Artif Intell Rev (2015) 43:593–600
DOI 10.1007/s10462-013-9398-7

Research on data stream clustering algorithms

Shifei Ding · Fulin Wu · Jun Qian · Hongjie Jia · Fengxiang Jin

Published online: 29 January 2013

Abstract Data stream is a potentially massive, continuous, rapid sequence of data information. It has aroused great concern and research upsurge in the field of data mining. Clustering is an effective tool of data mining, so data stream clustering will undoubtedly become the focus of the study in data stream mining. In view of the characteristic of the high dimension, dynamic, real-time, many effective data stream clustering algorithms have been proposed. In addition, data stream information are not deterministic and always exist outliers and contain noises, so developing effective data stream clustering algorithm is crucial. This paper reviews the development and trend of data stream clustering and analyzes typical data stream clustering algorithms proposed in recent years, such as Birch algorithm, Local Search algorithm, Stream algorithm and CluStream algorithm. We also summarize the latest research achievements in this field and introduce some new strategies to deal with outliers and noise data. At last, we put forward the focal points and difficulties of future research for data stream clustering.

Keywords Data mining · Data stream · Clustering · Data model

1 Introduction

In recent years, data stream has emerged as a new form of data and is attracting more and more attention. This kind of information has significant research value and practical meaning

S. Ding · F. Wu · J. Qian (✉) · H. Jia
School of Computer Science and Technology, China University of Mining and Technology,
Xuzhou 221116, China
e-mail: 791995837@qq.com

S. Ding
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Science, Beijing 100190, China

F. Jin
College of Geomatics, Shandong University of Science and Technology, Qingdao 266590, China

in all aspects of human lives, such as monitoring information for network media transmission, sensor transfer information in coal mines, access to website information, economic information produced by finance and securities companies, weather forecast information, and so on. Because of this form of data is massive and updated in real time, traditional clustering methods cannot be used to process it; so the need to discover new clustering methods is becoming more and more urgent.

Data stream mining is a real time process aimed at extracting interesting patterns from rapid and dynamic data. The biggest challenge is how to find valuable information in massive data streams during single scanning. As an effective tool for data stream mining, data stream clustering has aroused wide concern in recent years. Although the research of data stream clustering is still in its infancy, it has made great progress. Many valid data stream clustering algorithms have been proposed and the experimental results are quite satisfactory (Bifet et al. 2009; Sun et al. 2008). These algorithms can be applied in many fields, such as climate monitoring, agricultural, engineering control and so on (Talbot et al. 1999; Hui et al. 2008). The latest research achievements in this field need to be followed up and summarized. So this paper gives an overview of the various data stream clustering algorithms, and analyzes their pros and cons. At last we put forward some valuable research directions for data stream clustering, and point out the emphases and difficulties of the research.

This paper is organized as follows: the second part summarizes classical data stream clustering algorithms; the third part analyzes new data stream clustering algorithms; the fourth part describes some future works and finally makes a conclusion.

2 Classic data stream clustering algorithm

Compared with traditional data forms, data stream has its own characteristics. Traditional data is static and stable. It can be accessed at any time and processed more than once. Data stream is dynamic and it flows like a stream, which means that it is sequential and changes over time. “Real-time”, “continuous” and “ordered” are frequently used to describe data stream, meanwhile, “large amount of data”, “potentially unlimited”, “arrival rate uncertain” are also its obvious features. In view of these characteristics, traditional clustering algorithms no longer meet the requirements of data stream clustering, we have to improve traditional methods or even develop new algorithms. The concept of data stream clustering was first proposed in 2000, then researches in this field attracted broad attentions (Guha et al. 2003). Data stream clustering can be categorized in different ways. According to the characteristics of the data, it can be divided into the following categories (Ordonez 2003; Guha et al. 2003; Gaber et al. 2005; Babcock et al. 2002): density-based clustering, probability-based clustering, correlation-based clustering and mixed attribute-based clustering. There are different methods to process data streams, such as sliding window method, dynamic grid method, multi-classifier method, and dissimilarity matrix method. All of these methods have the same research goal that they are committed to find an effective way to deal with massive high-dimensional data streams, or try to solve the clustering problems that data information is uncertain, dynamic and distributed in arbitrary shape.

In the early development of data stream clustering, the focus of the research was to construct a suitable data structure for data streams based on traditional methods. Birch algorithm, LocalSearch algorithm, Stream algorithm and CluStream algorithm are classic data stream clustering methods proposed at the early time.

Birch algorithm (Zhang et al. 1996) is a typical integrated hierarchical clustering algorithm. Its basic idea is: calculating the distances between the new data point and all known data points; then comparing these distances with a threshold to determine the category of the new data point. Birch algorithm uses clustering feature tree *CF* and clustering features to group the data points. Clustering feature vector is defined as $CF = (n, LS, SS)$, where *LS* is linear sum, *SS* is square sum. The tree *CF* stores the characteristics of hierarchical clustering, and it has two parameters: branching factor B and threshold T. Tree *CF* can be constructed dynamically and does not require all data to be stored in advance, so it is very suitable for the clustering of data streams.

The complexity of the Birch algorithm is $O(n)$, and it can get good clustering results through only one traversal. But Birch algorithm does not work effectively for the data with arbitrary shape. When the sample data is non-convex, the performance of this algorithm is often unsatisfactory due to the poor clustering accuracy. In order to improve the deficiencies of Birch algorithm, researchers have done a lot of work, and M-Birch algorithm was proposed (Ling et al. 2007). M-Birch algorithm is an extension of Birch algorithm. M-Birch algorithm can dynamically adjust the threshold value, so that the clustering results can be optimized. Experiments show that M-Birch algorithm performs better than Birch algorithm. But it is still not able to essentially improve the quality of clustering. In 2003, Barbará proposed that data stream clustering algorithm must meet three requirements: compact expression, fast processing, dealing with outliers quickly and accurately (Barbará 2003). Most improved algorithms cannot satisfy these three requirements very well.

As early as around 2000, Guha et al. (2003) proposed LocalSearch algorithm. It uses divide and conquer strategy to get the clustering results of data streams. This algorithm can only describe data streams, but it cannot reflect the changes of data streams, still less can it forecast and analyze the tendency. So in 2002, the new algorithm—Stream was proposed by O'Callaghan et al. (2002). It is a continuation and development of LocalSearch algorithm. What's more, O'Callaghan also proved that in most cases, the clustering performance of Stream algorithm is better than Birch algorithm.

In 2003, Guha further improved Stream algorithm (Guha et al. 2003), but he ignored the fact that data stream is time-varying and it presents different patterns in different time. So his study did not achieve a breakthrough. Then (Aggarwal et al. 2003) proposed CluStream algorithm to solve the problem of data stream clustering. This algorithm is a framework, in which the clustering process is divided into two stages: online micro-clustering and offline macro-clustering. In this way, CluStream algorithm can cluster not only the recent data, but also the data of the specified time period. The first stage is online micro-clustering. Then store the micro-clusters in a structure called pyramid time frame. In this step, the idea of micro-clustering is similar to Birch algorithm that they both use eigenvalues to mark sub-clusters, so they will share the same shortcomings. The second stage is offline macro-clustering, in which the results of micro-clustering will be further analyzed according to the user's specific requirements. As the second step is based on the first step, the first step clustering results are vital to the entire data stream clustering.

CluStream algorithm can only use discrete attribute values. As for continuous attribute values, CluStream algorithm will lose some data information. So (Aggarwal et al. 2004) proposed HPstream algorithm, which can process high-dimensional data streams with continuous attribute values. HPsteam algorithm introduces the idea of subspace clustering, and creates a recession structure to classify data streams. When dealing with high dimensional data streams, the performance of HPstream algorithm is better than CluStream algorithm. HPstream algorithm abandons the snapshot storage technique used in CluStream framework, and uses projection method to reduce data dimensions. But it requires inputting the average clustering dimension first, and the value of this parameter is difficult to determine in practice. Besides it does not solve the clustering problem of non-convex data or the data with arbitrary shape. CD-Stream algorithm (Sun et al. 2004) proposed by Sun Huanliang is a new data stream clustering method based on space division. This algorithm is better than others on the data streams of non-convex shape. But it uses the whole space for clustering, so it cannot handle high-dimensional data streams very well.

There are many improved CluStream algorithms, and the typical one is A-Clustream (Zhu et al. 2006; Wu et al. 2009), which is an arbitrary shape clustering algorithm based on data stream proposed by Zhu Weiheng in 2006. This algorithm improves the time efficiency and clustering accuracy of CluStream algorithm. In 2007, Ni Weiwei improved existing algorithms and proposed CLUSMD algorithm (Ni et al. 2007; Yang et al. 2010). It uses k-means algorithm to partition data streams; then save the representative points of each partition, and give up the other data points to accommodate the capacity limit of the memory; finally classify these representative points according to their density. CLUSMD algorithm solves the problem that data streams are not evenly distributed in high-dimensional space. Yang and Zhou (2007) proposed mixed-attribute data stream clustering algorithm—HClustream. This improved algorithm has a good performance on standard data sets.

In 2004, Motoyoshi et al. (2004) presented a data stream clustering algorithm based on regression analysis. The advantage of this algorithm is that it doesn't depend on the initial values and merging clusters. However, this algorithm is only suit for the data streams in which partial regression exist. In 2005, Song proposed an online data stream clustering algorithm based on probability density (Song and Wang 2005). This algorithm builds a new date structure frame according to Gaussian mixture model theorem. The above algorithms can only solve certain problems, and they are only valid in specific areas. Therefore, these algorithms failed to be applied or concerned widely.

3 New data stream clustering algorithms

With the development of data stream clustering algorithms, the research focus on the following aspects: the first one is the clustering for irregular distributed data, which is also the most important direction of research in the field, and until now it hasn't achieved a breakthrough yet; the second is how to detect outliers and exclude the interference of noise data. Researchers have done a lot of work in these two areas, and proposed many efficient algorithms.

In 2005, Aggarwal tried to use sub-space projection technique (Aggarwal et al. 2005) to solve the problem of ”dimensions disaster”. After that, he proposed UMicro algorithm for clustering uncertain data streams (Aggarwal and Yu 2008a; Aggarwal et al. 2004). UMicro algorithm extends CF structures into ECF structure, which can describe the uncertain part in data streams. In China, Tang Changjie proposed probability intervention strategies (Wang et al. 2011, 2009) for uncertain data streams, and suggested that hot partitions should be separated and Kolmogorov complexity can be used in text data stream mining (Wang et al. 2011).

Based on these researches, in 2010, Zhang et al. (2010) put forward EMicro algorithm. This algorithm not only takes into account the distance between tuples and the attribute-level uncertainty of data tuples, but also emphasizes tuples' own characteristics - the existence-level uncertainty. As ECF cannot handle the problem of existence-level uncertain data streams, EMicro algorithm uses UCF (uncertain clustering feature) data structure (Chang et al. 2007; Cao et al. 2007) to deal with the probability of data flow. In addition, this algorithm also develops a novel method to detect outliers, and introduces new evaluation standards for clustering quality.

EMicro algorithm uses buffer mechanism to solve the problem of outliers, and its specific strategy is: keep two buffers in memory to store clusters and outliers respectively. The cluster selection method uses the principle of gravity, and provides the Find Optimal Cluster algorithm. EMicro algorithm takes into account both the distance factor and probability factor during the process of assigning a new tuple to an existing micro-clusters, and comes up with a new way to solve the problems of outliers and data storage structure. Compared with the previous uncertain data stream clustering algorithms, EMicro algorithm has a lot of improvements and performs well for data streams.

The data in data streams has uncertainty. If the data appears with a certain probability, this data flow can be defined as probabilistic data stream (Cormode and Garofalakis 2007; Jayram et al. 2008, 2007). In 2009, P-Stream, a new data stream clustering algorithm, was proposed by Dai et al. (2009). They come up with the concept of strong cluster, weak cluster and excessive cluster for the probability tuples in data streams. They also design a cluster selection strategy, which can effectively assign each arrived tuple to an appropriate cluster.

Given an uncertain tuple sequence $S = \{< \mathrm{v}^1, \mathrm{p}^2 >$...<$\mathrm{v}^k$, p^k> ...}, where v is the value of tuple, p is the probability of tuple, and p^i is in the range [θ,1]. Suppose a cluster of data tuples $C = \{< \mathrm{v}^1, \mathrm{p}^2 >$... <$\mathrm{v}^n$, p^n>}, its existing probability *EPC* is defined as the average existing probability of all tuples. For the given parameter $a(1 \geq a \geq 0)$, if cluster C satisfy EPC $\geq$ a, then cluster C can be called a strong cluster, and a is the lower bound of the strong cluster. P-Stream algorithm not only gives the concept of strong cluster, weaker cluster, excessive cluster, but also designs the method to find candidate clusters, and uses Model Outdated Process to deal with outliers. Find Candidate Cluster algorithm not only considers D_i ($D_i$is the distance between <v, p> and the center of candidate cluster C_i), but also takes the value of EP^{C_i} into account when candidate cluster is selected for <v, p>. CluStream algorithm only considers the distance factor in the selection of candidate clusters. In P-Stream algorithm, Find Candidate Cluster strategy first sorts $C_i$to form a cluster sequence $C' = \{C_{j1}, C_{jk}\}$, and then uses heuristic rules to determine the candidate clusters. It can be demonstrated that the new strategy can find more strong clusters than the original method when SSQ (sum of squared distance) is limited.

Due to the various flaws of CluStream algorithm for processing outliers, P-Stream algorithm uses adaptive model outdated process. In the offline clustering stage, the probability of data stream is transferred into statistical information by micro-cluster snapshots. When a clustering request (f, k) arrived, the stored time indexes will extract the closest micro-cluster snapshot to the moment t from disk. As merging strong clusters can still get a strong cluster, the clustering doesn't need to consider the relationship of the existence probability between clusters. Compared with CluStream algorithm, P-Stream can find a reasonable time to store micro-cluster snapshots, and when handing outliers, P-Stream algorithm is more optimized. So for the clustering of arbitrary shaped data streams, P-Stream algorithm should have larger research and development space.

Chen and Shi (2010) proposed wavelet synopsis based clustering of parallel data streams. This algorithm mentions two concepts: hierarchical amnesic summary (HAS) and discrete wavelet transform (DWT) (Guha and Harb 2005; Karras and Mamoulis 2005). Discrete wavelet transform is a data compression method, and the oblivion characteristic (Bulut and Singh 2003; Palpanas et al. 2004) of hierarchical amnesic structure can dynamically maintain the hierarchical data node of data streams. Using these two tools can construct synopsis structure W-HAS, which is wavelet based HAS in the hierarchical data stream. Then dynamically calculate the approximate distance and clustering center, and apply W-HAS to the on-line clustering of parallel data streams.

The idea of the dynamically maintenance is: continuously receive the arrival of data at a given level; if the data nodes reach a certain number, then merge the oldest N data nodes and generate a higher level data node. Obviously, the level lower and the data stream sequence shorter. The approximation between the corresponding summary information and the original information is better. However, as the hierarchy deepens, the approximation degree of summary information and the original information will be lower and lower. This also reflects the amnesic characteristic of this algorithm for data streams. But this feature should under control, otherwise the inaccuracy will be large, and the clustering results will be unsatisfactory. So researchers define the relative reconstruction error E of data information, and make $E < \varepsilon$ (ε is a parameter). W-HAS clustering algorithm consists of three steps: calculate the wavelet coefficient of clustering center; extract data nodes from original data streams, and calculating the wavelet coefficient of normalized data streams; dynamically update the wavelet coefficient of data streams and use k-means method to get the final clusters.

W-HAS structure and the additive data nodes reduce the space complexity of the algorithm. And with small space complexity, it can meet the environmental requirements of data stream clustering. What's more, because W-HAS saves the entire synopsis structure of data streams, we can select proper data nodes to analyze the required data streams, and are not limited to fixed length sliding window, which is an innovation compared with the data stream clustering based on sliding window.

In addition to the above researches, in China, Zhu et al. (2011) proposed a double-window-based classification algorithm for concept drifting data streams in 2011; Zhang et al. (2011) researched the continuous dynamic skyline queries over data streams; Han et al. (2011) explored load shedding strategies on sliding window joins over data streams. In foreign countries, Kavitha and Punithavalli (2010), committed to the study of time series data stream and proposed feasible data stream clustering algorithms. All of them have achieved remarkable achievements.

4 The further work

As we know, data stream clustering is a difficult task in the real world, and recently has caused wide public concern. The proposed studies provide a way to investigate the existing algorithms and techniques for time series data stream clustering and help to find the directions for future enhancement. Future research can be directed to the following aspects:

1) High dimensionality will affect the efficiency and accuracy of the algorithm, so for high dimensional data streams, it is necessary to find an effective dimension reduction method.
2) Adaptive partitioning and using slipping technique need a large amount of computation. How to reduce the computational complexity of data stream clearing is worthy of further research.
3) In most applications the data information are usually not sufficient for statistical analysis, so developing new methods to deal with uncertain data streams is also an important research aspect.
4) Time series data stream is very common in the real world, but its research has just begun. We should design an effective approach to predict the next value in time series.
5) Subspace clustering can effectively deal with massive high-dimensional data, so introducing subspace method into data stream clustering is also a good idea.
6) Combine data stream clustering with specific applications and explore the appropriate algorithm to solve practical problems. Test and improve data stream clustering algorithm in the process of practice.

5 Conclusion and prospect

Data stream clustering is the product of technological progress, and along with the development of information science. After ten years of research, it has made brilliant achievements. This paper provides an overview of various data stream clustering algorithms that proposed in recent years, and introduces the theoretical framework of these algorithms as well as their applications in our daily lives.

However, as a novel clustering method, data stream clustering is still immature and imperfect. Data information should be extracted more accurately and in real time, but it is hard for data stream clustering algorithm to meet this requirement. In reality, most data streams are of arbitrary shape and spread unevenly. Uncertain data and noise points in data streams will also interfere with the clustering process. These aspects are still the focal points and difficulties of future research for data stream clustering. In this paper, the uniqueness and drawbacks of past studies and some possible topics for further study are also discussed. In the future, we will research on data stream deeply, and try to develop an effective algorithm to solve the current problems of data stream clustering.

Acknowledgments This work is supported by the National Basic Research Program of China (No.2013CB 329502), the National Natural Science Foundation of China (No.41074003), and the Opening Foundation of Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences (No.IIP2010-1).

References

Aggarwal CC, Han J, Wang J et al (2003) A framewrok for clustering evolving data streams. In: Proceedings of VLDB 2003. pp 81–92

Aggarwal CC, Han J, Wang J, Yu PS (2004) A framework for projected clustering of high dimensional data streams. In: Proceedings of the 30th international conference on very large data bases. pp 852–863

Aggarwal CC, Yu PS (2008) A framework for clustering uncertain data streams. In: Proceeding of the 24th international conference on data engineering. pp 150–159

Aggarwal CC, Yu PS (2008) Outlier detection with uncertain data. In: Proceeding of the SIAM data mining conference pp 483–493

Aggarwal CC, Han J, Wang J et al (2005) On high dimension projected clustering of uncertain data streams. Data Min Knowl Discov 10(3):251–273

Babcock B, Babu S, Datar M, et al (2002) Models and issues in data streams. In: Proceedings of the 21th ACM symposium on principles of database systems. pp 1–16

Barbará D (2003) Requirements for clustering data streams. ACM SIGKDD Explor Newsl 3(2):23–27

Bifet A, Holmes G, Pfahringer B (2009) New ensemble methods for evolving data streams. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. pp 139–148

Bulut A, Singh AK (2003) SWAT: hierarchical stream summarization in large networks. In: Proceeding of the 19th international conference on data engineering. pp 303–314

Cao F, Zhou A (2007) Fast clustering of data stream using graphics processors. J Softw 18(2):291–304

Chang J, Cao F, Zhou A (2007) Clustering evolving data stream over sliding windows. J Softw 18(4):905–918

Chen H, Shi B (2010) Wavelet synopsis based clustering of parallel data streams. J Softw 21(4):644–658

Cormode G, Garofalakis M (2007) Sketching probabilistic data streams. In: Proceedings of the ACM SIGMOD international conference on management of data. pp 281–289

Dai D, Zhao W, Sun L (2009) Effective clustering algorithm for probabilistic data stream. J Softw 20(5):1313–1328

Dingi H, Trajcevski G, Scheuestern P, Xiaoyue W, Eamonn K (2008) Querying and mining of time series data: experimental comparison of representations and distance measures. In: ACM Proceedings of the VLDB endowment. 1(2):1542–1552

Gaber MM, Zaslavsky AB, Krishnaswamy S (2005) Mining data streams: a review. SIGMOD Rec 34(2):18–26

Guha S, Meyerson A et al (2003) Clustering datastreams: theory and practice. IEEE TKDE Special Issue Clust 3(2):37–46

Guha S, Meyerson A, Mishra N et al (2003) Clustering data streams: theory and practice. IEEE Trans Knowl Data Eng 15(3):505–528

Guha S, Harb B (2005) Wavelet synopsis for data streams: minimizing non-euclidean error. In: Proceeding of the 11th ACM SIGKDD international conference on knowledge discovery in data mining. pp 88–97

Guha S, Mishra N, Motwani R et al (2000) Clustering data streams. In: Proceedings of the 41st annual symposium on foundations of computer science. pp 359–366

Guha S, Mishra N, Motwani R et al (2000) Clustering data streams. In: Proceedings of the 41st annual symposium on foundations of computer science. Washington: IEEE Computer Society. pp 359–366

Han D, Gong P, Xiao C (2011) Load shedding strategies on sliding window joins over data streams. J Comput Res Dev 48(1):103–109

Jayram TS, Kale S, Vee E (2007) Efficient aggregation algorithms for probabilistic data. In: Proceeding of the 18th annual ACM-SIAM syrup. On discrete algorithms(SODA). pp 346–355

Jayram TS, McGregor A, Muthukrishan VE (2008) Estimating statistical aggregates on probabilistic data streams. ACM Trans Database Syst 33(4):26–30

Karras P, Mamoulis N (2005) One-pass wavelet synopses for maximum-errormetrics. In: Proceeding of the 31st international conference on very large data bases. pp 421–432

Kavitha V, Punithavalli M (2010) Clustering time series data stream—a literature survey. Int J Comput Sci Inf Secur IJCSIS 8(1):289–294

Mahdiraji AR (2009) Clustering data stream: a survey of algorithms. Int J Knowl Based Intell Eng Syst 12(2):39–44

Motoyoshi M, Miura T, Shioya I (2004) Clustering stream data by regression analysis. Duned Aust Comput Soc 32:115–120

Muthukrishnan S (2003) Data streams algorithms and applications. In: Proceeding of the 14th annual ACM-SIAM symposium on discrete algorithms. pp 13–413

Ni W, Lu J, Chen G, Sun Z (2007) Efficient data stream clustering algorithm based on k-means partitioning and density. J Chin Comput Syst 28(1):83–87

O'Callaghan L, Mishra N, Meyerson A et al (2002) Motwani. Streaming data algorithms for high-quality clustering. In: Proceedings of the 18th international conference on data engineering. pp 685–704

Ordonez C (2003) Clustering binary data streams with K- mean. In: Proceedings of DMKD'03. pp 12–19

Palpanas T, Vlachos M, Keogh E (2004) Online amnesic approximation of streaming time series. In: Proceeding of the 20th international conference on data engineering. pp 339–349

Song M, Wang H (2005) Highly efficient incremental estimation of gaussian mixture models for online data stream clustering. In: Proceeding of intelligence computing: theory and application. pp 174–183

Sun H, Zhao F, Bao Y (2004) CD-stream-a space partition based density clustering algorithm over data stream. J Comput Res Dev 41(suppl):289–294

Sun Y, Mao G, Liu X (2008) Ming concept drifts from data streams based on muti-classifiers. Acta Automatica Sinica 34(1):93–97

Talbot LM, Talbot BG, Peterson RE (1999) Application of fuzzy grade-of membership clustering to analysis of remotesensing data. J Clim 12:200–219

Wang XZ, Li RF (1999) Combining conceptual clustering and principal component analysis for state space based process monitoring. Ind Eng Chem Res 38:4345–4358

Wang Y, Tang CJ, Li C, Chen Y, Yang N, Tang R, Zhu J (2009) Intervention events detection and prediction in data streams. Lect Notes Comput Sci 5446:519–525

Wang Y, Tang CJ, Wang Y (2011) Mining hotspots from multiple text streams based on stream information distance. J Softw 22(8):1761–1770

Wu F, Zhong Y, Jin X (2009) Arbitrary shape clustering algorithm for evolving data stream over sliding windows. J Chin Comput Syst 30(5):887–890

Xin L, Ni Z, Huang L (2007) Modifiable Birch cluster algorithm used in data stream. Comput Eng Appl 43(5):166–169

Yang C, Zhou J (2007) A Heterogeneous data stream clustering algorithm. Chin J Comput 30(8):1364–1371

Yang N, Tang C, Wang Y (2010) Clustering algorithm on data stream with skew distribution based on density. J Softw 21(5):1031–1041

Yue Wang, Changjie Tang, Ning Yang (2011) Mining optimized probabilistic intervention strategy over uncertain data set. J Softw 22(2):285–297

Zhang C, Jin C, Zhou A (2010) Clustering algorithm over uncertain data stream. J Softw 21(9):2173–2181

Zhang L, Zou P, Jia Y (2011) Continuous dynamic skyline queries over data stream. J Comput Res Dev 48(1):77–85

Zhang T, Ramakrishnan R, Livny M (1996) Birch: an efficient data clustering method for very large databases. In: Proceeding of the SIGMOD. pp 103–114

Zhu W, Yin J, Xie Y (2006) Arbitrary shape cluster algorithm for clustering data stream. J Softw 17(3):379–387

Zhu Q, Zhang Y, Hu X (2011) A double-window-based classification algorithm for concept drifting data streams. Acta Automatica Sinica 37(9):1078–1084