ISSN 2394-187(Online)
ISSN 2394-7179 (Print)

Online Clustering Data Streams

M.Anitha[#1], S. Sridhar[*2]
[1]*Master Of Computer Applications, S.A. Engineering College, Chennai-77.*
anithaani1026@gmail.com
[2]*Asso. Prof., Department of Computer Applications, S.A. Engineering college, Chennai-77.*
sridhar@saec.ac.in

Abstract— We consider the problem of online clustering high speed data streams. A data stream can be thought of as a fleeting, incessantly rising succession of time-stamped data. To uphold the modern clustering construction, it is essential to examine the arriving data in an online manner, tolerating but a constant time delay. Purpose of this study is to analyze the working of popular algorithm on clustering data streams where make a comparative analysis.

Keywords— Data streams, Partional Clustering, Hierarchical Clustering.

1. Introduction

Online clustering data streams structure is forced by three foremost limited resources: time, remembrance and sample size. State of the art job on mining data streams contemplates on detain ing time-evolving drifts and patterns with "labeled" data. If there is neither intangible nor distributional modify, periodic inactive model revive and re-validation is dissipate of resources. Normally, the data streams have the subsequent distinctiveness:1) data are pending incessantly and at a very speedy rate,2) the extent of the data is limitless, and 3) data are instance unreliable. Because particularly the customary database executive system is intended for fixed and unrelenting datasets, the errands of managing data streams will provide the researchers with many of new challenges and the opportunities. During the recent years, the so called data streams where have attracted with considerable attention in the different fields of the computer science, such as e.g. database or distributed systems.

There are various applications in which the stream of this type are produced, such as the network monitoring, the telecommunication systems, the stock markets, and the customer click streams, for eg: any type of multi-sensor system. A data stream system where may be create continually a enormous quantities of the data. Imagine a multi-sensor system with 10,000 sensors, each of the system which sends a measurement every second of time. As concerned aspects of the data storage, the management and processing, the varying and of potentially unbounded streams raises new challenges and research problems. It has three main limited resources: time, memory and sample size. In traditional applications the of machine learning and the statistics, sample size tends to be the dominant limitation: the computational resources for a massive search where available, but carrying out of these search over the small samples are available often leads to the overfitting.

2. Data Stream Model

The data stream model assumes that the input data are not available for the random access from the disk or memory, but rather than arrive in the form disk or memory, but rather than disembark in the shape of one or more incessant information streams. The data stream model is differs from the standard relational model where are in the following ways.

The elements of the data stream that arrive in online (the stream is "active" in the sense that the incoming items are trigger operations on the data, rather than being send an request).

- The order in which elements of a stream arrive are not below the control of the structure.
- Data streams when potentially limitless in amount.
- Data stream elements that have been processed are either discarded or archived.
- Due to imperfect possessions (reminiscence) and severe time restraints, the dispensation of stream data will typically create estimated results.

3. Existing Work

In this for clustering we have introduced an online version of the k-means algorithm A key aspect of our method is an efficient incremental computation of the distance among such streams, using a DFT estimate of the unique data. This way, it becomes possible to cluster several thousands of data streams in real-time. In order to investigate the performance of our approach we are currently performing several controlled experiments with synthetic data. Especially, these experiments are meant to throw light on the scalability of the process and on the transaction between runtime intricacy and eminence of the clustering. All in all, our first outcomes demonstrate that the process realizes an tremendous achieve in intricacy at the cost of an acceptable (often very small) loss in quality. (In case we believe, the results of the article will be included in the final account of.)We are also expanding our process in numerous directions. Inter alia, we are executing a fuzzy version of *K*-means . A object can belong to more than one cluster in the fuzzy cluster and the degree of membership in each cluster can be characterized by means of a number in between 0 and 1. This extension of the standard approach appears particularly reasonable in the context of online grouping: If a data stream shifts from one group to an other cluster, it usually does so in a "smooth" rather than abrupt manner.

In the paper of dynamic clustering of high-speed data streams is fully about online clustering. There is some disadvantages. There is small delay in replying for Queries. Data streams should be adaptive. The data's cannot be retrieved easily. It will produce only approximate results. The order of data arrival is not under the control of the system. Here the measures are costly. Here also they used k-means algorithm and CURE algorithm. The research in this field is mainly done in the areas like modeling, querry processing and mining data streams.

4. Problem Identification

Data clustering symbolized most frequently encountered problems in the data analyses. The basic problem is the data points of "clusters" should be partition from a set of data, that are seems "close" for other data it is relatively "far from" it. A more general problem is the so-called cluster identification problem which is to identify "opaque" clusters in a perhaps loud backdrop.

5. Proposed Work

In this online clustering data streams there is little disadvantages. It will be fulfilled in this paper. In this data can be retrieved easily. The arrival of error is under the control of the system. Here the cost of measure will be reduced. It can be used in Marketing, Biology, Libraries, etc….In this results are up-to-date. This more efficiency. There is no time delay in replying of querries.

Our online clustering method has been implemented under K-means algorithm. The final is a JAVA records for inquiry dispensation urbanized and preserve, and in the meantime exploited by many frequent explores groups. Discretely from some index constructions, this records presents fixed operators for generate, uniting and dispensation queries. As the name proposes, the essential values of XXL is its flexibility and the possibility to easily extend the library. Currently, for the handling of active data streams a package is developed called PIPES. For our implementation we have used PIPES. Roughly speaking, the process of clustering corresponds to an machinist. As inputs this machinist obtains (blocks of) data items from the (active) data streams. It produces the current cluster number for each stream as the output. Basically, this tool shows a table whose rows correspond to the data streams. We can be seen here is the top of a list containing of 230 stock rates and whose overall length is 230. Each column will stands for one cluster, and a square indicates for stream's current cluster is. If the clustering structure changes the square moves horizontally each time .

We have to characterized each cluster by means of a prototypical stream. As we can be seen, in the example there are currently three types of stock rates where they showing a similar evolution of the size of the sliding window over the last 16 hours. The first kind, for occasion, reduced somewhat throughout the first half of that time period and then began to increase until so far. A second visualization tool, shows the recent history of the clustering structure, but it will not show only the existing state.

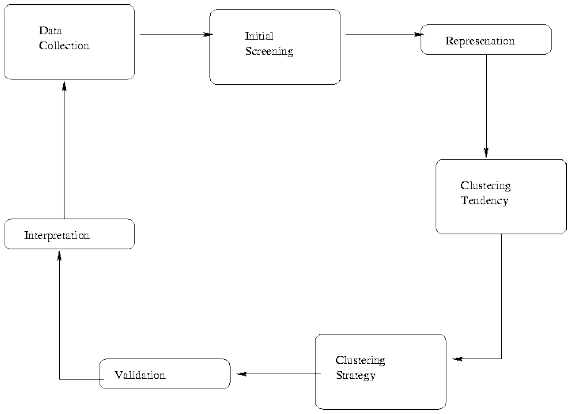

Fig.1: The steps of clustering

Data collection: This step requires careful recording of data. Initial screening: Before making the raw data ready for analysis usually it needs some massaging . The analysis techniques should be applied rather than cluster analysis if the facts cannot be revealed to have the propensity to huddle. *Clustering plan:* This system used to decide the precise clustering algorithm. Thought must be given to details such as matching the algorithm to the data, the presentation of results and the choice of parameters. *Validation:* This step changes the analysis into solid proof. constancy is one of the basis for contrasting clustering methods. *Interpretation:* Drawing conclusion from the analysis. This depends on the application..

6. Clustering

A group of data points is partitioned into a small number of clusters with the help of clustering process. Clustering is unsupervised learning. It is the procedure of grouping a collection of objects into classes or "clusters" such type of objects from different classes are dissimilar. To arrange the clusters into a natural hierarchy is the main goal. The basic three types of clustering algorithms are: Mixture modeling assumes that an underlying probabilistic model, is namely that the data were be generated by the probability density function, which is a mixture of the component density function. Marketing: Finding group of customers with the similar where behavior given in a large database of customer data containing with their properties and past buying records. Biology: It is used to the Classification of plants and animals given their features. Libraries: It is used for Book ordering in libraries.

7. Partitioning

The work of Partitioned clustering is to decomposes a set of clusters into a set of disjoint clusters. Given a database with a set of objects, a Partition clustering algorithm constructs a k partitions of the n^{th} data, where each cluster will optimizes a clustering criterion, such as the minimization of *sum of the squared distance from the mean of* within a each cluster.

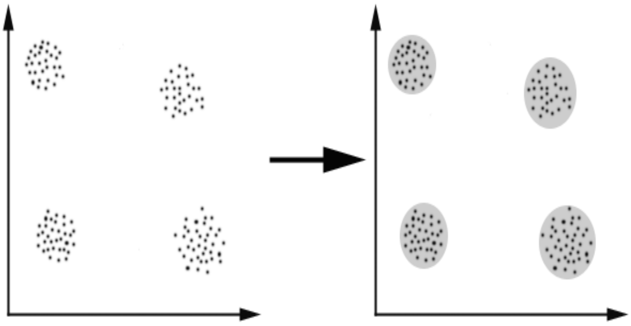

Fig.2: Clustering Diagram

Algorithm K - Means:
Input:
D = {t1,t2,…tn} // set of elements
K // Number of preferred clusters
Output:
K // set of clusters
K – Means algorithm:
Assign initial values for means m1,m2,…mk;

8. Hierarchical

It is a method of cluster analysis which seek to build a hierarchy of clusters. Hierarchical algorithms create a hierarchical putrefaction of the matters. They are also agglomerative (bottom-up) or divisive (top-down). *Repeat:* ti has the closest mean so it is assign each time to the cluster ;calculate the new mean for each cluster; *until* convergence criteria is met; Initiate the center of the clusters. Attribute the closest cluster to each data point. The position of each cluster have set to the mean of all the data points belonginig to that clusters.

9. Cure

To avoid problems with non-uniformed size (or) shaped clusters. CURE employees a hierarichal clustering algorithm that adopts a middle ground between the centroids based and all point extremes. The algorithm cannot be applied to large database because of high complexity.

Input:
D = {t1,t2,…tn } // set of elements
K // Desired number of clusters
Output:
Q // Heap containing k-clusters with
one entry for each cluster
CURE algorithm:
T = build(D);
Q = heapify(D); // Initially heap have to be build for one admission per thing;
replicate
u = min (Q);
erase (Q, u.close);
w = merge(u,v);
delete (T,u);
delete (T,v);
insert (T,w);
for each x ε Q do
x.close = find closest to x;
ifx is closest to w, then
w.close = x;
insert(Q,w);

Table.1: Cluster Result

No.of points	**1572**	**3568**	**7502**	**10256**
Time (in sec)				
Partition p=2	**6.4**	**7.8**	**29.4**	**75.7**
Partition p=3	**6.5**	**7.6**	**21.6**	**43.6**
Partition p=5	**6.1**	**7.3**	**12.2**	**21.2**

10. Conclusion

In this paper we have discussed about problem of online clustering data streams. The research in this field is mainly done in the areas like modeling, query processing and mining the data streams. Traditional data stream algorithms are challenged by the feature of data streams. So the conventional techniques for data mining needs to be molded according to the needs of data streams. The

ISSN 2394-187(Online)
ISSN 2394-7179 (Print)

properties like infinite data flow and drifting concepts and make the life of these researchers tough. To overcome these difficulties we have presented some of the recently developed and experimentally proved approaches for dealing with data streams.

References

[1] G Cormode, "Fundamentals of Analyzing andMining Data Streams", Workshop On Data Stream Analysis,Vol. 3, Issue 1, March, 2007, PP.15-16.

[2] B Babcock, S Babu, M Datar, R Motwani, and JWidom. "Models and issues in data stream systems", Proceedings of PODS, Vol.1 2002, PP.176-179.

[3] R.Agrawal and R. Srikant, ªFast Algorithms for Mining AssociationRules,º Proc. 1994 Int'l Conf. Very Large Data Bases, Santiago,Chile, Vol.2, Sept. 1994, PP. 487-499.

[4] R. Agrawal and R. Srikant, ªMining Sequential Patterns,º Proc. of Int. Conf. Data Eng., Taipei, Taiwan, Mar. 1995, PP. 3-14.

[6] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann, San Mateo,CA,1993, PP.67 .

[8] SeemaBandyopadhyay and Edward J. Coyle, "Anenergy efficient hierarchical clustering algorithm for wireless sensor networks", Technical Journal of Science, Vol.3, 2003, PP.1713–1723.

[9] Jason Cong and M'Lissa Smith, "A parallelbottom-up clustering algorithm with applications to circuit partitioning in VLSI design", In DAC'93: Proceedings of the 30th internationalconference on Design automation, New York, Vol.4 1, 993, PP.755–760.

[10] D.Rohini and R.Janaki, "Efficient Term Frequency and Optimal Similarity Measure of Snippet for Web Search Results", Engineering and Scientific International Journal, Vol 2, Iss 1, January - March 2015, PP.19-22

[11] A.S.Aneeshkumar and Dr. C.Jothi Venkateswaran, "A novel approach for Liver disorder Classification using Data Mining Techniques", Engineering and Scientific International Journal, Vol 2, Iss 1, January - March 2015, PP.15-19.