International Journal of Grid and Distributed Computing
Vol. 14, No. 1, (2021), pp. 738-741

Techniques for Data Stream Clustering: Comparative Analysis

[1]Ms. Mayuri G. Ghodmare , [2]Dr. M. N. Quadri , [3]Dr. S. B. Kishor

[1] Research Scholar, Department of Computer Science, Gondwana University, Gadchiroli

[2] HOD & Asst. Prof., Department of Computer Science, N.S. Sci. College, Bhadrawati, Dist. Chandrapur

[3] HOD & Asst. Prof., Department of Computer Science, Sardar Patel Mahavidyalaya, Chandrapur

Abstract

In the fast growing world applications are generating data in enormous volumes called data streams. Data stream is imaginably large, continual, rapid flow of information and in data mining the important tool is called clustering, hence data stream clustering (DSC) can be said as active research area. Recent attention of data stream clustering is through the applications that contain large amounts of streaming data. Data stream clustering is used in many areas such as weather forecasting, financial transactions, website analysis, sensor network monitoring, e-business, telephone records and telecommunications. In case of data stream clustering most popularly used heuristic is K-means and other algorithms like K-medoids and the popular BIRCH are developed. The aim of the abstract is to review the developments and trends of data stream clustering methods and analyze typical DSC algorithms proposed in recent years, such as BIRCH, STREAM, DSTREAM and some more algorithms.

Keywords: Data mining, Data stream, Clustering, classification, concept, drift

1. Introduction

Nowadays we have many applications with massive amount of data which are caused limitation in data storage capacity and processing time. Traditional data mining is not suitable for this kind of applications so they should be tuned and changed or designed with new algorithms. Besides of speed up and storage capacity, real-life concepts tend to change over time.

In data mining, data objects are represented as points. The classification of data into clusters is unsupervised and can be denoted as $X = C_1 \cup \ldots\ldots C_i \cup C_k$; $C_i \cap C_j = \phi$ (where $i \neq j$) where, X denotes original dataset, Ci, Cj are clusters formed in X, and k denotes the number of clusters formed [1]. In case of unsupervised knowledge, only former information of the domain and data are known, however structured characteristics are not known. It is very common that these unknown structural characteristics include the spatial distribution of the data. Characteristics such as volume, density, shape, or orientation are the basis of cluster formation. Clustering is a method to group unlabeled data sets based on "similarity" of the features extracted out of data items, and allocations of dissimilar data items in different groups.

Most of the algorithms on clustering data streams generate clusters over the whole data set. These algorithms consider clustering as a single-pass clustering method. For some applications, such clustering methods are useful, however, in the case of data streams, it is necessary to define the clustering problem carefully, as the data stream is an infinite process in which data is evolved with time. The clusters are also varying with respect to the time when they are calculated and the time over which they are measured.

2. Clustering Techniques

BIRCH can be considered a primitive method in this area [2]. In fact it has been designed for traditional data mining but it is suitable for very large data base so it has been applied for data stream mining. This method introduces two new concepts: micro clustering and macro clustering. Based on these two concepts it could overcome two main difficulties in agglomerative method in clustering: scalability and the inability to undo what was performed in the previous step. It works base on two steps: first it scans data base and builds a tree which is included information about data clusters. In second step BIRCH refines tree by removing sparse nodes as outliers and concrete original clusters. The main disadvantage of this method is the limitation in capacity of leaf. If clusters are not spherical in shape, BIRCH does not perform well because it uses the notion of radius or diameter to control the boundary of a cluster.

STREAM is the next main method which has been designed especially for data stream clustering [3]. In this method K-Medians is leveraged to cluster objects base on SSQ criterion for error measuring. In the first scan objects grouped and medians of each group is gathered and associated them a weight base on the number of objects in the cluster. In next step these medians is clustered until top tree.

The assignment of data points to (typically k) groups such that points within each group are more similar to each other than to points in different groups, is a very basic unsupervised data mining task. For static data sets, methods like k-means, k-medoids, hierarchical clustering and density-based methods have been developed among others Many of these

ISSN: 2005-4262 IJGDC

methods are available in tools like R, however, the standard algorithms need access to all data points and typically iterate over the data multiple times. This requirement makes these algorithms unsuitable for large data streams and led to the development of data stream clustering algorithms.

2.1 Hierarchical Method (HM)

Hierarchical clustering algorithm creates a hierarchy or tree of clusters for the data objects. This algorithm is a type of cluster analysis that tries to create or develop a hierarchy of clusters or a tree of clusters, also known as a Dendrogram, where each child clusters is contained in each cluster node, and the sibling clusters partition the points covered by their common parent.

Hierarchical clustering techniques are sub divided into two methods specifically agglomerative and divisive. First method combines a set of "n" objects to form a lot of general categories and the second method produces smaller clusters consecutively by dividing "n" objects. The main functionality of HM is to combine data objects to form a hierarchical tree of clusters. In hierarchical clustering algorithms objects in the clusters are connected based on their distance. Various clusters will form at various distances. By using cluster proximity as a measure hierarchical clustering techniques determine where to combine and where to split different clusters.

Advantages:

- It is capable of dealing with any kind of similarity or distance.
- Disadvantages:
- It is very ambiguous in case of termination criteria.
- High complexity.

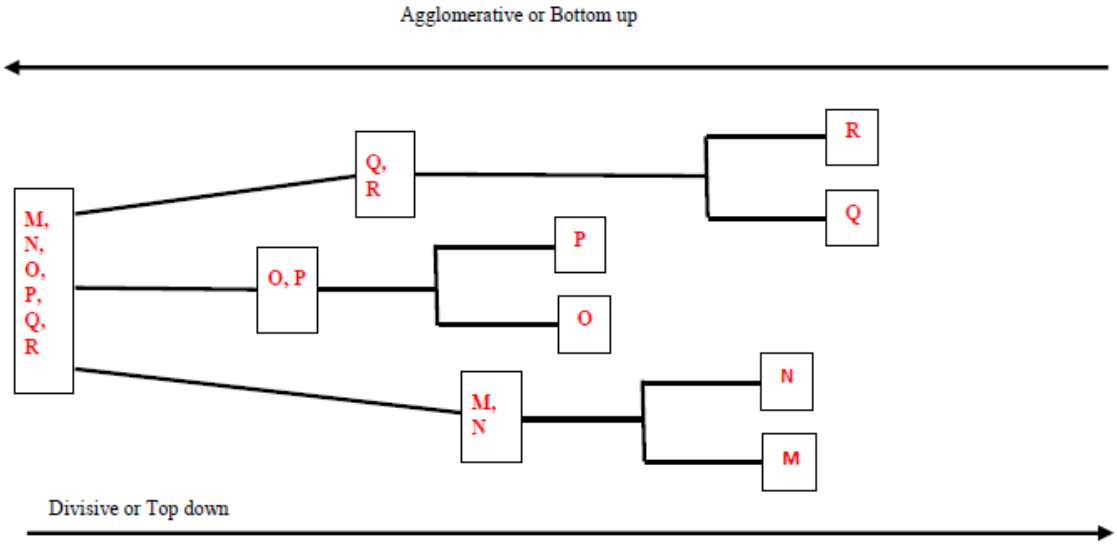

Fig. 1: Hierarchical Method

2.2 Partitioning Method (PM)

Partitioning clustering algorithms is an algorithm that groups objects into a specified partition where each partition reflects a cluster such that, the objects within that cluster are "similar" and "dissimilar" to objects in other clusters. Iterative relocation technique is then applied which tries to improve the grouping (partitioning) in the way of moving objects from a group to another hence partition of high quality is achieved suppose the objects in identical clusters are "close" or suchlike the other, and objects in distant clusters are "far apart" or very dissimilar.

K-median and K-means are the two techniques of the partitioning based data stream clustering. Partitioning algorithm categorise datasets into n clusters, where n is a parameter that is predefined. It continuously allocates objects from one group to another group to reduce the objective function. The widely used traditional clustering methods are k-means and k-medians. In partitioning clustering data is grouped as single partition rather than presenting it as a nested structure, therefore it is highly used when the data set is large.

ISSN: 2005-4262 IJGDC

The k-means partitioning algorithm functions properly only with numerical attributes but a single outlier negatively affects it. From the ongoing, this cannot be a good algorithm for mining numerical and object with large dataset as it becomes a challenge for K-means to handle data with outliers, clusters of differing sizes, densities and non-globular shapes.

Advantages:

- Implementation is easy.
- It creates clusters in an iterative manner.

Disadvantages:

- User need to predefine the number of clusters.
- Only spherical shaped clusters are determined here.

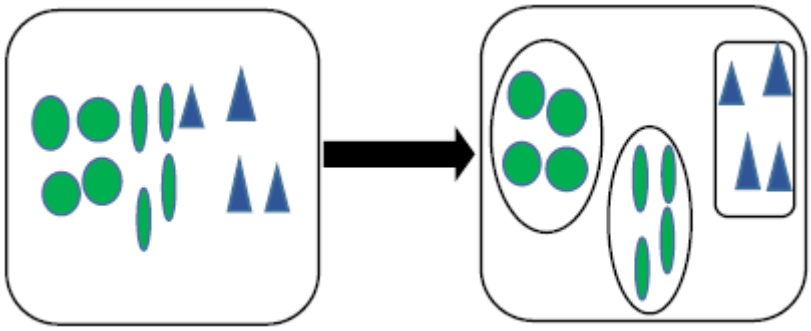

Fig. 2: Partitioning Method

Density-Based

Partitional clustering methods uses distance measure between the objects to cluster the objects. Generally, clusters generated out of this approach are spherical in shape. It may not be possible to discover arbitrary shaped cluster. The methods using density as basis for identifying clusters are categorizes as density-based methods. In these methods, cluster generated in all directions as long as density in the neighborhood exceeds some threshold. This method naturally protects data set from outlier. The overall density of a point is processed and analyzed in order to determine features or functions of dataset which influences a particular data point. The algorithms like DBSCAN, OPTICS, DBCLASD, DENCLUE and DenStream uses an approach such that noise (outliers) are automatically filtered out and cluster of arbitrary shape are constructed

Model-Based

Clustering methods use some predefined mathematical model to fit the data and then optimizes them. The basic assumption is that data is hybrid in terms of probability distributions. Model-based methods determine the number of clusters based on standard statistics. In order to have robust clustering method noise and outliers are considered while calculation of standard statistic. The issue in Clustering problem is to automatically determine the number of clusters based on standard statistic, taking outliers and noise into consideration. Therefore, these types of methods are very much robust methods with respect to noise and outliers. Based on approach used to generate clusters, these model-based methods are categorized into statistical and neural network approach methods. Algorithm like MCLUST [4] is model-based clustering algorithm. There are other model-based clustering algorithms like EM [5] (based on mixture density approach), COBWEB [6] (conceptual clustering), and neural network based methods such as self-organizing feature maps. Those model-based algorithms which are based on statistical approaches uses probability measures in determining clusters. In case of neural network approach, input and output are connected with units carrying weights. The Neural Network properties are useful in clustering problem; therefore neural network approach is much popular in clustering. The properties such as (1) Neural Network have in-built parallel and distributed computing architecture; (2) The interconnected weights are recursively adjusted so as to best fit the data. The weights are then normalized because of this recursive operation. The selection of feature is done on the basis of patterns; (3) Numerical data is processed or converted/transformed into quantitative features. Each cluster is represented as exemplar; which then acts as prototype of the cluster and does not belong to specific object. New incoming objects are assigned to cluster whose exemplar is similar, based on some distance measure.

2.3 Grid-Based Method (GBM)

ISSN: 2005-4262 IJGDC

The processing time in the grid-based algorithms is independent to the total number of data points hence these algorithms are faster. Here the object space splits into limited number of cells that indeed gives the grid structure where the clustering operations take place. The main intention of these algorithms is to compute the data sets into cells and then working on the objects that comes under these cells.

Advantages:

- This method can handle noises.
- Fast processing time.

Disadvantages:

- Here the grid sizes need to be predefined.
- GBM is not preferred for high dimensional data.

3. Conclusion

Various data stream clustering methods along with the mostly used data stream clustering algorithms are confer in this work. Extraction of data more precisely in real time plays a crucial role hence the scope of discovering the data stream clustering algorithms is high. In future research can take place in developing the most efficient Data Stream Clustering algorithms that help in solving the problems of data stream clustering.

4. References

[1]. Hahsler, M.; Bolanos, M.; Forrest, J. Introduction to stream: An Extensible Framework for Data Stream Clustering Research with R. J. Stat. Softw. 2017, 76, 1–50.
[2] Zhang, Ramakrishnan, and L. M., "BIRCH: An efficient data clustering method for very large databases " presented at ACM SIGMOD Conference on Management of Data, 1996.
[3] L. O Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, "Streaming-data algorithms for high-quality clustering," 2002.
[4]. Chris, F.; Raftery, A.E. MCLUST: Software for model-based cluster analysis. J. Classif. 1999, 16, 297–306.
[5]. Miin-Shen, Y.; Lai, C.; Lin, C. A robust EMclustering algorithm for Gaussian mixture models. Pattern Recognit. 2012, 45, 3950–3961.
[6]. Fisher, D.H. Knowledge acquisition via incremental conceptual clustering. Mach. Learn. 1987, 2, 139–172.

ISSN: 2005-4262 IJGDC