Network Big Data: A Literature Survey on Stream Data Mining

Yi Wu
School of Information Science and Technology, Heilongjiang University Harbin, Heilongjiang,150080, China
wy51cn@aliyun.com

Abstract—**With the rapid development of Internet, the internet of things and other information technology, big data usually exists in cyberspace as the form of the data stream. It brings great benefits for information society. Meanwhile, it also brings crucial challenges on big data mining in the data stream. Recently, academic and industrial communities have a widespread concern on massive data mining problems and achieve some impressive results. However, due to the big data mining in data stream, those researches are not enough. In this paper, we analyzed the characteristics of stream data mining under big data environment, discussed the challenges and research issues of big data mining in data stream and summarized research achievements on data mining of massive data.**

Index Terms—**big data, stream data, data mining, data usability**

I. Introduction

As the development of information technology, people gained the ability on collecting and using data which is unprecedented. Though the big data in cyberspace, people could get better understandings on surrounding environments, people and nature; people could sense the things outside themselves more precise, efficient and instantaneous and improve the quality of their lives and work. However, because of the massive amounts of data, the questions like how to manage and use these data for serving people's work and life become problems that demanding prompt solution.

Recently, stream data mining is one of the hot topics in academia. Especially under the context of big data, the way of speedily mining the valuable knowledge from massive instantaneous data becomes the priority among priorities. In the early stage of research on stream data mining, people mainly treated stream data as variations of static data. Hence, stream data did not get enough attention. In 1999, for the first time, Henzinger proposed stream data as a new data structure model and attracted broad scholars' attention gradually. Thereafter, the problems related to stream data mining become hot spot of research. The academia proposed a large amount of research results. At present, all kinds of stream data processing based on instantaneous data have become critical issues of Internet, internet of things, social networking and other technologies. For example, the famous global chain supermarket Wal-Mart needs to deal with more than 1 million pieces of user requests every hour, maintaining a database of more than 2.5 PB. The big data in cyberspace is often dynamically and quickly generated in the form of the data stream, featuring strong timeliness. As a steady flow of big data (data stream) arrives in the Internet, internet of things and other cyberspaces, the data processing system must provide fast response and timely mine valuable information from the data in order to create wealth for the information society. The big data in applications can be divided into two sorts. One is the data stream arriving at high-speed, and the other is the persistent historical data. To instantaneously complete analysis and processing of stream data, how to mine and obtain valuable information based on the high-speed stream data and massive historical data has become a new challenge in the field of data mining.

In fact, big data came out a new research mode on data mining research field. Researchers only need to research, analyze or dig the information and knowledge in demand; they even have no direct contact with the research object. There is a widespread demand on big data mining in cyberspace. However, this facet of study does not build a complete theoretical system. Furthermore, the Algorithms and paradigms with efficiently and rapidly processing, analyzing and mining are insufficient. This article focus on large-scale stream data mining in cyberspace, combs the representative research results of stream data mining in late years (Table I), and gives out the challenges and research problems of large-scale stream data mining in cyberspace, benefits on expanding the research work of large-scale stream data mining.

II. Characteristics, Challenges and Research Issues of Big Data-Oriented Stream Data Mining

2.1 Characteristics of Stream Data Mining

In addition to the characteristics of traditional stream data mining, the stream mining in the environment of big data has many unique characteristics. Here are several key characteristics:

(1) Usability

The large-volume stream data in cyberspace can be considered as infinite, and cannot be stored in hard disk or memory for further mining. Therefore, it is required to generate valuable information from data mining with one-time calculation, which requires the generated valuable information meets the quality requirements.

doi:10.4304/jsw.9.9.2427-2434

TABLE I.
THE LITERATURE SUMMARY ON STREAM DATA MINING

No	The content of representative research		
	Topic	Methods	Reference numbers
1	classification	the integrated learning	[3]…[6]
		incremental learning	[7][8][9]
		the concept drift detection	[10][11][12]
2	Clustering	high-dimensional;	[14]
		the wavelet	[15]
		Probability	[16]
		density-based distribute	[17]
3	frequent items	transaction amount	[18]…[22]
		Time-related	[23][24]
		Approximate	[25][26][27]
4	Other	compressing technology	[28][29][30]
		Privacy mining	[31][32][33]
		data mining quality	[34]…[37]

(2) Instantaneity

Since the stream data is generated at high-speed, it has high requirement for the efficiency of data mining. Compared with traditional stream data mining methods, the large-volume and high-speed massive stream data mining must have high instantaneity on the premise of meeting availability.

(3) Diversity of mode

Massive data streams are originated from the Internet, sensor networks and other cyberspaces. The entity data forms include texts, sounds, images, videos and other modes. It is required to conduct instantaneous identification according to the identity of the stream data entity to meet the requirement of valuable information mining.

(4) Multi-source heterogeneity

Massive stream data are originated from heterogeneously distributed Internet and sensor networks, comprising the heterogeneous communication networks, computing systems, storage systems and heterogeneous physical devices (such as different sensors and different file systems, etc.).

(5) Uncertainty

The arrival of massive stream data distributed in the cyberspace is not controlled by external factors, and it will change over time. It may be influenced by measurement errors, calculation model errors, environmental noise and other factors of instability. Therefore, data mining models are completely in a passive position.

(6) High cognition

There are great difficulties for cognition of the large-volume stream data entity generated in cyberspace, the Internet of business, social networking, entertainment, services, and other areas, or in the sensor networks that detect, observe and control the physical world. Stream data mining shall be capable of effectively analyzing, integrating, obtaining, and transforming entity data.

Combining with these characteristics, this paper will investigate the data mining issues of stream data in the environment of massive data and study data mining patterns, mining models, mining algorithms and other related issues.

2.2 Challenges and Research Issues of Stream Data Mining

According to the characteristic analysis of stream data mining in the environment of big data, the current research and practice in this area is facing a lot of challenges. In view of the large amount, fast data generation speed, complex data types, low value-density and other characteristics of big data, some main challenges and research issues are proposed as below.

2.2.1 Research on the pattern of stream data mining

Most of the traditional data mining processing methods are originated from the statistical area with progressive development and evolution, which tend to be more focused on the correctness and availability of the algorithm and lack in-depth study and attention on processing large-scale data sets, high-dimensional data processing capabilities and the execution efficiency of algorithms. In addition, there are no high standards on the space and time complexity of the algorithm.

With the development of information technology, big data problems appear gradually. It is necessary to process data with the grade of TB or even PB. Furthermore, the growth trend of big data will surpass the growth rate of corresponding data processing capacity. The big data streams arrived at high-speed in cyberspace are characterized by burstiness and instantaneity. Since the data volume is too large and some data even exist in distributed forms, it is difficult to concentrate and then process these data. Therefore, it is necessary to change the computing model of big data from the centralized and top-down model to a decentralized, bottom-up and self-organized computing model [2].

To solve these problems, further research is required to achieve algorithms that meet the requirement of linear and sub-linear computation complexity and to meet the needs of massive stream data mining in cyberspace.

2.2.2 Relative issues of stream data mining

First of all, the mining research of big data in the form of stream data includes classification mining, clustering mining, frequent item mining and other traditional stream data mining issues.

Secondly, with the rapid development of the Internet, sensor networks and other information technologies, the big data mining in the form of stream data will encounter a large number of new challenges in face of the growing social demands. For example, the radio frequency identification technology (RFID) is embedded into many devices, allowing automatic identification and acquisition of generated stream data. It is necessary to study new stream data mining methods that suit new technical conditions to achieve valuable information mining in applications.

In addition, according to the morphological characteristics of stream data and business features, the data mining research can also be carried out from the following aspects:

(1) Efficiency of one-time stream data calculation. To improve the instantaneity of stream data mining and overcome limited storage space, limited transmission channels and processors and other resource constraints, approximate calculation, compression algorithms and other mining algorithms suitable for handling massive data are applied to improve the processing efficiency of data mining on the premise of ensuring the availability of results.

(2) The privacy mining in cyberspace. As a number of social institutions take the privacy and commercial interests into account in network environment, the massive heterogeneous multi-source data to be processed have separate security mechanisms. The data access for these data can only be obtained with authorization or cipher texts. Therefore, how to mine massive stream data that have privacy protection in the network environment is a challenging research issue.

(3) Massive stream data mining in resource-constrained fields. For example, in the field of sensor networking, it is required to operate more calculations and minimize excessive communications for the power consumption constraints; the increase of bandwidth of the network will cause the expansion of computing performance of multi-core processing units; the appearance of memory wall phenomenon will lead to computing "bottleneck" and other problems. Therefore, the data mining algorithms that study computing, communications, storage and other resource constraints will be in critical demand.

III. Research Progress in Big Data-Oriented Stream Data Mining

The research on big data-oriented stream data mining has just started. Below we will introduce the development progress and trend in terms of the traditional stream data mining which is suitable for massive data, the stream data mining based on morphological characteristics and business characteristics, etc.

3.1 Stream Data Mining Based on Classification

(1) Data stream classification based on the integrated learning

The semi-supervised learning strategy is used in document [3] to reduce the influence of dimension on classifiers by conducting low-dimensional subspace mapping on the high-dimensional data streams. A classifier for each different data stream is built. These individual classifiers are established using the integrated learning (thinking) thought to establish an integrated multi-data stream classification model.

A semi-supervised ensemble approach for mining data streams is presented in document [4]. Data streams are divided into data chunks to deal with the infinite length. An ensemble classification model E is trained with existing labeled data chunks and decision boundary is constructed by using E for detecting novel classes. New labeled data chunks are used to update E while unlabeled ones are used to construct unsupervised models. Classes are predicted by a semi-supervised model Ex which is consist of E and unsupervised models in a maximization consensus manner, so better performance can be achieved by using the constraints from unsupervised models with limited labeled instances.

An integrated model is proposed in document [5] to study the classification and prediction of massive data streams. The proposed model is able to use the data labels with category and the category-missing labels for training. The results of individual classifiers are summarized to obtain the final classification results by adjusting the weight of different classifiers.

Aiming at customer data stream research, a customer data classification model based on integrated learning architecture is proposed in document [6]. This model can still meet the requirements of correct data classification and dynamic update in the uneven data environment by adjusting the traditional integrated model.

(2) Data stream classification based on incremental learning

In document [7], the traditional learning vector quantization (LVQ) algorithm is improved by using the incremental learning approach. With the incremental learning mechanism, not only the learned knowledge in the traditional algorithm (LVQ) is retained in the update process, new knowledge is also learned. It greatly improves the accuracy of the model classification.

In document [8], the HoeffdingTree algorithm which uses information gain as a selective measure of the attribute is proposed. It requires only a certain number of samples to conduct attribute split at a node. The number of split is calculated with the Hoeffding constraint equation. To increase the exponential order of the algorithm accuracy, it is only needed to linearly increase the number of sample data, and the samples shall only be scanned for once for the algorithm. The incremental establishment of HoeffdingTree can simultaneously classify the data when constructing classifiers. With the arrival of data streams and the growing increase of HoeffdingTree, its mining results are becoming more accurate.

In document [9], the incremental learning approach is used to improve the support vector machine model, which can meet the classification needs of large amount of Internet videos.

(3) Data stream classification based on the concept drift detection

Aiming at the concept drift problems of data streams, a method based on EWMA is proposed in document [10] by using ideas of parallel computing and increasing model monitoring layers on each computing node. This model can monitor the accuracy of the data stream classification model and define the concept drift with the decease of accuracy. Then, the model is further adjusted.

Multiple classifiers are simultaneously monitored to reduce the monitoring loads.

In document [11], the sliding window technology is introduced. On the basis of the mining performance improvement by VFDT, the response to dynamic trend of data is further strengthened. If there are concept drifts, an alternative attribute with higher information gain will appear and make some previous nodes which no longer meet the Hoeffding constraints. Then, the replacing sub-tree with better split attributes as root nodes begins to grow. The new growing sub-tree will replace the old sub-tree when its accuracy surpasses that of the old sub-tree. The accuracy of the algorithm has been greatly improved. This model has solved the "concept drift" to a certain extent and strengthened the adaptability of the algorithm for data streams.

In document [12], the concept drift mutation is taken as the target and analyzed, and a new concept detection method, which uses clustering analysis for the new data and the original data to determine whether the concept drift occurs, is designed based on the clustering algorithm. This method is supposed to analyze different concept drift types and propose detection method of concept drift with targeted characteristics.

(4) Summary

Currently, the research work on stream data classification mining of massive data includes several aspects, for example, using distributed classification processing, parallel processing and incremental learning methods to solve the instantaneous classification problems, applying the integrated learning classification method to solve the data problem of massiveness, and taking advantage of the concept drift detection classification method for dynamic change issues. With the development of cyberspace technology, the meaningful research direction of classification mining will be the mining combined with the diversity of modes and multi-source heterogeneity of entities and based on the multi-source and multi-mode entities of stream data.

3.2 Clustering Stream Data Mining

The HPStream algorithm which aims at high-dimensional clustering issues is proposed in document [13]. Its dynamic selection associates the dimensions of smallest clustering volume with cluster, which achieves a subspace clustering algorithm. In this algorithm, the historical data get increasingly decayed with decay factors over time. When there are too many clusters, those earliest clusters will be deleted. It is (a) one kind of incremental algorithm with good instantaneity.

The parallel data stream clustering methods are studied in document [14]. When data streams arrive synchronically, the exponential weight sliding window is used to maintain data streams; the correlation of stream data is taken as the clustering objective; the K-means clustering is implemented in interval and the discrete Fourier transform is calculated with incremental methods. It is also feasible to associate the distributed computing environment data streams, which means summarizing remote information by partial clustering and obtaining approximate clustering results through transforming the coefficients.

A parallel data stream clustering method based on the wavelet concept is proposed in document [15]. This method takes the data stream itself as the clustering object and uses Haar wavelet decomposition method combined with the nature of wavelet decomposition to compress data streams, and construct a relatively small structure compared with the primitive data stream scale. Then, this structure is used for storing main data stream information. The DWT hierarchy is used to dynamically maintain the structure of each data stream. Finally, the distance of data stream computing flows is replaced with these structures and the traditional K-means clustering is used to cluster these structures.

In document [16], the clustering algorithm UMicro orienting probabilistic data streams is proposed. First, the models of uncertain data streams are defined as error models, i.e. each data item of uncertain data streams is associated with a corresponding error subject to arbitrary distribution. Then, the clustering feature structure is extended to form the error clustering feature structure, which realizes the description of uncertain parts of data streams. UMicro algorithm consists of two phases: online maintenance of micro clusters and offline production of clustering results. In this algorithm, the expected similarity between the calculating points and the micro cluster center is used to determine which cluster the uncertain data should be distributed to. Furthermore, the expected radius of this cluster is calculated for further verification of the data attributed to this micro cluster. The expected distance and expected radius belong to statistical knowledge. In order to describe the dynamic evolution of data streams, the algorithm has defined the weight for each data point, indicating the influence of different data on establishing clusters. This method (can only) can be only used to restrict particular uncertain data models.

In document [17], the algorithm DB-DDSC (Density-Based Distribute Data Stream Clustering) was proposed. The algorithm consisted of two stages. First of all, it presented the concept of circular-point which is based on the representative points and designed the iterative algorithm to find the density connected circular-points, then generated the local model at the remote site. Secondly, it designed the algorithm to generate global clusters by combining the local models at coordinator site. The DB-DDSC algorithm could find clusters with different shapes under the distributed data stream environment, avoid frequently data sending by using the test-update algorithm and reduce the data transmission.

In recent years, the stream data clustering technology has achieved great development. However, depending on the high cognitive and uncertain characteristics of massive data mining in application fields, data entities tend to exhibit high dimensions, while the traditional clustering techniques commonly use distance to measure the similarity among data objects and conduct clustering, failing to achieve classification with the distance method in high-dimensional space. To solve these problems, a number of scholars have proposed many countermeasures, such as clustering techniques for uncertain data (probabilistic data) and parallel data stream clustering techniques based on the wavelet summary, setting examples for clustering mining research.

3.3 Stream Data Mining Based on Frequent Items

(1) Mining methods are based on the frequent item of transaction amount

In document [18], in order to handle large sliding windows of data streams, the window is divided into a plurality of blocks, and (a) one block of data is updated each time to improve efficiency. What's more, the frequent pattern is directly saved by using the pattern tree PT, which is based on the concept of FP-tree, thus verifying the frequent pattern according to the account of conditions. Accurate calculating results can also be obtained

In document [19] , the problem of finding frequent itemsets from uncertain data streams is proposed. an algorithm is proposed about UDS-FIM and a tree structure UDS-Tree. Firstly, UDS-FIM maintains probability values of each transactions to an array; secondly, compresses each transaction to a UDS-Tree in the same manner as an FP-Tree (so it is as compact as an FP-Tree) and maintains index of probability values of each transaction in the array to the corresponding tail-nodes; lastly, it mines frequent itemsets from the UDSTree without additional scan of transactions.

What proposed in document [20] is the algorithm Moment that mines the frequent closed item sets of data streams in sliding windows. In this algorithm, the tree is used as a summary data structure for data compression, dynamically maintaining the boundary region between frequent closed sets and infrequent closed sets. When a new data affair arrives, the corresponding node type is determined through checking frequent closed sets stored in tables and related nodes in the tree. As much information irrelative to the current frequent closed sets is saved in internal memory when mining, the memory consumption is large. In document [21], it is proposed to mine frequent closed pattern sets in data streams based on sliding windows. The tree is applied as the storage structure to maintain the frequent closed sets within the sliding windows. It has good time and space efficiency for dense data sets.

The distributed incremental algorithm of frequent item sets is proposed in document [22]. It can find the maximal frequent item set with the incremental approach in dynamic data and study frequent item set mining in a distributed environment. The local maximal frequent item sets of distributed nodes are exchanged to obtain the superset of all maximal frequent item sets. Supersets are exchanged among all nodes to obtain the local count. The exact maximal frequent item set for all is obtained with pruning operation.

(2) Time-related frequent item methods

In document [23], combined with the advantages of the sliding window, the tilt window is applied in the FP-stream algorithm to save the support of each frequent item set. To save space, the time period is divided into different time granularity in the tilt window. The closer the time period is to the current time point, the smaller (its) granularity it gets (is). In this way, the space cost reduces while meeting the time-related queries from customers.

In document [24], it is proposed to maintain the frequent pattern with the tilt time window technology, which, to a certain extent, solved the problem of instantaneity of data streams. First of all, a frequent mode tree is established to save mode information of data streams. Then, a tilt time window table is provided for each mode. Using the frequent mode tree to maintain and update the reached data streams can achieve approximate mining for frequent modes sensitive to time. The tilt time window can maintain frequent modes in different time granularity, and it has achieved good effects.

(3) Approximate method for frequent item mining

In document [25], the approximate mining of uncertain frequent items is carried out by calculating expected frequent items. It is defined that the uncertain frequent items of the sensor stream data record all possible world instance sets as the instance space. The expected item is determined according to the probability of the instance space, the number of appearance of the frequent item in the instance set, and the expected threshold. A method with linear complexity is given to solve the problem – the expected calculating complexity is exponential.

In document [26], the probability threshold frequent item query algorithm based on the sliding window is proposed for uncertain data streams. First, the incremental search algorithm bs-UFI is proposed to avoid double counting. Then, the pruning strategy is proposed which is based on Poisson distribution to effectively filter out a large number of items that is not necessary to calculate. Finally, the cp-list data structure is proposed to compress the candidate sets among different windows for reducing storage space.

To improve the mining efficiency of frequent modes in uncertain data sets, an approximate mining strategy which aims at the problem of massive calculation when creating sub-tables is proposed in document [27]. The mining efficiency of the whole algorithm is improved at the cost of losing a fraction of frequent item sets.

(4) Summary

The frequent item mining based on the number of transactions or time-association pays great attention to the frequent item mining issues with transaction or time as the basic unit. It is often difficult to meet the high efficiency requirement of massive stream data mining. The study on the approximate method for frequent item mining of massive data in cyberspace can improve the efficiency of frequent item mining of dynamic data streams, but the cost is losing a small part of frequent item sets. The research method for how to improve the mining efficiency on the premise of ensuring the availability of frequent item mining and allowing the loss of a small part of frequent item sets is pending and needs further study.

3.4 Other Mining Methods and Mining Quality Analysis

(1) Stream data compressing technology

In document [28], the application of Haar wavelet transform for data stream compression is analyzed. The multilevel decomposition algorithm of wavelet energy is also given in the document. A new method based on the data stream coupling characteristic and the multiple data stream compression of wavelet energy decomposition is proposed. It also analyzed the sequence characteristics of different strong coincident streams, and designed the multiple data compression methods with energy conservation and the compression algorithm of multi-scale energy decomposition. The efficiency has improved 2 to 4 times compared with that of the traditional wavelet compression methods.

In document [29], the proposed method is to use the function discovery approach of gene expression programming to compress data. It takes the advantage that the GEP mode neither needs prior knowledge nor pre-stored model to propose the DEF-GEP algorithm, which is much more effective than the traditional wavelet methods. The accuracy is much higher with reconstruction of data streams.

In document [30], the coupling relationship among multiple data streams is studied with the discrete Fourier transform technology, but there is no study on storing method of the historical information of data streams.

(2) Privacy mining technology

In document [31], the randomized method is proposed to achieve the data stream anonymous privacy protection. It is necessary to maintain a tree with tuple sets and generalized information as nodes, and to constantly update the tree with the newly arrived tuples. It is also required to release node-contained tuples that meet the anonymous requirement to achieve the anonymity of data streams. If the number of nodes is restricted, it is necessary to periodically delete the released nodes. If the release time is set as δ, the average time complexity of the algorithm would be $O(|S|\delta log\delta)$, a low degree of time complexity. The spatial complexity of the algorithm is O (δ), which is no more than one limit.

In document [32], it is proposed to use the clustering thought for data stream anonymous privacy protection. It achieves the quick anonymous protection of data streams through single scan of data identification and reuses (reusing) clusters that meet anonymous conditions. This method can simultaneously reduce the loss of information for generalization and inhibition processing while meeting the requirements of anonymity, and ensure high availability of data while quickly achieving data stream anonymity. Both of its time complexity and space complexity are linear.

In document [33], a privacy protection learning machine PPLM orienting massive data is proposed for the information disclosure problem of the core support vector machine in practical applications. This method uses the core vector machine to take samples of massive data. Then, two sample points are chosen from the core set and the normal plane where the straight line consisting of the two points locates is taken as the optimal classification surface. It effectively achieves the classification of massive data and ensures the privacy and security issue in the process of classification.

(3) Research of data mining quality

Document [34] studies how to obtain stream data from the sensor networking and ensure the availability of data quality at the same time. The Hermit interpolation method and cubic spline interpolation method are used to provide the variable frequency collection algorithm of stream data and ensure acquiring the maximum data of minimum data collection, and thus ensuring the quality of the stream data

Document [35] proposed the calculation rule of the minimum quality rule based on data item grouping and its credibility, the mining algorithm, and the idea of using quality rule to detect erroneous data. This data quality rule form has borrowed the credibility assessment mechanism of the association rule and the expression ability of conditional function dependency to uniformly describe functional dependency, conditional functional dependency, association rules, and so on. It features conciseness, objectivity, comprehensiveness, and detection accuracy of abnormal data.

Document [36] described the entire process for data generation and data evolution over time. It is used for data quality assessment, data verification, data recovery, data references, etc. The data lineage can be divided into the evolutionary process among different data sources and the internal evolutionary process of a data source, namely the model level and instance-level data evolution. In the document, the research development of data lineage is generally described by taking the model level and instance-level data lineage as expression and taking query as the main line. The model level lineage mainly introduced the lineage tracking technology of query rewriting and model mapping. The instance level lineage summarized the recent research progress in terms of relational data, XML data, and stream data.

Document [37] introduced basic concepts of big data availability, discussed the challenges of big data availability, explored the research of big data availability, and summarized the research results of data availability. The document also proposed the theoretical system of big data availability, theories and methods of high quality big data acquisition, efficient auto-discovery and auto-repair algorithm of big data errors, approximate calculation theories and algorithms of weakly-available big data, and the knowledge mining theories and algorithm of weakly-available big data.

(4) Summary

Aiming at the application needs of stream data in cyberspace, it is feasible and meaningful to conduct data mining research based on stream data characteristics and business characteristics. The examples are exploring multiple data stream compression methods combined with the coupling features among multiple data streams, and conducting deep research on privacy protection mining technology of stream data and techniques of outlier mining associated with the business requirements of data security in cyberspace. Although some research work related to stream data mining has been carried out, there are still many problems to be solved, especially the quality of stream data mining, which has made very few research achievements. However, there is an urgent need for the extraction of valuable data in the applications of data mining. Therefore, there are a lot of issues to be addressed in the field of stream data mining.

IV. Conclusion

The research on big data-oriented stream data mining is still in its infancy. Many studies are originated from traditional distributed stream data mining and parallel stream data mining. In recent years, as the constant development of dynamic, high-dimensional and complex theoretical connotations and practical applications of big data, numerous characteristics, such as the traditional data analysis, data mining and data processing ways, of massive stream data are no longer applicable. For example, the high speed and large-volume stream data are generally in the form of distribution; they are difficult to be handled together. Researches on the stream data mining of non-relational data, such as complex-type, semi-structured and unstructured big data, are also very scarce. Consequently, it is necessary to systematically carry out in-depth data mining research on stream data characteristics and business characteristics in terms of concurrent processing architecture technology of high-frequent stream data mining, stream data mining methods and technologies, the business areas of stream data mining and other aspects. It is also necessary to discuss the mining pattern, mining methods and mining technologies of stream data with characteristics of big data. In short, although there are no comprehensive and systematic studies in the field of stream data mining for big data and many problems remain to be solved, the application prospect of data mining in cyberspace is very broad, and the research work in this area is provided with high practical value and academic potential.

Acknowledgment

This work is supported by the Nature Science Foundation of Heilongjiang Province with the grant number: F201325. The author would like to thank for their support.

References

[1] Ricci E.,Rugini L.,Perfetti R. SVM-based CDMA receiver with incremental active learning[J]. Neurocomputing,2006, 69(13-15):1691-1696.

[2] WANG, Yuan-Zhuo, JIN, Xiao-Long CHENG, Xue-Qi. Network Big Data: Present and Futur[J]. Chinese Journal Of Computers,2013,36(6):1125-1138.

[3] Zhang P., Li J., Wang P., et al. Enabling fast prediction for ensemble models on data streams[C]. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and Data Mining, San Diego, CA, USA, Aug 21-24, 2011: 177-185.

[4] Jing Liu, Guo-sheng Xu, Da Xiao, Li-ze Gu, Xin-xin Niu. A Semi-supervised Ensemble Approach for Mining Data Streams[J]. Journal Of Computers, 2013,8(11):2873-2879.

[5] Xiao J. , Xie L., He C. Z. ,et al. Dynamic classifier ensemble model for customer classification with imbalanced class distribution[J]. Expert Systems with Applications, 2012,39(3):3668-3675.

[6] Xu Y., Shen F. R., Zhao J. X. An incremental learning vector quantization algorithm for pattern classification[J]. Neural Computing & Applications, 2012, 21(6):1205-1215.

[7] P.Domingos and G. Hulten. Mining high-speed data streams. Presented at the Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining,Boston,Massaehuserts,UnitedStates,2000.

[8] Ewerth R. ,Ballafkir K., Muhling M.. Long-Term Incremental Web-Supervised Learning of Visual Concepts via Random Savannas [J]. IEEE Transactions on Multimedia,2012,14(4): 1008—1020.

[9] Ross G. J ., Adams N. M. ,Tasoulis D. K. , et al. Exponentially weighted moving average charts for detecting concept drift[J]. Pattern Recognition Letters, 2012,33(2):191-198.

[10] G. Hulten, L. SPeneer, et al. Mining time-changing data streams. Presented at the Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining , San Francisco California 2001.

[11] Masud M., Gao J. , Khan L., et al. Classification and novel class detection in concept-drifting data streams under time constraints[J]. IEEE Transactions on Knowledge and Data Engineering, 2011,23(6): 859-874.

[12] B. Babeoek, M. Datar, et al. Maintaining variance and k-medians over data stream windows Presented at the Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, SanDiego, California,2003.

[13] Aggarw al C C, et al. A Framework for Projected Clustering of High Dimensional Data Streams. In: Proc of the 30th VLDB Conf,2004.852-863

[14] Beringer J, Hullermeier E. Online clustering of parallel data streams[J].Data and Knowledge Engineering,2005:180-204

[15] Chen Hua-Hui, Shi Bai-Le, Qian Jiang-Bo, Chen Ye-Fang. Wavelet Synopsis Based Clustering of Parallel Data Streams[J].Journal of Software,2010,(21)4:644−658.

[16] Aggarwal CC,Yu PS.A framework for clustering uncertain data streams[C]. Proceedings of the 24th IEEE International Conference on Data Engineering(ICDE).Cancun: IEEE 2008:150-159.

[17] Bing Gao, Jianpei Zhang. Density Based Distribute Data Stream Clustering Algorithm[J]. JOURNAL OF SOFTWARE, 2013,(8),2:435-442.

[18] Mozafari B ,Thakkar H ,Zaniolo C .Verifying and mining frequent patterns from large windows over data streams//Proceedings of the International Conference on Data Engineering.Cancun,Mexico,2008:179-188

[19] Le Wang, Lin Feng , Mingfei Wu. UDS-FIM An Efficient Algorithm of Frequent Itemsets Mining over Uncertain Transaction Data Streams[J]. Journal Of Software, 2014,9(1):44-56.

[20] C. Yun, W. Haixun, et al. Moment: maintaining closed frequent item sets over a stream sliding window. in Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on, 2004, pp.59-66.

[21] N. Jiang and L . Gruenwald. CFI-Stream: mining closed frequent item sets in data streams.presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, 2006

[22] Otey M,Wang C, Parthasarathy S, et al. Mining frequent item sets in distributed and dynamic databases[C] // Proceedings of the International Conference on Data Mining (ICDM).2003:617-620

[23] Giannella C, Han J, Pei J, Yan X, Yu P S. Mining frequent patterns in data streams at multiple time granularities//Kargupta H,J oshi A, Sivakumar K, Yesha Y eds. Next Generation Data Mining.AAAI/MIT,2003:191-210

[24] C. Giannella, Granularities.J. Han, et al.Mining Frequent Patteerns in Data Mining at Multiple Time Granularities.Next Generations on Data Mining. pp. 191-212 , 2003.

[25] Cormode G, Garofalakis M. Sketching probabilistic data streams//Proceedings of the 2007 ACM SIGMOD International Conference on Data Management. Beijing, China,2007:281-292.

[26] Wang Shuang, Wang Guo-Ren. Frequent Items Query Algorithm for Uncertain Sensing Dat. Chinese Journal Of Computers, 2013,36(3):571-581.

[27] WANG Shui, ZHU Kongtao, WANG Le. Approximation algorithm for frequent itemsets mining on uncertain dataset[J]. Application Research of Computers,2013,31(11).

[28] CHEN An-Long, TANG Chang-Jie, YUAN Chang-An, ZHU Ming-Fang, DUAN Lei. A Compression Algorithm for Multi-Streams Based on Wavelets and Coincidence[J]. *Journal of Software,* 2007,18(2):177-184.

[29] Ding Chao,Yuan Chang.A Compression Algorithm for Multi-Streams Based on GEP. Journal of Computer Research and Devolopment,2008, 45(sup):191-195.'

[30] Zhu YY, Shasha D. StatStream: Statistical monitoring of thousands of data streams in real time. In: Bressan S, Chaudhri AB, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases. New York: Springer-Verlag, 2003. 358-369.

[31] Zhou B,Han Y,Pei J,Jiang B,Tao YF,Jia Y. Continuous privacy preserving publishing of data stream. In: Proc. the 12th International Conference on Extending Database Technology: Advances in Database Technology. EDBT 2009. 648-659.

[32] GUO Kun, ZHANG Qi-Shan. Fast Clustering-Based Anonymization Algorithm for Data Streams[J].Journal of Software,2013,24(8):1852-1867.

[33] LIU Zhong-bao1,2 and WANG Shi-tong. Privacy Preserving Learning Machine for Large Scale Data sets. Journal of University of Electronic Science and Technology of China, 2013,(42)3:272-276.

[34] Cheng Siyao,Li jianzhong. O(ε)-Approximation to Physical World by Sensor Networks[C]. Proc of IEEE INFORCOM'13.Piscataway,NJ: IEEE,2013:3184-3192.

[35] LIU Bo,GENG Yin-Rong. Mining Method for Data Quality Detection Rules[J].PR & AI,2012,25(5) :835-844.

[36] GAO Ming, JIN Che-Qing, WANG Xiao-Ling,T IAN Xiu-Xia,ZHOU Ao-Ying. A Survey on Management of Data Provenance. Chinese Journal Of Computers, 2010,33(3):373-388.

[37] Li Jianzhong, Liu Xianmin. An Important Aspect of Big Data: Data Usability. Journal of Computer Research and Devolopment,2013, 50(6):1147-1162.

Yi Wu, born in 1963, professor.His current research interests include large scale data processing, data mining, sensor network, etc.