Национальный исследовательский университет «МЭИ»

Кафедра прикладной математики

Выпускная работа студента гр. А-13-08 Бочарова Ивана
на тему:
«Исследование и разработка методов классификации новостных текстов»

Руководитель работы: д.т.н., проф. Фальк В.Н.

Научный консультант: асс. Шаграев А.Г.

Москва, 2012

Цели и задачи

Целью данной работы является разработка модификации одного из классических методов классификации

Задачи:

- Исследование постановок задачи классификации, методов решения, способов оценки качества классификации
- Усовершенствование одного из классических методов
- Исследование качества классификации, получаемого при использовании разработанной модификации метода и его сравнение с уже имеющимися реализациями методов

План

1. Постановка задачи классификации
2. Метрики качества классификации и способы оценки качества классификации
3. Обзор методов классификации
4. Усовершенствованный метод
5. Вычислительные эксперименты
6. Заключение

План

1. Постановка задачи классификации
2. Метрики качества классификации и способы оценки качества классификации
3. Обзор методов классификации
4. Усовершенствованный метод
5. Вычислительные эксперименты
6. Заключение

Неформальная постановка задачи классификации

Пусть:

- X – множество классифицируемых объектов
- Y – конечное множество классов

Предполагается наличие целевой зависимости – отображения $y^*: X \to Y$, значения которой известны только на документах конечной обучающей выборки

$$X^l = \{< x_i, y_i > | y_i = y^*(x_i)\}_{i=1}^{l}$$

Требуется:

Построить решающую функцию $a : X \to Y$, способную классифицировать любой объект $x \in X$.

Вероятностная постановка задачи

Пусть:

X – множество классифицируемых объектов, Y - конечное множество классов,

На множестве $X \times Y$ определена функция плотности распределения:

$$p(< x, y >) = P(y)p(x|y)$$

Имеется конечная обучающая выборка

$$X^l =< x_i, y_i >_{i=1}^{l}, X^l \in (X \times Y)^l$$

Вероятности появления объектов каждого из классов $P_y = P(y)$ называются вероятностями классов. Плотности распределения $p_y(x) = p(x|y)$ называются функциями правдоподобия классов.

Необходимо:

- построить эмпирические оценки вероятностей классов $P(y)$ и функций правдоподобия $p_y(x)$
- построить классификатор $a : X \to Y$, минимизирующий вероятность ошибочной классификации.

Описание объектов

Ситуация, когда объекты используются для классификации в их первоначальном виде, довольно редка. Чаще всего формируется некоторое признаковое описание объекта.

Признак – результат измерения некоторой характеристики объекта.

Формально: $f: X \to D_f$, где D_f - множество допустимых значений признака.

Выделяют следующие типы признаков:

- бинарные ($D_f = \{0; 1\}$),
- номинальные (D_f - конечное),
- порядковые (D_f - конечное упорядоченное множество),
- количественные ($D_f = \mathbb{R}$).

План

1. Постановка задачи классификации
2. Оценка качества классификации
3. Обзор методов классификации
4. Усовершенствованный метод
5. Вычислительные эксперименты
6. Заключение

Метрики качества классификации

- Точность: $\rho = \frac{\sum_{d \in D}[a(d)=c \wedge y^*(d)=c]}{\sum_{d \in D}[a(d)=c]}$
- Полнота: $\pi = \frac{\sum_{d \in D}[a(d)=c \wedge y^*(d)=c]}{\sum_{d \in D}[y^*(d)=c]}$
- F-мера: $F_\beta = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho}, \beta^2 \in (0, +\infty)$

Усреднение метрик

- Макроусреднение

$$\pi_{macro} = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TP_i}{TP_i + FN_i}, \rho_{macro} = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TP_i}{TP_i + FP_i}$$

- Микроусреднение

$$\pi_{micro} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i}, \rho_{micro} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FP_i}$$

В данной работе усреднение производится методом макроусреднения, так как этот метод чувствителен к ошибкам классификации на малых классах

Скользящий контроль

Оценкой скользящего контроля по q разбиениям называется величина

$$CV_q(\mu, X^n) = \frac{1}{q}\sum_{i=1}^{q} Q(\mu(X^n \setminus X_i^n), X_i^n),$$

где:

- $X^n = X_1^n \sqcup X_2^n \sqcup \cdots \sqcup X_q^n$ – случайное разбиение выборки на q непересекающихся подмножеств мощности m;
- μ – метод обучения (отображение, ставящее в соответствие любой обучающей выборке решающую функцию $a : X \to Y$);
- Q – функционал качества.

Оценка скользящего контроля является случайной величиной, значение которой зависит от разбиения обучающей выборки.

Процедуру скользящего контроля также используют для построения доверительных интервалов, например:

$$P\left(Q\left(\mu\left(X^{l-m}\right), X^m\right) < \min_{1 \le i \le q} Q(\mu(X^n \setminus X_i^n), X_i^n)\right) = \frac{1}{q+1}$$

План

1. Постановка задачи классификации
2. Метрики качества классификации и способы оценки качества классификации
3. Обзор методов классификации
4. Усовершенствованный метод
5. Вычислительные эксперименты
6. Заключение

Наивный байесовский классификатор

Наивный байесовский классификатор – это один из методов решения задачи в вероятностной постановке.

Работа метода основана на теореме Байеса и («наивном») предположении о том, что признаки, которыми описывается объект, являются независимыми.

Достоинства метода:

- требуется малое количество данных для обучения
- высокая скорость работы
- легкость внесения в метод разного рода изменений

Байесовское решающее правило с использованием принципа максимизации апостериорной вероятности

$$a(d) = \arg\max_{c \in C} P(c|d)$$

Для вычисления $P(c|d)$ используют формулу Байеса:

$$P(c|d) = \frac{p(d,c)}{p(d)} = \frac{p_c(d)P_c}{\sum_{s \in C} p_s(d)P_c}$$

Для применения решающего правила необходимо получить оценки значений $p_c(d)$ и P_c

Оценки вероятностей в задаче классификации текстов

Для оценки вероятностей классов используется величина

$$\widehat{P}(c_i) = \frac{|c_i|}{|D|}, \text{ где}$$

$|c_i|$ – число документов, принадлежащих категории c_i, а $|D|$ – общее число документов в выборке.

В силу наивного предположения для оценки значений $p_c(d)$ в задаче классификации текстов необходимо оценить только значения $p(w_i|c)$, так как:

$$p(d|c) = p\left(w_1, \dots, w_i, \dots, w_{n_d}\middle|c\right) = p(w_1|c) \times \cdots \times p(w_i|c) \times \cdots \times p(w_n|c)$$

Их значения оцениваются по формуле:

$$\hat{p}(w_i|c) = \frac{T_{cw_i}}{\sum_{w'} T_{cw'}}, \text{ где:}$$

T_{cw_i} - число вхождений слова w_i в документы из обучающего множества, принадлежащие категории c.

Переход к суммированию

$$a(x) = \arg\max_{c \in C} p_c(d) P_c = \arg\max_{c \in C} (\ln P_c + \ln p_c(d))$$

$$= \arg\max_{c \in C} (\ln P_c + \ln \prod_{i=1}^{n_d} p(w_i|c))$$

$$= \arg\max_{c \in C} (\ln P_c + \sum_{i=1}^{n_d} \ln p(w_i|c))$$

Метод k ближайших взвешенных соседей

Метрический метод классификации. Предполагается , что близкие в смысле функции расстояния объекты принадлежат к одному классу.

Введем пороговую функцию :

$$[P] = \begin{cases} 1, P \\ 0, \bar{P} \end{cases}$$, где P – некое условие

Метод относит классифицируемый объект к тому классу, суммарный вес представителей которого среди k ближайших объектов является максимальным:

$$a(u) = \arg\max_{c \in C} \sum_{i=1}^{k} [c_u^{(i)} = c]\, w_i \ ,$$

где $c_u^{(i)}$ - категория, к которой принадлежит $d_u^{(i)}$ - i-й сосед объекта u.

Обычно: $w_i = q^i, q \in (0; 1)$

Машина опорных векторов (*SVM*)

Работа метода основана на понятии оптимальной разделяющей гиперплоскости.

Задача формулируется следующим образом: можем ли мы найти такую гиперплоскость, чтобы расстояние от нее до ближайшей точки было максимальным?

Если такая гиперплоскость существует, то она нас будет интересовать больше всего, она называется оптимальной разделяющей гиперплоскостью.

Достоинства метода:

- Обучение *SVM* сводится к задаче квадратичного программирования, допускающей эффективное вычисление единственного решения задачи;
- Решение обладает свойством «разреженности» – положение гиперплоскости определяется только небольшой частью выборки (именно они и называются опорными векторами);
- При помощи введения функций ядра этот метод изящно обобщается на случай нелинейных разделяющих поверхностей.

План

1. Постановка задачи классификации
2. Метрики качества классификации и способы оценки качества классификации
3. Обзор методов классификации
4. Усовершенствованный метод
5. Вычислительные эксперименты
6. Заключение

Базовый метод

В качестве базового метода был выбран наивный байесовский классификатор. Данный метод используется для решения задачи в вероятностной постановке.

Работа с новостными текстами ведется в рамках модели «мешок слов». В качестве признаков, описывающих документы, выбраны количества вхождений слов $w_i \in W$ в документ.

В задаче классификации текстов наивное допущение не является сильным, и его использование позволяет достигать высоких результатов.

Сглаживание вероятностей

Вообще говоря, непонятно, как оценивать значение $p(w_i|c)$, если слово w_i ни разу не встречалось в документах обучающей выборки.

Обычно поступают так. Предполагают существование некоторой априорной вероятности появления какого-либо слова. Рассмотрим, как применяется сглаживание в данной работе:

$$\hat{p}(w_i|c) = \frac{T_{cw_i}+\alpha}{\sum_{w'} T_{cw'}+|W|\alpha}$$, где:

$\alpha > 0$ – параметр сглаживания, $|W|$ - количество различных слов, встречавшихся в документах обучающей выборки

Предполагается, что каждое слово встречается хотя бы раз в каждом из документов выборки. После этого априорная вероятность корректируется в соответствии с содержанием документа.

Специфика метода

- Сделана попытка в явном виде учесть следующую особенность новостных текстов. Обычно новостная статья имеет очень содержательное начало, а вот к концу статьи ее содержательность может снижаться.

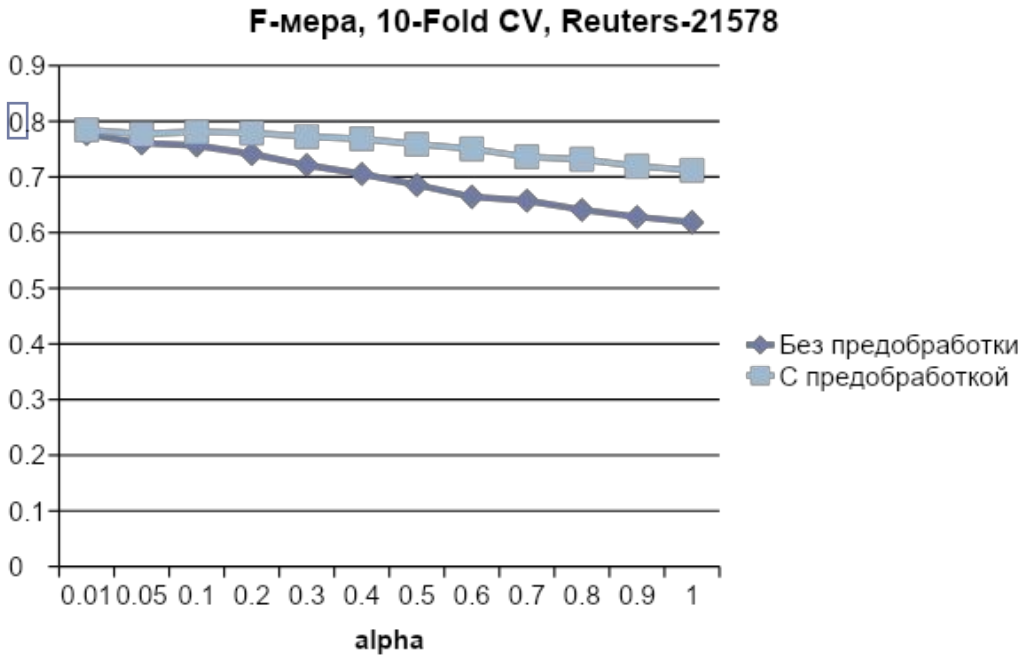
Предполагается, что, если проводить классификацию, скажем, по первым 150, 100, 50 и т.д. словам, а не по полному тексту, качество классификации ухудшится незначительно.

Так, w – число слов, по которым проводится классификация, становится еще одним параметром метода, наряду с α (параметром сглаживания)

План

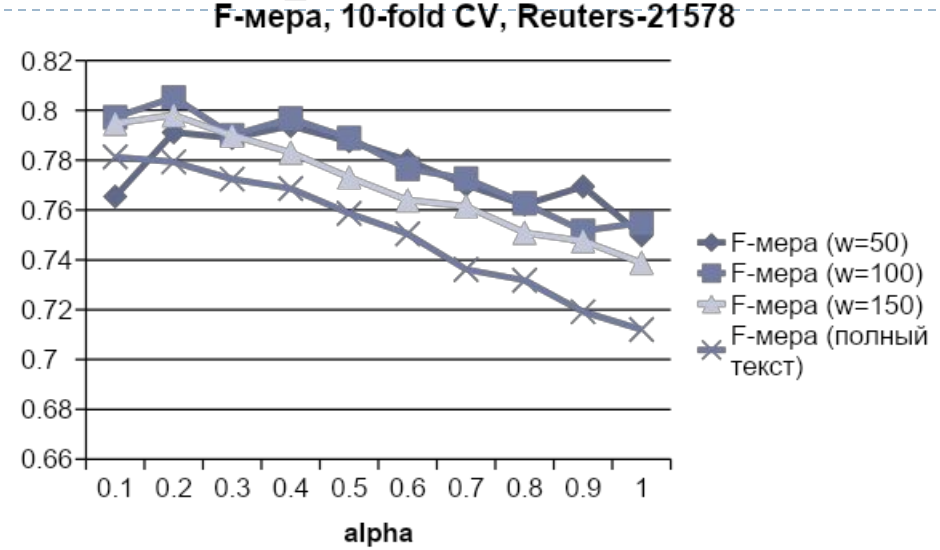
1. Постановка задачи классификации
2. Метрики качества классификации и способы оценки качества классификации
3. Обзор методов классификации
4. Усовершенствованный метод
5. Вычислительные эксперименты
6. Заключение

Эксперимент 1. Подбор параметра α. Влияние предобработки.



- Предварительная обработка (стемминг, удаление стоп-слов) положительно влияют на качество классификации
- В случае предварительно обработанных текстов, как и в случае необработанных текстов, рекомендуется выбирать значения параметра α, лежащие в окрестности точки 0.05, так как в окрестности этой точки полнота и точность классификации практически совпадают и при этом достаточно велики (около 0.8)

Эксперимент 2. Подбор параметра w



Выводы:

- Использование при классификации только начальной части текста (к примеру, первых 100 слов) улучшает качество классификации (особенно, при удачном выборе параметра α)
- Данный эксперимент показывает, что, проводить ли классификацию по первым 50 словам документа или по всему документу, практически не имеет значения. Было подтверждено предположение о высокой значимости начала новостной статьи
- Проведение классификации только по началу документа позволяет сократить время работы метода примерно на четверть (на стандартном разбиении *Reuters-21578* при $w = 50$ метод работал 6,53 секунды против 8,5 при полнотекстовой классификации)

5. Вычислительные

Эксперимент 3. Сравнение метода с *kNN (Reuters-21578)*

Метод	Точность	Полнота	F1-мера
	0,832	0,781	0,805
	0,792	0,678	0,75

Данные по методам *kNN* и *NewsNB* получены при помощи 10-кратного скользящего контроля.

Разработанная модификация метода работает лучше , чем метод k ближайших взвешенных соседей.

Эксперимент 4. Сравнение метода с *SVM(Reuters-21578, 20 Newsgroups)*

Метод	Точность	Полнота	F-мера	Время работы, с
SVM	0,795	0,636	0,6702	4,14
	0,915	0,896	0,908	7,25

Reuters-21578

Метод	Точность	Полнота	F-мера	Время работы, с
SVM	0,74	0,695	0,714	132,4
	0,816	0,810	0,813	148,33

20Newsgroups

- Разработанная модификация метода работает не хуже выбранной реализации *SVM*
- Использование только линейного ядра серьезно ухудшает качество работы алгоритма *SVM*
- Выбранная реализация SVM может работать быстрее разработанного метода по ряду причин:
  - При оценке времени работы авторского метода учитываются временные затраты на выделение признаков из текстов
  - Используемая реализация *SVM* написана а языке C, а авторский метод реализован на более «медленном» языке Python

План

1. Постановка задачи классификации
2. Метрики качества классификации и способы оценки качества классификации
3. Обзор методов классификации
4. Усовершенствованный метод
5. Вычислительные эксперименты
6. Заключение

Заключение

Основным результатом работы является разработанная модификация наивного байесовского классификатора.

Помимо этого:

- Изучена одна из возможных формальных постановок задачи классификации – вероятностная постановка.
- Проведено исследование алгоритмов классификации и методов предварительной обработки текста.
- Проведено достаточно большое количество вычислительных экспериментов, результаты которых подтверждают качество разработанного метода и позволяют говорить о том, что метод применим на практике.
- Разработан программный комплекс на ЯП Python, который позволяет проводить предварительную обработку текстов и осуществлять классификацию текстов при помощи модификации наивного байесовского классификатора.

Спасибо за внимание!