ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2, Issue 3 (July - Sept. 2014)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Outlier Detection Using High Dimensional Dataset for Comparision of Clustering Algorithms

[I]P. Sudha, [II]K. Krithigadevi
[I]Assistant Professor, Dept. of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi.
[II]Research Scholar, Sree Saraswathi Thyagaraja College, Pollachi.

Abstract
Data Mining is an extensively studied field of research area. Data mining is a mining of knowledge from large amount of data. There are lot of problems exists in large database such as data redundancy, missing data, invalid data etc., one of the major problem in data stream research area in handling high dimensional dataset. Outlier detection is a branch of Data Mining, which refers to the problem of finding objects in a large dataset that vary from other data objects. Outlier detection has been used to detect and remove unwanted data objects from large dataset. Clustering is the process of grouping a set of data objects into classes of similar data objects. The clustering techniques are highly helpful to detect the outliers so called cluster based outlier detection.
Outlier detection refers to the problem of finding patterns in data that do not conform to expected behavior. These unusual patterns are often known as outliers, anomalies, exceptions, aberrations, surprises in different domains of applications. There are three types of outlier detection approaches are used that is, the supervised anomaly detection, the semi supervised anomaly detection and the unsupervised anomaly detection.
The unsupervised anomaly detection approach is to detect anomalies in an unlabeled dataset under the assumption that the majority of the objects in the dataset are normal. This approach is applied for different kinds of outlier detection tasks and datasets.
But in this research work unsupervised anomaly detection approach is used to detect outlier in high dimensional datasets.

Keywords
Outlier, Dimensional, Data, Clustering, Detection, Mining.

I. Introduction

Outlier detection is a fundamental issue in Data Mining research areas. Outlier detection has been used to detect and remove unwanted anomalies objects from large dataset. Outlier detection is a necessary step in a variety of practical applications such as intrusion detection, health system monitoring and criminal activity detection in e-commerce and can also be used in scientific research for data analysis and knowledge discovery in the fields of chemistry, biology, astronomy etc.,
Outlier detection refers to the problem of finding patterns in data that do not conform to expected behavior. These unusual patterns are often known as outliers, anomalies, exceptions, aberrations, surprises in different domains of applications. This outlier technique can be used in a variety of applications such as credit card fraud detection, financial, weather, medical, business, social sciences, computer networks, traffic lines, insurance, fraud detection and intrusion detection. There are three types of outlier detection approaches are used that is, the supervised anomaly detection, the semi supervised anomaly detection and the unsupervised anomaly detection.
The supervised anomaly detection approach learns to classify using labeled objects belonging to normal and anomaly classes and assigns labels to test objects. In this approach K-Means clustering algorithm is supported.
The semi supervised anomaly detection approach learns to model representing normal behavior from a given dataset of normal objects and calculates the possibility of a test objects.
The unsupervised anomaly detection approach is to detect anomalies in an unlabeled dataset under the assumption that the majority of the objects in the dataset are normal. This approach is applied for different kinds of outlier detection tasks and datasets.
But in this research work unsupervised anomaly detection approach is used to detect outlier in high dimensional datasets.

II. Clustering Basics

There are several established and emerging standards related to data mining. These standards are for different components of the data mining systems. For instance, they are for,

- Models – to represent data mining and statistical data; for producing, displaying and for using the models, for analyzing and mining remote and distributed data.
- Attributes – to represent the cleaning, transforming and aggregating of attributes used as input in the models.
- Interfaces – to link to other languages and systems.
- Setting – to represent the internal parameters required for building and using the models.

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others.
The basic clustering steps are as follows,

Preprocessing and feature selection

Most clustering models assume that n-dimensional feature vectors represent all data items. This step therefore involves choosing an appropriate feature, and doing appropriate preprocessing and feature extraction on data items to measure the values of the chosen feature set. It will often be desirable to choose a subset of all the features available, to reduce the dimensionality of the problem space. This step often requires a good deal of domain knowledge

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

and data analysis.

Similarity Measures

Similarity measure plays an important role in the process of clustering where a set of objects are grouped into several clusters, so that similar objects will be in the same cluster and dissimilar ones in different cluster. In clustering, its features represent an object and the similarity relationship between objects is measured by a similarity function. This is a function, which takes two sets of data items as input, and returns as output a similarity measure between them.

Clustering Algorithm

Clustering algorithms are general schemes, which use particular similarity measures as subroutines. The particular choice of clustering algorithms depends on the desired properties of the final clustering. A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. The clustering algorithm also finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.

Result Validation

Do the results make sense? If not, we may want to iterate back to some prior stage. It may also be useful to do a test of clustering tendency, to try to guess if clusters are present at all; note that any clustering algorithm will produce some clusters regardless of whether or not natural clusters exist.

Result interpretation & application

Typical applications of clustering include data compression (via representing data samples by their cluster representative), hypothesis generation (looking for patterns in the clustering of data), hypothesis testing (e.g. verifying feature correlation or other data properties through a high degree of cluster formation), and prediction (once clusters have been formed from data and characterized, new data items can be classified by the characteristics of the cluster to which they would belong).

III. Clustering Algorithms

Categorization of clustering algorithms is neither straightforward, nor canonical. In reality, groups below overlap. For reader's convenience we provide a classification closely followed by this survey. Corresponding terms are explained below.

Clustering Algorithms

- Hierarchical Methods
- ➔ Agglomerative Algorithms.
- ➔ Divisive Algorithms.
- Partitioning Methods.
- ➔ Relocation Algorithms.
- ➔ Probabilistic Clustering.
- ➔ K-medoids Methods.
- ➔ K-means Methods.
- ➔ Density-Based Algorithms.
- - Density-Based Connectivity Clustering.
- - Density Functions Clustering.
- Grid-Based Methods.
- Methods Based on Co-Occurrence of Categorical Data.
- Constraint-Based Clustering.
- Clustering Algorithms Used in Machine Learning.
- ➔ Gradient Descent and Artificial Neural Networks.
- ➔ Evolutionary Methods.
- Scalable Clustering Algorithms.
- Algorithms For High Dimensional Data.
- ➔ Subspace Clustering.
- ➔ Projection Techniques.
- ➔ Co-Clustering Techniques.

Traditionally clustering techniques are broadly divided in hierarchical and partitioning. Hierarchical clustering is further subdivided into agglomerative and divisive. While hierarchical algorithms build clusters gradually, partitioning algorithms learn clusters directly. Partitioning algorithms of the second type are surveyed in the section Density-Based Partitioning. They try to discover dense connected components of data, which are flexible in terms of their shape. Density-based connectivity is used in the algorithms DBSCAN, OPTICS, DBCLASD, while the algorithm DENCLUE exploits space density functions. These algorithms are less sensitive to outliers and can discover clusters of irregular shapes. They usually work with low-dimensional data of numerical attributes, known as spatial data.

Some algorithms work with data indirectly by constructing summaries of data over the attribute space subsets. They perform space segmentation and then aggregate appropriate segments. Categorical data is intimately connected with transactional databases. The concept of a similarity alone is not sufficient for clustering such data. The idea of categorical data co occurrence comes to rescue. The algorithms ROCK, SNN, and CACTUS are surveyed in the section Co-Occurrence of Categorical Data.

Many other clustering techniques are developed, primarily in machine learning, that either have theoretical significance, are used traditionally outside the data mining community, or do not fit in previously outlined categories. The boundary is blurred.

Data Mining primarily works with large databases. Clustering large datasets presents scalability problems reviewed in the section Scalability and VLDB Extensions.

Clustering is a process of grouping objects with similar properties. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity. Clustering is an unsupervised learning i.e. it learns by observation rather than examples. There are no predefined class label exists for the data points. Cluster analysis is used in a number of applications such as data analysis, image processing, market analysis etc. Clustering helps in gaining, overall distribution of patterns and correlation among data objects. In this chapter describes about the general working behavior, the methodologies to be followed and the parameters which used in these clustering algorithms.

In this chapter is organized as follows gives an overview of different clustering algorithms. Then different clustering algorithms are used as follows Hierarchical clustering algorithms, K-means clustering algorithms, and Density Based Clustering Algorithm and the how the is methodology applied on these algorithms and the parameter used in these algorithms are described. Finally the conclusions are provided.

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

Hierarchical Clustering

Hierarchical clustering algorithm group's data objects to form a tree

shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, and centroid or wards method. In divisive approach all data points are considered as a single cluster and they are splited into number of clusters based on certain criteria, and this is called as top down approach.

1. Construct one cluster for each document.
2. Join the t most similar clusters.
3. Repeat 2 until a stopping criterion is reached.

K-Means Clustering Algorithm

K-means clustering is a partitioning method. *K-means* clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. The k –mean algorithm has the following important properties:

1. It is efficient in processing large data sets.
2. It often terminates at a local optimum.
3. It works only on numeric values.
4. The clusters have convex shapes.

A Novel K-Means clustering algorithm presents a method to use both advantages of HC and K-Means by introducing equivalency and compatible relation concepts. By these two concepts we defined similarity and our space and could divide our space by a specific criterion. Many directions exist to improve and extend the proposed method. Different applications can be used and examined the framework.

Density Based Clustering Algorithm

Density based algorithm continue to grow the given cluster as long as the density in the neighbourhood exceeds certain threshold [6]. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm.

1. Handles clusters of arbitrary shape.
2. Handle noise.
3. Needs only one scan of the input dataset.
4. Needs density parameters to be initialized.

The algorithms parametric the agglomerative method used in the pre-clustering step and the similarity metrics of interest. Despite its low complexity, qualitative results are very good and comparable with those obtained by state of the art clustering algorithms. Future work includes, among other topics, the investigation of similarity metrics particularly meaningful in high-dimensional spaces, exploiting summaries extracted from the regions associated to midpoints.

Model based Clustering

These methods attempt to optimize the fit between the given data and some mathematical models. Unlike conventional clustering, which identifies groups of objects; model-based clustering methods also find characteristic descriptions for each group, where each group represents a concept or class. The most frequently used induction methods are decision trees and neural networks.

Grid based Clustering

The main focus of these algorithms is spatial data, *i.e.*, data that model the geometric structure of objects in space, their relationships, properties and operations. The objective of these algorithms is to quantize the data set into a number of cells and then work with objects belonging to these cells. They do not relocate points but rather build several hierarchical levels of groups of objects. In this sense, they are closer to hierarchical algorithms but the merging of grids, and consequently clusters, does not depend on a distance measure but it is decided by a predefined parameter.

IV. Existing Methodology

This research work mainly deals to detect the outliers from high dimensional data set based on some existing clustering algorithms. The existing processes of clustering algorithms are K-Means, CLARA, CLARANS and CURE. An experiment is carried out in this research work to identify which clustering algorithm performance is good.

A. K-Means

The K-means algorithm is the best known partitioned clustering algorithm. The K-Means algorithm is a simple method for estimating the mean (vector) of set K groups. The most widely used K-Means among all clustering algorithms due to its efficiency and simplicity. The K-Means algorithm is as follows

Algorithm K-Means (k, D)
1 Chooses k data points as the initial centroids (cluster Centers)
2 Repeat
3 for each data point x ∈ D do
4 Compute the distance from x to each centered;
5 Assign x to the closest centered // a centered represents a cluster
6 end for
7 Re-compute the centered using the current cluster memberships
8 Using till the stopping criterion is met

B. Clara

CLARA stands for Clustering a Large Applications Algorithms. The focus is on clustering large number of objects rather than small number of objects in high dimensions. It works by clustering a sample from the data set and then assigns all objects in the data set. This algorithm relies on the sampling approach to handle large data sets. CLARA draws a small sample from data set and applies the PAM (Partitioning Analyzing Method) Algorithm. The CLARA Algorithm is as follows,

1. Draw a sample from the n objects and cluster it into k groups.
2. Assign each object in the dataset to the nearest group.
3. Store the average distance between the objects and their respective groups.
4. Repeat the process five times, selecting the clustering with the smallest average distance.
5. While assign a large number of objects to group.

C. Clarans

CLARANS stands for Clustering a Large Random Subset Searching Algorithm. CLARANS proceeds by searching a random subset of neighbors of a particular solution. This algorithm used two parameters of calculating solutions namely MAXneigh, the maximum number of neighbors of S to access, MAXsol, the maximum number of local solutions. The CLARANS Algorithm

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

is as follows,

1. Set S to be an arbitrary set of k representative objects.
   Set i =1
2. Set j = 1.
3. Consider a neighbour R of S at random. Calculate the total swap contribution of the two
   Neighbours.
4. If R has a lower cost, set R = S and go to Step 2.
   Otherwise increment j by one. If j ≤ MAXneigh goto Step 1.
5. When j > MAXneigh, compare the cost of S with the Best solution found so far. If the cost of
   S is less, record this cost and the representation. Increment i by one. If i > MAXsol stop,
   Otherwise go to Step 1.

D. Cure algorithm

CURE stands for **C**lustering **U**sing **Re**presentatives Algorithm. CURE is an efficient data clustering algorithm for large databases. It is processed using hierarchical methods to decompose a data set into tree like structures. It uses two clustering approaches: Partitioning clustering algorithm and Hierarchical clustering algorithm.

- When applied to Partitioning Clustering algorithm the sum of the squared errors are appeared in large differences in sizes or geometrics of different clusters.
- When applied to Hierarchical Clustering algorithm it measures the distance between (dmin, dmean) work with different shapes of clusters. So the running time is high when n is very large.

So, to avoid this problem with non uniform sized (or) shaped clusters of CURE hierarchical algorithm, the centroid points of clustering are merged at each step. This enables CURE to correctly identify the clusters and makes sensitive to outliers. The running time of the algorithm is **O (n2 log n)** and space complexity is **O (n)**. The CURE Algorithm is as follows,

CURE (no. of points, k)
Input: A set of points S
Output: k clusters
1. For every cluster u (each input point), in u.Mean and u.rep store the mean of the points in the cluster and a set of c representative points of the cluster initially c = 1 since each cluster has one data point also u, closest stores the cluster closest to u.
2. All the input points are inserted into a k-d tree T.
3. Treat each input point as separate cluster, compute u.closest for each u and then insert each
cluster into the heap Q.
4. While size(Q) > k.
5. Remove the top element of Q (say u) and merge it with its closest cluster u.closest (say v) and compute the new representative points for the merged cluster w. Also remove u and v from T and Q.
6. Also for all the clusters x in Q, update x.closest and relocate x.
7. Insert w into Q.
8. Repeat.

But this algorithm cannot be directly applied to large databases. So before applying do the following enhancements,

- Random Sampling
- Partitioning for speedup
- Labeling data on disk

To handle large data sets, do random sampling and draw a sample data set. The random sample data sets are stored in main memory. The basic idea is to partition the sample space into P partitions. The remaining data points are assigned to label data on disk. The main advantage of partitioning the input will reduce the execution time.

V. Results & Discusions

All the algorithms are implemented using MATLAB (R2010a). Two performance factors are considered namely Outlier Detection Accuracy and Clustering Accuracy for result analysis. For that two biological data sets are used one is Liver Disorders and another is Echocardiogram. These two biological data sets are very high dimensional.

A. Outlier detection accuracy

Outlier detection accuracy is calculated, in order to find out the number of outliers detected by the clustering algorithms namely K-Means, CLARA, CLARANS and CURE for Liver Disorders data set and Echocardiogram data set in three windows and five windows.

Detection rate

Detection rate refers to the ratio between the numbers of correctly detected outliers and to the total number of outliers**.** The detection rate is calculated using the formula**,**

$$d' = \frac{\mu_S - \mu_N}{\sqrt{\frac{1}{2}(\sigma_S^2 + \sigma_N^2)}}$$

The above formula provides the separation between the means of the signal and the noise distributions, compared against the standard deviation of the noise distribution. The distributed signal and noise with mean and the standard deviations are represented as μ_S ,σ_S, and μ_N andσ_N.

False alarm rate:

False alarm rate refers to the ratio between the numbers of normal objects that are misinterpreted as outlier to the total number of alarms.

The formula to calculate the false alarm rate is defined as an experiment from P positive instances and N negative instances for some condition. The four outcomes can be formulated in a 2×2 contingency table or confusion matrix, as follows:

- True positive (TP) = correctly identified
- False positive (FP) = incorrectly identified
- True negative (TN) = correctly rejected
- False negative (FN) = incorrectly rejected

These four matrix conditions values are used to calculate the outlier and clustering accuracy. The other name of false alarm rate is False Discovery Rate (FDR). The false alarm rate is calculated by using the formula,

$$FDR = FP/(TP + FP) = 1 - PPV$$

Where, False Positive (FP), True Positive (TP) and Positive Predictive (PPV) values are used to find the false alarm rate.

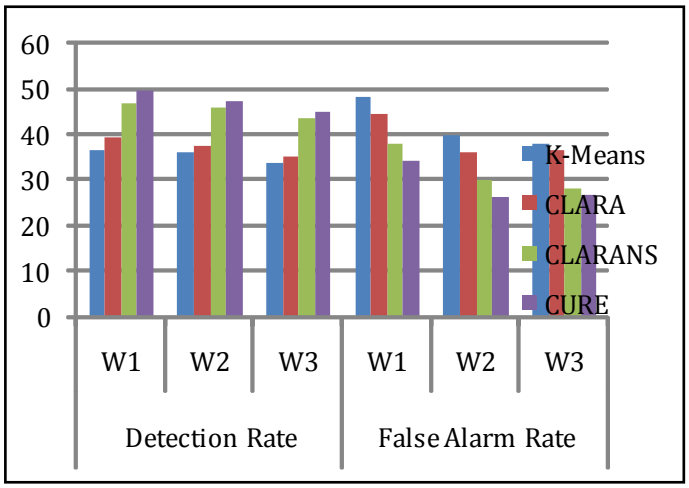

Fig. 5.1: Detection Rate and False Alarm Rate in Three Windows for Liver Disorders

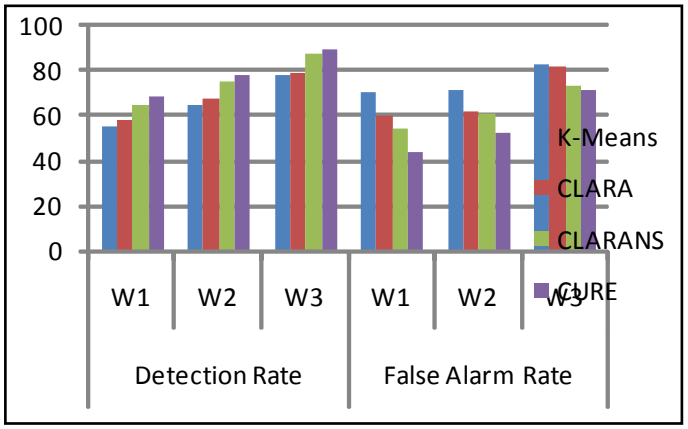

Fig. 5.2: - Detection Rate and False Alarm Rate in Three Windows for Echocardiogram

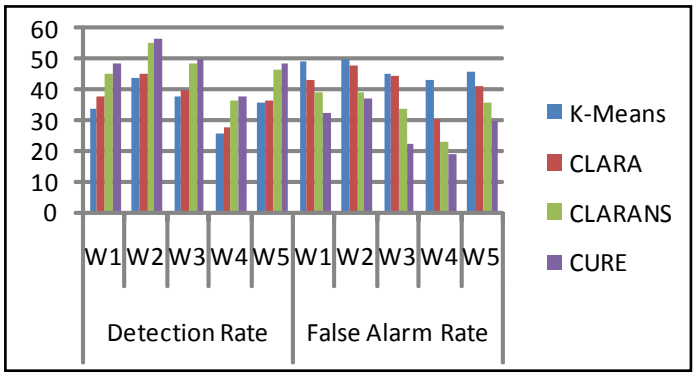

Fig. 5.3: Detection Rate and False Alarm Rate in Five Windows for Liver Disorders

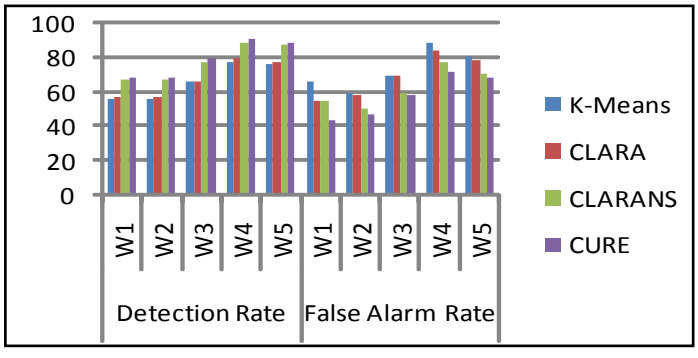

Fig. 5.4: Detection Rate and False Alarm Rate in Five Windows for Echocardiogram

Clustering accuracy is calculated by using two measures i.e., precision and recall. The clustering algorithms K-Means, CLARA, CLARANS and CURE are applied for Liver Disorders and Echocardiogram datasets to find the clustering accuracy in three windows and five windows. The clustering accuracy is calculated by the ways of Accuracy, Precision and Recall.

Accuracy

It determines how close the measurement comes to the true value of the quantity. So, it indicates the correctness of the result. Maximum effort has to be taken to acquire accuracy in data. The quality of the measurement depends on the accuracy of the entire data. It can be limited by factors like board resolution or environmental noise.

$$ACC = (TP + TN)/(P + N)$$

Where True Positive (TP), True Negative (TN), Positive (P) and Negative (N) values are used to calculate the clustering accuracy.

Precision

The other name of precision is Positive Predictive Value (PPV). The *precision* measurement reflects how exactly the result is determined without reference to what the result means. The *relative precision* indicates the uncertainty in the measurement as a fraction of the result.

$$PPV = TP/(TP + FP)$$

Where True Positive (TP) and False Positive (FP) values are used to find out the clustering accuracy of precision values.

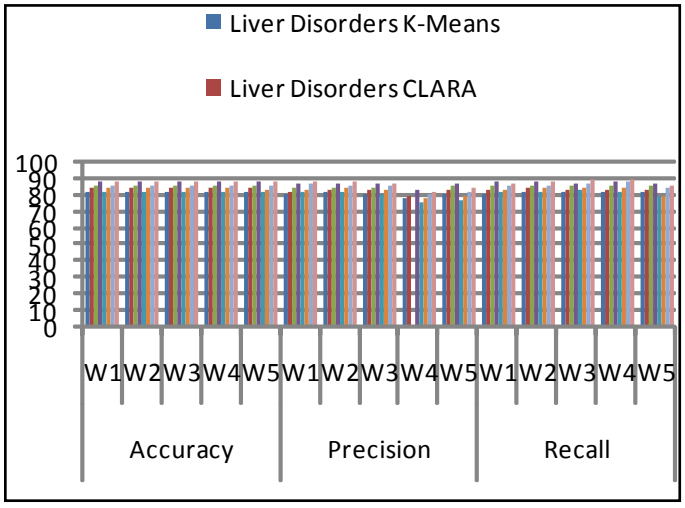

Fig. 5.5 : The Clustering Accuracy in Five Windows for Two Data Sets

From the above graph, it is observed that CURE clustering algorithm performs better than other clustering algorithms namely K-Means, CLARA and CLARANS in detecting outliers in both biological data sets such as Liver Disorders and Echocardiogram in three windows as well as in five windows. The CURE clustering algorithm performs well since it contains high Clustering Accuracy when compared to other clustering algorithms of K-Means, CLARA and CLARANS.

VI. Conclusions

The outlier detection is one of the challenging areas in Data Mining in different data streams. By using large data sets, hierarchical clustering and partitioning clustering are helpful to detect the

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

anomalies very efficiently. In this paper, the clustering and outlier performance are analyzed in K-Means, CLARA, CLARANS and CURE clustering algorithm for detecting outliers. To find out the best clustering algorithm for detecting outliers some important performance measures are used. From the Experimental results it is observed that the clustering and outlier detection accuracy is more efficient in CURE clustering when compare to another clustering algorithms of K-Means, CLARA and CLARANS.

It is concluded that in handling high dimensional data it is necessary to choose a proper algorithm according to the size of the dataset. If the dataset is low dimensional, use k-means algorithm. If the dataset is middle dimensional, use CLARA or CLARANS algorithm. If the data set is very high dimensional, better go for CURE algorithm. Once the correct algorithm is chosen properly then clustering process and outlier's detection will become easier. In future combinations of existing Algorithms will be experimented in order to produce even more accuracy in results.

References

[1] *C. Aggarwal, Ed., Data Streams – Models and Algorithms, Springer, 2007.*

[2] *C. Aggarwal, J. Han, J. Wang, P.S. Yu, "A framework for projected clustering of high dimensional data streams", in Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004, pp. 852-863.*

[3] *Han.J and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2006.*

[4] *Hendrik Fichtenberger, Marc Gillé , Melanie Schmidt ,in Algorithms –ESA 2013 , Volume 8125, 2013, pp 481-492*

[5] *Irad Ben-Gal, "outlier detection", Department of Industrial Engineering Tel-Aviv University Ramat-Aviv, Tel-Aviv 69978, Israel.*

[6] *Madjid Khalilian, Norwati Mustapha, "Data Stream Clustering: Challenges and Issues", IMECS 2010.*

[7] *Mahnoosh kholghi, Mohammadreza Keyvanpour, "An analytical framework of data stream Mining techniques based on challenges and requirements", IJEST, 2011.*

[8] *Silvia Nittel , Kelvin T. Leung, "Parallelizing Clustering of Geo scientific Data Sets using Data Streams" Spatial Information Science & Engineering University of Maine &California.*

[9] *Zhang, T., Raghu, R., Miron, L.: BIRCH: An Efficient Data Clustering Method for Very Large Databases. ACM SIGMOD Record, vol. 25(2), 103- 114 (1996)*

[10] *T.H. Cormen, C.E. Leiserson, and R.L. Rivest, Introduction to Algorithms, the MIT Press, Massachusetts (1990).*

[11] *Michael Steinbach, Levent Ertöz, and Vipin Kumar, "The Challenges of Clustering High Dimensional Data", army High Performance Computing Research Center, DAAD19-01-2-0014.*

[12] *Pavel Berkhin, "Survey of Clustering Data Mining Techniques", Accrue Software, Inc, 1045 Forest Knoll Dr., San Jose, CA, 95129.*

[13] *A.K. Jain and R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, New Jersey (1988).*

[14] *R. Motwani and P. Raghavan, Randomized Algorithms, Cambridge University Press (1995).*

[15] *R.T. Ng and J. Han, Efficient and effective clustering methods for spatial data mining, In Proceedings of the VLDB Conference, Santiago, Chile, pp. 144-155, Morgan Kaufmann (1994).*

[16] *C.F, Olson, Parallel Algorithms for Ifiemrchical Clustering Technical report, University of California at Berkeley (1993).*

[17] *H. Toivonen, Sampling large databases for association rules, In Proceedings of the VLDB Conference, Bombay, India, pp. 134-145, Morgan Kaufmann (1996).*