IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

A Comparative Analysis of Hierarchical and Partitioning Clustering Algorithms for Outlier Detection in Data Streams

Dr. T. Christopher[1], Mrs. T. Divya[2]

PG& Research Department of Computer Science, Government Arts College, Coimbatore, TN, India[1]

PG& Research Department of Computer Science, Government Arts College, Udumalpet, TN, India[2]

Abstract: Outlier Detection is a fundamental issue in Data Mining. Data stream clustering techniques are highly helpful to cluster the similar data items in data streams and also to detect the outliers, so they are called cluster based outlier detection. It has been used to detect and remove unwanted data objects from large dataset. The clustering techniques are highly helpful to detect the outliers called cluster based outlier detection. In this research work clustering algorithms namely K-Means with CURE, K-Means with BIRCH, CURE with CLARANS, CLARANS with BIRCH,CLARANS, E-CLARANS and analyzed for finding the best result of detecting outliers in data streams. Two performance factors such as clustering accuracy and outlier detection accuracy are used for observation using WEKA tool. Through examining the experimental results, it is observed that the E-CLARANS outperforms well then the K-Means with CURE, K-Means with BIRCH, CURE with CLARANS, CLARANS with BIRCH, CLARANS Algorithms. E-CLARANS clustering algorithm performance is more accurate results than the rest of the clustering algorithm.

Keywords: Data stream, Data stream Clustering, Outlier detection.

I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The Data Mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of the following functionalities: data collection and database creation, data management (including data storage and retrieval, and database transaction processing). For instance, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics.

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself. It also is a buzzword, and is frequently also applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence.

II. LITERATURE REVIEW

Sharma, M. Toshniwal [1] to work a clustering algorithm called CURE and it is used for detecting outliers. CURE achieves by representing point per cluster its allow CURE to adjust well to the geometry of non-spherical shapes and the reduction helps to reduce the effects of outliers. The combination of random sampling and partitioning and the experimental results confirm that the quality of clusters produced by CURE is much better than those found by existing algorithms. Moreover, the authors expressed the partitioning and random sampling enable CURE to not only outperform existing algorithms but also to scale well for large databases without sacrificing the quality of cluster.

Aggarwal and Yu [6] used a new technique for outlier detection which is especially suited to very high dimensional data sets. The method works by finding lower dimensional projections which are locally sparse, and cannot be discovered easily by brute force techniques

because of the number of combinations of possibilities. This technique for outlier detection has advantages over simple distance based outliers which cannot overcome the effects of the dimensionality curse. They illustrated how to implement the technique effectively for high dimensional applications by using an evolutionary search technique. This implementation works almost as well as a brute-force implementation over the search space in terms of finding projections with very negative sparsity coefficients, but at a much lower cost. The techniques discussed in this paper extend the applicability of outlier detection techniques to high dimensional problems; such cases are most valuable from the perspective of data mining applications.

Elahi, M. KunLi, Nisar, W. XinjieLv, HonganWang [2] discussed about a clustering based approach, it divide the stream into chunks and for cluster each chunk using k-mean in fixed number of cluster. In this research the author took the candidate outliers and mean value of every cluster for the next fixed number of data streaming chunks, to make sure that the detected candidate outliers are the real outliers. The mean value are used in the clusters of previous streaming chunk and the current chunk of mean values are taken to be the consideration, which is used to choose better outlierness for data stream objects. Quite a few experiments for different types of dataset confirm that the technique can find better outliers with low computational cost than the other existing distance based approaches of outlier detection in data stream.

Madjid Khalilian, Norwati Mustapha [3] has conversed about the algorithm of k means for clustering of data streams and detection of outliers. The technique which has been used for outlier detection is based on distance as well as on time, on which they arrive in the cluster. The author takes into account the selection of k centers and variable size of buckets with the help of which space can be effectively utilized during clustering. Most traditional algorithms makes very difficult problem in clustering by reducing their quality for a better efficiency. In this research the author indicates a small increase of time,due to this cause the cluster can efficiently cluster the data without much loss of quality of data.

D. Joice and K. Lakshmi and K. Thilagam [5] discussed about data stream is a new emerging research area in data mining. In this paper is to perform the clustering process in data streams and to detect the outliers in high dimensional data using the existing clustering algorithms like K-means, CLARA, CLARANS and CURE. The experimental result of this paper shows that CURE clustering algorithm yields best performance compared to other algorithms.

III. METHODOLOGY

Clustering and Outlier detection is one of the important tasks in data streams. Outlier detection is based on clustering approach and it provides new positive results. The main objective of this research work is to perform the clustering process in data streams and detecting the outliers in data streams. In this research work, six clustering algorithms are used for clustering the data items and finding the outliers in data streams.

A. Dataset

In order to compare the data stream clustering for detecting outliers, data sets were taken from UCI machine learning repository. Datasets namely Breast Cancer Wisconsin Dataset with 699 instances, 10 attributes and Pima Indian data set contain 768 instances and 8 attributes. These two biological data sets have numeric attributes which have been used in this research work. Data stream is an unbounded sequence of data as it is not possible to store complete data stream, for this purpose we divide the data into chunks of same size and each chunk size is specified by the user which depends upon the nature of data and finally we divided the data into chunks of same size in different windows.

B. Clustering

Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster[3]. 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 modelling 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 many others.

C. Outlier Detection

Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behaviour. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains. Outlier detection has been a widely researched problem and immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, intrusion detection for cyber security, fault detection in safety critical systems, military surveillance for enemy activities and many other areas.

D. BIRCH WITH K-MEANS

In Birch with k-means clustering technique, both hierarchical and partitioning clustering algorithms are combined. The BIRCH WITH K-MEANS clustering algorithm as follows,

Input: Represent the data sample, S is x1,x2,…xn
Output: Data point values are clustered & the outliers are detected.

1. Draw a original data set, sample [x1,x2……xn] and n=number of data points, and distance (D0,D1,D2,D3,D4).

2. Calculate the centroid of data point cf1,cf2 and also data object are placed in cluster having centroid nearest to all data object.
3. The clustering feature cf1,cf2 are updating the cluster by using centroid values.
4. Finally group the cluster and detect the outliers.

E. CURE WITH CLARANS

CURE stands for Clustering using Representatives Algorithm. CURE is an efficient data clustering algorithm for large databases. It is processed using hierarchical methods to decompose a dataset into tree like structures. It uses two clustering approaches namely Partitioning clustering algorithm and Hierarchical clustering algorithm.

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

So, to avoid this problem of 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 its sensitive to outliers. The running time of the algorithm us O(n2 log n) and space complexity is O(n). The CURE Algorithm is us follows,

CURE (no. of points, k)
Input: A set of points S
Output: k clusters

Procedure:

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.

F. CLARANS Clustering

This method involves partitioning clustering algorithm in data streams [9]. First the data's are splitted into chunks of same size in different windows, after that consider each database(s) into data point (dp), partition of size=s/p, along with max neighbor of k=3. Then the minimum cost for each data point (dp) identifies the neighbor value, and it follows the condition i=1and j=1.Then the distance for each data point is calculated and also choose maximum distance (n) for each data points, if (s) has a lower cost, set current to(s), are increment j by 1.when j > max neighbor, compare the cost of current with minimum cost. If the cost value is less than (<) min cost, set minimum cost to current of cost value. Finally group the cluster, in order to satisfy the threshold value≤ min cost.

Finally nodes are clustered and outliers are identified. The CLARANS algorithm as follows,

Input: Represent the database(S) into data point (dp) Partition size=S/P, with Max neighbor &K=3.
Output: Data point values are clustered & the outliers are detected

Procedure:

1. Input the parameters num local and max neighbor. Initialize i to 1, for mincost to a large number.
2. Calculating the distance between each data points and also choose n of max distance at each of the data points.
3. Consider a random neighbor S of current, and calculate the cost differential of the two nodes.
4. If S has a lower cost, set current to S, increment j by 1 and If j is max neighbor, when j > max neighbor, compare the cost of current with min cost.
5. If the cost value is less than (<) min cost, set min cost to the cost of current value and set best node value to current node.
6. Finally group the cluster, to satisfy the threshold value≤ min cost. Then nodes are updated & cluster data and return outliers.
7. Else, Repeat the step 3 to step 5 up to best minimum cost (dmin), are found to other samples.
8. Return, Best cluster and detect the outliers efficiently.

G. E-Clarans Clustering

In E-CLARANS, first the data are splitted into chunks of same size in different windows, after that consider each database(S) into data point (dp), partition of size=s/p, along with max neighbor of k=3. Then the minimum cost for each data point (dp) is identified the neighbor value, and it follows the condition i=1and j=1.Then calculate the distance for each data points and also choose maximum distance (n) for each data points. Set current to an arbitrary node in n: k, for each data point we have to set j to 1along with a random neighbor (s) of current value, and also calculate the cost differential of the two nodes. If (s) has a lower cost, set current to(s) is increment j by 1.

when j > max neighbor, compare the cost of current with minimum cost. If the cost value is less than (<) min cost, set minimum cost to current of cost value. Finally group the cluster, in order to satisfy the threshold value ≤ min cost. Then lastly nodes are clustered and detect outliers.

The E-CLARANS algorithm as follows,

Input: Represent the database(S) into data point (dp) Partition size=S/P, with Max neighbor &K=3.
Output: Data point values are clustered & the outliers are detected

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Procedure:

1. Input the parameters num local and max neighbor. Initialize i to 1, for mincost to a large number.
2. Calculating the distance between each data points and also choose n of max distance at each of the data points.
3. Set current to an arbitrary node in n: k and Set j to 1. Consider a random neighbor S of current, and calculate the cost differential of the two nodes.
4. If S has a lower cost, set current to S, are increment j by 1 and If j is max neighbor, when j > max neighbor, compare the cost of current with min cost.
5. If the cost value is less than (<) min cost, set min cost to the cost of current value and set best node value to current node.
6. Finally group the cluster, to satisfy the threshold value≤ min cost. Then nodes are updated & cluster data and return outliers.
7. Else, Repeat the step 3 to step 5 up to best minimum cost (dmin), are found to other samples.
8. Return, Best cluster and detect the outliers efficiently.

H. K-MEANS WITH CURE

CURE with K-Means clustering technique also combines both hierarchical and partitioning clustering algorithms. The cure with k-means algorithm is as follows

Input: Represent the database(s) into data point (dp), partition size=sp, with maximum neighbor &k=3.
Output: Data point values are clustered & the outliers are detected

Procedure:

1. Consider the sample input database(s) into data point (dp).
2. Partially cluster the data points is s/p is Q.
3. Then set the current of an arbitrary node isGn1rs=Σni=1 [dp*min cost]/n.
4. Calculate the cluster centroid values d(x,y) using distance function (i.e.)Euclidean distance
5. If the cluster centroid value ≤ /α , α (i.e.) threshold value. Partially updated the cluster data and return outlier data.
6. Else, Repeat the step 3 up to best minimum value.
7. Return, Best cluster and detect the outliers

IV. EXPERIMENTAL RESULTS

A. OUTLIER DETECTION ACCURACY AND RESULTS

Accuracy

Outlier detection accuracy is calculated in order to find out the number of outliers detected by the clustering algorithms.

Detection Rate

Detection rate refers to the ratio between the numbers of correctly detected outliers to the total number of outliers, the detection rate is calculated using the formula as in Eq.(1),

$$d' = \frac{\mu_{S-\mu_N}}{\sqrt{1/2((\sigma_S^2+\sigma_N^2))}} \qquad (1)$$

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 deviation are represented as μ_S and 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 other name for it is False Detection Rate. In order to calculate the false alarm rate the formula is in Eq.(2),

$$\mathbf{FDR = FP/(TP + FP)= 1\text{-}PPV} \qquad (2)$$

The above formula uses False Positive (FP), True Positive (TP) and Positive Predictive (PPV) values to find the false alarm rate.
Outlier detection accuracy is calculated, in order to find out number of outliers detected by the clustering Algorithms CURE with K-MEANS and CURE with CLARANS and BIRCH with CLARANS and K-MEANS with BIRCH and CLARANC and E-CLARANS for Pima Indian diabetes data set and Wiscosin-breast cancer data set Table I & Table II &Table III & Table IV show the number of outlier detection rate and false alarm rate in three windows and five windows. Pima Indian diabetes dataset for outlier accuracy as follows,

Table I: The Outlier Detection Accuracy in Three Windows for Pima Indian Diabetes

Outlier Accuracy	Window Size	BIRCH with Clarans	CURE with Clarans	K-Means with CURE	K-Means with BIRCH	Clarans	E-Clarans
Detection Rate	W1	36.80	39.48	36.72	32.00	35.82	37.00
	W2	35.70	37.54	36.12	33.00	42.90	44.00
	W3	33.00	35.00	33.80	31.00	38.00	50.00
False Alarm Rate	W1	45.00	34.42	38.10	50.00	32.00	30.00
	W2	33.76	26.08	30.00	36.06	41.00	40.00
	W3	35.00	26.68	27.91	40.00	35.00	34.00

DOI 10.17148/IJARCCE.2015.41263

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Table II: The Outlier Detection Accuracy in Five Windows for Pima Indian Diabetes

Outlier Accuracy	Window Size	BIRCH with Clarans	CURE with Clarans	K-Means With CURE	K-Means with BIRCH	Clarans	E-Clarans
Detection Rate	W1	33.82	37.79	33.89	32.40	30.00	34.00
	W2	41.28	45.22	44.03	39.40	35.00	40.00
	W3	36.69	39.84	37.61	33.50	30.00	47.00
	W4	25.78	27.96	25.54	24.57	30.00	35.00
	W5	34.00	36.44	35.50	33.60	40.00	45.00
False Alarm Rate	W1	33.82	32.22	38.88	44.44	30.00	28.00
	W2	41.28	37.03	38.00	50.00	30.76	28.00
	W3	36.69	22.22	34.00	51.12	40.00	37.00
	W4	25.78	19.11	23.00	27.00	27.77	25.00
	W5	34.00	30.00	35.55	44.00	30.76	25.00

Breast Cancer Dataset as follows,

Table III: The Outlier Detection Accuracy in Five Windows for Breast Cancer Wiscosin

Outlier Accuracy	Window Size	BIRCH with Clarans	CURE with Clarans	K-Means with CURE	K-Means with BIRCH	Clarans	E-Clarans
Detection Rate	W1	56.02	57.25	56.12	32.40	53.00	57.00
	W2	53.60	57.57	56.00	39.80	47.00	52.00
	W3	64.00	66.60	66.00	33.50	58.00	77.00
	W4	77.70	79.62	77.58	24.57	78.00	79.12
	W5	75.47	77.53	76.42	33.60	77.00	79.78
False Alarm Rate	W1	52.30	43.75	54.76	44.44	50.00	51.00
	W2	62.50	47.61	50.00	50.00	50.76	51.00
	W3	72.00	58.00	60.00	51.12	68.12	67.00
	W4	78.00	72.72	78.00	27.00	75.66	55.00
	W5	72.00	68.00	71.00	44.00	60.77	55.00

Table IV: The Outlier Detection Accuracy In Three Windows For Breast Cancer Wiscosin

Outlier Accuracy	Window Size	BIRCH with Clarans	CURE With Clarans	K-Means with CURE	K-Means with BIRCH	Clarans	E-Clarans
Detection Rate	W1	55.72	57.76	54.26	53.65	54.46	56.00
	W2	63.41	67.41	64.07	60.97	72.43	74.00
	W3	75.52	78.65	77.00	75.28	65.00	77.00
False Alarm Rate	W1	56.09	43.74	59.42	60.86	54.00	52.00
	W2	64.28	51.78	60.71	78.57	66.00	65.00
	W3	80.92	70.90	72.46	82.92	80.92	79.00

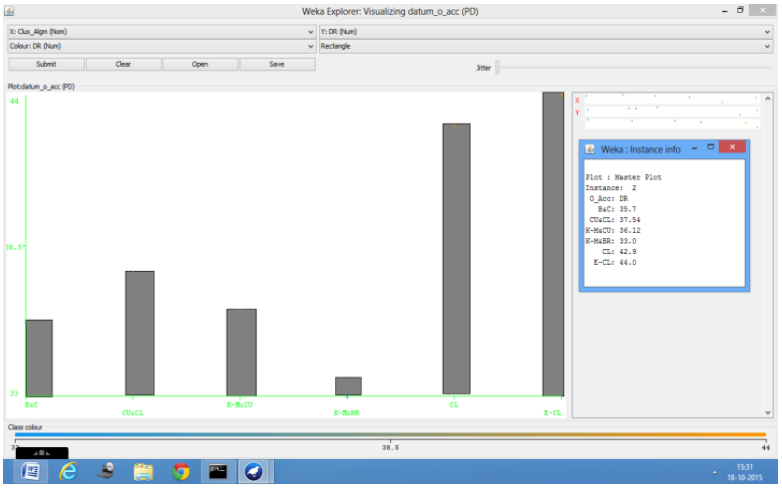

Fig 1: The outlier detection rate values in window size for Pima Indian Diabetes Dataset.

DOI 10.17148/IJARCCE.2015.41263

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

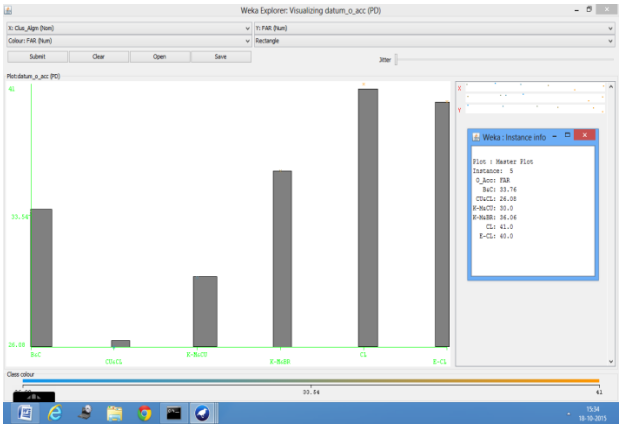

Fig 2: The outlier false alarm rate values in window size for Pima Indian Diabetes dataset

From the above graph, X-axis denotes the clustering algorithms are CURE with K-MEANS (K-M & CU) and CURE with CLARANS (CU & CL) and BIRCH with CLARANS (B&C) and K-MEANS with BIRCH (K- M & BR) and CLARANS (CL) and E-CLARANS (E-CL) and Y-axis denotes the Outlier Detection values for Pima Indian Diabetes dataset. It is observed that E-Clarans algorithm performs better than above clustering algorithms in Pima Indian Diabetes dataset for window size three and five. Therefore E-Clarans clustering algorithm performs well because it contains high outlier detection accuracy when compared to the above clustering algorithms.

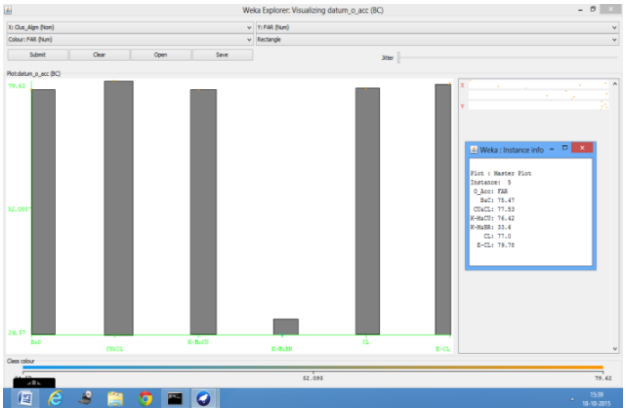

Fig 3: The outlier detection rate values in window size for Breast Cancer dataset

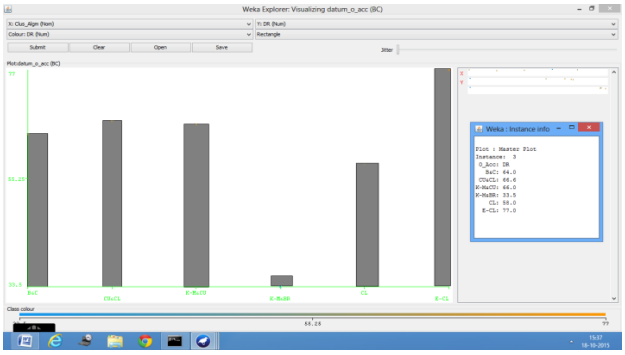

Fig 4: The outlier detection rate values in window size for Breast Cancer dataset

From the above graph, X-axis denotes the clustering algorithms are CURE with K-MEANS (K-M & CU) and CURE with CLARANS (CU & CL) and BIRCH with CLARANS (B&C) and K-MEANS with BIRCH (K-M & BR) and CLARANS (CL) and E-CLARANS (E-CL) and Y-axis denotes the outlier detection accuracy values for Breast Cancer dataset.

It is observed that E-Clarans algorithm performs better than above clustering algorithms in Breast Cancer dataset for window size three and five. Therefore E-Clarans clustering algorithm performs well because it contains high outlier detection accuracy when compared to the above clustering algorithms.

B. CLUSTERING ACCURACY AND RESULTS

Clustering accuracy is calculated using three measures i.e., Accuracy, Precision and Recall.

Accuracy

The accuracy determines how close the measurement comes to the true value of the quantity. So, it indicates the correctness of the result. The accuracy is calculated by using the formula is in Eq.(3),

$$\mathbf{ACC=(TP+TN)/(P+N)} \qquad (3)$$

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

Precision

The relative precision indicates the uncertainty in the measurement as a fraction of the result. The precision is calculated by using the formula is in Eq.(4),

$$\mathbf{PPV=TP/(TP+FP)} \qquad (4)$$

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

Recall

The recall relates to the test‘s ability to identify a condition correctly. The recall tests have few type II errors. The recall is calculated using the formula is in Eq.(5),

$$\mathbf{TPR=TP/P=TP/(TP+FN)} \qquad (5)$$

Where True Positive(TP), False Negative(FN) and Positive(P) values are used to found the recall values.

Pima Indian Diabetes dataset as follows,

From the below table clustering accuracy is calculated, by using two measures precision and recall. The clustering algorithms CURE with K-MEANS and CURE with CLARANS and BIRCH with CLARANS and K-MEANS with BIRCH and CLARANS and E-CLARANS for Pima Indian diabetes and Wiscosin-breast cancer data set.

Table V & Table VI & Table VII & Table VIII shows the clustering accuracy, precision and recall in three windows and five windows.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Table V: The Clustering Accuracy in three Windows for Pima Indian Diabetes

Clustering Accuracy	WZ	BIRCH With Clarans	CURE With Clarans	K-Means with CURE	K-Means with BIRCH	Clarans	E-Clarans
Accuracy	W1	76.17	88.28	82.03	70.31	87.50	90.00
	W2	76.20	88.32	82.10	70.03	87.50	90.00
	W3	76.17	88.28	82.03	70.31	87.60	90.00
Precision	W1	74.92	87.60	81.78	69.57	85.79	88.90
	W2	74.00	86.80	80.90	68.67	85.01	88
	W3	74.06	86.14	80.17	68.35	85.42	88.32
Recall	W1	74.01	87.60	81.58	69.57	85.31	88.12
	W2	75.89	87.25	81.35	69.28	86.89	89.90
	W3	76.73	88.82	82.89	70.68	87.96	90.64

Table VI: The Clustering Accuracy in Five Windows for Pima Indian Diabetes

Clustering Accuracy	WZ	BIRCH with Clarans	CURE with Clarans	K-Means With CURE	K-Means with BIRCH	Clarans	E-Clarans
Accuracy	W1	76.62	88.31	82.46	70.27	87.01	89.00
	W2	76.12	88.38	82.58	70.32	87.09	89.00
	W3	76.12	88.38	82.58	70.32	87.09	90.00
	W4	76.12	88.38	82.58	70.32	87.09	90.00
	W5	76.31	88.15	82.23	70.39	87.50	90.00
Precision	W1	75.22	86.81	81.32	69.45	86.43	88.54
	W2	76.18	87.96	80.90	70.39	87.12	88.92
	W3	74.39	87.36	80.86	68.75	85.32	86.78
	W4	69.50	83.86	78.67	65.32	80.50	83.5
	W5	74.84	87.07	76.86	68.32	85.34	88.04
Recall	W1	77.31	88.44	82.47	71.10	87	89.26
	W2	76.88	88.50	81.12	70.78	86.56	88.92
	W3	74.88	87.88	80.46	68.70	85.78	88.54
	W4	72.14	87.99	81.50	66.53	83.42	87.14
	W5	76.64	87.07	81.90	70.27	87.8	89.12

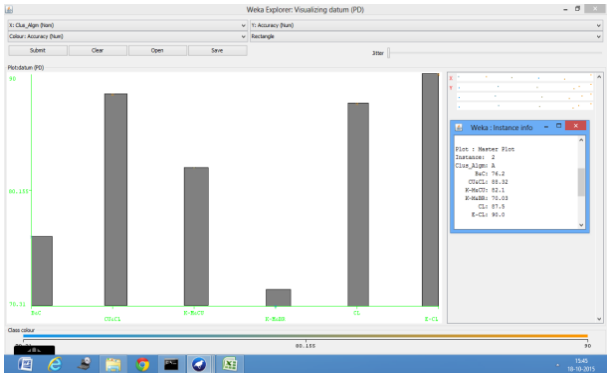

Fig 5: The clustering Accuracy values in window size for Pima Indian Diabetes dataset

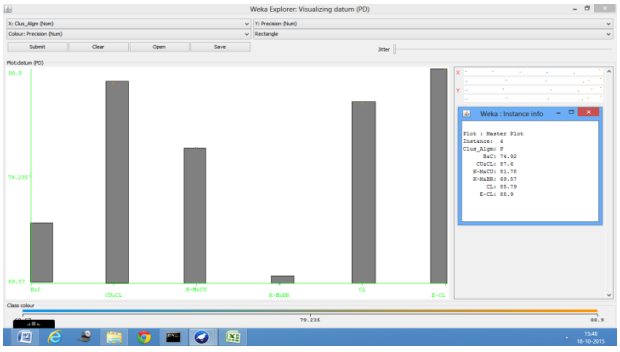

Fig 6: The clustering Precision values in window size for Pima Indian Diabetes dataset

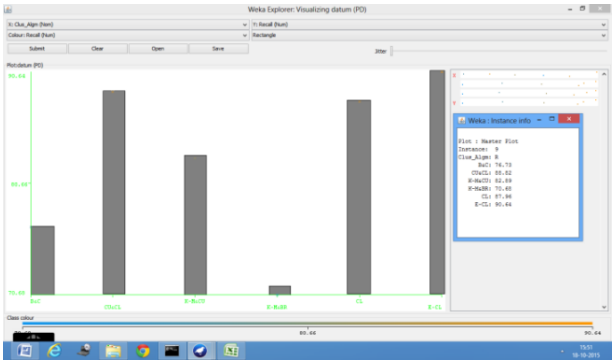

Fig 7: The clustering Recall values in window size for Pima Indian Diabetes dataset

From the above graph, X-axis denotes the clustering algorithms are CURE with K-MEANS (K-M & CU) and CURE with CLARANS (CU & CL) and BIRCH with CLARANS (B&C) and K-MEANS with BIRCH (K-M & BR) and CLARANS (CL) and E-CLARANS (E-CL) and Y-axis denotes the clustering accuracy values for Pima Indian Diabetes dataset. It is observed that E-Clarans algorithm performs better than above clustering algorithms in Pima Indian Diabetes dataset for window size three and five.

Therefore E-Clarans clustering algorithm performs well because it contains high accuracy when compared to the above clustering algorithms.

DOI 10.17148/IJARCCE.2015.41263

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Breast Cancer Dataset as follows,

Table VII: The Clustering Accuracy in Three Windows for Breast Cancer Wiscosin

Clustering Accuracy	WZ	BIRCH with Clarans	CURE with Clarans	K-Means With CURE	K-Means with BIRCH	Clarans	E-Clarans
Accuracy	W1	76.39	88.41	82.40	70.38	87.68	90.00
	W2	76.06	88.03	82.05	70.08	87.21	90.00
	W3	76.39	88.41	82.40	70.38	87.72	90.00
Precision	W1	76.11	88.28	82.98	70.24	86.47	89.28
	W2	74.79	86.80	81.38	68.54	85.01	88
	W3	69.55	83.00	76.68	66.06	82.28	85.00
Recall	W1	76.33	88.20	82.18	69.09	85.91	89.01
	W2	76.41	88.84	80.02	69.13	88.48	90.75
	W3	74.29	87.29	83.03	72.78	86.43	89.90

Table VIII: The Clustering Accuracy in Five Windows for Breast Cancer Wiscosin

Clustering Accuracy	WZ	BIRCH With Clarans	CURE With Clarans	K-Means With CURE	K-Means With BIRCH	Clarans	E-Clarans
Accuracy	W1	76.42	72.76	82.14	76.62	87.24	89.31
	W2	76.59	88.65	82.26	76.12	87.36	90.11
	W3	76.59	88.65	82.26	76.12	87.36	90.11
	W4	76.59	88.65	82.26	76.12	87.36	90.11
	W5	76.25	88.48	82.01	76.31	87.58	88.01
Precision	W1	76.18	88.99	81.69	75.45	87.89	87.56
	W2	76.40	88.66	80.92	76.01	87.32	88.32
	W3	75.12	87.09	79.85	73.18	85.02	86.54
	W4	71.20	82.70	76.75	68.12	79.67	83.08
	W5	69.82	84.18	79.29	73.92	82.64	86.32
Recall	W1	75.87	87.92	82.65	77.15	86	89.12
	W2	76.46	88.42	80.62	75.77	86.51	87.43
	W3	77.30	89.40	81.75	72.32	85.97	85.04
	W4	79.21	90.41	84.50	68.84	86.02	82.13
	W5	72.76	86.62	82.90	76.62	86.63	88.14

The graph denotes the X-axis are clustering algorithms namely CURE with K-MEANS (K-M & CU) and CURE with CLARANS (CU & CL) and BIRCH with CLARANS (B&C) and K-MEANS with BIRCH (K-M &BR) and CLARANS(CL) and E-CLARANS (E-CL) and Y-axis denotes the clustering accuracy values for Breast Cancer dataset.

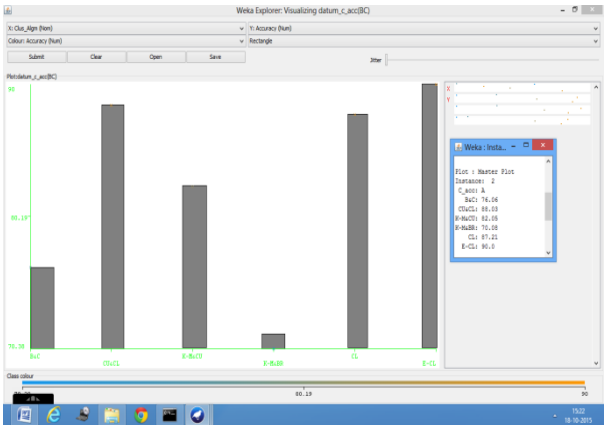

Fig 8: The clustering Accuracy values in window size for Breast Cancer dataset

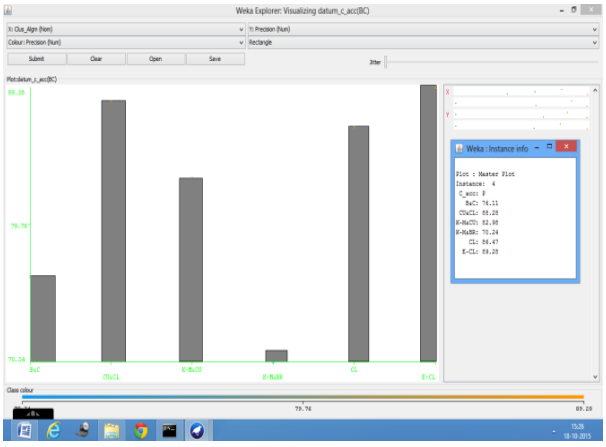

Fig 9: The clustering Precision values in window size for Breast Cancer dataset

From the graph it is observed that E-Clarans algorithm performs better than above clustering algorithms in Breast Cancer dataset for window size three and five. Therefore E-Clarans clustering algorithm performs well because it contains high accuracy when compared to the other clustering algorithm.

DOI 10.17148/IJARCCE.2015.41263

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

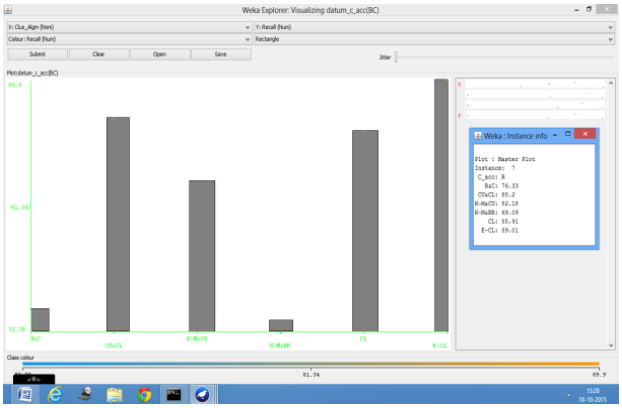

Fig 10: The clustering Recall values in window size for Breast Cancer dataset

V. CONCLUSION

Data streams are dynamic ordered, fast changing, massive, limitless and infinite sequence of data objects. Data streams clustering technique are highly helpful to handle those data. The outlier detection is one of the challenging areas in data stream. By using data stream hierarchical clustering and partition clustering are helpful to detect the outlier efficiently. In this paper we have analyzed the performance of CURE with K-MEANS and CURE with CLARANS and BIRCH with CLARANS and K-MEANS with BIRCH and CLARANS and E-CLARANS clustering algorithm for detecting the outliers. In order to find the best clustering algorithm for outlier detection two performance measures are used. From the experimental results it is observed that the outlier detection accuracy and clustering accuracy is more efficient in E-CLARANS clustering while compared to the above clustering algorithms.

REFERENCES

[1]. Sharma, M. Toshniwal, D, "Pre-clustering algorithm for anomaly detection and clustering that uses variable size buckets"**,** Published in Recent Advances in Information Technology (RAIT), 1st International Conference on 15-17 March 2012.

[2]. Elahi, M. KunLi, Nisar, W. XinjieLv, HonganWang, "Fuzzy Systems and Knowledge Discovery", Fifth International Conference on Vol.5, and Vol .3, pp. 23-27, 2002.

[3]. Madjid Khalilian, Norwati Mustapha , "Data Stream clustering-Challenges and issues", Proceedings of the International Multi Conference of Engineers and Computer Scientists Hong Kong ,Vol I,pp.17 - 19,March 2010.

[4]. Safal V Bhosale, "A Survey: Outlier Detection in Streaming Data Using Clustering Approached", International Journal of Computer Science And Information Technologies, Volume 5, Issue 5, 2014.

[5]. D.Joice, K. Lakshmi and K. Thilagam, "Comparison Of Cluster Based Algorithms For Outlier Detection In High Dimensional Dataset", Karpagam Journal of computer science, Volume 8, issue 3, April 2014.

[6]. Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection", The VLDB Journal, 2005, vol. 14, pp. 211–221.

[7]. Elahi, M. KunLi, Nisar, W. XinjieLv, HonganWang, "Fuzzy Systems and Knowledge Discovery", Fifth International Conference on Vol.5, and Vol .3, pp. 23-27, 2002.

[8]. Prakash Chandore, Prashant Chatur, "Outlier Detection Techniques Over Streaming Data in Data Mining: A Research Perspective", International Journal of Recent Technology and Engineering, Volume 2, Issue 1, March 2013.

[9]. Dr. Manju Kaushik, Mrs. Bhawana Mathur, "Comparative Study Of K-Means and Hierarchical Clustering Techniques", International Journal of Software & Hardware Research in Engineering, Volume 2, Issue 6, June 2014.

[10]. V. Suganthi, S.Tamilarasi, "A Study On Clustering High Dimensional Data Using Hubness Phenomenon", IOSR Journal Of Computer Engineering, Volume 16, Issue 2, Mar-Apr 2014.

[11]. Madjid Khalilian, Norwati Mustapha , "Data Stream clustering-Challenges and issues", Proceedings of the International Multi Conference of Engineers and Computer Scientists , Hong Kong ,Vol I,pp.17 - 19,March 2010.

[12]. Sharma, M. Toshniwal, D, " Pre-clustering algorithm for anomaly detection and clustering that uses variable size buckets"**,** Published in Recent Advances in Information Technology (RAIT), 1st International Conference on 15-17 March 2012.

[13]. Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, "CURE: an efficient clustering algorithm for large databases", ACM LIBRARY, 1999.