Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Comparison between Two Approach Based on Threshold and Entropy Based Approach

Navneet Kaur[1]
Computer Science, SGGSWU,
Fatehgarh Sahib, India

Prof. Kamaljit Kaur[2]
Computer Science, SGGSWU,
Fatehgarh Sahib, India

Abstract— Data mining refers to extracting or mining knowledge from large amounts of data. Organizing data into valid groupings is one of the most basic ways of understanding and learning. Cluster analysis is important for analysing the number of clusters of natural data in several domains. Outlier detection is a fundamental part of data mining. A key challenge with outlier detection is that it is not a well-formulated problem like clustering. This paper discussion on two different techniques and then comparison by analysing their different accuracy, mean squared error, time complexity. The techniques were: First is threshold based approach and second is entropy based approach. In order to find the best clustering algorithm for outlier detection several performance measures are used. The experimental results show that the outlier detection accuracy is very good in threshold approach clustering algorithm compared to the existing algorithms.

Keywords— clustering , outlier, entropy , threshold, mean squared error.

I. Introduction

Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. It is a useful technique for the discovery of data distribution and patterns in the original data. It is a method of unsupervised learning and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. It is an important technique used for outlier analysis.

One of the fundamental difficulties in data mining is outlier detection. Clustering is significant tool for outlier analysis. Outliers are normal elements when specified as input, but will load in inefficient outputs when processed with them. An outlier is data object that is different from the remaining dataset. According to Hawkins (1980), "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". In many applications outliers are more interesting than normal cases for example network intrusion detection, fault diagnosis in machines, credit card fraud detection, marketing, detecting outlying cases in wireless sensor network data. There are numerous deferent formulations of the outlier detection problem which have been explored in diverse disciplines such as statistics, machine learning, data mining, information theory, spectral decomposition.

Outlier detection is an important research problem that aims to find objects that are considerably dissimilar, exceptional and inconsistent in the database. There are large number of techniques are available to perform this task, and often selection of the most suitable technique poses a big challenge to the practitioner. Some of the outlier detection techniques are:

- Distance based outlier detection.
- Clustering based outlier detection.
- Density based outlier detection.
- Depth based outlier detection.

Outlier detection has direct applications in a wide variety of domains such as mining for anomalies to detect network intrusions, fraud detection in mobile phone industry and recently for detecting terrorism related activities [ ].The rest of the paper is organized as follows. Section .2 provides discussion on the previous works related to the topic. Section 3 describes briefly two approaches: threshold approach and entropy approach. Section .4 describes the experimental results. Conclusion and future works are given in Section 5.

II. Literature Review

- This paper a new method of clustering was proposed based on Multivariate outlier detection. Though several clustering procedures available in the literature, the proposed technique gives a unique idea to cluster the sample observations in a survey study based on the multivariate outliers. The feature of the proposed clustering technique was elaborately discussed and the authors also highlighted the application of the technique in a survey research. Based on the results derived, the proposed technique gives more insights to the researcher to cluster the sample observation at 5% and 1% significance level. The authors enlighten an idea for further research by

conducting simulation experiments for testing relationship between the significance level and the number of outlier clusters extracted. Moreover more rigorous experiments may conduct to identify the Multivariate outliers' inside the outlier clusters. [8]

- Sridhar presents the performance of K-Mean clustering algorithm, depending upon various mean values input methods. The mean values are the centroid of the specified number of cluster groups. The clustering algorithm consists to two stages with first stage forming the clusters-calculating centroid and the second stage determining the outliers. There are three methods for assigning the mean values in K-Mean clustering algorithm. [11]
  a. Taking the first 'k' values as centroid.
  b. Random centroid generation.
  c. User specified centroid.
- R.R. Rathod compares the results obtained with preprocessing by min-max normalization and without preprocessing the data set. The basic algorithm detects outliers in two steps. In first step clusters of original data are formed by using K-Mean clustering method. In the second step, it extracts the data elements from each cluster those are far away from their centers. These data elements are processed to determine their outlierness by using stastical measures. The experiments on different datasets confirm that preprocessing the data set refines the result. [12]
- Barkha.H.Desai and Nisha Shah compared the K-Mean types clustering algorithms like K-Mean, Weighted-K-Mean and Group-Weighted-K-Mean. A major problem of using the K-Mean type algorithm cannot select variables automatically because they treat all variables equally in the clustering process. Then we are faced the problems of outliers. To overcome this problem several algorithm has been proposed one of this algorithm is WK-Mean. WK-Mean add a new step to the basic K-Mean algorithm to update the variable weights based on the current partition of data. The variable weights produced by WK-Mean measure the importance of variables in clustering. The small weights reduce or eliminate the effect of noisy variables. WK-Mean, through it solves the problem of variable selection but it doesn't support for large dataset so another algorithm has been proposed by, named GWK-Mean. GWK-Mean is better than K-Mean and WK-Mean, because it reduced the effect of noise variables and improves the accuracy of clustering process. [13]
- A .Mira and S.Saharia developed a robust supervised outlier detection algorithm using hybrid approach (RODHA) which incorporates both the concept of distance and density along with entropy measure while determining an outlier. They have provided an empirical study of different existing outlier detection algorithms and established the effectiveness of the proposed RODHA in comparison to other outlier detection algorithms. The algorithm m is tested on synthetic and real-life datasets from UCI ML Repository. The detection performance of the algorithm is competing excellent than other existing algorithms. In the present work, the datasets on which the proposed technique is tested are of integer or real type. So, our work is undergoing to extend the algorithm on mixed data set. [14]

- S.Vijayarni and S.Nithya focused on outlier detection in health data sets such as Pima Indians Diabetes data set and Breast Cancer Wisconsin data set using partitioning clustering algorithms. The algorithms used in this research work are PAM, CLARA AND CLARANS and a new clustering algorithm ECLARANS is proposed for detecting outliers. In order to find the best clustering algorithm for outlier detection several performance measures are used. The experimental result shows that their algorithm ECLARANS improves the accuracy of detection and CLARANS reduces the time complexity when compared with other algorithms. Further work also lies in this application. We will use this detection of outliers for our future work and plan to reduce the time complexity of the proposed algorithm. [15]

- This paper is proposed a new efficient method for outlier detection. The proposed method is based on fuzzy clustering techniques. The c-means algorithm is first performed, and then small clusters are determined and considered as outlier clusters. Other outliers are then determined based on computing differences between objectives function values when points are temporarily removed from the data set. If a noticeable change occurred on the objective function values, the points are considered outliers. Test results were performed on different well-known data sets in the data mining literature. The results showed that the proposed method gave good results. The test results show that the proposed approach gave effective results when applied to different data sets. However, their proposed method is very time consuming. This is because the FCM algorithm has to run n times, where n is the number of points in a set. This will be our focus in the future work. [16]
- Jae–Gill Lee and Jiawei Han proposed a novel partition-and-detect framework for trajectory outlier detection, which partitions a trajectory into a set of line segments, and then, detects outlying line segments for trajectory outliers. Based on this partition-and-detect framework, we develop a trajectory outlier detection algorithm TRAOD .Our algorithm consists of two phases partitioning and detection. For the first phase, we proposed a two-level trajectory partitioning strategy that ensures both high quality and high efficiency. For the second phase, we present a hybrid of the distance-based and density-based approaches. The visual inspection results show that TRAOD effectively detects trajectory outliers with outlying t-partitions. They consider here only the spatial information. In future they are used some other datasets .[17]

- Deevi Radha Rani at al. presents a new K-Mean type clustering algorithm that can calculate weights to the variables. This method is efficient for dynamic data streams in order to overcome the global optimum problems. The variable weights produced by the algorithm measures the importance of variable in clustering and can be used in variable selection in which the data items with similar properties are grouped into clusters, the new approach of applying this weighted K-Mean on dynamic data streams is carried out in order to have efficient outlier detection within the user specific threshold value. [18]

- Yogita and Durga proposed a clustering based un-supervised outlier detection scheme for streaming data. In the proposed approach both densities based and partitioning clustering are combined to take advantage of both densities based and distance based outlier detection. The Partitioning clustering is also used to assign weights to attributes. Weighted attributes are helpful to reduce or remove the effect of noisy attributes. The proposed method gives higher outlier detection rate and lower false alarm rate than CORM. The proposed method has performed much better than CORM with increasing percentages of outliers. In future, this method extends for categorical and mixed data types. [19]

- This paper presents a very fast greedy algorithm for mining outliers under the same optimization model. The results on real datasets and large synthetic datasets that: [20]
  (1) Algorithm has comparable performance with respect to those states of art outlier
  (2) Algorithm can be an order of magnitude faster than LSA algorithm.
  In future work, they will study how to automatically determine an optimal number of outliers without human intervention. Furthermore, more efficient outlier mining algorithm under the optimization will be further addressed.

- S. D. Pachgade proposed Method for outlier detection uses hybrid approach. Purpose of approach is first to apply clustering algorithm that is K-Mean which partition the dataset into number of clusters and then find outliers from the each resulting clusters using distance based method. The principle of outliers finding depend on the threshold. Threshold is set by user. The main objective of the second stage is a finding out the objects, which are far away from their cluster centroid. Approach is only deals with numerical data, so future work requires modifications that can make applicable for textual mining also. The approach needs to be implemented on more complex datasets. Future work requires approach applicable for varying datasets. [21]

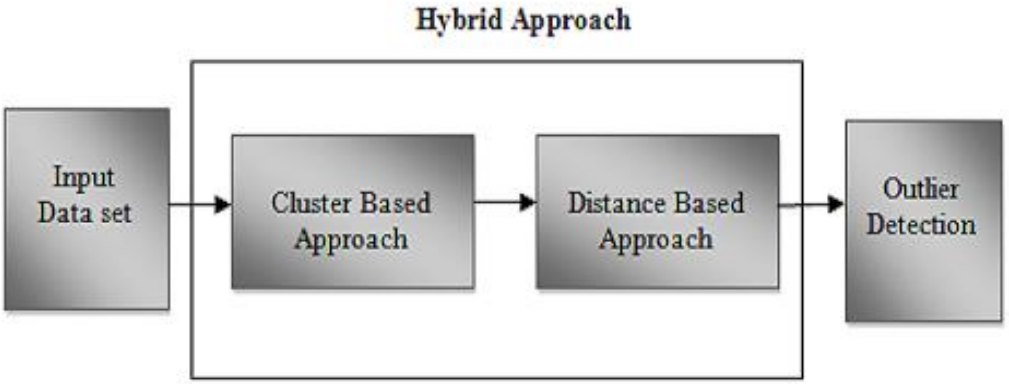

Figure: 2.1 hybrid approach for outlier detection

- This paper the researchers have introduced an Enhanced K-Mean with Greedy algorithm for Outlier Detection which is an improvement over the K-Mean algorithm for outlier detection. During the initialization phase of the greedy algorithm, each records is represented as non-outliers and hash tables for attributes are also built and updated .In the greedy procedure, the dataset is scanned for k times to discover exact k outliers that is, one outlier is found and removed in each pass .In every scan over dataset, read each record t that is represented as non-outliers, its label is changed to outlier and the changed entropy value is calculated. A record that accomplishes maximal entropy value impact is chosen as outliers in current scan and accumulated to the set of outliers. This hybrid approach is used to automatically detect and remove outliers and thus help in increasing the clustering accuracy. The proposed method has less Means Squared Error and execution time. Next time increase this method with mixed dataset.[22]

- This paper proposed a combined approach based on Minimum Spanning Tree based clustering and Density-based clustering for noise-free high density best number of clusters. Minimum Spanning Tree clustering algorithm is capable of detecting clusters with irregular boundaries. Their MSTDBCNFHDC clustering algorithm detects outliers from cluster and it uses a new cluster validation criterion based on the geometric property of partitioned regions/clusters to produce best number of "true" clusters with center for each of them.

The inter-cluster distances between centers of clusters/regions are used to find best number of noise-free clusters. All of these look nice from theoretical point of view. However from practical point of view, there is still some room for improvement for running time of the clustering algorithm. This could perhaps be accomplished by using some appropriate data structure. In the future they will explore and test our proposed clustering algorithm in various domains.[23]

- Neeraj Bansal compared the result of different Clustering techniques in terms of time complexity and proposed a new solution by adding fuzziness to already existing Clustering techniques. There are different algorithms exist for detecting outlier. As we have seen, ECLARANS is the best technique amongst them. It takes lesser amount of time to detect the outlier. As future lies, further advancement is going on in outlier detection methods. More work is being done on the basis of fuzzy approach in clustering techniques.

- H.S.Behera suggested a clustering based outlier detection algorithm for effective data mining which uses K-Mean clustering algorithm to cluster the data sets and outlier finding technique to find out outlier on the basis of density based and distance based outlier finding technique. From the experimental analysis of data sets both of lower dimension and higher dimension as in the case of Bupa dataset we can see that K-Mean can be used for outlier analysis. The outlier detection of the proposed algorithm in the Bupa data set has improved over the algorithm used by Moh"d Belal Al- Zoubi. K-Mean has sensitivity over outlier data but can be still used with OFT for the detection of outlier data. Many of the clustering techniques are being developed that are not affected by outliers and can be easily implemented to find outlier. K-Mean can be improved over noisy data that can be used to find outlier detection. [25]

- This paper proposed a new K-Mean type algorithm that can simultaneously weight variable groups and individual variable in clustering high dimensional data. Given high dimensional data sets with variable groups, GWK-Mean can weight both variable groups and individual variables simultaneously in the clustering process. With variable group weighting, important variable groups can be identified and effect of noise variables can be reduced. Therefore, better clustering results can be obtained from very high dimensional data with noise. We compared GWK-Mean to four clustering algorithms and the result has shown that the GWK-Mean algorithm outperformed other four clustering algorithms in clustering accuracy. Experiment results also show that GWK-Mean was less sensitive to the initial settings than other four algorithms. As such, it is a new tool for clustering very high dimensional data.[26]

III. Describes Briefly Two Approaches: Threshold Approach and Entropy Approach

Threshold Approach:

Threshold has a current *value*, the value ranges from 0 - 100%. The proposed method is using the advantages of existing outlier detection algorithms that are group weighted K-Mean and greedy algorithms. Purpose of new hybrid approach is first to apply the clustering algorithm that is GWK-Mean which partition the dataset into number of groups and second using greedy algorithm for detect outliers. The principal of outliers finding depend on the threshold. Threshold is set by user. The problem of this research work is to find out the outliers using a new hybrid approach on mixed type datasets and to verify the performance of existing clustering algorithms.

Entropy approach:

Entropy is a measure of the uncertainty in a random variable. The Greedy algorithm takes as an input the desired number of outliers (k). All points in the set are initially non-outliers. We formulate the set of outliers by conducting k scans over the dataset to iteratively determine the top k outliers. During each scan, we remove every non-outlier individually from the dataset and recalculate the total entropy of the system. The data point that has the maximum impact on the total entropy is the point that lowers the entropy the most when removed.

IV. Experimental Result

MATLAB (MATrix LABoratory) is a numerical computing environment and fourth-generation programming language. Matlab is a programming environment as well as a high level, interpreted, dynamically typed language, supporting functional, object oriented, and event driven paradigms. It is well suited for numerical computation, particularly computations involving matrix operations and linear algebra. Matlab has excellent support for data visualization and its concise and expressive syntax, as well as the plethora of predefined functions, results in a powerful environment excellent for rapid prototyping with minimal overhead. We use MATLAB tools for implementing our algorithms. We conducted all experiments on a Windows 7 Home Premium with Intel® Core™ i3 CPU M380 @ 2.53 GHz with 6.00 GB RAM.Experiments were conducted in Matlab 7.8.0 (R2009a) on various data sets.

4.1 Comparative Analysis of Threshold Methods with Entropy method

Entropy is a measure of the uncertainty in a random variable. The hybrid approach used to approach of threshold .The performance of the hybrid approach is evaluated against entropy method. Table shows the result of according to comparative analysis of threshold method with entropy method.

METHOD	ACCURACY	MSE	EXECUTION TIME
THRESHOLD	98%	0.081	1.0539
ENTROPY	92%	1.147	1.1157

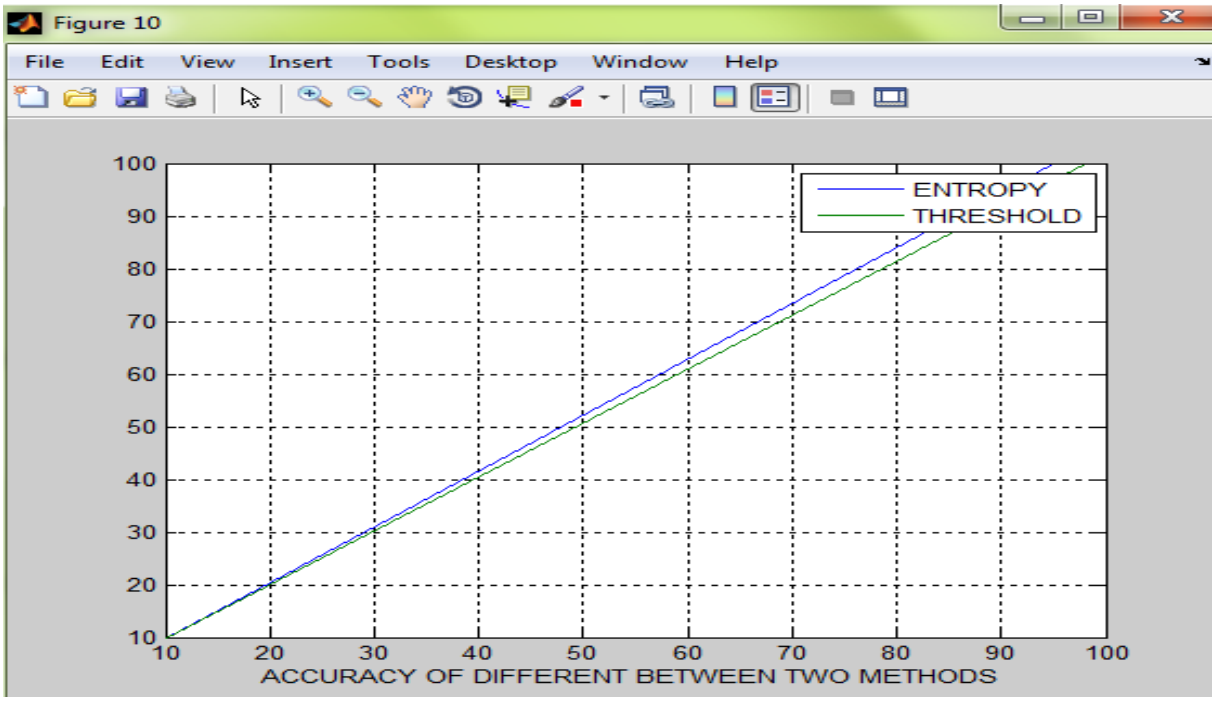

Figure: 4.1 Graphical Comparative Analysis of Threshold Methods with Entropy method In Iris Dataset

V. Conclusion

The most of the research has focused on numerical datasets, and not specifically on the detection of outliers in categorical data. The basic problem is to outlier detection using new hybrid approach on mixed type datasets. Purpose of threshold approach is first to apply the clustering algorithm that is GWK-Mean which partition the dataset into number of groups and second using greedy algorithm for detect outliers. The principal of outliers finding depend on the threshold. The threshold is set by user. The hybrid approach has two techniques that are combined to improve efficiently find outlier from the dataset .The hybrid approach introduced outlier detection algorithm is compared with the existing algorithms. The conclusion from this comparison is that the simple outlier detection algorithm that new Hybrid Approach is more existing strategies, and accurate in discovering outliers.

Future Scope

Although the proposed algorithm solve the problem of efficiency and accuracy to some extent, but more needs to be done to solve the problem of sub groups. How to overcome this problem is the topic of some of our future work. Otherwise hybrid approach is only deals with numerical or categorical data, so future work requires modifications that can make applicable for time series data or image mining also.

References

[1] Han, J. and Kamber, M., "Data Mining: Concepts and Techniques", Morgan Kaufman Publishers, 2006.
[2] Neelama Padhy and Pragnyaban Mishra, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.
[3] S.P.Deshpande , "Data Mining System and Application: A Review**",** International Journal of Distributed and Parallel System, Vol.1, September 2010.
[4] Prabdeep and Shubha Singh, "A Survey Of Clustering Techniques", International Journal of Computer Science, Vol.7, October 2010.
[5] Anshul Parash , " Survey of Different Partition Clustering Algorithm and Competitive Studies", International Journal of Advanced Computer Science, Vol.3, June 2012.
[6] Victoria J .Hodge and Jim Austin, "A survey of outlier detection methodologies".

[7] Irad Ben-Gal, " Outlier Detection", Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, ISBN 0-387-24435-2, Jan- 2005.
[8] Varun Chandola and Banerjee and Kumar," Outlier Detection: A Survey".
[9] Karanjit Singh and Shuchit," Outlier Detection: Applications", IJCT, Vol.9, Jan 2012.
[10] G.S.David," A New Procedure of Clustering Based on Multivariate Outlier Detection",Journal of Data Science 11(2013), 69-84.
[11] Ammar. W. Mohemmed," Particle Swarm Optimization for Outlier Detection", VUW ECSTR10-07 2010.
[12] R. R. Rathod and Dr. B. F. Momin," Performance evaluation of Outlier Detection with Normalized Data Set", Department of Information Technology Walchand College of Enginee- ing Sangli, Maharashtra State, India.
[13] H. Desai, "Comparative Study of K-means Type Algorithms", UNIASCIT, Vol. 2, 2011.
[14] A.Mira and S.Saharia," A Robust Outlier Detection Using Hybrid Approach", Aamerican Journal of Intelligent System 2012.
[15] S.Vijayarni and S.Nithya,"An Efficient Clustering Algorithm for Outlier Detection", (IJCS) Vol.32, October 2011.
[16] Mohd - Al-Zoubi," New Outlier Detection Method Based On Fuzzy Clustering", (IJAR) Vol.4, October 2010.
[17] Jae-Gil, "Trajectory Outlier Detection: A Partition-and-Detect Framework", Department Of Computer Science, University of Illinois at Urbana-Champaign Urbana, IL 61801, USA.
[18] Deevi Radha Rani and Naya Dhulipala , "Outlier Detection For Dynamic Data Streams Using Weighted K-Means" IJEST, Vol.3, October 2011.
[19] Yogita and Durga Toshniwal," Unsupervised Outlier Detection in Streaming Data Using Weighted Clustering", World Academy of Science, Engineering and Technology 2012.
[20] He. Xu, and S.Deng,"A Fast Greedy Algorithm for Outlier Mining, "PAKDD Conference, Singapore, 2006.
[21] Ms. S. D. Pachgade and Ms. S. S. Dhande," Outlier detection Over Data Set Using Cluster Based and Distance-Based Approach", (IJARCSSE), Volume 2, Issue6, June 2012.
[22] C.Sumithiradevi and Punithavalli,"Enhanced K-Means with Greedy Algorithm For Outlier Detection", IJARCS, Vol. 3, No.3, May-June 2012.
[23] S. John Peter," Hybrid Algorithm for Noise-free High Density Clusters with Detection Of Best Number of Clusters", (IJHIT) Vol. 4, No. 2, April, 2011.
[24] Neeraj Bansal," Differentiate Clustering Approaches for Outlier Detection", (IJIRCS), Vol. 1, Issue 2, April 2013.
[25] H.S.Behera," New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining Vol. 1, Issue 2, April 2013.
[26] H. Desai, "Comparative Study of K-means Type Algorithms", UNIASCIT, Vol. 2, 2011.

