КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2020 Т. 12 № 4 С. 921–933
DOI: 10.20537/2076-7633-2020-12-4-921-933

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 519.25

Сравнительный анализ статистических методов классификации научных публикаций в области медицины

Г. В. Данилов[1], В. В. Жуков[2], А. С. Куликов[1], Е. С. Макашова[1], Н. А. Митин[3,a], Ю. Н. Орлов[2,3]

[1] ФГАУ НМИЦ нейрохирургии им. ак. Н. Н. Бурденко,
Россия, 125047, г. Москва, 4-я Тверская-Ямская ул., д. 16

[2] Российский университет дружбы народов,
Россия, 117198, г. Москва, ул. Миклухо-Маклая, д. 6

[3] ФИЦ Институт прикладной математики им. М. В. Келдыша РАН,
Россия, 125047, г. Москва, Миусская пл., д. 4

E-mail: [a] mitin@keldysh.ru

Получено 25.03.2020, после доработки — 16.04.2020.
Принято к публикации 06.05.2020.

В работе проведено сравнение различных методов машинной классификации научных текстов по тематическим разделам на примере публикаций в профильных медицинских журналах, выпускаемых издательством Springer. Исследовался корпус текстов по пяти разделам: фармакология/токсикология, кардиология, иммунология, неврология и онкология. Рассматривались как методы поверхностной классификации, основанные на анализе аннотаций и ключевых слов, так и методы классификации на основе обработки собственно текстов. Были применены методы байесовской классификации, опорных векторов и эталонных буквосочетаний. Показано, что наилучшую точность имеет метод классификации на основе создания библиотеки эталонов буквенных триграмм, отвечающих текстам определенной тематики, а семантические методы уступают ему по точности. Выяснилось, что применительно к рассматриваемому корпусу текстов байесовский метод дает ошибку порядка 20 %, метод опорных векторов имеет ошибку порядка 10 %, а метод близости распределения текста к трехбуквенному эталону тематики дает ошибку порядка 5 %, что позволяет ранжировать эти методы для использования искусственного интеллекта в задачах классификации текстов по отраслевым специальностям. Существенно, что при анализе аннотаций метод опорных векторов дает такую же точность, что и при анализе полных текстов, что важно для сокращения числа операций для больших корпусов текстов.

Ключевые слова: машинное обучение, классификация медицинских текстов, статистический анализ

Работа выполнена при финансовой поддержке гранта РФФИ № 19-29-01174.

COMPUTER RESEARCH AND MODELING
2020 VOL. 12 NO. 4 P. 921–933
DOI: 10.20537/2076-7633-2020-12-4-921-933

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 519.25

Comparative analysis of statistical methods of scientific publications classification in medicine

G. V. Danilov[1], V. V. Zhukov[2], A. S. Kulikov[1], E. S. Makashova[1], N. A. Mitin[3,a], Yu. N. Orlov[2,3]

[1] Burdenko Neurosurgical Center,
16 4th Tverskaya-Yamskaya st., Moscow, 125047, Russia

[2] Peoples' Friendship University of Russia,
6 Miklukho-Maklaya st., Moscow, 117198, Russia

[3] Keldysh Institute of Applied Mathematics Russian Academy of Sciences,
4 Miusskaya square, Moscow, 125047, Russia

E-mail: [a] mitin@keldysh.ru

Received 25.03.2020, after completion — 16.04.2020.
Accepted for publication 06.05.2020.

In this paper the various methods of machine classification of scientific texts by thematic sections on the example of publications in specialized medical journals published by Springer are compared. The corpus of texts was studied in five sections: pharmacology/toxicology, cardiology, immunology, neurology and oncology. We considered both classification methods based on the analysis of annotations and keywords, and classification methods based on the processing of actual texts. Methods of Bayesian classification, reference vectors, and reference letter combinations were applied. It is shown that the method of classification with the best accuracy is based on creating a library of standards of letter trigrams that correspond to texts of a certain subject. It is turned out that for this corpus the Bayesian method gives an error of about 20%, the support vector machine has error of order 10%, and the proximity of the distribution of three-letter text to the standard theme gives an error of about 5%, which allows to rank these methods to the use of artificial intelligence in the task of text classification by industry specialties. It is important that the support vector method provides the same accuracy when analyzing annotations as when analyzing full texts, which is important for reducing the number of operations for large text corpus.

Keywords: machine learning, medicine texts classification, statistical analysis

Citation: *Computer Research and Modeling*, 2020, vol. 12, no. 4, pp. 921–933 (Russian).

The work was supported by the RFBR grant No. 19-29-01174.

1. Введение

В настоящее время существует острая необходимость в качественной машинной классификации научно-технических и иных специализированных текстов по определенной тематике исследований. Это связано как с растущим объемом научной печатной продукции, так и с анализом современных тенденций развития науки на основе различных наукометрических показателей. Особую важность в этом плане приобретает автоматическая классификация текстов в области медицины. Поскольку сегодня в мире поток научных публикаций по медицинской тематике составляет порядка 500 тыс. в год, проанализировать их вручную не представляется возможным. Тем не менее необходимо проводить хотя бы наукометрический анализ для целей стратегического планирования, существенной частью которого является формализация положения определенной медицинской отрасли в наукометрическом пространстве в виде классификационного перечня производства разных видов научной продукции и публикационной активности. Количественный результат такого анализа позволит обоснованно формализовать прогнозирование и поставить цель развития отрасли на некоторый горизонт. Анализируя публикации, можно понять, какие темы в какие периоды времени оказываются значимыми, какие направления поддерживаются финансирующими фондами, в каких журналах какие работы имеют в текущий момент большую вероятность публикации и т. д. Основная по объему часть соответствующей аналитической работы имеет рутинный характер и связана со сбором и классификацией информации. Разработка системы, помогающей решению таких аналитических задач посредством автоматизации механической части, является актуальной.

Заметим здесь, что у каждой научной статьи существуют классификационные коды, относящие ее к определенной тематике. Эти коды выбраны автором публикации. Они также позволяют провести тематическую классификацию текста. Однако интересующая нас проблема состоит в исследовании возможности машинной классификации статей на основе именно текстов. Предполагается, что журналы, публикующие статьи по определенной области науки, являются правильными фактическими классификаторами, поскольку члены редакционной коллегии представляют мнение экспертного сообщества в соответствующей области.

Таким образом, вопрос, изучаемый в настоящей работе, следующий: можно ли определить тематику журнала, в котором опубликована (или могла бы быть опубликована) данная статья, на основе машинной классификации текстов, напечатанных в этом журнале?

Одной из центральных проблем при разработке системы автоматической классификации текстов в соответствии с областью деятельности является выбор метода классификации. Несмотря на большое разнообразие статистических методов классификации, ни один из них не является универсальным. Подробный обзор основных из них содержится, например, в [Батура, 2017], где приведены также и примеры их использования в практических задачах. Однако следует заметить, что вычислительных экспериментов, доказывающих сравнительную эффективность разных методов, проведено еще очень мало с точки зрения анализа зависимости точности метода от размера и типа корпуса текстов. Результатом анализа обычно является одно число – точность классификации, полученная по результатам численного эксперимента применительно к определенному корпусу текстов определенным методом. Интерес же представляет многомерный объект: точность в зависимости от объемов обучающей и тестируемой выборок, количества классов, тематики корпусов текстов. В нашей работе предлагается сравнить в указанном смысле эффективность разных методов классификации медицинских научных публикаций в тематических журналах на английском языке.

Статистическая классификация имеет целью создание решающего правила, которое ставит в соответствие набору текстовых документов один или, возможно, несколько категорий (классов). Такое соответствие устанавливается на основе близости анализируемых элементов документа элементам класса. Близость формулируется в терминах расстояния в пространстве эталонов, составленных из элементов текстов обучающей выборки. Решающим правилом является обычно «ближайший эталон», т. е. документ считается принадлежащим тому классу, к эталону которого он ближе. Эталоны формируются после определенной предобработки текста и пред-

ставляют собой эмпирические частоты употребления элементов. Такими элементами могут выступать буквы и буквосочетания (метод *n*-грамм), слова в нормальной форме, словосочетания или связанные группы слов, иные символы, маркирующие документ.

Применительно к медицинской тематике методы машинной классификации востребованы во многих направлениях, которые являются основными источниками больших данных. Например, в области нейрохирургии такими направлениями работ являются нейровизуализация, секвенирование генома, инвазивные и неинвазивные биосенсоры, данные медицинских информационных систем. При этом анализу научной литературы с помощью методов искусственного интеллекта (ИИ) как в нейрохирургии, так и в иных областях медицины уделено существенно меньше внимания, чем другим задачам. Настоящая работа восполняет определенный пробел в этом направлении.

Современные технологии анализа текстов позволяют исследовать содержание медицинских записей, выделяя значимую информацию для формулирования и проверки научных гипотез. Технологии ИИ позволяют как извлекать данные из медицинских текстов, так и строить диагностические и прогностические модели на этих данных [Yoo, Song, 2008]. Источником данных для подобного рода анализа может стать архив электронных медицинских записей за большой период времени, которые в настоящий момент используются в ретроспективных исследованиях, но являются латентными информационными недрами медицины [Middleton et al., 2016]. Примером решения исследовательских задач с помощью технологий анализа естественного языка в нейрохирургии являются работы по прогнозированию внутрибольничных инфекций после нейрохирургических вмешательств [Campillo-Gimenez et al., 2013; Tvardik et al., 2018]. В работе [Cohen et al., 2016] было показано, что средства обработки естественного языка и методы машинного обучения могут способствовать улучшению отбора кандидатов для хирургического лечения эпилепсии и сокращению времени определения показаний к этому лечению. Тем не менее исследования, обобщающие факторы риска развития нейрохирургической патологии и прогноза вариантов ее течения с помощью методов анализа медицинских текстов, в настоящий момент являются единичными.

Важным источником больших данных для исследовательских задач являются электронные медицинские библиотеки. В настоящий момент обобщение и критическая оценка результатов исследований в соответствии с канонами доказательной медицины проводятся экспертами с помощью систематических обзоров и мета-анализов [Liu et al., 2018]. Такие обобщающие исследования, как правило, требуют больших трудозатрат и временного ресурса, дают максимально обоснованный ответ, однако на очень узкий круг вопросов. В то же время в научной литературе менее «высокого качества» с позиций доказательной медицины содержится большое количество информации, которая не подвергается критической оценке традиционными методами доказательной медицины. В хирургических областях медицины, включая нейрохирургию, количество исследований с высоким уровнем достоверности доказательств оказывается существенно меньшим, чем в терапевтических [Liu et al., 2018; Mansouri et al., 2016]. Именно поэтому аналитические инструменты, способные преодолевать ограничения систематических обзоров и наиболее полно использовать информацию из тысяч публикаций, в нейрохирургии приобретают особую значимость. В последние годы появились работы, в которых научная литература по нейрохирургии анализируется и обобщается с помощью ИИ [Buchlak et al., 2019; Hana et al., 2019].

В настоящей работе мы рассматриваем задачу классификации текстов для нескольких медицинских тематик. Рассматривается корпус текстов, кластеризованный по научным медицинским журналам, выпускаемым издательством Springer. Корпус состоит из 1440 научных статей в 18 журналах различных тематик по 80 статей в каждом, разделенных по двум дисциплинам (биомедицина и здравоохранение) и совокупно по пяти направлениям. Биомедицина в этом корпусе представлена направлениями «иммунология» (2 журнала), «неврология» (7 журналов) и «фармакология/токсикология» (3), здравоохранение — направлениями «кардиология» (3) и «онкология» (3). Для каждой тематики, а затем и направления проведено сравнение трех различных методов машинной классификации научных текстов. Классификатор должен был пра-

вильно указать журнал, в котором была опубликована та или иная статья. Поскольку корпус создавался как совокупность статей в определенных журналах, то классификационная принадлежность каждой статьи была достоверно известна изначально. Ошибкой классификатора в нашей задаче является доля статей, для которых неверно были определены журналы, в которых эти статьи были напечатаны.

2. «Наивный» Байес

Один из классических подходов к решению задачи классификации текстов состоит в применении «наивного» байесовского классификатора [Manning et al., 2008]. Данный подход достаточно прост в реализации и демонстрирует сравнительно высокую эффективность при работе со словарями и текстами небольшого размера. Набор ключевых слов статьи и ее аннотация как раз являются примерами таких небольших текстов.

Пусть имеется некоторое множество документов (текстов) D, которое надо распределить по некоторому множеству A классов. В байесовском методе каждый текст представляет собой набор слов, появление которых считается полностью независимым друг от друга (наивное предположение классификатора). В общем случае это, очевидно, не так, однако данное предположение сильно упрощает математическую модель классификатора, сохраняя приемлемую точность. Пусть $P(a|d)$ есть вероятность того, что документ d принадлежит классу a, $P(d|a)$ есть вероятность встретить документ d среди документов класса a, $p(a)$ — априорная вероятность встретить документ класса a во всем корпусе текстов, $q(d)$ — априорная вероятность встретить документ d во всем корпусе текстов. В основе метода лежит теорема Байеса:

$$P(a|d) = \frac{P(d|a)\,p(a)}{q(d)},\ a \in A,\ d \in D. \tag{1}$$

Пусть имеется документ «0» неизвестного класса. Для него вычисляются величины, входящие в правую часть формулы (1) для каждого класса. Если корпус текстов не варьируется, то $q(d)$ есть константа, которую можно далее опустить. Априорная вероятность класса оценивается частотой встречаемости документов в виде $p(a) = K(a)/K$, где $K(a)$ есть количество документов класса a, K — объем всего корпуса. Классификатор определяем по методу наибольшей вероятности, т. е. в виде

$$a^0 = \arg\max_{a \in A} P(0|a)\,p(a). \tag{2}$$

Предположение об условной независимости термов $\{t_i\}$ документа позволяет переписать вероятность $P(d|a)$ в виде

$$P(d|a) = \prod_i P(t_i|a). \tag{3}$$

Такая модель называется полиномиальным байесовским классификатором. В ней каждый документ представляется случайным полиномиальным вектором, компонентом которого является эмпирическая частота использования отдельного слова [Manning et al., 2008]. Данный метод является более эффективным по сравнению с классификатором, в основе которого лежит многомерная модель испытаний Бернулли [McCallum, Kamal, 1998].

Предобработка данных для классификатора состоит из следующих шагов:

1) токенизация — разбиение набора ключевых слов (аннотации) на отдельные слова, приведение слов к нижнему регистру;

2) очистка — удаление пунктуационных символов и шумовых слов (например, and, of, on и т. п.), наличие которых предполагается во всех документах корпуса;

3) стемминг — преобразование слова к его основе за счет удаления приставок, суффиксов и окончаний. При выполнении данного шага использовался стеммер Портера [Porter, 1980], реализованный в модуле Python NLTK (Natural Language Tool Kit).

Каждой статье был поставлен в соответствие вектор, содержащий частоты, с которыми различные слова встречались в данной статье (в ключевых словах или аннотации). Количество наиболее значащих слов является ключевым параметром. Оно определяет размер признакового пространства и напрямую влияет на скорость обучения алгоритмов и их тенденции к переобучению. Программная реализация алгоритма была разработана на языке Python с использованием модуля Python sklearn.

На рис. 1 представлена зависимость ошибки классификации от объема используемого словаря (т. е. от количества слов, обладающих максимальной взаимной информацией, вычисленной по обучающей выборке) для аннотаций и ключевых слов. Как видно, классификация на основе аннотаций дает более точные результаты, чем на основе ключевых слов, однако последнюю также можно считать достаточно высокой, учитывая, что классификация проводится по пяти тематикам.

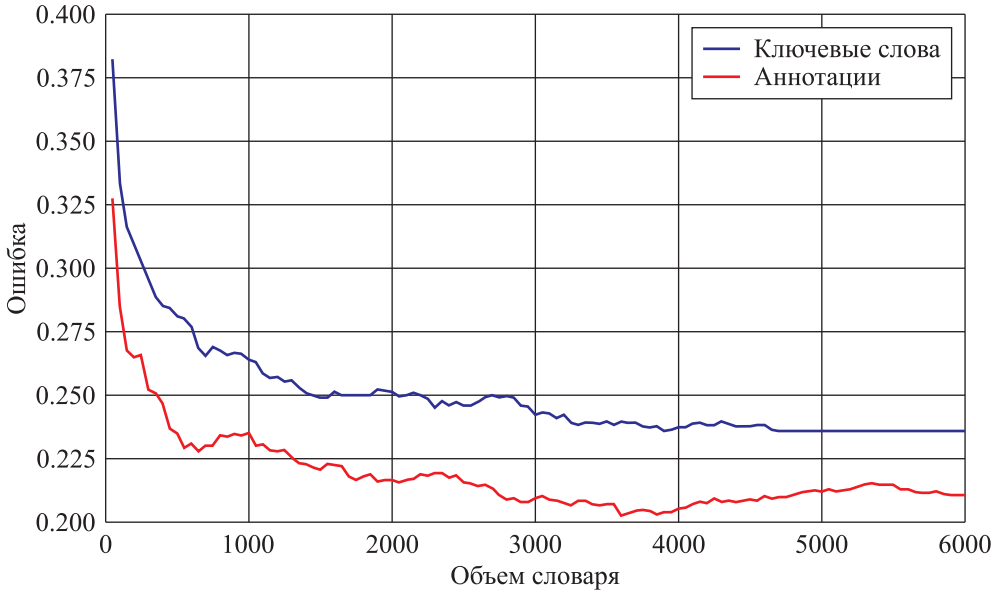

Рис. 1. Ошибка классификации тематики статьи в зависимости от объема словаря, байесовский классификатор

Следует отметить зависимость эффективности метода от объема словаря. Начиная с определенного порога, когда в модель включается почти все значимые слова, эффективность меняется не сильно (и может даже снижаться, что видно из графика по аннотациям). Эмпирически были определены оптимальные значения объема словаря для каждого вида исходных данных. Для ключевых слов это значение составило 3900, для аннотаций — 3600, для полных текстов — 5000 слов.

Результаты сравнения эффективности всех трех видов исходных данных в зависимости от объема обучающей выборки представлены на рис. 2. Если для аннотаций и ключевых слов характерна устойчивая зависимость от объема обучающей выборки, то классификация по полным текстам показывает высокие результаты уже при объеме обучающей выборки в 25 % от общего корпуса. Этот эффект можно объяснить тем, что в 25 % полных текстах корпуса уже содержатся практически все значимые слова, причем значения условных вероятностей принимают достаточно точные значения, характерные для всего корпуса в целом. Дальнейшее увеличение объема обучающей выборки позволяет уточнить значения условных вероятностей, что незначительно повышает эффективность метода.

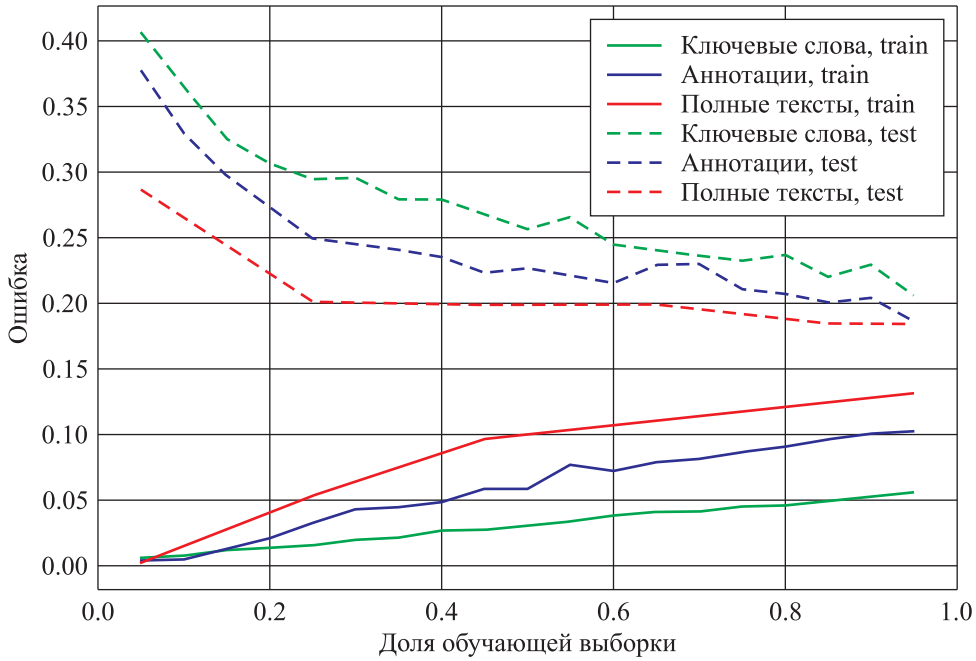

Рис. 2. Ошибка классификации тематики статьи в зависимости от доли обучающей выборки, байесовский классификатор

Следует отметить, что для байесовской модели точность классификации обучающей выборки существенно выше точности классификации тестовых данных. По мере увеличения доли обучающей выборки эта разница уменьшается, а точность предсказаний на тестовой выборке увеличивается. Для эффективной работы метода существенными аспектами являются однородность данных и корректные значения условных вероятностей всех значимых слов. Заметим также, что при должном объеме обучающей выборки (в данном случае ~70 % и выше) классификация на основе аннотаций обладает почти той же точностью, что и классификация на основе полных текстов, но при этом она значительно экономичнее по времени работы программы.

Одной из особенностей байесовского классификатора является то, что после процесса обучения алгоритма можно выделить слова, наиболее и наименее характерные для определенных тематик. В таблице 1 представлены 10 наиболее значимых слов, определенных для классификации между направлениями «кардиология» и «онкология». Так, например, встретить слово cardiolog в кардиологическом журнале в 33.4 раза вероятнее, чем в журнале, посвященном онкологии.

В целом можно сделать вывод, что байесовский метод дает ошибку порядка 20 % при классификации как полных текстов, так и аннотаций.

Таблица 1. Отношение условных вероятностей наиболее значимых слов при анализе полных текстов

Слово	Классы			Отношение условных вероятностей		
cardiolog	Кардиология	:	Онкология	33.4	:	1
cerebrovascular	Кардиология	:	Онкология	26.9	:	1
ventricl	Кардиология	:	Онкология	25.5	:	1
atherosclerosi	Кардиология	:	Онкология	25.5	:	1
atherosclerot	Кардиология	:	Онкология	24.8	:	1
metastasi	Онкология	:	Кардиология	20.2	:	1
lymph	Онкология	:	Кардиология	18.8	:	1
reproduct	Онкология	:	Кардиология	17.7	:	1
melanoma	Онкология	:	Кардиология	15.8	:	1
progesteron	Онкология	:	Кардиология	15.8	:	1

3. Опорные векторы (SVM)

Модель, основанная на применении машины опорных векторов [Вапник, Червоненкис, 1974], дает более точные результаты, чем «наивный» байесовский классификатор. На рис. 3 представлена зависимость ошибки классификации от количества компонент, выделенных после сингулярного разложения матрицы термы-документы. Сокращение размерности до 100 компонент становится достаточным для достижения практически той же точности, что и при выделении 250 или 500 компонент. Дальнейшее увеличение не дает существенного прироста эффективности (вообще говоря, применение сингулярного разложения для сокращения размерности теряет смысл при приближении количества выделяемых компонент к исходному количеству характеристик), увеличивая время, необходимое на обучение модели.

На рис. 4 представлена зависимость ошибки классификации от объема обучающей выборки для журналов, принадлежащих одному направлению из двух (например, «неврология» и «онкология»). Решение данной задачи отвечает на вопрос о том, можно ли терминологически разделить журналы даже в рамках одной дисциплины. Такое разделение позволит, например, авторам научных статей выбирать журналы для публикаций, которые в большой степени соответствуют их работе.

Выбор между двумя журналами решается классификатором достаточно хорошо, доводя точность до 92 %. Выбор между пятью вариантами представляется более сложной задачей. Здесь, во-первых, важную роль играет объем обучающей выборки. Во-вторых, даже при большом объеме точность классификации не превосходит 80 %.

Выбор между журналами, относящимися к разным направлениям, осуществить проще, поскольку терминологические и смысловые различия в текстах статей таких журналов достаточно велики (см. рис. 5).

Анализ ошибок для тестовых выборок позволяет сделать некоторые выводы об особенностях данного корпуса статей. Так, статьи, посвященные кардиологии или неврологии, выделяются достаточно хорошо. А отличия между статьями, посвященными онкологии или иммунологии, не всегда улавливаются моделью. Это объясняется важностью иммунологического аспекта, которому уделяется большое внимание в исследованиях рака. Такие статьи используют термины из обеих областей, что усложняет классификацию, в основе которой лежит именно поиск различий в используемых понятиях.

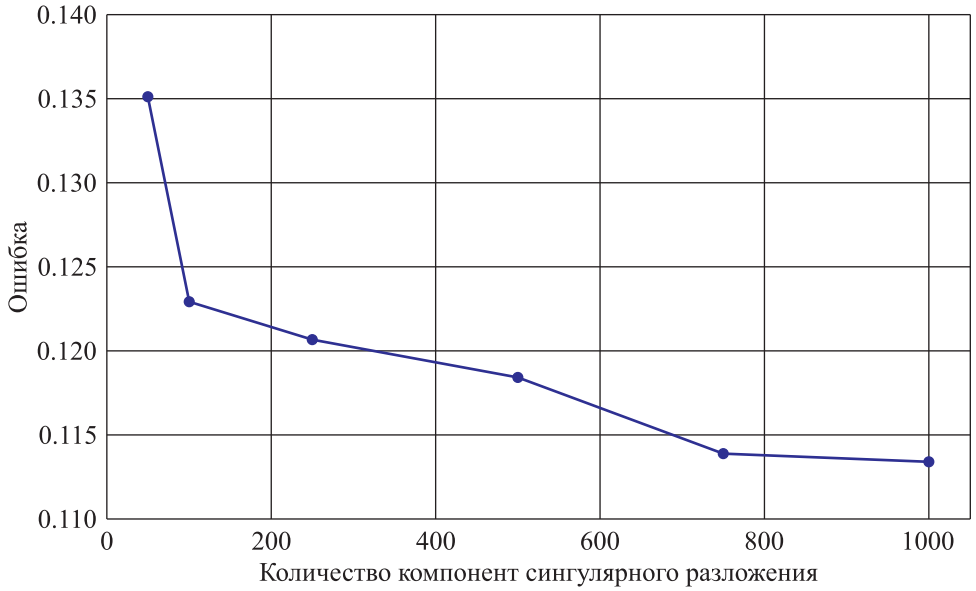

Рис. 3. Ошибка классификации тематики статьи в зависимости от количества компонент сингулярного разложения, машина опорных векторов

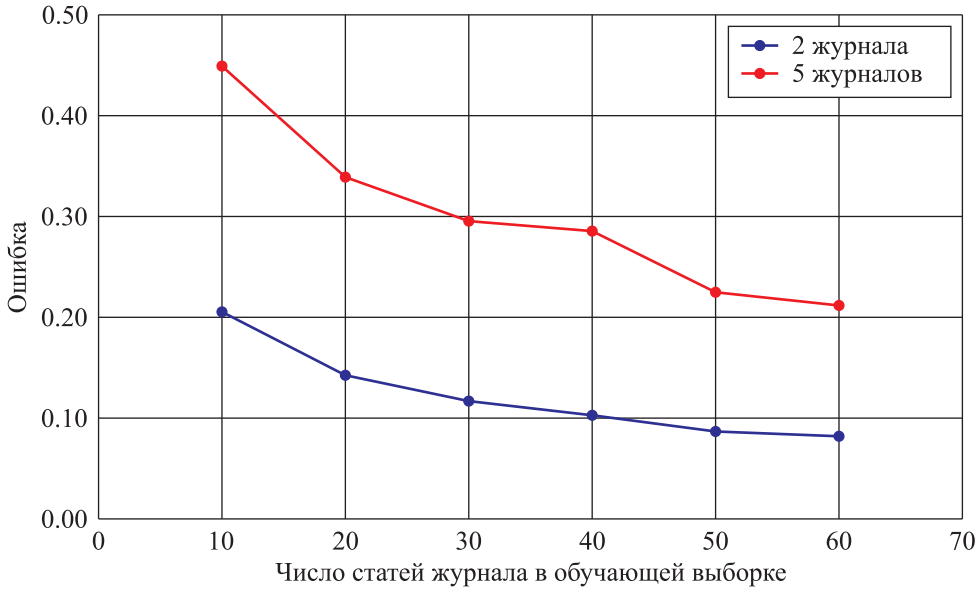

Рис. 4. Ошибка классификации статей относительно журналов в зависимости от объема обучающей выборки в рамках одного направления, машина опорных векторов

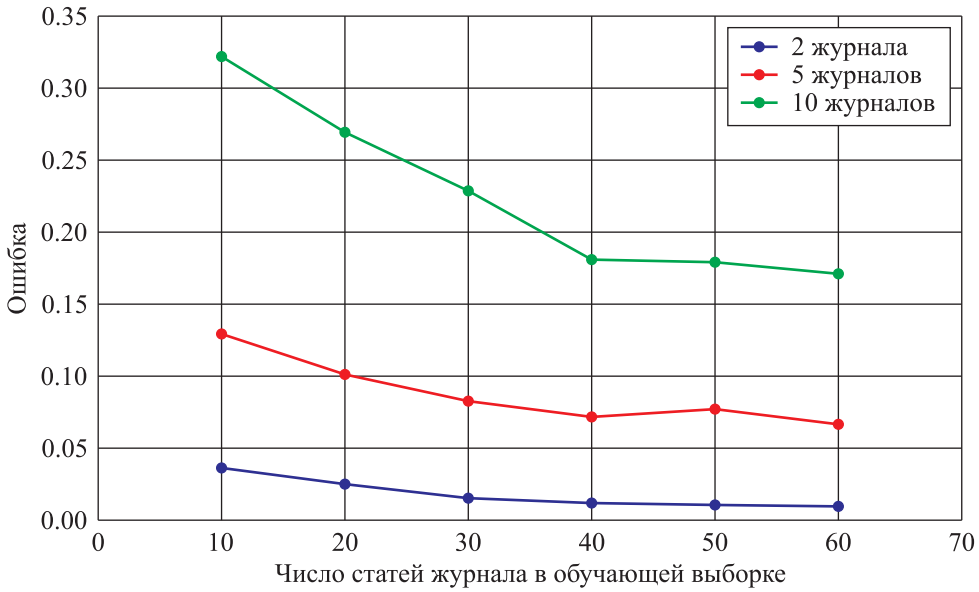

Рис. 5. Ошибка классификации статей в зависимости от объема обучающей выборки, разные направления, машина опорных векторов

4. Распределения буквосочетаний (*n*-граммы)

Опишем теперь методику определения тематики текста на основе составления эталонного распределения буквосочетаний, характерного для каждого научного направления. Этот метод показал высокую точность как при определении автора литературного произведения [Орлов, Осминин, 2012], так и при классификации научных текстов по отраслевым специальностям [Борисов и др., 2017]. Однако можно предположить, что определенная достаточно узкая тематика научного направления больше зависит от сочетания набора понятий, чем от авторского стиля, и может быть классифицирована этим методом. В данной работе мы впервые применяем

анализ вероятностного распределения буквосочетаний для классификации научных медицинских текстов по отдельным тематикам.

Формализуем задачу идентификации темы некоторого неизвестного текста. Она состоит в следующем. Имеется библиотека, содержащая тексты, представленные в виде плотностей функций распределения (далее ПФР) для A заданных тем. В нашем случае это тематики, которые поддерживаются определенными профессиональными журналами. Будем обозначать 1-ПФР — частоту использования одной определенной буквы в текстах данной тематики, 2-ПФР — частоту встречаемости пар букв, 3-ПФР — частоту встречаемости троек. Пусть K_a — имеющееся количество текстов *a*-й темы и $N_{i,a}$ — количество букв в *i*-м тексте на эту тему, $i = 1, 2, ..., K_a$. При этом из текстов удалены все цифры, формулы, рисунки, знаки препинания и пробелы. Считаем, что оставшаяся длина каждого из текстов достаточна для проведения статистического анализа. По оценкам [Орлов, Осминин, 2012], длина текста должна быть не меньше 7 тыс. знаков (букв). Обозначим $f_{i,a}(j)$ ПФР соответствующего текста, где j есть обозначение номера буквы или буквосочетания при их алфавитном упорядочении. Каждой теме поставим в соответствие средневзвешенную ПФР, представляющую эталонное распределение:

$$\overline{f}_a(j) = \frac{1}{N_a}\sum_{i=1}^{K_a} f_{i,a}(j) N_{i,a}, \quad N_a = \sum_{i=1}^{K_a} N_{i,a}. \tag{4}$$

В (4) мы пренебрегли единицей по сравнению с N_a при подсчете пар букв в тексте для 2-ПФР (или двойкой при подсчете троек, если речь идет о 3-ПФР, и т. д.). Введем норму ρ_{ik} как расстояние между ПФР текстов *i* и *k* в некоторой норме. Например, в норме L_1

$$\rho_{ik} = \left\| f_i - f_k \right\|_{L1} = \sum_{j=1}^{n} \left| f_i(j) - f_k(j) \right|, \tag{5}$$

где *n* есть количество типов анализируемых символов.

Пусть имеется текст «0» неизвестной тематики, который надо идентифицировать внутри данной библиотеки. Соответствующим атрибутом текста «0» считается та из тем *a*, для которой норма $\rho_a^0 = \left\| f_0 - \overline{f}_a \right\|$ разности между ПФР $f_0(j)$ текста «0» и эталонной ПФР $\overline{f}_a(j)$ минимальна:

$$\rho_a^0 = \left\| f_0 - \overline{f}_a \right\|, \; a^0 = \arg\min_a \rho_a^0. \tag{6}$$

Описанный метод оказался весьма точным в определении неизвестной тематики научного текста и достаточно простым при программной реализации. Независимо от тематической направленности журналов как внутри одного направления (например, «онкология»), так и при идентификации направлений из разных тематик (например, «иммунология» в сравнении с «кардиологией») двухбуквенное и трехбуквенное распределения показали достаточно высокий уровень правильной идентификации.

На рис. 6 показана ошибка идентификации тематики путем сравнения текста с эталонным журнальным распределением 2-ПФР в зависимости от количества разных тем (т. е. журналов) и в зависимости от длины обучающей выборки (количество статей, указанное по оси абсцисс). На рис. 7 показана аналогичная ошибка для 3-ПФР. Расчеты проводились с использованием программного комплекса [Орлов, Осминин, 2017].

По сравнению с графиками рис. 3 и рис. 5 графики на рис. 6 и рис. 7 еще не достигли уровня стабилизации. Отсюда следует, что у метода *n*-грамм еще есть резервы, позволяющие повысить точность классификации. Отметим также, что дальнейшее увеличение обучающей выборки для 3-ПФР ожидаемо приведет к более заметному улучшению качества классификации, чем для метода 2-ПФР, так как правая часть графика на рис. 7 менее пологая, чем на рис. 6.

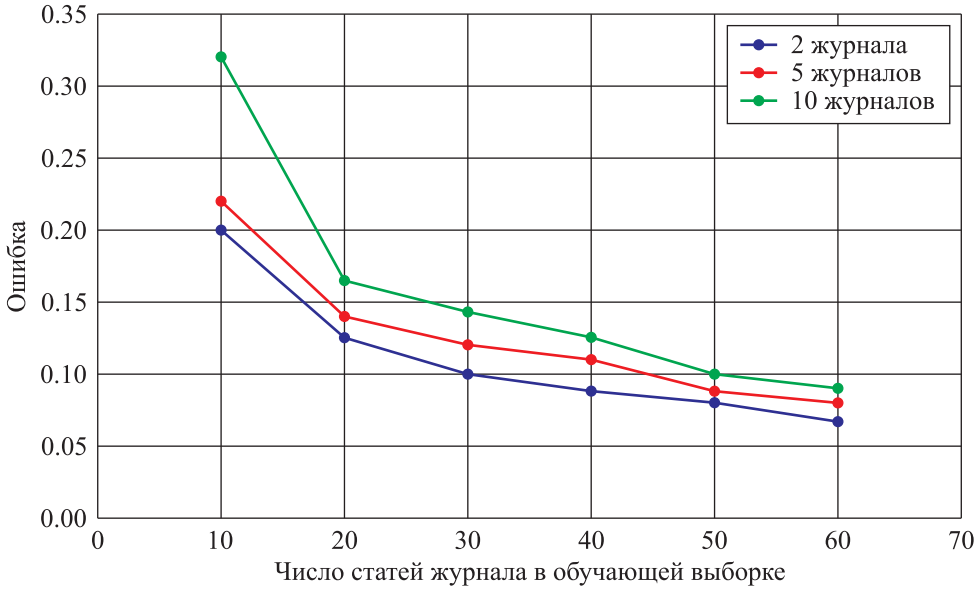

Рис. 6. Ошибка метода сравнения с эталоном 2-ПФР

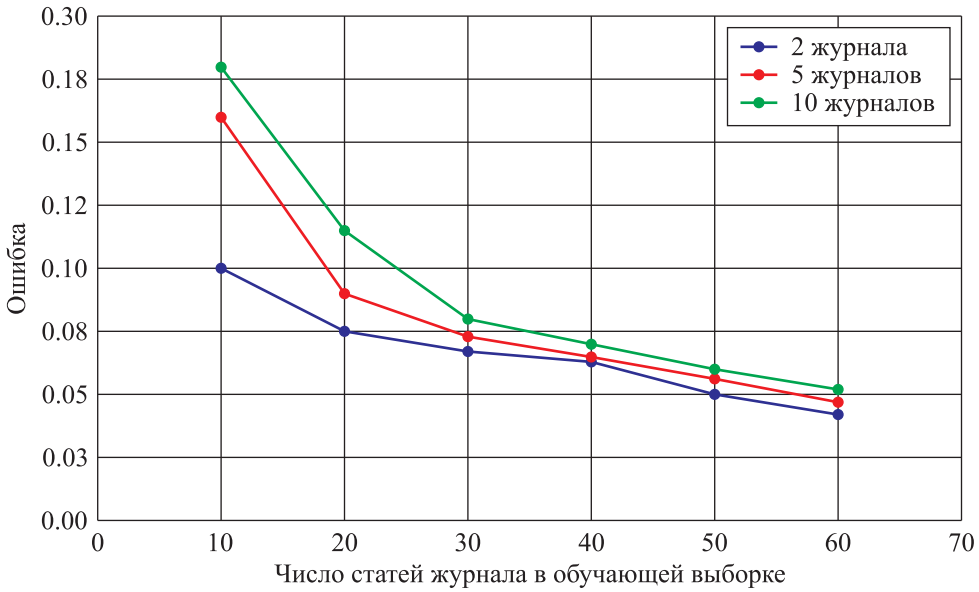

Рис. 7. Ошибка метода сравнения с эталоном 3-ПФР

5. Заключение

В работе проведено сравнение эффективности трех методов классификации медицинских текстов — байесовский, опорных векторов и по распределению n-грамм. Выяснилось, что применительно к данному конкретному корпусу текстов байесовский метод дает ошибку порядка 20 %, метод опорных векторов имеет ошибку порядка 10 %, а метод близости распределения текста к трехбуквенному эталону тематики дает ошибку порядка 5 %. Следовательно, методы SVM и 3-ПФР являются основными кандидатами для использования ИИ в задачах классификации. Существенно, что при анализе аннотаций метод 3-ПФР использовать не эффективно, поскольку для него нужны тексты достаточно больших длин, тогда как SVM может применяться почти без потери точности. В дальнейшем следует провести более детальный анализ этих методов для разных наборов научных тематик и размеров обучающей выборки.

Также следует отметить, что точность машинных методов классификации зависит от языка, на котором написаны тексты. По-видимому, семантические методы анализа проще применять к текстам на английском языке, чем, например, на русском, в силу специфики грамматики. Что касается метода частот буквосочетаний, то он представляется достаточно универсальным. Во всяком случае, применительно к распознаванию авторов текстов на русском языке [Орлов, Осминин, 2012] этот метод показал высокую точность. В то же время эффективность методов зависит и собственно от корпусов текстов. Поэтому исследование точности методов классификации в зависимости от языка профессионального общения (например, для журнальных текстов на русском языке и их переводных аналогов) также является важным планируемым этапом работ.

Список литературы (References)

Батура Т. В. Методы автоматической классификации текстов // Программные продукты и системы. — 2017. — Т. 30, № 1. — С. 85–99.
Batura T. V. Metody avtomaticheskoi klassifikatsii tekstov [Methods for automatic classification of texts] // Software products and systems [Programmnye produkty i sistemy]. — 2017. — Vol. 30, No. 1. — P. 85–99 (in Russian).

Борисов Л. А., Ивченко А. Ю., Митин Н. А., Орлов Ю. Н. Тематическая классификация текстов с помощью спектральных портретов // Препринты ИПМ им. М. В. Келдыша. — 2017. — № 106. — 22 с.
Borisov L. A., Ivchenko A. Yu., Mitin N. N., Orlov Yu. N. Tematicheskaya klassifikatsiya tekstov s pomoshch'yu spektral'nykh portretov [Classification of text information with the use of bigram analysis] // Preprinty IPM im. M.V. Keldysha [Keldysh Institute Preprints]. — 2017. — No. 106. — 22 p. (in Russian).

Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. — М.: Наука, 1974. — 416 с.
Vapnik V. N., Chervonenkis A. Ya. Teoriya raspoznavaniya obrazov [Theory of pattern recognition]. — Moscow: Nauka, 1974. — 416 p. (in Russian).

Орлов Ю. Н., Осминин К. П. Методы статистического анализа литературных текстов. — М.: Эдиториал УРСС / Книжный дом «ЛИБРОКОМ», 2012. — 312 с.
Orlov Yu. N., Osminin K. P. Metody statisticheskogo analiza literaturnykh tekstov [Methods of Statistical Analysis of literary texts]. — Moscow: Editorial URSS / Knizhnyi dom “LIBROKOM”, 2012. — 312 p. (in Russian).

Орлов Ю. Н., Осминин К. П. Программный комплекс TRIL для идентификации языка, автора и жанра литературного текста. Свидетельство о государственной регистрации № 2017611570 от 06.02.2017.
Orlov Yu. N., Osminin K. P. Svidetel'stvo o registratsii programmy dlya EVM “Programmnyi kompleks TRIL dlya identifikatsii yazyka, avtora i zhanra literaturnogo teksta”. Pravoobladatel': IPM im. M. V. Keldysha RAN. Svidetel'stvo o gosudarstvennoi registratsii No. 2017611570 ot 06.02.2017 [Certificate of registration of the computer program No. 2017611570 from 06.02.2017 “Software TRIL for identification of the language, author and genre of a literary text”] (in Russian).

Buchlak Q. D., Esmaili N., Leveque J.-C. et al. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review // Neurosurg Rev. — 2019. — DOI:10.1007/s10143-019-01163-8

Campillo-Gimenez B., Garcelon N., Jarno P., Chapplain J. M., Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France // Stud Health Technol Inform. — 2013. — Vol. 192. — P. 572–575.

Cohen K. B., Glass B., Greiner H. M. et al. Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning // Biomed Inform Insights. — 2016. — Vol. 8. — P. 11–18. — DOI:10.4137/BII.S38308

Hana T., Tanaka S., Nejo T. et al. Mining-Guided Machine Learning Analyses Revealed the Latest Trends in Neuro-Oncology // Cancers (Basel). — 2019. — Vol. 11 (2). — DOI:10.3390/cancers11020178

Liu W., Ni M., Jia W., Wan W., Tang J. Evidence-based medicine in neurosurgery: an academic publication view // Neurosurg Rev. — 2018. — Vol. 41 (1). — P. 55–65. — DOI:10.1007/s10143-016-0742-7

Manning C. D., Raghavan P., Schuetze H. Introduction to Information Retrieval. — Cambridge University Press, 2008. — P. 234–265.

Mansouri A., Cooper B., Shin S. M., Kondziolka D. Randomized controlled trials and neurosurgery: the ideal fit or should alternative methodologies be considered? // J Neurosurg. — 2016. — Vol. 124 (2). — P. 558–568. — DOI:10.3171/2014.12.JNS142465

Middleton B., Sittig D. F., Wright A. Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision // Yearb Med Inform. — 2016. — Suppl 1:S103-16. — DOI:10.15265/IYS-2016-s034

McCallum A., Kamal N. A comparison of event models for naive bayes text classification // AAAI-98 workshop on learning for text categorization. — 1998. — Vol. 752, No. 1. — P. 41–48.

Porter M. F. An algorithm for suffix stripping // Program. — 1980. — Vol. 14, No. 3. — P. 130–137.

Tvardik N., Kergourlay I., Bittar A., Segond F., Darmoni S., Metzger M.-H. Accuracy of using natural language processing methods for identifying healthcare-associated infections // Int J Med Inform. — 2018. — Vol. 117. — P. 96–102. — DOI:10.1016/j.ijmedinf.2018.06.002

Yoo I.-H., Song M. Biomedical ontologies and text mining for biomedicine and healthcare: A survey // J Comput Sci Eng. — 2008. — Vol. 2 (2). — P. 109–136.