КЛАССИФИКАЦИЯ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ КОМБИНИРОВАННОГО ПОДХОДА

Васильев В. (vvg_2000@mail.ru)

Давыдов С. (davydov_sergey@hotmail.com)

ЛАН-ПРОЕКТ

Элементы классификаторов

Название компоненты	Варианты
Модели текстов	Теоретико-множественная модель, деревья синтаксического анализа, комбинированные модели
Виды признаков	Все слова, Оценочные слова, Полярные слова, Знаки пунктуации, Хэш-теги, Валентности, Биграммы, Смайлы
Веса признаков	BNRY-COSN, TF-IDF, Delta TF-IDF
Снижения размерности	Селекция признаков, трансформация признаков
Методы классификации	SVM, Rule Based, k-NN, Combined SVM, байесовский классификатор

Классификация на основе правил 2011

Шаг 1. Проверка начала текста

Шаг 2. Проверка всего текста

Шаг 3. Сравнение весов положительных и отрицательных выражений

#define Good
(не ^:2 @@badmark) ^ : (не ^:2 @@goodmark) & ^ (не ^:2 @@isbad)

@@isbad ^ : ((не $DoubleQuotes) ^:2 @@isgood)

(не ^:2 (@@badmark @@isbad)) :\s ($comma ^:\s("не только" ^:\s ("$comma но" :10 (@@isgood @@goodmark))))

#set isgood
(
удовлетворяет :\s потребности
потрясающ*
прекрасн* :1^ день
великолепн*
"очень доволен" &7^ ("на тот момент" тогда "в то время")
"очень хорошая"
без ^:3 отличн*
советую :\s ^ (другие присматривать)
"цена соответствует качеству"
замечательный
поражать
чудесный
незаменимый
шедевр
увлекательный
остроумно
…

Информационные признаки

все слова - выделение всех слов и именных словосочетаний без учета их контекста употребления (с учетом стоп-словаря и списка частей речи);

контекстные правила – правила на специальном языке для отбора всех слов заданных частей речи с учетом наличия рядом оценочных слов из заданного словаря, усилительных слов, отрицаний;
(@@COND ^:5\s (@@NEG ^:5\s (@@INTENS ?:5\s ($Noun $Adj $Adv $Прич $Verb)))

оценочные слова – в правилах на специальном языке используются только слова из заранее построенного специального словаря из 1000 оценочных слов и словосочетаний для которых также учитываются ситуации наличия отрицательных, условных и усилительных слов;

комбинированный – использование объединенного множества признаков, включающего все слова и контекстные правила.

Информационные признаки

ОБЫЧНЫЕ – отбираются все слова заданных частей речи, перед которыми нет отрицательных и усилительных слов;

ОЧЕНЬ – отбираются все слова заданных частей речи, перед которыми нет отрицательных и есть усилительные слова;

НЕ - отбираются все слова заданных частей речи, перед которыми есть отрицательные и нет усилительных слов;

ОЧЕНЬ НЕ - отбираются все слова заданных частей речи, перед которыми есть отрицательные и есть усилительные слова.

Пример правила для книг

```
@@COND ^:5 (((@@NEG :5\s (@@INTENS :5\s
($Adj $Verb $Noun $Adv)))
&5\s ? @@OBJECT) >> ^ @@OBJECT)

#set OBJECT (
    книга     произведение
    роман     язык    роль
    идея     дочитать
    страница
    читать     чтение
    литература     произведение
    сочинение     прочтение
    прочитать     писать
    сюжет     образ
    глава     эпиграф
    написать     эпилог     повесть
)
```

Веса и снижение размерности

Вычисление весов
(TF,BNRY)-(IDF,IDFS)-(COSN)

Снижение размерности
- отбор признаков по частоте встречаемости – учитывались признаки, которые встречаются не менее чем в 2 текстах и не более чем в половине текстов;
- селекции признаков с использованием метода Хи-квадрат – отбирались первые 3000 признаков с наибольшими значениями веса;
- снижение размерности с использованием метода LSI – использовалось 100 факторов (данный метод использовался только в комбинации с методом на основе смеси вероятностных анализаторов главных компонент).

Методы классификации

SVMLIN - классификатор на основе машин опорных векторов с линейным ядром

SVMREG - регрессия на основе машин опорных векторов с линейным ядром

KNN - классификатор к-ближайших соседей

ROC - центроидный классификатор Роччио

BERN - байесовский классификатор на основе многомерных распределений Бернулли

GMM - байесовский классификатор многомерных нормальных распределений

VMFS - байесовский классификатор распределений фон Мизеса-Фишера

Параметры методов классификации

Method	**Features**	**Dimension Reduction**	**Parameter estimation**
VMF	TF-IDF	Document frequency	Robust estimate
GMM	TF-IDF	Document frequency, LSI	Robust estimate PPCA model
KNN	TF-IDF	Document frequency	Neighbors - 5
SVM	TF-IG	Document frequency	Linear
ROC	TF-IDF	Document frequency	Standard method
TREE	BNRY - IDF	Document frequency, Information Gain	C4.5

Решающие правила

Оценка новостей на три класса

1. Два бинарных классификатора

$$d(u) = \begin{cases} 1, & d_{pos}(u) > d_{neg}(u), \\ 2, & d_{neg}(u) > d_{pos}(u), \\ 3, & d_{pos}(u) = d_{neg}(u), \end{cases}$$

Оценка отзывов о книгах на два класса

1. Один классификатор для положительных текстов

$$d(u) = d_{pos}(u).$$

2. Два классификатора:

$$d(u) = \begin{cases} 1, & d_{pos}(u) > d_{neg}(u) | d_{pos}(u) = d_{neg}(u), w_{pos}(u) > w_{neg}(u), \\ 2, & d_{neg}(u) > d_{pos}(u) | d_{pos}(u) = d_{neg}(u), w_{pos}(u) \leq w_{neg}(u), \end{cases}$$

3. Логистическая регрессия

$$d_\tau(u) = \begin{cases} 1, & r(u) > \tau, \\ 2, & r(u) \leq \tau, \end{cases}$$

Оценка качества классификации новостей

Метод	Макро-Точность	Макро-Полнота	Макро-F1	Аккуратность
Q1 (COMB-ALL)	**0.57**	**0.56**	**0.57**	**0.57 (0.56, 0.59)**
Q2 (VMFS-ALL)	0.56	0.56	0.56	0.58 (0.57, 0.60)
Q3 (SVM-ALL)	0.57	0.56	0.56	0.57 (0.56, 0.59)
Q4 (COMB-RULE)	0.58	0.57	0.57	0.57 (0.56, 0.59)
Q5 (VMFS-ALL)	0.56	0.54	0.55	0.57 (0.55, 0.58)
Q6 (SVM-ALL)	0.57	0.56	0.56	0.57 (0.56, 0.58)
xxx-4	0.63	0.62	0.62	0.62 (0.60, 0.63)
Baseline	0.14	0.22	0.20	0.41 (0.40, 0.42)

Оценка качества классификации отзывов о книгах

Метод	Макро-Точность	Макро-Полнота	Макро-F1	Аккуратность
Q1 (COMB-ALL)	**0.67**	**0.75**	**0.70**	**0.82**
Q2 (VMFS-ALL)	0.59	0.68	0.63	0.48
Q3 (SVM-ALL)	0.61	0.70	0.65	0.74
Q4 (BERN-ALL)	0.43	0.50	0.46	0.87
Q5 (GMM-ALL)	0.48	0.49	0.49	0.81
Q6 (REG-ALL)	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

Оценка вероятностей ошибок экспертов

	Отрицательные	Положительные
Точность	1%-88%	12%-99%
Полнота	1%-91%	11%-98%

Очистка обучающей выборки от ошибок

Выбор состава положительных и отрицательных примеров

1. POS =1-6, NEG = 8-1

Очистка обучающей выборки от ошибок

1. Обучение на всем массиве примеров
2. Фильтрация примеров в зависимости от расстояния от гиперплоскости
3. Обучение на очищенном массиве примеров

Варианты порогов для фильтрации

Q1 – [0;2];

Q2 – $(-\infty; 0) \cup (2; +\infty)$;

Q3 – (0;+∞);

Q4 – среднее расстояние до гиперплоскости ± 1.

Оценка качества классификации отзывов о фильмах

Метод	Макро-Точность	Макро-Полнота	Макро-F1	Аккуратность
Q1	**0.70**	**0.73**	**0.71**	**0.81**
Q2	0.60	0.53	0.52	0.80
Q3	0.67	0.68	0.68	0.80
Q4	0.67	0.68	0.68	0.79
xxx-19	**0,70**	**0,72**	**0,71**	**0,81**
Baseline	0.40	0.50	0.45	0.81

Выводы

1. Имеются значительные неоднозначности в исходной оценке, что не позволяет достигнуть максимальных значений точности и полноты.
2. Требуется проведение перепроверки результатов экспертов и оценки степени совпадения их результатов.

3. Комбинированные методы предпочтительнее простых методов на правилах и на обучении на примерах на всех признаках.
4. Очистка обучающих примеров позволяет улучшить качество классификации.