МОДИФИЦИРОВАННЫЕ СТРАТЕГИИ ВЗВЕШИВАНИЯ TF-IDF ДЛЯ ТЕРМИНОВ ПРИ КЛАССИФИКАЦИИ ПУБЛИКАЦИЙ В МЕДИЦИНСКИХ ЖУРНАЛАХ

Сенчин Д.М.[1], Кайда А.Ю.[2]
[1]НИ ТПУ, ИШИТР, 8ПМ21, e-mail: dms14@tpu.ru
[2]НИ ТПУ, ИШИТР, ст. преподаватель ОИТ, e-mail: ayk13@tpu.ru

Введение

Управление массивами текстовых данных и интеллектуальный анализ обычно полагаются на технологию автоматической классификации текста [1]. Взвешивание терминов является основной проблемой при классификации текста и напрямую влияет на точность классификации. Поскольку традиционная TF-IDF (частота терминов и обратная частота документа) не полностью эффективна для классификации текста, обычно данный метод дополняют другими метриками. В этой статье проводятся сравнительные исследования различных схем взвешивания, основанных на классах документов рассматриваемых терминов, TF-IDF-D и TF-IDF-CF [2].

Целью данной работы является реализация и сравнение модифицированных алгоритмов TF-IDF для взвешивания терминов при классификации текстов.

Описание корпуса

Все документы предварительно обрабатываются и собираются в корпус P размерностью $d \times t$, где d – количество документов корпуса, t – количество уникальных терминов. Корпус P имеет c классов.

В рамках исследования был составлен корпус из введений 300 англоязычных исследовательских статей на медицинскую тематику с ресурса ScienceDirect. Общее количество слов корпуса составляет 87057 слов. В составленном корпусе выделено 5 классов по 60 документов, описывающих следующие темы исследований: cancer (онкологические заболевания), covid-19, stroke (сердечные приступы), diabetes (диабет), pneumonia (пневмония).

Предварительная обработка корпуса проводилась с помощью библиотеки nltk и состоит из следующей последовательности [3]:

- разбиение строки на вектор слов;
- перевод в нижний регистр;
- фильтрация знаков пунктуации;
- фильтрация потенциальных числовых данных;
- фильтрация стоп-слов;
- лемматизация.

По итогу была составлена частотная матрица из 300 документов и 4742 уникальных терминов.

Описание алгоритма

Для задач классификации документы должны представляться в подходящей векторной форме, состоящей из весов вида *W (i, j)*, где *i* – термин *j-го* документа. Данные веса могут быть вычислены по технике TF-IDF, 4 разновидности которой рассматриваются ниже [4].

Классическое представление TF-IDF:

$$W(i,j) = TF(i,j) * IDF(i). \quad (1)$$

Term Frequency (TF) – это отношение количества вхождений термина к общему количеству слов в документе:

$$TF(i,j) = \frac{n_i}{\sum_k n_k}. \quad (2)$$

Inverse Document Frequency (IDF) – инвертированная частота документов, содержащих термин, к общему количеству документов в корпусе:

$$IDF(i) = \log_{10} \frac{D}{d_i}. \quad (3)$$

Реализация TF-IDF из библиотеки scikitlearn использует формулу 1, но IDF вычисляется следующим образом:

$$IDF_{sklearn}(i) = \log_{10}(\frac{D}{d_i}) + 1. \quad (4)$$

Модификация, основанная межклассовой дисперсии, TF-IDF-D:

$$W(i,j) = TF(i,j) * IDF(i) * D(i), \quad (5)$$

$$D(i) = \frac{\sum_c (d_{ci} - M(i))^2}{c}, \quad (6)$$

$$M(i) = \frac{\sum_c d_{ci}}{c}. \quad (7)$$

Данная техника позволяет снизить вес терминов, которые распространены между классами, и повысить вес терминов, сосредоточенных в одном классе.

Модификация, учитывающая классовую частоту, TF-IDF-CF:

$$W(i,j) = TF(i,j) * IDF(i) * CF(i,j), \quad (8)$$

$$CF(i,j) = \frac{d_{ci}}{D_c}. \quad (9)$$

Техника TF-IDF-CF позволяет снизить вес терминов, которые не распределены внутри класса.

Результаты

Для оценки результатов были проведены с использованием моделей классификаторов из библиотеки scikitlearn [5]. Точность моделей для каждой техники взвешивания представлены в таблице 1 и на рисунке 1.

Таблица 1

Сравнение точности классификаторов

Классификатор	Техника взвешивания			
	Classic TF-IDF	Scikitlearn TF-IDF	TF-IDF-D	TF-IDF-CF
Multinomial Naive Bayes	20%	20%	**91.7%**	20%
Gaussian Naive Bayes	61.7%	68.3%	71.7%	**95%**
Bernoulli Naive Bayes	55%	55%	55%	55%
K-Neighbors	**96.7%**	**96.7%**	91.7%	**96.7%**
Linear Support Vector	36.7%	46.7%	**96.7%**	45%
Random Forest	68.3%	68.3%	68.3%	**91.7%**
Extra Trees	68.3%	68.3%	68.3%	**88.3%**
Decision Tree	93.3%	93.3%	93.3%	**96.7%**
AdaBoost	95%	95%	95%	**96.7%**

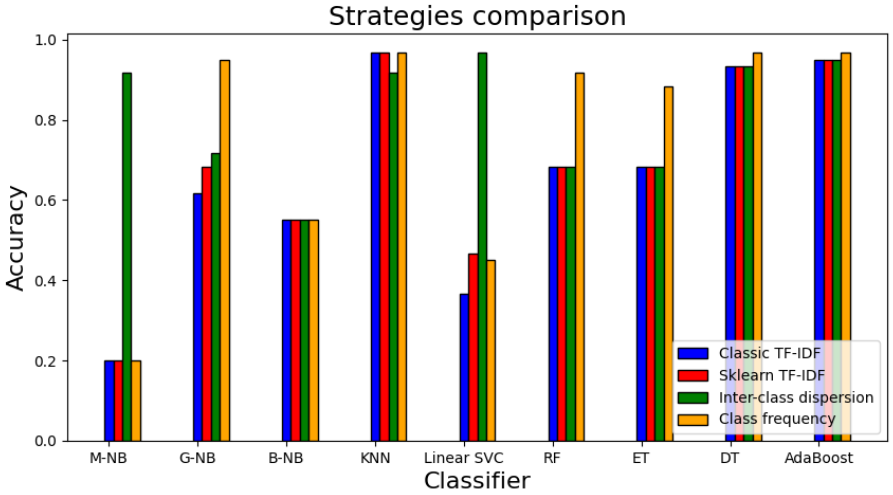

Рис. 1. Сравнение точности классификаторов

Заключение

Из полученных данных можно выявить влияние техники взвешивания точность классификатора. Значительное повышение точности, в сравнении с классическими техниками, наблюдается при использовании модифицированных алгоритмов на моделях: Multinomial Naive Bayes, Gaussian Naive Bayes, Linear Support Vector, Random Forest, Extra Trees.

Список использованных источников

1. Большакова Е.И. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017. — 269 с.
2. Turning from TF-IDF to TF-IGM for term weighting in text classification (2016). https://doi.org/10.1016/j.eswa.2016.09.009
3. Основы Natural Language Processing для текста. [Электронный ресурс]. – URL:https://habr.com/ru/company/Voximplant/blog/446738/ (дата обращения 18.02.2023).
4. Modified TF-IDF Term Weighting Strategies for Text Categorization (2018). DOI:10.1109/INDICON.2017.8487593
5. Блиц-проверка алгоритмов машинного обучения. [Электронный ресурс]. – URL: https://habr.com/ru/post/475552/ (дата обращения 18.02.2023).