УДК 519.24

РАЗРАБОТКА НОВЫХ МОДИФИКАЦИЙ ПРОФИЛЬНЫХ МЕТОДОВ КЛАССИФИКАЦИИ И ПОСТРОЕНИЕ КОЛЛЕКТИВОВ РЕШАЮЩИХ ПРАВИЛ

© **А. С. Мохов, В. О. Толчеев**[1]

Статья поступила 4 июля 2014 г.

Рассмотрена задача повышения точности классификации двуязычных (русско-английских) текстовых документов. На основе известных профильных методов разработаны новые модификации, которые используются для формирования коллективов решающих правил (КРП). Исследовано влияние точности и разнородности членов КРП на качество классификации. Результаты экспериментов на сформированных двуязычных выборках показали, что применение КРП позволяет снизить ошибку классификации.

Ключевые слова: обработка и анализ двуязычных текстовых документов; методы классификации; профильные методы; коллективы решающих правил; точность (ошибка) классификации.

Цель работы — повышение точности классификации двуязычных библиографических документов, заданных своими названиями и аннотациями на русском и английском языках. Это достигается за счет разработки новых модификаций профильных методов и построения коллективов решающих правил (КРП), использующих простое голосование для принятия решения. Рассмотрим возможности, которые предоставляют КРП, для снижения ошибки классификации.

Коллективы решающих правил

КРП — это несколько (обычно три и более) классификаторов (решающих правил, методов), объединенных для выработки общего решения [1, 2]. В случае простого голосования все члены коллектива при принятии решения имеют равный вес β_p. При этом соблюдается условие: $\sum_{p=1}^{m}\beta_p = 1$ (β_p — вес p-го классификатора, причем $\beta_p = 1/m$, $p = 1, ..., m$, m — количество классификаторов в КРП).

Распределение документов по классам осуществляется в два этапа. На первом из них для классификации нового наблюдения используются все методы, включенные в КРП. На втором этапе определяется класс, за который проголосовало большинство решающих правил.

В условиях практически полного отсутствия априорной информации о структуре документального массива КРП позволяют получать наиболее точное из возможных решений за счет использования дополняющих друг друга решающих правил и специальных стратегий обучения.

Утверждение. При объединении в КРП не менее трех равноточных и независимых классификаторов точность классификации повышается по сравнению с точностью индивидуальных решающих правил.

Данное утверждение доказывается на основе теоремы Бернулли [3]. Проведем расчет для простейшего случая, когда в коллектив объединены три независимых классификатора, имеющие одинаковую точность (вероятность безошибочной классификации) $p_1 = p_2 = p_3 = 0{,}7$.

Коллектив решающих правил проведет правильную классификацию, если два (из трех) или все три классификатора безошибочно определят класс документа. Согласно формуле Бернулли вероятность случая, когда два классификатора ($n = 2$) правильно оценили класс документа, а один ошибся, составит

$$P_1 = C_m^n p^n (1-p)^{m-n} = \frac{m!}{m!(m-n)!} p^n (1-p)^{m-n} =$$

$$= C_3^2 (0{,}7)^2 (0{,}3)^1 = 0{,}441. \qquad (1)$$

Вероятность безошибочной классификации для случая, когда все три классификатора ($n = 3$) правильно оценили класс документа,

$$P_2 = C_m^n p^n (1-p)^{m-n} = C_3^3 (0{,}7)^3 (0{,}3)^0 = 0{,}343. \qquad (2)$$

Общая точность КРП

$$P_{\text{безошиб}} \le P_1 + P_2 = 0{,}343 + 0{,}441 = 0{,}784. \qquad (3)$$

Рассчитанное $P_{\text{безошиб}}$ на 8,4 % выше точности исходных методов. Аналогичные расчеты, проведенные для $m = 5$, $m = 7$ и $m = 9$, дали следующие результаты: $P_{\text{безошиб}}(m = 5) \le 0{,}83$, $P_{\text{безошиб}}(m = 7) \le 0{,}87$, $P_{\text{безошиб}}(m = 9) \le 0{,}9$.

Если точность членов КРП более высокая, например $p_1 = p_2 = p_3 = 0{,}8$, получаем: $P_{\text{безошиб}}(m = 3) \le 0{,}896$; $P_{\text{безошиб}}(m = 5) \le 0{,}942$; $P_{\text{безошиб}}(m = 7) \le 0{,}966$; $P_{\text{безошиб}}(m = 9) \le 0{,}98$ [3, 4].

[1] Национальный исследовательский университет «МЭИ», Москва, Россия; e-mail: tolcheev@mail.ru

Из приведенных расчетов можно сделать вывод, что для формирования высокоточных КРП необходимо иметь от трех до девяти независимых и равноточных методов. Однако на практике оба эти требования достаточно сложно обеспечить. Так, точность наиболее известных процедур классификации (метода К-ближайших соседей — К-БС, наивного байесовского классификатора, метода центроидов) изменяется в достаточно широком диапазоне, что не позволяет рассматривать эти решающие правила как равноточные [5, 6]. Более того, обучение обычно проводится на одних и тех же выборках, что затрудняет получение независимых классификаторов. В связи с этим рассчитанные по формуле Бернулли оценки оказываются завышенными и указывают на «наилучшую» (теоретически достижимую) точность. Они способны выступать лишь в качестве предельных значений, «ориентиров» при разработке коллективов, состоящих из реальных (необязательно полностью независимых и равноточных) решающих правил.

Двуязычные текстовые документы

КРП применяются для решения широкого круга задач в области обработки и анализа информации. В данной работе, как отмечалось ранее, они разрабатывались для увеличения точности классификации двуязычных научных библиографических документов (название, аннотация, ключевые слова, представленные на русском и английском языках). Для проведения исследований формировались «непараллельные» текстовые выборки из электронной библиотеки eLibrary.ru.

«Непараллельными» двуязычными выборками называются выборки, состоящие из текстов, написанных одновременно на двух языках. «Непараллельность» означает, что перевод с одного языка на другой проводился автором или переводчиком, а не средствами автоматизированного перевода («параллельными» выборками называются выборки, в которых двуязычные документы получаются с помощью машинного перевода текстов с одного языка на другой).

Для математического описания выборки двуязычных текстовых документов используется расширенная матрица «документ — термин», строки которой представляют собой документы, а столбцы — (русскоязычные и англоязычные) термины, содержащиеся в этих документах [7, 8]:

$$\begin{array}{c}\text{Русские термины} \qquad \text{Английские термины} \\ p \neq M/2 \\ \mathbf{X} = \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(p)} & x_1^{(p+1)} & \dots & x_1^{(M)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N^{(1)} & \dots & x_N^{(p)} & x_N^{(p+1)} & \dots & x_N^{(M)} \end{bmatrix},\end{array} \quad (4)$$

где $x_j^{(i)}$ — вес термина i в документе j ($i = 1, ..., M$, M — общее количество терминов в смешанной выборке; $j = 1, ..., N$, N — количество документов; p — количество русских терминов, $(M - p)$ — количество английских терминов, в общем случае $p \neq M/2$).

Вес терминов в вышеприведенной матрице может рассчитываться разными способами, например, с помощью обычного *tf*-взвешивания

$$x_j^{(i)} = f_{ij} \quad (5)$$

или хорошо известной в специализированной литературе формулы *tf-idf*-взвешивания [7]

$$x_j^{(i)} = f_{ij} \log\left(\frac{N}{N_i}\right). \quad (6)$$

Здесь f_{ij} — частота слова i в документе j, N_i — общее количество документов выборки, содержащих слово i.

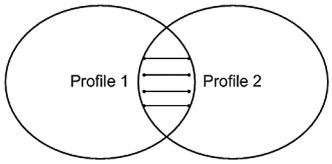
Главной особенностью двуязычных текстовых документов является то, что исследователь изначально располагает статьями на двух языках, т.е. каждую статью можно представить в виде двух документов, имеющих приблизительно одинаковую информационную ценность. При обработке русско-английских библиографических описаний в распоряжении у исследователя имеются наборы слов на каждом из языков. Эти наборы характеризуют и терминологически описывают одну и ту же проблему (задачу, решаемую в научной работе). По сравнению с анализом монолингвистических документов появляется дополнительная информация, которая может быть извлечена и задействована при обучении классификатора. Кроме того, появляется возможность настраивать параметры решающих правил, включаемых в КРП, на различных выборках (русскоязычных, англоязычных и смешанных — русско-англоязычных).

Новые профильные методы

Ранее отмечалось, что одним из «узких» мест при формировании КРП является отсутствие достаточного числа разработанных и исследованных классификаторов, которые можно было бы считать приблизительно равноточными. В данной работе предлагается ряд таких классификаторов, относящихся к профильным методам, и анализируется целесообразность их использования для построения КРП.

Профильными называются методы, в которых рассчитывается некоторый формальный объект — профиль класса, способный характеризовать все остальные элементы класса при классификации новых документов [9].

Профильные методы характеризуются наличием этапа обучения и этапа классификации. На этапе обучения происходят вычисление и построение профилей классов — выявление наиболее информативных терминов и расчет их весов. На этапе классификации новый текст анализируется, сравнивается с профилями, построенными при обучении, и принимается решение об отнесении документа к тому или иному классу. Ре-



Общие термины двух профилей

зультатом этапа обучения будет K профилей, рассчитанных для каждого класса, а результатом этапа классификации нового документа $\mathbf{X}_{N+1}$ — присвоенная ему метка класса ($\mathbf{X}_{N+1} \in Q_k$; $k = 1, ..., K$, K — количество классов).

Наиболее известным профильным методом является метод центроидов. Профильные методы, предлагаемые в данной работе, основаны на расчете таблиц сопряженности (табл. 1) и выявлении наиболее информативных классообразующих терминов.

В табл. 1 используются обозначения: A — число раз, когда термин $x^{(i)}$ и класс Q_k встречаются вместе; B — число раз, когда $x^{(i)}$ встречается без Q_k; C — число раз, когда Q_k встречается без $x^{(i)}$; D — число раз, когда ни Q_k, ни $x^{(i)}$ не встречаются; $A + B + C + D = N$ — общее количество документов в выборке.

В наших исследованиях применялись три способа выявления информативных терминов, из которых затем создавался соответствующий профиль: статистический (на основе χ^2-критерия), теоретико-информационный (на основе критерия взаимной информации), эвристический (на основе расчета коэффициентов ассоциативности). Профили вычисляются по следующим формулам [10] и имеют один настраиваемый параметр — длину профиля L — количество информативных терминов, включенных в профиль:

ФИ:
$$\Phi(x^{(i)}, Q_k) = \frac{\chi^2}{N} = \frac{(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}, \quad (7)$$

PO:
$$p(x^{(i)}, Q_k) = \sqrt{\frac{\chi^2}{N}} = \frac{(AD - CB)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}, \quad (8)$$

Таблица 1. Таблица сопряженности размера 2 × 2

Признак $x^{(i)}$	Класс Q_k	
	Принадлежность классу Q_k	Непринадлежность классу Q_k
Наличие признака $x^{(i)}$	A	B
Отсутствие признака $x^{(i)}$	C	D

Нормированный *МИ*-профиль (*НМИ*)

$$NMI(x^{(i)}, Q_k) = \frac{A \log \frac{AN}{(A+B)(A+C)}}{(A+B)\log \frac{N}{A+B}}. \quad (9)$$

Коэффициент ассоциативности Соукала – Сниса (*C-C*)

$$SS(x^{(i)}, Q_k) = \frac{2(A+D)}{2(A+D)+B+C}. \quad (10)$$

В работе были сформированы и исследованы новые профили — *UNI1*, *UNI2*, *UNI5*, *UNI6*. Для их разработки использовались различные комбинации *PO*-, *НМИ*- и *C-C*-профилей, имеющих различные принципы определения информативности терминов.

При построении новых процедур проверялись два эвристических предположения:

1) русско- и англоязычные термины обладают одинаковой информативностью и равнозначны при отборе в профиль;

2) русско- и англоязычные термины обладают различной информативностью и в профиль прежде всего имеет смысл включать более важные русскоязычные термины.

Для описания новых алгоритмов введем понятия «моноязычный» и «двузычный» профили. Моноязычный профиль в данной работе рассчитывается по документам одного языка и состоит только из русских (или английских) терминов. Двуязычный профиль определяется по смешанной (русско-английской) выборке. В состав такого профиля входят слова на обоих языках, причем соотношение терминов двух языков не фиксировано и зависит лишь от их весов.

В методе *UNI1* используется первое предположение. Исследуется целесообразность построения двуязычного профиля, в который включались самые информативные (имеющие наибольший вес) слова обоих языков, рассчитанные по формулам *PO*- и *НМИ*-профилей ($L = 200$). При этом в профиль отбирались только общие термины, присутствующие в двух исходных профилях (см. рисунок), и им присваивалось наибольшее значение из весов в *PO*- и *НМИ*-профилях.

Таким образом, *UNI1* содержит наиболее частотные термины, отобранные по формуле *PO*-профиля, а также специфические, достаточно редкие слова, отражающие терминологические особенности тематики и выявленные с помощью *НМИ*-профиля.

Алгоритм построения (обучения) профиля UNI1. Входными данными алгоритма являются: обучающая выборка документов, двуязычные *PO*- и *НМИ*-профили, рассчитанные по формулам (8) и (9). Длина *PO*- и *НМИ*-профилей равна L. Выходные данные: профили классов с упорядоченными по убыванию весами.

Шаг 1. Анализируются *PO*- и *НМИ*-профили, начиная с терминов с наибольшими весами. Выбирают-

ся общие (русские и английские) термины для обоих профилей.

Шаг 2. Каждому общему термину присваивается максимальный вес из значений в *PO*- и *НМИ*-профилях: $w_{UNI1} = \max(w_{PO}, w_{HMИ})$, где w_{UNI1} — вес термина в *UNI1*-профиле; w_{PO} — вес термина в *PO*-профиле; $w_{HMИ}$ — вес термина в *НМИ*-профиле.

Шаг 3. Полученные термины упорядочиваются по убыванию веса.

В методе *UNI2* проверяется второе предположение. Так как русский язык является «родным» для авторов статей, то можно предположить, что изложение материала на нем более квалифицированное и информативное, чем на английском (у авторов уровень знания иностранной терминологии ниже и, как следствие, описание темы менее качественное). В профиль *UNI2* включалось h общих терминов из русскоязычных *PO*- и *НМИ*-профилей, дополненных t общими англоязычными словами из соответствующих *PO*- и *НМИ*-профилей. Длина профиля $L = h + t$ (в наших исследованиях $h = 100$, $t = 100$ терминов).

Алгоритм построения (обучения) профиля UNI2. Входными данными алгоритма являются: обучающая выборка документов, моноязычные *PO*- и *НМИ*-профили, рассчитанные по формулам (8), (9) и имеющие длину L. Выходные данные: профили классов с упорядоченными по убыванию весами.

Шаг 1. Задаются параметры метода h и t.

Шаг 2. Среди русских *PO*- и *НМИ*-профилей выбираются h общих терминов с наибольшими весами, среди английских — t общих терминов.

Шаг 3. Каждому отобранному термину присваивается максимальный вес из значений моноязычных *PO*- и *НМИ*-профилей: $w_{UNI2} = \max(w_{PO}, w_{HMИ})$, где w_{UNI2} — вес термина в *UNI2*-профиле; w_{PO} — вес термина в *PO*-профиле, $w_{HMИ}$ — вес термина в *НМИ*-профиле.

Шаг 4. Полученные термины упорядочиваются по убыванию веса.

Модификацией метода *UNI2* являются процедуры *UNI5* и *UNI6*, в которых вместо выбора наибольшего веса рассчитывается сумма весов в исходных профилях. В *UNI5* для этого используются *НМИ*- и *C-C*-профили, в *UNI6* — *PO*-, *НМИ*- и *C-C*-профили.

Алгоритм построения (обучения) профилей UNI5 и UNI6. Входными данными алгоритма являются: обучающая выборка документов; моноязычные *C-C*-, *НМИ*- и *PO*-профили, рассчитанные соответственно по формулам (10), (9), (8) и имеющие длину L.

В процедуры *UNI5* и *UNI6* включались h классообразующих терминов из русскоязычных профилей, дополненных t наиболее информативными англоязычными словами.

Выходные данные: профили классов с упорядоченными по убыванию весами.

Шаг 1. Задаются параметры метода h и t.

Шаг 2. Среди русских *PO*-, *НМИ*- и *C-C*-профилей выбираются h общих терминов с наибольшими весами, среди английских — t общих терминов.

Шаг 3. Суммируются веса *PO*-, *C-C*- и *НМИ*-профилей для каждого общего термина: $w_{UNI5(6)} = \sum w_i$, где $w_{UNI5(6)}$ — вес термина в *UNI5* (*UNI6*)-профиле; w_i — вес термина в i-м профиле ($i = 1...3$).

Шаг 4. Полученные термины упорядочиваются по убыванию веса.

На этапе классификации рассчитываются значения весов классов для классифицируемого документа [9, 10]:

$$\omega_k = \sum_{i=1}^{M_k} tf_i \mathrm{Prof}(x^{(i)}, Q_k), \qquad (11)$$

где tf_i — частота встречаемости i-го слова в классифицируемом документе; M_k — количество наиболее информативных терминов, включенных в профиль k-го класса (в наших исследованиях все классы имели профиль одинакового размера $L = M_k$); $\mathrm{Prof}(x^{(i)}, Q_k)$ — вес i-го термина в профиле, вычисленном по одной из формул (8) – (10).

Решающее правило в профильных методах: классифицируемый документ $\mathbf{X}_{N+1}$ относится к тому классу ($\mathbf{X}_{N+1} \in Q_k$), которому соответствует наибольший вес — $\omega_k = \max \omega_k$ ($k = 1, ..., K$).

Описание выборок и экспериментальных исследований

Для проведения исследований были сформированы 33 обучающие и экзаменационные выборки, состоящие из семи классов каждая (11 — русскоязычных, 11 — англоязычных и 11 — смешанных). Обучающие и экзаменационные выборки между собой не пересекаются и состоят соответственно из 455 ($N = 455$) и 105 ($n = 105$) библиографических документов. Для формирования документальных массивов использовались публикации по различным тематикам Информатики (Computer Science), полученные по фиксированным запросам (ключевым словам) из электронной библиотеки eLibrary.ru. На этапе предварительной обработки данных были исключены стоп-слова, которые не несут смысловой нагрузки, и для обоих языков был проведен стемминг — выделение корня слова [9, 11].

На сформированных смешанных выборках были обучены и исследованы следующие классификаторы: *PO*-, *НМИ*-, *C-C*-, *UNI1*-, *UNI2*-, *UNI5*-, *UNI6*-профили. Результаты классификации на смешанных выборках приведены в табл. 2.

В работе также проводилось сравнение ошибок разработанных новых модификаций профильных методов с хорошо исследованным и часто применяемым на практике методом К-ближайших соседей (метод К-БС).

Как показали исследования, наиболее точными из профильных методов являются *PO*-профиль и синтезированный *UNI2*-профиль. Все остальные профильные процедуры показывают сопоставимые результаты, за исключением *C-C*-профиля, и их ошибки меньше ошибок метода К-ближайших соседей. Использование суммирования весов терминов в *UNI5* и *UNI6* не улучшает точность классификации по сравнению с *UNI2*, в котором выбирается наибольший вес из двух профилей (усреднение весов также малоэффективно).

Небольшие различия в точности методов *UNI1* и *UNI2* не позволяют выделить из двух сделанных ранее предположений то, которое наилучшим образом соответствует реальным выборкам. Это связано с тем, что наряду с анализом информативности терминов русского и английского языков необходимо исследовать еще ряд факторов, оказывающих существенное влияние на точность разрабатываемых профильных методов, таких как способ отбора термина (общие для обоих профилей или нет), способ присваивания веса (наибольшее или усредненное значение), учет местоположения терминов.

Итак, нами получена группа практически равноточных методов, основанных на различных принципах построения профиля класса и способных обучаться на русскоязычных, англоязычных и смешанных выборках, максимально используя всю имеющуюся в исходных данных информацию. Это позволяет рассматривать их в качестве кандидатов для построения коллективов решающих правил. При объединении в коллектив мы ожидаем, что профильные процедуры, основанные на разных (статистических, теоретико-информационных, эвристических) принципах принятия решений, будут «исправлять» ошибки друг друга, увеличивая результирующую точность классификации [12].

Сформированы и экспериментально проверены шесть следующих КРП, состоящих из разного количества членов (в скобках указаны входящие в состав классификаторы).

1. КРП1 (*PO*, *НМИ*, *C-C*), в который было включено три наиболее разнородных классификатора: статистический *PO*-профиль, теоретико-информационный нормированный *МИ*-профиль и эвристический *C-C*-профиль. Обучение проводилось на смешанных выборках, длина профиля $L = 200$.

2. КРП2 (*PO*, *НМИ*, *C-C*, *UNI2*, *UNI5*), представляющий собой КРП1, расширенный за счет включения *UNI2*- и *UNI5*-профилей. Обучение проводилось на смешанных выборках, длина профиля $L = 200$.

3. КРП3 ($НМИ_{рус}$, $PO_{рус}$, $C\text{-}C_{рус}$), состоящий из тех же трех классификаторов, что и КРП1, однако профильные методы обучались только на русском языке, длина профиля $L = 200$.

4. КРП4 ($НМИ_{рус}$, $PO_{рус}$, $C\text{-}C_{рус}$, *UNI2*, *UNI5*), включающий те же три классификатора, что и КРП2, за исключением того, что *PO*-, *НМИ*- и *C-C*-профили обучались на русской выборке, а *UNI2* и *UNI5* — на смешанной.

5. КРП5 (*НМИ*, *PO*, *C-C*, *UNI2*, метод К-БС), состоящий из наиболее точных профильных методов, дополненных точным классическим методом К-БС для достижения наибольшей разнородности. Все методы обучались на смешанной выборке. Для профильных методов длина профиля $L = 200$, размер русско-английского словаря для К-БС $M = 650$, использовалась косинусоидальная мера близости и пять ближайших соседей.

6. КРП6 (*PO*, *UNI1*, *UNI2*, *UNI5*, *UNI6*), составленный из *PO*-профиля и синтезированных методов. Обучение проводилось на смешанных выборках, длина профиля $L = 200$.

В табл. 3 приведены результаты классификации различными коллективами решающих правил и приведено их сравнение с наиболее точным *PO*-профилем.

Как показали проведенные исследования, КРП, состоящие из трех членов, не позволили обеспечить

Таблица 2. Ошибки классификации методов

Метод классификации	Средние ошибки классификации, %	Минимальное значение ошибки классификации, %	Максимальное значение ошибки классификации, %
PO	10,47	5,71	15,23
НМИ	11,54	9,52	14,28
C-C	13,93	9,52	19,04
UNI1	11,03	6,67	14,28
UNI2	10,23	6,67	12,37
UNI5	11,59	5,71	16,19
UNI6	11,31	4,76	17,14
К-БС	12,19	7,61	15,23

Таблица 3. Результаты классификации для коллективов решающих правил

Классификатор	Средние ошибки классификации, %	Минимальное значение ошибки классификации, %	Максимальное значение ошибки классификации, %
PO-профиль	10,47	5,71	15,23
КРП1(*PO*, *НМИ*, *C-C*)	11,25	7,61	15,23
КРП2(*PO*, *НМИ*, *C-C*, *UNI2*, *UNI5*)	9,61	5,71	13,33
КРП3 ($НМИ_{рус}$, $PO_{рус}$, $C\text{-}C_{рус}$)	11,94	8,57	17,14
КРП4 ($НМИ_{рус}$, $PO_{рус}$, $C\text{-}C_{рус}$, *UNI2*, *UNI5*)	10,90	6,67	13,33
КРП5(*PO*, *НМИ*, *C-C*, *UNI2*, К-БС)	9,17	5,71	12,38
КРП6 (*PO*, *UNI1*, *UNI2*, *UNI5*, *UNI6*)	9,87	5,71	13,33

более высокую точность, чем наилучший индивидуальный метод (*PO*-профиль). Это объясняется тем, что члены коллектива не обладают существенной разнородностью и их ошибки заметно различаются. К сожалению, не удалось также добиться более высокой разнородности методов и результирующей точности за счет обучения на различных (русскоязычных, англоязычных и смешанных) выборках.

Вместе с тем увеличение размера КРП до пяти методов обеспечивает прирост точности по сравнению с КРП, состоящими из трех методов, и *PO*-профилем, являющимся одним из наиболее точных индивидуальных классификаторов. Однако этот прирост значительно меньше теоретически ожидаемого, который можно было бы получить в случае применения полностью независимых (разнородных) и равноточных процедур.

Таким образом, разработаны и исследованы модификации профильных методов применительно к решению задачи классификации двуязычных (русско-английских) библиографических описаний. Эти процедуры используют различные эвристики при отборе и расчете весов информативных терминов и позволяют обеспечить высокую точность классификации.

Расширение семейства профильных методов за счет разработки модификаций позволило сформировать и исследовать новые коллективы решающих правил. Эксперименты показали, что объединение различных профильных методов в КРП чаще всего приводит к снижению ошибки классификации.

Как представляется, рассмотренные в работе подходы к построению профильных методов целесообразно дополнить следующими исследованиями:

разработать формулы взвешивания русско- и англоязычных терминов в зависимости от местоположения в документе (название, аннотация, ключевые слова);

проверить эффективность использования наряду с общими словами, которые имеются сразу в двух профилях, специфических терминов, присутствующих только в одном профиле (например, в *НМИ*-профиле, в котором редкие, но важные для предметной области слова получают наибольший вес);

провести дополнительную оценку разнородности методов, включенных в КРП (например, с помощью статистики Юла (*Q*-статистики) [12]);

использовать непараметрические статистические критерии для анализа результатов экспериментов и сопоставления точности классификаторов [13].

ЛИТЕРАТУРА

1. **Растригин Л. А., Эренштейн Р. Х.** Метод коллективного распознавания. — М.: Энергоиздат, 1981. — 79 с.
2. **Абусев Р. А., Лумельский Я. П.** Статистическая групповая классификация: Учебное пособие. — Пермь: ПГУ, 1987. — 92 с.
3. **Ruta D., Gabrys B.** A Theoretical Analysis of the Limits of Majority Voting Errors for Multiple Classifier Systems / Pattern Analysis and Applications. 2002. N 5. P. 333 – 350.
4. **Толчеев В. О.** Синтез коллективов решающих правил для проведения классификации текстовых документов / Информационные технологии. 2007. № 10. С. 32 – 38.
5. **Yang Y., Liu X.** A Re-Examination of Text Categorization Methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999. P. 42 – 49.
6. **Толчеев В. О.** Модели и методы классификации текстовой информации / Информационные технологии. 2004. № 5. С. 6 – 14.
7. **Сэлтон Г.** Автоматическая обработка, хранение и поиск информации. — М.: Советское радио, 1973. — 560 с.
8. **Aas K., Eikvil L.** Text Categorization: A Survey. — Oslo: Norwegian Computing Center, 1999. P. 1 – 37.
9. **Мохов А. С., Толчеев В. О.** Разработка методов высокоточной классификации двуязычных текстовых библиографических документов / Информационные технологии. 2014. № 5. С. 8 – 13.
10. **Толчеев В. О.** Основы теории классификации многомерных наблюдений. Учебное пособие. — М.: МЭИ, 2012. — 121 с.
11. **Мохов А. С., Толчеев В. О.** Разработка профильных методов классификации двуязычных текстовых документов. Материалы 6-й Всероссийской мультиконференции по проблемам управления — МКПУ-2013. — Дивноморское, 2013. Т. 1. С. 75 – 79.
12. **Kuncheva L. I., Whitaker C. J.** Measures of Diversity in Classifiers Ensembles and Their Relationship with the Ensemble Accuracy / Machine Learning. 2003. N 51. P. 181 – 207.
13. **Орлов А. И., Толчеев В. О.** Об использовании непараметрических статистических критериев для оценки точности классификации / Заводская лаборатория. Диагностика материалов. 2011. Т. 77. № 3. С. 58 – 66.