Классификация документов в векторном пространстве. Сравнение методов Роккио и метода *k*-ближайших соседей

Матяско А.А.; Хаустов В.А.
Кафедра информатики
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
e-mail: {alexander.matyasko, victor.khaustov}@gmail.com

***Аннотация*—классификации текстов является задачей информационного поиска, заключающаяся в отнесении документа к одной из нескольких категорий. В представленной работе описаны и проводится сравнение метода Роккио и метода *k*-ближайших соседей (kNN-классификация) для задачи классификации документов.**

Ключевые слова: классификация документов, метод Роккио, метод knn, информационный поиск

I. Введение

Одним из основных способов обмена информации является текст. Для работы с большими объемами текстовой информации были разработаны алгоритмы для её анализа и классификации. Задача классификации возникает во многих приложениях (автоматическое определение веб-спама, создание тематических каталогов, фильтрации контента и др.)

В данной работе рассматривается метод Роккио и *k*-ближайших соседей задачи векторной классификации документов и проводится их сравнение.

II. Векторная модель документа

В векторной модели документ представляется в виде вектора, где каждому термину сопоставлен некоторый вес:

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{nj})$$

В основе использования модели векторного пространства лежит гипотеза о компактности (классы образуют компактно локализованные подмножества в пространстве объектов) [1].

Рис.1. Классификация в векторном пространстве

На рисунке 1 представлен пример классификации на три класса в векторном пространстве. Документы с одинаковым классом представлены в виде одинаковых геометрических фигур. Звездочкой обозначен документ, подлежащий классификации.

Векторная классификация сводится к разработке алгоритма, вычисляющего "хорошие" границы, где термин "хорошие" означает высокую точность классификации на данных, не использованных в ходе обучения. Для векторной классификации документов необходимо определить границы между классами, поскольку именно они определяют результат классификации.

III. Метод Роккио

Наиболее известным методом определения границ классов является метод Роккио, в котором для идентификации границ используются центроиды. Центроид класса c вычисляется как усредненный вектор, или центр масс членов класса:

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D} \vec{v}(d) \qquad [2]$$

Здесь D_c — множество документов из пространства D, принадлежащих классу c: $D_c = \{d: <d,c> \in D\}$, $\vec{v}(d)$ — нормализованный вектор документа d.

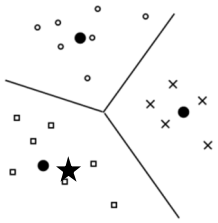

Рис.2. Классификация методом Роккио

Граница между двумя классами в методе Роккио представляет собой множество точек, равноудаленных от двух центроидов. На рисунке 2 центроиды представлены в виде жирных точек. Правило классификации в алгоритме Роккио заключается в определении области, в которую попадает точка. Это эквивалентно поиску центроида $\vec{\mu}(c)$, к которому точка лежит ближе, чем к другим центроидам, и приписыванию этой точки к классу c. На рисунке 3 представлен псевдокод алгоритма Роккио.

```
TrainRocchio(C,D)
1   for each c_j ∈ C
2   do D ← {d: <d,c_j> ∈ D}
3   μ_j ← (1/|D_j|) Σ_{d∈D_j} v(d)
4   return {μ_1, μ_2, ..., μ_j}
ApplyRocchio(C,D)
1   return arg min_j |μ_j − v(d)|
```

Рис. 3. Классификация Роккио. Обучение и тестирование

IV. Метод k-ближайших соседей

Метод *k*-ближайших, или kNN-классификации, является одним из самых изученных и высокоэффективных алгоритмов, используемых при создании автоматических классификаторов.

На основании гипотезы о компактности классифицируемый объект относится к тому классу, к которому принадлежат *k*-ближайшие к нему объекты обучающей выборки. В варианте 1NN каждый документ относится к определенному классу в зависимости от информации о его ближайшем соседе. В варианте kNN документ относится к преобладающему классу ближайших соседей, где k — параметр метода.

Разделяющие границы в методе 1NN представляют собой смежные сегменты диаграммы Вороного. Диаграмма Вороного разделяет плоскость на $|D|$ выпуклых многоугольников, каждый из которых содержит соответствующий документ.

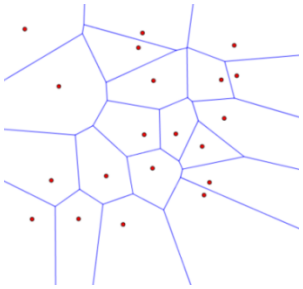

Рис. 4. Диаграмма Вороного и разделяющие границы в методе 1NN

Для произвольного параметра $k \in N$ в методе kNN область пространства, для которого множество *k*-ближайших соседей остается одинаковым, представляет собой выпуклый многоугольник, а пространство оказывается разделенным на выпуклые многоугольники, внутри каждого из которых множество *k*-ближайших соседей является инвариантным.

Метода 1NN не очень устойчив. Классификация каждого тестового документа зависит от класса, к которому относится отдельный обучающий документ, который может иметь неверную метку или вообще быть нетипичным. Метод kNN при k >1 является более устойчивым [3]. Он приписывает документы к превалирующему классу по k ближайшим соседям, случайным образом разрывая связи между ними.

Параметр k в методе kNN часто выбирается на основании опыта и знаний о решаемой задачи классификации. Желательно, чтобы параметр был нечетным, что уменьшить вероятность “ничьей”.

```
Train-knn (C, D)
1   D' ← Preprocess (D)
2   k ← Select-k{C,D'}
3   return D',k
Apply-kNN(C,D',k,d)
1   S ← ComputeNearestNeighbors(D',k,d)
2   for each c_j ∈ C
3   do p_j ← |S_k ∩ c_j|/k
4   return arg max_j p_j
```

Рис. 5. Метод kNN. Обучение и тестирование

V. Сравнение

Классификация Роккио проста в реализации и эффективна по скорости работы, но неточна, если классы далеки от сфер с примерно одинаковыми радиусами.

В классификации по методу kNN не производится оценка ни одного параметра, как в методе Роккио (центроиды).

Временная сложность тестирования в методе kNN линейно зависит от размера обучающего множества и имеет большую временную сложность, чем в методе Роккио. В свою очередь, продолжительность тестирования не зависит от количества классов *J*. Следовательно, при большом количестве классов *J* метод kNN имеет потенциальные преимущества.

Метод kNN не требует явного обучения и допускает использование обучающего множества в процессе классификации без предварительной обработки. Если обучающее множество велико, то метод kNN лучше справляется с несферическими и другими сложными классами, чем метод Роккио [4].

В таблице 3 приведены значения эффективности методов Роккио и kNN для десяти самых больших категорий, а также усредненное значение по всем 90 категориям коллекции Reuters-21578.

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, “Introduction to Information Retrieval”, Cambridge University Press, 2008.

[2] Thorsten Joachims, “A probabilistic analysis of the Rocchio algorithm with tfidf for text categorization”, Springer, 1997.

[3] Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Springer, 2001.

[4] Joachims Thorsten, “Learning to Classify Text Using Support Vector Machines”, Springer, 2002.

Табл. 1. Оценки временной сложности метода Роккио

Обучение	$\Theta(\|D\|L_{ave} + \|C\|\|V\|)$
Тестирование	$\Theta(L_a + \|C\|M_a) = \Theta(\|C\|M_a)$

Табл. 2. Оценки временной сложности классификатора kNN

Обучение	$\Theta(\|D\|L_{ave})$
Тестирование	$\Theta(L_a + \|D\|M_{ave}M_a) = \Theta(\|D\|M_{ave}M_a)$

Табл. 3. Результаты сравнения методов

	Роккио	kNN
earn	96,1	97,8
acq	90,7	91,8
grain	79,5	82,6
crude	81,5	85,8
trade	77,4	77,9
interest	72,5	76,7
ship	83,1	79,8
wheat	79,4	72,9
corn	62,2	71,4
Усреднее	79,9	82,6