ISSN No. 0976-5697

Volume 5, No. 6, July-August 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

High Dimensional Data & High Speed Data Streams – A Survey

Ms. A. Sheela
Research Scholar, Department of Computer Science
Sri Krishna Arts and Science College
Coimbatore, India

Mrs. C. Sunitha
HOD of CA &SS
Sri Krishna Arts and Science College
Coimbatore, India

Abstract: Clustering is used to grouping objects from the large database. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. It is a high dimension of the dataset, arbitrary shapes of clusters, scalability, input parameter, domain knowledge and noisy data. Large number of clustering algorithms had been proposed till date to address these challenges. There do not exist a single algorithm which can adequately handle all sorts of requirement. In this paper, we have discussed in K-means Clustering algorithm and Agglomerative clustering algorithm.

Keywords: K-means algorithm, agglomerative algorithm, scalability, High dimensional data streams, outliers.

I. INTRODUCTION

Data Stream is defined as a sequence of unbounded, real time data items with a very high data rate that can only read once. E.g.: computer Networks, traffic, phone conversations, web searches. In this paper, we examine the real process of web searches to be an example. Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining is the process of finding out what users are looking for on internet. Some users might be looking at only textual data whereas some other might want to get multimedia data. Web usage mining also helps finding the search pattern for a particular group of people belonging to a particular region. Web structure mining is the process of using graph theory to analyses the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds.

The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location. The other kind of the web structure mining is mining the document structure. It is using the tree-like structure to analyses and describes the HTML (Hyper Text Markup Language) or XML (extensible Markup Language) tags within the web page.

Web user session in clustering is a means of understanding user activity and interests on the World Wide Web. The period of time a user interfaces with an application. The user session begins when the user accesses the application and ends when the user quits the application. The session of activity that a user with a unique IP address spends on a Web site during a specified period of time is called a user session. The number of user sessions on a site is used in measuring the amount of traffic a Web site gets.

The site administrator determines what the time frame of a user session will be (e.g., 30 minutes). If the visitor comes back to the site within that time period, it is still considered one user session because any number of visits within those 30 minutes will only count as one session. If the visitor returns to the site after the allotted time period has expired, say an hour from the initial visit, then it is counted as a separate user session. Contrast with unique visitor, hit, click-through and page view, which are all other ways that site administrators measure the amount of traffic a Web site gets.

II.RELATED WORK

Clustering data streams is an interesting Data Mining problem. Detailed surveys of clustering algorithms may be found in three variants of the K-means algorithm to cluster binary data streams. The variants include On-line K-means, Scalable K-means, and Incremental K-means, a proposed variant introduced that finds higher quality solutions in less time [1], [2]. Clustering under the data stream model of computation, the data stream model is relevant to new classes of applications involving massive data sets, such as web click stream analysis and multimedia data analysis [3]. The nature of stream data makes it essential to use algorithms which require only one pass over the data. Recently, single-scan, stream analysis methods have been proposed in this context. However, a lot of stream data is high- dimensional in nature. High-dimensional data is inherently more complex in clustering, classification, and similarity search [4].

Recent research discusses methods for projected clustering over high-dimensional data sets [5]. Scalable is very important and difficult to handle the high dimensional data sets, previously they used to increase the efficiency in some algorithms like Scalable Advanced Massive Online Analysis (SAMOA), Gamma Region Density partition, and utilizes K-means - GARDEN-K-means[6],[7]. The famous K-means clustering algorithm is sensitive to the selection of the initial centroids and may converge to a local minimum of the criterion function value. A new algorithm for initialization of the K-means clustering algorithm is presented. The proposed initial starting centroids procedure allows the K-means algorithm to converge to a “better” local minimum.

© 2010-14, IJARCS All Rights Reserved

Finding frequent item sets from data streams is one of important tasks of stream data mining [8]. The ability to find clusters embedded in subspaces of high dimensional data, scalability, end-user comprehensibility of the results, non-presumption of any canonical data distribution, and insensitivity to the order of input records [9]. Data analysis problem from the astronomy simulation domain: massive-scale data clustering [10], [11].

Aim of this paper, K-means remains one of the most popular clustering algorithms used in practice. K-means to data stream clustering is how to maintain clusters so as to adapt to dramatic and gradual changes in streaming. To this end, we propose an agglomerative representation, where multiple spheres are dynamically maintained in a set. Each set describes the corresponding data domain presented in a data stream, k-means and agglomerative cluster patterns are discarded as outliers.

III.PARTITIONING METHODS

Partitioning clustering algorithms, such as K-means, K-medoids PAM, CLARA and CLARANS assign objects into k (predefined cluster number) clusters, and iteratively reallocate objects to improve the quality of clustering results. K-means is the most popular and easy-to understand clustering algorithm [13].

A. K-Means Algorithm:

K-means algorithm is implemented on preprocessed data for initial cluster technique. We start with a partitional algorithm with clusters, with significantly smaller than the total number of. To efficiently position, in our unidimensional metric space, the representatives at procedure startup, we evaluate the distance between any two adjacent samples. According to the distance metric, we take the farthest couples and determine intervals. Our goal is to exploit this property to group connections (objects) to identify user-sessions (clusters) in an automatic fashion. The following procedure for finding the k means:

a. Make initial guesses for the means $\mathbf{m}_1$, $\mathbf{m}_2$, ..., $\mathbf{m}_k$
b. Until there are no changes in any mean
a) Use the estimated means to classify the samples into clusters
b) For i from 1 to k
a. Replace $\mathbf{m}_i$ with the mean of all of the samples for cluster i
a) end_for
a. end_until

IV. HIERARCHICAL METHODS

Hierarchical algorithms create a hierarchical decomposition of the database. The algorithms iteratively split the database into smaller subsets, until some termination condition is satisfied [12]. The hierarchical algorithms do not need k as an input parameter, which is an obvious advantage over partitioning algorithms. However, the disadvantage of the hierarchical algorithm is that the termination condition is to be specified. Hierarchical decomposition can be represented as a dendrogram in two ways; i) Bottom-up (agglomerative) approach and ii) Top-down (divise) approach.

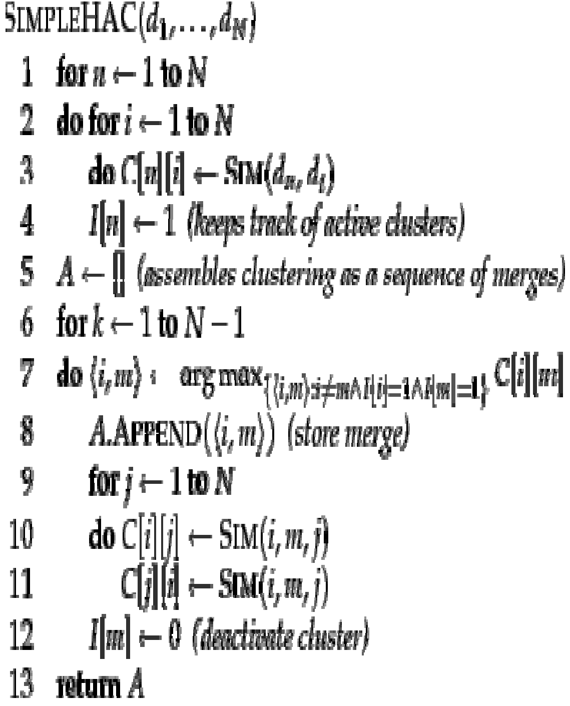
A. *An Hierarchical Agglomerative Clustering (HAC) Approach Agglomerative (bottom-up):*

Starts with each object forming a separate group. Merges the objects or groups close to one another until all of the groups are merged into one, or until a termination condition holds. The merge operation is based on the distance between two clusters.

There are three different notions of distance;

a. Single link
b. Average link
c. Complete link.

The Hierarchical Agglomerative Clustering Algorithm following as:



```
SIMPLEHAC($d_1, \ldots, d_N$)
for $n \leftarrow 1$ to $N$
do for $i \leftarrow 1$ to $N$
    do $C[n][i] \leftarrow$ SIM($d_n, d_i$)
    $I[n] \leftarrow 1$ (keeps track of active clusters)
$A \leftarrow []$ (assembles clustering as a sequence of merges)
for $k \leftarrow 1$ to $N - 1$
do $\langle i, m \rangle \leftarrow \arg\max_{\{\langle i,m \rangle : i \neq m \wedge I[i]=1 \wedge I[m]=1\}} C[i][m]$
    A.APPEND($\langle i, m \rangle$) (store merge)
    for $j \leftarrow 1$ to $N$
    do $C[i][j] \leftarrow$ SIM($i, m, j$)
        $C[j][i] \leftarrow$ SIM($i, m, j$)
    $I[m] \leftarrow 0$ (deactivate cluster)
return $A$
```

V. CONCLUSION

A partitional clustering procedure is run over the original data set, which includes all samples, using the optimal number of clusters determined so far and the same choice of cluster representatives adopted in the step. A fixed number of iterations are run to obtain a final refinement of the clustering definition. This method to be a very efficient to compared the previous methods. K-means and agglomerative cluster patterns are discarded as outliers. This paper shows the idea of how to these algorithms to be work and what are methods to be used in previous high dimensional data stream clustering. In future, we have to implemented these algorithms and discuss the results.

VI. REFERENCES

[1] Charu C. Aggarwal,” A Survey of Stream Clustering Algorithms” IBM T. J. Watson Research Center Yorktown Heights, NY 10598, pg: 229-256.

[2] C. Ordonez. Clustering Binary Data Streams with K-means. In Data Mining and Knowledge Discovery Workshop, 2003.

[3] S. Guha, N. Mishra, R. Motwani, and L. O’Callaghan. Clustering Data Streams. In IEEE FOCS Conference, 2000.

© 2010-14, IJARCS All Rights Reserved

[4] C. Aggarwal, J. Han, J. Wang, and P. Yu. A Framework for Projected Clustering of High Dimensional Data Streams. In VLDB Conference, 2004.

[5] C. Aggarwal, J. Han, J. Wang, and P. Yu. On High Dimensional Projected Clustering of Data Streams, Data Mining and Knowledge Discovery Journal, 10(3), pp. 251–273, 2005.

[6] F. Farnstrom, J. Lewis, and C. Elkan. Scalability for Clustering Algorithms Revisited. ACM SIGKDD Explorations, 2(1):pp. 51–57, 2000.

[7] P. Bradley, U. Fayyad, and C. Reina. Scaling Clustering Algorithms to Large Databases. In ACM KDD Conference, 1998.

[8] C. Zhang, M. Gao, and A. Zhou. Tracking High Quality Clusters over Uncertain Data Streams, ICDE Conference, 2009.

[9] Q. Zhang, J. Liu, and W. Wang. Approximate clustering of distributed data streams, ICDEConference, 2008.

[10] C. Aggarwal. On Change Dignosis in Evolving Data Streams. In IEEE TKDE, 17(5), 2005.

[11] C.Aggarwal, J. Han, J. Wang, and P. Yu. A Framework for Clustering Evolving Data Streams. In VLDB Conference, 2003.

[12] Neha Aggarwal,Kirti Aggarwal, Kanika gupta " Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining" International Journal of Scientific & Engineering Research, Volume 3, Issue 3, August-2012.

[13] Dr.N.Rajalingam, K.Ranjini "Hierarchical Clustering Algorithm - A Comparative Study" Volume 19– No.3, April 2011, International Journal of Computer Applications (0975 – 8887).

© 2010-14, IJARCS All Rights Reserved