это сообщение уже имело $i \geq 0$ неудачных попыток. Если сообщение отправлено успешно, то значение BC полагается равным нулю. Если же (при значении BC равном i) происходит конфликт, то значение BC увеличивается до $i + 1$. Для следующей отправки используется окно $W(i + 1) = [1, f(i + 1)]$ и т. д. В дальнейшем рассматриваются *ограниченный* и *неограниченный* протоколы. В ограниченном протоколе задана верхняя граница M значений BC: если сообщение не удалось отправить в течение M попыток, то оно выбрасывается из очереди. В неограниченном протоколе $M = \infty$. Далее подробно исследуется неограниченный протокол. Для ограниченного протокола достаточно использовать очевидную модификацию анализа, применяемого для неограниченного протокола, и поэтому в данном случае приведены лишь окончательные результаты.

Размер любого слота ограничен снизу величиной RTT (round trip time). Это необходимо, чтобы станции получали информацию об отсутствии столкновений в течение одного слота, поскольку RTT — это величина времени, за которое сигнал об отправке сообщения оповестит сеть и вернется на станцию отправки. Поэтому исключена ситуация, когда одна станция ведет отправку сообщения в течение нескольких слотов, в то время как другая станция уменьшила число *пустых слотов* до очередной попытки отправки сообщения. (Пустым слотом для данной станции является любой слот внутри окна передачи, когда станция не пытается передавать сообщение.) Кроме того, считается, что время передачи сообщения занимает один слот. Это ограничение основано в первую очередь на поведении сети Wi-Fi (протокол IEEE802.11). Например, для протокола IEEE802.11 время распространения сигнала в сети равно 1 мкс (т. е. $\mathrm{RTT} = 2$ мкс), а пустой слот равен 50 мкс [4, 10].

Для сети, где станции продолжают отсчитывать время до отправки, даже видя сеть занятой, рассматриваемая в данной статье модель верна, только если одно сообщение передается в течении одного слота, т. е. лишь при малых размерах отправляемых пакетов. Иначе после успешной передачи резко возрастает вероятность столкновения. Действительно, если передача сообщения требует *намного*

больше одного слота, то многие другие станции, исчерпав время до отправки, попытаются передать свои сообщения одновременно, сразу же после завершения данной передачи, что приведет к конфликтам. Заметим, что в сети Ethernet станции продолжают отсчитывать время (пустые слоты) до отправки, даже если видят сеть занятой. Поэтому, как было отмечено, предлагаемая модель применима в данном случае лишь при малых размерах отправляемых пакетов.

3 Математическая модель в дискретном времени и ее анализ

Суммируем предположения, положенные в основу модели. Они приняты в литературе, посвященной данному вопросу, и являются вполне естественными и согласованными с имеющимися данными о работе ВР (см., например, [3, 11, 12]).

Предположение 1. *Станции идентичны*.

Предположение 2. *Стационарность*: вероятность конфликта $p_c \in (0,1)$ постоянна и является одной и той же для каждой станции.

Предположение 3. *Условие насыщения*: каждая станция всегда имеет непустую очередь.

Предположение 4. Если конфликт не происходит в первый слот передачи сообщения, то данное сообщение передается успешно.

Предположение 5. Когда какая-либо станция начинает передавать сообщение, то либо 1) время до отправки на других станциях не уменьшается, либо 2) время отправки одного сообщения занимает один слот. При анализе протокола в терминах слотов оба эти предположения эквивалентны.

Далее, состоянием данной станции считается значение ее ВС. Если $\mathrm{BC} = i$, то

$$\mathsf{P}(D_i = k) = \frac{1}{f(i)}\,, \quad k = 1, \ldots, f(i)\,.$$

Поэтому среднее время пребывания станции в данном состоянии определяется по формуле $\mathsf{E}D_i = (f(i)+1)/2$. Далее, вероятность того, что значение $\mathrm{BC} = i$ и в этом состоянии произойдет успешная отправка, есть $P_i = (1-p_c)p_c^i$, $i \geq 0$. Назовем *циклом передачи сообщения* время с момента первой попытки отправки до успешной его отправки и обозначим его длительность через S.

Заметим, что стационарная вероятность γ_i того, что станция находится в состоянии i (в произвольный момент времени) равна доле времени, проводимом станцией в этом состоянии на одном цикле передачи. Это приводит к следующему соотношению:

$$\gamma_i = \frac{P_i \mathsf{E}D_i}{\sum\limits_{i=0}^{\infty} P_i \mathsf{E}D_i} = \frac{(f(i)+1)(1-p_c)p_c^i}{(1-p_c)F(p_c)+1}\,, \quad i \geq 0\,, \quad (1)$$

где использовано обозначение

$$F(p_c) = \sum_{i=0}^{\infty} f(i) p_c^i\,.$$

Функцию F, которая играет в дальнейшем анализе ключевую роль, назовем *функцией протокола*. Заметим, что отправка сообщения происходит (внутри окна передачи) в слоте с номером N_R, распределение которого имеет вид $\mathsf{P}(N_R = i) = P_i$, $i \geq 0$. Поэтому величина

$$\sum_{i=0}^{\infty} \mathsf{E}D_i P_i = \mathsf{E}D_{N_R}\,,$$

стоящая в знаменателе (1), является средней величиной окна, в котором происходит успешная отправка сообщения. В [12] доказано следующее равенство:

$$\frac{\gamma_i}{\mathsf{E}D_i} = \frac{P_i}{\mathsf{E}D_{N_R}}\,, \quad i \geq 0\,.$$

Из формулы полной вероятности следует такое выражение для стационарной вероятности успешной отправки сообщения (в произвольном слоте):

$$p_{\mathrm{st}} = \sum_{i=0}^{\infty} \frac{P_i}{\mathsf{E}D_{N_R}} = \frac{1}{\mathsf{E}D_{N_R}} = \frac{2}{(1-p_c)F(p_c)+1}\,. \quad (2)$$

Вероятность p_{st} можно получить иначе. Заметим, что средняя длина цикла передачи может быть записана таким образом:

$$\mathsf{E}\left(\sum_{i=0}^{N_R} D_i\right) = \sum_{i=0}^{\infty} \mathsf{E}D_i \mathsf{P}\left(N_R \geq i\right) =$$
$$= \frac{1}{2}\sum_{i\geq 0} \left(f(i)+1\right) p_c^i = \frac{1}{2}\left(F(p_c) + \frac{1}{(1-p_c)}\right) =$$
$$= \frac{(1-p_c)F(p_c)+1}{2(1-p_c)}\,. \quad (3)$$

Поскольку $P(N_R \geq i) = p_c^i$, то среднее число попыток до успешной отправки

$$\mathsf{E}N_R = \frac{p_c}{(1-p_c)}\,. \quad (4)$$

Понятно, что в стационарном режиме вероятность отправки (в произвольный момент времени) равна отношению среднего числа попыток к средней длине цикла передачи. Поэтому (3) и (4) дают выражение

$$p_{\mathrm{st}} = \frac{\mathsf{E}N_R}{\mathsf{E}\Bigl(\sum\limits_{i=0}^{N_R} D_i\Bigr)},$$

которое совпадает с (2). Учтем, что данная станция (как и остальные $N-1$ станций) находится в стационарном режиме. Поскольку конфликт возникает, если не менее двух станций пытаются отправить сообщения одновременно, то нетрудно получить следующее соотношение (см. также [9, 12]):

$$p_c = 1 - (1 - p_{\mathrm{st}})^{N-1}. \qquad (5)$$

Объединяя выражения (2) и (5), приходим к следующему выражению для функции протокола:

$$F(p_c) = \frac{1 + (1-p_c)^{1/(N-1)}}{(1-p_c)\left(1 - (1-p_c)^{1/(N-1)}\right)}. \qquad (6)$$

В [9] показано, что для существования единственного решения $p_c \in (0,1)$ уравнения (6) достаточно монотонного возрастания функции f. Заметим, что различные протоколы, определяемые различными функциями F, вообще говоря, приводят к различным решениям p_c.

Теперь найдем при каком значении p_c станция будет работать оптимально, т. е. обеспечит минимальную среднюю длину цикла передачи $\mathsf{E}S$. Используя независимость случайных величин N_R и D_i, можно записать

$$\begin{aligned}\mathsf{E}S = \mathsf{E}\left(\sum_{i=0}^{N_R} D_i\right) &= \\ = \frac{1}{2}\,\mathsf{E}\left(\sum_{i=0}^{\infty}(f(i)+1)\mathbb{1}(N_R \ge i)\right) &= \\ = \frac{1}{2}\left(F(p_c) + \frac{1}{1-p_c}\right), \qquad (7)\end{aligned}$$

где $\mathbb{1}$ обозначает индикатор.

Объединяя (6) и (7), получаем

$$\mathsf{E}S = \frac{1}{(1-p_c)\left(1 - (1-p_c)^{1/(N-1)}\right)}. \qquad (8)$$

Нетрудно проверить стандартным способом, что минимальное значение среднего цикла передачи $\mathsf{E}S$, являющегося функцией вероятности конфликта p_c, достигается в точке $p_c = p_c^*$, где

$$p_c^* = 1 - \left(1 - \frac{1}{N}\right)^{N-1}. \qquad (9)$$

Подстановка (9) в (8) дает

$$\mathsf{E}S = \frac{N}{(1-1/N)^{N-1}}. \qquad (10)$$

Если рассматривать вероятность конфликта как функцию числа станций N в сети, т. е. $p_c^* = p_c^*(N)$, то из (9) следует такой асимптотический результат:

$$p_c^* \to 1 - e^{-1}, \quad N \to \infty.$$

Подстановка (9) в (6) показывает, что протокол является оптимальным, если его функция F удовлетворяет условию

$$F(p_c^*) = \frac{2N-1}{(1-1/N)^{N-1}} := A(N). \qquad (11)$$

Отметим, что этому соотношению может удовлетворять, вообще говоря, не единственная функция расширения окна f (т. е. не единственный протокол), а целый класс функций f.

Рассмотрим класс наиболее важных с точки зрения практики *экспоненциальных протоколов*, где $f(i) = a^i$ для некоторого числа $a > 0$. В этом случае в предположении $ap_c^* < 1$ соотношение (11) принимает вид

$$\frac{1}{1 - ap_c^*} = A(N).$$

Таким образом,

$$a = \frac{A(N)-1}{A(N)p_c^*} = \frac{2N - 1 - (1-1/N)^{N-1}}{(2N-1)\left(1 - (1-1/N)^{N-1}\right)}.$$

Нетрудно показать, что при $N \to \infty$

$$a := a(N) \to \frac{1}{1-e^{-1}} \approx 1{,}58$$

и что $a(N)$ монотонно убывает с ростом числа станций N. Кроме того, $a(2) = 5/3 \approx 1{,}66$. Таким образом, *оптимальное значение параметра a в формуле* (11) *определяется однозначно для каждого N, находится в диапазоне* $[1{,}58,\ 1{,}66]$ *и с ростом числа станций приближается к 1,58 сверху*.

Коснемся условия стационарности системы, считая упрощенно, что суммарный входной поток в систему является процессом восстановления с интенсивностью $\lambda \in (0, \infty)$. Естественно считать, что тогда каждая станция получает входной поток интенсивности λ/N. Такое предположение вполне оправдано при большом числе идентичных

станций. С другой стороны, нетрудно показать (используя, скажем, аргументы из теории регенерирующих процессов), что условие стационарности (для каждой станции) имеет хорошо известный вид:

$$\frac{\lambda}{N}\mathsf{E}S < 1\,. \tag{12}$$

Такого рода условия *отрицательного сноса* хорошо известны в анализе стационарности марковских цепей [13]. Как правило, они исключают *уход процесса в бесконечность* и обеспечивают существование стационарного режима для широкого класса процессов, описывающих реальные коммуникационные системы. Отметим, что хотя $\mathsf{E}S \to \infty$ при $N \to \infty$ (см. (10)), однако интенсивность входного потока в отдельный узел пропорционально убывает, сохраняя неравенство (12). При оптимальном значении $p_c = p_c^*$ неравенство (12) принимает форму (см. (8), (9)):

$$\lambda < \left(1 - \frac{1}{N}\right)^{N-1}\,.$$

Это неравенство в пределе при $N \to \infty$ переходит в неравенство $\lambda < e^{-1}$, возникающее, например, при анализе стационарности синхронной системы ALOHA [14, 15]. Таким образом, при оптимальном выборе протокола и большом числе станций интенсивность входного потока на каждую станцию, обеспечивающая устойчивость, должна быть меньше e^{-1}. Следует ожидать, что для произвольного протокола эта область устойчивости еще меньше. Отметим, что проблема неустойчивости BP неоднократно отмечалась ранее (см. [6–8]).

4 Ограниченный протокол отсрочки

Проведенный выше анализ можно перенести на случай, когда число конфликтов ограничено величиной $M < \infty$. В этом случае формула (6), определяющая функцию протокола, преобразуется к виду:

$$F(p_c) := F_M(p_c) = \\ = \frac{\left(1 - p_c^{M+1}\right)\left(1 + (1-p_c)^{1/(N-1)}\right)}{(1-p_c)\left(1 - (1-p_c)^{1/(N-1)}\right)}\,, \tag{13}$$

а средняя длина цикла передачи принимает вид:

$$\mathsf{E}S = \frac{\left(1 - p_c^{M+1}\right)}{(1-p_c)\left(1 - (1-p_c)^{1/(N-1)}\right)}\,.$$

Рассматриваемое ограничение применяется, например, в алгоритме протокола Ethernet, где используется ограниченный бинарный экспоненциальный протокол. Более точно, $f(i) = 2^i$, $i = 1, \ldots, 10$ и $f(i) = 1024$, $i = 11, \ldots, 16$. После 16 конфликтов сообщение выбрасывается из очереди. Это правило дает следующую функцию протокола:

$$F_{16}(p_c) = \sum_{i=0}^{10} 2^i p_c^i + \sum_{i=11}^{16} 2^{10} p_c^i = \\ = \frac{1 - (2p_c)^{11}}{1 - 2p_c} + 2^{10} p_c^{11} \frac{1 - p_c^6}{1 - p_c}\,.$$

Поэтому уравнение (13) (относительно p_c) принимает вид

$$\frac{1 - (2p_c)^{11}}{1 - 2p_c} + 2^{10} p_c^{11} \frac{1 - p_c^6}{1 - p_c} = \\ = \frac{\left(1 - p_c^{17}\right)\left(1 + (1-p_c)^{1/(N-1)}\right)}{(1-p_c)\left(1 - (1-p_c)^{1/(N-1)}\right)}\,. \tag{14}$$

Решив это уравнение, можно подсчитать среднюю длину цикла передачи сообщения $\mathsf{E}S$ и вероятность потери сообщения p_c^{17}. Явное решение уравнения (14) в общем случае найти трудно, однако его можно исследовать численно. Некоторые численные результаты представлены в табл. 1. Заметим, что вероятность столкновения и вероятность потери возрастают с ростом числа станций. Так, при $N = 11$ вероятность столкновения $p_c = 0{,}621$ (и достаточно близка к оптимальной вероятности столкновения $p_c = p_c^* \approx 1 - e^{-1} = 0{,}6817$, полученной для неограниченного протокола). Вероятность потери в этом случае незначительна. При $N = 501$ и $N = 1001$ вероятность столкновения больше 0,9, а вероятность потери равна соответственно 0,35 и 0,81. При этом среднее время передачи сообщения в сети $\mathsf{E}S/N$ (в отличие от среднего цикла передачи $\mathsf{E}S$ для отдельной станции) практически не изменяется с ростом числа станций N. Причина такой устойчивости, по-видимому, состоит в том, что увеличение вероятности потери, вызванное ростом числа станций, сдерживает рост среднего времени передачи.

5 Модель протокола в непрерывном времени

Выше рассмотрена дискретная модель в терминах слотов, которые, вообще говоря, не равны по величине. Полезно рассмотреть модель в

Таблица 1 Численный анализ BEB

Число станций	Вероятность столкновения p_c	Среднее время передачи сообщения в сети $\mathsf{E}S/N$	Вероятность потери p_c^{17}
11	0,621	2,593	0,0003
51	0,743	2,828	0,0064
101	0,799	3,026	0,022
501	0,94	3,858	0,349
1001	0,988	3,515	0,809

непрерывном времени, что может привести к существенному изменению вида оптимизационной задачи. Для этого введем величину пустого слота T_i, величину слота T_c, где произошло столкновение, и величину T_s слота, где произошла успешная отправка. Если размеры слотов одинаковы (т. е. $T_c = T_s = T_i$), то среднее (физическое) время цикла передачи $\mathsf{E}_T S = \mathsf{E}S \cdot T_i$ и модель сводится к модели, рассмотренной выше.

В случае неограниченного протокола среднее число слотов в цикле передачи, равное $\mathsf{E}\left(\sum_{i=0}^{N_R} D_i\right)$, содержит в среднем $\mathsf{E}N_R + 1$ *непустых слотов*, причем в течение $\mathsf{E}N_R$ слотов происходят столкновения и один слот содержит успешную отправку. Пусть T_i^* — размер слота, в течение которого *данная произвольная станция* не отправляет сообщения (в то время как другие станции могут отправлять или не отправлять). Таким образом, средний цикл передачи

$$\mathsf{E}_T S = \mathsf{E}N_R T_c + \left(\mathsf{E}\left[\sum_{i=0}^{N_R} D_i\right] - \mathsf{E}N_R - 1\right) T_i^* + T_s\,. \quad (15)$$

С учетом (3) и (4) соотношение (15) принимает вид:

$$\mathsf{E}_T S = \frac{p_c}{1-p_c} T_c + \frac{1}{2(1-p_c)} \left((1-p_c)F(p_c) - 1\right) T_i^* + T_s\,. \quad (16)$$

Вообще говоря, $T_i^* \neq T_i$, поскольку величина пустого слота для данной станции *растягивается* на время возможных передач другими станциями. Для нахождения T_i^* используем формулу полной вероятности. Во-первых, $T_i^* = T_i$ с вероятностью $1 - p_c$, с которой остальные $N - 1$ станций не пытаются передавать (в произвольном слоте). Далее (предполагаем, что $N > 2$), лишь одна из $N - 1$ оставшихся станций передает сообщение (а остальные $N - 2$ станций молчат) в течение данного слота с вероятностью $(N-1)p_{\mathrm{st}}(1-p_{\mathrm{st}})^{N-2}$, и с этой вероятностью $T_i^* = T_s$. Оставшаяся вероятность соответствует событию, при котором, по крайней мере, две станции пытаются передавать сообщение в данном слоте, и тогда $T_i^* = T_c$. Таким образом, используя соотношения (2) и (5), имеем

$$T_i^* = (1-p_c)T_i + (1-p_c)\frac{2(N-1)}{(1-p_c)F(p_c)-1} T_s + \left(1 - (1-p_c) - (1-p_c)\frac{2(N-1)}{(1-p_c)F(p_c)-1}\right) T_c = (1-p_c)T_i + p_c T_c + \frac{2(1-p_c)}{(1-p_c)F(p_c)-1}(N-1)(T_s - T_c)\,.$$

Подставляя это выражение в (16), получаем

$$\mathsf{E}_T S = N(T_s - T_c) + \frac{1}{1-p_c} T_c + \frac{2(N-1)}{(1-p_c)F(p_c)-1} \left((1-p_c)\,T_i + p_c T_c\right)\,.$$

Используя явный вид (6) функции $F(p_c)$, окончательно получаем величину среднего цикла передачи в виде

$$\mathsf{E}_T S = N(T_s - T_c) + \frac{1}{1-p_c} T_c + (1-p_c)(N-1) \times \left((1-p_c)^{1-1/(N-1)} - 1\right) \left((1-p_c)T_i + p_c T_c\right)\,.$$

Таким образом, задача оптимизации сводится к отысканию минимума $\mathsf{E}_T S$ как функции p_c для заданных величин T_s, T_c и T_i. Заметим, что T_s не влияет на задачу оптимизации.

В работе [11] исследован частный случай данной функции (для бинарного экспоненциального протокола) при фиксированных величинах T_s, T_c и T_i, описанных в спецификации [4]. В частности, $T_i = 50$ мкс, средняя длина фрейма (пакета MAC-уровня) T_s определяется настройщиком сети, величина T_c определяется видом конкретно используемого протокола. Из работы [11] следует, что предлагаемая в данной статье модель адекватно описывает работу реальных сетевых протоколов.

6 Заключение

В статье предложен метод исследования сетевого протокола контроля несущей, в котором окно отсрочки передачи сообщения при конфликте сообщений, посланных разными станциями, меняется в соответствии с произвольной (не обязательно экспоненциальной) функцией f, монотонно возрастающей с ростом числа последовательных конфликтов при попытке передачи данного сообщения данной станцией. Метод базируется на нескольких естественных и неоднократно использованных ранее предположениях. В частности, предполагается, что станции сети идентичны и работают независимо друг от друга, находятся в состоянии насыщения (т. е. сообщения для передачи есть всегда), искомая вероятность конфликта p_c является стационарной (и одинаковой для всех станций). Кроме того, предполагается, что время отправки сообщения распределено равномерно внутри окна передачи. В отличие от предшествующих работ, исследование охватывает весь класс протоколов с расширяющимся окном. Ключевым элементом является введение функции протокола F, которая строится на базе функции f. С помощью установленных соотношений между вероятностью конфликта p_c и вероятностью отправки сообщения удалось найти явное выражение для F как функции вероятности p_c. В статье объясняются некоторые эффекты, обсуждавшиеся при исследовании данного протокола в ряде предшествующих работ (например, [3]). В частности, показано, что в нагруженном состоянии и при растущем числе станций в оптимальном режиме работы (т. е. при минимальном среднем цикле передачи $\mathsf{E}S$) вероятность конфликта $p_c \to 1 - e^{-1}$. При тех же условиях найдено, что область стационарности сети при суммарном входном потоке интенсивности λ имеет вид $\lambda < e^{-1}$.

Исследован протокол как с неограниченным, так и с ограниченным числом попыток передачи сообщения.

Литература

1. *Metcalfe R., Boggs D.* Ethernet: Distributed packet switching for local computer networks // Communications of the ACM, 1976. Vol. 19. No. 7. P. 395–404.
2. *Abramson N.* Development of the ALOHANET // IEEE Trans. on Inform. Theory, 1985. Vol. 31. No. 2. P. 119–123.
3. *Shoch J. F., Hupp J. A.* Measured performance of an Ethernet local network // Commun. ACM, 1980. Vol. 23. No. 12. P. 711–721.
4. *IEEE 802.11 Standard.* 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. — N.Y.: IEEE, 2007.
5. *Håstad J., Leighton T., Rogoff B.* Analysis of backoff protocols for multiple access channel // SIAM J. Comput., 1996. Vol. 25. No. 4. P. 740–774.
6. *Kelly F. P.* Stochastic models of computer communication systems //J. Roy. Statist. Soc. B, 1985. Vol. 47. P. 379–395.
7. *Kelly F. P., MacPhee I. M.* The number of packets transmitted by collision detect random access scheme // Annals of Prob., 1987. Vol. 15. P. 1557–1568.
8. *Aldous D. J.* Ultimate instability of exponential back-off protocol for acknowledgement-based transmission control of random access communication channel // IEEE Trans. on Information Theory, 1987. Vol. 33. No. 2. P. 219–223.
9. *Lukyanenko A., Gurtov A.* Performance analysis of general backoff protocols // J. Communications Software and Systems, 2008. Vol. 4. No. 1. P. 13–22.
10. *Вишневский В., Ляхов А., Портной С., Шахнович И.* Широкополосные беспроводные сети передачи информации. — М.: Техносфера, 2005.
11. *Bianchi G.* Performance analysis of the IEEE 802.11 Distributed Coordination Function // IEEE J. Selected Areas in Communications, 2000. Vol. 18. No. 3. P. 535–547.
12. *Kwak B., Song N., Miller L. E.* Performance analysis of exponential backoff // IEEE/ACM Transactions on Networking (TON), 2005. Vol. 13. No. 2. P. 343–355.
13. *Meyn S., Tweedie R. L.* Markov chains and stochastic stability. — New York: Cambridge University Press, 2009.
14. *Roberts L. G.* ALOHA packet system with and without slots and capture // SIGCOMM Comput. Commun., 1975. Vol. 5. No. 2. P. 28–42.
15. *Клейнрок Л.* Вычислительные системы с очередями. — М.: Мир, 1979.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 53–56

РАЗРАБОТКА ПАРАЛЛЕЛЬНЫХ ЭВРИСТИЧЕСКИХ АЛГОРИТМОВ ПОДБОРА ВЕСОВЫХ КОЭФФИЦИЕНТОВ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ

О. В. Крючин[1]

Аннотация: Описан градиентный алгоритм обучения искусственной нейронной сети (ИНС) и основанные на нем эвристические алгоритмы QuickProp и RProp. Рассмотрена возможность применения кластерных систем.

Ключевые слова: искусственная нейронная сеть; эвристические алгоритмы обучения; кластерные системы

1 Введение

Известно, что задача обучения ИНС сводится к минимизации функции

$$\varepsilon = \sum_{i=1}^{N-1}\sum_{j=0}^{P-1}(y_{i,j} - d_{i,j}) = \sum_{i=1}^{N-1}\sum_{j=0}^{P-1}(F(x,w,\mu)_j - d_{i,j}) \ , \quad (1)$$

где y, d — выходные значения ИНС и моделируемого объекта; N, P — число строк и столбцов (выходных данных) в обучающей выборке; $\overline{x}$ — вектор входных значений; $\overline{w}$ — вектор синоптических связей, μ — вектор активационных функций нейронов.

Таким образом, обучение состоит из двух частей: подбора активационных функций (в который также следует включать подбор структуры сети) и подбора весовых коэффициентов. В настоящее время не существует общего алгоритма определения структуры ИНС, подходящего для каждой рассматриваемой проблемы. Часто такую структуру выбирают методом проб и ошибок, который зачастую отнимает у исследователя много времени [1]. Поскольку для каждой структуры необходимо подбирать значения весовых коэффициентов, то ускорение этого подбора должно ускорить весь процесс обучения ИНС.

На данный момент создано огромное количество алгоритмов эвристического типа, представляющих собой в основном модификацию методов наискорейшего спуска или сопряженных градиентов. Подобные модификации широко известных алгоритмов связаны с внесением в них некоторых изменений, ускоряющих (по мнению авторов) процесс обучения. Как правило, такие методы не имеют серьезного теоретического обоснования, особенно это относится к процедуре подбора управляющих параметров. Однако в таких алгоритмах реализуется личный опыт работы авторов с нейронными сетями. К наиболее известным эвристическим алгоритмам относятся QuickProp С. Фальмана [2], а также RProp (Resilient Propagation) М. Ридмиллера и Х. Брауна [3].

Целью данной работы является разработка параллельных версий эвристических алгоритмов подбора весовых коэффициентов QuickProp и RProp.

2 Существующие варианты распараллеливания

Традиционно параллельность нейросетевых вычислений понимается как параллельное функционирование отдельных нейронов или групп нейронов. В лаборатории искусственных нейронных сетей ВНИИТФ предложен другой подход. Если в параллельной вычислительной системе имеется n процессоров, то обучающее множество делится на n равных частей. Процессоры взаимодействуют друг с другом по известной схеме «звезда», т. е. один из процессоров считается центральным. На очередном шаге обучения каждый процессор вычисляет частичный градиент ошибки для своего подмножества обучающих примеров. Затем эти частичные градиенты суммируются на центральном процессоре. Далее центральный процессор вычисляет поправку к весам сети по формулам того или иного оптимизирующего алгоритма, прибавляет ее и рассылает новое приближение весов всем остальным процессорам. Центральный процессор принимает

[1]Тамбовский государственный университет им. Г. Р. Державина, kryuchov@gmail.com

также решение о продолжении или останове итераций, о чем он сообщает всем задействованным процессорам.

Именно такой вариант распараллеливания обучения feed-forward сети был реализован в рамках комплекса нейросетевого моделирования Nimfa. Параллельные программы написаны с использованием стандарта MPI. Реализованы параллельные варианты двух алгоритмов оптимизации: стандартного градиентного спуска и RProp [4]. Для этих методов время подсчета поправки к весам мало по сравнению со временем вычисления суммарного градиента функции ошибки. Поэтому основным фактором, снижающим эффективность параллельной программы, становится межпроцессорный обмен данными [5].

3 Параллельная версия градиентного алгоритма

Алгоритмы QuickProp и RProp базируются на классическом градиентном алгоритме, называемом также алгоритмом наискорейшего спуска. Поэтому для разработки параллельных версий эвристических методов необходимо разработать параллельную версию этого алгоритма.

Суть метода заключается в вычислении вектора градиента и изменении весовых коэффициентов в направлении антиградиента [6, 7]

$$\vec{g} = \frac{\partial \varepsilon}{\partial w} = \left(\frac{\partial \varepsilon(w_0)}{\partial w_0}, \frac{\partial \varepsilon(w_1)}{\partial w_1}, \frac{\partial \varepsilon(w_2)}{\partial w_2}, \ldots, \frac{\partial \varepsilon(w_{l_w-1})}{\partial w_{w_{l_w-1}}} \right);$$

$$g_i = \frac{\partial \varepsilon(w_i)}{\partial w_i} \underset{\Delta w_i \to 0}{=} \frac{\varepsilon(w_i + \Delta w_i) - \varepsilon(w_i)}{\Delta w_i}.$$

Следовательно, для вычисления $\partial\varepsilon(w_i)/\partial w_i$ необходимо определить значение целевой функции при текущем значении весовых коэффициентов, а затем при измененном $w_i = w_i + \Delta w_i$. После вычисления вектора градиента происходит изменение весовых коэффициентов

$$w_i = w_i + \Delta w_i = w_i - s g_i = w_i - s \frac{\partial \varepsilon(w_i)}{\partial w_i}.$$

Как можно заметить, для вычисления нового значения одного из весовых коэффициентов используются значения остальных весовых коэффициентов, которые были на предыдущей итерации. Исходя из этого можно сделать вывод, что элементы вектора градиента могут быть вычислены одновременно, следовательно, этот вектор можно разделить на n частей (по числу процессоров), каждая из которых вычисляется на отдельном узле (для этого процессору необходимо лишь передать текущие значения весовых коэффициентов). После окончания вычисления процессоры не возвращают полученные результаты на ведущий, а изменяют значения приписанных к ним весовых коэффициентов, и уже после этого возвращают результат (новые весовые коэффициенты). Следовательно, каждый вычислительный узел вычисляет l_w/n весовых коэффициентов.

Таким образом, на каждой итерации происходит только две передачи данных — всех весовых коэффициентов на все процессоры и части из них со всех узлов на ведущий [8].

4 Параллельная версия алгоритма QuickProp

В алгоритме QuickProp изменение i-го весового коэффициента на k-й итерации производится согласно правилу:

$$\Delta w_i(k) = -s\left(g_i(k) + c_w w_i(k-1)\right) + q_i(k)\Delta w_i(k-1),$$

где $g_i(k) = \partial\varepsilon(w_i(k))/\partial w_i(k)$ — элемент вектора-градиента; s — коэффициент обучения; c_w — коэффициент минимизации значений весовых коэффициентов; $q_i(k)$ — коэффициент фактора момента.

Отличие от классического градиентного метода заключается в наличии двух слагаемых — минимизатора значений весовых коэффициентов $sc_w w_i(k-1)$ и фактора момента $q_i(k)\Delta w_i(k-1)$. Коэффициент минимизации c_w обычно принимает значение 10^{-4} и служит для ослабления весовых связей (вплоть до полного разрыва), а фактор момента необходим для адаптации алгоритма к текущим результатам обучения. Коэффициент q_i уникален для каждого весового коэффициента и вычисляется в два этапа. На первом определяется величина

$$q_i = \frac{g_i(k)}{g_i(k-1) - g_i(k)}, \qquad (2)$$

а на втором коэффициент момента принимает минимальное значение из q_i и $q_{\max}$. В качестве значения $q_{\max}$ С. Фальманом предложено 1,75 [2].

Существует также модифицированная форма алгоритма, отличающаяся уменьшением числа управляющих параметров без потери эффективности. В модифицированном алгоритме формула (1) заменяется на следующую:

$$\Delta w_i(k) = q_i(k)\Delta w_i(k-1) \quad \text{при } \Delta w_i(k-1) \neq 0;$$
$$\Delta w_i(k) = s g_i(k) \quad \text{при } \Delta w_i(k-1) = 0.$$

То есть в случае, когда элемент вектора градиента $g_i(k) = \partial\varepsilon(w_i(k))/\partial w_i(k)$ принимает нулевое значение, на следующей итерации значение соответствующего ему весового коэффициента вычисляется классическим градиентным методом [9].

Для распараллеливания была выбрана модифицированная версия алгоритма. Как уже отмечалось, отличие от классического градиентного метода заключается в том, что в случае, когда элемент вектора градиента не равен нулю, необходимо вычислять коэффициент момента $q = \min(q,\, q_{\max})$, где q определяется по формуле (2). Следовательно, кроме текущего значения элемента вектора-градиента, необходимо знать его предыдущее значение, поэтому принцип распараллеливания, используемый в классическом градиентном методе, пригоден и здесь.

5 Параллельная версия алгоритма RProp

Суть этого метода заключается в том, что игнорируются значения градиента, а учитывается исключительно знак. Изменение весового коэффициента вычисляется по формулам [3, 6]

$$\Delta w_i(k) = -s_i(k)\theta_g(i)\,;$$

$$s_i(k) = \begin{cases} \min\left(q_a s_i(k-1),\, q_{\max}\right), \\ \qquad\qquad g_i(k)g_i(k-1) > 0\,; \\ \max\left(q_b s_i(k-1),\, q_{\min}\right), \\ \qquad\qquad g_i(k)g_i(k-1) < 0\,; \\ s_i(k-1), \quad g_i(k)g_i(k-1) = 0 \end{cases}$$

$$\theta_g(i) = \begin{cases} 1, & g_i(k) > 0\,; \\ 0, & g_i(k) = 0\,; \\ -1, & g_i(k) < 0\,. \end{cases}$$

В работе [10] предложены следующие значения: $q_a = 1{,}2$, $q_b = 0{,}5$, $q_{\min} = 10^{-6}$, $q_{\max} = 50$ [11].

Как видно из приведенных выше формул, для вычисления нового значения весового коэффициента необходимо знать: значения градиента на текущей и предыдущей итерации, прошлое изменение весового коэффициента, величину коэффициента обучения (шага) на предыдущей итерации. Отсюда можно сделать вывод, что принцип распараллеливания, используемый для классического градиентного метода и алгоритма QuickProp, применим и к этому методу.

6 Теоретическая эффективность алгоритма

Время, необходимое для обучения на многопроцессорной машине, можно выразить формулой

$$\tau = \frac{t}{n} + \Psi\,,$$

где t — время обучения на однопроцессорной машине; n — количество вычислительных узлов.

Иными словами, существует некая величина Ψ, которая представляет собой разницу между тем вре́менным выигрышем, который можно было бы ожидать (в n раз), и тем, который реализуется на практике:

$$\Psi = \eta + \phi + \upsilon + \gamma\,,$$

где η — время межпроцессорной передачи данных; ψ — время, необходимое на подготовку данных; υ — время ожидания одних вычислительных узлов другими; γ — неучтенные временные затраты.

Присутствие этих величин обусловлено тем, что параллельные алгоритмы отличаются от последовательных. Поскольку очень редко удается полностью равномерно распределить нагрузку на вычислительные узлы, всегда на некоторых процессорах происходит ожидание окончания выполнения работы какого-нибудь другого вычислительного узла. Кроме того, средство межпроцессорной передачи данных — MPI — пересылает данные только определенного типа (массивы-указатели), которые отличаются от используемых в алгоритмах обучения высокоуровневых типов (контейнеры), и, соответственно, необходимо конвертирование одного типа в другой, что также занимает определенное время. К тому же большое влияние оказывает аппаратная платформа. Средства межпроцессорной коммуникации имеют различную скорость передачи, и поэтому выбор такого средства (оптоволокно, витая пара, wi fi) также влияет на значение величины Ψ.

Величина Ψ показывает лишь абсолютную разницу между одно- и многопроцессорными временными затратами, что не позволяет адекватно судить об эффективности параллельной версии того или иного метода обучения. Поэтому необходима новая величина, которую можно назвать коэффициентом эффективности параллельного алгоритма

$$\alpha = \frac{t}{n\tau}\,.$$

Эта величина показывает относительный временной выигрыш и для эффективных алгоритмов должна быть в диапазоне от 0,7 до 1.

Что касается сравнения эффективности описываемого способа распараллеливания с методом,

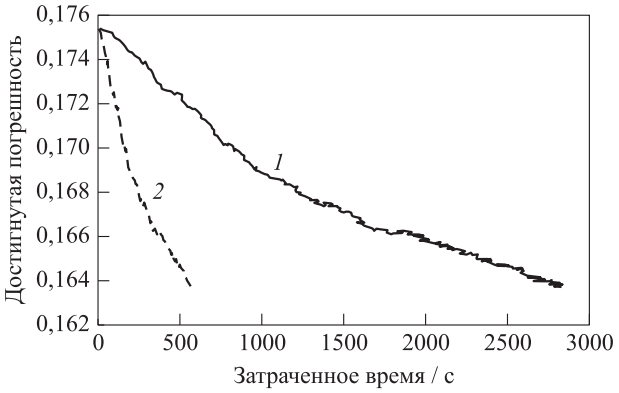

Рис. 1 Зависимость достигнутой погрешности от затраченного времени при подборе весовых коэффициентов последовательным методом QuickProp (*1*) и параллельным, использующим 6 процессоров (*2*)

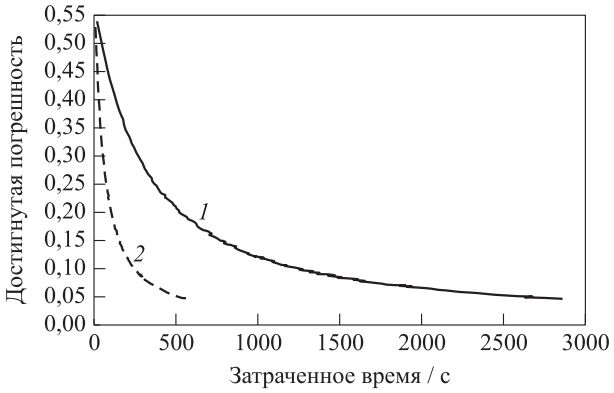

Рис. 2 Зависимость достигнутой погрешности от затраченного времени при подборе весовых коэффициентов последовательным методом RProp (*1*) и параллельным, использующим 6 процессоров (*2*)

разработанным лабораторией ВНИИТФ, то можно заметить, что данный способ требует меньшего объема межпроцессорных передач данных и в случае невысокой скорости интерконнекта является более эффективным.

7 Вычислительный эксперимент

Реализация описанных алгоритмов была выполнена в виде компьютерной программы, написанной на языке C++. В качестве средств межпроцессорной передачи данных была использована библиотека MPI. Эксперименты проводились на кластере Тамбовского государственного университета им. Г. Р. Державина.

Рисунки 1 и 2 показывают время, затраченное на обучение ИНС при использовании последовательных и параллельных алгоритмов подбора весовых коэффициентов. В случае параллельных версий используются 6 процессоров. Как можно видеть, в данном случае параллельные алгоритмы примерно в 5 раз быстрее последовательных, т. е. коэффициент эффективности $\alpha \approx 0{,}83$.

8 Заключение

Приведенные выше результаты показывают, что параллельные эвристические алгоритмы весьма эффективны и могут быть использованы для подбора весовых коэффициентов искусственных нейронных сетей.

Литература

1. *Арзамасцев А. А., Крючин О. В., Азарова П. А., Зенкова Н. А.* Универсальный программный комплекс для компьютерного моделирования на основе искусственной нейронной сети с самоорганизацией структуры // Вестн. Тамбовского университета. Сер. Естественные и технические науки. — Тамбов, 2006. Т. 11. Вып. 4. С. 564–570.
2. *Fahlman S. E.* An empirical study of learning speed in back-propagation networks. Technical report. CMU-CS-88-162. — Carnegie-Mellon University, 1988.
3. *Riedmiller M., Braun H.* RProp — a fast adaptive learning algorithms. Technical Report. — Karlsruhe: University Karlsruhe, 1992.
4. *Riedmiller M., Braun H.* A direct adaptive method for faster backpropagation learning: The RProp algorithm // Proceedings of the IEEE International Conference on Neural Networks (ICNN 93), 1993.
5. *Федорова Н. Н., Терехов С. А.* Параллельная реализация алгоритмов обучения нейронных сетей прямого распространения с использованием стандарта MPI. Адрес в Интернете: http://www.aconts.com/pub/archive/ijcnn99_p423_rus.pdf.
6. *Zadeh L. A.* The concept of linguistic variable and its application to approximate reasoning. Part 1–3 // Information Sci., 1975. P. 199–249.
7. *Gill P., Murray W., Wrights M.* Practical optimisation. — N.Y.: Academic Press, 1981.
8. *Крючин О. В., Арзамасцев А. А., Королев А. Н., Горбачев С. И., Семенов Н. О.* Универсальный симулятор, базирующийся на технологии искусственных нейронных сетей, способный работать на параллельных машинах // Вестн. Тамбовского университета. Сер. Естественные и технические науки. — Тамбов, 2008. Т. 13. Вып. 5. С. 372–375.
9. *Veith A. C., Holmes G. A.* A modified quickprop algorithm // Neural Computation, 1991. Vol. 3. P. 310–311.
10. *Осовский С.* Нейронные сети для обработки информации. — М.: Финансы и статистика, 2002. 344 с.
11. *Riedmiller M.* Untersuchungen zu konvergenz und generalisierungsverhalten uberwachter lernverfahren mit dem SNNS // Proceedings of the SNNS, 1993.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 57–62

ИСПОЛЬЗОВАНИЕ КООРДИНАТНОГО МЕТОДА ФРАГМЕНТАЦИИ КОММУТАТОРНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ СОКРАЩЕНИЯ ТРАФИКА

С. Ю. Степанов[1]

Аннотация: Описана проблема возрастания трафика при масштабировании коммутаторной нейронной сети, предложен метод и алгоритм ее решения. Приведен пример работы разработанного алгоритма.

Ключевые слова: коммутаторная нейронная сеть; масштабирование; трафик

1 Введение

Коммутаторная нейронная сеть — новая технология построения нейронных сетей, позволяющая создавать большие искусственные нейронные сети для задач управления сложными техническими объектами и обработки информации [1].

Традиционный способ построения нейронных сетей предусматривает, что такая сеть состоит только из одного типа элементов — нейронов. Нейрон имеет несколько входов и один выход. Каждому входу приписано некоторое число, называемое весом. Полученный на входные каналы сигнал называется входным сигналом. Он преобразуется в сигнал внутреннего возбуждения нейрона и рассчитывается как сумма произведений значений каждого сигнала на вес соответствующего канала. Выходной сигнал нейрона получается путем воздействия активационной функции на сигнал внутреннего возбуждения. Связываясь между собой, нейроны образуют нейронную сеть — совокупность нейронов, удовлетворяющую следующим требованиям:

- из совокупности всех входных каналов нейронов выделяется множество каналов, называемое множеством входных каналов сети;
- если вход нейрона не является входным каналом сети, то на него подается сигнал выходного канала одного и только одного нейрона этой сети;
- из совокупности всех нейронов сети выделяется некоторая подсовокупность, элементы которой называются выходными нейронами сети, выходные каналы которых образуют множество выходных каналов сети;
- если выходной канал нейрона сети не является выходным каналом сети, то его сигнал подается на входной канал хотя бы одного нейрона сети.

Таким образом, все нейроны выполняют две общие для них функции: обработку получаемой информации, масштабирование и передачу информации к соответствующим входам других нейронов сети.

При таком способе организации возникает ощутимая проблема масштабирования нейронной сети при значительном увеличении числа нейронов. Существенно усложняется работа нейрона по решению транспортной задачи из-за увеличения объема передаваемой информации и размера внутренних таблиц связи. При аппаратной реализации нейронной сети квадратично увеличивается число потенциально поддерживаемых линий передачи информации [2].

Коммутаторная нейронная сеть (рис. 1) отличается от традиционной нейронной сети тем, что задачи транспорта и обработки информации разделены. Такая сеть содержит два типа элементов (каждый из которых выполняет только одну функцию):

- нейрон, обрабатывающий информацию;
- коммутатор, обеспечивающий транспорт и масштабирование информации.

Нейрон коммутаторной нейронной сети — устройство, имеющее один вход и один выход и выполняющее функцию принятия решений. Нейрон изменяет свое состояние дискретно и может быть различного типа в соответствии с требуемой активационной функцией.

[1]ГОУ ВПО МГТУ «Станкин», cympak_shade@rambler.ru

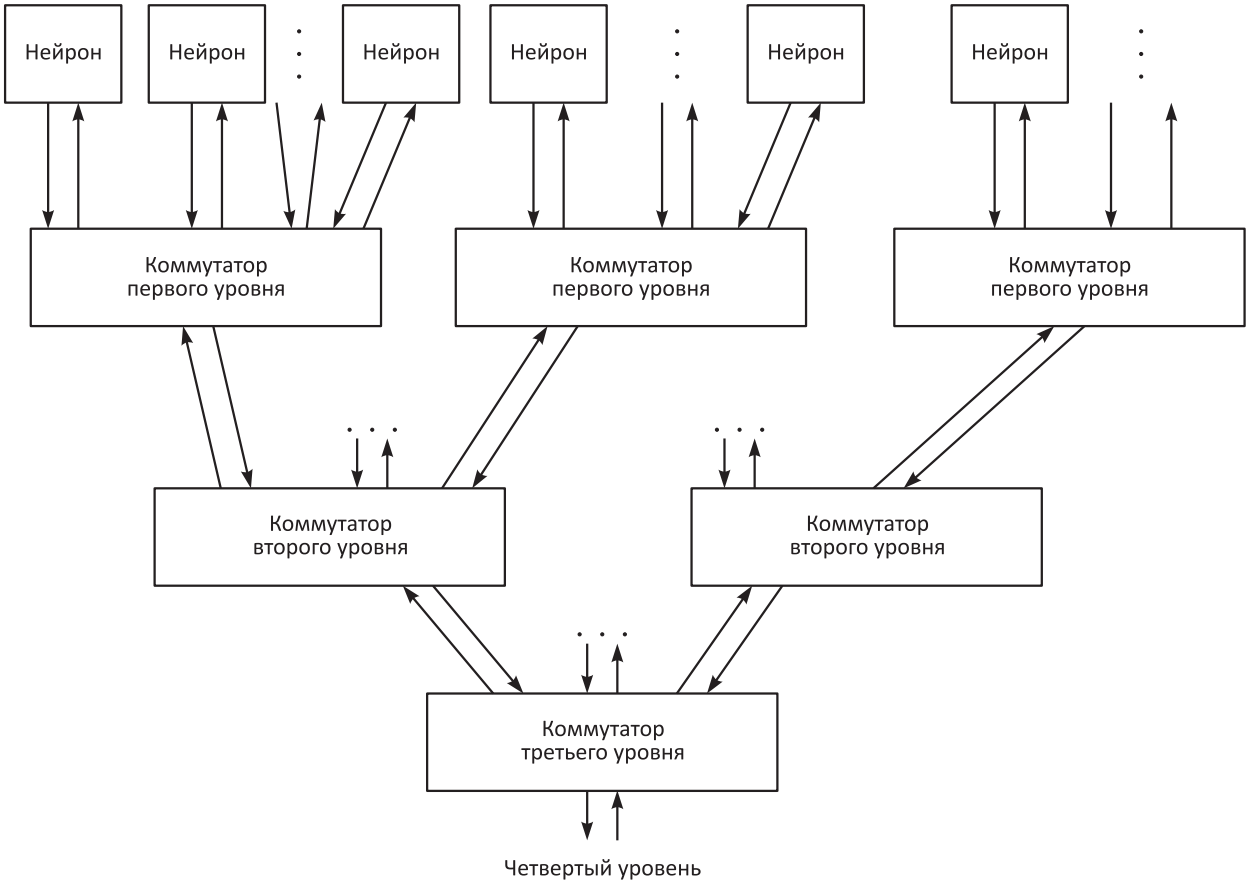

Рис. 1 Коммутаторная нейронная сеть

Коммутатор коммутаторной нейронной сети — устройство, имеющее много входов и много выходов и перераспределяющее информацию между другими коммутаторами и нейронами коммутаторной нейронной сети. Он состоит из таблицы связей нейронов, составляющих фрагмент сети, и устройства, масштабирующего и передающего информацию между нейронами на основе этой таблицы. Число входов и выходов коммутатора ограничено и определяется его реализацией. При обучении нейронной сети изменяется таблица связей нейронов в коммутаторе, а сами нейроны при этом не затрагиваются. В результате применения коммутаторов нейронная сеть приобретает иерархическую древовидную структуру [3].

Базовая нейронная сеть — нейронная сеть, которую обслуживает один коммутатор [4]. Масштабирование нейронной сети осуществляется увеличением числа базовых нейронных сетей. Для интеграции отдельных базовых сетей в единую нейронную сеть используются коммутаторы верхнего уровня.

2 Задача сокращения трафика в коммутаторной нейронной сети

При связи нейронов между собой путь передаваемой информации (трафик) существенно зависит от взаимного расположения нейронов в древовидной структуре [5]. Чем дальше нейроны расположены друг от друга, тем больше трафик информации. Поэтому правомерна задача оптимизации трафика в нейронной сети за счет рационального расположения нейронов на коммутаторах. Внутри одной базовой нейронной сети сократить трафик невозможно. Переставляя нейроны между различными базовыми нейронными сетями, можно обеспечить существенное снижение трафика в нейронной сети. Важным аспектом этой задачи является логическая группировка нейронов, по результатам которой возможно создание отдельных групп (подмножеств нейронов), в целом отвечающих за отдельные подзадачи общей задачи нейронной сети.

При решении оптимизационной задачи необходимо:

- определить, как по обученной нейронной сети выделить подмножества нейронов, отвечающих за отдельные подзадачи общей задачи нейронной сети;
- сгруппировать нейроны в базовые элементарные сегменты, сократив информационный трафик в сети;
- рационально расположить базовые элементарные сегменты на коммутаторах верхнего уровня, сократив информационный трафик в сети.

Группировку следует производить таким образом, чтобы связанные между собой нейроны были расположены максимально близко друг от друга, при этом не потребуется физически перемещать нейроны от одного коммутатора к другому. Достаточно изменить идентификаторы нейронов в дереве, содержащем информацию о структуре сети.

Элементарный сегмент обслуживается одним коммутатором нижнего уровня, емкость которого имеет определенные ограничения. Эти ограничения задают предельный размер базового элементарного сегмента.

Рациональная группировка нейронов позволит по обученной нейронной сети определить, какие подмножества нейронов отвечают за отдельные подзадачи, поскольку существование более или менее изолированной группы нейронов позволяет говорить, что эта группа выполняет более или менее изолированную подзадачу общей задачи. При этом чем меньше связей эта группа имеет с другими группами нейронов, тем более изолированную подзадачу она выполняет.

В полученных таким образом группах большинство связей между нейронами будут внутренними. Решение задачи оптимальной декомпозиции нейронной сети позволит перейти к доменной организации коммутаторной нейронной сети [6]. Для этого необходимо ввести дополнительный элемент — шлюз домена, отделяющий адресное пространство нейронов домена от адресного пространства всей нейронной сети. Доменная структура позволяет выделить автономный элемент, имеющий интерфейс ввода-вывода, облегчить проведение обучения больших нейронных сетей и анализ подзадач, которые они выполняют [7].

3 Математическая постановка задачи

Сопоставим нейронной сети ориентированный граф. Множеству вершин этого графа будет соответствовать множество нейронов сети, а множеству дуг — множество связей между нейронами. Направление дуг графа определяется направлением потока информации в нейронной сети. Граф необходимо разбить на фиксированное число подграфов таким образом, чтобы суммарное число дуг, связывающих подграфы, было минимальным. Все вершины графа должны войти в эти подграфы, причем каждая вершина должна войти только в один подграф.

В связи с физическим ограничением на число нейронов, одновременно подключенных к коммутатору первого уровня, следует ввести дополнительное условие: размер каждого подграфа должен быть меньше или равен числу нейронов, подключенных к коммутатору первого уровня.

Решение задачи разбиения графа на подграфы с учетом заданных условий позволит решить задачу оптимизации трафика в коммутаторной нейронной сети.

4 Метод решения задачи

Рассмотрим плоскость, на которой расположена сетка, в узлах которой в произвольном порядке размещена группа объектов, соответствующих вершинам данного графа (рис. 2). Введем импульс притяжения между объектами таким образом, что если между двумя вершинами в графе существует дуга, то эти две вершины притягиваются друг к другу. Импульс притяжения должен быть постоянным для каждой пары объектов (рис. 3).

Для предотвращения слияния объектов в общее ядро введем импульс отталкивания между всеми объектами. Этот импульс обратно пропорционален расстоянию между объектами. Для того чтобы объекты могли объединяться в группы, следует уменьшить импульс отталкивания, когда объекты сблизились на определенное расстояние.

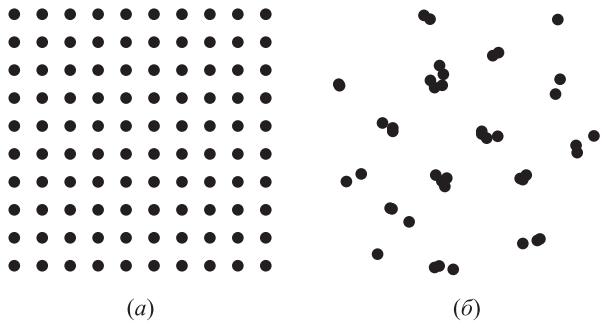

Рис. 2 Работа алгоритма: (*а*) исходное состояние; (*б*) результат группирования

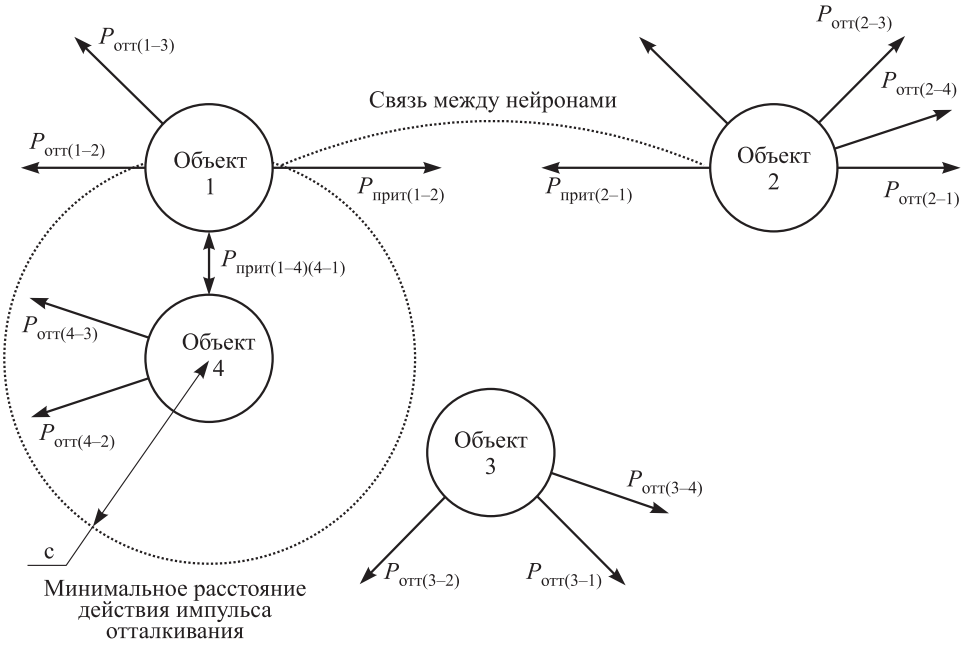

Рис. 3 Действующие на объекты импульсы

Определив результирующий импульс как сумму векторов всех импульсов, действующих на объект, получим направление движения для каждого объекта и значение перемещения. Через некоторое время объекты, имеющие между собой импульсы притяжения, должны оказаться рядом на плоскости. Изменяя параметры импульсов притяжения и отталкивания, можно варьировать количество и размеры получаемых элементарных сегментов.

Используя данный метод группирования многократно, можно решить задачу оптимизации трафика информации, разделив ее на следующие стадии:

– разбиение и группировка нейронов в группы;
– разбиение каждой группы на подгруппы определенного размера;
– группировка подгрупп между собой.

Разработанный алгоритм группирования можно описать следующим образом:

A — коэффициент притяжения, постоянный и положительный;

$W(i,k)$ — вес связи между нейронами i и k в рассматриваемой нейронной сети;

B — коэффициент отталкивания, постоянный и положительный;

$S(i,k)$ — расстояние между объектами i и k на плоскости;

$P_{\text{прит}} = A|W(i,k)|$ — импульс притяжения [кг·м/с];

$P_{\text{отт}} = B/S$ — импульс отталкивания [кг·м/с];

n — число нейронов в сети.

1. Для каждого нейрона сети провести следующие действия:
   1.1 для каждой пары нейронов:
      1.1.1 найти расстояние между соответствующими объектами на плоскости;
      1.1.2 рассчитать воздействующий на объекты импульс как разность $P_{\text{отт}}$ и $P_{\text{прит}}$;
      1.1.3 рассчитать $\Delta x(i,k)$ и $\Delta x(k,i)$ — перемещение каждого объекта из пары;
      1.1.4 рассчитать $\Delta y(i,k)$ и $\Delta y(k,i)$ — перемещение каждого объекта из пары.
2. Переместить каждый объект на плоскости в соответствии с рассчитанными $\Delta x(i)$ и $\Delta y(i)$.
3. Определить, пришла ли система из объектов на плоскости в состояние покоя, сравнив значения перемещений каждого объекта с заданной минимальной величиной.
4. Повторить пп. 1–3 m раз, пока система не придет в состояние равновесия.

Значение m зависит от числа нейронов, величины связей между ними, топологии сети и выбранных коэффициентов A и B. Для рассматриваемых в примерах сетей экспериментально определено, что значение коэффициента m сопоставимо с n^3.

Определим вычислительную трудоемкость разработанного алгоритма и сравним с вычислительной трудоемкостью алгоритма полного перебора [8].

Трудоемкость алгоритма: $O(m)(O(n^2) + O(n) + O(n)) = O(m(n^2 + n))$.

Определим вычислительную трудоемкость алгоритма перебора вариантов размещения нейронов в коммутаторной нейронной сети. Для этого формализуем алгоритм:

1. Для $n!$ возможных комбинаций расположения нейронов:
   - 1.1 подсчитать общее число линий передачи данных в коммутаторной нейронной сети для связи каждой пары нейронов друг с другом;
   - 1.2 определить оптимальный вариант из $n!$ комбинаций.

Трудоемкость алгоритма перебора: $O(n! \cdot n^2)$.

По полученной оценке трудоемкости алгоритмов можно сказать, что для фрагментации нейронных сетей, состоящих из большого числа нейронов (от сотен до десятков тысяч), применение координатного метода дает значительное преимущество по затратам на оптимизацию структуры сети.

5 Пример использования алгоритма группирования

Наиболее сложной структурой нейронной сети является полносвязная нейронная сеть, в которой выходные сигналы всех нейронов подаются на вход каждого нейрона сети, выходные каналы всех нейронов образуют выходной сигнал сети, а также каждый нейрон сети имеет как минимум один канал входной информации, являющейся внешней по отношению к этой сети. В реальных задачах используются сети с неполной топологией, поэтому для исследования работы алгоритма были выбраны полносвязные нейронные сети с произвольными весовыми коэффициентами связей, подвергнутые редукции связей.

Простейшим критерием редукции считается учет величины весов. Веса, которые значительно меньше средних значений (нулевые или близкие к нулю), оказывают незначительное влияние на общий уровень выходного сигнала связанного с ними

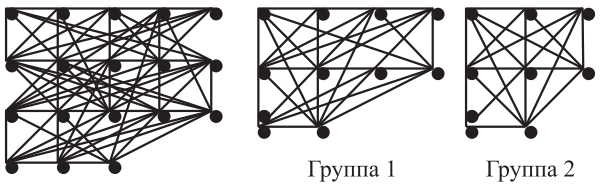

Рис. 4 Пример группирования: (*а*) исходная сеть; (*б*) результат группирования

нейрона. Поэтому их можно отсечь без существенного вреда для его функционирования [9].

Рассмотрим алгоритм подключения полученных групп на примере оптимизации работы тестовой полносвязной нейронной сети с редуцированным числом связей. Предположим, что коммутаторная нейронная сеть состоит из 18 нейронов, двух коммутаторов первого уровня с десятью входами и одним коммутатором второго уровня (рис. 4).

Исходя из структуры сети необходимо подобрать такие коэффициенты, чтобы число полученных групп равнялось 2, а число элементов в одной группе не превышало 10. В результате подбора и группирования были получены 2 группы размерами 10 и 8, что удовлетворяет требованиям к структуре сети.

Заметим, что по сравнению с первым результатом группирования уменьшилось число нейронов, связанных с другой группой, и стало равным 9.

Всего в рассматриваемой сети 126 связей между нейронами. Для сравнения объема трафика в коммутаторной нейронной сети до использования алгоритма группирования и после, возьмем первоначальную группу нейронов и произвольно разобьем ее на 2 подгруппы по 9 нейронов в каждой. Подключив эти подгруппы к коммутаторам первого уровня, получим следующий результат: в первой подгруппе будет 45 связей между элементами, во второй 29, а через коммутатор второго уровня пройдет 52 связи.

Затем подключим подгруппы, полученные в процессе работы алгоритма группирования. На первом коммутаторе первого уровня будет находиться 10 элементов и 65 связей между ними. На втором коммутаторе первого уровня будет находиться 8 элементов и 48 связей между ними. Через коммутатор второго уровня пройдет 13 связей.

В результате, если рассчитать S, суммарное число линий передачи информации коммутатор–коммутатор и коммутатор–нейрон, пройденных потоком информации, получим:

для первого случая $S_1 = (45+29)\cdot 2+52\cdot 4 = 356$;

для второго случая $S_2 = (65+48)\cdot 2+13\cdot 4 = 278$.

Таким образом, для рассматриваемого примера при использовании предложенного алгоритма оптимизации происходит сокращение трафика информации на 21,91% относительно первого варианта подключения.

На каждом шаге работы алгоритма число объектов, связанных между собой и принадлежащих различным группам, не увеличивается. Целью группирования является получение локального минимума суммарного числа связей между коммутаторами разных уровней, необходимых для работы сети. Таким образом, результаты работы предложенного алгоритма удовлетворяют решению поставленной задачи оптимизации трафика информации в коммутаторной нейронной сети.

Методом случайной генерации связей были получены три нейронные сети, состоящие из 169 нейронов, коэффициенты и наличие связей между двумя любыми объектами получены случайно с одинаковой для всех вероятностью [10]. В результате группирования каждой сети было получено около 30 групп с максимальным размером группы, равным 20. Суммарное число связей во всех случаях группирования каждой из рассматриваемых сетей находилось на отрезке [1254, 1302].

В среднем сокращение трафика информации при использовании предложенных алгоритмов по сравнению с произвольным подключением к коммутаторам для данных групп составило 5,5%.

Для множества сетей из 400 нейронов и приблизительно таким же распределением связей сокращение трафика составило около 9%.

Таким образом, даже для сети, по топологии близкой к полносвязной, в которой отсутствуют фрагменты, явно отвечающие за отдельные подзадачи, а соответственно, и более изолированные от всей сети, можно получить достаточно заметное сокращение трафика.

6 Заключение

Разработанный алгоритм может подвергаться масштабированию [11]. Для этого необходимо провести фрагментацию нейронной сети — произвольным образом разбить большую нейронную сеть на группы требуемого размера и проводить группирование каждой группы по отдельности. Затем провести замену полученных в результате работы алгоритма групп на объекты, имеющие соответствующие связи с другими группами сети. И в дальнейшем применять алгоритм к полученным объектам. Необходимым условием в таком случае является ограничение на размер получаемых групп, который должен быть при первичном группировании заведомо меньше размера базовой нейронной сети. В ситуации получения итогового размера групп нейронов значительно большего, чем размер базовой нейронной сети, следует дополнительно фрагментировать группы с использованием разработанного алгоритма.

Коммутаторная нейронная сеть дает возможность создавать нейронные сети любой архитектуры и практически любого объема, что позволяет, в свою очередь, создавать системы управления оборудованием и технологическими процессами, а также реализовывать системы искусственного интеллекта. Разработанный метод фрагментации коммутаторной нейронной сети позволяет оптимизировать ее структуру и сократить затраты на передачу информации между нейронами.

Литература

1. *Кабак И. С., Суханова Н. В.* Нейронная сеть. Патент на ПМ № 75247 РФ.
2. Теория нейронных сетей: Учеб. пособие для вузов / Под общ. ред. А. И. Галушкина. Кн. 1. — М.: ИПРЖР, 2000. 416 с.
3. *Кабак И. С., Суханова Н. В.* Большие нейронные сети в системах управления // Тр. XVI междунар. научно-технич. конф. «Информационные средства и технологии». Т. 3. — М.: МЭИ, 2008. С. 204–210.
4. *Кабак И. С.* Коммутаторная архитектура больших нейронных сетей // Тр. XV междунар. научно-технич. конф. «Информационные средства и технологии». Т. 3. — М.: МЭИ, 2007. С. 124–127.
5. *Кабак И. С., Степанов С. Ю.* Оптимизация трафика информации в коммутаторной нейронной сети // Тр. XIV междунар. научно-технич. конф. «Информационные средства и технологии». Т. 3. — М.: МЭИ, 2006. С. 163–167.
6. *Кабак И. С., Суханова Н. В.* Доменная нейронная сеть. Патент на ПМ № 72084 РФ.
7. *Кабак И. С.* Доменная организация коммутаторных нейронных сетей // Тр. XV междунар. научно-технич. конф. «Информационные средства и технологии». Т. 3. — М.: МЭИ, 2007. С. 128–131.
8. *Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К.* Алгоритмы: построение и анализ // Introduction to Algorithms / Под ред. И. В. Красикова. 2-е изд. — М.: Вильямс, 2005. 1296 с.
9. *Осовский С.* Нейронные сети для обработки информации / Пер. с польск. И. Д. Рудинского. — М.: Финансы и статистика, 2004. 344 с.
10. *Степанов С. Ю.* Группирование нейронов в двухуровневой коммутаторной нейронной сети // Тр. XV междунар. научно-технич. конф. «Информационные средства и технологии». Т. 3. — М.: МЭИ, 2007. С. 178–181.
11. *Степанов С. Ю.* Снижение трафика информации в коммутаторной нейронной сети на основе ее фрагментации // Тр. XVI междунар. научно-технич. конф. «Информационные средства и технологии». Т. 3. — М.: МЭИ, 2008. С. 236–241.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 63–74

О ПРЕДЕЛЬНОМ ПОВЕДЕНИИ МОЩНОСТЕЙ КРИТЕРИЕВ В СЛУЧАЕ РАСПРЕДЕЛЕНИЯ ЛАПЛАСА

В. Е. Бенинг[1], Р. А. Королев[2]

Аннотация: С использованием сходимости условных мер, зависящих от параметра, доказана формула для предела нормированной разности мощностей наилучшего и асимптотически оптимального критериев в случае распределения Лапласа. В связи с нерегулярностью распределения Лапласа логарифм отношения правдоподобия допускает нерегулярное стохастическое разложение; кроме того, для знаковой статистики, лежащей в основе асимптотически оптимального критерия, не выполняется аналог условия Крамера (С). Поэтому непосредственное применение теоремы 3.2.1 работы [1] или теоремы 2.1 работы [2] затруднительно и в настоящей работе их доказательства пересмотрены для случая распределения Лапласа.

Ключевые слова: функция мощности; условная вероятностная мера; условный момент; распределение Лапласа

1 Введение

В 1774 г. Пьер-Симон Лаплас в своей статье «Sur la probabilité des causes par les événements» (см. [3] и литературу там же) предложил естественный вероятностный закон для ошибки измерений в такой формулировке: логарифм частоты ошибки есть линейная функция абсолютного значения ошибки. Естественность этого вероятностного закона он объяснял так: «Чем дальше результат измерения от истинного значения, тем менее вероятным он должен быть, при этом такое уменьшение вероятности не может быть постоянным. Поскольку нет причин считать, что с ростом ошибки сами вероятности и последовательные разности между вероятностями убывают по-разному, то следует приравнять отношения двух бесконечно близких разностей вероятностей и двух бесконечно близких вероятностей. Интегрирование этого равенства показывает, что вероятность ошибки выражается как экспоненциальная функция самой ошибки независимо от ее знака».

Назвав этот закон первым законом для ошибки измерений, который исторически является первым вероятностным распределением с неограниченным носителем, Лаплас уже через 4 года в своей фундаментальной работе “Théorie Analytique” (см. [3] и литературу там же) рассматривает второй вероятностный закон, который гласит: логарифм частоты ошибки измерений есть квадратичная функция ошибки. Именно этот второй закон из-за хороших аналитических свойств будет детально исследоваться все последующее время, получит название «нормальное распределение» и займет главное место в теории вероятностей благодаря центральной предельной теореме. Лишь спустя почти 150 лет известный экономист и математик Дж. Кейнс (см. [3] и литературу там же) напомнит о существовании первого закона для ошибки измерений и получит его вновь из предположения, что наиболее вероятное значение измеряемой величины есть ее медиана. Следом за ним известный математик Э. Уилсон (см. [3] и литературу там же) с помощью непараметрических методов покажет на одном примере, что распределение отклонений от медианы измерений является скорее первым законом Лапласа, нежели нормальным законом. Спустя еще почти 50 лет в научной литературе (см. [3] и литературу там же) все чаще стали появляться призывы к использованию первого закона Лапласа в качестве основного распределения для экономических, биометрических и демографических данных в противовес нормальному распределению.

В наши дни первый закон Лапласа называют распределением Лапласа или двойным экспоненциальным распределением, указывая на возможность получения его как разности двух независимых одинаково распределенных экспоненциальных величин, которые часто используются для описания продолжительности жизни наблюдаемых объектов. Связь с χ^2-распределением также способствует распространению распределения Лапласа. Оно приковывает внимание многих математиков и находит все более широкое применение в прикладных областях

[1]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, bening@yandex.ru

[2]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, stochastique@gmail.com

экономики и науки. Нерегулярность распределения Лапласа долгое время вызывала трудности при получении результатов в задачах проверки статистических гипотез. Однако развитые в последние десятилетия асимптотические методы теории статистических гипотез (см. [1, 2, 4, 5] и литературу там же) позволяют теперь решать многие из таких задач. Одна из них — задача о потере мощности асимптотически оптимального критерия — рассматривается в данной работе.

В работе исследуется формула для предела нормированной разности мощностей критериев в случае распределения Лапласа (см. теорему 3.2.1 работы [1] и теорему 2.1 работы [2]). Следуя работам [6–8], рассмотрим задачу проверки простой гипотезы

$$\mathsf{H}_0 : \theta = 0 \tag{1}$$

против последовательности сложных близких альтернатив вида

$$\mathsf{H}_{n,1} : \theta = \frac{t}{\sqrt{n}}\,, \quad 0 < t \le C\,, \quad C > 0\,,$$

на основе выборки $\mathbf{X}_n = (X_1, \ldots, X_n)$, являющейся элементом выборочного пространства $(\mathcal{X}_n, \mathcal{A}_n)$ и состоящей из независимых и одинаково распределенных наблюдений с плотностью

$$p(x, \theta) = \frac{1}{2}\, e^{-|x-\theta|}\,, \quad x,\, \theta \in \mathsf{R}^1\,.$$

Обозначим через $\mathsf{P}_{n,0}$ и $\mathsf{P}_{n,1}$ распределения $\mathcal{L}(\mathbf{X}_n)$ при гипотезе H_0 и альтернативе $\mathsf{H}_{n,1}$, соответственно плотности распределений выборки будем обозначать $p_{n,0}(x)$ и $p_{n,1}(x)$.

В работе [8] на основании сходимости условных моментов и с помощью асимптотических разложений была доказана формула для предела нормированной разности мощности асимптотически оптимального критерия, основанного на статистике

$$S_n = \frac{t}{\sqrt{n}} \sum_{i=1}^{n} \operatorname{sign}\,(X_i) - \frac{t^2}{2}\,, \tag{2}$$

и мощности наилучшего критерия, основанного на логарифме отношения правдоподобия

$$\Lambda_n = \sum_{i=1}^{n} \left(|X_i| - |X_i - tn^{-1/2}| \right)\,. \tag{3}$$

При этом асимптотические разложения для функций распределения знаковой статистики S_n при гипотезе H_0 и альтернативе $\mathsf{H}_{n,1}$ были получены и для точек разрыва, что соответствовало статистической постановке задачи.

Естественным выглядит отказ от применения асимптотических разложений с целью формирования достаточных условий, подобных условиям основных теорем работ [1, 2], при доказательстве формулы для предела нормированной разности мощностей критериев (см. (6) и теорему 2.8).

В настоящей работе в доказательстве формулы (6) используются главным образом асимптотические свойства статистик Λ_n (см. (3)) и разности $\Delta_n = S_n - \Lambda_n$ (см. леммы 2.1–2.3) и методы сходимости условных моментов и условных мер, зависящих от параметра (см. [1, 2]). При этом асимптотические свойства логарифма отношения правдоподобия Λ_n описываются с помощью теорем об асимптотических разложениях функций распределения при гипотезе и альтернативе из работы [5].

Заметим, что теорема 3.2.1 работы [1] не может непосредственно быть применена к случаю распределения Лапласа, поскольку достаточное условие 3 (ii) (см. [1], с. 79) — аналог условия Крамера (C) — не выполняется для решетчатой статистики S_n. Также в связи с нерегулярностью распределения Лапласа статистика Λ_n допускает нерегулярное стохастическое разложение, поэтому скорость сходимости мощностей критериев есть $n^{-1/2}$ в отличие от случая n^{-1} в формулировке теоремы 2.1 работы [2]. Таким образом, в настоящей работе доказательства теоремы 3.2.1 работы [1] и теоремы 2.1 работы [2] пересматриваются для случая распределения Лапласа.

2 Формула для разности мощностей критериев

В этом разделе докажем формулу для предела нормированной разности мощностей критериев в случае распределения Лапласа.

Рассмотрим последовательность критериев

$$\Psi_n^*(\Lambda_n) = \begin{cases} 0\,, & \Lambda_n < c_n\,; \\ \gamma_n^*\,, & \Lambda_n = c_n\,; \\ 1\,, & \Lambda_n > c_n\,, \end{cases} \tag{4}$$

основанную на последовательности Λ_n (см. (3)), и последовательность критериев

$$\Psi_n(S_n) = \begin{cases} 0\,, & S_n < \bar{d}_n\,; \\ \gamma_n\,, & S_n = \bar{d}_n\,; \\ 1\,, & S_n > \bar{d}_n\,, \end{cases}$$

основанную на последовательности S_n (см. (2)), уровня α, $\alpha \in (0, 1)$, так что

$$\mathsf{E}_{n,0}\Psi_n^*(\Lambda_n) = \mathsf{E}_{n,0}\Psi_n(S_n) = \alpha + o\left(\tau_n^2\right)\,, \tag{5}$$

где $\tau_n \equiv n^{-1/4}$, а $\mathsf{E}_{n,0}$ и $\mathsf{E}_{n,1}$ — математические ожидания по отношению к $\mathsf{P}_{n,0}$ и $\mathsf{P}_{n,1}$ соответственно.

Обозначим через $\beta_n^* = \mathsf{E}_{n,1}\Psi_n^*(\Lambda_n)$ и $\beta_n = \mathsf{E}_{n,1}\Psi_n(S_n)$ мощности соответствующих критериев.

Докажем, что в случае распределения Лапласа справедлива формула для предела разности мощностей критериев r в терминах условной дисперсии

$$r \equiv \lim_{n\to\infty} \tau_n^{-2}\,(\beta_n^* - \beta_n) = \frac{1}{2}\,e^b \mathsf{D}\,[\Delta|\Lambda = b]\,p(b)\,, \quad (6)$$

где Δ и Λ — случайные величины, такие что

$$\mathcal{L}((\tau_n^{-1}\Delta_n, \Lambda_n)|\mathsf{H}_0) \to \mathcal{L}(\Delta, \Lambda)\,;$$

$$\Delta_n = \begin{cases} 0\,, & S_n = \Lambda_n = +\infty\,, \\ S_n - \Lambda_n\,, & \text{в остальных случаях}\,; \end{cases} \quad (7)$$

$$b = \Phi_1^{-1}(1-\alpha)\,. \quad (8)$$

Здесь $\Phi_1(x)$ обозначает предельную функцию распределения случайных величин Λ_n при гипотезе H_0 (см. (1)), $p(x)$ — ее функция плотности.

Учитывая, что в случае распределения Лапласа плотности $p_{n,0}(x)$, $p_{n,1}(x)$ не обращаются в нуль, т. е.

$$p_{n,0}(x) > 0\,,\ \ p_{n,1}(x) > 0\,,\ \ \forall x \in \mathcal{X}_n\,, \quad (9)$$

получаем из (2) и (3) для любого n

$$|S_n| < \infty\,, \quad |\Lambda_n| < \infty\,.$$

Тогда (7) запишется с точностью до множеств меры ноль (см. (3.4) работы [7])

$$\Delta_n = S_n - \Lambda_n = -\frac{t^2}{2} - \\ - 2\sum_{i=1}^{n}(X_i - tn^{-1/2})\mathbb{1}_{[0,tn^{-1/2}]}(X_i)\,.$$

Леммы 2.1 и 3.1 работы [7] утверждают, что

$$\mathcal{L}(\Lambda_n|\mathsf{H}_0) \to \mathcal{N}\left(-\frac{t^2}{2}, t^2\right)\,; \quad (10)$$

$$\mathcal{L}(\tau_n^{-1}\Delta_n|\mathsf{H}_0) \to \mathcal{N}\left(0, \frac{2t^3}{3}\right)\,;$$

$$\mathcal{L}\left((\tau_n^{-1}\Delta_n, \Lambda_n)\,|\mathsf{H}_0\right) \to \mathcal{N}_2\left(0, \frac{2t^3}{3}, 0, -\frac{t^2}{2}, t^2\right), \quad (11)$$

где $\mathcal{N}_2$ — двумерный нормальный закон с соответствующими параметрами.

Таким образом, необходимо показать, что формулы (6) и (8) справедливы с

$$b = tu_\alpha - \frac{t^2}{2}\,, \quad r = \frac{t^2}{3}\varphi(u_\alpha - t)\,,$$

где $\Phi(x)$ — функция распределения стандартного нормального закона, $\varphi(x)$ — его плотность, $\Phi(u_\alpha) = 1 - \alpha$.

Приведем некоторые свойства статистик Λ_n и Δ_n в случае распределения Лапласа (леммы 2.1–2.3). Из леммы 3.2 работы [6] следует

Лемма 2.1. *Для $\tau_n = n^{-1/4}$ и непрерывно дифференцируемой функции $\Phi_1(x)$, не зависящей от n и имеющей ограниченную производную $p(x) = \Phi_1'(x) > 0$, выполняется*

$$(i)\quad \sup_x \left|\mathsf{P}_{n,0}(\Lambda_n < x) - \Phi_1(x)\right| = \mathcal{O}(\tau_n^2)\,;$$

$$(ii)\quad \sup_{x\le x_0} \mathsf{P}_{n,1}\left(x \le \Lambda_n \le x + \tau_n^{2+\beta}\right) = o\left(\tau_n^2\right)$$

для некоторого $\beta > 0$ и любого $x_0 \in \mathsf{R}^1$.

В работе [6] (см. лемму 3.4) была доказана.

Лемма 2.2. *Для любого $x > 0$ справедливо следующее неравенство:*

$$\mathsf{P}_{n,0}\left(\tau_n^{-1}\,|\Delta_n| \ge x\right) \le Ce^{-x}\,,\ \ C > 0\,.$$

Отсюда, а также из (10) (см. лемму 2.2 работы [8]) следует

Лемма 2.3. *Для $\tau_n = n^{-1/4}$ и $\eta_n = n^{-1/8}$ выполняются соотношения:*

$$(i)\quad \eta_n^{-1}\mathsf{E}_{n,0}\Delta_n^2\mathbb{1}_{(\eta_n,\infty)}\left(|\Delta_n|\right) = o\left(\tau_n^2\right)\,;$$

$$(ii)\quad \mathsf{E}_{n,0}e^{\Lambda_n}\mathbb{1}_{(\eta_n,\infty)}\left(|\Delta_n|\right) = o\left(\tau_n^2\right)\,,$$

где $\mathbb{1}_A(\cdot)$ — индикаторная функция множества A.

Рассмотрим теперь свойства мощностей критериев Ψ_n и Ψ_n^* в случае распределения Лапласа (леммы 2.4, 2.6). Из леммы 2.1 следует

Лемма 2.4. *Для мощности β_n^* справедливо равенство*

$$\beta_n^* = \mathsf{P}_{n,1}\left(\Lambda_n > c_n\right) + o\left(\tau_n^2\right)\,.$$

Доказательство. Доказательство следует из равенства

$$\beta_n^* = \mathsf{E}_{n,1}\Psi_n^*(\Lambda_n) = \mathsf{P}_{n,1}\left(\Lambda_n > c_n\right) + \\ + \gamma_n^*\mathsf{P}_{n,1}\left(\Lambda_n = c_n\right)\,,$$

неравенства $0 \le \gamma_n^* \le 1$ и соотношения (см. лемму 2.1 (ii))

$$\mathsf{P}_{n,1}\left(\Lambda_n = c_n\right) \le \mathsf{P}_{n,1}\left(c_n \le \Lambda_n \le c_n + \tau_n^{2+\beta}\right) = \\ = o\left(\tau_n^2\right)\,.$$

Ограниченность c_n следует из леммы 2.1 (i). □

Введем в рассмотрение сглаженные случайные величины

$$\tilde{\Lambda}_n = \Lambda_n + \xi_n\,, \quad \xi_n \sim \mathcal{N}(0, \sigma_n^2)\,, \tag{12}$$

где

$$\sigma_n^2 = \mathcal{O}\left(\tau_n^{4+4\beta}\right)\,, \tag{13}$$

$\beta > 0$ — постоянная, такая же как в лемме 2.1, причем случайная величина ξ_n не зависит от $\mathbf{X}_n$ при H_0, и $\mathcal{N}(\mu, \sigma^2)$ — нормальный закон с параметрами (μ, σ^2).

Определим

$$\tilde{\beta}_n^* = \mathsf{E}_{n,0} e^{\tilde{\Lambda}_n} \mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right)\,;$$
$$\tilde{\beta}_n = \mathsf{E}_{n,0} e^{\tilde{\Lambda}_n} \Psi_n(S_n)\,.$$

Следующее утверждение хорошо известно, но ради полноты изложения приведем его доказательство.

Лемма 2.5. *Пусть $X \sim \mathcal{N}(0, \sigma^2)$, тогда для любого $a \in \mathsf{R}^1$*

$$\mathsf{E} e^{aX} = e^{a^2\sigma^2/2}\,.$$

Доказательство.

$$\mathsf{E} e^{aX} = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{\infty} \exp\left\{ax - \frac{x^2}{2\sigma^2}\right\} dx =$$
$$= \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\left(x^2 - 2xa\sigma^2 + a^2\sigma^4 - \right.\right.$$
$$\left.\left. - a^2\sigma^4\right)\right\} dx =$$
$$= e^{a^2\sigma^2/2} \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{\infty} e^{-(x-a\sigma^2)^2/(2\sigma^2)}\, dx = e^{a^2\sigma^2/2}\,. \square$$

Далее справедлива следующая

Лемма 2.6. *Для мощностей β_n и β_n^* справедливы равенства*

$$\beta_n = \tilde{\beta}_n + o\left(\tau_n^2\right)\,;$$
$$\beta_n^* = \tilde{\beta}_n^* + o\left(\tau_n^2\right)\,.$$

Доказательство. Учитывая (9), имеем

$$\mathsf{E}_{n,1} \Psi_n(S_n) = \mathsf{E}_{n,0} e^{\Lambda_n} \Psi_n(S_n)\,.$$

Тогда

$$\left|\beta_n - \tilde{\beta}_n\right| = \left|\mathsf{E}_{n,1}\Psi_n(S_n) - \mathsf{E}_{n,0} e^{\tilde{\Lambda}_n}\Psi_n(S_n)\right| =$$
$$= \left|\mathsf{E}_{n,0} e^{\Lambda_n}\Psi_n(S_n) - \mathsf{E}_{n,0} e^{\tilde{\Lambda}_n}\Psi_n(S_n)\right| \le$$
$$\le \mathsf{E}_{n,0} e^{\Lambda_n}\left|e^{\xi_n} - 1\right|\,.$$

В силу равенства

$$\mathsf{E}_{n,0} e^{\Lambda_n} = 1$$

и независимости $\mathbf{X}_n$ и ξ_n имеем

$$\left|\beta_n - \tilde{\beta}_n\right| \le \mathsf{E}\left|e^{\xi_n} - 1\right| \le \left(\mathsf{E}(e^{\xi_n} - 1)^2\right)^{1/2} =$$
$$= \left(\mathsf{E} e^{2\xi_n} - 2\mathsf{E} e^{\xi_n} + 1\right)^{1/2}\,.$$

Используя теперь лемму 2.5, получим неравенство

$$\left|\beta_n - \tilde{\beta}_n\right| \le \left(e^{2\sigma_n^2} - 2e^{\sigma_n^2/2} + 1\right)^{1/2} \le$$
$$\le \left(1 + 2\sigma_n^2 - 2 - \sigma_n^2 + 1 + o(\sigma_n^2)\right)^{1/2} =$$
$$= \mathcal{O}(\sigma_n) = o\left(\tau_n^2\right)\,.$$

Аналогично, используя лемму 2.4, имеем

$$\left|\beta_n^* - \tilde{\beta}_n^*\right| =$$
$$= \left|\mathsf{E}_{n,0} e^{\Lambda_n}\mathbb{1}_{(c_n,\infty)}(\Lambda_n) - \mathsf{E}_{n,0} e^{\tilde{\Lambda}_n}\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right)\right| +$$
$$+ o\left(\tau_n^2\right) \le \mathsf{E}\left|e^{\xi_n} - 1\right| +$$
$$+ \left|\mathsf{E}_{n,0} e^{\tilde{\Lambda}_n}\left(\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right) - \mathbb{1}_{(c_n,\infty)}(\Lambda_n)\right)\right| + o\left(\tau_n^2\right) \le$$
$$\le \left|\mathsf{E}_{n,0} e^{\tilde{\Lambda}_n}\left(\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right) - \mathbb{1}_{(c_n,\infty)}(\Lambda_n)\right)\right| + o\left(\tau_n^2\right)\,.$$

Докажем, что первое слагаемое в правой части последнего неравенства есть величина порядка $o\left(\tau_n^2\right)$. Заметим, что

$$\mathsf{P}(|\xi_n| > \tau_n^{2+\beta}) = \mathsf{P}\left(|\xi| > \frac{\tau_n^{2+\beta}}{\sigma_n}\right) =$$
$$= \mathsf{P}(|\xi_n| > \tau_n^{-\beta}) = 2\left(1 - \Phi\left(\tau_n^{-\beta}\right)\right)\,,$$

где $\xi \sim \mathcal{N}(0, 1)$.

Используя теперь лемму 7.1.2 из [9], получаем для любого $\lambda > 0$

$$\mathsf{P}\left(|\xi_n| > \tau_n^{2+\beta}\right) \le \frac{2}{\sqrt{2\pi}} \tau_n^{\beta} e^{-\tau_n^{-2\beta}/2} = o\left(\tau_n^{\lambda}\right)\,. \tag{14}$$

Запишем $\left|\mathsf{E}_{n,0} e^{\tilde{\Lambda}_n}\left(\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right) - \mathbb{1}_{(c_n,\infty)}(\Lambda_n)\right)\right| \equiv$ $\equiv I_1 + I_2$, где с учетом леммы 2.5 для $\lambda = 4$ имеем

$$I_1 = \left|\mathsf{E}_{n,0}e^{\tilde{\Lambda}_n}\left(\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right) - \mathbb{1}_{(c_n,\infty)}(\Lambda_n)\right)\mathbb{1}_{(\tau_n^{2+\beta},\infty)}(|\xi_n|)\right| \le \mathsf{E}_{n,0}e^{\Lambda_n}\mathsf{E}e^{\xi_n}\mathbb{1}_{\left(\tau_n^{2+\beta},\infty\right)}\left(|\xi_n|\right) \le$$
$$\le \left(\mathsf{E}e^{2\xi_n}\right)^{1/2}\left(\mathsf{P}\left(|\xi_n| > \tau_n^{2+\beta}\right)\right)^{1/2} = o\left(\tau_n^2\right) .$$

Далее

$$I_2 = \left|\mathsf{E}_{n,0}e^{\tilde{\Lambda}_n}\left(\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right) - \mathbb{1}_{(c_n,\infty)}\left(\Lambda_n\right)\right)\mathbb{1}_{\left[0,\tau_n^{2+\beta}\right]}\left(|\xi_n|\right)\right| =$$
$$= \left|\frac{1}{\sqrt{2\pi}\sigma_n}\int\limits_{-\tau_n^{2+\beta}}^{\tau_n^{2+\beta}} e^{y-y^2/(2\sigma_n^2)}\left(\mathsf{E}_{n,0}e^{\Lambda_n}\left(\mathbb{1}_{(c_n,\infty)}\left(\Lambda_n+y\right) - \mathbb{1}_{(c_n,\infty)}\left(\Lambda_n\right)\right)\right)\,dy\right| =$$
$$= \frac{1}{\sqrt{2\pi}\sigma_n}\left|\int\limits_{-\tau_n^{2+\beta}}^{\tau_n^{2+\beta}} e^{y-y^2/(2\sigma_n^2)}\left(\mathsf{P}_{n,1}\left(\Lambda_n > c_n - y\right) - \mathsf{P}_{n,1}\left(\Lambda_n > c_n\right)\right)\,dy\right| \le$$

$$\le \frac{1}{\sqrt{2\pi}\sigma_n}\int\limits_0^{\tau_n^{2+\beta}} e^{y-y^2/(2\sigma_n^2)}\mathsf{P}_{n,1}\left(c_n - y < \Lambda_n \le c_n\right)\,dy + \frac{1}{\sqrt{2\pi}\sigma_n}\int\limits_{-\tau_n^{2+\beta}}^{0} e^{y-y^2/(2\sigma_n^2)}\mathsf{P}_{n,1}\left(c_n < \Lambda_n \le c_n - y\right)\,dy =$$

$$= e^{\sigma_n^2/2}\frac{1}{\sqrt{2\pi}\sigma_n}\int\limits_0^{\tau_n^{2+\beta}} e^{-(y-\sigma_n^2)^2/(2\sigma_n^2)}\mathsf{P}_{n,1}\left(c_n - y \le \Lambda_n \le c_n\right)\,dy +$$

$$+ e^{\sigma_n^2/2}\frac{1}{\sqrt{2\pi}\sigma_n}\int\limits_{-\tau_n^{2+\beta}}^{0} e^{-(y-\sigma_n^2)^2/(2\sigma_n^2)}\mathsf{P}_{n,1}\left(c_n \le \Lambda_n \le c_n - y\right)\,dy\,.$$

Поскольку величина c_n ограничена (см. лемму 2.1 (i)), то с учетом леммы 2.1 (ii) имеем

$$\mathsf{P}_{n,1}\left(c_n - \tau_n^{2+\beta} \le \Lambda_n \le c_n\right) = o\left(\tau_n^2\right) .$$

Отсюда следует утверждение леммы. □

Докажем еще одну вспомогательную лемму.

Лемма 2.7. *В случае распределения Лапласа справедливо равенство*

$$\mathsf{E}_{n,0}\Psi_n^*(\tilde{\Lambda}_n) = \mathsf{E}_{n,0}\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right) = \alpha + o\left(\tau_n^2\right) .$$

Доказательство. В силу равенства (5) имеем

$$\mathsf{E}_{n,0}\Psi_n^*\left(\tilde{\Lambda}_n\right) = \mathsf{E}_{n,0}\Psi_n^*\left(\Lambda_n\right) + \mathsf{E}_{n,0}\left(\Psi_n^*\left(\tilde{\Lambda}_n\right) - \Psi_n^*\left(\Lambda_n\right)\right) = \alpha + \mathsf{E}_{n,0}\left(\Psi_n^*\left(\tilde{\Lambda}_n\right) - \Psi_n^*\left(\Lambda_n\right)\right) + o\left(\tau_n^2\right) .$$

Покажем, что второе слагаемое в правой части этого равенства есть $o\left(\tau_n^2\right)$. Имеем

$$\left|\mathsf{E}_{n,0}\left(\Psi_n^*\left(\tilde{\Lambda}_n\right) - \Psi_n^*\left(\Lambda_n\right)\right)\right| \equiv J_1 + J_2\,,$$

где с учетом (14) для $\lambda = 2$

$$J_1 = \left|\mathsf{E}_{n,0}\left(\Psi_n^*\left(\tilde{\Lambda}_n\right) - \Psi_n^*\left(\Lambda_n\right)\right)\mathbb{1}_{\left(\tau_n^{2+\beta},\infty\right)}\left(|\xi_n|\right)\right| \le \mathsf{P}\left(|\xi_n| > \tau_n^{2+\beta}\right) = o\left(\tau_n^2\right) ;$$

$$J_2 \le \frac{1}{\sqrt{2\pi}\sigma_n}\int\limits_{-\tau_n^{2+\beta}}^{\tau_n^{2+\beta}} e^{-y^2/(2\sigma_n^2)}\left|\mathsf{E}_{n,0}\left(\Psi_n^*\left(\Lambda_n+y\right) - \Psi_n^*\left(\Lambda_n\right)\right)\right|\,dy\,.$$

Из определения (4) критической функции Ψ_n^* следует, что

$$J_2 \le \mathsf{P}_{n,0}\left(\Lambda_n = c_n\right) + \frac{1}{\sqrt{2\pi}\sigma_n} \int\limits_{-\tau_n^{2+\beta}}^{\tau_n^{2+\beta}} e^{-y^2/(2\sigma_n^2)} \mathsf{P}_{n,0}(\Lambda_n = c_n - y)\, dy +$$

$$+ \frac{1}{\sqrt{2\pi}\sigma_n} \int\limits_{-\tau_n^{2+\beta}}^{\tau_n^{2+\beta}} e^{-y^2/(2\sigma_n^2)} \left|\mathsf{P}_{n,0}\left(\Lambda_n > c_n - y\right) - \mathsf{P}_{n,0}\left(\Lambda_n > c_n\right)\right| dy\,.$$

Поскольку $0 < c \le |c_n| \le C$, то с учетом леммы 2.1 имеем

$$\mathsf{P}_{n,0}\left(\Lambda_n = c_n\right) = \mathsf{E}_{n,1} e^{-\Lambda_n} \mathbb{1}_{\{c_n\}}\left(\Lambda_n\right) = e^{-c_n} \mathsf{P}_{n,1}\left(\Lambda_n = c_n\right) \le e^C \sup_{x \le C} \mathsf{P}_{n,1}\left(x \le \Lambda_n \le x + \tau_n^{2+\beta}\right) = o\left(\tau_n^2\right)\,.$$

Аналогично $\mathsf{P}_{n,0}\left(\Lambda_n = c_n - y\right) = o\left(\tau_n^2\right)$ при всех $|y| \le \tau_n^{2+\beta}$. Таким образом, имеем оценку

$$J_2 \le \frac{1}{\sqrt{2\pi}\sigma_n} \int\limits_{0}^{\tau_n^{2+\beta}} e^{-y^2/(2\sigma_n^2)} \mathsf{P}_{n,0}\left(c_n - y \le \Lambda_n \le c_n\right)\, dy +$$

$$+ \frac{1}{\sqrt{2\pi}\sigma_n} \int\limits_{-\tau_n^{2+\beta}}^{0} e^{-y^2/(2\sigma_n^2)} \mathsf{P}_{n,0}\left(c_n \le \Lambda_n \le c_n - y\right)\, dy + o\left(\tau_n^2\right)\,.$$

Но в силу леммы 2.1

$$\mathsf{P}_{n,0}\left(c_n - y \le \Lambda_n \le c_n\right) = \mathsf{E}_{n,1} e^{-\Lambda_n} \mathbb{1}_{[c_n - y, c_n]}\left(\Lambda_n\right) \le$$

$$\le e^{y - c_n} \mathsf{P}_{n,1}\left(c_n - y \le \Lambda_n \le c_n\right) \le e^{\tau_n^{2+\beta} + C} o\left(\tau_n^2\right) = o\left(\tau_n^2\right)\,.$$

Аналогично получаем $\mathsf{P}_{n,0}\left(c_n \le \Lambda_n \le c_n - y\right) = o\left(\tau_n^2\right)$. Отсюда следует утверждение леммы. □

Из лемм 2.6 и 2.7 и ограниченности $\bar{d}_n$ (см. лемму 3.5) разность $\beta_n^* - \beta_n$ может быть представлена в виде:

$$\beta_n^* - \beta_n = \tilde{\beta}_n^* - \tilde{\beta}_n + o\left(\tau_n^2\right) = \mathsf{E}_{n,0} e^{\tilde{\Lambda}_n} \left(\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right) - \Psi_n(S_n)\right) + o\left(\tau_n^2\right) =$$

$$= \mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{(c_n,\infty)}\left(\tilde{\Lambda}_n\right) - \Psi_n(S_n)\right) + o\left(\tau_n^2\right) \equiv \tilde{A}_n + \tilde{B}_n + o\left(\tau_n^2\right)\,,$$

где

$$\tilde{A}_n = \mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{\left(-\infty,\bar{d}_n\right)}\left(\tilde{\Lambda}_n\right) - \mathbb{1}_{(-\infty,c_n)}\left(\tilde{\Lambda}_n\right)\right)\,; \tag{15}$$

$$\tilde{B}_n = \mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(1 - \Psi_n(S_n) - \mathbb{1}_{\left(-\infty,\bar{d}_n\right)}\left(\tilde{\Lambda}_n\right)\right)\,. \tag{16}$$

Обозначим

$$D_n = c_n - \bar{d}_n\,. \tag{17}$$

В леммах 3.5–3.7 следующего раздела показано, что

$$D_n = -\tau_n \mathsf{E}\left[\Delta | \Lambda = b\right] + o\left(\tau_n\right)\,; \tag{18}$$

$$\tilde{A}_n = -\frac{1}{2} D_n^2 e^b p(b) + o\left(\tau_n^2\right)\,; \tag{19}$$

$$\tilde{B}_n = \frac{1}{2} \tau_n^2 e^b \mathsf{E}\left[\Delta^2 | \Lambda = b\right] p(b) + o\left(\tau_n^2\right)\,, \tag{20}$$

где $b = \Phi_1^{-1}(1-\alpha)$ (см. лемму 2.1).

Из (18)–(20) следует основная

Теорема 2.8. *В случае распределения Лапласа справедлива формула*

$$\lim_{n\to\infty} \tau_n^{-2}\left(\beta_n^* - \beta_n\right) = \frac{1}{2} e^b \mathsf{D}\left[\Delta | \Lambda = b\right] p(b)\,.$$

3 Схема доказательства основного результата

В этом разделе приведены формулировки вспомогательных лемм для случая распределения Лапласа, составляющие схему доказательства теоремы 2.8. Для удобства изложения доказательство вспомогательных лемм вынесено в следующий раздел.

Обозначим

$$\bar{Q}_{n,l}(x) = \int\limits_{|z|\le\eta_n\tau_n^{-1}} z^l \left[\mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n < z \middle| \tilde{\Lambda}_n = x - \tau_n z\right) - \mathbb{1}_{(0,\infty)}(z)\right] p_n\left(x-\tau_n z\right)\, dz\,;$$

$$\tilde{Q}_{n,l}(x) = \int\limits_{-\infty}^{\infty} z^l \left[\mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n < z \middle| \tilde{\Lambda}_n = x - \tau_n z\right) - \mathbb{1}_{(0,\infty)}(z)\right] p_n(x-\tau_n z)\, dz\,,$$

где $l = 0, 1$, а $p_n(x)$ — плотность случайной величины $\tilde{\Lambda}_n = \Lambda_n + \xi_n$ (см. (12)).

Лемма 3.1. *Справедливо соотношение* $\sup\limits_x \left|\bar{Q}_{n,l}(x) - \tilde{Q}_{n,l}(x)\right| \to 0$, $l = 0, 1$.

Введем обозначения для условных мер на R^1, зависящих от параметров $u \in \mathsf{R}^1$ и $t \in (0, C]$, $C > 0$. Положим для $B \in \mathcal{B}^1$

$$\tilde{Q}^*_{n,u,t}(B) = \mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n(t) \in B \middle| \tilde{\Lambda}_n(t) = u\right) p_n(u,t)\,;$$
$$Q^*_{u,t}(B) = \mathsf{P}\left(\Delta(t) \in B \middle| \Lambda(t) = u\right) p(u,t)\,;$$

их характеристические функции будем обозначать:

$$\tilde{q}^*_{n,u,t}(s) \equiv \int e^{isx}\tilde{Q}^*_{n,u,t}(dx) = \mathsf{E}_{n,0}\left[e^{is\tau_n^{-1}\Delta_n(t)} | \tilde{\Lambda}_n(t) = u\right]\, p_n(u,t)\,;$$
$$q^*_{u,t}(s) \equiv \int e^{isx}\tilde{Q}^*_{u,t}(dx) = \mathsf{E}\left[e^{is\Delta(t)} | \Lambda(t) = u\right]\, p(u,t)\,.$$

Здесь зависимость случайных величин и плотностей от параметра t указана в явном виде, в то время как в остальной части работы такая зависимость опускается.

Лемма 3.2. *Для любого* $s \in \mathsf{R}^1$ $\sup\limits_{0<t\le C}\sup\limits_u \left|\tilde{q}^*_{n,u,t}(s) - q^*_{u,t}(s)\right| \longrightarrow 0$ *при* $n \to \infty$.

Введем обозначения для функций распределения, зависящих от параметров u и t:

$$\tilde{F}^*_{n,u,t}(z) = \mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n(t) < z \middle| \tilde{\Lambda}_n(t) = u\right) p_n(u,t)\,;$$
$$F^*_{u,t}(z) = \mathsf{P}\left(\Delta(t) < z \middle| \Lambda(t) = u\right) p(u,t)\,.$$

Лемма 3.3. *Для любой последовательности* $\varepsilon_n \to 0$

$$\sup_{0<t\le C}\sup_u L\left(\tilde{F}^*_{n,u+\varepsilon_n,t},\, F^*_{u,t}\right) \longrightarrow 0 \text{ при } n \to \infty\,,$$

где $L(F_1,\, F_2)$ *— расстояние Леви между функциями распределения* F_1 *и* F_2 *на* R^1.

Будем обозначать для $l = 0, 1$

$$Q_l(x) \equiv p(x)\int z^l\left[\mathsf{P}\left(\Delta < z \middle| \Lambda = x\right) - \mathbb{1}_{(0,\infty)}(z)\right]\, dz = -\frac{1}{l+1}\mathsf{E}\left(\Delta^{l+1}\middle|\Lambda = x\right) p(x)\,.$$

Лемма 3.4. *Справедливо соотношение* $\sup\limits_x \left|\tilde{Q}_{n,l}(x) - Q_l(x)\right| \to 0$ *для* $l = 0, 1$.

Лемма 3.5 *Для величины* D_n (*см.* (17)) *справедливо представление* $D_n = -\tau_n\mathsf{E}\left[\Delta|\Lambda = b\right] + o\left(\tau_n\right)$ *и* $\bar{d}_n \to b$.

Лемма 3.6 *Для величины* $\tilde{A}_n$ (*см.* (15)) *справедливо следующее представление:*

$$\tilde{A}_n = -\frac{1}{2}\, D_n^2 e^{\bar{d}_n} p(\bar{d}_n) + o\left(\tau_n^2\right) .$$

Лемма 3.7 *Для величины* $\tilde{B}_n$ (*см.* (16)) *справедливо следующее представление:*

$$\tilde{B}_n = \frac{1}{2}\,\tau_n^2 e^{\bar{d}_n} \mathsf{E}\left[\Delta^2 | \Lambda = \bar{d}_n\right] p(\bar{d}_n) + o\left(\tau_n^2\right) . \qquad \square$$

4 Доказательство вспомогательных лемм

Доказательство леммы 3.1. Справедливо равенство

$$\left|\bar{Q}_{n,l}(x) - \tilde{Q}_{n,l}(x)\right| = \int\limits_{|z|\geq \eta_n/\tau_n} z^l \left[\mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n < z \middle| \tilde{\Lambda}_n = x - \tau_n z\right) - \mathbb{1}_{(0,\infty)}(z)\right] p_n(x-\tau_n z)\, dz \equiv I_{n,l}^+ + I_{n,l}^- ,$$

где

$$I_{n,l}^+ = \int\limits_{\eta_n/\tau_n}^{\infty} z^l \mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n \geq z \middle| \tilde{\Lambda}_n = x - \tau_n z\right) p_n(x-\tau_n z)\, dz ; \tag{21}$$

$$I_{n,l}^- = \int\limits_{-\infty}^{-\eta_n/\tau_n} |z|^l \mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n < z \middle| \tilde{\Lambda}_n = x - \tau_n z\right) p_n(x-\tau_n z)\, dz =$$

$$= \int\limits_{\eta_n/\tau_n}^{\infty} u^l \mathsf{P}_{n,0}\left(-\tau_n^{-1}\Delta_n > u \middle| \tilde{\Lambda}_n = x + \tau_n u\right) p_n\left(x+\tau_n u\right)\, du . \tag{22}$$

Последнее равенство получено из предшествующего путем замены $z = u$.

Оценим $I_{n,l}^+$ и $I_{n,l}^-$ с помощью неравенства Чебышёва вида:

$$\mathsf{P}(X > x | Y) \leq \frac{\mathsf{E}\left[|X|^{l+1}\mathbb{1}_{(x,\infty)}(|X|)|Y\right]}{x^{l+1}} , \;\; x > 0 .$$

Получим

$$I_{n,l}^+ \leq \int\limits_{\eta_n/\tau_n}^{\infty} z^{-1}\mathsf{E}_{n,0}\left[\left|\tau_n^{-1}\Delta_n\right|^{l+1} \mathbb{1}_{[z,\infty)}\left(\left|\tau_n^{-1}\Delta_n\right|\right) \middle| \tilde{\Lambda}_n = x - \tau_n z\right] p_n\left(x-\tau_n z\right)\, dz \leq$$

$$\leq \tau_n \eta_n^{-1} \int\limits_{\eta_n/\tau_n}^{\infty} \mathsf{E}_{n,0}\left[\left|\tau_n^{-1}\Delta_n\right|^{l+1} \mathbb{1}_{(\eta_n,\infty)}\left(|\Delta_n|\right) \middle| \tilde{\Lambda}_n = x - \tau_n z\right] p_n(x-\tau_n z)\, dz \leq$$

$$\leq \tau_n \eta_n^{-1} \int\limits_{-\infty}^{\infty} \mathsf{E}_{n,0}\left[\left|\tau_n^{-1}\Delta_n\right|^{l+1} \mathbb{1}_{(\eta_n,\infty)}\left(|\Delta_n|\right) \middle| \tilde{\Lambda}_n = x - \tau_n z\right] p_n\left(x-\tau_n z\right)\, dz =$$

(с заменой $v = x - \tau_n z$)

$$= \eta_n^{-1} \int\limits_{-\infty}^{\infty} \mathsf{E}_{n,0}\left[|\tau_n^{-1}\Delta_n|^{l+1}\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|) \middle| \tilde{\Lambda}_n = v\right] p_n(v)\, dv = \eta_n^{-1}\mathsf{E}_{n,0}|\tau_n^{-1}\Delta_n|^{l+1}\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|) .$$

Аналогично

$$I^-_{n,l} \le \int\limits_{\eta_n/\tau_n}^{\infty} u^{-1}\mathsf{E}_{n,0}\left[|\tau_n^{-1}\Delta_n|^{l+1}\mathbb{1}_{(u,\infty)}(|\tau_n^{-1}\Delta_n|)\middle|\tilde\Lambda_n = x+\tau_n u\right] p_n(x+\tau_n u)\,du \le$$

$$\le \tau_n\eta_n^{-1}\int\limits_{\eta_n/\tau_n}^{\infty}\mathsf{E}_{n,0}\left[|\tau_n^{-1}\Delta_n|^{l+1}\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|)\middle|\tilde\Lambda_n = x+\tau_n u\right] p_n(x+\tau_n u)\,du \le$$

$$\le \tau_n\eta_n^{-1}\int\limits_{-\infty}^{\infty}\mathsf{E}_{n,0}\left[|\tau_n^{-1}\Delta_n|^{l+1}\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|)\middle|\tilde\Lambda_n = x+\tau_n u\right] p_n(x+\tau_n u)\,du \le$$

(с заменой $\nu = x+\tau_n u$)

$$= \eta_n^{-1}\int\limits_{-\infty}^{\infty}\mathsf{E}_{n,0}\left[|\tau_n^{-1}\Delta_n|^{l+1}\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|)\middle|\tilde\Lambda_n = \nu\right] p_n(\nu)\,d\nu = \eta_n^{-1}\mathsf{E}_{n,0}|\tau_n^{-1}\Delta_n|^{l+1}\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|)\,.$$

Теперь для доказательства леммы достаточно показать, что правая часть последнего неравенства стремится к нулю. Используя (i) из леммы 2.3, получаем при $l=0$

$$\eta_n^{-1}\mathsf{E}_{n,0}|\tau_n^{-1}\Delta_n|\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|) = \eta_n^{-1}\tau_n^{-1}\mathsf{E}_{n,0}\frac{\Delta_n^2}{|\Delta_n|}\,\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|) \le$$

$$\le \eta_n^{-2}\tau_n^{-1}\mathsf{E}_{n,0}\Delta_n^2\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|) = o\left(n^{-1/8}\right)\to 0\,,$$

а при $l=1$ аналогично

$$\eta_n^{-1}\mathsf{E}_{n,0}(\tau_n^{-1}\Delta_n)^2\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|) = \eta_n^{-1}\tau_n^{-2}\mathsf{E}_{n,0}\Delta_n^2\mathbb{1}_{(\eta_n,\infty)}(|\Delta_n|) = o(1)\to 0\,. \qquad \square$$

Доказательство леммы 3.2. Обозначим для любого $t\in(0,C]$, $C>0$, характеристические функции случайных векторов $(\tau_n^{-1}\Delta_n(t),\Lambda_n(t))$ и $(\Delta(t),\Lambda(t))$

$$\tilde q_{n,t}(s,y) \equiv \int e^{iyu}\tilde q^*_{n,u,t}(s)\,du = \mathsf{E}_{n,0}e^{is\tau_n^{-1}\Delta_n(t)+iy\tilde\Lambda_n(t)} = \mathsf{E}e^{iy\xi_n}\mathsf{E}_{n,0}e^{is\tau_n^{-1}\Delta_n(t)+iy\Lambda_n(t)} \equiv \omega_n(y)q_{n,t}(s,y)\,;$$

$$q_t(s,y) \equiv \int e^{iyu}q^*_{u,t}(s)\,du = \mathsf{E}e^{is\Delta(t)+iy\Lambda(t)}\,.$$

Из (11) для $q^*_{u,t}(s)$ справедлива формула обращения

$$\tilde q^*_{u,t}(s) = \frac{1}{2\pi}\int e^{-iuy}q_t(s,y)\,dy\,.$$

Тогда для каждого $s\in\mathsf{R}^1$ рассмотрим оценку

$$\left|\tilde q^*_{n,u,t}(s) - q^*_{u,t}(s)\right| \le \int|\tilde q_{n,t}(s,y)-q_t(s,y)|\,dy = \int\limits_{|y|<\delta\sqrt{n}} + \int\limits_{\delta\sqrt{n}\le|y|<n} + \int\limits_{|y|>n}\,,$$

где $\delta>0$ выбирается для каждого $s\in\mathsf{R}^1$ так, чтобы в рамках центральной предельной теоремы (см. (11)) для каждого $s\in\mathsf{R}^1$ и из свойств случайной величины ξ_n (см. (13)) по схеме доказательства (6.15) из работы [2] с использованием теоремы о мажорируемой сходимости выполнялось

$$\int\limits_{|y|<\delta\sqrt{n}}|\tilde q_{n,t}(s,y)-q_t(s,y)|\,dy\to 0\,.$$

Применяя рассуждения для (6.16) из работы [2] к случаю распределения Лапласа (см. также пример 1.3 из [5]), получаем для каждого $s \in \mathsf{R}^1$

$$\int\limits_{\delta\sqrt{n}\le|y|<n} |q_{n,t}(s,y)|\, dy \longrightarrow 0\,.$$

Сходимость к нулю остальных интегралов следует из свойств случайной величины ξ_n (см. (13)) и случайного вектора (Δ, Λ) (см. (11)), что завершает доказательство леммы. □

Доказательство леммы 3.3. Из леммы 6.1 работы [4], леммы 3.2 и ограниченности ε_n следует, что

$$L(\tilde{Q}^*_{n,u+\varepsilon_n,t}, Q^*_{u,t}) \to 0$$

равномерно по u и t, поскольку семейство мер $\{Q^*_{u,t}\}$ плотно (см. лемму 6.2 из [2]) и ограничено по u и t (см. лемму 2.1). Тогда из свойств меры $Q^*_{u,t}$ и эквивалентного определения для слабой сходимости мер (см., например, теорему 2.1 из работы [10]) вытекает утверждение леммы. □

Доказательство леммы 3.4. Разобьем пределы интегрирования на три части:

$$\left|\tilde{Q}_{n,l}(x) \;-\; Q_l(x)\right| = \left|\int\limits_{-\infty}^{-a_n} + \int\limits_{-a_n}^{a_n} + \int\limits_{a_n}^{\infty}\right|,$$

$a_n \to +\infty$ будет выбрано далее. Поскольку $Q_l(x) \to 0$ на $(-\infty, -a_n)$, (a_n, ∞), то рассмотрим на этих интервалах $\tilde{Q}_{n,l}(x)$. Объединяя (21) и (22) и применяя неравенство Чебышёва, как в доказательстве леммы 3.1, получаем

$$\left|\int\limits_{a_n}^{\infty} z^l \mathsf{P}_{n,0}\left(\mp\tau_n^{-1}\Delta_n \ge z \middle| \tilde{\Lambda}_n = x \pm \tau_n z\right) p_n(x \pm \tau_n z)\, dz\right| \le$$

$$\le \left|\int\limits_{a_n}^{\infty} z^{-1} \mathsf{E}_{n,0}\left[|\tau_n^{-1}\Delta_n|^{l+1} \mathbb{1}_{[z,\infty)}(|\tau_n^{-1}\Delta_n|) \middle| \tilde{\Lambda}_n = x \pm \tau_n z\right] p_n(x \pm \tau_n z)\, dz\right| \le$$

(с заменой $x \pm \tau_n z = v$)

$$\le \tau_n^{-1} a_n^{-1} \left|\int\limits_{-\infty}^{\infty} \mathsf{E}_{n,0}\left[|\tau_n^{-1}\Delta_n|^{l+1} \mathbb{1}_{[a_n,\infty)}(|\tau_n^{-1}\Delta_n|) \middle| \tilde{\Lambda}_n = v\right] p_n(v) dv\right| =$$

$$= \tau_n^{-(l+2)} a_n^{-1} \mathsf{E}_{n,0}|\Delta_n|^{l+1} \mathbb{1}_{[a_n,\infty)}(|\tau_n^{-1}\Delta_n|) = \mathcal{O}(e^{-a_n}) \to 0\,,$$

где последовательность $a_n \to +\infty$ может быть выбрана любая. Последняя оценка получена из неравенства Гёльдера и леммы 2.2. Тогда имеем

$$\left|\tilde{Q}_{n,l}(x) - Q_l(x)\right| \le \left|\int\limits_{|z|\le a_n} z^l \left(\mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n < z \middle| \tilde{\Lambda}_n = x - \tau_n z\right) p_n(x - \tau_n z) - \mathsf{P}\left(\Delta < z \middle| \Lambda = x\right) p(x)\right) dz\right| +$$

$$+ \left|\int\limits_0^{a_n} z^l \left(p_n(x - \tau_n z) \;-\; p(x)\right) dz\right| + \mathcal{O}\left(e^{-a_n}\right).$$

Используя леммы 3.2 и 3.3 из [8] (см. также [5]) и технику доказательства локальной предельной теоремы, несложно показать, что $\kappa_n \equiv \sup\limits_{0<t\le C} \sup\limits_x |p_n(x) - p(x)| \to 0$. Из леммы 3.3 следует, что для $|z| \le a_n = o(\tau_n^{-1})$

$$\lambda_n \equiv \sup_{0<t\le C} \sup_x L\left(\tilde{F}^*_{n,x-\tau_n z,t},\, F^*_{x,t}\right) \to 0\,.$$

Теперь, если выбрать, например, $a_n = \min(\lambda_n^{-1/3}, \kappa_n^{-1/3}, n^{1/8})$, получим утверждение леммы. □

Доказательство леммы 3.5. Доказательство леммы проведено в работе [8] (см. там лемму 3.5).

Доказательство леммы 3.6. Доказательство леммы без изменений переносится из работы [1] (см. там лемму 3.4.3).

Доказательство леммы 3.7. Представим $\tilde{B}_n$ как

$$\tilde{B}_n = \mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{(-\infty,\bar{d}_n)}(\tilde{S}_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{\Lambda}_n\right)\right) + \rho_{n1} + \rho_{n2}\,,$$

где

$$\rho_{n1} \equiv \mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{[\bar{d}_n,\infty)}(S_n) - \Psi_n(S_n)\right) = (1-\gamma)\mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\mathbb{1}_{\{\bar{d}_n\}}(S_n)\,;$$
$$\rho_{n2} \equiv \mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{(-\infty,\bar{d}_n)}(S_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{S}_n\right)\right).$$

Имеем

$$|\rho_{n1}| \le \mathsf{E}_{n,0}\left|e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right|\mathbb{1}_{\{\bar{d}_n\}}(S_n) \le \mathsf{E}_{n,0}\left|e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right|\mathbb{1}_{\{\bar{d}_n\}}(S_n)\,\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|) +$$
$$+ \mathsf{E}_{n,0}\left|e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right|\mathbb{1}_{[n^{-1/8},\infty)}(|\Delta_n|) \le$$

(с учетом леммы 2.5)

$$\le e^{\bar{d}_n}\mathsf{E}\left|e^{\xi_n} - 1\right|\mathsf{E}_{n,0}e^{-\Delta_n}\mathbb{1}_{\{\bar{d}_n\}}(S_n)\,\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|) + e^{\bar{d}_n}\mathsf{E}_{n,0}\left|e^{-\Delta_n} - 1\right|\mathbb{1}_{\{\bar{d}_n\}}(S_n)\,\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|) +$$
$$+ e^{\sigma_n^2/2}\mathsf{E}_{n,0}e^{\Lambda_n}\mathbb{1}_{[n^{-1/8},\infty)}(|\Delta_n|) + e^{\bar{d}_n}\mathsf{E}_{n,0}\mathbb{1}_{[n^{-1/8},\infty)}(|\Delta_n|) \le$$

(с учетом неравенства Гёльдера и $|e^y - 1| < |y|e^{|y|}$ для любого ограниченного y)

$$\le e^{\bar{d}_n + n^{-1/8}}\mathsf{E}\left|e^{\xi_n} - 1\right| + e^{\bar{d}_n + n^{-1/8}}n^{-1/8}\mathsf{P}_{n,0}\left(S_n = \bar{d}_n\right) +$$
$$+ e^{\sigma_n^2/2}\left(\mathsf{E}_{n,0}e^{2\Lambda_n}\right)^{1/2}\left(\mathsf{P}_{n,0}\left(|\Delta_n| \ge n^{-1/8}\right)\right)^{1/2} + e^{\bar{d}_n}\mathsf{P}_{n,0}\left(|\Delta_n| \ge n^{-1/8}\right).$$

С учетом того, что $\mathsf{P}_{n,0}(S_n = \bar{d}_n) = \mathcal{O}(\tau_n^2)$, в связи с леммой 2.2 $\mathsf{P}_{n,0}\left(|\Delta_n| \ge n^{-1/8}\right) \le Ce^{-n^{-1/8}n^{1/4}} = Ce^{-n^{1/8}}$, из доказательства леммы 2.6 $\mathsf{E}\left|e^{\xi_n} - 1\right| = o\left(\tau_n^2\right)$ и (2.10) работы [7] (см. там лемму 2.1) для $s = -ix$

$$\mathsf{E}_{n,1}e^{x\Lambda_n} \longrightarrow e^{-t^2(-x^2-x)/2}$$

для любого фиксированного x, имеем $\rho_{n1} = o\left(\tau_n^2\right)$. Аналогичные рассуждения применим для оценки $|\rho_{n2}|$:

$$|\rho_{n2}| \le \left|\mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{(-\infty,\bar{d}_n)}(S_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{S}_n\right)\right)\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|)\right| +$$
$$+ \left|\mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{(-\infty,\bar{d}_n)}(S_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{S}_n\right)\right)\mathbb{1}_{[n^{-1/8},\infty)}(|\Delta_n|)\right| \le$$
$$\le \mathsf{E}_{n,0}\left|e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right|\left|\mathbb{1}_{(-\infty,\bar{d}_n)}(S_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{S}_n\right)\right|\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|)\,\mathbb{1}_{[0,\tau_n^{2+\beta})}(|\xi_n|) +$$
$$+ \mathsf{E}_{n,0}\left|e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right|\left|\mathbb{1}_{(-\infty,\bar{d}_n)}(S_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{S}_n\right)\right|\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|)\,\mathbb{1}_{[\tau_n^{2+\beta},\infty)}(|\xi_n|) + \mathcal{O}(e^{-n^{1/8}}) \le$$

(с учетом (14))

$$\le \mathsf{E}\left|e^{\xi_n} - 1\right|\mathsf{E}_{n,0}e^{\Lambda_n}\left|\mathbb{1}_{(-\infty,\bar{d}_n)}(S_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{S}_n\right)\right|\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|)\,\mathbb{1}_{[0,\tau_n^{2+\beta})}(|\xi_n|) +$$
$$+ e^{\bar{d}_n}\mathsf{E}_{n,0}\left|e^{\Lambda_n - \bar{d}_n} - 1\right|\left|\mathbb{1}_{(-\infty,\bar{d}_n)}(S_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{S}_n\right)\right|\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|)\,\mathbb{1}_{[0,\tau_n^{2+\beta})}(|\xi_n|) + o\left(\tau_n^2\right) \le$$
$$\le e^{\bar{d}_n}\mathsf{E}_{n,0}\left|e^{\Lambda_n - \bar{d}_n} - 1\right|\mathbb{1}_{[\bar{d}_n - \tau_n^{2+\beta},\bar{d}_n + \tau_n^{2+\beta}]}(S_n)\,\mathbb{1}_{[0,n^{-1/8})}(|\Delta_n|) + o\left(\tau_n^2\right) \le$$

(поскольку $\left|\Lambda_n - \bar{d}_n\right| \le \left|S_n - \bar{d}_n\right| + |\Delta_n| \le \tau_n^{2+\beta} + n^{-1/8}$)

$$\le Cn^{-1/8}\mathsf{P}_{n,0}\left(\bar{d}_n - \tau_n^{2+\beta} \le S_n \le \bar{d}_n + \tau_n^{2+\beta}\right) + o\left(\tau_n^2\right) = o\left(\tau_n^2\right) .$$

Теперь $\tilde{B}_n$ запишем как

$$\tilde{B}_n = \mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{S}_n\right) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{\Lambda}_n\right)\right)\mathbb{1}_{[0,\eta_n)}\left(|\Delta_n|\right) + \rho_{n3} + o\left(\tau_n^2\right) ,$$

где аналогично рассуждениям выше

$$\rho_{n3} \equiv \mathsf{E}_{n,0}\left(e^{\tilde{\Lambda}_n} - e^{\bar{d}_n}\right)\left(\mathbb{1}_{(-\infty,\bar{d}_n)}(\tilde{S}_n) - \mathbb{1}_{(-\infty,\bar{d}_n)}\left(\tilde{\Lambda}_n\right)\right)\mathbb{1}_{[\eta_n,\infty)}\left(|\Delta_n|\right) = o\left(\tau_n^2\right) .$$

Теперь

$$\tilde{B}_n = \tau_n e^{\bar{d}_n} \int\limits_{|z|\le\eta_n/\tau_n} \left(e^{-\tau_n z} - 1\right) \times \left[\mathsf{P}_{n,0}(\tau_n^{-1}\Delta_n < z \big| \tilde{\Lambda}_n = \bar{d}_n - \tau_n z) - \mathbb{1}_{(0,\infty)}(z)\right] \times p_n\left(\bar{d}_n - \tau_n z\right) dz + o\left(\tau_n^2\right) + \rho_{n4} ,$$

где

$$|\rho_{n4}| \le 2\tau_n e^{\bar{d}_n} \times \int\limits_{|z|\le\eta_n/\tau_n} \left|e^{-\tau_n z} - 1\right| \mathsf{P}_{n,0}\left(|\Delta_n| > \eta_n \big| \tilde{\Lambda}_n = \bar{d}_n - \tau_n z\right) p_n\left(\bar{d}_n - \tau_n z\right) dz \le 2\eta_n e^{\bar{d}_n + \eta_n}\mathsf{P}_{n,0}\left(|\Delta_n| > \eta_n\right) = o\left(\tau_n^2\right) .$$

Используя неравенство $|e^{-s} - 1 + s| \le (1/2)\gamma^2 e^{\gamma}$, $|s| \le \gamma$, где $s = \tau_n z$, можно представить $\tilde{B}_n$ как

$$\tilde{B}_n \equiv -\tau_n^2 e^{\bar{d}_n} \int\limits_{|z|\le\eta_n/\tau_n} z\left[\mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n < z \big| \tilde{\Lambda}_n = \bar{d}_n - \tau_n z\right) - \mathbb{1}_{(0,\infty)}(z)\right] p_n\left(\bar{d}_n - \tau_n z\right) dz + o\left(\tau_n^2\right) + \rho_{n5} ,$$

где

$$|\rho_{n5}| \le \frac{1}{2}\tau_n^2\eta_n e^{\bar{d}_n + \eta_n} \int\limits_{|z|\le\eta_n/\tau_n} |z| \left|\mathsf{P}_{n,0}\left(\tau_n^{-1}\Delta_n < z \big| \tilde{\Lambda}_n = \bar{d}_n - \tau_n z\right) - \mathbb{1}_{(0,\infty)}(z)\right| p_n\left(\bar{d}_n - \tau_n z\right) dz .$$

Последний интеграл равен $-\bar{Q}_{n,1}(\bar{d}_n)$, и из леммы 3.4 для $l = 1$ имеем $\rho_{n5} = o\left(\tau_n^2\right)$. Тогда $\tilde{B}_n = -\tau_n^2 e^{\bar{d}_n}\bar{Q}_{n,1}(\bar{d}_n) + o\left(\tau_n^2\right)$, и, поскольку

$$\bar{Q}_{n,1}(\bar{d}_n) \to Q_1(\bar{d}_n) \equiv -\frac{1}{2}\,\mathsf{E}\left[\Delta^2 \big| \Lambda = \bar{d}_n\right] p\left(\bar{d}_n\right) ,$$

отсюда и следует утверждение леммы. □

Литература

1. *Bening V. E.* Asymptotic theory of testing statistical hypotheses. — Utrecht: VSP, 2000. 277 p.
2. *Чибисов Д. М.* Вычисление дефекта асимптотически эффективных критериев // Теория вероятностей и ее применения, 1985. Т. 30. Вып. 2. С. 289–310.
3. *Kotz S., Kozubowski T. J., Podgorski K.* The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering, and finance. — Birkhauser, 2001. 349 p.
4. *Chibisov D. M.* An asymptotic expansion for distributions of C(α) test statistics // Lecture Notes in Statistics, 1980. Vol. 2. P. 63–96.
5. *Chibisov D. M., van Zwet W. R.* On the edgeworth expansion for the logarithm of the likelihood ratio. I // Теория вероятностей и ее применения, 1984. Т. 29. Вып. 3. С. 417–439.
6. *Королев Р. А., Бенинг В. Е.* Асимптотические разложения для мощностей критериев в случае распределения Лапласа // Вестник Тверского государственного университета. Сер. Прикладная математика, 2008. Вып. 3(10). № 26(86). С. 97–107.
7. *Королев Р. А., Тестова А. В., Бенинг В. Е.* О мощности асимптотически оптимального критерия в случае распределения Лапласа // Вестник Тверского государственного университета. Сер. Прикладная математика, 2008. Вып. 8. № 4(64). С. 5–23.
8. *Королев Р. А.* Формула для предела нормированной разности мощностей критериев в случае распределения Лапласа // Вестник Тверского государственного университета. Сер. Прикладная математика, 2010. В печати.
9. *Феллер В.* Введение в теорию вероятностей и ее приложения. Т. 1. — М.: Мир, 1984. 528 с.
10. *Billingsley P.* Convergence of probability measures. — Wiley, Canada, 1999. 278 p.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 75–82

УТОЧНЕНИЕ НЕРАВЕНСТВА КАЦА–БЕРРИ–ЭССЕЕНА*

М. Е. Григорьева[1], И. Г. Шевцова[2]

Аннотация: Уточнены верхние оценки абсолютной константы в неравенстве Каца–Берри–Эссеена для сумм независимых одинаково распределенных случайных величин с конечными абсолютными моментами порядка $2+\delta$, $0<\delta<1$. Предложена альтернатива неравенству Каца–Берри–Эссеена, имеющая более тонкую структуру, и построены верхние оценки входящих в уточненное неравенство констант.

Ключевые слова: центральная предельная теорема; неравенство Каца–Берри–Эссеена; дробь Ляпунова

1 Введение

При решении многих прикладных задач приходится учитывать эффекты, возникающие в результате суммарного воздействия большого числа случайных факторов, отдельный вклад каждого из которых в сумму пренебрежимо мал. Чаще всего в таких ситуациях статистические закономерности поведения суммы в силу центральной предельной теоремы аппроксимируются нормальным распределением вероятностей. При этом точность нормальной аппроксимации зависит от наличия у случайных слагаемых моментов достаточно высокого порядка или, другими словами, тяжестью (или легкостью) их «хвостов». Известно, что при некоторых достаточно общих условиях нормальная аппроксимация адекватна, если случайные слагаемые имеют моменты хотя бы второго порядка, причем чем выше порядок момента, тем, как правило, выше точность нормальной аппроксимации. При этом большой интерес представляет ситуация, когда случайные слагаемые имеют моменты, порядок которых заключен между двумя и тремя: с одной стороны, для распределений, имеющих моменты порядка, большего трех, скорость сходимости в центральной предельной теореме остается в общем случае такой же, как для распределений с третьими моментами; с другой стороны, во многих прикладных задачах важно оценивать точность нормальной аппроксимации, когда центральная предельная теорема все еще выполняется, но слагаемые имеют распределения со столь тяжелыми «хвостами», что третьего момента уже не существует. Такие задачи возникают, например, в страховании, когда речь заходит о маловероятных, но экстремально больших выплатах по тому или иному страховому случаю. Другие примеры связаны с практическим применением моделей типа распределения Парето с «хвостами», убывающими степенным образом, при анализе трафика в телекоммуникационных системах. Часто статистический анализ таких моделей позволяет сделать вывод, что показатель степени заключен между тремя и четырьмя, т. е. дисперсия существует, а третий момент отсутствует. Улучшению оценок точности нормальной аппроксимации именно для таких ситуаций и посвящена данная работа.

Для $0 \leqslant \delta \leqslant 1$ обозначим через $\mathcal{F}_{2+\delta}$ множество функций распределения с нулевым средним, единичной дисперсией и конечным абсолютным моментом $\beta_{2+\delta}$ порядка $2+\delta$. При $\delta = 0$ полагаем $\beta_2 = 1$ и $\mathcal{F}_2$ — класс всех распределений с нулевым средним и единичной дисперсией. Пусть $X_1, \ldots, X_n$ — независимые одинаково распределенные случайные величины с общей функцией распределения $F \in \mathcal{F}_{2+\delta}$, заданные на некотором вероятностном пространстве $(\Omega, \mathcal{A}, \mathsf{P})$. Обозначим

$$F_n(x) = F^{*n}(x\sqrt{n}) = \mathsf{P}\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}} < x\right);$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2}\, dt\,, \quad x \in \mathbb{R}\,.$$

Классическое неравенство Каца–Берри–Эссеена устанавливает существование конечной положительной постоянной $C_0 = C_0(\delta)$, зависящей только от δ, которая гарантирует справедливость неравенства

*Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты 08-01-00563, 08-01-00567, 08-07-00152 и 09-07-12032-офи-м), а также Министерства образования и науки (государственные контракты П1181 и П958, грант МК-581.2010.1).

[1]Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, maria-grigoryeva@yandex.ru

[2]Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, ishevtsova@cs.msu.su

$$\left.\begin{aligned} &\rho(F_n,\Phi)\equiv\sup_x|F_n(x)-\Phi(x)|\leqslant C_0(\delta)L_n^{2+\delta};\\ &L_n^{2+\delta}=\frac{\beta_{2+\delta}}{n^{\delta/2}}\end{aligned}\right\}\quad(1)$$

для всех $n\geqslant 1$ и $F\in\mathcal{F}_3$.

Для $\delta=1$ неравенство (1) было доказано независимо и одновременно Э. Берри [1] и К.-Г. Эссеном [2]. В 1960-е гг. разными авторами были предприняты успешные попытки обобщить результат Берри–Эссеена. Так, в 1963 г. М. Кац [3] доказал аналог (1) для независимых одинаково распределенных случайных величин с $\mathsf{E}X_1^2g(X_1)<\infty$ для функций g из некоторого класса, включающего $g(x)=|x|^\delta$. В 1965 г. В. В. Петров [4] обобщил неравенство Каца на разнораспределенные слагаемые. В 1966 г. А. Бикялис [5] доказал неравномерную оценку для разнораспределенных случайных величин, имеющих конечные абсолютные моменты порядка $2+\delta$, $0<\delta\leqslant 1$, из которой также вытекает неравенство (1). Точные формулировки упомянутых результатов вместе с их доказательствами можно найти, например, в монографии В. В. Петрова [6].

Относительно константы $C_0(1)$ известно, что

$$\frac{\sqrt{10}+3}{6\sqrt{2\pi}}\leqslant C_0(1)\leqslant 0{,}4784$$

(нижняя оценка найдена К.-Г. Эссееном [7], верхняя — В. Ю. Королевым и И. Г. Шевцовой [8]).

Верхние оценки величины $C_0(\delta)$ при некоторых $0<\delta<1$ впервые были получены в 1983 г. В. Тысиаком [9] (см. также [10]) и недавно были уточнены в работе [11]. В 1986 г. Г. Падитц [12] показал, что для всех $F\in\mathcal{F}_2$ и $n\geqslant 1$ имеет место неравенство

$$\rho(F_n,\Phi)\leqslant 3{,}51\mathsf{E}\left(X_1^2\min\left\{1,\frac{|X_1|}{\sqrt{n}}\right\}\right),$$

откуда вытекает равномерная по $\delta\in[0,1]$ оценка $C_0(\delta)\leqslant 3{,}51$, так как при любом $\delta\in[0,1]$ выражение в правой части последнего неравенства не превосходит $3{,}51\cdot L_n^{2+\delta}$.

Несмотря на то, что со времени первой публикации верхних оценок прошло более 25 лет, нижние оценки для величины $C_0(\delta)$ получены совсем недавно (см. [13]). Перечисленные оценки указаны в табл. 1: во втором столбце — верхние оценки Тысиака [9], в третьем — верхние оценки из работы [11], в пятом — нижние оценки из [13]; в четвертом же столбце указаны новые оценки, доказанные в данной работе. Для удобства в первой строке таблицы приведен год соответствующей публикации.

Из табл. 1 видно, что представленные в данной работе верхние оценки константы $C_0(\delta)$ не очень далеки от неулучшаемых: зазор между найденной

Таблица 1 История верхних оценок, а также нижние оценки константы $C_0(\delta)$

Год	1983	2009	2010	2010
δ	$C_0\leqslant$	$C_0\leqslant$	$C_0\leqslant$	$C_0\geqslant$
0,9	0,802	0,7671	0,5383	0,2133
0,8	0,812	0,7720	0,5723	0,2245
0,7	0,833	0,7876	0,6026	0,2376
0,6	0,863	0,8126	0,6276	0,2530
0,5	0,902	0,8454	0,6413	0,2715
0,4	0,950	0,8876	0,6342	0,2939
0,3	1,008	0,9407	0,6195	0,3220
0,2	1,076	1,0001	0,6094	0,3585
0,1	1,102	1,0739	0,6028	0,4092

Таблица 2 Двусторонние оценки константы $C_1(\delta)$

δ	$C_1\leqslant$	$C_1\geqslant$
0,9	0,3089	0,0323
0,8	0,3187	0,0356
0,7	0,3334	0,0396
0,6	0,3528	0,0444
0,5	0,3775	0,0503
0,4	0,4080	0,0575
0,3	0,4450	0,0665
0,2	0,4901	0,0780
0,1	0,5451	0,0939

мажорантой и соответствующей минорантой составляет всего 0,2–0,3, а их отношение колеблется в пределах 1,5–2,5 в зависимости от δ.

Наряду с уточнением константы $C_0(\delta)$ в (1) данная работа ставит своей целью уточнение и самой структуры неравенства (1). А именно, в качестве альтернативы предлагается рассмотреть неравенство

$$\rho(F_n,\Phi)\leqslant C_1(\delta)\frac{\beta_{2+\delta}+1}{n^{\delta/2}},\ \ n\geqslant 1,\ F\in\mathcal{F}_{2+\delta},\quad(2)$$

справедливость которого с некоторым положительным и конечным $C_1(\delta)$ вытекает тривиальным образом из (1) (например, с $C_1(\delta)=2C_0(\delta)$ в силу условия $\beta_{2+\delta}\geqslant 1$). Однако константа $C_1(\delta)$ в (2) оказывается гораздо меньше, чем $C_0(\delta)$ в (1), поэтому неравенство (2) при достаточно больших $\beta_{2+\delta}$ дает оценку, заведомо лучшую, чем (1), несмотря на то что для его справедливости необходима та же априорная информация о распределении F (а именно, только значение абсолютного момента $\beta_{2+\delta}$). Кроме того, более оптимистичными оказываются и нижние оценки $C_1(\delta)$, построенные в работе [13]. Например, для $\delta=1$ двусторонняя оценка имеет вид $0{,}2659\leqslant C_1(1)\leqslant 0{,}3041$ [13–16]. Для $0<\delta<1$ верхние оценки константы $C_k(\delta)$, устанавливаемые

в данной статье, и нижние, полученные в работе [13], приведены в табл. 2 во втором и третьем столбцах соответственно.

Рассуждения, приводящие к форме неравенства (2), основаны на используемых оценках для характеристических функций и подробно описаны в [8, 14].

2 Главный результат и основные идеи его доказательства

Теорема 1. *Для константы $C_k(\delta)$ в неравенстве*

$$\rho(F_n,\Phi)\leqslant C_k(\delta)\frac{\beta_{2+\delta}+k}{n^{\delta/2}}\,,\quad n\geqslant 1\,,\ F\in\mathcal{F}_{2+\delta}\,,$$

при $k=0$ и $k=1$ имеют место оценки, приведенные в табл. 3.

Таблица 3 Верхние оценки констант $C_0(\delta)$ и $C_1(\delta)$

δ	$C_0\leqslant$	$C_1\leqslant$
0,9	0,5383	0,3089
0,8	0,5723	0,3187
0,7	0,6026	0,3334
0,6	0,6276	0,3528
0,5	0,6413	0,3775
0,4	0,6342	0,4080
0,3	0,6195	0,4450
0,2	0,6094	0,4901
0,1	0,6028	0,5451

При доказательстве теоремы 1 будем придерживаться подхода, предложенного и развитого В. М. Золотарёвым в его работах [17–19]. Этот подход основан на применении неравенств сглаживания, которые позволяют оценить расстояние между функциями распределения через расстояние между соответствующими характеристическими функциями. В рамках этого подхода ключевыми моментами являются: (i) выбор надлежащего неравенства сглаживания; (ii) выбор в нем сглаживающего ядра; (iii) конструирование оценок для характеристических функций и ($i\nu$) выбор вычислительной оптимизационной процедуры. Опишем эти моменты в той последовательности, в которой они появляются при доказательстве неравенств (1) и (2). Соответствующие утверждения сформулируем в виде лемм.

Обозначим $f(t)$ и $f_n(t)$ характеристические функции случайных величин X_1 и стандартизованной суммы $(X_1+\ldots+X_n)/\sqrt{n}$ соответственно:

$$f(t)=\int\limits_{-\infty}^{\infty}e^{itx}\,dF(x)\,;$$

$$f_n(t)=\int\limits_{-\infty}^{\infty}e^{itx}\,dF_n(x)=\left(f\left(\frac{t}{\sqrt{n}}\right)\right)^n\,,\quad t\in\mathbb{R}\,.$$

Пусть

$$r_n(t)=\left|f_n(t)-e^{-t^2/2}\right|\,.$$

Лемма 1 (см. [20]). *Для произвольной функции распределения F при всех $n\geqslant 1$, $0<t_0\leqslant 1$ и $U>0$ имеет место неравенство*

$$\rho(F_n,\Phi)\leqslant 2\int\limits_0^{t_0}|K(t)|\,r_n(Ut)\,dt+$$
$$+2\int\limits_{t_0}^{1}|K(t)\,|\cdot|f_n(Ut)|\,dt+$$
$$+2\int\limits_0^{t_0}\left|K(t)-\frac{i}{2\pi t}\right|e^{-U^2t^2/2}\,dt+\frac{1}{\pi}\int\limits_{t_0}^{\infty}e^{-U^2t^2/2}\,\frac{dt}{t}\,,$$

где

$$K(t)=\frac{1}{2}\,(1-|t|)+\frac{i}{2}\left[(1-|t|)\cot\pi t+\frac{\operatorname{sign}t}{\pi}\right]\,,$$
$$-1\leqslant t\leqslant 1\,.$$

Перейдем теперь к оцениванию характеристических функций, фигурирующих в лемме 1. Пусть $\theta_0(\delta)$ — единственный корень уравнения

$$\frac{\delta\theta^2}{2}+\theta\sin\theta+(2+\delta)(\cos\theta-1)=0\,,\quad \pi\leqslant\theta\leqslant 2\pi\,;$$

$$\varkappa(\delta)\equiv\sup_{x>0}\frac{\left|\cos x-1+x^2/2\right|}{x^{2+\delta}}=$$
$$=\left.\frac{\cos x-1+x^2/2}{x^{2+\delta}}\right|_{x=\theta_0(\delta)}\,;$$

$$\gamma(\delta)=\sup_{x>0}\sqrt{\left(\frac{\cos x-1+x^2/2}{x^{2+\delta}}\right)^2+\left(\frac{\sin x-x}{x^{2+\delta}}\right)^2}\,,$$
$$0<\delta\leqslant 1\,.$$

Для $\varepsilon>0$ положим

$$\psi_\delta(t,\varepsilon)=\begin{cases}t^2/2-\varkappa(\delta)\varepsilon|t|^{2+\delta}\,, & |t|<\theta_0(\delta)\varepsilon^{-1/\delta}\,;\\ \dfrac{1-\cos\left(\varepsilon^{1/\delta}t\right)}{\varepsilon^{2/\delta}}\,, & \theta_0(\delta)\leqslant\varepsilon^{1/\delta}|t|\leqslant 2\pi\,;\\ 0\,, & |t|>2\pi\varepsilon^{-1/\delta}\,.\end{cases}$$

Несложно убедиться, что функция $\psi_\delta(t,\varepsilon)$ монотонно убывает по ε при каждом фиксированном $t\in\mathbb{R}$.

Ляпуновская дробь будет обозначаться $\ell = \beta_{2+\delta}n^{-\delta/2}$. Дополнительно обозначим

$$\ell_n = \ell + n^{-\delta/2}\,.$$

Лемма 2. *При всех $F\in\mathcal{F}_{2+\delta}$, $0<\delta\leqslant 1$, $n\geqslant 1$ и $t\in\mathbb{R}$ (если не оговорено иное) справедливы оценки*

$$|f_n(t)|\leqslant\left[1-\frac{2}{n}\psi_\delta(t,\ell_n)\right]^{n/2}\equiv f_1(t,\ell_n,n)\,;$$

$$|f_n(t)|\leqslant\exp\{-\psi_\delta(t,\ell_n)\}\equiv f_2(t,\ell_n)\,;$$

$$|f_n(t)|\leqslant\exp\left\{-\frac{t^2}{2}+\varkappa(\delta)\ell_n|t|^{2+\delta}\right\}\equiv f_3(t,\ell_n)\,;$$

$$r_n(t)\leqslant e^{-t^2/2}\left[\exp\left\{\gamma(\delta)\ell|t|^{2+\delta}-n\ln\left(1-\frac{t^2}{2n}\right)-\right.\right.$$
$$\left.\left.-\ \frac{t^2}{2}\right\}-1\right]\equiv r_1(t,\ell,n),\ \ |t|<\sqrt{2n};$$

$$r_n(t)\leqslant\left(\gamma(\delta)\ell|t|^{2+\delta}+\frac{|t|^4}{8n}\right)\times$$
$$\times\left(\max\left\{f_1(t,\ell_n,n),\,e^{-t^2/2}\right\}\right)^{(n-1)/n}\equiv r_2(t,\ell,n)\,;$$

$$r_n(t)\leqslant\frac{1}{2}\left(\gamma(\delta)\ell|t|^{2+\delta}+\frac{|t|^4}{8n}\right)\left(e^{-t^2/2}+\right.$$
$$\left.+\ \max\left\{f_1(t,\ell_n,n),\,e^{-t^2/2}\right\}\right)e^{t^2/(2n)}\equiv r_3(t,\ell,n)\,;$$

$$r_n(t)\leqslant f_1(t,\ell_n,n)+e^{-t^2/2}\equiv r_4(t,\ell,n)\,.$$

Замечание 1. Очевидно, $f_1(t,\varepsilon,n)\leqslant f_2(t,\varepsilon)$ при всех $n\geqslant 1$, $\varepsilon>0$ и $t\in\mathbb{R}$. Более того, из результата работы [21] вытекает, что $f_2(t,\varepsilon)\leqslant f_3(t,\varepsilon)$ для всех $\varepsilon>0$ и $t\in\mathbb{R}$, так что самую точную оценку для $|f_n(t)|$ дает $f_1(t,\ell_n,n)$, тогда как функции $f_j(t,\ell_n)$, $j=2,3$, обладают полезным свойством монотонности по ℓ_n, играющим важную роль в оптимизационной процедуре.

Доказательство. Первые три оценки (f_j, $j=1,2,3$) являются тривиальными следствиями оценок

$$|f(t)|^2\leqslant 1-2\psi_\delta(t,\beta_{2+\delta}+1)\,,\quad t\in\mathbb{R}\,;$$

$$|f(t)|^2\leqslant 1-t^2+2\varkappa(\delta)(\beta_{2+\delta}+1)|t|^{2+\delta}\,,\quad t\in\mathbb{R}\,,$$

полученных Шевцовой в [21]. Четвертая оценка (r_1) впервые объявлена В. М. Золотарёвым [17] для $\delta=1$ (без доказательства), ниже будет приведено полное доказательство для всех $0<\delta\leqslant 1$. Пятая (r_2) и шестая (r_3) оценки являются несложной комбинацией методов и результатов работ Правитца [22], Шевцовой [21], Гапоновой и Шевцовой [23]. Последняя оценка (r_4) тривиальна.

Докажем неравенство $r_n(t)\leqslant r_1(t,\ell,n)$, $|t|<\sqrt{2n}$. Из неравенств $|e^{ix}-1-ix|\leqslant|x|^2/2$ и $|e^{ix}-1-ix-(ix)^2/2|\leqslant\gamma(\delta)|x|^{2+\delta}$, $x\in\mathbb{R}$, с учетом моментных условий для распределения $F\in\mathcal{F}_{2+\delta}$ вытекают соотношения

$$|f(t)-1|\leqslant\frac{t^2}{2}\,,\quad |t|\leqslant\sqrt{2}\,,$$

$$f(t)=1-\frac{t^2}{2}+\theta_1\gamma(\delta)\beta_{2+\delta}|t|^{2+\delta}\,,\quad t\in\mathbb{R}\,,$$

с некоторым $\theta_1\in\mathbb{C}$, $|\theta_1|\leqslant 1$. Следовательно, для всех $|t|<\sqrt{2n}$ определен логарифм $\ln f(t)$ (условимся всегда выбирать главную ветвь логарифма) и

$$\left|\ln f(t)+\frac{t^2}{2}\right|=\left|\ln[1-(1-f(t))]+\frac{t^2}{2}\right|=$$
$$=\left|-\sum_{k=1}^{\infty}\frac{(1-f(t))^k}{k}+\frac{t^2}{2}\right|\leqslant$$
$$\leqslant\sum_{k=2}^{\infty}\frac{1}{k}\left(\frac{t^2}{2}\right)^k+\left|f(t)-1+\frac{t^2}{2}\right|\leqslant$$
$$\leqslant-\left[\ln\left(1-\frac{t^2}{2}\right)+\frac{t^2}{2}\right]+\gamma(\delta)\beta_{2+\delta}|t|^{2+\delta}\,,$$
$$|t|<\sqrt{2n}\,,$$

откуда с учетом неравенства $|e^z-1|\leqslant e^{|z|}-1$, $z\in\mathbb{C}$, получаем

$$r_n(t)=\left|f_n(t)-e^{-t^2/2}\right|=$$
$$=e^{-t^2/2}\left|\exp\left\{n\ln f\left(\frac{t}{\sqrt{n}}\right)+\frac{t^2}{2}\right\}-1\right|\leqslant$$
$$\leqslant e^{-t^2/2}\left(\exp\left\{n\left|\ln f\left(\frac{t}{\sqrt{n}}\right)+\frac{t^2}{2n}\right|\right\}-1\right)\leqslant$$
$$\leqslant e^{-t^2/2}\left(\exp\left\{\gamma(\delta)\frac{\beta_{2+\delta}|t|^{2+\delta}}{n^{\delta/2}}-n\ln\left(1-\frac{t^2}{2n}\right)-\right.\right.$$
$$\left.\left.-\ \frac{t^2}{2}\right\}-1\right)\equiv r_1(t,\ell,n)\,,$$

что и требовалось доказать.

Следующая лемма позволяет ограничить сверху множество рассматриваемых значений n при оценивании констант $C_k(\delta)$ в неравенстве (2) с $0\leqslant k\leqslant 1$.

Лемма 3. *Для любых положительных $k \leqslant 1$, T, ε,*

$$N_1 \geqslant N_1(T) \equiv \equiv T^2 \left(\frac{1}{8k\delta\gamma(\delta)} + \sqrt{1 + \left(\frac{1}{8k\delta\gamma(\delta)} \right)^2} \right)^2 ;$$

$$N_3 \geqslant N_3(T,\varepsilon) \equiv \left(\frac{T^{2-\delta}}{4\delta\gamma(\delta)} + \frac{\varepsilon T^2}{\delta} \right)^{2/(2-\delta)}$$

и таких, что $N_j \geqslant ((1+k)/\varepsilon)^{2/\delta}$, при всех $|t| \leqslant T$ справедливы оценки

$$\sup_{n \geqslant N_1} r_1\left(t, \varepsilon - kn^{-\delta/2}, n\right) \leqslant \leqslant e^{-t^2/2} \left(\exp\left\{ \gamma(\delta)\varepsilon|t|^{2+\delta} \right\} - 1 \right) \equiv \widetilde{r}_1(t,\varepsilon) ;$$

$$\sup_{n \geqslant N_3} r_3\left(t, \varepsilon - n^{-\delta/2}, n\right) \leqslant \leqslant \frac{\gamma(\delta)\varepsilon|t|^{2+\delta}}{2} \left(e^{\varkappa(\delta)\varepsilon|t|^{2+\delta}} + 1 \right) e^{-t^2/2} \equiv \widetilde{r}_3(t,\varepsilon) .$$

Доказательство. Для доказательства первой оценки запишем r_1 в виде

$$r_1\left(t, \varepsilon - kn^{-\delta/2}, n\right) = = e^{-t^2/2} \left(\exp\left\{ \gamma(\delta)\varepsilon|t|^{2+\delta} + g(n,|t|) \right\} - 1 \right) ,$$

где

$$g(n,|t|) = -\frac{k\gamma(\delta)}{n^{\delta/2}}|t|^{2+\delta} - n\ln\left(1 - \frac{t^2}{2n}\right) - \frac{t^2}{2} .$$

Тогда достаточно показать, что $g(x,t) \leqslant 0$ для всех $0 \leqslant t \leqslant T$ и $x \geqslant N_1(T,\varepsilon)$.

Используя разложение логарифма в степенной ряд, для всех $0 \leqslant t \leqslant \sqrt{2x}$ и $x > 0$ получаем

$$g(x,t) = -\frac{k\gamma(\delta)}{x^{\delta/2}}|t|^{2+\delta} + x\sum_{j=2}^{\infty} \frac{1}{j}\left(\frac{t^2}{2x}\right)^j \leqslant \leqslant -\frac{k\gamma(\delta)}{x^{\delta/2}}|t|^{2+\delta} + \frac{x}{2}\sum_{j=2}^{\infty}\left(\frac{t^2}{2x}\right)^j = = -\frac{k\gamma(\delta)}{x^{\delta/2}}|t|^{2+\delta} + \frac{t^4}{4(2x-t^2)} \equiv \widetilde{g}(x,t) .$$

Заметим, что для $0 \leqslant t \leqslant \sqrt{2x}$ и $x > 0$

$$\frac{\partial\widetilde{g}(x,t)}{\partial x} = \frac{k\delta\gamma(\delta)t^{2+\delta}}{2x^{1+\delta/2}} - \frac{t^4}{2(2x-t^2)^2} > 0 \iff \iff h(x,t) \equiv k\delta\gamma(\delta)(2x-t^2)^2 - t^{2-\delta}x^{1+\delta/2} > 0 .$$

Пусть T — произвольное число из интервала $(0, \sqrt{2x})$. Несложно видеть, что функция $h(x,t)$ монотонно убывает по t, поэтому для всех $0 \leqslant t \leqslant T$

$$h(x,t) \geqslant h(x,T) = x\left(4k\delta\gamma(\delta)x - 4k\delta\gamma(\delta)T^2 - - x^{\delta/2}T^{2-\delta}\right) + k\delta\gamma(\delta)T^4 > > x\left(4k\delta\gamma(\delta)x - 4k\delta\gamma(\delta)T^2 - x^{\delta/2}T^{2-\delta}\right) .$$

Для неотрицательности последнего выражения достаточно, чтобы

$$H(x) \equiv 4k\delta\gamma(\delta)(x - T^2) - x^{\delta/2}T^{2-\delta} \geqslant 0 .$$

Очевидно, для всех достаточно больших x функция $H(x)$ монотонно возрастает и не ограничена при $x \to \infty$. Следовательно, найдется такая точка $x_0 > 0$, что $H(x) \geqslant 0$ для всех $x \geqslant x_0$. Будем искать эту точку в виде $x_0 = zT^2$. Не ограничивая общности, можно считать, что $z > 1$, поскольку $H(T^2) = -T^2 < 0$. Имеем

$$H(zT^2) = 4k\delta\gamma(\delta)T^2(z-1) - T^2z^{\delta/2} > 0 \iff \iff 4k\delta\gamma(\delta)(z-1) - z^{\delta/2} > 0 .$$

Поскольку $z^{\delta/2} \leqslant \sqrt{z}$ при $z > 1$, для справедливости последнего условия достаточно, чтобы

$$4k\delta\gamma(\delta)z - \sqrt{z} - 4k\delta\gamma(\delta) > 0 ,$$

откуда, решив квадратное уравнение, получаем

$$\sqrt{z} > \frac{1 + \sqrt{1 + 64(\delta\gamma(\delta)k)^2}}{8\delta\gamma(\delta)k} \equiv z_0 .$$

Следовательно, $x_0 = z_0^2T^2$ и $H(x) > 0$ при

$$x \geqslant T^2 \left(\frac{1}{8\delta\gamma(\delta)k} + \sqrt{1 + \left(\frac{1}{8\delta\gamma(\delta)k} \right)^2} \right)^2 .$$

Таким образом, для всех $0 \leqslant t \leqslant T$ и $x \geqslant N_1(T,\varepsilon)$ имеем $h(x,t) > xH(x) > 0$, а значит $\widetilde{g}(x,t)$ монотонно возрастает по $x \geqslant N_1(T,\varepsilon)$ при каждом фиксированном $0 < t \leqslant T$ и для всех $N_1 \geqslant N_1(T,\varepsilon)$

$$\sup_{0\leqslant t\leqslant T}\sup_{x\geqslant N_1} g(x,t) \leqslant \sup_{0\leqslant t\leqslant T}\sup_{x\geqslant N_1} \widetilde{g}(x,t) = = \sup_{0\leqslant t\leqslant T}\lim_{x\to\infty} \widetilde{g}(x,t) = 0 ,$$

что и требовалось доказать.

Далее заметим, что в силу неравенства $f_1(t,\ell_n,n) \leqslant f_3(t,\ell_n)$, $t \in \mathbb{R}$, величину r_3 можно оценить следующим образом:

$$
\begin{aligned}
&r_3\left(t, \varepsilon - n^{-\delta/2}, n\right) \leqslant \\
&\leqslant \frac{1}{2}\left(\gamma(\delta)\varepsilon|t|^{2+\delta} - \frac{\gamma(\delta)|t|^{2+\delta}}{n^{\delta/2}} + \frac{t^4}{8n}\right)\left(e^{-t^2/2} + \right. \\
&\left. \exp\left\{-\frac{t^2}{2} + \varkappa(\delta)\varepsilon|t|^{2+\delta} - \frac{\varkappa(\delta)|t|^{2+\delta}}{n^{\delta/2}}\right\}\right) e^{t^2/(2n)} \leqslant \\
&\leqslant \frac{1}{2}\left(\gamma(\delta)\varepsilon|t|^{2+\delta} - \frac{\gamma(\delta)|t|^{2+\delta}}{n^{\delta/2}} + \frac{t^4}{8n}\right) \times \\
&\times \left(1 + e^{\varkappa(\delta)\varepsilon|t|^{2+\delta}}\right) e^{t^2/(2n) - t^2/2} = \\
&= \frac{|t|^{2+\delta}}{2}\left(\gamma(\delta)\varepsilon - \frac{\gamma(\delta)}{n^{\delta/2}} + \frac{|t|^{2-\delta}}{8n}\right) \times \\
&\times \left(1 + e^{\varkappa(\delta)\varepsilon|t|^{2+\delta}}\right) e^{t^2/(2n) - t^2/2} \equiv \\
&\equiv \frac{|t|^{2+\delta}}{2}\left(1 + e^{\varkappa(\delta)\varepsilon|t|^{2+\delta}}\right) \exp\left(-\frac{t^2}{2} + g(n,|t|)\right),
\end{aligned}
$$

где

$$
g(x,t) = \frac{t^2}{2x} + \ln\left(\gamma(\delta)\varepsilon - \frac{\gamma(\delta)}{x^{\delta/2}} + \frac{t^{2-\delta}}{8x}\right), \quad x > \left(\frac{2}{\varepsilon}\right)^{2/\delta}, \quad \varepsilon,\ t > 0\,.
$$

Заметим, что выражение под знаком логарифма положительно.

Покажем, что при всех фиксированных положительных ε и $t \leqslant T$ функция $g(x,t)$ монотонно возрастает по x при $x \geqslant N_3(T,\varepsilon)$. Вычислим производную

$$
\begin{aligned}
g'_x(x,t) &= -\frac{t^2}{2x^2} + \frac{(\delta/2)\gamma(\delta)x^{-1-\delta/2} - t^{2-\delta}/(8x^2)}{\gamma(\delta)\varepsilon - \gamma(\delta)x^{-\delta/2} + t^{2-\delta}/(8x)} = \\
&= \left(-8\gamma(\delta)\varepsilon x t^2 + 8\gamma(\delta)x^{1-\delta/2}t^2 - t^{4-\delta} + \right. \\
&\left. + 8\delta\gamma(\delta)x^{2-\delta/2} - 2xt^{2-\delta}\right) \Big/ \left(2x^2\left(8\gamma(\delta)\varepsilon x - \right.\right. \\
&\left.\left. - 8\gamma(\delta)x^{1-\delta/2} + t^{2-\delta}\right)\right).
\end{aligned}
$$

Знаменатель в последнем выражении совпадает с точностью до множителя $2x^3$ с выражением под знаком логарифма в определении $g(x,t)$, а следовательно, он положителен. В таком случае условие $g'_x(x,t) > 0$ равносильно неравенству

$$
\begin{aligned}
&x\left(8\delta\gamma(\delta)x^{1-\delta/2} - \left(2t^{2-\delta} + 8\gamma(\delta)\varepsilon t^2\right)\right) + \\
&+ 8\gamma(\delta)x^{1-\delta/2}t^2 - t^{4-\delta} \geqslant 0\,,
\end{aligned}
$$

для чего достаточно, чтобы

$$
\begin{aligned}
x^{1-\delta/2} &\geqslant \max\left\{\frac{t^{2-\delta}}{4\delta\gamma(\delta)} + \frac{\varepsilon t^2}{\delta},\ \frac{t^{2-\delta}}{8\gamma(\delta)}\right\} = \\
&= \frac{t^{2-\delta}}{4\delta\gamma(\delta)} + \frac{\varepsilon t^2}{\delta}\,.
\end{aligned}
$$

Таким образом, при всех $0 \leqslant t \leqslant T$ и

$$
\begin{aligned}
x &\geqslant N_3(T,\varepsilon) = \\
&= \max\left\{\left(\frac{2}{\varepsilon}\right)^{2/\delta}, \left(\frac{T^{2-\delta}}{4\delta\gamma(\delta)} + \frac{\varepsilon T^2}{\delta}\right)^{2/(2-\delta)}\right\}
\end{aligned}
$$

функция $g(x,t)$ монотонно возрастает по x, а следовательно, при всех $N_3 \geqslant N_3(T,\varepsilon)$

$$
\sup_{0\leqslant t\leqslant T}\ \sup_{n\geqslant N_3} g(n,t) = \sup_{0\leqslant t\leqslant T}\ \lim_{x\to\infty} g(x,t) = \ln(\gamma(\delta)\varepsilon)\,,
$$

что и требовалось доказать. Лемма доказана.

Наконец, правильно организовать процесс вычислительной оптимизации позволяют следующие утверждения.

Лемма 4 (см. [24]). *Для любого распределения F с нулевым средним и единичной дисперсией справедливо неравенство*

$$
\sup_x |F(x) - \Phi(x)| \leqslant \sup_{x>0}\left(\Phi(x) - \frac{x^2}{1+x^2}\right) = 0{,}54093654\ldots
$$

Лемма 5. (см. [23]). *Для любой функции распределения $F \in \mathcal{F}_{2+\delta}$ и всех $n \geqslant 2$ таких, что $(\beta_{2+\delta}+1)/n^{\delta/2} \leqslant 0{,}6$, справедливо неравенство*

$$
\rho(F_n, \Phi) \leqslant C'(\delta)\frac{\beta_{2+\delta}}{n^{\delta/2}} + \frac{C''(\delta)}{n^{\delta/2}}
$$

с $C'(\delta)$ и $C''(\delta)$, указанными в табл. 4.

Таблица 4 Значения $C'(\delta)$ и $C''(\delta)$ из леммы 5 при некоторых δ

δ	$C'(\delta)$	$C''(\delta)$
0,9	0,3085	0,2399
0,8	0,2987	0,2166
0,7	0,2912	0,1921
0,6	0,2852	0,1655
0,5	0,2800	0,1382
0,4	0,2765	0,1044
0,3	0,2776	0,0714
0,2	0,2915	0,0327
0,1	0,1500	0,0021

Из леммы 5 вытекает, что при всех n и $\beta_{2+\delta}$ таких, что $(\beta_{2+\delta}+k)/n^{\delta/2} < 0.3(1+k)$, неравенство (2) имеет место при любых $k \in [0,\,1]$ и $C_k(\delta) > 0$, удовлетворяющих условию $(k+1)C_k(\delta) \geqslant C'(\delta)+C''(\delta)$.

Подставляя оценки для характеристических функций из леммы 2 в правую часть неравенства сглаживания Правитца из леммы 1, получаем некоторую функцию $D(\ell, n, t_0, U)$, мажорирующую равномерное расстояние $\rho(F_n, \Phi)$ при всех $U > 0$, $t_0 \in (0, 1]$, $n \geqslant 1$ и F с фиксированной ляпуновской дробью $\beta_{2+\delta} n^{-\delta/2} = \ell$. Приведенные леммы дают основание ограничить область рассматриваемых значений величины $\varepsilon = (\beta_{2+\delta}+k)/n^{\delta/2}$ некоторым конечным отрезком, отделенным от нуля (подробнее об этом будет сказано ниже), и искать константу C_k при каждом $k \in [0,\,1]$ в виде

$$\left.\begin{aligned} C_k(\delta) &= \max_{\varepsilon} C_\delta(\varepsilon)\,; \\ C_\delta(\varepsilon) &= \frac{D_\delta(\varepsilon)}{\varepsilon}\,; \\ D_\delta(\varepsilon) &= \sup\{D_\delta(\varepsilon, n)\colon n \geqslant n_*\}\,, \end{aligned}\right\} \qquad (3)$$

где

$$D_\delta(\varepsilon, n) = \inf_{t_0,\,U>0} D\left(\varepsilon - \frac{k}{n^{\delta/2}}, n, t_0, U\right),$$

$$n_* = \max\left\{1,\, \left\lceil \left(\frac{1+k}{\ell}\right)^{2/\delta} \right\rceil\right\}.$$

Здесь $\lceil x \rceil$ — минимальное целое, не меньшее x. Условие $n \geqslant n_*$ является следствием неравенства $\beta_{2+\delta} \geqslant 1$. При этом для оценивания супремума по n вместо входящих в $D_\delta(\varepsilon, n)$ величин r_j, $j = 1, 2, 3, 4$, для достаточно больших n используются их монотонные мажоранты. Вычисление максимума по ε существенно опирается на свойство монотонного возрастания по ε всех используемых оценок для функций $|f_n(t)|$ и $r_n(t)$, а следовательно, и величины $D_\delta(\varepsilon) = \varepsilon C_\delta(\varepsilon)$. Это свойство позволяет оценить $\max\limits_{\varepsilon} C_\delta(\varepsilon)$ по значениям $C_\delta(\varepsilon)$ лишь в конечном числе точек. А именно имеет место

Лемма 6. *Для всех $\varepsilon_2 > \varepsilon_1 > 0$ имеет место неравенство*

$$\max_{\varepsilon_1 \leqslant \varepsilon \leqslant \varepsilon_2} C_\delta(\varepsilon) \leqslant C_\delta(\varepsilon_2)\frac{\varepsilon_2}{\varepsilon_1}.$$

Минимизация функции $D(\varepsilon - kn^{-\delta/2},\, n,\, t_0,\, U)$ по t_0 и U проводится численно с использованием стандартных процедур в системе Matlab 7.3 (R2006b).

Перейдем теперь к описанию алгоритма вычисления константы $C_0(\delta)$ в неравенстве (1). Положим

$$\varepsilon = \frac{\beta_{2+\delta}}{n^{\delta/2}}.$$

Таблица 5 Экстремальные значения n^*, $\varepsilon^* = (n^*)^{-\delta/2}$ и оптимальные t_0, U при вычислении константы $C_0(\delta)$

δ	$\varepsilon_{\max}$	n^*	ε^*	t_0	U
0,9	1,006	3	0,610	0,35	4,94
0,8	0,946	3	0,644	0,39	4,40
0,7	0,898	3	0,681	0,46	3,81
0,6	0,862	4	0,660	0,53	3,71
0,5	0,844	5	0,669	0,66	3,35
0,4	0,853	6	0,699	0,82	3,09
0,3	0,874	5	0,786	1,00	2,69
0,2	0,888	4	0,871	0,78	2,42
0,1	0,898	9	0,896	0,87	2,53

Тогда при $\varepsilon \leqslant 0.3$ неравенство (1) с $C_0(\delta)$, указанной в теореме 1, вытекает из леммы 5. С другой стороны, лемма 4 позволяет ограничить сверху область рассматриваемых значений ε величиной $0{,}5409\ldots/C_0(\delta) \equiv \varepsilon_{\max}(\delta)$, фиксированной при каждом δ. Таким образом, при вычислении $C_0(\delta)$ максимизация по ε в формулах (3) проводится на конечном отрезке $0{,}3 \leqslant \varepsilon \leqslant \varepsilon_{\max}(\delta)$. Значения правой границы $\varepsilon_{\max}(\delta)$ приведены в табл. 5. Для оценки характеристических функций при $n < 100$ используется f_1, а при $n \geqslant 100$ — функция f_2, монотонно убывающая по n, что позволяет при каждом ε оценивать супремум $D_\delta(\varepsilon, n)$ по значениям n лишь в конечном числе точек: $n_*, \ldots, \max\{n_*, 100\}$, где $n_* = \max\{1, \varepsilon^{-2/\delta}\}$. Максимум $C_\delta(\varepsilon) = D_\delta(\varepsilon)/\varepsilon$ по $0{,}3 \leqslant \varepsilon \leqslant \varepsilon_{\max}(\delta)$ оценивается с помощью леммы 6 и не превосходит тех значений $C_0(\delta)$, которые указаны в формулировке теоремы 1. Экстремальные значения $n = n^*$ и $\varepsilon = (\ell^*)^{-\delta/2}$ указаны в табл. 5 в третьем и четвертом столбцах, а соответствующие оптимальные значения параметров t_0 и U — в пятом и шестом столбцах. Отметим, что точке экстремума соответствует $\beta_{2+\delta} = 1$.

Пусть теперь $k = 1$. Обозначим

$$\varepsilon = \frac{\beta_{2+\delta}+1}{n^{\delta/2}}.$$

Тогда при $\varepsilon \leqslant 0{,}3$ неравенство (2) является следствием леммы 5, а при $\varepsilon \geqslant 0{,}5409\ldots/C_1(\delta) \equiv \equiv \varepsilon_{\max}(\delta)$ — следствием леммы 4. Таким образом, при вычислении $C_1(\delta)$ максимизацию по ε в формулах (3) достаточно проводить на отрезке $0{,}3 \leqslant \varepsilon \leqslant \varepsilon_{\max}(\delta)$. Значения $\varepsilon_{\max}(\delta)$ приведены в табл. 6. Из этой таблицы видно, что максимальное рассматриваемое значение ε не превосходит 1,76. Для вычисления супремума по $n \geqslant n_*$ используется лемма 3 с $T = 2{,}2$, соответствующие значения $N_1 = = N_1(2{,}2)$ и $N_3 = N_3(2{,}2,\, 1{,}76)$ приведены в табл. 6 (для $N_3(T, \varepsilon)$ взято «с запасом» значение $\varepsilon = 1{,}76$). Как видно, уже при $n \geqslant 184$ для всех рассматриваемых значений δ можно использовать оценки $\widetilde{r}_1(t, \varepsilon)$, $\widetilde{r}_3(t, \varepsilon)$ из леммы 3. Таким образом, супремум по n

Таблица 6 Экстремальные значения ε^* и оптимальные t_0, U при вычислении константы $C_1(\delta)$

δ	$\varepsilon_{\max}$	N_1	N_3	ε^*	t_0	U
0,9	1,752	38	142	1,061	0,38	2,14
0,8	1,698	37	110	1,108	0,40	2,12
0,7	1,623	36	91	1,135	0,43	2,09
0,6	1,534	36	79	1,143	0,44	2,06
0,5	1,434	37	73	1,136	0,46	2,02
0,4	1,326	40	71	1,114	0,47	1,99
0,3	1,213	49	75	1,079	0,48	1,96
0,2	1,099	72	90	1,032	0,49	1,93
0,1	0,985	184	144	0,976	0,49	1,90

достаточно оценивать по значениям n лишь в конечном числе точек: $n_*, \ldots, \max\{n_*, 184\}$, где $n_* = \max\{1, (2/\varepsilon)^{2/\delta}\}$. При этом экстремум целевой функции не превосходит значений, указанных в теореме 1 и достигается при $n \to \infty$ и $\varepsilon = \varepsilon^*(\delta)$ — указано в пятом столбце табл. 6. Соответствующие оптимальные значения t_0 и U приведены в шестом и седьмом столбцах табл. 6.

В заключение авторы выражают свою признательность В. Ю. Королеву за поддержку и постоянное внимание к работе.

Литература

1. *Berry A. C.* The accuracy of the Gaussian approximation to the sum of independent variates // Trans. Amer. Math. Soc., 1941. Vol. 49. P. 122–139.
2. *Esseen C.-G.* On the Liapunoff limit of error in the theory of probability // Ark. Mat. Astron. Fys., 1942. Vol. A28. No. 9. P. 1–19.
3. *Katz M.* A note on the Berry–Esseen theorem // Ann. Math. Statist., 1963. Vol. 34. P. 1107–1108.
4. *Петров В. В.* Одна оценка отклонения распределения суммы независимых случайных величин от нормального закона // ДАН СССР, 1965. Т. 160. Вып. 5. С. 1013–1015.
5. *Бикялис А.* Оценки остаточного члена в центральной предельной теореме // Литовский математический сб., 1966. Т. 6. Вып. 3. С. 323–346.
6. *Петров В. В.* Суммы независимых случайных величин. — М.: Наука, 1972.
7. *Esseen C.-G.* A moment inequality with an application to the central limit theorem // Skand. Aktuarietidskr., 1956. Vol. 39. P. 160–170.
8. *Королев В. Ю., Шевцова И. Г.* Уточнение неравенства Берри–Эссеена с приложениями к пуассоновским и смешанным пуассоновским случайным суммам // Обозрение прикладной и промышленной математики, 2010. Т. 17. Вып. 1. С. 25–56.
9. *Tysiak W.* Gleichmäßige und nicht-gleichmäßige Berry–Esseen–Abschatzungen. Dissertation. — Wuppertal, 1983.
10. *Paditz H.* On the error-bound in the nonuniform version of Esseen's inequality in the L_p-metric // Statistics, 1996. Vol. 27. P. 379–394.
11. *Гапонова М. О., Корчагин А. Ю., Шевцова И. Г.* Об абсолютных константах в равномерной оценке точности нормальной аппроксимации для распределений, не имеющих третьего момента // Сб. статей молодых ученых факультета ВМК МГУ. Вып. 6. — М.: Макс Пресс, 2009. С. 81–89.
12. *Paditz H.* Über eine Fehlerabschätzung im zentralen Grenzwertsatz // Wiss. Z. Hochschule für Verkehswesen "Friedrich List." — Dresden, 1986. Bd. 33. H. 2. S. 399–404.
13. *Шевцова И. Г.* Об асимптотически варл правильных постоянных в центральной предельной теореме // Теория вероятностей и ее применения, 2010 (в печати). Т. 55. Вып. 2.
14. *Королев В. Ю., Шевцова И. Г.* О верхней оценке абсолютной постоянной в неравенстве Берри–Эссеена // Теория вероятностей и ее применения, 2009. Т. 54. Вып. 4. С. 671–695.
15. *Шевцова И. Г.* Нижняя асимптотически правильная постоянная в центральной предельной теореме // Докл. РАН, 2010. Т. 430. Вып. 4. С. 466–469.
16. *Королев В. Ю., Шевцова И. Г.* Уточнение неравенства Берри–Эссеена // Докл. РАН, 2010. Т. 430. Вып. 6. С. 738–742.
17. *Золотарёв В. М.* Абсолютная оценка остаточного члена в центральной предельной теореме // Теория вероятностей и ее применения, 1966. Т. 11. Вып. 1. С. 108–119.
18. *Золотарёв В. М.* Некоторые неравенства теории вероятностей и их применение к уточнению теоремы А. М. Ляпунова // ДАН СССР, 1967. Т. 177. № 3. С. 501–504.
19. *Zolotarev V. M.* A sharpening of the inequality of Berry–Esseen // Z. Wahrsch. verw. Geb., 1967. Bd. 8. P. 332–342.
20. *Prawitz H.* Limits for a distribution, if the characteristic function is given in a finite domain // Scand. Aktuar Tidskr., 1972. P. 138–154.
21. *Шевцова И. Г.* Некоторые оценки для характеристических функций с применением к уточнению неравенства Мизеса // Информатика и её применения, 2009. Т. 3. Вып. 3. С. 69–78.
22. *Prawitz H.* On the remainder in the central limit theorem. I. Onedimensional independent variables with finite absolute moments of third order // Scand. Actuarial J., 1975. No. 3. P. 145–156.
23. *Гапонова М. О., Шевцова И. Г.* Асимптотические оценки абсолютной постоянной в неравенстве Берри–Эссеена для распределений, не имеющих третьего момента // Информатика и её применения, 2009. Т. 3. Вып. 4. С. 41–56.
24. *Бхаттачария Р. Н., Ранга Р. Р.* Аппроксимация нормальным распределением. — М.: Наука, 1982.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 2. С. 83–92

ЛИНГВИСТИЧЕСКИЕ ФИЛЬТРЫ В СТАТИСТИЧЕСКИХ МОДЕЛЯХ МАШИННОГО ПЕРЕВОДА

Е. Б. Козеренко[1]

Аннотация: Рассмотрены задачи создания лингвистических фильтров в статистических моделях машинного перевода и вопросы совершенствования механизмов выравнивания параллельных текстов для повышения точности и адекватности переводов. Приведены статистические и эвристические модели выравнивания и перевода. Предложены решения на основе гибридной грамматики, включающей лингвистические правила и вероятностные характеристики языковых структур.

Ключевые слова: статистические модели; машинный перевод; параллельные тексты; выравнивание; лингвистические фильтры

1 Введение

В машинном переводе, основанном на статистических методах, или статистическом машинном переводе (СМП), задача перевода с одного естественного языка на другой рассматривается как задача машинного обучения. Это означает, что через обучение на очень большом числе образцов переводов, выполненных людьми-переводчиками, алгоритмы СМП осваивают правила перевода автоматически. Машинный перевод на основе статистики был впервые предложен в [1, 2]. Применение статистических моделей чрезвычайно продвинуло развитие машинного перевода за последние два десятилетия, однако в дальнейшем появляются новые идеи и методы, направленные на создание систем, эффективно сочетающих разные модели.

В последние годы все более отчетливой становится тенденция к выработке новых подходов к построению моделей машинного перевода, в которых учтены сильные стороны как статистических моделей, так и подходов на основе лингвистических правил. Встречное движение к созданию гибридных моделей обработки естественного языка идет как со стороны статистических подходов к машинному переводу, когда в чисто статистические модели встраиваются элементы грамматик, так и со стороны традиционных методов машинного перевода, в которых правила анализа и перевода дополняются статистическими данными. Эти данные учитываются при принятии решений «машиной перевода», особенно для разрешения неоднозначности языковых объектов.

Истоки стохастической исследовательской парадигмы применительно к задачам обработки естественного языка находятся в проектах разработки алгоритмов распознавания речи, символов, исправления орфографии. Основным методом решения многих задач, в частности определения и разметки частей речи, вероятностного грамматического разбора, является правило Байеса. В архитектуре стохастических систем в основном используется алгоритм динамического программирования.

Машинное обучение в значительной степени основано на стохастической исследовательской парадигме. Алгоритмы обучения могут быть двух типов: неуправляемые и управляемые. Неуправляемый алгоритм должен вывести модель, пригодную для обобщения новых данных, которые ему ранее не предъявлялись, и этот вывод должен быть основан только на данных. Управляемый же алгоритм обучается на множестве правильных ответов на данные из обучающей выборки таким образом, что выведенная модель дает более точные решения. Целью машинного обучения является автоматический вывод модели для некоторой области на основе данных из этой области. Таким образом, система, обучаемая, например, синтаксическим правилам, должна быть обеспечена базовым набором правил фразовых структур. В последнее время исследователи стали уделять больше внимания построению N-граммов, отражающих сложности синтаксических и семантических структур [3, 4], применению N-граммов переменной длины [5], включению семантической информации в N-граммы. В работе [6] дается детальное описание подхода к созданию статистического машинного перевода, основанного на N-граммах двуязычных единиц, называемых «кортежами», а также четырех специальных атрибутных функций.

[1]Институт проблем информатики Российской академии наук, kozerenko@mail.ru

Статистические модели перевода строятся на основе данных, получаемых из корпусов параллельных текстов на разных языках. Обычно сравнение текстов производится для языковых пар. Текст на языке, с которого необходимо осуществить перевод, называют исходным, а текст, который является его переводом, называют целевым. Соответственно говорят об исходном языке и целевом языке (или языке перевода).

Основным способом получения данных о соответствии между исходным и целевым текстами и языками служит процедура выравнивания текстов. Результат этой процедуры также называется выравниванием и обозначается через A (alignment — выравнивание). Вероятностные характеристики выравниваний используются для создания алгоритмов перевода в статистических моделях машинного перевода. Таким образом, выравнивания и распределения вероятностей являются ключевыми понятиями при описании этих моделей.

В данной статье используются следующие нотации: символ P используется для обозначения распределений вероятности в самом общем смысле, а символ p используется для обозначения распределений вероятности на основе некоторой особой модели. Основное внимание уделяется описанию различных методов выравнивания параллельных текстов на разных языках, поскольку результаты выравнивания определяют точность и адекватность перевода. При этом все более возрастает роль *лингвистических фильтров*, которые вводятся в виде структур данных и правил в статистические модели перевода.

Рассматриваемые модели иллюстрируются исходя из двуязычной ситуации: в фокусе внимания находится языковая пара *русский–английский*. Однако сходные методы применимы и при рассмотрении выравниваний и переводов с русского языка на французский, немецкий и другие европейские языки.

2 Методы выравнивания параллельных текстов

Статистические подходы к выравниванию параллельных текстов направлены на то, чтобы найти наиболее вероятный вариант выравнивания A для двух заданных параллельных текстов S и T:

$$\arg\max_A P\left(A|S,T\right) = \arg\max_A P\left(A,S,T\right) .$$

Для того чтобы оценить значения вероятностей, указанных в этом выражении, чаще всего применяются методы, которые представляют параллельные тексты в виде последовательности выравниваемых цепочек предложений $(B_1,\ldots,B_K)$. При этом предполагается, что вероятность одной цепочки не зависит от вероятностей других цепочек, а зависит только от предложений в данной цепочке [7]. Тогда

$$P(A,S,T) \approx \prod_{k=1}^{K} P(B_k) .$$

Этот метод просто учитывает длину предложения на исходном языке и на языке перевода, измеренную в символах. Предполагается, что более длинное предложение одного языка будет соответствовать более длинному предложению другого языка. Такой подход дает вполне устойчивые результаты для сходных языков и буквального перевода.

Более тонкие механизмы сопоставления обеспечиваются методами лексического выравнивания. Так, в работе [8] представлен метод выравнивания посредством создания модели последовательного пословного перевода. Наилучшим результатом выравнивания будет тот, который максимизирует вероятность порождения корпуса при заданной модели перевода. Для выравнивания двух текстов S и T необходимо разбить их и представить в виде последовательности цепочек предложений. Цепочка содержит ноль или более предложений на каждом из языков, а последовательность цепочек покрывает весь корпус

$$B_k = (S_{a_k},\ldots,S_{b_k};\, t_{c_k},\ldots,t_{d_k}) .$$

Затем наиболее вероятное выравнивание $A = B_1,\ldots,B_{m_A}$ данного корпуса определяется следующим выражением (при этом цепочки предложений не зависят друг от друга):

$$\arg\max_A P\left(S,T,A\right) = \arg\max_A P\left(L\right) \prod_{k=1}^{m_A} P(B_k) ,$$

где $P(L)$ означает вероятность того, что порождается выравнивание L цепочек. Модель перевода, используемая при этом подходе, предельно упрощена и не учитывает фактор порядка слов в предложении и возможность того, что слову в исходном тексте может соответствовать более чем одно слово в тексте перевода. В этой модели используются цепочки слов, при этом они ограничены соответствиями 1:1, 0:1 и 1:0. Суть модели заключается в том, что если некоторое слово обычно переводится словом другого языка, то вероятность соответствия цепочек слов 1:1 будет высокой — значительно выше, чем произведение вероятностей соответствий 1:0 и 0:1 цепочек слов, использующих это рассматриваемое

слово. При этом программа выбирает наиболее вероятный вариант выравнивания.

Модель перевода, основанная на пословном выравнивании (например, русского и английского параллельных текстов) будет выглядеть следующим образом:

$$P\left(r|e\right) = \frac{1}{Z}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{j=1}^{m}P\left(r_j|e_{a_j}\right)\ ,$$

где e — предложение на английском языке; l — длина e, выраженная в словах; r — предложение на русском языке; m — длина r; r_j — j-е слово в r; a_j — позиция в e, с которой выравнивается r_j; $P(w_r|w_e)$ — вероятность перевода, т. е. вероятность того, что w_r окажется в предложении на русском языке, если соответствующее w_e встречается в английском предложении; Z — константа нормализации.

Для конкретного выравнивания перемножаются m вероятностей переводов, при этом отдельные переводы не зависят один от другого. Так, если необходимо вычислить вероятность

$$P(\text{Колумб}|\text{Columbus})\times P(\text{открыл}|\text{discovered})\times \\ \times P(\text{Америку}|\text{America})$$

для выравнивания P(Колумб|Columbus), (открыл|discovered), (Америку|America) следует перемножить вероятности этих трех переводных соответствий. При этом для каждого выравнивания делаются два упрощающих допущения: каждое русское слово порождается ровно одним английским словом (которое может быть нулевым, т. е. отсутствовать) и порождение каждого русского слова не зависит от других порождаемых слов в русском предложении.

Однако выше описанный подход, основанный на пословном сопоставлении и никак не учитывающий связи между словами и фразами, не дает оптимальных результатов при выравнивании русско- и англоязычных текстов, поскольку между этими языками имеются определенные структурные различия и при переводе могут осуществляться значительные трансформации.

Если рассматриваемые языки структурно отличаются, применяются методы, ориентированные на привлечение грамматических знаний, например используются методы выравнивания по словам, относящимся к значимым частям речи [9]. При этом служебные слова не учитываются. Для использования этих методов необходимо произвести разметку параллельных текстов по частям речи.

Методы выравнивания параллельных текстов для создания статистических моделей перевода, как правило, разрабатывались на основе сопоставления слов: каждому слову в цепочке на исходном языке необходимо было найти соответствующее слово в цепочке на целевом языке (т. е. языке перевода) и в обратном порядке. Однако довольно часто бывает трудно определить, какие слова в целевой цепочке и цепочке исходной соответствуют друг другу. Особые проблемы возникают при попытках выравнивания слов внутри идиом, в случае переводческих трансформаций, при вольном переводе и при опущении служебных слов.

Выравнивание двух цепочек слов может происходить весьма сложным образом. Очень часто приходится учитывать различные перестановки слов, опущения, вставки и межуровневые выравнивания «слово–фраза», когда одному слову в исходном тексте соответствует целая фраза в целевом тексте, или наоборот. Самое общее определение пословного выравнивания приводится в [10]. Пусть даны две цепочки слов: одна на исходном языке (например, русском —r) $r_1^J = r_1,\ldots,r_j,\ldots,r_J$, а другая — на целевом языке (английском — e) $e_1^I = e_1,\ldots,e_i,\ldots,e_I$, и для этих цепочек необходимо установить выравнивание. Выравнивание между двумя цепочками слов — это подмножество декартова произведения позиций слов, т. е. выравнивание A определяется как

$$A \subseteq \{(j,i):\ j=1,\ldots,J;\ i=1,\ldots,I\}\ .$$

В машинном переводе, основанном на статистических методах, делается попытка построения модели вероятности перевода $P(r_1^J|e_1^I)$, которая описывает соотношение между некоторой цепочкой r_1^J на исходном языке и цепочкой e_1^I на целевом языке. В статистических моделях выравнивания текстов $P(r_1^J,\,a_1^J|e_1^I)$ вводится «скрытое» выравнивание a_1^J, которое описывает отображение из исходной позиции j в целевую позицию a_j. Соотношение между моделью перевода и моделью выравнивания задается следующим образом:

$$P\left(r_1^J|e_1^I\right) = \sum_{a_1^J} P\left(r_1^J,\,a_1^J|e_1^I\right)\ .$$

Выравнивание a_1^J может содержать выравнивания $a_j = 0$ с пустым словом e_0 для тех слов исходного языка, которые не были выравнены ни с каким словом целевого языка.

В целом статистическая модель зависит от множества неизвестных параметров θ, которые извлекаются из обучающего набора данных в процессе обучения. Для того чтобы выразить зависимость модели от множества параметров, используется следующее представление:

$$P\left(r_1^J,\,a_1^J|e_1^I\right) = p_\theta\left(r_1^J,\,a_1^J|e_1^I\right)\ .$$

Искусство статистического моделирования состоит в том, чтобы разработать специфические статистические модели, которые бы отражали наиболее важные свойства рассматриваемой проблемной области. Так, статистическая модель выравнивания должна адекватно описывать соотношение между цепочкой на исходном и целевом языках.

Для выявления неизвестных параметров θ задается обучающий корпус параллельных текстов, содержащий S пар предложений $\{(r_s, e_s) : s = 1, \ldots, S\}$. Для каждой пары (r_s, e_s) переменная выравнивания обозначается как $a = a_1^J$. Неизвестные параметры определяются путем максимизации сходства параллельных текстов в корпусе:

$$\hat{\theta} = \arg\max_{\theta} \prod_{s=1}^{S} \sum_{a} p_\theta\left(r_s,\, a|e_s\right) .$$

Как правило, для подобных моделей максимизация осуществляется на основе алгоритма максимизации ожидания [11] или ему подобных. Такой алгоритм полезен для решения задачи оценки параметров, но не является абсолютно необходимым для статистического подхода.

Итак, несмотря на то что для некоторой заданной пары предложений существует большое число выравниваний, всегда можно найти наилучшее выравнивание

$$\hat{a}_1^J = \arg\max_{a_1^J} p_{\hat{\theta}}\left(r_1^J,\, a_1^J|e_1^I\right) .$$

Выравнивание $\hat{a}_1^J$ также называется выравниванием *Витерби* для пары предложений (r_1^J, e_1^I). Оценка качества выравнивания Витерби осуществляется путем сравнения с некоторым эталонным выравниванием, проведенным вручную. Параметры статистических моделей выравнивания оптимизируются с учетом критерия максимального правдоподобия, который далеко не всегда отражает качество выравнивания.

3 Сопоставление статистических и эвристических моделей выравнивания

Наиболее известными статистическими моделями выравнивания параллельных текстов являются модели, приведенные в [1, 2, 10], а также скрытая марковская модель выравнивания, описанная в [12]. Каждая из этих моделей предлагает свое, отличное от других, разложение вероятности $P(r_1^J,\, a_1^J|e_1^I)$.

Значительно более простые методы нахождения выравниваний по словам (получившие название *эвристических*) используют функцию сходства между типами двух языков [13–15]. В качестве такой функции сходства используются варианты коэффициента Дайса [16]. Для каждой пары предложений формируется матрица, в которой представлены меры ассоциаций между каждым словом в каждой позиции:

$$\mathrm{dice}(i, j) = \frac{2C(e_i, r_j)}{C(e_i)C(r_j)} ,$$

где $C(e, r)$ обозначает частоту совместной встречаемости e и r в обучающем корпусе параллельных текстов. Соответственно $C(e)$ означает частоту появления e в предложениях на целевом языке, а $C(r)$ — частоту r в предложениях на исходном языке. Выравнивание слов из такой матрицы мер ассоциаций получают посредством применения подходящего эвристического метода. Один из таких методов заключается в том, чтобы выбрать в качестве выравнивания $a_j = i$ для позиции j такое слово, у которого мера ассоциации является наибольшей:

$$a_j = \arg\max\left\{\mathrm{dice}(i, j)\right\} .$$

Развитие этого метода дается в [15]. Суть его состоит в итерационном нахождении выравниваний с наибольшей мерой ассоциативности на каждом шаге.

Основное достоинство эвристических моделей — их простота, поэтому они широко применяются для пословного выравнивания и результаты подробно описаны в литературе. Однако то обстоятельство, что в эвристических моделях функция сходства задается весьма произвольно, делает их менее состоятельными, чем статистические модели.

Наиболее распространенной статистической моделью, которая применяется для выравнивания параллельных текстов, является скрытая марковская модель. Модель выравнивания $P(r_1^J,\, a_1^J|e_1^I)$ можно структурировать без потери общности следующим образом:

$$\begin{aligned} P\left(r_1^J,\, a_1^J|e_1^I\right) &= \\ &= P\left(J|e_1^I\right) \prod_{j=1}^{J} P\left(r_j,\, a_j|r_1^{j-1},\, a_1^{j-1},\, e_1^I\right) = \\ &= P\left(J|e_1^I\right) \prod_{j=1}^{J} P\left(a_j|r_1^{j-1},\, a_1^{j-1},\, e_1^I\right) \times \\ &\qquad \times P\left(r_j|r_1^{j-1},\, a_1^j,\, e_1^I\right) . \end{aligned}$$

При использовании этого разложения получаются три различные вероятности: длины $P(J|e_1^I)$, выравнивания $P(a_j|r_1^{j-1},\, a_1^{j-1},\, e_1^I)$ и лексикона $P(r_j|r_1^{j-1},\, a_1^j,\, e_1^I)$. В скрытой марковской модели выравнивания предполагается зависимость первого порядка для выравниваний a_j, а также то, что вероятность лексикона зависит только от слова в позиции a_j:

$$P\left(a_j|r_1^{j-1},\, a_1^{j-1},\, e_1^I\right) = p\left(a_j|a_{j-1},\, I\right) ;$$

$$P\left(r_j|r_1^{j-1},\, a_1^j,\, e_1^I\right) = p\left(r_j|e_{a_j},\right) .$$

Если принять простую модель длины $P(J|e_1^I) = p(J|I)$, то для $p(r_1^J|e_1^I)$ получается следующее разложение на основе скрытой марковской модели:

$$p\left(r_1^J|e_1^I\right) = p\left(J|I\right) \sum_{a_1^J} \prod_{j=1}^{J} \left[p\left(a_{j-1},\, I\right) p\left(r_j|e_{a_j}\right)\right]$$

с вероятностью выравнивания $p(i|i',\, I)$ и вероятностью перевода $p(r|e)$. Для того чтобы сделать параметры выравнивания независимыми от абсолютных значений позиций слов, принимается, что вероятности выравниваний $p(i|i',\, I)$ зависят только от ширины шага $(i-i')$. Используя множество неотрицательных параметров $\{c(i-i')\}$, можно записать вероятности выравнивания в следующем виде:

$$p\left(i|i',\, I\right) = \frac{c(i-i')}{\sum\limits_{i''=1}^{I} c(i''-i')} .$$

Этот вид обеспечивает то, что вероятности выравнивания удовлетворяют ограничению нормализации для каждой обусловливающей позиции слова i', $i' = 1, \ldots, I$. Эта модель называется также однородной скрытой марковской моделью [12]. Подобная идея была предложена в работе [17].

В исходной формулировке скрытой марковской модели выравнивания отсутствует пустое слово, которое порождает слово исходного текста, не имеющее прямого соответствия в виде слова в целевом тексте (т. е. у этого слова нет непосредственного пословного выравнивания).

В работе [18] вводится пустое слово и расширяется сеть скрытой марковской модели посредством I пустых слов e_{I+1}^{2I}. Целевое слово e_i имеет соответствующее пустое слово e_{i+1} (т. е. позиция пустого слова кодирует ранее пройденное целевое слово). На переходы в сети скрытой марковской модели $(i \le I,\, i' \le I)$ накладываются следующие ограничения, в которых участвует пустое слово e_0:

$$p\left(i+I|i',\, I\right) = p_0\delta\left(i,\, i'\right) ;$$
$$p\left(i+I|i'+I,\, I\right) = p_0\delta\left(i,\, i'\right) ;$$
$$p\left(i|i'+I,\, I\right) = p\left(i|i',\, I\right) ,$$

где $\delta(i, i')$ — функция Кронекера, которая равна единице, если $i = i'$, и нулю — в противном случае. Параметр p_0 — это вероятность перехода к пустому слову, которая должна оптимизироваться на предоставляемых данных. В экспериментах, описанных в [18], устанавливается $p_0 = 0{,}2$.

Скрытая марковская модель основана на зависимостях первого порядка $p(i = a_j|a_{j-1},\, I)$ для распределения выравнивания. Также довольно распространенными являются две модели, использующие зависимости нулевого порядка $p(i = a_j|j,\, I,\, J)$:

(1) в модели 1 используется единообразное распределение $p(i|j,\, I, J) = 1/(I+1)$:

$$P\left(r_1^J,\, a_1^J|e_1^I\right) = \frac{p(J|I)}{I+1} \prod_{j=1}^{J} p\left(r_j|e_{a_j}\right) ,$$

из чего следует, что порядок слов не влияет на вероятность выравнивания;

(2) в модели 2 распределение представлено следующим образом:

$$P\left(r_1^J,\, a_1^J|e_1^I\right) = \\ = p(J|I) \prod_{j=1}^{J} \left[p\left(a_j|j,\, I,\, J\right) p\left(r_j|e_{a_j}\right)\right] .$$

Для того чтобы сократить число параметров выравнивания, в модели не учитывают зависимость от J и используют распределение $p(a_j|j,\, I)$ вместо $p(a_j|j,\, I,\, J)$.

4 Методы перевода на основе выравниваний по фразам

Модель перевода по фразам, или модель перевода на основе шаблонов выравнивания [19], и другие сходные модели очень сильно продвинули развитие технологии машинного перевода за счет расширения базовых единиц перевода от слов к фразам, т. е. подстрокам произвольного размера. Однако далеко не всегда фразы этой модели СМП являются фразами в смысле какой-либо существующей синтаксической теории или формальной грамматики, например, фразой может считаться "*alignments the*" и т. п. Однако переход к уровню фраз при формировании данных для СМП позволил представить в

модели локальные перестановки, переводы многословных выражений, вставки и опущения, которые определяются локальным контекстом. Все эти возможности делают модель, основанную на фразах, простым и очень мощным инструментом перевода. Базовая модель фразового перевода является частным случаем модели канала с шумами [2]. Так, перевод предложения с русского языка на английский будет представлен следующим образом:

$$\arg\max_e P\,(e|r) = \arg\max_e P\,(e,\,r) = \\ = \arg\max_e \left(P(e)P(r|e)\right)\,.$$

Модель перевода по фразам $P(r|e)$ «кодирует» e в r, выполняя следующие шаги:

1) сегментирует e на фразы $\overline{e}_1 \ldots \overline{e}_I$, обычно единообразным распределением по всем сегментациям;

2) перегруппировывает $\overline{e}_i$ в соответствии с некоторой моделью искажения;

3) переводит каждую из $\overline{e}_i$ на русский язык в соответствии с моделью $P(\overline{r}|\overline{e})$, которая оценивается на основании обучающих данных.

Другие модели, основанные на фразах, представляют совместное распределение $P(e, r)$ [20] или превращают $P(e)$ и $P(r|e)$ в атрибуты модели [21]. Но базовая архитектура сегментации (или порождения), переупорядочивания фраз и фразового перевода остается такой же. Модели на основе фраз могут надежно выполнять переводы, которые локализованы в подцепочках и которые являются достаточно частотными, чтобы их можно было обнаружить в процессе обучения.

На современном этапе развития исследований и разработок в области машинного перевода и обработки текстовых знаний все больше назревает потребность в подходах, основанных на лингвистических знаниях. Это осознается и сторонниками статистических подходов. Например, в работе [22] описан двухступенчатый метод автоматического извлечения переводных шаблонов из параллельных текстов на английском и китайском языках. Этот метод основан на алгоритме индукции грамматики и алгоритме выравнивания с использованием скобочной трансдукционной грамматики. Однако сами авторы указывают, что, поскольку в данной модели не заложены предварительные синтаксические знания, грамматическая правильность результата не может быть гарантирована.

Основные подходы к статистическому машинному переводу, в которых используется синтаксис, различаются по тем синтаксическим теориям, которых они придерживаются, и системам аннотирования, построенным в соответствии с той или иной теорией.

5 Вероятностные методы грамматического разбора предложений

Статистические методы обработки естественного языка расширяют схему основных существующих подходов к машинному переводу — прямого перевода, переноса (трансфера) и подхода на основе языка-посредника (интерлингвы) [23, 24].

Значения вероятностей для каждого возможного варианта грамматического разбора (т. е. развертывания нетерминального узла) вычисляются на основе частот встречаемости таких вариантов разбора в существующих текстовых корпусах с синтаксической разметкой (treebanks). Значения вероятностей для вариантов разбора могут быть также получены и в виде лингвистических экспертных оценок.

Для любой системы обработки естественного языка необходимо проектирование модуля определения и разметки частей речи (тэггера). Стохастические тэггеры появились в 1980-е гг. Их общая идея заключается в выборе наиболее вероятного тэга (т. е. частеречной метки) для данного слова. Чаще всего для вероятностных тэггеров используются марковские модели. К примеру, для некоторого предложения или последовательности слов выбирается последовательность тэгов, которая максимизирует следующую формулу:

$$P(\text{слово}|\text{тэг}) \times P(\text{тэг}|\text{предыдущие } n \text{ тэгов})\,.$$

Еще один подход к машинному обучению, основанный на правилах и стохастическом тэггировании (разметке частей речи), известен как обучение, основанное на трансформациях (Transformation-Based Learning, TBL — метод управляемого обучения с использованием заранее размеченного обучающего корпуса).

Для вероятностного грамматического разбора применяются стохастические грамматики:

– *вероятностная контекстно-свободная грамматика* определяется в виде $G = (N, T, P, S, D)$, где N — множество нетерминальных символов; T — множество терминальных символов; P — множество продукций вида $A \to b$; A — это нетерминальный символ; b — цепочка символов; S — специальный исходный символ;

D — функция, приписывающая значения вероятности каждому правилу из множества P.

Как получить необходимые данные для вероятностной контекстно-свободной грамматики? Один из путей — использование корпуса синтаксически размеченных предложений. Такой корпус называется банком синтаксических деревьев (treebank). Например, Penn Treebank [25] содержит деревья разбора для ряда текстовых корпусов (Brown Corpus, Switchboard corpus). Если задан банк деревьев разбора, то вероятность каждой развертки некоторого нетерминального узла может быть вычислена путем подсчета частоты ее встречаемости с последующей нормализацией:

$$P(\alpha \to \beta|\alpha) = \frac{\mathrm{Count}(\alpha \to \beta)}{\sum\limits_{\gamma} \mathrm{Count}(\alpha \to \gamma)} = \frac{\mathrm{Count}(\alpha \to \beta)}{\mathrm{Count}(\alpha)};$$

– *вероятностная грамматика замещения деревьев* является обобщением вероятностной контекстно-свободной грамматики, при этом более мощной стохастически, поскольку она дает возможность приписывать значения вероятности фрагментам или даже целым схемам разбора.

Рассмотрим, каким образом значения вероятности используются в процессе грамматического разбора. Например, вероятностная контекстно-свободная грамматика (PCFG — Probabilistic Context Free Grammar) и вероятностная грамматика замещения деревьев (PTSG — Probabilistic Tree Substitution Grammar) присваивают вероятность P каждому дереву разбора T (т. е. каждому деривату) предложения S. Эта информация является ключевой для разрешения неоднозначности синтаксических структур. Вероятность каждого возможного дерева разбора T определяется как произведение вероятностей всех правил r, используемых для развертывания каждого узла n в дереве разбора

$$P(T,S) = \prod_{n \in T} p(r(n)). \qquad (1)$$

Вероятность однозначного предложения (т. е. предложения, где не требуется разрешать неоднозначность) равна вероятности единственного дерева разбора для этого предложения, т. е. $P(T,S) = P(T)$. Вероятность же неоднозначного предложения равна сумме вероятностей всех возможных деревьев разбора $\tau(S)$ данного предложения

$$P(S) = \sum_{T \in \tau(S)} P(T,S) = \sum_{T \in \tau(S)} P(T).$$

Вероятность полного разбора предложения вычисляется с учетом категориальной информации для каждой головной вершины каждого узла. Пусть n — синтаксическая категория некоторого узла N; $h(n)$ — головная вершина узла n; $m(n)$ — материнский узел для n. Тогда будем вычислять вероятность $p(r(n)|n, h(n))$, а для этого преобразуем выражение (1) таким образом, чтобы каждое правило обусловливалось своей головной вершиной:

$$P(T,S) = \prod_{n \in T} p(r(n)|n,\, h(n)) \times p(h(n)|n,\, h(m(n))).$$

Одна из центральных проблем повышения качества машинного перевода — это включение в модель таких языковых фактов, как *перефразы* [26], возможные отношения синонимии синтаксических структур, в частности глагольных фраз в активном и пассивном залоге, номинализация [27]. В указанных работах предлагаются решения для отдельных видов структур, однако необходима возможно полная типизация синонимических языковых структур для выравнивания параллельных текстов и извлечения новых правил на основе методов машинного обучения.

6 Лингвистические фильтры на основе грамматики когнитивного трансфера

В сформулированной ранее *когнитивной трансферной грамматике* (КТГ) [28–30] функциональные значения языковых структур определяются категориальными значениями головных вершин. Вероятностные характеристики вводятся в правила унификационной грамматики в виде весов, присваиваемых деревьям разбора.

В КТГ элементарными структурами являются *трансфемы* [29]. *Трансфема* — это единица когнитивного переноса, устанавливающая функционально-семантическое соответствие между структурами исходного языка L_S и структурами целевого языка L_T. Для выравнивания параллельных текстов трансфемы задаются как правила переписывания, в которых в левой части стоит нетерминальный символ, а в правой — выравненные пары цепочек терминальных и нетерминальных символов, принадлежащих исходному и целевому языкам:

$$T \to \langle \rho,\, \alpha,\, \sim \rangle\,,$$

где T — нетерминальный символ; ρ и α — цепочки терминальных и нетерминальных символов, принадлежащих русскому и английскому языкам, $\sim$ — символ соответствия между нетерминальными символами, входящими в ρ, и нетерминальными символами, входящими в α. При выравнивании параллельных текстов на основе когнитивной трансферной грамматики процесс деривации начинается с пары связанных исходных символов S_ρ и S_α, далее на каждом шаге связанные нетерминальные символы попарно переписываются с использованием двух компонентов единого правила.

Для автоматического извлечения правил из параллельных текстов на основе когнитивной трансферной грамматики необходимо предварительно выравнять тексты по предложениям и словам. Извлекаемые правила опираются на пословные выравнивания таким образом, что вначале идентифицируются исходные фразовые пары с использованием такого же критерия, как и большинство статистических моделей перевода, основанных на фразовом подходе [19]: должно быть хотя бы одно слово внутри фразы на одном языке, выравненное с некоторым словом внутри фразы на другом языке, но никакое слово внутри фразы на одном языке не может быть выравнено ни с каким словом за пределами парной ей фразы на другом языке.

Определение 1. *Пусть дана пара предложений $\langle r,\, e,\, \sim \rangle$, выравненных по словам, пусть r_i^j обозначает подцепочку r от позиции i до позиции j включительно, а соответственно, $e_{i'}^{j'}$ — подцепочку e от позиции i' до позиции j' включительно. Тогда правило $\langle r_i^j,\, e_{i'}^{j'},\, \sim \rangle$ — это исходная фразовая пара, если и только если:*

1) $r_k \sim e_{k'}$ для некоторого $k \in [i, j]$ и $k' \in [i', j']$;
2) $r_k \not\sim e_{k'}$ для всех $k \in [i, j]$ и $k' \notin [i', j']$;
3) $r_k \not\sim e_{k'}$ для всех $k \notin [i, j]$ и $k' \in [i', j']$.

Для того чтобы продолжить извлечение правил из выделенных фраз, находим фразы, которые содержат другие фразы, и заменяем их нетерминальными символами. Таким образом осуществляется механизм вложенности правил, отображающий иерархическую структуру естественного языка.

Определение 2. *Множество правил $\langle r,\, e,\, \sim \rangle$ — это наименьшее множество, удовлетворяющее следующим условиям:*

1. *Если $\left\langle r_i^j,\, e_{i'}^{j'} \right\rangle$ — это исходная пара фраз, то $X \to \left\langle r_i^j, e_{i'}^{j'} \right\rangle$ является правилом из $\langle r,\, e,\, \sim \rangle$.*
2. *Если $X \to \langle \rho,\, \alpha \rangle$ — это правило из $\langle r,\, e,\, \sim \rangle$, а $\left\langle r_i^j,\, e_{i'}^{j'},\, \sim \right\rangle$ — это исходная пара фраз такая, что $\rho = \rho_1 r_i^j \rho_2$ и $\alpha = \alpha_1 e_{i'}^{j'} \alpha_2$, то $X \to \left\langle \rho_1 X_{\boxed{k}} \rho_2,\, \alpha_1 X_{\boxed{k}} \alpha_2 \right\rangle$ — это правило из $\langle r,\, e,\, \sim \rangle$ и k — это индекс, не используемый ни в ρ, ни в α.*

Следующий шаг — формирование системы правил в нотации КТГ. Когнитивная трансферная грамматика — это унификационно-порождающая грамматика, имеющая иерархическую структуру и отражающая значительную часть языковых трансформаций, производимых при переводе с одного языка на другой. При этом на основе полученных авторами экспериментальных данных в правила КТГ включены веса различных вариантов их развертки.

Определение 3. *Грамматикой когнитивного трансфера G_{CT} будем называть множество*

$$G_{CT} = \\ = \{T_{L_1}, T_{L_2}, N_{L_1}, N_{L_2}, P_{CA}, P_{CT}, S_{L_1}, S_{L_2}, M, D\}\,,$$

где T_{L_1} и T_{L_2} — множества терминальных символов языков L_1 и L_2; N_{L_1} и N_{L_2} — множества нетерминальных символов языков L_1 и L_2; P_{CA} и P_{CT} — правила анализа и синтеза на основе когнитивного трансфера; S_{L_1} и S_{L_2} — пара исходных символов языков L_1 и L_2, с которых начинается процесс анализа и выравнивания предложений; M — функция установления соответствия между языковыми структурами L_1 и L_2; D — функция, приписывающая значения вероятности каждому правилу из множеств P_{CA} и P_{CT}.

Неоднозначность является коренным свойством естественного языка и вызывает основные затруднения при создании систем машинного перевода. Неоднозначные и многозначные синтаксические структуры учитываются авторами в дальнейшем развитии КТГ — многовариантной когнитивной трансферной грамматике (МКТГ) — и ее реализациях в виде структур данных, которые могут быть использованы в качестве лингвистических фильтров при построении статистических моделей перевода. Эти структуры данных назовем многовариантными когнитивными трансферными структурами (МКТС).

В общем виде синтаксис МКТС может быть представлен следующим образом:

МКТС{МКТС ⟨идентификатор⟩ МКТС ⟨вес⟩ МКТС ⟨метка⟩} →
⟨Входная фразовая структура & набор атрибутов-значений⟩ →
⟨Схема трансфера, управляемого головной вершиной⟩ →
⟨Генерируемая фразовая структура & набор атрибутов-значений 1⟩⟨вес1⟩
⟨Генерируемая фразовая структура & набор атрибутов-значений 2⟩⟨вес2⟩ . . .
. . . ⟨Генерируемая фразовая структура & набор атрибутов-значений N⟩⟨весN⟩ .

В новом варианте МКТГ отражено явление многозначности синтаксических структур и предусмотрены основные механизмы разрешения неоднозначности посредством включения в систему правил разбора и перевода статистической информации о возможных контекстах языковых структур. Многовариантная когнитивная трансферная грамматика обеспечивает расширяемую платформу для разработки систем машинного перевода и извлечения знаний из текста. В настоящее время основные правила когнитивного трансфера сформулированы также для русско-французской и русско-немецкой языковых пар. На основе МКТГ формируется новый гибридный подход к построению моделей для систем машинного перевода и обработки знаний. Продолжают формироваться обучающие наборы данных для расширения и модификации правил. Для сокращения числа избыточных правил (которые неизбежно возникают на этапе обучения) формируются лингвистические фильтры на основе пространств когнитивного трансфера.

7 Заключение

Актуальность проблемы создания новых гибридных методов представления языковых объектов обусловлена тем, что на современном этапе исследований встает задача оптимального сочетания сильных сторон двух исследовательских парадигм: логико-лингвистического моделирования, использующего правила, и стохастического подхода. Особое значение предлагаемая работа имеет для решения проблемы структурного анализа и компьютерного моделирования полнотекстовых научных, финансово-экономических и патентных документов.

Необходимость моделирования языковых трансформаций для систем машинного перевода и извлечения знаний из текстов обусловлена тем, что до сих пор эти явления мало исследованы с точки зрения возможностей их компьютерной реализации и, соответственно, недостаточно учтены в действующих системах МП, а правила, задающие функциональную синонимию языковых конструкций, позволяют извлечь необходимую информацию из параллельных текстов и избежать формирования избыточных правил и «шумов».

При формировании процедур снятия неоднозначности были использованы статистические данные, полученные при изучении параллельных текстов научных и патентных документов собственного экспериментального корпуса. Так, при ранжировании предпочтительности вариантов трансфера использовалась статистика соотношения личных и неличных форм глаголов и оборотов с ними, приоритетности употребления активных или пассивных конструкций, номинализаций и вербализаций в русском, английском, французском и немецком дискурсах.

Дальнейшие исследования будут связаны с расширением числа типов трансформаций в лингвистических представлениях для многоязычной ситуации, созданием инженерно-лингвистической среды исследований и разработок в области машинного перевода и извлечения знаний из текстов на разных языках. Как составная часть инженерно-лингвистической среды разрабатывается многоязычная лингвистическая база знаний, основанная на сочетании методов функционально-семантического анализа фразовых структур и статистических моделях языков, включенных в базу.

Полученные результаты используются также для выработки инновационных подходов к преподаванию курсов перевода и переводоведения, компьютерной лингвистики и когнитивных технологий.

Литература

1. *Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S.* A statistical approach to machine translation // Comput. Linguistics, 1990. Vol. 16. P. 79–85.
2. *Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L.* The mathematics of statistical machine translation: Parameter estimation // Comput. Linguistics, 1993. Vol. 19. No. 2. P. 263–311.
3. *Rosenfeld R.* A maximum entropy approach to adaptive statistical language modeling // Computer Speech Language, 1996. Vol. 10. P. 187–228.

4. *Niesler T. R., Woodland P. C.* Modelling word-pair relations in a category-based language model // IEEE ICASSP-99, 1999. P. 795–798.
5. *Ney H., Essen U., Kneser R.* On structuring probabilistic dependencies in stochastic language modeling // Computer Speech Language, 1994. Vol. 8. P. 1–38.
6. *Marino J. B., Banchs R. E., Crego J. M., de Gispert A., Lambert P., Fonollosa J. A. R., Costa-Jussa M. R.* N-gram-based Machine Translation // Comput. Linguistics, 2006. Vol. 32. No. 4. P. 527–549.
7. *Gale W. A., Church K. W.* A program for aligning sentences in bilingual corpora // Comput. Linguistics, 1993. Vol. 19. P. 75–102.
8. *Chen S. F.* Aligning sentences in bilingual corpora using lexical information // 31st Annual Conference of the Association for Computational Linguistics Proceedings, 1993. P. 9–16.
9. *Masahiko H., Yamazaki T.* High-performance bilingual text alignment using statistical and dictionary information // ACL 34, 1996. P. 131–138.
10. *Och F. J., Ney H.* A comparison of alignment models for statistical machine translation // COLING'00: The 18th Conference (International) on Computational Linguistics. Saarbrucken, Germany, 2000. P. 1086–1090.
11. *Dempster A. P., Laird N. M., Rubin D. B.* Maximum likelihood from incomplete data via the EM algorithm // J. Roy. Statistical Soc. Ser. B, 1977. Vol. 39. No. 1. P. 1–22.
12. *Vogel S., Ney H., Tillmann Ch.* HMM-based word alignment in statistical translation // COLING'96: The 16th Conference (International) on Computational Linguistics Proceedings. Copenhagen, Denmark, 1996. P. 836–841.
13. *Smadja F., McKeown K. R., Hatzivassiloglou V.* Translating collocations for bilingual lexicons: A statistical approach // Comput. Linguistics, 1996. Vol. 22. No. 1. P. 1–38.
14. *Ker S. J., Chang J. S.* A class-based approach to word alignment // Comput. Linguistics, 1997. Vol. 23. No. 2. P. 313–343.
15. *Melamed I. D.* Models of translational equivalence among words / Comput. Linguistics, 2000. Vol. 26. No. 2. P. 221–249.
16. *Dice L. R.* Measures of the amount of ecologic association between species // J. Ecology, 1945. Vol. 26. P. 297–302.
17. *Dagan I., Church K. W., Gale W. A.* Robust bilingual word alignment for machine aided translation // Workshop on Very Large Corpora Proceedings. Columbus, Ohio, 1993. P. 1–8.
18. *Och F. J., Ney H.* A systematic comparison of various statistical alignment models // Comput. Linguistics, 2003. Vol. 29. No. 1. P. 19–51.
19. *Och F. J., Ney H.* The alignment template approach to statistical machine translation // Comput. Linguistics, 2004. Vol. 30. P. 417–449.
20. *Marcu D., Wong W.* A phrase-based, joint probability model for statistical machine translation // EMNLP Proceedings. Philadelphia, PA, 2002. P. 133–139.
21. *Och F. J., Ney H.* Discriminative training and maximum entropy models for statistical machine translation // 40th Annual Meeting of the ACL Proceedings. Philadelphia, PA, 2002. P. 295–302.
22. *Hu R., Zong Ch., Xu B.* An approach to automatic acquisition of translation templates based on phrase structure extraction and alignment // IEEE Transactions on Audio, Speech and Language Processing, 2006. Vol. 14. No. 5. P. 1656–1663.
23. *Dorr B., Habash N.* Interlingua approximation: A generation-heavy approach // AMTA-2002 Interlingua Reliability Workshop. Tiburon, California, USA, 2002.
24. *Voss C., Dorr B. J.* Toward a lexicalized grammar for interlinguas // Machine Translation, 1995. Vol. 10. No. 1–2. P. 139–180.
25. *Marcus M. P., Santorini B., Marcinkiewicz M. A.* Building a large annotated corpus of English: The Penn Treebank // Comput. Linguistics. 1993. Vol. 19. No. 2. P. 313–330.
26. *Callison-Burch Ch., Cohn T., Lapata M.* ParaMetric: An automatic evaluation metric for paraphrasing // 22nd Conference (International) on Computational Linguistics (Coling 2008) Proceedings. Manchester, 2008. P. 97–104.
27. *Dagan I., Bar-Haim R., Szpektor I., Greental I., Shnarch E.* Natural language as the basis for meaning representation and inference // Computational Linguistics and Intelligent Text Processing: 9th Conference, CICLing 2008 Proceedings. Haifa, Israel, 2008. — Springer, 2008. P. 151–170.
28. *Kozerenko E. B.* Cognitive approach to language structure segmentation for machine translation algorithms // Conference (International) on Machine Learning, Models, Technologies and Applications Proceedings. Las Vegas, USA, 2003. — CSREA Press, 2003. P. 49–55.
29. *Козеренко Е. Б.* Лингвистическое моделирование для систем машинного перевода и обработки знаний // Информатика и её применения, 2007. Т. 1. Вып. 1. С. 54–65.
30. *Kozerenko E.* Features and categories design for the English-Russian transfer model // Advances Natural Language Processing Applications Research Comput. Sci., 2008. Vol. 33. P. 123–138.

ABSTRACTS

ON TASK FLOW PLANNING IN COMPUTATIONAL RESOURCE SYSTEMS

M. G. Konovalov

IPI RAN, mkonovalov@ipiran.ru

The problem of flow distribution analysis, optimization, and pricing in shared computational resource systems is examined. The appropriate literature review is given. An approach to mathematical models construction is suggested where task flows being described as dynamic balance equations and quality of service relations are used. The subjects in the systems possess own strategies of behavior and purpose individual aims in terms of cost and quality of service. Distributed gradient algorithm is one of possible system member strategies. A numerical example is given and model development and employment for future trends are discussed.

Keywords: computational resource systems; flow distribution; quality of service; cooperative behavior

NONPARAMETRIC ESTIMATION OF BAYESIAN CLASSIFIER ELEMENTS

M. Krivenko

IPI RAN, mkrivenko@ipiran.ru

The problem of constructing an empirical Bayesian classifier, providing recognition of the text, where some symbols have different picture sizes, is considered. A combined method of constructing an evaluation of Bayesian classifier is proposed. The method includes nonparametric kernel estimation and parametric estimation with the help of the density of normal distribution. This combined assessment allows to deal effectively with the task of handling small amounts of training set. Productivity of the proposed ideas is illustrated by an example of recognizing the real text.

Keywords: Bayesian classifier; combined multivariate density estimation; adaptive kernel estimation; text recognition

SOLVABILITY PROBLEMS IN THE PROTEIN SECONDARY STRUCTURE RECOGNITION

K. V. Rudakov[1] and I. Yu. Torshin[2]

[1]Computing Center of RAS; Moscow Institute of Physics and Technology, rudakov@ccas.ru
[2]Russian Center of the Trace Element Institute for UNESCO, tiy135@yahoo.com

The purpose of the work is to develop a formalism for the application of the algebraic approach to recognition of the protein secondary structure. Paper presents rigorous formal description of the problem and considers its solvability, regularity, and locality. Key terms for the analysis of locality, such as a neighborhood, mask, mask system, monotony, and irreducibility of the mask systems were proposed. An algorithm for constructing nonredundant mask systems was formulated. The formalism has allowed to formulate a correct description of the hypothesis of the local character of the dependence of the secondary structure on the primary and to obtain constructive criteria of the solvability of the problem.

Keywords: algebraic approach; the secondary structure of protein; bioinformatics

ASYMPTOTIC PROPERTIES OF RISK ESTIMATE OF WAVELET-VAGUELETTE COEFFICIENTS THRESHOLDING IN TOMOGRAPHIC RECONSTRUCTION PROBLEM

A. V. Markin[1] and O. V. Shestakov[2]

[1] Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, artem.v.markin@mail.ru
[2] Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, oshestakov@cs.msu.su

Tomographic image reconstruction problem using wavelet-vaguelette decomposition is considered. Consistency and asymptotic normality of risk estimate of vaguelette coefficients thresholding are studied.

Keywords: wavelets; tomography; thresholding; risk estimate; limit distribution

ANALYSIS OF A LINK PROTOCOL WITH A GENERAL CONTENTION WINDOW BACKOFF FUNCTION

A. S. Lukyanenko[1], E. V. Morozov[2], and A. Gurtov[3]

[1] Helsinki Institute for Information Technology HIIT, Aalto, Finland, firstname.secondname@hiit.fi
[2] Institute of Applied Mathematical Research, Karelian Research Centre RAS; Petrozavodsk State University, emorozov@krc.karelia.ru
[3] Helsinki Institute for Information Technology HIIT, Aalto, Finland, gurtov@hiit.fi

A set of medium access (backoff) protocols, where the collision resolution window for a station depends on the number of successive collisions, is analyzed. Under mild common assumptions for the network properties, a general backoff protocol is studied. An optimal criterion yields a backoff protocol possessing the minimal service time. An asymptotic analysis is offered for the unboundly growing number of stations. Bounded and unbounded protocols are considered in the analysis. Finally, a continuous time model is introduced as an extension for the slotted model, which allows slots of different sizes.

Keywords: data communications; performance analysis; backoff protocol; stability; medium access control

DEVELOPMENT OF PARALLEL HEURISTIC ALGORITHMS OF WEIGHTS COEFFICIENTS SELECTION FOR ARTIFICIAL NEURAL NETWORK

O. V. Kryuchin

G. R. Derzhavin Tambov State University, kryuchov@gmail.com

A gradients alogirithm of artificial neural network and heuristic algoritms QuickProp and RProp which are based on it are described. Possible applications of cluster systems have been considered.

Keywords: artificial neural network; heuristic algorithms of teaching; cluster systesm

APPLICATION OF THE COORDINATE METHOD OF COMMUTATED NEURAL NETWORK FRAGMENTATION FOR TRAFFIC REDUCTION

S. Y. Stepanov

Moscow State Technological Institute STANKIN, cympak_shade@rambler.ru

The problem of increasing traffic in scaling commutated neural network, a method and algorithm of its solution are outlined. An example of the developed algorithm is also presented.

Keywords: commutated neural network; scaling; traffic

ON ASYMPTOTIC BEHAVIOR OF THE POWERS OF THE TESTS FOR THE CASE OF LAPLACE DISTRIBUTION

V. E. Bening[1] and R. A. Korolev[2]

[1] Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, bening@yandex.ru

[2] Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, stochastique@gmail.com

A formula for the limit of the normalized difference between the power of the asymptotically most powerful test and the power of the asymptotically optimal test for the case of Laplace distribution was proved. Due to the nonregularity of the Laplace distribution, the logarithm of the likelihood ratio admits nonregular stochastic expansion, and an analog of Cramér condition is not valid for the sign statistic which is the basis of the asymptotically optimal test. Then direct use of theorem 3.2.1 from [1] or theorem 2.1 from [2] is difficult, and in the present paper, their proofs for the case of Laplace distribution are revisited.

Keywords: power function; conditional probability measure; conditional moment; Laplace distribution

AN IMPROVEMENT OF THE KATZ–BERRY–ESSEEN INEQUALITY

M. E. Grigorieva[1] and I. G. Shevtsova[2]

[1] Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, maria-grigorieva@yandex.su

[2] Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, ishevtsova@cs.msu.su

The upper estimates of the absolute constant in the Katz–Berry–Esseen inequality for sums of independent identically distributed random variables with finite absolute moments of order between 2 and 3 are sharpened and an alternative inequality with sharpened structure and evaluated constants is proposed.

Keywords: central limit theorem; Katz–Berry–Esseen inequality; Lyapounov fraction

LINGUISTIC FILTERS IN STATISTICAL MACHINE TRANSLATION MODELS

E. B. Kozerenko

IPI RAN, kozerenko@mail.ru

The paper focuses on the problems of linguistic filters development for statistical machine translation and advancement of parallel texts alignment methods for enhancing precision and adequacy of translations. Statistical and heuristic models of alignment and translation are considered. The solutions proposed are based on the hybrid grammar formalism comprising the linguistic rules and probability characteristics of language structures.

Keywords: statistical models; machine translation; parallel texts; alignment; linguistic filters

Об авторах

Бенинг Владимир Евгеньевич (р. 1954) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; старший научный сотрудник ИПИ РАН

Григорьева Мария Евгеньевна (р. 1986) — аспирантка кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Гуртов Андрей Валерьевич (р. 1979) — доктор философии, ведущий научный сотрудник Хельсинкского института информационных технологий; профессор университета Оулу, Финляндия

Козеренко Елена Борисовна (р. 1959) — кандидат филологических наук, заведующая лабораторией компьютерной лингвистики и когнитивных технологий обработки текстов ИПИ РАН

Коновалов Михаил Григорьевич (р. 1950) — доктор технических наук, заведующий сектором ИПИ РАН

Королев Роман Анатольевич (р. 1977) — аспирант факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова; ассистент, Российский университет дружбы народов

Кривенко Михаил Петрович (р. 1946) — доктор технических наук, профессор, ведущий научный сотрудник ИПИ РАН

Крючин Олег Владимирович (р. 1986) — аспирант кафедры «Компьютерное и математическое моделирование» Тамбовского государственного университета им. Г. Р. Державина

Лукьяненко Андрей Сергеевич (р. 1982) — аспирант, научный сотрудник Хельсинкского института информационных технологий, Финляндия

Маркин Артём Васильевич (р. 1985) — аспирант кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Морозов Евсей Викторович (р. 1947) — доктор физико-математических наук, ведущий научный сотрудник Института прикладных математических исследований Карельского Научного центра РАН; профессор Петрозаводского государственного университета

Рудаков Константин Владимирович (р. 1954) — доктор физико-математических наук, член-корреспондент РАН, заведующий отделом вычислительных методов прогнозирования Вычислительного центра РАН имени А. А. Дородницына, заведующий кафедрой «Интеллектуальные системы» МФТИ

Степанов Сергей Юрьевич (р. 1984) — аспирант кафедры «Компьютерные системы управления», ГОУ ВПО МГТУ «Станкин»

Торшин Иван Юрьевич (р. 1972) — кандидат химических наук, ведущий научный сотрудник Российского отделения Института микроэлементов ЮНЕСКО, сотрудник ООО «Центр систем прогнозирования и распознавания»

Шевцова Ирина Геннадьевна (р. 1983) — кандидат физико-математических наук; ассистент кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

Шестаков Олег Владимирович (р. 1976) — кандидат физико-математических наук, старший преподаватель кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова

About Authors

Bening Vladimir E. (b. 1954) — Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University

Grigorieva Maria E. (b. 1986) — PhD student; Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University

Gurtov Andrei V. (b. 1979) — PhD, Principal Scientist at Helsinki Institute for Information Technology; Professor at University of Oulu, Finland

Konovalov Mikhail G. (b. 1950) — Doctor of Science in technology, Head of Laboratory, Institute of Informatics Problems, Russian Academy of Sciences

Korolev Roman A. (b. 1977) — PhD student, M. V. Lomonosov Moscow State University; assistant, Peoples' Friendship University of Russia

Kozerenko Elena B. (b. 1959) — Candidate of Science (PhD) in linguistics, Head of Laboratory for Computational Linguistics and Cognitive Text Processing Technologies, Institute of Informatics Problems, Russian Academy of Sciences

Krivenko Mikhail P. (b. 1946) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Russian Academy of Sciences

Kryuchin Oleg V. (b. 1986) — postgraduate student, G. R. Derzhavin Tambov State University, Department of Computer and Mathematical Simulation

Lukyanenko Andrey S.(b. 1982) — PhD student, researcher at Helsinki Institute for Information Technology, Finland

Markin Artem V. (b. 1985) — PhD student, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University

Morozov Evsey V. (b. 1947) — Doctor of Science in physics and mathematics, leading researcher, Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences; professor, Petrozavodsk State University

Rudakov Konstantin V. (b. 1954) — Doctor of Science in physics and mathematics, Corresponding Member of the Russian Academy of Sciences, Head of Department of Computational Methods of Prediction at Computing Centre of the Russian Academy of Sciences, Head of Department of Intelligent Systems at Moscow Institute of Physics and Technology

Shestakov Oleg V. (b. 1976) — Candidate of Science (PhD) in physics and mathematics, senior lecturer, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University

Shevtsova Irina G. (b. 1983) — PhD in physics and mathematics; assistant professor; Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University

Stepanov Sergey Yu. (b. 1984) — postgraduate student, Moscow State Technological Institute STANKIN

Torshin Ivan Yu. (b. 1972) — PhD, leading researcher at the Russian Center of the Trace Elements for UNESCO, researcher at "Center of Forecasting Systems and Recognition," LLC

Правила подготовки рукописей статей 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. 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 результаты, ранее не опубликованные и не предназначенные к одновременной публикации в других изданиях. Публикация не должна нарушать закон об авторских правах. Направляя свою рукопись в редакцию, авторы автоматически передают учредителям и редколлегии неисключительные права на издание данной статьи на русском языке и на ее распространение в России и за рубежом. При этом за авторами сохраняются все права как собственников данной рукописи. В связи с этим авторами должно быть представлено в редакцию письмо в следующей форме: Соглашение о передаче права на публикацию:
   «Мы, нижеподписавшиеся, авторы рукописи « », передаем учредителям и редколлегии журнала «Информатика и её применения» неисключительное право опубликовать данную рукопись статьи на русском языке как в печатной, так и в электронной версиях журнала. Мы подтверждаем, что данная публикация не нарушает авторского права других лиц или организаций. Подписи авторов: (ф. и. о., дата, адрес)».
   Редколлегия вправе запросить у авторов экспертное заключение о возможности опубликования представленной статьи в открытой печати.
2. Статья подписывается всеми авторами. На отдельном листе представляются данные автора (или всех авторов): фамилия, полные имя и отчество, телефон, факс, e-mail, почтовый адрес. Если работа выполнена несколькими авторами, указывается фамилия одного из них, ответственного за переписку с редакцией.
3. Редакция журнала осуществляет самостоятельную экспертизу присланных статей. Возвращение рукописи на доработку не означает, что статья уже принята к печати. Доработанный вариант с ответом на замечания рецензента необходимо прислать в редакцию.
4. Решение редакционной коллегии о принятии статьи к печати или ее отклонении сообщается авторам. Редколлегия не обязуется направлять рецензию авторам отклоненной статьи.
5. Корректура статей высылается авторам для просмотра. Редакция просит авторов присылать свои замечания в кратчайшие сроки.
6. 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 25 страниц указанного формата. Ознакомиться с шаблонами, содержащими примеры оформления, можно по адресу в Интернете: http://www.ipiran.ru/journal/template.doc.
7. К рукописи, предоставляемой в 2-х экземплярах, обязательно прилагается электронная версия статьи (как правило, в форматах MS WORD (.doc) или LaTex (.tex), а также — дополнительно — в формате .pdf) на дискете, лазерном диске или по электронной почте. Сокращения слов, кроме стандартных, не применяются. Все страницы рукописи должны быть пронумерованы.
8. Статья должна содержать следующую информацию на русском и английском языках: название, Ф.И.О. авторов, места работы авторов и их электронные адреса, аннотация (не более 100 слов), ключевые слова. Ссылки на литературу в тексте статьи нумеруются (в квадратных скобках) и располагаются в порядке их первого упоминания. Все фамилии авторов, заглавия статей, названия книг, конференций и т. п. даются на языке оригинала, если этот язык использует кириллический или латинский алфавит.
9. Присланные в редакцию материалы авторам не возвращаются.
10. При отправке файлов по электронной почте просим придерживаться следующих правил:
    - указывать в поле subject (тема) название журнала и фамилию автора;
    - использовать attach (присоединение);
    - в случае больших объемов информации возможно использование общеизвестных архиваторов (ZIP, RAR);
    - в состав электронной версии статьи должны входить: файл, содержащий текст статьи, и файл(ы), содержащий(е) иллюстрации.
11. Журнал «Информатика и её применения» является некоммерческим изданием, и гонорар авторам не выплачивается.

Адрес редакции: Москва 119333, ул. Вавилова, д. 44, корп. 2, ИПИ РАН
Тел.: +7 (499) 135-86-92 Факс: +7 (495) 930-45-05 E-mail: rust@ipiran.ru