International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

Outlier Detection Method for Data Set Based on Clustering and EDA Technique

Pranali K. Bhowate
Department of computer sci& engg
P.R.Patil college of engg Amravati

Prof. Vijay B. Gadicha
Department of computer sci& engg
P.R.Patil college of engg Amravati

***Abstract* - The proposed system is clustering based and distance based method to capture the outliers. Outlier is a point of data that does not belong to group of data. The proposed system apply Bisecting K-means algorithm for clustering a data and Euclidean distance based algorithm for find the outliers within the data set. This work is to identify the points which are not outliers using clustering and distance function and prune out those points. Next calculate a distance based measure for all remaining points, which is used as a parameter to identify a point to be outlier or not. And according to an outliers scores, declare the top N points are outliers with the highest score. It is work as the functions of sampling, scale, and flagging.**

Keywords—outlier, distance based approach

I. INTRODUCTION

Outlier is a point of data that does not belongs to group of data also it is a data point that does not conform to the normal points characterizing the data set [1].Outlier detection is an integral part of data mining and has attracted much attention recently [2].It is very vigorous problem to Find anomalous points among the data points is the basic idea to find out an outlier. Outlier detection signals out the objects mostly deviating from a given data set.

Data mining, in general, deals with the discovery of non-trivial, hidden and interesting knowledge from different types of data. With the development of information technologies, the number of databases, as well as their dimension and complexity, grow rapidly. It is necessary what we need automated analysis of great amount of information. The analysis results are then used for making a decision by a human or program. One of the basic problems of data mining is the outlier detection [9].

Outlier detection as a branch of data mining has many important applications and deserves more

attention from data mining community [3]. Detecting outliers has important applications in data cleaning as well as in the mining of abnormal points for fraud detection, stock market analysis, intrusion detection, marketing, network sensors [1].

Outliers may be erroneous or real in the following sense. Real outliers are observations whose actual values are very different than those observed for the rest of the data and violate plausible relationships among variables. Erroneous outliers are observations that are distorted due to misreporting or miss recording errors in the data-collection process. Outliers of either type may exert undue influence on the results of statistical analysis, so they should be identified using reliable detection methods prior to performing data analysis [9].

The proposed system uses a clustering based and distance based technique to find the outliers within the data set. In proposed outlier detection technique ,first apply bisecting K–Means clustering algorithm to divided a data set into cluster and then the point which are lying near the centroid of the cluster are not probable candidate for outlier and that points can be prune out from each cluster. Next calculated a distance based outlier score for remaining points. Based on this outlier score declare the top n points with the highest score as outliers. This proposed system functions like Sampling, scaling and flagging.

The salient approaches to outlier detection can be classified as either distribution-based, depth based, clustering, distance-based or density-based [2]. In this proposed work there are two techniques are used which is cluster based and distance based, For clustering based approach uses the bisecting K-Means algorithm and for distance based approach uses the EDA (Euclidean Distance Algorithm).

II. Related Work

The first used technique for outlier detection is the MDL(minimum description length) and smoothing factor to find the outliers [4]. Then *Breunig et.al* proposed a local outlier factor(LOF) for each object I the data set, indicating its degree of outlierness. The LOF value of an object is obtained by comparing its density with those in its neighborhood [5].

After that the proposed system used a local distance based outlier factor (LDOF) method to find outlier from data set.

which determines the degree to which object deviates from its neighborhood. Calculating LDOF for all points in the data set, makes overall complexity O(N2), where N is

the number of points in the data set [6]. Next system *Rajendra pamula* proposed for outlier detection is the micro clustering based local outlier mining algorithm which is distribution based and depth based [7].*Knorr and Ng* [8] were the first to introduce distance based outlier detection techniques. An object p in a data set DS is a DB (q,dist)-outlier if at least fraction q of the objects in DS lie at a greater distance than dist from p. This definition is well accepted, since it generalizes several [8].

Clustering methods like CLARANS, DBSCAN, BIRCH and CURE may detect outliers. However, since the main objective of a clustering method is to find clusters, they are developed to optimize clustering, and not to optimize outlier detection. The definitions of outlier used are subjective to the clusters that are detected by these algorithms. While definitions of distance-based outliers are more objective and independent of how clusters in the input data are identified [1]., used K-means algorithm for clustering and LDOF for finding the distance based outlier score. This method used to find out the outlier score from that find the top-n points with the highest score of outlier. LDOF define as *ldof*(*p*) := *dp*/ *Dp* [1].

III. PROPOSED METHODOLOGY

The main idea underlying the pruning-based algorithm is to first cluster the data set into clusters, and then prune the points in different clusters if determined that they cannot be outliers. Since *n* (number of outliers) will typically be very small, this additional preprocessing step helps to eliminate a significant number of points which are not outliers. We describes our method to find out outliers.

We briefly describe the steps need to be performed by our pruning based algorithm.

(1) **Generating clusters:** Initially, we cluster the entire dataset into *N* clusters using Bisecting K-means clustering algorithm and calculate radius of each cluster.
(2) **Clusters having less number of points:** If a cluster contains less number of points than the required number of outliers, the radius pruning is avoided for that cluster.

(3) **Pruning points inside each cluster:** Calculate distance of each point of a cluster from the centroid of the cluster. If the distance of a point is less than the radius of a cluster, the point is pruned.

(4) **Computing outlier points:** Calculate distance for all the points that are left un pruned in all the clusters. Then *n* points with high Distance values are reported as outliers.

A. Clustering Algorithm

In the proposed system we have used the bisecting k-mean algorithm to cluster the data set.

Basic Bisecting K-means Algorithm for finding K clusters:-

(1) Pick a cluster to split.

(2) Find 2 sub-clusters using the basic k-Means algorithm (*Bisecting step*)

(3) Repeat step 2, the bisecting step, for time and take the split that produces the

clustering with the highest overall similarity.

(4) Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

B. Outlier Detection Methods

Outlier detection is very essential of any modelling exercise. A failure to detect outliers or their ineffective handling can have serious ramifications on the strength of the inferences drained from the exercise. 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.

There is no standard technique for outlier detection. Some of the outlier detection techniques are:

(1) Distance based outlier detection
(2) Clustering based outlier detection
(3) Density based outlier detection
(4) Depth based outlier detection

Each of these techniques has its own advantages and disadvantages. In general, in all these methods, the technique to detect outliers consists of two steps. The first identifies an outlier around a data set using a set of inliers (normal data). In the second step, a data request is analyzed and identified as outlier when its attributes are different from the attributes of inliers. All these techniques assume that all normal instances will be similar, while the anomalies will be different.

In the proposed methodology we have used the distance based outlier detection technique in the we have used EDA i.e. Euclidean Distance Algorithm. Which perform the task of finding the distance between the point and the centroid and pruned out such points which is larger than the defined distance.
The methods used in EDA as:-

$$d^2_{rs} = (x_r - x_s)(x_r - x_s)' \quad (1)$$

C. System Architecture

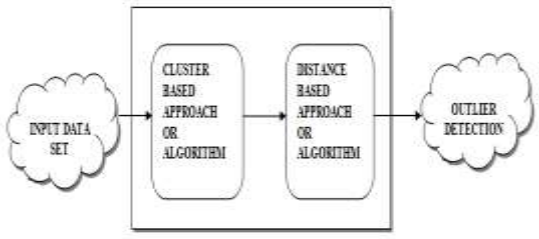

Fig 1. Outlier detection system

C. *Following steps need to be performed by our pruning based algorithm:-*

- **Input Data Set:** A data set is an ordered sequence of objects X1, ..,Xn.

- **Cluster Based Approach:** Clustering technique is used to group similar data points or objects in groups or clusters. Clustering is an important tool for outlier analysis. For this proposed system uses the bisecting K-Mean algorithm

- **Distance Based Approach:** This technique is highly dependent on the parameters provided by the users. Given any distance measure, objects that have distances to their nearest neighbor that exceed a specific threshold are considered potential anomalies. Proposed system uses the Euclidean distance Algorithm (EDA) EDA used the method as :-

$$d^2_{rs} = (x_r - x_s)(x_r - x_s)' .$$

- **Outlier Detection:** Outlier detection is an extremely important task in a wide variety of application domains. Outlier detection is a task that finds objects that are dissimilar or inconsistent with respect to the remaining data or which are far away from their cluster centroid.

D. *Proposed Clustering and distance based Algorithm*

Generating cluster: Bisecting K-means clustering is a partitioning method .Initially, cluster the entire dataset into k cluster using bisecting K-mean clustering and calculate centroid of each cluster.
Clustering: Given k, the bisecting k-means algorithm is implemented in four steps:

(a) Select k observations from data matrix X at random

(b) Calculate distance with each instances (with respect to randomly selected instances)

(c) Assign each instance to the cluster with the nearest seed

(d) Go back to Step b, stop when no instance to move group

(1) Calculate Distance for each cluster (using EDA)

---------- finding min max values from each clusters
----------finding maximum distance from centroid
---------- take distance from user
---------- find distance value for each cluster

(2) Calculate distance of each point of cluster from centroid of the cluster. If the distance of
points is greater than distance given by user then it will declare as —"outlier".

The proposed system is perform better than existing system on the basis of performance because existing system used the K-mean clustering and whose complexity is $O(N^2)$ and the proposed system will used the bisecting k-means clustering whose complexity is O(N).so that it is proposed that the proposed system will fast detect the outliers in data sets.

IV. CONCLUSION

In this paper, we have proposed a method to detect the outlier from data set. We have used the clustering technique such as bisecting k-mean and it is compared with the k-mean clustering technique and for finding the outlier used the distance based approach .using and combining these two method we proposed that the system perform better than existing system. Bisecting k- means work better than k-means and we can find the outlier fast.

V. REFERENCES

1. Rajendra pamula,Jatindra Kumar Deka,Sukumar Nandi, An Outlier Detection Method based on clustering ,Second international conference On Emerging application of information technology,2011.
2. Spiros Papadimitriou, Hiroyuki Kitagawa, Phillip B. Gibbons, LOCI: Fast Outlier Detection Using the Local Correlation Integral ,I data engineering ,Proceeding,19th international conference on data engineering ,March 2003
3. Ms. S. D. Pachgade, Ms. S. S. Dhande, Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012
4. A.arning,R.Agrawal and P.Raghavan, A linear method for deviation detection in large database,1996.
5. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: identifying density-based local outliers, SIGMOD,2000
6. K. Zhang, M. Hutter, and H. Jin. ,A new local distance-based outlier detection approach for scattered real-world data,In PAKDD '09: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining,2009
7. W. Jin, A. K. H. Tung, and J. Han., Mining top-n local outliers in large databases, In KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001.
8. E. M. Knorr and R. T. Ng, Algorithms for mining distance based outliers in large datasets, In Proc. 24th Int. Conf. Very Large Data Bases, 1998.
9. Svetlana Cherednichenko, Outlier Detection in Clustering,2005