International Journal of Advanced Information in Engineering Technology (IJAIET) ISSN: 2454-6933
Vol.4, No.11, November 2017

An Outlier Detection using K-median Clustering Algorithm

[1]B.Angelin M.Sc, [2] Dr.D.Devakumari MCA, MPhil, Ph.D.,
[1]Research Scholar (M.Phil), [2]Assistant Professor in Computer Science,
[1, 2,] Pg&Research Department of Computer Science,
[1, 2] Government Arts College (Autonomous),
Coimbatore-18.
Mail: angelindia93@gmail.com , ramdevshri@gmail.com

ABSTRCT- 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 research initially reviewed more logics on clustering techniques and outlier detection techniques. Secondly, the problem is identified in K means clustering algorithm for outlier detection. Finally, the solution is proposed from k-median based clustering and compared it with the traditional weight based K-means clustering algorithm. To test the algorithm in outlier detection, the Iris dataset is considered. All the above process are carried out in MATLAB simulations. From the experimental result it is concluded that the proposed method named as K-median weight based approach has the maximum weight value in sepal frame.

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

1. INTRODUCTION

Outlier detection is an imperative branch in information pre-handling and data mining, as this stage is required in elaboration and mining of information originating from numerous application fields, for example, modern procedures, transportation, biology, open security, climatology. Exceptions are information which can be viewed as a typical because of a few causes (e.g. mistaken estimations or abnormal process conditions). Anomaly discovery procedures are utilized, for example, to limit the effect of special cases in the last model to make, or as a preliminary pre-taking care of stage before the information cruised on by a flag is explained.[1][2]

2. LITRATURE SURVEY

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. Exception identification has been looked into inside different application spaces and learning disciplines. This review gives a far reaching outline of existing anomaly identification methods by arranging them along various measurements. [6][8][9]

In this segment different ideas utilized as a part of this investigation are characterized on the premise of definitions found in course books and prior examinations. The exact investigations identified with the anticipating money related time arrangement information are audited in this part are assembled in different classifications as takes after:

- Studies identified with Outlier based Methods.
- Studies concerning bunch based approach.
- Studies identified with time arrangement estimating models for the estimation of securities exchange unpredictability and forecasting efficiency.[12][9][7]

3. PROPOSED K-MEDIAN OUTLIER ALGORITHM

K-Medians Instead of the mean, in k-medians clustering the median is calculated for each dimension

in the data vector. Finding the cluster centroid. The centroid of a cluster can be defined in different ways. For k-means clustering, the centroid of a cluster is defined as the mean over all items in a cluster for each dimension separately. For robustness against outliers, in k-medians clustering the median is used instead of the mean. In k-medoids clustering, the cluster centroid is the item with the smallest sum of distances to the other items in the cluster.[3][10][13]

In this research, a modified clustering-based technique is proposed to identify the outliers and simultaneously produce data clustering. Proposed outlier detection process at the same time is effective for extracting clusters and very efficient in finding outliers.[4]

3.1 Proposed K-Median Outlier Algorithmic steps

Input:

Iris Dataset $\{i_1, i_2, i_3, \ldots..,i_n\}$, where, n is the number of images Cluster centre C= $\{c_1,c_2,.....c_k)$, where ci is the cluster centre and k is the number of cluster centres.

Output:A set of K-clusters without outliers.

Step 1: Select k observations from data set using Divide-and-Conquer method.

Step 2: Calculate distance with each data instances using List Data structure.

Step3: Assign each instance to the cluster with the nearest seed.

Step 4: Find the cluster matrix with the utilization of K-Median outlieralgorithm.

Step 5: Separate the input data matrix and initialize the cluster labels

Step 6: Reducethe chance of numerical problems and calculate the euclidean distance matrix to remove the empty clusters.

Step 7: Process all the outlier withmedian and its standard deviation.

Step 8:Calculate the weight based centre approach

Step 9: Calculate the Mean and median of each cluster centre

Step 10: Calculate the maximum and minimum value of each cluster K

Step 11: Compare each data item in a cluster with the threshold value

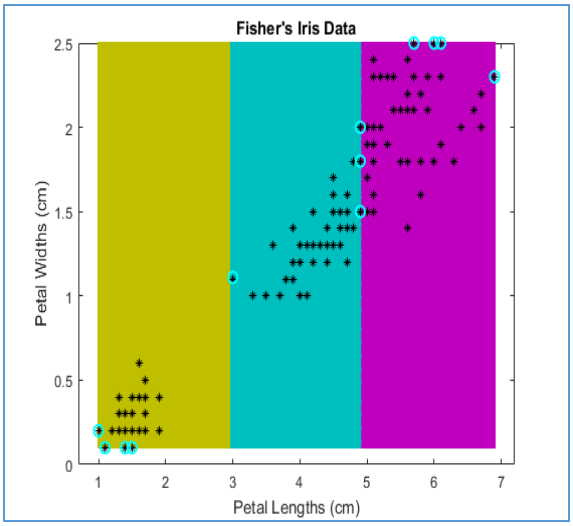
Step 12: If the data value identified is very minimum then the given data item in cluster k is the outlier, where x=1,2,…,n.

Step 13: Remove the outlier data items from the cluster.

Step 14: Repeat step 7-13 for each resultant cluster.

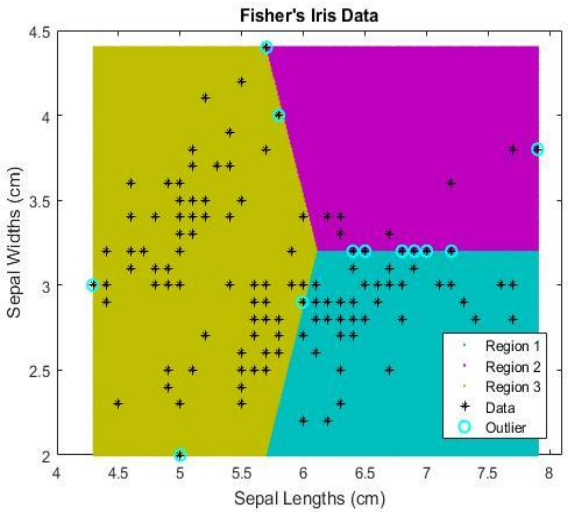
The proposed K-Median Outlier algorithm flow is represented in the recognizes the nearby exceptions amid the clustering, and thus expels their impact on the group centroids.

The calculation finds the groups with non-round shapes by combining at least two bunches, which are in reality having a place with one common group of non-spherical shape. The calculation considers the densities of the groups while recognizing the nearby anomalies furthermore, blending bunches.[17][19] From there on, the calculation bolsters the revelation of bunches of various densities.



The points of each band are grouped into a separate cluster which is clearly visible from the all the points of dissimilar clusters are shown by altered colours. The proposed method also located the outliers with a small number of points with different colours. The significance of all the above experiments can't be overlooked. Overall, all the optimizers have successfully brought about an appreciable improvement in reaching better centroids after a few iterations. As the sizes of data sets continues to grow

Rapidly these days with easy availability of relevant data for analysis.[5][9]



CONCLUSION

In this work, a data mining based hybrid outlier detection system has been designed for distinguishing normal events. The major contributions of this research work are the proposal of a hybrid architectural framework for effective outlier detection, the detection techniques use clustering algorithms to enhance the performance of the outlier detection system. The performance of the proposed hybrid outlier detection system has been evaluated in terms of weight based cluster. The Iris Dataset has been used to test both methods namely, Enhanced K-means Clustering Algorithm using weight based center approach and Proposed K-Median based Clustering algorithm using weight based centre approach respectively.

In the present study, the main focus is application of iris Sets in data mining. Hence, the proposed algorithms are compared only with the traditional hybrid K-means weigh based clustering approaches contributed in the respective areas.[13][20] It is interesting if these methods are compared with other existing data mining methods. In order to conduct various experiments, in this study, only limited size datasets are used. Because of the explosive growth of available information, a series of experiments and investigations are necessary to establish the potential utility of the proposed methods in large datasets.

REFERENCES

[1]. Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S., & Simon, I. (2003). Continuous representations of time-series gene expression data. Journal of Computational Biology, 10(3-4), 341-356.

[2]. Barnett, V., & Lewis, T. (1994). Outliers in statistical data (Vol. 3, No. 1). New York: Wiley.

[3]. Bidari, P. S., Manshaei, R., Lohrasebi, T., Feizi, A., Malboobi, M. A., &Alirezaie, J. (2008, October). Time series gene expression data clustering and pattern extraction in arabidopsis thaliana phosphatase-encoding genes. In *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on* (pp. 1-6). IEEE.

[4]. Chatterjee, S., &Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. Statistical Science, 379-393.

[5]. Chernick, M. R., Downing, D. J., & Pike, D. H. (1982). Detecting outliers in time series data. Journal of the American Statistical Association, 77(380), 743-747.

[6]. Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman and Hall.

[7]. Denning, D. E. (1987). An intrusion-detection model. IEEE Transactions on software engineering, (2), 222-232.

[8]. 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.

[9]. Fisher,R.A. "The use of multiple measurements in tSaxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).

[10]. Fox, A. J. (1972). Outliers in time series. Journal of the Royal Statistical Society. Series B (Methodological), 350-363.

[11]. 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).

[12]. Georgiadis, D., Kontaki, M., Gounaris, A., Papadopoulos, A. N., Tsichlas, K., &Manolopoulos, Y. (2013, June). Continuous outlier detection in data streams: an extensible framework and state-of-the-art algorithms.

In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 1061-1064). ACM.

[13]. Guyon, I., &Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[14]. Guyon, I., Matic, N., &Vapnik, V. (1996). Discovering Informative Patterns and Data Cleaning.

[15]. Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., &Tatham, R. L. (1998). Multivariate data analysis (Vol. 5, No. 3, pp. 207-219). Upper Saddle River, NJ: Prentice hall.

[16]. Hautamaki, V., Nykanen, P., &Franti, P. (2008, December). Time-series clustering by approximate prototypes. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on (pp. 1-4). IEEE.

[17]. Hawkins DM. Identification of outliers?, New York, NY: Chapman and Hall; 1980.

[18]. Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

[19]. Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*(2), 85-126.

[20]. Jixue, D. (2009, May). Data mining of time series based on wave cluster. In Information Technology and Applications, 2009. IFITA'09. International Forum on (Vol. 1, pp. 697-699). IEEE.