электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.054/[[004.023+004.8]/[005:[62::811]]]

Сравнение эффективности некоторых статистических методов классификации на примере технических статей

***Васнецов А.Г.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»

Научный руководитель: Самарев Р.С., к.т.н, доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»
bauman@bmstu.ru

Введение

Классификация документов – одна из задач информационного поиска, заключающаяся в отнесении документа к одной из нескольких категорий на основании содержания документа. Классификация документов находит свое применение в таких областях как: ограничение области поиска в поисковых системах, автоматического составления аннотации, составление интернет-каталогов, фильтрация спама и т.д.

В данной работе исследовалась возможность применения некоторых наиболее популярных статистических алгоритмов классификации для классификации статей технического характера. Экспериментально оценена точность классификации при различных признаковых описаниях документов.

Задача классификации

Задача классификации определяется следующим образом. Имеется некоторое множество объектов, разделенное произвольным образом на непересекающиеся группы (классы). Для некоторого конечного подмножества объектов данного множества известно к каким классам они принадлежат. Это подмножество называется выборкой. Классовая принадлежность остальных объектов неизвестна. Задача заключается в построении алгоритма, способного классифицировать произвольный объект из исходного множества.

Подготовка выборки

В данной работе в качестве классифицируемого множества объектов

рассматриваются статьи, взятые с Интернет-ресурса http://habrahabr.ru/. Особенностью этого ресурса является то, что каждая статья отнесена самим автором как минимум к одной из множества заранее определенных категорий. В данном случае под категорией понимается множество статей, объединенных общей тематикой. Такая особенность делает возможным автоматизировать процесс подготовки материала для тестирования классификаторов.

В качестве обучающей выборки было взято по 100 статей каждой из следующих пяти категорий:

1) программирование;
2) алгоритмы;
3) гаджеты;
4) информационная безопасность;
5) DIY или сделай сам.

В связи с тем, что указанный Интернет-ресурс позволяет отнести каждую статью к нескольким категориям, а задача классификации подразумевает назначение единственного класса для статьи, то возникает неоднозначность при ее классификации. В этом случае можно говорить о погрешности выборки. Оценим вероятность появления погрешности при классификации. Для этого посчитаем количество статей в выборке, которые относятся более чем к одной категории:

$$c = \sum_{a \in A : |h_a \cap H| > 1} 1 .$$

Где A — выборка.

h_a — множество категорий, ассоциированных со статьей $a, a \in A$.

H — множество всех рассматриваемых категорий (в данном случае 5 вышеперечисленных).

Так как в выборке присутствуют статьи, относящиеся одновременно к нескольким рассматриваемым категориям, то количество уникальных статей в выборке будет меньше чем суммарное количество статей в каждой рассматриваемой категории. Оценим долю погрешности в выборке как отношение количества статей, относящихся более чем к одному классу к общему числу уникальных статей $|A_u| = 451$:

$$E = \frac{c}{|A_u|} = 0.07317 .$$

Эксперимент заключается в сравнении различных метрик для классификации на основе некоторых элементов исходного множества, для которых известна их классовая принадлежность, но которые не присутствуют в обучающей выборке. После классификации результат сравнивается с заранее определенными классами элементов. Считается доля правильно классифицированных элементов. Для проведения эксперимента было взято по 20 статей из каждой категории.

Описание объектов

Для описания объектов-статей использовалось признаковое описание. Признаковое описание — это вектор, составленный из значений фиксированного набора признаков данного объекта. Признаки в общем случае могут иметь различные типы.

Далее не будет делаться различия между объектом и его признаковым описанием. В данной работе были произведены эксперименты с несколькими вариантами признакового описания V :

1) Наличие слова в тексте статьи $V^A = (V_i^A), V_i^A \in \{0,1\}, i = \overline{1..N}$, где N — количества уникальных слов.

2) Количество слов в тексте статьи $V^B = (V_i^B), V_i^B \in , i = \overline{1..N}$, где N — количества уникальных слов.

3) Наличие в тексте статьи множества часто совместно встречающихся слов, полученных алгоритмом FPGrowth [2] $V^C = (V_i^C), V_i^C \in \{0,1\}, i = \overline{1..N}$, где N — количество множеств часто совместно встречающихся слов.

4) Признаки TF-IDF [4] $V^D = (V_i^D), V_i^D \in , i = \overline{1..N}$, где N — количества уникальных слов.

5) Признаки IDF $V^F = (V_i^F), V_i^F \in , i = \overline{1..N}$, где N — количества уникальных слов.

Приведение слов к исходной форме

При подготовке текстов статей к классификации слова приводились к начальной форме с помощью программы mystem от Yandex [3]. Это делалось для исключения ситуаций, когда одно и о же слово в разных формах считалось бы классификатором разными словами.

Классификация на основе определения ближайшего класса в пространстве признаков

Каждому классу ставится в соответствие точка в пространстве признаков, координаты которой равны математическому ожиданию значения данного признака у объекта, принадлежащего рассматриваемому классу $F : C \to \overline{X}, X_i = M[V_i], V \in C, i = \overline{1 \ldots |V|}$. Классифицируемому объекту также ставится в соответствие точка в этом пространстве, координаты которой равны значению соответствующих параметров этого объекта. Затем рассчитывается расстояние от точки объекта до всех точек классов в какой-либо метрике. Объект считается принадлежащим к тому классу, расстояние до точки которого оказалось минимальным.

Евклидова метрика. При использовании Евклидовой метрики расстояние между точками *x, y* определяется с помощью формулы:

$$d(\bar{x}, \bar{y}) = \sqrt{\sum_{i=1}^{|V|} (y_i - x_i)^2} .$$

Расстояние Махаланобиса — мера расстояния между векторами случайных величин, обобщающая понятие евклидова расстояния. С помощью расстояния Махаланобиса можно определять сходство неизвестной и известной выборки. Оно отличается от расстояния Евклида тем, что учитывает корреляции между переменными и инвариантно к масштабу. Расстояние между точками определяется с помощью формулы:

$$d(\bar{x}, \bar{y}) = \sqrt{(\bar{x} - \bar{y}) S^{-1} (\bar{x} - \bar{y})},$$

где S - ковариационная матрица.

Косинусная мера. Расстояние между точками определяется как 1 минус косинус угла между векторами, идущими из начала координат в данные точки.

$$d(\bar{x}, \bar{y}) = 1 - \frac{\bar{x}\bar{y}}{|\bar{x}||\bar{y}|} .$$

Результаты экспериментов

В ходе эксперимента оценивалась эффективность классификации, где под эффективностью понимается отношение правильно классифицированных тестовых статей, к общему числу тестовых статей. Экспериментальные результаты оценки эффективности использования байесовского классификатора приведены в

нижеследующей таблице.

Пространство признаков	Классификатор на основе поиска ближайшего класса в пространстве признаков		
	Евклидова метрика	Метрика Махаланобиса	Косинусная мера
FPGrowth	0,45	0,41	0,48
Наличие слова	0,62	0,64	0,78
Кол-во слов	0,51	0,61	0,76
TFIDF	0,69	0,59	0,69
IDF	0,69	0,64	0,7

Заключение

Экспериментально показана неэффективность использования пространства признаков на основе множества часто встречающихся слов по сравнению с пространством признаков, основанным на наличии слова независимо от используемой метрики.

Наибольшую точность продемонстрировала косинусная мера в пространстве признаков, основанном на наличии слова.

Список литературы

1. Воронцов К.В. Машинное обучение: курс лекций. Режим доступа: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения: 14.04.2014).
2. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Режим доступа: http://pdf.aminer.org/000/303/388/reducing_the_frequent_pattern_set.pdf (дата обращения 14.04.2014).
3. О программе mystem. Режим доступа: http://company.yandex.ru/technologies/mystem/ (дата обращения 14.04.2014).
4. TF-IDF. Википедия, свободная энциклопедия.. Режим доступа: http://ru.wikipedia.org/wiki/TF-IDF (дата обращения 14.04.2014).
5. K-nearest neighbors algorithm. Википедия, свободная энциклопедия. Режим доступа: http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm (дата обращения 14.04.2014).