ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 88-91

An Efficient Hierarchical Clustering Using Tree Observation Technique

[1] Dr.V.Kavitha , [2]S.Subhasini,[3] J.S.Anithalilly,[4]Krithik Roshan.P

[1]Associate Professor, [2]Assistant professor, [3]Assistant professor, [4]Student
[1]Department of MCA, [2]Department of computer applications (BCA), [3]Department of computer applications (BCA), [4]Department of computer applications (BCA)
[1]Hindusthan College of arts and science, [2]Hindusthan College of arts and science, [3]Hindusthan College of arts and science, [4]Hindusthan College of arts and science
[1]Coimbatore,India ,[2]Coimbatore,India,[3]Coimbatore,India. [4]Coimbatore,India

Abstract: *Data stream is an ordered sequence of data objects that can be read only once or a small number of times.The characteristics of data stream are very large, continuous, high dimensional, immeasurable, dynamically high speed and massive amount of data in offline and also in online and there is not sufficient time to rescan the entire database. Data stream are required to store vast amounts of data that are continuously inserted and queried. Due to the above features of data stream, obtaining the fruitful information is a critical process. Hence, analyzing huge data sets and extracting valuable pattern in many applications are interesting for researchers. Moreover, the dynamic high speed of time series data stream is controlled when using clustering technique. And also clustering trace out the minimal variation in data stream which leads to save the quality of the cluster structure and restricting is not required. Hence, the technique of tree observation is applied and implemented with hierarchical clustering and it performed and compared with some of the performance factors and prove that its efficiency.*

I. Hierarchical Clustering

Hierarchical clustering technique performs by assembling time series stream data objects into a tree structure form of clusters. Hierarchical Clustering is subdivided into agglomerative hierarchical clustering and divisive hierarchical clustering. Agglomerative hierarchical clustering approach is based on the fashion of bottom up merging process. Whereas the next criteria of divisive hierarchical clustering is belongs to the strategy of top down splitting process. The hierarchical clustering process is suffered from the its incapability to carry out amendment once a merge or split decision has been executed. Therefore the pure hierarchical clustering performance leads to poor quality. That is, if a specific merge or split assessment later turns out to have been a poor selection and it cannot do the back tracking process and corrects it.

Hierarchical clustering techniques suffer from the fact that once a step that is merge or split is completed, it can never be undone. This inflexibility is informative in that it leads to minor commencement costs by not having to be troubled about a combinatorial number of various choices. Conversely, such methods are unable to correct erroneous assessment. Hence, the hierarchical clustering approach needs to concentrate the two performance factors for improving cluster quality.

1) Carry out careful investigation of the stream data object, “linkages” at each hierarchical partitioning.
2) Incorporate hierarchical agglomeration and other approaches by first using a hierarchical agglomerative technique to cluster stream data objects into micro cluster and then achieving macro clustering on the micro cluster using any other clustering techniques.

Hence, the performance of hierarchical clustering technique is enhanced with the help of any other clustering approaches like partitioning clustering, grid clustering or density subspace clustering. This kind of integration leads to improve the cluster quality.

II. Comparing Hierarchical and Non hierarchical approach

The clustering techniques are generally classified into two approaches. Namely

i) Hierarchical Clustering Approach
ii) Non Hierarchical Clustering Approach

Hierarchical Approach

The hierarchical approach the algorithm should constructs a tree structure or a hierarchy to view the relationship between the database entities. Observations or individuals are denoted as entities. Diameter is measured in this hierarchical approach. It denotes the maximum distance among the two data points of the specified cluster. Based on the diameter only the tree like structure will start to grow.

Non Hierarchical Approach

In the non hierarchical approach the clustering will be formed based on the centroid point of the cluster. And then distance is considered from that centroid point. Non hierarchical approach is having less population due to fixing the optimal central point position is a major challenge.

Related Clustering Technique

The time series data are extensively available in several real world domains for instance financial, medical, and science etc. Time series data are obviously very large for data analysis and also it is having added number of observations. Hence, the traditional clustering algorithms will not support the new arrival of fast time series data stream. For that reason the innovative clustering algorithms are needed to process the nature of biomedical applications. One of the mining clustering technique involves vast amount of time series data stream and coined the best outcome by using the proposed system.

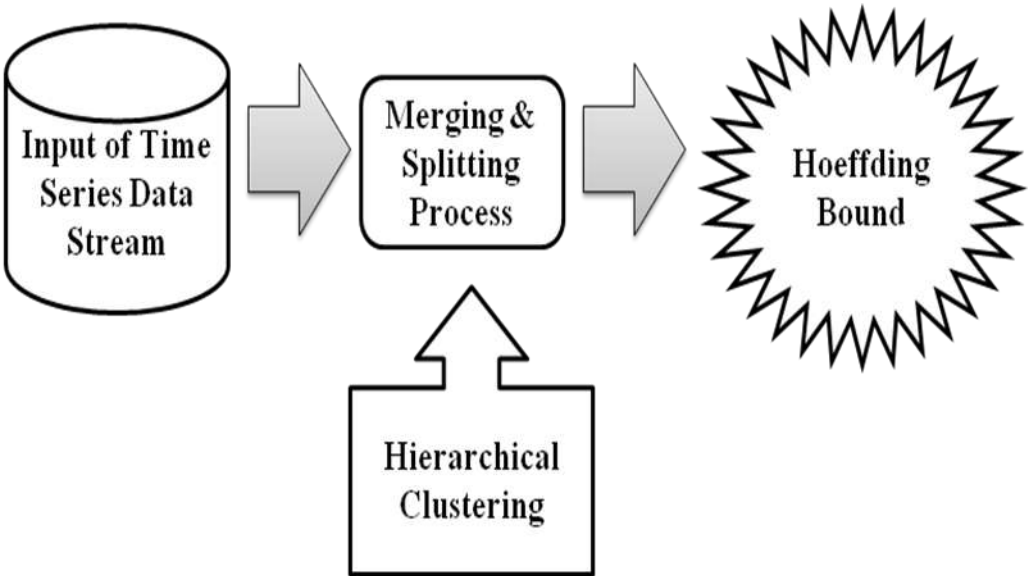

Figure 1 System of Hierarchical Clustering

In figure 1 describes, the existing method is named as ODAC algorithm (Online Divisive Agglomerative Clustering) which is an incremental monitoring and hierarchical clustering based algorithm. This algorithm used to clustering the data stream and it assembles a tree structure shaped cluster using hierarchical clustering approach. The algorithm shelter a distance incremental measure and carry out the aggregation and expansion procedures of the tree shaped structure. In this approach the clustering is formed with the help of manipulating diameters that is discovered by the distance similarity measures. This hierarchical tree structure continuously monitoring the already existing clusters diameter with in a time interval. Diameter is nothing but the maximum distance between the two time series data points of the specified cluster. At the time of growing the hierarchical tree, the system will discover the diameter of the previous cluster for each and every level. According to the best diameter value the algorithm will be decided whether it is necessary to do the splitting process or merging task.

III. Tree Observation Hierarchical Clustering

In this Tree Observation Hierarchical Clustering system the time series data set and number of clusters are input parameters. Hierarchical Clustering is based on incremental clustering. It is used for analyzing and extracting the knowledge from the fast moving time series continuous data streams. This algorithm constructs a hierarchical tree shaped structure of cluster by using a top down strategy.

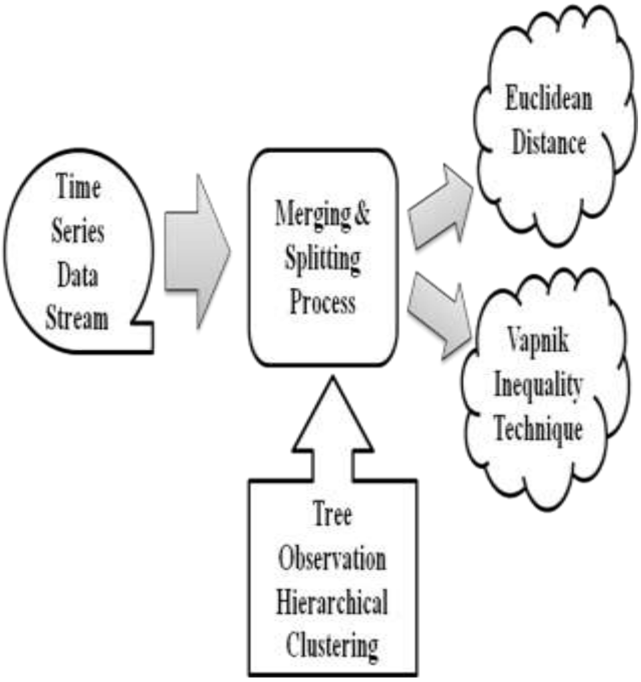
Time Series Data Stream
Merging & Splitting Process
Euclidean Distance
Vapnik Inequality Technique
Tree Observation Hierarchical Clustering

Figure 2 Tree Observation Hierarchical Clustering

In figure 2 explains about the main hypothesis of the system is that decisions are taken over the data set under certain conditions and to monitor the growth of hierarchical tree structure. The system continuously monitors the clusters diameter over time. The diameter of the cluster is the maximum distance between the variables of that cluster. At a specific time, if a given condition is met on this diameter, the system will split the cluster and assign each of the chosen variables for a new cluster.

On stationary data stream, the overall intra cluster similarity should decrease with each split. In this way the system will decide whether to use the split or merge function to grow the hierarchical tree structure. Tree observation hierarchical clustering system must analyze distance between the time series data points. These distances must be computed incrementally. The diameter between the data points will be scaled using the distance measure. The performances of various kinds of distance measures are analyzed and among them the Euclidean distance measure is preferred for determining the cluster quality.

Since, the system has to make decisions with statistical support using Vapnik Inequality technique which leads to observe the growth of cluster tree structure. Threshold value will be predetermined in the existing technique for all iterations of the tree structure whereas it is dynamic in Vapnik Inequality technique which signifies that the threshold value will be varying for all iteration of the tree structure. The first objective “To formulate an efficient hierarchical clustering technique for discovering the best diameter is very much essential for fruitful clustering. Moreover this approach enhances the extraction of information” is achieved.

IV. Cluster Quality Measure

Quality of the cluster is categorized into two techniques, which is used to measure with the assistance of distance measure between the stream data objects. The two techniques of cluster quality are,

- Inter Cluster Quality
- Intra Cluster Quality

Inter Cluster Quality

The distance between the different type of various cluster will be measured in this inter cluster quality. The state of this cluster distance must be maximized then only that measure will be recommended as a optimal distance. The strategy of inter cluster quality will be depicted in the following figure. The general formula to measure the inter cluster distance is,

$$\text{Inter Cluster} = \sum_{k=1}^{n} \sum_{i,j=1}^{N} \frac{|d_{i,k} - d_{j,k}|}{N * n}$$

Where N denotes a number of points inside the cluster. And n is termed as number of clusters.

Intra Cluster Quality

The distance between the cluster stream data objects with in a cluster will be measured in this intra cluster quality. The state of this cluster distance must be minimized, and then the measure will be suggested as a optimal distance.

The general formula to measure the intra cluster distance is,

$$\text{Intra Cluster} = \sum_{i,j=1}^{n} \frac{|C_i - C_j|}{n}$$

Where n is number of data objects within a cluster. Ci and Cj are each and every isolated data point of the cluster.

V. Conclusion

Data stream are required to store vast amounts of data that are continuously inserted and queried. Due to the above features of data stream, obtaining the fruitful information is a critical process. Hence, analyzing huge data sets and extracting valuable pattern in many applications are interesting for researchers. Moreover, the dynamic high speed of time series data stream is controlled when using clustering technique. And also clustering trace out the minimal variation in data stream which leads to save the quality of the cluster structure and restricting is not required. Hence, the technique of tree observation is applied and implemented with hierarchical clustering and it performed and compared with some of the performance factors and prove that its efficiency. The performance factor of cluster quality which means inter cluster and intra cluster and time complexity is compared with various clustering techniques and concluded that tree observation hierarchal clustering technique is optimal one.

References

[1]. Pantelis n.Karamolegkos, Charalampos Z.Patrikakis Nikolaos D.Doulamis Panagiotis, “An Evaluation Study of Clustering Algorithms in the Scope of user Communities Assessment” Computers $ Mathematics with Applications, Elsevier, Vol No 58, issue no 8, October 2009, Pages 1498 - 1519.

[2]. Man Abdel - Maksoud, Mohammed Elmogy, Rashid Al-Awadi, “Brain Tumor Segmentation Based on a Hybrid Clustering Technique”, Egyptian Informatics Journal, Vol No 16, Issue no 1, March 2005, Pages 1 - 81.

[3]. Madjid Khalilian, Norwati Mustapha, Data Stream Clustering: Challenges and Issue, Proceedings of the International Multi conference of Engineers and Computer Scientists 2010 Vol No1, IMECS 2010,March 17-19 2010.

[4]. Maryam Mousavi1 , Azuraliza Abu Bakar, and Mohammadmahdi Vakilian, “Data Stream Clustering Algorithms: A Review”, International Journal of Advance Soft Computer Applications Vol o 7, Issue No 3, November 2015, ISSN 2074-8523.

[5]. Jose R. Fernandez,” A Framework and Algorithm for Data Stream Cluster Analysis”, International Journal of Advanced Computer Science and Applications, Vol No 2, Issue No11, Pages 87, 2011.

[6]. Twinkle B Ankleshwaria, Twinkle B Ankleshwaria, Mining Data Streams: A Survey, International Journal of Advance Research in Computer Science and Management Studies, Vol No 2, Issue No 2, Feb 2014, ISSN: 2321-778.

[7]. Amineh Amini, Teh Ying Wah, “Density Micro-Clustering Algorithms on Data Streams: A Review”, Proceedings of the International MultiConference of Engineers and Computer Scientists 2011 Vol No 1, IMCES 2011, March 16-18, 2011.

[8]. Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., de Carvalho, A. C. P. L. F., and Gama, J, “Data stream clustering: A survey”, ACM Computing Surveys, Vol No 46, Issue No1, Article 13, October 2013, Pages 31.

[9]. DoniaAugustine, “A Survey on Density based Micro-clustering Algorithms for Data Stream Clustering”, International Journal of Advanced Research in Computer Science and Software Engineering Research, Vol No 7, Issue No 1, January 2017.

[10]. Dure Supriya Suresh, Prof. Wadne Vinod, “Survey Paper on Clustering Data Streams Based on Shared Density between Micro-Clusters”, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Vol No 04 ,Issue No 01, January 2017.

[11]. Amini A, Wah TY, Saboohi H, “On density-based data streams clustering algorithms: A survey”,Journal of Computer Science and Technology, Pages 116–141, January 2014, DOI 10.1007/s11390-013-1416-3.

[12]. Safal V Bhosale, “A Survey: Outlier Detection in Streaming Data Using Clustering Approache”, International Journal of Computer Science and Information Technologies, Vol No 5, 2014, 6050-6053 ISSN 0975 - 9646.

[13]. Prashant V. Desai, Vilas S. Gaikawad, “Novel approach for data stream clustering through micro-clusters shared Density”,International Journal of Computer Sciences and Engineering Volume-5, Issue-1 E-ISSN: 2347-2693.

[14]. M.S.B.PhridviRaj, C.V.GuruRao, “Data Mining - Past, Present and Future - A Typical Survey on Data Streams”, Elsevier Procedia Technology”, Vol No 12, 2014, Pages 255 - 263.

[15]. Yisroel Mirsky, Bracha Shapira, Lior Rokach, and Yuval Elovici, “pcStream: A Stream Clustering Algorithm for Dynamically Detecting and Managing Temporal Contexts”, Springer International Publishing Switzerland 2015, PAKDD 2015, Part II, LNAI 9078, pp. 119–133, 2015. DOI: 10.1007/978-3-319-18032-8_10.

[16]. Shufeng Gong, Yanfeng Zhang, Ge Yu1, “Clustering Stream Data by Exploring the Evolution of Density Mountain”, PVLDB, 11(4) 2017. DOI: 10.1145/3164135.3164136.