Survey on Outlier Detection in Data Mining

Janpreet Singh
Mtech. Research Scholar
Department of Computer Science and Engineering
Sri Guru Granth Sahib World University.
(SGGSWU)
Fatehgarh Sahib, Punjab, India.

Shruti Aggarwal
Assistant Professor
Department of Computer Science and Engineering
Sri Guru Granth Sahib World University.
(SGGSWU)
Fatehgarh Sahib, Punjab, India.

ABSTRACT
Data Mining is used to extract useful information from a collection of databases or data warehouses. In recent years, Data Mining has become an important field. This paper has surveyed upon data mining and its various techniques that are used to extract useful information such as clustering, and has also surveyed the techniques that are used to detect the outliers. This paper also presents various techniques used by different researchers to detect outliers and present the efficient result to the user.

Keywords: Data Mining, Clustering, Outlier, Outlier Detection

1. INTRODUCTION
Data Mining is the task of extracting useful knowledge from a collection of data bases or data warehouses, nowadays data is stored in various formats such as documents, images, audio, videos, scientific data, etc. [1]. The data collected from different applications require proper mechanism of extracting knowledge/information from large repositories for better decision making. Knowledge Discovery in Databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data. [2].

1.1 Data Preprocessing
Preprocessing is the first step of Knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared from the data warehouses and other information repositories [3][4] as shown in fig.1. The figure shows the KDD process that shows how is knowledge formed from the raw data.

1. *Data Cleaning*: In data cleaning noise is removed from the data, such as removing fields or attribute or variables that are irrelevant.
2. *Data Integration*: In this step data is collected and combined from multiple heterogeneous resources.
3. *Data Selection*: Relevant data is selected according to user need.
4. *Data Transformation*: Data is transformed into appropriate form. It involves smoothing, generalization.

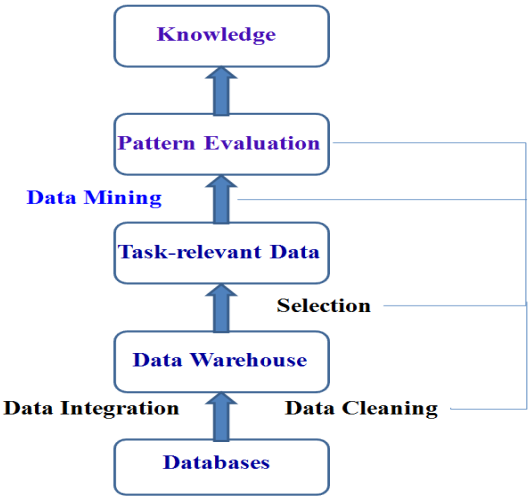

Fig.1 Data Mining Knowledge Discovery Process

1.2 Data Mining Functionalities
Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. There are various types of databases and information repositories on which data mining can be performed. There are different data mining functionalities such as, 1. Concept/Class Description: Characterization and Discrimination, 2. Classification and Prediction, 3. Cluster Analysis, 4. Evolution and Deviation Analysis, 5.Outlier Analysis. [3].

2. CLUSTERING
Clustering is the process of grouping similar objects that are different from other objects. Clustering is an unsupervised classification technique, which means that it does not have any prior knowledge of its data and results before classifying the data [5]. For example: if we want to arrange the books on the book shelf and want to retrieve them quickly and easily then we can group the books in such a way that similar book form a one group and other from another group, such grouping is known as clustering. Cluster analysis is used in a number of applications such as data analysis, image processing, market analysis etc [6]. The term clustering is also used by several research communities to describe the method of grouping unlabeled data. Clustering is used to improve the efficiency of the result by making groups of the data. So to cluster the data means specifying the data objects to a specific cluster which has similar objects or a group of objects.

2.1 Clustering Methods
Clustering is used to classify the data into different clusters. There are various clustering methods used today are:

A. Hierarchical Clustering Method

B. Density Based Clustering Method

C. Partition Based Clustering Method

D. Grid Based Clustering Method

E. Model Based Clustering Method.

A. Hierarchical Clustering Method

In hierarchical clustering algorithm data objects are grouped to form a tree shaped structure there are two types of hierarchical clusters, agglomerative hierarchical cluster and divisive hierarchical clusters. Agglomerative hierarchical cluster, is bottom-up approach and the divisive hierarchical clusters, is top-down approach. Some scalable clustering methods are BRICH (Balance Iterative Reducing and Clustering using Hierarchies) [7], CURE (Cluster Using Representatives) [3][6][8].

B. Density Based Clustering Method

In density Based Method clusters are made according to the density of the data The general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold value[3].

C. Partitioning Methods

In this method a database of n objects or data tuples, are partitioned into k number of partitions, where each partition represents a cluster. The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects of different clusters are "far apart" or very different. There are few popular heuristic methods, such as (1) the k-means algorithm, where each cluster is represented by the mean value of the objects in the cluster, and (2) the k-medoid algorithm, where each cluster is represented by one of the objects located near the center of the cluster [3].

D. Grid Based Method

In Grid-based methods the object is placed into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure. The main advantage of this approach is its fast processing time [3].

E. Model Based Method

In Model-based methods each of the clusters is best fitted to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points [3].

3. OUTLIER DETECTION

Outlier detection is a fundamental issue in data mining; specifically it has been used to detect and remove anomalous objects from data [9]. Outliers are the objects that are different from the rest of data set. Outlier detection is used in many fields and applications to detect anomalous objects, such as to detect outlier from computer network intrusions, credit card fraud detection, criminal activity in e-commerce, etc. Outlier detection can also be defined as finding of the unrealistic data that is of no use with rest of the data set.

3.1 Outlier Detection Methods

A. Density Based Outlier Detection
B. Distance Based Outlier Detection
C. Clustering Based Outlier Detection
D. Partition Based Outlier Detection

A. Density Based Outlier Detection

In density based method outlier are detected after clustering the data. The data objects that do not fit into the density of the cluster are declared as the outlier. Markus M. Breunig et al. has proposed a method in which outlier is find on the bases of the local outlier factor that how much the object is different from the other data objects with respect to the surrounding neighborhood [10]. Raghuvira Pratap et al. have used a method based on density in which an efficient density based k-medoids clustering algorithm has been used to overcome the drawbacks of DBSCAN and k-medoids clustering algorithms [11][16].

B. Distance Based Outlier Detection

In distance Based method outliers are found according to the distance between the data objects from the centroid or the center point of the cluster. Moh'd Belal Al- Zoubi has proposed a method in which first they perform the Partitioning Around Medoids clustering algorithm. Small clusters are then determined and considered as outlier clusters and after that the rest of outliers (if any) are then detected based on calculating the absolute distances between the medoid of the current cluster and each one of the points in the same cluster [13]. Ms. S. D. Pachgade and Ms. S. S. Dhande has used a k-mean clustering algorithm to cluster the dataset and used Euclidean distance to find outlier by finding the distance between the objects [15][16].

C. Clustering Based Outlier Detection

In clustering based technique outliers are found by dividing the data set into clusters by some methods such as k-mean or k-medoid. H.S.Behera et al. has proposed a method based on Clustering based outlier detection for effective data mining which uses k-means clustering algorithm to cluster the data sets and outlier finding technique (OFT) to find out outlier on the basis of density based and distance based outlier finding technique [9]. Parneeta Dhaliwal et al. have proposed a cluster based approach in which they have used a weighted k-median technique to detect outlier [14].

D. Partition Based Outlier Detection

K-means is a well known partitioning based clustering technique that attempts to find a user specified number of clusters represented by their centroids [12]. Rajashree Dash et al. has proposed a hybridized k-mean clustering approach for high dimensional dataset in which they have used Principal Component Analysis (PCA) method with an improved initial centroids finding technique with k-mean clustering [12].

4. COMPARATIVE STUDY OF OUTLIER METHODS

Outlier detection methods are used to detect the outliers from the various database systems and application software's. There are various methods used now days to detect outlier so here we will have a Table 1 which shows the comparative study of different algorithms used by different researchers. Much of the researchers have used partition based algorithms to divide the data set into k number of clusters to detect outliers with different outlier detection method, such as distance based, density based etc.

Table 1. Comparative study of different algorithms by different researchers.

Outlier Detection Methods	Proposed Algorithm	Researcher
Density Based Outlier Detection	Finding local outlier factor which is different from the other data objects with respect to the surrounding neighborhood.	Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander
	K-medoid algorithm with DBSCAN	Raghuvira Pratap, Suvarna, J Rama Devi, Dr.K Nageswara Rao
Distance Based Outlier Detection	Partitioning Around Medoids with absolute distance	Moh'd Belal Al- Zoubi
	K-mean Clustering algorithm with Euclidean distance	Ms. S. D. Pachgade and Ms. S. S. Dhande
Clustering Based Outlier Detection	Hybridized K-mean clustering algorithm with distance based and density based	H.S.Behera, AbhishekGhosh, SipakkuMishra
	Weighted K-median clustering algorithm	Parneeta Dhaliwal, MPS Bhatia and Priti Bansal
Partition Base Outlier Detection	Hybridized K-mean clustering algorithm with Principal Component Analysis (PCA) method.	Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath, MiluAcharya

Here Table1 shows the comparative study of the different methods used by different researchers. In the table1 much of the researchers have used the k-mean algorithm because it is the most popular partition based algorithm which divide the data in to k number of clusters and detect the outliers. Some of the researches have improved the k-mean algorithm according to their need to detect outlier and uses additional algorithms to detect hidden outliers from the data set. Outlier detection is used on various types of dataset, such as graphical dataset, numerical dataset, Text dataset, and can also be used on the pictures etc.

5. CONCLUSION

In today's life data mining is used in various fields, due to the nature of extracting useful data from a collection of databases or data warehouses, data mining is used, with various algorithms and techniques to extract useful data from the databases. Clustering is the technique of extracting useful data from databases, but with the extraction of the object from the dataset an unwanted data also comes that is known as outlier. To detect outlier there are various methods. In recent years outlier detection techniques are used in various field and applications such as in credit card fraud detection etc. Due to the increase of data on the web outlier detection has become an important part of the data mining. So to detect outlier from different data sets different outlier detection techniques are used with different clustering algorithms. Most popular clustering algorithm is k-mean algorithm and is widely used to cluster the data set and for outlier detection and can be improved according to need to detect outliers.

6. REFERENCES

[1] Venkatadri.M, Dr. Lokanatha C. Reddy"A Review on Data mining from Past to the Future" International Journal of Computer Applications (0975 –8887) Volume 15– No.7, February 2011.

[2] Heikki, Mannila.1996" Data mining: machine learning, statistics, and databases" IEEE.

[3] Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2nd Ed.

[4] Dhanashree S. Deshpande "A survey on web Data Mining Application" Emerging Trends in Computer Science and Information Technology -2012.

[5] H.S Behera, Rosly Boy,Lingdoh, Diptendra Kodama singh "An Improved hybridized k-means clustering algorithm for high dimensional data set & it's performance analysis" International Journal on Computer Science and Engineering (IJCSE).

[6] M.Vijayalakshmi, M.Renuka Devi "A Survey of different issue of different clustering algorithms used in large data sets" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3,March2012.

[7] M. Livny, R.Ramakrishnan, T. Zhang, 1996." BIRCH: An Efficient Clustering Method for Very Large Databases". Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.

[8] S. Guha, R. Rastogi, and K. Shim, 1998. "CURE: An Efficient Clustering Algorithm for Large Databases". Proc. ACM Int'l Conf. Management of Data.

[9] H.S.Behera, AbhishekGhosh, SipakkuMishra "A New Hybridized K-Means Clustering Based Outlier Detection Technique for Effective Data Mining "International Journal of Advanced Research in Computer Science and Software Engineeing, Volume 2, Issue 4, April 2012.

[10] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander" LOF: Identifying Density-Based Local Outliers".

[11] Raghuvira Pratap, K Suvarna, J Rama Devi, Dr.K Nageswara Rao "Efficient Density based Improved K-

Medoids " International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[12] RajashreeDash, Debahuti Mishra, Amiya Kumar Rath, MiluAcharya " A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology, Vol. 2, No. 2, 2010.

[13] Moh"d Belal Al-Zoubi "An Effective Clustering-Based Approach for Outlier Detection" European Journal of Scientific Research Vol.28 No.2 (2009).

[14] Parneeta Dhaliwal, MPS Bhatia and Priti Bansal "A Cluster-based Approach for Outlier Detection in Dynamic Data Streams" Vol. 2, Issue 2, Feb 2010.

[15] Ms. S. D. Pachgade, Ms. S. S. Dhande "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, June 2012.

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