КС-классификатор и дорожка классификации веб-сайтов РОМИП'2010.

© Маслов М. Ю., Пяллинг А.А.

Яндекс
{maslov, pyal}@yandex-team.ru

Аннотация

В настоящей работе исследуются особенности метрик РОМИП для дорожки классификации веб-сайтов. Проводится оценка качества работы КС-классификатора [1] в сравнении с алгоритмами других участников дорожки. Предлагается новый метод автоматического отбора терминов.

1. Введение

Алгоритм [1], применяемый для автоматической классификации сайтов в Яндексе, участвует в дорожке классификации веб-сайтов РОМИП впервые. Свойства метрик рассматриваются в подразделах Введения и в разделе 4.2. В разделе 2 описывается КС-классификатор. В разделе 3 описывается метод отбора признаков, который, по нашему мнению, более адекватен специфике веб-данных, чем широко известные методы. Лучшие результаты участников дорожек РОМИП прошлых лет анализируются в разделе 4.2. В разделе 5 даются результаты дорожки 2010 года. В Разделе 6 проводится анализ результатов.

1.1 Свойства обучающей выборки РОМИП

В дорожке классификации веб-сайтов организаторы в качестве обучающего множества предлагают выборку, построенную на основе русскоязычного подраздела World -> Russian интернет-каталога Dmoz. Всего даётся 340 тыс. страниц с 2.2 тыс сайтов. Сайты относятся к 247 категориям.

Заметим, что эта выборка для обучения очень невелика. На рубрику приходится в среднем порядка 10 сайтов. Однако, если поставить в соответствие рубрики РОМИП рубрикам Dmoz, то видно, что можно было бы получить обучающие выборки гораздо больших размеров. Для этого достаточно учесть иерархическую структуру рубрик Dmoz, т.е. отнести к рубрике не только сайты, описанные непосредственно в ней, но и сайты всех ее «детей». Можно было бы получить выборку из 40-70 тыс. сайтов, что раз в 20-30 больше, чем в обучающем множестве РОМИП.

1.2 Свойства тестовых выборок РОМИП

Организаторы дают задание участникам провести классификацию порядка 20 тыс. сайтов из домена .by. Далее каждый год выбирается 15-20 рубрик (для каждого года с 2007 по 2009 свой набор рубрик, не пересекающееся с наборами рубрик других лет). Формируется “общий котел” ответов систем из этих рубрик, и отдается асессорам для оценки согласно процедуре [10]

У задачи классификации веб-сайтов есть характерная особенность. В качестве "учителя" естественно брать интернет-каталог, такой как dmoz.org, yahoo.com, list.mail.ru, yaca.yandex.ru и т.п. Но возникает вопрос о репрезентативности такого "учителя". Дело в том, что в интернет-каталогах редакторы описывают в основном известные и/или богатые содержанием сайты. Однако задача классификации сайтов состоит в том, чтобы автоматически соотнести с рубриками интернет-каталога сайты, которых в каталоге нет. А это в основном сайты менее известные, менее наполненные содержанием и т.д. То есть задача такова, что при измерении качества решения множество объектов, на котором строится метрика качества, должно по своей природе отличаться от обучающего множества вышеописанным образом.

Чтобы оценить отличия этих множеств, и понять соотношение свойств обучающего и «измеряющего» множеств РОМИП, мы сделали следующее.

1. Образовали четыре множества сайтов:
   a. Я.Каталог – сайты из Каталога Яндекса (порядка 100 тыс.)
   b. Дополнение до Я.Каталога – сайты «рунета», проиндексированные поисковой системой Яндекса, не считающиеся поисковым спамом, с ненулевой цитируемостью, имеющие количество проиндексированных

документов не меньше двух, и не описанные в каталоге Яндекса (порядка 1.5 млн.)
   c. РОМИП-обучение – сайты из обучающей выборки РОМИП (порядка 2 тыс.)
   d. РОМИП-тест – сайты из выборки By.Web (порядка 20 тыс.)
2. Для этих четырех выборок получили средние значения по следующим показателям:
   a. ТИЦ – средняя тематическая цитируемость сайта
   b. Размер – среднее количество проиндексированных Яндексом документов сайта
   c. Показы – среднее количество показов сайта в выдаче поиска Яндекса за период с декабря 2009 г. по август 2010 г.

Данные, описанные выше, представлены в **Таблице 1**.

Таблица 1. Сравнение показателей для четырех выборок сайтов

	ТИЦ	Размер	Показы
Я.Каталог	162,5	19196,9	150406
Дополнение до Я.Каталога	11,4	267,1	2783,4
РОМИП-обучение	250,6	27174,2	228612
РОМИП-тест	10,0	1434,4	5114,5

В **Таблице 1** видно, что выборки отчетливо разделяются на две группы: РОМИП-обучение и Я.Каталог с одной стороны, и РОМИП-тест и Дополнение до Я.Каталога с другой. Т.е. этим показателям тестовая выборка РОМИП для рассматриваемой дорожки близка к множеству сайтов, отсутствующих в интернет-каталоге и не похожа на множество сайтов, которые в каталоге есть. По нашему мнению, это свойство выгодно отличает метрику РОМИП от метрик, которые можно построить, например, на основе деления интернет-каталога на обучающую и тестовую части.

При экспериментах с КС-классификатором на таблицах релевантности [2] обнаружилось еще одно существенное свойство тестовых выборок РОМИП. Но об этом ниже.

2. КС-классификатор

Алгоритм работы КС классификатора описан в [1]. Отличием данной реализации было использование url имен и заголовков страниц в качестве дополнительных признаков. Каждое слово из url и заголовков страниц учитывалось три раза. Слова заголовков, в отличие от слов в “теле” текста, лемматизировались.

Как и в работе [1] рассчитывался вес слова $W_{ij}(L)$ - логарифм вероятности того, что документ относится к заданной рубрике j, при условии, что в документе длиной L встретилось данное слово i:

$$W_{ij}(L) = \ln\left(\frac{P_{ij}(L)}{\overline{P_{ij}(L)}}\right), \quad P_{ij} = \frac{N_{ij}}{N^j}, \; P_{ij}(L) = 1 - (1 - P_{ij})^L$$

где N_{ij} - сколько раз встретилось слово i в документах, относящихся к рубрике j; N^j - сколько всего слов в рубрике j; $\overline{P_{ij}(L)}$ – считается на множестве документов, которые не относятся к рубрике j.

Веса слов рассчитывались при параметре длины документа L = 50. Использование формул, более точно учитывающих длину документа, даёт незначительный прирост качества, но существенно замедляет обучение и классификацию.

Признаки сортировались по весу. Оставлялись только R_j^g признаков с максимальным положительным весом (ключевых слов) и R_j^b признаков с максимальным отрицательным весом. Вес остальных признаков считался равным 0.

Для классификации документа для каждой темы j вычислялось

$$F_j = \frac{\sum_i W_{ij} \cdot N_i}{\sum_i N_i}$$

где F_j - среднее значение логарифма вероятности того, что документ соответствует рубрике j, N_i - сколько раз i-ый признак встретился в документе.

Считалось, что документ относится к рубрике, если $F_j > R_j^{\min}$, и документ не относится к рубрике, если $F_j < R_j^{\max}$.

Если число документов сайта, относящихся к рубрике, превышало число документов, не относящихся к рубрике, то сайт относился к рубрике.

Параметры алгоритма R_j^g, R_j^b, $R_j^{\min}$, $R_j^{\max}$ оптимизировались последовательно для получения максимального качества классификации (F1-меры рубрики j).

3. Отбор признаков

Решение задачи выбора признаков для классификации в широко известных методах (см. напр. [3,4]) основывается на сравнении признаков между собой и отборе наиболее информативных признаков по их численным характеристикам на обучающем множестве. Однако мы обнаружили, что при классификация Веба такие методы недостаточно учитывают присущий этой задаче шум. Случайные слова, частотные на одном сайте (слово с ошибкой, ник пользователя, многократное (тысячи раз) цитирование одной и той же случайной фразы, нередко по общеизвестным критериям отбора признаков оказываются значимыми для рубрик, к которым относятся эти сайты.

Эти наблюдения навели на мысль построить алгоритм отбора признаков, основанный на кросс-валидации дешево вычисляемых показателей признаков, таких как абсолютная и относительная частоты, веса W_i^j и т.п.

Из исходного множества документов извлекаются все признаки, которые встретились больше 10 раз. Далее исходное множество сайтов делится на два – обучающее train и проверочное valid. Разбиение проводится посайтово, документы одного сайта могут быть или в обучающем или в проверочном подмножестве.

Для каждого слова i и для всех рубрик j (включая тему «все») собирается статистика встречаемости слова.

Для каждого признака также вычисляется максимальный вес на множестве рубрик. Выбираются признаки с максимальным весом больше $L_M \approx 0.5$. Параметр L_M подбирается для фиксированного множества рубрик.

Далее слова фильтруются по следующим правилам.

1. **Если** /**для рубрики* в *одном множестве слово редкое, в другом частотное*/*
$$\left(\mathrm{N}_{\mathrm{ij}}^{\mathrm{tr}} < L_{rare} \quad \& \quad \mathrm{N}_{\mathrm{ij}}^{\mathrm{vld}} > L_{freq}\right) \mid \left(\mathrm{N}_{\mathrm{ij}}^{\mathrm{tr}} > L_{rare} \quad \& \quad \mathrm{N}_{\mathrm{ij}}^{\mathrm{vld}} < L_{freq}\right)$$
то *слово фильтруется*;

2. **Если** /**в обоих множествах слово редкое* */
$$\left(\mathrm{N}_{\mathrm{i}}^{\mathrm{tr}} < L_{rare} \quad \& \quad \mathrm{N}_{\mathrm{i}}^{\mathrm{vld}} < L_{rare}\right)$$
то *слово фильтруется*;

3. **Если** /**для рубрики слово редкое в обоих множествах, и хотя бы в одном вес положительный*/*
$$\left(\mathrm{N}_{\mathrm{ij}}^{\mathrm{tr}} < L_{rare} \quad \& \quad \mathrm{N}_{\mathrm{ij}}^{\mathrm{vld}} < L_{rare}\right) \quad \& \quad \left(W_{ij}^{tr} > 0 \quad \mid \quad W_{ij}^{vld} > 0\right)$$
то *слово фильтруется*;

4. **Если** */*слово частотно в обоих множествах*/*
$$\left(\mathrm{N}_{\mathrm{ij}}^{\mathrm{tr}} > L_{freq} \quad \& \quad \mathrm{N}_{\mathrm{ij}}^{\mathrm{vld}} > L_{freq}\right)$$
то
если */*вероятность встретить слово в рубрике на train и valid сильно отличается*/*
$$\frac{\left|\mathrm{P}_{\mathrm{ij}}^{\mathrm{tr}} - P_{\mathrm{ij}}^{\mathrm{vld}}\right|}{\sqrt{P_{\mathrm{ij}}\left(1-P_{\mathrm{ij}}\right)\left(\frac{1}{N^{\mathrm{j\,vld}}} + \frac{1}{N^{j\ \mathrm{tr}}}\right)}} > L_P$$
или */*вес слова для рубрики на train и valid сильно отличается*/*
$$\frac{\left|\mathrm{W}_{\mathrm{ij}}^{\mathrm{tr}} - \mathrm{W}_{\mathrm{ij}}^{\mathrm{vld}}\right|}{\min\left(\left|\mathrm{W}_{\mathrm{ij}}^{\mathrm{tr}}\right|, \left|\mathrm{W}_{\mathrm{ij}}^{\mathrm{vld}}\right|\right)} > L_W$$
или */*для рубрики на одном из множеств слово характерно, на другом нет*/*
$$\mathrm{W}_{\mathrm{ij}}^{\mathrm{tr}} \cdot \mathrm{W}_{\mathrm{ij}}^{\mathrm{vld}} < 0$$
то *слово фильтруется*;

5. **Если** /**слово характерно для рубрики, но недостаточно часто встречается во всей коллекции*/*

$$\left(\mathrm{W}_{\mathrm{ij}}^{\mathrm{tr}} > 0 \quad | \quad \mathrm{W}_{\mathrm{ij}}^{\mathrm{vld}} > 0\right) \quad \& \quad \left(N_i < N^j \cdot L_N\right)$$

то *слово фильтруется*;

6. **Если** /**слово в рубрике встречается лишь однажды, и на всей коллекции слишком редкое*/*

$$N_i^j < 2 \quad \& \quad \frac{N^j}{\overline{N^j}} N_i < 2$$

7. Для каждого признака вычисляется максимальный вес на множестве рубрик. Выбираются признаки с максимальным весом больше $L_M \approx 0.5$. Параметр L_M подбирается для фиксированного множества рубрик.

Здесь $L_{rare} = 3, L_{freq} = 50,$ $L_P = 30, L_W = 3, L_N = 10^{-4}$; $\mathrm{N}_{\mathrm{ij}}^{\mathrm{tr}}, \mathrm{N}_{\mathrm{ij}}^{\mathrm{vld}}$ – сколько раз слово i встретилось в рубрике j на обучающем и проверочном множестве, $\mathrm{W}_{\mathrm{ij}}^{\mathrm{tr}}, \mathrm{W}_{\mathrm{ij}}^{\mathrm{vld}}$ – вес слова i для рубрики j на обучающем и проверочном множестве, $\mathrm{W}_{\mathrm{ij}} = \mathrm{W}_{\mathrm{ij}}(50)$

В выборке из 3 млн. документов содержится порядка 3 млн. признаков с встречаемостью более 10. После вышеописанного отбора остается около 100 тыс. признаков для русскоязычных текстов (среди которых встречаются «латинические» термины), либо около 50 тыс. признаков для англоязычных текстов. Кроме того, для рубрик в среднем ~90% признаков получаются нулевыми, что позволяет ускорить классификацию и уменьшить требуемую для хранения правил память в десять раз. Т.е. для 500 тематических рубрик объем требуемой памяти снизился с сотен до десятков мегабайт, а скорость классификации документов увеличилась с тысяч до десятков тысяч в секунду. Это существенно облегчило встраивание классификатора в механизм индексирования Веба Яндексом. После отбора признаков показатели качества классификации практически не снижаются.

4. Эксперименты

Для обучения классификатора исходное множество сайтов разбивается на три – обучающее, проверочное, тестовое (ОПТ). Отбор признаков проводится на ОП подмножествах. Далее, на обучающем множестве собирается статистика для отобранных признаков. На проверочном множестве оптимизируются параметры классификатора R_j^g, R_j^b, $R_j^{\min}$, $R_j^{\max}$. Для классификатора используется статистика признаков, собранная на объединенном ОП множестве и и параметры классификатора, оптимизированные на проверочном множестве. Тестовое множество используется для контроля качества полученного классификатора.

Для обучения важно именно количество сайтов в обучающей выборке, а не количество страниц. При постраничной разбивке на три подмножества результаты тестирования получаются существенно завышенными (10-15%), алгоритмы классификации – переобученными. Качество классификации оказывается хуже, чем при обучении на подмножествах с сайтовой разбивкой.

4.1 Эксперименты на обучающей выборке и данных каталога DMOZ

В связи маленькими размерами обучающих выборок вопрос о качественном разбиении обучающего множества на подмножества становится крайне важным.

Было сделано три разных разбиения – random, balanced и full.

1. Разбиение random: хосты равновероятно случайным образом относятся к ОПТ подмножествам.
2. Разбиение balanced: в каждой рубрике сайты отсортированы по алфавиту, далее, идя по списку, мы последовательно относим к ОПТ подмножествам. Эта схема гарантирует, что если рубрика представлена 3*n хостами, то в ОПТ множествах будет по n хостов.
3. Разбиение full: аналогично balanced, но подмножества только два – ОП.

В **Таблице 2** представлены результаты прогонов КС-классификатора на вышеупомянутых трех схемах разбиения. Полным множеством называется тестирование на всем множестве хостов, данном для обучения. Самопроверка подразумевает

тестирование на специально выделенной тестовой части обучающего множества, которая не использовалась в обучении.

Таблица 2. Сравнение вариантов разбиения на обучающей выборке

Разбиение	Самопроверка		Полное множество	
	MicroF1	MacroF1	MicroF1	MacroF1
Random	22.3	12.8	н/д	н/д
Balanced	30	18.5	53.6*	47.8*
Full	н/д	н/д	64.3*	56.3*

* помечены результаты тестирования, когда часть тестовых данных использовалась для обучения.

Эксперимент позволил сделать следующие выводы.

1. Разбиение balanced существенно – в 1.5 раза – превосходит схему random по F1-мере на множестве Самопроверка. Было решено, что из этих двух схем в дорожке будет участвовать только balanced.
2. Разбиение full на Полном множестве показывает заметно более высокий результат, чем balanced. Однако проверка на Полном множестве не дает информации о степени переобученности метода. Поэтому было решено сравнить с помощью дорожки РОМИП обе этих схемы разбиения.
3. Показатели качества КС-классификатора, измеренные при тестировании на части обучающей выборки, оказалось низким по сравнению с лучшими показателями участников РОМИП за предыдущие годы.

Третий вывод побудил нас попытаться использовать дополнительные данные. Мы сформировали обучающую выборку на основе каталога Dmoz с учетом иерархической структуры его рубрик, как это было описано в п. 1.1. Был получен список из 40 тыс. сайтов. С них собрано 2 млн. документов. Процедура выбора документов с сайта подобна описанной в [9]. Разбиение ОПТ делалась согласно схеме random. Обученный на такой выборке классификатор получил название dmoz.

Полученные показатели качества dmoz-а оказались неоднозначными. С одной стороны, при самопроверке для dmoz

получились такие метрики: MicroF1 = 62.7, MacroF1=35.3. Т.е. dmoz по Самопроверке оказался лучше, чем balanced, практически в 2 раза. С другой стороны, при измерении по схеме Полное множество для dmoz-а получилось: MicroF1 = 38.0, MacroF1=32.1. Т.е. метрики dmoz-а на Полном множестве оказались гораздо хуже, чем метрики и для balanced, и для full, причем в случае full хуже почти в 2 раза (см. **Таблицу .2**). Чтобы разрешить эту неоднозначность, мы решили измерить с помощью дорожки РОМИП также и dmoz.

4.2 Эксперименты на таблицах релевантности прошлых РОМИП-ов

Для того, чтобы лучше понять свойства метрик РОМИП, мы скачали таблицы релевантности РОМИП [2], собрали информацию об оценках систем в дорожках, и о лучших результатах участников дорожек за прошлые годы. Далее мы измерили показатели качества для трех вышеописанных алгоритмов согласно методикам [5]. Результаты представлены в **Таблицах 2** и **3**. В ней жирным курсивом помечены показатели с максимальной среди рассматриваемых классификаторов мерой.

Таблица 3. Тестирование AND (MacroF1 MicroF1)

Год	Тип	bestRomip	Источник	balanced	Full	Dmoz
2007	Macro	32	УИС [6]	35.1	41.2	***53.5***
	Micro	28	УИС [6]	37.4	41.8	***59.2***
2008	Macro	***38***	RCO [7]	27.5	31.9	20.6
	Micro	***38***	RCO [7]	31.5	39.9	21.3
2009	Macro	***39***	RCO [8]	33.1	36.8	17.8
	Micro	***51***	RCO [8]	39.9	45.2	21.1

Таблица 4. Тестирование OR (MacroF1 MicroF1)

Год	Тип	bestRomip	Источник	balanced	Full	dmoz
2007	Macro	48	УИС [6]	29.4	39.9	***50.7***
	Micro	46	УИС [6]	27.6	37.3	***53.8***
2008	Macro	***43***	RCO [7]	25.6	29.5	15
	Micro	***44***	RCO [7]	29.2	30.1	14.4
2009	Macro	***41***	RCO [8]	25.9	27.7	16.4
	Micro	***53***	RCO [8]	34.9	40.6	14.1

Здесь можно обратить внимание на следующее.

- F1-мера алгоритмов balanced и dmoz оказалась заметно лучше, чем на Самопроверке, но для всех трех алгоритмов заметно хуже, чем у лучших результатов РОМИП.
- Сильные и слабые метрики для bestRomip сходятся: в 2007 году различались в полтора раза, а в 2009-м почти одинаковы

Однако наиболее интересным оказалось сопоставление полноты и точности для balanced, dmoz и full на Самопроверке и оценках за прошлые годы – см. **Таблицу 5**.

Таблица 5.

	Класси-фикатор	Balanced		Dmoz		full	
	Выборка	Само-проверка	РОМИП 2007_OR	Само-проверка	РОМИП 2007_OR	Само-проверка	РОМИП 2007_OR
Micro	Precision	36.3	87.7	55.9	85.5	н/д	79.2
	Recall	25.5	16.4	71.3	39.2	н/д	24.4
	F1	30.0	27.6	62.7	53.8	н/д	37.3
Macro	Precision	21.3	71.2	37.8	81.1	н/д	89.0
	Recall	16.3	18.5	33.1	36.9	н/д	25.7
	F1	18.5	29.4	35.3	50.7	н/д	39.9

Отметим следующие изменения показателей:

- Точность значительно выросла – в 2-3 раза
- Микро-полнота заметно упала – в 1.5-1.8 раз
- Макро-полнота незначительно увеличилась

Эти изменения, по-видимому, говорят о том, что эталонная классификация объектов из обучающего множества весьма существенно отличается от эталонной классификации на тестирующем множестве. Мы предполагаем, что разница обусловлена существенными различиями в подходе к классификации сайтов у построивших обучающее множество редакторов Dmoz, и асессоров, выполняющих задание по оценке дорожек классификации сайтов РОМИП.

Мы решили построить и измерить на дорожках РОМИП еще 5 вариантов классификации, в которых полнота увеличена. Для этого в классификаторах full и dmoz пороговые числа ключевых слов R_j^q для всех рубрик были увеличены в 3 раза (full3 и dmoz3), и, соответственно, в 5 раз (dmoz5). После этого решения

классификаторов full3 и dmoz3 объединили в sum3, решение full3 и dmoz5 объединили в sum5.

Результаты измерения для full3, dmoz3, dmoz5, sum3 и sum5 приведены в Таблицах 6 и 7.

Таблица 6. Тестирование AND (MacroF1, MicroF1)

Год	Тип	Best Romip	Источник	full3	dmoz3	sum3	sum3*	sum5	sum5*
2007	Macro	32	УИС [6]	66.4	63.1	***72.7***	73.3	69.2	***73.6***
	Micro	28	УИС [6]	51.3	***54.5***	51.5	51.9	48	48.8
2008	Macro	38	RCO [7]	42.6	28.3	51.1	52.6	***52.5***	***54.2***
	Micro	38	RCO [7]	48.8	27.2	***49.9***	***50.6***	48.7	48.8
2009	Macro	39	RCO [8]	52	33	58.7	60.9	***59***	***66.4***
	Micro	51	RCO [8]	61.6	41.3	***63.5***	63.3	63.1	62.8

Таблица 7. Тестирование OR (MacroF1, MicroF1)

Год	Тип	Best Romip	Источник	full3	dmoz3	sum3	sum3*	sum5	sum5*
2007	Macro	48	УИС [6]	61.5	64.2	***72.3***	75	70.2	***76.7***
	Micro	46	УИС [6]	55.2	65.7	***65.7***	68.5	62.1	***68.5***
2008	Macro	43	RCO [7]	43.9	24.8	49.3	50.4	***52.2***	***53.7***
	Micro	44	RCO [7]	45.7	26.4	49.3	50.1	***50.6***	***51.5***
2009	Macro	41	RCO [8]	44.3	31	***51.8***	55.9	50.9	***61.7***
	Micro	53	RCO [8]	59.6	36	62.3	62.9	***62.8***	***64.6***

* помечены результаты тестирования, где не делалось ограничение в 5 рубрик – максимальное число рубрик для хоста согласно правилам оценки дорожки.

Классификаторы sum3 и sum5 показали наилучшие результаты при тестировании на оценках за прошлые годы. Однако эти классификаторы дали в среднем около 8 отнесений к классам. При этом в методике измерения дорожки есть ограничение в 5 отнесений к классам на хост. Это ограничение потребовало дополнительной фильтрации рубрик. Фильтрация рубрик проводилась по следующему алгоритму:

1. Если отнесений к рубрикам для хоста не больше 5, то оставляем все отнесения.
2. Для оставшихся хостов из списка отнесений убираем отнесения к маленьким рубрикам – с количеством менее 5-и хостов в обучающем множестве. Заметим, что в оценках из дорожек прежних лет отнесения к таким рубрикам отсутствовали.

3. Оставшийся список рубрик хоста сортируем по качеству классификации, оставляем 5 рубрик с лучшим качеством.

В **Таблицах 6** и **7** жирным курсивом помечены показатели с максимальной мерой. При этом не учитывались классификаторы без фильтрации, помеченные звездочкой. Их результаты, если они максимальные, выделены серым жирным курсивом. Видно, что фильтрация заметно сказалась на результатах классификаторов sum3 и sum5. F1-меры упали на 1% -10%

5. Результаты дорожки РОМИП'2010

Показатели метрик в дорожке 2010 года в варианте judged-only для представленных нами классификаторов приведены в **Таблице 9**.

Таблица 9. Метрики вариантов КС-классификатора по оценкам РОМИП-2010 – judged-only

Вар	Тип	Best Romip	Balanced	full	full3	dmoz	dmoz3	sum3	sum3*	sum5	sum5*
AND	Macro	40.5	21.5	22.7	39.3	27.5	36.6	***47.9***	***49.9***	42.7	48.6
	Micro	44.8	23.2	26.3	46.1	25.8	41.7	***52.2***	***54.1***	47.8	52.2
OR	Macro	***52.9***	22.1	19.8	36.3	22.7	34.4	46.6	52.3	46.2	***55.7***
	Micro	***56.2***	17.7	22.5	42.4	21.7	37.4	50.9	53.9	53	***59.5***

* помечены результаты тестирования, где не делалось ограничение в 5 рубрик – максимальное число рубрик для хоста.

Видно, что ограничение 5 рубрик на хост существенно сказалось на конечном результате, привело к снижению лучших результатов от 2% до 9%.

Показатели judged-only для всех систем приведены на **Рисунке 1**. Метрики наших алгоритмов в РОМИП'2010 получили номера с первого по восьмой. Они сгруппированы в левой части гистограммы. Метрики алгоритмов других участников дорожки сгруппированы справа. Полные метрики приведены на **Рисунке 2** в аналогичном порядке.

К сожалению, при формировании результатов мы допустили существенную ошибку: рубрики РОМИП были идентифицированы нашими «внутренними» номерами. Таким образом, ответы КС-классификаторов фактически не участвовали в «общем котле» оценок. Это хорошо видно при сопоставлении judged only и полных оценок.

Рисунок 1. Показатели качества участников дорожки РОМИП-2010 – метрики judged-only

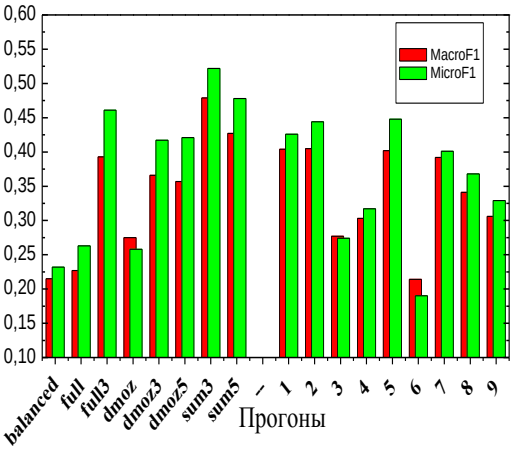
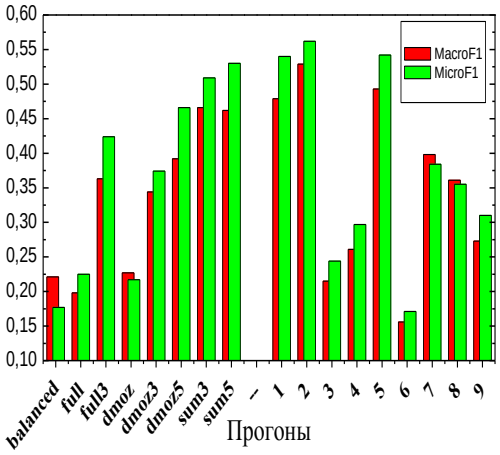
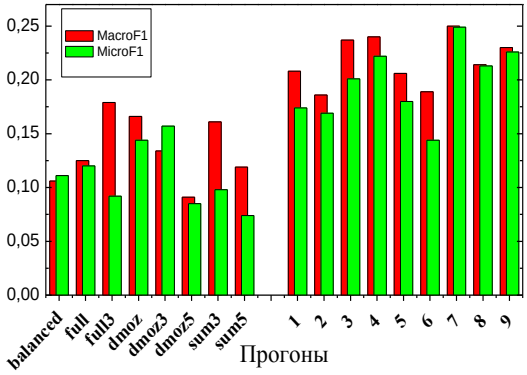
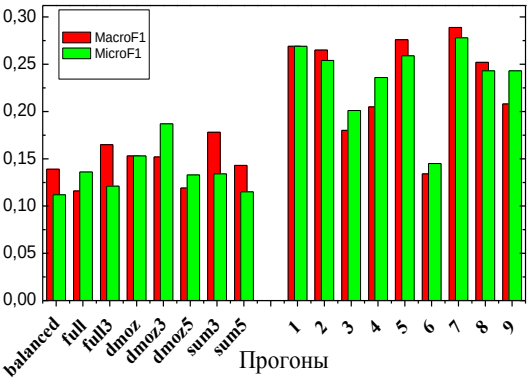
	**Judged only AND**
	**Judged only OR**

Рисунок 2. Показатели качества участников дорожки РОМИП’2010 – полные метрики

	**Full AND**
	**Full OR**

6. Анализ результатов дорожек

1. Завышение полноты, по-видимому, сработало на дорожке РОМИП'2010: классификаторы full3, dmoz3, dmoz5, sum3, sum5 показывают примерно такое же увеличение F1-мер, по сравнению с исходными алгоритмами, как и на таблицах релевантности за прошлые годы. Можно сделать два вывода:
   - разница свойств обучающего и тестового множеств РОМИП, обнаруженная на данных прошлых лет, проявляется и в дорожке 2010 года
   - КС-классификатор позволяет достаточно эффективно обменивать точность на полноту
2. Метод full в целом дает показатели лучшие, чем balanced. Можно сделать вывод, что выделение тестового подмножества из обучающих данных приводит к ухудшению качества классификации на 2% за счет уменьшения множества, на котором строятся правила, на треть.
3. Полные данные из интернет-каталога Dmoz оказались весьма полезными, даже несмотря на то, что они, по-видимому, использовались нами не лучшим образом. Их применение дает существенную прибавку в показателях – порядка 10%.
4. По метрикам AND judged only три прогона КС-классификатора показали лучшие результаты. В то же время по OR judged only ни один из прогонов КС-классификатора лучшего результата не показал. Видимо, алгоритм фильтрации результатов, описанный в п.4.2, оказался не вполне удачным. Он на метрике OR сказывается сильнее – это видно из **Таблицы 9** и сопоставления **Таблиц 6** и **7**.

7. Заключение

В заключение хотелось бы высказать пожелания по совершенствованию обучающих и тестовых наборов данных РОМИП для дорожки классификации веб-сайтов.

Во-первых, хотелось бы увеличить обучающее множество, включив в примеры для рубрик Dmoz сайты из соответствующих подрубрик. Если есть проблемы с объёмами данных, то можно увеличить количество сайтов в обучающей выборке, компенсировав

объём меньшим количеством страниц, отбираемых с сайта. По нашему мнению, таким образом увеличится разнообразие объектов в обучающей выборке, поскольку страницы с разных сайтов как правило более различаются по лексике и другим признакам, чем страницы с одного сайта.

Во-вторых, хорошо бы, чтобы классификация объектов в тестовом множестве была более похожа на классификацию в обучающем множестве, т.е. чтобы асессоры классифицировали сайты примерно так, как это делают редакторы каталога Dmoz. Сейчас, по-видимому, имеются существенные различия. Это, по нашему мнению, приводит к тому, что метрики измеряют скорее качество настройки алгоритмов на особенности коллекции, чем качество самих алгоритмов. Мы предполагаем, что можно подправить методику оценки и инструкцию асессора так, что при этих же асессорах может получиться более близкий к обучающему множеству результат оценки.

Литература

1 М.Ю. Маслов, А.А. Пяллинг, С.И. Трифонов. Автоматическая классификация веб-сайтов. RCDL, Дубна, Россия, (2008)с. 230-235. (http://rcdl2008.jinr.ru/pdf/230_235_paper27.pdf)

2 Таблицы релевантности для дорожек РОМИП за разные годы. http://romip.ru/relevance-tables/ru/index.html

3. S. Chakrabarti. Mining the web. Discovering knowledge from hypertext data. Elsevier 2003.

4. I. H. Witten, E. Frank. Data mining: practical machine learning tools and techniques. Elsevier 2005.

5. М. Агеев, И. Кураленок, И.Некрестьянов. Официальные метрики РОМИП. РОМИП Петрозаводск (2009). (http://romip.ru/romip2009/20_appendix_a_metrics.pdf)

6. М.С. Агеев, Б.В. Добров, П.В. Красильников Н.В. Лукашевич, А.М. Павлов, А.В. Сидоров, С.В. Штернов. УИС РОССИЯ в РОМИП 2007: поиск и классификация. РОМИП, Екатеринбург, (2007), стр. 199-220 (http://romip.ru/romip2007/2007_20_uis.pdf)

7. П.Ю.Поляков, В.В.Плешко RCO на РОМИП 2008. РОМИП, Дубна, (2008), стр. 96-107. (http://romip.ru/romip2008/2008_08_rco.pdf)

8. П.Ю.Поляков, В.В.Плешко, А.Е. Ермаков. RCO на РОМИП 2009. РОМИП Петрозаводск (2009) стр. 122-134. (http://romip.ru/romip2009/11_rco.pdf)
9. Описание наборов данных для участников программы научных стипендий Яндекса 2005 г. http://company.yandex.ru/grant/datasets_description.xml
10. Инструкция асессора для дорожки Веб классификации для РОМИП'2009 http://romip.ru/romip2009/22_appendix_C_WC.pdf
11. Описание дорожки по классификации Веб-сайтов http://romip.ru/ru/2010/tracks/web-classification.html

KC-classifier and RIRES web-site classification track

Maslov M., Pyalling A.

The article presents an investigation of RIRES web-site classification track metrics. KC-classifier is compared with another members of the track. New method feature selection suggested.