Outlier Detection in Dataset using Hybrid Approach

Shivani P. Patel
Research Scholar
G.H. Patel College of Engg and Tech.
Vallabh Vidyanagar, India

Vinita Shah
Asst. Prof., IT Dept.
G.H. Patel College of Engg and Tech.
Vallabh Vidyanagar, India

Jay Vala
Asst. Prof., IT Dept.
G.H. Patel College of Engg and Tech.
Vallabh Vidyanagar, India

ABSTRACT

Outlier is a data point that deviates too much from the rest of dataset. Most of real-world dataset have outlier. Outlier analysis is one of the techniques in data mining whose task is to discover the data which have an exceptional behavior compare to remaining dataset. Outlier detection plays an important role in data mining field. Outlier Detection is useful in many fields like Medical, Network intrusion detection, Credit card fraud detection, medical, fault diagnosis in machines, etc. In order to deal with outlier, clustering method is used. Outlier detection contains clustering and finding outlier by applying any outlier detection technique. For that K-mean is widely used to cluster the dataset. Different techniques like statistical-based, distance-based, and deviation-based and density based methods are used to detect outlier. The experiment result shows that existing algorithm perform better than proposed cluster-based and distance-based Algorithm.

General Terms

Clustering, Outlier Detection.

Keywords

Data Mining, Outlier, Clustering Approach, k-mean Algorithm, Distance Based Approach

1. INTRODUCTION

Data mining is one of the steps of Knowledge Discovery from Dataset process. This process recognize interesting pattern from large dataset by performing data cleaning, integration, selection, mining, pattern evaluation and knowledge presentation. Data mining involves three common tasks are association rule mining, clustering and classification.

An outlier is an observation of data point that deviates too much from other point that they are generated because of faulty condition in experiment. Outlier can be caused by measurement or execution error. For example, display person’s age as (-49) or 554. Different applications of outlier detection are credit card fraud detection, network intrusion detection, detecting outlying in wireless sensor network data, fault diagnosis in machines, stoke market analysis, etc. In order to deal with outlier, different outlier detection methods are used.

In order to deal with outlier clustering is also useful. Clustering is process of grouping similar objects of a dataset into one cluster or class. For example, in general store if we want to retrieve items easily and quickly, we can group the items in such way that similar items put into one group and another items into different group, and such grouping can be known as clustering. Now a day’s most popular and widely used clustering algorithm for outlier detection is k-mean algorithm.

Objective of this Research work is that, Different dataset have different characteristics regarding a specific variable, such as height data not stratified by gender. Outliers can be caused by incorrect measurements, including data entry errors, or by coming from a different population than the rest of the data. To detect such outlier is useful in many applications. There are different approaches available for outlier detection such as distance-based, density-based, statistical-based, and deviation-based approach. Objective of this research work is to detect outliers from dataset by implementing clustering k-mean with determining initial centroid and by using distance based outlier detection approach to get more accurate result for outlier detection.

Section 2 includes literature survey in which different authors were introducing different technique to detect outlier from different dataset. Section 3 discusses methodology used for proposed algorithm. Also discuss existing algorithm which include cluster based k-mean and distance-based approach for outlier detection. Section 4 defines steps of Proposed Algorithm. Section 5 shows the experiment result. And the last section 6 provides the conclusion and future work.

2. RELATED WORK

S. Vijayarani and S. Nithya [5] used clustering algorithm are PAM, CLARA, CLARANS and they proposed E-CLARANS algorithm for outlier detection. They define the problem of CLARANS algorithm and overcome that by selecting proper arbitrary nodes instead of random selection. And conclude that proposed algorithm perform better but it take more computation time than CLARANS.

RajendraPamula, Jatindra Kumar Deka, and Sukumar Nandi [10] used local distance based outlier factor(LDOF) to find outlier. But this method is computationally expensive because for each point we need to find LDOF value. To overcome this problem author first apply k-mean clustering algorithm to partition dataset and prune the cluster having less number of point. Then apply LDOF on rest of point in remaining cluster. And detect point as outlier whose LDOF value is high.

Juntao Wang and Xiaolong Su [2] used density based outlier detection technique to improve the accuracy of k-mean clustering algorithm. They remove the abnormal point from dataset by using LOF of point, if lof of point is large than 1 than that point is remove. After this process a new set of normal point is generated. And author determines the initial center as calculating mean of new dataset. And then perform the simple k-mean algorithm until no changes in center.

Vijay Kumar, Sunil Kumar, and Ajay Kumar Singh [6] first used PAM clustering algorithm to cluster dataset and then apply distance based approach to find outlier. For distance based approach used threshold parameter as absolute distance between point and medoid multiply by 1.5. Points having greater value than this threshold are considered as outlier.

Ms. S. D. Pachgade and Ms. S. S. Dhande [7] first define the problem of distance based outlier detection approach that it required increase number of pair wise distance calculation. For this solution they first apply k-mean clustering method and then apply distance based approach. For finding outliers, threshold value is taken from the user.

In our research work, we discussed about existing methodology of outlier detection over dataset using cluster-based and distance-based approach [7], in that they select initial center randomly in k-mean clustering and threshold parameter is taken from user.

3. METHODOLOGY

The Proposed Method use hybrid approach which includes Cluster Based and Distance Based approaches. Proposed method used clustering k-mean algorithm with dynamically selection of initial center instead of random selection of center. And generate dynamic threshold value instead of taken from user as in existing algorithm.

3.1 Cluster Based Approach:

Clustering process first partition the dataset into group based on similarity and then assigns labels to number of groups. Clustering is unsupervised technique that means there no need of prior knowledge of data. Different Clustering methods available like partition-based, hierarchical, density-based, and model-based. Where partition based method partition the dataset D of n objects into a set of k clusters. Partition based algorithm are k-means and k-medoids. In k-means each cluster are represented by the center of the cluster. The variant of k-means algorithm is k-modes, which cluster categorical data by replacing mean of cluster with modes. One of the most popular and widely studied clustering methods for objects in Euclidean space is called k-means clustering.

Basic k-mean Algorithm:
Input: Random selection of number of clusters K and data set D having n objects.
Output: A set of k clusters.
Method:

1. Arbitrarily select k initial cluster center from dataset;
2. Repeat, assign each data point to the cluster that are most similar according to the mean value of data point in the cluster;
3. Update cluster by calculating mean value of objects for each cluster;
4. Repeat Until no change in centroid.

3.2 Distance Based Approach:

There are different outlier detection approaches available such as statistic based, distance based, and density based and deviation based outlier detection approach. In which distance based approach introduces to overcome the main disadvantage of statistical approach, which is this approach work with multi-dimensional analysis. A distance based outlier can be define as, Object in the dataset D, is outlier if object lie at a distance greater than the some distance parameter from neighbor. To find Distance between points with its neighbor, the different dissimilarity measure used are Euclidean distance, cosine distance, city block distance, etc. This approach does not require any a priori knowledge of data distributions as the statistics methods. But in this approach need to define the threshold parameter.

4. PROPOSED APPROACH

Flow of Proposed Work:

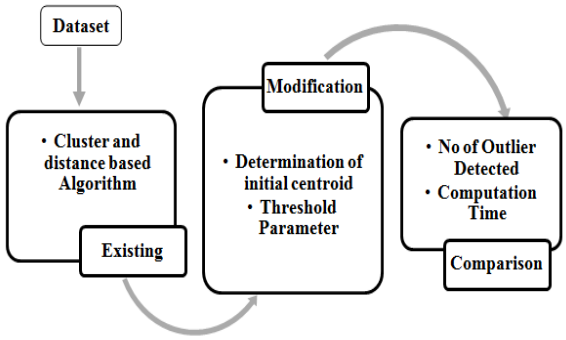

Fig 1: System Architecture

System architecture start with take dataset, next implement existing algorithm, then modify existing algorithm by determine initial centroid and calculating threshold at last compare both existing and proposed algorithm with number of outlier detected and computation time.

Steps of proposed algorithm:

```
Step I: Input dataset D, Initialize number of cluster K
Step II: Determine initial centroid from dataset
    A: Set m=1; calculate the distance between each
    object data and all other object data in D.
    B: Find closest pair distance of object from
    dataset D, form a dataset Sm, add closest point in
    set Sm and delete it from D
    C : If m<K, then increment m as m=m+1, find
    closest pair of object data whose distance is
    shorter, to form another set Sm
    D : For each set Sm, find the mean of data point in
    Sm, and mean will be the initial centroid.
Step III: Calculate the distance between each object
data and each center of cluster, and then assign each
data point to the closest cluster.
Step IV: Repeat until No changes in centroid.
Step V: Calculate the distance of each object data of
cluster from cluster centroid.
Step VI: Take threshold T value as Average (ADPC).
Step VII: If Distance > T than point is declare as
"Outlier".
```

First four steps are of clustering approach and last three are of distance based approach. Second step contain again four steps that are of selecting initial centroid. Here in distance based approach threshold value is taken as average of absolute distance between point and cluster centroid. And the distance having value is greater than threshold value are declared as outlier.

5. EXPERIMENTAL RESULT

For Experiment result I used MATLAB tools for implementing algorithm. All experiment on windows 7 with Intel CORE TMi3, 2.53 GHz with 3 GB RAM. Experiment conducted in MATLAB 7.8.0 (R2009a) on different dataset.

1) Bupa Liver Disorder Dataset: Taken from UCI machine learning repository. It contains 345 No of instance with 7 no of attribute, and is in .xls file format.

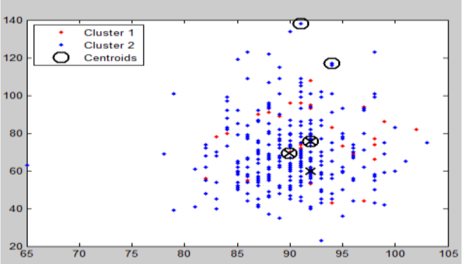

Fig.2 No of Outlier Detected in Existing Method

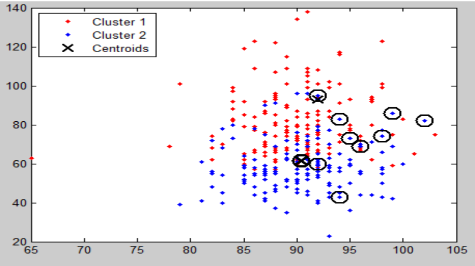

Fig.3 No of Outlier Detected in Proposed Method

Table 1: Comparative analysis of no of outlier detected on BUPA Liver dataset

Method		Existing Algorithm (T=75%)	Proposed Algorithm
Number Of Outliers	Cluster 1	1	1
	Cluster 2	2	9
	Total	3	10

Table 2: Computation time in second on BUPA Liver dataset

Elapsed Time (Second)	
Existing Algorithm	Proposed Algorithm
1.039322	0.921767

2) Breast Cancer Dataset: Taken from UCI machine learning repository. It contains 569 No of instance with 32 no of attribute, and is in .xls file format.

Table 3: Comparative analysis of no of outlier detected on Breast Cancer dataset

Method		Existing Algorithm (T=75%)	Proposed Algorithm
Number Of Outliers	Cluster 1	2	12
	Cluster 2	14	15
	Total	16	27

Table 4: Computation time in second on Breast Cancer dataset

Elapsed Time (Second)	
Existing Algorithm	Proposed Algorithm
2.109881	2.098539

6. CONCLUSION

This paper discusses about the concept of outlier, clustering, and outlier detection approaches. There are different approaches available for outlier detection. For this task clustering is also played an important role in data mining. From the literature review it is found that k-means algorithm is most widely used. The advantage of K-mean algorithm is easy to implement and unsupervised learning method. But K-mean has some limitation are initialize number of cluster and random selection of initial centroid. Here Existing algorithm is modified by determining initial centroide instead of random selection and by taking threshold value dynamically. Experiment result concludes that outlier detection of the proposed algorithm in dataset has improved over existing algorithm. And proposed method take less computation time.

7. REFERENCES

[1] Dantong Yu*, Gholamhosein Sheikholeslami and Aidong Zhang, *"FindOut: Finding Outliers in Very Large Datasets"*, Knowledge and Information Systems, 31 May 2001,pp. 387-412

[2] Juntao Wang, Xiaolong Su, *"An improved K-Means clustering algorithm"*, 2011, IEEE, pp.44-46

[3] Janpreet Singh, Shruti Aggarwal, *"Survey on Outlier Detection in Data Mining"*, International Journal of Computer Application, (0975 – 8887) Volume 67– No.19,April 2013

[4] Karanjit Singh and Dr. Shuchita Upadhyaya, *"Outlier Detection: Applications And Techniques"*, International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012,pp.307-323

[5] S. Vijayarani, S. Nithya, *"An Efficient Clustering Algorithm For Outlier Detection"*, International Journal of Computer Application, (0975 – 8887) Volume 32– No.7,October 2011

[6] Vijay Kumar, Sunil Kumar, Ajay Kumar Singh, *"Outlier Detection: A Clustering-Based Approach"*, International Journal of Science and Modern Engineering, Volume-1, Isaue-7, June 2013, pp.16-19

[7] Ms. S. D. Pachgade, Ms. S. S. Dhande, *"Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach"*, International Journal of Advance Research in Computer science and Software Engineering, Volume 2, Issue 6,June 2012,pp.12-16

[9] Jingke Xi, *"Outlier Detection Algorithm in Data Mining"*, Second International Symposium on Intelligent Information Technology Application, 2008 IEEE, pp.94-97

[10] RajendraPamula, Jatindra Kumar Deka, Sukumar Nandi, *"An Outlier Detection Method based on Clustering"*, Second International Conference on Emerging Applications of Information Technology, 2011 IEEE, pp.253-256