5J-1

A Proposal for Outlier Detection in High Dimensional Space

Zhana† Wataru Kameyama‡

GITS, Waseda University †‡

1. Introduction

The outlier detection is an important part in data mining field. It aims to find observations that they are very different from other observations in a dataset. The different concept of outliers is used to identify different observations that deserve special treatment. The well known idea is Hawkins definition of outliers: an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism[1]. The outlier detection is applied in many fields, such as intrusion detection, fraud detection, medical and public health anomaly detection, industrial damage detection, image processing, and text processing, and etc. The outliers can be viewed not only as points that are different from other points, but as points that they are distant from them. When the distance is a poor measure of difference, the space or the distance metric can be changed to make the distance more appropriate. The default distance metric is Euclidean distance. However, a problem arises in high dimensionality, which is in the most common distributions of data, i.e. the contrast of the distances between any pair of points in the dataset approaches zero as the dimensionality increases. This is known as "the curse of dimension", i.e. no points are very distant from the rest of the dataset in high dimensional data. This means that there are no outliers by distance. In the previous researches, two methods are proposed to challenge the curse of dimensionality: one is to use a more robust distance function to find full-dimensional outliers such as LOF[2], ABOD[3]; the other is to find outliers in projections (subspace) from the original feature space such as grid-based and SOD[4].

2. Concepts and Proposed Method

After analyzing the characteristic of high dimensional space and the observation of outlier feature, it is found that normal points are always clustered together in all dimensional spaces, while outlier points seem deviated from normal points in some dimensions or clustered with different points in different dimensions. Based on this assumption, RPGS (Rim Projected Grid Statistic) algorithm is proposed where projected data point to each dimension, then statistics of each dimension and difference between each two different dimensions are used.

All ever-proposed algorithms try to solve this problem in one step. However, the complexity of outlier detection in a high dimensional space is extremely difficulty to overcome.

In this proposal, we take three steps to detect outliers: finding outliers in each dimension in 1st step, finding outliers between two different dimensions in 2nd step, then summarize the result for each point. The outliers are the points whose values are obviously lower than most of values. To realize this method, the original data are decomposed into point information and section information. The new data structure is like a grid, including the same section number and dimension number, so we can calculate the value in dimension or between different dimensions.

Some important concepts are introduced here:

Section: the range of data is divided into same number of equi-width parts in each dimension, which is called a section. The width of section is determined heuristically.

Density: average number of points in one section, e.g. the density of i^{th} dimension is called d_i.

PtVal: the record of the point's value ratio against density.

The main data structure is listed below:

PointInfo[Dimension ID, Point ID]:point in section ID

SectionInfo[Dimension ID, Section ID]: #points in the section

We calculate the PointVal with SectionInfo, and compare SectionInfo change between different dimensions with PointInfo. To explain the measurement of outlier detection, 3 definitions are listed below:

Definition 1 (Ratio of Density Function)

$$PtVal_1(x_{i,j}) = \frac{Density_{x_{i,j}}}{\overline{Density_j}} \quad (1)$$

The $\overline{Density}_j$ refers to the average number of points in one section by all points in j^{th} dimension; $Density_{xi,j}$ refers to number of points of the section where the points exist in j dimension.

Definition 2 (Center Section Function)

$$Sec_{center} = \frac{1}{k}\sum_{i=1}^{k} SectionID(x_i) \quad (2)$$

When the k points of same section in one dimension are projected to another dimension, Sec_{center} refer to the average section ID where these k points are.

Definition 3 (Center Section Function)

$$PtVal_2(x_i) = \frac{SecID(x_i) - Sec_{center}}{MaxSec - Sec_{center}} \quad (3)$$

$$PtVal_2(x_i) = \frac{Sec_{center} - SecID(x_i)}{Sec_{center} - MinSec} \quad (4)$$

Among these sections, MinSec is the minimum section ID; MaxSec is the maximum section ID. If SecID > Sec_{center}, Equation (3) is used, or else Equation (4) is used.

Definition 4 (Grid Statistic)

$$SI(x_i) = \sum_{j=1}^{m} \log_2(PtVal_1(x_{i,j})) + \sum_{j=1}^{m} \log_2(PtVal_2(x_{i,j})) \quad (5)$$

Summarize all dimensional result of each point in log function, the points whose SI values are much less than most points are outlier points.
The RPGS algorithm is illustrated in Fig.1 where the data-set has m dimensions and n points:

Begin
Initialize(PointInfo, SectionInfo)
For i=1 to m
  d_i=n/length(SectionInfo[i,]!=0)
  For j=1 to n
    Get $PtVal_1$[i,j] with Equation (1)
  End
End
For i=1 to m
  For j=1 to n
    Ptsecid= PointInfo[i,j]
    SecID = PointInfo[i+1, PointInfo[i,]=Ptsecid]
    Get Sec_{center} with Equation(2)
    Get $PtVal_2$ with Equation(3) or (4)
Get SI value with Equation(5) for each point
Output points the value << normal value

Fig.1 RPGS Algorithm

In the proposed algorithm, the concept of the session density is used instead of calculating Euclidean distances between points. Therefore, the curse of dimension won't occur in the processing of data when dimension is very high.

3. Experiments

To evaluate the proposed algorithm, a simple data set is designed in Fig.2. It is 2 dimension artificial data of total 23 points including 2 clusters and 3 outlier points.

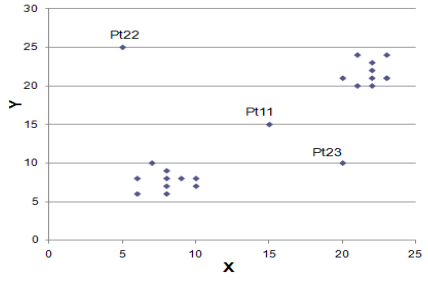

Fig.2 2-D Sample

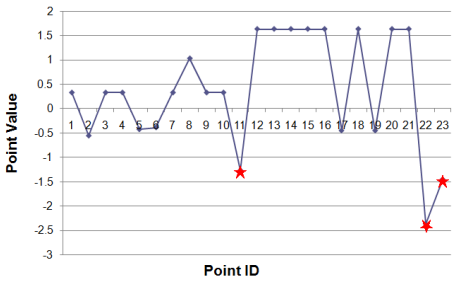

Fig.3 2-D Sample Result

Compared with normal points, outlier points which are labeled with star in Fig.3 have less value.
Another artificial data set with 1000 dimension and 100 data points is evaluated with this algorithm. The data set include 990 normal points of 5 clusters and 10 outlier points. The distribution of point with dimension 1 and dimension 2 are shown in Fig.4.

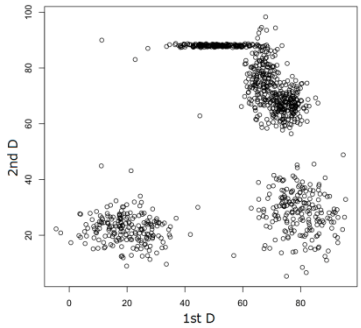

Fig.4 1_{st}-2_{nd}-D Sample

The result is shown in Fig.5, which also demonstrates a good result. Both evaluations show that this proposed algorithm works even in high dimensional space.

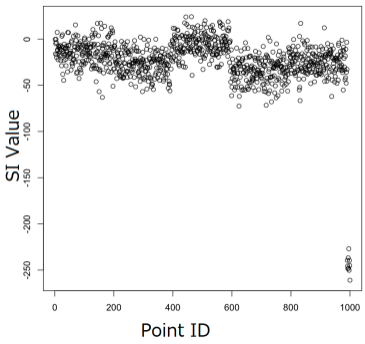

Fig.5 100-D Sample Result

4. Conclusions

The proposed RPGS algorithm is based on a new dimension projection and statistical analysis. It works well in low dimensional and high dimensional space with simple data-set tests. However, more works needs to evaluate effectiveness and efficiency compared with ever-proposed methods, e.g. how to reduce the noise by marginal points of the normal cluster and clusters-overlap is a topic needs to be considered in future.

References

1. D. Hawkins. Identification of Outliers. Chapman and Hall, London, 1980.

2. Markus M.Breunig, Hans-Peter Kriegel, Raymond T.Ng, Jorg Sander. LOF: Indetify density-based local outliers. Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data.

3. Hans-peter Kriegel, Matthias Schubert, Arthur Zimek. Angle-based outlier detection in high dimensional data. KDD2008.

4. Feng chen, Chang-Tien Lu, Arnold P. Boedihardjo. GLS-SOD: a generalized local statistical approach for spatial outlier detection. KDD2010.

5. Das, K. and Schneider, J. Detecting anomalous records in categorical datasets. KDD 2007.