УДК 004.89

ПАРАЛЛЕЛЬНЫЙ ВЫБОР ПАРАМЕТРОВ КЛАССИФИКАТОРА ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ

Е.В. Котельников, Т.А. Пескишева, О.А. Пестов

ФГБОУ ВПО «Вятский государственный гуманитарный университет», г. Киров

Рецензент д-р пед. наук, профессор С.М. Окулов

Ключевые слова и фразы: анализ тональности; метод опорных векторов; метод скользящего контроля; параллельные алгоритмы; текстовая классификация.

Аннотация: Представлен параллельный алгоритм настройки параметров текстового классификатора, основанный на методе скользящего контроля. Экспериментальное тестирование осуществлено на коллекции семинара РОМИП для задачи анализа тональности. В результате проведенных экспериментов определен оптимальный набор параметров классификатора для данной задачи и подтверждена эффективность разработанного алгоритма.

Введение

Многие задачи интеллектуальной обработки текстов, такие как классификация, кластеризация, аннотирование и др., связанные с использованием методов машинного обучения, требуют значительных вычислительных ресурсов. Например, в системах тематической текстовой классификации, в которых применяется метод опорных векторов (Support Vector Machine, **SVM**) [7], весьма трудоемкими являются этапы формирования векторной модели текста и обучения SVM-классификатора [10]. Точность работы классификатора зависит от правильности выбора его параметров, что также требует высоких вычислительных затрат. Одним из подходов, применяемых для преодоления проблемы вычислительной сложности, является параллельная реализация трудоемких алгоритмов.

Котельников Евгений Вячеславович – кандидат технических наук, доцент, старший научный сотрудник кафедры прикладной математики и информатики; Пескишева Татьяна Анатольевна – старший преподаватель кафедры прикладной математики и информатики; Пестов Олег Александрович – ассистент кафедры прикладной математики и информатики, e-mail: oleg.pestov@gmail.com, ФГБОУ ВПО «Вятский государственный гуманитарный университет», г. Киров.

В данной работе предлагается параллельный алгоритм настройки параметров классификатора, используемого для решения задачи анализа тональности текстов.

Задача анализа тональности

Тональность текста – это эмоциональное отношение, выраженное в тексте. Анализ тональности (sentiment analysis) подразумевает наличие определенной шкалы в диапазоне от отрицательной тональности до положительной. Количество значений на этой шкале может быть равно двум (положительная/отрицательная), трем (добавляется нейтральная тональность) или более (выделяются различные степени положительной и отрицательной тональностей).

Задача анализа тональности текста в простейшем случае заключается в автоматическом определении для данного текста значения на шкале тональности. Анализ тональности находит применение во многих сферах, например, в электронных обучающих средах, интерфейсах на естественном языке, маркетинговых исследованиях, мониторинге СМИ и социальных сетей и др.

Активные исследования по данной проблематике за рубежом начались в 2000-х гг. [5]. В России таких работ до последнего времени было крайне мало; только в 2012 году оценка тональности текста была выбрана одной из главных тем конференции по компьютерной лингвистике «Диалог–2012» [8].

В настоящее время в теории анализа тональности имеются существенные пробелы, а соответствующие программные продукты на рынке практически не представлены. Одной из причин, препятствующих развитию методов и инструментов анализа тональности, является трудность выбора параметров классификатора. Связано это, во-первых, с тем, что тональность текста выражается более завуалированными способами, чем тематика, во-вторых, тональность текста в высокой степени зависит от предметной области [5]. Поэтому основным способом настройки параметров является полный перебор, который, при наличии достаточно большого количества параметров и их возможных значений, на практике оказывается неприменим.

В данной работе предлагается для подбора параметров использовать параллельные алгоритмы и многопроцессорные вычислительные системы, а процесс подбора осуществлять на основе скользящего контроля.

Метод скользящего контроля

Среди методов оценки качества обучения классификатора наиболее распространенным и эффективным является метод скользящего контроля (cross-validation) [3]. Целью использования метода в текстовой классификации является оценка качества классификатора при заданных параметрах для данной обучающей коллекции текстов.

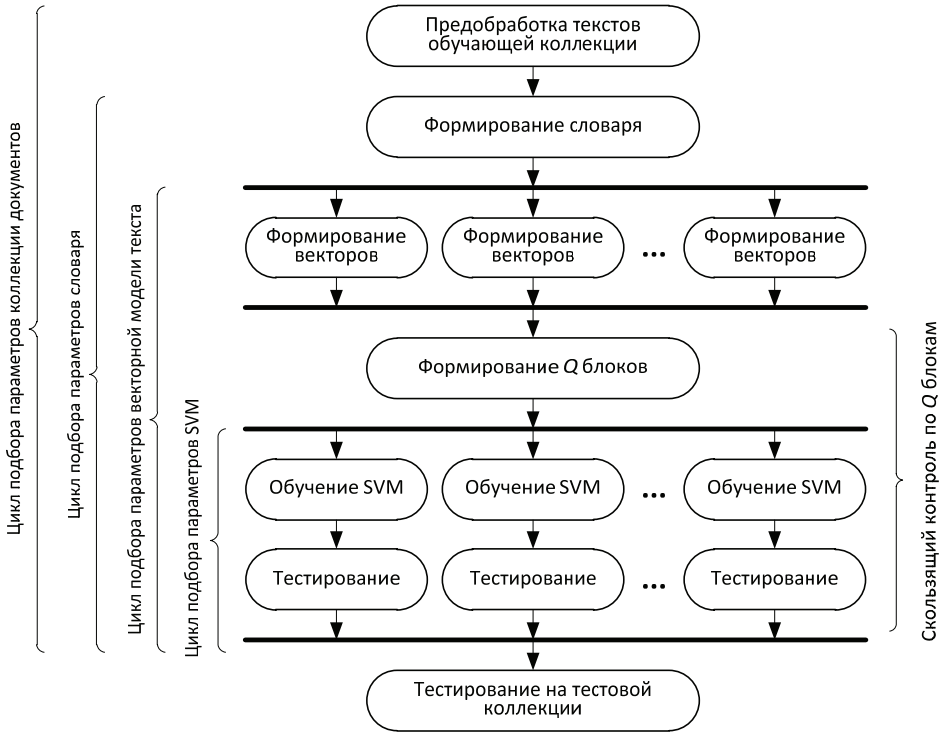
В этом методе обучающая коллекция текстов случайным образом разбивается на Q непересекающихся блоков, приблизительно совпадающих по размеру. Важное условие заключается в сохранении пропорций количества текстов разных классов во всех множествах. Процесс скользящего

контроля по Q блокам (Q-fold cross-validation) включает Q шагов. На каждом шаге один из блоков выбирается в качестве тестового, а классификатор с заданными параметрами обучается на остальных $Q-1$ блоках. Обученный таким образом классификатор проверяется на тестовом блоке, результат проверки – оценка качества – фиксируется. После Q шагов определяется среднее значение Q оценок качества, которое и принимается за результат классификатора при данных параметрах. Имея множество таких результатов для разных наборов параметров, возможно выбрать оптимальный набор параметров.

Параллельный алгоритм настройки параметров

Параллельный алгоритм настройки параметров классификатора основан на методе скользящего контроля по Q блокам (рисунок). Особенностью алгоритма является использование перебора для разных групп параметров: словаря коллекции, векторной модели текста, метода опорных векторов. Для каждой группы параметров используется свой цикл перебора. При этом параллельная реализация была использована для наиболее трудоемких этапов алгоритма – формирования векторной модели и обучения SVM-классификатора.

Результатом работы алгоритма является оптимальный набор параметров классификатора, то есть такой набор параметров, при котором достигаются максимальные значения характеристик качества обучения (точности, полноты, F-меры) [1].



Параллельный алгоритм настройки параметров классификатора

Для *словаря коллекции текстов* настраиваются следующие параметры:

– параметр удаления стоп-слов определяет, удалять или нет стоп-слова (слова, не несущие смысловой нагрузки для задачи тематической классификации);

– группа параметров, отвечающих за учет различных частей речи; эти параметры определяют, удалять или нет (отдельно) существительные, прилагательные, глаголы, наречия, другие части речи;

– параметр, ограничивающий минимально допустимую длину слов;

– параметр, определяющий минимальную частоту слова в текстах коллекции, при которой данное слово попадает в словарь.

Для *векторной модели текста* оптимизируются следующие параметры (в соответствии с подходом TF-IDF) [2, 6]:

– метод локального взвешивания – используются бинарный (**BNRY**), частотный (**FREQ**) и логарифмический (**LOGA**) методы;

– метод глобального взвешивания – используются методы IDF, IDFP, GF-IDF, IGFL и IGFS;

– метод нормализации – либо используется косинусная, либо нормализация отсутствует.

Для *метода опорных векторов* подбираются следующие параметры [7]:

– тип применяемого ядра (линейное, полиномиальное, гауссово);

– регулирующий параметр C;

– для полиномиального ядра – показатель степени;

– для гауссова ядра – параметр γ.

Во время подбора параметров метода опорных векторов для каждого проверяемого набора параметров работает метод скользящего контроля по Q блокам.

По окончании всех циклов определяется оптимальный набор параметров, классификатор обучается на всей обучающей коллекции и тестируется на тестовой коллекции. Итоговым результатом работы, кроме набора оптимальных параметров, становятся характеристики качества обучения для тестовых документов.

Обучающая и тестовая коллекции

Для экспериментов в данной работе в качестве обучающей коллекции использовалась одна из коллекций Российского семинара по оценке методов информационного поиска РОМИП–2011 (http://romip.ru) – набор отзывов пользователей рекомендательного портала Imhonet.ru на различные фильмы [1]. Каждый отзыв имеет оценку автора отзыва от 1 до 10 (чем выше оценка, тем более позитивный отзыв). В данной работе исследовалась задача классификации с двумя значениями на шкале тональности (положительная/отрицательная), поэтому десятибалльная шкала отображалась в двухбалльную: отзывы с оценками от 6 до 10 обозначались как положительные, с оценками от 1 до 5 – как отрицательные. Общее количество отзывов – 14813, из них положительных – 11680, отрицательных – 3133.

Тестовая коллекция в задаче анализа тональности на семинаре РОМИП–2011 представляет собой набор из 329 отзывов на фильмы, оцененных независимо двумя экспертами. Каждый эксперт оценивал отзыв как положительный или отрицательный. Для 312 отзывов оценки совпали и именно эти отзывы использовались в нашей тестовой коллекции (схема оценки AND).

Для оценки эффективности классификатора использовалась основная метрика семинара РОМИП–2011 – F1-мера (F1-*measure*) [1]. Для усреднения результатов по классам применялся подход макро-усреднения (*macro-averaging*) [1].

Особенности реализации

Эксперименты проводились на многопроцессорной вычислительной системе (два 6-ядерных процессора Intel Xeon 2,66 ГГц, 48 Гбайт ОЗУ). Программная реализация была выполнена на языке C#; для обучения SVM-классификатора использовалась библиотека LIBSVM [4]. Морфологический анализ выполнялся при помощи программы *Mystem* версии 2.0 компании Яндекс [9]. Для реализации параллельных этапов работы алгоритма применялся класс *System.Threading.Tasks.Parallel* платформы Microsoft .NET Framework 4.0.

Результаты экспериментов

В результате экспериментов в процессе работы алгоритма настройки параметров на обучающей коллекции был получен набор оптимальных параметров классификатора, включающий три группы параметров, рассмотренных ниже. Затем классификатор с полученными оптимальными параметрами обучался на обучающей коллекции и тестировался на тестовой коллекции. Все оценки качества далее приводятся для тестовой коллекции.

Параметры словаря коллекции:

– результаты исследования влияния частей речи на качество классификации приведены в таблице, из которой видно, что наибольшее влияние

Влияние частей речи на качество классификации

Существительные (12145 слов)	Прилагательные (4338 слов)	Наречия (781 слов)	Глаголы (5173 слов)	Прочие (245 слов)	F1, %
+	+	+	+	+	65,49
–	+	+	+	+	59,07
+	–	+	+	+	64,17
+	+	–	+	+	61,58
+	+	+	–	+	65,75
+	+	+	+	–	62,62

на оценку тональности оказывают существительные и наречия, а наименьшее – глаголы;

– удаление стоп-слов, не несущих смысловой нагрузки для задачи определения тематики, приводит к небольшому снижению (2 %) качества классификации в задаче анализа тональности;

– слова длиной менее 4 символов не оказывают влияния на качество классификации. При удалении слов длиной 4 символа качество существенно падает (на 9 %);

– низкочастотные слова, которые встречаются менее 5 раз во всей обучающей коллекции, не влияют на качество классификации. Если удаляются слова с частотой встречаемости равной 5, качество ухудшается на 2 %.

Параметры векторной модели текста.

Оптимальная комбинация методов локального и глобального взвешивания оказалась следующей: бинарный метод без вычисления глобальных весов с косинусной нормализацией.

Применение инверсной документной частоты (Inverse Document Frequency) также демонстрирует хорошие результаты, но при комбинации с бинарным взвешиванием оказывается несколько хуже (на 2 %), чем использование только бинарных весов.

Параметры метода опорных векторов.

В результате экспериментов наилучшие результаты показало линейное ядро. Для классификации на два класса использовались два различных классификатора, параметры которых подбирались независимо. Оказалось, что оптимальные значения регулирующего параметра C отличаются для разных классов: для текстов отрицательной тональности $C = 10$, для текстов положительной тональности $C = 1$.

Таким образом, набор оптимальных параметров, подобранных при помощи параллельного алгоритма настройки параметров для коллекции отзывов по фильмам, оказался следующим:

− для классификации следует использовать все части речи (возможно, кроме глаголов), а также стоп-слова;

− слова длиной менее 4 символов и частотой встречаемости в коллекции менее 5 можно не учитывать;

− для взвешивания следует использовать бинарный метод с косинусной нормализацией;

− в SVM-классификаторе нужно выбирать линейное ядро с различными регулирующими параметрами C для разных классов: для отрицательной тональности $C = 10$, для положительной тональности $C = 1$.

Классификатор с приведенными значениями параметров обеспечил меру $\mathrm{F1} = 65{,}75\ \%$.

В процессе экспериментов, кроме выбора оптимальных параметров оценивалась также эффективность параллельного алгоритма. В среднем значение ускорения для 12-ядерной вычислительной системы равно 9, что дает эффективность (как отношение ускорения к количеству ядер) равную 0,75. Применение многопроцессорной вычислительной системы и парал-

лельного алгоритма позволило сократить временные затраты на выбор оптимальных параметров с 90 до 10 часов.

Заключение

Таким образом, в работе был предложен параллельный алгоритм настройки параметров классификатора. Для тестирования алгоритма использовались коллекции задачи анализа мнений Российского семинара по оценке методов информационного поиска РОМИП–2011. Применение параллельного алгоритма для настройки параметров классификатора позволило, во-первых, определить оптимальные значения параметров для задачи анализа тональности, во-вторых, подтвердить эффективность разработанного алгоритма.

В дальнейших исследованиях авторы предполагают использовать более обширные коллекции, в том числе англоязычные, а также применять вычислительные системы с распределенной памятью, что потребует некоторой коррекции алгоритма.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект 12-07-97000-р_поволжье_а).

Список литературы

1. Chetviorkin, I.I. Sentiment Analysis Track at ROMIP 2011 / I.I. Chetviorkin, P.I. Braslavskiy, N.V. Loukachevitch // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог», Бекасово, 30 мая – 3 июня 2012 г. В 2 т. Т. 2. Доклады специальных секций / Рос. гос. гуманитар. ун-т. – М., 2012. – Вып. 11(18). – С. 1–14.

2. Chisholm, E. New Term Weighting Formulas for the Vector Space Method in Information Retrieval : Technical Report Number ORNL-TM-13756 / E. Chisholm, T.G. Kolda. – Oak Ridge National Laboratory, Oak Ridge, TN, March 1999 [Электронный ресурс]. – Режим доступа : http://130.203.133.150/ viewdoc/versions?doi=10.1.1.40.3899. – Загл. с экрана.

3. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection / R. Kohavi // 14th International Joint Conference on Artificial Intelligence / Palais de Congres Montreal. – Quebec, Canada, 1995. – P. 1137–1145.

4. LIBSVM – A Library for Support Vector Machines [Электронный ресурс]. – Режим доступа : http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (дата обращения 01.12.2011). – Загл. с экрана.

5. Pang, B. Opinion Mining and Sentiment Analysis [Электронный ресурс] / B. Pang, L. Lee. – Режим доступа : http://www.nowpublishers.com/product.aspx?product=INR&doi=1500000011. – Загл. с экрана.

6. Salton, G. Term-Weighting Approaches / G. Salton, C. Buckley // In Automatic Text Retrieval. Information Processing & Management. – 1988. – Vol. 24, No. 5. – P. 513–523.

7. Vapnik, V. Statistical Learning Theory / V. Vapnik. – New York : Wiley, 1998. – 740 p.

8. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог», Бекасово, 30 мая – 3 июня 2012 г. Вып. 11(18). В 2 т. / редкол.: А.Е. Кибрик (гл. ред.) [и др.]. – М. : Изд-во Рос. гос. гуманитар. ун-та, 2012. – 2 т.

9. Морфологический анализатор Mystem от компании Yandex [Электронный ресурс]. – Режим доступа: http://company.yandex.ru/technology/mystem (дата обращения 01.12.2011). – Загл. с экрана.

10. Пескишева, Т.А. Параллельный алгоритм обучения текстового классификатора для многопроцессорной системы с иерархической архитектурой / Т.А. Пескишева, Е.В. Котельников, О.А. Пестов // Вопр. соврем. науки и практики. Ун-т им. В.И. Вернадского. – 2011. – № 3(34). – С. 105–112.

Parallel Choice of Classifier Parameters for Text Sentiment Analysis

E.V. Kotelnikov, T.A. Peskisheva, O.A. Pestov

Vyatka State University of Humanities, Kirov

Key words and phrases: cross-validation; parallel algorithms; sentiment analysis; support vector machines; text categorization.

Abstract: The paper presents the parallel algorithm of adjustment of parameters of text classifier based on cross-validation. Experimental testing was held on text collection of seminar ROMIP for sentiment analysis task. As the result of the experiment optimal set of parameters of classifier for given task was selected and the effectiveness of the introduced parallel algorithm was proved.
