В. В. Азарко

СРАВНЕНИЕ АЛГОРИТМОВ ВЕКТОРИЗАЦИЯ ТЕКСТА В СИСТЕМАХ КЛАССИФИКАЦИИ

Рассматривается задача векторизации текстов в системах классификации. Сравнивается подход к классификации текста на основе подхода Word2Vec и Bag of Words.

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. Первым этапом решения задачи классификации текстов является представление текста в виде вектора. От качества решения данной задачи зависит время и качество решения задачи классификации целиком. Среди методов векторизации текстов, разработанных на данных момент, самыми распространёнными являются Word2Vec и Bag of Words. В связи с важностью качественного представления текста в виде вектора, актуальным является вопрос выбора лучшего из них.

I. Постановка задачи

Формально постановку задачи векторизации можно записать следующим образом. Имеются текстовый документ состоящий из множества слов $D = d_1, \ldots, d_n$. Необходимо построить функцию, которая преобразует множество слов D в числовой вектор $X = x_1, \ldots, x_n$.

II. Критерии сравнения методов решения задачи

Для оценки качества векторизации текстовых документов в системах классификации можно оценить качество работы алгоритмов классификации при использовании исследуемых алгоритмов векторизации. Для выполнения сравнения использовались метод опорных векторов и классификатор Байеса. Основным критерием при оценке качества классификации является комбинация точности и полноты. 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 с использованием подхоов Word2Vec и Bag of Words. Точность и полнота работы классификаторов с использованием подхода Bag of Words описана в таблице 1, с использованием Word2Vec – в таблице 2.

Таблица 1 – Результаты эксперимента с использованием алгоритма Bag of Words

Метод	Точность	Полнота
Метод опорных векторов	80-85%	83-87%
Классификатор Байеса	80-95%	70-85%

Таблица 2 – Результаты эксперимента с использованием алгоритма Word2Vec

Метод	Точность	Полнота
Метод опорных векторов	85-89%	89-92%
Классификатор Байеса	80-85%	74-80%

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. Использование алгоритмов векторизации значительно влияет на качество работы классификаторов, однако для каждого конкретного классификатора целесообразно выбирать подходящий алгоритм классификации.

1. A Text classification problem and features set / Polyakov I.V. // Vestn. NGU 2015,

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

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