Сравнительный анализ алгоритмов классификации и способов представления Web-документов

© Максаков Алексей

ВМиК МГУ
bruzz@yandex.ru

Аннотация

Данная статья посвящена двум основным проблемам рубрикации текстов: выбору алгоритма классификации и способам предварительной обработки текста. На основе экспериментов в рамках семинара РОМИП'2005 был проведен сравнительный анализ рассматриваемых подходов и предложены способы решения обнаруженных проблем.

1. Введение

Данная работа проводится в рамках исследования способов обеспечения периодического тематического поиска, что включает в себя и разработку такого рода системы. Одним из основных этапов работы этой системы является этап классификации текстов. Поскольку качество рубрикации во многом определяет качество итогового результата, исследование алгоритмов классификации и способов предварительной обработки документов, является важнейшей задачей.

Алгоритмы классификации работают с некоторой математической моделью представления экземпляра (в данном случае – текстового документа). Наиболее распространенной моделью является представление в виде набора признаков, которое и будет рассмотрено в дальнейшем. Определение признаков и сопоставление им весов является существенно неформальным шагом и во многом влияет на результат классификации. По этой причине этап предварительной обработки документа рассматривается отдельно.

2. Рассматриваемые алгоритмы

В ходе экспериментов были рассмотрены три представителя семейства линейных алгоритмов: метод опорных векторов (SVM) [1], алгоритм PrTFIDF[2] и модифицированный наивный алгоритм Байеса[3].

Алгоритмы PrTFIDF и наивный алгоритм Байеса основаны на статистической модели и имеют много общего. Данные алгоритмы представляют интерес, поскольку они хорошо масштабируемы и обладают высокой производительностью. Известным их недостатком является сравнительно низкая точность классификации, особенно в случае бинарной классификации. Алгоритм PrTFIDF рассматривается в качестве базового алгоритма по ряду причин:

- Экспериментально[2] показана более высокая точность классификации по сравнению с наивным Байесом и TFIDF[4]. Также это было подтверждено и в других экспериментах, выходящих за рамки данной статьи.
- Данный алгоритм применим для анализа большого количества документов и позволяет использовать большое количество признаков. Это важно, поскольку алгоритм планируется применять для обработки больших объемов данных.

Алгоритм Байеса в последнее время оценивается как сравнительно низкокачественный алгоритм. Основными причинами являются проблемы, связанные с принципом независимости признаков и некорректной оценкой априорной вероятности в случае существенно неравномощных обучающих выборок. Предложив ряд эмпирических модификаций алгоритма или добавив дополнительные признаки, например, на основе выбора фраз в документе, можно попытаться решить существующие проблемы и получить более качественный алгоритм, сохранив его простоту, производительность и масштабируемость.

Результаты классификации методом опорных векторов в последнее время оцениваются[5] как лучшие или одни из лучших. Однако, скорость обучения данного алгоритма сравнительно низка ($O(|D|^a$, где $a>1,2$[5])) и он требует большого объема памяти, что снижает его масштабируемость. Тем не менее, данный алгоритм можно использовать в качестве эталона с точки зрения качества классификации. Также были предложены модификации оценки весов признаков, которые будут рассмотрены позднее.

Таким образом, требования к алгоритму классификации в рамках решаемой задачи можно сформулировать следующим образом:

1. Качество классификации должно быть сравнимо с качеством метода опорных векторов.
2. Алгоритм должен обладать низкой вычислительной сложностью и является хорошо масштабируемым.

Рассмотрим предлагаемые модификации к существующим алгоритмам.

2.1 Методика предварительной оценки модификаций

Модификации алгоритмов требуют экспериментальной проверки для оценки их влияния на алгоритм. В рамках данной работы предварительная оценка производилась на двух тестовых наборах: Newsgroup-20[6] и обучающей коллекции нормативных документов. Для второй коллекции в качестве обучающей выборки было выбрано 40% документов случайным образом, остальные документы использовались для оценки точности классификации. Вторая выборка интересна сильной неравномерностью распределения документов по классам и большим количеством классов.

2.2 Модификации наивного алгоритма Байеса

Правило определения класса для документа в алгоритме Байеса можно представить следующим образом:

$$C(d) = \arg\max_{C}[\log(p(C) + \sum_{w \in d} f_w \log p_{Cw}],$$

где f_w - количество вхождений лексемы w в документ,

$$p_{Cw} = p(w \mid C)$$

Для борьбы с некорректным определением априорной условной вероятности признаков в случае неравномощных обучающих выборок, предлагается использовать парадигму класса-дополнения, то есть вместо вероятности принадлежности лексемы классу оценивать вероятность принадлежности лексемы классу-дополнению C’ (следует учесть, что p(w|C) ~ 1-p(w|C’)).Используя принцип сглаживания параметров по Лапласу, получаем следующее правило:

$$C(d) = \arg\max_{C}[\log(p(C) - \sum_{w \in d} f_w \log(\frac{\overline{N}_{Cw} + 1}{\overline{N}_C + |V|})]$$

где $\overline{N}_{Cw}$- количество лексем во всех классах кроме данного, $\overline{N}_C$ -общее количество лексем в классе-дополнении, $|V|$ - размерность словаря лексем.

Следует отметить, что данная эвристика работает только в том случае, если количество классов N >> 2.

Для дальнейшего улучшения качества классификации предлагаются следующие приемы:

- Логарифмическое сглаживание частоты признаков
- Нормализация весов признаков в документе по его длине
- Использование инверсной частоты признака (IDF и IDF'[2])
- Нормализация логарифмов весов признаков ($\log(p_{Cw})$)

Предварительные эксперименты показали улучшение точности классификации при включении всех эвристик, кроме логарифмического сглаживания и использования инверсной частоты. В итоге перед окончательным прогоном алгоритма (далее ModBayes) ухудшающие качество эвристики были отключены. Точность алгоритма при предварительном тестировании оказалась сравнимой с точностью SVM, при этом точность базового алгоритма Байеса была близка к нулю.

2.3 Модификации алгоритма SVM

Рассматриваемая модификация алгоритма сводится к тривиальному эмпирическому изменению оценки веса признаков. Изначальные предпосылки обусловлены следующим:

- Лексемы с высокой инверсной частотой возможно более значимы, и соответственно должны иметь больший вес, аналогично предположениям алгоритма TFIDF.
- Если лексема часто встречается в документах одного класса, но редко в документах другого, то эта лексема также возможно более значима, чем лексема, встречающаяся в малом количестве документов, но во многих классах. В качестве примера можно привести две ситуации: лексема встречается в десяти документах одного класса, а другая по два раза в обоих классах. С точки зрения инверсной частоты вторая лексема будет иметь больший вес, но фактически первая гораздо более значима для качественного разделения двух классов.

Таким образом был предложен следующий модификатор веса лексемы:

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

В ходе предварительных экспериментов на тестовой коллекции Newsgroup-20 применение этой эвристики привело к небольшому увеличению точности классификации. При прогоне алгоритма SVM на коллекции нормативных документов эта эвристика была включена.

3. Предварительная обработка документов

Задачей этапа предварительной обработки документов является выделение признаков документа и сопоставления им весов. В простейшем случае мультиномиальной модели набором признаков документа будет содержащийся в нем набор лексем, а в качестве веса используется количество вхождений лексемы в документ.

Недостатком такого подхода является то, что практически не учитываются особенности естественного языка, а также структурированность документа и связи между документами в случае Web-страниц.

3.1 Обработка текстов на естественном языке

При обработке текста можно выделить несколько этапов:

1. Лексический анализ
2. Морфологический анализ
3. Синтаксический и пост-морфологический анализ
4. Выделение фраз (n-грамм)
5. Устранение стоп-слов

Первые два этапа достаточно очевидны: задачей первого является выделение лексем, а второй на основе набора правил и внутреннего словаря сопоставляет каждой лексеме набор возможных словооснов с их грамматическими характеристиками.

Использование синтаксического анализа позволяет разрешить значительную часть случаев омонимии. Также синтаксический анализ может позволить обеспечить более точную фильтрацию стоп-слов и построение фраз на основе синтаксически связанных лексем, что существенно сокращает их количество по сравнению с полным перебором соседей лексемы.

3.1.1 Синтаксический анализ

Проблемами существующих решений в области синтаксического анализа (например, LinkParser [7], Диалинг[8]) являются достаточно низкая скорость обработки текста и чувствительность к некорректным синтаксическим конструкциям. Эти проблемы следуют из областей применения этих решений – проверка правописания и машинный перевод. В задаче классификации текстов требования к синтаксическому анализу несколько другие: высокая скорость обработки текстов и работа с синтаксически неполными фрагментами текста, при этом допустимо некоторое увеличение погрешности анализа.

Разработанный в ходе работы синтаксический анализатор имеет много общего с алгоритмом, используемым в системе Диалинг. Отличия заключаются в изменении списка правил и существенном упрощении фрагментационного анализа. В результате алгоритм более корректно разбирает синтаксически неполные фрагменты, а скорость обработки текста выросла примерно на порядок.

Результатом работы синтаксического анализатора является устранение морфологических неоднозначностей и построение набора синтаксически связанных фраз. Пост-морфологический анализ позволяет более точно определить часть речи лексемы, соответственно этап фильтрации стоп-слов производится после синтаксического анализа.

3.1.2 Выбор фраз

В предыдущем пункте мы рассмотрели выбор фраз на основе синтаксического анализа. Также существуют алгоритмы выбора фраз, основанные на статистическом анализе. Следует отметить, что при использовании таких алгоритмов в чистом виде анализируется чрезвычайно большое количество фраз, что затрудняет их применение при обработке большого количества документов.

Рассмотрим два базовых алгоритма отбора фраз. Суть первого алгоритма заключается в том, что фразы рассматриваются как некоторый контекст для наиболее весомых лексем в рамках некоторой тематики. Таким образом, фраза считается «контекстной», если она содержит хотя бы один из наиболее весомых термов, предварительно отобранных по обычным алгоритмам отбора признаков.

Второй алгоритм основывается на следующем: если данная фраза является «устойчивой», то среди множества документов, в которых встречаются термы фразы, должно быть и документы, в

которых присутствует фраза. Таким образом, отбор «устойчивых фраз» в рамках некоторой тематики сводится к следующему:

Для каждой фразы рассчитывается количество документов N_p, в которых она встречается.

Затем рассчитывается количество документов N_t, в которых встречаются все термы фразы.

Фраза считается устойчивой, если $N_p * K \geq N_t$, где K – некоторый коэффициент стабильности фразы, определяемый экспериментально.

На практике при анализе большого количества документов приходится совмещать оба алгоритма. Однако, более перспективным представляется совместное использование синтаксического отбора фраз и фильтрации фраз на основе принципа “устойчивости”.

3.2 Обработка Web-страниц

Используемый при тестах анализ Web-страниц достаточно прост. В частности, необходимо решать проблему автоматического определения кодировки, поскольку явно она указывается далеко не во всех документах. Решения этой задачи производилось в два этапа:

- Анализ частоты вхождения наиболее часто используемых литер
- В случае, когда частотный анализ не позволяет сделать определенного вывода, производится проверка наличия части выделенных лексем в морфологическом словаре

В случае, если частотный или словарный анализ показали несоответствие кодировки, текст документа конвертируется в другую кодировку.

Также при обработке документа производилось увеличения веса лексем, входящих в заголовки, название, ключевые слова, текст ссылки и т.п.

4. Результаты экспериментов

4.1 Дорожка классификации Web-страниц

В этой дорожке был проведен один прогон. В качестве алгоритма классификации использовался модифицированный алгоритм PrTFIDF. В ходе предварительной обработки текстов использовался

морфологический анализ на основе словарей ISpell и анализ структуры Web-страницы. Основной целью данного прогона было сравнить алгоритм с другими на большом объеме реальных данных.

Полученные результаты оказались хуже, чем у других участников семинара, что во многом объясняется слабостью алгоритма в случае неравномощных обучающих выборок и показывает фактическую неприменимость этого алгоритма для такого рода задач. Это подтвердилось и при тестировании на обучающей коллекции нормативных документов.

4.2 Дорожка классификации нормативно-правовых документов

В рамках дорожки классификации нормативных документов было проведено четыре прогона:

Прогон 1. алгоритм PrTFIDF.

Прогон 2. алгоритм PrTFIDF со статистическим выбором фраз.

Прогон 3. модифицированный наивный алгоритм Байеса с использованием пост-морфологии и частичным выбором фраз

Прогон 4. модифицированный алгоритм SVM с использованием пост-морфологии и частичным выбором фраз

Целью прогонов было определить степень влияния выбора фраз на качество классификации, а также сравнительную оценку алгоритмов PrTFIDF, модифицированного алгоритма Байеса и SVM.

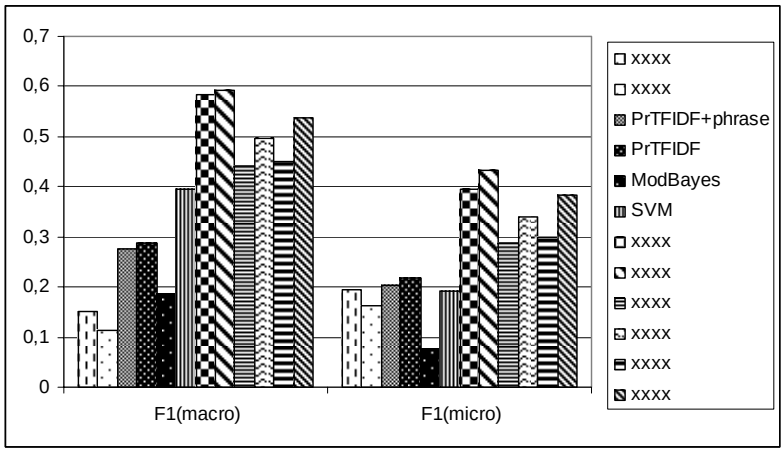

Рис.1 Сравнение качества классификаторов

Результаты прогонов оказались существенно отличными от результатов, полученных в ходе предварительных экспериментов и их достаточно сложно интерпретировать, особенно средневзвешенную по документам оценку F1. В частности, можно обнаружить, что алгоритм PrTFIDF превосходит SVM, что можно объяснить разве что неудачным подбором модификаторов весов. Тем не менее, в случае предварительного тестирования на обучающей выборке коллекции нормативных документов выигрыш SVM был более чем существенным.

Также следует отметить отсутствие выигрыша при использовании выбора фраз, что ставит под сомнение использование статистического выбора в чистом виде. В предварительных экспериментах статистический (на наборе Newsgroup-20) и синтаксический (на обучающей коллекции) выбор фраз показывали улучшение точности для всех анализируемых алгоритмах в диапазоне от 1 до 4%.

Итоговые результаты, в том числе и по модифицированному алгоритму Байеса, существенно расходятся с ожидаемыми, что требует дополнительного исследования.

4.3 Эксперименты на обучающем наборе коллекции нормативно-правовых документов

В ходе анализа результатов классификации был обнаружен ряд слабостей используемых алгоритмов. Для решения найденных проблем был предложен ряд модификаций к алгоритму Байеса, а также предложен алгоритм ModSimpl, основанный на построении нескольких разделяющих гиперплоскостей, соответствующих дискриминанту Фишера.

Эксперименты, проведенные на обучающем наборе нормативных документов (вне рамок семинара РОМИП) показали результаты, позволяющие говорить о перспективности этих алгоритмов.

	NB	PrTFIDF	ModBayes	ModSimpl	SVM
точность	< 10%	< 10%	45,46%	44,54%	47,83%

Таблица 1. Сравнение точности алгоритмов

Использование синтаксического анализа фраз также обеспечило небольшой прирост точности классификации.

5. Дальнейшие направления работы

Анализируя результаты, можно сделать выводы о необходимости более детальной проработки как предварительного анализа документов, в особенности для Web-коллекции, так и самих используемых алгоритмов. Можно выделить следующие направления дальнейшей работы:

1. Доработка вероятностных алгоритмов для решения задачи рубрикации с большим количеством неравномощных классов.
2. Исследование и доработка алгоритма ModSimpl. Данный алгоритм, в отличие от вероятностных,, показал хорошие результаты и при решении задачи бинарной классификации
3. Доработка синтаксического анализатора, учет частей речи и прочих характеристик лексем при сопоставлении весов и отборе признаков
4. Совместное использование синтаксического и статистического выбора фраз
5. Анализ блоков Web-страниц и устранение шумовых элементов
6. Анализ контекста ссылок на данный документ
7. Использование словарей синонимов и, возможно, адаптированного вероятностного латентно-семантического анализа

6. Заключение

В данной работе был рассмотрен ряд алгоритмов классификации и вопросы предварительной обработки текстов. На основе анализа результатов экспериментов был предложен ряд усовершенствований классификаторов и выделены основные направления дальнейшего развития.

Литература

[1] T. Joachims. Making large-scale SVM learning practical// Advances in kernel methods: support vector learning, MIT Press, 1999

[2] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization// Proceedings of ICML-97, 14th International Conference on Machine Learning, pages 143-151 // Morgan Kaufmann Publishers, 1997

[3] 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
[4] G.Salton. Developments in Automatic Text Retrieval// Science, vol 253, pages 974-979, 1991
[5] S. Chakrabarti. Mining The Web Discovering Knowledge From Hypertext Data// Morgan Kaufmann Publishers, 2004
[6] Home Page for 20 Newsgroups Data Set. http://people.csail.mit.edu/jrennie/20Newsgroups/
[7] D. Temperley, J. Lafferty , D. Sleator. Link Grammar Parser http://www.link.cs.cmu.edu/link
[8] А.Сокирко. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ)// Диссертация. http://www.aot.ru/docs/sokirko/sokirko-candid-1.html

On Comparative Analysis of Classification Algorithms and Web documents representation

Alexey Maksakov

Two main problems in text rubrication are reviewed in this article: classification algorithm choice and text preprocessing methods. Based on experiments held on ROMIP'2005 collections, methods were compared and solutions to revealed problems were proposed.