A Novel Spatial Clustering Method based on Wavelet Network and Density Analysis for Data Stream

Chonghuan Xu
College of Business Administration, Zhejiang Gongshang University, Hangzhou 310018, China
Center for Studies of Modern Business, Zhejiang Gongshang University, Hangzhou 310018, China
Email: talentxch@gmail.com

***Abstract*—With the limited memory and time, a fast and effective clustering can't be achieved for massive, high-speed data stream, so this paper mainly studies the key method of data stream clustering under the restriction of resource, and then proposes a dynamic data stream clustering algorithm (D-DStream) based on wavelet network and density, which uses sliding window to process data stream. Firstly, apply wavelet network to compress data stream and build a much smaller synopsis data structure to save major characteristics of data stream, then cluster with two-phase density clustering algorithm. The results of experiment show that the D-DStream algorithm can successfully solve clustering problems caused by STREAM or others, also has high time efficiency and high clustering quality.**

***Index Terms*—Data Stream Clustering, Wavelet Network, Two-phase Density Clustering, Sliding Window**

I. Introduction

Researches on data streams are motivated by emerging applications involving continuous massive data sets such as customer click streams, E-commerce, wireless sensor network, network monitor, telecommunication system, stock market and meteorological data. For the data stream applications, the volume of data is usually too large to be stored or scanned more than once. Furthermore, the data objects can be only sequentially accessed in the data streams, random data access techniques are not practical, so data stream clustering is a challenging area of research that attempts to extract useful information from continuously arriving data. A well designed data stream approach allows the processing of potentially infinite amounts of data. It scans the stream ideally in a single pass, keeping just the necessary data in the main memory. The elimination of random access is the great benefit that allows even gigantic amounts of data to be processed. However, a specialized data stream algorithm is necessary. The aim of our work is to develop an algorithm that can handle with data stream clustering effectively.

This paper is organized as follows. In section 2, the related works of clustering algorithm are given. In section 3, the novel dynamic data stream clustering algorithm is presented. The experimental results of comparing the algorithm proposed in this paper with other algorithms are also presented in section 4. Finally, our work of this paper is summarized in the last section.

II. Related Work

Data stream clustering has been heavily investigated in recent years, several important algorithms have been actively introduced. At present data steam clustering algorithms[1-8]are mainly proposed and improved by Guha, Aggarwal and others, like LOCALSEARCH algorithm which uses K-means to cluster data steam by a continuous iterative process in limited space based on the idea of partition. STREAM algorithm [9], which is improved on the basis of LOCALSEARCH algorithm, is a single-scanning stream clustering algorithm based on K-means, and is proved to be better than BIRCH algorithm. STREAM considers neither the evolution of data nor the variation of time granularity and clustering may be controlled by historical data. In the applications, it can effectively conquer the effects brought by noise data, but it only offers a description of current data stream rather than the changes of data stream. Then the CluStream [10] is proposed. It is an algorithm for clustering incoming data streams based on user-specified, online clustering queries. It divides the process of clustering into online and offline components. The online component computes and stores summary statistics of the data stream with micro-cluster, while the offline component performs macro-cluster and responds various user queries with the stored summary statistics. The amount of information archived is controlled by a user specified maximum number of micro-clusters with the algorithm attempting to capture as much detail as memory constraints allow. Its biggest disadvantage is that the radius of clustering continuously increases with the inflowing of data, and as it doesn't eliminate "old data" online, more and more data will increase the cost of process. HPStream [11], a modification of CluStream to enable clustering of high-dimensional data was proposed. The algorithm employs a data projection method to reduce the dimensionality of the data stream to a subset of dimensions that minimize the radius of cluster

doi:10.4304/jcp.8.8.2139-2143

groupings. It was demonstrated that by projecting data onto a smaller set of dimensions both synthetic and real world data sets could be more accurately processed. As with CluStream, however, the underlying assumption remains that clusters in the projected space remain spherical in nature. How best to classify incoming data using the CluStream and HPStream frameworks was discussed in literature [12]. Birch [13] is a well known hierarchical clustering algorithm that incrementally updates summary cluster information for offline analysis. Clusters suitable for classification are then extracted using the summary information via a second pass over the data. The algorithm was later adapted for online clustering and classification by combining the secondary offline phase with the incremental update component. Tu *et al.*[14] proposed a novel density-based hierarchical clustering scheme for streaming data to improve both accuracy and effectiveness; The method is based on the agglomerative clustering framework. Traditionally, clustering algorithms for streaming data often use the cluster center to represent the whole cluster when conducting cluster merging, which may lead to unsatisfactory results. They argued that even if the data set is accessed only once, some parameters, such as the variance within cluster, the intra-cluster density and the inter-cluster distance, could be calculated accurately. Giuseppe *et al.*[15] proposed a fully distributed *K*-means algorithm (*Epidemic K*-means) which does not require global communication and is intrinsically fault tolerant. The proposed distributed *K*-means algorithm provided a clustering solution which could approximate the solution of an ideal centralized algorithm over the aggregated data as closely as desired. Meanwhile, some outstanding algorithms have been proposed in China, for example, Chang [16] proposed an evolved data stream clustering based on sliding window, Zhu *et al.*[17] proposed a clustering algorithm based on random shape of data stream and so on.

Faced with these shortcomings, this paper proposes a clustering model called D-DStream based on wavelet network and two-phase density clustering. D-DStream uses sliding window to handle with data, then uses wavelet network to compress data stream in basic window and applies density clustering to process data stream, after that, conducts secondary clustering and updates clusters on the basis of synopsis data. Through the analysis of theory and experiment, we know that D-DStream model has effective improvement in execution time, memory space, clustering quality and so on.

III. The Method Based On Wavelet Network and Two-Phase Desity Clustering

A. Data Stream Management Model

Let t denote any of the timestamp, a_t denote the data element arrived at time t, then data stream can be represented as infinite collection $\{\ldots,a_{t-1}, a_t, a_{t+1},\ldots\}$. Generally, it uses Turnstile model to describe the data element a_t of data stream.

Liu *et al.*[18] described three window models as follows: landmark window, sliding window and damped window. Through analyzing all the historical data, landmark window model gets overall frequent pattern, its window's size increases with data stream inflows. Sliding window is suitable to the application which is only interested in current data, the algorithm only keeps and mines data in fixed window breadth, the window position slides as data stream flows. In damped window, each transaction corresponds to a weight, and thus weight will reduces as time increases.

Set t as current time stamp which describes the latest n data in data stream and its searching range is $\{a_{t-n+1},\ldots,a_t\}$. With the continuous arrival of data, old data shift out from one side of the window and new data shift in from another side. Generally, we divide the sliding window into several basic windows, and update the data sequence in a basic window every time. Figure 1 shows the process of sliding window and basic window. Divide the sliding window with a width of w into k basic windows by time, each basic window contains $b=w/k$ transaction data.

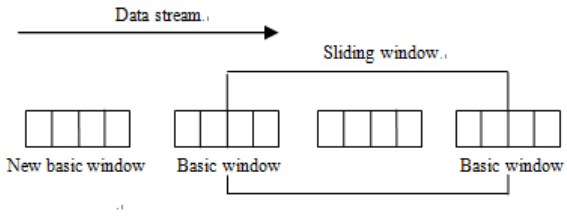

Figure 1. Sliding window and basic window.

B. Data Compression based on Wavelet Network

DWT (Discrete wavelet transform)[19] is an important data compression method which saves a part of important wavelet coefficients through the wavelet transformation of original data sets and approximately restores the original data sets. Haar wavelet is the simplest one of DWT, and is widely applied in the field of data compression and others because it's simple, effective and easily produced. One-dimensional Haar wavelet decomposition transforms the vector $A=\{x_1,x_2,\ldots,x_n\}$ into n wavelet coefficients $\{c_1,c_2,\ldots,c_n\}$.

The data compression based on wavelet transformation makes use of a good property of wavelet decomposition: if we keep r ($r<n$) most important wavelet coefficients (set other coefficients as 0), we can reconstruct good approximate results of original sequence. Then these r coefficients are known as the wavelet synopsis of original sequence.

The structure of wavelet neural network is achieved by common neural network that introduces wavelet transformation, in which, single-input and single-output BP neural network are used as the foundation. Neuron function in input layer is $y(t)=t$, the number of hidden layer neurons is 12, wavelet function $h((t-b_t)/a_t)$ is selected as neurons function in hidden layer, the net weight between input layer and hidden layer is set to 1.

The net weight between hidden layer and output layer is denoted as $w_1,w_2,\ldots,w_k$, and single neuron in output layer is used to conduct summation operator on the signals through the transformation of hidden layer neurons. This is because the signal formulation based on wavelet neural network is achieved through the linear superposition of selected wavelet basis. Set output signal as $s(t)$, then come up the formula $s(t)=\sum_{k=1}^{K} w_k h((t-b_k)/a_k)$.

In which, w_k, a_k, b_k are net weight, wavelet expansion factor and shift factor, K is the number of wavelet basis, that is, the number of hidden layer neurons. The structure of single layer network with an input and output node is shown in Figure 2.

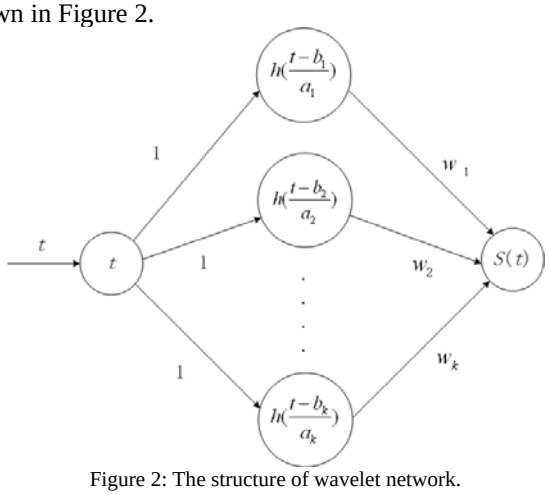

Figure 2: The structure of wavelet network.

C. Density-based Spatial Clustering

DBScan [20] is a classic clustering algorithm based on density, which is used to filter the outlier data and discover clusters of random shape, and its main idea is to cluster when the density of nearby area (the number of objects) is more than a certain threshold, that is, every object of the given cluster should contain certain objects in a specific area.

DBScan algorithm

1 The area in radius ε of given object is the ε -neighborhood of the object.
2 If the ε -neighborhood of an object contains *MinPts* objects at least, then these objects are core objects.
3 For a given object set D, an object p is directly density-reachable from a core object q if it is part of its ε -neighborhood.
4 Object p is called density-reachable from q about ε and *MinPts*, if there is an object chain $p_1,\ldots,p_n$, $p_1=q$, $p_n=p$ and for $p_i \in D(1 \le i \le n)$, p_{i+1} is directly density-reachable from p_i about ε and *MinPts*.
5 Object p and q are density-connected if there is an object o in object set D such that both p and q are density reachable from o about ε and *MinPts*.

D. The Description of D-DStream Algorithms

D-DStream executes incremental processing on data stream, that is, adds new generated synopsis data, removes old data and continuously updates clusters along with the increase or deletion of data stream. This paper uses two-phase density clustering algorithm to cluster data streams, so the building of synopsis data is different from others. It applies wavelet transformation on the data in basic windows and conducts one-phase density clustering to build synopsis data. Synopsis data is a structure that can continuously update a representative dataset feature in memory which is much smaller than data scale, D-DStream is used for the approximate processing and online analysis of the synopsis data. The description of the algorithm is as follows:

Divide the data stream sequence fragment of window w into m basic windows at t, apply technology of wavelet transformation on the data of these m data fragments and transform the high-dimensional data structure into other low-dimensional one. Then apply one-phase TDBScan algorithm on the processed data stream and get the synopsis data. At last, take the generated synopsis data as static clustering data source, and then conduct the secondary destiny clustering method and update clusters, that is, produce k clusters as requirement. The design of D-DStream algorithm embeds DWT algorithm and TDBScan algorithm, the description of TDBScan is as follows:

1 Build *R**-tree for the new data in basic window and find corresponding distance when k-dist curve turns steep to smooth, then determine the reasonable value of *Eps*;
2 Select any point p from data set and query it in basic window;
3 If p is the core point, then find all the point directly density-reachable from p and form a cluster contains p;
4 Otherwise, label p as noise point;
5 If there is not any marked point in the data set of basic window, then randomly select a point and repeat the operation mentioned above;
6 Otherwise, get initial clusters;
7 Get synopsis data;

The whole process of clustering algorithm is shown in Figure 3.

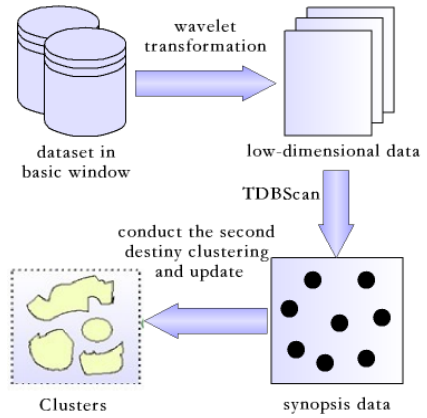

Figure 3: The flow diagram of clustering algorithm.

E. Algorithm Algorithms

Both building R*-tree and drawing k-dist graph are very time-consuming, especially of large-scale database. Besides, users need trial and error to select appropriate k-dist value, so we should preprocess the data stream before clustering.

DBScan algorithm can execute clustering with random shape, but its fatal shortcoming is that analysts need to set the value of ε and *MinPts* by subjective judgment and that clustering result is hypersensitive to the parameters. Obviously it's unrealistic to fix more than two parameters before clustering because the real high dimensional data is often unevenly distributed and uneven data need continuously varying parameters. So we can dynamically design ε and *MinPts* as the functions of the ratio of the amount of data and the distribution area of data, such as: $\varepsilon = f_1(C/S), MinPts = f_2(C/S)$, in which C denotes the amount of data and S denotes the area around cure object o. Then make comparison of the D-DStream, STREAM and CluStream.

The clustering result gotten by STREAM may be controlled by historical data, while D-DStream algorithm uses sliding window to process data stream, applies wavelet network to compress data stream and builds a much smaller synopsis data structure to save main characteristics of data stream, then clusters with two-phase density clustering algorithm, which has a good impact on real-time update of data, solves the shortcoming of STREAM. As CluStream algorithm doesn't eliminate “old data” on line and causes the increasing cost of process, while in the process of secondary clustering, D-DStream algorithm processes on the basis of synopsis data and gets new clusters, compare the clusters with origins, the number of data won’t increase, on the contrary, D-DStream algorithm may cause the smaller of density threshold and merge some clusters far away, thus will reduce the number of data and get better result.

IV. EXPERIMENTAL RESULT

The program is written in Matlab under the Matlab 7.9 running on Windows server 2008. The tests were performed on a Core(TM) i7 2.67GHz with 4 GMB Memory and 500GB Hard disk. The experimental data is derived from the network intrusive data stream of KDD-CUP’99 which is widely used in many data stream clustering literatures.

A. The Comparison of Execution Time

In the process of experiment, we need to read 20000 data from dataset and select 34 data attribute values, then set the number of clusters as 8 and make comparative test. Use equal proportion (for example 10%-20%, 20%-30% and so on) to randomly exert data from dataset, and record the cost of time. The result of experiment is shown in Figure 4.

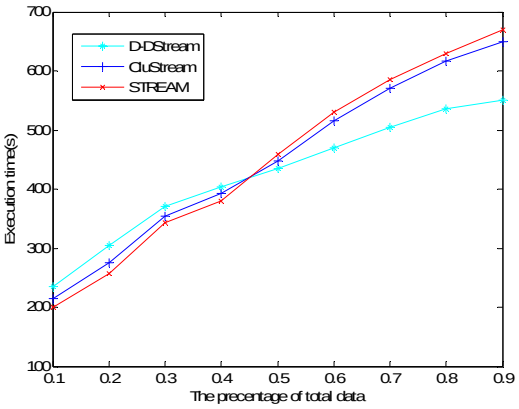

Figure 4: The comparison of execution time.

What we can know from Figure 4 is that short execution time and high processing speed of STREAM and CluStream can be achieved when the amount of data is small, but with the increasing data, D-DStream shows its advantage on both aspects, execution time rises slowly instead of rapidly and processing speed is faster. It illustrates that D-DStream is more appropriate in processing large amounts of data, effectively overcomes the shortcoming of inherence and gets the expected results.

B. The Comparison of Clustering Quality

Use SSQ(Sum of Square Distance) to compare the clustering qualities of D-DStream, STREAM and CluStream on the basis of real dataset. SSQ is a method to compare k-partition clustering quality, which measures the quality of k-partition of algorithm by calculating the distance between every point and its clustering center, smaller the value of SSQ is, better the clustering quality of algorithm is. For D-DStream, SSQ is the quadratic sum of the distance between the point and corresponding cure object. Then we can know from Figure 5 that the clustering quality of D-DStream is better than STREAM and CluStream which rely too much on historical data and neglect new data, while D-DStream introduces synopsis data to solve the problem.

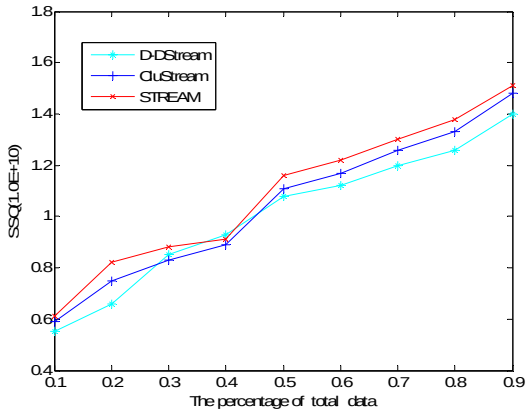

Figure 5: The comparison of clustering quality

V. CONCLUSIONS

This paper proposes a data stream clustering algorithm (D-DStream) based on wavelet network and two-phase density. In essence, wavelet network applies the technology of wavelet transformation to compress data which flows into time window, also it has been strictly proved and is able to get good results, while TDBScan is improved on the basis of density-based spatial clustering (DBScan).Because STREAM is controlled by historical data and CluStream is difficult to describe nonspherical cluster or eliminate "old data" on line, D-DStream introduces method of wavelet network and two-phase density to solve these problems, meanwhile, it gets clustering results with high quality and consumes little time and memory. This single-scan, incremental update algorithm can process noise and old data by adjusting the parameter of density, and is demonstrated effective and accurate on the basis of real dataset. For a future work, this paper suggests studies on how to accurately design the varying synopsis data in sliding window and simplify the structure of D-DStream algorithm.

ACKNOWLEDGMENT

This work was supported in part by NSFC of China under Grant No. 71071140, 60905026 and 71071141, Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20103326110001, Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20093326120004), Zhejiang Science and Technology Plan Project (No. 2010C33016, 2012R10041-09), and the Key Technology Innovation Team Building Program of Zhejiang Province (No. 2010R50041) as well as Zhejiang Provincial Natural Science Foundation of China under Grant No. Z1091224 ,LQ12G01007 and Y6110628.

REFERENCES

[1] Jingli Zhou,Xuejun Nie, Leihua Qin, Jianfeng Zhu. Web clustering based on tag set similarity. Journal of computers.vol.6, no.1, 2011,pp.59-66.
[2] Xie Juanying,Jiang shuai, Xie weixin, Gao xinbo. An efficient global K-means clustering algorithm.Journal of computers. vol.6, no.2, 2011,pp.271-279.
[3] Huang Chenghui, Yin Jian, Hou Fang.. Text clustering using a suffix tree similarity measure. Journal of computers.vol.6,no.10, 2011,pp.2180-2186.
[4] Xie Linquan,Wang Ying,Yu Fei,Xu Chen, Yue Guangxue.. Research on intrusion detection model of heterogeneous attributes clustering. Journal of software. vol.7,no.12, 2012,pp.2823-2831.
[5] A.Alzghoul and M.Löfstrand. Increasing availability of industrial systems through data stream mining .Comput Ind Eng. vol.60,no.2, 2011,pp.195-205.
[6] Chen,L., Zou,L.J and Tu,L. A clustering algorithm for multiple data streams based on spectral component similarity. Inform Sciences. vol.183, no.1,2012,pp.35-47.
[7] Lei Dajiang,Zhu Qingsheng, Chen Jun, Lin Hai, Yang Peng.Automatic PAM clustering algorithm for outlier detection. Journal of software.vol.7,no.5.2012.pp.1045-1051.
[8] Niu Qiang,Huang Xinjian.An improved fuzzy C-means clustering algorithm based on PSO.Journal of Software. vol.6,no.5,2011.pp.873-879.
[9] S.Guha, N.Mishra, R.Motwani and L.O'Callaghan. Clustering data streams: Theory and Practice. IEEE T Data En. vol.15,no.3,2003,pp.515-528.
[10] C.C.Aggarwal, J.Han, J.Wang and P.Yu. Proceedings of the 29th International Conference on Very Large Data Bases, September 12-13;Berlin,Germany,2003.
[11] C.C.Aggarwal, J.Han, J.Wang and P.S.Yu..Proceedings of the 30th International Conference on Very Large Data Bases, August 31 - September 3;Toronto, Canada,2004.
[12] C.C.Aggarwal, J.Han and P.S.Yu. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 22 -25; Washington, USA,2004.
[13] V.Ganti., J.Gehrke and R.Ramakrishnan,.DEMON mining and monitoring evolving data. IEEE T Data En. vol.13,no.1,2001,pp.50-63.
[14] Q.Tu,J.F.Lu,B.Yuan,J.B.Tang,J.Y.Yang.Density-based hierarchical clustering for streaming data. Pattern Recogn Lett. vol.33, no.5,2012, pp.641-645.
[15] Giuseppe Di Fatta, Francesco Blasa, Simone Cafiero, Giancarlo Fortino. Fault tolerant decentralised K-Means clustering for asynchronous large-scale networks. J papallel distr com.vol.73,no.3, 2013.pp.317-329.
[16] Chang,J.L., Cao,F and Zhou,A.Y. Clustering Evolving Data Streams over Sliding Windows. Chinese Journal of Software.vol.18,no.4,2007,pp.905-918.
[17] Zhu,W.H., Yin,J and Xie,Y.H. Arbitrary Shape Cluster Algorithm for Clustering Data Stream. Chinese Journal of Software. vol.17,no.3,2006,pp.379-387.
[18] Liu, X., Mao,G.J., Sun,Y and Liu,C.N. An Algorithm to Approximately Mine Frequent Closed Itemsets from Data Streams. Chinese Acta Electronica sinica. vol.35,no.5,2007,pp.900-905.
[19] Chen,H.H., Shi,B.L., Qian, J.B and Chen,Y.F. Wavelet Synopsis Based Clustering of Parallel Data Streams. Chinese Journal of Software. vol.21,no.4,2010,pp.644-658.
[20] Feng,S.R and Xiao,W.J.An Improved DBSCAN Clustering Algorithm. Journal of China University of Mining and Technology. vol.37,no.1,2008,pp.105-111.

Chonghuan Xu received his B.S. and M.S. degrees in Computer and Information Engineering from Zhejiang Gongshang University, Hangzhou. Now he is a lecturer in College of Business Administration, ZheJiang Gongshang University. His research interests include electronic commerce, data mining. He has published over 10 publications in academic journals and conference proceedings.