International Journal of Advance Engineering and Research Development (IJAERD)
Volume 1,Issue 6,June 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

K-means based data stream clustering algorithm extended with no. of cluster estimation method

Makadia Dipti[1], Prof. Tejal Patel[2]

[1]*Information and Technology Department, G.H.Patel Engineering College, V.V.Nagar, India*
dipa.makadia@gmail.com
[2]*Information and Technology Department, G.H.Patel Engineering College, V.V.Nagar, India*
tejalrpatel@gcet.ac.in

Abstract: Data stream are generated from many sources. This Data streams are needed to be transformed into significant information to take more effective decisions. Clustering is the best way for analyzing data streams. The material on clustering is very large.Many clustering algorithms are available for data stream which uses k-means algorithm as a base. Clustream algorithm is one of the examples of it. Main drawback of such k-means based data stream clustering algorithm (Clustream) is that user has to give no. of cluster (k) in advance. Many times it happens that user does not know detail about the data and gives value of k randomly. In this type of case we will not get satisfactory result. i.e. we can't get proper quality of clusters. To tackle the above mentioned problem, we have proposed the framework. According to it, we will use another algorithm to find appropriate no. of clusters in advance. Here we used Bisecting k-means algorithm to find no. of clusters for data stream. So we have combined the clustream algorithm with bisecting approach for finding best quality clusters without interference of user to fix value of no. of cluster at user side.

I. INTRODUCTION

So many resources such as real-time surveillance systems, communication networks, Internet traffic, on-line transactions in the financial market or retail industry, electric power grids, industry production processes, scientific and engineering experiments, remote sensors, and other dynamic environments generate High volume and potential infinite data streams. In comparison with traditional data sets, data stream are dynamic in nature. Traditional data sets are easy to store but data stream are massive, so it is not easy to store. Many data mining techniques for streaming data are available like clustering, classification, frequent pattern mining, outlier detection etc. In this paper we will focus on clustering technique only.

Now let's start with properties of data stream. Fundamental Data Stream Properties are as per given: [2]

Unboundedness

Data streams are potentially unbounded and can thus generate an infinite amount of data. That means it is not possible to store stream entirely.

High data generation rate

In some applications, stream elements are generated at a rapid rate. The elements thus have to be processed in a timely manner in order to keep up with the stream rate. Usually a single scan of such data stream is necessary.

Evolving nature

In most applications, the characteristics of the data stream as well as its elements evolve over time. This property is referred to as temporal locality and adds an inherent temporal component to the data stream mining process. Stream elements should thus be analyzed in a time-aware manner to accommodate the changes in stream characteristics.

Due to huge amount and high storage cost, it is impossible to store an entire data streams or to scan through it multiple times. So it makes so many challenges in storage, computational and communication capabilities of computational systems. Because of high volume and speed of input data, it is needed to use semi-automatic interactional techniques to extract embedded knowledge from data. There is a need of effective and efficient data mining clustering techniques for streaming data which can handle the challenges associated with streaming data [1].

In computer science, data stream clustering is defined as the clustering of data that arrive constantly for instance telephone records, multimedia data, financial transactions etc. Data stream clustering is generally studied as a streaming algorithm and the goal is, for given sequence of points, to create a good clustering of the stream, using a small amount of memory and time. Data stream clustering has recently attracted attention for up-and-coming applications that involve large amounts of streaming data.

II. REVIEW OF ALGORITHMS USED

Clustream Algorithm

CluStream [4] is proposed by Aggarwal et al. in 2003. Before the idea of that is proposed there were problems in data stream clustering algorithm in following manner. a)we get poor quality clusters when the data evolves considerably over time.b)data stream algorithm requires much better functionality in discovering and exploring clusters over different portions of the stream. This paper discusses a basically different philosophy for data stream clustering. The idea is divide the clustering process into two components i.e. a) online component and b) offline component. An online component which periodically stores detailed summary statistics and an online component which uses only this summary statistics. The offline component is used by the analyst who can use a wide variety of inputs (such as time horizon or number of clusters) in order to provide a quick understanding of the broad clusters in the data stream. The problems of effective choice, storage, and use of this statistical data for a fast data stream turns out headed for be quite tricky. For this purpose, authors have used the concepts of a pyramidal time frame in conjunction with a micro-clustering approach. Statistical information about the data locality is maintained in terms of micro-clusters. These micro-clusters are defined as the temporal extension of the cluster feature vector. The micro-clusters are stored at snapshots in time following pyramidal pattern. This type of pattern provides an efficient trade-off between the storage requirements and the ability to recall summary statistics from different time horizons. This summary information in the micro-clusters is used by an offline component which is dependent on a wide variety of user inputs such as the time horizon and no. of k using k-means algorithm.In this paper, obtained result shows that CluStream can achieve higher accuracy than STREAM algorithm.

No. of Cluster Estimation method

Bisecting Algorithm

Main steps of Bisecting Algorithm is as per given below. **[11]**
Step-1 initially starts with single cluster formed by all the data set objects.
Step-2 Find 2 sub cluster using basic k-means algorithm.
Step-3 Repeat above mentioned bisecting step, for ITER times and select the split that produces the clustering with highest overall similarity.
Step-4 Repeat 1st, 2nd,3rd steps until the $K\max$ is reached.

There are many ways to choose which cluster to split. For e.g., we can choose the largest cluster at each step for split. We will consider $K\max = \sqrt{N}$ where N=no. of data objects. In order to estimate best value of K from set {2,...., $K\max$ } Silhouette coefficient is used.

III. EXPERIMENTAL SETUP

For Data Stream Clustering purpose, a framework for stream learning evaluation was recently introduced, called Massive Online Analysis (MOA).We have implemented our algorithms in java and evaluated using MOA tool.

Real Data Set

We have used real dataset to obtain results. We have used Covertype dataset to test the results. It is easily available at UCI Machine Learning Repository.

This dataset contain 5,81,012 data points and 54 numeric attributes. This data set is converted into data stream by giving data input in chunk of data.

Evaluation Parameter Used

We have used two parameters to evaluate the algorithm. These two parameter are :

I. Silhouette co-efficient
II. Processing Rate

I. Silhouette co-efficient [10]

The Simplified Silhouette criterion measures how compact and separate the clusters of a given partition are. In this sense, there is a synergy between this index and the k-means algorithm, as both favours the same kind of partition.

To explain this index, let us consider an object Xj belong to cluster Ci. The dissimilarity between Xj and the centroid $\overline{Xi}$ of Ci is denoted by $a(Xj)$, whereas the dissimilarity between Xj and the centroid $\overline{X_l}$ of another cluster C_l is termed $d(Xj, \overline{X_l})$.After computing $d(Xj, \overline{X_l})$ for all clusters the lowest one is retained and termed $b(Xj)$ i.e., $b(Xj) = \min d(Xj, \overline{X_l})$.

This value represents the dissimilarity between Xj and it's nearest neighbouring cluster. Once $a(Xj)$ and $b(Xj)$ have been introduced, the Simplified Silhouette $s(Xj)$ can thus be defined as:

$$s(X_j) = \frac{b(X_j) - a(X_j)}{\max\{a(X_j), b(X_j)\}}$$

It's now easy to define that value of silhouette co-eff. should be between [0, 1] for compact cluster. If the value is near to 1 then it will have good quality cluster.

II. Processing Rate

Processing rate can be defined as the no. of objects processed per second.
We will use processing rate as the evaluation parameter of the algorithm to find performance of it.

IV. RESULTS

As we have already discussed ,we are using real dataset Covertype. we have considered only numeric attributes of data set. we have compared clustream algorithm and clustream combined with bisecting method .As we know, clustream algorithm asks for no. of k at user side. when clustream algorithm extended with bisecting k-means method then it will not asks for no. of k at user side. Here results shows clustream (with different values of k=10,25,50) and clustream combined with bisecting approach.)

As shown in Fig.1 Clustream Algorithm has different values for different no. of clusters (k) input. It's highest value is about 0.7.but when clustream is combined with Bisecting Approach it's silhouette value becomes about 0.85.it shows that quality of clusters improves .

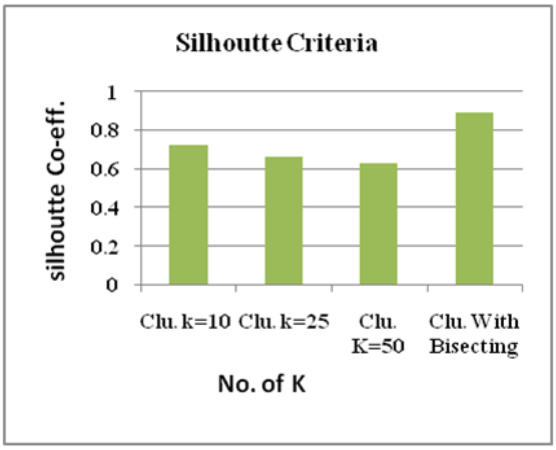

Fig.1 Silhouette Criteria Comparison

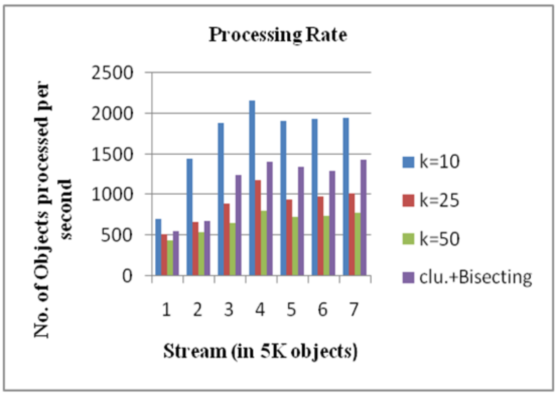

Fig. 2 Processing Rate Comparison

Though quality of clusters improves while combining Clustream with Bisecting, it also gives good processing rate as shown in Fig.2.

V. CONCLUSION

Now a days, hardware technology become more advance. So many resources such as real-time surveillance systems, communication networks, Internet traffic, on-line transactions in the financial market or retail industry, electric power grids, industry production processes, scientific and engineering experiments, remote sensors, and other dynamic environments generate high volume and probably infinite data streams. Many algorithms for clustering data streams based on the k-Means have been proposed up till now. But, most of the literature assume that number of clusters, k, is known and fixed in advance at the user side. But this type of assumption is not reasonable in real life application. So, we have combine data stream clustering Clustream algorithm with no. of cluster estimation bisecting method. So such type of framework will provide best value of k automatically estimated from data at user side. Though it improve the quality of the clustering in terms of silhouette criteria without lowering the processing rate.

REFERENCES

[1] 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.

[2] Conny Franke ,Adaptivity in Data Stream Mining ,2009

[3] L. callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani,"Streaming-Data Algorithms for High-Quality Clustering," inProceedings of IEEE International Conference on Data Engineering,2001.

[4] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the 29th international conference on Very large data bases - Volume 29, ser. VLDB '03. VLDB Endowment, 2003, pp. 81–92.

[5] Chunyu Yang , Jie Zhou "HClustream: A Novel Approach for Clustering Evolving Heterogeneous Data Stream"in Procedding of Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06),2006.

[6] M. R. Ackermann, C. Lammersen, M. M¨artens, C. Raupach, C. Sohler, and K. Swierkot, "StreamKM++: A Clustering Algorithms for Data Streams," in Proc. of the ALENEX, 2010, pp. 173–187.

[7] Yogita, Durga Toshniwal,"Clustering Techniques for Streaming Data–A Survey"in proc. Of the IEEE,2012.

[8] Philipp Kranen, Hardy Kremer, Timm Jansen, Thomas Seidl, Albert Bifety, Geoff Holmesy and Bernhard Pfahringery," Clustering Performance on Evolving Data Streams:Assessing Algorithms and Evaluation Measures within MOA" in IEEE International Conference on Data Mining Workshops,2010

[9] Shifei Ding ,Fulin Wu ,Jun Qian , Hongjie Jia ,Fengxiang Jin," Research on data stream clustering algorithms"in Springer Science+Business Media,2013

[10] M.C. Naldi,R.J.G.B. Campello,E.R. Hruschka, A.C.P.L.F. Carvalho "Efficiency issues of evolutionary k-means" in Applied Soft Computing(Elsevier), 2010.

[11] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. of the KDD Workshop on Text Mining,2000, pp. 109–111.