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 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.

4 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

5 ABOD algorithm

ABOD is stands for angle-based outlier detection. The basic idea about that is very clear in Fig 1 and Fig 2.

6 The basic idea of ABOD

- 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

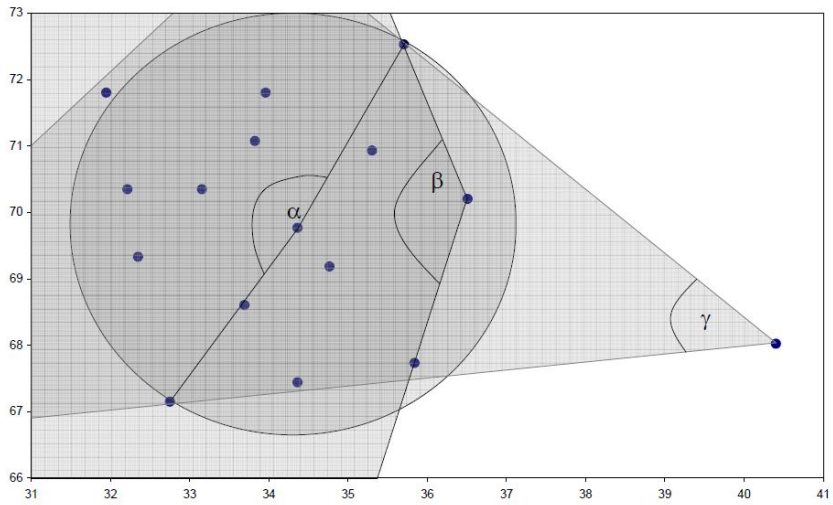

Figure 1: Intuition of angle-based outlier detection.

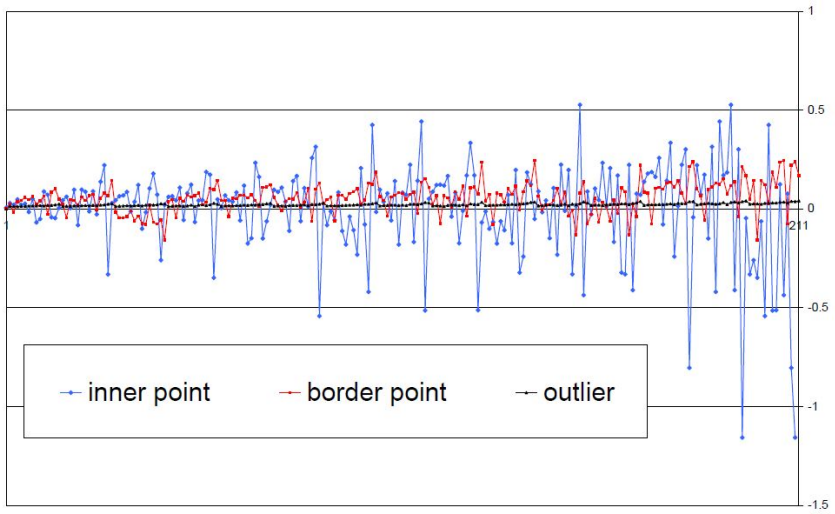

Figure 2: Spectra of angles for different types of points.

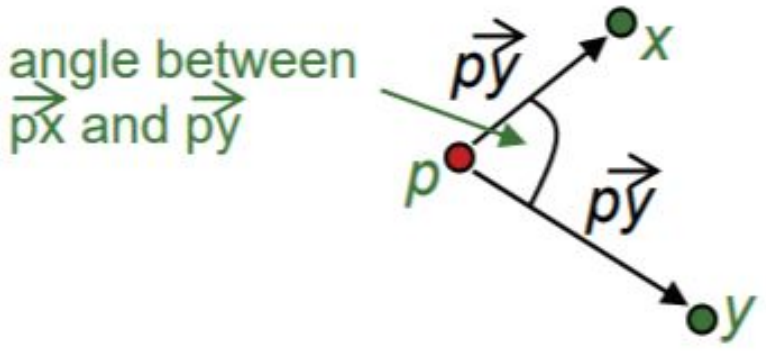

Figure 3: Angle between two vectors

  - 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

7 Summary

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.