Методы классификации в условиях противоречивого обучающего множества

© А.Ю. Колесов

Факультет Информатики и вычислительной техники
ЯрГУ им. П.Г. Демидова
kolesov.ay@ya.ru

Аннотация

В работе рассматривается один из подходов улучшения качества систем автоматической классификации в условиях неполного обучающего множества, основанный на модификации обучающей выборки. Рассмотрены несколько таких методов, которые используют метод Soft-supervised learning и метод k взвешенных ближайших соседей. Описаны эксперименты, показывающие применимость данного подхода к реальным задачам категоризации текстов с большим количеством рубрик.

1. Введение

Для структурирования больших объемов данных часто применяют рубрикацию. Для решения этой задачи предложено большое количество методов. Тем не менее, качество классификации для многих задач является низким (менее 50% полноты и точности). Для коллекций с большим количеством текстовых документов обычно задается большое число классов-рубрик, имеющих пересечения, т.е. требуется решать задачу мультиклассовой классификации (один объект может относиться к нескольким рубрикам). Это усложняет и делает более дорогим процесс создания обучающей выборки. Для того, чтобы среди большого числа рубрик отметить полный набор классов, релевантных документу, требуется много трудозатрат одного эксперта, либо участие нескольких экспертов. Как показывает практика, не всегда эти условия являются приемлемыми и выполняются. Поэтому для реальных задач зачастую документы из обучающего множества имеют неполный набор отмеченных категорий. Отсутствие у документа из обучающей выборки метки класса A, который является релевантным данному документу, приводит к тому, что при обучении он попадает в число отрицательных примеров для класса A. Это искажает решающее правило, что отрицательно влияет на качество классификации.

В данной работе предложен подход к решению задачи классификации в условиях неполного набора меток объектов из обучения, основанный на модификации обучающего множества. Этот способ позволяет уменьшить противоречия в ней. На реальной коллекции данных показано, что этот подход повышает эффективность применения одного из самых эффективных методов машинного обучения (SVM). Указаны условия применимости данного метода.

2. Предлагаемый подход

2.1 Методы модификации обучающей выборки

Существует несколько подходов к модификации обучающей выборки. Самые известные из них удаление выбросов (см. [11]) и отбор эталонных объектов [2]. Отметим, что применимость этих подходов к реальным задачам классификации текстов в настоящее время недостаточно исследована.

Алгоритмы, предложенные в данной работе, основываются на трех предположениях относительно данных, которые подаются на вход классификатору. Во-первых, утверждается, что существуют объекты из обучающего множества, составленного экспертами, имеющие неполный набор меток, и их достаточно много. Полный набор меток определим как некий идеальный консенсус экспертов, заключающийся в том, что для всех документов коллекции больше нельзя ни добавить, ни удалить ни одной метки-рубрики. Во-вторых, предполагается, что эксперты не ошибаются: любая метка, которую эксперт проставил для документа, правильна. Т.е. в процессе создания обучающей выборки возможны только ошибки 1 рода (при проверке гипотезы «объект d принадлежит рубрике c»). В-третьих, исходим из гипотезы компактности, которая гласит: «Схожие объекты гораздо чаще находятся в одном классе, чем в разных; или, другими словами, что классы образуют компактно локализованные подмножества в пространстве объектов». Отсюда ясно, что для решения задачи

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

классификации в указанных предположениях для каждой рубрики алгоритму требуется найти те документы, которые, исходя из геометрии данных, релевантны рубрике. Т.к. алгоритмы, основанные, на вышеперечисленных подходах не предусматривают таких предположений, то необходим другой подход, который и описан ниже.

Для первоначальных экспериментов были исследованы следующие два алгоритма:

1) на базе одного из новых методов частичного обучения ([9]), назовем его SoftSL;
2) на основе k взвешенных ближайших соседей.

Оба алгоритма используют такую структуру, как множество ближайших соседей для документа, что соответствует третьему предположению.

Схема метода модификации следующая. На первом шаге с помощью одного из указанных алгоритмов получаем возможные пары релевантности документ-рубрика, которые не отметили эксперты. Обозначим это множество так: $probably\_couples = PC = \{(d,c) \mid \psi(d,c) = 1\}$, где d – документ, c – рубрика, ψ – соответствующий алгоритм. Далее выбираем способ использования множества PC. Рассмотрим два варианта:

1) добавляем множество PC в обучение;
2) для пар (d,c) из PC – при обучении для рубрики c документ d не попадет в отрицательные примеры для этой рубрики.

Т.о. оставляем в обучающей выборке те объекты, которые согласно алгоритму ψ не принадлежат выборке, если их отметил эксперт, т.к. предполагаем, что эксперт не ошибается.

Далее опишем алгоритмы обнаружения недостающих меток для документов.

2.2 Поиск недостающих меток на основе алгоритма SoftSL

В статье [9] описан новый графовый метод частичного обучения (semi-supervised learning), основанный на минимизации расстояний Кульбака-Лейблера [12] между распределениями вероятностей принадлежности классам, определенными для каждого документа. В дальнейшем будем называть его Soft-supervised learning method (SoftSL). Одним из достоинств алгоритма является то, что его можно напрямую применять к мультиклассовой задаче. В [10] показано превосходство SoftSL перед другими современными методами.

Кратко опишем идею алгоритма. Пусть имеется множество, состоящее из размеченных и неразмеченных объектов $D = \{D_l, D_u\}$. Здесь $D_l = \{(x_i, y_i)\}_{i=1}^{l}$, $D_u = \{x_i\}_{i=l+1}^{n}$, где x_i – входные векторы, соответствующие объектам для классификации, y_i – метки классов-категорий. Определим ненаправленный граф $G = (V, E)$ с весами, где $V = \{1, \ldots, n\}$, n – количество объектов в множестве $E = V \times V$. Обозначим $w_{ij} \in W$ – вес ребра между объектами i и j. Вес ребра определяется так: $w_{ij} = sim(x_i, x_j)\delta(j \in K(i))$, где $K(i)$ – множество k ближайших соседей объекта x_i.

Для каждого объекта $d_i \in D$ сопоставляется набор вероятностей принадлежности каждому классу $p_i = (p_i^t)_{t=1}^{m}$, где m – количество классов. Для каждого из размеченных объектов имеется также известный набор вероятностей $r_i = (r_i^t)_{t=1}^{m}$ – то, что разметили эксперты.

Задача сводится к минимизации функционала по наборам вероятностей p:

$$\min_p C_1(p), \text{ where } C_1(p) = \sum_{i=1}^{l} D_{KL}(r_i \parallel p_i) + \\ + \mu \sum_{i=1}^{n} \sum_{j \in K(i)} w_{ij} D_{KL}(p_i \parallel p_j) - \nu \sum_{i=1}^{n} H(p_i)$$

$D_{KL}(p_i \parallel p_j)$ – расстояние Кульбака-Лейблера. $H(p_i)$ – энтропия.

Первый член отвечает за то, чтобы результирующие наборы вероятностей не сильно уклонялись от вероятностей, заданных экспертами. Второй член позволяет учитывать геометрию графа, т.е. близкие (по мере близости) объекты должны иметь сходные распределения вероятностей по рубрикам. Третий член – регуляризация, приближает распределения по рубрикам к равновероятным в случае, если другие члены формулы не предпочитают обратного. Численно задача решается методом alternating minimization (AM). Подробнее см. [9].

При использовании алгоритма для рассматриваемой задачи неразмеченных данных D_u нет. После минимизации функционала $\min_p C_1(p)$ мы получаем для каждого документа $d_i \in D$ набор вероятностей $p_i = (p_i^1, \ldots, p_i^m)$. Выбрав правило определения принадлежности документа рубрике, например, порог $T \in [0,1]$, можно выделить рубрики, релевантные документу согласно алгоритму. В случае порога: документ $d_i \in c_j$, если $p_i^j \geq T$.

2.3 Поиск недостающих меток на базе алгоритма weighted-kNN

Алгоритм k-взвешенных ближайших соседей [8] также напрямую позволяет решать мультиклассовую задачу. Кратко опишем, как он работает.

Пусть определена функция расстояния $\rho(d,d')$ между документами. Определим функцию принадлежности документа d рубрике c:

$$S(d,c)=\frac{\sum_{d'\in KNN(d)}\rho(d,d')\phi(d',c)}{\sum_{d'\in KNN(d)}\rho(d,d')},$$

где $KNN(d)$ – множество k ближайших соседей в обучающей выборке для документа d, $\phi(d',c)=0$, если $d'\notin c$, $\phi(d',c)=1$, если $d'\in c$.

Вводится порог T: если $S(d,c)\geq T$, то документ $d\in c$.

Алгоритм состоит в подсчете величины $S(d,c)$ для всевозможных (d,c). Все пары, для которых $S(d,c)\geq T$, определяем как недостающие метки. На основе этих пар документ-рубрика можно модифицировать обучающую выборку.

2.4 Схема предлагаемого алгоритма

Опишем схему работы предлагаемого алгоритма. В качестве входных данных имеем обучающее множество (пары документ-рубрика). Выполняем следующие шаги:

1) модификация обучающей выборки (см. 2.1);
2) применение метода обучения к обновленной обучающей выборке.

На выходе получается решающее правило $a(x)$ для рубрикации новых объектов. Например, при подходе one-vs-rest для алгоритма SVM $a(x)=(a_1(x),\ldots,a_m(x))$, где m – количество рубрик. Далее, с помощью этого решающего правила можно делать предсказания классов для новых объектов.

3. Эксперименты

Эксперименты проводились с использованием вычислительных ресурсов суперкомпьютера СКИФ МГУ «Чебышев», а также кластера ЯрГУ.

Данные для воспроизведения экспериментов на Agingportfolio можно получить в разделе wiki на сайте [4] или по запросу у автора статьи.

Эксперименты проводились на двух описанных ниже коллекциях.

3.1 Данные о научных проектах Agingportfolio

Для проведения экспериментов использовались данные о научных проектах Agingportfolio [4]. Система содержит базу данных проектов, связанных со старением и финансируемых Национальным институтом здоровья (NIH) и Европейской комиссии (EC CORDIS). В настоящее время в базе Agingportfolio имеется более 1 млн. 100 тыс. проектов, на которые были выделены гранты. По проектам имеется следующая информация: авторы, название, краткое описание мотивации и целей исследований, организация, период, на который получен грант. Для некоторых проектов были даны тэги – термины, характеризующие проект. Вся информация на английском языке. Средняя длина описания проекта – 100 слов. В экспериментах использовались только поля с названием, кратким описанием и тегами.

Имеется таксономия категорий, состоящая из 335 рубрик на 6 уровнях иерархии. Подробно с ней можно ознакомиться на сайте [4]. Тренировочное и тестовое множества состоят из вручную размеченных по категориям этой таксономии документов. Для тестового множества разметка производилась особым образом. Для каждого документа имелось два набора категорий, составленных разными экспертами. В качестве меток взято объединение этих наборов. Это позволяет получить более полный набор категорий. Обучающее множество составлено с меньшим контролем, разными людьми, в том числе пользователями ресурса Agingportfolio. Как показывает визуальный анализ, в обучающем множестве довольно большое количество проектов имеет неполный набор категорий. Об этом свидетельствует, например, следующий факт. Среднее число рубрик на проект в обучающей выборке составляет $4,36$, а на тестовой – $9,79$.

Тестовое множество состояло из 750 проектов. Обучающее множество – из 3144 проектов

3.2 Коллекция ROMIP Legal2007

Также описанный подход был исследован на коллекции ROMIP Legal2007, которая содержит 348410 документов Законодательства Российской Федерации, Москвы и Санкт-Петербурга. Множество рубрик, по которым требовалось выполнить классификацию, состоит из 726 рубрик 2-4 уровня, являющихся подмножеством большого иерархического рубрикатора нормативных документов. Для каждой рубрики предоставлено некоторое множество примеров документов, относящихся к рубрике. Минимальное количество примеров для рубрики – 50, максимальное – 615. Из 726 рубрик для оценки организаторами выбраны 75 случайных рубрик. Тестовое множество состояло из 29261 документов, обучающее множество – из 22195 документов.

3.3 Предобработка данных

Для обеих коллекций использовалась векторная модель представления текстов. Признаками являлись слова, приведенные к нормальной форме.

Веса признакам присваивались согласно правилу TFIDF в формулировке INQUERY (см. [3]). Вектора, представляющие документы, приводились к норме 1 в евклидовой метрике n-мерного пространства.

3.4 Алгоритм обучения

В качестве алгоритма классификации использовался линейный SVM [13], который по сути является двуклассовым классификатором. При наличии более двух классов используются различные способы для мультиклассовой категоризации на базе SVM. Одним из традиционных путей реализации мультиклассовой классификации является подход one-vs-rest (см. [8]). Он и был использован в наших экспериментах. Т.е. для каждой рубрики c строилось свое решающее правило: $w_c x + b_c > 0$

Программа Liblinear [5][6] была выбрана в качестве реализации SVM.

3.5 Подбор параметров SVM

Для каждого способа векторного представления документов подбираются следующие параметры:

- параметр C (характеризует компромисс между максимизацией разделяющей полосы и минимизацией суммарной ошибки, см. работу [7]);
- параметр b (порог классификации в решающем правиле, см. также [3]).

Подбор параметров осуществляется методом скользящего контроля с разбиением множества на 5 частей согласно следующей стратегии. Параметр C подбираем по сетке и сразу же подбираем параметр b для каждого C из сетки и каждой рубрики, оптимальной считается пара параметров $(C,(b_i))$, доставляющая максимум F-мере при макроусреднении по всем рубрикам (C – число, единое для всех рубрик, (b_i) – набор параметров b для каждой рубрики, оптимальных для параметра C).

В силу несбалансированности обучающей выборки на коллекции Legal2007 (см. [1]) подбор параметра b приводит к переобучению и ухудшает качество классификации. Поэтому для нее подбирался только параметр C, общий для всего обучающего множества и доставляющий максимум при макроусреднении по всем рубрикам.

3.6 Модификация обучающего множества

Модификация обучающего множества производилась методами на основе SoftSL (так же обозначено в таблице) и k взвешенных ближайших соседей (обозначение w-kNN). Стратегию добавления множества PC в обучающую выборку обозначим «add+». Стратегию удаления документа d из отрицательных примеров класса c, если пара (d,c) содержится в множестве PC, обозначим «del+».

Параметры k, T, μ выбирались по сетке следующими способами.

Для коллекции Agingportfolio при подборе использовалось множество из 200 документов (обозначим его D), которое было получено таким же способом, как и тестовое множество. На всей обучающей выборке обучались, на D тестировались.

Для коллекции Legal2007 обучающее множество разбивалась на две части 70%+30%, на 70% обучались, на 30% тестировались.

Набор параметров считался оптимальным, если он максимизировал F_1-меру при макроусреднении. Параметры выбирались для каждого способа модификации обучающей выборки. Для SoftSL параметр μ подбирался уже для оптимальных k и T, подобранных для метода w-kNN. Параметр ν для SoftSL был взят равным 0, поскольку все документы в экспериментах имели метки и регуляризации не требовалось. Порог принадлежности рубрике для модификации с помощью SoftSL был фиксирован и различен для разных коллекций: для Agingportfolio $T = 0,005$, для Legal2007 $T = 0,01$.

3.7 Базовые алгоритмы

Для того, чтобы судить об улучшении/ухудшении качества классификации, опишем базовые алгоритмы, с результатами которых будем сравнивать эффективность других методов. Первый из них это SVM с параметрами по умолчанию (обозначим noFit-SVM), второй – SVM с подбором параметров (обозначим SVM). Таким образом, можно будет сравнить, какой вклад в улучшение качества классификации вносит подбор параметров, и насколько улучшает результаты модификация обучающего множества каждый из рассматриваемых способов.

Кроме того, для коллекции Legal2007 произведено сравнение с лучшими результатами участников РОМИП’2007.

3.8 Рассматриваемые алгоритмы

Согласно схеме из 2.4 эксперименты с модификацией обучающей выборки производились следующим образом. В соответствии с заданным алгоритмом (SoftSL, w-kNN) подбирались оптимальные параметры для них, после чего находилось множество PC. Затем применялась заданная («add+», «del+») стратегия использования PC. На полученном обучающем множестве запускался SVM (Liblinear) с подбором параметров (см. 3.5). Далее полученное алгоритмом SVM решающее правило применялось к тестовому

множеству и подсчитывались метрики качества (полнота, точность, F_1-мера; см. [8]).

4. Результаты и выводы

В этом разделе приведем таблицы с результатами экспериментов, с некоторыми данными для их анализа, а также сделаем выводы по каждой таблице.

	macro_f	micro_f
noFit-SVM	2,02%	22,64%
SVM	14,61%	28,47%
del+w-kNN	16,76%	24,34%
add+w-kNN	20,15%	39,43%
del+SoftSL	15,09%	20,37%
add+SoftSL	16,11%	31,13%

Таблица 1. Результаты на Agingportfolio

В Таблице 1 приведены значения F_1-меры для базовых алгоритмов и при применении SVM с подбором параметров на модифицированной обучающей выборке коллекции Agingportfolio (обозначения способов модификации см. выше). Как видно из Таблицы 1:

- Подбор параметров SVM значительно повышает F_1-меру.
- Модификация обучающего множества улучшает качество классификации.
- Результаты при модификации «добавление релевантных документов в обучающее множество» превосходит способ «удаление из отрицательных примеров релевантных документов». Этот факт также подтверждает, что предположение неполноты наборов меток для документов в обучающем множестве верно. Т.е. в исходной обучающей выборке объекты, очень важные для построения решающих правил для рубрик, имеют противоположную релевантность.
- Модификация обучающего множества на базе w-kNN работает значительно лучше.

	macro_f	micro_f
ROMIP_best	45,09%	48,34%
noFit-SVM	32,38%	40,61%
SVM	51,48%	55,25%
del+w-kNN	51,43%	55,60%
add+w-kNN	50,89%	54,31%
del+SoftSL	46,89%	51,74%
add+SoftSL	46,75%	50,71%

Таблица 2. Результаты на Legal2007

В Таблице 2 отражены результаты аналогичных экспериментов на коллекции Legal2007. По таблице можно сделать следующие выводы:

- Подбор параметра C алгоритма SVM значительно повышает качество классификации.
- Модификация обучающего множества не улучшает качество рубрикации.
- Модификация по принципу «удаление из отрицательных примеров релевантных документов» дает результаты, несколько лучшие, чем стратегия «добавление релевантных документов в обучающее множество».
- Лучшие результаты участников РОМИП'2007 превзойдены только благодаря подбору параметра C.
- Результаты модификации обучающего множества на базе w-kNN превосходит результаты, которые дает модификация на основе SoftSL.

Для коллекции Legal2007 модификация обучающей выборки не приносит улучшения качества классификации по одной из двух причин. Или документы коллекции хорошо выверены и размечены достаточно полным набором категорий, или улучшения нет, поскольку и тестовое и обучающее множества размечены одинаково неполно.

В Таблицах 3 и 4 содержится следующая информация об обучающих множествах: среднее количество рубрик на документ (avg_rubr_cnt) и среднее количество документов на рубрику (avg_doc_cnt). Эти сведения даны для исходного обучающего множества (обозначение no_add_modif; то же количество сохраняется при отсутствии добавления документов в обучение после модификации), а также для модифицированных обучающих множеств.

	avg_rubr_cnt	avg_doc_cnt
no_add_modif	4,36	44,35
add+w-kNN	11,86	120,62
add+SoftSL	10,78	111,05

Таблица 3. Среднее количество рубрик на документ и документов на рубрику в различных обучающих множествах. Коллекция Agingportfolio.

По таблице 3 можно сказать следующее:

- При использовании принципа «добавление релевантных документов в обучающее множество» количество положительных примеров для каждой рубрики значительно возрастает. Вероятно, это и приводит к увеличению F_1-меры (см. Таблицу 1).
- При использовании алгоритма SoftSL количество добавленных в обучающее множество пар документ-рубрика несколько меньше, чем при использовании алгоритма w-kNN, как и результаты по F_1-мере (см. Таблицу 1).

- Значение среднего количества рубрик на документ на обновленном обучающем множестве больше, чем на тесте (9,79 против 11,86 и 10,78). Это может свидетельствовать о том, что, возможно, есть потенциал для увеличения качества классификации даже с использованием рассмотренных в этой работе методов.

	avg_rubr_cnt	avg_doc_cnt
no_add_modif	3,19	129,64
add+w-kNN	3,21	130,31
add+SoftSL	3,53	143,41

Таблица 4. Среднее количество рубрик на документ и документов на рубрику в различных обучающих множествах. Коллекция Legal2007.

По таблице 4 отметим следующее:

- Обучающая выборка изменяется незначительно после модификации.
- Алгоритм SoftSL находит большее количество релевантных пар документ-рубрика, но для этого способа модификации получаются более низкие результаты (см. Таблицу 2). Это может свидетельствовать о том, что или в этом случае данные в обучающей выборке становятся более рассогласованными с тестовой выборкой (что очень вероятно, т.к. методика ручной классификации для Legal2007 была одна и та же), или среди добавляемых пар достаточно много ложно релевантных.

Итак, эксперименты показывают, что предположения относительно обучающей выборки коллекции Agingportfolio были верны. Модификация дает довольно много дополнительных меток. Благодаря этому для каждой рубрики растет число положительных примеров при обучении. В итоге наблюдается улучшение качества классификации (по F_1-мере на 38%). Отметим также, что полнота и точность после модификации обучающего множества становятся боле сбалансированными (см. Таблицу 5).

	micro_recall	micro_prec
noFit-SVM	13,00%	87,77%
SVM	22,86%	37,73%
del+w-kNN	50,08%	16,07%
add+w-kNN	39,39%	39,46%
del+SoftSL	39,65%	25,62%
add+SoftSL	46,90%	13,01%

Таблица 5. Полнота и точность результатов классификации. Коллекция Agingportfolio.

	micro_recall	micro_prec
ROMIP_best	36,68%	70,86%
noFit-SVM	29,14%	66,95%
SVM	54,38%	56,14%
del+w-kNN	58,00%	53,39%
add+w-kNN	57,63%	51,35%
del+SoftSL	61,93%	42,93%

Таблица 6. Полнота и точность результатов классификации. Коллекция Legal2007.

Сделаем также выводы относительно методики экспериментов и данных, к которым применим рассматриваемый подход. Как видно по таблицам, на коллекции Legal2007 предлагаемый подход не дает улучшения качества классификации. На наш взгляд, это связано, в первую очередь, с тем, что для разметки и тестового и обучающего множеств использовалась одна методика.

Если в обучающем множестве документы имеют неполные наборы рубрик, то и на тесте документы имеют неполные наборы рубрик. Соответственно, при увеличении «реальной» полноты на обучении предсказание на тестовых данных имеет большую «реальную» полноту, а поскольку это не согласуется с разметкой тестового множества, то при подсчете метрик снижается точность. Проще говоря, если в тестовых данных у документа d пропущена метка c, а классификатор верно отнес документ к рубрике c, то это приведет к снижению точности при подсчете метрик. Поэтому для адекватной оценки качества классификации объекты на тесте должны быть более полно размечены, желательно участие нескольких экспертов в разметке каждого документа. Фраза «реальная» полнота характеризует степень соответствия меток документов из множества полным наборам меток (определение в разделе 2.1) для этих документов. Чем больше средняя мощность пересечений множеств меток для документов с соответствующими полными наборами меток, тем больше «реальная» полнота, и наоборот. Проще говоря, чем меньше пар документ-рубрика после просмотра множества документов группа экспертов сможет добавить к обучающему множеству, тем выше «реальная» полнота.

Если же, наоборот, и тестовое, и тренировочное множества хорошо размечены, то опять же после модификации обучающего множества получим рассогласование в «реальной» полноте. Поэтому на данных, которые размечены полно, не имеет смысла применять этот метод.

5. Заключение

В данной работе представлен подход для повышения эффективности применения алгоритма SVM в задаче классификации в условиях неполного набора меток объектов из обучения. Идея заключается в модификации обучающего множества. Предложены несколько способов модификации.

Эффективность подхода показана на задаче классификации научных грантов с большим числом рубрик. Метод на основе w-kNN и стратегии «добавление релевантных документов в обучающее

множество» дает улучшение качества классификации по сравнению с базовым (SVM) по F_1 -мере на 38% и при макроусреднении, и при микроусреднении. Другие алгоритмы также улучшают F_1 -меру, но не столь значительно.

На коллекции Legal2007 модификация обучающего множества не дает улучшения F_1 -меры. Указаны возможные причины такого результата. Попутно путем подбора параметра C алгоритма SVM на этой коллекции получено улучшение по F_1 -мере на 14% и при макроусреднении, и при микроусреднении по сравнению с лучшими результатами участников РОМИП'2007.

В дальнейшем планируется опробовать другие методы поиска противоречий в обучающей выборке, исследовать влияние модификации обучающей выборки на качество классификации при использовании различных алгоритмов категоризации, а также проверить методы на новых коллекциях.

Литература

[1] М.С. Агеев, Б.В. Добров, Н.В. Лукашевич, С.В. Штернов. УИС РОССИЯ в РОМИП 2008: поиск и классификация нормативных документов. РОМИП'2008

[2] И. А. Борисова, В. В. Дюбанов, Н. Г. Загоруйко, О. А. Кутненко Сходство и компактность // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 89–92.

[3] А.Ю. Колесов Влияние векторного представления на качество классификации документов. Труды Третьей Российской конференции молодых ученых по информационному поиску. Петрозаводский государственный университет. 11-16 сентября 2009 г.

[4] AgingPortfolio Web site, 2011. - http://agingportfolio.org/

[5] Chih-Jen Lin's Home Page. - http://www.csie.ntu.edu.tw/~cjlin/index.html

[6] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R.Wang, C.J. Lin (2008). LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9: 1871-1874.

[7] T.Joachims Text Categorization with Support Vector Machines: Learning with Many Relevant Features. // Proceedings of ECML-98, 10th European Conference on Machine Learning — 1998.

[8] C.D. Manning, P. Raghavan, H. Schutze (2009). An Introduction to Information Retrieval, Cambridge, England: Cambridge University Press.

[9] A. Subramanya and J. Bilmes, "Soft-supervised text classification", in EMNLP, 2008.

[10] A. Subramanya and J. Bilmes, "Entropic graph regularization in non-parametric semi-supervised classification," in NIPS, 2009

[11] J. Tang, Z. Chen, A. Fu, and D. Cheung, Capabilities of outlier detection formulation schemes, framework and methodologies, Knowledge and Information Systems, Vol. 11(1) 45-84, January 2007. Springer.

[12] S. Kullback, R.A. Leibler On information and sufficiency. Annals of Mathematical Statistics 22: 79–86.

[13] V.Vapnik The Nature of Statistical Learning Theory. — Springer-Verlag — New York, 1995.

Classification Methods with an Inconsistent Learning Set

© Anton Y. Kolesov

In this paper we consider an approach to improve the performance of classification with incomplete training set. This approach is based on the process of modifying the training set. We consider several such methods. They use two different methods: soft-supervised learning method and weighted k-nearest neighbors. Experiments show the applicability of this approach to a real text categorization problem with many classes.