National Conference on Recent Research in Engineering and Technology (NCRRET-2015)
International Journal of Advance Engineering and Research Development (IJAERD)
e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

A SURVEY OF OUTLIER DETECTION IN DATA MINING

SHIVANI P. PATEL
Research Scholar,
G.H.Patel College of
Engg. and Tech.
Vallabh Vidyanagar, India
ptlshivani312@gmail.com

VINITA SHAH
Ass. Prof. , IT Dept.,
G.H.Patel College of
Engg. and Tech.
Vallabh Vidyanagar, India
vinitashah@gcet.ac.in

JAY VALA
Asst. Prof. , IT Dept.,
G.H.Patel College of
Engg. and Tech.
Vallabh Vidyanagar, India
jayvala@gcet.ac.in

***Abstract*—Outlier is a data point that deviates too much from the rest of dataset. Most of real-world dataset have outlier. Outlier detection plays an important role in data mining field. Outlier Detection is useful in many fields like Network intrusion detection, Credit card fraud detection, stoke market analysis, detecting outlying in wireless sensor network data, fault diagnosis in machines, etc. This paper is a survey on different Outlier detection approaches, which are statistical-based approach, deviation-based approach, distance-based approach, density-based approach. In order to deal with outlier, clustering method is used. For that K-mean is widely used to cluster the dataset then we can apply any technique for finding outliers.**

Keywords: Data Mining, Clustering, Outlier, Outlier Detection

1. INTRODUCTION

Data mining is one of the steps of "Knowledge Discovery from Dataset" process.This process discover interesting pattern from large dataset by performing data cleaning, integration, selection, mining, pattern evaluation and knowledge presentation. The overall goal of data mining is to extract information from a dataset and transform it into an understandable structure for further use. However there are many problems exist in mining data in large datasets such as data redundancy, value of attribute is not specific; data is not complete [8].
An outlier is an observation of data point that deviates too much from other point that they are generated because of faulty condition in experiment. Different applications of outlier detection are credit card fraud detection, network intrusion detection, detecting outlying in wireless sensor network data, fault diagnosis in machines, stoke market analysis, etc. where as in credit card fraud detection, credit card owner's purchasing behavior is usually changed when the card is stolen we can consider this information as outlier.

In order to deal with outlier, clustering method is used. Clustering is process of grouping similar objects of a dataset into one cluster or class. For example, in general store if we want to retrieve items easily and quickly, we can group the items in such way that similar items put into one group and another items into different group, and such grouping can be known as clustering. Cluster analysis is used in many applications such as digital image processing, data analysis, market data analysis etc. Now a day's most popular and widely used clustering algorithm for outlier detection is k-mean algorithm. Authors in [2] have used k-mean algorithm and proposed improve k-means algorithm.
This research paper discusses and compares outlier detection approach and determines better approach for outlier detection. Section 2 defines Outlier. Section 3 discusses the different clustering method. Section 4 discusses different outlier detection approaches. Section5provides the conclusion and future work.

2. OUTLIERS IN DATASET

Outliers are the abnormal data object having different behavior than normal object in the data set. Outlier can be caused by measurement or execution error. For example, person's age display as -445.

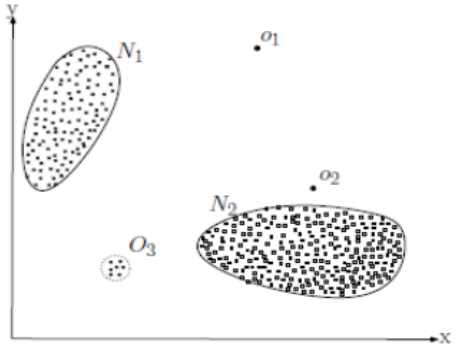

Figure 1: An Example of outlier in two-dimensional dataset [4]

Figure 1 illustrates outlier in two-dimensional data set. The dataset is grouped into three regions, N1, N2 and O3. In which most of data point are lies in two regions N1 and N2, and point that are far away from these region such as O1, O2, and data point in regions O3 are consider as Outlier [4]. Authors in [3] discuss different algorithm proposed by different researchers for detecting outliers.

3. CLUSTERING

Clustering is process of grouping similar objects in the same cluster. Clustering is one of the well-known techniques with successful application on large domain for finding patterns. The major clustering methods are: Partitioning method, Hierarchical method, Density-based method, Grid-based method, Model-based method [1].

3.1 Partitioning method: Construct a partition of a dataset D of n objects into a set of k clusters. Partition based algorithm are k-means andk-medoids. In k-means each cluster are represented by the center of the cluster. The variant of k-means algorithm is k-modes, which cluster categorical data by replacing mean of cluster with modes. K-medoids algorithms are PAM, CLARA, and CLARANS in which each cluster is represented by one of the object in the cluster. Authors in [5] have used partitioning clustering algorithm for detecting outliers.

3.2 Hierarchical method: Create a hierarchical decomposition of the set of data using some criteria. Hierarchical method is classified into agglomerative or divisive depending on whether hierarchy is formed in top-down or bottom-up form. Agglomerative is bottom-up strategy that merge cluster into larger cluster until all object are into one cluster or until some condition for termination are satisfied. Divisive is top-down strategy and is reverse of agglomerative which subdivide the cluster into small cluster.For given input set S,the goal is to produce hierarchies in which nodes represent subset of S. This method form tree structure of cluster. Each level of the tree represents a partition of input data into several cluster or group. Hierarchical clustering algorithm are BIRCH (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives). Strength of this method is no need to assume or define number of cluster initially.

3.3 Density-based method: Most of the partition based clustering methods are based on distance. Distance based method deal with only spherical-shaped cluster and difficult for arbitrary shapes. Density based method use connectivity and density function to make cluster. Density-based algorithm is DBSCAN (Density-Based Spatial Clustering of Application with Noise). This algorithm is used to discover cluster of arbitrary shapes.

3.4 Grid-Based Method: Assign object to the appropriate grid cell and compute the density of each cell. Eliminate cells, whose density is below a certain threshold. Advantages of this method are fast processing, independent of the number of data object and there is no need of distance computation.This method is depends on number of cells. Grid-based algorithms are STING (Statistical Information Grid approach) and CLIQUE (Clustering in Quest).

3.5 Model-Based Method: Model-based methods hypothesize a model for each of the cluster and find the best fit of the data to the given model [1]. Model-based method is EM (Expectation-Maximization).

The Choice of clustering algorithm depends on the type of dataset and application.

4. OUTLIER DETECTION APPROACH

In this section we discuss different outlier detection approaches. These can be categorized into: Statistical Distribution-Based approach, Distance-based approach, Deviation-based approach, density-based approach [1] [8].

4.1 Statistical distribution-based Outlier detection: This approach assumes a distribution or probability model of given dataset and identifies outlier by using discordancy test [1].Discordancy test depends on data distribution, distribution parameter and number of expected outlier. There are certain kinds of statistical distribution such as Gaussian. In which parameters are computed by assuming all data point have been generated by statistical distribution such as mean and standard deviation. Outlier is the point that has low probability to be generated by overall distribution. Disadvantage of this approach are most tests are for single attribute and require the knowledge about data distribution parameter.

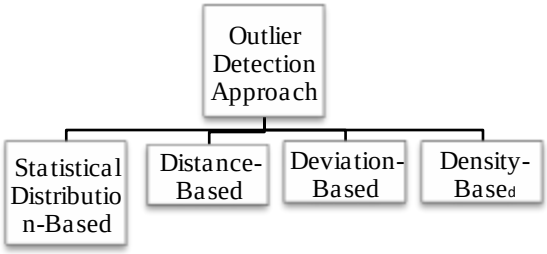

Figure 2: Outlier Detection Approaches

4.2 Distance-based Outlier detection: This approach introduces to overcome the main disadvantage of previous statistical approach, which is this approach work with multi-dimensional analysis. A distance based outlier can be define as, An object *o*, in the dataset, *D*, is a distance-based outlier with parameters *pct* and *dmin*, that is, a *DB (pct, dmin)*-outlier, if at least a fraction, *pct*, of the object in *D* lie at a distance greater than *dmin* from o [1].To find Distance between point with its neighbor, the different dissimilarity measure used are *Euclidean distance, cosine distance, city block distance*, etc. Authors in [6], used k-means algorithm to form cluster and distance-based outlier detection approach to detect outliers. They use Euclidean distance for dissimilarity measure [6],

$$d_{rs}^2 = (x_r - x_s)(x_r - x_s)'$$

Where,

$$\overline{x_r} = \frac{1}{n}\sum_j x_{rj}$$

And

$$\overline{x_s} = \frac{1}{n}\sum_j x_{sj}$$

Authors in [9] proposed k-means algorithm to cluster the dataset and use local distance-based outlier factor (LDOF) to detect outlier.

4.3 Deviation-based Outlier Detection: This approach identifies outlier by observing main characteristics of object in data set. The object that deviates too much from these feature are consider as outliers. Two technique used for deviation-based outlier detection are sequential exception technique and OLAP data cube technique. Sequential exception technique selects the sequence of subsets from the set for analysis and determines dissimilarity difference for each subset according to the previous subset in the sequence. OLAP data cube technique uses data cube to identify regions of anomalies in large multidimensional data [1].

4.4 Density-based Outlier Detection: Distance-based outlier detection approach have problem with different densities. The basic idea of this approach compares the density around point with the density around its local neighbors. The relative density of a point compared to its neighbors is computed as an outlier score. The density around a normal data object is similar to the density around its neighbors. The density around an outlier is different to the density around its neighbors. To define Local Outlier Factor (LOF), we need the concept of *k*-distance, *k*-distance neighborhood, reach ability distance, and local reach ability density [1]. K-nearest neighborhood distance of object *p*, is denoted by $N_{k\text{-}distance\ (p)}(p)$ from this we get $N_{minpts}(p)$. The reach ability distance of an object *p* with respect to object *0*, is define as $reach\_dist_{minpts}(p,o)=max\{MinPtsdistance(o),d(p,o)\}$. Local reach ability distance (lrd) of point p,inverse of the average reach-dists of the kNNs of *p* [1],

$$\mathrm{lrd(p)} = \frac{|\mathrm{Nminpt\ (p)}|}{\sum_{o\ \in \mathrm{Nminpt\ (p)}} \mathrm{reach-dist\ (p,o)}} \quad \ldots (1)$$

And Local Outlier Factor (LOF) of *p*, average ratio of neighbors of *p* and lrd of *p* [1],

$$\mathrm{LOF(p)} = \frac{\sum_{o\in \mathrm{Nminpt\ (p)}} \frac{\mathrm{lrd\ (o)}}{\mathrm{lrd\ (p)}}}{|\mathrm{Nminpt\ (p)}|} \quad \ldots (2)$$

We can determine whether a point p is a local outlier based on the computation of $LOF_{MinPts}(p)$.
Authors in [7] have used density-based outlier detection approach. They first apply density-based approach to remove noise data and then apply k-means to cluster data.

5. CONCLUSION

This paper discusses about the concept of outlier. Then we discuss different method used for clustering the dataset. We conclude that k-mean algorithm is most widely used for clustering the dataset. Next, this paper alsodescribes and compares different approaches of outlier detection which are statistical approach, distance-based approach, density-based approach, deviation-based approach. Most of researchers use distance based approach and density based approach to detect the outlier. So, in future we can combine two or more approaches to get more accuracy for detecting outliers.

REFERENCES

[1] Data mining Concepts and Techniques; Jiawei Han and Micheline Kamber; Morgan kaufmann publishers.

[2] Juntao Wang, Xiaolong Su, "*An improved K-Means clustering algorithm*", 2011, IEEE, pp.44-46

[3] Janpreet Singh, Shruti Aggarwal, "*Survey on Outlier Detection in Data Mining*", International Journal of Computer Application, (0975 – 8887) Volume 67– No.19,April 2013

[4] Karanjit Singh and Dr. Shuchita Upadhyaya, "Outlier Detection: Applications And Techniques", International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012,pp.307-323

[5] S. Vijayarani, S. Nithya, "*An Efficient Clustering Algorithm For Outlier Detection*", International Journal of Computer Application, (0975 – 8887) Volume 32– No.7,October 2011

[6] Ms. S. D. Pachgade, Ms. S. S. Dhande, "*Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach*", International Journal of Advance Research in Computer science and Software Engineering, Volume 2, Issue 6,June 2012,pp.12-16

[7] Juntao Wang, Xiaolong Su, "*An improved K-Means clustering algorithm*", 2011, IEEE, pp.44-46

[8] Jingke Xi, "*Outlier Detection Algorithm in Data Mining*", Second International Symposium on Intelligent Information Technology Application, 2008 IEEE, pp.94-97

[9] RajendraPamula, Jatindra Kumar Deka, Sukumar Nandi, "*An Outlier Detection Method based on Clustering*", Second International Conference on Emerging Applications of Information Technology, 2011 IEEE, pp.253-256