Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Южно-Уральский государственный университет

(национальный исследовательский университет)»

Высшая школа электроники и компьютерных наук

Кафедра «Системное программирование»

**ОТЧЁТ**

по лабораторной работе №12

на тему «Поиск точечных аномалий»

|  |  |
| --- | --- |
|  | Выполнил  Студент группы КЭ-120  \_\_\_\_\_\_\_\_\_\_\_\_\_ Д.А. Снегирева  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2021 г. |
|  | Email: dashasneg@mail.ru |

Челябинск 2021

**ЗАДАНИЕ**

Выполните поиск точечных аномалий (выбросов) в двух различных наборах одномерных данных с помощью двух любых приемов из следующего множества: метод максимального правдоподобия, оценка 𝜒2, построение гистограмм. Выполните визуализацию полученных результатов в виде точечных графиков, использующих два цвета для отражения нормальных/аномальных точек.

СОДЕРЖАНИЕ

[ЗАДАНИЕ 2](#_Toc68258457)

[1 КРАТКИЕ СВЕДЕНИЯ О НАБОРАХ ДАННЫХ И СРЕДСТВАХ РЕАЛИЗАЦИИ 4](#_Toc68258459)

[2 ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ 5](#_Toc68258460)

# КРАТКИЕ СВЕДЕНИЯ О НАБОРАХ ДАННЫХ И СРЕДСТВАХ РЕАЛИЗАЦИИ

В данной работе был использован набор данных содержащий транзакции, совершенные по кредитным картам в сентябре 2013 года европейскими держателями карт (https://www.kaggle.com/mlg-ulb/creditcardfraud). В этом наборе данных представлены транзакции, произошедшие за два дня, где 492 мошенничества из 284 807 транзакций. Набор данных сильно не сбалансирован, положительный класс (мошенники) составляет 0,172% от всех транзакций, состоящий из десяти классов.

Также, был использован набор данных сотрудников торгового центра (https://www.kaggle.com/akram24/mall-customers).

В качестве средств реализации были использована библиотека scikit-learn.

Репозиторий задания: https://github.com/DasHaSneg/BigDataMiningCourse

Каталог задания: 12 detect anomalies 1

ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

На рис. 1 приведена визуализация первого набора данных.



Рис. 1. Первый набор данных

Из рисунка видно, что большая часть данных находится в пределах от -400 до 400.

Затем был выполнен поиск аномалий с помощью метода максимального правдоподобия (Max likelihood estimation, MLE) и построения гистограмм (Histogram-based outlier score, HBOS). Результаты приведены на рис. 2.



Рис. 2. Результаты поиска для первого набора данных

Из рисунка видно, что более правдоподобный результат дал первый метод, но он также определил аномальными те данные, которые находятся достаточно близко к области, содержащей большую часть экземпляров.

На рис. 3 приведена визуализация второго набора данных.



Рис. 3. Второй набор данных

Из рисунка видно, что большая часть данных находится в пределах от -400 до 400.

Затем был выполнен поиск аномалий с помощью метода максимального правдоподобия (Max likelihood estimation, MLE) и построения гистограмм (Histogram-based outlier score, HBOS). Результаты приведены на рис. 4.

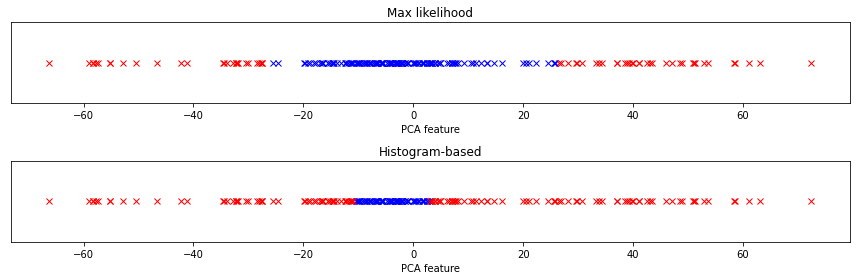


Рис. 4. Результаты поиска для второго набора данных

Из рисунка видно, что особенно узкую область нормальных точек дал алгоритм HBOS.

Исходя из результатов можно сделать вывод, что оба метода можно использовать для поиска аномалий в одномерных наборах данных.