ISSN No. 0976-5697

Volume 4, No. 2, Jan-Feb 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Cluster Based and Distance Based Approach for Outlier Detection

Ms.Niketa V. Kadam*
Information Technology
P.R.M.I.T.&R. Badnera,Amravati
niketak39@gmail.com

Prof. M.A.Pund
M.C.A.
P.R.M.I.T.& R. Badnera,Amravati
mapund@mitra.ac.in

Abstract: Outlier Detection now a days become very vast area. Outlier is nothing but the abnormal result or the any kind of uncertainity among the others. The various techniques are used for outlier detection . The hybrid approach i. E. The combination of the cluster based and distance based approach provides the efficient solution to the problem of the outlier detection under certain conditions.

Keywords: outlier, cluster based, distance based

I. INTRODAUCTION

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.

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. [1] [2]

II. LITERATURE REVIEW/SURVEY

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:

A. 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]

a. 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.

B. Connectedness Approach:

In domains where objects are linked (social networks, biological networks), objects with few links are considered potential anomalies. [3]

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

C. Distance-Based Approach:

Given any distance measure, objects that have distances to their nearest neighbors 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. [4]

D. 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. [5]

E. 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. [6]

III. SYSTEM PLANNING AND DESIGN

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:

a. 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.

a) 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.

The K means algorithm will do the following steps:

B. Generating clusters:

Iterate until stable (= no object move group)

- i. Determine the centroid coordinate
- ii. Determine the distance of each object to the centroid
- iii. 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.

a. 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 outlier.

a) 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.

Distance-based Algorithm steps:

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

C. System Design:

- ***a.*** ***Input Data Set:*** A data set is an ordered sequence of objects *X*1, ..,*Xn*. Applications, such as fraud detection, network flow monitoring, telecommunications, data management, etc., where the data arrival is continuous.
- ***b.*** **Cluster Based Approach:** Clustering is a popular technique used to group similar data points or objects in groups or clusters. Clustering is an important tool for outlier analysis. This technique relies on the key assumption that normal objects belong to large and dense clusters, while outliers form very small clusters.

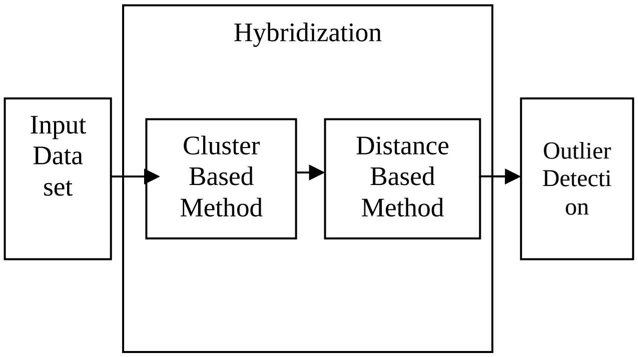

Figure 1: Process of Outlier Detection

- ***c.*** ***Distance Based Approach:*** These techniques are 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.
- ***d.*** **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. Outliers are often considered as error or noise and are removed once detected. Examples include skewed data values resulting from measurement error, or erroneous values resulting from data entry mistakes.

D. Architecture of system:

It can be divided in 4 steps:

- a. Partition the data set into number of chunks and each chunk contain set of data.
- b. Over each chunk, apply clustering method to figure out candidate outliers and safe region.
- c. • Apply distance based outlier detection algorithm over clusters with respect to centroid of cluster.
- d. 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.

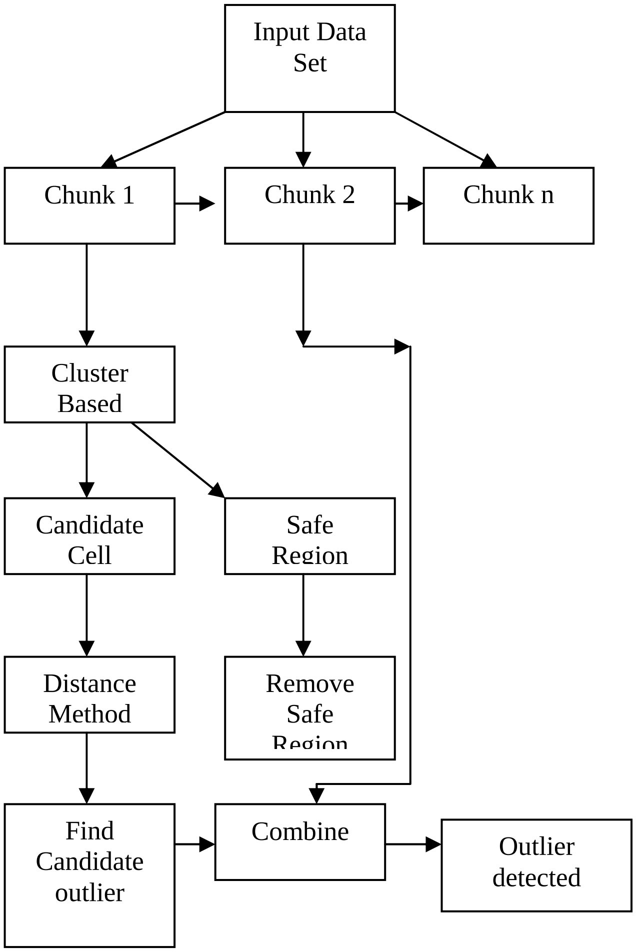

Figure.2 System Architecture

IV. CONCLUSION

The dissertation title is formed by exploring the need and available techniques of Outlier Detection. The problem is defined by taking literature survey in related topics of different approaches of Outlier Detection. Also identifying the parameters and features of techniques used like cluster based and distance based approaches. The ybrid model will provide solution to the problem of outlier detection.

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

[3]. 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.

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

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

[6]. 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.