КЛАССИФИКАЦИЯ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ КОМБИНИРОВАННОГО ПОДХОДА

Васильев В. Г. (vvg_2000@mail.ru),
Давыдов С. Ю. (davydov_sergey@hotmail.com)

ООО «ЛАН-ПРОЕКТ», Москва, Россия

В работе проводится сравнительный анализ эффективности использования различных подходок к классификации отзывов пользователей. Рассматриваются варианты комбинированного использования различных подходов. Приводятся результаты экспериментов в рамках соответствующей дорожки РОМИП 2012.

Ключевые слова: анализ отзывов пользователей, классификация, фрагментные правила

SENTIMENT CLASSIFICATION BY COMBINED APPROACH

Vasilyev V. G. (vvg_2000@mail.ru),
Davydov S. Yu. (davydov_sergey@hotmail.com)

LAN-PROJECT, Moscow, Russia

In this paper various approaches to sentiment classification are compared. Known and new combined approaches are described. Training sets, evaluation metrics and experiments are used according to ROMIP 2012 sentiment analysis track.

Keywords: sentiment analysis, classification, fragment rules

1. Введение

В настоящее время в связи с активным развитием социальных сетей, форумов и блогов вопросы автоматизации анализа мнений пользователей сети по различным вопросам (отношение к товарам и услугам, событиям,

высказываниям, сообщениям) вызывают большой интерес у многих организаций, что приводит к активизации научных исследований и экспериментов в данной области. В 2011 году в рамках семинара РОМИП-2011 и конференции Диалог-2012 были открыты соответствующие дорожки по оценке мнений пользователей о книгах, фильмах и цифровых камерах [1]. Рассматривались ситуации с классификацией на 2, 3 и 5 классов.

Участниками использовались подходы к построению классификаторов, как основанные на обучении на примерах, так и основанные на задании правил вручную. Наилучшие результаты при этом были достигнуты при применении первого подхода. Соответствующие методы характеризуются следующими основными структурными элементами: моделью представления текстов, видом информационных признаков, методами вычисления весов признаков, методами снижения размерности, моделями и методами классификации. В следующей таблице приводится сводное описание вариантов элементов, которые были выбраны в качестве лучших различными участниками.

Таблица 1. Варианты задания элементов классификаторов при классификации отзывов

Название компоненты	**Варианты**
Модели представления текстов	Теоретико-множественная модель (Bag of Words)
Виды информационных признаков	Все слова Оценочные слова Полярные слова Знаки пунктуации Полярные слова
Методы вычисления весов признаков	BNRY-COSN (бинарные признаки с косинусной нормализацией) TF-IDF TF-IDF с глобальными весами от другого массива TF-IDF с весами для оценочных слов
Методы снижения размерности	Не используется
Методы классификации	Машины опорных векторов с линейным ядром

Целью настоящей работы является исследование эффективности использования стандартных и комбинированных методов классификации, в которых совместно используется несколько методов или подходов. Оценка эффективности рассматриваемых методов производится в рамки дорожки классификации отзывов пользователей о книгах на два класса и дорожки классификации прямой и косвенной речи в новостных сообщениях.

2. Описание используемых подходов

Для задания правил в данной работе применяется подход, описанный в работе [8].

Операции для задания правил можно разбить на следующие группы: элементарные — выделяют фрагменты, соответствующих отдельным словам; сложные — выделяют сложные многословных выражений; определяющие — задание общих понятий и множеств; управляющие — задают параметры классификации и обучения на примерах.

В настоящее время разработано большое количество алгоритмов машинного обучения для решения задач классификации текстов. В данной работе был решено провести тестирование следующих стандартных алгоритмов[10]:

- алгоритм к-ближайших соседей;
- алгоритм построение деревьев решений C4.5;
- алгоритм на основе машин опорных векторов;
- байесовский классификатор на основе смеси многомерных нормальных распределений;
- байесовский классификатор на основе смеси распределений фон Мизеса-Фишера;
- центроидный классификатор Роччио.

Общая схема алгоритма обучения в данном случае является достаточно стандартной и имеет следующий вид.

Обучение классификатора на примерах

- Формирование векторного представления текстов в рамках модели «BagOfWords».

1. Снижение размерности (селекция признаков по частоте) и вычисление весов признаков (TF_IDF).
2. Обучениеи оценка классификатора на обучающей выборкес использование 5-шаговой процедуры кросс-проверки.

3. Эксперименты

3.1. Новости

3.1.1. Описание тестовых массивов и показателей качества

Эксперименты по оценке качества проводились в рамках дорожки РОМИП 2012 классификации тональности прямой и косвенной речи в новостных сообщениях на 3 класса: положительный, отрицательный и нейтральный (нет оценки).

Данная дорожка содержала один обучающий массив текстов: массив прямых и косвенных речей из новостей, 4260 текстов, каждый текст оценен одной из четырёх оценок: +, −, +−, 0.

Для тестирования использовался набор из 124 648 текстов, содержащих прямую речь из новостных сообщений. Задачей дорожки было отнести каждый текст к классу положительных, отрицательных, либо нейтральных отзывов. Оценка качества классификации тестового набора производилась путем экспертной оценки 4573 случайно выбранных сообщений из него.

Для оценки качества работы классификаторов в настоящей работе использовались следующие стандартные показатели качества: макро-точность, макро-полнота, макро-F1-мера и аккуратность.

3.1.2. Результаты экспериментов

Эксперименты проводились в два этапа. На первом этапе была выполнена самооценка качества классификации с использованием обучающего множества текстов, предоставленного организаторами дорожки. На втором этапе была выполнена обработка тестового множества текстов с использованием отдельных классификаторов и получены оценки качества от организаторов дорожки.

В следующих таблицах приведены результаты самооценки качества с использованием метода перепроверки.

Таблица 2. Показатели качества классификации с использованием всех признаков

Метод	Макро-Точность	Макро-Полнота	Макро-F1
Gmm	0,55	0,55	0,55
Knn	0,54	0,51	0,51
Mixvmfs	**0,82**	**0,83**	**0,82**
Mixberns	0,72	0,72	0,71
Roccio	0,47	0,34	0,15
Svm	**0,94**	**0,95**	**0,94**

Таблица 3. Показатели качества классификации с использованием всех существительных, прилагательных, глаголов, наречий, причастий, отобранных правилами

Метод	Макро-Точность	Макро-Полнота	Макро-F1
Gmm	0,53	0,53	0,52
Knn	0,54	0,51	0,51
Mixvmfs	**0,83**	**0,83**	**0,83**
Mixberns	0,72	0,72	0,72
Roccio	0,34	0,33	0,13
Svm	**0,94**	**0,94**	**0,94**

Таблица 4. Показатели качества классификации с использованием всех существительных, прилагательных, наречий, причастий, отобранных правилами

Метод	Макро-Точность	Макро-Полнота	Макро-F1
Gmm	0,52	0,51	0,51
Knn	0,53	0,50	0,49
Mixvmfs	**0,79**	**0,80**	**0,79**
Mixberns	0,68	0,68	0,67
Roccio	0,47	0,33	0,14
Svm	**0,92**	**0,93**	**0,92**

Таблица 5. Показатели качества классификации с использованием всех существительных, отобранных правилами

Метод	Макро-Точность	Макро-Полнота	Макро-F1
Gmm	0,51	0,51	0,51
Knn	0,52	0,50	0,50
Mixvmfs	**0,75**	**0,76**	**0,75**
Mixberns	0,65	0,65	0,64
Roccio	0,51	0,34	0,14
Svm	**0,89**	**0,90**	**0,89**

В результате самопроверки наилучшие результаты показали алгоритмы svm и mixvmfs при классификации с использованием всех признаков и при классификации с использованием всех существительных, прилагательных, глаголов, наречий, причастий, отобранных правилами. В связи с этим, было решено использовать при классификации тестового множества именно эти методы и комбинированный классификатор, состоящий из всех методов.

Результаты экспериментов по классификации тестового множества приведены в следующей таблице, в которую включена наилучшая оценка по дорожке и результаты оценки качества для 6 прогонов:

- Q1 — комбинированный автоматический классификатор с использованием всех признаков;
- Q2 — классификатор фон Мизеса-Фишера (mixvmfs) с использованием всех признаков;
- Q3 — классификатор машин опорных векторов (SVM) с использованием всех признаков;
- Q4 — комбинированный автоматический классификатор с использованием правил;
- Q5 — классификатор фон Мизеса-Фишера (mixvmfs) с использованием правил;
- Q6 — классификатор машин опорных векторов (SVM) с использованием правил.

Таблица 6. Результаты оценки качества

Метод	Макро-Точность	Макро-Полнота	Макро-F1	Аккуратность
Q1	0,58	0,56	0,57	0,57
Q2	0,56	0,56	0,56	**0,58**
Q3	0,57	0,56	0,56	0,57
Q4	**0,58**	**0,57**	**0,57**	0,57
Q5	0,56	0,54	0,55	0,57
Q6	0,57	0,56	0,56	0,57
xxx-4	**0,63**	**0,62**	**0,62**	**0,62**
Baseline	0,14	0,22	0,20	0,41

В целом все алгоритмы, участвовавшие в тестировании показали примерно одинаковые результаты. Наилучшее качество среди отправленных методов показал комбинированный алгоритм классификации с использованием всех существительных, прилагательных, глаголов, наречий, причастий в качестве признаков, отобранных правилами.

3.2. Книги

3.2.1. Описание тестовых массивов и показателей качества

Эксперименты по оценке качества проводились в рамках дорожки РОМИП 2012 классификации отзывов на два класса. Данная дорожка содержала обучающий массив текстов: массив отзывов о книгах из 24 159 текстов, предоставленных онлайновой службы рекомендаций IMHONET, каждый отзыв оценен по 10 бальной шкале.

Для тестирования использовался набор из 60 737 текстов, содержащих описание различных объектов интереса пользователей. Задачей дорожки было отнести каждый текст к классу положительных, либо к классу отрицательных отзывов.

Для оценки качества работы классификаторов в настоящей работе использовались следующие стандартные показатели качества: макро-точность, макро-полнота, макро-F1-мера и аккуратность.

3.2.2. Результаты экспериментов

Результаты экспериментов по классификации тестового множества приведены в следующей таблице, в которую включена наилучшая оценка по дорожке и результаты оценки качества для 6 прогонов:

- Q1 — комбинированный автоматический классификатор с использованием всех признаков;
- Q2 — классификатор фон Мизеса-Фишера (mixvmfs) с использованием всех признаков;
- Q3 — классификатор машин опорных векторов (SVM) с использованием всех признаков;
- Q4 — классификатор mixberns с использованием всех признаков;
- Q5 — байесовский классификатор на основе смеси многомерных нормальных распределений GMM с использованием всех признаков;

- Q6 — регрессионный классификатор с использованием всех признаков.

Таблица 7. Результаты оценки качества

Метод	Макро-Точность	Макро-Полнота	Макро-F1	Аккуратность
Q1	**0,67**	**0,75**	**0,70**	**0,82**
Q2	0,59	0,68	0,63	0,48
Q3	0,61	0,70	0,65	0,74
Q4	0,43	0,50	0,46	**0,87**
Q5	0,48	0,49	0,49	0,81
Q6	0,50	0,51	0,51	0,32
xxx-17	**0,75**	**0,68**	**0,71**	**0,88**
Baseline	0,43	0,50	0,47	0,87

В целом результаты экспериментов показывают, что качество классификации данной дорожки по показателю Аккуратность находится на уровне Baseline, что связано с сильным отличием размеров классов (87% наблюдений относятся к одному классу). Макро оценки не учитывают размер классов и по ним алгоритмы классификации уже оказываются выше базового уровня.

Наилучшее качество среди отправленных методов показал комбинированный алгоритм с использованием всех признаков, в котором для каждой рубрики был выбран свой метод классификации.

4. Выводы

Таким образом, в настоящей работе рассмотрены несколько подходов к классификации отзывов пользователей.

Литература

1. *Chetviorkin I. I., Braslavski P. I., Loukachevitch N. V.* (2012) SENTIMENT ANALYSIS TRACK AT ROMIP 2011. Computational Linguistics and Intellectual Technologies:Papers from the Annual International Conference “Dialogue” (2012), Issue 11, Volume 2 of 2, 2012, pp. 1–14.
2. *Vasilyev V. G.* (2011) Fragment extraction and text classification by logical rules [Klassifikatsija i vydelenie fragmentov v tekstah na osnove logicheskih pravil] Digital libraries: Advanced Methods and Technologies, Digital Collections RCDL’2011, Voronezh, 2011, pp. 133–139.
3. *Vasilyev V. G.* (2008) Complex technology of automatic text classification [Kompleksnaja tehnologija avtomaticheskoj klassifikacii tekstov]. Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2006” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2008”]. Bekasovo, 2008, pp. 83–90.