An Efficient Clustering Algorithm for Outlier Detection

S.Vijayarani
Assistant Professor,
School of Computer Science and Engg.
Bharathiar University
Coimbatore

S.Nithya
M.Phil Research Scholar
School of Computer Science and Engg.
Bharathiar University,
Coimbatore

ABSTRACT
With the help of data mining, an important and valuable knowledge is extracted from the large massive collection of data. 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. Outlier Detection is a very important research problem in data mining. Clustering algorithms are used for detecting the outliers efficiently. In this research paper, we focused on outlier detection in health data sets such as Pima Indians Diabetes data set and Breast Cancer Wisconsin data set using partitioning clustering algorithms. The algorithms used in this research work 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. The experimental results show that the outlier detection accuracy is very good in the proposed ECLARANS clustering algorithm compared to the existing algorithms.

General Terms
Data mining, Clustering, Outlier detection

Keywords
Data Mining, Clustering, PAM, CLARA, CLARANS and ECLARANS, Outlier Detection.

1. INTRODUCTION
Data mining is the non-trivail method of identifying valid, novel, potentially useful, and finally understandable patterns in data [1]. Now, data mining is becoming an important tool to convert the data into 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. It can be used to uncover patterns in data but is often carried out only on sample of data. The mining process will be ineffective if the samples are not good representation of the larger body of the data. The discovery of a particular pattern in a particular set of data does not necessarily mean that pattern is found elsewhere in the larger data from which that sample was drawn. An important part of the method is the verification and validation of patterns on other samples of data. A primary reason for using data mining is to assist in the analysis of collection of observations of behaviour.

Cluster analysis or clustering is the assignment of a set of observations into subsets called clusters so that observations in the same clusters are similar in some sense. It is a useful technique for the discovery of data distribution and patterns in the original data. The goal of clustering technique is to find out both the dense and the sparse region in a data set. It is a method of unsupervised learning and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. It is an important technique used for outlier analysis. Outlier detection based on clustering approach provides new positive results. Clustering algorithms are used for outlier detection, where outliers (values that are "far away" from any cluster) may be more interesting than common cases. Clustering is a challenging field of research in which its potential applications pose their own requirements.

Outliers detection is an outstanding data mining task, referred to as outlier mining. Outliers are objects that do not comply with the general behaviour of the data. By definition, outliers are rare occurrences and hence represent a small portion of the data. Outlier detection has direct applications in a wide variety of domains such as mining for anomalies to detect network intrusions, fraud detection in mobile phone industry and recently for detecting terrorism related activities [2].

The rest of the paper is organized as follows. Section 2 explains a brief discussion about the outlier detection. Section 3 provides discussion on the previous works related to the topic. Section 4 describes the existing approaches of outlier detection using PAM, CLARA AND CLARANS clustering algorithms and the proposed algorithm ECLARANS clustering algorithm for outlier detection. Conclusion and future works are given in Section 5.

2. OUTLIER DETECTION
Outlier detection is very essential of any modelling exercise. A failure to detect outliers or their ineffective handling can have serious ramifications on the strength of the inferences drained from the exercise. There are large number of techniques are available to perform this task, and often selection of the most suitable technique poses a big challenge to the practitioner. There is no standard

technique for outlier detection. Some of the outlier detection techniques are:

- Distance based outlier detection
- Clustering based outlier detection
- Density based outlier detection
- 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. All these techniques assume that all normal instances will be similar, while the anomalies will be different.

Outlier can appear about by any option in any distribution, but they are often suggestive either of measurement error or that the population has a serious-tailed distribution. In the past case one wish to dispose of them or use statics that are robust to outlier. In data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated on assumed family of probability distributions or it may be that some comments are distant from the centre of the data.

Outliers, being the most excessive observations, may include the sample minimum or sample maximum or both depending on whether they extremely high or low. However the sample minimum or sample maximum is not always outliers because they may not be abnormally distant from other comments. Many statistical techniques are sensitive to the occurrence of outliers. Checking for outliers should be a usual part of any data analysis.

3. RELATED WORKS

The paper [6] describes a methodology for the application of hierarchical clustering methods to the task of outlier detection. The methodology is tested on the problem of cleaning official statistics data. The goal of this paper is the detection of erroneous foreign trade transactions in data collection. 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. The authors compared several alternative hierarchical clustering methodologies for this task. The results they have obtained here confirmed the validity of the use of hierarchical clustering techniques for this task. Their comparison results show that their methodology improves previous results by keeping similar number of erroneous transactions identified with significantly.

The paper [4] generalizes local outlier factor of object and proposed a clustering-based outlier detection scheme (CBOD). The method consists of two phases, the first phase cluster dataset by one-pass clustering algorithm and second phase decide outlier cluster by outlier factor. The time difficulty of CBOD is almost linear with the size of dataset and the number of attributes, which results in good scalability and proper to large dataset. The theoretic study and the experimental results show that the detection process is effective and feasible.

The paper [5] discussed about the Minimum Spanning Tree based clustering algorithm for detecting outliers. They mentioned Minimum Spanning Tree based clustering algorithm is capable of detecting clusters with irregular boundaries .The 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, where as the second stage of the algorithm detect outliers. The key feature of their algorithm is it finds noise-free/error-free clusters for a given dataset without using any input parameters.

The paper [3] 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 paper [10] 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.

The paper [11] describes 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.

The paper [7] 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 CLARA and PAM.

In this research work, we have discussed about the existing work of Improved Hybrid Clustering and Distance Based Technique for Outlier Removal [7] and we have proposed a new clustering algorithm for the detection of outliers. Our proposed algorithm experimental result improves the accuracy of detection while compared with the existing approach results.

4. PROBLEM DEFINITION & THE PROPOSED METHODOLOGY

Outlier detection is a very important research work in the field of data mining. The problem statement of this research work is to find out the outliers using different clustering algorithms and to verify the performance of the clustering algorithms.

SYSTEM ARCHITECTURE

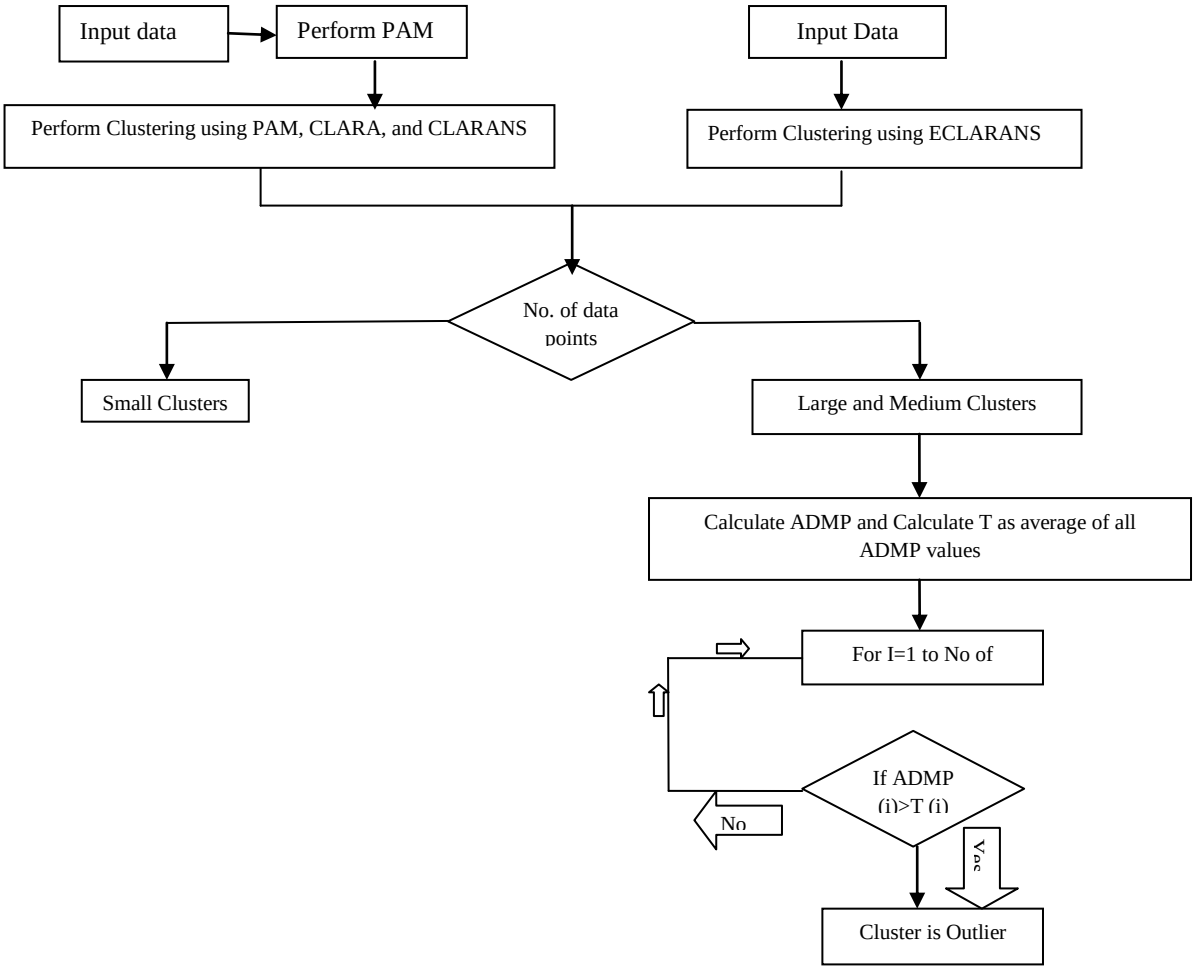

Fig1. System Architecture

Proposed Solution

1. Outlier Detection

   1.1 Existing Clustering Algorithm
   PAM
   CLARA
   CLARANS
   1.2 Proposed Clustering Algorithm
   ECLARANS
2. Performance Factors
   2.1 Outlier Accuracy
   2.2 Time Complexity
3. Best Clustering Algorithm for Outlier Detection

1.1 Outlier Detection

Outliers detection is an outstanding data mining task, referred to as outlier mining. Outliers are objects that do not comply with the general behaviour of the data. By definition, outliers are rare occurrences and hence represent a small portion of the data. Murugavel. P. et al [13] the algorithm first performs clustering using one of the algorithms PAM,CLARA,CLARANS and ECLARANS .The algorithm produces a set of clusters and a set of medoids (cluster centers). In the next step, the average number of points in 'k' cluster is calculated (AKN) and the clusters are segregated as small and large clusters. All those clusters which have less than half of AKN are declared as small cluster. These small clusters are removed from the datasets as outliers or noise. The outliers in the large clusters are then detected using the following procedure. First, the Absolute Distances between the Medoid (μ) (ADMP) of the current cluster and each one of the points (pi) is calculated using Equation 2. A threshold value is calculated as the average of all ADMP values of the same cluster multiplied by 1.5. When the ADMP value of a cluster is greater than T, then it is an outlier, else it is an inlier.

1.1 Existing Clustering Algorithm

PAM (Partitioning around Medoid): PAM uses a k-medoid method for clustering. It is very robust when compared with k-means in the presence of noise and outliers. Mainly in contains two phases Build phase and Swap phase [1].

Build phase: This step is sequentially select k objects which is centrally located. This k objects to be used as k-medoids.

Swap phase: Calculates the total cost for each pair of selected and non-selected object.

PAM 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.

CLARA (Clustering Large Applications): CLARA is introduced to overcome the problem of PAM. This works in larger data set than PAM. This method takes only a sample of data from the data set instead of taking full data set. It randomly selects the data and chooses the medoid using PAM algorithm [1].

CLARA Procedure:

1. Input the dataset D
2. Repeat n times
3. Draw sample S randomly from D
4. Call PAM from S to get medoids M.
5. Classify the entire dataset D to Cost1.....costk
6. Calculate the average dissimilarity from the obtained clusters

*CLARANS (*Clustering Large Applications Based on Randomized Search): This method is similar to PAM and CLARA. It starts with the selection of medoids randomly. It draws the neighbour dynamically. It checks "maxneighbour" for swapping. If the pair is negative then it chooses another medoid set. Otherwise it chooses current selection of medoids as local optimum and restarts with the new selection of medoids randomly. It stops the process until returns the best.

CLARANS Procedure:

1. Input parameters numlocal and maxneighbour.
2. Select k objects from the database object D randomly.
3. Mark these K objects as selected Si and all other as non-selected Sh.
4. Calculate the cost T for selected Si
5. If T is negative update medoid set. Otherwise selected medoid chosen as local optimum.
6. Restart the selection of another set of medoid and find another local optimum.
7. CLARANS stops until returns the best.

1.2 Proposed Clustering Algorithm

ENHANCED CLARANS (ECLARANS): This method is different from PAM, CLARA AND CLARANS.Thus method is produced to improve the accuracy of outliers. ECLARANS is a new partitioning algorithm which is an improvement of CLARANS to form clusters with selecting proper arbitrary nodes instead of selecting as random searching operations. The algorithm is similar to CLARANS but these selected arbitrary nodes reduce the number of iterations of CLARANS

ECLARANS Procedure

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

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.

2. Performance Factors

The performance of clustering algorithms is presented in this section. Two Health Datasets namely Pima Indians Diabetes Data Set with 8 attributes, 768 instances and Breast Cancer Wisconsin Dataset with 10 attributes, 699 instances collected from http*://archive.ics.uci.edu/ml/datasets.html* data sets. Table1 shows the results of these algorithms for the accuracy of detection of outliers.

2.1 Outlier Accuracy

Outlier detection accuracy is calculated, in order to find out more number of outliers detected by the existing clustering algorithms PAM, CLARA, CLARANS and the proposed clustering algorithm ECLARANS.

2.1.1Table1 Number of outliers detected.

DATASET	PAM	CLARA	CLARANS	ECLARANS
PIMA INDIAN DIABETES	6	38	219	286
WISCONSIN BREAST CANCER	109	130	314	319

It could be seen that the ECLARANS method is better when compared with PAM, CLARA and CLARANS. In Pima, PAM detected 6 outliers, CLARA detected 38 outliers, CLARANS detected 219 outliers, and ECLARANS detected 286 outliers, the same algorithms are used for detecting outliers in cancer dataset is also shown in Table1.Thus it could be shows that the ECLARANS algorithm improves the accuracy of detecting the outliers. We have implemented all our algorithms in MATLAB 7.10(R2010a)

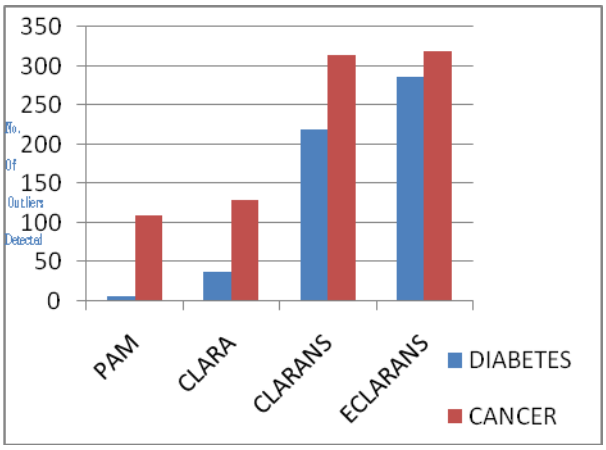

Fig2: Outlier Accuracy

The chart shows that the number of outliers detected by the existing clustering algorithms PAM, CLARA, CLARANS and the proposed clustering algorithm ECLARANS. The new proposed clustering technique ECLARANS has detected more number of outliers compared to the existing techniques.

2.2. Time Complexity

The Time complexity performance factor is measured in terms of the time required for detecting the outliers by the existing clustering algorithms PAM, CLARA, CLARANS and the proposed clustering algorithm ECLARANS.

2.2.1Table II Time Efficiency

DATASETS	PAM	CLARA	CLARANS	ECLARANS
PIMA INDIAN DIABETES	238.92	269.41	3.73	30.97
WISCONSIN BREAST CANCER	374.52	206.32	3.47	16.06

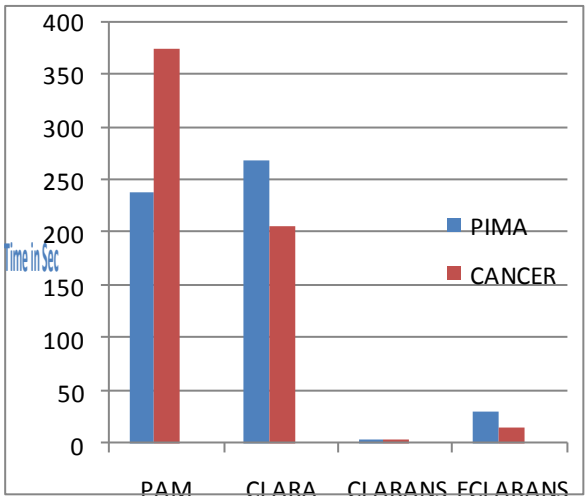

Fig3: Time Complexity

Comparing the performance of time complexity of the existing and the proposed clustering algorithms the CLARANS algorithm has taken less time.

1.3 Best Clustering Algorithm for Outlier Detection

The results of the existing algorithms and the proposed algorithm for outliers detection is shown above. From the experimental results we found that the outlier accuracy is high in the proposed ECLARANS algorithm and the time requirement is very less in CLARANS algorithm.

5. CONCLUSION

Data mining is the method of extracting patterns from data. In data mining, clustering is the process of grouping the data that have high similarity in comparison to one another. In this paper, we have discussed about the different clustering techniques for outlier detection. We have proposed a new methodology for outlier detection. The experimental result shows that our algorithm ECLARANS improves the accuracy of detection and CLARANS reduces the time complexity when compared with other algorithms. Further work also lies in this application. We will use this detection of outliers for our future work and plan to reduce the time complexity of the proposed algorithm.

6. REFERENCES

[1] Arun K Pujari: Data Mining Techniques, Universities Press (India) Private Limited 2001.

[2] Ajay Challagalla,S.S.Shivaji Dhiraj ,D.V.L.N Somayajulu,Toms Shaji Mathew,Saurav Tiwari,Syed Sharique Ahmad " Privacy Preserving Outlier Detection Using Hierarchical Clustering Methods,2010 34th Annual IEEE Computer Software and Applications Conference Workshops.

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

[4] Jiang, S. And An, Q. (2008) Clustering Based Outlier Detection Method, Fifth International Conference on Fuzzy Systems and Knowledge Discovery.

[5] John Peter.S., Department of computer science and research center St.Xavier's College, Palayamkottai, An Efficient Algorithm for Local Outlier Detection Using Minimum Spanning Tree, International Journal of Research and Reviews in Computer Science (IJRRCS), March 2011.

[6] Loureiro, A., Torgo, L. And Soares, C. (2004) Outlier Detection using Clustering Methods: A Data Cleaning Application, in Proceedings of KDNet Symposium on Knowledge-Based Systems for the public Sector. Bonn, Germany.

[7] Murugavel. P. et al, Improved Hybrid Clustering And Distance-Based Technique for Outlier Removal, International Journal on Computer Science and Engineering (IJCSE), 1 JAN 2011

[8]Ng, R. and Han, J. (1994) Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Conf.

[9]Ng, R. and Han, J. (2002) CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE Transactions on Knowledge and Data Engineering.

[10]Outlier Detection Algorithms in Data Mining Systems. I. Petrovskiy, Department of Computational Mathematics and Cybernetics, Moscow State University, Vorob'evy gory, Moscow, 119992 Russia.e-mail: michael@cs.msu.suReceived February 19, 2003.

[11]OUTLIER DETECTION, Irad Ben-Gal**,** Department of Industrial Engineering**,**Tel-Aviv University**,**Ramat-Aviv, Tel-Aviv 69978, Israel.**,**bengal@eng.tau.ac.il.

[12]Velmurugan, T. and Santhanam, T. (2011) A survey of partition based clustering algorithms in data mining: An experimental approach, Inform. Technol. J.,