Volume 2, Issue 6, June 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach

Ms. S. D. Pachgade, Ms. S. S. Dhande
Computer Science & Engineering
Sipna's College of Engg& Tech.
Amravati, Maharashtra.
India
sdpachgade@gmail.com

***Abstract*— Outlier detection is currently very active area of research in data set mining community. Finding outliers in a collection of patterns is a very well-known problem in the data mining field. An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the dataset. Proposed Method for outlier detection uses hybrid approach. Purpose of approach is first to apply clustering algorithm that is kmeans 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 centroids. In proposed approach, two techniques are combining to efficiently find the outlier from the data set. The experimental results using real dataset demonstrate that proposed method takes less computational cost and performs better than the distance based method. Proposed algorithm efficiently prunes of the safe cells (inliers) and save huge number of extra calculations.**

***Keywords*— Outlier, Cluster-based, Distance-based.**

I. INTRODUCTION

Data mining is a process of extracting hidden and useful information from the data and the knowledge discovered by data mining is previously unknown, potentially useful, and valid and of high quality. Finding outliers is an important task in data mining. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. In recent years, conventional database querying methods are inadequate to extract useful information, and hence researches nowadays are focused to develop new techniques to meet the raised requirements. It is to be noted that the increase in dimensionality of data gives rise to a number of new computational challenges not only due to the increase in number of data objects but also due to the increase in number of attributes. Outlier detection is an important research problem that aims to find objects that are considerably dissimilar, exceptional and inconsistent in the database. Medical application is a high dimensional domain hence determining outliers is found to be very tedious due to the Curse of dimensionality. There are various origins of outliers. With the growth of the medical dataset day by day, the process of determining outliers becomes more complex and tedious. Efficient detection of outliers reduces the risk of making poor decisions based on erroneous data, and aids in identifying, preventing, and repairing the effects of malicious or faulty behavior. Additionally, many data mining and machine learning algorithms and techniques for statistical analysis may not work well in the presence of outliers. Outliers may introduce skew or complexity into models of the data, making it difficult, if not impossible, to fit an accurate model to the data in a computationally feasible manner. For example, statistical measures of the data may be skewed because of erroneous values, or the noise of the outliers may obscure the truly valuable information residing in the data set. Accurate and efficient removal of outliers may greatly enhance the performance of statistical and data mining algorithms and techniques [6]. Detecting and eliminating such outliers as a pre-processing step for other techniques is known as data cleaning. As can be seen, different domains have different reasons for discovering outliers: They may be noise that we want to remove.

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. Distance based techniques use the distance function for relating each pair of objects of the data set. Distance based definition (these definitions are computationally efficient) [7, 10] represent a useful tool for data analysis [8].

In this work, we are introducing clustering method that will reduce size of datasets, and groups the data having similar characteristic. Next we apply distance based to detect the outliers as per given threshold. Within a cluster get outliers, that are far from their cluster centroid.

II. OBJECTIVES OF STUDY

Basic aims to reduce the number of pair wise distance calculations, to let user free to provide sensitive parameters .We are first testing with distance based approach; this approach applies to all data, then testing with hybrid approach. In that we are first partition the data in to number of clusters and then we apply distance based approach. The principle of outlier's detection depends on the threshold. This approach takes less computational time than distance based method. **1)**

III. RELATED WORK

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 [2]: 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.

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

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

Density-Based [3]: 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.

Distance-Based [1]: 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.

Cluster based approach [4]: 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.

K-Nearest Neighbor Based Approach [12]: K-nearest neighbor based schemes analyses each object with respect to its local neighborhood. The basic idea behind such schemes is that an outlier will have a neighborhood where it will stand out, while a normal object will have a neighborhood where all its neighbors will be exactly like it. The obvious strength of these techniques is that they can work in an unsupervised mode, i.e. they do not assume availability of class labels.

IV. PROPOSED WORK

1. System architecture

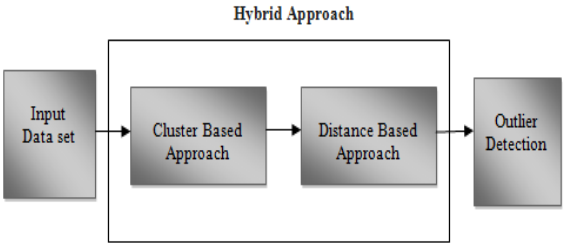

Figure 1: System Architecture

Input Data Set: Collecting dataset from UCI Machine learning repository [13].

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. Cluster based approach is here act as data reduction. First, clustering technique is used to groups the data having similar characteristics. And calculate the centroids for each group.

Distance Based Approach: Distance based technique is used to calculate maximum distance value for each cluster. If this maximum distance is greater than some threshold then it will declare as "outlier" otherwise as a real object or inliers. Threshold is given by user.

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 centroids.

2. Distance based Algorithm

This method is highly dependent on parameter provided by the users and computationally expensive when applied unbounded data set. With the development of information technologies, the number of databases, as well as their dimensions and complexity grow rapidly. With high dimensional dataset calculate distance with each instances will increase the computational cost. We are comparing distance based method with proposed method.

Pairwise distance computes the Euclidean distance between pairs of objects in n-by-p data matrix X. Rows of X correspond to observations; columns correspond to variables. y is a row vector of length n(n–1)/2, corresponding to pairs of observations in X. The distances are arranged in the order (2,1), (3,1), ..., (n,1), (3,2), ..., (n,2), ..., (n,n–1)). y is commonly used as a dissimilarity matrix in clustering or multidimensional scaling.

Euclidean distance

$$d_{rs}^2 = (x_r - x_s)(x_r - x_s)'$$

Where,

$$\bar{x}_r = \frac{1}{n}\sum_j x_{rj} \quad \text{and}$$

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

1) Calculate pairwise distance that is computing the Euclidean distance between pairs of object.
2) Take square distance. Calculate maximum values from square distance values
4) Take threshold from user.
5) If distance > threshold value that will be the outliers.

3. Proposed Clustering and Distance-Based Algorithm

Generating cluster: *K*-means clustering is a partitioning method .Initially, cluster the entire dataset into k cluster using K-mean clustering and calculate centroid of each cluster.

Kmean Clustering: Given *k*, the *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
2) Calculate Threshold % for each cluster
-- finding min max values from each clusters
-- finding maximum distance from centroid
-- take threshold from user
-- find threshold % value for each cluster
3) Calculate distance of each point of cluster from centroid of the cluster. If the distance is greater than threshold then it will declare as "outlier".

V. EXPERIMENTAL RESULTS

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. Data is collected from UCI machine learning repository that provided various types of datasets. This dataset can be used for clustering, classification and regression. Dataset has multiple attribute and instances. A repository of databases, domain theories and data generators are used by the machine learning community for the empirical analysis of machine. Data File Format is in .data and .xls excel file or .txt or .csv file format. This data file will be taken to find the outlier.

1) Medical Diagnosis Data Set: In real-world data repositories, it is hard to find a data set for evaluating outlier detection algorithms, because only for very few real-world data sets it is exactly known which objects are really behaving differently. In this experiment, we use a medical data set, WDBC (Diagnosis), which has been used for nuclear feature extraction for breast tumor diagnosis. The data set contains 428 medical diagnosis records (objects), each with 32 attributes (ID, diagnosis, 30 real-valued input features).The diagnosis is binary: Benign and Malignant. There are 2 types of datasets so we are dividing dataset into 2 numbers of clusters.

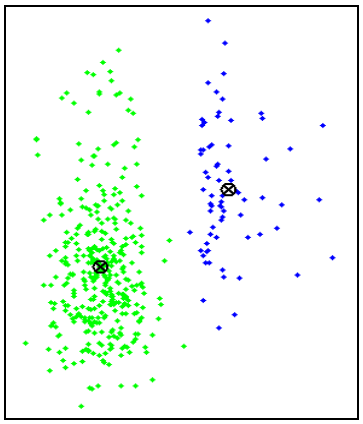

Figure 2: Clustering and Outliers for Cancer Dataset

Number of Data Points in each cluster for Cancer dataset	
1 Cluster	71
2 Cluster	357

Table 1: Shows number of data points for cancer Dataset

Here, we get groups of 71 instances in first cluster and 357 instances in second cluster among 428 instances. Group of 71 belongs to malignant record and group of 357 belongs to benign record. From first cluster 10 numbers of outliers detected at 75 percent of threshold. From second cluster 21 outliers detected at 75 percent of threshold.

Threshold %		75	80	85	90	95
No. of Outliers	1 Cluster	10	8	7	6	6
	2 Cluster	21	19	17	11	9

Table 2: Test on different threshold value for cancer dataset and getting variations in number of outliers

Elapsed Time	
Distance Based Approach	Proposed Approach
0.29246s	0.0954145s

Table 3: CPU Time in Second for Cancer Dataset

2) Bupa liver disorder datasets: which refers to the first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the bupa Data file constitutes the record of a single male individual. It appears that drinks > 5 is some sort of a selector on this database. Selector field used to split data into two sets. The Dataset contains 345 instances and 7 attributes. In this dataset only few numbers of outliers are detected at 75%. From first cluster only one outlier is detected and in second cluster two outliers detected.

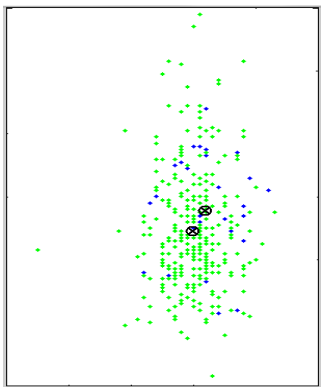

Figure 3: Clustering and Outliers for Liver Disorder Dataset

Number of Data Points in each cluster for Liver Disorder		Number of outliers at 75 %
1 Cluster	37	1
2 Cluster	308	2

Table 4: Shows number of data points and outliers for liver Disorder Dataset

Elapsed Time	
Distance Based Approach	Proposed Approach
0.307235s	0.10975s

Table 5: CPU Time in Second for Liver Disorder Dataset

VI. DISCUSSION

Finding outliers is an important task in data mining. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. Comparison between Distance based approach and proposed approach are as follows:

Distance-Based Method

- Operate on whole data. Cannot give number of clusters.
- Computation time will increases
- Give only one value as most expected outlier

Clustering and Distance-Based

- Can group the data in to number of clusters
- Reduce the size of database that will reduces computation time
- To each cluster user can give certain radius to find outliers.

VII. CONCLUSION

This papers aims to detect outliers is the task that finds objects that are dissimilar or inconsistent with respect to remaining data. We proposed an efficient outlier detection method. We first groups the data (having similar characteristics) in to number of clusters. Due to reduction in size of dataset, the computation time reduced considerably. Then we take threshold value from user and calculate outliers according to given threshold value for each cluster. We get outliers within a cluster. Hybrid approach takes less computation time.

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.

REFERENCES

[1] F. Angiulli and F. Fassetti, "Detecting Distance-based Outliers in Streams of Data," In Proceedings of CIKM'07, Pages 811-820, November 6-10 2007.

[2] F. J. Anscombe and I. Guttman, "Rejection of Outliers," Techno metrics, vol. 2, Pages 123-147, May 1960.

[3] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "OPTICS-OF: Identifying Local Outliers," In

Proceedings of PKDD'99, Pages 262- 270, September 15-18 1999.

[4] Parneeta Dhaliwal, MPS Bhatia and Priti Bansal," A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner)" JOURNAL OF COMPUTING, VOLUME 2, ISSUE 2, FEBRUARY 2010, ISSN: 2151-9617.PAGES 74-80.

[5] Manzoor Elahi, KunLi, Wasif Nisar, Xinjie Lv, Hongan Wang, "Efficient Clustering-Based Outlier Detection Algorithm for Dynamic Data Stream" In Proc .of Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD.2008),ISBN: 978-0-7695-3305-6/08, pages 298-304.

[6] Hadi A.S., A.H.M.R. Imon, and M. Werner, "Detection of outliers," *Computational Statistics*, vol. 1, 2009, 57-70.

[7] 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, VLDB, pages 392–403, 1998.

[8] M. Knorr and R. T. Ng. "Finding intentional knowledge of distance-based outliers" In VLDB '99: Proceedings of the 25thInternational Conference on Very Large Data Bases, pages 211–222, 1999.

[9] Rajendra Pamula,Jatindra kumar Deka,Sukumar Nandi."An Outlier Detection Method based on Clustering", Second International Conference on Emerging Applications of information Technology,2011. ISBN: 978-0-7695-4329-1/11, Pages 253-256.

[10] Ramaswamy, R. Rastogi, and K. Shim. "Efficient algorithms for mining outliers from large data sets" pages 427–438, 2000.

[11] 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, Pages 535-548, May 6-8 2002.

[12] Peng Yang; Biao Huang;" KNN Based Outlier Detection Algorithm in Large Dataset" International Workshop on Education Technology and Training, ISBN: 978-0-7695-3563-0, Pages 611 – 613, 2008.

[13] http://archive.ics.uci.edu/ml/

