Оценка эффективности масштабируемых алгоритмов классификации текстов

© Максаков А.В

Московский Государственный Университет им. М.В. Ломоносова
bruzz@yandex.ru

Аннотация

Статья посвящена исследованию алгоритмов, обладающих низкой вычислительной сложностью обучения и классификации. Представлены результаты экспериментов в рамках дорожек по тематической классификации семинара РОМИП'2006 и проведено сравнение качества предложенных алгоритмов.

1. Введение

Целью повторного участия в семинаре является продолжения исследования алгоритмов классификации, в том числе и применительно к решению задачи периодического тематического поиска. Итоговое качество поиска гибридной системы, представляющей собой комбинацию системы поиска по ключевым словам и классификатора, напрямую зависит от качества классификации. Также важными требованиями к используемым алгоритмам в рамках этого подхода является низкая вычислительная сложность классификации и обучения алгоритмов. Таким образом, целью участия в семинаре РОМИП-2006 было сравнение *линейно масштабируемых* алгоритмов с другими алгоритмами по качеству рубрикации. Под линейно масштабируемыми алгоритмами понимаются алгоритмы, сложность обучения и классификации которых линейно зависит от количества документов в обучающей выборке и размерности пространства признаков.

Наряду с характеристиками самого алгоритма классификации, на итоговое качество рубрикации оказывают влияние и способ предва-

рительной обработки текстового документа, осуществляющий переход от текстового представления к представлению в виде математической модели, используемой непосредственно алгоритмом классификации. По этой причине, также интересно исследование зависимости качества классификации от выбранного способа такого перехода.

2. Описание рассматриваемых алгоритмов

В ходе экспериментов рассматривалось три алгоритма классификации: метод опорных векторов (алгоритм SVM), модифицированный алгоритм Байеса и метод построения нескольких разделяющих гиперплоскостей. Алгоритм SVM (с линейным ядром) использовался в качестве базового, в связи с тем, что он признается одним из лучших алгоритмов с точки зрения качества классификации [5]. Преимуществом двух других алгоритмов является их более низкая вычислительная сложность обучения (O(N) против $O(N^a)$, где N – число документов в обучающей выборке, *a* оценивается больше 1,7 [1]) и меньшие требования к оперативной памяти при обучении.

2.1 Модифицированный алгоритм Байеса

Данный алгоритм построен на основе давно известного алгоритма Байеса [4]. Экспериментальные исследования поведения алгоритма позволили обнаружить два систематических недостатка, сильно понижающих качество классификации:

- Предпочтение классификатором классов, содержащих большее количество примеров в обучающей выборке.
- Предпочтение классификатором классов, в которых содержится большее количество взаимно зависимых признаков (не выполняется предположение о независимости признаков).

Для борьбы с некорректным определением параметров, в случае неравномощных обучающих выборок классов, предлагается использовать парадигму класса-дополнения, описанную в прошлогодней статье [8]. Для частичной компенсации использования принципа независимости признаков, производится нормализация весов признаков

$$\theta_{cw} = \frac{\log(\theta'_{cw})}{\sum_{w \subset C} |\log(\theta'_{cw})|}, \text{ где } \theta'_{Cw} = \frac{\overline{N}_{Cw} + 1}{\overline{N}_C + |V|},$$

где $\overline{N}_{Cw}$- количество вхождений признака во все классы кроме данного, $\overline{N}_C$ - общее количество вхождений всех признаков в класс-дополнение, $|V|$ - размерность словаря признаков. Следует отметить, что данный алгоритм наиболее эффективен при решении задачи классификации на непересекающихся множествах классов, что не относится к дорожкам семинара РОМИП, поэтому ожидалось некоторое ухудшение итогового качества.

2.2 Алгоритм построения разделяющих гипреплоскостей

Для задачи бинарной классификации внесенные модификации не позволяют приблизить алгоритм Байеса по качеству к лучшим показателям (парадигма классов-дополнений не вносит никаких изменений), поэтому в таких случаях предлагается использовать алгоритм с условным названием ModFisher. Идея алгоритма состоит в последовательном нахождении направлений (как правило, не более 3-4), соответствующих *дискриминанту Фишера*[1,2], максимизирующему так называемый *индекс Фишера*

$$J(a) = \frac{(\frac{1}{|x|}\sum_{x\in X}(x,a) - \frac{1}{|y|}\sum_{y\in Y}(y,a))^2}{\frac{1}{|x|}\sum_{x\in X}(x,a)^2 - (\frac{1}{|x|}\sum_{x\in X}(x,a))^2 + \frac{1}{|y|}\sum_{y\in Y}(y,a)^2 - (\frac{1}{|y|}\sum_{y\in Y}(y,a))^2}$$

Вдоль такого направления можно эффективно разделить часть обучающих примеров. В дальнейшем использовать точки отсечения для положительных и отрицательных экземпляров вдоль каждого направления и процесс повторяется для оставшихся примеров.

Схема обучения алгоритма выглядит следующим образом:

1. Методом градиентного спуска находим локальный максимум *J(a)*.
2. Проецируем все обучающие экземпляры на полученное направление и запоминаем точку оптимального разделения классов, а также полупрямые, содержащие только положительные и отрицательные экземпляры.
3. Отбрасываем корректно классифицированные экземпляры на данном направлении и повторяем шаги 1-3 до достижения пустого множества экземпляров или фиксированного числа итераций.

Классификация экземпляра производится по следующему алгоритму:

цикл i = 1... количество направлений

Анализируем *i*-ое направление:

если документ находится на полупрямой положительных или отрицательных документов, выдаем соответствующую метку и выходим из цикла

если данное направление последнее, определяем метку экземпляра с помощью точки оптимального разделения классов

конец цикла

К слабым сторонам алгоритма следует отнести плохую устойчивость к ошибкам в обучающей выборке.

2.3 Метод опорных векторов

В ходе экспериментов использовалась реализация метода svm_light [3]. В связи с большим временем обучения на RBF-ядре, было решено от него отказаться, при прогонах использовалось линейное ядро. Также использовался ранее описанный [8] способ определения веса признака, учитывающий распределение признака по классам в обучающей выборке:

$$w_i = \ln(TF) \cdot IDF_{new}, \quad (3.4.1)$$

где IDF_{new} определяется согласно формуле

$$IDF_{new} = \sqrt{\max_{C' \in C} TF(w, C')} * IDF',$$

$$IDF' = \sqrt{\frac{|D|}{\sum_{C' \in C} \frac{TF(w, C')}{\sum_{w' \in F} TF(w', C')}}}$$

3. Предварительная обработка документов

3.1 Обработка текста на естественном языке

При обработке текста использовался словарный морфологический анализ, а также постморфологический анализ, основанный на использовании синтаксического анализатора. Также производилась фильтрация стоп-слов на основе заданного словаря и метки части речи, полученной на этапе лексического анализа.

3.2 Определение пространства признаков

Сокращение пространства признаков позволяет сократить вычислительную сложность задачи классификации, при этом во многих случаях наблюдается лишь незначительное понижение качества классификации или его улучшение [6,7]. При экспериментах рассматривался отбор признаков по критерию соотношения порядков (*Odds ratio*) [2]:

$$OR(t_k, C_i) = \frac{P(t_k, C_i)(1 - P(t_k, \overline{C_i}))}{(1 - P(t_k, C_i))P(t_k, \overline{C_i})},$$

где $P(C_i) = 1 - P(\overline{C_i}) = \frac{|D \subset C_i|}{|D|}$

$$P(t_k) = 1 - P(\overline{t_k}) = \frac{|D : t_k \in D|}{|D|}$$

$P(t_k, C_i)$ - вероятность совместного появления признака t_k и метки класса C_i в одном документе.

4. Оценка результатов экспериментов

В рамках семинара РОМИП было представлено 11 прогонов:

- 3 прогона для дорожки классификации Web-сайтов
- 2 прогона для дорожки классификации Web-страниц
- 6 прогонов для дорожки классификации нормативно-правовых документов

Для дорожек классификации Web-сайтов и страниц основной целью было оценить эффективность сокращения размерности пространства признаков на обучающих коллекциях, содержащих малорелевантные документы. На коллекции нормативно-правовых документов было интересно сравнить рассмотренные масштабируемые алгоритмы с методом опорных векторов.

В таблицах и на гистограммах использовались следующие обозначения:

- **MNB** – модифицированный алгоритм Байеса (*ModBayes*)
- **MF** – метод построения нескольких разделяющих гиперплоскостей (*ModFisher*).
- **SVM** – метод опорных векторов.

Также использовались следующие суффиксы:

- **oddsN** – сокращение размерности пространства признаков до N с использованием критерия *OddsRatio*.
- **doccnt** – сокращение размерности пространства признаков по количеству документов, содержащих данный признак.
- **log** – логарифмическое сглаживание количества вхождений признаков для алгоритмов ModBayes и SVM
- **idf'** – оценка веса признака $ln(TF)IDF_{new}$ для SVM

На дорожках, использующих коллекцию Web-документов, сокращение пространства признаков с помощью критерия *OddsRatio* позволило сохранить качество классификации при существенном сокращении времени обучения и классификации. Тем не менее, общие результаты говорят о необходимости более агрессивного отсеивания "мусорных" признаков и поиска способов выделения значимых признаков и оценки степени их значимости.

	F1 (micro)	F1
xxxx-1	0,490	0,486
xxxx-2	0,207	0,130
MNB-doccnt	0,109	0,094
odds10000	0,135	0,111
odds50000	0,108	0,089

Таблица 1. Оценки качества классификации на дорожке классификации Web-сайтов со слабыми требованиями релевантности

	F1 (micro)	F1
xxxx-1	0,541	0,444
xxxx-2	0,648	0,548
odds10000	0,043	0,044
odds40000	0,042	0,043
xxxx-5	0,039	0,021
xxxx-6	0,044	0,026
xxxx-7	0,028	0,018

Таблица 2. Оценки качества классификации на дорожке классификации Web-страниц со слабыми требованиями релевантности

Прогоны, сделанные для дорожки классификации нормативно-правовых документов с одной стороны показали применимость алгоритмов ModBayes и ModFisher для решения задачи, но с другой стороны очевидна необходимость устранения их недостатков: неустойчивости к ошибкам алгоритма ModFisher и неадапти рованности алгоритма ModBayes к решению задачи классификации “один-против-всех”. Также следует отметить ощутимое снижение качества классификации при отборе признаков, что связано с характером коллекции.

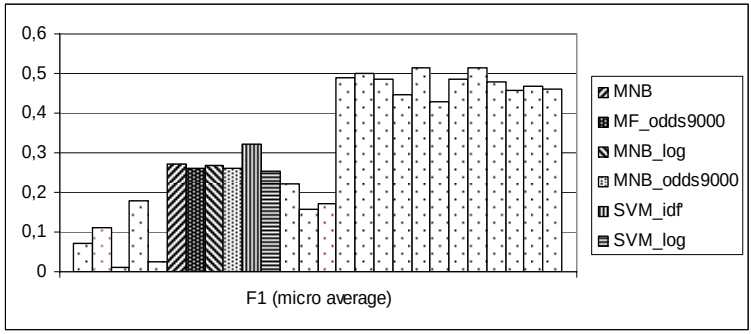

Рис.1. Значения меры F1-micro на дорожке классификации нормативно-правовых документов

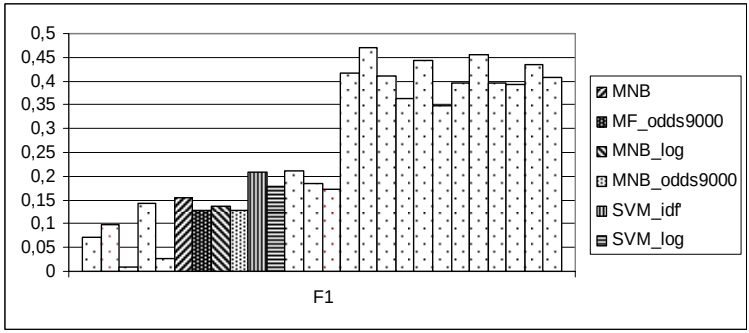

Рис.2. Значения меры F1 на дорожке классификации нормативно-правовых документов

	F1 (micro)	F1
MNB_log	0,271	0,153
MF_odds9000	0,260	0,128
MNB	0,269	0,138
MNB_odds9000	0,260	0,128
SVM_idf'	0,322	0,208
SVM_log	0,254	0,179

Таблица 3. Сравнение качества классификации представленных прогонов

5. Заключение

В рамках семинара РОМИП-2006 рассматривались масштабируемые алгоритмы классификации и некоторые способы сокращения пространства признаков. Предложенные алгоритмы показали приемлемые результаты сравнительно с алгоритмом SVM при одинаковом подходе к обработке текстов. В дальнейшем планируется развитие этих алгоритмов с целью устранения их недостатков. Также необходимо провести большую работу над предварительной обработкой документов и способами выделения значимых признаков.

Литература

[1] S. Chakrabarti. Mining The Web Discovering Knowledge From Hypertext Data. Morgan Kaufmann Publishers, 2004

[2] R. Fisher. The use of multiple measurements in taxonomic problems. *Eugen.*, 7:179-188, 1936.

[3] T. Joachims. Making Large-Scale SVM Learning Practical. *Advances in Kernel Methods. Support Vector Learning*. MIT-Press, 1999.

[4] D.Lewis. Naive Bayes at forty: The independence assumption in information retrieval. Proceedings of ECML-98, 10th European Conference on Machine Learning, pages 4-15, 1998

[5] Yang Y., Pedersen J. A comparative study on feature selection in text categorization. // In: Proc. of ICML-97, 14th International Conf. On machine Learning — Nashville, USA, 1997. — pp. 412-420.

[6] F. Sebastiani, Machine Learning in Automated Text Categorization. ACM Computing Surveys, vol.1,pp. 1-47, 2002.

[7] А. Максаков. Сравнительный анализ алгоритмов классификации и способов представления документов. Труды третьего российского семинара РОМИП'2005, Ярославль, 2004. с.63-73.
[8] А. Максаков. Исследование способов уменьшения набора характеристик в алгоритмах классификации текстов. Труды Всероссийской научной конференции "Методы и средства обработки информации" -М.: Издательский отдел факультета ВМиК МГУ, 2003. - С. 234-240.

Evaluating efficiency of scalable classification algorithms

© A. Maksakov

bruzz@yandex.ru

In this paper two algorithms characterized by low learning and classification computing costs are reviewed. Results of assessments in classification tracks of RIRES-2006 seminar are presented and analyzed.