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 системного программирования

ОТЧЁТ ПО ЗАДАНИЮ №12

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Точечные аномалии

Выполнил
студент группы КЭ-120
Глизница Максим Николаевич
E-mail: letadllo@mail.ru

Челябинск 2021

1. Задание

Выполните поиск точечных аномалий (выбросов) в двух различных наборах одномерных данных с помощью двух любых приемов из следующего множества: метод максимального правдоподобия, оценка χ^2, построение гистограмм.

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

2. Краткие сведения о наборах данных

Использованные наборы данных:

Swiss banknote counterfeit detection (https://www.kaggle.com/chrizzles/swiss-banknote-conterfeit-detection). Содержит различные данные о банкнотах, такие как длина, ширина, отступы и т.д. Оригинальный набор содержит данные о 200 банкнотах, среди которых 100 – фальшивые, но для выявления аномалий было взято подмножество, содержащее 100 подлинных и 10 фальшивых банкнот.

Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Содержит информацию о произведённых партиях красного вина "Vinho Verde" (такую как pH и содержание хлоридов), а также оценку качества от 1 до 10 (реально в наборе данных присутствуют значения от 3 до 8). Как предлагается в пояснении к набору данных, высококачественным можно считать вино с оценкой 7 и выше. Таким образом, в наборе данных присутствуют 217 партий высокого качества и 1382 партий низкого качества.

Оба набора данных были приведены к одному измерению с помощью PCA.

3. Краткие сведения о средствах реализации

Для реализации методов была использована библиотека scikit-learn, включающая в себя множество алгоритмов для анализа данных, и библиотека PyOD, содержащая различные алгоритмы для выявления аномалий.

Репозиторий по дисциплине: https://github.com/Airpllane/DAAlgorithms.

Каталог для задания: 12. Outliers

4. Визуализация

На рис. 1 приведена визуализация первого набора данных.

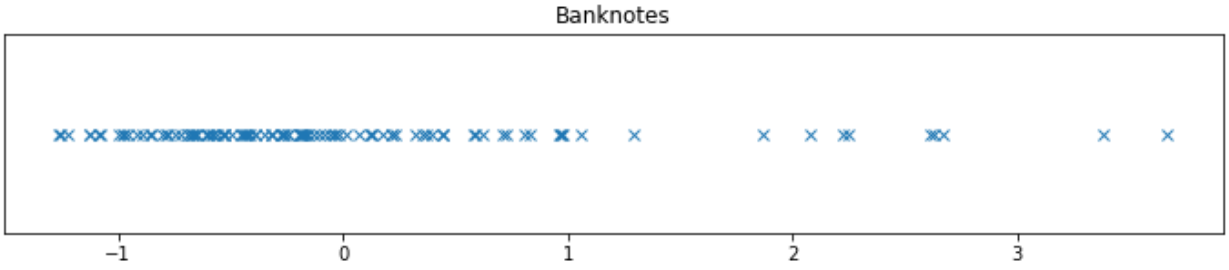

Рис. 1. Визуализация набора данных

Из рисунка можно увидеть, что большая часть набора находится в пределах около [-1;1], в то время как некоторые экземпляры находятся далеко от этого промежутка.

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

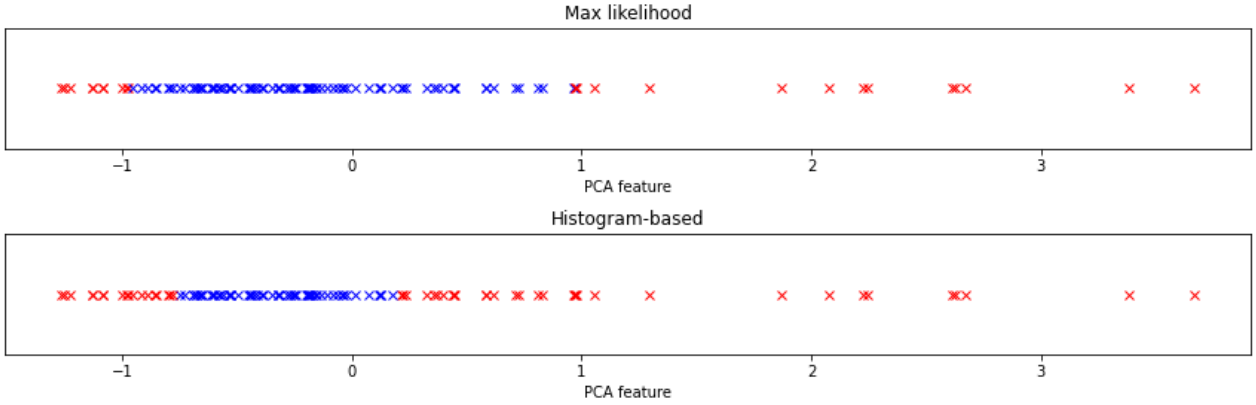

Рис. 2. Результаты поиска аномалий с помощью MLE и HBOS

Из рисунка можно увидеть, что методы поиска смогли корректно определить область, содержащую большую часть экземпляров, но выделили множество достаточно близких к этой области экземпляров как аномалии. Особенно узкую область нормальных точек дал алгоритм HBOS.

Далее был загружен второй набор данных, визуализация которого приведена на рис. 3.

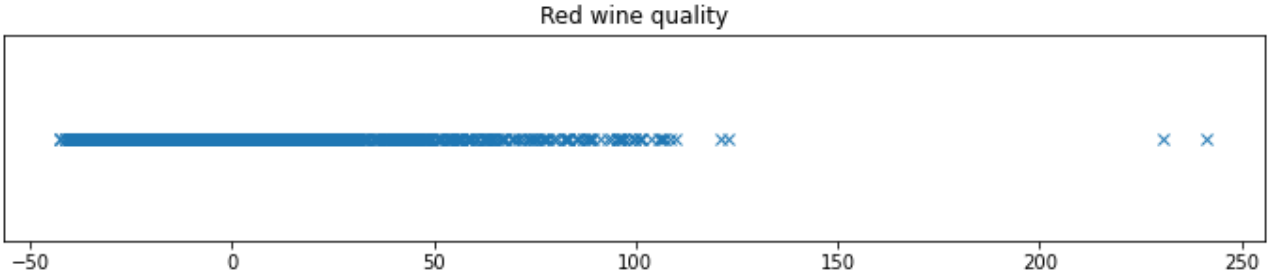

Рис. 3. Визуализация набора данных

Видно, что набор содержит очень много плотно расположенных экземпляров в пределах около [-50;100], при этом есть небольшое количество экземпляров, находящихся очень далеко от этого промежутка.

Далее был выполнен поиск аномалий в этом наборе данных с помощью тех же алгоритмов. Результаты поиска приведены на рис. 4.

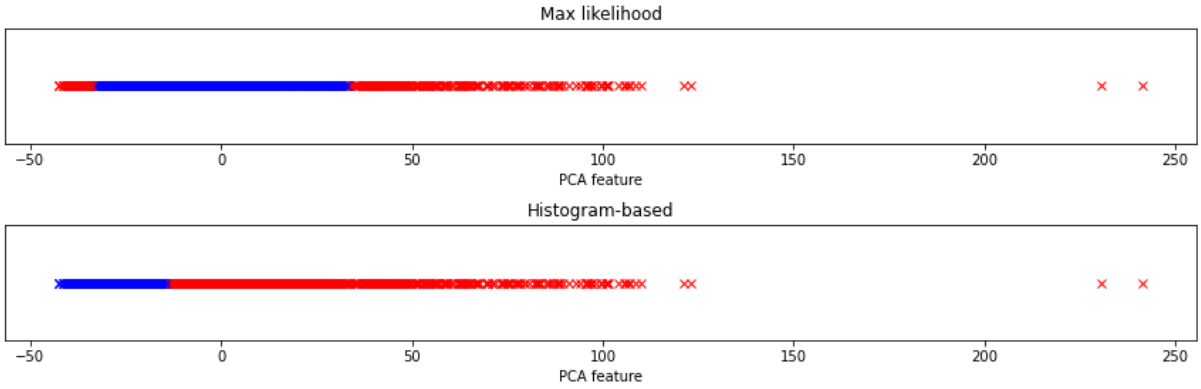

Рис. 4. Результаты поиска аномалий с помощью MLE и HBOS

Из рисунка видно, что метод MLE показал результаты, похожие на предыдущий пример – множество нормальных экземпляров были отмечены как аномалии. Алгоритм HBOS показал несколько другой результат – в нём область, содержащая наибольшее количество нормальных экземпляров, оказалась смещена вправо.

Таким образом, можно сделать вывод, что оба рассмотренных алгоритма можно использовать для поиска аномалий в одномерных наборах данных, но при этом существует риск отметить множество нормальных экземпляров как аномалии.