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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

Hierarchical and Partitioning Clustering Algorithms for Detecting Outliers in Data Streams

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

Assistant Professor, Department of Computer Science, School of Computer Science and Engineering,

Bharathiar University, Coimbatore, Tamilnadu, India[1]

M.Phil Research Scholar, Department of Computer Science, School of Computer Science and Engineering,

Bharathiar University, Coimbatore, Tamilnadu, India[2]

Abstract: The data stream is a new arrival of research area in data mining where as data stream refers to the process of extracting knowledge structures from unlimited and fast growing data records. Future applications involved in data streams are motivated by many researchers involving continuous and massive data sets such as telecommunication system, customer click streams, ecommerce, meteorological data, network monitor, stock market and wireless sensor network. For handling this type of stream data, the recent data mining methods are not sufficient and equipped to deal with them, for this reason it leads to a numerous computational and mining challenges due to shortage of hardware limitations. Nowadays many researchers have focused on mining data streams and they proposed many techniques for data stream classification and clustering, as well as mining frequent items from data streams. Data stream clustering and outlier detection provides a number of unique challenges in evolving data stream environment. Data stream clustering algorithms are highly used for detecting the outliers in efficient manner. The main purpose of this research work is to perform the clustering process and detecting the outliers in data streams. In this research work, two types of clustering algorithms namely BIRCH with K-Means and CURE with K-means is used for finding the outliers in data streams. Two performance factors such as clustering accuracy and outlier detection accuracy are used for observation. Through examining the experimental results, it is observed that the CURE with K-Means clustering algorithm performance is more accurate than the BIRCH with K-Means algorithm.

Keywords: Data stream, Data stream Clustering, Outlier detection, CURE, K-Means, BIRCH

I. INTRODUCTION

Data mining is broadly studied field of research area, where most of the work is emphasized over, in that data stream is one of the research areas in data mining because in data streams, [2] data are massive, fast changing, unlimited, continuous flow and endless. Applications of data streams can diverge from scientific and astronomical applications to important business and financial ones therefore real-time analysis and mining of data streams have attracted substantial amount of researches. Data stream clustering [1] is a sub-area of mining data streams, since clustering algorithms arrange a dataset into several disjoint groups, such that points in the same group are related to each other and are unrelated to other groups, according to a few relationship metrics. In order to use clustering in data streams, the requirements are to be generated for overall high-quality clusters without seeing the old data, high quality, efficient incremental clustering algorithms and analysis in multi-dimensional space. Hierarchical clustering and partitioning clustering algorithms [3] are highly helpful for outlier detection. The hierarchical algorithms create a decomposition of the objects and they are either agglomerative, divisive top down or bottom up. Agglomerative algorithms usually start with each object, and successively unite groups according to a distance measure, where as clustering may stop when all objects are in a single group or at any other point the user wants and these methods defined as greedy bottom up merging. Divisive algorithms follow the reverse approach; it starts with single group of all objects and successively split groups into minor items, until all object falls into single cluster, are to be preferred. The partitioning algorithm constructs various partitions for the data elements and then evaluates them by some criteria. Data stream clustering methodologies are highly helpful to detect outliers and outlier detection is one of the data mining tasks and it is otherwise called as outlier mining. An outlier detection, streaming data is one of the active research area from data mining that aims to detect object which have different actions exceptional than normal object. An outlier is an object [8] that is significantly dissimilar or inconsistent to other data object where as click stream, fraud detection, web logs, and web documents are the application of outlier detection in data streams area. There are many algorithms for outlier detection in static and stored data sets which are based on a variety of approaches like nearest neighbour based, density based outlier detection, distance based outlier detection and clustering based outlier detection.

The rest of this paper is organized in the following manner. Section 2 illustrates the review of literature; Section 3 described how Cure with K-Means and Birch with K-Means clustering algorithms are used for detecting

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

outliers in data streams. Section 4 is discussed the experimental results and Conclusions are given in Section 5.

II. LITERATURE REVIEW

Luis Torgo, et.al [7] put forth a methodology for the application of hierarchical clustering methods to perform the task of outlier detection. This methodology is tested on the official statistics data and the foreign trade transactions data, where the data collected from the Statistics Institute. In this research work the authors discussed the outlier ranking method (LOF) and it achieved better results.

Thankran.Y, et.al [13] discussed an unsupervised outlier detection method for streaming data. This is an unsupervised data mining task for clustering based method and it does not require labelled data. In this method density and partitioning based clustering methods are combined and they assigned weights to each attributes depending upon their respective relevance in mining task and weights are adaptive in nature. Their proposed method is incremental and adaptive to concept evolution and it also relevant to the decrease effect of noisy attributes.

Sharma.M, et.al [12] has presented the algorithm of k-means for clustering and outlier detection for data streams. Most traditional algorithms makes very difficult problem in clustering by reducing their quality for a better effectiveness. In this research the author designates a small increase of time, due to this cause the cluster can efficiently cluster the data without much loss of quality of data.

Hendrik Fichtenberger et.al, [5] proposed a k-means based data stream clustering algorithm called BICO (BIRCH Meets Core sets for k-Means Clustering), BICO computes high quality solutions in a time short and also it computes a summary S of the data with a provable quality. In this research, the authors compared BICO experimentally with popular BIRCH and Mac Queen Algorithm along with approximation algorithms as Stream km++ and Stream LS.

Irad Ben Gal, et.al [6] discussed about several methods, and techniques for outlier detection and how the outliers are distinguished between uniform variate vs. multivariate techniques and parametric vs. non-parametric procedures. This paper also discussed how a mixture of supervised, semi-supervised and unsupervised techniques has been used for outlier detection and their strengths and weaknesses.

III. METHODOLOGY

Clustering and Outlier detection is one of the important tasks in data streams. Outlier detection supports clustering approaches which provides new optimistic results. The major intention of this research work is to perform the clustering method and detecting the outliers in data streams. In this research work, clustering algorithms namely BIRCH with K-Means and CURE with K-Means is used for clustering the data items and finding the outliers in data streams. Figure 1 shows the system architecture of the research work.

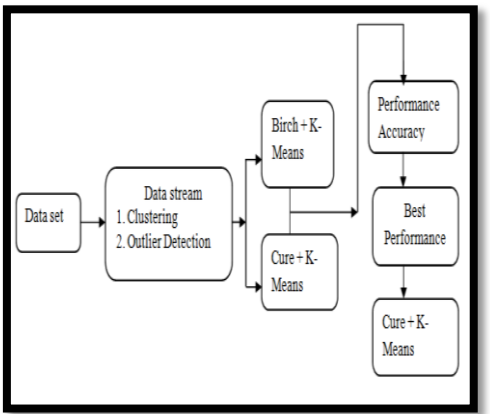

Fig 1: System Architecture

A. Dataset

In order to compare the data stream clustering algorithms for detecting outliers, data sets were taken from UCI machine learning repository [4], from which Pima Indian data set is used for this work. It contains 768 instances and 8 attributes. Data stream is an unbounded sequence of data as it is not possible to store complete data stream, for this purpose we divided the data into chunks of same size in different windows.

B. Clustering

The clustering algorithm is used to group objects into significant subclasses [2]. The clustering algorithms for data streams should be adaptive in the sense that up-to-date clusters are obtainable at any time, allows new data items as soon as they arrive. There are different types of clustering algorithms namely hierarchical clustering algorithm, partitioning clustering algorithm, grid and density based clustering algorithms. Clustering is defined as an unsupervised problem and where as there are no predefined class labels exist for the data points. Cluster analysis is used in number of applications such as image processing, data analysis, stock market study etc.

C. Outlier Detection

Outlier detection has a wide range of applications such as fraud detection, intrusion detection, and credit card analysis. It is further complicated by the fact that in many cases outliers have to be detected from a large volume of data growing at an unlimited rate. Traditional outlier detection algorithms cannot be functional to data stream efficiently, since the data stream is potentially infinite and evolving continuously. It has to be routed within an exact time constraint and limited space, thus outlier detection in data stream imposes great challenges [6] are followed. The cluster based outlier detection is a best technique to supervise this problem. For our research work we have used cluster based outlier detection as BIRCH with K-Means and CURE with K-Means.

D. BIRCH with K-MEANS clustering

In Birch with k-means clustering technique, both hierarchical and partitioning clustering algorithms are combined.

Input: Represent the data sample, S is $x_1, x_2, \ldots x_n$

Output: Data point values are clustered & the outliers are detected. Procedure:

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

1. Draw a original data set sample $[x_1,x_2.....x_n]$ and n=number of data points, and distance (D_0,D_1,D_2,D_3,D_4).
2. Calculate the centroid of data point cf_1,cf_2 and also data object are placed in cluster having centroid nearest to all data object.
3. The clustering feature cf_1,cf_2 are updating the cluster by using centroid values.
4. Finally group the cluster and detect the outliers.

E. *CURE with K-MEANS clustering*

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 is $Gn_1rs=\sum^{n}_{i=1}[dp*min\ cost]/n$.
4. Calculate the cluster centroid values d(x,y) using distance function (i.e.)Euclidean distance
5. If the cluster centroid value $\leq /\alpha$, 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

The evaluation was performed on PC Intel Pentium processor, 2GB RAM, OS Windows 7 Ultimate 32-bit. We have implemented our algorithms in MATLAB 7.10(R2010a). For performance evaluation, two performance factors such as clustering accuracy and outlier detection accuracy are used. The outlier detection accuracy is calculated by using two measures they are detection rate and false alarm rate. Pima Indian diabetes data set is used in this work. Pima Indian diabetes data set contains 8 attributes and 768 instances.

A. Clustering accuracy

TABLE I
THE CLUSTERING ACCURACY FOR BIRCH +K-MEANS, CURE+ K-MEANS IN THREE AND FIVE WINDOW

Clustering Accuracy	Window size	No. of windows	BIRCH+ K-Means	CURE + K-Means
Accuracy	Three	w1	70.31	82.03
		w2	70.03	82.10
		w3	70.31	82.03
	Five	w1	70.27	82.46
		w2	70.32	82.58
		w3	70.32	82.58
		w4	70.32	82.58
		w5	70.39	82.23

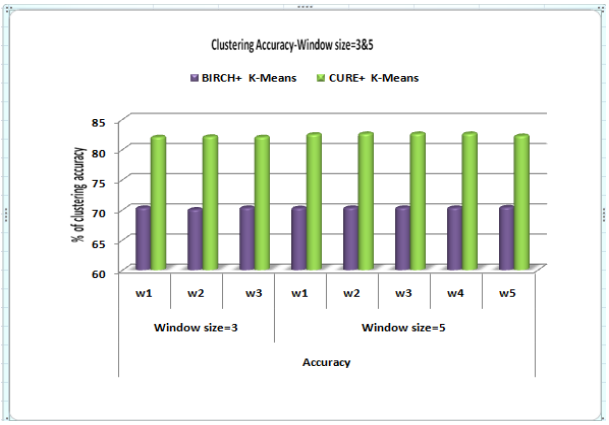

Fig 2: The clustering accuracy for BIRCH +K-Means And CURE+K-Means

From the above figure 2, it is observed that CURE with K-Means clustering algorithm performs better than BIRCH with K-Means algorithms for detecting outliers in Pima Indian Diabetes dataset for both window size as five and three. Therefore the CURE with K-Means clustering algorithm performs well because it contains high clustering accuracy when compared to BIRCH with K-Means.

B. Outlier Accuracy

TABLE II
THE OUTLIER ACCURACY FOR BIRCH +K-MEANS, CURE+K-MEANS IN THREE AND FIVE WINDOWS

Outlier accuracy	Window size	No. of windows	BIRCH+ K-MEANS	CURE + K-MEANS
Detection rate	Three	W1	32.00	36.72
		W2	33.00	36.12
		W3	31.00	33.80
	Five	W1	32.40	33.89
		W2	39.40	44.03
		W3	33.50	37.61
		W4	24.57	25.54
		W5	33.60	35.50
False alarm rate	Three	W1	50.00	38.10
		W2	36.06	30.00
		W3	40.00	27.91
	Five	W1	44.44	38.88
		W2	50.00	38.00
		W3	51.12	34.00
		W4	27.00	23.00
		W5	44.00	35.55

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

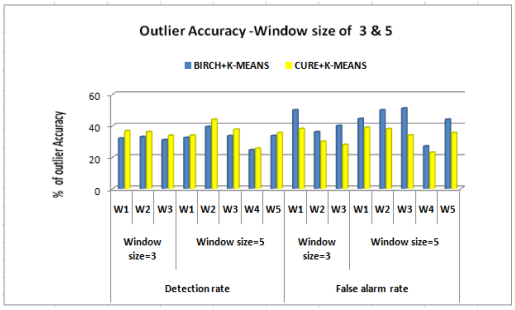

Fig 3: The outlier accuracy for, BIRCH +K-Means and CURE+K-Means

V. CONCLUSION

Data streams are active ordered, fast changing and gigantic, immeasurable and infinite sequence of data objects. Data streams clustering technique are highly helpful to handle those data and outlier detection. Hierarchical clustering and partition clustering are helpful to detect the outliers in a prominent manner. In this paper, we have analysed the performance of BIRCH with K-Means and CURE with K-Means clustering algorithms for detecting the outliers. In turn to find the efficient clustering algorithm for outlier detection two performance measures are carried out. From the experimental results, it is observed that the accuracies of clustering and outlier detection is more efficient in CURE with K-Means clustering while compared to BIRCH with K-Means clustering.

From the above figure 2, it is observed that CURE with K-Means clustering algorithm performs better than BIRCH with K-Means algorithms for detecting outliers in Pima Indian Diabetes dataset for both window size as five and three. Therefore the CURE with K-Means clustering algorithm performs well because it contains high clustering accuracy when compared to BIRCH with K-Means.

REFERENCES

[1] Aggarwal.C, Ed., "Data Streams – Models and Algorithms", Springer, 2007.
[2] Aggarwal.C.C, J. Han, J. Wang, and P. S. Yu,"A framework for clustering evolving data streams," In Proc. of VLDB, pages 81-92, 2003.
[3] Chandrika.J, Dr. K.R. Ananda Kumar, "Dynamic Clustering Of High Speed Data Streams", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.
[4] C. J. Merz and P. M. Murph, UCI Repository of Machine Learning Databases Univ. of CA,Dept. of CIS, Irvine.
[5] Hendrik Fichtenberger, Marc Gillé , Melanie Schmidt ,in Algorithms – ESA 2013 , Volume 8125, 2013, pp 481-492.
[6] Irad Ben-Gal, "Outlier Detection", Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer , Academic Publishers, 2005.
[7] Luis Torgo, Carlos soares, "Resource-bounded Outlier Detection using Clustering Methods", proceedings of the conference on data mining for business applications, 2010.
[8] Larose D.T, "Discvoery knowledge in data-Introduction to Data mining, ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.
[9] Mahnoosh Kholghi, Mohammadreza Keyvanpour," An Analytical Framework For Data Stream Mining Techniques Based On Challenges And Requirements" in International Journal of Engineering Science and Technology,2011.
[10] Madjid Khalilian, Norwati Mustapha "DataStream clustering-Challenges and issues",Proceedings of the International Multi Conference of Engineers and ComputerScientists 2010 Vol I,IMECS 2010,March 17 -19, 2010, Hong Kong.
[11] Neha Gupta,"Stream Data Mining:A Survey",Indrjeet Rajput ,International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 1, January -February 2013, pp.1113-1118 .
[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] Thakran. Y, Toshniwal .D, " Unsupervised outlier detection in streaming data using weighted clustering", Intelligent Systems Design and Applications (ISDA), 2012.
[14] Yi-hong lu, Yan huang, "Mining DataStreams Using Clustering", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics,vol.4, pp. 18-21,2005.
[15] Yogita, Durga Toshniwal, "Clustering Techniques for Streaming Data–A Survey" in proc. Of the IEEE, 2012.