Semester report

László Kiss

June 2019

This document is created for summary of second semester of 2018/19 school year, which is my fourth semester. I was examining about the outlier detection in high dimensions datasets. I picked the Angle-Based outlier detection (ABOD) method. This paper is a short summary about that.

1 Abstract

Outlier detection is very useful in many applications, such as fraud detection and network intrusion. The angle-based outlier detection (ABOD) method, proposed by Kriegel, plays an important role in identifying outliers in high-dimensional spaces. However, ABOD only considers the relationships between each point and its neighbors and does not consider the relationships among these neighbors, causing the method to identify incorrect outliers.

2 Introduction

The general idea of outlier detection is to identify data objects that do not fit well in the general data distributions. This is a major data mining task and an important application in many fields such as detection of errors in data sets in telecommunications or the identification of measurement errors in scientific data.

3 The angle-based outlier detection (ABOD)

3.1 Definition of outlier

- An outlier is "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (by Hawkins, 1980)
- Outliers appear everywhere Intrusions in network traffic, credit card fraud, defective products in industry, medical diagnosis from X-ray images
- Outliers should be detected and removed

- Outliers can cause fake results in subsequent analysis

3.2 The basic idea of ABOD

In detection process of outliers comparing distances becomes more and more meaningless with increasing data dimensionality. Mining high dimensional dataset requires a new approaches to the discover patterns. This method not use the distance between points in the vector space but primarily the directions of distance vectors. Comparing the angles between pairs of distance vectors to other points helps to discern between points similar to other points and outliers. Consider the dataset which is plotted in Figure 1. For a point within a cluster, the angles between difference vectors to pairs of other points differ widely. The variance of the angles will become smaller for points at the border of a cluster. However, even here the variance is still relatively high compared to the variance of angles for real outliers. Here, the angles to most pairs of points will be small since most points are clustered in some directions.

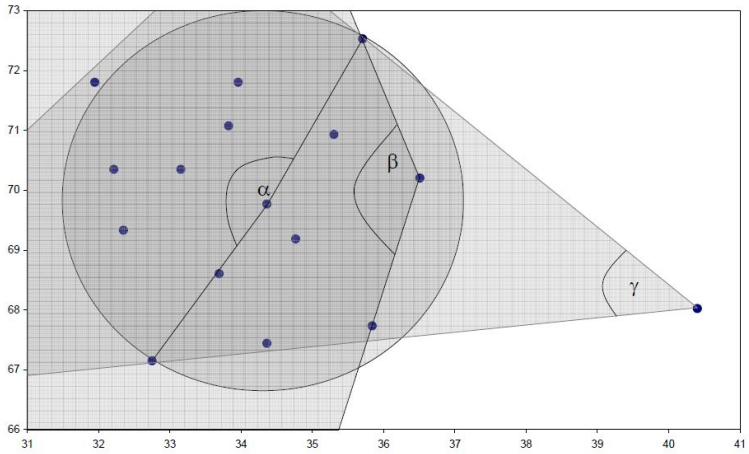

Figure 1: : Intuition of angle-based outlier detection.

The corresponding spectra for these three types of points are illustrated for a sample data set in Figure 2.

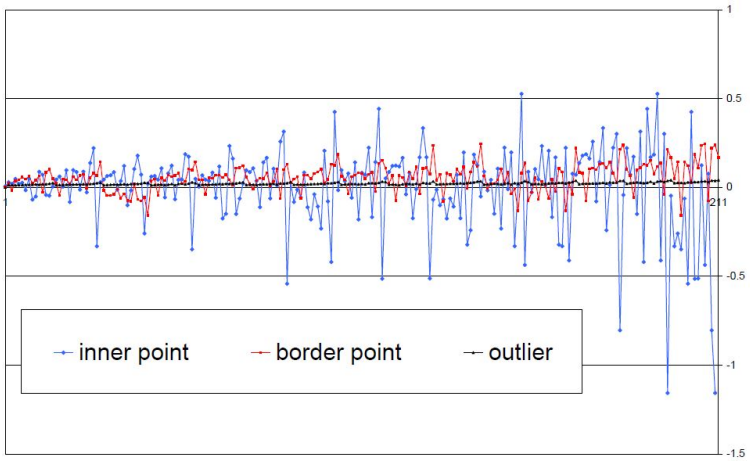

Figure 2: : Spectra of angles for different types of points.

As the graph shows, the spectrum of angles to pairs of points remains rather small for an outlier whereas the variance of angles is higher for border points of a cluster and very high for inner points of a cluster. As a result of these considerations, an angle-based outlier factor (ABOF) can describe the divergence in directions of objects relatively to one another.

3.3 Fast ABOD

There is a basic issue about the original approach and it is obvious: the time complexity is in $O(n^n)$ which is not attractive. There is an approximation algorithm called Fast ABOD. Their approximation based on a sample from database. It uses the pairs of points with the strongest weight in variance, e.g. pairs between the k nearest neighbors. It can be a very good method because the nearest neighbors have a largest weight in the ABOD. Employing the nearest neighbors might result in a better approximation, especially in data sets of low dimensionality where the distance is more meaningful. We can see test result with this two methods in Figure 3.

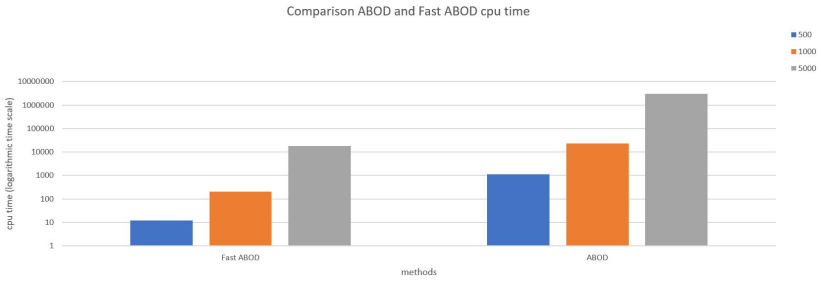

Figure 3: : Comparison between ABOD and Fast ABOD.

3.4 Short summary about ABOD

- If x is an outlier, the variance of angles between pairs of the remaining objects becomes small.

  The score

  $$ABOF(x) := Var_{y,z \epsilon X} s(y - x, z - x)$$

  where

  - s(x, y) is the similarity between vectors x and y, e.g. the cosine similarity
  - s(z - x,y - x) correlates with angle of y and z with the reference to coordinate origin x

- Rational

  - Angles are more stable than distances in high dimensional spaces
  - Object o is an outlier if most other objects are located in similar directions
  - Object o is no outlier if many other objects are located in varying directions

- Basic assumption
  - Outliers are at the border of the data distribution
  - Normal points are in the center of the data distribution
- Model
  - Consider for a given point p the angle between $\overrightarrow{px}$ and $\overrightarrow{py}$ for any two x,y from the database (Fig 3.)
  - Consider the spectrum of all these angles
  - The broadness of this spectrum is a score for the outlierness of a point

3.5 Challenges

The dataset is high dimensional - more than 600 attributes collected in every second - so we have to face some issues regarding that.

1. Challenges
   - Curse of dimensionality
     - Relative contrast between distances decreases with increasing
     - Data is very sparse, almost all points are outliers
     - Concept of neighborhood becomes meaningless
2. Solutions
   - Use more robust distance functions and find full-dimensional outliers
   - Find outliers in projections (subspaces) of the original feature space

With this method we would like to analyze the data set about the incidents in R language based environment. This PoC could be a very fundamental base for the final analysis which can be the part of the production system.

4 The dataset

This project was created as a PoC (Proof of Concept) to make sure this algorithm can handle this very large data set which describes the incidents related to eNodeB. The eNodeB-s physically are the "mobile-towers". It is the hardware that is connected to the mobile phone network that communicates directly wirelessly with mobile handsets. The actual dataset is a matrix their rows represent the type of mobile devices and their columns is the typ of eNodeB-s. During the data preprocess we aggregated the number of calls and incidents by this groups and after we can determine the incident ratio with this formula:

$$inc\_ratio = \frac{num\_of\_inc}{num\_of\_calls}$$

where num_of_inc stands for number of telecommunication incidents and num_of_calls stands for number of connection - not just for the voice calls - between the eNodeB-s and mobile devices. Regarding the actual regulation of Ericsson I changed the metadata and data about this data set but after the changing it remains suitable for this.

We can see the part of this data set by vizualied in the Figure 4.

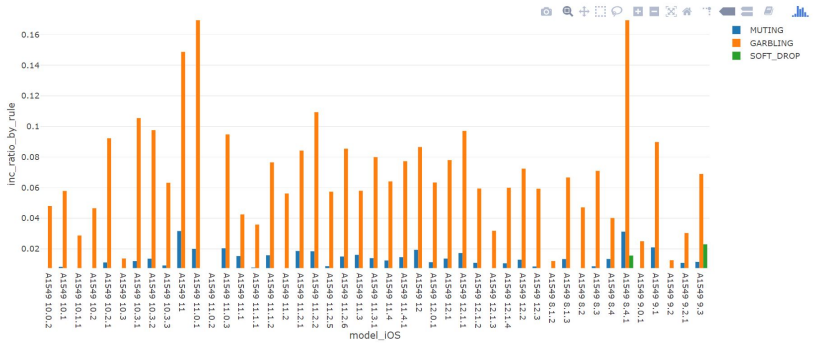

Figure 4: : The part of the examined data set.

5 The R project

Based on previous data set I made an analysis as a PoC (Proof of Concept), to support a further examination in a product environment. In this R project I used the Fast ABOD method with the Func.FBOD function from the HighDimOut r package. This package contains a few algorithm to support the outlier detection in high dimensional data set. This function requires 3 parameters as follows: Func.FBOD(data, iter, k.nn)

The "data" parameter stands for the name of the examined data set, the "iter" is the number of the iteration and the k.nn is the value used for calculating the LOF score.

The function returns a vector containing the FBOD outlier scores for each observation. Regarding this score values we can filter for the outliers.

We can see the results - outliers - a three dimensions vector space in Figure 5.

6 Summary

7 References

Angle-Based Outlier Detection in High-dimensional Data
Hans-Peter Kriegel, Matthias Schubert, Arthur Zimek, 2008

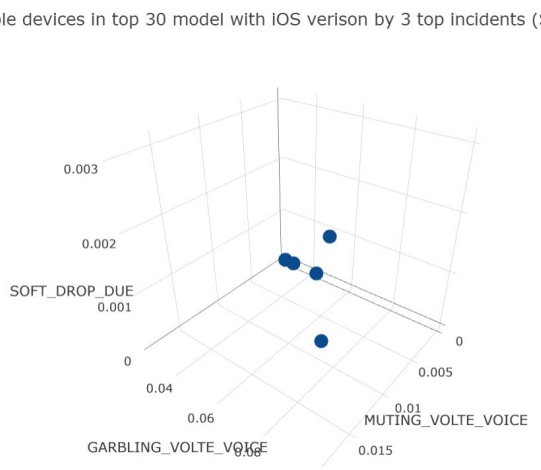

Figure 5: : The outliers in three dimensions vector space.

A Near-linear Time Approximation Algorithm for Angle-based Outlier Detection in High-dimensional Data
Ninh Pham, Rasmus Pagh