ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A New Approach for Detecting Outliers in Data Streams

Dr. S. Vijayarani[*1], Ms. P. Jothi[2]

[*1] Assistant Professor, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India

[2] M.Phil Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India

vijimohan_2000@yahoo.com

Abstract

In modern years, data streams have become an increasingly important research area, where as data stream refers to continuous flow of data and it is a process of extracting knowledge structure from continuous, rapid data records and it can be considered as a subfield of data mining. Data Stream can be classified into two types they are offline and online streams. Online data stream used in an amount of real world appliances, including network traffic monitoring, intrusion detection, credit card and fraud detection and offline data stream are used in reports based on web log streams. Data size is extremely huge and potentially infinite and it's not possible to store all the data, so it leads to a mining challenge where shortage of limitations occurs in hardware and software. Data mining techniques are newly proposed for data streams they are highly helpful to mine the data are data stream clustering, data stream classification, frequent pattern technique, sliding window techniques and so on. For outlier detection data stream clustering technique is highly desirable one. The main objective of this research work is to perform the clustering process in data streams and detecting the outliers in data streams. Two types of clustering algorithms namely FUZZY C-MEANS and CLARANS are used for finding the outliers in data streams. The two performance factors such as clustering accuracy and outlier detection accuracy are used for analysis. By analyzing the experimental results, it is observed that the CLARANS clustering algorithm performance is more accurate than the FUZZY C-MEANS.

Keywords: Data stream, Data stream Clustering, Outlier detection, BIRCH, CURE, CLARANS

Introduction

Data mining is one of the research areas where data mining is termed as extracting hidden and useful information from the data, but as now recent days it is observed that enormous research activity is transferred in the format of data streams [1]. A data stream is an unremitting, immediate, stream flow of sequence of items and it is not possible to control the order in which data item arrive, nor is it feasible to locally store a stream in its entirety. The applications of data streams are generated like sensor networks, , traffic management, call detail records, blogging and twitter posts has been led to encourage by the study of data stream. Due to lack of resources where as this type of large data, the current data mining systems are not sufficient and equipped to deal with them. A model which is developed from data stream for clustering is an appropriate method for handling huge volume of updatable data [2]. Data stream clustering is a well-known task in mining data stream, clustering is known as grouping related objects into a cluster. With the help of data stream clustering method, we can detect the outliers, and the outliers are nothing but it is one of the data mining jobs, and it is termed as outlier mining. An outlier is an object that does not fulfill with the behavior of normal data objects. Applications of outlier detection are telecommunication, web logs, fraud detection and web document [14]. Clustering based outlier mining methods are called as unsupervised in nature and its main objective is to find the outlier from the data stream using partitioning cluster based method. The cluster based outlier detection approaches are termed as a cluster based technique for detecting outlier where it finds closely related objects. The object which does not belong to any cluster or belongs to a small cluster is stated as outlier, and the outlier detection process highly depends upon the clustering technique.

The remaining section of this paper is followed as; Section 2 illustrates the review of literature. Section 3 describes the FUZZY C MEANS and CLARANS clustering algorithms used to detect outliers in data

ISSN: 2277-9655
Impact Factor: 1.852

streams. Experimental results are discussed in section 4 and conclusions are given in section 5.

Literature Review

Jurgen Beringer, et.al [8] discussed about the problem of clustering parallel streams of real-valued data in continuously evolving time series of grouping data streams the evolution of over time. In order to maintain an up-to-date clustering structure, the authors had been necessary to analyze the incoming data in an online manner and also tolerating in a constant time delayed. For this purpose, the authors develop an efficient online version of the fuzzy C-means clustering algorithm. Depending on the point of view and on the assumptions on the data generating process, it can even be argued that preprocessing can improve the quality by removing noise in the original streams. Finally the authors concluded the paper the FCM-DS does not necessarily produce the end product of a data mining process and actually, it had been considered as transforming a set of data streams into a new set of streams the elements of which are clustered memberships and the cluster data streams can be evaluated by means of other data mining tools.

P.hore, et.al,[11]proposed an online fuzzy c means algorithm for streaming data or very large data and the results on a number of large data sets and as well as millions of examples showed the new algorithm produced a partitions which are very closes to clustered of all the data at one time arrival. The experiments on the online Fuzzy c mean algorithm output performed a streaming FCM in processing of streaming data. Online FCM can process the streaming data as it comes and also it accessed the single pass of FCM requires random reordering to achieve the reasonable results.

Raymond T. Ng ,et.al, [13] presented a clustering algorithm called CLARANS which is based on randomize search. The authors had developed two spatial data mining algorithms SD (CLARANS) and NSD (CLARANS). The experimental results and analysis indicated that both algorithms are effective, and can lead to discoveries that are difficult to obtain with existing spatial data mining algorithms. Finally, they had a presented experimental results showed that CLARANS itself is more efficient than existing clustering methods. Hence, CLARANS has established itself as a very promising tool for efficient and effective spatial data mining.

Methodology

In data stream, the clustering technique is applied for grouping the data items and also detecting the outliers. Clustering and Outlier detection are most important problems in data streams. The main objective of this research work is to analyse the performance of the two clustering algorithms namely FUZZY C-MEANS and CLARANS for detecting the outliers. The system architecture of the research work is as follows as

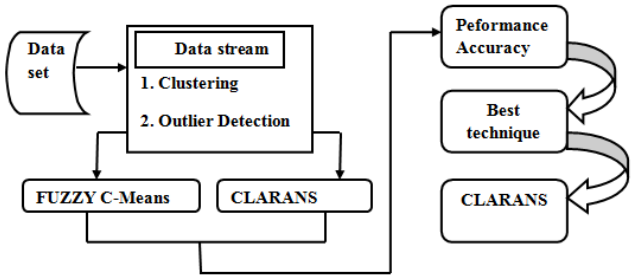

Fig 1: System Architecture

A. Dataset

Dataset which have been used in this research work is Pima Indian data set; it contains 768 instances and 8 attributes. The dataset is taken UCI machine learning repository [3]. Data stream is an abundant flawless sequence of data and it is not possible to store the complete data stream, due to this reason 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

B. Clustering

Cluster analysis is used in a various number of applications they are stock market analysis, data analysis, image processing and financial market analysis etc [1]. In data streams the clustering is one of the sub- process areas which are used to group the objects as well as it is used to detect the outliers efficiently and also clustering is one of the unverified action in data streams. The data stream clustering are different types of approaches they are distance based, grid based, partition based, hierarchical based and so on.

C. Outlier detection

Outlier detection over streaming data is active research area from data mining that aims to detect object which have different [5] behavior, exceptional than normal object. An outlier is an object that is significantly unrelated or incompatible to other data object whereas weblogs click stream telecommunication, fraud detection, documents of web are the application areas of outlier detection in data streams. The other names of outlier detection are termed as noise, anomalies, indifferent, not catchable to the related object, unknown and so on. The clustering based outlier detection is a best technique to manage this problem. For our research we have used cluster based outlier detection algorithms FUZZY C -Means and CLARANS.

D. FUZZY C-Means clustering

Fuzzy c-means is a clustering FCM method [9] the clustering allows single piece of data which belong to two or more clusters and this method was developed

ISSN: 2277-9655
Impact Factor: 1.852

by Bezdek and Dunn; it has been often used in pattern recognition. Based on minimization of the subsequent intention function, were m is any real number greater than 1 and u_{ij} is the degree of membership of x_i in the cluster j along with x_i is the ith of d-dimensional measured data, at last c_j is the d-dimension center of the cluster, and operator ||*|| is any norm to the center point of data. Fuzzy partitioning is agreed out through an iterative optimization of the objective functions, along with the in formable and update of membership u_{ij} and the cluster centers c_j by, formulae, the Fuzzy c-Means algorithm followed as

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, \quad 1 \le m < \infty$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

1. This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k is the iteration steps. And this procedure converges to a local minimum or add point of J_m.
2. The algorithm is compose Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
3. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$
4. Update $U^{(k)}, U^{(k+1)}$
5. If $\| U^{(k+1)} - U^{(k)}\| < \varepsilon$ then STOP; otherwise return to step 2.

E. CLARANS

CLARANS [12] is abbreviated as Clustering Large Application Based upon Randomized Search and it use random search, to generate neighbors by starting with arbitrary node and randomly check max-neighbors. CLARANS are similar to PAM and CLARA while it starts with the selection of mediod at randomly and it describes the neighbor dynamically and it checks max neighbor for swapping and if the pair is negative then it chooses another medoid set or otherwise it chooses current selection of medoids as local optimum and restarts with the new selection of medoids randomly and it stops the process until returns the best. If the neighbor represent as better partition the process continue with new node otherwise local minimum is found and algorithm restart until num local minima is found the value of num local is=2 recommended then the best node return resulting partition. CLARANS take a random dynamic selection of data

1. Randomly choose k mediod
2. Randomly consider the one of mediod swapped with non mediod
3. If the cost of new configuration is lower repeat step 2 with new solution
4. If the cost higher repeat step 2 with different non mediod object unless limit has been reached
5. Compare the solution keeps the best
6. Return step 1 unless limit has been reached (set to the value of 2).

at each step of process and thus the same sample set is not used throughout in the clustering process and CLARANS is accurately detecting outlier than CLARA and it is much less affected by increasing dimensionally and draw the sample of neighbors in each step of search this is benefit of confining the search localize area.

Experimental Results

We have implemented the two algorithms in MATLAB 7.10 (R2010a). In order to evaluate the performance of the algorithms, the two factors namely clustering accuracy and outlier accuracy are used. The different sizes of the window are 5 and 10.

Clustering Accuracy

Clustering accuracy is calculated, by using two measures Precision and recall, and the clustering algorithms FUZZY C-Means and CLARANS used for Pima Indian diabetes. Table 1 & Table 2 illustrate the clustering accuracy, precision and recall in five windows and ten windows.

Table 1: The clustering accuracy in five windows for Pima Indian diabetes dataset

Clustering Accuracy	No. of windows	FUZZY C Means	CLARANS
Accuracy	w1	83.7	87.01
	w2	83.2	87.09
	w3	83.22	87.09
	w4	83.22	87.09
	w5	83.55	87.5
Precision	w1	81.4	85.49
	w2	83.18	86.68
	w3	81.92	85.91
	w4	77.55	85.39
	w5	81.14	87.03
Recall	w1	83.17	86.59
	w2	84.06	86.97
	w3	83.43	86.86
	w4	81.13	87.98
	w5	83.5	85.32

ISSN: 2277-9655
Impact Factor: 1.852

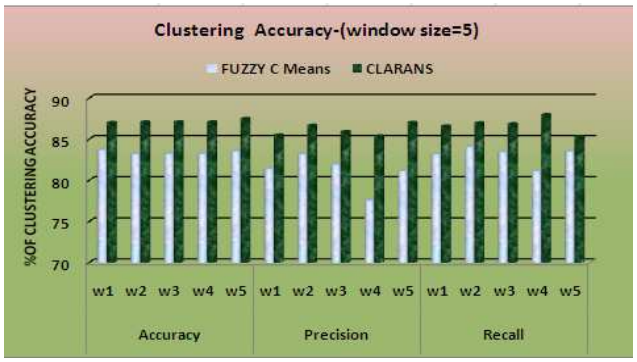

Fig 2: The clustering accuracy in five windows for Pima Indian Diabetes dataset

Table 2: The clustering accuracy in ten windows for Pima Indian Diabetes dataset

Clustering Accuracy	No. of windows	Accuracy	Precision	Recall
Fuzzy C Means	w1	83.11	82.84	83.75
	w2	83.33	79.93	83.12
	w3	83.33	82.55	83.42
	w4	83.33	82.71	82.33
	w5	83.33	81.59	83.77
	w6	83.33	82.83	84.58
	w7	83.33	75.23	83.6
	w8	83.33	80.7	82.45
	w9	83.33	80.71	82.45
	w10	84	83.62	82.78
CLARANS	w1	87.01	86.62	87.53
	w2	87.17	84.1	87.11
	w3	87.17	86.98	86.01
	w4	87.17	86.62	86.42
	w5	87.17	85.84	85.64
	w6	87.17	86.86	85.83
	w7	87.17	80	92.06
	w8	87.17	85.28	88.28
	w9	87.17	84.97	86.34
	w10	88	87.87	89.44

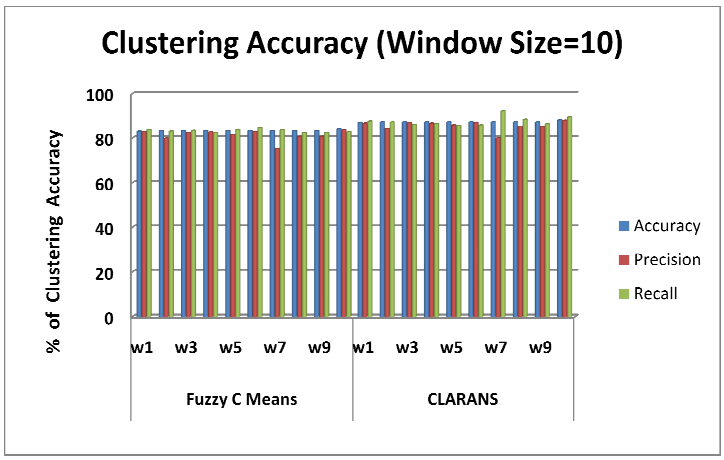

Fig 3: The clustering accuracy in ten windows for Pima Indian Diabetes dataset

From the above graph, it is observed that FUZZY C-Means clustering algorithm performs better than CLARANS clustering algorithm in Pima Indian Diabetes dataset for both window size as five and ten. Therefore the CLARANS clustering algorithm performs well because it contains high clustering accuracy when compared to FUZZY C-Means.

Outlier Accuracy

Detection Rate and False Alarm Rate for Pima Indian Diabetes

Outlier detection accuracy is calculated, in order to finding the number of outliers detected by the clustering algorithms FUZZY C-Means and CLARANS for Pima Indian diabetes data set. Table 3& Table 4 show the number of outlier detection rate and false alarm rate in five windows and ten windows.

Table 3: Detection rate and false alarm rate in five windows-Pima Indian diabetes

Outlier Accuracy	No. of windows	FUZZY C Means	CLARANS
Detection rate	w1	34.55	35.82
	w2	40.14	42.95
	w3	35.65	36.29
	w4	23.70	24.81
	w5	34.64	35.97
False alarm rate	w1	38.88	30.00
	w2	55.55	30.76
	w3	42.30	40.00
	w4	34.61	27.77
	w5	40.00	30.76

ISSN: 2277-9655
Impact Factor: 1.852

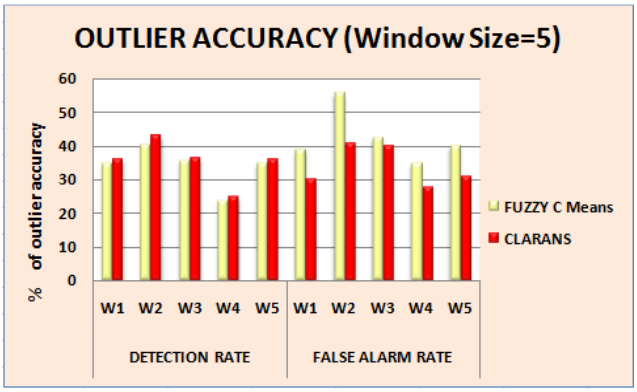

Fig 4: Detection rate and False alarm rate in five windows- Pima Indian diabetes

Table 4: Outlier Accuracy for Ten windows- Pima Indian diabetes

Outlier Accuracy	No of Windows	FUZZY C Means	CLARANS
Detection Rate	W1	41.79	42.64
	W2	26.47	30.43
	W3	43.47	45.58
	W4	36.61	40.57
	W5	33.8	34.78
	W6	36.22	38.23
	W7	36.22	20.12
	W8	32.39	33.84
	W9	32.39	34.78
	W10	39.13	39.39
False Alarm Rate	W1	40.00	33.33
	W2	46.15	22.22
	W3	55.55	40.00
	W4	61.53	33.33
	W5	42.85	33.33
	W6	53.84	40.00
	W7	20.11	11.11
	W8	28.57	23.07
	W9	28.57	11.11
	W10	50.00	44.44

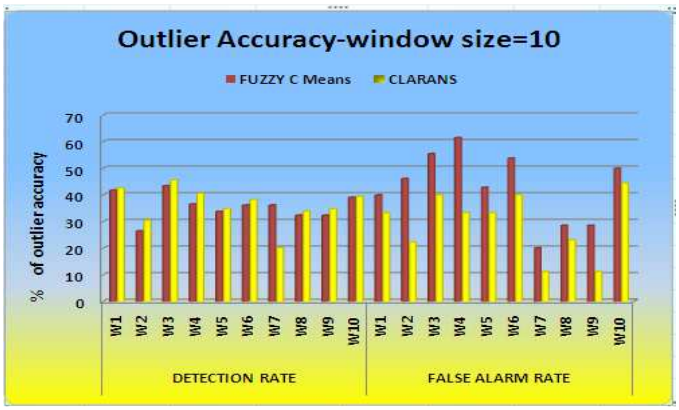

Fig 5: Detection rate and false alarm rate in Ten windows- Pima Indian diabetes

From the above graph, it is observed that CLARANS clustering algorithm performs better than FUZZY C-MEANS algorithms for detecting outliers in Pima Indian Diabetes dataset for both window size as five and ten.

Conclusion

Data streams are fast and limitless arrival of ordered and unordered data, by using of data streams clustering technique we can handle those data. Detecting the outlier in data stream is one of the challenging research problems. In this paper, we have analyzed the performance of FUZZY C-Means and CLARANS clustering algorithm for detecting the outliers. In order to find the best clustering algorithm for outlier detection several performance measures are used. From the experimental results it is come to know that the outlier detection and clustering accuracies are more efficient in CLARANS while compared to FUZZY C-Means clustering.

References

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," In Proc. of VLDB, pages 81-92, 2003.

[2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional data streams," In Proc. of VLDB, pages 852-863, 2004.

[3] C. J. Merz and P. M. Murphy. UCI Repository of Machine Learning Databases Univ. of CA. Dept. of CIS, Irvine, CA.

[4] D. Hawkins, "Identification of outliers- Monographs on statistics and applied probability", First edition, pages-188, Springer published in 1980.

[5] Hossein Moradi Koupaie, Suhaimi Ibrahim, Javad Hosseinkhani, "Outlier Detection in Stream Data by Clustering Method"

International Journal of Advanced Computer Science and Information Technology (IJACSIT), BVol. 2, No. 3, 2013, Page: 25-34, ISSN: 2296-1739.

[6] Irad Ben-Gal, "Outlier Detection", Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer , Academic Publishers, 2005

[7] J. Yang, "Dynamic clustering of evolving streams with a single pass," In Proc. of ICDE, pages 695-697, 2003.

[8] Jurgen Beringer and Eyke Hullermeier, "Online clustering of parallel data streams," Data Knowl. Eng, 58:180–204, 2006.

[9] Maciej Jaworski, Piotr Duda, Lena Pietruczuk, "Artificial Intelligence and Soft Computing", Lecture Notes in Computer Science Volume 7268, 2012, pp 82-91.

[10] P. Hore, L. O. Hall and D. B. Goldgof, "Creating Streaming Iterative Soft Clustering Algorithms," NAFIPS 07, San Diego, pages 484-488, 2007.

[11] P.hore, L.O.hall, D.B. Goldgof, W. cheng, **"**Online Fuzzy C Means ", IEEE, Page(s):1 – 5,E-ISBN :978-1-4244-2352-1,Print ISBN:978-1-4244-2351-4,2008.

[12] R. J. Hathaway and J. C. Bezdek, "Extending Fuzzy and Probabilistic Clustering to Very Large Data Sets," Journal of Computational Statistics and Data Analysis, pages 215-234, 2006.

[13] Raymond T. Ng and J. Han. Efficient and effective clustering method for spatial data mining, VLDB'94.

[14] Sudipto Guha, Adam Meyerson , Nine Mishra and Rajeev Motwani, "Clustering Data Streams: Theory and practice," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, pp. 515-528, May/June,2003.