УДК 004.9 https://doi.org/10.26160/2618-7493-2021-4-6-8

КЛАССИФИКАЦИЯ ДОКУМЕНТОВ ВУЗА ГИБРИДНЫМ МЕТОДОМ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

Ткаченко А.Л.
Сибирский государственный автомобильно-дорожный университет, г.Омск

Ключевые слова: классификация текстов, обработка текстов, машинное обучение.
Аннотация. Рассмотрен гибридный метод интеллектуального анализа документов вуза, использующий опорные векторы в качестве исходных данных для классификации документов. Описаны этапы классификации документов вуза, а также приведены результаты классификации рассмотренным методом.

CLASSIFICATION OF UNIVERSITY DOCUMENTS BY THE INTELLECTUAL ANALYSIS HYBRID METHOD

Tkachenko A.L.
Siberian State Automobile and Highway University, Omsk

Keywords: text classification, text processing, machine learning.
Abstract. An intellectual analysis hybrid method of university documents, using support vectors as source data for document classification, is considered. The classification stages of university documents are described. The classification results by the considered method are presented.

В настоящее время одной из основных задач инновационного развития высших учебных заведений (вузов) является модернизация процессов документооборота. Отмечается [1], что автоматизация процессов документооборота возможна за счет внедрения технологий роботизации бизнес-процессов (RPA), которые призваны автоматизировать ручной монотонный труд сотрудников вуза. Одной из повседневных задач вуза, которую можно решить средствами RPA, является задача классификации документов, на решение которой сотрудниками вуза ежедневно тратиться большое количество времени. Для решения этой задачи широко используются методы машинного обучения [2].

Предлагается гибридный метод интеллектуального анализа документов вуза, использующий опорные векторы в качестве исходных данных для классификации документов.

Этапы классификации документов предложенным алгоритмом представлены на рисунке 1.

Рис. 1. Этапы классификации документов вуза методом SVM-kNN

На *первом этапе* происходит предварительная обработка документов: все символы текстов документов переводятся в нижний регистр, тексты разбиваются на слова, из текстов документов удаляются шумовые слова (слова, не обладающие семантической нагрузкой), оставшиеся слова текстов приводятся к нормальной форме.

После предварительной обработки документов начинается *второй этап* – числовое представление документов. На данном этапе строится числовая модель документов bag of words, которая представляет документ в виде вектора, состоящего из входящих в документ слов [3]. Каждому слову сопоставляется его вес – количественная оценка значимости слова в документе. Для ее вычисления используется метод TF-IDF, согласно которому слова, наиболее распространенные в конкретном документе и менее распространенные в остальных документах, получают больший вес [4]:

$$W_{t,d} = TF \cdot IDF ,$$

где $TF = n_{t,d} / n_d$ – частота, с которой слово встречается в пределах одного документа, $n_{t,d}$ – число вхождений слова t в документ d; n_d – количество всех слов в документе d; $IDF = \log(|D|/D_t)$ – обратная частота документа, $|D|$ – количество документов в наборе данных; D_t – число документов из набора данных, в которых встречается слово t.

После построения числовой модели текстов на *третьем этапе* набор документов разбивается на обучающую и тестовую выборки. В процессе обучения классификатора используется обучающая выборка в объеме 80% от общего числа документов, остальные 20% используются при тестировании работы классификатора.

Четвертый этап включает построение и обучение классификатора методом опорных векторов (SVM), суть которого состоит в построении гиперплоскости, максимально разделяющей набор текстов на классы [5]. С помощью метода SVM сужается пространство документов, участвующих в классификации новых объектов, до пространства опорных векторов – документов, ближе остальных лежащих к поверхности решений [6]. Для построения гиперплоскости используется радиальная базисная функция Гаусса, которая вычисляется по формуле [5]:

$$K(d_i, d_j) = \exp^{-\gamma\|d_i - d_j\|^2},$$

где d_i и d_j – векторные представления документов; γ – параметр ядра.

Классификация документов происходит на *пятом этапе* методом *k*-ближайших соседей (kNN), суть которого заключается в том, что документу *d* присваивается тот класс *c*, к которому принадлежит большинство из *k* ближайших соседей документа, вычисленных с помощью метрики расстояния. При расчете расстояния в методе *k*-ближайших соседей использована косинусная мера, которая рассчитывается по формуле:

$$k(d_i, d_j) = \frac{d_i d_j}{\| d_i \| \| d_j \|},$$

где d_i и d_j – векторные представления документов.

После построения классификатора происходит его обучение с использованием обучающей выборки документов, начинается *шестой этап* классификации – оценка работы классификатора с использованием тестовой выборки документов.

Для вычисления оценок классификации используется матрица несоответствий, представленная в таблице 1 [4].

Табл. 1. Матрица несоответствий предсказаний классификатора

Класс c_j		Принадлежность классу	
		Положительная	Отрицательная
Оценка классификатора	Положительная	G_p^+	G_p^-
	Отрицательная	G_n^-	G_n^+
G_p^+ – классификатор верно определил класс документа; G_n^+ – классификатор верно определил, что документ не относится к классу c_j; G_p^- – классификатор неверно определил класс документа; G_n^- – классификатор неверно определил, что документ не относится к классу c_j			

В исследовании оценка построенного классификатора происходит по значению *F*-меры – гармонического среднего значения между точностью и полнотой классификации. *F*-мера (J_F) вычисляется по следующей формуле [6]:

$$J_F = 2 \times \frac{J_p \times J_r}{J_p + J_r},$$

где $J_p = \frac{G_p^+}{G_p^+ + G_p^-}$ – точность классификации, показывает долю документов, отнесенных классификатором к рассматриваемому классу, относительно всех документов этого класса [6]; $J_r = \frac{G_p^+}{G_p^+ + G_n^-}$ – полнота классификации, показывает долю документов действительно принадлежащих к рассматриваемому классу относительно всех документов, которые классификатор отнес к этому классу [6].

На основании полученного значения *F*-меры оценивается работа классификатора, делается вывод о приемлемости использования метода классификации.

В качестве исходных данных для автоматической классификации документов вуза использован набор документов отдела организации практики и содействия трудоустройству выпускников СибАДИ. Набор документов состоит из 291 документа, каждый из которых принадлежит к одной из четырех категорий: приказ, распоряжение, письмо, извещение о вакантном месте.

В результате проведенного исследования получено, что рассмотренный метод SVM-kNN классифицирует документы с точностью 96,57%, что сравнимо с точностью классификации аналогами метода. Отсюда следует вывод, что рассмотренный метод может быть использован для решения задачи классификации документов вуза.

Список литературы

1. Ткаченко А.Л., Мещеряков В.А. Анализ технологий роботизации бизнес-процессов // Сборник материалов IV Международной научно-практической конференции «Архитектурно-строительный и дорожно-транспортный комплексы: проблемы, перспективы, инновации». Омск, 2019. С. 536-540.
2. Ткаченко А.Л. Обзор методов интеллектуального анализа документов // Материалы XI Всероссийской научно-практической конференции студентов, аспирантов, работников образования и промышленности «Информационные технологии и автоматизация управления». Омск, 2020. С. 218-227.
3. Xiang Zhang, Junbo Zhao, Yann LeCun Character-level convolutional networks for text classification // Neural Information Processing Systems. 2015. Vol. 28. P. 649-657.
4. Ткаченко А.Л. Решение задачи классификации документов вуза на основе методов интеллектуального анализа // Вестник кибернетики. 2021. №1. С. 12-19.
5. Le Thi Minh Nguyen Text classification based on support vector machine // Dalat University Journal Of Science. 2019. Vol. 9, Iss. 2. P. 3-19.
6. Хайкин С. Нейронные сети: полный курс. 2-е изд. – М.: Изд. дом «Вильямс», 2006. – 1104 с.

Сведения об авторе:
Ткаченко Анастасия Леонидовна – инженер-программист, преподаватель, СибАДИ, Омск.