International Core Journal of Engineering Volume 8 Issue 6, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0012

A Review: Analysis and Thinking of Concept Drift Data Stream Clustering

Hui Liu

College of Information Engineering, Shanghai Maritime University, Shanghai 200135, China

*huiliuliu@126.com

Abstract

A large amount of data is generated every day in life, and it is generated at an unlimited rate. This type of data related to timestamps is called a data stream, and a data stream is also a type of time series. The clustering of many stream data has been widely studied in the field of data mining. There are many traditional algorithms for solving such problems, such as k-means, density-based algorithms, distance-based algorithms, and probability clustering methods. But these algorithms are not practical for clustering streaming data. This paper introduces the comparative analysis between clustering of stream data and traditional clustering, specifically explains the problems to be solved in data mining for stream data, and analyzes the reasons why traditional algorithms cannot be applied to clustering of stream data. Then, the algorithms in two recent articles on dynamic data flow clustering are described in detail, and the shortcomings in the algorithms are analyzed.

Keywords

Datastream; Cluster; Algorithm.

1. Introduction

With the development of information technology such as communication and computer, the data generated in many application fields such as financial market, network monitoring, telecommunication data management, sensor network, etc. belong to the data stream. Data streams were first proposed in 1998 by Henzinger et al. in the paper "Computing on Data Streams". The fast, continuous, infinite and dynamic characteristics of data stream, coupled with the limitation of data collection time and analysis processing speed, leading to traditional data management and analysis techniques ineffective or need improvement. Data stream has become a new research field and hotspot.

2. Data Stream Correlation Comparison

2.1 Stream Data and Traditional Data

Traditional data has many characteristics, such as large amount of data, complex data distribution and various data types. Therefore, it is necessary to process the data through corresponding algorithms and outlier point processing to really extract valuable information from the data.

• Boundless

In the process of data processing, the data flow has been continuously generated without boundaries.

• Time-dependent

The time definition needs to be recorded on the data point. The data stream is a type of time series. The generation of the data stream is incrementally generated with time, and the importance of the data stream will gradually decay with time, which has no practical significance for its data analysis.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0012

• Single liquidity

Due to the large amount of mobile data, the continuous generation of data, and the limited storage, the data stream only has a single liquidity, and the data stream can only appear once and cannot be obtained randomly.

Therefore, the mining of data stream has practical significance. Clustering of data stream is an important direction and research point of mining to cluster data according to the similarity of data stream.

2.2 Data Stream Clustering and Dynamic Clustering

The traditional clustering methods are all applied to static existing data. Generally, the data is stored first, and then the macro clustering process is performed. With the increasing number of data, it is difficult to store all data. Therefore, the dynamic clustering method of gradually classifying data has gradually entered people's attention, while the clustering of data stream is based on dynamic clustering. deeper promotion.

(1) Dynamic clustering: For clusters that are roughly pre-classified, when a new data point is entered or a data point is deleted, it gradually adjusts and re-clusters all data to obtain a clustering result. This process represents dynamic clustering. Compared with the previous systematic clustering, dynamic clustering has less computation and less storage space, and is more suitable for clustering large sample data.

(2) Stream clustering: that is, the clustering of data stream. The key to stream clustering is the timestamp information. The data changes gradually with time. In other words, when the newly entered data stream points are re-added to the clustering process, some outdated data points are also deleted while dynamic clustering is performed. This clustering process is called stream clustering. In this stream clustering process, there is no need to specify the number of clusters in the final cluster in advance, and the clustering result can be clusters of any shape, and can handle outliers well.

3. Data Stream Clustering Challenges and Main Problems

3.1 Data Stream Clustering Challenge

(1) There is a huge amount of information in the data stream, and it is impossible to store all the data. Therefore, for the data that flows at one time, scan all the data, and set up a check mechanism to retain the data that is helpful for the clustering results. The check mechanism needs to select updates within a short period of time.

(2) The data stream changes with time which means that the clustering of the data stream needs to be updated all the time, and the real-time clustering results can be fed back according to the demand.

(3) The clustering of data stream has high requirements on space, and how to store the necessary data streams efficiently and in a space-saving manner is a problem that must be considered in data stream clustering.

3.2 Main Problems

In recent years, the development of hardware and software has produced a large number of data streams, such as network streams, graph streams, short text streams, digital streams, event streams, semantic concept streams, web click streams, and so on.

The ultimate goal of data stream clustering is to effectively divide the data objects arriving in a stream way into several clusters, and to capture the conceptual drift of the data stream by observing the changes of clusters under limited computing resources. At the same time, limited by time and space, the key to data stream clustering is how to cluster online, how to track changes in clusters, and how to change clusters in real time.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0012

3.3 Disadvantages of Traditional Clustering Methods

Compared with other data, the data stream has the limitation of one-time passing of data, so it is difficult to use traditional algorithms to calculate clusters flexibly, multiple times, and arbitrarily for incoming data at any time point. For example, if we use the most familiar K-means algorithm to solve the clustering of data streams, if we want to obtain real-time results of clustering at any time point, the calculation process is very large, so the online clustering process cannot be carried out. Moreover, sometimes it is necessary to compare the clustering results at a certain time point in the past with the current clustering, and how to record the clustering results of the data flow at different time nodes is also difficult to deal with. Therefore, it is difficult for traditional clustering algorithms to realize the clustering of data streams.

4. Clustering Algorithms for Data Streams

4.1 Hierarchical Thinking

Of course, many people make improvements based on traditional classical algorithms to cluster data streams. For example, the Stream algorithm [2] that uses the idea of hierarchical layering, Stream is to process the data stream into blocks.Firstly perform first-level clustering on each block, and the algorithm used for clustering is the K-median algorithm that we are familiar with. The center point of the weight of the data point is subjected to secondary clustering, and so on, until the center point of the upper layer is clustered. Obviously, the higher the level of clustering, the slower the number of center points grows, until the last time K-median clustering is used, the data of almost all blocks are integrated together. The Stream algorithm can save the space for storing data streams, and does not need to store all the data, and finally the clustering results can be obtained. However, there are still many shortcomings in the algorithm. The algorithm adopts a hierarchical structure and is not sensitive to the evolution process of data flow clustering. It cannot track the changing process of clustering in real time, and cannot obtain clustering results at any time.

4.2 Micro-clustering Framework

There is also a widely recognized micro-clustering framework for data stream clustering, which divides the clustering of data streams into online and offline states. Typical representatives include CluStream [3]. The CluStream algorithm hopes to realize the summary information of the data stream through the offline process, generate the micro-cluster structure, and select the appropriate time to save the micro-cluster structure through the pyramid time frame structure. In the offline state, according to the user's needs, specify the clustering results of a certain period of time, and quickly generate the clustering results. The micro-cluster framework structure can obtain the clustering structure at the specified time according to the user's needs, but there is still a problem of real-time clustering. The algorithm itself may be affected by the speed of the data stream, and the speed of the data stream will lead to very different clustering results.

4.3 Cluster Merge

The above two algorithms were proposed relatively early, and many people have made improvements in the follow-up. Below I will introduce the algorithms proposed in recent years to solve the problem of data stream clustering.

The SOStream algorithm [1] introduces a density-based adaptive clustering algorithm. The adaptation in this article refers to automatically adjusting the position of the clusters. The algorithm will merge the overlapping clusters according to the input vector, and finally achieve the real-time online clustering effect of the data stream. The SOStream algorithm is based on the idea of the SOM algorithm proposed before. When the SOM algorithm is initialized, it uses a typical clustering algorithm to cluster the incoming data stream to obtain several clusters. When a new input vector enters, it will be compared with the incoming data stream. The nearest cluster is defined as the winner cluster, the input vector is merged into the winner cluster, the neighboring clusters of the winner cluster are obtained according to the threshold radius, and the positions of the neighboring clusters

ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0012

are updated and adjusted so that the neighboring clusters are closer to the winner cluster. In this way, when a new similar input vector comes in, it is easier for neighboring clusters to find the winner cluster. However, the update of the adjacent clusters of the winner cluster will inevitably lead to the overlap between the clusters, and then the clusters need to be merged. Compared with other algorithms solved by online merging, the algorithm has higher purity of data in each cluster. The purity refers to the data points in the cluster that really affect the results of the cluster, but this algorithm cannot solve the separation of clusters. And there is no special treatment for outliers, and the outliers in the data stream also have a great impact on the clustering results, and the algorithm does not realize the processing of outliers.

4.4 Dynamic Clustering

In the "Clustering Method of Data Streams Based on Density Mountains" published on VLDB in 2017, the article proposed the EDMStream algorithm [4], a relatively new algorithm to solve the dynamic data stream clustering problem. In the article, the evolution process of clusters in the clustering process of the data stream is described in depth, and the attenuation function data points are included, which makes the clustering results more accurate. The EDMStream algorithm is optimized based on the DP algorithm. The DP algorithm is based on the DBSCAN algorithm to find out the dependencies between points and then obtain the dependencies between clusters. The difference from DBSCAN is that the DP algorithm relies on the data that is closest to itself and whose density is greater than its own. point. The DP-Tree structure is proposed in the article. The tree structure is used to save the dependencies between the above-mentioned points, and the strong dependency subtree is found according to the set threshold, and all data points on the strong dependency subtree are found. It is the final clustering result. The DP algorithm cannot dynamically cluster the data stream, because the real-time nature of the data is not considered in the algorithm, and the timestamp information of the data is not introduced. Therefore, the improved EDMStream algorithm introduces the freshness decay model of data points, directly associates the density of data points in the data stream with the decay function, and then clusters the data. At this time, the data stream can be well resolved. dynamic clustering problem. In addition, the structure of DP-Tree has been improved accordingly. The original DP algorithm uses each data node as a node of the tree, and the connection between nodes represents the dependency between points. The concept of cluster cells is introduced into the EMDStream algorithm of , which is very similar to the micro-cluster model. Each cluster cell is regarded as each node in the tree, and the connection between the nodes represents the dependency between the cluster cells. The algorithm not only solves the online real-time clustering of data by introducing the freshness of data points, but also reduces the storage space by changing the nodes of the tree, and at the same time can reduce the amount of calculation, and finally achieve real-time high data flow. quality clustering.

Of course, the time complexity of the EDMStream algorithm is relatively high. The main problem is that every time a new data point enters, the tree structure will be reconstructed, which greatly affects the time complexity of the algorithm. I personally think that if the number of tree reconstructions can be reduced, or the tree reconstruction can be avoided by adjusting some pointers in the tree, but in the end it is still possible to find the largest strongly dependent subtree that obtains clusters in the EDMStream algorithm. Will be a big innovation point. This algorithm also has a problem that we usually need to solve when designing the algorithm. There are too many parameters set in the algorithm design. If the parameters that appear can be adaptive, the adjustment parameters can be automatically set through learning to avoid The impact of manual intervention in the clustering process on the final result of clustering will also be a great improvement to the EDMStream algorithm.

4.5 Multiview Clustering

In February this year, a dynamic clustering algorithm for multi-view data streams was published on IEEE. The previous problems of data stream clustering were based on single data stream problems, but this article proposes a multi-view clustering algorithm MVStream [5] , the algorithm first integrates the data in multiple directions in a streaming way, abstracts the overall statistical

ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0012

information of historical view data objects, designs a multi-view cluster labeling method, finds clusters of any shape from each view, and tracks relative The labels between neighboring clusters finally get the evolution information of multi-view clusters to realize clustering.

5. Conclusion

The data mining of stream has always been an important branch of time series since its inception. Its future research will mainly focus on the following aspects: (1) Adaptive real-time data stream clustering based on resource constraints. Mainly in terms of space constraints and computational constraints. (2) Clustering of high-dimensional real-time data streams. Most real data streams have high-dimensional characteristics. Objects in high-dimensional space are sparsely distributed, and noise is difficult to identify, which is a difficult problem to solve; (3) Real-time clustering of multiple data streams in a distributed environment.

References

[1] C. Isaksson, M. H. Dunham, and M. Hahsler.SOStream: Self Organizing Density-Based Clustering Over Data Stream. Springer Berlin Heidelberg, 2012.

[2] L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani. Streaming-data algorithms for high-quality clustering. In ICDE Conference, pages 685–694, 2002.

[3] Aggarwal CC,Han Jia-Wei,Wang Jian-Yong. A framework for clustering evolving data streams [C].Proceedings of the 29th International Conference on Very Large Data Bases.Berlin,Germany,2003: 81-92.

[4] Shufeng Gong, Yanfeng Zhang, Ge Yu. Clustering Stream Data by Exploring the Evolution of Density Mountain. PVLDB, 11(4):393-405, 2017.DOI: 10.1145/3164135.3164136.

[5] Ling Huang , Chang-Dong Wang , and Hong-Yang Chao. MVStream: Multiview Data Stream Clustering. IEEE 2162-237X © 2019.

[6] Miller Z , Dickinson B , Deitrick W , et al. Twitter spammer detection using data stream clustering[J]. Information Sciences, 2014, 260:64-73.