БЕЛОРУССКИЙ ГОСУДАРСАТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра математического моделирования и анализа данных

Аннотация к дипломной работе

«АЛГОРИТМЫ ОБНАРУЖЕНИЯ АНОМАЛИЙ В ТЕКСТВОЫХ ДАННЫХ»

Карабанович Дмитрий Павлович

Научный руководитель – кандидат физ.-мат. наук, доцент В.И. Малюгин

Минск, 2020

РЕФЕРАТ

Дипломная работа, 53 с., 26 рис., 3 табл., 2 приложения, 18 источников.

Ключевые слова: АНОМАЛИИ В ТЕКСТОВЫХ ДАННЫХ, МЕТОДЫ ОБНАРУЖЕНИЯ АНОМАЛИЙ, СПАМ, СПАМ-ФИЛЬТР, НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР, СЛУЧАЙНЫЙ ЛЕС

Объект исследования – задача обнаружения аномалий типа «спам» в текстовых данных.

Цель работы: исследование алгоритмов для решения задачи обнаружения аномалий типа «спам» в текстовых данных и написание программного обеспечения для решения поставленной задачи.

Основные методы исследования: методы теории вероятности, математической статистики и машинного обучения.

За время работы были решены следующие задачи:

– Подготовлен обзор основных методов выявления аномалий в многомерных данных, допускающих кластерную неоднородность структуры;
– Приведено описание решаемой проблемы (задача обнаружения аномалий типа «спам» в текстовых данных) и показана её актуальность на сегодняшний день;
– Для решения задачи обнаружения аномалий типа «спам» в текстовых данных описаны алгоритмы: наивный байесовский классификатор, случайный лес, изолирующий лес;
– Написана программа для решения поставленной задачи с помощью алгоритма наивный байесовский классификатор на языке программирования C#, используя набор данных содержащий текст электронных сообщений;
– Написана программа для решения поставленной задачи с помощью алгоритмов случайный лес и изолирующий лес на языке программирования Python, используя набор данных построенный относительно результатов работы наивного байесовского классификатора;
– Проведен сравнительный анализ работы всех вышеизложенных алгоритмов.

Полученные результаты и код написанных программ могут использоваться для создания новых спам-фильтров, основанных на полученных результатах.

ABSTRACT

Graduate work, 53 p., 26 pic., 3 tables, 2 attachments, 18 sources.

Keywords: ANOMALIES IN TEXT DATA, METHODS FOR DETECTING ANOMALIES, SPAM, SPAM FILTER, NAIVE BAYESIAN CLASSIFIER, RANDOM FOREST

The object of research is the task of detecting "spam" type anomalies in text data.

The purpose of the work: to analyze algorithms for solving the problem of detecting anomalies of the "spam" type in text data and to write software for solving this problem.

Key research methods: methods of probability theory, mathematical statistics, and machine learning.

During the work the following tasks were solved:

- An overview of the main methods for detecting anomalies in multidimensional data that allow for cluster heterogeneity of the structure has been prepared;
- The description of the problem being solved (the problem of detecting anomalies of the "spam" type in text data) is given and its relevance to date is shown;
- For solving the problem of detecting "spam" type anomalies in text data, algorithms are described: naive Bayesian classifier, random forest, isolating forest;
- A program is written to solve the problem using the naive Bayesian classifier algorithm in the C# programming language, using a data set containing the text of electronic messages;
- A program is written to solve the problem using random forest and isolating forest algorithms in the Python programming language, using a data set constructed relative to the results of a naive Bayesian classifier;
- A comparative analysis of the work of all the above algorithms is carried out.

The results obtained and the code of the programs written can be used to create new spam filters based on the results obtained.