NATURALISTA CAMPANO
ISSN: 1827-7160
Volume 28 Issue 1, 2024

A Review on an Outlier Detection Using Clustering Algorithms and Optimization Techniques

Mrs M.Hemalatha[1], Dr.N.Kamaraj[2]

[1]Ph.D Research Scholar Department of Computer Science SRMV College of Arts and Science Coimbatore, India.
[2]Assistant Professor Department of Computer Science Coimbatore, India

Email: hema212latha@gmail.com[1], nkamaraj17@gmail.com[2]

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 moresimilar 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 detectionis an important issue in data mining, where it is used to identify and eliminate anomalous data objects from given data set. Outlier detection is an essential step in the data mining process. Its main purpose to remove the incompatible data from the original data. This purpose helps in the removal of data which necessary for carrying out to speed up the applicationslike classification, data perturbation and compression. It plays an important role in the weather forecasting, performance analysis of sports person and network intrusion detection systems. The outlier for the single variable can be easily observedbut for the n-variable it becomes a tedious process. To enhance the performance of outlier detection in n-variable or attributes several methods were discussed. This paper provides a brief survey on clustering techniques and outlier detection techniques.Particularly the k-means outlier detectionand Optimization Techniques for outlier detection is discussed.

Keywords: Data Mining, Clustering, Outlier Detection, K-Means Algorithm.

1. Introduction

Data Mining is a vast field to explore and exploit the larger datasets to determine ameaningful pattern or rules. It is differ from the normal prediction because it find out the future outcomes based on the mining process. The role of data mining is important in the followingfield like card fraud detection, bio medical field, Intruder detection in networks and Qualitative data assessment. The advantages of data mining are automated decision making, predicting the future results and cost reduction. But it also faces problems like big data sets, over fitting models and privacy and security[1].

The big data problem can be overcome by using the cloud and artificial intelligence techniques. The privacy and security problem isovercome by the advanced encryption techniques. But the major drawback is the over fitting models. Because in over-fitting models the larger dataset training results in mere prediction and smaller datasets training results in false prediction. This problem can be overcome with the help of Outlier detection. Outlier detection is a process of detecting the irrelevant information that differs from the remaining dataset. Outlier is also defined in twoways which is mentioned in [2] as follows. Firstdefinition of outlier is defined as theappearance of data which is differed from the remaining set of data. Second definition of Outlier is an observation which appears to be aninconsistent to the remaining set of data.Generally, the outliers are classified into three types namely point outlier, Context outlier and collective outlier. Point outlier means a datapoint which differ from the remaining set of data. Context outlier is one in which the behavior or the attributes of the data is differed. Collective outlier is one in which a group of data is differed from the other groups of data. The detection of outlier is classified into two types as classic outlier and spatial outlier. The classic outlier has four categories namely statistical based approach, deviation based approach, distance based approach and density based approach.

The spatial outlier detection has two categories namely space based and Graph based approach [2]. The statistical outlier detection is used in the Wireless sensor networks for the node information and intruder detection

[3]. The statistical approach is applied on determining the correlation property on the temporal and spatial properties of the WSN data. [4]
Each of these methodologies presents focal points and impediments, hence in the current years numerous commitments have beenproposed to beat them and enhance the natureof the information. Established strategies areregularly not reasonable to treat some specific databases, along these lines late examinations have been directed on exception location forthese sort of datasets. Specifically, a high number of commitments in view of counterfeit consciousness, hereditary calculations and picture preparing as to grow new effective exceptions discovery strategies that can be appropriate in a wide range of utilizations.[5][6][10]
Outlier detection has been an imperative idea in the domain of information investigation. As of late, a few application areas have understood the immediate mapping between exceptions in information and certifiable abnormalities that are of extraordinary enthusiasm to an investigator. Exceptionidentification has been looked into inside different application spaces and learning disciplines.[7][8] This review gives a far reaching outline of existing anomaly identification methods by arranging them along various measurements. [9]

An Overview of OutlierDetection
Outlier detection is an active part forresearch in dataset mining community. Detection outliers and examining large data set can lead to discovery of behavior in tele communication, web logs, and web document etc. a lot of outlier detection procedures exist and most of them are based on distancemeasure. Identifying outlier within data led to the discovery of useful and meaningful knowledge or improve data analysis for additional discovery within numerous applications domains. Its 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.[11]

An Outlier Detection UsingClustering Algorithms
Clustering based outlier detection is an unsupervised outlier detection technique inwhich 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 method.
The data streams was separate into chunks of data and mined for temporal outliers. Number of phases were tested for temporal outliers. Number of phases were tested for being final outlier. Declare a points 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 asthe 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 todetect and eliminate anomalous objects fromdata. Many researchers whether clustering algorithms are an appropriate special for outlier detection. Outlier detection technique of findingoutlier over clustering.[12][13]

Challenges in OutlierDetection
A key test in exception discovery is that it includes investigating the concealed space. As said before, at a theoretical level, an exception can be characterized as an example that does not fit in with expected typical conduct. A clear approach will be to characterize a district speaking to typical conduct and announce any perception in theinformation which does not have a place with this ordinary locale as an exception. In any case, a few variables make this obviously basic approach exceptionally difficult. [14][15].

Defining a typical area which includes each conceivable ordinary conduct is exceptionally troublesome. Often time's typical conduct continues developing and a current idea of ordinary conduct won't be adequately illustrative later on. The limit amongst typical and distant conduct is frequently fluffy. In this way a remote perception which lies near the limit can be really ordinary and the other way around.
The correct idea of an exception is distinctive for various application spaces. Each application space forces an arrangement of necessities and imperatives offering ascend to a particular issuedetailing for anomaly location.
Availability of named data forpreparing/approval is frequently a noteworthy issue while building up an anomaly

locationprocedure.

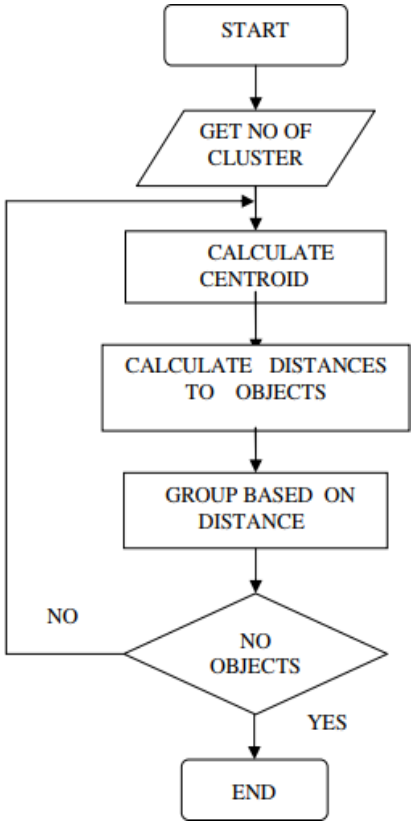

Fig: 1 Detecting outlier

In a few cases in which exceptions are the after effect of noxious activities, the vindictive foes adjust to mention the peripheral objective facts seem like typical, consequently makingthe undertaking of characterizing ordinary conduct more troublesome.
Often the information contains hugeamount which is like the genuine exceptionsand consequently is hard to recognize and expel.

1. Assign each observation to the cluster withthe nearest center
2. Update each cluster center as the mean forpoints in that cluster. [16][17][18].

Dragon Fly K-Means Clustring Algorithm for Outlier Detection
K-means clustering techniques is a widely used most standard clustering tool used in scientific and industrial applications. Cluster analysis goals to partition 'n' observations intok clusters. K-means is the most popular partitioning technique of clustering. K-means isa representative objects based clusteringalgorithm. K-means is a prototype based humble partition clustering technique which attempts to find a user sepecified k number of clusters. Original k-means algorithm selectinitial centroids and medoids randomly that affect the excellence of the resulting clusters. The new approach for the K-means 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 [19].

The K-means algorithm is a well-knownpartition 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. Kmeans algorithm which removes the problem of generation of empty clusters and increase the efficiency of traditional Kmeans algorithm. Choose initial K- centroids phase calculate the distance phase and recalculating cluster center phase center have achieved using divide and conquer method [20].
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 theclustering center through the local outlier index calculation and clustering accuracy as a whole [21].

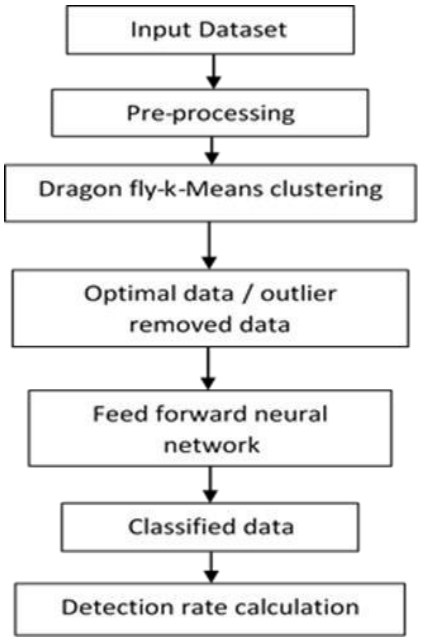

Fig 2: Dragon fly K-means Outlier Detection

2. Conclusion

In the present study, the main focus is application data mining. Outlier detection techniques are studied. 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 alsostudied. It is interesting if these methods are compared with other existing methods.
In order to conduct various experiments, in this study, only limited size datasets are used. Because of the explosive growth of available information, aseries of experiments and investigations are necessary to establish the potential utility ofthe proposed methods in real time datasets. Here, taking real time covid dataset using k means clustering algorithm to detect the outliers such as vaccinated, non vaccinated.

3. References

1. Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education India.
2. Bansal, R., Gaur, N., & Singh, S. N. (2016, January). Outlier detection: applications and techniques in data mining. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 373-377). IEEE.
3. Zhang, Y., Hamm, N. A., Meratnia, N., Stein, A., Van De Voort, M., & Havinga P. J. (2021). Statistics

based outlier detection for wireless sensor networks. International Journal of Geographical Information Science, 26(8), 1373-1392.

4. Wahid, A., & Rao, A. C. S. (2019). A distance-based outlier detection using particle swarm. optimization technique. In Information and Communication of Computer Science and Mobile Computing (2320-088X) Vol. 5, Issue. 4,April 2016, PP 453- 464.
5. Kamalov, F., & Leung, H. H. (2020). PP 7- 10.Outlier detection in high dimensional
6. J. James Manoharan1, Dr.S.Hari Ganesh2 Miner)" JOURNAL OF COMPUTING Ph.D.,Dr. J.G.R. Sathiaseelan3, "Outlier (2151-9617) VOLUME 2, ISSUE 2, Detection Using Enhanced K-Means FEBRUARY 2010, PP 74-80. Clustering Algorithm And Weight Based.
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 onInformation 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, "OutlierDetection 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 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.
12. Parneeta Dhaliwal , MPS Bhatia and Priti data. Journal of Information & Bansal, "A Cluster-based Approach for Knowledge Management, 19(01), Outlier Detection in Dynamic Data 2040013. Streams KORM: k-median Outlier.
13. Parmeet Kaur1, Kanwarpreet Kaur "A Center Approach" International Journal Review on Outlier Detection for Data
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. Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman and Hall.
17. Denning, D. E. (1987). An intrusion- detection model. IEEE Transactions onsoftware engineering, (2), 222-232.
18. Duda,R.O., &Hart,P.E. (1973) Pattern Classification and Scene Analysis.(Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. See page 218.
19. Fisher,R.A. "The use of multiple measurements in Saxonomic problems"Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Technology for Competitive Strategies Journal of Computer Applications (0975- (pp. 633-643). Springer, Singapore. 8887) Volume 65– No.11, March 2013, Mathematical Statistics" (John Wiley, NY, 1950).
20. Fox, A. J. (1972). Outliers in time series. Journal of the Royal Statistical Society. Series B (Methodological), 350-363.
21. Fu, T. C., Chung, F. L., Ng, V., &Luk, R. (2001, August). Pattern discovery from stock time series using self-organizing maps. In Workshop Notes of KDD2001 Workshop on Temporal Data Mining (pp. 26-29).