Applied Mathematics & Information Sciences
An International Journal

A Novel Method of Data Stream Clustering Based on Wavelet Timing Series Tree Synopsis

Dongsheng Liu[1], *Chonghuan Xu* [2,*], *and Shujiang Fan*[3]

[1]College of Computer Science & Information Engineering, Zhejiang Gongshang University, Hangzhou 310018 P. R. China
[2]College of Business Administrator, Zhejiang Gongshang University, Hangzhou 310018 P. R. China
[3]College of Computer Science & Information Engineering, Zhejiang Gongshang University, Hangzhou 310018 P. R. China

Received: 29 Aug. 2012, Revised: 10 Nov. 2012, Accepted: 17 Jan. 2013
Published online: 1 May 2013

Abstract: For the difficulty of obtaining cluster result fast and effectively under the limitations of bounded memory and time, this paper proposes a novel data stream clustering method based on wavelet timing series tree synopsis to solve the problem. The proposed method considers the attenuation characteristic of data stream, which combines the dynamic maintenance of wavelet coefficient and attenuation feature of wavelet coefficients of data stream, and can achieve approximate representation of data stream fragment information and dynamic maintenance of its synopsis structure. The proposed method employs wavelet timing series tree synopsis method to compress data stream fragment, then adopts two-phase density clustering algorithm to cluster. Detailed experiments show that the proposed method can get high compression quality, good space and time efficiency and good clustering results.

Keywords: synopsis structure, attenuation feature, wavelet transformation, density clustering

1. Introduction

Existed methods 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 data stream applications, the volume of data is usually too large to be stored or scanned more than once. A good data stream approach allows processing of potentially infinite amounts of data. It scans the stream ideally in a single pass, keeping just 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.

Clustering is the task of assigning a set of objects into groups (called clusters) so that objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. It is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Hence data stream clustering is a challenging area of research that attempts to extract useful information from continuously arriving data [1–3].

The aim of our work is to develop a method that can handle with data stream clustering effectively. In order to solve the problem of getting cluster result fast and effectively under the limitations of bounded memory and time of existing data stream clustering method, a novel data stream clustering method based on wavelet timing series tree synopsis is proposed. We fully consider the attenuation characteristic of data stream. The proposed method combines the dynamic maintenance of wavelet coefficient and attenuation feature of wavelet coefficients of data stream, and achieves the approximate representation of data stream fragment information and the dynamic maintenance of its synopsis structure. First, we use wavelet timing series tree synopsis method to compress data stream fragment, and then two-phase density clustering algorithm is adopted to cluster. Detailed simulation analysis demonstrates that the presented method achieves high efficiency of space and time and is more stable. The rest of the paper is organized as follows. Review of related work and description of knowledge are briefly given in Section 2. In Section 3, the

* Corresponding author e-mail: talentxch@gmail.com

proposed method to process on the data stream with wavelet timing series tree synopsis W-TWS is presented. In Section 4, the improved density clustering algorithm based on W-TWS is proposed. Section 5 provides experimental results of W-TWS-DStream on simulation datasets. Finally, we conclude in Section 6.

2. Background

2.1. Related work

Data stream clustering has been comprehensively investigated in recent years; several important algorithms [4–14] have been actively introduced. At present data steam clustering algorithms are mainly proposed and improved by Guha, Aggarwal and others, such as 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 [15], 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 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 [16] is proposed, which is an algorithm for clustering incoming data streams based on user-specified, online clustering queries. CluStream 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 process cost. HPStream[17], a modification of CluStream to enable clustering of high-dimensional data was proposed, which employs a data projection method to reduce the dimensionality of data stream to a subset of dimensions that minimize the radius of cluster 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 [18]. Birch[19] 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. In literature [6], a validity index based method of adaptive feature selection is proposed, incorporating with which a new text stream clustering algorithm is developed. During the clustering process, threshold of cluster valid index is used to automatically trigger feature re-selection in order to ensure the validity of clustering. J.Skla and I.Kolingerov [7] proposed a novel approach to handle large amounts of geometric data. A data stream clustering is used to reduce the amount of data and build a hierarchy of clusters. The cluster hierarchy is then used in a dynamic triangulation to create a multiresolution model. It allows for the interactive selection of a different level of detail in various parts of the data. The clustering and the triangulation are supplemented by an elliptical metric to handle data with anisotropic properties. Ling Chen et al.[9] proposed a new algorithm to cluster multiple and parallel data streams using spectral component similarity analysis, a new similarity metric. This algorithm performs auto-regressive modeling to measure the lag correlation between the data streams and uses it as the distance metric for clustering. And the algorithm uses a sliding window model to continuously report the most recent clustering results and to dynamically adjust the number of clusters. Beringer and Huellermeier[10] proposed an online version of the K-means clustering algorithm, which uses a discrete Fourier transforms (DFT) approximation of the original data and utilizes the low-frequency (instead of all) coefficients to compute the distance between two streams. The problems caused by process are as follows: 1) The length of window is fixed and can't be changed by user requirements; 2) For high-speed data stream, memory space might not save all the data of certain length of the sliding window. Yang J [11] used the sum of snapshot difference with weight as the measure of the distance between two data streams, but it could not reflect the similarity of changing tendency among data streams. Yeh MY et al[12] used Piecewise Linear Approximation (PLA) method to compress data streams, analyzed the correlation coefficient between two data streams, then clustered them by measuring the similarity among them. Its shortcoming is that it can't select a certain segment according to the user requirements. Tu Li et al.[13] proposed a multiple data streams clustering algorithm based on correlation analysis, which compresses the raw data of multiple data streams as a synopsis and clusters them. This method has fine efficiency, but yet has some problems in the process of compressing synopsis. It just overlaps the attributes of data stream and saves them. In addition, it needs to build a new matrix Cij which consumes too much storage space. Lhr and Lazarescu[14] presented an incremental

graph-based clustering algorithm whose design was motivated by a need to extract and retain meaningful information from data streams produced by applications such as large scale surveillance, network packet inspection and financial transaction monitoring.

2.2. Discrete Wavelet Transform

The main principle of DWT (Discrete Wavelet Transform) [20] is that saves the wavelet coefficient which plays an important role on reconstruction data via wavelet transforming to the original data, so as to achieve the purpose of data compression. Haar wavelet is one of the simplest DWT. One dimensional Haar wavelet decomposition can transform the time series $X = \{x_1, x_2, ..., x_n\}$ into r wavelet coefficients $\{y_1, y_2, ..., y_r\}$, and this wavelet coefficient sequences can restore the original data well. For example, we assume that series $X = \{8,3,4,5,6,3,6,6\}(n = 8)$, the Haar wavelet processes on X are showing in Table 1, finally we get the wavelet coefficient sequences:
$\{5, 0, 0.5, -1, 2.5, -0.5, 1, 0\}$.

Table 1 The Haar wavelet processing on series X

level	Time series	Wavelet coefficient
l=3	{8,3,4,5,6,3,6,6}	-
l=2	{5.5,4.5,4.5,6}	{2.5,-0.5,1,0}
l=1	{5,5}	{0.5,-1}
l=0	{5}	{0}

Data compression based on wavelet transform takes advantage of wavelet decomposition, that just reserves the most $r(r < n)$ important wavelet coefficient, and it will reconstruct the original sequence approximatively. We called these r coefficients as wavelet synopsis of the original series. Wavelet transform based tree attenuation synopsis can be described as error tree structure intuitively, as shown in Figure 1. The nodes in the tree $w_i(i = 0, 1, ...)$ are corresponding to wavelet coefficients and leaf nodes $x_i(i = 0, 1, ...)$ are corresponding to the original data. In the tree T and the nodes $w_k(k = 1, 2, , n-1)$, we let$leaves_k$ denote all leaf nodes collection of root node $w_k, path_k$ represent the set of nonzero coefficients on the route from the w_k to leafs in the tree $T, lleaves_k$denote all leaf nodes collection of left subtree of root node w_k and $rleaves_k$ denote all leaf nodes collection of right subtree of root node $w_k.w_0$ is the mean of all data and r_k is the mean of $rleaves_k$ and l_k is the mean of $lleaves_k$, so $w_k = (l_k - r_k)/2$. From the forming processes of error tree we can know that reconstruction of the original data $x_k$only related to the coefficient of $path_k$. It means that $x_i = \sum_{w_j \in path_k} \varepsilon_{ij} w_j$, if $x_i \in lleaves_j$ or $j = 0$, then $\varepsilon_{ij} = +1$; if $x_i \in rleaves_j$, then $\varepsilon_{ij} = -1$.

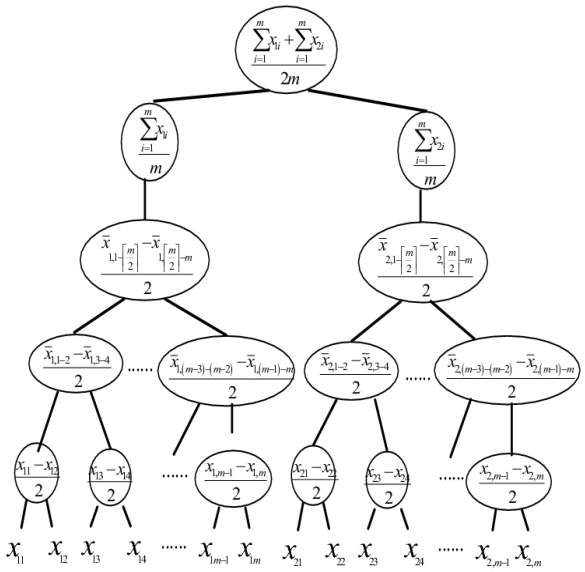

Figure 2.1 Structure of error tree synopsis

2.3. Density-based spatial clustering

DBScan [21] is a classic clustering algorithm based on density, which is used to filter 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*.

3. W-TWS based on wavelet timing series tree synopsis

With continuous arrival of data stream, the older the data are, the higher levels will the data be compressed. Moreover the higher level of the data, the lower weight of the data node. It means that the higher lever of the data have a high degree of attenuation. We called this synopsis structure as wavelet transform based tree attenuation synopsis (W-TWS). Generally, when we reconstruct the original data sequence, the longer the data in the sequence are, the greater error is. However the feature of W-TWS makes the longer data have a greater attenuation, and these data have a little influence on reconstruction of the total data stream. Hence it can ensure the quality of reconstruction.

3.1. Main idea of W-TWS

Definition 1: let four units $(t,m,\bar{X},\Phi)$ denote the synopsis information of data node P, in which t is the time stamp of the data node, it means the arrival time of last data in D; m is the number of the data in D;$\bar{X}$ is the mean of the data and Φ is a set of the r most important wavelet coefficients $\{w_1,w_2,...,w_r\}$.

The main idea of W-TWS is as follows:

(1) Preprocess the continuous arriving data stream and adopt the idea of sliding window to control the amount of processing data every time.

(2) Set the pretreated data stream segment as the $0th$ level, do compress operation on each m data, then generate the $1th$ level of compressed data nodes $P_i, i=(1,2,3...)$. Introduce nonlinear attenuation function f to weight P_i, the reason is that the nonlinear attenuation function is more conform to the principles of people's psychology and can be understand easily. In this paper, we construct a nonlinear attenuation function $f_i=-\alpha e^{-(t_i-t_0)}+\beta$ based on the forgetting curve which was proposed by Hermann Ebbinghaus. In the attenuation function, α,β are attenuation factors and $\alpha+\beta=1,\alpha>\beta$,$\alpha,\beta\in[0,1]$ and they can be adjusted to improve the precision of function. t_0 denotes the initial moment, when at the moment t_i the weighted value of the data node P_i is P_if_i. Adopt DWT transform to compress data nodes, where Φ is used to preserve r most important wavelet coefficient after transformation.

As different functions of different wavelet coefficients in reconstruction, we need to normalize the wavelet coefficients. Generally we divide every coefficient by $\sqrt{2^l}$, so the coefficient transform into $w_1/\sqrt{2^l},w_2/\sqrt{2^l},...,w_r/\sqrt{2^l}$, where l is the level. The metric of reconstruction error determines the selection of the most important wavelet coefficients. The screen process of wavelet coefficient should obey following principles: 1) Missing the coefficient with big absolute value has greater influence on reconstruction of related data. 2) We can know from the structure tree that if the coefficient close to root node influences more data in the process of reconstruction, then it owns greater importance. In this paper, we adopt the sum squared error $sse(D,D')=\sum_{i=1}^{n}(x_i-x_i')^2$ to screen the wavelet coefficient, and the measure of importance of the coefficient based on the minimize of the sse. Hence $\Phi$is used to preserve r greatest absolute value of the coefficient.

(3) With the arrival of new data, the compressed data node P in the $1th$ level increases continuously. When reach a certain number, the m oldest data nodes will be merged into a data node in the $2th$ level and calculate the synopsis information of the data node. Do like this layer by layer so that make sure the data stream can be compressed into several hierarchical data nodes. In the whole calculation processes, memory only store n data nodes and weed out the old data constantly.

Additivity[22] of the compressed data nodes ensure the synopsis information of data stream segment obtains from merger and the synopsis information obtains from the original data directly are consistent, so each level of the tree structure can be dynamic maintenance. The lower level of data node corresponding to the shorter segment of data stream and have a better degree of approximation in the process of reconstruction. Conversely, the higher level of data node corresponding to the longer segment of data stream has a rough degree of approximation in the process of reconstruction. Figure. 3.1 shows the construct process of wavelet tree attenuation synopsis and dynamic maintenance, in which l_i denotes the layer of the tree, and merge m sequence data nodes in each layer and weed out the old data timely. *Count* is used to count the data nodes, and it will reset after weed out m data nodes.

3.2. Concrete steps of W-TWS

Step 1. Data pretreatment : the pretreatment of data mainly include the processing on missing value of data and outliers elimination. Assume the number of attributes of data stream is a, if the number of missing attributes $a'\geq a/2$, we think the data stream is invalid and filtrate directly. Or the missing value of attributes will be replaced by the average value of all attributes.

Step 2. Assume the data stream is stable and set the data arrive continuously as the $0th$ level, the data stream arrive at the same time, on average, include m data. Assume every m data make up a segment and record it as di, the number of subsegments in the $0th$ level will be $[\frac{n}{m}]$, in which n is the total number of data entering the memory.

Step 3. Compression: we use DWT to compress the segment d_i that consists of m data from the $0th$ level and generate some data nodes of the $1th$ level, and let P_i denote the data nodes.P_i will preserve the synopsis information of d_i. Meanwhile, we introduce attenuation function f to weight P_i and get P_i'. With the arrival of new

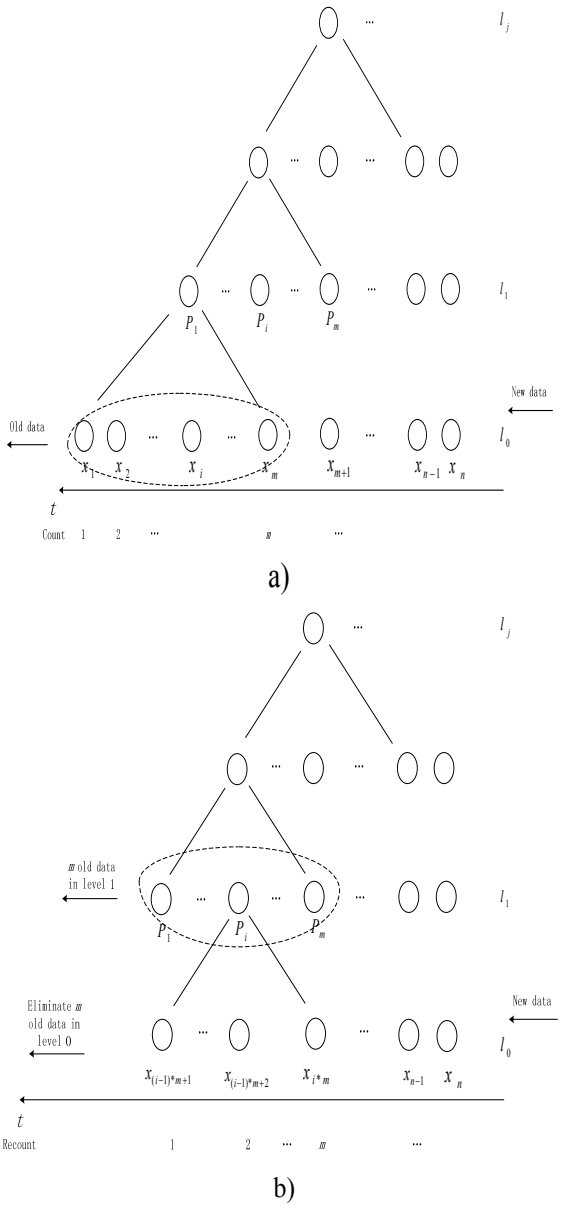

Figure 3.1 he construct process of wavelet tree attenuation synopsis and dynamic maintenance

data, the compressed data node in the $1th$ level increases continuously, and when it reaches a certain number, the m oldest data nodes $\{P'_1, P'_2, ..., P'_m\}$ will be merged and become a data node of the $2th$ level, and then calculate the synopsis information of the data node, and so on. At last, the data stream will be compressed as several hierarchical data notes, and a synopsis tree save wavelet coefficients will be constructed.

Step 4. Screening of wavelet coefficients, select r most important wavelet coefficients based on the minimization principle of sum squared error (SSE).

Step 5. Dynamic maintenance of the W-TWS, update the nodes of the synopsis tree dynamically and merge the m oldest nodes of the l_i level into l_{i+1}level according to the additivity of the data nodes, and then recalculate the rest nodes of the l_i level, meanwhile weed out these m old data nodes from memory.

4. Improved density clustering algorithm based on W-TWS

4.1. W-TWS-DStream algorithm

W-TWS-DStream is used for the approximate processing and online analysis of synopsis data. It 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. Synopsis data is a structure that can continuously updates a representative dataset feature in memory which is much smaller than data scale. In W-TWS-DStream, we use two-phase density clustering algorithm to cluster data stream.

As raw data don't be saved after wavelet transform, this paper calculates the weight Φ of data node P according to the characteristics of distance keeping of wavelet transform, and then clusters. The experimental result in section 5 proves that the calculation of weight is effective, and the cluster quality is great.

The description of the algorithm is as follows: Apply W-TWS on data fragments. Then use one-phase TDBScan algorithm to address the processed data stream and build 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 W-TWS-DStream algorithm embeds W-TWS algorithm and TDBScan algorithm. The description of TDBScan is as follows:

1 Build R^*-tree for new data in basic window and find corresponding distance when k-dist curve turns steep to smooth, and 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 points 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, randomly select a point and repeat the operation mentioned above;
6 Otherwise, get initial clusters;
7 Get synopsis data;

4.2. Algorithm analysis

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.

W-TWS-DStream 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 W-TWS-DStream, STREAM and CluStream.

The clustering result gotten by STREAM may be controlled by historical data, while W-TWS-DStream algorithm applies W-TWS to compress data stream and builds a much smaller synopsis data structure to save main characteristics of data stream, and then clusters with two-phase density clustering algorithm, which has a good impact on real-time update of data, and 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, W-TWS-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, W-TWS-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.

5. Experimental result and analysis

The program is written in Matlab and java under the Matlab 7.9 running on Windows 2003, because Matlab do well in calling java programs. The tests were performed on a Core(TM) i7 2.67GHz with 4 GMB Memory and 500GB Hard disk. These data are from UCI dataset of time series- the dataset of Ozone Level Detection[23]. It collects detected data of Houston, Galveston and Brazoria from 1998 to 2004 and contains 2536 data and each one has 73 properties.

Choose sum square error as screening index. First make pretreatment to get the first layer dataset with standardization, good attribute independence and no vacancy values and abnormal values. When we apply W-TWS to handle the original data, if the data number is small than 2^n, it can use parallel continuation method for supplement. We segment data stream to reduce the wavelet decomposition calculation. The amount of original data for every compute is 1024. For example, the first layer data fragment is $X_1 = \{x_1, x_2, ..., x_{1024}\}$, $X_2 = \{x_{1025}, x_{1026}, ..., x_{2048}\}$....Data node of each layer is merged by two lower data nodes. By that analogy, data synopsis extraction is continuous.

Figure 5.1 shows the reconstruction error comparison between W-TWS-sse and Haar-sse. Set the parameters $\alpha = 0.5, \beta = 0.5$ in W-TWS-sse. From the figure, we can see that W-TWS-sse is more superior when important wavelet coefficient r is larger.

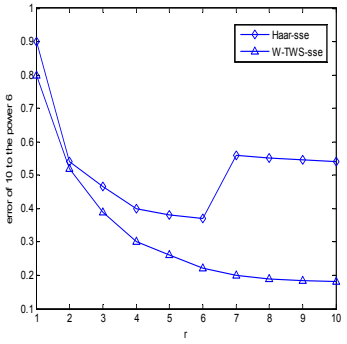

Figure 5.1 Reconstruction error comparison between W-TWS-sse and Haar-sse

Experiment 1: use SSQ (sums of the squares of the distances between data points and the centers of clusters) to test quality of two-phase density clustering algorithm based on Wavelet tree synopsis compression, smaller the SSQ is, better the clustering quality is. This paper compares W-TWS-Dstream , CluStream and STREAM. Figure. 5.2 shows that W-TWS-DStream's clustering quality is better than other two algorithms.

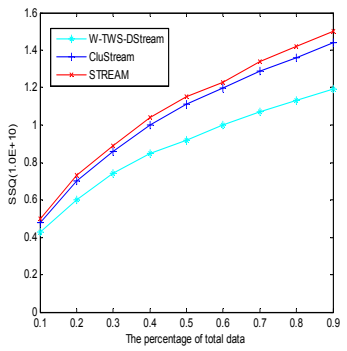

Figure 5.2 The clustering quality of three algorithms

Experiment 2: test time and space efficiency of two-phase density clustering algorithm based on Wavelet tree synopsis compression. This paper compares W-TWS-Dstream , CluStream and STREAM for their advantages or disadvantages through the two indexes of execution time and memory consumption. The results are as follows:

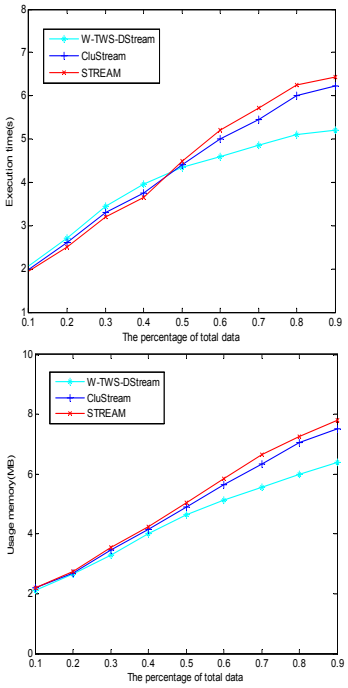

Figure 5.3 The comparison of time and space efficiency between three algorithms

What we can know from Figure. 5.3 is 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, W-TWS-DStream shows its advantage on both aspects, execution time rises slowly instead of rapidly and processing speed is faster. It illustrates that W-TWS-DStream is more appropriate in processing large amounts of data, effectively overcomes the shortcoming of inherence and gets the expected results. In aspect of memory consumption, W-TWS-DStream always has good advantage.

Besides this paper adopts external criteria to test clustering effect from the aspects of accuracy, precision and recall. According with the known data structure ,we use the criteria to assess how well the clustering result fits the known classification. Accuracy (AC), Precision(PE), Recall(RE) are defined as follows[24]:

$$AC = \frac{\sum_{i=1}^{k} a_i}{n}, PE = \frac{\sum_{i=1}^{k} \frac{a_i}{a_i+b_i}}{k}, RE = \frac{\sum_{i=1}^{k} a_i/(a_i+c_i)}{k}.$$

where n denotes the number of objects of dataset, a_i denotes the number of objects which assigned to class i correctly, b_i denotes the number of objects which assigned to class i incorrectly. c_i denotes the number of objects which should be assigned to class i but not. k denotes clustering number.

Contrast experiment: we make a comparison to CluStream algorithm, two-phase density clustering algorithm based on Haar transform and two-phase density clustering algorithm based on W-TWS. Table 2 shows the comparison of clustering effect.

Table 2 Comparison of clustering effect

Validation Measure	CluStream	Haar-DStream	W-TWS-DStream
AC	0.562	0.641	0.744
PE	0.528	0.623	0.726
RE	0.547	0.630	0.731

From the Table 2, we can see that the clustering effect of W-TWS-DStream is better than traditional one. The result is average value of five independent experiments.

Experiment 3: investigate parameters influence of W-TWS. The proposed W-TWS in this paper is mainly referred to two parameters of nonlinear attenuation function:$f_i = -\alpha e^{-(t_i - t_0)} + \beta$, that is attenuation rate factor:$\alpha, \beta, \alpha + \beta = 1, \alpha > \beta, \alpha, \beta \in [0,1]$. This parameter is used to increase the accuracy through the adjustment for attenuation function.

Figure. 5.4 shows that it has he best accuracy when $\alpha = 0.6, \beta = 0.4$.

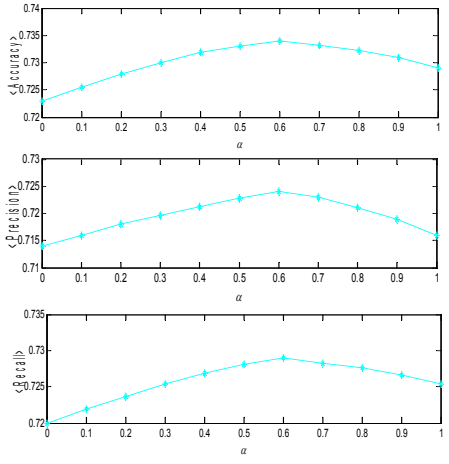

Figure 5.4 Parameters influence

6. Conclusions

This paper studies existing data stream clustering algorithms, aim at their problem of getting cluster result fast and effectively under the limitations of bounded memory and time. And we proposed a novel data stream clustering method based on wavelet timing series tree synopsis, which fully considers the attenuation characteristic of data stream, combines the dynamic maintenance of wavelet coefficient and attenuation feature of wavelet coefficients of data stream, and achieves the approximate representation of data stream fragment information and the dynamic maintenance of its synopsis structure. First, we use wavelet timing series tree synopsis method to compress data stream fragment, and then use two-phase density clustering algorithm to cluster. Detailed simulation analysis demonstrates that the presented method shows high compression quality, good space and time efficiency, and gets good clustering results. In the future work, we will focus on the following aspects: study other data stream synopsis constructor method and introduce excellent ideas into the structure of wavelet tree to further improve the efficiency and quality, and launch the processing of data stream concept drift based on synopsis structure, to satisfy the needs of practical application better.

Acknowledgement

This research was supported by National Natural Science Foundation of China (Grant No. 71071140 and 71071141), Natural Science Foundation of Zhejiang Province (Grant No.Z1091224, Y1090617 and LQ12G01007) as well as Major Technology Plan Project of Hangzhou (Grant No. 20112311A09),Key Innovation Team of Zhejiang Province (Grant No. 2010R50041).
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] B. Babcock, S. Babu, M. Datar, R. Motwani, J. Widom, Proceedings of the 21st ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems.ACM Press, Madison, Wisconsin, USA **21**, 1-16 (2002).
[2] P. Domingos, G. Hulten, Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Santa Barbara, California, USA. 1-5 (2001).
[3] Huang, Shih-Chang, Applied Mathematics & information sciences **6**, 331S-337S (2012).
[4] Nam Hun Park, Won Suk Lee, Data & Knowledge Engineering **63**, 528-549 (2007).
[5] Nam Hun Park, Sang Hyun Oh, Won Suk Lee, Information Sciences **180**, 2375-2389 (2010).
[6] Linghui Gong, Jianping Zeng, Shiyong Zhang, Expert Systems with Applications **38**, 1393-1399 (2011).
[7] J.Skála, I.Kolingerová, Computers & Geosciences **37**, 1092-1101 (2011).
[8] Ahmad Alzghoul, Magnus L fstrand. Computers & Industrial Engineering **60**, 195-205 (2011).
[9] Ling Chen, LingJun Zou, Li Tu, Information Sciences. **183**, 35-47 (2012).
[10] J. Beringer, E. Huellermeier, Data and Knowledge Engineering **58**, 180-204 (2006).
[11] Yang J, In: Proc. of the 19th IEEE Int'l Conf. on Data Engineering(ICDE 2003). Bangalore **19**, 695-697 (2003).
[12] Yeh MY, Dai Biru, Chen M S, IEEE Transactions on Knowledge and Data Engineering **19**, 1349-1362 (2007).
[13] TU Li, CHEN Ling, ZOU Ling-Jun, Journal of Software in Chinese **20**, 1756-1767 (2009).
[14] Sebastian Lühr, Mihai Lazarescu, Data and Knowledge Engineering **68**, 1-27 (2009).
[15] S.Guha, N.Mishra, R.Motwani, L.O'Callaghan. IEEE Transactions on Knowledge and Data Engineering **15**, 515-528 (2003).
[16] C.C.Aggarwal, J.Han, J.Wang, P.Yu. Proceedings of the 29th International Conference on Very Large Data Bases. **29** (2003).
[17] C.C.Aggarwal, J.Han, J.Wang, P.S.Yu. Proceedings of the 30th International Conference on Very Large Data Bases, Toronto, Canada. **30** (2004).
[18] C.C.Aggarwal, J.Han, P.S.Yu. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **10** (2004).
[19] V.Ganti, J.Gehrke, R.Ramakrishnan. IEEE Transactions on Knowledge and Data Engineering **13**, 50-63 (2001).
[20] Cormode G, Muthukrishnan S, Journal of Algorithms **55**, 58-75 (2005).
[21] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) **2**, 226-231 (1996).
[22] Huahui Chen, Bole Shi. Journal of computer research and development **46**, 268-279 (2009).
[23] Ozone Level Detection Data Set, 2008. http://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection.
[24] Liang Jiye, Bai Liang, Cao Fuyuan, Journal of computer research and development **47**, 1749-1755 (2010).

Dongsheng Liu Dongsheng Liu received his PHD degree in the school of information engineering, in 2008, Zhejiang Gongshang University (ZJGSU), China. He is currently an associate Professor in the school of information engineering at ZJGSU. His research interests include data mining, electronic commerce and wireless network.

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.

Shujiang Fan was born in Jiangxi province in 1987, a graduate student in Zhejiang Gongshang University (ZJGSU) China., His research interests include Electronic Commerce, Logistic Optimization, context aware and Data mining.