RCO на РОМИП 2006

© Авторы

Поляков П.Ю., Плешко В.В.
rco@metric.ru

Аннотация

Настоящая работа является отчетом об экспериментах, проведенных авторами в рамках цикла семинара РОМИП 2006 года. Проведены исследования различных факторов, влияющих на качество тематической классификации методом опорных векторов [1]. Исследованы различные типы ядра, а также способы отбора классификационных признаков и их взвешивания.

1. Введение

В настоящей статье приводится отчет об экспериментах авторов по выполнению заданий по тематической классификации нормативно-правовых документов.

В 2005 году в работе [2] был представлен отчет об экспериментах с различными типами ядра в методе опорных векторов. Однако представленное исследование нам показалось недостаточно подробным и систематичным. Целью настоящей работы является восполнение данного пробела.

2. Классификация нормативно-правовых документов

2.1 Постановка задачи

Участникам был предложен перечень категорий первого уровня классификатора справочно-правовой системы «Кодекс» (173 категории), а в качестве обучающей выборки было отобрано подмножество документов (всего 13771), входящих в категории данного классификатора. Задача состояла в отнесении оставшейся части коллекции документов к категориям классификатора.

Каждому документу допускалось присвоить не более 5 категорий. Также допускалось не присваивать документу ни одной категории.

2.2 Общий подход

Исследования проводились в рамках векторной модели представления документов. Во всех прогонах использовался только метод опорных векторов.

Исследовались следующие параметры построения классификаторов:

1. Способ выбора терминов для профиля категории;
2. Способ вычисления весов терминов в документе;
3. Тип ядра в методе опорных векторов.

2.3 Отбор терминов

Отбор терминов проводился из набора положительных примеров для каждой из категорий. Было исследовано два способа выделения терминов:

1. Однословные термины;
2. Теоретико-множественное объединение однословных и многословных терминов.

В качестве однословных терминов выделялись все слова документа. Многословные термины выделялись при помощи алгоритма синтактико-семантического анализа [3]. Настройки алгоритма приведены в [4].

Выделенные термины подвергались фильтрации для каждой категории отдельно. Фильтрация производилась по документной частоте и по информационной значимости термина.

При фильтрации термины с документной частотой менее 3 отбрасывались из рассмотрения.

За основу информационной значимости термина был выбран коэффициент IG (information gain, см. например [5]):

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_i, \bar{t}_i\}} P(t,c) \cdot \log\left(\frac{P(t,c)}{P(t)P(c)}\right). \quad (1)$$

Нами были взяты только первое и четвертое слагаемые, характеризующие описательную способность термином рубрики. Фильтрация по признаку информационной значимости проводилась следующим образом:

1. Термины упорядочивались по убыванию информационной значимости;

2. Далее были отобраны первые N терминов, сумма информационной значимости которых составила 80% от общей суммы по всем терминам.

Заметим, что для большинства категорий величина N составила около 20% от общего числа терминов.

2.4 Взвешивание терминов

Было опробовано три способа взвешивания терминов при построении векторов документов:

1. Бинарное (1 – если термин встретился в документе, 0 – в противном случае), обозначим – «B»;
2. Частотное, пропорционально частоте термина, обозначим – «F»;
3. Информационно-поисковое, обозначим – «TFIDF».

При расчете весов терминов в документе создавался вектор, ненулевыми элементами которого служили перечисленные характеристики для всех найденных в нем терминов из множества отобранных для категории. Затем вектор документа приводился к единичной длине.

2.5 Тип ядра

Для апробации метода опорных векторов была взята его реализация SVMLight [6]. Метод чувствителен к отношению весов ошибок I и II рода (параметр j). В проводимых экспериментах было взято значение $j = 10$. Опробовались три типа ядра:

1. Линейное ядро;
2. Полиномиальное ядро;
3. Гауссово ядро.

Линейному ядру соответствует обычный линейный классификатор:

$$K(x, y) = \vec{x} \cdot \vec{y}. \quad (2)$$

Полиномиальное ядро характеризуется параметрами s, d и c:

$$K(x, y) = (s\vec{x} \cdot \vec{y} + c)^d. \quad (3)$$

Было проведено эмпирическое исследование зависимости качества результатов от перечисленных параметров. Значения каждого из параметров рассматривалось в следующих пределах: $0.001 \le s \le 10$, $2 \le d \le 5$, $0 \le c \le 10$. Оказалось, что значения параметров существенно влияют на качество. Наилучшие результаты были показаны при значениях $s = 10$, $d = 3$, $c = 10$.

Гауссово (radial-based) ядро описывается параметром g:

$$K(x, y) = \exp(-g\|\vec{x} - \vec{y}\|^2). \quad (4)$$

Значения параметра варьировалось в следующих пределах: $0.005 \le g \le 1$. Наилучшие результаты были показаны при $g = 0.2$.

Полученные выше оптимальные значения параметров ядер использовались в дальнейших экспериментах. Следует также отметить, что разброс между лучшим и худшим результатами в процессе подбора параметров составлял около 30%. Кроме того, нет гарантии, что найденные параметры дадут наилучший результат на других коллекциях.

2.6 Описание прогонов

План эксперимента составлял из двух этапов. На первом этапе исследовались выбор ядра и способ взвешивания. При этом в качестве терминов использовалось объединение однословных и многословных терминов. На этом этапе было сделано 9 прогонов (3*3).

На втором этапе исследовался эффект от расширения словаря путем добавления многословных терминов к однословным. Было сделано 3 прогона для линейного ядра.

2.7 Результаты оценки

Все прогоны предварительно проверялись на полных матрицах релевантности ideal40 (40 категорий, содержащие не менее 10 положительных примеров в обучающей выборке), использованных на семинаре РОМИП в 2004, 2005 и 2006 годах.

Сводные результаты по всем участникам в 2006 году по показателям F1(micro) и F1(macro) представлены на рисунке 1. Наши прогоны имеют номера от 15 до 26, и их расшифровка приведена в таблице 1.

Таблица 1. Расшифровка прогонов. Во второй строке указан тип взвешивания терминов, расшифровка обозначений в п.2.4.

линейное ядро			полиномиальное			гаусово ядро			Линейное ядро без многословных терминов		
TFIDF	F	B	TFIDF	F	B	TFIDF	F	B	TFIDF	F	B
15	16	17	18	19	20	21	22	23	24	25	26

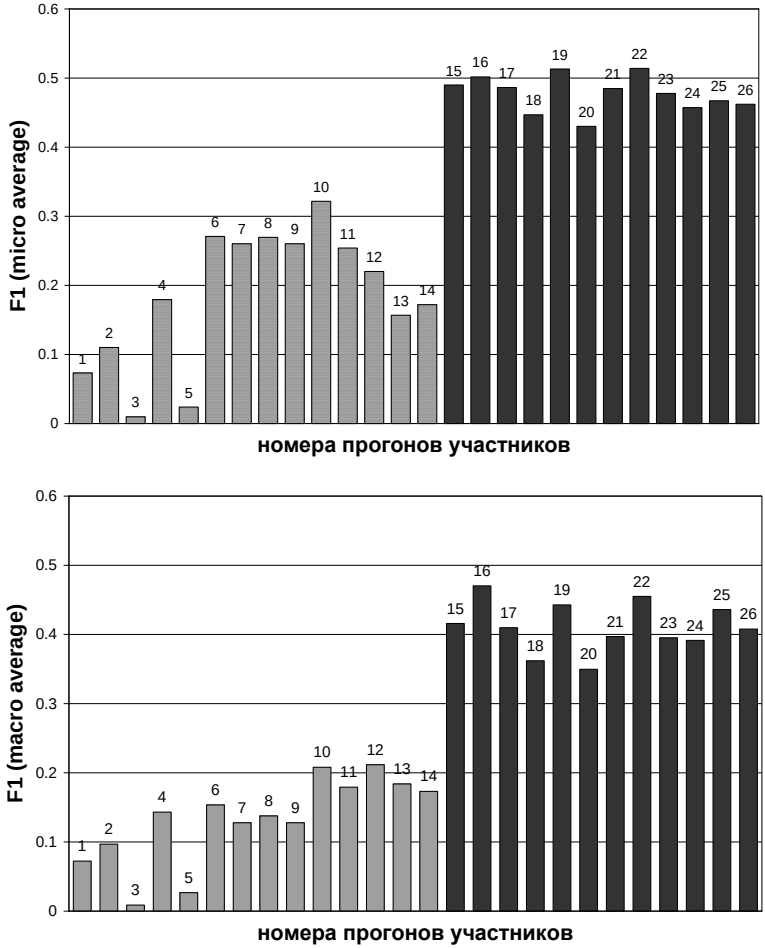

Рисунок 1. Значения F1(micro) и F1(macro) для участников дорожки классификации нормативно-правовых документов по ideal40.

В таблицах 2 и 3 приведены результаты оценки прогонов на матрицах релевантности 2004, 2005 и 2006 для показателей F1(micro) и F1(macro) соответственно. В последней колонке каждой из таблиц приведено наибольшее значение показателя среди прогонов, оцененных в соответствующем году.

Таблица 2. F1(micro) по матрицам релевантности 2004, 2005, 2006.

	линейное ядро			полиномиальное			гаусово ядро			лучший результ.
	TFIDF	F	B	TFIDF	F	B	TFIDF	F	B	
2004	0.574	0.563	0.569	0.547	0.598	0.557	0.577	0.585	0.577	0.467
2005	0.591	0.577	0.584	0.568	0.610	0.566	0.594	0.599	0.589	0.592
2006	0.490	0.502	0.486	0.447	0.513	0.430	0.485	0.514	0.478	0.514

Таблица 3. F1(macro) по матрицам релевантности 2004, 2005, 2006.

	линейное ядро			полиномиальное			гаусово ядро			лучший результ.
	TFIDF	F	B	TFIDF	F	B	TFIDF	F	B	
2004	0.484	0.520	0.448	0.411	0.491	0.384	0.466	0.517	0.430	0.349
2005	0.423	0.476	0.397	0.350	0.443	0.326	0.407	0.472	0.377	0.432
2006	0.416	0.470	0.410	0.362	0.443	0.350	0.397	0.455	0.395	0.470

На основании результатов, представленных в таблицах 2 и 3, можно сделать следующие предварительные выводы. Выбор ядра незначительно влияет на качество результата. Данный фактор оказывает существенно меньшее влияние чем, например, способ взвешивания. Частотное взвешивание терминов дает наилучший результат (прогоны 16, 19, 22, 25).

В силу того, что выбор ядра несильно влиял на результат, второй этап экспериментов был проведен только с линейным ядром.

В таблицах 4 и 5 сопоставлены показатели для прогонов с использованием линейного ядра с добавлением многословных терминов и без добавления.

Таблица 4. F1(micro).

	линейное ядро с многословными терминами			линейное ядро без многословных терминов		
	TFIDF	F	B	TFIDF	F	B
2004	0.574	0.563	0.569	0.530	0.530	0.528
2005	0.591	0.577	0.584	0.546	0.541	0.539
2006	0.490	0.502	0.486	0.457	0.467	0.462

Таблица 5. F1(macro).

	линейное ядро с многословными терминами			линейное ядро без многословных терминов		
	TFIDF	F	B	TFIDF	F	B
2004	0.484	0.520	0.448	0.469	0.502	0.452
2005	0.423	0.476	0.397	0.410	0.456	0.407
2006	0.416	0.470	0.410	0.391	0.436	0.408

Из таблиц 4 и 5 видно, что относительный прирост качества классификации при добавлении многословных терминов составляет около 6%.

3. Заключение

В работе показано, что линейное ядро дает результаты сравнимые по качеству с нелинейными ядрами. При этом выбор параметров нелинейных ядер существенно влияет на результат, а оптимальные значения параметров в общем случае могут зависеть от коллекции.

В наши дальнейшие планы входит продолжение исследований методов отбора и взвешивания классификационных признаков, а также апробация полученных результатов на соседних дорожках РОМИП, посвященных классификации веб-сайтов и веб-страниц, в которых мы в этом году, к сожалению, не смогли принять участие.

Литература

[1] *Burges C.J.C.* A Tutorial on Support Vector Machines for Pattern Recognition // Data Mining and Knowledge Discovery – 1998, V.2, No.2 – pp.121-167.

[2] *Федоровский А., Костин М, Проскурин А.* Mail.Ru на РОМИП-2005 // Труды третьего российского семинара РОМИП'2005. (Ярославль, 6 октября 2005г.). - Санкт-Петербург: НИИ Химии СПбГУ - 2005 - с. 106-124.

[3] *Ермаков А.Е.* Значимость элементов текста в свете теории синтаксической парадигмы // Русский язык: исторические судьбы и современность. II Международный конгресс исследователей русского языка. Труды и материалы. - Москва: МГУ - 2004.

[4] *Плешко В.В., Ермаков А.Е., Голенков В.П, Поляков П.Ю.* RCO на РОМИП 2005 // Труды третьего российского семинара

РОМИП’2005. (Ярославль, 6 октября 2005г.). - Санкт-Петербург: НИИ Химии СПбГУ - 2005 - с. 106-124.
[5] *H. Avancini, A. Lavelli, F. Sebastiani, R. Zanoli*. Automatic expansion of domain-specific lexicons by term categorization // ACM Transactions on Speech and Language Processing (TSLP) Discovery – 2006, V.3, No.1 – pp.1-30.
[6] *Joachims T.* Making large-scale support vector machine learning practical // Advances in Kernel Methods: Support Vector Machines / B.Scholkopf, C. Burges, A. Smola (eds.) - MIT Press: Cambridge, MA" – 1998.

RCO at RIRES 2006

Polyakov P.Yu., Pleshko V.V.

This article presents report on experiments in IR that were driven as a part of RIRES seminar. The research was taken on different factors that affect quality of SVM method for document classification task. Factors that were under consideration are kernel type, feature selection and weighting methods.