International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.6, No.10, October 2017 DOI:10.15693/ijaist/2017.v6i10.1-5

Outlier Detection using Clustering Algorithm A Survey

B. Angelin M.Sc.,
Research Scholar (M.Phil.),
Pg.& Research Department of Computer Science,
Government Arts College (Autonomous),
Coimbatore-18.
Mail: angelindia93@gmail.com

Dr. D. Devakumari MCA, M.Phil., Ph.D.,
Assistant Professor in Computer Science,
Pg. & Research Department of Computer Science,
Government Arts College (Autonomous),
Coimbatore-18.
Mail: ramdevshri@gmail.com

Abstract- Clustering is the task of assigning a set of data objects into groups called clusters so that the objects in the same cluster are more similar in some sense to each other than to those in other cluster. Data items whose values are different from rest of the data or whose values fall outside the described range are called outliers. Outlier detection is an important issue in data mining, where it is used to identify and eliminate anomalous data objects from given data set. This paper provides a brief survey on clustering techniques and outlier detection techniques. Particularly the K means clustering algorithm for outlier detection is discussed.

Index Terms- Data Mining, Clustering, Outlier Detection, K-means algorithm**.**

1. Introduction

Clustering is an important technique in data analysis and data mining applications. Clustering is the partition of data into groups in a way that objects in the same group are more similar to each other and different from objects of other groups. These groups are called clusters. Data mining is an activity which is carried out in different steps. These are anomaly detection, association, classification and clustering. Data mining is the searching and training of large data sets, in order to find out significant patterns and rules. Data mining is one of the best ways to illustrate the difference among data and information. It converts data into information. [4][18].

There are two learning approaches used, supervised learning and unsupervised learning.

1.1 Supervised learning:

In supervised learning, also called direct data mining, the variables under investigation are divided into groups. The aim of this analysis is to specify a relation between the dependent variable and explanatory variables. The values of the dependent variable must be known for a sufficiently large part of the data set to continue with directed data mining techniques. Supervised learning requires that target variable should be well defined and that an appropriate number of its values are given.

1.2 Unsupervised learning:

In unsupervised learning, all the variables are treated in similar way, there is no distinction contrast to the term undirected data mining, still there is some target to achieve. This target strength can be data reduction as general or more specific like clustering. The allotting line between unsupervised learning and supervised learning is the same that distinguishes discriminant analysis after cluster analysis. Unsupervised learning is typically such that either the target variable is absent or is a descriptive variable.

2. Clustering

A good clustering method will produce high quality clusters with high intra-clusters parallel and low inter cluster parallel. In well-separated clusters points are nearer to every other point in the cluster than to any point not in the cluster. Center based: It resides then the other clusters then it is named as center based cluster. Contiguous cluster: Nearer to one or more other points in the cluster than to any point not in the cluster. Density based cluster: Low density states, from other regions of high density. This type of cluster used only when the clusters are irregular or intertwined and when sound outliers are present. Shared property-share some common property or represents a particular notion [1] [2] [18].

2.1 Types of data clustering techniques:

2.1.1 Hierarchical technique:

This approach constructs the cluster by recursively by partitioning the instances in either a top-down or bottom-up fashion. There are two types of hierarchical clustering.

(i) Agglomerative hierarchical clustering:

In which each object primarily represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained.

(ii) Divisive hierarchical clustering:

In which all objects primarily belong to one cluster. Then the cluster is divided into sub-clusters, which are

Successively divided into their own sub-cluster. This process endures until the desired cluster structure is obtained.

2.1.2 Partitioning technique:

Partitioning methods consist of a set of M cluster and each object belongs to one single cluster. Partition clustering algorithms split the objects into number of clusters. This method creates various partitions [18].

2.1.3 Density Based techniques:

Density based algorithm finds the clusters according to the regions which are developed with high density rather than total data set. The objects in these sparse areas which are required to separate cluster in these sparse areas are generally considered to be noise and border points. Clusters are recognized by looking at the density of points [7],[13][18].

2.1.4 Grid-Based technique:

This clustering method uses a multi resolution grid data structure. It partitions the space of objects into fixed number of cells that form a grid structure on which all clustering actions are performed. The main advantage of this approach is its fast processing time. A grid-based clustering algorithm consists of the following five steps.

(i) Partitioning data space into a limited cells.
(ii) Computing the cell density for each cell.
(iii) Arrange the cells according to their density.
(iv) Finding cluster centers.
(v) Traversal of neighbor cells [18].

3. Outlier detection

Outlier detection is an active part for research in data set mining community. Detecting outliers and examining large data set can lead to discovery of behavior in telecommunication, web logs, and web document etc. A lot of outlier detection procedures exist and most of them are based on distance measure. Identifying outlier within data led to the discovery of useful and meaningful knowledge or improve data analysis for additional discovery within numerous applications domains. It also helps to avoid a wrong conclusion. Outlier patterns in data are those that do not imitate a well-defined notation of normal behavior. Effective outlier detection needs the construction of a model that accurately represent the data [17].

3.1 Existing outlier detection algorithm:

One of the simple problems of data mining is the outlier detection. In this section different existing outlier detection techniques have been discussed that are used for detection and deduction of outliers.

3.1.1 Statistical outlier detection:

Statistical model can only handle one attribute and it can switch multi attributes and handle data efficiently up to the k<4. Two actions have been described for the outlier detection. They are parametric method and nonparametric method [7], [13].

3.1.2 Distribution outlier detection:

The outlier is detected on the basis of the probability circulation. Method of detecting outlier based on the general pattern within data points is measured quite efficiently [7].

3.1.3 Depth based outlier detection:

The detection of the outlier is done on the root of the depth. The data objects that have lesser depth have high probability of being considered as outlier. It sanctions the processing of the multidimensional data objects [7].

3.1.4 Distance based outlier detection:

This is one of the algorithms of outlier detection that is in need of the distance between the points. The neighbors of the point are designated and checked in this method. This technique is quite efficient as there is no need of describing the explicit distribution that defines the peculiarity. In distance based method k-nearest neighbor distance from the original data points are measured for calculating outlier score instead of pre aggregated data so outlier detection is performed at a final granularity than other methods like clustering or density based method [7][13].

3.1.5 Density based outlier detection:

In density based outlier detection, density around a data point is compared through the density around its local neighbors. The relative density of a point compared to its neighbors is calculated as an outlier score. The density based outlier detection process are more effective than distance based method. But they are more complicated and computationally expensive because they contain density of both the point and its neighbors also. Detecting the outlier each object that is present is qualified a local outlier factor. The local outlier factor is basically the degree allocated to object. The object with high local outlier factor is termed as outlier and the objects having low local outlier factor are reflected to be normal [3][7][13].

3.1.6 Sliding windows based outlier detection:

Streaming data uses sliding window concept which is used for keeping the statistical information in data stream. The window is classifying two sliding end points. Bothe ends are active. During the moving method, both ends are moving in

The same direction and flowing the same units. Let W be the window size so only the last W records to arrive in data stream are related at any point of time. Choosing accurate window size in sliding window based outlier detection is mandatory. The selection of the window is not dependent on the data point used for execution which gives poor result over outlier detection. Some outlier were considered as inliers in other window, so this process efficient [7], [13].

4. Outlier detection using clustering

Clustering based outlier detection is an unsupervised outlier detection technique in which class label as "normal" or "outlier" are not presented. Clustering means learning by observation rather than learning by samples. Clustering based outlier detection technique for evolving data stream that allots weight to attribute according to its relevance in mining task. Outlier detection technique is quite effective as the data from the database is initially segmented into clusters. In every cluster each data point is approved as a degree of the membership. The outlier is detected without any interference in the clustering method. Clustering on streaming data is characterized by grid based and k-means/k median method.

The clustering techniques are greatly helpful to detect the outlier and they are called cluster based outlier detection. The clustering based technique involve a clustering step which partitions the data into groups which contains parallel objects. Clustering is used to increase the efficiency of the result by making groups of the data. The goal of clustering algorithms is to group objects into meaningful sub classes. In order to use clustering in data streams the requirements are to be created for overall high quality clusters. There are numerous types of clustering techniques useful for outlier detection. The clustering algorithm for data streams should be adaptive in the sense that up to data cluster are available at any time taking new data items. There are no predefined class label exist for the data points. Outlier detection exactness is calculation in order to find out the number of outlier detected by the clustering algorithms.

The data streams was separated into chunks of data and mined for temporal outliers. Number of phases were tested for being final outlier. Declare a point as an outlier for the data stream but call it temporal outlier for the certain chunk of data. This point may be an outlier for the present data but may not be an outlier for the fallowing data chunk as the data stream is dynamic. We declare it as an outlier for the data stream and are not included additional for clustering. It has been used to detect and eliminate anomalous objects from data. Many researchers whether clustering algorithms are an appropriate special for outlier detection. Outlier detection technique of finding outlier over clustering [4][17][8][11].

5. K-means Clustering Algorithm for Outlier detection

K means clustering technique is a widely used most standard clustering tool used in scientific and industrial Applications. Cluster analysis goals to partition 'n' observations into k clusters. K means is the most popular partitioning technique of clustering [2], [16]. K means is a representative objects based clustering algorithm. K means is a prototype based humble partition clustering technique which attempts to find a user specified k number of clusters. Original K-mean algorithm select initial centroids and medoids randomly that affect the excellence of the resulting clusters. The new approach for the K-mean algorithm removes the deficiency of existing K-mean. K-mean is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The main hint is to define K centroids one for each cluster. These centroids should be placed in a cunning way because different locations produced different results. As a result of this loop we may notice that the K centroids change their location stage by stage until no more changes are done [10].

The K-means algorithm is a well-known partition based unsupervised clustering algorithm. The K-means discovers a locally optimal solution by reducing a distance measure between each data and its nearest cluster center. K-means algorithm which removes the problem of generation of empty clusters and increase the efficiency of traditional K-means algorithm. Choose initial K-centroids phase calculate the distance phase and recalculating cluster center phase center have achieved using divide and conquer method [6]. K-means algorithm helps to avoid the formation of unfilled cluster using data structure. Adaptive k-means clustering method goals to overcome the dependence of traditional K-means on the selection of the number of clusters. Calculate the distance between data in each group and the value calculated and then the result will be stored lastly calculate mean for each row this value will be taken as initial centroids.

Local outlier index calculation increase the operand, so the next step should research the method of filtering data with pertinence and discuss the method of decreasing the operand of algorithm so as to improve the arithmetic speed. K-means clustering algorithm is superior to the average accuracy of the traditional algorithm. The developed algorithm can optimize the clustering center through the local outlier index calculation and clustering accuracy as a whole [19]. Data items whose values are altered from rest of data or whose values falls outside the described range are called outlier. Detecting outlier has main applications in data cleaning as well as in the mining of abnormal points for fraud detection, intrusion detection, network sensors, stock market analysis and marketing. Data points are the basic hint to find out an outlier. Outliers may

Present skew or complexity into models of the data, making it difficult, if not impossible to fit an accurate model to the data in a computationally feasible manner. Outliers are ordinary elements when elements specified as input but will direct in inefficient outputs with them. [5].

CONCLUSION

In this paper, various data clustering algorithms like Hierarchical, Partitioning, Density based and Grid based are discussed. Also existing outlier detection techniques namely, Statistical, Distribution, Depth based, Distance based etc are analyzed. The K means clustering algorithm based outlier detection is found to be an easy and simple method that is normally used. The possibilities for enhancing the K means algorithm for outlier detection is also studied.

REFERENCES

[1] Mr.Anand Khandare1, Dr. A.S. Alvi2 "Efficient Clustering Algorithm with Improved Clusters Quality" IOSR Journal of Computer Engineering (2278-8727) Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), PP 15-19.

[2] Anshul yadav, Sakshi dhingra "A Review on K-means Clustering Technique" International Journal of Latest Research in Science and Technology (2278-5299) volume 5, issue 4, July-Augest2016, PP 13-16.

[3] H.S.Behera Abhishek Ghosh. Sipak ku. Mishra, "A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering (2277-128X) Volume 2, Issue 4, April 2012, PP 287-292.

[4] Dr.T.Christopher, T.Divya, "A Study of Clustering Based Algorithm for Outlier Detection in Data Streams" Proceeding of the UGC Sponsored

[5] National Conference on Advanced Networking and Applications, 27th march 2015, PP 194-197.

[6] J. James Manoharan1, Dr.S.Hari Ganesh2 Ph.D., Dr. J.G.R. Sathiaseelan3, "Outlier Detection Using Enhanced K-Means Clustering Algorithm And Weight Based Center Approach" International Journal of Computer Science and Mobile Computing (2320-088X) Vol. 5, Issue. 4, April 2016, PP 453-464.

[7] J. James Manoharan1, S. Hari Ganesh2, "A FRAMEWORK FOR ENHANCING THE EFFICIENCY OF K-MEANS CLUSTERING ALGORITHM TO AVOID FORMATION OF EMPTY CLUSTERS" International Journal on Information Sciences and Computing Vol. 10 No. 2 July 2016, PP 22-31.

[8] Kamaljeet Kaur, Atul Garg "Comparative Study of Outlier Detection Algorithms" International Journal of Computer Applications (0975-8887) Volume 147 – No. 9, August 2016.

[9] Manish Gupta, Jing Gao, member IEEE, caru c. Aggarwal, fellow, IEEE, and Jiawei Han, fellow IEEE, "Outlier Detection for Temporal Data: A Survey" IEEE Transactions on knowledge and data engineering , Volume 26, no 9, September 2014, PP 2250-2267.

[10] Mr. Mukesh K.Deshmukh 1, Prof. A. S. Kapse 2 "A Survey On Outlier Detection Technique In Streaming Data Using Data Clustering Approach" International Journal Of Engineering And Computer Science (2319-7242) Volume 5 Issue 1 January 2016, PP 15453-15456.

[11] Pallavi Purohit, Ritesh Joshi "A New Efficient Approach towards k-means Clustering Algorithm" International Journal of Computer Applications (0975-8887) Volume 65– No.11, March 2013, PP 7-10.

[12] Parneeta Dhaliwal , MPS Bhatia and Priti Bansal, "A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner)" JOURNAL OF COMPUTING (2151-9617) VOLUME 2, ISSUE 2, FEBRUARY 2010, PP 74-80.

[13] Parmeet Kaur1, Kanwarpreet Kaur "A Review on Outlier Detection for Data Cleaning in Data Mining" International Journal of Innovative Research in Computer and Communication Engineering (2320-9798) Vol. 4, Issue 7, July 2016, PP 14373-14376.

[14] Pooja Thakkar, Jay Vala, Vishal Prajapati "Survey on Outlier Detection in Data Stream" International Journal of Computer Applications (0975-8887) Volume 136 – No.2, February 2016, PP 13-16.

[15] Mr. Raghav M. Purankar 1 , Prof. Pragati Patil 2 "A Survey paper on An Effective Analytical Approaches for Detecting Outlier in Continuous Time Variant Data Stream" International Journal Of Engineering And Computer Science (2319-7242) Volume 4 Issue 11 Nov 2015, PP 14946-14949

[16] Rishikesh Suryawanshi Shubha Puthran "Review of Various Enhancement for Clustering Algorithms in Big Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering (2277-128X) Volume 5, Issue 4, 2015, PP 742-747.

[17] G. Sathiya and P. Kavitha "An Efficient Enhanced K-Means Approach with Improved Initial Cluster Centers" Middle-East Journal of Scientific Research 20 (4), (1990-9233) 2014, PP 485-491.

[18] Sreevidya s s, "Detection of Outliers in Data Stream using Clustering Method" international journal of science, engineering and technology research (2278-7798) volume 4, issue 3, march 2015, PP 559-563.

[19] Sukhvir Kaur, "SURVEY OF DIFFERENT DATA CLUSTERING ALGORITHMS" International Journal of Computer Science and Mobile Computing (2320-088X) Vol.5 Issue.5, May- 2016, PP 584-588.

[20] Xiang Li1,2,3,*, Zhenwei Wei2,3,4 and Lingling Li "An Improved K-means Clustering Algorithm Based on Meliorated Initial Centre" Advances in Intelligent Systems Research, volume 133, 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016), PP 73-76.

Author Profile

Dr. D. Devakumari has received M. Phil degree from Manonmaniam Sundaranar University in 2003 and Ph.D from Mother Teresa Womens' University in 2013. Currently she is working as Assistant Professor in the PG and Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India. Her research papers have been published in International journals including Inderscience, Springer etc. She has presented papers in National and International Conferences. Her research interests include Data Pre-processing and Pattern Recognition.

Miss. B.Angelin has received BCA degree from Pioneer college of Arts and Science and M.Sc from Government Arts College, Coimbatore. Pursuing her M.Phil degree from Government Arts College, Pg and Research Department of computer Science Coimbatore, India. Her Research interested area is Data Mining.