A Survey on Outlier Detection Techniques Useful for Financial Card Fraud Detection

V.Kathiresan
Assistant Professor, Department of Computer Science and Engineering, Coimbatore Institute of Engineering and Technology, Coimbatore, TamilNadu.

Dr.N.A.Vasanthi
Professor and Head, Department of Information Technology, Dr.N.G.P Institute of Technology , Coimbatore, TamilNadu.

Abstract - Every data set contains its own outlier .some time outliers are detected as a noise and removed from the data to increase the quality of data; outliers are detected and considered in many situations like fraud detection, health care management etc.. Among all financial card transactions 0.1% is resulted as fraud even though it causes huge financial loss, based on 1999 survey one transaction turned out to be fraudulent out of 1200 transactions.Based on the trend in financial card fraud there is a huge need for financial card fraud detection techniques.Statistical based outlier detection techniques place huge role in financial card fraud detection. In this paper we introduces a survey of statistical based outlier detection techniques those are useful for financial card fraud detection. Also we analyze and identify the pros and cons of various financial card fraud detection techniques.

Keywords – Outlier, dataset, fraud detection, financial card, statistical techniques

I. INTRODUCTION

1.1. Defining outlier and financial card fraud

Based on the definition of Hawkins “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”

By considering objects some objects properties are deviated from other objects those different behavior objects are called outliers.

Financial card fraud is fraud exists or occurred by using payment card or financial card, such as a credit card or debit card, fraudulent source of funds plays a role in transaction.other than fraud detection outlier detection attains many more applications like unusual symptoms detection in health care, sports statistics, detecting measurement errors etc..

1.2 Data Used in outlier detection

Based on the properties of data it basically classified into three those are univariate, bivariate and multivariate these are also called respectively one-dimensional, two dimensional and multi-dimensional. Normally univariate data contains only one variable that is it contains one piece of information. Height of a group of students is an example for univariate. Bivariate data contains two variables that is it contains two pieces of information Age and height of a group of person is an example for bivariate. Data contains multiple variables is multivariate.

1.3 Application Scenario in outlier detection

Basically applications are classified into three based on their scenario supervised, semi-supervised and unsupervised. The training data is used both normal and abnormal cases in supervised scenario. The training data is used for only normal or abnormal case in semi supervised scenario. There is no training data is used in unsupervised scenario. Mostly unsupervised scenario is used in most of the applications.

1.4 various approaches for outlier detection

1.4.1. Labeling Approach

Labeling with respect to outlier detection normally contains two labels one label for normal data another for outlier.

1.4.2. Scoring Approach

In scoring approach each data object is assigned by a probability score based on the probability we can detect the object is outlier or normal. Sometimes scoring approach provides binary output also by setting threshold limit.by scoring method user can able to identify top n outlier objects according to the application or need.

1.4.3. Model Based Approach

Data objects are represented by a model, the data objects those are not fit in the model are consider as an outlier. Probabilistic test based on statistical model, Depth Based Approaches, Deviation Based Approaches are example for model based approach.

1.4.4. Proximity Based Approach

In proximity based approach Spatial proximity of each object is identified in the data space, the data objects which deviates more from the proximity of other data objects are considered as outliers. Distance based approaches and density based approaches are the example for proximity based approach.

II. GAUSSIAN DISTRIBUTION BASED STATISTICAL TEST FOR OUTLIER DETECTION

Gaussian distribution helps outlier detection, normally most of the data points lies in the normal distribution curve the data objects those are lies away from the curve are considered as outliers, probability of each and every data object can be calculated by probability density function. [4]

$$N(\mu,\sigma) = \frac{e^{-1/2\left(\frac{x_i-\mu}{\sigma}\right)}}{\sqrt{2\pi}\sigma}$$

$$\mu = \frac{\sum x_i}{n}$$

$$n = number\ of\ data\ objects$$

$$x^i = \ i\ th\ data\ object$$

$$sample\ variance\ s^2 = \sum \frac{(x_i-\mu)^2}{n-1}$$

$$standard\ deviation\ \sigma = \sqrt{variance}$$

2.1 Outlier identified by setting threshold limit

Every data object has its own probability, the data objects which contains less or higher probability than threshold are considered as outliers

2.2 identification of fixed number of outliers

N number of outliers is identified with the help of probability, first n number of lower probability or higher probability data objects are considered as an outlier.

III. DEPTH BASED OUTLIER DETECTION

Depth based outlier detection is come under another statistical model based outlier detection mechanism, FDC algorithm is a good example for depth based outlier detection mechanism. FDC (First K 2D depth contours) detect first K 2D depth contours. In FDC data objects are plotted in covex Hull, convex hull is formed by convex polygon.

Convex polygon is a polygon with all its interior angles less than 180º.This means that all the vertices or data objects of the the polygon should point outwards, away from the interior of the shape.

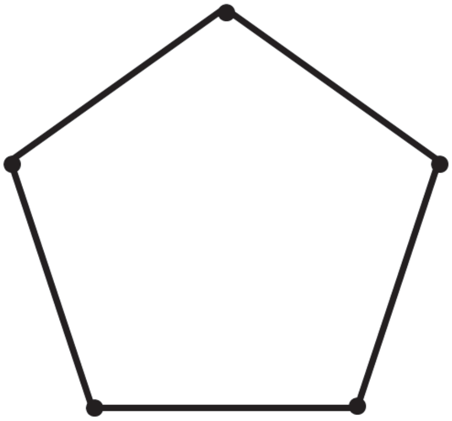

Fig 3.1Convexpolygon

The convex hull of a set of points Q is the smallest convex polygon P for which each point in set Q is either on the boundary of P or in its interior.

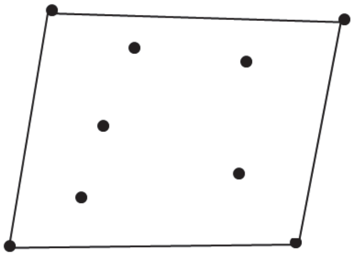

Fig 3.2 Convex hull

We can set depth also in convex hull, Outer most layers is in depth 1, obviously inner most layer attains high depth.

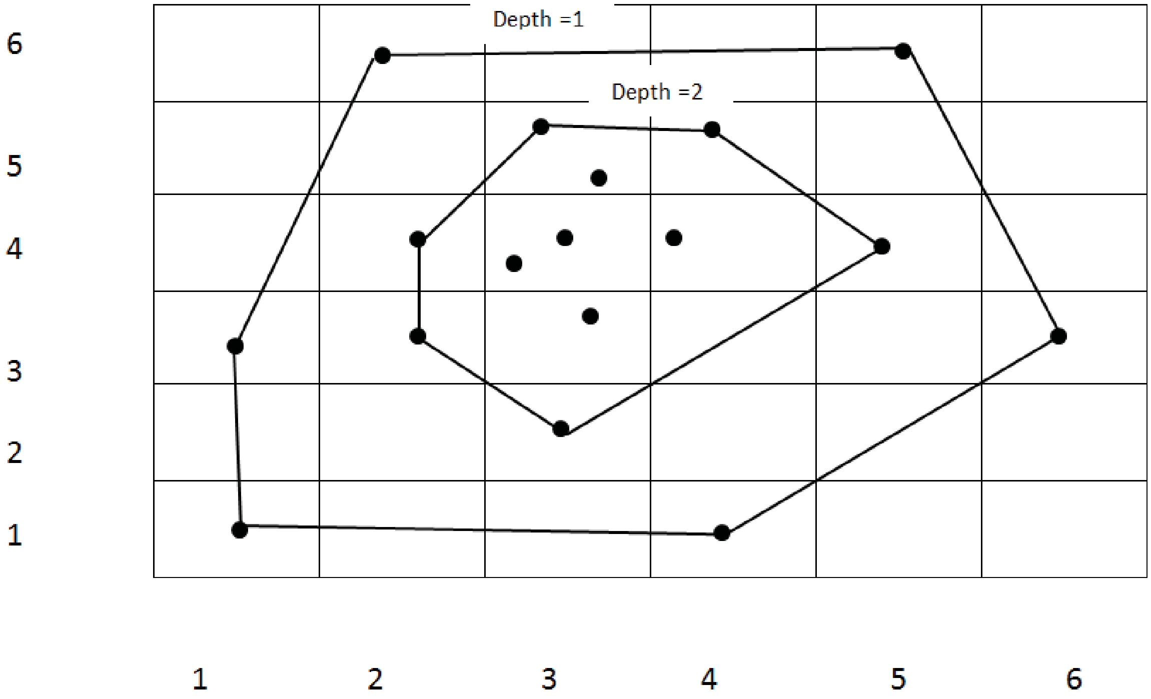

Fig3.3 Depth in convex hull

3.1. Outlier detection with convex hull

Mostly outer most data objects in convex hull are considered as outliers that is the data object or vertices in depth 1. Otherwise according to the users need or application scenario first K –depth data objects are considered as outliers. Interior data objects are considered as normal data objects.[5]

IV. DEVIATION BASED OUTLIER DETECTION

Deviation based outlier detection also come under another kind of statistical based outlier detection model. In deviation based approach smoothing factor of every data object is calculated, outlier is detected based on smoothing factor, outlier is identified based on how much variance is decreased when the particular data object is removed from the data set.

The data object which contains high smoothing factor or high decrease in variance when the particular data object is removed from the data set is considered as an outlier. If more than one data object have equal decrease in variance in the sense, we can make an exception set with those objects, outliers are the elements of the exception set, but the condition is smoothing factor of exception set must be greater than equal to smoothing factor of all other individual data objects.

V. DISTANCE BASED OUTLIER DETECTION

Distance based outlier detection come under the category of model based on spatial proximity. According to the model proposed by knorr and Ng 1997[6].

Radius ε of the data and the percentage π has been given. Here a point P is considered as an outlier if at most π percentage of all other points has a distance to P less than ε

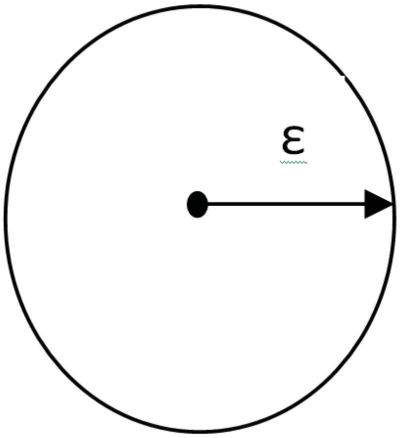

Fig-5.1 Distance based outlier detection

ε- radius

π- percentage (Ex:40%)

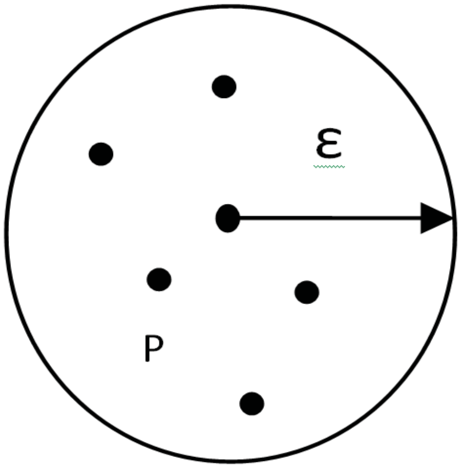

Fig-5.2 Distance based outlier detection

If this P is an outlier if at most 40 percentage of the other objects have distance to P less than ε.

Index based algorithm for distance based approach:

In index based algorithm distance is calculated based on spatial index structure. Normally it gets all the objects whose distance from the current object or current location is less than a specified value. Other objects are consider as an outliers.[6]

5.1. Computation of distance based on spatial index structure

Mostly in spatial indexing data's are organised with the help of R-Tree data structure, Data's are represented by minimum bounding box or minimum bounding rectancle(MBR).

Each node in the index bounds its children. A node can have many objects in it. The leaves point to the actual objects.

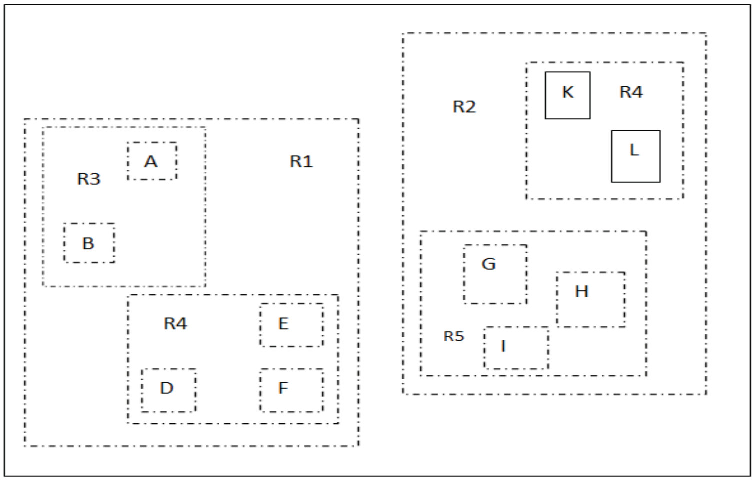

Fig 5.3 spatial data distribution

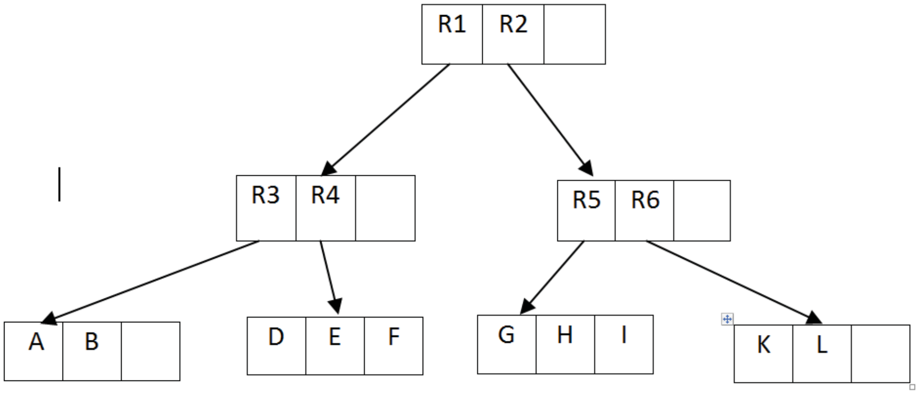

Fig 5.4 R-Tree based spatial data index

Spatial index locate the nearest neighbour to an object. It gets all objects whose distance from the current location is less than a specified value.

VI. NESTED LOOP BASED OUTLIER DETECTION

Nested loop based outlier detection is come under the concept of distance based outlier detection. In nested loop based outlier detection buffer of the data object is divided into two parts ,The data objects in first part is compare with the data objects in the second part with the help of nested loop.[6]

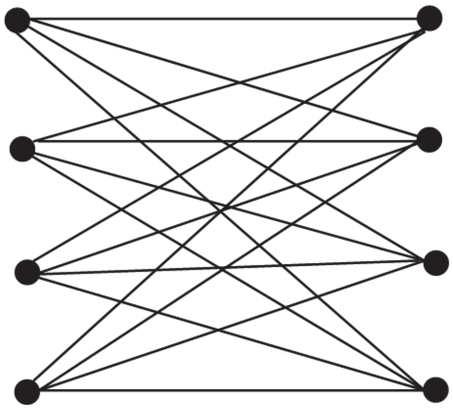

Fig 6.1 Nested Loop Based Outlier Detection

VII.GRID BASED APPROACH

Data objects are grouped into grids where two data objects from same grid cell have a distance of at most ε to each other(ε is a variable). Points need only compared with points from neighbouring cells.Outlier can identified according to the application need.[6]

7.1. Outlier Scoring based on KNN distances

Here KNN distance of a data object is its own outlier score 1NN,2NN. . . .K NN is calculated for every object, aggregation of 1NN to KNN is the outlier score of a perticular data object, the data objects those are having higher outlier score are consider as an outlier.

7.1.1. Working Priciple of KNN algorithm

In given N trainning data is classified into C classes , KNN algorithm identifies the K-nearest neighbours of new data object according to that object is assigned to any one class among C classes.

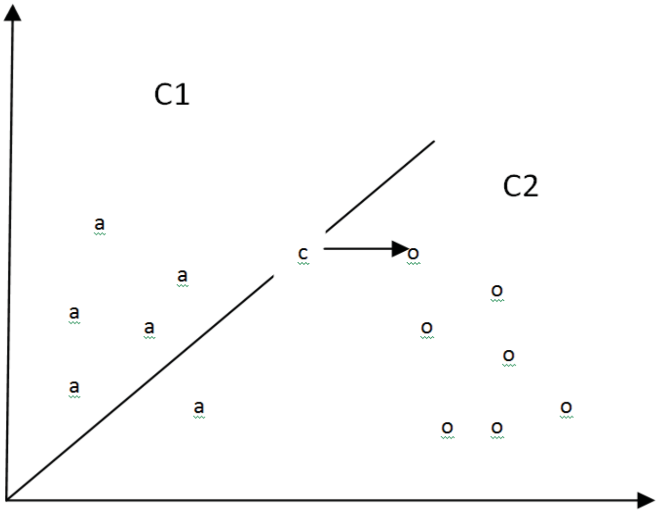

Fig 7.1 working priciple of KNN algorithm

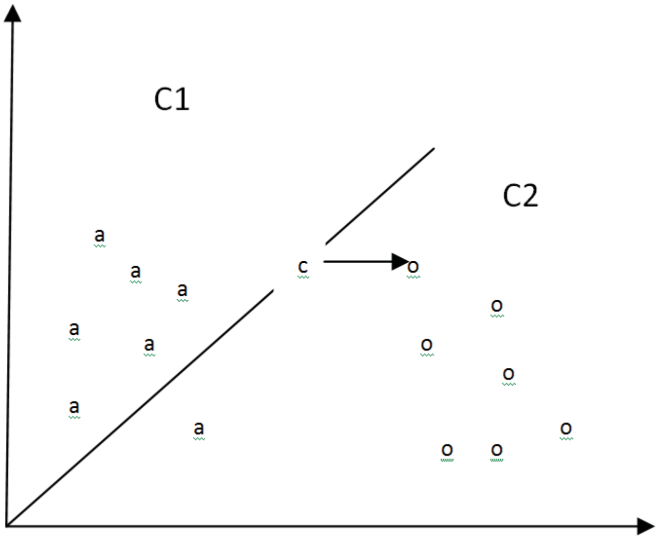

Fig 7.2 KNN neighbour node detection

We need to identify whether ‘c’ belongs to the data object category represented by ‘a’ or data object category represented by ‘o’. If k=3 c belongs to ‘o’ object category because 2 ‘o’ object is nearest to C.

7.2. Nested Loop Algorithm and Linearization for Computing top-n Outliers

With the help of nested loop algorithm we can find the distance of every data object to every other data object in a simple way.

Linearization is smoothing the multidimensional data or plotting the multidimensional data objects in straightline. After smoothing the data objects outliers are detected based on KNN distance model.

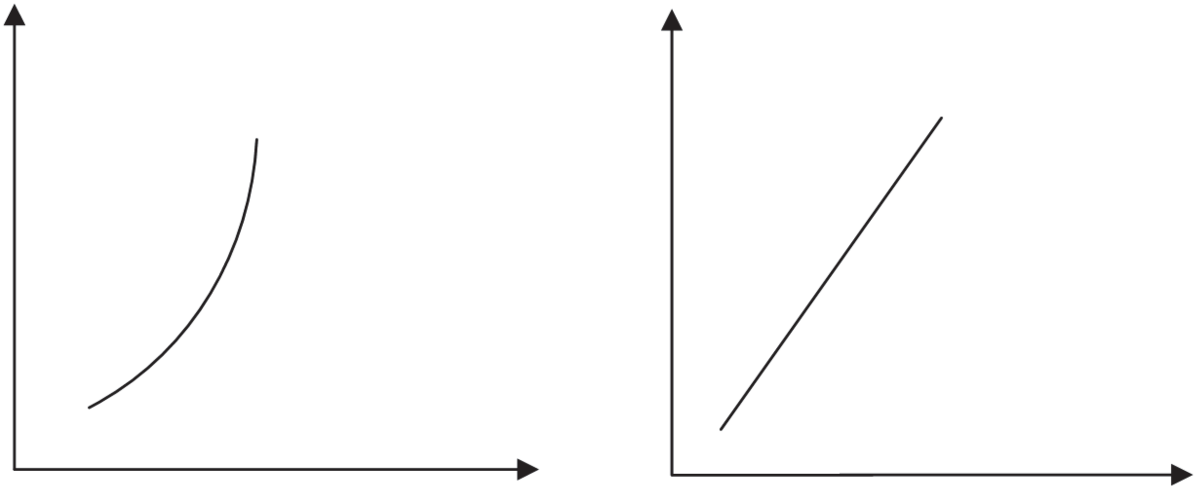

Fig 7.3 Linearization

7.3.ORCA – for finding top –n Outliers

ORCA is used NL algorithm(Nested Looping algorithm) for outlier detection , The main idea in NL algorithm is that for each data point or data object in dataset D. We keep track of its K-closest neighbours as we scan the dataset. When a data point’s or data objects kth closest neighbour has a distance that is less than the cut-off threshold C , then the data point is No longer an outlier, if any one data object‘s kth closest neibhour has a distance higher than the threshold or cut-off, then the data object is condider as an outlier.

7.4. Sample Algorithm RBRB

RBRB(Recursive Binning and Re-Projection) algorithm is used to find the outlier in multidimensional data set, RBRB scales log-linearly as a function of the number of data points and linearly as a function of the number of dimensions.

Recursive Binning and Re-Projection is used 2-phase algorithm to make the outlier mining process more fast.

7.5 Variant – Outlier detection using In-degree number

Variant is one kind of distance based outlier detection mechanism which identifies outlier based on in-degree of vertex or data object.

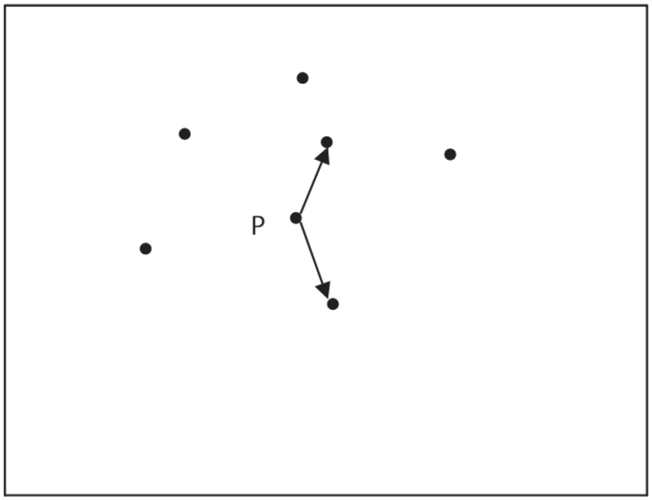

Fig 7.5.1 Indegree of P is 2

7.5.1 Steps involved in outlier detection using In-degree Number

i) KNN graph for a data set is constructed ,All data point or data objects are considered as vertices.
ii) Edge is constructed based on the following condition

Edge: If $q \in KNN(p)$ then there is a directed edge from p to q

iii)The vertices those have indegree less than the user defined threshold are considered as an outlier.[7]

VIII. DENSITY BASED OUTLIER DETECTION APPROACH

Density of a particular data object is compared with density of the neibhours of that particular data object.

The relative density of a data object compared to its neighbours is computed as an outlier score. Various methods are available to compute density, The data objects those have higher outlier score are consider as an outliers.

IX. CONCLUSION

Outlier detection is an era which contains number of techniques and approaches like labeling, scoring, model based and proximity based approaches. In outlier detection, the developer or user should select an approach or algorithm that is suitable for the following factors those are type of application, distribution and other charecteristics of dataset.

Usage or benefits of outlier is completely based on the developer or user, In some context outliers are completely removed from the dataset to have future process in smooth.

Some specific applications identify outliers from the dataset and outliers are considered and processing in future rather than normal data, for example fraud detection in finantial card transaction, here outliers are take vital role rather than normal data.

The scenario like loan sanction in banking, bankers need to identify the geniune customers and doubtful customers, Here bankers consider only outlier in dataset and suspect those are doubtful customers based on the variables or parameters.

REFERENCES

[1] Hodge, Victoria J., and Jim Austin. "A survey of outlier detection methodologies." Artificial Intelligence Review 22.2 (2004): 85-126.

[2] Hassibi PhD, Khosrow (2000). Chapter 9 on "Detecting Payment Card Fraud with Neural Networks in book "Business Applications of Neural Networks". Singapore-New Jersey-London-Hong Kong: World Scientific. pp. 141–158. ISBN 978-9810240899.

[3] Allen, George D., and Sarah Hawkins. "Phonological rhythm: Definition and development." Child phonology 1 (1980): 227-256.

[4] Casella, George, and Roger L. Berger. Statistical inference. Vol. 2. Pacific Grove, CA: Duxbury, 2002.

[5] Johnson, T., Kwok, I., and Ng, R.T. 1998. Fast computation of 2-dimensional depth contours. In Proc. Int.Conf. on Knowledge Discovery and Data Mining (KDD), New York, NY.

[6] Knorr, Edwin M., and Raymond T. Ng. "A Unified Notion of Outliers: Properties and Computation." KDD. 1997.

[7] Hautamäki, Ville, Ismo Kärkkäinen, and Pasi Fränti. "Outlier Detection Using k-Nearest Neighbour Graph." ICPR (3). 2004.

[8] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann, 2006.

[9] Hodge, Victoria J., and Jim Austin. "A survey of outlier detection methodologies." Artificial Intelligence Review 22.2 (2004): 85-126.

[10] Kriegel, Hans-Peter, and Arthur Zimek. "Angle-based outlier detection in high-dimensional data." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.`

[11] Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." ACM Sigmod Record. Vol. 30. No. 2. ACM, 2001.

[12] Zimek, Arthur, Erich Schubert, and Hans-Peter Kriegel. "Outlier Detection in High Dimensional Data." Tutorial at the 12th International Conference on Data Mining (ICDM), Brussels, Belgium. Vol. 10. 2012.