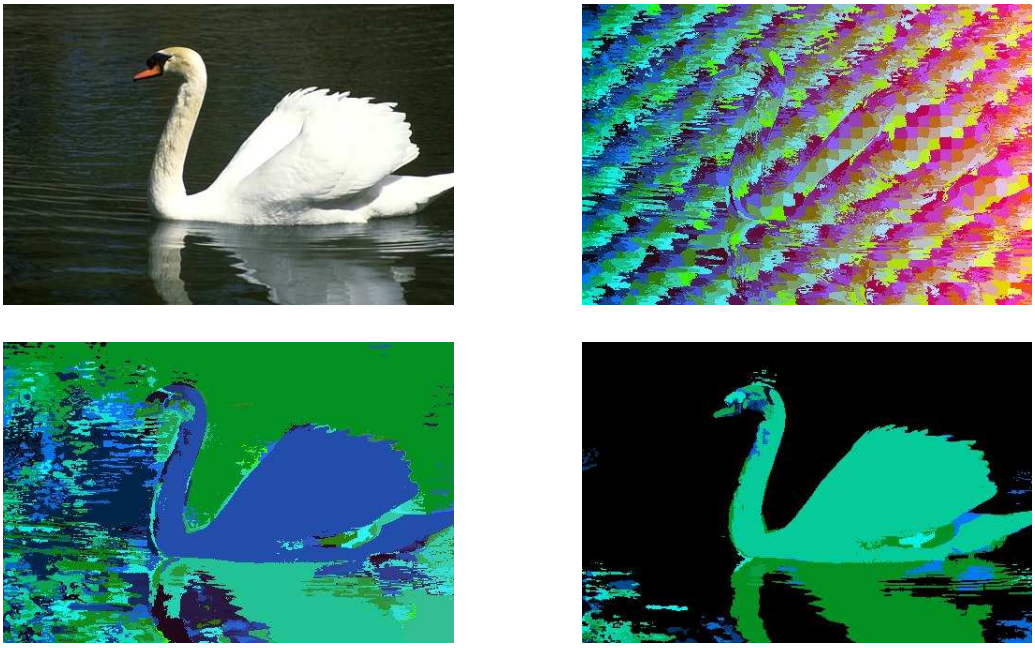
ISSN 2223-3792

Машинное обучение и анализ данных

2016 год **Том 2, номер 2**



Машинное обучение и анализ данных

Журнал «Машинное обучение и анализ данных» публикует новые теоретические и обзорные статьи с результатами научных исследований в области искусственного интеллекта, теоретических основ информатики и приложений. Цель журнала — развитие теории машинного обучения, интеллектуального анализа данных и методов проведения вычислительных экспериментов. Принимаются статьи на английском и русском языках.

Журнал включен в российский индекс научного цитирования РИНЦ. Информация о цитировании статей находится на сайте Российского индекса научного цитирования, ISSN 2223-3792, номер свидетельства о регистрации ЭЛ № ФС 77-55486. Журнал зарегистрирован в системе Crossref, doi `http://dx.doi.org/10.21469/22233792`.

- Новостной сайт http://jmlda.org/
- Электронная система подачи статей http://jmlda.org/papers/
- Правила подготовки статей http://jmlda.org/papers/doc/authors-guide.pdf

Тематика журнала:
— классификация, кластеризация, регрессионный анализ;
— алгебраический подход к проблеме синтеза корректных алгоритмов;
— многомерный статистический анализ;
— выбор моделей и сложность;
— глубокое обучение;
— статистическая теория обучения;
— методы прогнозирования временных рядов;
— методы обработки и распознавания сигналов;
— методы оптимизации в задачах машинного обучения и анализа данных;
— методы анализа больших данных;
— методы визуализации данных;
— обработка и распознавание речи и изображений;
— анализ и понимание текста;
— информационный поиск;
— прикладные задачи анализа данных.

Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Москва, 2016

Journal of Machine Learning and Data Analysis

The journal Machine Learning and Data Analysis publishes original research papers and reviews of the developments in the field of artificial intelligence, theoretical computer science and its applications. The journal aims to promote the theory of machine learning and data mining and methods of conducting computational experiments. Papers are accepted in English and Russian.

The journal is included in the Russian science citation index RSCI. Information about citation to articles can be found at the Russian science citation index website. ISSN 2223-3792. Mass media registration certificate ЭЛ № ФС 77-55486. The Crossref journal doi is http://dx.doi.org/10.21469/22233792.

- Journal news and archive http://jmlda.org/
- Open journal system for papers submission http://jmlda.org/papers/
- Style guide for authors http://jmlda.org/papers/doc/authors-guide.pdf

The scope of the journal:
— classification, clustering, regression analysis;
— multidimensional statistical analysis;
— Bayesian methods for regression and classification;
— model selection and complexity;
— deep learning;
— Statistical Learning Theory;
— time series forecasting techniques;
— methods of signal processing and speech recognition;
— optimization methods for solving machine learning and data mining problems;
— methods of big data analysis;
— data visualization techniques;
— methods of image processing and recognition;
— text analysis, text mining and information retrieval;
— applied data analysis problems.

Dorodnicyn Computing Centre FRC CSC RAS
Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics
Division "Intelligent Systems"

Moscow, 2016

Содержание

Contents

Группировка признаков на основе оптимальной последовательности миноров корреляционной матрицы*

С. Д. Двоенко, Д. О. Пшеничный

dsd@tsu.tula.ru; denispshenichny@yandex.ru

[1]Тульский государственный университет, Россия, г. Тула, пр. Ленина, 92

При решении задачи группировки возникает проблема содержательной интерпретации полученных факторов и групп признаков. Тем не менее факторы групп является синтетическими признаками, интерпретация которых может быть затруднена, поэтому после выделения групп признаков и построения соответствующих им факторов в каждой группе обычно определяется ее представитель как наиболее сильно коррелирующий с фактором группы признак. Тогда оказывается возможным содержательно интерпретировать результат группировки прямо в терминах исходных признаков. Предложен новый подход для выбора подмножества признаков, способных адекватно представить скрытые факторы, без определения собственных или центроидных направлений в качестве промежуточных преобразований. Данный подход основан на построении оптимальной последовательности значений главных миноров корреляционной матрицы признаков. В начале такой оптимальной последовательности расположены наименее коррелированные друг с другом и с остальными признаки, а к ее концу выстраиваются все более коррелированные с остальными признаки, выбранные в последнюю очередь. Показано, что предложенный подход позволяет формировать начальное решение для других алгоритмов группировки и также может применяться самостоятельно для оценки числа групп и построения содержательных группировок.

Ключевые слова: *группировка; кластер; метрика; корреляция; собственное число; собственный вектор; детерминант*

DOI: 10.21469/22233792.2.2.1

1 Введение

Считается, что задачи анализа данных в первую очередь возникают на ранних этапах исследования изучаемого явления, когда еще не построена его модель и еще рано говорить о задаче ее идентификации. В этом случае необходимо накопить и изучить как можно больше разнородной информации о наиболее существенных свойствах изучаемого явления. Вынужденность и противоречивость такого подхода вполне очевидна: вообще говоря, неизвестно, какие свойства наиболее существенны, и, как следствие, неизвестно, какие сведения накапливать.

Таким образом, интеллектуальные методы анализа данных должны устранить указанное выше противоречие, сконцентрировав в описании изучаемого явления наиболее существенную и адекватную информацию о нем. В основе такого подхода лежат достаточно естественные предположения, сформулированные в виде так называемых гипотез. Таких гипотез, по сути, всего две: компактности и скрытых факторов.

*Работа выполнена при частичной финансовой поддержке РФФИ, проекты №№ 15-07-02228, 15-07-08967, 14-07-00527 и 14-07-00964.

В интеллектуальном анализе данных обычно предполагается, что экспериментальные сведения об изучаемом явлении представлены как результаты измерений в виде матрицы данных $X(N, n)$, где N — число измерений; n — число измеряемых характеристик. Каждый акт измерения характеристик изучаемого явления рассматривается как объект $\omega_i \in \Omega$, который процессом измерения помещен в n-мерное признаковое пространство и представлен в нем вектором-строкой $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$, $i = 1, \ldots, N$. Матрица данных представляет собой множество из N строк $X(N, n) = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$, расположенных друг под другом.

Согласно гипотезе компактности, предполагается, что объекты образуют K локальных сгущений (классы, кластеры, таксоны), которые следует выделить (отделить друг от друга), так как они, предположительно, характеризуют различные состояния изучаемого явления.

С другой стороны, совокупность измерений одной характеристики образует вариационный ряд, т. е. признак, представленный наблюдениями $X_j = (x_{1j}, \ldots, x_{Nj})^{\mathrm{T}}$. Тогда матрица данных представляет собой множество из n вариационных рядов-столбцов $X(N, n) = (X_1, \ldots, X_n)$.

Согласно гипотезе скрытых факторов, считается, что их поведение определяет соответствующие «глубинные» свойства объекта исследования, которые проявляются через измеренные признаки как его реакции на внешние воздействия. Факторы проявляются через измеряемые признаки и различным образом влияют на эти признаки. Зависимость признаков от некоторого фактора определяет похожесть их поведения, т. е. похожесть изменений значений соответствующих вариационных рядов. Предполагается, что существует L таких факторов F_i, которым должны соответствовать группы признаков G_i, $i = 1, \ldots, L$.

Очевидно, что объективная закономерность, скрыто присутствующая в изучаемом явлении, обязательно проявится в результатах обработки различными методами и алгоритмами, основанными на различных предположениях о нем. Комплексирование таких сведений позволит в итоге адекватно решить поставленную задачу [1].

Таким образом, необходимо поддерживать и расширять разнообразие интеллектуальных методов обработки данных. В данной работе предпринята такая попытка. Актуальность таких попыток вполне очевидна, особенно в связи с накоплением больших объемов экспериментальных данных и развитием методов обработки данных, представленных парными сравнениями.

2 Задача группировки признаков

Задача группировки признаков имеет самостоятельное значение и может решаться разными способами.

Относительно факторов делается важное предположение, что в идеале они независимы. Статистический смысл независимости факторов означает, что соответствующие вариационные ряды наблюдений, будучи построенными, окажутся некоррелированными. Это означает, что такие скрытые признаки могут быть представлены наблюдениями $F_i = (f_{1i}, \ldots, f_{Ni})^T$, $i = 1, \ldots, L$, которые в соответствующем пространстве формируют систему ортогональных векторов.

Если сначала определяются факторы, то потом определяются признаки, подверженные их влиянию в наибольшей степени (задача факторного анализа и проблема вращения для определения факторных нагрузок и получения так называемой «простой» факторной структуры).

Известная проблема заключается в том, что ортогональное вращение факторов не совсем адекватно решает проблему получения простой структуры, поэтому приходится применять косоугольное вращение, что усложняет модель факторного анализа, так как факторы уже не являются независимыми [2, 3].

Если же сначала выделять группы сильно коррелирующих признаков, где признаки из разных групп почти не коррелируют, то потом можно построить представляющие эти группы факторы. При таком решении, в частности, проблема простой факторной структуры для косоугольной системы факторов решается автоматически, хотя сами факторы несколько отличаются от классической факторной модели. В этом случае решается, например, известная задача экстремальной группировки [4]. Следует отметить, что данная задача решается также и для центроидных направлений.

Отметим, что в обоих вариантах задачи группировки возникает проблема содержательной интерпретации полученных факторов или соответствующих групп признаков. Опыт показывает, что признаки, объединяемые в группы, часто можно совместно содержательно интерпретировать. С другой стороны, фактор группы все-таки является синтетическим признаком, интерпретация которого может быть затруднена, поэтому часто применяется следующий прием.

После выделения групп признаков и построения соответствующих им факторов в каждой группе определяется так называемый «представитель» группы как наиболее сильно коррелирующий с фактором группы признак. Далее рассматривается только множество таких признаков-представителей.

Очевидно, что в этом случае задача группировки также решает и другую известную задачу сокращения размерности признакового пространства. Эта задача также имеет самостоятельное значение. В данном случае получается сокращенное и содержательно интерпретируемое признаковое пространство. Важное свойство такого подпространства очевидно: эти реальные признаки коррелируют между собой в наименьшей степени и лучше всего могут представить скрытые факторы. Совсем упрощая, их даже часто рассматривают как факторы.

Легко увидеть, что при таком подходе все преобразования, выполняемые в соответствии с факторной моделью, являются промежуточными, так как в итоге выбираются некоторые исходные признаки.

Можно ли предложить другой подход, который позволит выбрать подмножество из исходных признаков, обладающих аналогичными свойствами, не требуя построения собственных направлений (а также и центроидных) в качестве промежуточного этапа преобразований? Ниже рассмотрен один из возможных подходов.

3 Метричность конфигурации элементов и ее нарушения

Следует отметить, что задача группировки (выделения факторов) решается для матрицы взвешенных скалярных произведений признаков X_j, $j = 1, \ldots, n$, т. е. для матрицы $R(n, n)$ корреляций вариационных рядов наблюдений. Для определения свойств факторов сами наблюдения $X(N, n)$ уже не нужны. Именно поэтому в факторном анализе оценка значений факторов как восстановленных наблюдений является отдельной и дополнительной задачей.

Это замечание особенно актуально в связи с развитием современных подходов, опирающихся на данные об объектах исследования, представленных только лишь в виде парных сравнений. В этом случае предполагается, что существует гипотетическая система признаков или что реальные признаки существуют, но для измерения уже недоступны.

Считается, что от измеренных признаков остались лишь матрица расстояний $D(N,N)$ или скалярных произведений $C(N,N)$ между объектами и матрица корреляций $R(n,n)$ между признаками.

Развитие этих методов показывает, что нужно обеспечить вложенность экспериментальных наблюдений в соответствующее метрическое (евклидово) пространство признаков и предложить модификации алгоритмов кластер-анализа и группировки, не требующих явного наличия матрицы данных X.

Проблема метричности конфигурации элементов известна и рассматривается, например, в задаче шкалирования [5]. Ее конечной целью является восстановление хорошо интерпретируемых признаков в явном виде как представленных соответствующими измерениями. Если этого не требуется, то известные задачи кластеризации и группировки можно решить и без непосредственного восстановления собственно значений признаков. В частности, такой подход позволяет для решения задач кластеризации и группировки применять одни и те же алгоритмы, рассматривая объекты или признаки просто как элементы множества, погруженные в соответствующее метрическое пространство [6].

Если элементами множества являются признаки, то исходя из смысла похожести вариационных рядов рассматривают модули или квадраты коэффициентов корреляций в матрице $R(n,n)$. Кроме того, если изначально рассматривается некоторая функция парных сравнений, имеющая смысл близости $s_{ij} \geqslant 0$; $i,j = 1,\ldots,n$, то ее можно рассматривать как положительные вариации (или корреляции, если они нормированы).

На практике часто в полученных конфигурациях элементов имеются метрические нарушения. Причины этого различны. Поэтому одной из актуальных задач современного анализа данных является восстановление метричности данных. Именно в этом случае применение упомянутых выше алгоритмов является математически корректным.

Известно, что нарушения метричности конфигураций приводят к появлению отрицательных собственных чисел в матрице скалярных произведений между элементами множества. В случае множества признаков это относится к матрице корреляций $R(n,n)$. Если ее собственные числа упорядочить по убыванию $\lambda_1 > \cdots > \lambda_n$, то можно считать, что пространства размерностей, соответствующих отрицательным собственным числам, не существуют в том смысле, что в них для наблюдений не выполняется, например, теорема Пифагора или, в общем случае, теорема о косинусах, могут быть нарушены неравенства треугольника и т. д. Это ведет к тому, что интуитивно понятные аналогии нашему обычному трехмерному пространству, полезные в анализе данных, становятся некорректными. Но тогда и результаты обработки, вообще говоря, следует признать недостаточно корректными, где уровень некорректности определяется математической некорректностью результата.

Как известно, величина дисперсии данных — это размерность n пространства признаков. Для устранения в матрице $R(n,n)$ отрицательных собственных чисел обычно можно применить известное дискретное разложение Карунена–Лоэва, а именно: из всех элементов матрицы $R(n,n)$ «послойно» исключить вклады собственных векторов (направлений), соответствующих отрицательным собственным числам (также будем говорить, что это — «вклады» собственных чисел).

Заметим, что в факторном анализе матрица так называемых «остаточных» корреляций $R_q(n,n)$ определяется после послойного устранения вкладов первых q собственных векторов, соответствующих собственным числам, упорядоченным по убыванию. Естественно, что $\det R_q(n,n) = 0$.

Удобно также применить этот термин к результату устранения вкладов q отрицательных собственных чисел, которые оказываются последними в упорядочении. Очевидно, у такой матрицы остаточных корреляций $R_{q-}(n,n)$ также $\det R_{q-}(n,n)=0$.

Легко увидеть, что после устранения вкладов q отрицательных собственных чисел матрица остатков $R_{q-}(n,n)$ становится ненормированной (и более того, некорректной), где $r_{ii}>1,\ \ i=1,\ldots,n$. Но тогда, строго говоря, в данных «ниоткуда» появляется добавочная дисперсия, так как $\sum_{i=1}^{n} r_{ii}>n$. Формально из $R_{q-}(n,n)$ можно получить корректную корреляционную матрицу с единичной главной диагональю, просто пронормировав ее.

Очевидно, что нормировка в этом случае уничтожает сведения о доле внесенной дисперсии, поэтому в общем случае такой процесс появления новой дисперсии в данных после нормировок матрицы корреляций уже невозможно проконтролировать. Это нежелательно, например, когда решается задача группировки признаков.

По-видимому, эта проблема не столь принципиальна при наличии признакового пространства, т. е. матрицы данных $X(N,n)$. В этом случае можно построить матрицу так называемых «вычисленных признаков» $Y(N,m)$, где $m=n-q<n$ и $\lambda_1>\cdots>\lambda_m>0$, как проекций векторов-объектов из X на m первых собственных направлений. В пространстве вычисленных признаков наблюдения-строки $\mathbf{y}_i=(y_{i1},\ldots,y_{i\,m})$ образуют метрическую конфигурацию, что позволяет корректно решать задачи группировки, кластеризации, визуализации и т. д.

С другой стороны, в линейной факторной модели существует известная проблема определения общностей (вкладов общих факторов в дисперсию данных). Например, в методе главных факторов после редукции $R(n,n)$ с целью устранения дисперсий характерных факторов редуцированная матрица $\bar{R}(n,n)$ оказывается ненормированной, так как $r_{ii}<1,\ \ i=1,\ldots,n$, из-за уменьшенных значений ее диагональных элементов. Это известная в факторном анализе проблема определения общностей, теоретического решения которой не предлагалось. Есть лишь эмпирические рекомендации по оценке величины общностей.

Здесь следует отметить следующее. Следование эмпирическим рекомендациям часто приводит к появлению отрицательных собственных чисел в редуцированной матрице $\bar{R}(n,n)$, т. е. к нарушению метричности конфигурации элементов множества. Чтобы избежать этого, при построении главных факторов придется лишь «слегка» редуцировать диагональные элементы корреляционной матрицы, обычно в значительно меньшей степени, чем по эмпирическим рекомендациям. Вообще-то, это означает, что доля дисперсии в данных, объясняемая общими факторами, очень высока. Как в этом случае интерпретировать соотношения общностей и характерностей с точки зрения факторной модели — это другая проблема.

Следует также отметить, что проблема общностей возникает и при построении центроидных факторов. Центроидные направления отличаются от собственных направлений, но эмпирический принцип выбора общностей также приводит к появлению отрицательных собственных чисел, т. е. к нарушению метричности конфигурации элементов множества.

4 Оптимальная последовательность признаков и выбор их подмножества

В отличие от процедуры Карунена–Лоэва авторами был предложен другой метод так называемой «индивидуальной» корректировки лишь некоторых (или всех) парных сравнений некоторых элементов множества с остальными элементами для восстановления нарушенной метрической конфигурации, при котором сохраняется дисперсия данных [7, 8].

В данном методе наличие собственных чисел в матрице $S(n,n)$ взвешенных скалярных произведений, где $s_{ii}=1,\ i=1,\ldots,n$, связывается не с послойным ее разложением на вклады соответствующих собственных векторов (чисел), а с индивидуальными вкладами самих элементов множества. В качестве такой матрицы можно взять, например, матрицу $R(n,n)$ корреляций, модулей или квадратов корреляций признаков.

Пусть дана симметричная нормированная матрица $S(n,n)$. Согласно критерию Сильвестра [9], матрица квадратичной формы положительно определена, если все ее главные миноры $S_k=S(k,\ k),\ k=1,\ldots,n$, положительны: $\det S_k>0$, где $S_1=S(1,1)=s_{11}=1$. Согласно следствию из закона инерции Сильвестра, число q отрицательных собственных чисел совпадает с числом смен знаков детерминантов в последовательности $S_0=1,\ S_1,\ S_2,\ldots,S_n=S(n,n)$. Легко увидеть, что значения главных миноров (их детерминанты) в нормированной S убывают, начиная с единицы. При наличии отрицательных собственных чисел последовательность главных миноров оказывается знакопеременной, где значения главных миноров постепенно уменьшаются по модулю.

Известно, что одновременная перестановка двух строк и двух соответствующих столбцов в S не изменяет ее собственных чисел. Такая перестановка соответствует перестановке двух элементов множества. Определим такой порядок элементов множества, чтобы смены знаков значений главных миноров в последовательности $S_k,\ k=1,\ldots,n$, происходили в ее конце. Если матрица $S(n,n)$ ранга n имеет q отрицательных собственных чисел, то тогда в идеальном случае главный минор S_{n-q+1} впервые окажется отрицательным: $\det S_{n-q+1}<0$, а знаки последующих миноров будут чередоваться.

Естественно считать, что именно в этот момент: $k=n-q+1$ очередной элемент множества ω_k, представленный своими парными сравнениями $s_{ki}=s_{ik},\ i=1,\ldots,n$, с остальными, внес метрическое нарушение в уже построенную конфигурацию. Нарушение можно устранить одним из предложенных нами ранее способов коррекции его парных сравнений, получив положительное значение текущего главного минора S_k [7, 8]. Следующий минор S_{k+1} снова окажется отрицательным и потребует исправления. Всего потребуется скорректировать парные сравнения для q элементов множества. В этом смысле отрицательные собственные числа оказываются связанными с конкретными элементами множества или, другими словами, оказываются «локализованными» в матрице парных сравнений.

Рассмотрим процедуру, которая позволит получить оптимальную последовательность элементов множества. Известно, что определитель матрицы $S(n,n)$ равен произведению ее собственных чисел. Если он отрицателен, то количество собственных чисел нечетно, если положителен, то четно.

Рассмотрим главные миноры $S_k,\ k=n,\ldots,1$, в обратном порядке. Определим в матрице S_k такую строку и столбец i, что значение дополнительного минора $(S_k)_i^i$, $1\leqslant i\leqslant k$, образованного при их удалении, сменит знак по сравнению с S_k и окажется максимальным по модулю. Если знак не изменяется, то просто найдем такой дополнительный минор без смены знака. Пусть u — общее число таких шагов без смены знака дополнительного минора до локализации всех q смен знаков главных миноров.

Последовательность поочередно отброшенных строк и столбцов формирует оптимальную последовательность главных миноров $S_k,\ k=1,\ldots,n$ (и элементов множества, последовательно формирующих текущие миноры), в которой впервые отрицательный минор встретится не ранее, чем в момент $n-q-u+1$. Это означает, что полученная перестановка формирует такую матрицу $S(n,n)$, у которой придется корректировать парные сравнения не более, чем у $q+u$ последних элементов множества в оптимальной последовательности.

В общем случае при неоптимальной последовательности элементов приходится корректировать значительно большее число элементов множества, так как каждая очередная коррекция обычно порождает шлейф дополнительных коррекций.

Легко увидеть, что при отсутствии метрических нарушений будет получена локально оптимальная последовательность главных миноров $S_k,\ k = 1, \ldots, n$, где их значения, оставаясь неотрицательными, убывают наиболее медленно (почти).

Рассмотрим матрицу корреляций $R(n, n)$. Можно заметить, что значение $\det R$ зависит от степени «ортогональности» конфигурации системы признаков: чем «ортогональнее» система признаков, тем ближе значение детерминанта к единице, и к нулю — в противном случае. Для $n = 2$ это очевидно, так как $\det R = 1 - r^2$. Для $n = 3$ в этом нетрудно убедиться, так как $\det R = 1 + 2r_{12}r_{13}r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2$, рассмотрев возможные значения парных коэффициентов корреляций, которые соответствуют конфигурациям без метрических нарушений, и т. д. С увеличением размерности n это эмпирическое свойство преимущественно сохраняется в целом, но, естественно, появляются возможности для взаимной компенсации достаточно высоких корреляций в усложняющихся формулах вычислений детерминантов, тем более для корреляций со знаками.

В этих условиях оказывается, что для метрически корректной матрицы $R(n, n)$ оптимальная последовательность главных миноров $S_k,\ k = 1, \ldots, n$, где $S_1 = 1$ и $S_n = R(n, n)$, определяет локально оптимальную последовательность вложенных подмножеств «наиболее ортогональных» признаков. В начале такой оптимальной последовательности расположены «наиболее ортогональные» друг к другу и к остальным признаки, а к концу последовательности выстраиваются все «менее ортогональные» к остальным признаки, выбранные в последнюю очередь.

Корреляционная матрица $R(n, n)$ имеет статистический смысл, поэтому будем говорить об оптимальной последовательности вложенных подмножеств наименее коррелированных признаков. Отсюда легко увидеть, что первые m признаков в оптимальной последовательности должны образовать наименее коррелирующее подмножество из всех признаков, которое содержательно удобно интерпретировать как множество представителей m групп признаков.

Таким образом, процедура построения локально-оптимальной последовательности признаков позволяет решить задачу группировки на m групп (редукции размерности) без построения собственных направлений (для квадратов корреляций) или без построения центроидных направлений (для модулей корреляций).

5 Проблема начального разбиения в алгоритмах группировки

При решении задач группировки авторами было показано, что известный алгоритм экстремальной группировки на модулях коэффициентов корреляций («модуль») эквивалентен алгоритму k-средних, который представлен в модифицированной форме для обработки близостей [6]. В свою очередь, было показано, что такая модификация эквивалента классическому алгоритму k-средних для матрицы данных X в том смысле, что «внезапное» погружение элементов множества в пространство признаков не изменит результат разбиения. Такая кластеризация решает задачу построения центроидных факторов в задаче экстремальной группировки признаков (алгоритм «модуль»).

В свою очередь, алгоритм экстремальной группировки на квадратах коэффициентов корреляций («квадрат») решает задачу построения первых главных компонент для каждой группы сильно коррелирующих признаков. Тем самым решается известная задача факторного анализа как задача построения главных компонент или главных факторов

для, соответственно, нередуцированной $R(n,n)$ или редуцированной $\bar{R}(n,n)$ матриц корреляций.

Как и все процедуры кластер-анализа и группировки, процедура построения оптимальной последовательности также является локальной. Эксперименты показывают, что локальность процедуры построения оптимальной последовательности того же свойства, что и у процедур кластер-анализа и группировки. В частности, ожидаемым свойством оптимальной последовательности признаков обычно является устойчивое выделение от двух до пяти наименее (почти) коррелирующих признаков.

Таким образом, в силу локальности свойств известных процедур экстремальной группировки процедура построения оптимальной последовательности имеет самостоятельное значение в задаче группировки признаков.

Известно, что в процедурах с локальными свойствами важной проблемой является поиск начального решения (разбиения). Например, в задаче экстремальной группировки считается, что центроидные решения являются хорошим началом для группировок по собственным направлениям. Поэтому оптимальная последовательность признаков может рассматриваться как другой способ получения начального решения для алгоритмов экстремальной группировки, которое для заданной матрицы $R(n,n)$ является единственным. Это свойство представляется нам наиболее интересным.

6 Эксперименты

6.1 Программа экспериментов

Пусть L — число групп признаков G_i, $i = 1,\ldots,L$, где $|G_i| = n_i$, $\sum_{i=1}^{L} n_i = n$ и $r(X_j,F_i)$ — корреляция фактора $F_i = (f_{1i},\ldots,f_{Ni})^{\mathrm{T}}$ с признаком $X_j = (x_{1j},\ldots,x_{Nj})^{\mathrm{T}}$. Для количественной оценки качества группировок рассмотрим известные критерии I_Q для алгоритма «квадрат» и I_M для алгоритма «модуль»:

$$I_Q = \sum_{i=1}^{L}\sum_{j\in G_i} r^2(X_j,F_i)\,; \quad I_M = \sum_{i=1}^{L}\sum_{j\in G_i} |r(X_j,F_i)|\,.$$

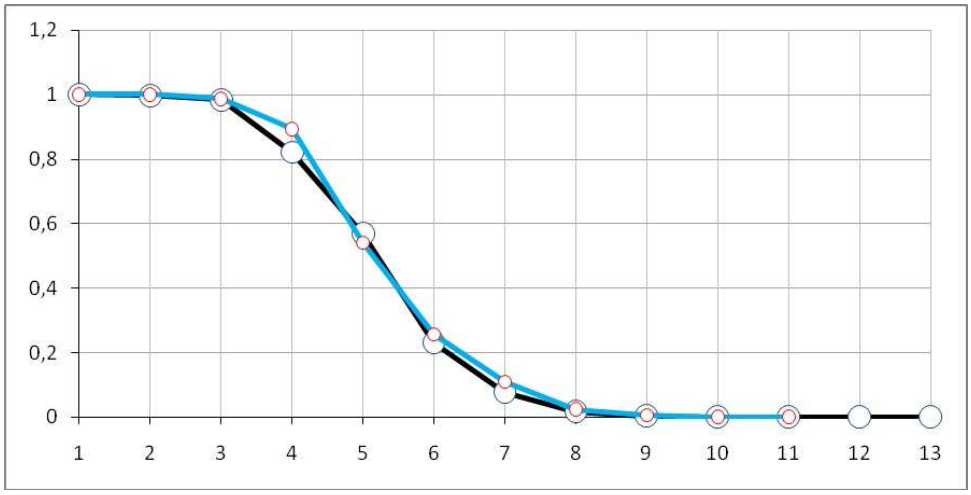
Алгоритмы экстремальной группировки имеют следующий вид.

Начальный шаг. Для L групп найти начальное разбиение каким-либо способом.
Шаг k.

1. В каждой группе G_i, $i = 1,\ldots,L$, образующей подматрицу $R(n_i,n_i)$, $i = 1,\ldots,L$, построить фактор $F_i = (f_{1i},\ldots,f_{Ni})^{\mathrm{T}}$ как главный или центроидный фактор или как первую главную компоненту.
2. Просмотреть все признаки и перенести каждый из них в ту группу, с фактором которой он коррелирует сильнее всего: $X_j \in G_p$, если $s(X_j,F_p) > s(X_j,F_i)$, $i = 1,\ldots,L$. Здесь $s(X_j,F_p) = |r(X_j,F_i)|$ или $s(X_j,F_p) = r^2(X_j,F_i)$.
3. Повторять шаги 1 и 2 до тех пор, пока группы не перестанут изменяться.

В данной работе программа экспериментов была направлена на исследование свойств оптимальной последовательности признаков, представляющей как начальное решение для алгоритмов экстремальной группировки, так и применяемой самостоятельно.

Рассматривались следующие начальные решения: первые L признаков в оптимальной последовательности; L минимально коррелирующих признаков как явная классическая альтернатива им; просто первые L признаков; случайно отобранные L признаков. Эти начальные решения порождали начальные разбиения для алгоритмов «квадрат» и «модуль»



Оптимальные последовательности значений главных миноров корреляционных матриц: 13 экономических показателей и 11 электрокардиограмм

в смысле критериев I_Q и I_M. Очевидно, что первые два начальных решения являются хорошими, вторые два — плохими. Эксперименты это подтвердили, поэтому здесь далее рассматриваются результаты только для хороших начальных решений.

Определение числа кластеров K является хорошо известной проблемой кластер-анализа. Одним из известных эвристических приемов для его выбора является определение границы, начиная с которой убывание критерия кластеризации до нуля (средневзвешенная дисперсия кластеров) при изменении K от 1 до N резко замедляется, где N — число объектов. Аналогично рассуждают и в задаче группировки. Очевидно, что при изменении числа групп $L = 1, \ldots, n$ критерии группировки I_Q и I_M возрастают до n, где n — число признаков. В этом случае также рассматривают границу, начиная с которой возрастание этих критериев резко замедляется.

Рассмотрим оптимальную последовательность главных миноров $S_k,\ k = 1, \ldots, n$, где их значения, оставаясь неотрицательными (если потребовалось, то после коррекции), убывают наиболее медленно. Оказывается, что в этом случае график изменения их значений приобретает характерный вид (см., например, рисунок).

В этом случае аналогичное эвристическое предположение о числе групп признаков предполагает рассмотрение области резкого падения значений главных миноров.

Также следует предположить, что при оптимальном числе групп должны получиться хорошо содержательно интерпретируемые факторы, характеризующие их.

Здесь представлены результаты экспериментов с некоторыми массивами данных.

6.2 Экономические показатели

Первый массив представляет собой данные Организации экономического сотрудничества и развития (Organization for Economic Cooperation and Development, OECD) [10] из сводного отчета за 2013 г. (Fastbook Country Statistical Profiles, 2013 edition) по 13 экономическим показателям 13 стран мира: Австралия, Франция, Германия, Италия, Япония, Корея, Мексика, Турция, США, Китай, Индонезия, Россия, ЮАР. Представлены следующие показатели:

1. Валовой внутренний продукт (ВВП) на душу населения (долл.)
2. Рост реального ВВП (%).
3. Прибыль, полученная в сельском хозяйстве, охоте и лесном хозяйстве, рыбалке (%).

4. Прибыль, полученная в промышленности, включая энергетические отрасли (%).
5. Прибыль, полученная в оптовых и розничных продажах, отелях, ресторанах, ремонте, транспорте (%).
6. Прибыль, полученная в финансовом посредничестве, недвижимости, арендных и деловых услугах (%).
7. Реальная прибыль, полученная в сельском хозяйстве, охоте и лесном хозяйстве, рыбалке (%).
8. Реальная прибыль, полученная в промышленности, включая энергетические отрасли (%).
9. Реальная прибыль, полученная в оптовых и розничных продажах, отелях, ресторанах, ремонте, транспорте (%).
10. Реальная прибыль, полученная в финансовом посредничестве, недвижимости, арендных и деловых услугах (%).
11. Общее потребление энергии (ТВт-ч).
12. Электричество, производимое ядерной энергетикой (ТВт-ч).
13. Доля электричества, производимого ядерной энергетикой, от общего объема (%).

Значения показателей прибыли в различных сферах активности представлены как с учетом общего уровня цен (реальная прибыль) в результате процессов инфляции-дефляции, так и без учета общего уровня цен (прибыль). Другие показатели связаны с уровнем ВВП и потреблением энергии. Статистические связи между уровнем ВВП, прибылью и энергетическими затратами представлены корреляционной матрицей $R(13, 13)$.

Как было сказано выше, на рисунке видно, что для экономических показателей число предполагаемых групп составляет 4–5.

Оптимальная последовательность признаков, построенная по матрице квадратов корреляций экономических показателей, имеет вид: [8, 5, 13, 9, 11, 4, 7, 12, 3, 1, 10, 6, 2].

Оптимальная последовательность признаков, построенная по матрице модулей корреляций экономических показателей, имеет вид: [10, 8, 5, 13, 12, 4, 11, 9, 2, 7, 3, 6, 1].

Результаты группировок показаны в табл. 1 и 2. Для каждого числа групп показаны начальные решения как представители, выбранные по разным принципам (минимально коррелирующие признаки и первые признаки из оптимальной последовательности). Также показаны результирующие группы и их представители. Изменение хотя бы одного начального представителя после перегруппировки означает, что начальное разбиение было улучшено. Если представители не изменились, то начальное разбиение не улучшилось. Номера признаков-представителей выделены жирным шрифтом.

Таблица 1 Группировки экономических показателей по критерию I_Q

Число групп	Минимальная корреляция	Представители	Группы	Оптимальная последовательность	Представители	Группы
	7	7	**7** 8 13	8	2	**2 8**
3	11	11	5 **11** 12	5	5	**5** 11
	6	10	1 2 3 4 **6** 9 **10**	13	3	1 **3** 4 6 7 9 10 12 **13**
	8	8	**8**	8	8	**8**
	5	5	**5**	5	5	**5**
5	7	7	**7** 13	13	13	7 12 **13**
	6	10	1 2 3 4 **6** 9 **10**	9	10	1 2 3 4 6 **9 10**
	11	11	**11** 12	11	11	**11**

Таблица 2 Группировки экономических показателей по критерию I_M

Число групп	Минимальная корреляция	Представители	Группы	Оптимальная последовательность	Представители	Группы
3	6	10	1 2 3 4 **6** 9 **10**	10	1	**1** 2 3 4 6 9 **10** 12 13
	7	7	**7** 8 3	8	7	**7 8**
	11	11	5 **11** 12	5	5	**5** 11
4	6	10	1 2 3 4 **6** 9 **10**	10	10	1 2 3 4 6 9 **10**
	7	7	**7** 8 13	8	8	**8**
	5	5	**5**	5	5	**5** 11
	11	11	**11** 12	13	13	7 12 **13**
5	6	10	1 2 3 4 **6** 9 **10**	10	10	1 2 3 4 6 9 **10**
	8	8	**8**	8	8	**8**
	5	5	**5**	5	5	**5**
	7	7	**7** 13	13	7	7 **13**
	11	11	**11** 12	12	11	**11 12**

Рассмотрим табл. 1. Для разбиения на три группы признаков результат по критерию I_Q неудовлетворителен, а именно: для начального разбиения по минимальным корреляциям признаки 2 и 8 после перегруппировки попали в разные группы. В то же время для начального разбиения по оптимальной последовательности эти два признака после перегруппировки оказались вместе в одной отдельной группе. Такая же ситуация сохраняется и для четырех групп признаков (в таблице не показана).

Но для пяти групп признаков результаты двух группировок практически одинаковы, а именно: представители разных групп для обоих вариантов начального разбиения обязательно входят в состав разных групп и после перегруппировки. В частности, признаки 2 и 8 также входят в составы разных групп в обеих группировках. В обоих случаях начальные группировки были улучшены, причем начальные представители остались в своих группах, даже если для них после перегруппировки были выбраны новые представители. Сами результирующие группировки минимально отличаются друг от друга признаком 12.

Такой результат хорошо соответствует ранее сделанному формальному предположению о пяти группах экономических показателей. Состав полученных групп позволяет содержательно интерпретировать их следующим образом (в порядке перечисления в табл. 1 показаны признаки, присутствующие в составе соответствующих групп одновременно в обоих разбиениях):

1) прибыль в промышленности с учетом энергозатрат (8);
2) прибыль в торговых и транспортных услугах (5);
3) прибыль в производстве натуральной продукции с учетом энергозатрат (7, 13);
4) ВВП и прибыль во всех сферах активности (1, 2, 3, 4, 6, 9, 10);
5) общее потребление энергии (11).

Рассмотрим табл. 2. Для разбиения на три группы по критерию I_M результаты похожи в том смысле, что все представители разных групп находятся в разных группах до и после перегруппировки. Для разбиения на четыре группы результат неудовлетворителен, так как в одной группировке признаки 5 и 11 представляют разные группы, а в другой группировке признаки 5 и 11 располагаются вместе и образуют отдельную группу.

Для пяти групп признаков разбиения полностью совпадают там, где признак 12 находится в одной группе вместе с признаком 11. Таким образом, и в этом варианте под-

Таблица 3 Качество группировок экономических показателей по критерию I_Q

Число групп	Минимальные корреляции		Оптимальная последовательность	
	Начальное разбиение	Результат	Начальное разбиение	Результат
3	5,1537	7,1308	3,5312	6,4304
5	7,0055	8,8924	**7,1196**	8,8002

Таблица 4 Качество группировок экономических показателей по критерию I_M

Число групп	Минимальные корреляции		Оптимальная последовательность	
	Начальное разбиение	Результат	Начальное разбиение	Результат
3	9,0739	9,0739	8,5066	8,8205
4	9,7296	9,7296	9,9426	9,9426
5	10,521	10,521	**10,521**	10,521

твердилась ранее предложенная интерпретация групп признаков (порядок перечисления групп соответствует табл. 1).

Рассмотрим качество группировок. Таблица 3 показывает, что для пяти групп начальное разбиение, полученное по оптимальной последовательности признаков, лучше, чем по минимальным корреляциям. Этот результат имеет самостоятельное значение, если экстремальная группировка по критерию «квадрат» не применяется. С другой стороны, здесь продемонстрирована локальность рассматриваемых процедур, заключающаяся в том, что лучшее по качеству начальное решение (оптимальная последовательность) не всегда обеспечивает лучший результат.

Таблица 4 также показывает, что для 4 и 5 групп экстремальная группировка по критерию «модуль» не улучшила начальное разбиение. В этом случае разбиение, полученное по оптимальной последовательности признаков, также имеет самостоятельное значение, так как сразу формирует окончательную группировку. Отметим, что в данном случае оптимальная последовательность формирует множество признаков, наиболее адекватно соответствующих предположению о наименьшей коррелированности. В целом можно отметить более сложное поведение критерия I_Q.

6.3 Электроэнцефалограммы

Второй массив представляет собой электроэнцефалограммы (ЭЭГ) биоритмов головного мозга, представленные корреляционной матрицей статистических взаимосвязей между энергетическими свойствами биоритмов для 11 частот: три частоты (1–3) представляют тета-ритмы, две частоты (4, 5) представляют низкочастотные (НЧ) альфа-ритмы, две частоты (6, 7) представляют высокочастотные (ВЧ) альфа-ритмы, четыре частоты (8–11) представляют бета-ритмы. Электроэнцефаллограммы были получены В. Д. Небылицыным в ходе исследований по эффекту навязывания ритма при светослуховом воздействии [11] на испытуемых.

Для ЭЭГ биоритмов головного мозга получается аналогичный график (см. рисунок) изменения значений главных миноров корреляционной матрицы $R(11, 11)$ биоритмов в оптимальной последовательности. Оказалось, что число предполагаемых групп также со-

Таблица 5 Группировки ЭЭГ по критериям I_Q и I_M

Число групп	Минимальная корреляция	Представители	Группы	Оптимальная последовательность	Представители	Группы
4	7	6	**6 7**	7	6	**6 7**
	2	3	1 **2 3**	2	3	1 **2 3**
	4	4	**4** 5	4	4	**4** 5
	10	10	8 9 **10** 11	8	10	**8** 9 **10** 11

Таблица 6 Качество группировок ЭЭГ по критериям I_Q и I_M

Критерий	Число групп	Минимальные корреляции		Оптимальная последовательность	
		Начальное разбиение	Результат	Начальное разбиение	Результат
Квадрат	4	6,5210	7,8048	6,1681	7,8048
Модуль	4	9,0942	9,0942	**9,0942**	9,0942

ставляет 4–5. Также очевидно, что содержательно ЭЭГ биоритмов должны быть представлены четырьмя группами ритмов разных типов (тета-, НЧ альфа-, ВЧ альфа- и бета-).

Оптимальная последовательность признаков, построенная по матрице корреляций ЭЭГ, имеет вид: [7, 4, 8, 1, 3, 11, 5, 9, 6, 10, 2].

Оптимальная последовательность признаков, построенная по матрице для квадратов корреляций ЭЭГ, имеет вид: [7, 2, 4, 8, 1, 5, 11, 3, 6, 9, 10].

Оптимальная последовательность признаков, построенная по матрице для модулей корреляций ЭЭГ, имеет вид: [7, 2, 4, 8, 1, 5, 11, 3, 10, 6, 9].

Отметим, что оптимальные последовательности признаков во всех случаях сразу же выделяют в качестве представителей четырех групп по одной частоте каждого типа. В качестве представителей пятой и т. д. групп выделяются частоты уже ранее представленного типа. В соответствии со смыслом критериев I_Q и I_M использовались оптимальные последовательности признаков для квадратов и модулей корреляций.

Результаты группировок показаны в табл. 5. Разбиения на четыре группы при разных способах получения начальных разбиений, а также по обоим критериям оказались полностью идентичными. Здесь сразу же показаны результаты только для четырех групп, так как для меньшего числа групп они не соответствуют содержательному смыслу, а для большего числа групп просто происходит расщепление содержательных групп ритмов разных типов.

Рассмотрим качество группировок. Таблица 6 показывает, что для 4 групп группировка по критерию «модуль» не улучшила начальное разбиение. В этом случае разбиение, полученное по оптимальной последовательности признаков, наиболее адекватно соответствует предположению о наименьшей коррелированности и также имеет самостоятельное значение, так как сразу формирует окончательную группировку.

В свою очередь, группировки по критерию «квадрат» были улучшены для обоих видов начальных разбиений, сформировав одну и ту же группировку. Таким образом, разбиение, полученное по оптимальной последовательности признаков, также адекватно соответствует предположению о наименьшей коррелированности признаков в искомой группировке. И снова продемонстрирована локальность рассматриваемых процедур, заключающаяся

в том, что лучшее по качеству начальное решение (минимальные корреляции) не всегда обеспечивает лучший результат. В целом также можно отметить более сложное поведение критерия I_Q.

7 Заключение

При решении задачи группировки возникает проблема содержательной интерпретации полученных факторов и групп признаков. Признаки, объединенные в группы, обычно поддаются содержательной совместной интерпретации. Тем не менее факторы групп являются синтетическими признаками, интерпретация которых может быть затруднена.

Часто после выделения групп признаков и построения соответствующих им факторов в каждой группе определяется ее представитель как наиболее сильно коррелирующий с фактором группы признак. Если рассматривать только множество таких признаков-представителей, то оказывается возможным содержательно интерпретировать результат группировки прямо в терминах исходных признаков.

При таком подходе все преобразования, выполняемые в соответствии с факторной моделью, являются промежуточными, так как в итоге выбираются представители из исходных признаков.

В данной работе предложен подход, позволяющий выбрать подмножество из исходных признаков, способных адекватно представить скрытые факторы, не требуя построения собственных и центроидных направлений в качестве промежуточного этапа преобразований. Данный подход основан на построении оптимальной последовательности признаков. В начале такой оптимальной последовательности расположены наименее коррелированные друг с другом и с остальными признаки, а к концу последовательности выстраиваются все более коррелированные с остальными признаки, выбранные в последнюю очередь.

Показано, что предложенный подход позволяет формировать начальное решение для известных алгоритмов группировки и также может применяться самостоятельно для оценки числа групп и построения содержательных группировок.

Литература

[1] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, 1978. Т. 33. С. 5–68.

[2] *Lawley D.N., Maxwell A.E.* Factor analysis as a statistical method. — 2nd ed. — London: Butterworth, 1971. 117 p.

[3] *Harman H.H.* Modern factor analysis. — 3rd ed. — University of Chicago Press, 1976. 508 p.

[4] *Lumelskii V. Ya.* Parameter grouping on the basis of the square coupling matrix // Automat. Rem. Contr., 1970. No. 1. P. 117–127.

[5] *Cox T.F., Cox M.A.A.* Multidimensional scaling. — 2nd ed. — Chapman and Hall/CRC, 2000. 328 p.

[6] *Двоенко С. Д.* Кластеризация множества, описанного парными расстояниями и близостями между его элементами // Сиб. журн. индустр. матем., 2009. Т. 12. № 1. С. 61–73.

[7] *Двоенко С. Д., Пшеничный Д. О.* Устранение метрических нарушений в матрицах парных сравнений // Известия Тульского государственного университета. Технические науки, 2013. № 2. С. 96–104.

[8] *Двоенко С. Д., Пшеничный Д. О.* О локализации отрицательных собственных значений в матрицах парных сравнений // Известия Тульского государственного университета. Технические науки, 2013. № 9(2). С. 94–102.

[9] *Гантмахер Ф.Р.* Теория матриц — М.: Наука, 1988. 552 с.

[10] OECD statistics. OECD, 2013–2014. http://stats.oecd.org/.

[11] *Небылицын В.Д.* Основные свойства нервной системы человека — М.: Просвещение, 1966. 384 с.

Поступила в редакцию 19.08.2016

Feature grouping based on the optimal sequence of correlation matrix minors*

S. D. Dvoenko and D. O. Pshenichny

dsd@tsu.tula.ru; denispshenichny@yandex.ru

Tula State University, 92 Lenina pr., Tula, Russia

Background: It is known that data analysis problems usually arise in early stages of investigations, when a model of a phenomenon in researching has not been developed yet. Hence, it is too early to introduce a problem of a model identification. It needs to collect and study a lot of miscellaneous information about the most significant characteristics of a phenomenon under investigation in this case. Such a situation forces one to use inconsistent approach, since it is unknown what characteristics are important and what knowledge needs to be collected. Therefore, data analysis methods should resolve the contradiction and focus on the correct description of the phenomenon. The problem of informal interpretation of factors and groups arises in the grouping problem. Factors are synthetic features and difficulties can arise in informal interpretation of them. Therefore, groups and corresponding factors have been built, the representative usually is defined for each group as a feature, the most correlated with the group factor. As a result, it is possible to name groups informally as such initial features.

Methods: The new approach to specify a feature subset is proposed to represent correctly hidden factors. In this approach, it does not need to define eigenvectors or centroid ones as intermediate transformations. It is based on the optimal sequence of correlation matrix minors, since the less correlated features are placed at the beginning of the sequence and the more correlated ones are placed closer to the end of it.

Results: As it is shown, the proposed approach can produce initial partitioning for other grouping algorithms and additionally can be used to evaluate a number of groups and to get informal partitions.

Concluding Remarks: As it is evident, the natural hidden regularity in the phenomenon under investigation appears undoubtedly because of processing data by different techniques and algorithms targeted to uncover it. All such results as a whole will support the correct result. Therefore, it needs to support and develop the diversity of data processing intelligent methods. In this paper, an attempt to do it is presented. It is the relevant attempt since large volume of experimental data has been collected and methods for pairwise comparisons have been developed.

Keywords: *grouping; cluster; metrics; correlation; eigenvalue; eigenvector; determinant*

DOI: 10.21469/22233792.2.2.1

*The research was partially supported by the Russian Foundation for Basic Research (grants 15-07-02228, 15-07-08967, 14-07-00527, and 14-07-00964).

References

[1] Zhuravlev, U. I. 1978. Ob algebraicheskom podhode k resheniyu zadach raspoznavaniya ili klassifikatsii [About algebraic approach to solving problems of recognition or classification]. *Problemy kibernetiki* [Cybernetics Problems] 33:5–68.

[2] Lawley, D. N., and A. E. Maxwell. 1971. *Factor analysis as a statistical method*. 2nd ed. London: Butterworth. 117 p.

[3] Harman, H. H. 1976. *Modern factor analysis*. 3rd ed. University of Chicago Press. 508 p.

[4] Lumelskii, V. Ya. 1970. Parameter grouping on the basis of the square coupling matrix. *Automat. Rem. Contr.* 1:117–127.

[5] Cox, T. F., and M. A. A. Cox. 2000. *Multidimensional scaling*. 2nd ed. Chapman and Hall/CRC. 328 p.

[6] Dvoenko, S. D. 2009. Clustering and separating of a set of members in terms of mutual distances and similarities. *Trans. Machine Learning Data Mining* 2(2):80–99.

[7] Dvoenko, S. D., and D. O. Pshenichny. 2013. Ustranenie metricheskikh narusheniy v matritsakh parnykh sravneniy [The removing of metric violations in matrixes of pair comparisons]. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tehnicheskie nauki* [Proceedings of the Tula State University. Engineering Sciences] 9(2):96–104.

[8] Dvoenko, S. D., and D. O. Pshenichny. 2013. O lokalizatsii otritsatel'nykh sobstvennykh znacheniy v matritsakh parnykh sravneniy [On localization of the negative eigenvalues for matrices of pairwise comparisons]. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tehnicheskie nauki* [Proceedings of the Tula State University. Engineering Sciences] 9(2):94–102.

[9] Gilbert, G. T. 1991. Positive definite matrices and Sylvester's criterion. *Am. Math. Mon.* 98(1):44–46. doi: 10.2307/2324036.

[10] OECD. 2013–2014. OECD statistics. Available at: `http://stats.oecd.org/` (accessed August 15, 2016).

[11] Nebylitsyn,V. D. 1966. *Osnovnye svoystva nervnoy sistemy cheloveka* [Main characteristics of the nervous system of a person]. 1966. Moscow: Prosveshchenie. 384 p.

Received August 19, 2016

Определение области затенения радужки кластеризацией, основанной на локальных текстурных признаках*

К. И. Талипов[1,2]***, И. А. Матвеев***[1,2]
kamiltalipov@gmail.com; ivanmatveev@mail.ru
[1]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
[2]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 40

Решается задача выделения точек затенения области радужки различными объектами. Исходными данными является изображение радужки глаза человека и окружности, аппроксимирующие границы зрачок–радужка и радужка–склера. В качестве метода решения предлагается использовать расчет локальных текстурных признаков и кластеризацию полученного вектора признаков. Целью работы является построение эффективного алгоритма, выделяющего точки затенения, и исследование возможности сегментации затенений радужки без априорно заданной модели ее текстуры. Работа алгоритма проиллюстрирована примерами на данных из баз изображений радужки.

Ключевые слова: *локальные текстурные признаки; кластеризация; сегментация изображений*

DOI: 10.21469/22233792.2.2.02

1 Введение

Распознавание радужки глаза — один из наиболее точных способов идентификации человека, имеющий выжные практические приложения. При использовании любого метода идентификации важна его точность. На практике изображение радужки глаза часто перекрывается (затенено) различными объектами: блики, веки, ресницы, тени от век и ресниц. Шумы такого рода снижают точность распознавания, поэтому задача выделения затененных областей изображения является важной для обеспечения высокой точности идентификации.

Можно выделить следующие основные используемые подходы к решению этой задачи: анализ направлений градиента, анализ границ радужки, кластеризация векторов локальных признаков точки, сравнение последовательных кадров одного и того же глаза. Более подробно методы описаны в работе [1]. Кроме того, в этой же работе производится сравнение эффективности различных алгоритмов нахождения затенений.

В качестве базового алгоритма предлагается метод, описанный в [2]. Для каждого пикселя области радужки рассчитывается набор локальных текстурных характеристик, которые составляют вектор признаков. На полученных векторах выполняется процедура кластеризации. Класс, содержащий максимальное число элементов, считается классом пикселей радужки, остальные — различного рода помехами. В [2] этот метод используется для сегментации областей радужки глаза. Результаты позволяют сделать вывод о возможности применения рассматриваемого метода для решения поставленной задачи. Однако разброс точности сегментирования говорит о необходимости подбора локальных текстурных характеристик и способа кластеризации полученных объектов. Целью работы является создание алгоритма опредения точек затенения без заранее заданной модели текстуры радужки.

*Работа выполнена при финансовой поддержке РФФИ проект №15-01-05552

В работе [3] описан другой алгоритм выделения затененных областей, основанный на анализе освещенности блоков малого размера и отсеивания шумов посредством винеровского оценивания. Он показал точность в 98,52% на 756 тестах. Представляется разумным сравнить точность нахождения затененных точек для различных алгоритмов. В качестве изображений радужки для тестирования предлагается использовать изображения из баз данных [4–7].

2 Постановка задачи

Дано растровое монохроматическое изображение глаза человека $I(x, y)$ являющееся матрицей целочисленных значений: $I(x, y) \in [0; 255]$, размером $W \times H$, т. е. $x \in [1; W]$, $y \in [1; H]$. Кроме того, известны окружности, аппроксимирующие границы зрачок–радужка и радужка–склера, определенные автоматическими методами [8]. Полагается, что кольцо, заключенное между этими окружностями, является радужкой. Каждая окружность задана своим центром и радиусом. На рис. 1 приведен пример изображения с выделенными окружностями. Область радужки часто перекрыта (затенена) различными объектами: блики, веки, ресницы, тени от век и ресниц. Требуется выделить точки затенения в области радужки.

Рассматривается задача классификации, в которой объектами являются точки изображения радужки глаза человека. Для каждой точки априорно не известен класс, к которому она принадлежит. Класс точек без помех обозначается меткой $y = 0$, затененых — $y = 1$. Каждой точке приписывается n-мерный набор локальных текстурных признаков $\mathbf{x}_i \in [0; 1]^n$, полученный на основе анализа изображения.

Предполагается, что точки радужки имеют «схожее» признаковое описание, «отличающееся» от характеристик затененных точек. Другими словами, считается, что выполнена гипотеза компактности. Понятие «схожести» удобно ввести в терминах функции расстояния, определенной на множестве объектов $\mathbf{X}$:

$$\rho(\mathbf{x}_i, \mathbf{x}_k) : \mathbf{X} \times \mathbf{X} \longrightarrow \mathbb{R}_+.$$

Исходя из вышеуказанного предположения для классификации точек используется кластеризация вектора локальных текстурных признаков. Класс, содержащий максимальное число точек, считается радужкой, остальные — точками затенения. Пусть функция

$$a : (\mathbf{u}, \mathbf{X}) \mapsto t \in \{0, 1\},$$

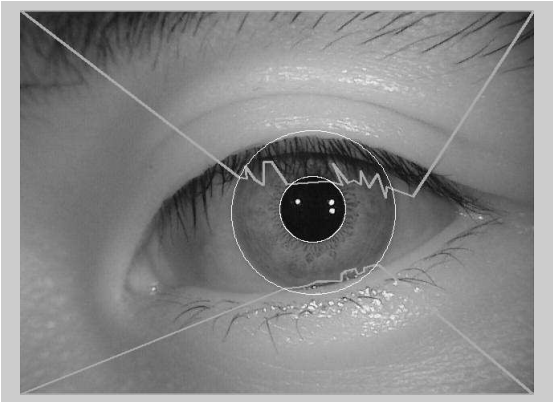


Рис. 1 Пример изображения глаза с областями затенений, выделенными окружностями и экспертной разметкой

где $\mathbf{u} \in \mathbf{X}$, а $y = [y_1, ..., y_M]^{\mathrm{T}} \in \{0,1\}^M$ — вектор ответов, осуществляет класификацию выборки из M точек.

Для контроля качества алгоритма на изображении рассчитываются два критерия:

1) относительная ошибка первого рода — доля точек радужки, которые алгоритм ошибочно классифицировал как затененные:

$$\mathbf{E}_1(I) = \frac{1}{M_0}\sum_{m=1}^{M}[a(\mathbf{x}_m, \{\mathbf{X}\backslash\mathbf{x}_m\}) \neq y_m][y_m = 0],$$

где M_0 — число незатенненых точек радужки в выборке $\mathbf{X}$;

2) относительная ошибка второго рода — доля затенненых точек, которых алгоритм ошибочно классифицировал как точек радужки:

$$\mathbf{E}_2(I) = \frac{1}{M_1}\sum_{m=1}^{M}[a(\mathbf{x}_m, \{\mathbf{X}\backslash\mathbf{x}_m\}) \neq y_m][y_m = 1],$$

где M_1 — число затененных точек в выборке $\mathbf{X}$.

Качество решения определяется как минимум суммы относительных ошибок первого и второго рода по тестовому набору изображений, для которых известны области затенения. Также рассматривается доля точек, для которых алгоритм верно определил класс.

3 Метод решения

Для удобства работы с изображением производится полярное преобразование изображения, также называемое нормализацией, по методу, предложенному в работе [9] (рис. 2):

$$I(x(r,\theta), y(r,\theta)) \longrightarrow I(r,\theta).$$

Нормализованное изображение радужки $P(\theta, r)$ представляет собой прямоугольник $F \times R$ пикселей, где $P(\theta, r) \in [0, 255]$. Для удобства работы будем считать что по оси F точки зациклены, а по оси R отзеркалены, т. е. $P(F+1, R+1) = P(1, R-1)$. Так же производится повышение контрастности изображения путем изменения диапазона интенсивностей исходного изображения (метод *imadjust*).

В качестве локальных текстурных признаков точки используются:

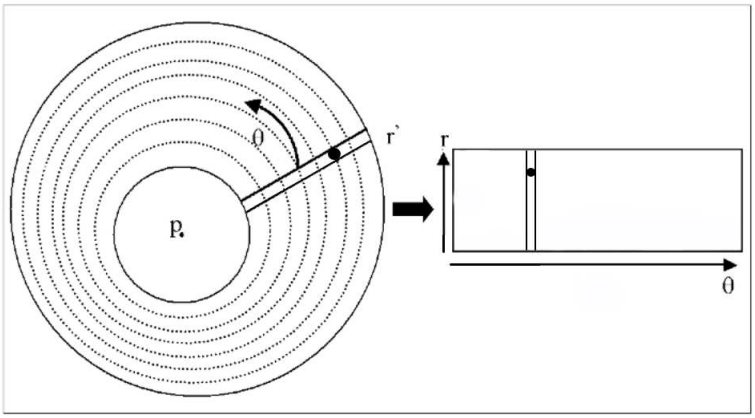


Рис. 2 Полярное преобразование

- первый момент яркости в окрестности 7×7:

$$B_1(\theta, r) = \frac{1}{49} \sum_{\triangle\theta=-3}^{3} \sum_{\triangle r=-3}^{3} P(\theta + \triangle\theta, r + \triangle r);$$

- второй момент яркости в окрестности 7×7:

$$B_2(\theta, r) = \frac{1}{49} \sum_{\triangle\theta=-3}^{3} \sum_{\triangle r=-3}^{3} P(\theta + \triangle\theta, r + \triangle r)^2;$$

- стандартное отклонение в окрестности 3×3:

$$D(\theta, r) = \frac{1}{9} \left(\left(\sum_{\triangle\theta=-1}^{1} \sum_{\triangle r=-1}^{1} P(\theta + \triangle\theta, r + \triangle r)^2 \right) + \left(\sum_{\triangle\theta=-1}^{1} \sum_{\triangle r=-1}^{1} P(\theta + \triangle\theta, r + \triangle r) \right)^2 \right);$$

- перепад яркости (разница между максимальной яркостью и минимальной) в окрестности 3×3:

$$M(\theta, r) = \max_{(v,u)\in\Omega} P(v, u) - \min_{(v,u)\in\Omega} P(v, u),$$

где $\Omega = [\theta - 1; \theta + 1] \times [r - 1; r + 1]$;
- расстояние до зрачка, нормированное на радиус радужки:

$$A(\theta, r) = \frac{d(\theta, r)}{R},$$

где $d(\theta, r)$ — евклидово расстояние до зрачка; R — радиус радужки;
- главные компонененты с 90%-ной значимостью для матрицы данных $V = \{\boldsymbol{v}_{\boldsymbol{\theta_r}} | x \in [1, F], y \in [1, R]\}$, где $\boldsymbol{v}_{\boldsymbol{\theta_r}}$ — это значения яркости в окрестности 7×7 точки (θ, r), рассматриваемые как вектор;
- случайное марковское поле в окрестости 7×7. Для каждой точки окрестности считается индикатор $I(\theta, r) = [P(\theta, r) < A]$, где A — значение средней интенсивности. Обозначим за $T(x, y)$ вероятность перехода при осущетвлении обхода из ячейки со значением индикатора x в клетку со значением индикатора y, где $x \in \{0, 1\}, y \in \{0, 1\}$. Тогда вектор $(T(0, 0), T(0, 1), T(1, 0), T(1, 1))$ будет марковским случайным полем.

Так как каждый признак имеет разные максимальные и минимальные значения, то происходит процедура нормировки, которая нормирует значения признаков так, чтобы они имели среднее значение, равное 0, и стандартное отклонение, равное 1. После нормировки векторов признаков для каждой точки производится кластеризация этих векторов с использование некоторой метрики. Для классификации объекта $\mathbf{u} \in \mathbf{X}$ предлагается применить следующие алгоритмы: метод k-средних и метод k-медоидов. Вычислительный эксперимент производится при параметре k, равном 2 и 3.

4 Вычислительный эксперимент

Вычислительный эксперимент проводился на двух тестовых выборках. В качестве первой используется $S_1 = 110$ изображений радужки, взятых случайным образом из баз данных [4–7]. Затенения на этих изображений имеют различный характер, площадь затенений

и геометрическое расположение затенений также различны. На этих данных протестирована точность (доля правильно распознанных точек) алгоритма при использовании различных метрик и методов кластеризации. Вторая выборка представляет расширенную версию первой. Она состоит из $S_2 = 950$ изображений, взятых случайно из баз данных [4–7]. Показавшие наиболее высокие результаты на первой выборки алгоритмы тестируются на второй. Затенения на этих изображений имеют различный характер, площадь затенений и геометрическое расположение затенений также различны. На этих данных протестирована точность (доля правильно распознанных точек) алгоритма при использовании различных метрик и методов кластеризации.

В качестве функций расстояния между векторами признаков используются расстояния:

- квадратичное Евклидово расстояние:

$$d(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|^2;$$

- расстояние Евклида (только для k-medoids):

$$d(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|;$$

- нормализованное евклидово расстояние (только для k-medoids):

$$d(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{\sum_{i=1}^{n} \left(\frac{u_i}{s_i} - \frac{v_i}{s_i} \right)^2},$$

  где s_i — стандартное отклонение i-й компоненты вектора по всей выборке;
- расстояние городских кварталов:

$$d(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{n} |u_i - v_i|;$$

- метрика Минковского с $p = 2$ (только для k-medoids):

$$d(\boldsymbol{u}, \boldsymbol{v}) = \left(\sum_{i=1}^{n} |u_i - v_i|^p \right)^{1/p};$$

- расстояние Чебышёва (только для k-medoids):

$$d(\boldsymbol{u}, \boldsymbol{v}) = \max_i |u_i - v_i|;$$

- расстояние Махаланобиса (только для k-medoids):

$$d(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{(\boldsymbol{u} - \boldsymbol{v}) V^{-1} (\boldsymbol{u} - \boldsymbol{v})^T},$$

  где V — матрица ковариации;
- косинусное расстояние:

$$d(\boldsymbol{u}, \boldsymbol{v}) = 1 - \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|},$$

  где $\boldsymbol{u} \cdot \boldsymbol{v}$ — скалярное произведение;

- корреляционное расстояние:

$$d(\boldsymbol{u}, \boldsymbol{v}) = 1 - \frac{(\boldsymbol{u} - \overline{\boldsymbol{u}})(\boldsymbol{v} - \overline{\boldsymbol{v}})}{\|\boldsymbol{u} - \overline{\boldsymbol{u}}\|\|\boldsymbol{v} - \overline{\boldsymbol{v}}\|},$$

где $\overline{\boldsymbol{u}}$ — среднее значение элементов $\boldsymbol{u}$.

Производится сравнение точности алгоритма в зависимости от выбранной метрики и метода кластеризации.

Результаты вычислительного эксперимента на первой тестовой выборки при $k = 2$ и 3 показаны в табл. 1 и 2 соответсвенно, где точность

$$Q = \frac{1}{S}\sum_{i=1}^{S} Q_i = \frac{1}{S}\sum_{i=1}^{S} Q_i \frac{W \times H - \mathbf{E}_1(I_i) - \mathbf{E}_2(I_i)}{W \times H},$$

усредненная по всем изображениям из тестовой выборки доля точек изображения радужки, распознанных алгоритмом корректно, а $\triangle Q$ — дисперсия точности.

Видно что k-means с расстоянием городских кварталов, k-medoids с нормализованное Евклидовом расстоянии и иерархическая с нормализованное Евклидовом расстоянии показали наиболее высокие результаты и именно они будут протестированы на второй выборки (табл. 3). Кроме того, для проверки адекватности оценки алгоритма средним значением качества на выборки были построены графики распределения качества (рис. 3).

Таблица 1 Результаты тестирования при $k = 2$

Метод кластеризации	Расстояние	$Q \pm \triangle Q$
k-means	квадратичное Евклидово	0,707 ± 0,006
	городских кварталов	0,683 ± 0,008
	косинусное	0,718 ± 0,008
	корреляционное	0,718 ± 0,008
k-medoids	нормализованное Евклидово	0,721 ± 0,008
	квадратичное Евклидово	0,703 ± 0,006
	Евклида	0,685 ± 0,008
	городских кварталов	0,684 ± 0,008
	Минковского	0,680 ± 0,008
	Чебышёва	0,685 ± 0,008
	Махаланобиса	0,652 ± 0,011
	косинусное	0,719 ± 0,009
	корреляционное	0,718 ± 0,009
Иерархическая	нормализованное Евклидово	0,704 ± 0,035
	квадратичное Евклидово расстояние	0,696 ± 0,067
	расстояние Евклида	0,670 ± 0,033
	расстояние городских кварталов	0,671 ± 0,084
	метрика Минковского	0,665 ± 0,055
	расстояние Чебышёва	0,675 ± 0,047
	расстояние Махаланобиса	0,649 ± 0,061
	косинусное расстояние	0,662 ± 0,072
	корреляционное расстояние	0,655 ± 0,077

Таблица 2 Результаты тестирования при $k = 3$

Метод кластеризации	Расстояние	$Q \pm \triangle Q$
k-means	квадратичное Евклидово	0,784 ± 0,003
	городских кварталов	0,785 ± 0,004
	косинусное	0,764 ± 0,007
	корреляционное	0,760 ± 0,007
k-medoids	нормализованное Евклидово	0,804 ± 0,005
	квадратичное Евклидово	0,779 ± 0,004
	Евклида	0,779 ± 0,004
	городских кварталов	0,787 ± 0,004
	Минковского	0,782 ± 0,004
	Чебышёва	0,779 ± 0,004
	Махаланобиса	0,775 ± 0,006
	косинусное	0,762 ± 0,007
	корреляционное	0,755 ± 0,008
Иерархическая	нормализованное Евклидово	0,721 ± 0,035
	квадратичное Евклидово расстояние	0,716 ± 0,053
	расстояние Евклида	0,681 ± 0,036
	расстояние городских кварталов	0,691 ± 0,078
	метрика Минковского	0,675 ± 0,049
	расстояние Чебышёва	0,681 ± 0,052
	расстояние Махаланобиса	0,658 ± 0,058
	косинусное расстояние	0,672 ± 0,070
	корреляционное расстояние	0,662 ± 0,073

Таблица 3 Результаты тестирования на большей выборке при $k = 3$

Метод кластеризации	Расстояние	$Q \pm \triangle$ Q
k-means	городских кварталов	0,771 ± 0,009
k-medoids	нормализованное Евклидово	0,792 ± 0,011
Иерархическая	нормализованное Евклидово	0,703 ± 0,012

Видно, что усредненное значение качества хорошо описывает оведение алгоритма в целом.

Иллюстрации различного качества распознавания точек затенения. Изображения рис. 4, *а* и 5, *а* — визуализация результата работы алгоритма в полярном представлении (черные точки — точки радужки, белые точки — точки затенения). Изображение рис. 4, *б* и 5, *б* — эталонный ответ в полярном представлении (черные точки — точки радужки, белые точки — точки затенения). Изображение рис. 4, *в* и 5, *в* содержит экспертную разметку (линии серого цвета) и точки, которые алгоритм считает точками затенения (белый цвет).

4.1 Заключение

Наибольшая точность распознавания в $79{,}2\% \pm 1{,}1\%$ получена при использовании нормализованного Евклидова расстояния как метрики и k-medoids, как метода кластеризации при параметре $k = 3$. Вектор локальных текстурных признаков состоял из первых и вто-

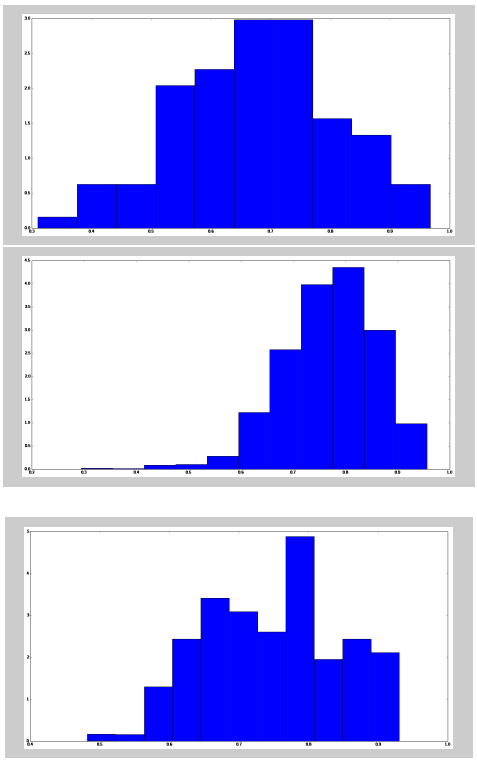
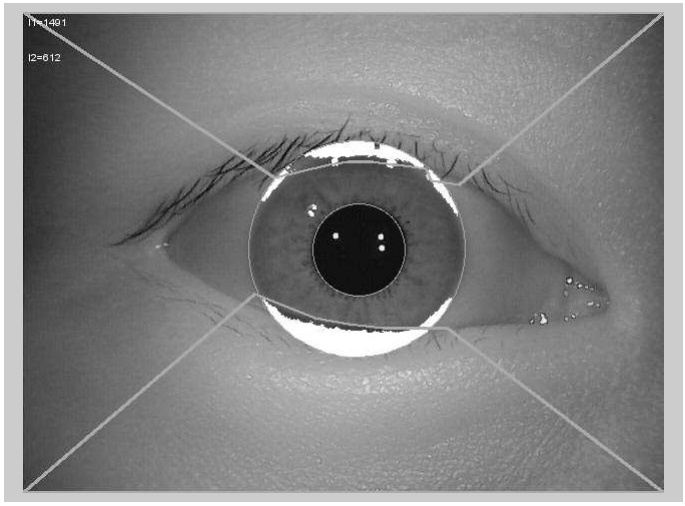


Рис. 3 Распределение точностей некоторых запусков

(*а*)

(*б*)



(*в*)

Рис. 4 Пример высокой точности алгоритма

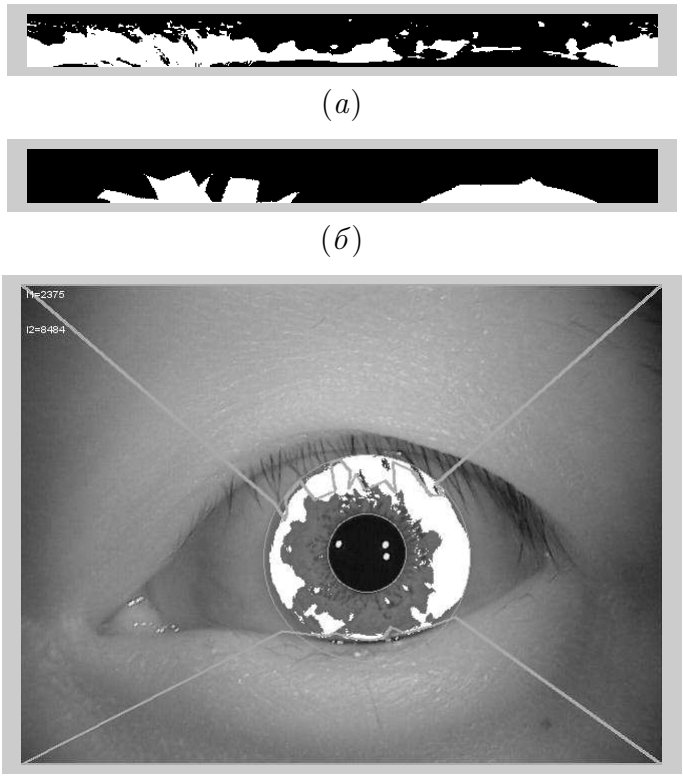


Рис. 5 Пример низкой точности алгоритма

рых моментов яркости в окрестности, стандартное отклонение в окрестности, перепад яркости в окрестности, главные компоненты в окрестности, марковское поле в окрестности и расстояние до зрачка, нормированное на радиус радужки.

Из полученых результатов можно сделать вывод о возможности использования метода кластеризации векторов локальных признаков для решения задачи нахождения точек затенения изображения глаза. Однако невысокая точность распознавания по сравнению с алгоритмом, представленным в работе [3], делает необходимым дальнейшее изучение способов улучшения алгоритма. Исходя из этого факта и из зависимости точности распознавания от метода кластеризации и метрики, представляется разумным провести больше вычислительных экспериментов, взяв дополнительные локальные текстурные признаки и проверить точность распознавания на различных методах кластеризации. Кроме того, стоит протестировать работу алгоритма при неравных весах признаков. Также заметна взаимосвязь точности алгоритма и количества кластеров, на которые происходит кластеризация. Поэтому стоит рассмотреть работу алгоритма при более высоких значениях параметра k.

Литература

[1] *Min T.-H., Park R.-H.* Comparison of eyelid and eyelash detection algorithms for performance improvement of iris recognition // Pattern Recogn. Lett., 2009. Vol. 30. No. 12. P. 1138–1143. doi: 10.1109/ICIP.2008.4711740.

[2] *Proenca H., Alexandre L.* Iris segmentation methodology for non-cooperative recognition // IEE Proc. Vis. Image Sign., 2006. Vol. 153. P. 199–205. doi: 10.1049/ip-vis:20050213.

[3] *Xu G., Zhang Z., Ma Y.* Improving the performance of iris recognition system using eyelids and eyelashes detection and iris image enhancement // IEEE Conference (International) on Cognitive Informatics Proceedings, 2006. P. 871–876. doi: 10.1109/COGINF.2006.365606.

[4] Chinese Academy of Sciences Institute of Automation. Iris image database, ver. 3. 2005. http://biometrics.idealtest.org/dbDetailForUser.do?id=3.

[5] *Monro D. M., Rakshit S., Zhang D.* Iris image database. U.K.: University of Bath, 2005. http://www.bath.ac.uk/elec-eng/research/sipg/irisweb/.

[6] MMU Iris Image Database. Multimedia University. http://pesonna.mmu.edu.my/ccteo/.

[7] *Phillips P., Scruggs W., O'Toole A., et al.* FRVT 2006 and ICE 2006 large-scale experimental results // IEEE Trans. Pattern Anal., 2010. Vol. 5. No. 32. P. 831–846. doi: 10.1109/TPAMI.2009.59.

[8] *Ганькин К.А., Гнеушев А.Н., Матвеев И.А.* Сегментация изображения радужки глаза, основанная на приближенных методах с последующими уточнениями // Известия РАН. Теория и системы управления, 2014. № 2. С. 80–94. doi: 10.7868/S0002338814020097.

[9] *Daugman J.* High confidence visual recognition of persons by a test of statistical independence // IEEE Trans. Pattern. Anal., 1993. Vol. 15. No. 11. P. 1148–1161. doi: 10.1109/34.244676.

Поступила в редакцию 27.08.2016

Eyelids and eyelash detection based on clusterization of vector of local features*

***K. I. Talipov*[1,2] *and I. A. Matveev*[1,2]**

kamiltalipov@gmail.com; ivanmatveev@mail.ru

[1]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow, Russia

[2]Federal Research Center "Computer Science and Control" of RAS, 44/2 Vavilova Str., Moscow, Russia

An attempt has been done to solve the problem of extracting areas where the iris is occluded by various objects. Initial data consist of an image of iris and a circle approximating the boundary between the sclera, the iris, and the pupil. Calculation of local texure features and clusterizing the data based on the extracted information is proposed as a solution method. Two main goals of this particular work are to introduce an effective algorithm for occluded point detection and to study the possibility of their segmentation without a preset texture model. The algorithm's performance is illustrated with the results on various iris image datasets.

Keywords: *local texture feature; clustering; image segmentation*

DOI: 10.21469/22233792.2.2.02

References

[1] Min, T.-H., and R.-H. Park. 2009. Comparison of eyelid and eyelash detection algorithms for performance improvement of iris recognition. *Pattern Recogn. Lett.* 30(12):1138–1143. doi: 10.1109/ICIP.2008.4711740.

[2] Proenca, H., and L. Alexandre. 2006. Iris segmentation methodology for non-cooperative recognition. *IEE Proc. Vis. Image Sign.* 153:199–205. doi: 10.1049/ip-vis:20050213.

*The research was supported by the Russian Foundation for Basic Research (grant 15-01-05552).

[3] Xu, G., Z. Zhang, and Y. Ma. 2006. Improving the performance of iris recognition system using eyelids and eyelashes detection and iris image enhancement. *IEEE Conference (International) on Cognitive Informatics Proceedings*. 871–876. doi: 10.1109/COGINF.2006.365606.

[4] Chinese Academy of Sciences Institute of Automation. 2005. Iris image database, ver. 3. Available at: `http://biometrics.idealtest.org/dbDetailForUser.do?id=3` (accessed November 18, 2016).

[5] Monro, D. M., S. Rakshit, and D. Zhang. 2005. Iris image database. U.K.:University of Bath. Available at: `http://www.bath.ac.uk/elec-eng/research/sipg/irisweb/` (accessed August 26, 2008).

[6] MMU Iris Image Database. Multimedia University. Available at: `http://pesonna.mmu.edu.my/ccteo/` (accessed October 13, 2013).

[7] Phillips, P., W. Scruggs, A. O'Toole, *et al.* 2010. FRVT 2006 and ICE 2006 large-scale experimental results . *IEEE Trans. Patter. Anal.* 5(32):831–846. doi: 10.1109/TPAMI.2009.59.

[8] Gankin, K. A., A. N. Gneushev, and I. A. Matveev. 2014. Segmentatsiya izobrazheniya raduzhki glaza, osnovannaya na priblizhennykh metodakh s posleduyushchimi utochneniyami. *Izvestiya RAN. Teoriya i sistemy upravleniya* [Herald of RAS. Theory and Control Systems] 2:80–94. doi: 10.7868/S0002338814020097.

[9] Daugman, J. 1993. High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. Pattern Anal.* 15(11):1148–1161. doi: 10.1109/34.244676.

Received August 27, 2016

Быстрый алгоритм поиска границ зрачка и радужной оболочки глаза*

***В. В. Чигринский*[1], *Ю. С. Ефимов*[1], *И. А. Матвеев*[2]**
chigrinskiy.viktor@phystech.edu; yuri.efimov@phystech.edu; matveev@ccas.ru
[1]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
[2]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Решается задача поиска границ зрачка и радужной оболочки на изображении глаза. Определяются параметры аппроксимирующих окружностей, а именно: координаты центров и радиусы. Для решения задачи выполняется последовательность шагов: морфологическая обработка и бинаризация входного изображения, определение параметров зрачка, выделение множества граничных точек с помощью оператора Кэнни и определение параметров радужной оболочки с помощью плотности распределения точек по их расстояниям до найденного центра зрачка. Для тестирования алгоритма используется смесь из 2331 изображения радужки.

Ключевые слова: *компьютерное зрение; аппроксимация окружностями; морфологическая обработка изображений; оператор Кэнни; определение границ радужной оболочки*

DOI: 10.21469/22233792.2.2.03

1 Введение

Решается задача выделения радужной оболочки на изображении глаза. Решение может быть использовано в биометрии и медицине. Требуется аппроксимировать границы зрачка и радужной оболочки окружностями. В процессе решения возникают трудности, связанные с нечеткостью изображений, наличием бликов, посторонних объектов или шумов.

Ранее были предложены различные методы, например метод оптимального кругового пути [1, 2], основной идеей которого является рассмотрение изображения в качестве графа. Другой подход базируется на преобразовании Хафа [3, 4], которое позволяет находить на монохромном изображении параметрически заданные кривые. Идея метода — поиск локальных максимумов в фазовом пространстве параметров. В работе [4] предложен метод парных градиентов, основанный на преобразовании Хафа и сочетающий в себе несколько подходов, который давал приемлемые результаты как с точки зрения точности, так и с точки зрения времени работы.

Формально, поставленная задача — это задача поиска параметрически заданных кривых на изображении, различные подходы к решению которой подробно описаны в [5]. Для решения используются такие методы, как математическая морфология и выделение граничных точек изображения с помощью оператора Кэнни. Основные операции математической морфологии: наращивание, эрозия, замыкание и размыкание различными структурными элементами. Подробно об этих операциях можно прочитать в [6]. Оператор Кэнни [7] сглаживает изображение для подавления шума и отмечает границы в тех точках, в которых градиент яркости принимает максимальное значение. В поставленной задаче морфологическая обработка представляет собой последовательное выполнение эрозии и наращивания, что позволяет удалить блики и посторонние шумы на изображении,

*Работа выполнена при финансовой поддержке РФФИ, проект № 16-07-01171.

а оператор Кэнни необходим для выделения граничных точек, соответствующих искомым границам зрачка и радужной оболочки.

2 Постановка задачи

Входные данные — растровое монохромное изображение $\mathbf{I}_0$ размера $W \times H$. Типичный размер — 640×480, однако, возможны и другие варианты. Изображение получено путем фотографирования в ближнем инфракрасном (ИК) диапазоне (850–880 нм) открытого глаза камерой, расположенной приблизительно на его оптической оси. Примеры входных данных приведены на рис. 1.

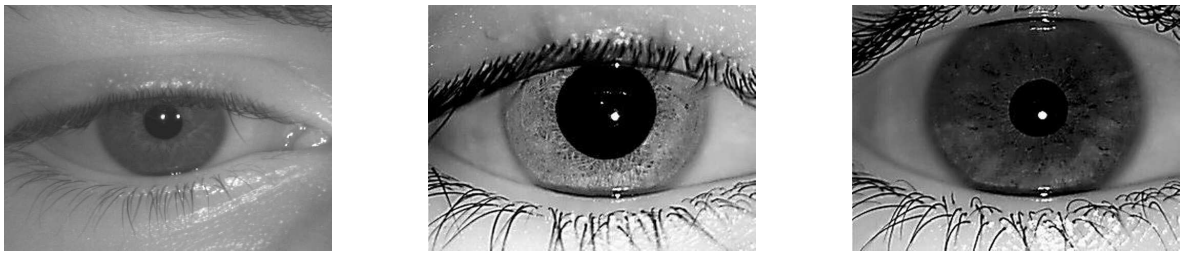


Рис. 1 Примеры входных данных алгоритма

Каждый пиксель входного изображения описывается целым числом от 0 до 255, определяя одну из градаций серого. Требуется найти параметры двух окружностей, аппроксимирующих границы зрачка и радужки, а именно координаты центров и радиусы: $\{\xi_{\text{pupil}}, \eta_{\text{pupil}}, \rho_{\text{pupil}}\}$ и $\{\xi_{\text{iris}}, \eta_{\text{iris}}, \rho_{\text{iris}}\}$.

Качество работы метода оценивается на основе сравнения полученных результатов с «истинными» параметрами, определенными экспертом: $\{\tilde{\xi}_{\text{pupil}}, \tilde{\eta}_{\text{pupil}}, \tilde{\rho}_{\text{pupil}}\}$ и $\{\tilde{\xi}_{\text{iris}}, \tilde{\eta}_{\text{iris}}, \tilde{\rho}_{\text{iris}}\}$.

Абсолютной ошибкой определения искомых параметров считается максимум среди модулей отклонений найденных значений от экспертных:

$$S = \max\left\{\left|\xi_{\text{pupil}} - \tilde{\xi}_{\text{pupil}}\right|, \left|\eta_{\text{pupil}} - \tilde{\eta}_{\text{pupil}}\right|, \left|\rho_{\text{pupil}} - \tilde{\rho}_{\text{pupil}}\right|, \left|\xi_{\text{iris}} - \tilde{\xi}_{\text{iris}}\right|, \left|\eta_{\text{iris}} - \tilde{\eta}_{\text{iris}}\right|, \left|\rho_{\text{iris}} - \tilde{\rho}_{\text{iris}}\right|\right\}, \quad (1)$$

Относительной ошибкой e считается отношение абсолютной (1) к радиусу радужной оболочки $\tilde{\rho}_{\text{iris}}$:

$$e = \frac{S}{\tilde{\rho}_{\text{iris}}}. \quad (2)$$

На отдельном изображении эксперимент считается успешным, если значение относительной ошибки (2) не превосходит допустимого значения δ, определенного экспертом. Критерием качества описанного алгоритма в целом считается доля изображений, на которых относительная ошибка не превосходит δ.

3 Описание алгоритма

Решение задачи осуществляется в два этапа. На первом этапе ведется поиск границы зрачка. Для этого изображение сначала подвергается морфологической обработке, а за-

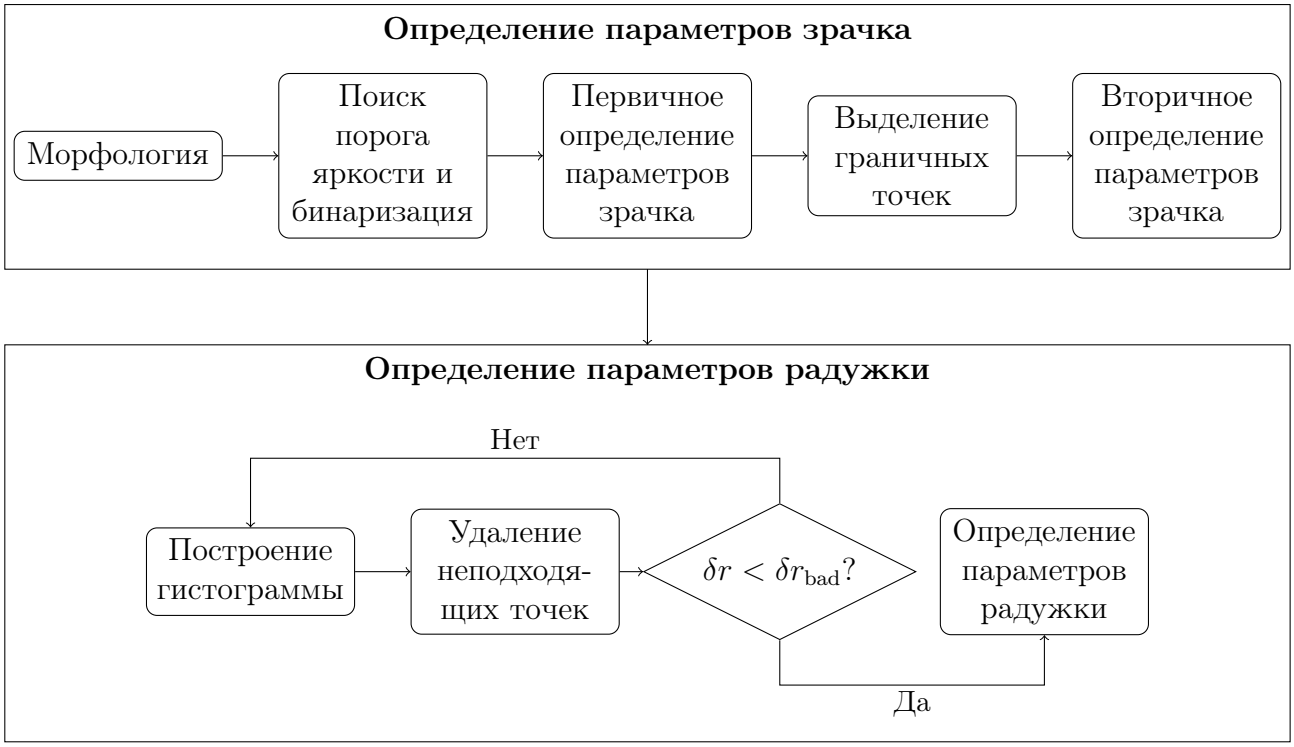


Рис. 2 Общая блок-схема алгоритма

тем бинаризуется, и в бинаризованном изображении выделяется наиболее похожая на круг компонента. Ее граничные точки аппроксимируются окружностью методом наименьших квадратов, и параметры зрачка первично полагаются равными параметрам этой окружности. Затем на морфологически обработанном изображении выделяются граничные точки с помощью оператора Кэнни и по этим граничным точкам окончательно определяется аппроксимирующая зрачок окружность.

На втором этапе осуществляется поиск границы радужной оболочки. Для этого по граничным точкам изображения строится гистограмма численного приближения плотности распределения этих точек по расстояниям до найденного ранее центра зрачка. Максимум такой гистограммы находится в окрестностях радиуса границы радужной оболочки. Затем итеративно удаляются шумовые точки, пока среднеквадратичное отклонение δr аппроксимации оставшихся точек окружностью не станет ниже заданного значения δr_{bad}. После окончания итеративной процедуры определяются параметры радужной оболочки. Общая блок-схема алгоритма приведена на рис. 2.

3.1 Первичное определение границ зрачка

Исходное изображение глаза $\mathbf{I}_0$ подвергается морфологической обработке, а именно: последовательному выполнению эрозии и наращивания, — для избавления от бликов и посторонних мелких шумов. Далее морфологически обработанное изображение $\mathbf{I}_{\text{morph}}$ бинаризуется по порогу яркости $\mathcal{T}$, а именно: строится изображение $\mathbf{B}(\mathcal{T};\xi,\eta)$ следующим образом

$$\mathbf{B}(\mathcal{T};\xi,\eta) = [\mathbf{I}_{\text{morph}}(\xi,\eta) \leqslant \mathcal{T}]\,. \tag{3}$$

Здесь и далее квадратными скобками $[\cdot]$ обозначена нотация Айверсона: индикаторная функция, определенная на множестве всех логических выражений $\mathfrak{B}$, принимающая значение 1 на истинных и 0 на ложных, формально:

$$[\cdot] : \mathfrak{B} \to \{0, 1\}; \quad [\beta] = \begin{cases} 1, & \text{если } \beta \text{ истинно;} \\ 0, & \text{если } \beta \text{ ложно.} \end{cases}$$

Определение порога яркости $\mathcal{T}$ осуществляется перебором. Производится бинаризация по всем различным значениям пикселей изображения. В бинаризованном по некоторому порогу τ изображении выделяются компоненты связности. В предположении, что радиус зрачка не меньше, чем

$$r_{\text{small}} = 0{,}01\,\frac{H + W}{2},$$

где H и W — высота и ширина изображения соответственно, удаляются все компоненты связности, в которых число пикселей меньше πr^2_{small}.

Пусть число оставшихся компонент $N_{\text{cc}}(\tau)$. Если $N_{\text{cc}}(\tau) = 0$, качество данного порога τ полагается равным нулю. Иначе, для каждой компоненты связности определяется эффективный радиус:

$$r_{\text{eff}}\,(\tau; i) = \frac{1}{2}\max\left\{\xi_{\max}\,(\tau; i) - \xi_{\min}\,(\tau; i)\,, \eta_{\max}\,(\tau; i) - \eta_{\min}\,(\tau; i)\right\}, \quad i = 1, \ldots, N_{\text{cc}}(\tau), \tag{4}$$

где $\xi_{\max}\,(\tau; i)$ и $\xi_{\min}\,(\tau; i)$, $\eta_{\max}\,(\tau; i)$ и $\eta_{\min}\,(\tau; i)$ — максимальная и минимальная абсциссы, максимальная и минимальная ординаты точек, относящихся к i-й компоненте связности на изображении $\mathbf{B}(\tau; \xi, \eta)$ соответственно. Далее, каждой компоненте приписывается качество:

$$q(\tau; i) = 1 - \left|1 - \frac{S(\tau; i)}{\pi r^2_{\text{eff}}(\tau; i)}\right|, \tag{5}$$

где $S(\tau; i)$ — число пикселей в i-й компоненте связности на изображении $\mathbf{B}(\tau; \xi, \eta)$. Чем ближе значение $q(\tau; i)$ к единице, тем больше i-я компонента похожа на круг. Значению порога τ, по которому происходила бинаризация, в этом случае приписывается качество, равное максимальному среди качеств всех его компонент. В общем случае, качество порога яркости τ определяется следующим образом:

$$Q(\tau) = [N_{\text{cc}}(\tau) > 0] \max_i q(\tau; i)\,. \tag{6}$$

Искомый порог яркости $\mathcal{T}$ определяется как

$$\mathcal{T} = \arg\max_{\tau} Q(\tau). \tag{7}$$

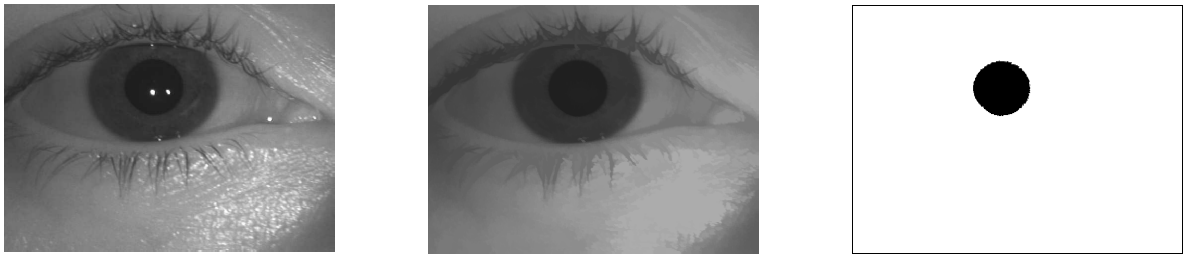


Рис. 3 Морфологическая обработка и бинаризация изображения

Пример бинаризации изображения приведен на рис. 3. Здесь и далее все бинарные изображения инвертированы для большей наглядности.

После бинаризации изображения по найденному значению $\mathcal{T}$ из него выделяется компонента связности с наибольшим значением $q(\mathcal{T}; i)$, т. е. наиболее похожая на круг. Граничные точки этой компоненты аппроксимируются окружностью. Параметры зрачка ξ_{pupil}, η_{pupil} и ρ_{pupil} первично полагаются равными параметрам этой окружности.

3.2 Вторичное определение границ зрачка

К изображению $\mathbf{I}_{\text{morph}}$ — результату морфологической обработки исходного изображения — применяется оператор Кэнни, и получается изображение его граничных точек $\mathbf{I}_{\text{edge}}$. Для каждой граничной точки определяется расстояние до найденного центра зрачка:

$$\rho = \sqrt{(\xi - \xi_{\text{pupil}})^2 + (\eta - \eta_{\text{pupil}})^2}. \tag{8}$$

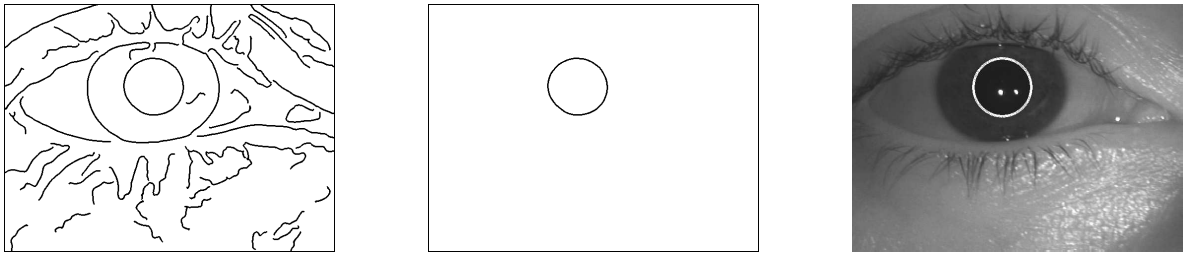


Рис. 4 Определение границ зрачка

Поскольку вычисленные ранее параметры зрачка зачастую оказываются неточными, например из-за недостаточно открытых век, накрашенных ресниц или теней на фотографии, иногда требуется обновить их. Для этого строится изображение $\mathbf{I}_{\text{pupil}}$, состоящее из точек изображения $\mathbf{I}_{\text{edge}}$, предположительно относящихся к границе зрачка, и только них:

$$\mathbf{I}_{\text{pupil}}(\xi, \eta) = \left[\rho < \frac{5}{4}\rho_{\text{pupil}}\right] \mathbf{I}_{\text{edge}}(\xi, \eta). \tag{9}$$

Константа $5/4$ взята из соображения, заключающегося в том, что отношение радиуса радужной оболочки к радиусу зрачка не превосходит этой константы, согласно [8]. Далее на построенном изображении $\mathbf{I}_{\text{pupil}}$ удаляются достаточно малые компоненты связности: такие, для которых

$$S < \rho_{\text{pupil}}, \tag{10}$$

где S — число пикселей в некоторой компоненте связности. Оставшиеся точки аппроксимируются окружностью. Параметры этой окружности — вторичные параметры зрачка. Пример определения границ зрачка приведен на рис. 4.

3.3 Определение границ радужной оболочки. Общая идея

Известно [8], что диапазон значений отношения радиуса радужной оболочки к радиусу зрачка лежит в интервале $(5/4, 5)$. Поэтому точки изображения $\mathbf{I}_{\text{edge}}$, которые могут соответствовать границе радужной оболочки, удовлетворяют системе неравенств:

$$\frac{5}{4}\rho_{\text{pupil}} < \rho < 5\rho_{\text{pupil}}.$$

В силу этого наблюдения строится изображение $\mathbf{I}^{(0)}_{\text{iris}}$, состоящее из тех и только тех точек изображения $\mathbf{I}_{\text{edge}}$, которые могут относится к границе радужной оболочки:

$$\mathbf{I}^{(0)}_{\text{iris}}(\xi,\eta) = \left[\frac{5}{4}\rho_{\text{pupil}} < \rho < 5\rho_{\text{pupil}}\right] \mathbf{I}_{\text{edge}}(\xi,\eta). \tag{11}$$

Построенное таким образом изображение охватывает слишком большое количество точек, большинство из которых являются шумовыми. Для избавления от таковых предлагается итеративная процедура, основанная на том, что плотность распределения точек по их расстоянию до центра зрачка (8) имеет характерный локальный максимум в окрестности значения искомого радиуса радужной оболочки в силу того, что зрачок и радужка являются приближенно концентрическими. Однако искомый максимум плотности распределения может затеряться среди других, соответствующих большому количеству шума за пределами радужки, поэтому реальную плотность распределения $f_{\text{real}}(\rho)$ необходимо отнормировать на наиболее вероятное значение радиуса радужки. Таким образом, получается эффективная плотность распределения:

$$f(\rho) = \frac{f_{\text{real}}(\rho)\nu(\rho_{\text{pupil}};\rho)}{\int_{-\infty}^{+\infty} f_{\text{real}}(\rho')\nu(\rho_{\text{pupil}};\rho')\,d\rho'}\,; \tag{12}$$

$$\nu(\rho_{\text{pupil}};\rho) \sim \mathcal{N}\left(\mu,\sigma^2\right); \quad \mu = \frac{5}{2}\rho_{\text{pupil}}; \quad \sigma = \frac{3}{10}\rho_{\text{pupil}}, \tag{13}$$

константы 5/2 и 3/10 получены путем анализа большой базы изображений радужки, для которой известна экспертная разметка.

k-й шаг итеративной процедуры заключается в следующем: сначала точки изображения $\mathbf{I}^{(k-1)}_{\text{iris}}$ аппроксимируются окружностью, вычисляется среднеквадратичное отклонение такой аппроксимации $\delta r^{(k)}$. В случае, если среднеквадратичное отклонение оказывается меньше некоторого заданного параметра δr_{bad}, процедура прекращается (в работе полагается $\delta r_{\text{bad}} = 0{,}05$). В противном случае определяется реальная плотность распределения точек изображения по расстояниям до центра зрачка, затем с помощью (12) и (13) строится эффективная плотность распределения и, наконец, изображение $\mathbf{I}^{(k)}_{\text{iris}}$ следующим образом:

$$\mathbf{I}^{(k)}_{\text{iris}}(\xi,\eta) = \left[\forall\,\lambda \in [0,1] \quad f\left(\lambda\rho + (1-\lambda)\arg\max_{\rho} f(\rho)\right) > \frac{1}{\ell}\right] \mathbf{I}^{(k-1)}_{\text{iris}}(\xi,\eta), \tag{14}$$

где

$$\ell = \rho_{\max} - \rho_{\min}. \tag{15}$$

Другими словами, выделяется максимальный по вложению промежуток значений ρ, содержащий в себе значение, доставляющее максимум плотности распределения, и такой, что каждая его точка удовлетворяет условию:

$$f(\rho) > \frac{1}{\ell}\,.$$

После окончания процедуры в текущем изображении $\mathbf{I}^{(k)}_{\text{iris}}$ удаляются достаточно малые (10) компоненты связности. Получается изображение $\mathbf{I}_{\text{iris}}$, точки которого аппроксимируются окружностью. Параметры этой окружности — искомые параметры радужной

Рис. 5 Определение границ радужной оболочки

оболочки. В случае, если расстояние между найденными центрами зрачка и радужки оказывается достаточно большим, т. е. удовлетворяющим условию

$$\sqrt{(\xi_{\text{pupil}} - \xi_{\text{iris}})^2 + (\eta_{\text{pupil}} - \eta_{\text{iris}})^2} > \frac{1}{5}\rho_{\text{pupil}},$$

точки $\mathbf{I}_{\text{iris}}$ вновь аппроксимируются окружностью, но уже с фиксированным центром, совпадающим с центром зрачка, и параметры границы радужной оболочки обновляются. Пример определения границ радужной оболочки приведен на рис. 5.

3.4 Определение границ радужной оболочки. Численная реализация

В действительности изображение, с которым ведется работа, является не непрерывной функцией двух переменных, а матрицей конечных размеров. В связи с такой дискретизацией оказывается затруднительным реализовать описанную выше процедуру на вычислительной машине, а именно — построить эффективную плотность распределения. Поэтому предлагается следующая численная реализация построения этой плотности: значения ρ, соответствующие всем пикселям текущего изображения, округляются до целых. Для всех целочисленных $\tilde{\rho}$ строится функция $n(\tilde{\rho})$, определяющее число точек, соответствующих данному $\tilde{\rho}$. Численное приближение плотности распределения определяется как

$$\tilde{f}_{\text{real}}\,(\tilde{\rho}) = \frac{n(\tilde{\rho})/(2\pi\tilde{\rho})}{\sum n(\tilde{\rho}')/(2\pi\tilde{\rho}')}, \tag{16}$$

где знаменатель определяет нормировочный коэффициент, а суммирование осуществляется по всем целочисленным $\tilde{\rho}'$, соответствующим точкам текущего изображения. Эффективное значение плотности распределения получается из (12) и (13) с заменой реальной плотности распределения $f_{\text{real}}(\rho)$ ее численным приближением (16) и интеграла суммой:

$$\tilde{f}(\tilde{\rho}) = \frac{\tilde{f}_{\text{real}}(\tilde{\rho})\nu(\rho_{\text{pupil}};\tilde{\rho})}{\sum \tilde{f}_{\text{real}}(\tilde{\rho}')\nu(\rho_{\text{pupil}};\tilde{\rho}')}. \tag{17}$$

На текущем этапе график $\left(\tilde{\rho}, \tilde{f}\left(\tilde{\rho}\right)\right)$ представляет собой пилообразную гистограмму. Для хорошей работы вышеописанного алгоритма на остальных шагах эта гистограмма подвергается процессу сглаживания, чтобы график наиболее сильно походил на график плотности распределения точек на изображении, рассматриваемом как непрерывная функция. Для сглаживания применяется метод скользящего среднего с шириной окна $2h+$ $+\,1$: значению функции в некоторой точке x_0 присваивается ее среднее значение по всем точкам отрезка длины $2h+1$ с центром в x_0:

$$\tilde{f}_{\text{smooth}}(h;\tilde{\rho}) = \frac{1}{2h+1}\sum_{s=-h}^{h}\tilde{f}(\tilde{\rho}+s). \tag{18}$$

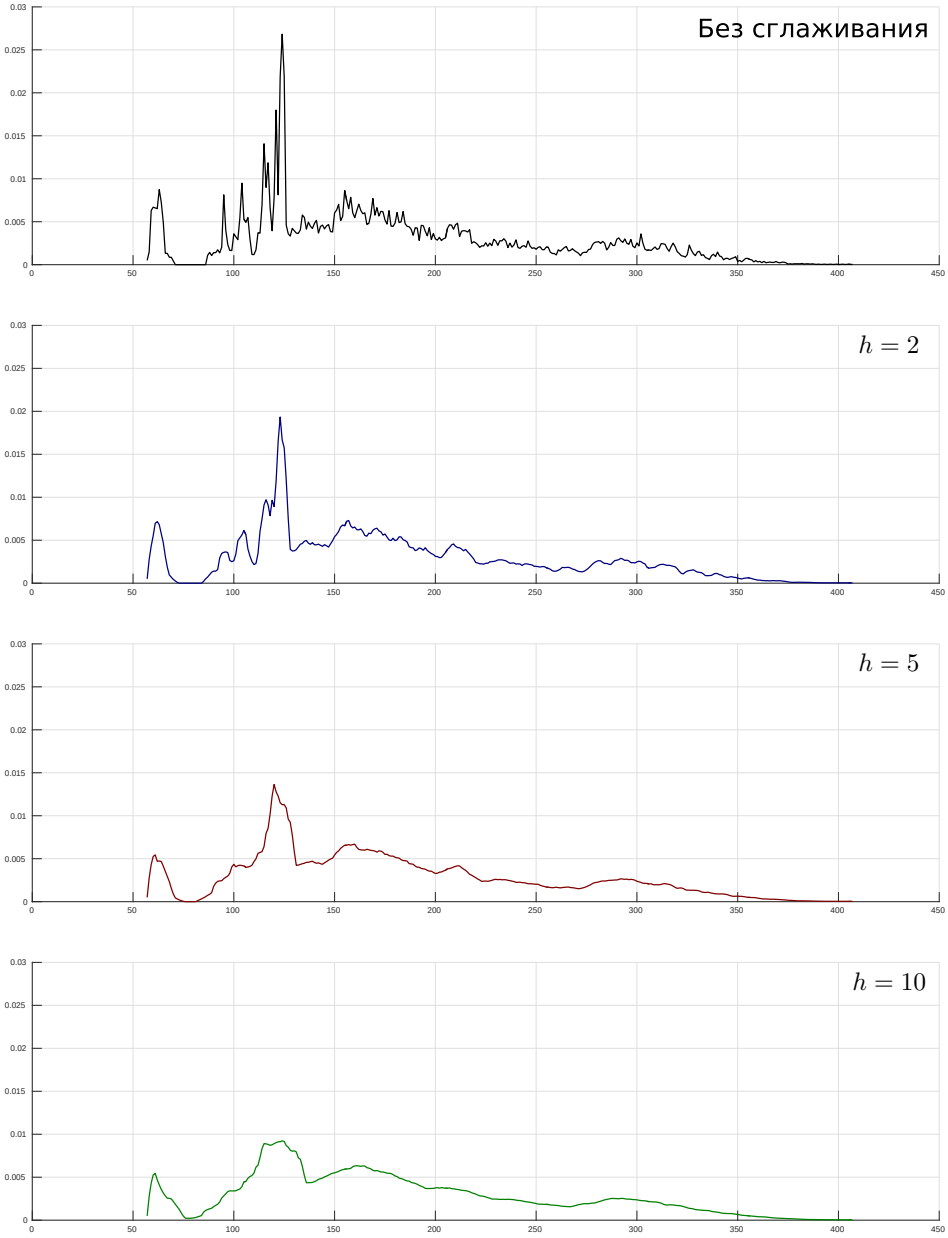


Рис. 6 Сглаживание гистограммы при различных значения полуширины окна

Иллюстрация сглаживания гистограммы приведена на рис. 6.

Краткое описание всего вышеизложенного приведено в алгоритме 1.

Алгоритм 1 Определение границ зрачка радужной оболочки глаза на изображении

Вход: изображение $\mathbf{I}_0$.
Выход: параметры окружностей $\xi_{\text{pupil}}, \eta_{\text{pupil}}, \rho_{\text{pupil}}, \xi_{\text{iris}}, \eta_{\text{iris}}, \rho_{\text{iris}}$.
  $\mathbf{I}_{\text{morph}} \leftarrow \mathsf{morphology}(\mathbf{I}_0)$ // *Морфологическая обработка*
  для всех значений τ пикселей в $\mathbf{I}_{\text{morph}}$
    $\mathbf{B}(\tau) \leftarrow \mathsf{binary}(\mathbf{I}_{\text{morph}}; \tau)$ // *Бинаризация по порогу* τ
    Удалить малые компоненты связности
    для всех компонент связности i
      $r_{\text{eff}}(\tau; i) \leftarrow (4)$
      $q(\tau; i) \leftarrow (5)$
    $Q(\tau) \leftarrow (6)$
  $\mathcal{T} \leftarrow (7)$
  $\mathbf{B} \leftarrow (3)$
  Выделить в $\mathbf{B}$ компоненту связности с наибольшим значением q
  Выделить граничные точки $\mathbf{I}_{\text{pupil}}$ этой компоненты
  $\xi_{\text{pupil}}, \eta_{\text{pupil}}, \rho_{\text{pupil}} \leftarrow \mathsf{OLS}(\mathbf{I}_{\text{pupil}})$ // *Метод наименьших квадратов*
  $\mathbf{I}_{\text{edge}} \leftarrow \mathsf{Canny}(\mathbf{I}_{\text{morph}})$ // *Оператор Кэнни*
  $\mathbf{I}_{\text{pupil}} \leftarrow (9)$
  Удалить малые компоненты связности
  $\xi_{\text{pupil}}, \eta_{\text{pupil}}, \rho_{\text{pupil}} \leftarrow \mathsf{OLS}(\mathbf{I}_{\text{pupil}})$
  $\mathbf{I}_{\text{iris}}^{(0)} \leftarrow (11)$
  $k = 0$
  $\delta r^{(0)} \leftarrow \mathsf{OLS}(\mathrm{I}_{\text{iris}}^{(0)})$
  пока $\delta r^{(k)} > \delta r_{\text{bad}}$
    $f_{\text{real}} \leftarrow (16)$
    $f \leftarrow (17)$
    $f_{\text{smooth}} \leftarrow (18)$
    $k \leftarrow k + 1$
    $\mathbf{I}_{\text{iris}}^{(k)} \leftarrow (14)$
    $\delta r^{(k)} \leftarrow \mathsf{OLS}(\mathbf{I}_{\text{iris}}^{(k)})$
  Удалить малые компоненты связности
  $\mathbf{I}_{\text{iris}} \leftarrow \mathbf{I}_{\text{iris}}^{(k)}$
  $\xi_{\text{iris}}, \eta_{\text{iris}}, \rho_{\text{iris}} \leftarrow \mathsf{OLS}(\mathbf{I}_{\text{iris}})$

4 Вычислительный эксперимент

Целью вычислительного эксперимента является проверка работы алгоритма на реальных данных, а также сравнение точности результатов и времени работы предлагаемого метода с этими же показателями метода парных градиентов, предложенного в работе [4].

Вычисления производились на персональном компьютере с четырехядерным процессором Intel Core i7 3630QM с частотой 2,4 ГГц, оперативная память 8 ГБ, в системе MATLAB.

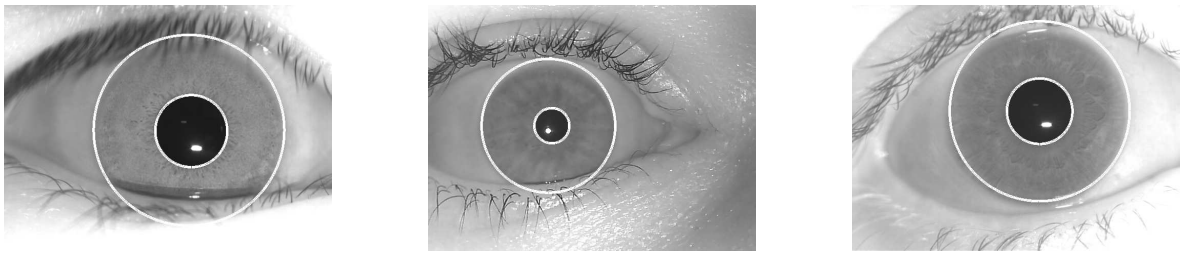


Рис. 7 Примеры корректной работы алгоритма

Для тестирования алгоритма использовалась смесь различных изображений радужки, включающая в себя такие базы, как BATH, ICE, NDIRIS, UBI, CASIA [9], и состоящая из 2331 изображения различного разрешения и качества. Для каждого изображения экспертом были определены истинные значения искомых параметров, записанные в файл разметки, и с учетом этих значений и результатов работы алгоритма были рассчитаны величины относительных ошибок (2). Примеры корректной работы алгоритма приведены на рис. 7.

4.1 Анализ точности и времени работы

Для выявления наиболее хорошего в плане точности результатов и времени работы параметра метода сглаживания был осуществлен перебор при различных значениях этого параметра и сравнение всех полученных результатов между собой и с результатами метода парных градиентов, которое приведено в таблице, где t — время работы в среднем на одно изображение, $\tilde{\ell}$ — количество точек гистограммы, определяется аналогично с (15):

$$\tilde{\ell} = \tilde{\rho}_{\max} - \tilde{\rho}_{\min},$$

а случай $h = 0$ соответствует отсутствию сглаживания.

Как видно из таблицы, предлагаемый алгоритм, независимо от параметра, тратит в среднем на одно изображение около 0,25 с, что почти в 2 раза меньше того же показателя алгоритма, с которым производилось сравнение. Точность вычислений оказалась также практически не зависящей от параметра, однако наиболее предпочтительные результаты были получены при значении $h = 0{,}02\tilde{\ell}$. Более подробное сравнение результатов,

Результаты работы алгоритмов

$h/\tilde{\ell}$	$e < 0{,}02$, %	$e < 0{,}03$, %	$e < 0{,}05$, %	$e < 0{,}07$, %	$e < 0{,}1$, %	t, с
Предлагаемый метод						
0	25,91	46,68	71,39	80,82	87,22	0,246
0,005	26,21	46,85	71,47	81,34	86,96	0,250
0,01	28,01	49,72	73,06	82,11	87,34	0,253
0,015	28,61	50,58	74,05	82,15	87,26	0,254
0,02	**29,56**	**52,25**	**73,62**	**81,72**	**86,53**	**0,254**
0,025	30,12	52,47	73,14	81,25	85,89	0,254
0,03	29,56	51,91	72,89	80,48	84,98	0,254
Метод парных градиентов						
—	11,71	28,87	53,41	68,00	77,43	0,432

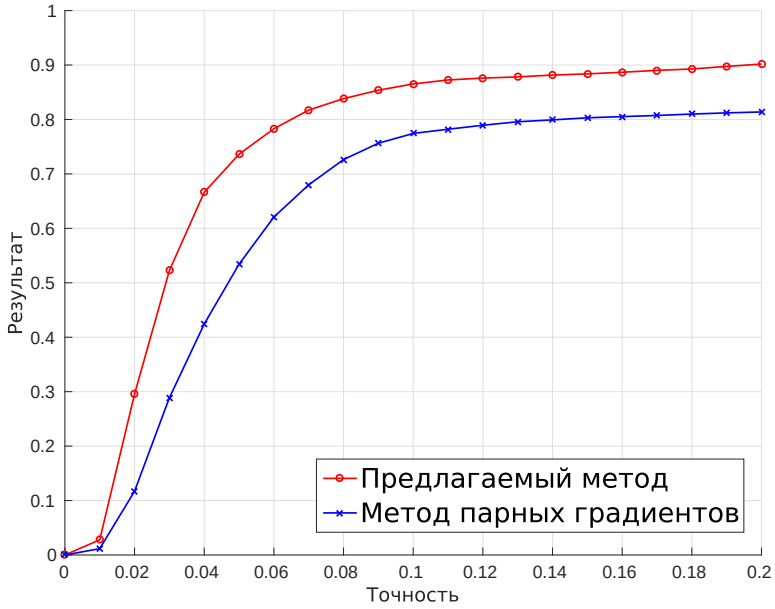


Рис. 8 Сравнение предлагаемого метода и метода парных градиентов

даваемых предлагаемым методом и методом парных градиентов, осуществлено при этом значении параметра и приведено на рис. 8.

4.2 Анализ ошибок

При проведении вычислительного эксперимента были выявлены некоторые недостатки предлагаемого алгоритма. При сильном затемнении изображения в окрестности зрачка после бинаризации изображения не удается выделить круглую компоненту, в связи с чем не удается правильно определить положение зрачка (рис. 9).

При недостаточно темном зрачке на фотографии не удается найти его после бинаризации по более низкому порогу яркости (рис. 10).

При сильной зашумленности на изображении после выделения граничных точек из-за высокой плотности точек, находящихся на периферии изображения, зачастую не удается верно определить радиус радужной оболочки (рис. 11).

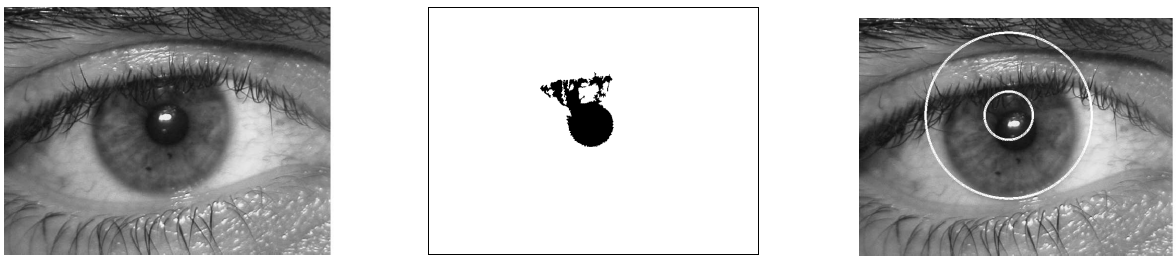


Рис. 9 Сильное затемнение изображения в окрестности зрачка

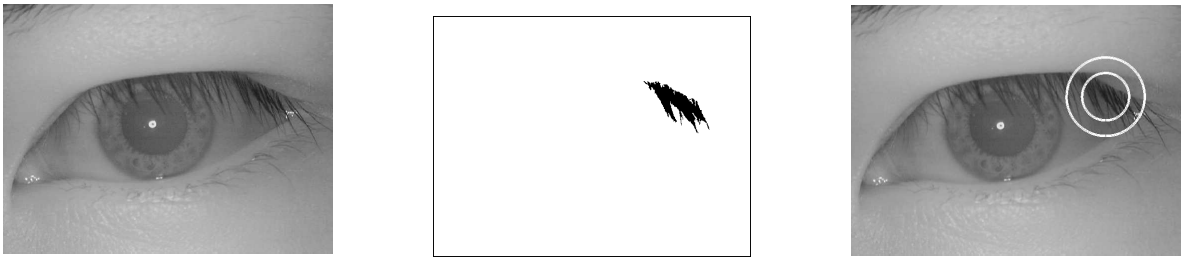


Рис. 10 Недостаточно темный зрачок

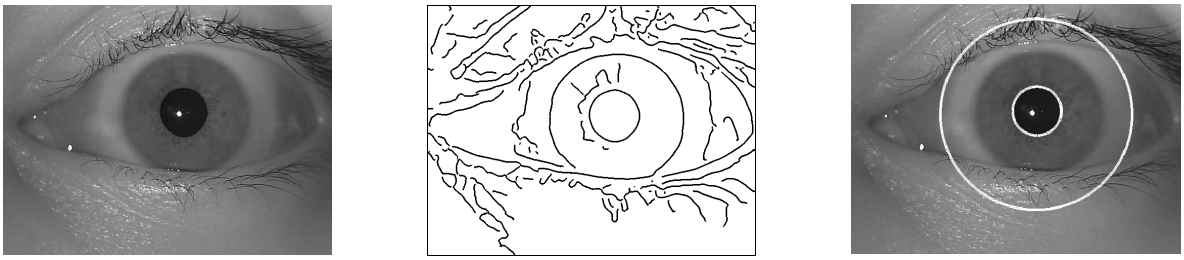


Рис. 11 Сильная зашумленность граничных точек

5 Заключение

Предложен быстрый алгоритм поиска границ зрачка и радужной оболочки. Проведен вычислительный эксперимент, проверяющий работоспособность предложенного алгоритма на реальных данных, результаты которого сведены в таблицу. Приведены примеры верной работы алгоритма, а также его ошибок. Произведено сравнение предложенного метода с методом парных градиентов по показателям качества и времени работы, результаты сравнения приведены в виде графика. Предложенный метод оказывается заметно более быстрым и точным, чем метод парных градиентов. Основным недостатком предлагаемого алгоритма является большое количество параметров, в данной работе грубо оцениваемых некоторыми константами, полученными эмпирическим путём, а также зависимость выбора этих параметров от входного изображения.

Литература

[1] *Матвеев И. А.* Оптимизация кругового пути как метод выделения и уточнения границ радужки на изображении глаза // Известия РАН. Теория и системы управления, 2011. Т. 50. № 5. С. 778–784. doi: 10.1134/S1064230711050157.

[2] Matveev I. A., Simonenko I. V. Detecting precise iris boundaries by circular shortest path method // Pattern Recogn. Image Anal., 2014. Vol. 24. No. 2. P. 304–309. doi: 10.1134/S1054661814020126.

[3] *Матвеев И. А., Ганькин К. А., Гнеушев А. Н.* Сегментация изображения радужки глаза, основанная на приближенных методах с последующими уточнениями // Известия РАН. Теория и системы управления, 2014. Т. 53. № 2. С. 224–238. doi: 10.1134/S1064230714020099.

[4] *Ефимов Ю. С., Матвеев И. А.* Поиск внешней и внутренней границ радужной оболочки на изображении глаза методом парных градиентов // Машинное обучение и анализ данных, 2015. Т. 1. № 14. С. 1991–2002. doi: 10.21469/22233792.1.14.08.

[5] Stockman G., Shapiro L. G. Computer vision. — L.: Prentice Hall PTR, 2001. P. 322–340.

[6] Serra J. Image analysis and mathematical morphology. — Upper Saddle River: Academic Press, Inc., 1982. P. 424–478.

[7] Canny J. F. A computational approach to edge detection // IEEE Trans. Pattern Anal. Machine Intelligence, 1986. Vol. 8. No. 6. P. 679–698. doi: 10.1109/TPAMI.1986.4767851.

[8] *Гиляров М. С.* Биологический энциклопедический словарь. — М.: Советская энциклопедия, 1986. 218 с.

[9] Casia iris image database. Ver. 2. http://biometrics.idealtest.org/dbDetailForUser.do?id=2.

Поступила в редакцию 29.08.2016

Fast algorithm for determining pupil and iris boundaries*

***V. V. Chigrinskiy*[1], *Y. S. Efimov*[1], *and I. A. Matveev*[2]**

chigrinskiy.viktor@phystech.edu; yuri.efimov@phystech.edu; matveev@ccas.ru

[1]Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow, Russia

[2]Federal Research Center "Computer Science and Control" of RAS

44/2 Vavilova Str., Moscow, Russia

The paper presents a method of pupil and iris boundaries determining on eye images. The aim is to find out the parameters of approximating circles, namely, the coordinates of centers and radiuses. To solve the problem, several steps are implemented: morphological processing and a binarization of the input image, determining the pupil parameters, detecting the image edges with the Canny edge detector, and determining the iris parameters using a density of a points distribution by these distances to the just found pupil center. The mixture of the 2331 different iris images is used to test the algorithm.

Keywords: *computer vision; circles approximating; morphological image processing; Canny edge detector; iris boundaries determining*

DOI: 10.21469/22233792.2.2.03

References

[1] Matveev, I. A. 2011. Circular shortest path as a method of detection and refinement of iris borders in eye image. *J. Comput. Syst. Sci. Int.* 50(5):778–784. doi: 10.1134/S1064230711050157.

[2] Matveev, I. A., and I. V. Simonenko. 2014. Detecting precise iris boundaries by circular shortest path method. *Pattern Recogn. Image Anal.* 24(2):304–309. doi: 10.1134/S1054661814020126.

[3] Matveev, I. A., K. A. Gankin, and A. N. Gneushev. 2014. Iris image segmentation based on approximate methods with subsequent refinements. *J. Comput. Syst. Sci. Int.* 53(2):224–238. doi: 10.1134/S1064230714020099.

[4] Efimov, Y. S., and I. A. Matveev. 2015. Iris border detection using a method of paired gradients. *J. Machine Learning Data Anal.* 1(14):1991–2002. doi: 10.21469/22233792.1.14.08.

*The research was supported by the Russian Foundation for Basic Research (grant 16-07-01171).

[5] Stockman, G., and L. G. Shapiro. 2001. *Computer vision*. London: Prentice Hall PTR. 322–340.

[6] Serra, J. 1982. *Image analysis and mathematical morphology*. Upper Saddle River: Academic Press, Inc. 424–478.

[7] Canny, J. F. 1986. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Machine Intelligence* 8(6):679–698. doi: 10.1109/TPAMI.1986.4767851.

[8] Gilyarov, M. S. 1986. *Biological encyclopedia*. Moscow: Sovetskaya entsiklopedia. 218 p.

[9] Casia iris image database. Ver. 2. Available at: http://biometrics.idealtest.org/dbDetailForUser.do?id=2 (accessed December 8, 2016).

Received August 29, 2016

Мультимодальные тематические модели для разведочного поиска в коллективном блоге*

А. О. Янина[1,2]***, К. В. Воронцов***[1,2]
yanina-n@yandex-team.ru; vokov@forecsys.ru
[1]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
[2]Яндекс, Россия, г. Москва, ул. Льва Толстого, 16

Разведочный информационный поиск нацелен на приобретение и систематизацию профессиональных знаний в отличие от поисковых систем, отвечающих на короткие запросы массовых пользователей. Для него характерно отсутствие как точной формулировки запроса, так и единственного правильного ответа. В данной работе предлагается технология тематического разведочного поиска. Рассматривается задача поиска тематически близких документов по текстовому запросу произвольной длины. Применение аддитивной регуляризации тематических моделей (ARTM — additive regularization for topic modeling) позволяет комбинировать требования различности тем и разреженности векторных тематических представлений документов, а также учитывать дополнительные данные об авторах и категориях документов. Для построения тематических моделей используется библиотека с открытым кодом BigARTM. Предлагается методика оценивания точности и полноты тематического поиска на основе оценок асессоров. Эксперименты на данных коллективного блога habraharb.ru показывают, что качество тематического поиска сравнимо с качеством асессорского поиска и даже несколько превосходит его по критерию полноты, при этом асессоры тратят в среднем по 30 мин на каждый тематический запрос, тогда как тематическая поисковая система выдает результат практически мгновенно.

Ключевые слова: *информационный поиск; разведочный поиск; тематическое моделирование; аддитивная регуляризация тематических моделей; BigARTM*

DOI: 10.21469/22233792.2.2.04

1 Введение

Современные поисковые системы отвечают на короткие четко сформулированные запросы массового пользователя. Исследовательский или *разведочный поиск* (exploratory search) — это относительно новая парадигма в информационном поиске, нацеленная на самообразование, приобретение и систематизацию знаний [1, 2]. Потенциальные пользователи разведочного поиска — исследователи, преподаватели, студенты, специалисты различных профессий, работа которых связана с накоплением и анализом информации. Переход к обществу, основанному на знаниях, приводит к расширению информационных потребностей людей и необходимости создания принципиально новых инструментов поиска.

Основной особенностью разведочного поиска является отсутствие точной формулировки запроса и отсутствие единственного ответа. Когда пользователь плохо ориентируется в терминологии или слабо представляет себе структуру предметной области, его первейшей информационной потребностью становится получение «дорожной карты» предметной области, определение наиболее важных тем, систематизация и визуализация релевантной информации по этим темам. В этих случаях трудно или вообще невозможно сформулировать запрос в виде короткой текстовой строки. Проще наметить направление поиска,

*Работа выполнена при финансовой поддержке РФФИ, проекты 16-37-00498, 14-07-00847 и 14-07-00908.

задав в качестве запроса большой фрагмент текста, документ или подборку документов. Целью разведочного поиска является получение ответов на сложные вопросы: «какие темы представлены в тексте запроса», «что читать в первую очередь по этим темам», «что находится на стыке этих тем со смежными областями», «какова структура данной предметной области», «как она развивалась во времени», «каковы последние достижения», «где находятся основные центры компетентности», «кто является экспертом по данной теме» и т. д. Пользователь обычной поисковой системы вынужден итеративно переформулировать свои короткие запросы, расширяя зону поиска по мере усвоения терминологии предметной области, периодически пересматривая и систематизируя результаты поиска. Это требует больших затрат времени и высокой квалификации. При отсутствии инструмента для получения «общей картины» всегда остается сомнение, что какой-то важный аспект изучаемой проблемы так и не был найден. Если образно представить итеративный поиск как блуждание по лабиринту знаний, то разведочный поиск — это средство автоматического построения карты любой части этого лабиринта.

Исследования разведочного поиска можно условно разделить на несколько направлений: изучение поведения пользователей обычных поисковых систем, разработка системных архитектур, сценариев и средств визуализации для разведочного поиска, развитие и применение методов кластеризации, категоризации и семантического анализа текстов.

Отдельным направлением работ в области разведочного поиска является создание размеченных коллекций для оценивания качества поиска [3–5]. В [3] методы оценивания качества разведочного поиска делятся на две большие группы: user-centered и system-centered. Подходы, учитывающие пользовательское поведение в процессе поиска, являются наиболее сложными и ресурсоемкими, но позволяют более точно оценивать качество поиска. Например, в [5] с помощью машинного обучения строится предсказательная модель действий пользователя в ходе разведочного поиска. Признаки для обучения классификатора генерируются из данных об информационной потребности пользователя и о полноте доступной информации по заданной теме. Информационная потребность пользователя определяется по числу запросов, длине каждого запроса, числу слов в запросе, энтропии запроса. Покрытие определяется тремя признаками: числом посещенных веб-страниц; числом страниц, на которых пользователь провел больше 30 с; числом сохраненных страниц. Кроме того, учитываются такие признаки, как «эффективность пользователя» и «интерпретируемость запроса». Далее по этим признакам настраивается предсказательная модель близости пользователя к требуемому результату поиска. Качество разведочного поиска оценивается по шкале от 1 до 4.

В данной работе рассматриваются методы *тематического разведочного поиска*. В их основе лежат следующие предположения: (1) в коллекции текстов, написанных на естественном языке, можно выделить относительно небольшое число тем, много меньшее числа слов и числа документов; (2) каждая тема представляется своим лексиконом — семантически однородным частотным словарем слов и выражений; (3) семантика каждого документа представляется частотным списком тем. *Вероятностные тематические модели* (probabilistic topic models) формализуют эти предположения, представляя каждую тему дискретным распределением вероятностей на множестве слов, а каждый документ — дискретным распределением вероятностей на множестве тем [6–8].

Векторные семантические описания позволяют решать перечисленные выше задачи тематического поиска. Одна из основных — задача поиска тематически близких документов. Системы полнотекстового поиска основаны на инвертированных индексах, в которых для каждого слова хранится список содержащих его документов [9]. Поисковая система ищет

документы, содержащие все слова запроса, поэтому по длинному запросу, скорее всего, ничего не будет найдено. Тематическая поисковая система обходит эту проблему. Она строит тематическую модель коллекции документов, преобразуя каждый документ в относительно короткий частотный список тем. Текст запроса, каким бы длинным он ни был, также преобразуется в короткий список тем. Таким образом, для поиска документов схожей тематики применимы те же механизмы индексирования, поиска и ранжирования, только в роли слов выступают темы.

В тематическом разведочном поиске нет итеративного переформулирования запросов, поэтому нет необходимости в сложных методиках оценивания поведения пользователей. В данной работе для измерения качества поиска используются обычные критерии точности и полноты на основе оценок асессоров и предлагается методика формирования выборки запросов для тематического поиска.

В литературе по разведочному поиску тематическое моделирование стали использовать относительно недавно [10–13], а многие обзоры о нем вообще не упоминают [14–19]. В недавней статье [13] важными преимуществами тематических моделей называются гибкость, возможности визуализации и навигации. В то же время, в качестве недостатков отмечаются проблемы с интерпретируемостью тем, трудности с модификацией тематической модели при поступлении новых документов и высокая вычислительная сложность. Эти проблемы относятся к устаревшим методам и успешно решены в последние годы: десятки новых моделей разработаны для улучшения интерпретируемости; онлайновые алгоритмы способны обрабатывать большие коллекции и потоки документов за линейное время [20–22]. С другой стороны, в работах по тематическому моделированию разведочный поиск часто называют одним из важнейших приложений, а оценки качества поиска используют для валидации моделей [23, 24]. Однако эти исследования пока не привели к созданию общедоступных систем разведочного поиска. Тенденция к сближению этих двух научных направлений наметилась лишь в последние годы.

В системах разведочного поиска к тематическим моделям предъявляется нетривиальная совокупность требований. Они должны автоматически строить существенно различающиеся и хорошо интерпретируемые темы; определять оптимальное число тем или иерархически разбивать темы на подтемы; учитывать не только отдельные слова, но и тематически значимые словосочетания; учитывать разнородные метаданные документов: авторство, время, категории, теги. Этим требованиям по отдельности удовлетворяют различные байесовские тематические модели [8, 25]. Однако комбинирование моделей в байесовском подходе наталкивается на значительные технические трудности. Поэтому в данной работе используется небайесовский многокритериальный подход *аддитивной регуляризации тематических моделей*, ARTM [26].

Для комбинирования моделей, разнородных требований и источников данных в ARTM ставится задача оптимизации взвешенной суммы критериев правдоподобия и регуляризаторов. Независимо от выбранного сочетания регуляризаторов, задача решается одним и тем же регуляризованным EM (expectation-maximization) алгоритмом. Это позволило сочетать модульную технологию тематического моделирования с высокоэффективным параллельным онлайновым EM-алгоритмом в библиотеке с открытым кодом BigARTM (`bigartm.org`) [27]. В предшествующих работах было показано, что ARTM позволяет улучшать интерпретируемость тем одновременно с разреживанием модели и выделением слов общей лексики [28, 29], отбрасывать зависимые и неинформативные темы [30], обрабатывать разнородные документы, содержащие наряду со словами токены различных модальностей [22], использовать словари ключевых слов для выделения узко специализирован-

ных тем, в частности для изучения межнациональных отношений по данным социальных сетей [31].

В данной работе предлагается мультимодальная регуляризованная тематическая модель для разведочного поиска. Для построения модели используется комбинация регуляризаторов, встроенных в библиотеку BigARTM. Предлагается методика оценивания точности и полноты тематического поиска на основе оценок асессоров. С помощью данной методики обосновывается выбор числа тем и дополнительных модальностей.

2 Вероятностное тематическое моделирование

Пусть D — конечное множество (коллекция) текстовых документов, T — конечное множество тем, M — конечное множество модальностей. Каждой модальности $m \in M$ соответствует словарь — конечное множество токенов W_m. Примерами модальностей являются слова, биграммы, теги, категории, авторы, метки времени. Обозначим через W объединение непересекающихся множеств W_m по всем модальностям $m \in M$. Каждый документ $d \in D$ представляет собой последовательность токенов $w_1, \ldots, w_{n_d}$ из W, где n_d — длина документа. Принимая «гипотезу мешка слов», будем считать, что последовательность токенов не важна, и учитывать только число вхождений n_{dw} токена w в документ d.

Вероятностная тематическая модель описывает условную вероятность появления токенов w в документе $d \in D$ как вероятностную смесь распределений $\varphi_{wt} = p(w|t)$ токенов в темах $t \in T$ с коэффициентами $\theta_{td} = p(t|d)$, зависящими от документов:

$$p(w|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}, \quad w \in W_m,\ d \in D.$$

Матрицы $\Phi = (\varphi_{wt})_{W \times T}$ и $\Theta = (\theta_{td})_{T \times D}$ будем использовать для обозначения параметров тематической модели.

В вероятностном латентном семантическом анализе (PLSA — probabilistic latent semantic analysis) используется единственная модальность терминов (как правило, отдельных слов) и ставится задача максимизации логарифма правдоподобия модели $p(w|d)$ при ограничениях неотрицательности и нормировки столбцов матриц Φ и Θ [6].

В аддитивной регуляризации тематических моделей (ARTM) критерий логарифма правдоподобия вводится для каждой модальности и максимизируется их взвешенная сумма [22, 29]. В общем случае данная задача имеет бесконечно много решений, поэтому к этой сумме добавляются дополнительные критерии-регуляризаторы R_i:

$$\sum_{m \in M} \frac{\tau_m}{n_m} \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta},$$

где $n_m = \sum_{d \in D} \sum_{w \in W_m} n_{dw}$ — нормировочный множитель для балансировки модальностей. Задача оптимизации решается с помощью регуляризованного EM-алгоритма [22, 29]. Веса модальностей τ_m и коэффициенты регуляризации τ_i подбираются в эксперименте.

Регуляризатор сглаживания вводит в модель требование, чтобы распределения φ_{wt} и θ_{td} были похожи на заданные распределения β_w и α_t соответственно [28]:

$$R(\Phi, \Theta) = \beta \sum_{m \in M} \sum_{t \in T} \sum_{w \in W^m} \beta_w \ln \varphi_{wt} + \alpha \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Регуляризатор разреживания имеет такой же вид, но коэффициенты регуляризации β и α отрицательны, что способствует появлению нулевых элементов в распределениях φ_{wt} и θ_{td}. Эксперименты показывают, что в тематических моделях возможно сильное разреживание матриц Φ и Θ, до 90%–98%, практически без потери качества модели [29].

Разреживание матрицы Θ, сохраняющее остальные критерии качества модели, важно для тематического поиска, так как это позволяет находить более компактные семантические представления документов и запросов.

Регуляризатор декоррелирования минимизирует ковариации между вектор-столбцами матрицы Φ, повышая различность тем и улучшая интерпретируемость модели [32]:

$$R(\Phi) = -\tau \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws} \to \max .$$

Этот регуляризатор имеет побочный эффект разреживания матрицы Φ, поэтому отдельный регуляризатор разреживания для Φ совместно с ним можно не применять.

3 Выбор стратегии регуляризации

Эксперименты проводились на коллекции из 132 157 статей коллективного блога `habrahabr.ru`. Кроме основной модальности терминов (52 354 слова) использовались следующие модальности: 524 авторов статей, 10 000 комментаторов (авторов комментариев к статьям), 2546 тегов, 123 хаба (категории).

Терминами считались слова длиной больше двух букв. Из числа терминов были исключены слова общей лексики — 5% самых высокочастотных слов в коллекции. Предварительная обработка текстов включала в себя удаление пунктуации, приведение слов к нижнему регистру, замену буквы «ё» на букву «е», лемматизацию при помощи морфологического анализатора pymorphy2.

Всего на Хабрахабре 693 509 пользователей, но из них большая часть только читает и комментирует статьи, не размещая собственные статьи в блоге. Поэтому в качестве комментаторов были выбраны 10 000 активных пользователей следующих трех категорий: авторы хотя бы одной статьи; авторы не менее десяти комментариев с лайками других пользователей; пользователи из групп «старожилы», «звезды», «легенды» и «авторы», составляющие ядро аудитории Хабрахабра.

Тематическая модель строилась с помощью библиотеки BigARTM. Столбцы матрицы Φ инициализировались по умолчанию случайными распределениями, столбцы матрицы Θ — равномерными. В каждую тематическую модель были включены три регуляризатора: декоррелирование распределений терминов в темах (с коэффициентом τ), разреживание распределений тем в документах (с коэффициентом α), сглаживание распределений терминов в темах (с коэффициентом β). Регуляризаторы добавлялись в модель в указанном порядке один за другим. При добавлении каждого регуляризатора его коэффициент регуляризации выбирался из заданной сетки значений по нескольким критериям качества. Для каждого значения коэффициента регуляризации производилось 8 итераций EM-алгоритма. Из всех значений выбиралось то, при котором улучшался хотя бы один из критериев без существенного ухудшения остальных. При этом коэффициенты предыдущих регуляризаторов не изменялись.

Для оценивания модели использовались следующие критерии качества [29]: перплексия, разреженность распределений тем в документах, разреженность распределений токенов в темах для каждой из пяти модальностей — термины, авторы, комментаторы, теги, хабы. Под разреженностью понимается доля нулевых элементов в матрице распределений.

На рис. 1 показаны зависимости перплексии и разреженности от числа итераций при различных значениях коэффициентов регуляризации. В результате была выбрана совокупность коэффициентов регуляризации $\tau = 10^8$, $\alpha = -1{,}5$, $\beta = 0{,}5$. Жирной кривой выделена наилучшая траектория регуляризации.

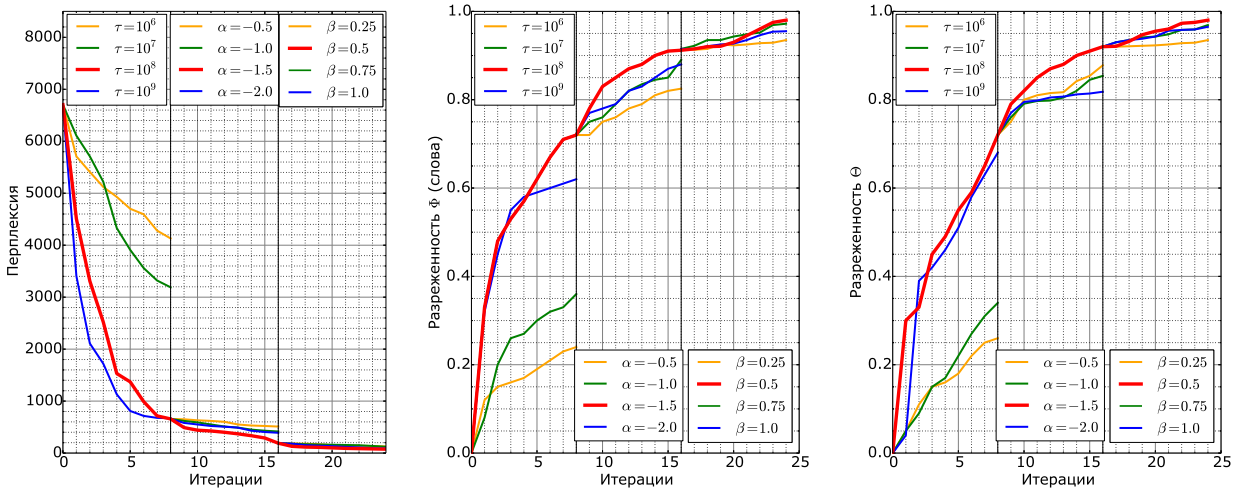


Рис. 1 Зависимости перплексии и разреженности матриц Θ и Φ (только по модальности терминов) от числа итераций и коэффициентов регуляризации.

Веса модальностей τ_m также подбирались по сетке методом проб и ошибок, по критериям перплексии, разреженности и качества тематического поиска (см. ниже). В итоге были подобраны следующие значения τ_m: 1,0 для терминов; 0,5 для авторов; 0,75 для комментаторов; 15,0 для тегов; 10,0 для хабов.

4 Разведочный тематический поиск

Допустим, тематическая модель коллекции уже построена, имеется матрица Φ распределений терминов в темах и текст запроса $q = (w_1, \ldots, w_{n_q})$. Построим для него распределение $\theta_{tq} = p(t|q)$, запустив тематическое моделирование документа q при фиксированной матрице Φ (библиотека BigARTM поддерживает такой режим запуска). Отранжируем документы коллекции $d \in D$ по убыванию косинусной меры близости к запросу q:

$$\text{cosine\_similarity}(q, d) = \frac{\sum_t \theta_{tq}\theta_{td}}{\left(\sum_t \theta_{tq}^2\right)^{1/2}\left(\sum_t \theta_{td}^2\right)^{1/2}}.$$

В качестве выдачи тематического поиска возьмем k документов с векторами θ_d, самыми близкими к вектору запроса θ_q. Число k является параметром поискового механизма или процедуры оценивания качества поиска.

Для оценивания качества тематического поиска предлагается следующая методика, основанная на асессорских оценках релевантности.

Шаг 1. Организатор тестирования составляет множество запросов, соответствующих тематике коллекции. Запросы могут формироваться либо из фрагментов документов коллекции, либо из сторонних текстов — это два альтернативных варианта методики. Каждый запрос должен быть достаточно кратким, чтобы асессор мог быстро понять его смысл, в то же время достаточно полным, чтобы между асессорами не возникало расхождений в его интерпретации. Примерный объем текста запроса — одна страница формата А4. Запрос может иметь заголовок. Текст запроса должен быть информативнее заголовка в том смысле, что обычные поисковые системы не должны давать удовлетворительно полного ответа по тексту заголовка. Вместе с коллекцией запросов асессорам может сообщаться модельная ситуация поиска. Например, в случае новостной коллекции запросом

может быть текст одного или нескольких новостных сообщений, а релевантными документами — тексты новостей, необходимые для восстановления цепочки связанных событий. В случае коллективного блога Хабрахабр запросом может быть несколько текстовых фрагментов, отобранных из внешних источников, а релевантными документами — все статьи Хабрахабра по соответствующей тематике.

Шаг 2. Асессорам раздаются запросы и инструкция, объясняющая поисковое задание и модельную ситуацию поиска. По каждому запросу асессор должен найти в коллекции как можно больше релевантных документов и предоставить список найденных документов. Он может пользоваться любыми доступными ему средствами поиска. Также замеряется время, потраченное асессором на обработку запроса. Число асессоров m, обработавших каждый запрос, является параметром методики. Чем больше m, тем объективнее будут оценки полноты поиска.

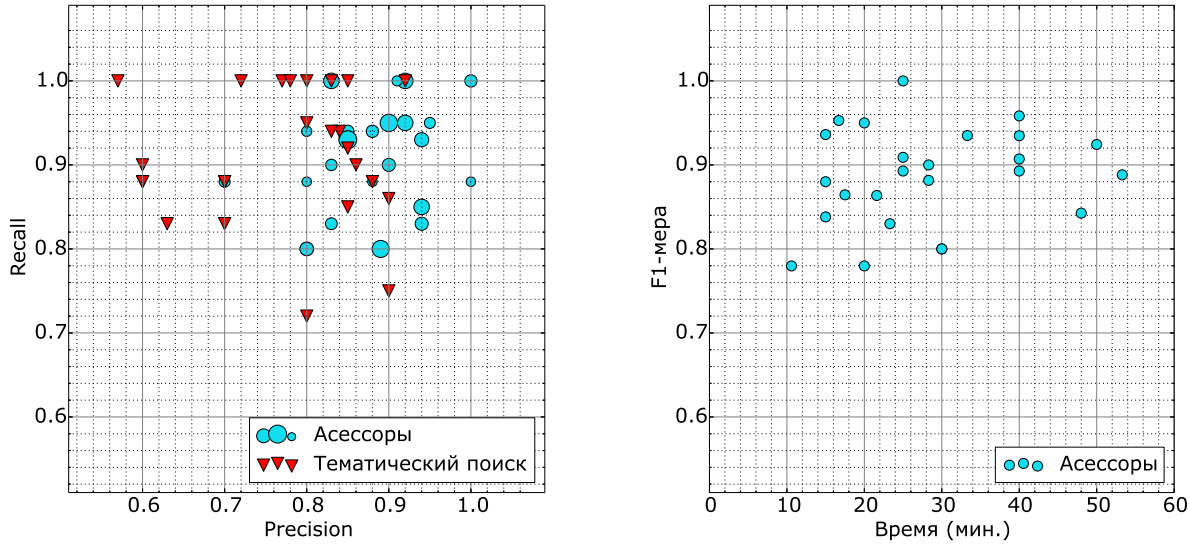
Шаг 3. На том же запросе асессору дается второе задание — разметить результаты тематического поиска. Для каждого пункта поисковой выдачи асессор ставит оценку релевантности в бинарной или порядковой шкале. Документ считается релевантным запросу, если хотя бы один асессор нашел этот документ или если этот документ был найден тематическим поиском и хотя бы n из m асессоров отметили его как релевантный. Число n также является параметром методики.

Для каждого запроса определим две меры качества поиска: *точность* Precision@k — доля релевантных документов среди первых k найденных; *полнота* Recall@k — доля k первых найденных релевантных документов среди всех релевантных. Для измерения качества тематического поиска точность и полнота усредняются по всем запросам. При измерении качества асессорского поиска точность и полнота усредняются еще и по асессорам. Агрегированная оценка качества поиска F_1-*мера* определяется как среднее гармоническое точности P и полноты R: $F_1 = (P + R)/(2PR)$.

Описанная методика была применена к коллекции Хабрахабра. Для эксперимента были составлены 25 запросов по тематике коллективного блога путем копирования текстовых фрагментов из различных внешних источников. Тексты запросов не превышали одной страницы формата A4. Примеры заголовков запросов представлены в табл. 1. Полный список использованных запросов и инструкцию для асессоров можно найти на странице русскоязычного вики-ресурса MachineLearning.ru «Оценивание качества разведочного поиска (эксперимент)». Каждый запрос обрабатывался $m = 3$ асессорами. Результат тематического поиска считался релевантным, если хотя бы $n = 2$ асессора отметили его как релевантный.

Таблица 1 Заголовки запросов для разведочного поиска

Алгоритмы раскраски графов	IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	AB-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Self-driving Google car	Контекстная реклама
Криптосистемы с открытым ключом	Марсоход Curiosity
Обзор платформ онлайн-курсов	Видеокарты NVIDIA
Data Science Meetups в Москве	Распознавание образов
Образовательные проекты mail.ru	Сервисы Google scholar
Межпланетная станция New horizons	MIT MediaLab Research
Языковая модель word2vec	Платформа Microsoft Azure



(*а*) точность и полнота поиска

(*б*) время и F_1-мера (асессоры)

Рис. 2 Качество разведочного поиска по 25 запросам для асессоров и тематического поиска

Результаты эксперимента показаны на рис. 2. Точки на графиках соответствуют запросам. На графике 2, *а* сравниваются точность и полнота поиска, выполненного асессорами, и тематического разведочного поиска. Видно, что точность в среднем выше у асессорского поиска, а полнота — у тематического. Полнота тематического поиска принимала наивысшее значение 1,0 для 8 из 25 запросов. Размер точек пропорционален времени, в среднем затраченному асессорами на обработку данного запроса. График 2, *б* показывает, что нет прямой зависимости между временем, затраченным асессором, и качеством поиска. В среднем на обработку одного запроса асессоры тратили около 30 мин.

Таким образом, тематический поиск позволяет находить релевантные документы полнее и значительно быстрее, чем это делают асессоры, ценой незначительного ухудшения точности (появления нерелевантных документов в результатах поиска).

5 Выбор тематической модели по критерию качества поиска

Множества релевантных документов, найденные асессорами для каждого запроса, позволяют оценивать точность и полноту тематического поиска для новых тематических моделей. Появляется возможность сравнивать тематические модели по критериям качества поиска. Были проведены два таких эксперимента, их результаты сведены в табл. 2. В случаях, когда с помощью новых моделей алгоритм тематического поиска находил новые релевантные документы, расширяли множества релевантных документов и пересчитывали оценки точности и полноты для всех моделей.

В первом эксперименте сравнивались мультимодальные модели с различными сочетаниями модальностей (термины, авторы, комментаторы, теги, хабы) и с числом тем $|T| = 200$. Совместное использование всех модальностей значительно улучшает полноту и точность поиска. Основной вклад вносят модальности терминов и тегов. Модели без терминов, а также унимодальная модель, учитывающая только термины, показывают заметно худшие результаты.

Во втором эксперименте сравнивались модели с числом тем $|T| = 100$, 200, 300, 400 и 500. Оптимальное качество поиска достигается при 200 темах, дальнейшее увеличение числа тем не ведет к повышению точности и полноты. Таким образом, качество поиска

Таблица 2 Сравнение асессорского и тематического поиска по критериям Precision@k и Recall@k: для моделей с разными сочетаниями модальностей (Слова, Комментаторы, Теги, Хабы) при числе тем $|T| = 200$ и для моделей со всеми пятью модальностями и с разным числом тем $|T|$

Критерий	Асессоры	Модальности						Число тем				
		С	К	ТХ	СТ	СХ	СТХ	100	200	300	400	500
Precision@5	0,82	0,63	0,54	0,59	0,74	0,73	0,73	0,61	**0,74**	0,71	0,69	0,59
Precision@10	0,87	0,67	0,56	0,58	0,77	0,74	0,75	0,65	**0,77**	0,72	0,67	0,61
Precision@15	0,86	0,65	0,53	0,55	0,67	0,67	0,68	0,67	**0,68**	0,67	0,65	0,62
Precision@20	0,85	0,64	0,53	0,54	0,66	0,67	0,68	0,64	**0,68**	0,67	0,64	0,60
Recall@5	0,78	0,77	0,63	0,69	0,82	0,81	0,82	0,62	**0,82**	0,80	0,72	0,63
Recall@10	0,84	0,79	0,64	0,71	0,88	0,82	0,87	0,63	**0,88**	0,81	0,75	0,64
Recall@15	0,88	0,82	0,67	0,73	0,90	0,84	0,89	0,67	**0,90**	0,82	0,77	0,67
Recall@20	0,88	0,85	0,68	0,74	0,91	0,85	0,89	0,69	**0,91**	0,85	0,77	0,68

является эффективным внешним критерием для определения числа тем, в то время как внутренние критерии, такие как перплексия, не позволяют судить о числе тем в коллекции [30].

6 Заключение

Конечной целью разведочного информационного поиска является интенсификация и автоматизация процессов приобретения и систематизации знаний людьми. Тематическое моделирование рассматривается как одна из его ключевых технологий. В данной работе исследуется тематический поиск по длинным текстовым запросам на примере коллекции статей коллективного блога Хабрахабр.

Тематическая модель строится с помощью библиотеки с открытым кодом BigARTM, которая позволяет оптимизировать одновременно несколько критериев качества и находить сжатые векторные тематические представления статей и запросов. Для подбора коэффициентов регуляризации использована «жадная» стратегия последовательного добавления регуляризаторов в тематическую модель. Тематический поиск реализуется путем сравнения тематических векторов запроса и статей по косинусной мере близости.

Для оценивания качества поиска разработана специальная коллекция запросов — заданий разведочного поиска, которые сначала выполняются людьми (асессорами), затем системой тематического поиска, затем релевантность найденных ею документов снова оценивается асессорами. Данная методика позволяет, единожды сделав разметку результатов поиска, многократно вычислять оценки качества тематического поиска для различных тематических моделей и механизмов поиска.

На данных Хабрахабра показано, что тематический поиск находит релевантные документы полнее и намного быстрее, чем это делают асессоры. Подбор тематической модели по критериям точности и полноты поиска показал, что использование категоризации статей по тегам и хабам улучшает качество поиска существеннее, чем использование метаинформации об авторах статей и комментариев.

Литература

[1] *Marchionini G.* Exploratory search: From finding to understanding // Commun. ACM, 2006. Vol. 49. No. 4. P. 41–46.

[2] *White R. W., Roth R. A.* Exploratory search: Beyond the query-response paradigm. — Morgan and Claypool Publs., 2009. 98 p.

[3] *Kraaij W., Post W.* Task based evaluation of exploratory search systems // SIGIR Workshop on Evaluating Exploratory Search Systems Proceedings. — ACM, 2006. P. 24–27.

[4] *Potthast M., Hagen M., Völske M., Stein B.* Exploratory search missions for TREC topics // EuroHCIR, 2013. Vol. 1033. P. 7–10.

[5] *Shah C., Hendahewa C., Gonzalez-Ibanez R.* Rain or shine? Forecasting search process performance in exploratory search tasks // J. Assoc. Inform. Sci. Technol., 2016. Vol. 67. No. 7. P. 1607–1623.

[6] *Hofmann T.* Probabilistic latent semantic indexing // 22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings. — New York, NY, USA: ACM, 1999. P. 50–57.

[7] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // J. Machine Learn. Res., 2003. Vol. 3. P. 993–1022.

[8] *Blei D. M.* Probabilistic topic models // Commun. ACM, 2012. Vol. 55. No. 4. P. 77–84.

[9] *Manning C. D., Raghavan P., Schütze H.* Introduction to information retrieval. — New York, NY, USA: Cambridge University Press, 2008. 504 p.

[10] *Scherer M., von Landesberger T., Schreck T.* Topic modeling for search and exploration in multivariate research data repositories // Research and Advanced Technology for Digital Libraries: Conference (International) on Theory and Practice of Digital Libraries Proceedings / Eds. T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, C. J. Farrugia. — Berlin–Heidelberg: Springer, 2013. P. 370–373.

[11] *Grant C. E., George C. P., Kanjilal V., Nirkhiwale S., Wilson J. N., Wang D. Z.* A topic-based search, visualization, and exploration system // FLAIRS Conference. — AIAA Press, 2015. P. 43–48.

[12] *Rönnqvist S.* Exploratory topic modeling with distributional semantics // 14th Symposium (International) on Advances in Intelligent Data Analysis Proceedings / Eds. E. Fromont, T. De Bie, M. van Leeuwen. — Saint Etienne, France: Springer International Publs., 2015. P. 241–252.

[13] *Veas E. E., di Sciascio C.* Interactive topic analysis with visual analytics and recommender systems // 2nd Workshop on Cognitive Computing and Applications for Augmented Human Intelligence, Joint Conference (International) on Artificial Intelligence. — Aachen, Germany: CEUR-WS.org, 2015.

[14] *Feldman S. E.* The answer machine // Synthesis Lectures Inform. Concepts Retrieval Services, 2012. Vol. 4. P. 1–137.

[15] *Rahman M.* Search engines going beyond keyword search: A survey // Int. J. Comput. Appl., 2013. Vol. 75. No. 17. P. 1–8.

[16] *Singh R., Hsu Y.-W., Moon N.* Multiple perspective interactive search: A paradigm for exploratory search and information retrieval on the Web // Multimedia Tools Appl., 2013. Vol. 62. No. 2. P. 507–543.

[17] *Jiang T.* Exploratory search: A critical analysis of the theoretical foundations, system features, and research trends // Library and information sciences: Trends and research / Eds. C. Chen, R. Larsen. — Berlin–Heidelberg: Springer, 2014. P. 79–103.

[18] *Marie N., Gandon F.* Survey of linked data based exploration systems // 3rd Workshop (International) on Intelligent Exploration of Semantic Data co-located with the 13th Semantic Web Conference (International) Proceedings. — Riva del Garda, Italy, 2014.

[19] *Jacksi K., Dimililer N., Zeebaree S. R. M.* A survey of exploratory search systems based on LOD resources // 5th Conference (International) on Computing and Informatics Proceedings. — Malaysia: School of Computing, University Utara, 2015. P. 501–509.

[20] *Mimno D., Hoffman M., Blei D.* Sparse stochastic inference for latent Dirichlet allocation // 29th Conference (International) on Machine Learning Proceedings / Eds. J. Langford J. Pineau. — New York, NY, USA: Omnipress, 2012. P. 1599–1606.

[21] *Bassiou N., Kotropoulos C.* Online PLSA: Batch updating techniques including out-of-vocabulary words // IEEE Trans. Neural Networks Learning Syst., 2014. Vol. 25. No. 11. P. 1953–1966.

[22] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A.* Non-Bayesian additive regularization for multimodal topic modeling of large collections // Workshop on Topic Models: Post-Processing and Applications Proceedings. — New York, NY, USA: ACM, 2015. P. 29–37.

[23] *Yi X., Allan J.* A comparative study of utilizing topic models for information retrieval // Adv. Inform. Retrieval, 2009. Vol. 5478. P. 29–41.

[24] *Andrzejewski D., Buttler D.* Latent topic feedback for information retrieval // 17th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2011. P. 600–608.

[25] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: A survey // Frontiers Comput. Sci. China, 2010. Vol. 4. No. 2. P. 280–301.

[26] *Vorontsov K. V.* Additive regularization for topic models of text collections // Dokl. Math., 2014. Vol. 89. No. 3. P. 301–304.

[27] *Frei O., Apishev M.* Parallel non-blocking deterministic algorithm for online topic modeling // 5th Conference (International) on Analysis of Images, Social Networks and Texts, 2016.

[28] *Vorontsov K. V., Potapenko A. A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // Analysis images, social networks and texts / Eds. D. I. Ignatov, M. Yu. Khachay, M. Y. Panchenko, *et al.* — Communications in computer and information science ser. — Springer International Publs., 2014. Vol. 436. P. 29–46.

[29] *Vorontsov K. V., Potapenko A. A.* Additive regularization of topic models // Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications, 2015. Vol. 101. No. 1. P. 303–323.

[30] *Vorontsov K. V., Potapenko A. A., Plavin A. V.* Additive regularization of topic models for topic selection and sparse factorization // 3rd Symposium (International) on Learning and Data Sciences Proceedings / Ed. A. Gammerman. — U.K.: University of London, 2015. P. 193–202.

[31] *Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Additive regularization for topic modeling in sociological studies of user-generated text content // 15th Mexican Conference (International) on Artificial Intelligence Proceedings, 2016.

[32] *Tan Y., Ou Z.* Topic-weak-correlated latent Dirichlet allocation // 7th Symposium (International) Chinese Spoken Language Processing Proceedings, 2010. P. 224–228.

Поступила в редакцию 02.09.2016

Multimodal topic modeling for exploratory search in collective blog*

***A. O. Ianina*[1,2] *and K. V. Vorontsov*[1,2]**

yanina-n@yandex-team.ru; vokov@forecsys.ru

[1]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia

[2]Yandex LLC, 16 Leo Tolstoy Str., Moscow, Russia

Exploratory Search is a new paradigm in information retrieval focused on the acquisition and systematization of knowledge by professionals, unlike major Web search engines that answer short text queries of mass users. An exploratory search engine has been developed based on probabilistic topic modeling for seeking information thematically relevant to the long text queries. *Additive Regularization for Topic Modeling* (ARTM) was used to combine many requirements such as sparsity, diversity, and interpretability of topics and to incorporate heterogeneous modalities such as authors, tags, and categories into the model. The parallelized online implementation of ARTM was used in open source library *BigARTM* (`bigartm.org`). The thematic search is implemented by maximizing cosine similarity between query and document both represented by their sparse distributions over topics. The authors evaluate precision and recall of the thematic search by a two-step procedure. First, human assessors perform exploratory search tasks manually using any available search utilities (it takes them about 30 min per task in average). Second, they evaluate the relevance of search results found by the present thematic search engine for the same tasks. The experiments on the collection of 132 000 articles from `habrahabr.ru` collective blog showed that thematic search provided comparable precision and better recall, also reducing search time from half an hour to seconds. With data labeled by assessors, the optimal number of topics was determined and it was shown that the joint use of all modalities (authors of articles, authors of comments, tags, and hub categories) significantly improves the search quality.

Keywords: *information retrieval; exploratory search; topic modeling; additive regularization for topic modeling; BigARTM*

DOI: 10.21469/22233792.2.2.04

References

[1] Marchionini, G. 2006. Exploratory search: From finding to understanding. *Commun. ACM* 49(4):41–46.

[2] White, R. W., and R. A. Roth. 2009. *Exploratory search: Beyond the query-response paradigm*. Morgan and Claypool Publs. 98 p.

[3] Kraaij, W., and W. Post. 2006. Task based evaluation of exploratory search systems. *SIGIR Workshop on Evaluating Exploratory Search Systems Proceedings*. ACM. 24–27.

[4] Potthast, M., M. Hagen, M. Völske, and B. Stein. 2013. Exploratory search missions for TREC topics. *EuroHCIR* 1033:7–10.

[5] Shah, C., C. Hendahewa, and R. Gonzalez-Ibanez. 2016. Rain or shine? Forecasting search process performance in exploratory search tasks. *J. Assoc. Inform. Sci. Technol.* 67(7):1607–1623.

*The research was supported by the Russian Foundation for Basic Research, grants 16-37-00498, 14-07-00847, and 14-07-00908.

[6] Hofmann, T. 1999. Probabilistic latent semantic indexing. *22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings*. New York, NY: ACM. 50–57.

[7] Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.

[8] Blei, D. M. 2012. Probabilistic topic models. *Commun. ACM* 55(4):77–84.

[9] Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to information retrieval*. New York, NY: Cambridge University Press. 504 p.

[10] Scherer, M., T. von Landesberger, and T. Schreck. 2013. Topic modeling for search and exploration in multivariate research data repositories. *Research and Advanced Technology for Digital Libraries: Conference (International) on Theory and Practice of Digital Libraries Proceedings*. Eds. T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, and C. J. Farrugia. Berlin–Heidelberg: Springer. 370–373.

[11] Grant, C. E., C. P. George, V. Kanjilal, S. Nirkhiwale, J. N. Wilson, and D. Z. Wang. 2015. A topic-based search, visualization, and exploration system. *FLAIRS Conference*. AAAI Press. 43–48.

[12] Rönnqvist, S. 2015. Exploratory topic modeling with distributional semantics. *14th International Symposium on Advances in Intelligent Data Analysis Proceedings*. Eds. E. Fromont, T. De Bie, and M. van Leeuwen. Saint Etienne, France: Springer International Publs. 241–252.

[13] Veas, E. E., and C. di Sciascio. 2015. Interactive topic analysis with visual analytics and recommender systems. *2nd Workshop on Cognitive Computing and Applications for Augmented Human Intelligence, Joint Conference (International) on Artificial Intelligence*. Aachen, Germany: CEUR-WS.org.

[14] Feldman, S. E. 2012. The answer machine. *Synthesis Lectures Inform. Concepts Retrieval Services* 4(3):1–137.

[15] Rahman, M. 2013. Search engines going beyond keyword search: A survey. *Int. J. Comput. Appl.* 75(17):1–8.

[16] Singh, R., Y.-W. Hsu, and N. Moon. 2013. Multiple perspective interactive search: A paradigm for exploratory search and information retrieval on the Web. *Multimedia Tools Appl.* 62(2):507–543.

[17] Jiang, T. 2014. Exploratory search: A critical analysis of the theoretical foundations, system features, and research trends. *Library and information sciences: Trends and research*. Eds. C. Chen and R. Larsen. Berlin–Heidelberg: Springer. 79–103.

[18] Marie, N., and F. Gandon. 2014. Survey of linked data based exploration systems. *3rd Workshop (International) on Intelligent Exploration of Semantic Data co-located with the 13th Semantic Web Conference (International) Proceedings*. Riva del Garda, Italy.

[19] Jacksi, K., N. Dimililer, and S. R. M. Zeebaree. 2015. A survey of exploratory search systems based on LOD resources. *5th Conference (International) on Computing and Informatics Proceedings*. Malaysia: School of Computing, Universiti Utara. 501–509.

[20] Mimno, D., M. Hoffman, and D. Blei. 2012. Sparse stochastic inference for latent Dirichlet allocation. *29th Conference (International) on Machine Learning Proceedings*. Eds. J. Langford and J. Pineau. New York, NY: Omnipress. 1599–1606.

[21] Bassiou, N., and C. Kotropoulos. 2014. Online PLSA: Batch updating techniques including out-of-vocabulary words. *IEEE Trans. Neural Networks Learning Systems* 25(11):1953–1966.

[22] Vorontsov, K., O. Frei, M. Apishev, P. Romov, M. Suvorova, and A. Yanina. 2015. Non-Bayesian additive regularization for multimodal topic modeling of large collections. *Workshop on Topic Models: Post-Processing and Applications Proceedings*. New York, NY: ACM. 29–37.

[23] Yi, X., and J. Allan. 2009. A comparative study of utilizing topic models for information retrieval. *Advances in information retrieval*. Berlin–Heidelberg: Springer. 5478:29–41.

[24] Andrzejewski, D., and D. Buttler. 2011. Latent topic feedback for information retrieval. *17th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings*. 600–608.

[25] Daud, A., J. Li, L. Zhou, and F. Muhammad. 2010. Knowledge discovery through directed probabilistic topic models: A survey. *Frontiers Comput. Sci. China* 4(2):280–301.

[26] Vorontsov, K. V. 2014. Additive regularization for topic models of text collections. *Dokl. Math.* 89(3):301–304.

[27] Frei, O., and M. Apishev. 2016. Parallel non-blocking deterministic algorithm for online topic modeling. *5th Conference (International) on Analysis of Images, Social Networks and Texts*.

[28] Vorontsov, K. V., and A. A. Potapenko. 2014. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. *Analysis images, social networks and texts*. Eds. D. I. Ignatov, M. Yu. Khachay, M. Y. Panchenko, *et al.* Communications in computer and information science ser. Springer International Publs. 436:29–46.

[29] Vorontsov, K. V., and A. A. Potapenko. 2015. Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications* 101(1):303–323.

[30] Vorontsov, K. V., A. A. Potapenko, and A. V. Plavin. 2015. Additive regularization of topic models for topic selection and sparse factorization. Ed. A. Gammerman. *3rd Symposium (International) on Learning and Data Sciences Proceedings*. University of London, U.K. Switzerland: Springer International Publs. 193–202.

[31] Apishev, M., S. Koltcov, O. Koltsova, S. Nikolenko, and K. Vorontsov. 2016. Additive regularization for topic modeling in sociological studies of user-generated text content. *15th Mexican Conference (International) on Artificial Intelligence Proceedings*.

[32] Tan, Y., and Z. Ou. 2010. Topic-weak-correlated latent Dirichlet allocation. *7th Symposium (International) Chinese Spoken Language Processing Proceedings*. 224–228.

Received September 2, 2016

Additive regularization for hierarchical multimodal topic modeling*

***N. A. Chirkova*[1,2] *and K. V. Vorontsov*[3]**
nadiinchi@gmail.com; vokov@forecsys.ru
[1]JSC Antiplagiat, 33 Nagatiskaya Str., Moscow, Russia
[2]Lomonosov Moscow State University, 1 Leninskie Gory, Moscow, Russia;
[3]Federal Research Center "Computer Science and Control" of RAS
44/2 Vavilova Str., Moscow, Russia

Probabilistic topic models uncover the latent semantics of text collections and represent each document by a multinomial distribution over topics. Hierarchical models divide topics into subtopics recursively, thus simplifying information retrieval, browsing and understanding of large multidisciplinary collections. The most of existing approaches to hierarchy learning rely on Bayesian inference. This makes difficult the incorporation of topical hierarchies into other types of topic models. The authors use non-Bayesian multicriteria approach called Additive Regularization of Topic Models (ARTM), which enables to combine any topic models formalized via log-likelihood maximization with additive regularization criteria. In this work, such formalization is proposed for topical hierarchies. Hence, the hierarchical ARTM (hARTM) can be easily adapted to a wide class of text mining problems, e. g., for learning topical hierarchies from multimodal and multilingual heterogeneous data of scientific digital libraries or social media. The authors focus on topical hierarchies that allow a topic to have several parent topics which is important for multidisciplinary collections of scientific papers. The regularization approach allows one to control the sparsity of the parent–child relation and automatically determine the number of subtopics for each topic. Before learning the hierarchy, it is necessary to fix the number of topics for each layer. The additive regularization does not complicate the learning algorithm; so, this approach is well scalable on large text collections.
Keywords: *topic modeling; ARTM; topic hierarchies; regularization*

DOI: 10.21469/22233792.2.2.05

1 Introduction

Topic modeling is a popular technique for semantic analysis of text collections. A probabilistic topic model defines each topic by a probability distribution over words and describes each document by a probability distribution over topics. In large text collections such as digital libraries or social media archives, the topics are usually organized in a hierarchy. Topic hierarchy helps user to navigate through the collection: going down the hierarchy, user chooses subtopics and finds a small subset of documents to read.

In last years, a lot of research was done about topic hierarchies learning. There is no common definition and common quality measure of topic hierarchy in the literature. Also, there is still no common hierarchy learning approach [1].

It is difficult to combine existing approaches with other modifications of topic models: spatiotemporal [2], short text [3], multilingual [4], multimodal [5], semisupervised [6], decorrelated [7], sparse [8], etc. On the other hand, there is a general approach for combining different types of topic models called additive regilarization of topic models [9, 10]. This framework is

*The research was supported by the Russian Foundation for Basic Research (grants 16-37-00498, 14-07-00847, and 14-07-00908).

well scalable for large collections [10] and is implemented in open-source topic modeling library BigARTM.

The goal of this work is to propose a method of learning topic hierarchies via topic model regularization and integrate it with ARTM.

Let us focus on hierarchies as multipartite (multilevel) directed acyclic graph of topics rather than a topic tree. While the last definition is a mainstream in literature, an assumption that a topic can inherit from several parent topics looks more reasonable. It is a common case in any field of knowledge when specific topic occurs on the edge of two or even more parent topics. For example, bioinformatics combines applied mathematics and computer science to solve the biology problems. This situation is called multiple inheritance. The presented approach supports multiple inheritance and controllable sparsing of topic graph and automatically determines the number of subtopics for each topic.

The reminder of the paper is organized as follows. In section 1, an overview of existing approaches for learning hierarchies is presented. In section 2, a formal problem statement is given and then, in section 3, the present authors' approach is described and in section 4, its implementation in BigARTM is presented. The last two sections are about experiments and discussion.

2 Related Work

Two basic topic modeling techniques are probabilistic latent semantic analysis (PLSA) [11] and its Bayesian extension latent Dirichlet allocation (LDA) [12]. A lot of LDA modifications were developed to meet applications tasks [13].

Additive regularization of topic models [10] is non-Bayesian extension that allows to impose additional, problem-specific criteria on topic model parameters. Many of LDA expansions can be interpreted as regularization criteria, this allows to combine several modifications in a single model.

In hierarchal models, the topics are linked by parent–child relations. Topic hierarchies are usually constructed in two ways: via generative model complication or as a combination of several tied flat models. Hierarchical LDA (hLDA) [14] and hierarchical Pachinko allocation model (hPAM) [15] are the examples of generative models. As other LDA extensions, these models are trained using time consuming Gibbs sampling that limits available collection size [16] and integration with other types of topic models. Hierarchical LDA is a tree structure and hPAM is a directed acyclic multilevel graph with no tools for edges number reduction.

The second group is split into top-down and bottom-up approaches. Tree structured hierarchies are often learned top-down recursively: firs, a flat model with few topics is learned and then, process repeats for each subtopic. SplitLDA splits documents between topics accordingly to the distribution over topics for each document–word pair [17]. Constructing A Topical HierarchY (CATHY) approach [18] operates with phrases rather than with words and divides them between subtopics. In Scalable and Robust Construction of Topic Hierarchies (STROD) [16], each topic distribution over words can be expanded to a mixture of suptopics distributions using tensor decomposition algorithm. The drawback of recursive approaches is that they need heuristics to determine the number of subtopics in each topic. On the other hand, recursive learning is usually fast, STROD is proven [16] to be the fastest on large collections.

Multiple inheritance supporting hierarchies are usually learned level by level. In [19], the hierarchy is learned in two steps: first, flat LDA models are learned for each level and next, topics between levels are linked using special subsumption criteria. An advantage is that changing a threshold in subsumption criteria controls the hierarchy sparsity. The disadvantage

is that specific topics are modeled independently from their parent topics. Also, the authors propose a simple agglomerative clustering based method for determining the number of topics in levels.

In [1], the hierarchy is constructed by bottom-up strategy. The last level of topics is learned first and then, these topics are treated as pseudodocuments and the next level model is learned from them. In this case, subtopic-pseudodocument proportions specify the topic graph structure and there is no ability to control the graph sparsity.

Almost all hierarchical topic models are based on Bayesian inference, it makes difficult to combine other topic model modifications with hierarchy. The present authors propose a top-down hierarchy learning framework based on ARTM that incorporates few reasonable ideas from other approaches.

3 Problem Statement

Let D denote the text collection. Documents $d \in D$ may contain not only words but also other elements such as tags, links, location marks, etc. Let us refer to such types of elements as modalities. For example, a scientific paper usually contains three modalities: text, keywords, and references. Let M denote a set of all modalities in the collection. Modalities $m \in M$ are defined by disjoint dictionaries $W = \bigsqcup_{m \in M} W^m$.

A document $d \in D$ is a sequence of n_d elements: $(w_1, w_2, w_3, \dots)$, $w_i \in W$. In this paper, an order of elements is not important. Thu, collection can be represented as a counters matrix $\{n_{dw}\}_{D \times W}$ where n_{dw} is the number of w occurencies in d.

Given the text collection, the goal is to organize its documents into comprehensive hierarchical structure. Let us define a *topic hierarchy* as an oriented multiparticle (multilevel) acyclic graph of topics so that the edges connect the topics from the neighboring levels. If there is an edge $a \to t$ in the hierarchy, then the topic a is called *parent*, or *ancestor*, topic and t is called *child topic*, or *subtopic*. The parent topic is divided into several more specific child topics. The number of topics in each following (child) level is usually greater than in the previous (parent) level. Zero level consists of only one topic called *root*. An example of topic hierarchy is given in Fig. 1.

Each topic in the hierarchy is associated with distributions over each modality dictionary. This allows one to represent a topic by a top of most probable words saying what this topic is about. The same can be done with other modalities.

To construct the hierarchy, let us learn several flat topic models and tie them via regularization.

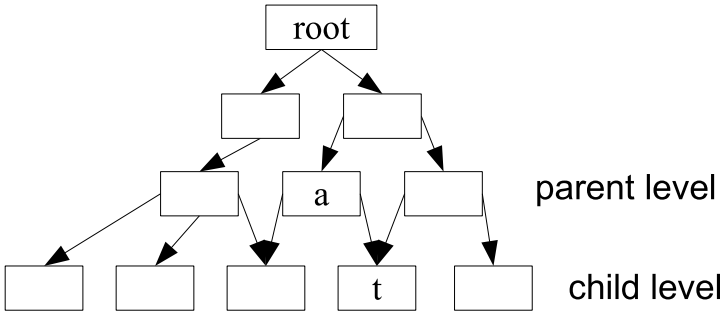


Figure 1 An example of topic hierarchy

In the rest of the paper, an operator $\underset{i\in I}{\text{norm}}[y_i] = \max\{y_i, 0\}/\sum_{i'\in I}\max\{y_{i'}, 0\}$ transforming real vector $(y_i)_{i\in I}$ to a probability distribution is used.

4 hARTM framework

4.1 ARTM: flat topic models

The flat topic model describes collection D by finite topics set T. In ARTM [10], document distribution over each modality is modeled as a mixture of topic distributions:

$$p(w|d) \approx \sum_{t\in T} p(w|t)p(t|d) \quad d \in D,\, w \in W^m.$$

In other words, for each modality m, the topic model is a low-rank approximation

$$F^m \approx \Phi^m\Theta$$

of the frequency matrix $F^m = \{f_{wd}\}_{W^m\times D}$ where $f_{wd} = \underset{w\in W^m}{\text{norm}}[n_{dw}]$ is the frequency $p(w|d)$. The model parameters are the matrices $\Phi^m = \{\varphi_{wt}\}_{W^m\times T}$ with $\varphi_{wt} = p(w|t)$ and $\Theta = \{\theta_{td}\}_{T\times D}$ with $\theta_{td} = p(t|d)$, Φ and Θ being the stochastic matrices:

$$\sum_{w\in W^m} \varphi_{wt} =; \quad \sum_{t\in T} \theta_{td} = 1. \tag{1}$$

For brevity, let is denote vertically stacked Φ^m and F^m, $m \in M$, by Φ and F, respectively. Then, the topic model in an approximate matrix factorization $F \approx \Phi\Theta$.

Let us maximize the weighted sum of modality log-likelihoods and regularizers R_i to learn Φ and Θ :

$$\sum_{m\in M} \varkappa_m \sum_{d\in D}\sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta}. \tag{2}$$

Weights $\varkappa_m$ are used to balance log-likehood of modalities. Regularizers R_i impose additional problem-specific criteria on the model parameters. Regularizer coefficients τ_i balance the importance of regularizers and log-likehoods. If the regularizer term $R = \sum_i \tau_i R_i(\Phi,\Theta)$ equals zero and there is only text modality, then described model simplifies to PLSA.

Theorem 1 (see [10]). *If all regularizers are continuously differentiable on Φ and Θ, then the stationary point of the problem* (2) *with constrains* (1) *satisfies the following system yielding expectation-maximization (EM) algorithm for model training:*

$$\left.\begin{aligned}
&\textit{E-step}: \quad p(t|d,w) = \underset{t\in T}{\text{norm}}[\varphi_{wt}\theta_{td}], \quad w \in W,\, d \in D;\\
&\textit{M-step}: \quad \varphi_{wt} = \underset{w\in W^m}{\text{norm}}\left[n_{wt} + \frac{\partial R}{\partial \varphi_{wt}}\varphi_{wt}\right], \quad n_{wt} = \sum_{d\in D} n_{dw}p(t|d,w),\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad w \in W^m, t \in T, m \in M;\\
&\qquad\quad \theta_{td} = \underset{t\in T}{\text{norm}}\left[n_{td} + \frac{\partial R}{\partial \theta_{td}}\theta_{td}\right], \quad n_{td} = \sum_{w\in W} n_{dw}p(t|d,w), \quad t \in T, d \in D.
\end{aligned}\right\} \tag{3}$$

The EM-algorithm is obtained by applying the fixed point iteration method to the system. Matrices Φ and Θ are initialized randomly.

Sparsing regularizers. Frequently used sparsing regularizer [10] causes distributions $p(w|t)$ and $p(t|d)$ to be sparse meaning the majority of distribution domain elements have zero probability. To do this, Kullback–Leibler divergence between specified distribution $\boldsymbol{\alpha}$, usually uniform, and target distribution is maximized. For instance, Θ-sparsing regularizer:

$$\sum_{d\in D} KL(\boldsymbol{\alpha}\|\boldsymbol{\theta}_d) \to \max_{\Theta} \quad \Leftrightarrow \quad R_1(\Theta) = -\sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td} \to \max_{\Theta},$$

$\boldsymbol{\theta}_d$ denotes Θ column and for uniform distribution, $\alpha_t = 1/|T|$. Similarly, for Φ^m sparsing with uniform specified distribution,

$$R_2(\Phi^m) = -\sum_{t\in T}\sum_{w\in W^m} \frac{1}{|W^m|} \ln \varphi^m_{wt} \to \max_{\Phi^m}.$$

Modified M-step formulas for parameters update:

$$\varphi_{wt} = \underset{w\in W^m}{\operatorname{norm}}\left[n_{wt} - \frac{\tau_1}{|W^m|}\right]; \quad \theta_{td} = \underset{t\in T}{\operatorname{norm}}\left[n_{td} - \frac{\tau_2}{|T|}\right]. \tag{4}$$

Hyperparameters of flat topic model are number of topics $|T|$, weights $\{\varkappa_m\}_{m\in M}$, and regularization coefficients $\{\tau_i\}_i$. While learning topic hierarchy, flat topic model is trained for each level of hierarchy, every time with new hyperparameters settings.

4.2 hARTM: Top-down hierarchy learning

Since topic hierarchy is a multilevel graph, let us consider each level as a flat topic model. The authors propose top-down, level by level hierarchy learning algorithm. Zero level is associated with the whole collection. The first level contains small number of major topics. Starting from the second level, it is necessary not only to model the topics, but also to establish parent–child topic relations. To do this, the authors introduce two additional matrix factorization problems and propose two new interchangeable regularizers based on them.

Assume one has already learned $\ell \geqslant 1$ hierarchy levels. Now, let us learn $(\ell+1)$th level that is a child level for the ℓth ancestor level. Not to confuse levels, let us denote parent level topics $a \in A$ and parameters Φ^ℓ and Θ^ℓ instead of $t \in T$, Φ, and Θ used for child level. Note that Φ^ℓ and Θ^ℓ are already modeled.

Φ interlevel regularizer. Let us model parent topic distribution over words and other modalities as a mixture of child topics distributions:

$$p(w|a) = \sum_{t\in T} p(w|t)p(t|a), \quad w \in W^m,\ a \in A.$$

This means an approximation

$$\Phi^\ell \approx \Phi\Psi \tag{5}$$

with new parameters matrix $\Psi = \{\psi_{ta}\}_{T\times A}$, $\psi_{ta} = p(t|a)$ containing *interlevel distributions* of children topics t in parent topics a. This gives the following regularizaion criteria:

$$\sum_{a\in A} n_a\, KL(\boldsymbol{\varphi}^{\ell,m}_a \| \Phi^m\, \boldsymbol{\psi}_a) \to \min_{\Phi^m,\Psi}$$

or, equivalently,

$$R_3(\Phi^m, \Psi) = \sum_{a\in A}\sum_{w\in W^m} n_{wa} \ln \sum_{t\in T} \varphi_{wt}\psi_{ta} \to \max_{\Phi^m,\Psi},$$

$\boldsymbol{\varphi}_a^{\ell,m}$ and $\boldsymbol{\psi}_a$ denote columns of $\Phi^{\ell,m}$ and Ψ, respectively. Weights $n_a = \sum_{w \in W^m} n_{wa}$ are imposed to balance parent opics proportionally to their size and to scale regularization criteria up to the log-likelihood, n_{wa} being the parent topic counters from the EM-algorithm. Regularizer criterias are weighted by the modality weights:

$$R_3(\Phi, \Psi) = \sum_{m \in M} \varkappa_m R_3(\Phi^m, \Psi).$$

This regularizer is equivalent to adding $|A|$ pseudodocuments represented by $\{n_{wa}\}_{W \times A}$ columns. Then, the matrix Ψ forms $|A|$ additional columns to the matrix Θ corresponding to pseudodocuments. Note than child level could not be trained only on pseudodocuments because internal dimension in approximation (5) is higher than the minimum dimension of Φ^ℓ and Φ will just copy columns of Φ^ℓ.

Θ interlevel regularizer. The same idea can be applied for regularizing Θ instead of Φ. Then, for each document, distribution over parent topics is modeled by the mixture of topic distributions:

$$p(a|d) = \sum_{t \in T} p(a|t) p(t|d).$$

Additional matrix approximation looks like

$$\Theta^\ell \approx \widetilde{\Psi} \Theta$$

with interlevel distributions $\widetilde{\Psi} = \{\tilde{\psi}_{at}\}_{A \times T}$, $\tilde{\psi}_{at} = p(a|t)$. This means that parent topic's documents set is a union of children's documents sets. Regularizer criteria is

$$R_4(\Theta, \widetilde{\Psi}) = \sum_{a \in A} \sum_{d \in D} \theta_{ad}^\ell \ln \sum_{t \in T} \tilde{\psi}_{at} \theta_{td} \to \max_{\widetilde{\Psi}, \Theta}.$$

To train child model with the regularizer, let us add a new modality $\tilde{m}$ corresponding to parent topics and consider document counters for this modality θ_{ad}^ℓ. The Θ-regularizer coefficient becomes the modality weight and $\widetilde{\Psi}$ corresponds to the matrix $\Phi^{\tilde{m}}$.

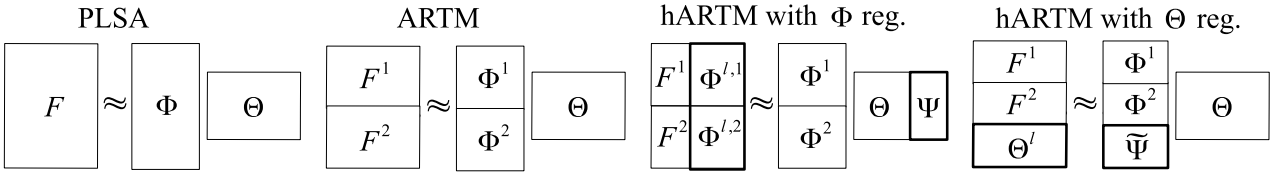


Figure 2 An illustration of child level regularization

An illustration of manipulating with pseudodocuments and new modality while the regularization of child level is given in Fig. 2.

Hierarchy sparsing regularizers. When the topics are allowed to inherit from a number of parents, it is assumed that this number will not be large, i. e., 1–3, rarely greater parents. Such hierarchy is called the *sparse* one. In other words, we want distributions $p(a|t)$ to be sparse. The regularization allows us to achieve this requirement.

Since in Θ interlevel regularization approach $\widetilde{\Psi}$ is a child $\Phi^{\tilde{m}}$ and its columns represent distributions $p(a|t)$, one can use Φ-sparsing regularizer described above to make the hierarchy

sparse. Let us rewrite (4) replacing $\varphi \to \tilde{\psi}$, $w \to a$, and $W^m \to A$ to show how $\widetilde{\Psi}$ updates on each iteration:

$$\tilde{\psi}_{at} = \underset{a \in A}{\text{norm}}\left[n_{at} - \frac{\tau_1}{|A|}\right].$$

In case of Φ interlevel regularization, Ψ columns represent $p(t|a)$ distributions that can be converted to $p(a|t)$ using Bayes formula. Following the idea of other parsing regularizers, let us maximize KL-divergence between uniform distribution $\boldsymbol{\gamma} = \{1/|A|\}_{a \in A}$ and the arget one $\tilde{\boldsymbol{\psi}}_t = \{p(a|t)\}_{a \in A}$:

$$\sum_{t \in T} \mathrm{KL}\,(\boldsymbol{\gamma} \| \tilde{\boldsymbol{\psi}}_t) \to \max_{\Psi}$$

or, equivalently,

$$R_5(\Psi) = \sum_{t \in T} \sum_{a \in A} \frac{1}{|A|} \ln p(a|t) = \frac{1}{|A|} \sum_a \sum_t \ln \frac{\psi_{ta}\, p(a)}{\sum_{a'} \psi_{ta'}\, p(a')} \to \min_{\Psi}.$$

Probabilities $p(a)$ are counted from Θ^ℓ.

To show how Ψ updates, let us rewrite M-step formula in (**??**) replacing $\theta \to \psi$ and $d \to a$ and taking derivatives of $R_5(\Psi)$ with respect to ψ_{ta}:

$$\psi_{ta} = \underset{t \in T}{\text{norm}}\left[n_{ta} - \tau_5\left(\frac{1}{|A|} - p(a|t)\right)\right].$$

For each topic t, parent topics a with high $p(a|t)$ get higher and parents with low $p(a|t)$ get lower. Note that R_5 cannot zeroize all components of Ψ column whereas R_1 can do this with $\widetilde{\Psi}$ column.

Hierarchy learning scenario. Thus, hyperparameters of topic hierarchy are the number of levels, the number of topics on each level, modalities weights, and regularization coefficients. One can learn hierarchy level by level, on each level finding parents for topics from previous level using Φ or Θ interlevel regularizer. If sparse hierarchy is desired, hierarchy sparsing regularizer should also be used. The process of training levels is stopped when the topics on the last level are highly specialized.

Regularization coefficients may be tuned for each level individually or used the same for all levels. Note that when learning the $(\ell + 1)$th level, only ℓth level's topics are used for regularization, not all previous levels' topics.

When hierarchy is learned, the topics on each level are represented by their distributions over words and other modalities. The documents on each level are assigned to several topics with proportions specified in this level's Θ matrix. The hierarchy structure is defined by interlevel distributions. To draw the topic graph, one may impose a threshold on $p(a|t)$ or $p(t|a)$.

5 Implementation in BigARTM

BigARTM is an open-source topic modeling library with C++ kernel [20]. BigARTM provides command line, C++ and python interfaces, and rich built-in library with regularizers and scores. BigARTM takes multimodal input data in a range of formats and transforms it into a series of *batches*, internal format. All batches store about the same number of documents, each batch is assigned a float weight (default 1.0).

BigARTM provides offline and online multithread learning algorithms. Offline algorithm performs a number of scans over the entire collection. During one scan, each thread processes one batch at a time, calculating n_{td} and θ_{td} (applying Θ-regularizers) and contributing local, batch-specific n_{wt} multiplied by batch weight to global n_{wt} counters. After that, the scan algorithm applies Φ-regularizers to global n_{wt} and normalizes them to calculate Φ. Online algorithm improves the convergence rate by recalculating Φ after every portion of batches.

The hierarchy learning is implemented as a wrapper over library interface without changing the kernel. To use Φ interlevel regularizer, an additional batch is created from parent Φ matrix, the weight of this batch equals to regularization coefficient. This parent batch is appended to the collection batches during the learning of child level, it does not affect algorithm efficiency.

To use Θ interlevel regularizer during child level learning, each batch should be appended the new modality corresponding to current batch parent Θ. This is time consuming operation. In experiments, it will be shown that two proposed interlevel regularizers are interchangeable; so, there is no need to use ineffective algorithm.

The Ψ sparsing regularizer is implemented as usual Θ regularizer since Ψ is the parent batch Θ.

6 Experiments

In this section, two proposed interlevel regularizers will be compared and the properties of the present hierarchy construction method will be studied.

Datasets. Let us run the experiments on two text datasets:

1) English Wikipedia dump (08.12.2014): $|D| = 3665223$ and $|W| = 100000$ after lemmatization and filtering words by frequency; and
2) dump of http://postnauka.ru site (scientific lectures in Russian): $|D| = 1728$ and $|W| = 38467$ after lemmatization.

Let us use only the text modality.

Regularizers comparison. Since both proposed regularizers impose additional matrix factorization task, compare the quality of this approximation varying $|W|/|D|$ proportion. Let us measure Hellinger distance between two stochastic matrices $A_{n\times m}$ and $B_{n\times m}$:

$$H(A,B) = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{a_{ij}} - \sqrt{b_{ij}}\right)^2}.$$

Two-level hierarchy was learned with 50 and 250 topics in each level on Wikipedia subset $D' \subset D$ several times. For $|D'| = 1$, 10, 50, 100, and 200, second level was learned twice: with Φ and Θ interlevel regularizer, respectively. For each run, $H(\Phi^\ell, \Phi\Psi)$ and $H(\Theta^\ell, \widetilde{\Psi}\Theta)$, $\ell = 1$, were measured. Coefficients of interlevel regularizers are set so that Θ^ℓ is approximated at the same rate with both regularizers, there is no hierarchy sparsing. The results are given in Fig. 3.

The graphic shows that with described coefficients, setup strategy Φ^ℓ is also approximated at the same rate for any $|W|/|D|$ proportion. In other words, both regularizers approximate both matrices Φ^ℓ and Θ^ℓ. Moreover, Φ-regularizer approximates both matrices a bit better than its counterpart. Also, remember Φ-regularizer allows more efficient realization. Hence, the authors recommend to use it instead of Θ-regularizer. Let us run the following experiments with Φ interlevel regularizer.

Children number study. One can trust children topics number estimated by the proposed method if these estimates are robust. In this experiment, one can see how the number of children topics and its deviation depend on hierarchy sparsing regularizer coefficient τ_5.

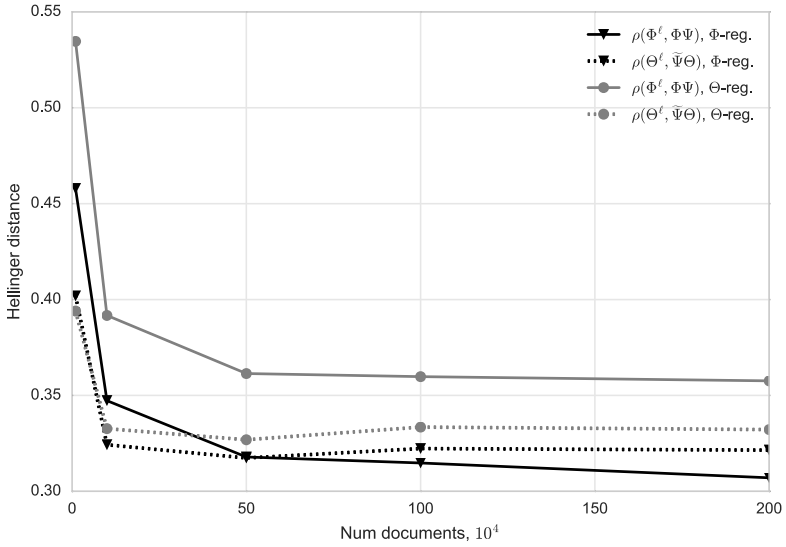


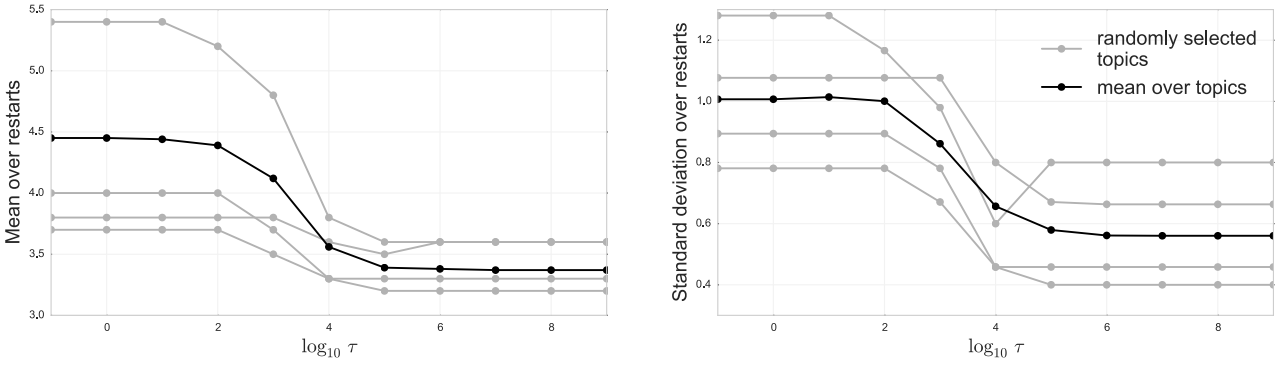
Figure 3 Interlevel regularizers comparison

Postnauka hierarchy was learned with two levels T_1 and T_2, $|T_1| = 10$ and $|T_2| = 30$. The first level was modeled once and fixed. Then, for each $\tau = \tau_5 = 0.1, 1, \ldots, 10^9$, second level with 10 restarts was learned from different random initializations. The mean m_t^τ and the standard deviation v_t^τ of children number were counted for each topic $t \in T_1$ and they were averaged over children topics:

$$m^\tau = \frac{1}{|T^1|} \sum_{t \in T^1} m_t^\tau; \quad v^\tau = \frac{1}{|T^1|} \sum_{t \in T^1} v_t^\tau.$$

We set a threshold on ψ_{ta} as maximum so that the hierarchy is still a connected graph.

Figure 4 shows that the number of children topics and its variance falls with certain hierarchy sparsing regularizer coefficient τ. For some topics, there is global minimum of m_t^τ and v_t^τ in $\tau_* = 10^4$ or 10^5. For $\tau \ll \tau_*$, regularizer affects Ψ weakly; for $\tau \gg \tau_*$, it zeroizes some ψ_{ta}

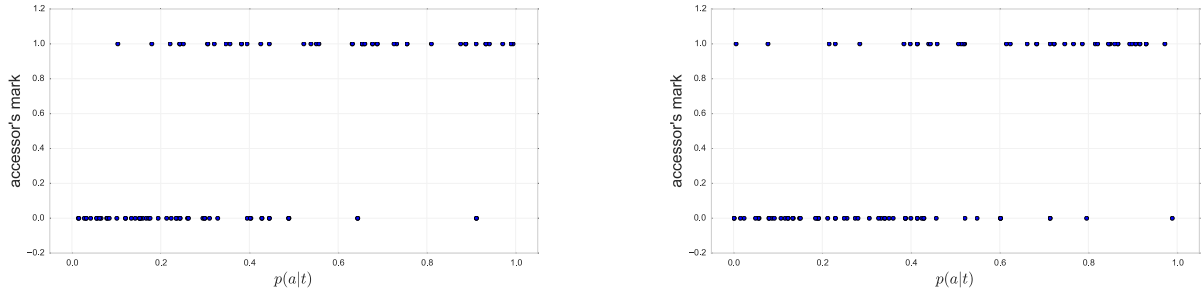


(a) (b)

Figure 4 Children number study. Dependences m_t^τ vs. τ (a) and v_t^τ vs. τ (b) for 4 randomly selected topics are in gray, black line displays m^τ vs. τ and v^τ vs. τ

immediately after initialization and hides real parent–child relations. The variance of children number is small (0.55–1), it proves that this estimate is robust. Mean parents number that equals $(|T_1|/|T_2|)m^\tau = 1.16$ for $\tau = 10^4$ is also small showing that the topic graph is close to a tree.

Parent–child relations study. In this experiment, see how well probabilities $p(a|t)$ reflect the existence of parent–child relations. We constructed three-level Postnauka hierarchy with 10, 30, and 90 topics in each level and chose 100 random pairs a–t. We asked an expert to mark each pair as interpretable (relation exists) or uninterpretable. Figure 5 shows the scatter of expert mark vs. $p(a|t)$.



(a) (b)

Figure 5 The difference in $p(a|t)$ between pairs a–t with 0 and 1 expert marks: (*a*) no sparsing and (*b*) with hierarchy sparsing

The bias between 0 and 1-marked pairs is greater for the sparse hierarchy. Although there is no explicit gap, one can impose a threshold on $p(a|t)$ so that the majority of 0- and 1-marked pairs are determined right.

Few randomly selected braches of a topic hierarchy learned from Wikipedia are shown in Fig 6.

7 Discussion

In this paper, a method of constructing topic hierarchy via regularization of the flat topic model is proposed. An experiment showed that both described regularizers do the same work; so, the more efficient one has been chosen. An idea of this regularizer is based on the assumption that a parent topic is a mixture of children topics. Some other works [16] make this assumption as well.

The authors suggest to learn hierarchy top down, level by level, not the whole hierarchy at once. Thus, the quality of topics is controled on the higher levels before splitting them into subtopics and the hierarchy is preserved from having uninterpretable branches. While other top-down approaches are recursive and split each topic node into subtopics, the suggested algorithm determines parent–child relations during child level learning and allows topics to have more than one parents. At the same time, it determines children number of each parent topic. An experiment shows that these estimates are robust. To the best of our knowledge, it is the first top-down approach with multiple inheritance support.

An open question is how to specify the number of topics on each level. To do this, one can apply the clustering technique proposed in [19]. Another way is to use a topic selection regularizer [21] that chooses as possible linearly independent topics from certainly excess topics

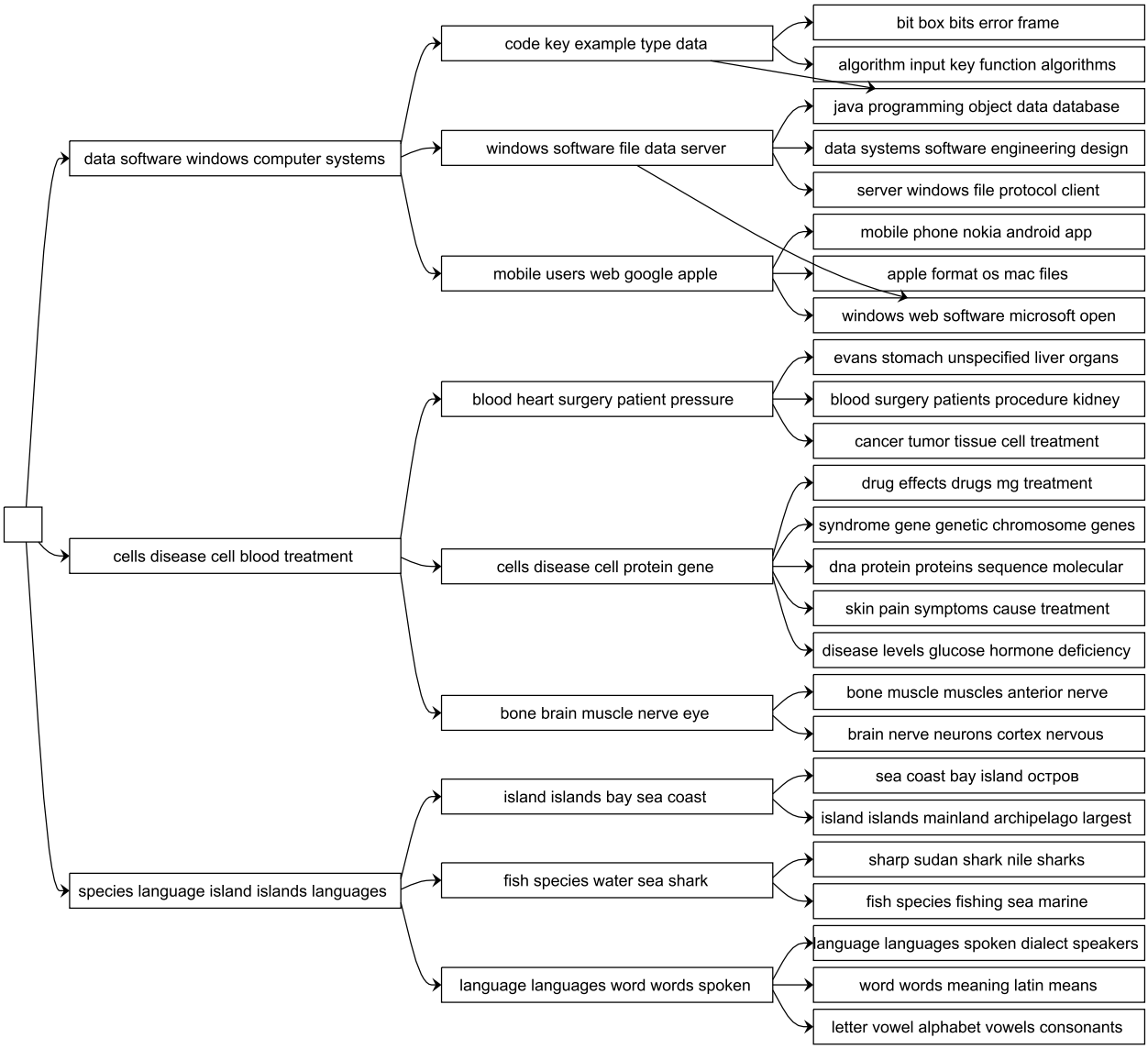


Figure 6 A part of Wikipedia hierarchy

set. The regularizer coefficients and modalities weights are usually tuned to maximize particular criteria or visual hierarchy interpretability.

References

[1] Zavitsanos, E., G. Paliouras, and G. A. Vouros. 2011. Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes. *J. Mach. Learn. Res.* 12:2749–2775.

[2] Blei, D. M., and J. D. Lafferty. 2006. Dynamic topic models. *23rd Conference (International) on Machine Learning Proceedings*. New York, NY: ACM. 113–120.

[3] Yan, X., J. Guo, Y. Lan, and X. Cheng. 2013. A biterm topic model for short texts. *22nd Conference (International) on World Wide Web Proceedings*. New York, NY: ACM. 1445–1456.

[4] Mimno, D., H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. 2009. Polylingual topic models. *Conference on Empirical Methods in Natural Language Processing Proceedings*.

Stroudsburg, PA: Association for Computational Linguistics. 2:880–889. Available at: http://dl.acm.org/citation.cfm?id=1699571.1699627 (accessed December 22, 2016).

[5] Virtanen, S., Y. Jia, A. Klami, and T. Darrell. 2012. Factorized multi-modal topic model. *28th Conference on Uncertainty in Artificial Intelligence Proceedings*. Eds. N. de Freitas and K. Murphy. Corvallis, OR: AUAI Press. 843–851.

[6] Wang, D., M. Thint, and A. Al-Rubaie. 2012. Semi-supervised latent Dirichlet allocation and its application for document classification. *IEEE/WIC/ACM Conferences (International) on Web Intelligence and Intelligent Agent Technology*. 3:306–310.

[7] Blei, D. M., and J. D. Lafferty. 2006. Correlated topic models. *23rd Conference (International) on Machine Learning Proceedings*. MIT Press. 113–120.

[8] Than, K., and T. B. Ho. 2012. Fully sparse topic models. *European Conference on Machine Learning and Knowledge Discovery in Databases Proceedings*. Berlin–Heidelberg: Springer-Verlag. I:490–505.

[9] Vorontsov, K. V. 2014. Additive regularization for topic models of text collections. *Dokl. Math.* 89(3):301–304.

[10] Vorontsov, K., O. Frei, M. Apishev, P. Romov, M. Suvorova, A. Yanina. 2015. Non-Bayesian additive regularization for multimodal topic modeling of large collections. *Workshop on Topic Models: Post-Processing and Applications Proceedings*. New York, NY: ACM. 29–37.

[11] Hofmann, T. 1999. Probabilistic latent semantic indexing. *22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings*. New York, NY: ACM. 50–57.

[12] Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

[13] Blei, D. M., and J. D. Lafferty. 2009. Topic models. *Text mining: Classification, clustering, and applications*. Eds. A. N. Srivastava and M. Sahami. Chapman & Hall/CRC data mining and knowledge ser. CRC Press. 71–94.

[14] Blei, D. M., T. Griffiths, M. I. Jordan, and J. B. Tenenbaum. 2003. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*. Eds. S. Thrun, L. K. Saul, and P. B. Schölkopf. NIPS. 8 p.

[15] Mimno, D., W. Li, and A. McCallum. 2007. Mixtures of hierarchical topics with Pachinko allocation. *24th Conference (International) on Machine Learning Proceedings*. ACM. 633–640.

[16] Wang, C., X. Liu, Y. Song, and J. Han. 2014. Scalable and robust construction of topical hierarchies. arXiv:1403.3460.

[17] Pujara, J., and P. Skomoroch. 2012. Large-scale hierarchical topic models. *NIPS Workshop on Big Learning*.

[18] Wang, C., M. Danilevsky, N. Desai, *et al.* 2013. A phrase mining framework for recursive construction of a topical hierarchy. *19th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings*. New York, NY: ACM. 437–445.

[19] Srivastava, N. 2010. Learning size and structure of document ontologies using generative topic models. Available at: http://www.cs.toronto.edu/∼nitish/iitk_page/updated_cs397.pdf (accessed December 22, 2016).

[20] BigARTM. Available at: http://bigartm.org (accessed December 22, 2016).

[21] Vorontsov, K., and A. Potapenko. 2015. Additive regularization of topic models. *Machine Learn.* 101(1):303–323. doi: http://dx.doi.org/10.1007/s10994-014-5476-6.

Received September 3, 2016

Аддитивная регуляризация мультимодальных иерархических тематических моделей*

Н. А. Чиркова[1,2]***, К. В. Воронцов***[3]
nadiinchi@gmail.com; vokov@forecsys.ru
[1]ЗАО Антиплагиат, Россия, г. Москва, ул. Нагатинская, 33
[2]МГУ им. М. В. Ломоносова, Россия, г. Москва, Ленинские горы, 1
[3]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Вероятностные тематические модели выявляют семантику текстовых коллекций, описывая каждый документ дискретным распределением вероятностей на множестве тем. Иерархические модели рекурсивно делят темы на подтемы, что упрощает информационный поиск и навигацию по большим мультидисциплинарным коллекциям. В большинстве работ по иерархическому тематическому моделированию применяется байесовский вывод, что затрудняет введение тематических иерархий в тематические модели других видов. Не-байесовская аддитивная регуляризация тематических моделей, наоборот, позволяет комбинировать любые тематические модели, если их специфические особенности формализуемы в виде критериев-регуляризаторов. Однако до сих пор иерархические модели не имели такой формализации. Предлагаются регуляризаторы тематических иерархий, адаптируемые для широкого класса задач, в частности для тематизации мультимодальных и мультиязычных данных научных электронных библиотек и социальных сетей. Рассматриваются иерархии, в которых каждая подтема может иметь несколько родительских, что особенно актуально для междисциплинарных коллекций научных статей. Предлагаемый подход позволяет контролировать разреженность отношения тема–подтема и автоматически определять число подтем каждой темы. При построении модели задается только число тем на каждом уровне иерархии. Аддитивная регуляризация не усложняет процесс обучения тематической модели, что делает данный подход масштабируемым на большие текстовые коллекции.

Ключевые слова: *тематическое моделирование; APTM; тематические иерархии; регуляризация*

DOI: 10.21469/22233792.2.2.05

Литература

[1] *Zavitsanos E., Paliouras G., Vouros G. A.* Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes // J. Mach. Learn. Res., 2011. Vol. 12. P. 2749–2775.

[2] *Blei D. M., Lafferty J. D.* Dynamic topic models // 23rd Conference (International) on Machine Learning Proceedings. — New York, NY, USA: ACM, 2006. P. 113–120.

[3] *Yan X., Guo J., Lan Y., Cheng X.* A biterm topic model for short texts // 22nd Conference (International) on World Wide Web Proceedings. — New York, NY, USA: ACM, 2013. P. 1445–1456.

[4] *Mimno D., Wallach H. M., Naradowsky J., Smith D. A., McCallum A.* Polylingual topic models // Conference on Empirical Methods in Natural Language Processing Proceedings. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. Vol. 2. P. 880–889. URL: http://dl.acm.org/citation.cfm?id=1699571.1699627.

*Работа выполнена при финансовой поддержке РФФИ, проекты №№ 16-37-00498, 14-07-00847 и 14-07-00908.

[5] *Virtanen S., Jia Y., Klami A., Darrell T.* Factorized multi-modal topic model // 28th Conference on Uncertainty in Artificial Intelligence Proceedings / Eds. N. de Freitas, K. Murphy. — Corvallis, OR, USA: AUAI Press, 2012. P. 843–851.

[6] *Wang D., Thint M., Al-Rubaie A.* Semi-supervised latent Dirichlet allocation and its application for document classification // IEEE/WIC/ACM Conferences (International) on Web Intelligence and Intelligent Agent Technology, 2012. Vol. 3. P. 306–310.

[7] *Blei D. M., Lafferty J. D.* Correlated topic models // 23rd Conference (International) on Machine Learning Proceedings, 2006. P. 113–120.

[8] *Than K., Ho T. B.* Fully sparse topic models // European Conference on Machine Learning and Knowledge Discovery in Databases Proceedings. — Berlin–Heidelberg: Springer-Verlag, 2012. Part I. P. 490–505.

[9] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Докл. Акад. наук, 2014. Т. 455. №3. С. 268–271.

[10] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A.* Non-Bayesian additive regularization for multimodal topic modeling of large collections // Workshop on Topic Models: Post-Processing and Applications Proceedings. — New York, NY, USA: ACM, 2015. P. 29–37.

[11] *Hofmann T.* Probabilistic latent semantic indexing // 22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings. — New York, NY, USA: ACM, 1999. P. 50–57.

[12] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // J. Mach. Learn. Res., 2003. Vol. 3. P. 993–1022.

[13] *Blei D. M., Lafferty J. D.* Topic models // Text mining: classification, clustering, and applications / Eds. A. N. Srivastava, M. Sahami. — Chapman & Hall/CRC data mining and knowledge ser. — CRC Press, 2009. P. 71–94.

[14] *Blei D. M., Griffiths T., Jordan M. I., Tenenbaum J. B.* Hierarchical topic models and the nested chinese restaurant process // Advances in neural information processing systems / Eds. S. Thrun, L. K. Saul, P. B. Schölkopf. — NIPS, 2003. 8 p.

[15] *Mimno D., Li W., McCallum A.* Mixtures of hierarchical topics with Pachinko allocation // 24th Conference (International) on Machine Learning Proceedings. — ACM, 2007. P. 633–640.

[16] *Wang C., Liu X., Song Y., Han J.* Scalable and robust construction of topical hierarchies // CoRR, 2014. arXiv:1403.3460.

[17] *Pujara J., Skomoroch P.* Large-scale hierarchical topic models // NIPS Workshop on Big Learning, 2012.

[18] *Wang C., Danilevsky M., Desai N., et al.* A phrase mining framework for recursive construction of a topical hierarchy // 19th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings. — New York, NY, USA: ACM, 2013. P. 437–445.

[19] *Srivastava N.* Learning size and structure of document ontologies using generative topic models, 2010. http://www.cs.toronto.edu/~nitish/iitk_page/updated_cs397.pdf.

[20] BigARTM. `http://bigartm.org`.

[21] *Vorontsov K., Potapenko A.* Additive regularization of topic models // Machine Learning, 2015. Vol. 101. No. 1. P. 303–323. doi: `http://dx.doi.org/10.1007/s10994-014-5476-6`.

Поступила в редакцию 03.09.2016

Параметрический подход к построению синтаксических деревьев для частично формализованных текстовых документов*

К. В. Чувилин[1,2]
kirill@chuvilin.pro
[1]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
[2]ИФТИ, Россия, Московская область, г. Протвино, Заводской проезд, 6

Данная работа посвящена исследованию возможности автоматического построения логической структуры (абстрактного синтаксического дерева) для текстовых документов, формат которых не является полностью определенным стандартами или другими общими для всех документов правилами. В отличие от синтаксисов, описываемых формальными грамматиками, в таких случаях нет возможности в автоматическом режиме построить синтаксический анализатор. Типичными примерами таких форматированных документов с не полностью формализованным синтаксисом разметки являются текстовые файлы в формате LaTeX. В данной работе они используются как ресурсы для практической реализации разрабатываемых алгоритмов. Актуальность анализа именно LaTeX-документов обусловлена тем, что многие научные издательства и конференции используют систему верстки LaTeX, и это порождает важные прикладные задачи по автоматизации рубрикации, коррекции, сравнения, сбора статистики, отображения для WEB и т. п. При синтаксическом анализе документов в формате LaTeX требуется дополнительная информация о стилях: символах, командах и окружениях. В данной работе предлагается метод их описания в формате JSON, который позволяет задавать не только информацию, необходимую для синтаксического анализа, но и метаинформацию, упрощающую дальнейший интеллектуальный анализ. Такой подход использован впервые. Описываются разработанные алгоритмы построения синтаксического дерева документа в формате LaTeX, использующие такую информацию как внешний параметр. Полученные результаты успешно применены в задачах сравнения, автоматической коррекции и рубрикации научных статей. Реализация разработанных алгоритмов доступна в виде набора библиотек, распространяемых по лицензии LGPLv3. Ключевыми особенностями предлагаемого подхода являются гибкость (в рамках рассматриваемой задачи) и простота описания параметров. Предложенные подходы позволяют решить задачу синтаксического анализа документов в формате LaTeX. Но для широкого практического использования разработанных алгоритмов требуется сформировать базу описаний элементов стилей.

Ключевые слова: *абстрактное синтаксическое дерево; дерево; интеллектуальный анализ текстов; синтаксический анализ; JSON; LaTeX*

DOI: 10.21469/22233792.2.2.06

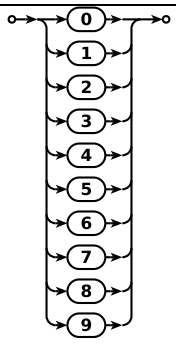
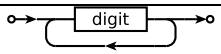
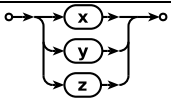
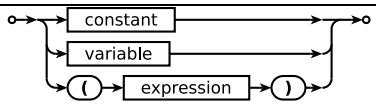
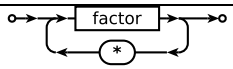
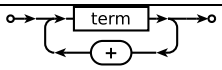
1 Введение

1.1 Способы описания и анализа форматированных текстовых документов

В современных информационных технологиях широко распространено применение текстовых файлов: для хранения и передачи данных (XML, JSON), для отображения данных (HTML, CSS, Markdown, BBCode, Textile), для обработки данных (C/C++, Python,

*Работа выполнена при финансовой поддержке РФФИ, проекты № 16-37-60049 и № 16-07-01267.

Таблица 1 Пример описания грамматики: простые арифметические выражения. Терминальные символы: `0`, `1`, `2`, `3`, `4`, `5`, `6`, `7`, `8`, `9`, `x`, `y`, `z`, `(`, `)`, `*`, `+`

Символ	РБНФ	Синтаксическая диаграмма
`digit`	`''0'' | ''1'' | ''2'' | ''3'' | ''4'' | ''5'' | ''6'' | ''7'' | ''8'' | ''9''`	
`constant`	`digit , {digit}`	
`variable`	`''x'' | ''y'' | ''z''`	
`factor`	`constant | variable | ''('' , expression , '')''`	
`term`	`factor | term, ''*'' , factor`	
`expression`	`term | expression, ''+'' , term`	

JavaScript, C# и многие другие языки программирования). Содержимое таких файлов структурировано с помощью специальной *разметки*.

Файлы каждого типа описываются собственным способом разметки. Чтобы понять, какая именно информация содержится в файлах, требуется знать правила такой разметки. Обычно подобные правила называются *форматом* или (компьютерным) *языком*. Примерами как раз и являются языки программирования, языки разметки, языки спецификаций и т. д. Формат отвечает за то, как размещаются и оформляются, за что отвечают элементы информации в текстовых документах.

Но синтаксис (или *грамматику*) языка тоже нужно как-то описать. Зачастую он допускает довольно большое количество возможных конструкций, которые, тем не менее, состоят из повторяющихся блоков. Благодаря такой специфике возможно описание устройства одних блоков через другие с допустимой рекурсией. Этот подход положен в основу формальных систем определения синтаксиса, таких как: форма Бэкуса–Наура (БНФ) [1], расширенная БНФ (РБНФ) [2], синтаксические диаграммы [3]. БНФ и РБНФ описывают грамматику с помощью текстовых конструкций, синтаксические диаграммы являются визуальной интерпретацией РБНФ.

В табл. 1 приведен пример грамматики языка, с помощью которого можно строить арифметические выражения. Есть набор терминальных (конечных) символов: цифры, знаки переменных, операции, скобки. Остальные конструкции определяются через них и через друг друга: цифра, константа, переменная, множитель, терм, выражение.

Формальные грамматики описаны для большинства популярных языков разметки и программирования. Примеры таких описаний:

- C++ (ISO/IEC 14882:1998(E)): `http://www.externsoft.ch/download/cpp-iso.html`;
- C# 1.0/2.0/3.0/4.0: `http://www.externsoft.ch/download/csharp.html`;
- ECMAScript (JavaScript): `antlr3.org/grammar/1153976512034/ecmascriptA3.g`;
- JSON: `http://rfc7159.net/rfc7159`;
- XML: `https://www.w3.org/TR/REC-xml/#sec-notation`;
- HTML 5: `https://gist.github.com/tkqubo/2842772`.

Наличие формальной грамматики для языка позволяет не только получить стандартизацию формата файлов, но и автоматизировать процесс их анализа.

В данной работе *структурированными текстовыми документами* называются текстовые документы, для которых можно построить абстрактное синтаксическое дерево. Анализ документа подразумевает построение такого дерева. Алгоритм, который строит синтаксическое дерево по структурированному текстовому документу, называется *программой грамматического разбора* или *парсером*.

Оказывается, что наличие формальной грамматики позволяет автоматически построить парсер [4]. Известны два подхода использования формальных грамматик для построения синтаксических деревьев: нисходящий и восходящий синтаксические анализы. При нисходящем синтаксическом анализе правила формальной грамматики применяются начиная со стартового символа до тех пор, пока не будет получена требуемая последовательность. Такой подход реализуют рекурсивный нисходящий парсер и LL-анализатор. При восходящем синтаксическом анализе происходит восстановление выражений до стартового символа. Соответствующие алгоритмы: LR-анализатор и GLR-парсер.

Таким образом, задача построения синтаксического дерева для файлов, описываемых языком с формализованной грамматикой, может считаться решенной.

1.2 Проблемы при работе с документами в формате LaTeX

В общем виде проблему можно сформулировать следующим образом: для документов в формате LaTeX нет формальной грамматики. Это означает и отсутствие строгой стандартизации, и отсутствие возможности автоматически построить парсер известными методами. Источником такой проблемы являются четыре факта:

1) сигнатура команд и символы LaTeX не определены в общем виде. И количество параметров, и способы отделения параметров для разных команд могут существенно отличаться. В большинстве команд параметры выделяются фигурными скобками. Но в некоторых случаях есть необязательные параметры, которые могут быть в квадратных скобках. Кроме того, можно определить команду, в которой параметры будут разделяться, например, точкой или любым другим знаком;
2) и сигнатура, и набор команд и символов определяются стилевыми файлами. Стилевые файлы LaTeX могут содержать правила форматирования и оформления, переопределенные символы, команды и окружения. Например, есть общепринятая команда `\author{Имя автора}` для указания автора документа. Но некоторые стилевые файлы переопределяют ее так, что появляется необязательный параметр: `\author{Имя автора}[Имя для колонтитулов]`;
3) и сигнатура, и набор доступных команд и символов определяются контекстом. Например, есть общеупотребительный символ тире: `---`. Он работает независимо от используемого языка, но отображается с отбивками, не принятыми в русской типографике. Для получения правильных отбивок в публикации нужно использовать символ `''---`,

но он недоступен, если не был выбран русский язык. Также сильно разнятся наборы команд и символов внутри формул и вне;

4) TEX не подразумевает наличие синтаксического дерева. LATEX является наиболее популярным пакетом макрорасширений для TEX — системы компьютерной верстки, разработанной Дональдом Кнутом [5, 6]. TEX предоставляет средства для структуризации и оформления текстов, но только в LATEX появляются команды и окружения, совокупность которых позволяет формировать абстрактное синтаксическое дерево. В LATEX-документах допустимы фрагменты на «чистом» TEX, но такие прецеденты являются скорее исключением и при качественной верстке должны быть перемещены в стилевой файл. В рамках данного исследования подобные фрагменты исходного кода могут быть интерпретированы как отдельные терминальные вершины синтаксического дерева.

Таким образом, задача построения синтаксического дерева для документов в формате LATEX разрешима, но требует отдельных исследования и алгоритмов.

1.3 Актуальность анализа документов в формате LATEX

Многие научные издательства и конференции используют LATEX для подготовки публикаций. Как следствие, есть ряд практических задач, связанных с обработкой документов в таком формате, и на этапе подготовки документов, и для интеллектуальной обработки существующих коллекций: автоматизация коррекции, статистический анализ, извлечение информации, преобразование форматов.

Самой первой задачей из тех, с которыми работал автор, требующей синтаксического разбора документов LATEX, оказалась задача автоматизации коррекции *типографических ошибок*. Такие ошибки связаны с оформлением отступов, формул, штрифтами и т. п. При текущем уровне технологий их исправление производится корректорами вручную, что порождает проблемы, связанные с качеством и временем обработки. Процесс коррекции научных статей изображен на рис. 1. Оказалось, что эту задачу можно решить методами машинного обучения, при этом обучающая выборка составляется из пар синтаксических деревьев документов до обработки профессиональным корректором и после [7].

Опубликованные статьи, как правило, попадают в одну из систем цитирования. Для русскоязычных статей создан РИНЦ (Российский индекс научного цитирования), который управляется eLIBRARY. Это национальная библиографическая база данных научного цитирования, аккумулирующая более 9 млн публикаций российских авторов, а также информацию о цитировании этих публикаций из более чем 6000 российских журналов [8]. Процесс загрузки статей в базу данных РИНЦ изображен на рис. 2. На данный момент одним из этапов является ручная обработка в системе Articulus, которая принимает только документы в формате HTML. В данном случае качественно преобразовать код LATEX в HTML можно, только анализируя синтаксическое дерево для выделения логических бло-

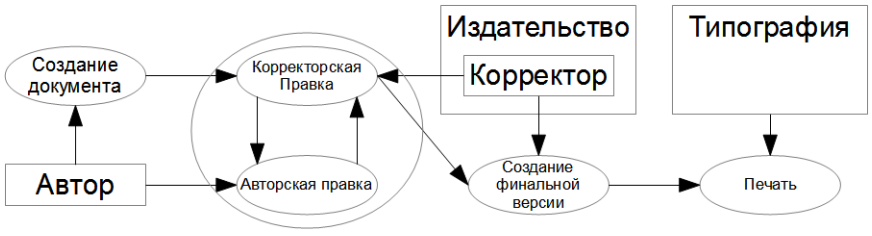


Рис. 1 Бизнес-процесс коррекции научных статей

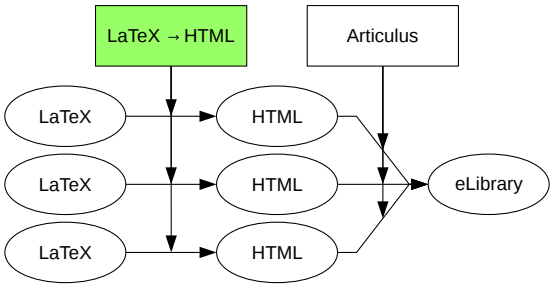


Рис. 2 Бизнес-процесс добавления статей в РИНЦ

ков. Более того, подобный анализ откроет возможность избавления от ручной обработки статей.

На самом деле, задача преобразования формата возникает не только в контексте РИНЦ. Формат HTML также удобно использовать для отображения статей на WEB-ресурсах. Бывает, что издательства требуют документ в формате DOC или DOCX, а материалы, в силу своей научной природы, готовы только в LaTeX. Для хранения и обработки структурированной информации удобно использовать XML. В каждом из случаев требуется корректно выделять логические блоки документов, чего нельзя добиться без качественного анализа синтаксической структуры, т. е. построения синтаксического дерева.

Выделение логических блоков и удобное их представление также полезны в задачах интеллектуального анализа текста, таких как сбор статистики по авторам и издательствам, каталогизация, построение выборки для тематического моделирования.

Таким образом, задача построения синтаксического дерева для документов в формате LaTeX актуальна для практического использования.

1.4 Известные реализации и их недостатки

Исследования, связанные с синтаксическим анализом документов LaTeX, уже проводились, и существует ряд реализаций, среди которых: на Python — plasTeX [9], на Perl — LaTeX::Parser [10], на Java — SnuggleTeX [11]. В силу особенностей анализа формата в каждом из них используются внешние ресурсы — идея, активно применяющаяся в данной работе. Но при этом все эти проекты игнорируют логический смысл элементов синтаксического дерева. Например, нельзя определить, является ли математический символ оператором или просто буквой, а команда в тексте — тэгом разметки или изменением состояния. Такие свойства важны не только для корректного анализа синтаксической структуры документов LaTeX, но и для последующего интеллектуального анализа синтаксического дерева. Самое простые примеры: не зная, какие символы являются буквами или цифрами, нельзя сгруппировать их, соответственно, в слова или числа.

Таким образом, имелась потребность в новом исследовании, на удовлетворение которой и нацелена данная работа.

2 Структура документа LaTeX

В этом разделе описывается то, как воспринимаются документы формата LaTeX в данной работе. Это хорошо зарекомендовавшая себя эвристика, позволяющая формализовать понятие синтаксического дерева для таких документов.

Отдельно нужно сказать: в исходном коде LaTeX допустимы комментарии — части строк, начинающиеся с неэкранированного символа %. Для упрощения описания алгоритмов будем считать, что комментарии удаляются в ходе предобработки.

Весь исходный код файла состоит из блоков, каждый из которых может быть *символом*, *командой* или *окружением*.

Символ LaTeX— это произвольный набор последовательных знаков. Символ может быть терминальным или содержать параметры. Типичными представителями терминальных символов являются цифры и буквы. Знак тире `''---` также является терминальным. В качестве примера символа с параметром выступает обозначение формулы в нотации `$$#1$$`. Где `#1` означает параметр, в данном случае это — тело формулы.

Отдельно нужно отметить символ пробела. Ему эквивалентно любое количество подряд идущих знаков пробела или табуляции и, быть может, одного знака переноса строки. Если в наборе подряд идущих пробельных знаков больше одного знака переноса строки, такой набор эквивалентен двум переносам строки и воспринимается как символ разделения абзацев. Это общеизвестная специфика верстки с издательской системой LaTeX.

Команда LaTeX — это последовательность знаков вида `\имя_команды шаблон`. Обязательное `имя_команды` должно представлять собой последовательность латинских букв, которая может заканчиваться знаком звездочки. Необязательный `шаблон` имеет тот же формат, что и символы LaTeX, и так же может содержать параметры. Пример команды, которая используется для выделения текста полужирным: `\textbf#1`.

Окружение LaTeX — это последовательность знаков вида:
`\begin{имя_окружения}шаблон_команды_начала`
`тело_окружения`
`\end{имя_окружения}шаблон_команды_конца`.
`имя_окружения` имеет тот же формат, что и `имя_команды`. Для документов формата LaTeX общепринято, что весь выводимый контент помещается внутрь окружения `document`.

Независимо от своей природы (символ, команда или окружение) каждый блок документа обладает назначением (логическим смыслом), которое в данной работе называется *типом лексемы*. Выделяемые типы лексем представлены в табл. 2.

Учет логического смысла элементов важен не только при последующем интеллектуальном анализе файлов, но и используется во время синтаксического анализа. Он позволяет определять контекст параметров для символов и команд и ограничивать перебор возможных вариантов. В рамках подобных исследований такой подход применен впервые.

Контекст также определяется состоянием анализатора. Это аналогично поведению компилятора TeX в зависимости от того, какой режим активирован. Список поддерживаемых режимов доступен в табл. 3.

Режимы могут изменяться в любом месте документа по отдельности или группами. Кроме того, может быть начата группа локального определения режимов. После завершения группы состояния режимов восстанавливаются на значения, которые были до начала группы.

3 Описание стилевых элементов LaTeX

Как было указано ранее, наборы доступных символов, команд и окружений LaTeX определяются стилевыми файлами. Это специфика, которая приводит к тому, что описания используемых команд нужно каким-то образом получать. Задача извлечения этой информации из кода стилевых файлов крайне нетривиальна и сравнима, если не превосходит, сложность реализации компилятора TeX. Поэтому в данной работе предлагается использовать внешним образом сформированные описания символов, команд и окружений, которые потом будут переданы как параметр алгоритму синтаксического анализа.

Таблица 2 Типы лексем LaTeX

Тип лексемы	Комментарий
BINARY_OPERATOR	бинарный математический оператор
BRACKETS	логические скобки
CELL_SEPARATOR	разделитель ячеек таблицы
CHAR	знак
DIGIT	цифра
DIRECTIVE	директива LaTeX
DISPLAY_EQUATION	выключенная формула
FILE_PATH	путь файловой системы
FLOATING_BOX	плавающий блок
HORIZONTAL_SKIP	горизонтальный интервал
INLINE_EQUATION	включенная формула
LABEL	идентификатор метки
LENGTH	линейное измерение
LETTER	буква
LINE_BREAK	разрыв строки
LIST_ITEM	элемент списка
LIST	список
NUMBER	последовательность цифр
PARAGRAPH_SEPARATOR	разделитель абзацев
PICTURE	картинка
POST_OPERATOR	математический постоператор
PRE_OPERATOR	математический преоператор
RAW	необрабатываемая часть исходника
SPACE	пробел
SUBSCRIPT	нижний индекс
SUPERSCRIPT	верхний индекс
TABLE	таблица
TABULAR_PARAMETERS	параметры таблицы
TAG	тэг форматирования
UNKNOWN	неизвестный элемент
VERTICAL_SKIP	вертикальный интервал
WORD	последовательность букв
WRAPPER	обертка

Таблица 3 Режимы LaTeX

Режим	Комментарий
LIST	внутри списка элементов
MATH	внутри математического выражения
PICTURE	внутри описания изображения
TABLE	внутри таблицы
TEXT	обычный текст (по умолчанию)
VERTICAL	между абзацами

Далее будут представлены основные структуры, с с помощью которых формируется информация о стилевых файлах.

Operation — операция над состоянием LaTeX:

- `directive` — директива: `BEGIN` (начать) или `END` (закончить);
- `operand` — режим LaTeX или `GROUP` (группа локального определения режимов).

Операция описывает процесс изменения режимов. `BEGIN` означает активацию режима, `END` — деактивацию.

Parameter — параметр символа или команды:

- `lexeme` — тип лексемы (логический смысл), необязательно указывается;
- `modes` — режимы, в которых определен;
- `operations` — набор операций, выполняемых перед параметром.

Symbol — символ LaTeX:

- `lexeme` — тип лексемы (логический смысл);
- `modes` — режимы, в которых определен;
- `operations` — набор операций, выполняемых после символа;
- `parameters` — описание параметров;
- `pattern` — шаблон LaTeX.

Шаблон описывает сигнатуру символа, в которой `#номер_параметра` обозначает положение параметров, а остальные знаки соответствуют знакам символа в исходном коде документа. Пример на языке JSON описания символа включенной формулы:

```
{
    "lexeme": "WRAPPER",
    "modes": ["TEXT"],
    "operations": [{
        "directive": "END",
        "operand": "MATH"
    }],
    "parameters": [{
        "operations": [{
            "directive": "BEGIN",
            "operand": "MATH"
         }]
    }],
    "pattern": "$#1$"
}.
```

Command — команда LaTeX:

- `lexeme` — тип лексемы (логический смысл);
- `modes` — режимы, в которых определена;
- `operations` — набор операций, выполняемых после команды;
- `parameters` — описание параметров;
- `pattern` — шаблон LaTeX;
- `name` — имя команды.

От описания символа отличается только тем, что добавлено имя. Пример на языке JSON описания двух команд вставки информации об авторе документа с одним и тем же именем, но разным набором параметров:

```
{
    "lexeme": "TAG",
    "modes": ["TEXT"],
    "parameters": [{ }, { }],
    "pattern": "[#1]#2",
    "name": "author"
},
{
    "lexeme": "TAG",
    "modes": ["TEXT"],
    "parameters": [{ }],
    "pattern": "#1",
    "name": "author"
}.
```

Environment — окружение LaTeX:

- `lexeme` — тип лексемы (логический смысл);
- `modes` — режимы, в которых определено;
- `name` — имя окружения.

Пример на языке JSON описания окружения для горизонтального выравнивания по центру:

```
{
    "lexeme": "WRAPPER",
    "modes": ["TEXT"],
    "name": "center"
}.
```

Предложенный способ описания стилевых элементов является простым для понимания: он не требует глубоких навыков программирования или знания теории формальных языков. Описания команд могут быть составлены и пользователями LaTeX, обладающими опытом верстки с помощью этой системы. Достаточно знания допустимых синтаксисов символов и команд и понимания логического смысла элементов. В то же время, этот способ достаточно гибок, поскольку позволяет описывать не только синтаксические конструкции, но и задавать метаинформацию (режимы LaTeX, типы лексем) для управления контекстом, а сам набор описаний не является фиксированным и может быть модифицирован при замене или добавлении стилевого файла.

4 Синтаксическое дерево документа LaTeX

Взаимное расположение символов, команд и окружений LaTeX образует древовидную структуру, корнем которой является окружение `document`. Узлы этой структуры называются *токенами*. При синтаксическом анализе документа в формате LaTeX элементы исходного кода, в зависимости от лексем и контекста, могут порождать токены определенных типов, список которых представлен в табл. 4.

Полученное дерево удобно использовать для преобразования документа LaTeX в другой формат, интерпретируя отдельные токены. И, поскольку каждому токену соответствует тип лексемы, такая структура оказывается дополнительно информативной для интеллектуального анализа.

Таблица 4 Типы токенов синтаксического дерева LaTeX

Тип токена	Пример исходного кода
LaTeX environment body	\begin{tabular}{c|c} height & 1.2m \end{tabular}
LaTeX command	\includegraphics [width=10cm] {../figure.eps}
LaTeX environment	\begin{tabular}{c|c} height & 1.2m \end{tabular}
Label	\ref{equation1}
Linear dimension	\textwidth=10cm
Number	height 1.2\,m
Paragraph separator	Paragraph □ New paragraph
Filesystem path	\includegraphics [width=10cm] {../figure.eps}
Space	height□1.2\,m
Symbol	height 1.2 \, m
Tabular parameters	\begin{tabular}{c|c} height & 1.2m \end{tabular}
Word	height 1.2 \,m
Raw char sequence	\verb|complex source|

5 Алгоритмы синтаксического анализа элементов документа LaTeX

В этом разделе описываются разработанные в ходе описываемого исследования алгоритмы синтаксического анализа отдельных фрагментов исходного кода документа в формате LaTeX: шаблона, символа, команды, окружения. Допускаются рекурсивные вызовы одним алгоритмом других. Это необходимо, поскольку, вообще говоря, нет ограничений на типы и глубину вложенных токенов, а предлагаемый метод разбора исходного кода подобен рекурсивному нисходящему парсеру.

Алгоритм 1 описывает процесс синтаксического анализа шаблона команды или символа LaTeX. У него два основных предназначения: выбрать, какой синтаксис применим, и рекурсивно собрать набор токенов параметров. Если алгоритм возвращает TRUE, предложенный синтаксис применим, и в $parameterTokens$ будет список полученных токенов для параметров в соответствующем порядке. Если алгоритм возвращает FALSE, в данном месте не может быть использована команда или символ с предлагаемым синтаксисом.

Тонкости, которые не вошли в описание алгоритма: если параметр имеет особенный тип лексемы, соответствующий, например, параметрам таблицы или пути в файловой системе, используется специальный алгоритм синтаксического анализа, которые возвращает токен соответствующего типа.

Алгоритм 1 Синтаксический анализ шаблона команды или символа

Вход:
W — строка для анализа, pos — текущая позиция,
W_p — шаблон LaTeX, $pos_p = 0$ — текущая позиция в шаблоне,
$style$ — описание символа или команды, чей шаблон анализируется,
$parameterTokens$ — стек токенов параметров (изначально пустой)
Выход: TRUE, если код соответствует шаблону; FALSE, в противном случае

```
1: пока pos_p не в конце W_p
2:   если W_p[pos_p] — пробел то
3:     если не удается считать пробел из W в позиции pos то
4:       return FALSE
5:     передвинуть pos к концу пробела
6:     pos_p = pos_p + 1
7:   иначе если W_p[pos_p] == # то
8:     pos_p = pos_p + 1
9:     номер параметра = W_p[pos_p]
10:    получить свойства параметра из style
11:    если не удаётся считать параметр из W в позиции pos то
12:      очистить parameterTokens
13:      return FALSE
14:    добавить считанный токен параметра в parameterTokens
15:    передвинуть pos к концу параметра
16:    pos_p = pos_p + 1
17:  иначе
18:    если W[pos]! = W_p[pos_p] то
19:      return FALSE
20:    pos = pos + 1
21:    pos_p = pos_p + 1
22: return TRUE
```

Алгоритм 2 описывает процесс синтаксического анализа символа LaTeX. По исходному коду в текущей позиции и активным режимам LaTeX выбираются символы с допустимыми шаблонами. Все они перебираются до тех пор, пока не будет найден подходящий. Если один из допустимых шаблонов оказался применим, то генерируется токен соответствующего символа, а стек полученных при анализе шаблона параметров формирует набор дочерних токенов. Если ни один из допустимых шаблонов не подошел, возвращается токен символа с неизвестным описанием. Таким образом, алгоритм всегда возвращает положительный результат.

Тонкости, которые не вошли в описание алгоритма: если полученный токен имеет тип лексемы пробела, разделителя абзацев, буквы или цифры, он преобразуется в токен соответствующего типа.

Алгоритм 2 Синтаксический анализ символа

Вход: W — строка для анализа, pos — текущая позиция
Выход: токен символа

```
1: сохранить текущее состояние
2: получить описания символов для текущего состояния, начинающиеся с W[pos]
3: для всех полученных описаний символов
4:    если W, начиная с позиции pos, соответствует шаблону то
5:       t = токен символа с текущим описанием
6:       дочерние токены t = стек токенов параметров, полученных при анализе шаблона
7:       return t
8:    восстановить сохраненное состояние
9: return токен символа с неизвестным описанием
```

Алгоритм 3 описывает процесс синтаксического анализа команды LaTeX. Он практически повторяет логику алгоритма 2 за тем исключением, что допустимые команды определяются по имени, и в случае, если исходный код в текущей позиции не содержит команду, алгоритм не возвращает токен.

Алгоритм 3 Синтаксический анализ команды

Вход: W — строка для анализа, pos — текущая позиция
Выход: токен команды или ничего, если в текущей позиции нет команды

```
1: если W начинается не с \имя_команды то
      выход
2: сохранить текущее состояние
3: получить имя команды
4: получить описания команд с полученным именем для текущего состояния
5: для всех полученных описаний команд
6:    если W, начиная с позиции pos, соответствует шаблону то
7:       t = токен команды с текущим описанием
8:       дочерние токены t = стек токенов параметров, полученных при анализе шаблона
9:       return t
10:   восстановить сохраненное состояние
11: return токен команды с неизвестным описанием
```

Алгоритм 4 описывает процесс синтаксического анализа окружения LaTeX. Если исходный код в текущей позиции не содержит начало окружения, ничего не возвращается. В противном случае возвращается токен, соответствующий окружению, описание которого извлекается по имени.

Тонкости, которые не вошли в описание алгоритма: при некорректном документе формата LaTeX конец окружения может отсутствовать (в этом случае будет автоматически сформирована команда конца при завершении исходного кода), а если описания окружения с данным именем нет, то будет сформирован токен окружения с неизвестным описанием.

Алгоритм 4 Синтаксический анализ окружения

Вход: W — строка для анализа, pos — текущая позиция
Выход: токен окружения или ничего, если в текущей позиции нет начала окружения

если W начинается не с `\begin{имя_окружения}` **то**
  выход
получить имя окружения
t = токен окружения, соответствующего имени
сдвинуть pos концу `\begin{имя_окружения}`
считать шаблон команды с именем «`имя_окружения`»
записать соответствующий токен, как токен команды начала t
пока с позиции pos в W не стоит `\end{имя_окружения}`
  считать дочерний токен t
сдвинуть pos концу `\end{имя_окружения}`
считать шаблон команды с именем «`endимя_окружения`»
записать соответствующий токен, как токен команды конца t
return t

Алгоритм 5 описывает процесс получения очередного токена. Априори не известно, какого типа токен находится в текущей позиции исходного кода. Поэтому перебираются возможные варианты: пробел, окружение, команда, символ. Токен символа всегда может быть получен, поэтому каждая итерация «поглощает» некоторый ненулевой фрагмент исходного кода. Таким образом, алгоритм выполним в любом случае.

Алгоритм 5 Синтаксический анализ кода LaTeX

Вход: W — строка для анализа, $pos = 0$ — текущая позиция
Выход: $tokens$ — последовательность считанных токенов

пока pos не в конце W
  сохранить текущее состояние
  если удалось считать считать пробел в W с позиции pos **то**
    записать токен пробела в $tokens$
    сдвинуть pos к концу пробела
    перейти к новой итерации
  восстановить сохраненное состояние
  если удалось считать считать окружение в W с позиции pos **то**
    записать токен окружения в $tokens$
    сдвинуть pos к концу окружения
    перейти к новой итерации
  восстановить сохраненное состояние
  если удалось считать считать команду в W с позиции pos **то**
    записать токен команды в $tokens$
    сдвинуть pos к концу команды
    перейти к новой итерации

восстановить сохраненное состояние
считать символ в W с позиции pos
записать токен символа в $tokens$

Описанные алгоритмы в совокупности с комментариями о тонкостях охватывают все возможные ситуации в используемой интерпретации синтаксиса LaTeX.

6 Реализация

Описанные в данной работе идеи были реализованы автором несколько раз. Впервые они успешно использовались для задач автоматической коррекции документов [12] и построения XML-описаний статей для применения в тематическом моделировании. Реализация была сделана с помощью Qt/C++, исходные коды публично не распространялись.

С 2016 г. автором запущен проект по реализации синтаксического анализатора LaTeX на JavaScript [13]. Это набор библиотек, распространяемых по лицензии LGPLv3. Основной их задачей является прозрачное внедрение инструментов анализа документов в формате LaTeX для всех окружений, поддерживающих JavaScript, в том числе для WEB.

Доступны следующие библиотеки:

- `Latex.js`. Содержит основные определения структур LaTeX, такие как лексемы, режимы и операции;
- `LatexStyle.js`. Содержит определения структур стиля LaTeX: символы, команды, окружения. Также предоставляет инструменты для работы с коллекциями таких структур в формате JSON;
- `LatexTree.js`. Содержит определения структур синтаксического дерева LaTeX: токены всех типов, само дерево;
- `LatexParser.js`. Предоставляет класс синтаксического анализатора, принимающего описания стилей и позволяющего по исходному тексту документа в формате LaTeX получать синтаксическое дерево.

На данный момент принципы работы всех библиотек соответствуют алгоритмам, приведенным в этой работе. С их использованием разрабатывается проект по отображению LaTeX-документов в браузере средствами HTML.

7 Заключение

В данной работе рассматривалась задача синтаксического анализа текстовых документов, формат которых не является полностью определенным стандартами или другими общими для всех документов правилами. В качестве образца таких файлов были выбраны документы в формате LaTeX. Было показано, что для языка их разметки нет формализованной грамматики, поэтому требуется отдельный подход, учитывающий внешние ресурсы, соответствующие подгружаемым стилевым файлам. Существующие решения не используют логические значения элементов синтаксиса, поэтому не могут быть применены для интеллектуального анализа текстов.

Предложен подход, который позволяет с помощью несложных конструкций описать стилевые файлы LaTeX. Описаны разработанные алгоритмы синтаксического анализатора, принимающего такие описания. Они реализованы в виде набора библиотек на языке JavaScript, распространяемого по лицензии LGPLv3.

Таким образом, можно считать, что задача синтаксического анализа документов LaTeX решена в рамках рассматриваемого в этой работе представления о синтаксическом дереве таких документов.

Но остается важная проблема, которая препятствует быстрому внедрению разработанного синтаксического анализатора — необходимость ручного описания элементов стиля: символов, команд и окружений LaTeX. Это несложный, но трудоемкий процесс, требующий навыков верстки от исполнителей. Продолжение исследований по рассматриваемой тематике может быть направлено на автоматизацию этого процесса.

Литература

[1] Programming systems and languages / Ed. S. Rosen. — McGraw Hill computer science ser. — New York, NY, USA: McGraw Hill, 1967. 734 p.

[2] ISO/IEC 14977:1996. Information technology — syntactic metalanguage — extended BNF. http://www.iso.org/iso/catalogue_detail?csnumber=26153.

[3] Syntax diagram. Wikipedia. https://en.wikipedia.org/wiki/Syntax_diagram.

[4] *Ахо А. В., Лам М. С., Сети Р., Ульман Д. Д.* Компиляторы: принципы, технологии и инструментарий / Пер. с англ. — 2-е изд. — М.: Вильямс, 2008. 1184 с. (*Aho A., Lam M., Sethi R., Ullman J.* Compilers: Principles, techniques, and tools. — 2nd ed. — Prentice Hall, 2006. 1000 p.)

[5] *Lamport, L.* LaTeX: A document preparation system. — Reading, MA, USA: Addison-Wesley Professional, 1994. 273 p.

[6] *Кнут Д.* Всё про TeX / Пер. с англ. — М.: Вильямс, 2003. 560 с. (*Knuth D. E.* The TeX book. — Reading, MA: Addison-Wesley Professional, 1984. 496 p.)

[7] *Chuvilin K.* Machine learning approach to automated correction of LaTeX documents // 18th FRUCT & ISPIT Conference Proceedings. — Saint-Petersburg: Technopark of ITMO University. P. 33–40. http://fruct.org/publications/fruct18/files/Chu.pdf.

[8] eLIBRARY.RU — Российский индекс научного цитирования. http://elibrary.ru/project_risc.asp.

[9] plasTeX — a Python framework for processing LaTeX documents. http://plastex.sourceforge.net/plastex/index.html.

[10] *Heinicke S.* LaTeX-Parser-0.01. http://search.cpan.org/~svenh/LaTeX-Parser-0.01/.

[11] SnuggleTeX — overview & features. http://www2.ph.ed.ac.uk/snuggletex/documentation/overview-and-features.html.

[12] *Чувилин К. В.* Автоматический синтез правил коррекции текстовых документов формата LaTeX. — М.: Вычислительный центр им. А. А. Дородницына Российской академии наук, 2013. Дисс. ... канд. техн. наук.

[13] texnous latex-parser — Bitbucket. https://bitbucket.org/texnous/latex-parser/.

Поступила в редакцию 1.09.2016

Parametric approach to the construction of syntax trees for partially formalized text documents*

K. V. Chuvilin[1,2]
kirill@chuvilin.pro
[1]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow, Russia
[2]ICPT, 6 Zavodskoy proezd, Protvino, Moscow Region, Russia

This article investigates the possibility of logical structure (abstract syntax tree) automatic construction for text documents, the format of which is not fully defined by standards or other rules common to all the documents. In contrast to the syntax described by formal grammars, in such cases, there is no way to build the parser automatically. Text files in LaTeX format are the typical examples of such formatted documents with not completely formalized syntax markup. They are used as the resources for the implementation of the algorithms developed in this work. The relevance of LaTeX document analysis is due to the fact that many scientific publishers and conferences use LaTeX typesetting system, and this gives rise to important applied task of automation for categorization, correction, comparison, statistics collection, rendering for WEB, etc. The parsing of documents in LaTeX format requires additional information about styles: symbols, commands, and environments. A method to describe them in JSON format is proposed in this work. It allows to specify not only the information necessary to pars, but also meta information that facilitates further data mining. This approach is used for the first time. The developed algorithms for constructing a syntax tree of a document in LaTeX format that use such information as an external parameter are described. The results are successfully applied in the tasks of comparison, autocorrection, and categorization of scientific papers. The implementation of the developed algorithms is available as a set of libraries released under the LGPLv3. The key features of the proposed approach are flexibility (within the framework of the problem) and simplicity of parameter descriptions. The proposed approach allows one to solve the problem of parsing documents in LaTeX format. But it is required to form the base of style element descriptions for widespread practical use of the developed algorithms.

Keywords: *abstract syntax tree; JSON; LaTeX; parsing; text mining; tree*

DOI: 10.21469/22233792.2.2.06

References

[1] Rosen, S., ed. 1967. *Programming systems and languages*. McGraw Hill computer science ser. New York, NY: McGraw Hill. 734 p.

[2] ISO/IEC 14977:1996. Information technology — syntactic metalanguage — extended BNF. Available at: http://www.iso.org/iso/catalogue_detail?csnumber=26153 (accessed August 31, 2016).

[3] Syntax diagram. Wikipedia. Available at: https://en.wikipedia.org/wiki/Syntax_diagram (accessed August 31, 2016).

[4] Aho, A., M. Lam, R. Sethi, and J. Ullman. 2006. *Compilers: Principles, techniques, and tools*. 2nd ed. Prentice Hall. 1000 p.

[5] Lamport, L. 1994. *LaTeX: A document preparation system: User's guide and reference*. 2nd ed. Reading, MA: Addison-Wesley Professional. 273 p.

[6] Knuth, D. E. 1984. *The TeXbook*. Reading, MA: Addison-Wesley Professional. 496 p.

*The research was supported by the Russian Foundation for Basic Research (grants 16-37-60049 and 16-07-01267).

[7] Chuvilin, K. V. Machine learning approach to automated aorrection of LaTeX documents. *18th FRUCT & ISPIT Conference Proceedings*. Saint-Petersburg: Technopark of ITMO University. 33–40. Available at: `http://fruct.org/publications/fruct18/files/Chu.pdf` (accessed August 31, 2016).

[8] eLIBRARY.RU — Rossiyskiy indeks nauchnogo tsitirovaniya [eLIBRARY.RU — Russian Science Citation Index]. Available at: `http://elibrary.ru/project_risc.asp` (accessed August 31, 2016).

[9] plasTeX — a Python framework for processing LaTeX documents. Available at: `http://plastex.sourceforge.net/plastex/index.html` (accessed August 31, 2016).

[10] Heinicke, S. LaTeX-Parser-0.01. Available at: `http://search.cpan.org/~svenh/LaTeX-Parser-0.01/` (accessed August 31, 2016).

[11] SnuggleTeX — overview & features. Available at: `http://www2.ph.ed.ac.uk/snuggletex/documentation/overview-and-features.html` (accessed August 31, 2016).

[12] Chuvilin, K. V. 2013. Automatic synthesis of correction rules for text documents in the LaTeX format. Moscow: Dorodnicyn Computing Center of the Russian Academy of Sciences. PhD Diss. (In Russian.)

[13] texnous latex-parser — Bitbucket. Available at: `https://bitbucket.org/texnous/latex-parser/` (accessed August 31, 2016).

Received September 1, 2016

Оптимальный выбор параметров для восстановления спектров морского волнения по аэрокосмическим изображениям*

***В. Г. Бондур*[1], *А. Б. Мурынин*[1,2], *В. Ю. Игнатьев*[1,2]**
vgbondur@aerocosmos.info; amurynin@bk.ru; vladimir.ignatiev.mipt@gmail.com
[1]НИИ аэрокосмического мониторинга «АЭРОКОСМОС», Россия, г. Москва, Гороховский пер., 4
[2]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Рассматривается проблема восстановления спектров морской поверхности по аэрокосмическим изображениям в широком спектральном диапазоне длин волн. В рамках описанной нелинейной модели поля яркости, регистрируемого аппаратурой дистанционного зондирования, предложена модификация восстанавливающего оператора, действующего во всей пространственно-спектральной области. Описан итерационный процесс выбора оптимальных значений параметров модифицированного оператора с использованием подспутниковых измерений для валидации. Представлены результаты проверки работоспособности построенного оператора для различных условий регистрации изображений морской поверхности.

Ключевые слова: *пространственный спектр; восстанавливающий оператор; оптимизация по параметрам*

DOI: 10.21469/22233792.2.2.07

1 Введение

Различные процессы, происходящие на взволнованной поверхности океана, наиболее полно описываются с использованием пространственно-частотных, пространственных и частотных спектров волнения [1–3]. Эти спектры позволяют получать важную информацию о явлениях, происходящих на поверхности и в приповерхностном слое морей и океанов, об энергетических особенностях морских волн, о характеристиках приводного слоя атмосферы и ветровом режиме [2]. Подход, основанный на пространственно-частотном и частотном спектральном анализе, позволяет выявлять зоны негативных естественных и антропогенных воздействий на водную среду [1, 4].

Для получения информации о состоянии границы раздела океан–атмосфера и о спектрах поверхностного волнения на больших площадях с различным пространственным разрешением в любых, в том числе труднодоступных регионах океана, перспективно использование аэрокосмических методов дистанционного зондирования и методов обработки аэрокосмических изображений [1, 4–10]. Для регистрации двумерных и одномерных пространственных спектров волнения целесообразно применение оптических аэрокосмических изображений высокого пространственного разрешения, позволяющих регистрировать мгновенные распределения полей яркости, которые несут информацию о пространственной структуре морских волн [1, 4–6]. Для адекватной оценки двумерных и одномерных спектров поверхностного волнения по оптическим изображениям должны использоваться специальные методы восстановления характеристик границы раздела атмосфера–гидросфера по данным дистанционного зондирования [1, 4–6, 11–14]. При этом необходимо применять восстанавливающие операторы, которые строятся на основе учета различных

*Работа выполнена при поддержке РФФИ, гранты № 14-05-91759 , № 16-51-55019.

условий формирования изображений и характеристик аппаратуры [1, 5, 6, 10–14]. В настоящей работе описывается метод построения операторов для восстановления пространственных спектров морского волнения по спектрам оптических аэрокосмических изображений и приводятся результаты его применения.

2 Постановка задачи

Взволнованная морская поверхность представляет собой случайное поле возвышений (волновых аппликат) [1]:

$$z = \zeta(x, y, t), \tag{1}$$

где $\zeta(x, y, t)$ — случайная функция возвышений морской поверхности (поле возвышений); (x, y, z) — прямоугольная декартова система координат, в которой плоскость (x, y) совпадает с уровнем спокойной (невзволнованной) водной поверхности; t — время.

Фиксируя в (1) момент времени $t = t_0$, получим двумерную случайную функцию пространственных координат:

$$z = \zeta(x, y, t)|_{t=t_0} = \xi(x, y)\,.$$

Для исследований характеристик поля возвышений морской поверхности в фиксированный момент времени $z = \xi(x, y)$ используются аэрокосмические изображения, которые регистрируются дистанционными методами. Двумерные поля сигналов, которые представлены на аэрокосмических изображениях, связаны с полем возвышения морской поверхности и могут использоваться для оценки значимых характеристик этой поверхности.

Так как поле возвышений морской поверхности $\xi(x, y)$ является гауссовским квазистационарным полем, то она достаточно полно описывается спектральной плотностью [2]:

$$\Psi(\mathbf{k}) = \hat{S}[\xi](\mathbf{k}),$$

где $\hat{S}$ — оператор спектральной плотности, пропорциональный квадрату модуля Фурье-преобразования поля $\xi(x, y)$; $\mathbf{k} = (k_x, k_y)$ — волновой вектор.

Поскольку оптические изображения морской поверхности формируются в результате отражения и преломления света по законам геометрической оптики, то при их анализе структуру морской поверхности наряду с полем возвышений $\xi(x, y)$ удобно характеризовать полями уклонов (или градиентов) вдоль осей [2]

$$\xi_\alpha(x, y) = \frac{\partial \xi(x, y)}{\partial \alpha}\,, \quad \alpha = x, y\,. \tag{2}$$

Поле уклонов морской поверхности в произвольном направлении φ (отличном от направлений осей координат) с учетом (2) можно выразить следующим образом:

$$\beta_\varphi(x, y) = \cos\varphi\xi_x(x, y) + \sin\varphi\xi_y(x, y).$$

Учитывая свойства преобразования Фурье, можно связать спектр такого поля уклонов $\Phi(\mathbf{k}) = \hat{S}[\beta_\varphi](\mathbf{k})$ со спектром поля возвышений $\Psi(\mathbf{k})$:

$$\Phi(\mathbf{k}) = (\cos\varphi k_x + \sin\varphi k_y)\Psi(\mathbf{k}). \tag{3}$$

Пространственный спектр $\Psi(\mathbf{k}, \varphi)$ в полярных координатах (k, φ) описывает распределение волновой энергии по волновым числам $k = |\mathbf{k}|$ и направлениям, задаваемым волновым азимутом $\varphi = \arctan(k_x, k_y)$.

Рассмотрим физические модели формирования поля яркости, фиксируемого на аэрокосмических изображениях. Поле яркости взволнованной морской поверхности формируется в результате отражения от нее излучения, приходящего из верхней полусферы и преломления на ней восходящего светового потока, образующегося при рассеянии в водной толще.

Поле яркости $L(x, y)$ в фиксированный момент времени состоит из нескольких составляющих, формируемых различными физическими процессами [1]:

$$L(x, y) = L^{(1)}(x, y) + \lfloor L^{(2)}(x, y) + L^{(3)}(x, y) \rfloor \tau_a,$$

где $L^{(1)}(x, y)$ — яркость, обусловленная рассеянием в атмосфере в направлении приемника; $L^{(2)}(x, y)$ и $L^{(3)}(x, y)$ — яркости, обусловленные отражением от поверхности и излучением, выходящим из-под воды (рассеяние молекулами воды и взвешенными веществами); τ_a — коэффициент пропускания атмосферы. Компоненты $L^{(2)}(x, y)$ и $L^{(3)}(x, y)$ отражаются и преломляются от элементов морской поверхности в соответствии с законами геометрической оптики, поэтому яркость L элемента поверхности, имеющего координаты (x, y), зависит от направления вектора нормали $\mathbf{n}$ к поверхности в точке (x, y), который в свою очередь связан с локальными уклонами ξ_x, ξ_y в этой точке:

$$\mathbf{n} = \frac{(-\xi_x, -\xi_y, 1)}{\sqrt{1 - \xi_x^2 - \xi_y^2}}.$$

При разработке математической модели регистрируемого на оптическом изображении сигнала целесообразно разделить в нем составляющие, различным образом связанные с уклонами морской поверхности. Необходимость такого разделения следует из решаемой задачи восстановления спектров уклонов и возвышений поверхности по оптическим изображениям. В соответствии с этим принципом представим сигнал $L(x, y)$ в виде следующей суммы [1, 11, 12]

$$L(x, y) = L_\wedge\Big(x, y, \xi_x(x, y), \xi_y(x, y)\Big) + N\Big(x, y, \xi_x(x, y), \xi_y(x, y)\Big),$$

где $L_\wedge$ и N — линейная и нелинейная по уклонам составляющие соответственно, формируемые световыми полями в верхней и нижней полусферах,отражаемыми и преломляемыми элементами морской поверхности.

Поле яркости, регистрируемое аппаратурой дистанционного зондирования в фиксированный момент времени, может быть разложено в степенной ряд по уклонам поверхности и представлено в виде [1, 11, 12]:

$$L(x, y) = C_0 + C_x \xi_x(x, y) + C_y \xi_y(x, y) + N\Big(x, y, \xi_x(x, y), \xi_y(x, y)\Big),$$

где N — нелинейная составляющая сигнала, содержащая члены, пропорциональные $\xi_x^2(x, y)$, $\xi_y^2(x, y), \xi_x(x, y), \xi_y(x, y)$ и т. д.; C_0, C_x и C_y — коэффициенты линейной части разложения; ξ_x и ξ_y — поля уклонов (градиентов поля возвышений) морской поверхности. Вклад в регистрируемый сигнал нелинейной составляющей $N(x, y, \xi_x, \xi_y)$ определяется рядом параметров: условиями освещения, состоянием волнения, характеристиками регистрирующей аппаратуры. Аналитические оценки нелинейной составляющей затруднительны, поэтому для решения поставленной задачи используется метод численного моделирования [1, 11–13]. Введем определение восстанавливающего оператора $\mathbf{R}$, позволяющего перейти от спектра оптического изображения $S(\mathbf{k})$, полученного при известных условиях,

к спектру уклонов морской поверхности $\Phi(\mathbf{k})$ в направлении, определяемом этими условиями:

$$\Phi(\mathbf{k}) = \mathbf{R}S(\mathbf{k})\,. \tag{4}$$

При таком определении оператор $\mathbf{R}$ зависит от многомерного вектора $\mathbf{W}_R$, компонентами которого являются параметры условий получения оптического изображения: $\mathbf{R} = \mathbf{R}(\mathbf{k}, \mathbf{W}_R)$. В линейной модели сигнала при $N = L_f = 0$ восстанавливающий оператор тождественно равен константе: $\mathbf{R} = C^{-2}$.

3 Метод формирования восстанавливающего оператора

Разработанный метод построения восстанавливающего оператора является развитием метода, предложенного в работах [1, 5, 6, 10–14]. Для построения восстанавливающего оператора, соответствующего определенным условиям получения изображений, выполнялось прямое численное моделирование оптических изображений при заданном комплексе условий [8–11], после чего строилась аппроксимация пространственно-частотного фильтра [11–14], позволяющего получить пространственный спектр уклонов морской поверхности из спектра аэрокосмического изображения. Для валидации восстанавливающего оператора проводилось сопоставление с данными контактных измерений и стереосъемки морской поверхности со стационарной платформы [3, 5, 6].

Существенное ограничение использовавшегося подхода состояло в том, что аппроксимация восстанавливающего оператора, верифицированная с использованием данных комплексных экспериментов, относилась только к ограниченной пространственно-частотной области, а именно: к высокочастотной области степенного спадания спектральной плотности морского волнения. Предлагаемый метод построения модифицированного восстанавливающего оператора должен обеспечить получение аппроксимации восстанавливающего оператора во всей пространственно-частотной области, где имеются данные контактных измерений.

Модифицированный восстанавливающий оператор $\mathbf{R}(\mathbf{k})$ можно представить в виде суперпозиции высокочастотного и низкочастотного операторов [6]:

$$\mathbf{R}(\mathbf{k}) = \mathbf{R}_{\text{low}}(\mathbf{k})\mathbf{R}_{\text{high}}(\mathbf{k}),$$

где $\mathbf{R}_{\text{high}}(\mathbf{k})$ — восстанавливающий оператор в области высоких частот; $\mathbf{R}_{\text{low}}(\mathbf{k})$ — восстанавливающий оператор в области низких частот. При этом как $\mathbf{R}_{\text{high}}$, так и $\mathbf{R}_{\text{low}}$ будут зависеть от некоторых свободных параметров, которые представим в виде вектора параметров $\mathbf{a}$. В качестве $\mathbf{R}_{\text{high}}(\mathbf{k})$ будем использовать оператор в известной форме [13, 14]:

$$\mathbf{R}_{\text{high}}(\mathbf{k}) = a_0\Big((\cos(\varphi - \varphi_c))^{a_3} k^{a_1 + a_2\cos(\varphi - \varphi_c)}\Big)\,. \tag{5}$$

Низкочастотную составляющую восстанавливающего оператора $\mathbf{R}_{\text{low}}(\mathbf{k})$ будем строить в виде, характерном для большинства аппроксимаций спектров морского волнения [6]:

$$\mathbf{R}_{\text{low}}(\mathbf{k}) = \exp(a_4 k^{a_5}), \tag{6}$$

где a_4 и a_5 — свободные настраиваемые параметры, которые в общем случае зависят от гидрометеорологических условий.

Пространственный спектр уклонов поверхностного волнения $\Phi(\mathbf{k})$ получается путем применения восстанавливающего оператора к двумерному спектру аэрокосмического изображения [1, 10–14]. Для получения значений компонент вектора $\mathbf{a}$ будем использовать

данные о спектрах морского волнения, получаемых с помощью контактных измерений с использованием решетки струнных волнографов [5], при этом будем учитывать дисперсионное соотношение для гравитационных волн [2]:

$$\omega^2 = gk.$$

Дисперсионное соотношение связывает циклическую частоту $\omega = 2\pi/T$ волны, распространяющейся по водной поверхности с временны́м периодом T, и пространственную циклическую частоту $\omega = 2\pi/\Lambda$. Диапазон пространственных частот, соответствующих частотам, измеряемых контактным методом, в общем случае отличается от диапазона пространственных частот в спектре, восстанавливаемом по спектру аэрокосмического изображения. Область на плоскости пространственных частот (k_x, k_y), в которой возможно измерение поверхностных волн как контактным, так и дистанционным методами, назовем общей пространственно-частотной областью Ξ_{com}.

Обозначим пространственный спектр уклонов морских волн в общей пространственно-частотной области $\Phi_{\text{com}}(\mathbf{k})$, а соответствующий спектр изображения — $S_{\text{com}}(\mathbf{k})$. Записав выражение для $\Phi_{\text{com}}(\mathbf{k})$ в полярных координатах (k, φ), где $\varphi = \arctan(k_y/k_x)$, с учетом (4), получим:

$$\Phi_{\text{com}}(k, \varphi, \mathbf{a}) = \mathbf{R}(k, \varphi, \mathbf{a}) S_{\text{com}}(k, \varphi). \tag{7}$$

Тогда спектр возвышений морской поверхности $\Psi(\mathbf{k})$, связанный со спектром аэрокосмического изображения $S(\mathbf{k})$ соотношением (3), в области низких частот определяется по формуле:

$$\Psi_{\text{com}}(k, \varphi, \mathbf{a}) = \frac{\mathbf{R}(k, \varphi, \mathbf{a}) S_{\text{com}}(k, \varphi)}{(\cos\varphi k_x + \sin\varphi k_y)^2}, \quad k \in \Xi_{\text{com}}. \tag{8}$$

Поскольку волнографом регистрируются одномерные частотные спектры $\Psi^{\text{конт}}(\omega)$, то для калибровки двумерных спектров, получаемых по аэрокосмическим изображениям, необходимо сначала перейти к одномерному пространственному спектру:

$$\chi^{\text{дист}}(k) = C \iint \Psi(k, \varphi) k \, dk d\varphi \,, \tag{9}$$

а затем, воспользовавшись дисперсионным соотношением $\omega^2 = gk$ и условием равенства энергии в элементарном объеме $\psi(\omega)d\omega = \chi(k)dk$, к частотному спектру:

$$\psi^{\text{дист}}_{\text{low}}(\omega) = \chi^{\text{дист}}_{\text{low}}(k) \frac{2\omega}{g}. \tag{10}$$

В качестве меры различия спектров, полученных дистанционным и контактным методами, используем функцию [6]

$$\text{dist}\left(\psi^{\text{дист}}_{\text{com}}(\omega), \psi^{\text{конт}}_{\text{com}}(\omega)\right) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\psi^{\text{дист}}_{\text{com}}(\omega_i) - \psi^{\text{конт}}_{\text{com}}(\omega_i)}{\psi^{\text{конт}}_{\text{com}}(\omega_i)} \right)^2}, \tag{11}$$

где $\omega_i, i = 1, \ldots, n$, — значения частоты по калибровочным данным; $\psi^{\text{дист}}_{\text{com}}(\omega)$ и $\psi^{\text{конт}}_{\text{com}}(\omega)$ — частотные спектры, полученные на основе спутниковых и контактных данных. Оптимальные значения компонент вектора параметров $\mathbf{a} = (a_0, a_1, a_2, a_3, a_4, a_5)$ находятся путем минимизации функции (11):

$$\mathbf{a} = \arg\min_{\mathbf{a}} \operatorname{dist}\Big(\psi_{\text{com}}^{\text{дист}}(\omega), \psi_{\text{com}}^{\text{конт}}(\omega)\Big). \tag{12}$$

Опишем процедуру построения модифицированного восстанавливающего оператора $\mathbf{R}(\mathbf{k})$. В области высоких частот восстанавливающий оператор полностью определяется формулой (5), а в области низких частот он определяется формулой (6). При этом оптимальные значения параметров $\mathbf{a}$ могут быть вычислены в результате оптимизационного процесса, который может быть представлен в виде последовательности этапов. Блок-схема метода формирования модифицированного восстанавливающего оператора для получения спектров морского волнения по спектрам аэрокосмических изображений приведена на рис. 1. Входными данными процедуры реализации метода являются контактные данные, получаемые от решетки струнных волнографов, аэрокосмические оптические изображения, а также набор параметров оптимизационного процесса:

– начальные значения компонент вектора параметров $\mathbf{a}$;
– интервалы сходимости этих параметров $a_i \in (a_{i,\min}, a_{i,\max}), i = 0, \ldots, 5$;
– шаги дискретизации по компонентам параметра a_i;
– параметры, определяющие границы области Ξ_{com} на плоскости k_x, k_y;
– начальное значение ошибки $\text{err}_0 = 1$.

Вычислительная процедура реализации метода состоит из нескольких рабочих процессов [6].

В **1-м рабочем процессе** выполняется анализ данных наземных контактных измерерний. В этом процессе выполняются следующие вычислительные операции (см. рис. 1):

1) по аэрокосмическому изображению определяются значения $\omega_i, i = 1, \ldots, n$, для области Ξ_{com};
2) производится аппроксимация данных, полученных при контактных измерениях частотного спектра волнения $\psi_{\text{low}}^{\text{конт}}(\omega)$, и вычисляются значения аппроксимирующей кривой в точках $\omega_i, i = 1, \ldots, n$.

Эти параметры используются далее в **3-м рабочем процессе** процедуры построения восстанавливающего оператора, описанном ниже.

Во **2-м рабочем процессе** выполняется обработка аэрокосмических изображений. На вход **2-го процесса** поступают: фрагменты аэрокосмических изображений морской поверхности и начальные значения параметров $a_{i,0}$. В процессе выполняются следующие операции:

1) вычисляется двумерный пространственный спектр аэрокосмического изображения $S(k, \varphi)$;
2) по формуле (7) вычисляется начальная оценка спектра уклонов в области Ξ_{com} для начальных параметров $a_{i,0}$.

Эти характеристики используются в **3-м рабочем процессе** процедуры построения восстанавливающего оператора. В **3-м процессе** осуществляется перебор значений свободных параметров восстанавливающего оператора, задаваемого формулой (6), с заданными шагами в заданных диапазонах их изменения. При этом выполняются следующие операции:

1) на каждом шаге по формуле (7) вычисляется пространственный спектр уклонов Φ_{com} для области Ξ_{com};
2) по формуле (8) вычисляется двумерный пространственный спектр возвышений $\Psi_{\text{com}}(k, \varphi, \mathbf{a})$ для области Ξ_{com};

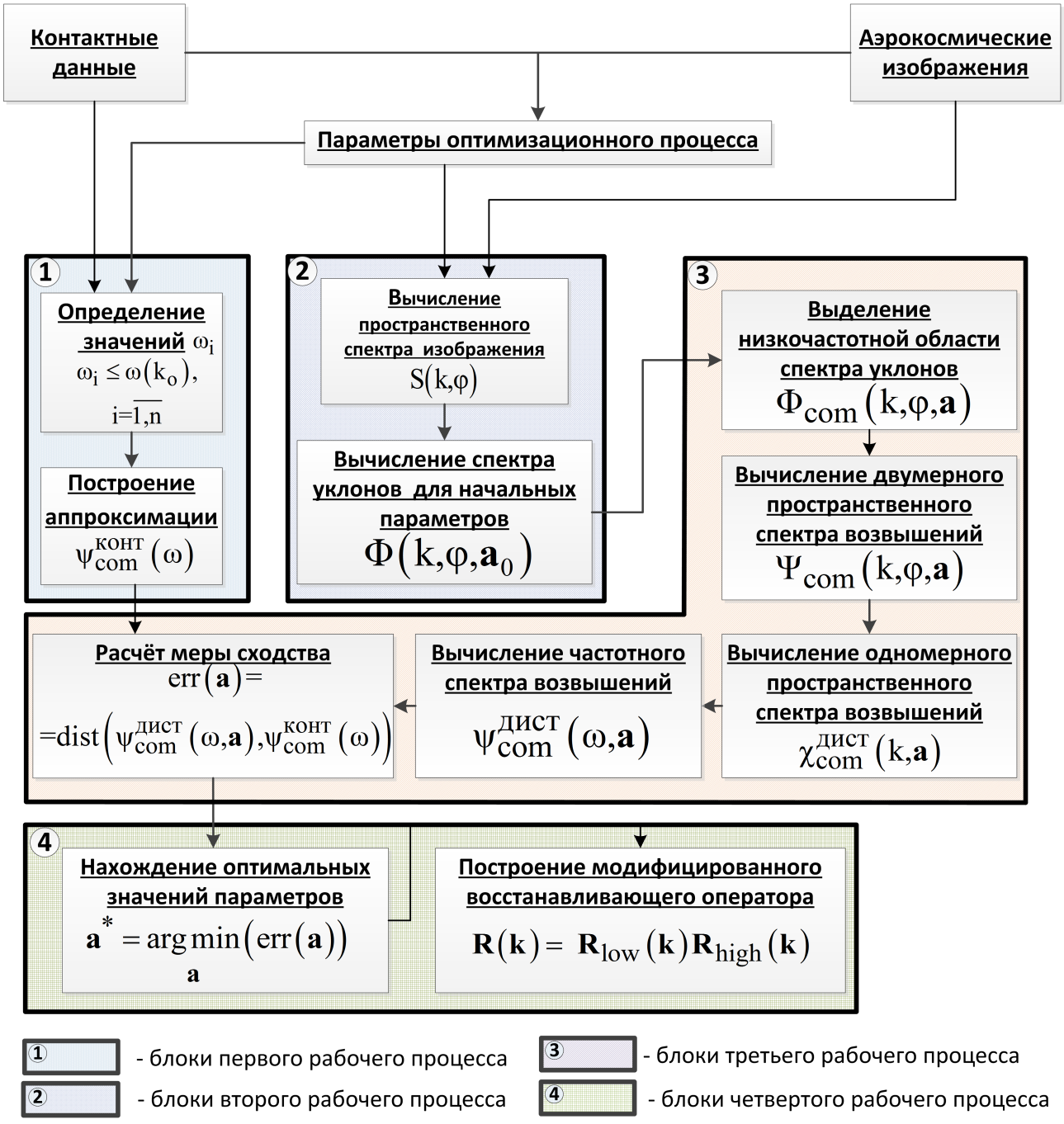


Рис. 1 Блок-схема метода формирования модифицированного восстанавливающего оператора для получения спектров морского волнения по спектрам аэрокосмических изображений

3) по формуле (9) вычисляется одномерный пространственный спектр возвышений $\chi_{\text{low}}^{\text{дист}}(k, \mathbf{a})$;
4) по формуле (10) вычисляется интегральный частотный спектр для области низких частот $\psi_{\text{low}}^{\text{дист}}$;
5) по формуле (11) рассчитывается мера различия спектров, построенных по результатам контактных измерений и по дистанционным данным $\text{err} = \text{dist}\,(\psi_{\text{com}}^{\text{дист}}(\omega), \psi_{\text{com}}^{\text{конт}}(\omega))$.

В завершение **3-го процесса** проверяется условие остановки итерационного процесса, соответствующее минимальной ошибке err_0. В **4-м рабочем процессе** процедуры построения восстанавливающего оператора определяются значения компонент вектора параметров, которые минимизируют меру различия (11) между $\psi_{\mathrm{com}}^{\mathrm{дист}}(\omega)$ и $\psi_{\mathrm{com}}^{\mathrm{конт}}(\omega)$. Затем строится модифицированный восстанавливающий оператор.

4 Вычислительные эксперименты

Проверка работоспособности разработанного метода проводилась с использованием результатов комплексных экспериментальных исследований. При этом выполнялась совместная обработка космических изображений высокого пространственного разрешения и данных, полученных при синхронных контактных измерениях с помощью решетки струнных волнографов. Оптимальные значения параметров восстанавливающих операторов рассчитывались описанным выше методом. Изучалась также применимость разработанного метода для измерений характеристик ветрового волнения при различных условиях волнообразования. В процессе исследований использовались результаты подспутниковых измерений частотных спектров возвышений морской поверхности с помощью решетки струнных волнографов, полученные в ходе экспериментов, проведенных в акватории Черного моря в районе пос. Кацивели со стационарной океанографической платформы. В ходе этих экспериментов были получены космические изображения исследуемой акватории с помощью оптической аппаратуры спутника GeoEye с пространственным разрешением 0,5 м. Экспериментальные исследования проводились при различных условиях [6]

- Эксперимент №1 — для случая слабого ветрового волнения (скорость ветра $w_{\mathrm{в}}$ от 0 до 2 м/с) в присутствии волн зыби (эксперимент проведен 16 сентября 2012 г.).
- Эксперимент №2 — для случая развитого ветрового волнения при скорости ветра $w_{\mathrm{в}} = 10$–11 м/с (эксперимент проведен 24 сентября 2015 г.).

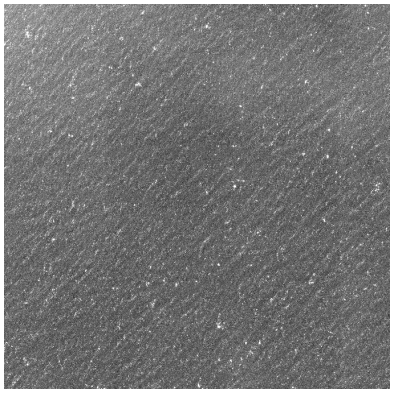
На рис. 2 представлены некоторые результаты восстановления спектров уклонов и возвышений ветрового волнения для эксперимента №2, с использованием восстанавливающего оператора $\mathbf{R}(\mathbf{k})$. По спектру фрагмента космического изображения (см. рис. 2, *а*) восстановлен двумерный спектр уклонов, показанный на рис. 2, *в*. По восстановленному двумерному спектру уклонов рассчитан одномерный частотный спектр возвышений, который сопоставлен одномерным частотным спектрам возвышений, рассчитанным по данным струнных волнографов на рис. 2, *б*. Наблюдается хорошее соответствие между данными дистанционных и контактных измерений во всем частотном диапазоне.

На рис. 2, *б* приведена также известная аппроксимация Тоба для интервала равновесия частотного спектра морского волнения, которая описывается формулой [15]:

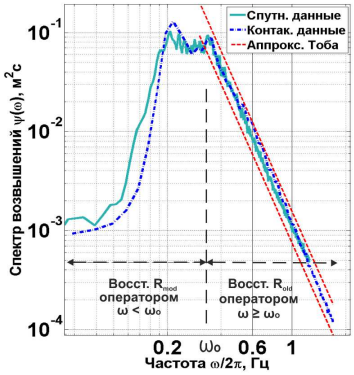
$$S(\omega) = \alpha g u_* \omega^{-4},$$

где u_* — динамическая скорость; α — коэффициент, определяемый эмпирически и равный 0,06 и 0,11 для двух показанных пунктирных линий.

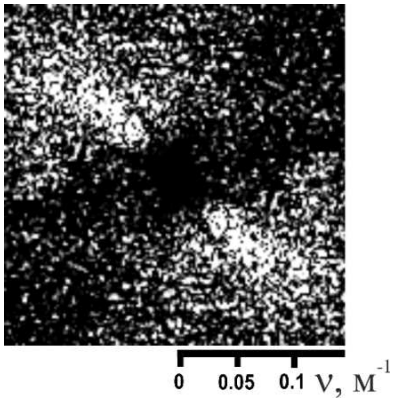
Далее на рис. 2г для сравнения приведен двумерный пространственный спектр уклонов, восстановленный оператором $\mathbf{R}_{\mathrm{high}}$, построенным с использованием первых четырех компонент вектора $\mathbf{a}$ и действующим только в области степенного спадания спектральной мощности уклонов. В восстановленном двумерном спектре уклонов на рис. 2, *г* наблюдаются явные искажения в низкочастотной области, что подтверждается ходом одномерного частотного спектра возвышений, сопоставленного на рис. 2, *д* со спектром, полученным по



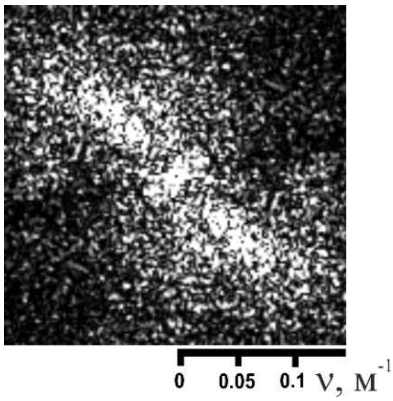
(*а*) Фрагмент космического изображения



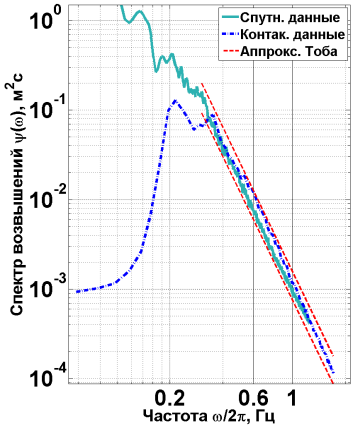
(*б*) Сопоставление одномерных частотных спектров возвышений, полученные из двумерного спектра уклонов с данными струнных волнографов



(*в*) Восстановленный двумерный спектр уклонов



(*г*) Двумерный пространственный спектр уклонов, восстановленный восстанавливающим оператором $\mathbf{R}_{\text{high}}$, в котором наблюдаются искажения в низкочастотной области

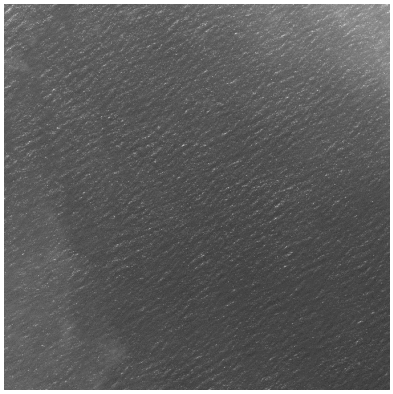


(*д*) Сопоставление с данными струнных волнографов одномерных частотных спектров возвышений, полученные из двумерного спектра уклонов при наличии искажений в низкочастотной области

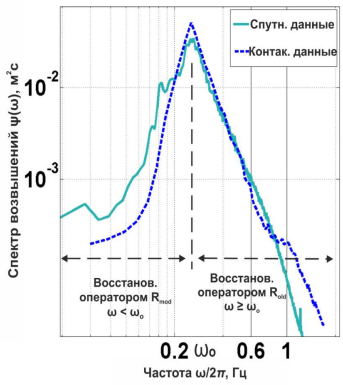
Рис. 2 Сопоставление восстановленных спектров ветрового волнения с данными струнных волнографов

контактным данным. Расхождение спектров в низкочастотной области значительно, что свидетельствует о необходимости включения низкочастотной компоненты восстанавливающего оператора $\mathbf{R}_{\text{low}}(k)$.

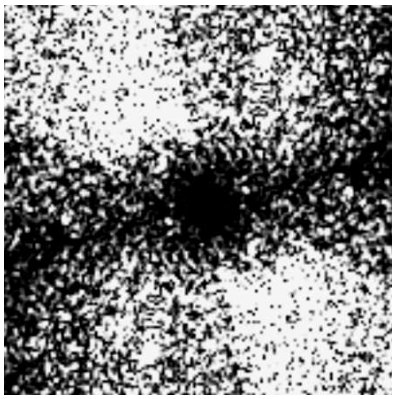
На рис. 3 представлены некоторые результаты восстановления спектров уклонов и возвышений ветрового волнения в присутствии волн зыби (эксперимент № 1), аналогичные приведенным на рис. 2 [6]. Сопоставление одномерного частотного спектра возвышений, рассчитанного по восстановленному спектру уклонов, с одномерным частотным спектром возвышений, рассчитанным по данным струнных волнографов, также демонстрирует хо-



(*а*) Фрагмент космического изображения



(*б*) Сопоставление одномерных частотных спектров возвышений, полученные из двумерного спектра уклонов с данными струнных волнографов



(*в*) Восстановленный двумерный спектр уклонов

Рис. 3 Сопоставление восстановленных спектров ветрового волнения с данными струнных волнографов [6]

Параметры аппроксимации восстанавливающего оператора, полученные разработанным методом

a_0	a_1	a_2	a_3	a_4	a_5
0,0005	−0,43	0,29	0,23	−0,3	−0,8

рошее соответствие между данными дистанционных и контактных измерений во всем частотном диапазоне.

В таблице приведены оптимальные значения компонент вектора **a** параметров аппроксимации восстанавливающего оператора, полученные в результате оптимизации по формуле (12). Начальные значения $\mathbf{a}_0$ в процедуре оптимизации задавались следующим образом. В качестве параметров аппроксимации $a_{1,0}$, $a_{2,0}$ и $a_{3,0}$ использовались значения, полученные в работе [6]. Начальные значения параметров низкочастотной части оператора задавались эмпирически: $a_{4,0} = 0{,}5$ и $a_{5,0} = 0{,}5$. Для $a_{0,0}$ использовалась эмпирическая зависимость дисперсии уклонов от скорости приповерхностного ветра [16]. Скорость ветра определялась по экспериментальным данным, полученным на океанографической платформе.

5 Заключение

Разработан метод построения оператора для восстановления спектров уклонов и возвышений морского волнения по спектрам аэрокосмических оптических изображений в широком диапазоне частот. Оптимальные параметры такого восстанавливающего оператора определяются итерационным путем при сопоставлении спектров аэрокосмических изображений со спектрами характеристик морского волнения, измеренными с высокой точностью струнными волнографами в контролируемых условиях.

В результате численной оптимизации подобраны значения параметров нелинейных восстанавливающих фильтров, работающих в различных условиях, как для развивающегося

волнения, так и в присутствии волн зыби. Расхождение спектров волнения, восстановленных по спутниковым изображениям высокого пространственного разрешения и подспутниковым данным при оптимальных значениях параметров, невелико и составляет $\sim 0{,}1$, что свидетельствует об адекватности предложенного метода построения восстанавливающего оператора.

Разработанный метод может использоваться при дистанционных исследованиях состояния поверхностного волнения, в том числе при космическом мониторинге естественных и антропогенных воздействий на морские акватории.

Работа выполнена при поддержке Министерства образования и науки РФ (проект №2015/Н8).

Литература

[1] *Бондур В. Г.* Аэрокосмические методы в современной океанологии. — В кн. “Новые идеи в океанологии” М.: Наука, Т1: Физика. Химия. Биология, 2004. С. 55-117.

[2] *Давидан И. Н., Лопатухин Н. И., Рожков В. А.* Ветровое волнение как вероятностный гидродинамический процесс. — Л.: Гидрометеоиздат, 1978. 284 с.

[3] *Барановский В. Д., Бондур В. Г., Кулаков В. В., Малинников В. А., Мурынин А. Б.* Калибровка дистанционных измерений двумерных пространственных спектров волнения по оптическим изображениям // Исследование Земли из космоса, 1992. № 2. С. 59–67.

[4] *Бондур В. Г., Филатов Н. Н., Гребенюк Ю. В., Долотов Ю. С., Здоровеннов Р. Э., Петров М. П., Цидилина М. Н.* Исследования гидрофизических процессов при мониторинге антропогенных воздействий на прибрежные акватории (на примере бухты Мамала, о. Оаху, Гавайи)// Океанология, 2007. № 6. С. 827–846.

[5] *Бондур В. Г., Дулов В. А., Мурынин А. Б., Юровский Ю. Ю.* Исследование спектров морского волнения в широком диапазоне длин волн по спутниковым и контактным данным // Исследование Земли из космоса, 2016. № 1-2. С. 7–24. doi: 10.7868/S0205961416010048.

[6] *Бондур В. Г., Дулов В. А., Мурынин А. Б., Игнатьев В. Ю.* Восстановление спектров морского волнения по спектрам космических изображений в широком диапазоне частот // Известия РАН. Физика атмосферы и океана, 2016. Т. 52. № 6. С. 717–728.

[7] *Бондур В. Г., Старченков С. А.* Методы и программы обработки и классификации аэрокосмических изображений // Известия высших учебных заведений. Геодезия и аэрофотосъемка. 2001. № 3. С. 118–143.

[8] *Бондур В. Г., Савин А. И.* Принципы моделирования полей сигналов на входе аппаратуры ДЗ аэрокосмических систем мониторинга окружающей среды // Исследование Земли из космоса, 1995. № 4. С. 24–33.

[9] *Бондур В. Г.* Методы моделирования полей излучения на входе аэрокосмических систем дистанционного зондирования // Исследование Земли из космоса, 2000. № 5. С. 16–27.

[10] *Бондур В. Г.* Моделирование двумерных случайных полей яркости на входе аэрокосмической аппаратуры методом фазового спектра // Исследование Земли из космоса, 2000. № 5. С. 28–44.

[11] *Мурынин А. Б.* Восстановление пространственных спектров морской поверхности по оптическим изображениям в нелинейной модели поля яркости // Исследования Земли из космоса, 1990. № 6. С. 60–70.

[12] *Бондур В. Г., Мурынин А. Б.* Восстановление спектров поверхностного волнения по спектрам изображений с учетом нелинейной модуляции поля яркости // Оптика атмосферы и океана, 1991. Т. 4. № 4. С. 387–393.

[13] *Мурынин А. Б.* Параметризация фильтров, восстанавливающих пространственные спектры уклонов морской поверхности по оптическим изображениям // Исследования Земли из космоса, 1991. № 5. С. 31–38.

[14] *Бондур В. Г., Мурынин А. Б.* Методы восстановления спектров морского волнения по спектрам аэрокосмических изображений // Исследования Земли из космоса, 2015. № 6. С. 3–14. doi: 10.7868/S0205961415060020.

[15] *Toba J.* Local balance in the air–sea boundary process // Oceanogr. Soc. Japan, 1973. Vol. 29. Iss. 5. P. 209–225. doi: 10.1007/BF02108528.

[16] *Cox C., Munk W.* Measurement of the roughness of the sea surface from photographs of the Sun's glitters // J. Opt. Soc. Amer., 1954. Vol. 44. Iss. 11. P. 838–850. doi: 10.1364/JOSA.44.000838.

Поступила в редакцию 30.08.2016

Parameters optimization in the problem of sea-wave spectra recovery by airspace images*

***V. G. Bondur*[1], *A. B. Murynin*[1,2], *and V. Yu. Ignatiev*[1,2]**

vgbondur@aerocosmos.info; amurynin@bk.ru; vladimir.ignatiev.mipt@gmail.com

[1]ISR "AEROCOSMOS," 4 Gorokhovskii per., Moscow, Russia

[2]Federal Research Center "Computer Science and Control" of RAS, 44/2 Vavilova Str., Moscow, Russia

The problem of the sea surface spectra reconstruction on aerospace images over a wide wavelength range is considered. Within the described nonlinear model of the brightness field, registered by remote sensing equipment, a modification of recovery operator, which acts in the whole spatiospectral domain has been proposed. The iterative process of selecting the optimal values of the modified parameters for the operator is presented using the ground truth measurements for validation. The results of the test performance for constructed operator under different registration conditions of the sea surface images are analyzed.

Keywords: *spatial spectrum; recovery operator; optimization of the parameters*

DOI: 10.21469/22233792.2.2.07

References

[1] Bondur, V. G. 2004. *Aerospace methods in modern oceanology*. New Ideas in Oceanology. V.1. Physics. Chemistry. Biology. Moscow: Nauka. P. 55-117.

[2] Davidan, I. N., N. I. Lopatukhin, and V. A. Rozhkov. 1978. *Vetrovoe volnenie kak veroyatnostnyy gidrodinamicheskiy protsess*. Leningrad: Gidrometeoizdat. 284 p.

[3] Baranovskii, V. D., V. G. Bondur, V. V. Kulakov, V. A. Malinnikov, and A. B. Murynin. 1992. Calibration of remote measurements of two-dimensional spatial spectra of wind waves by optical images. *Issledovanie Zemli iz kosmosa* [Earth Observation and Remote Sensing] 2:59–67.

[4] Bondur, V. G., N. N. Filatov, Yu. V. Grebenyuk, Yu. S. Dolotov, R. E. Zdorovennov, M. P. Petrov, and M. N. Tsidilina. 2007. Studies of hydrophysical processes during monitoring of the anthropogenic impact on coastal basins using the example of Mamala Bay of Oahu Island in Hawaii. *Oceanology* 47(6):769–787.

*The research was supported by the Russian Foundation for Basic Research (grants 14-05-91759 , 16-51-55019).

[5] Bondur, V. G., V. A. Dulov, A. B. Murynin, and Yu. Yu. Yurovskiy. 2016. A study of sea-wave spectra in a wide wavelength range from satellite and in-situ data. *Izvestia, Atmospheric and Oceanic Physics* Vol.52, 9:888–903 doi: 10.1134/S0001433816090097.

[6] Bondur, V. G., V. A. Dulov, A. B. Murynin, and V. Yu. Igantiev. 2016. Retrieving sea wave spectra using satellite imagery spectra in a wide range of frequencies. *Izvestiya, Atmospheric and Oceanic Physics* 52(6):637–648.

[7] Bondur, V. G., and S. A. Starchenkov. 2001. Metody i programmy obrabotki i klassifikatsii aerokosmicheskikh izobrazheniy [Methods and programs for treatment and classification of space images]. *Izvestiya vysshikh uchebnykh zavedeniy. Geodesiya i aerofotos"emka* 3:118–143. (In Russian.)

[8] Bondur, V. G., and A. I. Savin. 1995. Printsipy modelirovaniya poley signalov na vkhode apparatury DZ aerokosmicheskikh sistem monitoringa okruzhayushchey sredy [Principles of modeling of field signals at input of remote sensing instrumentation of aerospace environmental monitoring systems]. *Issledovaniya Zemli iz kosmosa* [Earth Observation and Remote Sensing] 4:24–33.

[9] Bondur, V. G. 2000. Metody modelirovaniya poley izlucheniya na vkhode aerokosmicheskikh sistem distantsionnogo zondirovaniya [Methods of the emission model field formed at the entrance of airspace remote sensing system]. *Issledovanie Zemli iz kosmosa* [Earth Observation and Remote Sensing] 5:16–27.

[10] Bondur, V. G. 2000. Modelirovanie dvumernykh sluchaynykh poley yarkosti na vkhode aerokosmicheskoy apparatury metodom fazovogo spektra [Simulation of two-dimensional random fields on the input brightness aerospace equipment by phase spectrum]. *Issledovanie Zemli iz kosmosa* [Earth Observation and Remote Sensing] 5:28–27.

[11] Murynin, A. B. 1990. Vosstanovlenie prostranstvennykh spektrov morskoy poverkhnosti po opticheskim izobrazheniyam v nelineynoy modeli polya yarkosti [Restoration of the spatial spectrum of the sea surface from the optical images in a nonlinear model of brightness field]. *Issledovanie Zemli iz kosmosa* [Earth Observation and Remote Sensing] 6:60–70.

[12] Bondur, V. G., and A. B. Murynin. 1991. Vosstanovlenie spektrov poverkhnostnogo volneniya po spektram izobrazheniy s uchetom nelineynoy modulyatsii polya yarkosti [Recovery spectra of surface waves on the spectra of the image based on the brightness of the nonlinear field modulation]. *Optika atmosfery i okeana* [Atmospheric and Ocean Optics] 4(4):387–393.

[13] Murynin, A. B. 1991. Parametrizatsiya fil'trov, vosstanavlivayushchikh prostranctvennye spektry uklonov morskoy poverkhnosti po opticheskim izobrazheniyam [Parametrization of filters restoring spatial spectra of slopes of sea surface using optical images]. *Issledovanie Zemli iz kosmosa* [Earth Observation and Remote Sensing] 5:31–38.

[14] Bondur, V. G., and A. B. Murynin. 2015. Methods for retrieval of sea wave spectra from aerospace image spectra. *Izvestiya, Atmospheric and Oceanic Physics* Vol.52, 9:877–887. doi: 10.1134/S0001433816090085.

[15] Toba, J. 1973. Local balance in the air–sea boundary process. *Oceanogr. Soc. Japan* 29(5):209–225. doi: 10.1007/BF02108528.

[16] Cox, C., and W. Munk. 1954. Measurement of the roughness of the sea surface from photographs of the Sun's glitters. *J. Opt. Soc. Amer.* 44(11):838–850. doi: 10.1364/JOSA.44.000838.

Received August 30, 2016

Применение теоретико-информационного подхода для сегментации изображений*

Д. М. Мурашов
d_murashov@mail.ru
ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Рассматривается задача разработки метода обеспечения наилучшего качества сегментации цифровых изображений. Метод ориентирован на применение модифицированного суперпиксельного алгоритма сегментации. В известных работах для оценки качества сегментации использовался «взвешенный показатель недостоверности», вычисляемый через значения нормализованной взаимной информации цветовых каналов входного и сегментированного изображений. Зависимость показателя недостоверности от параметра алгоритма сегментации монотонна, что потребовало обучения алгоритма и разработки итерационной процедуры выбора параметра. В данной работе в качестве критерия для оптимизации качества сегментации предлагается применять меру избыточности информации. Такой критерий обеспечивает лучший результат с точки зрения визуального восприятия. Показано, что предложенный способ построения меры избыточности позволил получить экстремальные свойства. Эксперимент, проведенный на изображениях из базы Berkeley Segmentation Dataset, подтвердил, что сегментированное изображение, соответствующее минимуму меры избыточности, дает минимальное различие по теоретико-информационной мере при сравнении с исходным изображением. Кроме того, выбранный с помощью предложенного критерия вариант сегментации дает наибольшее сходство с эталонами, имеющимися в базе.

Ключевые слова: *сегментация изображений; теоретико-информационная модель; мера избыточности; суперпиксельный алгоритм*

DOI: 10.21469/22233792.2.2.08

1 Введение

Рассматривается задача разработки метода обеспечения наилучшего качества сегментации цифровых изображений и определения соответствующего параметра модифицированного суперпиксельного алгоритма SLIC (Simple Linear Iterative Clustering) [1, 2].

В работе [3] под сегментацией понимается процесс разделения изображения, представляемого как область Ω, на n непересекающихся связных подобластей (сегментов) $\Omega_1, \Omega_2, ..., \Omega_n$, элементы которых схожи по какому-либо признаку и отличаются от элементов соседних областей. Строгое определение сегментации дано в работе [4]. При сегментации изображений возникает проблема выбора параметров применяемых алгоритмов. Параметры выбираются исходя из наилучшего качества сегментации. При этом для разных задач анализа изображений выбирается свой критерий качества. Это может быть визуальная оценка эксперта или какой-либо количественный показатель. В исследованиях по сегментации результат обычно сравнивается с изображением, сегментированным экспертом и принятым в качестве эталона [5, 6]. Возможно наличие нескольких эталонов, принятых разными экспертами.

*Работа выполнена при финансовой поддержке РФФИ, проекты № 15-07-09324 и № 15-07-07516.

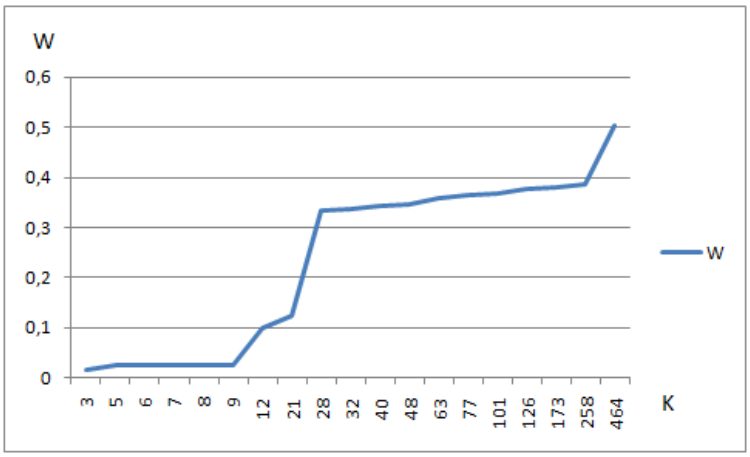


Рис. 1 Зависимость показателя недостоверности W от числа сегментов K, выделенных на изображении из базы BSDS500

Если операция сегментации рассматривается как процесс кластеризации пикселей, то применяются теоретико-множественные, статистические и теоретико-информационные меры [7], используемые для сравнения результатов кластеризации данных. Наиболее часто применяются меры хи-квадрат; индекс Рэнда (Rand Index) [8] и его варианты; мера Фаулкса–Мэллоуза (Fowlkes–Mallows) [9]; взаимная информация и варианты нормализованной взаимной информации [10]; вариация информации (variation of information) [11, 12]. Эти меры позволяют сравнить разные варианты разбиения изображений на непересекающиеся сегменты. В работе [5] отмечается, что стандартная методика для оценки эффективности алгоритмов сегментации еще не выработана.

В работе [13] предложен другой подход. При выборе параметров алгоритма оценивалось не сходство сегментированного изображения с эталоном, а сходство результата сегментации с исходным изображением. В качестве меры сходства исходного и сегментированного изображений предложено использовать «взвешенный показатель недостоверности» (weighted uncertainty index), вычисляемый через значения нормализованной взаимной информации соответствующих цветовых каналов входного и сегментированного изображений [10]. Авторы предлагают выбирать значения параметра, при которых получается наилучший с точки зрения визуального восприятия результат сегментации. Кривая зависимости показателя недостоверности от параметра и, соответственно, от числа выделенных подобластей (классов), построенная для одного изображения, не имеет экстремумов (см. [13] и рис. 1). Поэтому по экспертным оценкам результатов сегментации фрагментов серии изображений, выполненной при различных значениях параметра, на координатной плоскости с помощью классификатора, аналогичного SVM (support vector machine), выделены области «недосегментации», «пересегментации» и оптимальной сегментации. При сегментации изображений параметр алгоритма для каждой точки (x, y) обрабатываемого изображения выбирается с помощью итерационной процедуры на основе graph-cut алгоритма [14]. При этом сначала вычисляются значения показателя недостоверности для двух граничных значений параметра, а затем происходит итерационная корректировка параметра до попадания в область оптимальной сегментации. К недостаткам такого подхода следует отнести субъективность экспертных оценок, а также то, что обученный алгоритм будет давать приемлемые результаты только для тех классов изображений, которые использовались при обучении.

В данной работе лучшим вариантом сегментации считается тот, который дает приемлемое значение теоретико-информационной меры сходства при сравнении с исходным изображением, т. е. при сегментации не происходит потери значительной части информации. Принимается, что лучшее качество сегментации соответствует минимуму теоретико-информационной меры различия при сравнении с исходным изображением. Наилучшая сегментация содержит информацию только о наиболее важных деталях исходного изображения и, как и в [13], является лучшим с точки зрения визуального восприятия. В работе [15] предложена теоретико-информационная модель зрительной системы человека. В качестве основы модели использована гипотеза Барлоу [16] о минимизации избыточности информации на ранних стадиях обработки сигнала в зрительной системе человека. Предполагается, что на ранних стадиях зрительного восприятия происходит сокращение избыточности информации, поступающей от сетчатки через зрительный нерв. В данной работе, исходя из принципа минимизации избыточности информации [15], для определения наилучшего варианта сегментации изображений (и определения соответствующего параметра алгоритма) в качестве критерия предлагается применять меру избыточности информации. В предлагаемой работе установлено, что при определенном способе формирования теоретико-информационной модели системы сегментации мера избыточности обладает экстремальными свойствами. Сегментированное изображение, соответствующее минимуму меры избыточности, дает минимум меры различия при сравнении с исходным изображением. Эксперимент, проведенный на изображениях из базы Berkeley Segmentation Dataset BSDS500 [5], показал, что выбранный с помощью предложенного критерия вариант сегментации дает наибольшее сходство с эталонами.

2 Модификация алгоритма сегментации SLIC

Исследования по разработке метода выбора наилучшего варианта сегментации проводились на базе суперпиксельного алгоритма SLIC [1], дополненного предлагаемой ниже процедурой постобработки. Этот алгоритм простой и позволяет достаточно быстро получить серию сегментированных изображений при вариации параметра процедуры постобработки. В следующем подразделе дается краткое описание используемого суперпиксельного алгоритма.

2.1 Алгоритм сегментации SLIC

Основная идея алгоритма сегментации SLIC заключается в кластеризации пикселей в ограниченных областях, на которые регулярным образом разбивается анализируемое изображение. Каждая точка изображения характеризуется пятимерным вектором $\boldsymbol{p} = (c_1, c_2, c_3, x, y)^{\mathrm{T}}$, где c_1, c_2 и c_3 — координаты точки в выбранном цветовом пространстве; x и y — пространственные координаты точки изображения. Авторы метода использовали цветовое пространство CIE Lab.

Алгоритм включает следующие шаги.

1. Изображение разбивается на K фрагментов размера $a \times a$, которые задают начальное приближение кластеров-суперпикселей. В качестве начальных центров суперпиксельных фрагментов выбираются их геометрические центры C_k.
2. Корректируются координаты центров фрагментов из условия минимального значения цветового градиента в 3×3 окрестности геометрического центра.
3. Формирование локальных кластеров в $2a \times 2a$ окрестности центров C_k аналогично методу k-средних. Расстояние D между центром и точками фрагмента вычисляется

как комбинация евклидовых расстояний по цветовой d_c и пространственной d_s составляющим описания точки $\boldsymbol{p}$:

$$d_c = \sqrt{(c_{j1} - c_{i1})^2 + (c_{j2} - c_{i2})^2 + (c_{j3} - c_{i3})^2}\,; \tag{1}$$

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}\,; \tag{2}$$

$$D = \sqrt{d_c^2 + \left(\frac{d_s}{a}\right)^2 m^2}, \tag{3}$$

где m — параметр, задающий соотношение вкладов двух составляющих описания изображения в величину расстояния D; i и j — номера точек, между которыми вычисляется расстояние.

4. Определение новых центров кластеров и вычисление смещений центров.
5. Повтор шагов 3 и 4 до тех пор, пока смещение центров между итерациями не станет меньше заданного значения.

Чтобы выделить однородные области, соответствующие объектам, зафиксированным на изображении, необходимо объединить отдельные суперпиксельные кластеры. Для этого предлагается процедура постобработки, описанная в следующем подразделе.

2.2 Процедура постобработки

В представляемой работе для сегментации цветного изображений предлагается применять алгоритм SLIC с двухступенчатой постобработкой. Целью постобработки является объединение полученных суперпикселей в однородные области, соответствующие объектам исходного изображения.

На первой ступени предлагается производить объединение соседних суперпиксельных областей. Для принятия решения об объединении используется пороговое решающее правило, разрешающее объединение, если выполняется неравенство:

$$d\,(C_i C_j) \leqslant \Delta_1\,; \tag{4}$$

$$d\,(C_i C_j) = \sqrt{(c_{1j} - c_{1i})^2 + (c_{2j} - c_{2i})^2 + (c_{j3} - c_{3i})^2}\,, \tag{5}$$

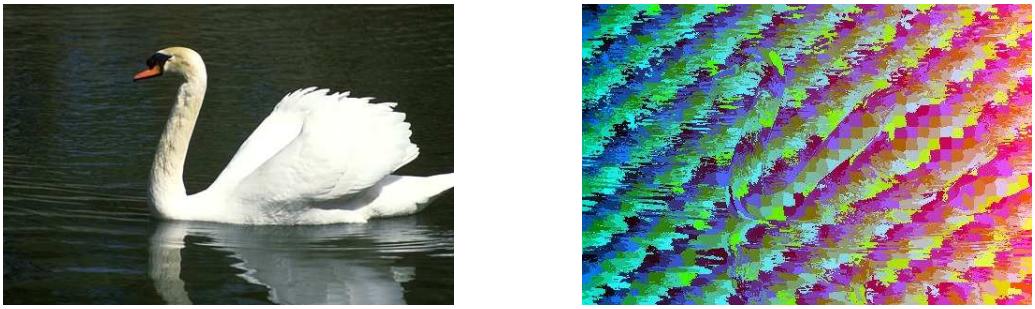
где $d\,(C_i C_j)$ — расстояние между центрами соседних суперпикселей с номерами i и j в выбранном цветовом пространстве; c_{1k}, c_{2k} и c_{3k} — координаты центра k; Δ_1 — пороговое значение.

На второй ступени предлагается объединять суперпиксельные кластеры в пределах всего изображения. Для принятия решения об объединении, так же как и на первой ступени, используется пороговое решающее правило, разрешающее объединение, если выполняется неравенство:

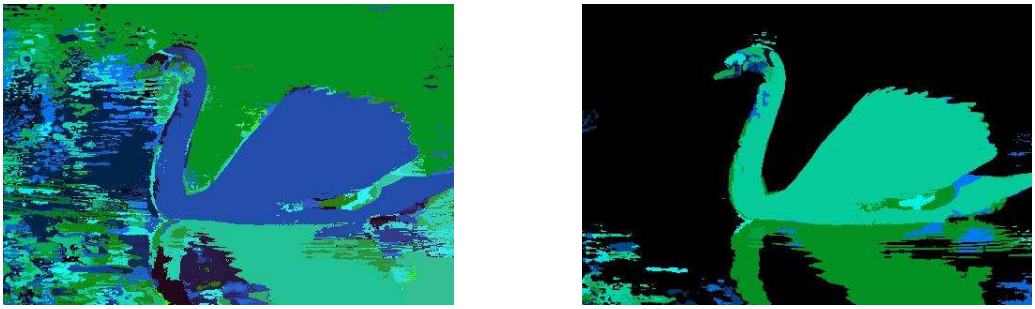
$$d\,(C_i C_j) \leqslant \Delta_2, \tag{6}$$

где Δ_2 — пороговое значение.

Процедура включает следующие шаги: (а) просмотр массива центров суперпиксельных кластеров изображения и формирование матрицы объединения соседних суперпикселей по правилу (4)–(5); (б) объединение соседних суперпиксельных кластеров; (в) нахождение центров новых кластеров; (г) просмотр массива центров кластеров для формирования матрицы объединения суперпикселей по правилу (6); (д) объединение схожих суперпиксельных кластеров. Результаты сегментации изображения из базы BSDS500 показаны на рис. 2.



(*а*) (*б*)



(*в*) (*г*)

Рис. 2 Результат работы алгоритма сегментации с двухступенчатой постобработкой: (*а*) исходное изображение; (*б*) результат сегментации без постобработки; (*в*) и (*г*) изображения, полученные после первой и второй ступеней постобработки

Таким образом, предлагаемая процедура на базе алгоритма суперпиксельной сегментации SLIC содержит четыре параметра: начальная длина стороны суперпиксельной области a, параметр m, задающий соотношение вкладов цветовой и пространственной составляющих описания точек изображения в величину расстояния (1)–(3) и два пороговых значения Δ_1 и Δ_2 в решающих правилах (4)–(6).

В следующем разделе предлагается метод, позволяющий выбрать наилучший вариант сегментации и определить соответствующее значение порога первой ступени процедуры постобработки.

3 Выбор наилучшего варианта сегментации и определение параметров алгоритма

Выбор параметров алгоритма осуществляется следующим образом. Начальный размер a суперпикселей задается по размеру наименьших деталей изображения, которые должны быть выделены. Параметр m специфичен для каждой решаемой задачи. Наибольший интерес представляет выбор порогов Δ_1 и Δ_2 правил (4)–(6) процедуры постобработки.

Для применения теоретико-информационного подхода необходима вероятностная модель связи между исходным и сегментированным изображениями. При оценивании качества сегментации будем использовать каналы цветового пространства CIE Lab (например, L).

Пусть исходное и сегментированное изображения являются входом и выходом стохастической информационной системы. Уровни светлоты L изображений описываются

непрерывными случайными переменными U и V с плотностями вероятности $p(u)$ и $p(v)$, где u и v — значения переменных U и V. Операция сегментации будет представлена моделью информационного канала:

$$V = F(U + \eta), \tag{7}$$

где U — сигнал на входе канала; V — выход канала; F — функция преобразования; η — шум канала. Предполагается, что шум является гауссовой случайной переменной с нулевым средним и дисперсией σ^2; переменные V и η независимы.

В качестве критерия качества сегментации изображений предлагается использовать меру избыточности канала, определяемую выражением [15]:

$$R = 1 - \frac{I(U;V)}{C(V)}, \tag{8}$$

где $I(U;V)$ — взаимная информация между входом и выходом; $C(V)$ — пропускная способность канала.

Положим $C(V) = H(V)$, где $H(V)$ — энтропия выхода. Тогда, учитывая, что $I(U;V) = H(V) - H(V \mid U)$, выражение (8) примет вид:

$$R = \frac{H(V \mid U)}{H(V)}, \tag{9}$$

где $H(V \mid U)$ — условная энтропия выхода канала при условии, что вход равен U.

Покажем, что величина меры избыточности системы сегментации, описываемой моделью (7)–(9), зависит от количества сегментов и имеет минимум.

В соответствии со способом представления сегментированного изображения плотность вероятности выхода системы может быть представлена суммой:

$$p(v) = \sum_{k=1}^{K} P(v_k)\,\delta(v - v_k), \tag{10}$$

где $P(v_k)$ — вероятность появления значения v_k; $\delta(v - v_k)$ — дельта-функция; K — количество выделенных сегментов. В случае непрерывной модели (7) дифференциальная энтропия выхода системы с учетом (10) равна:

$$H(V) = -\int_{-\infty}^{+\infty} p(v) \log p(v)\, dv = -\sum_{k=1}^{K} P(v_k) \log P(v_k)\ . \tag{11}$$

Пусть все значения V равновероятны: $P(v_k) = 1/K$. Тогда из (11) получим

$$H(V) = \log K. \tag{12}$$

Далее найдем значение условной дифференциальной энтропии $H(V \mid U)$. Условная энтропия $H(V \mid U)$ является мерой информации о шуме η сигнала, которая измеряется на выходе системы. В этом случае можно принять [17]:

$$H(V \mid U) = H(\eta). \tag{13}$$

Дифференциальная энтропия гауссова шума равна [17]:

$$H(\eta) = \frac{1}{2}\left[\log e + \log\left(2\sigma_\eta^2\right)\right], \tag{14}$$

где σ_η^2 — дисперсия шума системы. Тогда, подставляя (12)–(14) в (9), получим

$$R = \frac{\log e + \log\left(2\sigma_\eta^2\right)}{2\log K}. \tag{15}$$

Таким образом, мера избыточности линейно зависит от логарифма дисперсии шума и обратно пропорциональна логарифму количества полученных сегментов. Функция $R(K)$ будет иметь минимум, например, если при малых значениях K дисперсия шума близка к нулю, а при увеличении K резко возрастает. Предложена модель дисперсии шума от количества сегментов на выходе системы сегментации, которая описывается полиномом шестой степени. Модель позволяет получить минимум функции (15) и согласуется с данными эксперимента на большинстве тестовых изображений.

Учитывая зависимость избыточности от количества сегментов, выбор наилучшего варианта сегментации производится следующим образом. Исследуемое изображение U сегментируется с помощью модифицированного алгоритма SLIC с различными значениями параметра Δ_1 условия (4) процедуры постобработки. В результате сегментации изображения получено множество из Q изображений $\mathfrak{V} = \{V_1, V_2, \ldots, V_Q\}$. Далее сегментированные изображения $V_q, q = 1, 2, \ldots, Q$, сравниваются с исходным изображением U и выбирается сегментация $V_q = V_{R\min}$, при которой $R(V_q) = R_{\min}$. Выбранному изображению, состоящему из $K_{\min}$ сегментов, соответствует значение параметра $\Delta_1 = \Delta_{1\min}$.

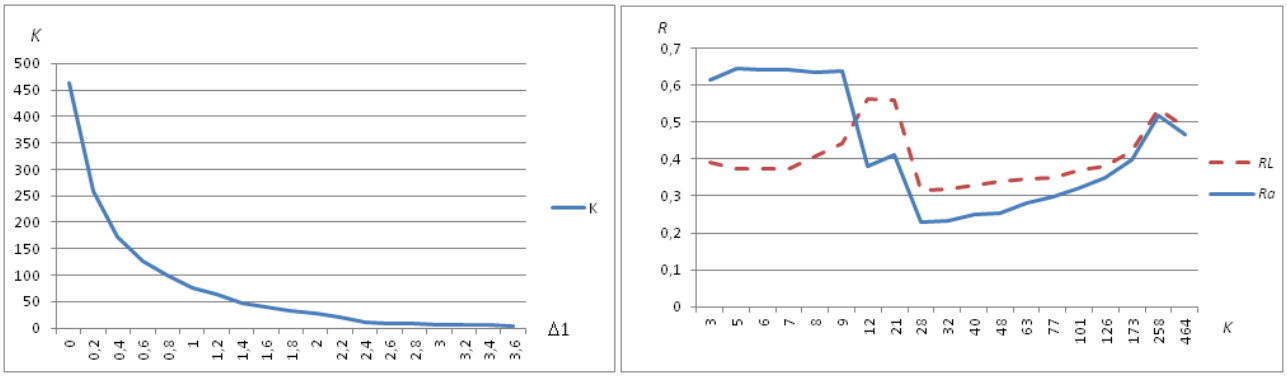
4 Вычислительный эксперимент

В данной работе при выборе параметров использовались изображения из базы BSDS500 университета Беркли [5]. Каждое из исследуемых изображений сегментировалось с помощью модифицированного алгоритма SLIC с различными значениями параметра Δ_1 условия (4) процедуры постобработки. В результате сегментации изображения U получено множество из Q изображений $\mathfrak{V} = \{V_1, V_2, \ldots, V_Q\}$. Далее сегментированные изображения сравнивались с исходным изображением U и с вариантами эталонной сегментации $V_t^{GT}, t = 1, 2, \ldots, T$ из базы BSDS500. Для сравнения изображений применялась вариация информации (variation of information) — теоретико-информационная мера, предложенная в работах [11, 12]. Вариация информации является метрикой и обладает свойствами, полезными для сравнения результатов кластеризации данных. В рассматриваемом случае вариация информации характеризует различие (расстояние) между двумя версиями сегментации (или между исходным и сегментированным изображением) и определена следующим образом:

$$VI(S, S') = H(S) + H(S') - 2I(S; S'), \tag{16}$$

где $VI(S, S')$ — вариация информации; $H(S)$ и $H(S')$ — энтропии сравниваемых изображений S и S'; $(S; S')$ — взаимная информация сравниваемых изображений. С целью подтверждения того, что сегментированное изображение, обеспечивающее минимальное значение меры избыточности, является наилучшим в смысле сходства с входным изображением и сходства с эталонными вариантами сегментации, решались следующие задачи. Во-первых, для исходного изображения U и каждого из изображений V_q, $q = 1, 2, \ldots, Q$,

Рис. 3 Изображение из базы BSDS500

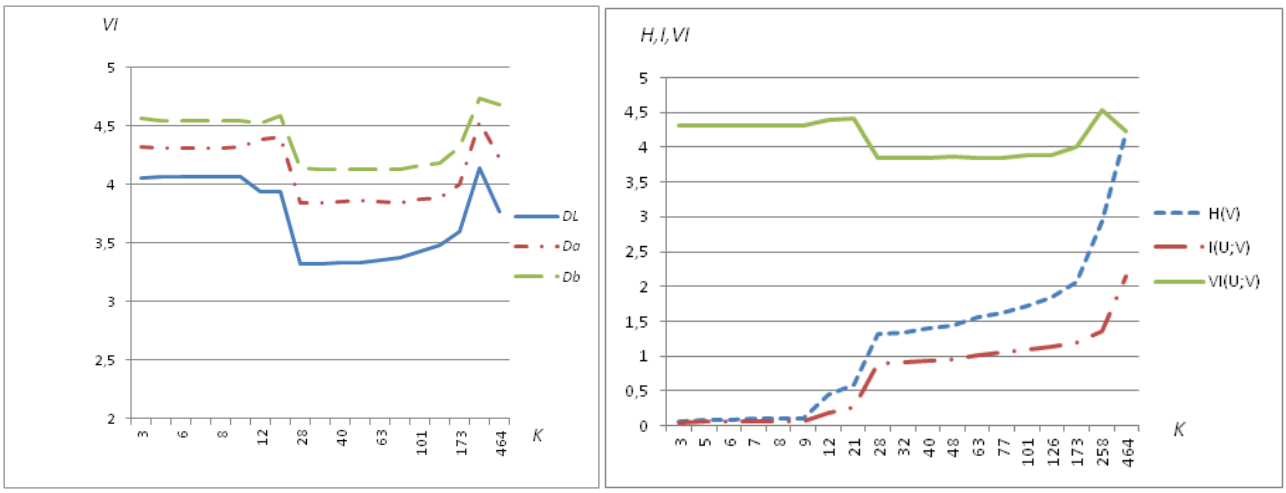


(*а*) (*б*)

Рис. 4 Результат работы алгоритма сегментации: (*а*) соотношение между значением порога Δ_1 и количеством сегментов K; (*б*) зависимости меры избыточности R от числа сегментов K в каналах L (RL) и a (Ra) цветового пространства CIE Lab для изображения, показанного на рис. 3

полученных с помощью модифицированного алгоритма сегментации при различных значениях параметра, вычислены значения меры избыточности R. Затем среди изображений V_q находилось изображение $V_{R\min}$, соответствующее глобальному минимуму R. Во-вторых, сравнивалось расстояние от входного изображения U до найденного варианта сегментации $V_{R\min}$ по выбранной мере (9) с расстояниями до других изображений V_q. В-третьих, эталонные сегментации V_t^{GT} сравнивались с входным изображением U. В-четвертых, сравнивались расстояния между каждым из эталонов V_t^{GT} и вариантом сегментации $V_{R\min}$ с расстояниями между эталонами и сегментациями V_q. В эксперименте использовалось 20 изображений. Сегментация изображений производилась в цветовом пространстве CIE Lab. Результаты эксперимента будут продемонстрированы на изображении 118035.jpg (рис. 3).

Получены следующие результаты. С помощью алгоритма SLIC с процедурой постобработки произведена сегментация изображений из базы при значениях порога Δ_1 в интервале $0 \leqslant \Delta_1 \leqslant 3{,}6$. Соответствие между значением порога и количеством сегментов на полученных изображениях показано на рис. 4, *а*.

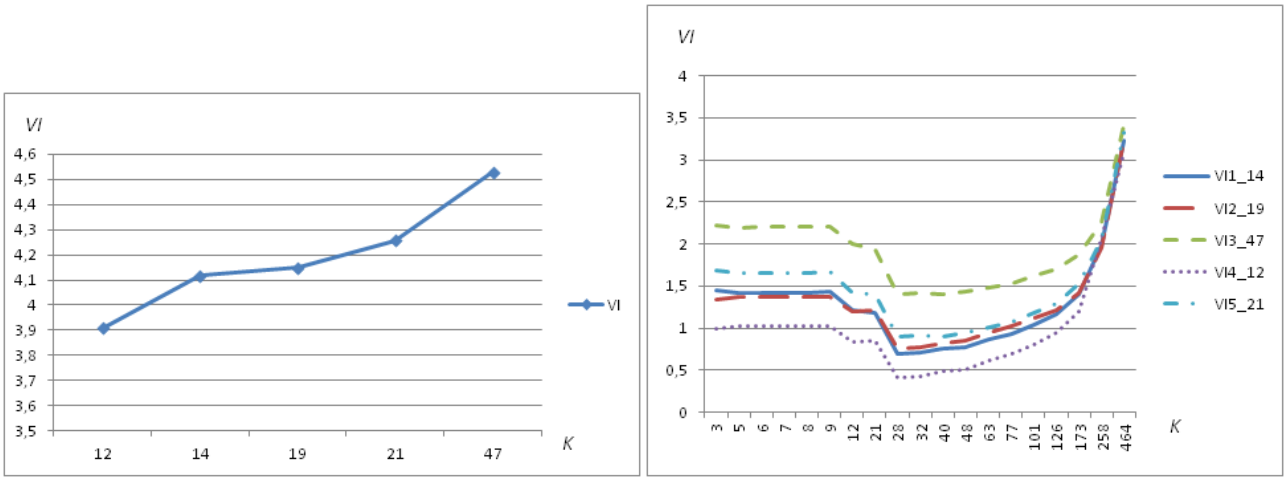


(*а*) (*б*)

Рис. 5 Результат работы алгоритма сегментации: (*а*) изменение вариации информации VI в каналах L, a и b при сравнении результатов сегментации с исходным изображением (см. рис. 3); (*б*) энтропия выхода $H(V)$, взаимная информация $I(U;V)$ и вариация информации $VI(U,V)$ в системе сегментации при разном количестве сегментов K

Для цветовых каналов исходного изображения U и каждого из полученных при сегментации изображений V_q, состоящих из K сегментов, вычислен показатель избыточности R. Полученные значения меры избыточности в каналах L и a представлены в виде графика на рис. 4, *б*. Минимум показателя избыточности достигается во всех цветовых каналах при $K = 28$, что соответствует $\Delta_1 = 2$.

По результатам сравнения исходного изображения U с полученными при значениях порога $0 \leqslant \Delta_1 \leqslant 3{,}6$ сегментированными изображениями V_q построены графики в пространстве «число сегментов K — величина вариации информации VI в каналах L, a и b», показанные на рис. 5, *а*. На рисунке область $K < 12$ соответствует значениям $2{,}4 \leqslant \Delta_1 \leqslant 3{,}6$. В этой области при увеличении параметра Δ_1 K уменьшается, так как происходит интенсивное слияние сегментов, исчезают мелкие и даже относительно крупные детали изображения, образуются большие однородные сегменты. Значение вариации информации в этой области диаграммы достаточно велико, так как сегментированное изображение сильно упрощается и сходство с исходным изображением слабое. При уменьшении порога Δ_1 и увеличении числа сегментов наблюдается резкое падение значения вариации информации до минимума при $\Delta_1 = 2$ (что соответствует $K = 28$) и затем ее плавный рост до $K = 173$ (соответственно, $\Delta_1 = 0{,}4$). В этом интервале значений параметра не происходит значительных изменений сегментированного изображения, все основные детали изображения сохраняются, происходит слияние малых сегментов. Далее при $K > 173$, что соответствует уменьшению Δ_1 до 0, значение VI резко возрастает. Это связано с тем, что с уменьшением размера сегментов до $a \times a$ (см. подразд. 2.1) и появлением на сегментированном изображении большого количества мелких деталей, видимых на исходном изображении, рост энтропии выхода системы $H(V)$ существенно превосходит рост взаимной информации $I(U;V)$ между выходом и входом при постоянной энтропии входа $H(U)$ (рис. 5, *б* и определение вариации информации (16)). Таким образом, из рис. 4 и 5 следует, что наибольшее сходство между входным и сегментированным изображениями достигается при значении параметра (и, соответственно, количестве сегментов), обеспечивающих минимум показателя избыточности R.



(*а*) (*б*)

Рис. 6 Результат работы алгоритма сегментации: (*а*) величина вариации информации VI, вычисленная в канале a пространства CIE Lab при сравнении исходного изображения, показанного на рис. 3, с эталонными сегментациями, состоящими из разного числа сегментов K; (*б*) вариация информации VI, вычисленная при сравнении эталонных вариантов сегментации и результатов сегментации изображения, показанного на рис. 3, при различном количестве сегментов K

(*а*) (*б*)

Рис. 7 Варианты сегментации изображения, показанного на рис. 3: (*а*) результат сегментации с постобработкой, соответствующий минимуму избыточности при $K = 28$; (*б*) эталонная сегментация, число сегментов $K = 12$

Далее для входного изображения и пяти эталонных (groundtruth) сегментаций, взятых из базы, вычислены значения вариации информации и построена кривая зависимости вариации информации числа сегментов, представленная на рис. 6, *а*. Из рисунка следует, что наименьшее значение меры различия соответствует эталонной сегментации с числом классов $K = 12$.

Следующая задача эксперимента заключалась в сравнении эталонных сегментаций исследуемого изображения и серии сегментированных изображений, полученных с помощью алгоритма суперпиксельной сегментации с процедурой постобработки. Результат сравнения в канале L показан на рис. 6, *б*. Все эталонные сегментации дают минимум расстояния по метрике (16) (наибольшее сходство) с изображением, полученным при $\Delta_1 = 2$ (28 сегментов), на котором достигается минимум меры избыточности R (см. рис. 4, *б*). Наблюдается наибольшее сходство указанного варианта сегментации с эталонным вариантом, состоящим из 12 сегментов (см. рис. 6, *б*). Сегментированное изображение с числом сегментов $K = 28$, соответствующее минимуму избыточности, и эталонное изображение с числом сегментов $K = 12$ показаны на рис. 7.

Таким образом, эксперимент показал, что из полученного множества сегментаций входного изображения, полученных при различных значениях параметра алгоритма, лучшим вариантом в смысле минимума меры различия входного и сегментированного изображений является изображение, обеспечивающее минимум меры избыточности информации для модели системы сегментации вида (7). Выбранный вариант сегментации наиболее близок к эталонным сегментированным изображениям базы BSDS500. Следует отметить, что приемлемыми результатами сегментации изображения, показанного на рис. 3, могут быть изображения с количеством сегментов от 28 до 173, соответствующие участку плавного роста меры различия (см. рис. 4, *б*).

5 Заключение

Рассмотрена задача получения наилучшего качества сегментации изображений. Предложена двухшаговая процедура постобработки для суперпиксельного алгоритма сегментации SLIC. Разработан метод выбора оптимального варианта сегментации и определения параметра процедуры на основе принципа минимума избыточности информации. В качестве критерия предложено применять меру избыточности информации, позволяющую получить наилучший с точки зрения визуального восприятия результат сегментации. Показано, что при предложенном способе формирования критерия он обладает экстремальными свойствами в отличие от известного критерия в виде взвешенного показателя недостоверности. Результаты вычислительного эксперимента показали, что сегментированное изображение, соответствующее минимуму меры избыточности, дает минимум различия при сравнении с исходным изображением. Выбранный с помощью предложенного критерия вариант сегментации показал наибольшее сходство с эталонами.

Дальнейшие исследования будут направлены на изучение и уточнение модели дисперсии шума в теоретико-информационной модели сегментации, а также уточнению границ применимости метода.

Литература

[1] *Achanta R., Shaji A., Smith K., Lucchi A., Fua P., Susstrunk S.* SLIC superpixels. Lausanne: TEPFL, 2010. Technical Report.

[2] *Achanta R., Shaji A., Smith K., Lucchi A., Fua P., Susstrunk S.* SLIC superpixels compared to state-of-the-art superpixel methods // IEEE Trans. Pattern Anal., 2012. Vol. 34. No. 11. P. 2274–2282.

[3] *Haralick R. M., Shapiro L. G.* Image segmentation techniques // Comput. Vision Graph., 1985. Vol. 29. No. 1. P. 100–132.

[4] *Gonzale R. C., Woods R. E.* Digital image processing. — 3rd ed. — Upper Saddle River, NJ, USA: Pearson Education Inc., 2008. 954 p.

[5] *Arbelaez P., Maire M., Fowlkes C., Malik J.* Contour detection and hierarchical image segmentation // IEEE Trans. Pattern. Anal., 2011. Vol. 33. No. 5. P. 898–916.

[6] Berkeley segmentation data set and benchmarks 500 (BSDS500). `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/BSR/BSR_bsds500.tgz`.

[7] *Wagner S., Wagner D.* Comparing clusterings — an overview. Universität Karlsruhe: TH, 2007. Technical Report 2006-04.

[8] *Rand W. M.* Objective criteria for the evaluation of clustering methods // J. Am. Stat. Assoc., 1971. Vol. 66. No. 336. P. 846–850.

[9] *Fowlkes E. B., Mallows C. L.* A method for comparing two hierarchical clusterings // J. Am. Stat. Assoc., 1983. Vol. 78. No. 383. P. 553–569. doi: 10.2307/2288117.

[10] *Ana L. N. F., Jain A. K.* Robust data clustering // CVPR Proceedings. — IEEE, 2003. P. 111–122.

[11] *Meila M.* Comparing clusterings by the variation of information // Learning theory and kernel machines / Eds. B. Schoelkopf, M. Warmuth. — Lecture notes in artificial intelligence ser. — Berlin–Heidelberg: Springer-Verlag, 2003. Vol. 2777. P. 173–187.

[12] *Meila M.* Comparing clusterings: An axiomatic view // 22nd Conference (International) on Machine Learning Proceedings, 2005. P. 577–584.

[13] *Frosio I., Ratner E. R.* Adaptive segmentation based on a learned quality metric // VISAPP2015 Proceedings. — INSTICC, 2015. Vol. 1. P. 283–291.

[14] *Felzenszwalb P. F., Huttenlocher D. P.* Efficient graph-based image segmentation // Int. J. Comput. Vision, 2004. Vol. 59. No. 2. P. 167–181. doi: 10.1023/B:VISI.0000022288.19776.77.

[15] *Atick J. J., Redlich A. N.* Towards a theory of early visual processing // Neural Comput., 1990. Vol. 2. No. 3. P. 308–320. doi: 10.1162/neco.1990.2.3.308.

[16] *Barlow H. B.* Possible principles underlying the transformations of sensory messages // Sensory communication / Ed. W. A. Rosenblith. — Cambridge: M.I.T. Press, 1961. P. 217–234.

[17] *Haykin S.* Neural networks: A comprehensive foundation. — 2nd ed. — Upper Saddle River, NJ, USA: Prentice Hall Inc., 1999. 869 p.

Поступила в редакцию 01.09.2016

Application of information-theoretical approach for image segmentation*

D. M. Murashov

d_murashov@mail.ru

Federal Research Center "Computer Science and Control" of RAS,

44/2 Vavilova Str., Moscow, Russia

A problem of segmentation quality of digital images is considered. The developed technique is based on the information-theoretical approach and applied to a modified superpixel segmentation algorithm. In one of the conventional techniques, the weighted uncertainty index is used for measuring segmentation quality. The index is calculated using normalized mutual information of color channels in given and segmented images. The uncertainty index varies monotonously depending on the parameter of the segmentation algorithm. This caused application of learning technique and iterative procedure for choosing parameter value. In this work, information redundancy measure is proposed as a criterion for optimizing segmentation quality. This criterion provides the best result in terms of visual perception. It is shown that proposed method of constructing the redundancy measure provides it with extremal properties. The experiment was conducted using the images from the database Berkeley Segmentation Dataset. The experiment confirmed that the segmented image corresponding to a minimum of redundancy measure produces the minimum difference in the information-theoretical dissimilarity measure when compared with the original image. In addition, the segmented image that was selected using the proposed criteria, gives the highest similarity with the groundtruth segmentations, available in the database.

*The research was supported by the Russian Foundation for Basic Research (grants 15-07-09324 and 15-07-07516).

Keywords: *image segmentation; information-theoretical model; redundancy measure; superpixel algorithm*

DOI: 10.21469/22233792.2.2.08

References

[1] Achanta, R., A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. 2010. SLIC superpixels. Lausanne: TEPFL. Technical Report.

[2] Achanta, R., A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal.* 34(11):2274–2282.

[3] Haralick, R. M., and L. G. Shapiro. 1985. Image segmentation techniques. *Comput. Vision Graph.* 29(1):100–132.

[4] Gonzale, R. C., and R. E. Woods. 2008. *Digital image processing*. 3rd ed. Upper Saddle River, NJ: Pearson Education Inc. 954 p.

[5] Arbelaez, P., M. Maire, C. Fowlkes, and J. Malik. 2011. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal.* 33(5):898–916.

[6] Berkeley segmentation data set and benchmarks 500 (BSDS500). Available at: `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/BSR/BSR_bsds500.tgz` (accessed August 30, 2016).

[7] Wagner, S., and D. Wagner. 2007. *Comparing clusterings — an overview*. Universität Karlsruhe: TH. Technical Report 2006-04.

[8] Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* 66(336):846–850.

[9] Fowlkes, E. B., and C. L. Mallows. 1983. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* 78(383):553–569. doi: 10.2307/2288117.

[10] Ana, L. N. F., and A. K. Jain. 2003. Robust data clustering. *CVPR Proceedings*. IEEE, 2003. 111–122.

[11] Meila, M.. 2003. Comparing clusterings by the variation of information. *Learning theory and kernel machines*. Eds. B. Schoelkopf and M. Warmuth. Lecture notes in artificial intelligence ser. Berlin–Heidelberg: Springer-Verlag. 2777:173–187.

[12] Meila, M. 2005. Comparing clusterings: An axiomatic view. *22nd Conference (International) on Machine Learning Proceedings*. 577–584.

[13] Frosio, I., and E. R. Ratner. 2015. Adaptive segmentation based on a learned quality metric. *VISAPP2015 Proceedings*. INSTICC. 1:283–291.

[14] Felzenszwalb, P. F., and D. P. Huttenlocher. 2004. Efficient graph-based image segmentation. *Int. J. Comput. Vision* 59(2):167–181. doi: 10.1023/B:VISI.0000022288.19776.77.

[15] Atick, J. J., and A. N. Redlich. 1990. Towards a theory of early visual processing. *Neural Comput.* 2(3):308–320. doi: 10.1162/neco.1990.2.3.308.

[16] Barlow, H. B. 1961. Possible principles underlying the transformations of sensory messages. *Sensory communication*. Ed. W. A. Rosenblith. Cambridge: M.I.T. Press. 217–234.

[17] Haykin, S. 1999. *Neural networks: A comprehensive foundation*. 2nd ed. Upper Saddle River, NJ: Prentice Hall Inc. 869 p.

Received September 1, 2016

Применение обучения с подкреплением для одновременного выбора модели алгоритма классификации и ее структурных параметров*

В. А. Ефимова, А. А. Фильченков, А. А. Шалыто
efimova@rain.ifmo.ru; afilchenkov@corp.ifmo.ru; shalyto@mail.ifmo.ru
Университет ИТМО, Россия, г. Санкт-Петербург, Кронверкский проспект, 49

Существует множество алгоритмов машинного обучения, однако для эффективного решения задачи интеллектуального анализа данных необходимо не только выбрать один из них, но и настроить его структурные параметры. В настоящей работе ставится задача одновременного автоматического выбора алгоритма классификации и настройки его структурных параметров и предлагается ее решение на основе решения задачи о многоруком бандите. Описываются эксперименты, проведенные на множестве реальных наборов данных. Продемонстрировано, что предложенный подход обеспечивает более высокую точность классификации по сравнению с существующими методами.

Ключевые слова: *выбор алгоритма; настройка структурных параметров; настройка гиперпараметров; оптимизация; многорукий бандит; обучение с подкреплением*

DOI: 10.21469/22233792.2.2.09

1 Введение

Совместный выбор модели алгоритма и ее структурных параметров (гиперпараметров) для обработки набора данных считается трудной задачей, на текущий момент не получившей решения. Фактически на настоящий момент задача разбивается на две подзадачи, решаемые независимо: выбор алгоритма с фиксированными структурными параметрами из конечного множества (портфолио) алгоритмов и оптимизацию структурных параметров конкретной модели алгоритма. Для того чтобы снять терминологическую неопределенность, далее в этой работе *алгоритмами* будем называть алгоритмы машинного обучения, для которых заданы структурные параметры, если же структурные параметры не заданы, такой алгоритм будем называть *моделью алгоритма*.

Первая подзадача в подавляющем большинстве случаев решается с помощью перебора. В одних из первых работ [1, 2], посвященных проблематике выбора алгоритма из портфолио, исследовалось влияние задачи классификации на выбор алгоритма и использовались решающие правила. Другими решениями первой подзадачи выступают следующие подходы к выбору алгоритма: случайным образом; на основе каких-либо эвристических правил, выработанных исследователем; помощью *k-стратового скользящего контроля* (*k-fold cross-validation*) [3]. Последний из указанных подходов предполагает запуск всех алгоритмов и последующее сравнение, что требует значительного времени, а остальные подходы не универсальны, т. е. не могут быть применены во всех случаях. Существуют более эффективные подходы, такие как, например, мета-обучение [4]. Его целью является решение задачи выбора алгоритма из портфолио алгоритмов для решения поставленной задачи без непосредственного применения каждого из них [4]. Решение этой задачи в рамках

*Работа выполнена при финансовой поддержке Правительства Российской Федерации, грант 074-U01, и РФФИ, проект №16-37-60115.

мета-обучения сводится к задаче обучения с учителем. Для этого используется заранее отобранное множество наборов данных D. Для каждого набора данных $d \in D$ вычисляется вектор мета-признаков, которые описывают свойства этого набора данных. Ими могут быть: число категориальных или численных признаков объеков в d, число возможных меток, размер d и многие другие [5]. Каждый алгоритм запускается на всех наборах данных из D. После этого вычисляется эмпирический риск, на основе которого формируются метки классов. Затем мета-классификатор обучается на полученных результатах. В качестве описания набора данных выступает вектор мета-признаков, а в качестве метки — алгоритм, оказавшийся самым эффективным с точки зрения заранее выбранной меры качества. Данная подзадача исследовалась во многих работах, например в [6, 7].

Вторая подзадача — оптимизация структурных параметров — состоит в поиске параметров, характеризующих модель алгоритма, при которых алгоритмы этой модели достигают наилучшего результата с точки зрения заранее выбранной меры качества. Например, для *метода опорных векторов* (SVM — support vector machine) структурным параметром выступает функция ядра, а при использовании *нейронной сети* — число скрытых слоев и число нейронов в них. Для ряда моделей (преимущественно статистических и регрессионных) структурные параметры подбирают аналитически или сводят задачу к более простой задаче оптимизации, как, например, это сделано в [8]. Однако в общем случае этот подход неприменим. В настоящее время разработаны алгоритмы, автоматически решающие данную задачу: *поиск по решетке* (*Grid Search*) [9], *случайный поиск* (*Random Search*) [10], *стохастический градиентный спуск* [11], *древовидный парзеновский оценщик* (*Tree-structured Parzen estimator*) [12] и *байесовская оптимизация*, к которой относится алгоритм *последовательной оптимизации по модели* (*Sequental Model-Based Optimization*, далее SMBO) [13]. В работе [14] был предложен алгоритм *последовательной конфигурации алгоритма по модели* (*Sequential model-based algorithm configuration*, далее SMBA), работающий на основе SMBO. Рассмотрим его работу. В некоторый момент времени для каждой модели алгоритма уже известно множество структурных параметров, с которыми на текущий момент она работает оптимальным образом. С помощью локального поиска в это множество добавляются наборы структурных параметров, отличающиеся от оптимальных в одной позиции и улучшающие эффективность алгоритма. Кроме того, в это множество добавляется некоторое число случайных наборов структурных параметров. Выбранные конфигурации (модели алгоритмов с заданными структурными параметрами) сортируются по *ожидаемому улучшению* (*expected improvement*), а после этого запускаются несколько лучших.

Единственным совмещающим решение двух подзадач и реализованным на данный момент подходом является полный перебор. Существует открытая библиотека для настройки алгоритмов машинного обучения Auto-WEKA [15]. Она позволяет автоматически выбирать из 27 базовых алгоритмов, 10 мета-алгоритмов и 2 ансамблевых алгоритмов лучший, одновременно настраивая его структурные параметры. Решение достигается полным перебором: оптимизация структурных параметров запускается на всех алгоритмах по очереди. Подробно работа библиотеки Auto-WEKA описана в [15]. Недостатком такого подхода является слишком большое время работы.

Решение проблемы одновременного выбора модели алгоритма и структурных параметров может также осуществляться на основе мета-обучения. Например, в [16] предложено 292 комбинации моделей алгоритмов и их структурных параметров для 6 изученных моделей алгоритмов. Однако выделение алгоритмов осуществляется эмпирическим путем, что обусловливает методологическую несостоятельность подхода для решения рассматри-

ваемой задачи, поскольку подход не гарантирует, что среди перебираемых решений будет оптимальное.

Цель данной работы — предложить метод одновременного выбора модели алгоритма и ее структурных параметров на основе конкурентного распределения времени между настройкой структурных параметров для различных моделей, который окажется эффективнее, чем существующие.

2 Постановка задачи

Для того чтобы поставить задачу одновременного выбора модели алгоритма и ее структурных параметров, сначала формально поставим две подзачади, перечисленные во Введении. Будем следовать обозначениям из [17].

Пусть A — модель алгоритма, характеризующаяся структурными параметрами $\lambda = \{\lambda_1, \ldots, \lambda_m\}, \lambda_1 \in \Lambda_1, \ldots, \lambda_m \in \Lambda_m$. Тогда с ней связано пространство структурных параметров $\Lambda = \Lambda_1 \times \cdots \times \Lambda_m$. За A_λ обозначим алгоритм, т. е. модель алгоритма, для которой задан вектор структурных параметров $\lambda \in \Lambda$.

Для обеих подзадач необходимо зафиксировать меру качества работы алгоритма. За меру качества в данной работе примем *отложенный эмпирический риск* (*hold-out empirical risk*). Эффективность алгоритма A_λ оценивается с помощью разделения набора данных на обучающую и тестовую выборку с последующим подсчетом эмпирического риска, достигаемого на тестовой выборке:

$$Q(A_\lambda, D) = \frac{1}{|D|} \sum_{x \in D} [A_\lambda(x) \neq y(x)],$$

где $[A_\lambda(x) \neq y(x)]$ — функция потерь для задачи классификации, определяющая размер ошибки при запуске алгоритма A_λ на объекте x набора данных D.

Подзадача выбора лучшего алгоритма из портфолио формулируется следующим образом. Дано некоторое множество алгоритмов с фиксированными структурными параметрами $\mathcal{A} = \{A^1_{\lambda_1}, \ldots, A^m_{\lambda_m}\}$ и обучающая выборка $D = \{d_1, \ldots, d_n\}$. Здесь $d_i = (x_i, y_i) \in X \times Y$, где X — множество признаков, описывающих объекты, а Y — конечное множество меток. Требуется выбрать алгоритм $A^*_{\lambda^*}$, который окажется наиболее эффективным с точки зрения меры качества Q. В рассматриваемом случае это сводится к задаче минимизации эмпирического риска:

$$A^*_{\lambda_*} \in \arg\min_{A^j_{\lambda_j} \in \mathcal{A}} Q(A^j_{\lambda_j}, D).$$

Подзадача оптимизации структурных параметров заключается в подборе таких $\lambda^* \in \Lambda$, при которых заданная модель алгоритма A будет наиболее эффективна. Запишем в виде формулы:

$$\lambda^* \in \arg\min_{\lambda \in \Lambda} Q(A_\lambda, D).$$

Структурные параметры для модели алгоритма подбирается в процессе их настройки. Формализуем это понятие. *Процесс настройки структурных параметров* для алгоритма A^i:

$$\pi_i(t, A^i, \{\lambda_j\}_{j=0}^k) \to \lambda^i_{k+1} \in \Lambda^i.$$

Его можно описать как работу алгоритма настройки структурных параметров, запущенного на некоторой модели алгоритма A^i с ограничением по времени t и хранящего историю

изменения вектора лучших найденных на данный момент (за k итераций) структурных параметров $\{\lambda_j\}_{j=0}^{k}$.

В данной работе предлагается одновременно решать задачи выбора модели алгоритма и оптимизации структурных параметров: дан набор моделей алгоритмов $\mathcal{A} = \{A^1, \ldots, A^k\}$, с каждым из которых связано пространство структурных параметров $\Lambda^1, \ldots, \Lambda^k$ соответственно. Необходимо найти алгоритм $A^*_{\lambda^*}$, минимизирующий эмпирический риск:

$$A^*_{\lambda^*} \in \underset{A^j \in \mathcal{A}, \lambda \in \Lambda^j}{\arg\min} Q(A^j_\lambda, D).$$

Предположим, что дан один или некоторое множество алгоритмов настройки структурных параметров для моделей алгоритмов из $\mathcal{A}$, задающих соответствующие процессы π_i. Тогда предыдущая задача сводится к запуску этих процессов. Но возникает новая задача — задача минимизации времени, потраченного на поиск лушчего алгортима. С практической точки зрения больший интерес представляет похожая задача, а именно: задача поиска лучшего алгоритма за фиксированное время. Эта задача и решается в данной работе. Формализуем ее.

Пусть задан некоторый временной бюджет T на поиск лучшего алгоритма $A^*_{\lambda^*}$. Требуется разбить его на интервалы $T = t_1 + \cdots + t_m$ таким образом, что при запуске процессов π_i с ограничением по времени t_i соответственно получим минимальный эмпирический риск:

$$\min_j Q(A^j_{\lambda_j}, D) \xrightarrow[(t_1,\ldots,t_m)]{} \min,$$

где $A^j \in \mathcal{A}, \lambda_j = \pi_j(t_j, A^j, \varnothing)$ и $t_1 + \cdots + t_m = T$; $t_i \geqslant 0 \, \forall i$.

3 Предлагаемый подход

В рассматриваемой задаче ключевым ресурсом является время T, отведенное на оптимизацию структурных параметров. Для удобства разобьем его на q небольших равных интервалов t, которые будем называть *временными бюджетами*. Теперь перейдем к задаче распределения временных бюджетов. Рассматриваемую задачу можно представить в следующем виде: через равные временны́е интервалы необходимо решать, какой процесс настройки структурных параметров будет выполняться в течение следующего интервала.

Заведомо неизвестно, какое качество будут показывать алгоритмы той или иной модели при решении конкретной задачи. Выделяя время только на модель, алгоритмы которой показывают лучшие результаты («эксплуатируя» соответствующую модель), можно упустить хорошие алгоритмы других моделей. Наоборот: при равномерном распределении времени между моделями настройка структурных параметров для моделей алгоритмов, неэффективных при решении данной задачи, занимает слишком много времени. Поэтому задача представляет собой поиск компромисса между исследованием (поочередным запуском настройки структурных параметров для всех моделей алгоритмов) и эксплуатацией (запуском настройки только для одной модели алгоритма). Его поиск производится с помощью обучения с подкреплением, частным случаем которого является задача о многоруком бандите [18]. Остальные подвиды обучения с подкреплением решают задачу в начальных предположениях, несколько отличающихся от условий задачи одновременного выбора модели алгоритма и ее структурных параметров, например они предполагают, что среда изменяется в зависимости от выбранного действия. Поскольку в данной задаче это не так, остальные виды обучения с подкреплением рассматриваться не будут.

В задаче о многоруком бандите рассматривается бандит с N ручками, дернув за любую из которых можно получить выигрыш некоторого размера, определяемый случайным

распределением, ассоциированным с данной ручкой. В каждый момент времени k агент выбирает ручку a_i и получает выигрыш $r(i,k)$. Цель агента — минимизировать суммарные потери (по сравнению с лучшей стратегией) за конечное время T. В данной работе использовались следующие алгоритмы решения задачи о многоруком бандите, описанные в [18]:

1) ε-жадный — на каждой итерации находит для каждой ручки a средний выигрыш $\bar{r}_{a,t}$, затем с вероятностью $1-\varepsilon$ выбирает ручку с максимальным средним выигрышем, а с вероятностью ε выбирает случайную ручку. Если перейти к пределу, то каждая ручка будет выбрана бесконечное число раз, так что средние выигрыши с вероятностью 1 сойдутся к настоящему выигрыш;
2) UCB1 — на стадии инициализации агент выбирает все ручки по очереди. Далее на итерации t выбирает ручку a_t для которой выполняется:

$$a_t = \underset{i=1,\ldots,N}{\arg\max}\, \bar{r}_i + \sqrt{\frac{2\ln t}{n_i}},$$

где $\bar{r}_i$ — средний выигрыш при выборе ручки i; n_i — число раз, которое выбиралась ручка i. Стоит отметить, что данный алгоритм прост в реализации;
3) Softmax — на стадии инициализации агент выбирает все ручки по очереди. Далее на итерации t выбирает ручку a_i с вероятностью:

$$p_{a_i} = \frac{e^{\bar{r}_i/\tau}}{\sum_{j=1}^{N} e^{r_j/\tau}},$$

где τ — положительный параметр, называемый температурой. При $\tau \to 0$ Softmax подобен жадному алгоритму.

В качестве ручек будем рассматривать процессы настройки структурных параметров $\{\pi_i(t, A^i, \{\lambda_k\}_{k=0}^q) \to \lambda_{q+1}^i \in \Lambda^i\}_{i=0}^m$ для множества алгоритмов $\mathcal{A} = \{A_{\lambda_1}^1, \ldots, A_{\lambda_m}^m\}$. После выбора ручки $i = a_k$ на итерации k будем выдавать процессу π_{a_k} некоторый промежуток времени t на настройку структурных параметров, в конце которого получим вектор структурных параметров λ_k^i. После завершения работы выбранного процесса оценку результата произведем, вычисляя эмпирический риск для процесса π_i на итерации k как $Q(A_{\lambda_k^i}^i, D)$.

Функцию выигрыша $r(i,k)$ можно определить двумя способами. В первом (простейшем) случае наградой будет выступать разница между оптимальным эмпирическим риском, найденным в ходе предыдущих итераций, и текущим эмпирическим риском.

Для второго способа воспользуемся особенностями работы алгоритма SMAC, описанного во Введении. Для нашей цели воспользуемся математическим ожиданием эмпирического риска на шаге k, с помощью которого вычисляется ожидаемое улучшение: $E_t(Q(A_{\lambda_k^i}^i, D))$, где $Q(A_{\lambda_k^i}^i, D)$ — эмпирический риск, достигаемый процессом p_i на наборе данных D в момент времени k.

Заметим, что процесс π_i решает задачу минимизации эмпирического риска, тогда как задача о многоруком бандите — задачи о максимизации выигрыша. Поэтому функцию среднего выигрыша определим следующим образом:

$$\bar{r}_{i,(k)} = \frac{Q_{\max} - E_{(k)}(Q(A_{\lambda_k^i}^i, D))}{Q_{\max}},$$

где $Q_{\max}$ — максимальный эмпирический риск, достижимый при решении данной задачи.

4 Экспериментальное исследование

Выше было рассмотрены существующие подходы к решению задачи. Поскольку на данный момент наиболее полно задачу можно решить с помощью библиотеки Auto-WEKA, для сравнения выбрана именно она.

Исследования проводились на десяти наборах реальных данных, находящихся в репозитории UCI и доступных по ссылке `http://www.cs.ubc.ca/labs/beta/Projects/autoweka/datasets/`, описание которых представлено в табл. 1. Предложенный подход позволяет использовать любой метод настройки структурных параметров, но для более корректного сравнения с библиотекой Auto-WEKA был выбран метод, реализованный в ней, а именно: SMBO. Рассматриваются 6 известных моделей алгоритмов классификации: *метод ближайших соседей (kNN)*, *метод опорных векторов*, *логистическая регрессия* (*Logistic Regression*), *случайный лес* (*Random Forest*), *персептрон* (*Perceptron*), *дерево принятия решений* (*C4.5 Decision Tree*). В табл. 2 приведено число категориальных и численных структурных параметров для каждой из рассмотренных моделей алгоритмов классификации.

В поставленной задаче временной бюджет выделяется на ее полное решение, тогда как в предложенном алгоритме производится разбиение общего временно́го отрезка на равные более мелкие отрезки, которые по одному выдаются процессам на каждой итерации.

Таблица 1 Описание использованных наборов данных

Название набора данных	Число категориальных признаков	Число численных признаков	Число классов	Число объектов в обучающей выборке	Число объектов в тестовой выборке
Dexter	0	20000	2	420	180
German Credit	13	7	2	700	300
Dorothea	0	100000	2	805	345
Yeast	0	8	10	1039	445
Secom	0	590	2	1097	470
Semeion	0	256	10	1116	477
Car	6	0	4	1210	518
KR-vs-KP	36	0	2	2238	958
Waveform	0	40	3	3500	1500
Shuttle	38	192	2	35000	15000

Таблица 2 Число категориальных и численных структурных параметров, настраиваемых у моделей алгоритмов, между которыми производился выбор

Модель алгоритма	Категориальные	Численные
Метод ближайших соседей	4	1
Метод опорных векторов	4	6
Логистическая регрессия	0	1
Случайный лес	2	3
Перцептрон	5	2
Дерево принятия решений C4.5	6	2

Таблица 3 Сравнение предложенной активной стратегии с библиотекой Auto-WEKA по наименьшему достигнутому эмпирическому риску Q

Набор данных	AutoWEKA	UCB1	0,4-жадный	0,6-жадный	Softmax	$\text{UCB1}_{E(Q)}$	$\text{Softmax}_{E(Q)}$
Car	0,3305	**0,1836**	**0,1836**	**0,1836**	**0,1836**	**0,1836**	**0,1836**
Yeast	34,13	**29,81**	**29,81**	**33,65**	**29,81**	**29,81**	**29,81**
KR-vs-KP	0,2976	**0,1488**	**0,1488**	**0,1488**	**0,1488**	**0,1488**	**0,1488**
Semeion	4,646	**1,786**	**1,786**	**1,786**	**1,786**	**1,786**	**1,786**
Shuttle	**0,00766**	0,0115	0,0115	**0,00766**	0,0115	**0,0076**	**0,0076**
Dexter	7,143	2,38	2,381	2,381	2,381	2,381	**0,16**
Waveform	11,28	**8,286**	**8,286**	**8,286**	**8,286**	**8,286**	**8,286**
Secom	4,545	**3,636**	4,545	4,545	**3,636**	**3,636**	**3,636**
Dorothea	6,676	4,938	4,958	4,938	4,938	4,32	**2,469**
German Credits	19,29	**14,29**	**14,29**	15,71	**14,29**	**14,29**	**14,29**

Авторами было проведено исследование поведения предложенного алгоритма для различных значений временно́го бюджета, выделяемого на одну итерацию, чтобы найти значение, при котором результат оптимален. Были рассмотрены временны́е бюджеты от 10 до 60 с, с шагом в 3 с. Предложенный алгоритм был запущен на трех наборах данных из описанных выше: Car, German Credits и KRvsKP. Для решения задачи о многоруком бандите использовались алгоритмы UCB1, 0,4-жадный, 0,6-жадный и Softmax. Запуск каждой конфигурации производился трижды. В результате был выбран временной бюджет в 30 с на итерацию.

Рассматривалась работа предложенного алгоритма с алгоритмами решения задачи о многоруком бандите UCB1, 0,4-жадный, 0,6-жадный, Softmax с наивной функцией выигрыша, а также алгоритмами $\text{UCB1}_{E(Q)}$, $\text{Softmax}_{E(Q)}$ с описанной выше функцией выигрыша. На итерацию выделялось по 30 с, а общий временной бюджет составил 3 ч (10 800 с). Каждый из алгоритмов запускался 12 раз со случайными начальными значениями генератора псевдослучайных чисел, которые требует библиотека. Время работы библиотеки Auto-WEKA тоже ограничивалось 3 ч, а выбор производился из тех же алгоритмов классификации. Результаты приведены в табл. 3. Жирным выделены наименьшие значения в строке как оптимальные для задачи минимизации эмпирического риска.

Как видно из таблицы, предложенный метод оказался не хуже, а во многих случаях и существенно лучше библиотеки Auto-WEKA на всех десяти наборах данных, так как позволил достичь меньшей ошибки. Результаты для различных алгоритмов решения задачи о многоруком бандите не сильно отличаются, однако наименьшей ошибки достигли алгоритмы $\text{UCB1}_{E(Q)}$ и $\text{Softmax}_{E(Q)}$, использующие предложенную функцию выигрыша. Как показали эксперименты, предложенный метод позволяет улучшить существующее решение задачи совместного выбора алгоритма классификации и его структурных параметров, так как поиск производится на всем пространстве структурных параметров каждой модели алгоритма. Существенно, что за фиксированное время предложенный метод позволяет найти решения, по мере качества не уступающие конфигурациям, найденным с помощью Auto-WEKA.

Для того чтобы показать статистическую значимость того, что ариации предложенного метода достигли меньшей ошибки, был проведен тест Уилкоксона. Он применим для оценки полученных результатов, так как имеется 10 наборов данных, на каждом из ко-

торых запускались библиотека Auto-WEKA и описанные выше вариации предложенного алгоритма. Получили 6 проверок критерия: сравнение библиотеки Auto-WEKA с каждым вариантом. При $n = 10$ выборках значимые результаты получаются при сумме нетипичных рангов $T < T_{0,01} = 5$. Так как в данном случае решается задача минимизации, проверили критерий для лучшего из 12 запусков на каждом наборе данных. Для ε-жадных алгоритмов получили $T = 3$, для остальных — $T = 1$, что показывает статистическую значимость полученных результатов.

5 Заключение

В данной работе был предложен и исследован метод решения актуальной задачи совместного выбора алгоритма классификации и его структурных параметров, основанный на сведении задачи к задаче о многоруком бандите. Кроме того, предложена функция выигрыша, позволяющая лучше адаптировать решение задачи о многоруком бандите для применения к решению данной задачи. Как показали эксперименты, предложенная стратегия позволяет улучшить существующее решение задачи совместного выбора алгоритма классификации и его структурных параметров, так как поиск производится на всем пространстве структурных параметров каждой модели алгоритма.

Метод можно улучшить, изначально отсортировав процессы настройки структурных параметров с помощью мета-обучения. Его также можно улучшить, если с помощью мета-обучения подбирать в пару алгоритму классификации алгоритм настройки структурных параметров. Кроме того, можно ввести контекст процесса настройки структурных параметров и воспользоваться алгоритмом для решения задачи о многоруком контекстном бандите. В качестве контекста можно использовать значение эмпирического риска, полученного при запуске алгоритма классификации на наборах данных, отобранных с помощью мета-обучения.

Литература

[1] *Rendell L., Cho H.* Empirical learning as a function of concept character // Mach. Learn., 1990. Vol. 5. P. 267–298.

[2] *Aha D. W.* Generalizing from case studies: A case study // 9th Conference (International) on Machine Learning Proceedings, 1992. P. 1–10. doi: 10.1016/B978-1-55860-247-2.50006-1.

[3] *Rodrigues J. D., Perez A., Lozano J. A.* Sensitivity analysis of k-fold cross validation in prediction error estimate // IEEE Trans. Pattern Anal., 2010. Vol. 32. No. 3. P. 569–575.

[4] *Giraud-Carrier C., Vilalta R., Brazdil P.* Introduction to the special issue on meta-learning // Mach. Learn., 2004. Vol. 54. No. 3. P. 187–193. doi: 10.1023/B:MACH.0000015878.60765.42.

[5] *Filchenkov A., Pendryak A.* Datasets meta-feature description for recommending feature selection algorithm // Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference Proceedings. — IEEE, 2015. P. 11–18. doi: 10.1109/AINL-ISMW-FRUCT.2015.7382962.

[6] *Lim T.-S., Loh W.-Y., Shih Y.-S.* A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms // Mach. Learn., 2000. Vol. 40. No. 3. P. 203–228. doi: 10.1023/A:1007608224229.

[7] *Ali S., Smith K. A.* On learning algorithm selection for classification // Appl. Soft Comput., 2006. Vol. 6. No. 2. P. 119–138. doi: 10.1016/j.asoc.2004.12.002.

[8] *Зайцев А. А., Стрижов В. В., Токмакова А. А.* Оценка гиперпараметров линейных регрессионных моделей методом максимального правдоподобия // Информационные технологии, 2013. Вып. 2. С. 11–15.

[9] *Bergstra J., Bengio Y.* Random search for hyper-parameter optimization // J. Mach. Learn. Res., 2012. Vol. 13. No. 1. P. 281–305.

[10] *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning: Data mining, inference and prediction // Math. Intell., 2005. Vol. 27. No. 2. P. 83–85. doi: 10.1007/978-0-387-84858-7.

[11] *Bottou L.* Online learning and stochastic approximations // On-Line Learning Neural Networks, 1998. Vol. 17. No. 9. P. 142–177.

[12] *Bergstra J., Bardenet R., Bengio Y.* Algorithms for hyper-parameter optimization // Advances in Neural Information Processing Systems Proceedings, 2011. P. 2546–2554.

[13] *Snoek J., Larochelle H., Adams R. P.* Practical Bayesian optimization of machine learning algorithms // Advances in Neural Information Processing Systems Proceedings, 2012. P. 2951–2959.

[14] *Hutter F., Hoos H. H., Leyton-Brown K.* Sequential model-based optimization for general algorithm configuration. — University of British Columbia, Computer Science, 2010. TR-2010-10.

[15] *Thornton C., Hutter F., Hoos H.* Auto-WEKA: Automated selection and hyper-parameter optimization of classification algorithms // 19th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2013. doi: 10.1145/2487575.2487629.

[16] *Leite R., Brazdil P., Vanschoren J.* Selecting classification algorithms with active testing // Machine learning and data mining in pattern recognition. — Springer, 2012. P. 117–131.

[17] *Воронцов К. В.* Математические методы обучения по прецедентам (теория обучения машин). `http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf`.

[18] *Sutton R. S., Barto A. G.* Reinforcement learning: An introduction. — Cambridge, MA, USA: MIT Press, 1998. 342 p.

Поступила в редакцию 18.08.2016

Reinforcement-based simultaneous classification model and its hyperparameters selection*

V. A. Efimova, A. A. Filchenkov, and A. A. Shalyto

`efimova@rain.ifmo.ru; afilchenkov@corp.ifmo.ru; shalyto@mail.ifmo.ru`

ITMO University, 49 Kronverksky pr., St. Petersburg, Russia

Many algorithms for data analysis exist, especially for the classification problem. To solve a data analysis problem, a proper algorithm should be chosen, and also, its hyperparameters should be selected. These two problems, algorithm selection and hyperparameter optimization, are commonly solved independently. The full-model selection process requires unacceptable time budgets. Thus, this is one of the factors preventing the spread of automated model selection methods. The goal of this work is to suggest a method for simultaneous algorithm and its parameters selection to reduce full-model election time. In order to do so, this problem was reduced to a multiarmed bandit problem. An algorithm is presented as an arm and algorithm of hyperparameters search during a fixed time is presented as the corresponding arm play. Also, several reward functions are described. The experiments have been held on 10 popular labeled datasets from the UCI repository. To compare the proposed method, 10 several well-known classification algorithms from WEKA library and algorithm for hyperparameter optimization

*The research was supported by the Russian Government (grant 074-U01) and the Russian Foundation for Basic Research (project No. 16-37-60115).

from Auto-WEKA library have been used. The proposed method has been compared with the brute force search implemented in WEKA library and a random time budget assignment policy. The results show significant time reduction of selecting proper algorithm and its hyperparameters for processing given dataset. The proposed method often produces classification results much better than Auto-WEKA state-of-the-art automatic algorithm selection and hyperparameter optimization tool.

Keywords: *algorithm selection; hyperparameter optimization; multiarmed bandit; reinforcement learning*

DOI: 10.21469/22233792.2.2.09

References

[1] Rendell, L., and H. Cho. 1990. Empirical learning as a function of concept character. *Mach. Learn.* 5:267–298.

[2] Aha, D. W. 1992. Generalizing from case studies: A case study. *9th Conference (International) on Machine Learning Proceedings*. 1–10. doi: 10.1016/B978-1-55860-247-2.50006-1.

[3] Rodrigues, J. D., A. Perez, and J. A. Lozano. 2010. Sensitivity analysis of k-fold cross validation in prediction error estimate. *IEEE Trans. Pattern Anal.* 32(3):569–575.

[4] Giraud-Carrier, C., R. Vilalta, and P. Brazdil. 2004. Introduction to the special issue on meta-learning. *Mach. Learn.* 54(3):187–193. doi: 10.1023/B:MACH.0000015878.60765.42.

[5] Filchenkov, A., and A. Pendryak. 2015. Datasets meta-feature description for recommending feature selection algorithm. *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference Proceedings*. IEEE. 11–18. doi: 10.1109/AINL-ISMW-FRUCT.2015.7382962.

[6] Lim, T.-S., W.-Y. Loh, and Y.-S. Shih. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.* 40(3):203–228. doi: 10.1023/A:1007608224229.

[7] Ali, S., and K. A. Smith. 2006. On learning algorithm selection for classification. *Appl. Soft Comput.* 6(2):119–138. doi: 10.1016/j.asoc.2004.12.002.

[8] Zaytsev, A. A., V. V. Strijov, and A. A. Tokmakova. 2013. Estimation regression model hydroparameters using maximum likelihood. *Informatsionnye Tekhnologii* [Information Technologies] 2:11–15.

[9] Bergstra, J., and Y. Bengio. 2012. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13(1):281–305.

[10] Hastie, T., R. Tibshirani, and J. Friedman. 2005. The elements of statistical learning: data mining, inference and prediction. *Math. Intell.* 27(2):83–85. doi: 10.1007/978-0-387-84858-7.

[11] Bottou, L. 1998. Online learning and stochastic approximations. *On-Line Learning Neural Networks* 17(9):142–177.

[12] Bergstra, J., R. Bardenet, and Y. Bengio. 2011. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems Proceedings*. 2546–2554.

[13] Snoek, J., H. Larochelle, and R. P. Adams. 2012. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems Proceedings*. 2951–2959.

[14] Hutter, F., H. H. Hoos, and K. Leyton-Brown. 2010. *Sequential model-based optimization for general algorithm configuration*. University of British Columbia, Computer Science. TR-2010-10.

[15] Thornton, C., F. Hutter, and H. Hoos. 2013. Auto-WEKA: Automated selection and hyperparameter optimization of classification algorithms. *19th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings*. doi: 10.1145/2487575.2487629.

[16] Leite, R., P. Brazdil, and J. Vanschoren. 2012. Selecting classification algorithms with active testing. *Machine learning and data mining pattern recognition*. Springer. 117–131.

[17] Vorontsov, K. V. *Matematicheskie metody obucheniya po pretsedentam (teoriya obucheniya mashin)* [Mathematical methods of training on precedents (machine learning theory)]. `http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf` (accessed November 22, 2016).

[18] Sutton, R. S., and A. G. Barto. 1998. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press. 342 p.

Received August 18, 2016