В. В. Азарко

СРАВНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ

Рассматривается задача классификации текстов. Описываются критерии сравнения классификаторов. Приводятся результаты эксперимента по сравнению методов.

Введение

Классификация текстов является одной из основных задач компьютерной лингвистики, так как к ней сводятся некоторые другие задачи: определение темы текстов, автора текста, эмоциональной окраски и др. Среди методов классификации текстов, разработанных на данных момент, выделяют группу методов, основанную на алгоритмах машинного обучения. Так как это достаточно большая группа методов, актуальным является вопрос выбора лучшего из них.

I. Постановка задачи

Формально постановку задачи классификации можно записать следующим образом. Имеются множество документов $D = d_1, \ldots, d_n$ и множество возможных категорий (классов) $C = c_1, \ldots, c_k$. Неизвестная целевая функция задается формулой:

$$F(c_i, d_j) = \begin{cases} 1, d_j \in c_i \\ 0, d_j \notin c_i \end{cases}$$

Требуется построить классификатор F‘, максимально близкий к F.

II. Критерии сравнения методов решения задачи

Основным критерием при оценке качества классификации является комбинация точности и полноты. Точность классификации в пределах класса – это доля найденных классификатором документов, действительно принадлежащих данному классу, относительно всех документов, которые система отнесла к этому классу. Полнота классификации – это доля найденных классификатором документов, действительно принадлежащих классу, относительно всех.

Пусть TP – это истинно положительное решение; TN – это истинно отрицательное решение; FP – ложно положительное решение; FN – ложно отрицательное решение. Тогда точность вычисляется по формуле:

$$p = \frac{TP}{TP + FP}$$

Полнота вычисляется следующим образом:

$$r = \frac{TP}{TP + FN}$$

III. Эксперименты по сравнению методов

В статье [1] описаны результаты многих экспериментов по сравнению вышеописанных классификаторов. Результаты представлены в таблице 1.

Таблица 1 – Результаты эксперимента

Метод	Точность	Полнота
Метод опорных векторов	80-85%	83-87%
Свёрточная нейронная сеть	80-95%	70-85%
Классификатор Байеса	70-91%	80-90%

Особенностью свёрточных нейронных сетей является то, что они требуют большое количество данных для обучения. В статье [2] исследователи несколько источников структурированных текстов. Обучающие выборки содержали от 120 000 до 3 600 000 текстов. В результате их исследования, точность работы классификатора, основанного на свёрточных нейронных сетях, достигла 95%. Из таблицы 1 можно заметить, что оценки классификатора Байеса являются сопоставимыми с результатами двух других методов. Однако разброс этих оценок в различных экспериментах достаточно велик.

IV. Выводы

В соответствии с результатами различных исследований, наилучшими методами для классификации текста по критериям точность и полнота являются свёрточные нейронные сети и метод опорных векторов. Остальные методы могут показывать сопоставимые результаты, однако их точность и полнота сильно отличается в различных экспериментах.

1. Батура, Т. В. Методы автоматической классификации / Т. В. Батура // Программные продукты и системы. 2017. Т. 30. № 1. С. 85–99
2. Character-level convolutional networks for text classification. / Xiang Zhang [et al.] // NIPS 2015. Montreal, Canada, 2015.

Азарко Владислав Вячеславович, магистрант кафедры информационных технологий автоматизированных систем БГУИР, azarkovlad@gmail.com.

Научный руководитель: Гуринович Алевтина Борисовна, заместитель декана ФИТУ, кандидат технических наук, доцент, gurinovich@bsuir.by