На правах рукописи

Бородкин Артем Александрович

РАЗРАБОТКА МЕТОДА ПОВЫШЕНИЯ БЫСТРОДЕЙСТВИЯ НЕПАРАМЕТРИЧЕСКИХ КЛАССИФИКАТОРОВ БИБЛИОГРАФИЧЕСКИХ ТЕКСТОВЫХ ДОКУМЕНТОВ

Специальность 05.13.01 – системный анализ, управление и обработка информации (энергетика, приборостроение, информатика, производственные процессы)

Автореферат диссертации на соискание ученой степени кандидата технических наук

Москва – 2012

Работа выполнена на кафедре Управления и информатики ФГБОУ ВПО «Национальный исследовательский университет «МЭИ»

Научный руководитель: доктор технических наук
доцент
Толчеев Владимир Олегович

Официальные оппоненты: **Ковшов Евгений Евгеньевич**
доктор технических наук
профессор
заведующий кафедрой «Управление и информатика в технических системах» ФГБОУ ВПО МГТУ «СТАНКИН»

Орлов Александр Иванович
доктор технических наук
доктор экономических наук
профессор
профессор кафедры «Экономика и организация производства» ФГБОУ ВПО МГТУ им. Н.Э. Баумана

Ведущая организация: ФГБУН Институт проблем управления им. В.А. Трапезникова РАН

Защита состоится "24" мая 2012 г. в 16 часов 00 мин. на заседании диссертационного совета Д 212.157.08 при НИУ МЭИ по адресу: 111250, Москва, ул. Красноказарменная, д. 14, Малый актовый зал.

С диссертацией можно ознакомиться в библиотеке НИУ МЭИ.

Отзывы в двух экземплярах, заверенные печатью, просьба направлять по адресу: 111250, Москва, ул. Красноказарменная, д. 14, Ученый совет НИУ МЭИ.

Автореферат разослан " " _____ 2012 года

Ученый секретарь
диссертационного совета Д 212.157.08
кандидат технических наук, доцент Д.Н.Анисимов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Стремительный рост объемов текстовых документов, в том числе научно-технических статей, существенно увеличил потребности пользователей в эффективных программно-алгоритмических средствах анализа документальной информации. Одним из наиболее востребованных на практике направлений обработки текстовых данных является классификация, которая позволяет упорядочивать большие документальные массивы и снизить информационную нагрузку на пользователя.

Анализ российских и зарубежных публикаций показывает, что основные усилия исследователей сконцентрированы на построении классификаторов, обладающих высокой точностью. Однако при разработке методов классификации текстовых данных, имеющих высокую размерность (большое число терминов, описывающих документ), особое внимание требуется уделять также вопросам быстродействия (т.е. уменьшению времени, затрачиваемого на отнесение документа к одному из классов). Обеспечение высокого быстродействия важно при решении таких задач как обработка коротких новостных сообщений в информационных агентствах, анализ вопросов, поступающих в ходе on-line конференций, автоматизированное разнесение документов по классам в больших цифровых библиотеках, организация эффективного документооборота в крупных компаниях, отслеживание и анализ публикаций на сайтах научных журналов по заданным тематикам.

На практике реализация мер, направленных на увеличение точности классификации, обычно приводит к снижению быстродействия. Существует лишь небольшое число методов классификации, для которых могут быть разработаны специальные процедуры, позволяющие повысить быстродействие практически без потерь в точности. Прежде всего, к таким классификаторам относятся *непараметрические методы* (метод ближайшего соседа и его модификации, метод потенциальных функций). Непараметрические методы обеспечивают достаточно высокую точность, однако затрачивают значительное время на классификацию новых документов. В специализированной литературе предлагаются различные модификации непараметрических классификаторов с целью увеличения быстродействия. Эти модификации можно разделить на две группы: *методы ускоренного поиска ближайшего соседа,* использующие упорядочивание обучающей выборки, и *методы ре-*

дукции (сокращения) размеров обучающих выборок. При этом вопросам разработки методов редукции в литературе уделяется значительно меньше внимания, чем построению методов ускоренного поиска ближайшего соседа. В большинстве известных работ рассматриваются методы редукции выборок, которые содержат фактографическую информацию. Вместе с тем при классификации больших массивов неструктурированных текстовых данных, обладающих высокой размерностью, особо важно использовать процедуры, «ускоряющие» непараметрические классификаторы и практически не изменяющие их точность.

Необходимо отметить, что в крупных хранилищах текстовых данных в свободном (бесплатном) доступе имеются документы, чаще всего представленные в виде библиографических описаний, т.е. состоящие из названия, аннотации, ключевых слов, фамилий авторов и другой вспомогательной информации. Доступ к полнотекстовым версиям обычно реализуется на коммерческой основе. В связи с этим обработку и анализ научных статей (например, публикаций в ведущих профессиональных изданиях) целесообразно проводить по их библиографическим описаниям.

Объектом исследований в данной работе являются системы обработки и анализа текстовых документов, позволяющие проводить классификацию документальной информации.

Предметом исследований в диссертации являются методы редукции обучающих выборок и непараметрические методы классификации библиографических текстовых документов.

Цель работы: увеличение быстродействия непараметрических классификаторов библиографической текстовой информации без существенного снижения их точности на основе разработки метода редукции обучающей выборки.

Для достижения указанной цели необходимо:

1. Сформулировать целевой показатель редукции, учитывающий требования по точности и быстродействию.
2. Провести комплексный сравнительный анализ известных методов редукции.
3. С позиций сформулированного целевого показателя разработать метод редукции обучающих выборок, позволяющий увеличить быстродействие непараметрических методов классификации без существенных потерь в точности.

4. Исследовать предложенный метод редукции на различных выборках, состоящих из библиографических текстовых документов.
5. Разработать и применить комплексную методику выбора процедур (и параметров) обработки и анализа текстовых данных на основе статистических непараметрических критериев.
6. На основе предложенных процедур и известных методов разработать программный комплекс для обработки и анализа массивов библиографических документов.

Методы исследования. Полученные в диссертации результаты основываются на применении методов теории вероятностей, математической статистики, линейной алгебры, теории множеств, вычислительной геометрии, теории алгоритмов.

Научная новизна.

1. Обоснован и исследован критерий выявления “внутренних” документов, основанный на новой формуле линейного взвешивания *k*-ближайших соседей.

2. Разработан новый метод редукции, основывающийся на критерии выявления “внутренних” документов, алгоритме выбора радиуса окрестности для каждого класса и модифицированном методе прототипов для объединения “внутренних” документов. Даны рекомендации по выбору настраиваемых параметров разработанного метода, приведены оценки вычислительной сложности.

3. В результате исследований на различных выборках было установлено, что разработанный метод редукции удовлетворяет сформулированному целевому критерию и в среднем на 19 процентов увеличивает быстродействие и практически не изменяет ошибку классификации метода *k*-ближайших соседей.

4. С помощью разработанной методики, использующей статистические непараметрические критерии, обоснован выбор использованных в работе процедур предварительной обработки текстовых документов, определены значения настраиваемых параметров методов классификации и редукции.

Практическая ценность результатов. Разработан учебно-исследовательский программный комплекс (*УИПК*), позволяющий проводить эффективную предварительную обработку, редукцию обучающих выборок и классификацию библиографической текстовой информации. В *УИПК* наряду с алгорит-

мами известных методов редукции включены алгоритмы, предложенные автором. Разработанное программное обеспечение может быть адаптировано к различным предметным областям и требованиям пользователя, при необходимости оно может дополняться новыми модулями. *УИПК* предназначен для широкого круга исследователей, не имеющих специальных знаний в области программирования и теории классификации. *УИПК* позволяет успешно решать как прикладные научно-исследовательские, так и учебные задачи.

Достоверность и обоснованность научных положений, рекомендаций и выводов подтверждается результатами экспериментальных исследований, проведенных на различных англоязычных и русскоязычных выборках библиографических текстовых документов, а также сопоставлением собственных результатов с результатами ранее выполненных работ по разработке методов редукции фактографических и документальных данных.

Реализация результатов. Программные модули *УИПК* были использованы при реализации проекта по созданию информационно-аналитической системы Института проблем химической физики РАН (ИПХФ РАН). Эффективность практического применения разработанного программно-алгоритмического обеспечения подтверждается актом об использовании результатов диссертационной работы в ИПХФ РАН. *УИПК* внедрен в учебный процесс кафедры управления и информатики МЭИ, на его базе проводится 3 лабораторные работы по курсу «Интеллектуальные информационные системы». Применение разработанного программного комплекса в учебном процессе подтверждено актом о внедрении.

Апробация работы. Материалы диссертации докладывались на четырех конференциях “Информационные средства и технологии” (2007, 2008, 2009, 2010 гг., Москва, МЭИ), на Научной сессии МИФИ (2008 г., Москва, МИФИ), на двух научно-технических семинарах “Современные технологии в задачах управления, автоматики и обработки информации” (2007, 2011 гг., Алушта, МАИ).

Публикации. По теме диссертации опубликовано 10 работ, в том числе 2 статьи в журналах из Перечня ВАК.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы, содержащего 129 наименований, 3-х прило-

жений. Основной текст диссертации излагается на 150 машинописных страницах и содержит 34 рисунка и 17 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность решаемой в диссертации проблемы, формулируется цель и задачи работы. Приводятся основные результаты, выносимые на защиту и определяющие новизну работы, излагается практическая ценность проведенных исследований.

В первой главе вводятся ключевые термины и определения, используемые в работе, формулируется задача классификации и редукции обучающих выборок. С учетом специфики решаемой задачи описываются основные этапы обработки и анализа текстовой информации. Для этапа сбора данных излагаются рекомендации по выбору объема и способа формирования обучающих выборок. На этапе начальной и содержательной обработки массив документов описывается в виде матрицы «документ-термин». При этом отдельный документ представляется вектором (точкой) в *M*–мерном пространстве терминов. Для получения координат этого вектора проводится взвешивание терминов (применяется формула *tf*-*idf*-взвешивания или ее модификации). Для снижения размерности многомерного пространства в работе используется процедура выделения корней слов (стемминг). На этапе разведочного анализа текстовых данных рассматривается целесообразность выявления глобальных выбросов с помощью принципа расстояния и построения диаграмм размаха Тьюки. Для настройки параметров классификаторов и методов редукции в работе используются обучающие и тестовые выборки. Оценка ошибок классификации осуществляется на экзаменационных выборках.

Основное внимание в первой главе уделяется анализу непараметрических классификаторов, прежде всего, метода ближайшего соседа, метода *k*-ближайших соседей (метод *k*-БС) и взвешенного метода *k*-ближайших соседей. Для взвешенного метода *k*-ближайших соседей автором предложены и исследованы две новые формулы линейного взвешивания:

$$\omega_j = \begin{cases} \dfrac{d_k - d_j + d_1}{d_k d_1} & , d_j \neq d_1 \\ 1 \big/ d_1 & , d_j = d_1. \end{cases} \quad (1) \quad \text{и} \quad \omega_j = \begin{cases} \dfrac{d_k - d_j + d_1}{d_k} & , d_j \neq d_1 \\ 1 & , d_j = d_1. \end{cases} \quad (2)$$

Здесь d_j –расстояние между классифицируемым документом и его *j*-м ближайшим соседом, найденное с помощью евклидовой метрики расстояния (*j* = 1,…, *k*); d_1 и d_k – расстояния между классифицируемым документом и соответственно его 1-м и *k*-м ближайшими соседями; ω_j – вес *j*-го ближайшего соседа.

В первой главе также уточняется постановка задачи исследования и формулируется целевой показатель редукции. Отмечается, что при проведении «агрессивной» редукции (более 30 процентов) для существенного увеличения быстродействия метода *k*-БС наблюдается быстрый рост ошибки. Однако снижение точности при «агрессивной» редукции более чем на 5 процентов может сделать нецелесообразным использование метода *k*-БС для решения ряда прикладных задач. Достижению цели диссертационной работы наилучшим образом соответствует «умеренная» редукция, которая предусматривает сокращение элементов обучающей выборки в диапазоне от 10 до 30 процентов от исходного размера, и чаще всего лишь незначительно ухудшает точность классификации.

Для проведения «умеренной редукции» в работе формулируется следующий целевой показатель: метод редукции должен обеспечивать сокращение исходного размера обучающей выборки более чем на 10 процентов при условии, что ошибка классификации метода *k*-ближайших соседей после проведения редукции увеличится менее чем на 3 процента.

Вторая глава посвящена разработке метода редукции обучающих выборок, удовлетворяющего сформулированному целевому показателю. Проводится сравнительный анализ известных методов редукции (метод нахождения прототипов, сжатый метод ближайшего соседа, редактируемый метод ближайшего соседа, выборочный метод ближайшего соседа, методы редукции *DROP1,…,DROP4*). Делается вывод, что ряд методов редукции позволяют обеспечить существенное сокращение объема обучающих выборок (более чем на 40%), однако при этом достаточно сильно снижается точность классификации, что делает нецелесообразным использование непараметрических классификаторов.

Для разработки метода редукции, позволяющего сохранять точность классификации практически без изменений, в работе предлагается в обучающей выборке провести выявление “внутренних” документов – документов, которые находятся в окружении преимущественно документов своего класса и вдалеке от элементов

других классов. Затем “внутренние” документы проходят проверку на возможность их объединения с точки зрения выполнения требований целевого показателя и осуществляется редукция выборки. Отнесение документа $\vec{X}_j$ из обучающей выборки к группе “внутренних” документов проводится на основе анализа совпадения метки класса документа $\vec{X}_j$ с метками классов у документов, попадающих в гиперсферу радиуса R_j с центром в $\vec{X}_j$.

В работе исследуется шесть критериев определения “внутренних” документов: *критерий* расчета соотношения числа “своих” документов (одного класса с $\vec{X}_j$) и числа “чужих” документов (из других классов) внутри гиперсферы радиуса R_j; *критерий* вычисления соотношения евклидовых расстояний (расстояний между $\vec{X}_j$ и “своими” документами в гиперсфере радиуса R_j и расстояний между $\vec{X}_j$ и “чужими” документами в гиперсфере); *критерий* рангового взвешивания для определения весов документов в гиперсфере (при этом ближайший документ к $\vec{X}_j$ имеет наибольший ранг, а самый дальний документ в гиперсфере - наименьший ранг); *критерий* линейного взвешивания документов в гиперсфере (исследовалось пять формул взвешивания, две из которых были предложены автором – формулы (1) и (2)); *комбинированный критерий*, использующий расчет соотношения евклидовых расстояний и ранговое взвешивание; *обобщенный критерий*, включающий вычисление соотношения числа “своих” и “чужих” документов, расстояний, рангов и линейное взвешивание.

Проведенные исследования показали, что с точки зрения сформулированного целевого показателя редукции предпочтительно использовать критерий линейного взвешивания на основе предложенной автором формулы взвешивания (2), имеющий вид:

$$\gamma = \frac{w^+}{w^+ + w^-} \geq 1 - \delta \qquad (3)$$

Здесь w^+ – сумма весов ближайших соседей документа $\vec{X}_j$, принадлежащих g-му классу, w^- – сумма весов ближайших соседей документа $\vec{X}_j$, не принадлежащих g-му классу (веса вычисляются по формуле (2)), δ - пороговое значение, позволяющее регулировать степень редукции и выбираемое из интервала [illegible];0,5].

Так как значение δ в формуле (3) заранее неизвестно, то для применения критерия на практике требуется определить два параметра: *пороговое значение* δ и *радиус гиперсферы* R_j. Такие задачи обычно решаются в ходе экспериментальных исследований путем фиксирования величины одного из параметров и настройки другого. В данной диссертационной работе настройка параметра R_j при фиксированном δ проводится так, чтобы получить наибольшее количество "внутренних" документов, к которым применима операция объединения.

Алгоритм расчета радиуса окрестности

Входные данные: обучающая выборка размера N (массив векторов $\{\vec{X}_j\}$, $j = 1,...,N$); значение δ из диапазона [0;0,5), критерий γ .

Выходные данные: радиусы окрестности для каждого класса $R_1,\ldots,R_g,\ldots,R_G$; массив "внутренних" документов (массив векторов $\{\vec{X}_j\}$, $j = 1,...,N_1$; $N_1 \le N$), массив попарных расстояний $\{W\}$.

Описание алгоритма:

1. Рассчитываются попарные расстояния между всеми документами выборки и сохраняются в массиве $\{W\}$. Находится минимальное и максимальное значение расстояний. Вычисляются значения радиусов окрестностей:

$$r_i = d_{\min} + i\frac{d_{\max} - d_{\min}}{100},\ i = 0,...,99. \qquad (4)$$

2. Для каждого значения радиуса r_i ($i = 0,...,99$) по заданному критерию γ определяются "внутренние" документы. Подсчитывается количество "внутренних" документов внутри g-го класса ($g = 1,...,G$).
3. В качестве радиуса окрестности R_g выбирается такое значение r_i, при котором количество "внутренних" документов g-го класса – максимально ($g = 1,...,G$).

Важным результатом выполнения данного алгоритма является получение массива "внутренних" документов, являющихся кандидатами для объединения. Объединение выявленных "внутренних" документов проводится согласно разработанному автором модифицированному методу прототипов.

Алгоритм модифицированного метода прототипов

Входные данные: обучающая и тестовая выборка, массив "внутренних" документов (массив векторов $\{\vec{X}_j\}$, $j = 1,\ldots,N_1$); значения радиусов окрестности для каждого класса $R_1,\ldots,R_g,\ldots,R_G$, критерий γ, массив попарных расстояний $\{W\}$.

Выходные данные: редуцированная обучающая выборка, полученная за счет слияния "внутренних" документов (размер редуцированного множества $N_2 \le N_1$).

Описание алгоритма:

1. Для всех "внутренних" документов $\vec{X}_j$ выполняются следующие операции:
   a. В массиве попарных расстояний $\{W\}$ для $\vec{X}_j$ находятся "свои" соседи ($\vec{X}_m \in Q_g$) и "чужие" соседи ($\vec{X}_p \notin Q_g$), попавшие в гиперсферу радиуса R_g.
   b. Рассчитывается значение разницы *Z* между количеством "своих" и "чужих" соседей.
2. Составляется список документов, упорядоченный по убыванию разницы *Z*.
3. Из упорядоченного списка выбираются *S* документов, принадлежащих разным классам ($S \le G$). Для каждого из выбранных документов выполняется объединение (усреднением) $\vec{X}_j$ с его ближайшим "своим" соседом. Для нового элемента, полученного усреднением, находится значение критерия γ. Объединение признается успешным, если для всех S новых документов выполняется условие $\gamma \ge 1-\delta$ (то есть документы по-прежнему относятся к категории "внутренние"), в противном случае осуществляется иной выбор документов. В случае успешного объединения множество векторов $\vec{X}_j$ сокращается на *S* документов. Если ни для одного класса не удается найти документ, удовлетворяющий указанному условию, то осуществляется переход к шагу 5.
4. Если ошибка классификации тестовой выборки методом *k*-БС при обучении на редуцированном множестве не превосходит ошибку классификации тестовой выборки при обучении на исходном обучающем множестве более чем на 3%, то проводится пересчет матрицы попарных расстояний и возврат к шагу 1 для выбора новых элементов для объединения.

5. Вывод редуцированного множества внутренних документов.

Таким образом, разработанный *метод редукции* состоит из следующих этапов:

1. Задается целевой показатель, критерий определения “внутренних” документов, непараметрический классификатор.
2. По алгоритму расчета радиуса окрестности вычисляются радиусы окрестностей для каждого класса $R_1,\ldots,R_g,\ldots,R_G$, выбирается значение порога δ, выявляется массив “внутренних” документов.
3. Согласно модифицированному методу прототипов на основе объединения “внутренних” документов формируется редуцированная обучающая выборка.
4. На экзаменационных выборках, которые (в отличие от обучающих и тестовых выборок) не участвовали в процессе настройки параметров редукции, оценивается точность и быстродействие непараметрического классификатора до и после редукции.

Проведем оценку вычислительной сложности и быстродействия классификации методов семейства *k*-ближайших соседей при обучении на редуцированных выборках.

Исходные решающие правила при использовании непараметрических методов для классификации редуцированных выборок не изменяются. Эффект увеличения быстродействия достигается за счет сокращения размера обучающей выборки и, как следствие, снижения числа вычислительных операций, которые необходимы для определения класса нового документа. Для метода *k*-ближайших соседей справедливо: $O_{k-БС}^{до\ редукции} = N \cdot O(M)$. Здесь $O_{k-БС}^{до\ редукции}$ – число вычислительных операций, необходимых для определения метки класса документа; $O(M)$ – вычислительная сложность наиболее ресурсозатратной операции расчета (евклидова) расстояния, N – размер обучающей выборки до редукции, M – размер словаря терминов (количество терминов в документе).

Выигрыш в быстродействии метода *k*-ближайших соседей обеспечивается в том случае, если в ходе выполнения редукции были объединены документы из обучающего массива (т.е. $N_2 < N$).

Во второй главе также дается описание выборок, применяемых для исследований. В экспериментах используется девять выборок – по три выборки из англоязычных библиографических баз данных *ACM* (выборки обозначаются *A1,A2,A3*), *Compendex* (*C1,C2,C3*), *ResearchIndex* (*R1,R2,R3*). Обучающие выборки содержат 700 документов, тестовые и экзаменационные выборки включают 140 документов, во всех выборках документы поровну распределены по 7 классам. Кроме того, в экспериментальных исследованиях используется шесть выборок библиографических документов из *русскоязычной цифровой библиотеки eLibrary*. Три из них (*V1,V2,V3*) имеют одинаковое количество документов в классах (объем обучающих выборок $N = 200$, размер тестовых и экзаменационных $n = 50$, количество классов в выборках $G = 5$), три другие – (*V4,V5,V6*) имеют размер обучающих выборок $N = 266$ и неодинаковое распределение документов по классам ($N_1 = 70,\ N_2 = 63,\ N_3 = 53,\ N_4 = N_5 = 40$), размер тестовых и экзаменационных выборок $n = 68$. На основе проведенных экспериментов даются рекомендации по выбору параметров предложенного метода.

На рисунке 1 приведена зависимость минимальной степени редукции, полученной на девяти англоязычных выборках, от величины δ при использовании различных критериев выявления “внутренних” документов: а) *критерия* вычисления соотношения евклидовых расстояний; б) критерия линейного взвешивания документов по формуле (2); в) обобщенного критерия.

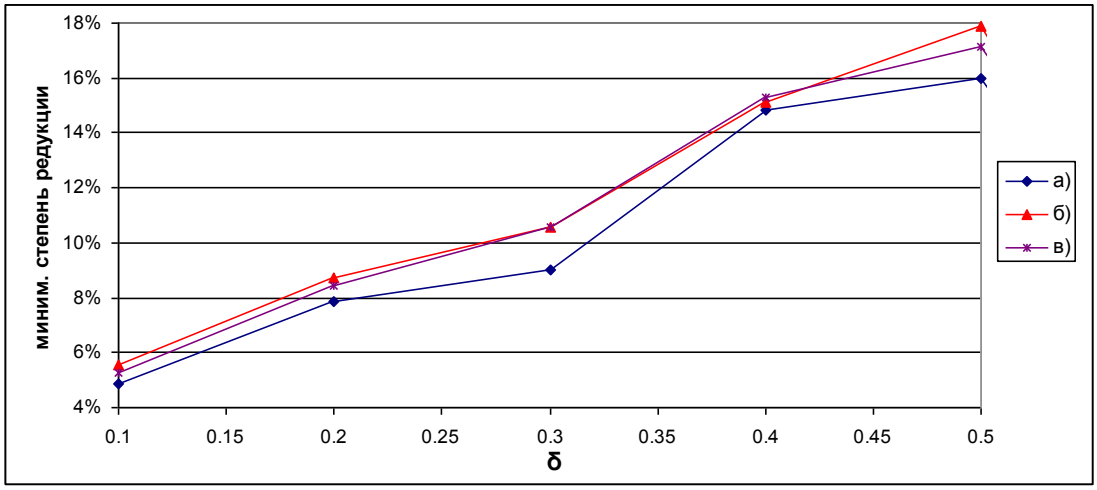

Рисунок 1. Зависимость минимальной степени редукции от значений δ при использовании различных критериев выявления “внутренних” документов

На рисунке 2 приводятся результаты расчета изменения ошибки метода *k*-БС на экзаменационных выборках после проведения редукции девяти англоязычных обучающих выборок.

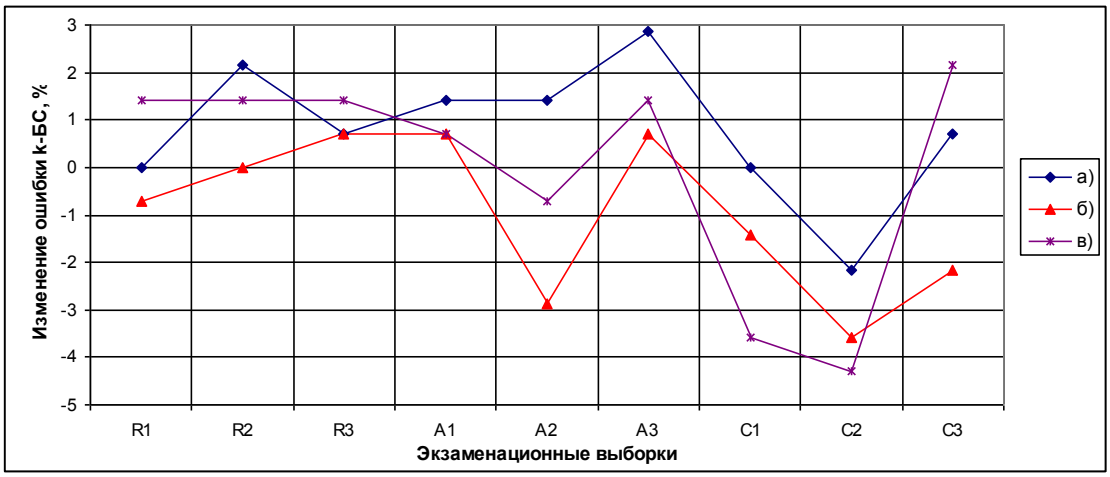

Рисунок 2. Изменение ошибок метода *k*-БС на экзаменационных выборках при использовании различных критериев выявления “внутренних” документов

Изменения ошибок классификации $\Delta\varepsilon$, представленные на рисунке 2, могут быть как положительными, так и отрицательными. Положительные значения $\Delta\varepsilon$ свидетельствуют об ухудшении точности классификации после редукции обучающей выборки, а отрицательные значения $\Delta\varepsilon$ – об улучшении точности классификации на редуцированных выборках.

Анализ результатов, приведенных на рисунках 1 и 2, позволяет сделать вывод, что «умеренной» редукции соответствует интервал варьирования δ от 0,35 до 0,4 (при уменьшении значений δ будет проводиться «мягкая» редукция, а при увеличении – «агрессивная»). Отметим также, что критерий линейного взвешивания документов по формуле (2) обеспечивает практически такую же степень редукции, как и более сложный для расчета обобщающий критерий (при этом критерий линейного взвешивания на большинстве редуцированных выборок обеспечивает более высокую точность классификации).

На рисунке 3 показана зависимость количества “внутренних” документов от радиуса окрестности R_g (для класса «Control systems synthesis», при $\delta = 0.4$).

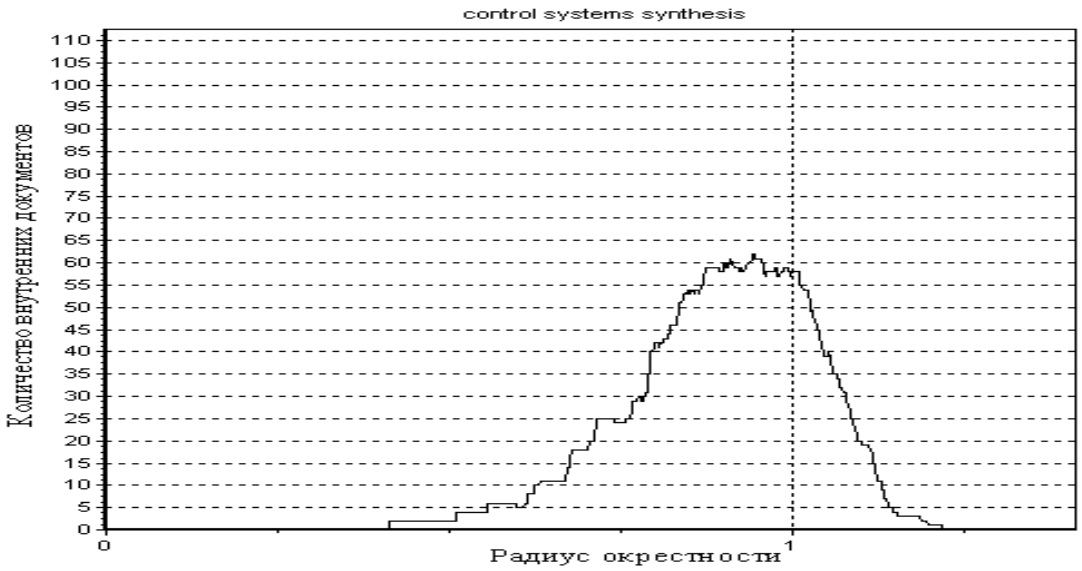

Рисунок 3. Зависимость количества “внутренних” документов от радиуса окрестности в классе «Control systems synthesis» при использовании весового критерия.

На основе проведенных экспериментальных исследований определено, что для вышеуказанного диапазона изменения δ средние значения R_g будут изменяться в интервалах: $R_g \in [0{,}87;\ 1{,}18]$.

Третья глава посвящена разработке и применению методики выбора процедур (и параметров) обработки и анализа текстовых данных на основе непараметрических критериев Вилкоксона и Фридмана. Разработанная методика позволяет устранить субъективность выбора процедур и параметров при проведении предварительной обработки, классификации и редукции библиографических текстовых выборок. С ее помощью в работе обоснован выбор количества информативных терминов, способа взвешивания терминов и меры близости, определены значения настраиваемых параметров процедуры редукции.

Основное внимание в главе уделено исследованию разработанного метода редукции на различных выборках и сопоставлению результатов с методом *DROP4*, обладающим наилучшим соотношением «точность-степень редукции» среди известных процедур.

В таблице 1 представлены результаты расчета изменения ошибок и быстродействия разработанного метода и метода *DROP4* после проведения редукции англоязычных и русскоязычных выборок.

Таблица 1

Англоязычные выборки					Русскоязычные выборки				
	DROP-4		Разработанный метод редукции			*DROP-4*		Разработанный метод редукции	
Выборка	$\Delta\varepsilon$	Δt	$\Delta\varepsilon$	Δt	Выборка	$\Delta\varepsilon$	Δt	$\Delta\varepsilon$	Δt
A1	-4.29	66	0.71	23	V1	8	42	2	21
A2	2.14	58	-2.86	23	V2	0	39	-2	24
A3	0	54	0.71	22	V3	2	41	0	19
C1	0	51	-1.43	18	V4	4.41	44	2.94	23
C2	0	56	-3.57	18	V5	7.35	45	1.47	17
C3	-2.86	52	-2.15	18	V6	8.82	47	2.94	12
R1	-0.72	55	-0.72	17					
R2	0.71	44	0	18					
R3	2.14	54	0.71	15					

В таблице 1 использованы обозначения: $\Delta\varepsilon$ - изменение ошибки метода *k*-ближайших соседей на редуцированных выборках (в %), Δt - увеличение быстродействия классификации на редуцированных выборках (в %).

Исследование точности классификации метода *k*-ближайших соседей на редуцированных выборках, полученных с помощью разработанного метода редукции и метода *DROP4*, показало, что разработанный метод обеспечивает большую точность, чем *DROP4*, и на всех выборках удовлетворяет целевому показателю, обеспечивая на англоязычных и русскоязычных выборках увеличение быстродействия в среднем на 19%. Дополнительные исследования разработанного метода редукции продемонстрировали его устойчивость к незначительным изменениям структуры выборок (вариациям размера выборок и количества документов в классах). Эксперименты на выборках из БД *ResearchIndex* с разным количеством документов в обучающей выборке показали, что разработанный метод редукции, в отличие от *DROP4*, на всех выборках удовлетворяет требованиям целевого показателя и обеспечивает более высокую точность.

В четвертой главе дается краткий обзор существующих программных решений для обработки и анализа текстовой информации. Приводится структура и функциональные возможности разработанного учебно-исследовательского программного комплекса. Описывается разработанный комплекс лабораторных работ по курсу «Интеллектуальные и информационные системы» и демонстрируется применение *УИПК* для решения прикладных задач.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Проведен обзор и сравнительный анализ известных методов редукции. Предложен целевой показатель, который предусматривает сокращение размера обучающей выборки более чем на 10 процентов при условии допустимого увеличения ошибки классификации менее чем на 3 процента.

2. Рассмотрены известные непараметрические методы классификации, проанализированы их преимущества и недостатки. Проанализированы способы устранения выявленных недостатков. Предложены две новые формулы линейного взвешивания *k*-ближайших соседей, применение которых не требует экспериментальной настройки дополнительных параметров и позволяет при расчете весов наиболее полно учитывать структуру выборки.

3. На основе разработанной автором формулы линейного взвешивания предложен новый критерий выявления “внутренних” документов. Экспериментально исследовано шесть критериев выявления “внутренних” документов и обосновано применение нового критерия для проведения редукции обучающих выборок.

4. Разработан новый метод редукции, основывающийся на предложенном автором критерии выявления “внутренних” документов, алгоритме выбора радиуса окрестности для каждого класса и модифицированном методе прототипов для объединения “внутренних” документов. Даны рекомендации по выбору настраиваемых параметров разработанного метода, приведены оценки вычислительной сложности.

5. Показано, что разработанный метод редукции удовлетворяет сформулированному целевому показателю и при практически неизменной ошибке классификации в среднем на 19 процентов сокращает размер англоязычных и русскоязычных обучающих выборок (соответственно также в среднем на 19 процентов увеличивает быстродействие метода *k*-ближайших соседей). Разработанный метод обладает устойчивостью по отношению к небольшим изменениям структуры выборок (размера выборки и количества документов в классах).

6. Разработана и обоснована методика использования статистических непараметрических критериев для выбора наиболее подходящих процедур обработки и анализа текстовых данных. Предложенная методика применена на практике для выбора алгоритмов предварительной обработки текстовых данных, параметров методов классификации и редукции.

7. Разработан, апробирован и внедрен в учебный процесс учебно-исследовательский программный комплекс. Наряду с известными алгоритмами обработки и анализа в *УИПК* включены процедуры, разработанные автором. Данный программный комплекс может быть адаптирован к различным предметным областям и требованиям пользователей, при необходимости он может дополняться новыми модулями. Продемонстрированы возможности *УИПК* по проведению комплексных исследований методов обработки текстовой информации и решению прикладных и образовательных задач.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК:

1. Бородкин А.А., Толчеев В.О. Разработка учебно-исследовательского программного комплекса для обработки и анализа библиографических текстовых документов. Вестник МЭИ №1 2010, с. 96-102
2. Бородкин А.А., Толчеев В.О. Разработка комплексной процедуры редукции для увеличения быстродействия непараметрических методов классификации текстовых документов. Заводская лаборатория. Диагностика материалов. №11 2011, с.64-69.

Другие статьи и материалы конференций:

3. Бородкин А.А., Толчеев В.О. Об оценке точностных и временных характеристик методов классификации библиографических текстовых документов. Научная сессия МИФИ 2008. Том 11. М. МИФИ, 2008, стр. 152-153.
4. Бородкин А.А., Толчеев В.О. Исследование влияния структуры выборки и процедур предварительной обработки на точность классификации текстовой информации. Международная конференция “Информационные средства и технологии”. Том 2. МЭИ. Изд-во Станкин, 2007, с. 33-34.
5. Бородкин А.А. Комплексная процедура редукции выборок текстовых документов // Международный форум информатизации МФИ-2010. Труды XVIII международной научно-технической конференции «Информационные средства и технологии». Т.3. - М.:МЭИ, 2010 – с. 251-254
6. Бородкин А.А., Толчеев В.О. Методы удаления нерелевантных документов из обучающих выборок. Международный форум информатизации МФИ-2009.

Труды XVII международной научно-технической конференции «Информационные средства и технологии». Т.3. - М.:МЭИ, 2009 – с. 169-173

7. Бородкин А.А., Толчеев В.О., Часовский А.В. Исследование зависимости точности классификации от структуры выборки// Современные технологии в задачах управления, автоматики и обработки информации: труды XVI Международного научно-технического семинара. Сентябрь 2007 г., Алушта. – Тула: Изд-во ТулГУ, 2007 – с. 244-245
8. Бородкин А.А., Толчеев В.О. Структура и функциональные возможности учебно-исследовательского программного комплекса. Международный форум информатизации МФИ-2008. Труды XVI международной научно-технической конференции «Информационные средства и технологии». Т.3. - М.:МЭИ, 2008 – с. 85-87
9. Бородкин А.А. Толчеев В.О. Применение метода потенциальной функции для классификации библиографических текстовых документов // Научная сессия МИФИ-2008. Сборник научных трудов. Т.11. Технологии разработки программных систем. Информационные технологии. – М.: МИФИ, 2008 – с. 150-151.
10. Бородкин А.А., Дербенев Н.В., Толчеев В.О. Программно-алгоритмические средства обработки и анализа библиографической текстовой информации. Современные технологии в задачах управления, автоматики и обработки информации: тезисы докладов XX Международного научно-технического семинара (г. Алушта, 18-24 сентября 2011 г.) – Пенза: Изд-во ПГУ, 2011 – с. 267-268.