IJETST- Vol.||02||Issue||10||Pages 3265-3270||October||ISSN 2348-9480 2015

Open access Journal International Journal of Emerging Trends in Science and Technology

DOI: http://dx.doi.org/10.18535/ijetst/v2i10.05

An Influential Algorithm for Outlier Detection

Authors
Sachin Yele[1], Nidhi Sharma[2]
[1]Assistant Professor Sanghvi Institute of Management and Science Indore, India
[2]M.Tech Research Scholar Sanghvi Institute of Management and Science Indore, India

ABSTRACT

The Subject of Data Mining, which is very vast in the direction of data analyzing , data preprocessing, data extracting, knowledge data discovery, extracting hidden data and many more. Here we discuss one of the parts of the data mining. Which is clustering, in the clustering there is the major issue to eradicate outliers from the data. Here we do work for detect outliers from the data sets.

In the paper we had detailed studied about the different-different ways for detecting outliers. Likewise cluster based, distance based, density based etc. this approaches are used in the different-different methods of outlier detection (PAM, CLARA, CLARANCE, ECLARA). Here we choose the best one for further enhancement on the basis of their performance. new hybrid algorithm proposed with the classical clustering method for the improving performance on the basis of accuracy, error rate, time complexity.

Keywords: *K-Mean, Clustering, PAM, CLARA, CLARANs*

INTRODUCTION

Data Mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications.

In the data mining there is two popular approaches are mainly used supervised and unsupervised. this approaches are used to for detection of unuseful data. This approaches are involving two content classification and clustering, Where classification is a part of supervised clustering approach and another one is clustering which is belongs to the unsupervised clustering approach. Here when the number of normal attributes is more than the abnormal behavioral attributes, a clustering-based approach to outlier detection provides more positive results. In these situations, the key assumption made here is that large and dense clusters have normal data and the data which do not belong to any cluster or small clusters (low dense clusters) are considered as outliers.[2]

Cluster is a group of objects which belongs to the same class. Clustering is a part of data mining for data extraction. Mainly clustering is a process of making a group of abstract objects into classes of similar objects. A cluster of data objects can be treated as one group. While doing cluster analysis, we first perform partition the set of data into groups which is based on data similarity and then assign the labels to the group.

The main benefit of clustering over classification is that, it is adaptable to changes and helps single out useful feature. As a data mining function cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Clustering is also used in outlier detection. In this duration there is outlier detection is major issue in clustering: those data or object whose different in behavior or nature are declared as a outlier and outlier detection technique is a process to find inconsistent or dissimilar data from the remaining data. an *outlier* is a data point that significantly differs from the other data points in a sample. Often, outliers in a data set can alert statisticians to experimental abnormalities or errors in the measurements taken, which may cause them to omit the outliers from the data set. If they *do* omit outliers from their data set, significant changes in the conclusions drawn from the study may result. Which are mainly used in various ways for such as detection of credit card fraud, network intrusion detection, data analysis, pattern recognition. The problem arise how to find outlier from the large data sets which are actually difficult task for the data analyzer.

Before deciding whether or not to omit outlying values from a given data set, first, obviously, we must identify the data set's potential outliers. Generally speaking, outliers are data points that differ greatly from the trend expressed by the other values in the data set - in other words, they lie outside the other values. It's usually easy to detect this on data tables or (especially) on graphs. If the data set is expressed visually on the graph, outlying points will be "far away" from the other values. If, for instance, the majority of the points in a data set form a straight line, outlying values will not be able to be reasonably construed to conform to the line.

Let's consider a data set that represents the temperatures of 12 different objects in a room. If 11 of the objects have temperatures within a few degrees of 70 degrees Fahrenheit (21 degrees Celsius), but the twelfth object, an oven, has a temperature of 300 degrees Fahrenheit (150 degrees Celsius), a cursory examination can tell you that the oven is a likely outlier.

TYPES OF CLUSTERING

Cluster-based methods either belong to semi-supervised or supervised categories. In semi-supervised techniques, the normal data is clustered to create modes of normal behavior and instances which are not close to any clusters are identified as outliers. In unsupervised techniques, a post-processing step is included after clustering to determine the size of the clusters. The distance from the clusters is then calculated, using which the outliers are detected. Furthermore, depending on the method adopted to define clusters, the techniques can be further grouped as partitioned clustering, hierarchical clustering, density-based clustering and grid based clustering [1].

Distance Based Outlier Detection: Distance-based method was originally proposed by Knorr and Ng. [13] Distance based approach is used to outlier detection according to given threshold value. This is given by user. This technique is used to calculate the maximum distance value for each cluster if the maximum distance of cluster is greater than threshold value then the cluster will we declared as outlier [3].

Clustering Based Outlier Detection: Clustering based approach is used when the number of normal attribute is more than abnormal behavior attribut e .this technique provides more positive result. This approach is used in those situation when large and dense cluster have normal data and data which does not belong to any cluster or small cluster (low dense cluster) are consider as outlier. [4]

Density Based Outlier Detection: Density Based approach method involve the investigation not only local density but also studied local density of its nearest neighbors [5]. This method identify the outlier by checking the main features or characteristics of object in database the object that are deviate from these feature are consider as outlier. [8]

Distribution Based Outlier Detection: It Develop statistical models from the given data and then apply a statistical test to determine if an object belongs to this model or not. Objects which have low probability to belong to the statistical model are declared as outliers. However, Distribution-based approaches cannot be applied in multidimensional scenarios because they are univariate in nature. Most distribution models typically apply directly to the future space and are univariate i.e. having very few degrees of freedom. Thus, they are unsuitable even for moderately high-dimensional data sets . [12]

In the paper we had discussed and detailed study about the most popular approaches of clustering. Where we observe the problem regarding the outliers (unwanted data). This is creating the data inconsistency. here to detect outliers is a very difficult task from the large data sets. and no. of large methods are used somewhere this approaches are provide inappropriate result. When we investigated on the topic of outliers then there is a problem of time complexity, inconsistency, data accuracy and data efficiency occur.

BACKGROUND

A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. the main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns. In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations. Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location. Clustering also helps in classifying documents on the web for information discovery. Clustering is also used in outlier detection applications such as detection of credit card fraud. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

The work on Clustering in the data mining for detect outliers used various approaches like distance based outlier detection, clustering based outlier detection, density based outlier detection, distribution based outlier detection. The most common method is used for outlier detection is partitions based clustering algorithm. Partition based clustering create partition of data. It consist heuristic function for calculation of object values. Mainly heuristic method is used in two ways K-Mean and K-Medoids.

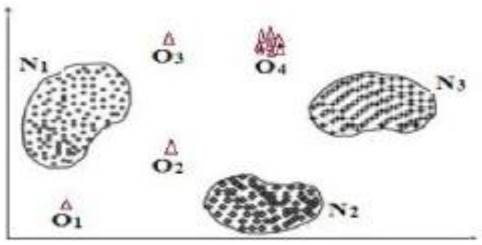


Fig. Outlier's detection in two dimensional dataset [13]

Illustrating the outlier in two dimensional dataset. There are N1, N2, and N3 are the three normal regions. Points that are sufficiently far away from the normal region such as points O1, O2, O3 and points in O4 regions are outliers.

K-Mean: K-Mean is an iterative relocating technique. Randomly chosen (user define no.) objects as the initial cluster centers. Until no change, do Reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster Update the cluster means, i.e., calculate the mean value of the objects for each cluster.

K-Medoids or PAM (partition around medoids): each cluster is represented by one of the objects in the cluster. randomly choose k (user define no.) objects as the initial medoids Until no change, do Reassign each object to the cluster to which the nearest medoid randomly select a non-medoid object, compute the total cost of swapping medoids with later selected mediod If total cost will be minimum then swap previous mediod with the new randomly chosen mediod to form the new set of k medoids. It is very robust when compare with k-mean in the presence of noise or outlier [4]. This algorithm performs well with the small data sets but not good for the large data sets. Procedure of PAM is:

1. Initialize: randomly select *k* of the *n* data points as the medoids
2. Associate each data point to the closest medoid.
3. For each medoid m calculate distance For each non-medoid data point o. Swap m and o and compute the total cost of the configuration
4. Select the configuration with the lowest cost.
5. Repeat steps 2 to 4 until there is no change in the medoid.

CLARA (Clustering Large Application): CLARA is one of the most popular clustering algorithm. It works on randomly selected subset of the original data and produces near accurate results at a faster rate. also based on partitioning method. but its deals with samples for averaging of simulation result to reduce erroe. Although the effectiveness or efficiency of CLARA is totally depend on the method of the sampling and it's size.

Procedure of CLARA:

1. Input the data set D
2. Draw sample S randomly from D
3. Call PAM to get mediod
4. Classify entire data set to cost1…cost k.
5. Calculate the average dissimilarity from obtained cluster.[3]

CLARANS (Clustering Large Application Based Upon Randomized Search): To improve the quality and scalability of CLARA, another clustering algorithm called Clustering Large Applications based upon Randomized Search (CLARANS) was proposed in [14]. When searching for a better centre, CLARANS tries to find a better solution by randomly choosing object from the other (n-k) objects. If no better solution is found after a certain number of attempts, the local optimal is assumed to be reached. CLARANS has been experimentally shown to be more efficient than both PAM and CLARA.

1. Randomly select a node (ie., k medoids)
2. Consider a set of randomly chosen neighbor nodes as candidate of new medoids
3. Moves to the neighbor node if the neighbor is a better choice for medoids. Otherwise, a local optima is discovered.

The entire process is repeated multiple times to find better local optima

RECENT STUDY PARTITION BASED CLUSTERING SCHEME

Boomija M,D. MCA, M.PHIL, et.al.[Dec-20008], the idea define in the journal of computer , vol-1, no. 4. Here had studied about clustering algorithm, which depends on type of data available on the particular purpose and application small to medium sized of database, these partition based algorithm performed well with the different shaped of cluster.

Vijayarani.S,Asssitant Professor, et.al.[Oct-2011], presented in the journal of computer application(0975-8887), Vol 32, no. 7. They discussed over the different clustering method for outliers detection and they believed for arbitrary node instead of randomly selecting node operations, this article joint to the existing one and performed for the improvement in terms of accuracy etc.

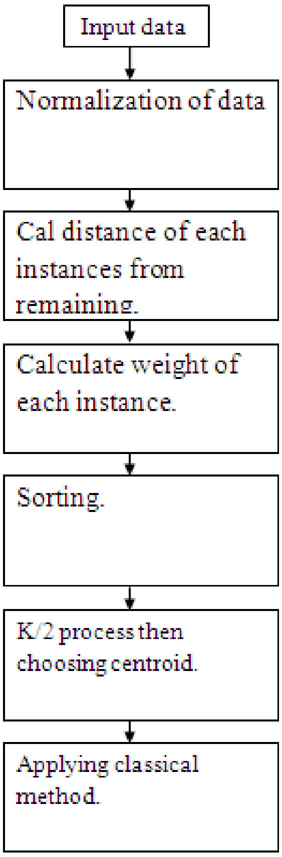
Aggrwal Shruti,Assitant Professor, et.al.[jun-2013] here defines there idea in international journal for advanced research in engineering and technology. Before proposing the their idea they studied over the old classical methods and then they perform with their the algorithm and done the comparative study along with this operation.

Wang Juntao, Su Xiaolong et.al.[yr.2011], presented in the 978-1-61284-486-2/11 © 2011 IEEE, they took basic idea of the k-mean clustering algorithm and from that they enhance the performance of the clustering for eliminating noise(i.e. outliers), in that section verified effectiveness and feasibility of proposed algorithm. Later they showed experimental results on the basis of their performance where they exclude the interference of the outliers.

Shivaram.K. et.al.[2013], defines logical and experimental idea in global journal of advanced engineering technology, vol-2, issue 1. Where used unsupervised method to detect the outliers between the medoids i.e. there is no desired results are present before for checking the accuracy of the output data. They have studied of all classical methods like PAM/CLARA/CLARANS etc. measuring absolute distance between medoids while using any one of them methods and shows the experimental comparative results of different-different datasets.

PROPOSED METHOD

In the paper we are improving the efficiency of data, verify the effectiveness and feasibility of the data by the proposed algorithm. Concentrate to manage two most important factor which is time complexity and data inconsistency. The perform the normalization of the variable for the data scaling. the calculates the distance and weights between the instances. Then here feed the concepts of sorting for reducing time. k/2 process are work on the sorted instance list. From here choose the medoid (cluster center) as per user choice. There after applying the old classical method of clarans. Algorithm for single iteration to create the find cluster and to compare with old approach. This clustering algorithm based on k-means, clarans algorithms by this we eradicate the interference of outliers firstly, result of this hybrid algorithm is more efficient and higher accurate. The formal architecture of proposal shown below.


Input data
Normalization of data
Cal distance of each instances from remaining.
Calculate weight of each instance.
Sorting.
K/2 process then choosing centroid.
Applying classical method.

CONCLUSION

In this swction our work is discussed about the different clustering techniques for outlier detection we proposing new methodology for eradicating outliers. These hybrid algorithms use the concepts of k-mean with the clarans. And also with them the Sorting concepts is used with the k/2 process which is highlight of the technique. The experimental result will show that the algorithms of eclara improves the accuracy of the detection on the other hand clarans reduces the time complexity which is going to be compares with the all other algorithms. Further work also lies in this application. We will use this detection of outliers for future work and plan to reduce the time complexity, error rate, memory and increase the accuracy rate and performance factor by the proposing algorithm.

REFERENCE

1. Osama, A., Erna, V., Eric, P. and Marc, R. (2004) Exploring anthropometric data through cluster analysis, Society of Automotive Engineers, New York, NY, ETATS-UNIS (1927-200), Vol. 113, No. 1,Pp. 241-244.
2. Department of Information and Technology, SRM University, Kattankulathur, Chennai, T.N., *INDIA*, An Effective Algorithm For Outlier Detetion Vol2-Issue1-2013.
3. Deepak Soni, Naveen Jha, Deepak Sinwar," Discovery of Outlier from Database using different Clustering Algorithms", Indian J. Edu. Inf. Manage., Volume 1, pp 388-391, September 2012.
4. P. Murugavel, Dr. M. Punithavalli," Improved Hybrid Clustering and Distance-based Technique for Outlier emoval", International Journal on Computer Science and Engineering, Volume 3, pp 333-339, 1 January 2011.
5. S.Vijayarani, S.Nithya," Sensitive Outlier Protection in Privacy Preserving Data Mining", International Journal of Computer Applications, Volume 33, pp 19-27, November 2011.
6. Aggrwal Shurti, Kaur Prabhdip " Survey of Partition based Clustering Algorithm Used For Outlier Detection", International Journal for Advance Research In Engineering And Technology, Volume 1, Issue V,June 2013.
7. Boomija M.D., "Comparison Of Partition Based Clustering Algorithms", Journal Of Computer Applications, Volume 1, No. 4, Oct-Dec 2008.
8. Periklis Andritsos," Data Clustering Techniques", pp 1-34, March 11, 2002.
9. Wang Juntao, Su Xiaolong, School Of Coputer Science And Technology, "An Improved K-Means Clustering Algorithm", 978-61284-486- 2/11©2011 IEEE.
10. S.Vijayarani, S.Nithya, "An Efficient Clustering Algorithm For Outlier Detection". International Journal Of Computer Applications (90975-8887),Volume 32, No. 7, October 2011.
11. Arthur Zimek, Matthew Gaudet, Ricardo J.G.B. Campello, Jorg Sander, KDD'13, August 11-14, 2013, Chicago, Illinois, USA, ©2013 ACM 978-1-4503-2174-7/13/08
12. Silvia Cateni, Valentina Colla ,Marco Vannucci Scuola Superiore Sant Anna, Pisa, " Outlier Detection Methods for Industrial Applications", ISBN 78-953-7619-16-9, pp. 472, October 2008
13. A. Mira, D.K. Bhattacharyya, S. Saharia," RODHA: Robust Outlier Detection using Hybrid Approach", American Journal of Intelligent Systems, volume 2, pp 129-140, 2012
14. Ng, R. and Han, J. (1994) Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Conf. Very Large Databases, Pp. 144–155.