Повышение качества классификации текстов путем модификации обучающего множества

© А.Ю. Колесов
Ярославский Государственный Университет им. П.Г. Демидова,
Ярославль
kolesov.ay@ya.ru

Аннотация

Автоматическую классификацию текстов часто используют для структурирования больших объемов данных. В этой работе предложен новый метод повышения качества классификации путем модификации обучающего множества. Метод опробован на общедоступной коллекции текстов, где один документ может относится к нескольким классам.

1 Введение

Рассмотрим задачу классификации, когда имеется много рубрик, один документ может относится к одной или нескольким рубрикам (multi-label задача). В работе [1] представлены методы повышения качества классификации в условиях неполноты обучающего множества. Это условие означает, что для некоторых объектов из обучения могут быть не проставлены некоторые метки классов (рубрик).

В этой работе мы, во-первых, смоделируем ситуацию неполноты меток на хорошо размеченной коллекции биомедицинских исследовательских статей, представленной на конкурсе JRS'12 [2] и доступной на сайте конкурса. Тем самым еще раз покажем, что методы из статьи [1] хорошо работают. Во-вторых, исходя из результатов экспериментов, предложим метод повышения качества классификации для multi-label задач без предположения неполноты данных (т. е. для хорошо размеченных данных).

2 Материалы и методы

2.1 Данные JRS'12

Коллекция данных JRS'12 представляет из себя набор из 20000 биомедицинских статей, доступных на PubMed Central [3]. Каждая из статей была размечена экспертами Pubmed в области биомедицины по MeSH (Medical Subject Headings) [4]. Каждый документ имеет 25640 атрибутов, значение каждого из которых означает, насколько соответствующий ему биомедицинский термин характеризует данный документ.

2.2 Предобработка данных

Данные представляют из себя матрицу документ — вес атрибута. Веса атрибутов предоставлены организаторами JRS'12. Для проведения экспериментов мы нормировали строки матрицы по норме l2.

2.3 Метрики качества классификации

Пусть N — количество тестовых документов. $TrueTopics_i$ — множество верных (отмеченных экспертом) меток рубрик для i-ого документа. $PredTopics_i$ — множество меток рубрик, которые выдает классификатор для i-ого документа. Определим следующие метрики качества классификации, подсчитываемые для i-ого документа:

$$Precision_i = \frac{|TrueTopics_i \cap PredTopics_i|}{|PredTopics_i|}$$

$$Recall_i = \frac{|TrueTopics_i \cap PredTopics_i|}{|TrueTopics_i|}$$

$$Fscore_i = 2 \cdot \frac{Precision_i \cdot Recall_i}{Precision_i + Recall_i}$$

Для каждой метрики будем рассчитывать усредненные метрики:

$$AvgMetric = \frac{\sum_{i=1}^{N} Metric_i}{N}$$

Подставляя вместо $Metric_i$ соответствующую определенную выше метрику (например, $Fscore_i$), получаем ее усреднение по тесту.

2.4 Эксперименты

Установки экспериментов классификации точно такие же как в работе [1]. Здесь приведены результаты только с использованием модификации обучаю-

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

Таблица 1 *Значения метрик классификации*

Доля удаленных меток	Параметры w-kNN		До модификации обучающего множества			После модификации обучающего множества		
	Optimal k	Optimal T	precision	recall	F-score	precision	recall	F-score
0	-	-	0,444	0,6377	0,5235	-	-	-
0,1	5	0,1	0,46	0,5742	0,5108	0,5151	0,577	0,5443
0,2	15	0,1	0,3679	0,5239	0,4322	0,508	0,5702	0,5373
0,4	15	0,05	0,3204	0,4455	0,3727	0,4788	0,5741	0,5222
0,6	35	0,05	0,1469	0,338	0,2048	0,5387	0,4559	0,4939

.

щего множества на основе метода *k*-взвешенных ближайших соседей (w-kNN).

Опишем, как мы моделировали неполноту меток в обучении. Для этого задается доля удаляемых меток. Затем случайным образом отбирается заданное количество меток, но так, чтобы ни один документ не остался без меток и ни одна рубрика не осталась без документов. Затем проводится обучение/классификация на полученном обучающем множестве и на его модификации.

3 Результаты

В Таблице 1 представлены результаты экспериментов. В ячейке первого столбца указана доля удаленных меток при моделировании. Например, 0,2 — означает, что 20% исходных меток (т. е. пар документ-рубрика были удалены из обучающего множества). Во втором и третьем столбце указаны подобранные (см. [1]) оптимальные значения для алгоритма w-kNN, где k — количество используемых ближайших соседей, T — порог принадлежности рубрики документу. В следующих столбцах содержатся значения метрик до и после модификации обучающего множества.

Как и ожидалось при удалении меток результаты по F-мере ухудшаются (чем больше меток удаляем, тем хуже качество классификации). После модификации обучающего множества F-мера значительно увеличивается. Более того, для значений доли удаленных меток 0,1 и 0,2 усредненная F-мера превосходит усредненную F-меру при классификации по исходному обучающему множеству.

Таким образом, метод, предлагаемый для улучшения качества классификации, заключается в следующем. Удаляем случайным образом небольшое количество меток из обучающего множества. Применяем метод модификации обучающего множества (из работы [1]). Обучаемся на модифицированном обучающем множестве.

4 Выводы

Применение предложенного метода повышения качества классификации позволило улучшить результаты по F-мере с 52,35% до 54,43%, т. е. В относительном выражении на 4%. Отметим, что значение 54,43% превосходит лучший результат участников конкурса JRS'12 (53,579%).

Также мы подтвердили результаты работы [1] путем моделирования на хорошо размеченной коллекции данных.

Описанный в работе эффект требует дополнительного изучения. Требуется определить правило для вычисления, какую долю меток удалять. Работает ли этот метод на multi-class задачах? Это дело дальнейшей работы.

Литература

[1] Колесов А.Ю. Методы классификации в условиях противоречивого обучающего множества. Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2011. Воронеж: 2011, с. 140-146.

[2] JRS 2012 Data Mining Competition: Topical Classification of Biomedical Research Papers. Http://tunedit.org/challenge/JRS12Contest/JRS12Contest

[3] Home - PubMed – NCBI. http://www.ncbi.nlm.nih.gov/pubmed

[4] Medical Subject Headings - Home Page. http://www.nlm.nih.gov/mesh/

Improvement of text classification performance by modifying the training set

Anton Kolesov

Automatic data classification methods are frequently employed for structuring large amounts of data. In this paper, we propose a new method to increase the performance of classification of data by modifying the training set. The method is tested on publicly available multi-label collection of texts.