An Approach of Modified ECLARANS for Efficient Outlier Detection

[1]Monika Kanojiya, [2]Prateek Gupta
[1,2]Shriram Institute of Science and Technology Jabalpur, MP, India

Abstract
There are several techniques and algorithms are used for extracting the hidden patterns from the large data sets and finding the relationships between them. Clustering is one of the important techniques in data mining. Clustering algorithms are used for grouping the data items based on their similarity the goal of clustering is to group sets of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate clusters. Outlier Detection is a very important research problem in data mining. Clustering algorithms are used for detecting the outliers efficiently The algorithms used in this research work are PAM (Partitioning around Medoid), CLARA (Clustering Large Applications) AND CLARANS (Clustering Large Applications Based on Randomized Search) and a new clustering algorithm ENHANCED CLARANS for detecting outliers. In order to find the best clustering algorithm for outlier detection several performance measures are used. The experimental results show that the outlier detection accuracy is very good in the ECLARANS clustering algorithm compared to the existing algorithms. It has a very high accuracy but still it takes time to be accurate. So by this research work this can also be done. The aim of this research is to reduce the time complexity of the ECLARANS.

Keywords
Data mining, Clustering, Outlier Detection

I. Introduction

Data mining is the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a relatively young and interdisciplinary field of computer science, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Now, data mining is becoming an important technique to convert the data into valuable information. It is commonly used in a wide series of profiling practices, such as marketing, fraud detection and scientific discovery. Data mining is the method of extracting patterns from data. The mining process will be ineffective if the samples are not good representation of the larger body of the data. Therefore, it is important to detect whether the extracted is either useful to us or not. Outliers are the set of objects that are considerably dissimilar from the remainder of the data. Outlier detection is an extremely important problem with a direct application in a wide variety of application domains, including fraud detection, identifying computer network intrusions and bottlenecks, criminal activities in e-commerce and detecting suspicious activities. Different approaches have been proposed to detect outliers.

Outlier Detection

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. A failure to detect outliers or their ineffective handling can have serious impact on the strength of the inferences drained from the exercise. There are large number of techniques are available to perform this task, since there is no standard algorithm exist for detecting it.

Activities 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. The methodology for the application of clustering methods to the task of outlier detection. The methodology is tested on the problem of cleaning official statistics data.

Indeed, for cluster analysis to work effectively, some key issues are there like, whether there exists a natural notion of similarities among the "objects" to be clustered, whether clustering a large number of objects can be efficiently carried out. Traditional cluster analysis algorithms are not designed for large data sets, with say more than 1,000 objects.

Till now some of the algorithms which are being used for outlier detection are-

- PAM(Partitioning around medoids)
- CLARA(Clustering large applications)
- CLARANS(Clustering large applications by randomized search)
- ECLARANS(Enhanced Clarans)

A. PAM (Partitioning Around Medoids)

PAM (Partitioning Around Medoids) was developed by Kaufman and Rousseeuw . To find k clusters, PAM'sapproach is to determine a representative object for each cluster.This representative object, called a medoid, is meantto be the most centrally located object within the cluster.Once the medoids have been selected, each nonselected object is grouped with the medoid to which it is the most similar.

Procedure-

1. Input the dataset D
2. Randomly select k objects from the dataset D
3. Calculate the Total cost T for each pair of selected Si and non-selected object Sh
4. For each pair if T si < 0, then it is replaced Sh
5. Then find similar medoid for each non-selected object 6. Repeat steps 2, 3 and 4, until find the medoids.

B. Clustering Algorithm Based on Randomized Search

It gives higher quality clusterings than CLARA, and CLARANS requires a very small number of searches. We now present the details of Algorithm CLARANS.

Procedure of CLARANS-

1. Input parameters numlocal and maxneighbor. Initialize i to 1, and mincost to a large number.
2. Set current to an arbitrary node in n:k.
3. Set j to 1.

4. Consider a random neighbor S of current, and based on 5, calculate the cost differential of the two nodes.
5. If S has a lower cost, set current to S, and go to Step 3.
6. Otherwise, increment j by 1. If j maxneighbor, go to Step 4.
7. Otherwise, when j > maxneighbor, compare the cost of current with mincost. If the former is less than mincost, set mincost to the cost of current and set bestnode to current.
8. Increment i by 1. If i > numlocal, output bestnode and halt. Otherwise, go to Step 2. Steps 3 to 6 above search for nodes with progressively lower costs. But, if the current node has already been compared with the maximum number of the neighbors of the node (specified by maxneighbor) and is still of the lowest cost, the current node is declared to be a "local" minimum. Then, in Step 7, the cost of this local minimum is compared with the lowest cost obtained so far. The lower of the two costs above is stored in mincost. Algorithm CLARANS then repeats to search for other local minima, until numlocal of them have been found.

As shown above, CLARANS has two parameters: the maximum number of neighbors examined (maxneighbor) and the number of local minima obtained (numlocal). The higher the value of maxneighbor, the closer is CLARANS to PAM, and the longer is each search of a local minima. But, the quality of such a local minima is higher and fewer local minima needs to be obtained.

ECLARANS (Enhanced Clarans)

1. Input parameters numlocal and maxneighbour. Initialize i to 1, and mincost to a large number.
2. Calculating distance between each data points
3. Choose n maximum distance data points
4. Set current to an arbitrary node in n: k
5. Set j to 1.
6. Consider a random neighbor S of current, and based on 6, calculate the cost differential of the two nodes.
7. If S has a lower cost, set current to S, and go to Step 5.
8. Otherwise, increment j by 1. If j maxneighbour, go to Step 6.
9. Otherwise, when j > maxneighbour, compare the cost of current with mincost. If the former is less than mincost, set mincost to the cost of current and set best node to current.
10. Increment i by 1. If i > numlocal, output best node and halt. Otherwise, go to Step 4.

The goal of this research is the detection of outliers with high accuracy and time efficiency. The methodology discussed here is able to save a large amount of time by selecting a small subset of suspicious transactions for manual inspection which includes most of the erroneous transactions.

II. Literature Survey/Brief Review

According to Data Mining concepts and Techniques by Jiawai Han and Micheline Kamber clustering algorithm partition the dataset into optimal number of clusters. Small clusters are then determined and considered as outliers. The rest of the outliers (if any) are then detected in the remaining clusters based on temporary removing an edge (Euclidean distance between objects) from the data set and recalculate the weight function. They introduce a new cluster validation criterion based on the geometric property of data partition of the dataset in order to find the proper number of clusters. The algorithm works in two stages. The first stage of the algorithm creates optimal number of clusters, whereas the second stage of the algorithm detects outliers.

According to paper [3], a new efficient method for outlier detection is proposed. The proposed method is based on fuzzy clustering techniques. The c-means algorithm is first performed, and then small clusters are determined and considered as outlier clusters. Other outliers are the determined based on computing differences between objective function values when points are temporarily removed from the data set.

By the paper [5] some of clustering algorithms are PAM, CLARA AND CLARANS and a new clustering algorithm ECLARANS is proposed for detecting outliers. In order to find the best clustering algorithm for outlier detection several performance measures are used.

In this paper, a new proposed method based on clustering algorithms for outlier detection is proposed.

By paper [1] we first perform the PAM clustering algorithm. Small clusters are then determined and considered as outlier clusters. The rest of outliers are then found (if any) in the remaining clusters based on calculating the absolute distances between the medoid of the current cluster and each of the points in the same cluster. The test results show that the proposed approach gave effective results when applied to different data sets.

The key feature of their algorithm is it finds noise-free/error-free clusters for a given dataset without using any input parameters. The research proposes a method based on clustering approaches for outlier detection. They first perform the PAM clustering algorithm in that, small clusters are detected in the remaining clusters based on calculating the absolute distances between the results show that their method works well. The experimental results show that the proposed approaches give effective results when applied to different data sets. The research discusses outlier detection algorithms used in data mining system. Fundamental approaches currently used for solving this problem are considered, and their advantages and disadvantages are discussed. A new outlier detection algorithm is recommended. It is based on methods of fuzzy set theory and the use of kernel functions and possesses a number of advantages compared to the existing methods. The presentation of the algorithm suggested is studied by the example of the applied problem of anomaly finding arising in computer security systems, the so-called intrusion detection systems.PAM,CLARA algorithms describe about the outlier detection. It is a primary step in many data-mining applications. They present several methods for outlier detection, while distinguishing between Univariate and multivariate techniques and parametric vs. nonparametric procedures. In presence of outliers, special concentration should be taken to assure the strength of the used estimators. Outlier detection for data mining is repeatedly based on distance measures, clustering and spatial methods. There compared three partition based algorithms with k-medoid distance based method for outlier detection. Here they improve the time efficiency and accuracy of detection. The main advantages of all these approaches is that they are all Unsupervised methods, which means new data can be added to the database can be tested for outliers in future in an efficient manner. Experiments showed that CLARANS is the best algorithm while considering outlier detection, followed by PAM and CLARA.

Our proposed performance correction in the algorithm experimental result improves the accuracy of detection along with the time efficiency. while compared with the existing approach results.

III. Proposed Solution

The goal of this research work is to study about different outlier detection algorithms and extract a method for detection of outliers

with high accuracy and time efficiency by changing the existing algorithm of ECLARANS. The proposed methodology is able to save a large amount of time by selecting the maximum distance data points by calculating cost between each data point and taking the highest distance data points rather than taking arbitrary data points.

IV. Implementation

This research has been implemented using Net Beans IDE version 1.7.3. For calculation and comparison "WHO "database (with 8000 objects) has taken.

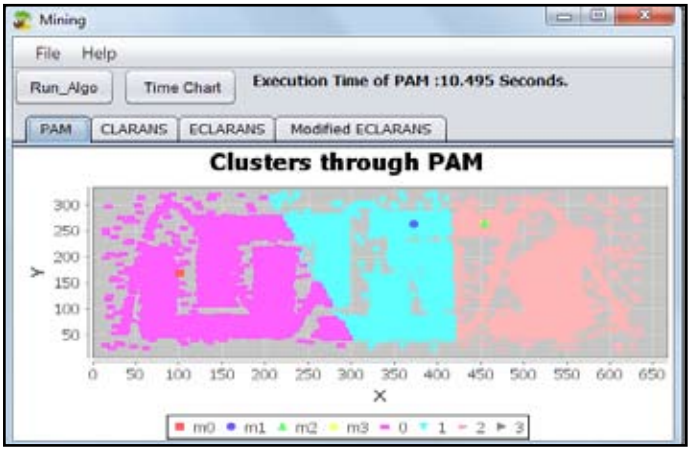

Fig. 1:

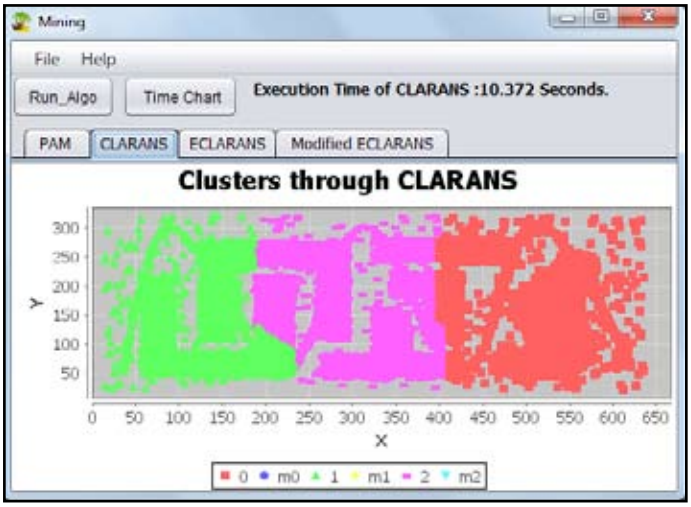

Fig. 2:

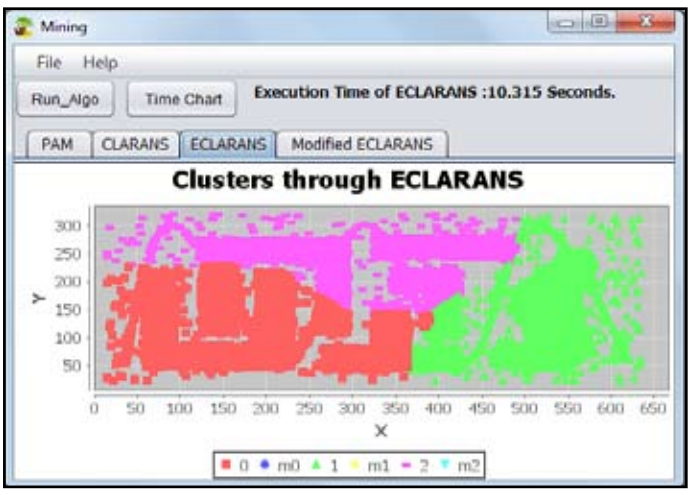

Fig. 3:

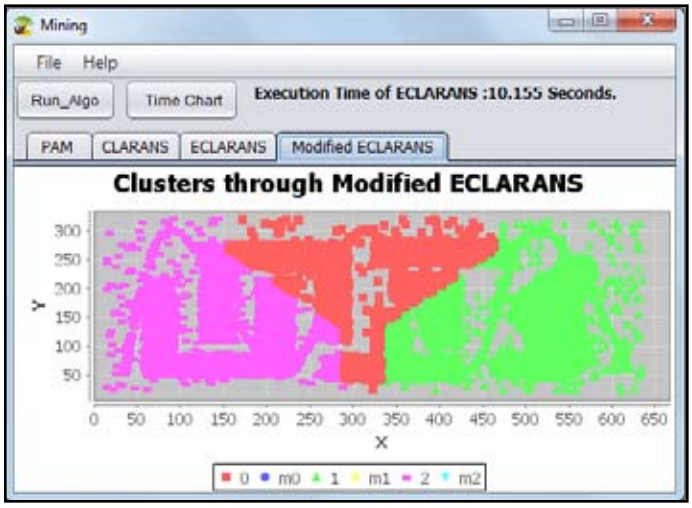

Fig. 4:

V. Comparison Study

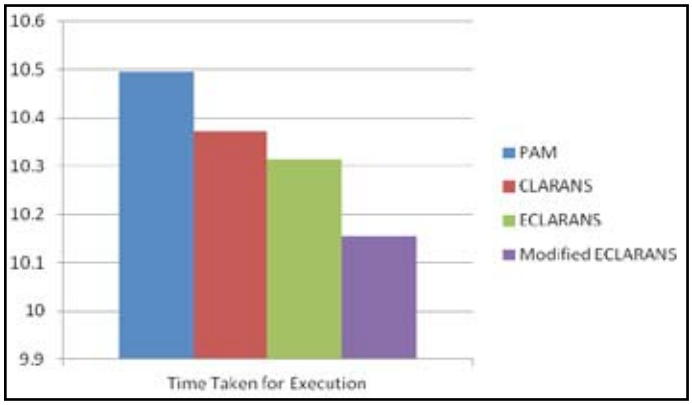

Fig. 4:

Table 1:

NAME OF THE ALGORI-THM	PAM	CLARANS	E-CLARNS	Modifi-ed Eclarans
Time Taken for Execution(in Sec)	10.495	10.372	10.355	10.155

VI. Conclusion

As we have seen the previously existing algorithms for outlier detection, it was concluded that CLARANS was better than PAM for outlier detection and ECLARANS is best in terms of accuracy but not in terms of time efficiency for this we have changed some factors of the algorithm by changing the approach of implementation. This work has been completed and the time efficiency of ECLARANS has been improved by this research work.

References

[1] Al-Zoubi M.,"An Effective Clustering-Based Approach for Outlier Detection, European Journal of Scientific Research, 2009.

[2] Deepak Soni, Asst. Prof Naveen Jha, Deepak Sinwar, "Discovery of Outlier from Database using different Clustering Algorithms, Indian J. Edu. Inf. Manage., Vol. 1, No. 6, 2012.

[3] Moh'd Belal Al-Zoubi, Ali Al-Dahoud, Abdelfatah A. Yahya,"New Outlier Detection Method Based on Fuzzy

Clustering", WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS

[4] Raymond T. Ng, Jiawei Han, Member, IEEE Computer Society, CLARANS: A Method for Clustering Objects for Spatial Data Mining(2002), IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 14, No. 5, SEPTEMBER/OCTOBER 1003

[5] Vijayarani S, Nithya S,"An Efficient Clustering Algorithm for Outlier Detection", International Journal of Computer Applications.

Monika Kanojiya, M.Tech (CSE) student of Shriram Institute Of Science And Technology, Jabalpur, MP, India.