Available at: www.researchpublications.org

Joint Approach for Outlier Detection

Niketa V. Kadam[#] , Prof. M. A. Pund[*]
[#] *M.E., Information Technology,P.R.M.I.T.& R,Badnera* - [*] *Information Technology,P.R.M.I.T.& R,Badnera*

[1]niketak39@gmail.com
[2]mapund@mitra.ac.in

***Abstract*: Outlier is nothing but the unusual result in the dataset. Now a days outlier detection becomes the need. The serious problem arises with segmenting the data into the number of chunks. To improve the effectiveness of outlier detection, we need proper partition of the data. For the detection of the outlier first the data must be segmented into the number of chunks and after that each chunk is compared with the another one for getting the candidate outlier From the computational point of view, the main obstacle to find the outlier is the vast dataset. In realistic settings, all of the above complicating factors do not appear in isolation, but contribute collectively to increasing the complexity of the comparison problem.**

There are the different logical approaches to detect outlier data values, such as clustering algorithms like, K-mean or K-median, which can produce good clustering results and at the same time deserves good scalability. Finally, distance based technique is used to find the distance from centroid to candidate outlier. So that, both approaches take less computational cost. In brief, to avoid pair wise distance calculations, to detect better outlier even if the evolution of data set change, to let user free to provide sensitive parameters, and to mine Data set even in limited memory resources here we propose a clustering based method because; the clustering methods have good space and time complexity.

Keywords: **Outlier, Cluster based, Distance based.**

I. INTRODUCTION

Outlier detection is currently very active area of research in data set mining community. However, earlier research for the problem of outlier detection is suitable for disk resident datasets where the entire dataset is available in advance and algorithms can operate in more than single passes. But, outlier detection over data set is a challenging task because data is continuously updated and flowing. Outlier is a data point that does not conform to the normal points characterizing the data set. 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. Finding anomalous points among the data points is the basic idea to find out an outlier.

Finding outliers in a collection of patterns is a very well-known problem in the data mining field. An outlier is unusual pattern with respect to the rest of the patterns in the dataset. Depending upon the application domain, outliers are of particular interest. In some cases presence of outliers are adversely affect the conclusions drawn out of the analysis and hence need to be eliminated beforehand. There are varied reasons for outlier generation in the first place. For example outliers may be generated due to measurement impairments, rare normal events exhibiting entirely different characteristics, deliberate actions etc. Detecting outliers may lead to the discovery of truly unexpected behavior and help avoid wrong conclusions etc. Most of the existing work for outlier detection over the data set only focus on detection rate of outliers while ignoring the most important issue of data set mining like, low memory requirements and high speed algorithms to keep pace with high speed unbounded data set. [1] [2][3]

In this work, we identify the points which are not outliers using clustering and distance functions, and prune out those points. Next, we calculate a distance-based measure for all remaining points, which is used as a parameter to identify a point to be an outlier or not. These techniques were highly dependent on the parameters provided by the users and were computationally expensive when applied to unbounded data streams.[4]

II. RELATED WORK

Existing approaches to the problem of outlier Detection are summarized as follows. Outlier detection (deviation detection, exception mining, novelty detection, etc.) is an important problem that has attracted wide interest and numerous solutions. These solutions can be broadly classified into several major ideas:

Model-Based Approach:

The complete processing of this approach is based on the model. An explicit model of the domain is built (i.e., a model of the heart, or of an oil refinery), and objects that do not fit the model are flagged. Means depending upon the model which we select for the processing on that only the complete process depends. [2]

Disadvantage: Model-based methods require the building of a model, which is often an expensive and difficult enterprise requiring the input of a domain expert.

Connectedness Approach:

In domains where objects are linked (social networks, biological networks), objects with few links are considered potential anomalies. [5][6]

Disadvantage: Connectedness approaches are only defined for datasets with linkage information.

Distance-Based Approach:

Given any distance measure, objects that have distances to their nearest neighbours that exceed a specific threshold are considered potential anomalies. In contrast to the above, distance-based methods are much more flexible and robust. They are defined for any data type for which we have a distance measure and do not require a detailed understanding of the application domain. [7][8][9]

Cluster Based Approach:

The clustering based techniques involve a clustering step which partitions the data into groups which contain similar objects. The assumed behavior of outliers is that they either do not belong to any cluster, or belong to very small clusters, or are forced to belong to a cluster where they are very different from other members. Clustering based outlier detection techniques have been enveloped which make use of the fact that outliers do not belong to any cluster since they are very few and different from the normal instances. [10][11][12]

Density-Based Approach:

In the Density Based approach, author Breunig et al described one technique for the outlier detection. In which the outlier detection is depend upon the density. Here Objects in low-density regions of space are flagged.

Disadvantage: Density based models require the careful settings of several parameters. It requires quadratic time complexity.

It may rule out outliers close to some non-outliers patterns that has low density.[13]

II. CLUSTER BASED and DISTANCE BASED APPROACH

The proposed system will be identified to provide a solution to the problem of outlier detection. Outlier detection i.e. searching for abnormal values. As an Example, we are considering 1000 data elements in the data set. In first stage, Partition the data set into number of chunks and each chunk contain set of data. Suppose we made partitions the data set in to 10 number of chunks each with 100 elements as P1 - - - - P10. In second stage, over each chunk, apply clustering method to figure out candidate outliers and safe region i.e. grouping the data elements with each chunk. In the third stage, applying distance based outlier detection algorithm (For detecting outliers) over clusters with respect to centroid of cluster. In the fourth stage giving a chance to the candidate outlier to survive in next set, and allow it for appropriate number of set chunks, and then declare candidate outliers as real outliers or inliers.

A.Techniques Used:

I) Cluster-based approach:

Cluster based approach is here used to reduce the size of dataset i.e., act as data reduction. First, cluster based technique is used to form cluster of dataset. Once cluster are formed, centriod of each cluster are calculated. Remove the data up to certain radius as a real data. After removing the real data, remaining data are the candidate outlier. Candidate outliers are the temporary outlier. Figure 1 shows Cluster–Based Approach.

Clustering algorithm (K-mean):

K-number of cluster, we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids. Clustering is nothing but the grouping the data.

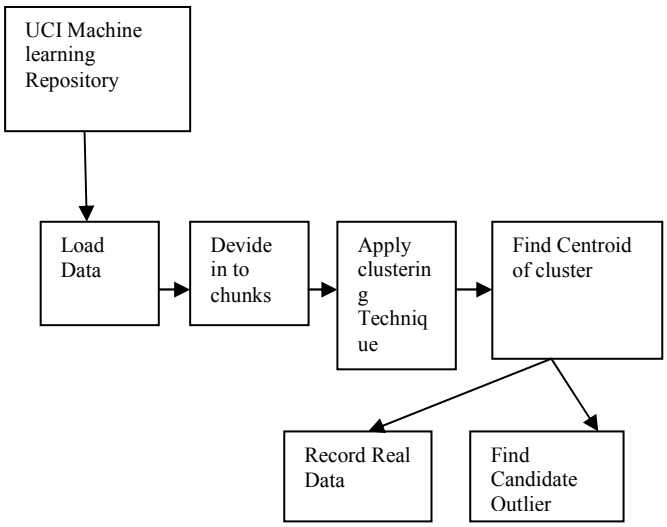

Fig. 1 Cluster Based Approach

The K means algorithm will do the following steps:

Generating clusters:

Iterate until stable (= no object move group)

Determine the centroid coordinate

Determine the distance of each object to the centroid

Group the object based on minimum distance

By this way we can cluster the entire dataset in to number of clusters and calculate centroid of each cluster.

Find Candidate cells:

Remove the data up to certain radius as a real data. After removing real data rest of the data will be candidate outlie

II) Distance-based approach:

Distance based technique is used to find the distance from centroid to candidate outlier. If this distance is greater than some threshold then it will declare as “outlier” otherwise as a real object.

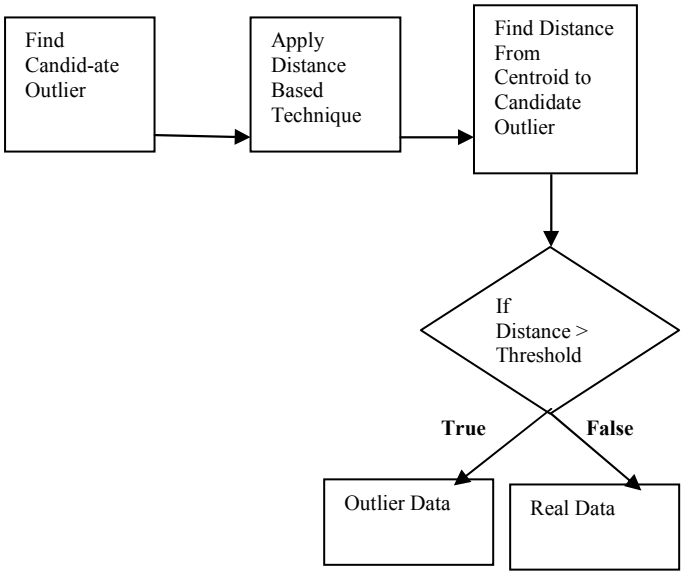

Fig. 2 Distance Based Approach

Distance-based Algorithm steps:
i) Centroid of each cluster is calculated
ii) Calculate distance of each point (candidate outlier) from centroid of the cluster.
ii) If Distance >Threshold then it will declare finally as "outlier" otherwise as a "real" data.

B. Joint approach

This work will prove to be most efficient for the problem of outlier detection as this approach simply combines the cluster based and distance based approaches that can be used for outlier detection more efficiently.
It can be divided in 4 steps:
I) Partition the data set into number of chunks and each chunk contain set of data.
II) Over each chunk, apply clustering method to figure out candidate outliers and safe region.
III) Apply distance based outlier detection algorithm over clusters with respect to centroid of cluster.
IV) Give a chance to the candidate outlier to survive in next set, and allow it for appropriate number of set chunks, and then declare candidate outliers as real outliers or inliers.

IV. CONCLUSION:

The joint approach for outlier detection Detection presents many interesting advantages w. r. t. previous proposals in the field of data mining.This approach is form just by combining the cluster based and distance based approach.Both approaches take less computational cost. In brief, to avoid pair wise distance calculations, to detect better outlier even if the evolution of data set change, to let user free to provide sensitive parameters, and to mine There are two algorithms which are used for the efficient outlier detection they are k-means and k-median. So this hybrid model will provide solution to the problem of outlier detection.

REFERENCES

[1] Zang et al., 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.

[2] Fayyad et al.,U.M.; Piatetsky-Shapiro, G.; Smyth, P."The KDD Process for Extracting Useful Knowledge from Volumes of Data" Communications Of The ACM,1996.

[3] M. Knorr and R.T.Ng. "Finding intentional knowledge of distance-based outliers" *In VLDB '99: Proceedings of the 25th International Conference on Very Large Data Base,*1999.

[4] Elahi et al.,ManzoorElahi, Kun Li, WasifNisar, XinjieLv, Hongan Wang, "Efficient Clustering-Based Outlier Detection Algorithm for Dynamic Data Stream" In Proc.of Fifth International Conference on Fuzzy Systems and Knowledge Discovery ,2008.

[5] Yongzhen Zhuang and Lei Chen Hong Kong, In-network Outlier Cleaning for Data Collection in Sensor Networks" Yongzhen Zhuang and Lei Chen Hong Kong University of Science and Technology fcszyz, leicheng,2008.

[6] Parneeta Dhaliwal , MPS Bhatia and Priti Bansal, "A Cluster-based Approach for Outlier Detection in Dynamic Data Streams" (KORM: k-median OutlieR Miner) Parneeta Dhaliwal , MPS Bhatia and Priti Bansal,2010.

[7] Niennattrakul etal Vi Niennattrakul, Eamonn Keogh, Chotirat Ann Ratanamahatana, "Data Editing Techniques to Allow the Application of Distance-Based Outlier Detection to Streams", *IEEE International Conference on Data Mining (ICDM) 2010.*

[8] Anscombe&Guttman, F. J. Anscombe and I. Guttman, "Rejection of Outliers," Technometrics, vol. 2, pp. 123-147, May 1960.

[9] Tang et al., J. Tang, Z. Chen, A. W.-C. Fu and D. W.-L. Cheung, "Enhancing Effectiveness of Outlier Detections for Low Density Patterns," *In Proceedings of PAKDD'02*, May 6-8 2012.

[10] Angiulli&Fassetti, F. Angiulli and F. Fassetti, "Detecting Distance-based Outliers in Streams of Data," *In Proceedings of CIKM'07*, November 6-10 2007.

[11] Barnett and Lewis, Barnett V., Lewis T., Outliers in Statistical Data.John Wiley, 1994.

[12] Dhaliwal et al., ParneetaDhaliwal, MPS Bhatia and PritiBansal," A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner)" JOURNAL OF COMPUTING, VOLUME 2, ISSUE 2, 2010.

[13] Yang & Huang, KNN Based Outlier Detection Algorithm in Large Dataset" *International Workshop on Education Technology and Training,* 2008.