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 №13

на тему «Поиск коллективных аномалий»

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

Челябинск 2021

**ЗАДАНИЕ**

1. Выполните поиск коллективных аномалий (выбросов) в двух различных наборах 2-х или 3-мерных данных с помощью двух любых приемов из следующего множества: метод вложенных циклов, метод решеток, кластеризация.

Выполните визуализацию полученных результатов в виде точечных графиков, использующих два цвета для отражения нормальных/аномальных точек.

2. Выполните поиск коллективных аномалий (выбросов) во временном ряде на основе понятия диссонанса.

Выполните визуализацию полученных результатов в виде точечных графиков, использующих два цвета для отражения нормальных/аномальных подпоследовательностей.

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 набора данных с временными рядами был взять набор данных «Дневные минимальные температуры в Мельбурне, Австралия, 1981-1990 гг» (https://www.kaggle.com/paulbrabban/daily-minimum-temperatures-in-melbourne).

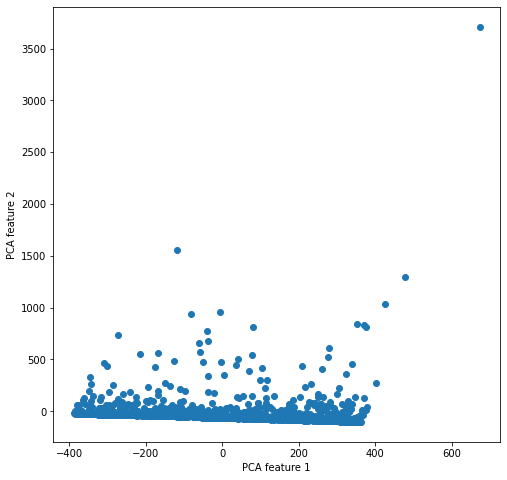
В качестве средств реализации были использована библиотека scikit-learn.

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

Каталог задания: 13 detect anomalies 2

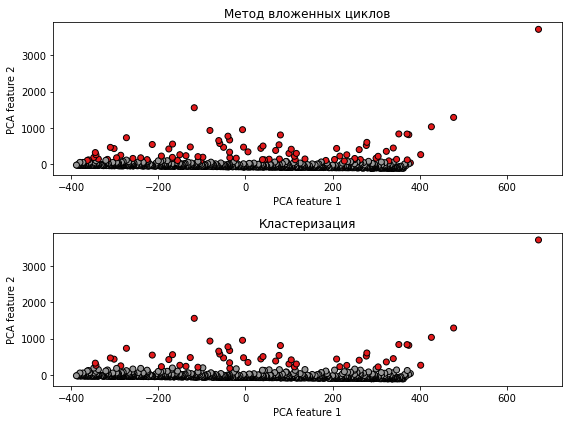
ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

На рисунке 1 приведен первый набор данных.



**Рис. 1.** Визуализация первого набора данных

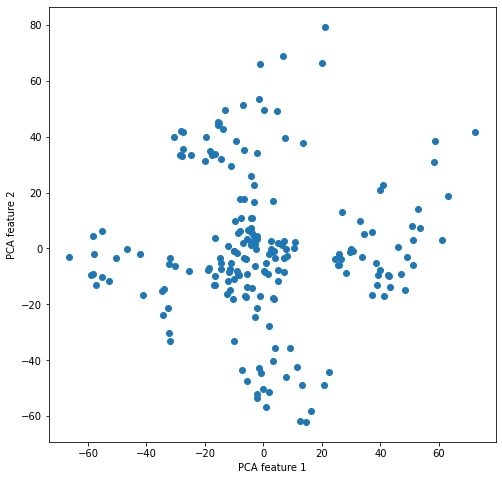
Был выполнен поиск аномалий в данном наборе данных с помощью метода вложенных циклов и плотностной кластеризации. Результаты на рисунке 2.



**Рис. 2.** Результаты поиска аномалий в первом наборе данных

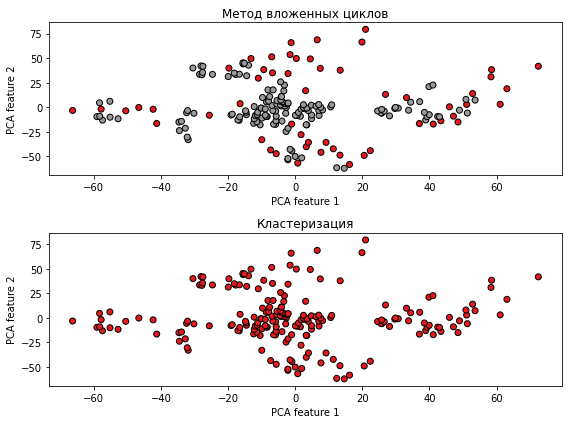
На рисунках видно, что методы обозначили аномалиями как далеко расположенных от основных данных значения, так и находящиеся относительно близко.

Далее был загружен и визуализирован второй набор данных, изображенный на 3 рисунке.



**Рис. 3.** Визуализация второго набора данных

Затем был выполнен поиск аномалий в данном наборе данных с помощью метода вложенных циклов и плотностной кластеризации. Результаты на рисунке 4.

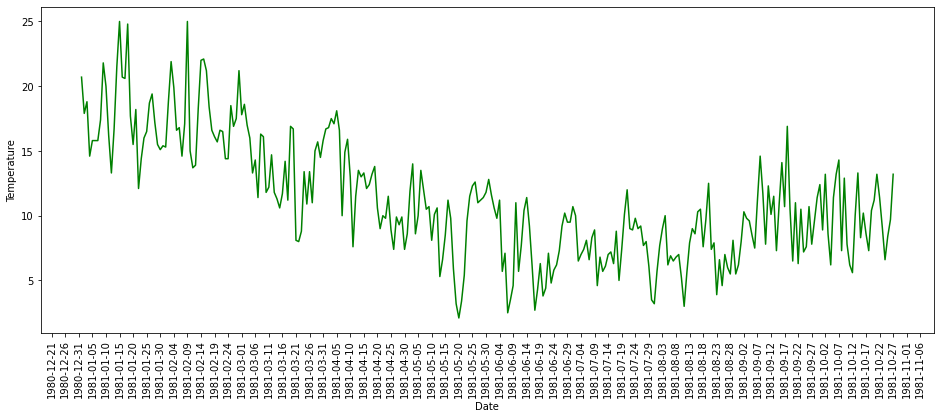


**Рис. 4.** Результаты поиска аномалий во втором наборе данных

Из рисунка видно, что первый способ смог выделить аномалии корректно, тогда как второй выделил в аномалии весь набор данных.

Можно сделать вывод, что для данного набора данных подходит метод вложенных циклов.

Далее был загружен третий набор данных. Результат визуализации первых 100 строк данных на рисунке 5.



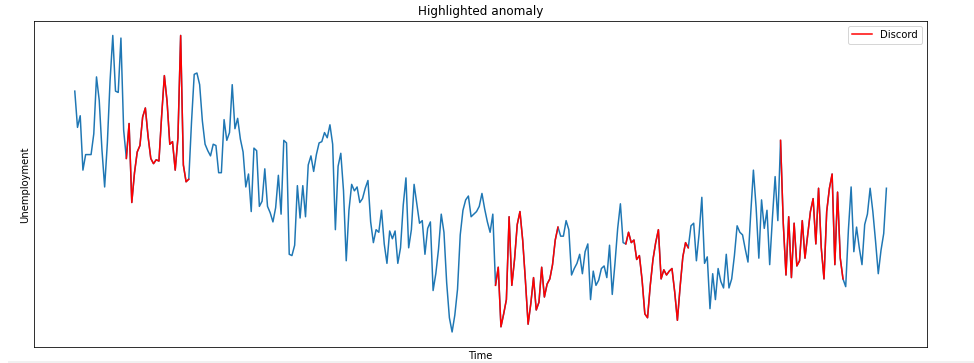
**Рис. 5.** Визуализация третьего набора данных

Затем был выполнен поиск аномалий на основе диссонанса. Количество аномалий, которое нужно найти было установлено на 4. Результаты приведены на рисунке 6.



**Рис. 6.** Результат поиска аномалий временного ряда.

На рисунке видно, что диссонансы были отмечены приблизительно в момент резкого изменения температуры.



**Рис. 7.** Временные рамки аномалий

На рисунке видно, что были отмечены аномалии на момент резкого увеличение и понижения температуры.

Исходя из полученных результатов, можно сделать вывод, что для поиска аномалий во временных рядах можно использовать графики Matrix Profile.