Сравнительный анализ качества классификации в зависимости от используемой системы признаков

Сфера: текстовый анализ данных

СТУДЕНТ: КОЗЛОВ И. А.
ГРУППА: А-03-19
РУКОВОДИТЕЛЬ ВКР: ПРОФЕССОР ТОЛЧЕЕВ В.О.

ВВЕДЕНИЕ

Технологии Машинного обучения развиваются с каждым годом и становятся всё актуальнее.

Сегодня более 60% населения планеты и более 90% населения России использует интернет. Для комфортного и эффективного пользования всемирной сетью активно применяют технологии обработки естественного языка **(Natural Language Processing).**

- *Разделение web-страниц на разделы*
- *Боты в мессенджерах*
- *Таргетинговая реклама*
- *Сбор статистики*
- *Определение языка текста*
- *Поисковые системы*

ПОСТАНОВКА ЗАДАЧИ

В работе рассматриваются две выборки научных документов, бинарная и многоклассовая, целью является рассмотрение двух признаковых пространств: библиографическое описание (названия, аннотации и ключевые слова) и только названия документов.

Необходимо изучить возможность использования небольших признаковых пространств в случае текстовых данных.

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ТЕКСТОВЫХ ДАННЫХ

В работе предварительная обработка данных состоит из следующих этапов

- Преобразование в нижний регистр
- Удаление знаков препинания
- Удаление стоп-слов
- Токенизация (разделение текста на слова и словосочетания)
- Лемматизация
- Векторное представление слов с помощью TF-IDF взвешивания

$$\boldsymbol{TF - IDF = TF \cdot IDF}$$

$$TF(d,t) = \frac{\text{количество вхождений терма } t \text{ в документе } d}{\text{общее количество термов в документе } d}$$

$$IDF(t) = log\left(\frac{\text{общее количество документов в коллекции}}{\text{количество документов, содержащих терм } t}\right)$$

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

Матрица неточностей

	Оценка эксперта	
Оценка классификатора	Положительная	Отрицательная
Положительная	TP	FP
Отрицательная	FN	TN

На её основе вычисляются следующие метрики

$$Precision = \frac{TP}{(TP + FP)} \qquad F1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Recall = \frac{TP}{TP + FN} \qquad Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Для многоклассовой классификации используют средние значения метрик

macro avg $= \frac{1}{n}\sum_{i=1}^{n} x_i$ **– среднее арифметическое**

weighted avg $= \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ **– среднее взвешенное**

МЕТОДЫ КЛАССИФИКАЦИИ

- **Логистическая регрессия** — статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём его сравнения с логистической кривой.
- **Алгоритм k-ближайших соседей (k-NN).** При классификации объекта алгоритм находит k ближайших к нему объектов из обучающей выборки и присваивает объекту тот класс, который наиболее часто встречается среди этих k ближайших соседей.
- **Деревья решений** — это алгоритм машинного обучения, который используется в задачах классификации и регрессии. Он представляет собой древовидную структуру, в которой каждый узел представляет собой проверку значения одного из признаков объекта, а каждое ребро, исходящее из узла, соответствует одному из возможных значений этого признака. Листья дерева содержат метки классов или числовые значения для задачи регрессии.
- ***Алгоритм случайного леса*** – это алгоритм машинного обучения, который использует ансамбль решающих деревьев для решения задач классификации и регрессии.

ИСПОЛЬЗУЕМОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Python – высокоуровневый язык программирования, имеющий большое число библиотек для машинного обучения.

- **Scikit-learn**
- **Matplotlib**
- **NumPy**
- **Pandas**
- **Pymorphy2**
- **NLTK**

БИНАРНАЯ КЛАССИФИКАЦИЯ

В выборке содержится 3267 элементов, каждый из которых относится к классу «Интеллектуальный анализ данных» (ИАД) или «не ИАД». БО состоят из названия, аннотации, года выпуска, автора и ключевых слов.

Классы сбалансированы

- Количество текстов по теме ИАД 1528
- Количество текстов по теме не ИАД 1739

Размерность признаковых пространств составляет

- По названиям **6393**
- По ключевым словам **12185**
- По библиографическим описаниям **25123**

ПРИМЕР ОБЪЕКТА ВЫБОРКИ

Пример БО:

- Название:

ПЕРСПЕКТИВЫ ВНЕДРЕНИЯ ТЕХНОЛОГИЙ DATA MINING В ТАМОЖЕННУЮ ДЕЯТЕЛЬНОСТЬ

- **Аннотация:**

В статье проведен анализ перспективных направлений внедрения технологий Data Mining в деятельность таможенных органов. Рассмотрены классификационные методы машинного обучения с учителем и без учителя, применение которых может автоматизировать решение сложных задач по отнесению поставок товаров к рисковым или выявлению потенциальных рисков. Особое внимание уделено кластерному анализу и программным платформам, которые поддерживают его реализацию.

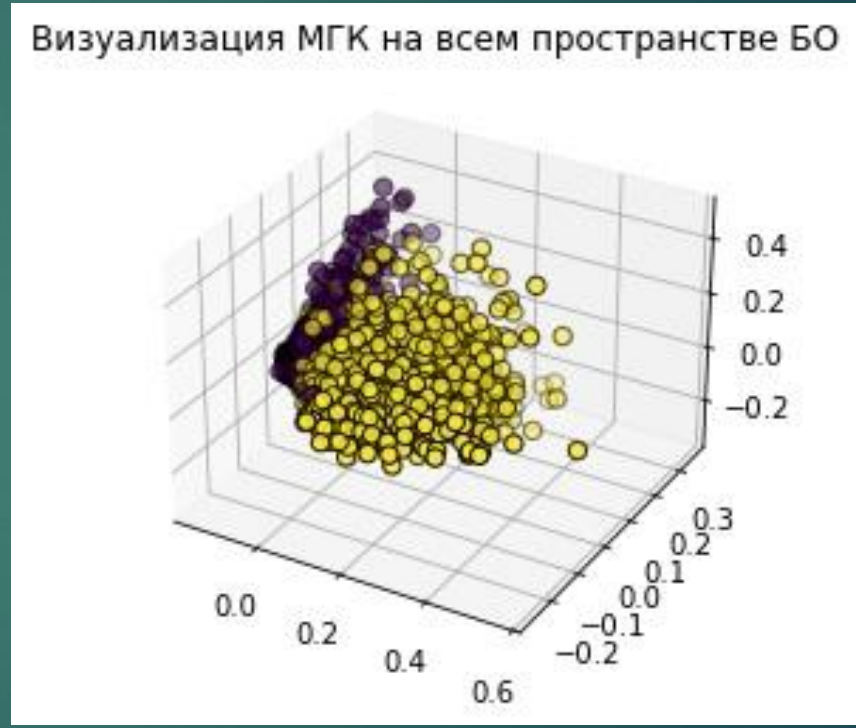
- **Ключевые слова:**

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ МАШИННОЕ ОБУЧЕНИЕ ТАМОЖЕННЫЕ РИСКИ ТАМОЖЕННАЯ ДЕЯТЕЛЬНОСТЬ DATA MINING MACHINE LEARNING CUSTOMS RISK CUSTOMS ACTIVITY

ВИЗУАЛИЗАЦИЯ

Здесь желтые точки – ИАД, фиолетовые точки – не ИАД

Для визуализации использовался ***метод главных компонент***

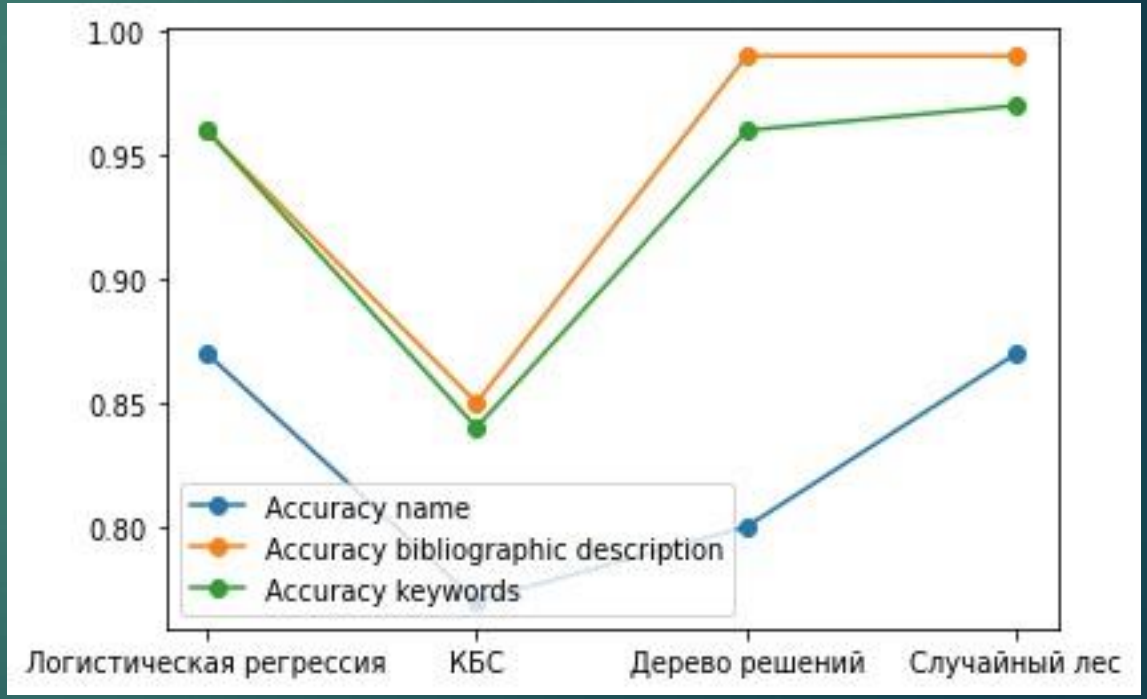


СРАВНЕНИЕ РЕЗУЛЬТАТОВ КЛАССИФИКАЦИИ

Таблица сравнения методов и признаковых пространств

	Названия	Библиографическое описание	Ключевые слова
Метрика / Метод	Accuracy	Accuracy	Accuracy
Логистическая регрессия	0.87	0.96	0.96
КБС	0.77	0.85	0.84
Дерево решений	0.80	0.99	0.96
Случайный лес	0.87	0.99	0.97

Визуальное отображение



МНОГОКЛАССОВАЯ КЛАССИФИКАЦИЯ

В этой главе мы рассмотрим выборку из 10 классов, содержащую 19027 объектов, сформированную А.В. Кононенко и состоящую из документов научно-технической литературы.

Классы можно считать сбалансированными

- computer vision (компьютерное зрение) - 1797
- text mining (интеллектуальный анализ текста) - 1933
- control systems (системы управления) - 1452
- cyber security (кибербезопасность) - 2965
- information retrieval (информационный поиск) - 1944
- fuzzy (нечеткие системы) - 1776
- neural nets (нейросети) - 1840
- database (базы данных) - 1936
- robotic (роботизированные системы) - 1550
- expert (экспертные системы) – 1834

Размерность признаковых пространств

- **По названиям 13330 терминов**
- **По БО 42962 термина**

ПРИМЕР ОБЪЕКТА ВЫБОРКИ

Пример БО:

- **Название:**

Combining ensemble methods and social network metrics for improving accuracy of OCSVM on intrusion detection in SCADA systems

- **Аннотация:**

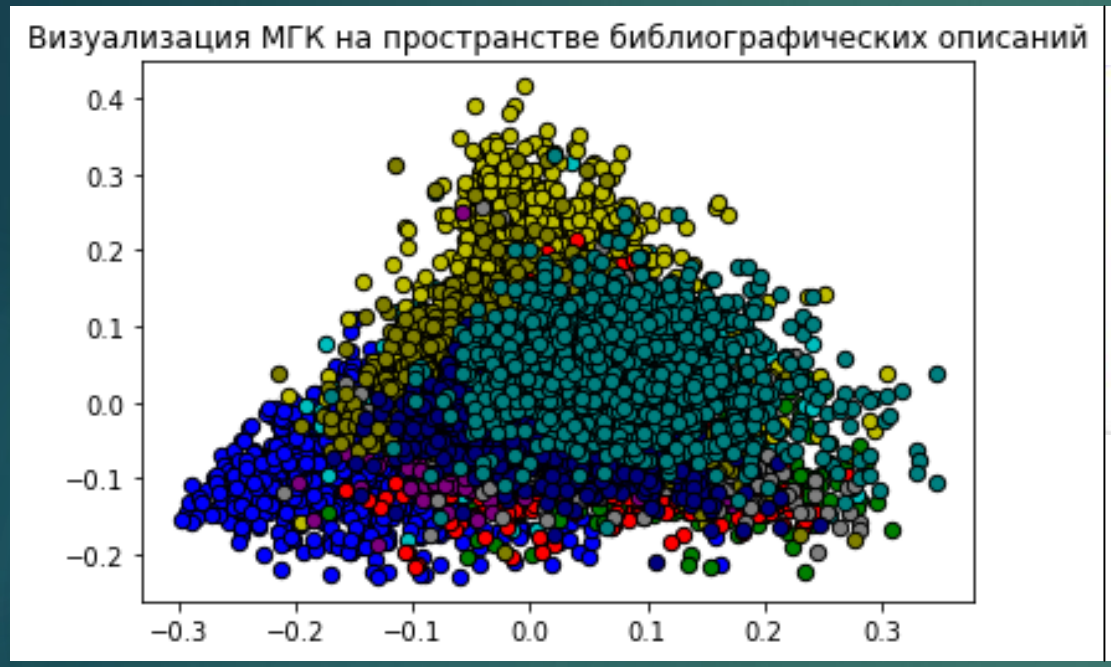
Modern Supervisory Control and Data Acquisition SCADA systems used by the electric utility industry to monitor and control electric power generation, transmission and distribution are recognized today as critical components of the electric power delivery infrastructure. SCADA systems are large, complex and incorporate increasing numbers of widely distributed components. The presence of a real time intrusion detection mechanism, which can cope with different types of attacks, is of great importance, in order to defend a system against cyber attacks This defense mechanism must be distributed, cheap and above all accurate, since false positive alarms, or mistakes regarding the origin of the intrusion mean severe costs for the system. Recently an integrated detection mechanism, namely IT-OCSVM was proposed, which is distributed in a SCADA network as a part of a distributed intrusion detection system (IDS), providing accurate data about the origin and the time of an intrusion. In this paper we also analyze the architecture of the integrated detection mechanism and we perform extensive simulations based on real cyber attacks in a small SCADA testbed in order to evaluate the performance of the proposed mechanism.

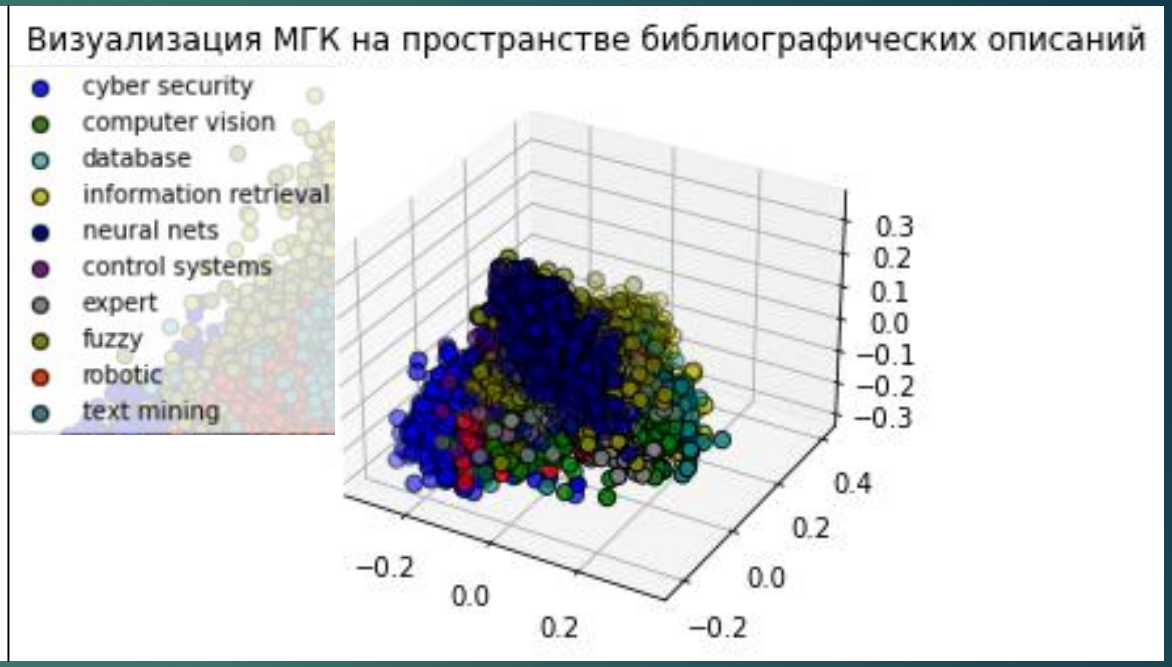
- **Ключевые слова:**

Cryptography and Security

ВИЗУАЛИЗАЦИЯ

Для визуализации использовался ***метод главных компонент***



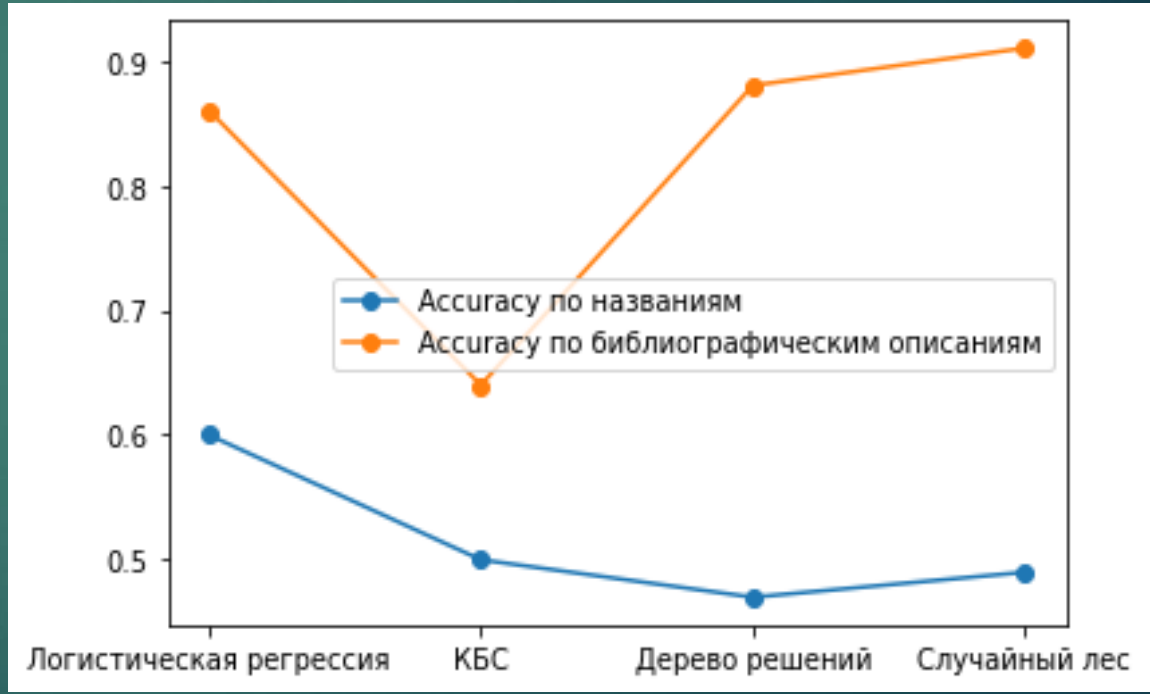


СРАВНЕНИЕ РЕЗУЛЬТАТОВ КЛАССИФИКАЦИИ

Таблица сравнения методов и признаковых пространств

	Названия	Библиографическое описание
Метрика / Метод	Accuracy	Accuracy
Логистическая регрессия	0.60	0.86
КБС	0.50	0.64
Дерево решений	0.47	0.88
Случайный лес	0.49	0.91

Визуальное отображение



ВЫВОД ПО РАБОТЕ

В ходе работы проведено сравнение качества классификации на библиографических описаниях и названиях. Были рассмотрены способы обработки текстовых данных, построены классификаторы и настроены их гиперпараметры.

Лучшие результаты на обеих выборках показал метод случайного леса, однако в рамках задачи лучшие результаты имеет логистическая регрессия, так как на меньшем признаковом пространстве она меньше всех теряет в качестве.

Можно сделать вывод о целесообразности использования короткого признакового пространства в случае бинарной классификации и нецелесообразности в случае многоклассовой классификации.