DE GRUYTER
OPEN

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **16**, No 5
Special Issue on Application of Advanced Computing and Simulation in Information Systems

Sofia • 2016 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2016-0053

Noval Stream Data Mining Framework under the Background of Big Data

Wenquan Yi[1]*, Fei Teng*[2]*, Jianfeng Xu*[2]

[1]*Jiang Xi Vocational College of Finance and Economic, Jiu Jiang, China*
[2]*Software College, Nanchang University, China*
Emails: ywq@jxvc.jx.cn tengfei_ncu@163.com jianfeng_x@ncu.edu.cn

Abstract: *Stream data mining has been a hot topic for research in the data mining research area in recent years, as it has an extensive application prospect in big data ages. Research on stream data mining mainly focuses on frequent item sets mining, clustering and classification. However, traditional steam data mining methods are not effective enough for handling high dimensional data set because these methods are not fit for the characteristics of stream data. So, these traditional stream data mining methods need to be enhanced for big data applications. To resolve this issue, a hybrid framework is proposed for big steam data mining. In this framework, online and offline model are organized for different tasks, the interior of each model is rationally organized according to different mining tasks. This framework provides a new research idea and macro perspective for stream data mining under the background of big data.*

Keywords: *Stream data, data mining, clustering, classification, framework.*

1. Introduction

In recent years, with continuous development of Internet and Database Technology against the background of the big data, various complicated types of data are experiencing high-speed growth. Stream data with 4V characteristic (Volume, Variety, Value, Velocity) is a unique one and exists widely in reality, such as the data stream of phone records in communication area and data stream of user clicks on the Web, etc. Different from previous static data, these stream data have the basic characteristics of high speed, high dimension and fast changing [1]. As for these types of data, the traditional data mining method is not applicable. Therefore, a novel framework for stream data mining must be build.

2. Summary of stream data mining

Stream data refers to the data which is dynamically streaming over time with unlimited growth. Stream data mining is used to reveal the experience and knowledge hiding behind the massive data, and to find the meaningful patterns, rules or abnormal points [2]. As shown in Table 1, compared with traditional data, stream data is chronological, rapidly changing, massive and unlimited, which makes the most of the traditional data mining methods not applicable to big data stream mining. Therefore, we need to apply special techniques to mining big data stream.

Table 1. The difference between traditional data mining and stream data mining

Data parameters	Traditional data mining	Stream data mining
Data type	static	dynamic
Data length	bounded	unbounded
Arrival	once	many
Update speed	slow	fast
Scanning times	multiple	single
Response time	non real-time	real-time
Time and space complexity	not strictly	strict

Random sampling technique. Random sampling for stream data and knowledge mining from these sampled data. Namely, a small part of the data is randomly selected from the data set, and the approximate query results are obtained according to the sample set.

Sliding window technique. Sliding window technique is a method of data mining based on the data of recent period time. Historical data no longer affects the result of current data as they are sliding out of the window. Assuming that *w* is the length of the sliding window, then the data element of time *t* is useless in time $t+w$.

Histogram technique. Data stream is distributed in a series of adjacent buckets based on data partitioning techniques, then, the distribution of the data is approximated by means of calculating approximate probability distribution.

The wavelet [3] transform technique. According to grid size, data is modulated to different resolution. The nearest classification block would be found by serial layer resolution scanning, and the influence of old data on classification accuracy can be reduced through setting the threshold and degradation coefficients.

Multiresolution technique. Multiresolution technique is a data protocol method. It can handle stream data with noise and concept drift by changing data granularity according to the actual situation. When no concept drift is generated, the particle size threshold can be increased to allow more data passing. When there is noise and concept drift, the threshold will be diminished slowly to explore the trend of change.

Outline data structure technology [5]. Outline data structure technology is used to balance the contradiction between the accuracy of the data stream processing and the storage structure of the data.

3. Current status of data stream mining technology

3.1. Mining frequent items set

With the continuous expansion of the application field of data mining, more and more attention has been placed on frequent items set mining in database. The researchers have designed and implemented a lot of classic algorithms such as Apriori [6], FP-growth [7], and CLOSET [8]. However, these algorithms cannot dynamically update the mining result, which make them difficult to apply to data stream mining directly. For frequent items set mining in data stream, there are many related researches in recent years. In order to deal with the dynamic characteristics of data stream, a novel window mode named weighted sliding window [9] is proposed in 2013 and it can through improvement of WSW algorithm to get more efficient result. In the same year, multi-thread mining is mentioned in reference [10]. In reference [10], linear linked list structure is used to store the current candidate set and transaction information of windows, and multi-thread method was used to generate frequent pattern. From the perspective of practical application, frequent items set mining problems of WSN and high speed network flow are studied in [13] of 2013 and [14] of 2014 and it achieved good results.

3.2. Clustering

Clustering is also called unsupervised learning [15], it divides data into different data clusters according to different data characteristics, which makes the distance between elements in same cluster as small as possible and the distance between elements in different cluster are as big as possible.

Stream [14] and Clustream [15] algorithms are the most classical algorithm in data stream clustering algorithm. On the basis of k-Means algorithm, stream algorithm uses dividing and conquering method for clustering and ensures minimum error sum of squares of elements in one cluster. However, the disadvantage of this algorithm is that it cannot detect the dynamic change of stream data. Clustream is a first two stage framework for data stream clustering algorithm and proposed in 2003. This algorithm is divided into two parts: online micro clustering and offline macro clustering. The algorithm of online phase is responsible for online statistics data stream feature vector to generate synopsis data, and store these feature vectors regularly with adopting pyramid time frame model; in the off-line phase, according to clustering parameters of user request, the more accurate clustering result is obtained by analysing synopsis data of micro-cluster of online phase. But there are still many problems with this algorithm, for example, the clustering effect of non-spherical clusters is not good, the number of clusters in the online phase requires manual intervention, input data sequence is not sensitive and the ability of dealing with noise is weak.

Although Clustream can monitor data stream with dynamic change, it has a poor ability to handle multi-dimensional data. Therefore, some scholars put forward the EWSSC [16] algorithm in 2013. This algorithm retains the characteristics of the traditional soft subspace clustering algorithm and uses the fuzzy scalable clustering strategy to apply the soft subspace clustering algorithm into the clustering analysis of stream data. In 2013, literature [17] summarizes the current data stream clustering algorithms, and proposes two kinds of common data stream clustering framework: the data stream clustering framework based on object and the data stream clustering framework based on attribute; object based clustering framework is mainly divided into data abstract layer (online part) and clustering layer (offline part). The abstract layer dynamic maintains the data structure with statistical characteristics, clustering layer uses above data structure to divide data. Attribute-based clustering framework is also called variable clustering, mainly used to deal with data streams with high dimensional attributes, such as DGClust [18] algorithm. In 2014, two stages clustering thought is used in [19] and it can solve the clustering efficiency optimization problem of Clustream algorithm used in the super large scale data stream. In this algorithm, firstly, it uses time slider window as time unit to real-time acquire stream data, and in the online phase, information of stream data is used to miniaturize clusters in real-time manner and these miniaturized clusters are maintained by updates and deletion. Then, in the off-line phase, data set of miniaturized cluster sample is weighed according to their attribute information contribution and distance in sample category. Finally, real-time clustering is conducted for these miniaturized clusters.

Analysing from clustering analyzed objects and data characteristics, the current data stream clustering algorithms are mainly focused on the improvement of the accuracy of clustering results, which includes the reduction of time and space complexity. In 2014, reference [21] is proposed in order to study data clustering problem with large-scale distribution and it gets better clustering effect. From the view of application, a k-Means optimization clustering algorithm based on slope sorting is proposed in 2014 [22], and it is practically used in data clustering optimization and mechanical fault diagnosis, which reflects good guiding significance and practical value. In the same year, the practical application of WSN [23] and Internet users click [24] are used to data stream clustering research.

In addition, in the stream data clustering analysis, some scholars have analyzed the concept drift of data stream, according to the phenomenon that the data characteristics change along with the change of the data stream. In 2014, three dissimilarity measurement methods [25] are proposed to study data similarity and dissimilarity measurement problem of data clustering with concept drift, which is respectively named dissimilarity measurement between point and cluster based on information entropy, dissimilarity measurement between clusters and dissimilarity measurement between clusters is based on sample standard deviation. This algorithm has a good effect on clustering result generation and concept drift detection.

3.3. Classifications

Classification algorithm is a supervised learning algorithm, the classification algorithm can obtain a learning model through the learning of an existing training set, and it can give the forecast classification information [26]. In the classification problem, there are two main research areas: classification based on concept drift, classification based on sample classifier.

For classification based on sample classifier, a lot of scholars focus on the research of sample time efficiency and memory utilization. Such as VFDT algorithm [27], it is proposed by Domingos and Hulten [25] in 2000 and has many advantages in terms of speed and accuracy. VFDT algorithm uses information entropy to select attributes and uses Hoeffding tree to make decision support and constraint handles data stream with high speed. It solves the problem of a high precision decision tree obtained with single pass on the data stream. In addition, there are NIP, VFDTc, etc. These algorithms improve the time efficiency and memory utilization, but helpless for concept that gradual or abrupt change. For this problem, Hulten, Spencer and Domingos [26] propose a method named CVFDT. CVFDT retains the quick and accurate characteristic of VFDT, but also can adapts the change of concept through continuously updating subtree or generating fungible subtree in sliding window over time. The limitation of CVFDT is consuming a large amount of memory if subtree growing continuously, and it has poor robust for data with large quantity of noise and frequent concept drift. In addition to concept drift tracking and model updating, in data stream classification still exist real-time processing massive data, the model stability and robustness problems. Therefore, in recent years, many scholars conduct related research from different ways.

Furthermore, many current algorithms need training set with experts label to train classifier, which is not realistic for stream data with high speed due to high time and cost expenses for instance labelling. At this time, if the supervised learning method is used to train the classifier, the sparsely labelled data can be obtained with a weak classifier. Therefore, in 2015, a classification algorithm based on active learning [29] is proposed. It selects small part samples to label and these samples have low classification confidence coefficients, thus, it can greatly reduce the number of instance with manual labelling.

For classification based on concept drift, classification features or classification model will change over data streaming. Concept drift is the most difficult problem in the process of data classification, currently, most methods use algorithms to detect concept drift phenomenon and then update classifier model. For example, method in [30] determines the occurrence of concept drift according to the difference of feature set of adjacent data, and then use adaptive fast decision tree with the ability of handling concept drift and online Bayes to build heterogeneous integrated classifier, which makes it can handle high dimensional data stream. In 2015, paper [31] studies the application of data flow classification with concept drift, and proposes an incremental data flow classification model based on sample uncertainty selection strategy.

There are many related algorithms of data stream classification, but they all have some common flaws: redundant attributes are not deleted on the whole and the concept drift is detected by using external attributes (such as the use of external data classification accuracy). Based on the basic principle and basic method of rough set and F-rough set, a parallel reduction and deleting redundant attributes method [32] is proposed to detect concept drift.

The abovementioned three kinds of mining methods are the main research direction in the stream data mining, in the practical application, outlier detection, intrusion detection can be achieved by using clustering, classification and other stream mining techniques. A method in reference makes full use of dissimilarity between data and frequent patterns to measure outlier degree, it can dynamically update the information of outliers in data stream by constructing the ODDS-Tree tree.

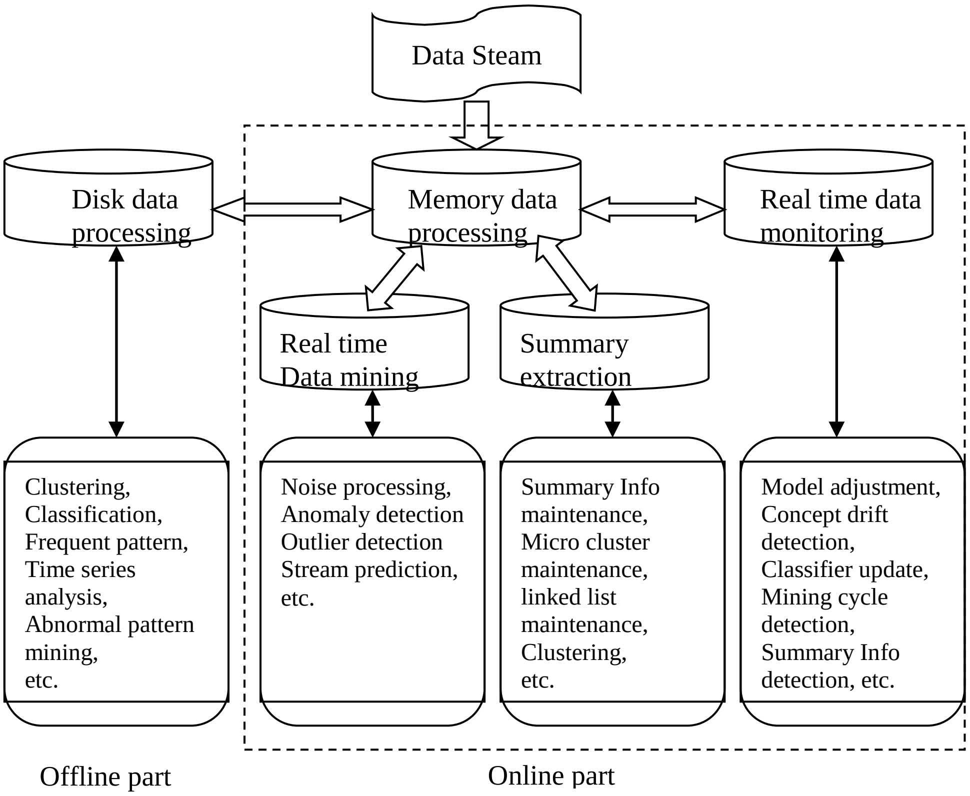
Table 2. Research and Analysis of main algorithms

Parameters	Reference	On-line	Off-line	Data scale	Sample speed	Mining mode	Publish year
Cluster	[23]	No	Yes	Small	Small	Off-line	2014
	[16]	Yes	No	Big	Big	On-line	2013
	[19]	No	Yes	Big	Big	Off-line	2014
	[21]	Yes	No	Big	Big	On-line	2014
Classification	[30]	Yes	No	Small	Small	On-line	2014
	[31]	Yes	No	Small	Small	On-line	2015
	[10]	Yes	No	Small	Small	On-line	2013
Frequent item sets	[11]	Yes	No	Big	Big	On-line	2013
	[12]	Yes	No	Big	Big	On-line	2014

4. The shortcomings of current research on streaming data mining and a new framework model

From journals and conference papers in recent years, data stream mining has become a research hotspot in the field of data mining. Many scholars propose particular mining algorithm for uncertainty of data stream mining. Considering algorithms of recent years and the characteristics of data stream, a summary table of portion algorithm is shown in Table 2.

From Table 2, the above data mining algorithms are mainly separated into two kinds: one is online mining; the other one is offline mining. However, under the background of big data, data has characteristics like Volume, Variety, Value and Velocity. These algorithms in Table 2 cannot completely adapt to the data stream with 4V characteristics. For example, some algorithms [30, 31] are still stuck in the small sample data set and difficult to deal with big data set, some algorithms [23, 16, 10] only can deal with data stream with low velocity, but it cannot handle data stream with high speed. With the development of streaming data mining under the background of big data, some researches [19, 23, 21, 12, 11] have already studied stream data mining with high speed under the super large scale. It can be seen that mining these new type data stream with 4V characteristic has become a trend. Based on the above research status, a hybrid framework model of big data stream mining is proposed and shown in Fig. 1.



Data Steam
Disk data processing
Memory data processing
Real time data monitoring
Real time Data mining
Summary extraction
Clustering, Classification, Frequent pattern, Time series analysis, Abnormal pattern mining, etc.
Noise processing, Anomaly detection Outlier detection Stream prediction, etc.
Summary Info maintenance, Micro cluster maintenance, linked list maintenance, Clustering, etc.
Model adjustment, Concept drift detection, Classifier update, Mining cycle detection, Summary Info detection, etc.
Offline part
Online part

Fig. 1. Hybrid framework model for data stream mining

In this model, the processing of data stream is divided into off-line and on-line parts. The online part can be processed on one or more data streams, the memory data processing module can choose different task modules according to the different mining request. Real time mining module is mainly aimed at the mining algorithm with high real-time requirement, such as abnormal monitoring and traffic flow prediction. As the existence of data stream's uncertainty and concept drift, the real-time data monitoring module is necessary to monitor the online data and adjust the mining strategy and information maintenance of online part according with detection result. According to the request of users, the offline part can deal with the data structure which is stored in the online section, and then return an approximate query result to the user.

5. Summaries

In this paper, a detailed study of related research of data stream mining field is presented. Many scholars propose algorithms for stream data clustering, classification and frequent items set mining, and a comparison between these algorithms is used to show currently research situation. Meanwhile, aiming at the shortcomings with large scale and high speed data set of the current stream data mining algorithms under the big data background, a hybrid framework model of stream data mining is proposed, which can be widely used in the big data stream mining field. This model synoptically summarizes the mining algorithms of data stream mining, and provides a research way to deal with a new type of data stream with 4V characteristics.

Acknowledgements**:** Authors would like to thank the anonymous referees for their constructive comments and valuable suggestions. This work is supported by the National Science Foundation of China (No 61273304).

References

1. Wu, Q., H. Sui, B. Yang et al. Research Progress of Distributed Data Stream Mining. – Computer Science, Vol. **39**, 2012, No 1, pp. 1-8.
2. Du, Z. Research on Mining Algorithm Based on Data Stream. Xi'an University of Science and Technology, 2012.
3. Law, Y. N., C. Zaniolo. An Adaptive Nearest Neighbour Classification Algorithm for Data Streams. Knowledge Discovery in Databases: PKDD 2005. Berlin, Heidelberg, Springer, 2005, pp. 108-120.
4. Aggarwal, C. C., J. Han, J. Wang et al. A Framework for Projected Clustering of High Dimensional Data Streams. – In: Proc. of 30th International Conference on Very Big Data Bases, Volume **30**.VLD B Endowment, 2004, pp. 852-863.
5. Golab, L., M. T. Özsu. Issues in Data Stream Management. – ACM SIGMOD Record, Vol. **32**, 2003, No 2, pp. 5-14.
6. Margahny, M. H., A. A. Mitwally. Fast Algorithm for Mining Association Rules. – In: Proc. of Conference AIML, CICC (pp. 36-40), Cairo, Egypt. 2005, pp. 19-21.
7. Han, J., J. Pei, Y. Yin et al. Mining Frequent Patterns Without Candidate Generation: A Frequent-Pattern Tree Approach. – Data Mining and Knowledge Discovery, Vol. **8**, 2004, No 1, pp. 53-87.
8. Fang, G., Y. Wu, M. Li et al. An Efficient Algorithm for Mining Frequent Closed Itemsets. – Information, Vol. **39**, 2015, No 1.
9. Wang, Z., X. Wang. Improvement of Data Stream Frequent Item Set Mining Algorithm WSW-Imp. – Computer Engineering and Applications, Vol. **49**, 2013, No 8.
10. Zhou, X., J. Lu, J. B. Tang. Data Stream Frequent Pattern Mining Based on Multi Thread Technology. – Computer Application, Vol. **33**, 2013, No A01, pp. 69-72.
11. Hong, Y. Frequent Item Set Mining Algorithm for Distributed Data Streams in Sensor Networks. – Computer Science, Vol. **40**, 2013, No 2, pp. 58-60.
12. Zhao, X., J. Xia, K. Fu. High Speed Network Flow Frequent Item Mining Algorithm. – Computer Research and Development, Vol. **51**, 2014, No 11, pp. 2458-2469.
13. Kim, Y., W. Kim, U. Kim. Mining Frequent Itemsets with Normalized Weight in Continuous Data Stream. – Journal of Information Processing Systems, Vol. **6**, 2010, No 1, pp. 79-90.
14. Callaghan, L. O., N. Mishra, A. Meyerson et al. Streaming-Data Algorithms for High-Quality Clustering. – In: Proc. of 18th International Conference on Data Engineering, San Jose, IEEE, 2002.
15. Aggarwal, C. J., P. S. Han. A Framework for Clustering Evolving Data Streams – In: Proc. of International Conference on Very Big Data Bases, San Francisco, CA, Morgan Kaufmann, 2003.
16. Zhu, L., J. Lei, Z. Bi et al. A Soft Subspace Clustering Algorithm Based on Data Stream. – Journal of Software, Vol. **24**, 2013, No 11, pp. 2610-2627.

17. Silva, J. A., E. R. Faria, R. C. Barros et al. Data Stream Clustering: A Survey. – ACM Computing Surveys (CSUR), Vol. **46**, 2013, No 1, 13.
18. Gan, G., C. Ma, J. Wu. Data Clustering: Theory, Algorithms, and Applications. Siam, 2007.
19. Zhang, Y., W. Wei. Research on the Clustering Method of Ultra Large Scale Data Streams under Time Series. – Computer Simulation, Vol. **31**, 2014, No 4, pp. 273-276.
20. Hu, W. An Improved Dynamic k-Mean Clustering Algorithm. – Computer System Application, 2013, No 5, pp. 116-121.
21. Candice, C. Big Data Clustering Design Based on MIP and Improved Fuzzy k-Means – Algorith. – Computer Measurement and Control, Vol. **22**, 2014, No 004, pp. 1270-1272.
22. Gao, Z., G. Chen, G. Hu et al. Tilt Factor k-Average Optimization Data Clustering and Fault Diagnosis Research. – Computer and Digital Engineering, Vol. **42**, 2014, No 1, pp. 14-18.
23. Li, Y., G. Wang. Research on Data Stream Clustering Algorithm in Data Stream Mining. – Intelligent Computer and Application, 2014, No 5, 003.
24. Wu, Y. Discussion on Data Stream Mining. – Office Automation (Integrated Edition), 2011, No 4, pp. 8-10. DOI:10.3969/j. ISSN:1007-00X-B.2011.04.002.
25. Domingos, P., G. Hulten. Mining High-Speed Data Streams. – In: Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2000, pp. 71-80.
26. Hulten, G., L. Spencer, P. Domingos. Mining Time-Changing Data Streams. – In: Proc. of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2001, pp. 97-106.
27. Xiong, Z., X. Zhou, Y. Zhang. The Lack of Labeled Data Stream Classifier. – Computer Engineering and Applications, Vol. **51**, 2015, No 6.
28. Zhang, Y., S. Liu. Data Stream Ensemble Classification Based on Feature Drift. – Computer Engineering and Science, Vol. **36**, 2014, No 5, pp. 977-985.
29. Sun, Na. Data Flow Classification Model Based on Concept Drift Detection Algorithm. – Computer Engineering and Design, Vol. **34**, 2013, No 9, pp. 3141-3145.
30. Sun, Z., T. Liu, S. Liu. Research on Incremental Data Flow Classification Based on Sample Uncertainty. – Small and Micro Computer Systems, Vol. **36**, 2015, No 2, pp. 193-196.
31. Deng, D., X. Xu, H. Huang. Concept Drift Detection Based on Parallel Reduction. – Computer Research and Development, Vol. **52**, No 5, pp. 1071-1079.
32. Tang, X., G. Li, Y. Yang. A Fast Algorithm for Mining Outliers in Data Streams. – Small and Micro Computer Systems, Vol. **32**, 2011, No 1, pp. 9-16.