REVIEW OF COMPUTER ENGINEERING STUDIES

ISSN: 2369-0755 (Print), 2369-0763 (Online)
Vol. 4, No. 1, March, 2017, pp. 31-35
DOI: 10.18280/rces.040106

A publication of IIETA
International Information and Engineering Technology Association
Advancing the World of Information and Engineering
http://www.iieta.org/Journals/RCES

# Data mining techniques for data streams mining

V. Sidda Reddy[1*], T.V. Rao[2], Govardhan A.[3]

[*] [1]Professor in CSE, Sai Tirumala NVR Engineering College
[2]Professor in CSE, K.L.University, Guntur
[3]Professor in CSE and Principal, JNTUHCE, Hyderabad

Email:siddareddy.v@gmail.com

## ABSTRACT

In resent years data stream mining plays an important role in real-time applications that generate gigantic of data needed intelligent data processing and on-line data analysis. The source of high-speed data streams may include video surveillance systems, stock markets, internet traffic, tweets etc. Traditional data mining techniques can't feasible for the data stream mining due to unique characteristics of data streams such as high dimensional, continuous flow, high-speed and fast changing. It necessitates building new data mining techniques or modifying existing ones to mine data streams. The main challenges include that the data stream mining needs to handle data distribution and concept drifting. This paper analyzes the challenges involved in designing data mining techniques for mining data streams besides evaluating various existing techniques and their preprocessing methods. The evaluation results reveal which methods are feasible and which methods are not feasible in real-time data streaming applications.

**Keywords:** Data Mining, OLAP, Concept Drifting, Data Streams, Data Stream Mining.

## 1. INTRODUCTION

With the advent of real time online applications, data repositories in World Wide Web are growing faster than before. As the data is exponentially increased the applications started using data mining techniques that analyze the huge amount of data in order to bring about trends or patterns which are required for business intelligence that leads to making well informed decisions. In real-time decision making, mining data streams become an important active research work and more widespread in several fields of computer science and engineering. Thus, data mining techniques effectively handle the challenges pertaining to storing and processing the huge amount of data [1]. Recently data mining techniques were proposed to process streaming data which is very challenging. Data streams can be conceived as sequences of training examples that arrive continuously at high-speed from a one of more sources [8], [9]. Data stream mining is a process of mining continuous incoming real time streaming data with acceptable performance [2]. Across wide range of real time applications such as network intrusion detection, stock market analysis, analysis of online click-streams, and web personalization data stream mining is essential [4]. There are many challenges in mining such streaming data in real time as developing techniques for the purpose is difficult [3]. Traditionally Online Analytical Processing (OLAP) systems involve in scanning data one or more times if needed for processing the data into information. This is not feasible for data stream mining [5] due to unique characteristics. Therefore, it is very important to modify the traditional data mining techniques in order to handle steaming data which comes from diverse sources over network. Processing streaming data in order to discover knowledge is given much importance recently as such data is made available through rich internet applications. There are two challenges in developing new techniques that could handle streaming data [6], [7], [9]. The first challenge is to design fast mining method for handling streaming data while the second challenge is detecting data distribution and changing concepts in a highly dynamic environment. This paper presents a comprehensive study of data stream mining challenges, mining techniques, their advantages and limitations.

The rest of the paper is organized as follows. Section II provides information about data stream mining and general data stream mining approach. Section III focuses on the data stream mining challenges. Section IV describes about data stream mining techniques. Section V evaluates the methods of mining streaming data while section VI concludes the paper.

## 2. MINING DATA STREAMS

Data stream is a high-speed continuous flow of data from diverse resources. The sources might include remote sensors, scientific processes, stock markets, online transactions, tweets, internet traffic, video surveillance systems etc. Generally these streams come in high-speed with a huge volume of data generated by real-time applications. Data streams have unique characteristics when compared with traditional datasets. They include potentially infinite, massive, continuous, temporarily ordered and fast changing. Storing such streams and then process is not viable as that needs a lot of storage and processing power. For this reason they are to be processed in real-time in order to discover knowledge from them instead of storing and processing like traditional data mining. Thus the processing of data streams throw challenges in terms of memory and processing power of systems. General procedure for processing streaming data is presented in Figure 1.

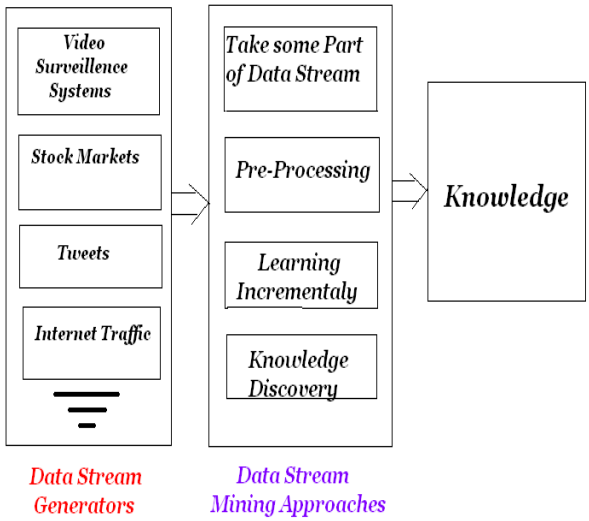


**Figure 1**. General data stream mining procedure

From the above Figure 1, the data streams are generated by various sources that include video surveillance systems, stock markets, sensor networks, internet traffic etc. The generated data streams are taken as input by stream data mining methods. The data stream mining procedure includes selecting a part of stream data, preprocessing, incremental learning and extraction of knowledge in a single pass. The result of data stream mining is the knowledge that can help in taking intelligent decisions. The data stream mining method analyzes the data which is high-dimensional, fast changing. Such methods should be able to work on streams and also large volumes of data. Memory related issues can be overcome using summarization techniques. Time and space efficient algorithms can be chosen from computation theory. Existing data mining techniques can also be used for data stream mining with some required changes [11].

## 3. CHALLENGES

A data stream refers to huge volume of data generated by rapidly in real-time applications. Traditional data mining techniques are challenged by two most important features of data streams are huge volume of data and concept drifting. When the volume of the underlying data is very large, high-speed and continuous flow it leads to number of computational and mining challenges listed below.

(1) Data contained in data streams is fast changing, high-speed and real-time.

(2) Multiple or random access of data streams is in expensive rather almost impossible.

(3) Huge volume of data to be processed in limited memory.

(4) Data stream mining system must process high-speed and gigantic data within time limitations.

(5) The data arriving in multidimensional and low level so techniques to mine such data needs to be very sophisticated.

(6) Data stream elements change rapidly overtime. Thus, data from the past may become irrelevant for the mining.

Out of all these challenges optimization of memory space is an important one as memory management is essential while mining streams. This is particularly an issue in many applications where the nodes are provided limited memory space. For instance, in wireless sensor networks where nodes are resources constrained, it is not possible to have algorithms that consume huge amount of memory and processing power. Therefore, it is essential to make use of summarization techniques in order to collect data from data streams [11]. Out of all the phases of data stream mining procedure as presented in Fig. 1, preprocessing is the phase that consumes more resources. Therefore, a technique which is lightweight is desired. Such technique gives good quality results. Integrating such technique with the stream mining approach is also a challenge. Data structures are to be used keeping the size of memory and the huge amount of streaming data in mind. The memory issues in data stream mining are explored in [12], [13]. To overcome the memory problem [13] introduced a runtime parameter to control the result as per the memory available. In [14] an algorithm is proposed which works with available limited resources consuming less memory and processing power. The research issues associated with identified challenges respectively are memory management, data preprocessing, compact data structure, resource aware and visualization of results. The next sub section provides various techniques that can address these research issues.

## 4. DATA STREAM MINING TECHNIQUES

In the recent past many data stream mining techniques came into existence. They mine frequent patterns in stream data to discover knowledge from huge amount of data for data analysis and decision making (business intelligence). Some data stream mining algorithms have preprocessing phase while some other algorithms do not have it. A survey of literature and analysis of methods used for knowledge discovery from continuous, high-speed data streams listed below.

**A. Discovering frequent patterns with preprocessing**

1. Clustering
   - STREAM and LOCAL SEARCH [24]
   - VFKM [25,26,27]
   - CluStream [28]
2. Classification
   - GEMM and FOCUS [15]
   - OLIN [16]
   - VFDT and CVFDT [17]
   - LW Class [18]
   - On-demand [20]
   - Ensemble-SCALLOP ANNCAD based [21]

**B. Discovering frequent patterns without preprocessing**

3. Clustering
   - D-Stream [29]
   - HP Stream [31]
   - AWSOM [30]
4. Classification
   - SCALLOP [23]
   - ANNCAD [22]
   - CDM [19]

**C. Frequency Counting and Time Series Analysis**

- Approximate Frequent Counts [32]
- FP Stream [33]
-

**D. Preprocessing Techniques for Data stream mining**

5. Storing some portions of summarized data.
   - Sampling
   - Load shedding
   - Sketching
6. Choosing a subset of incoming stream
   - Synopsis data
   - Aggregation
7. Without needing to store
   - Approximation Algorithms
   - Sliding windows
   - Algorithm Output Granularity

As can be seen from above clustering and classification techniques work with preprocessing and also without preprocessing. Frequency counting and time series analysis techniques are without preprocessing phase. Classification techniques include GEMM, FOCUS, OLIN, VFDT, CVFDT, LWClass, CDM, on demand stream classification, ensemble based classification, ANNCAD, and SCALLOP. The clustering techniques for data stream mining include Stream and Locale Search, VFKM, CluStream, D-Stream, AWSOM and HPStream. The data streaming techniques pertaining to frequency counting and time series analysis include FPStream and Approximate Frequent Counts. As seen in Fig. 2 (b) preprocessing techniques are of two types. They are techniques that store some portion of summarized data and the techniques that do not need to store. Sampling, load shedding, sketching are techniques that summarize whole dataset while synopsis data and aggregating techniques choose a subset of incoming stream. Preprocessing techniques that do not need storing data include approximation algorithms.

## 5. EVALUATION

This section evaluates all data stream mining algorithms of all types. With respect to clustering algorithms HPStream [31] is a projection based clustering algorithm. It exhibits high scalability, uses an incremental update, and is efficient for high dimensional data. However, it is highly complex. CluStream [28] follows a micro clustering approach in addition to the concepts of pyramidal time frame. It is time and space efficient, can detect concept drifts, and highly accurate in nature. However, it supports only offline clustering. Search and Locale Search [24] algorithms are K-Medians that make use of incremental learning. It is faster but exhibits low clustering quality and accuracy. VFKM [25], [26], [27] is a K-Means algorithm which is very faster and uses less memory storage. However, it needs to multiple passes to complete processing. D-Stream [29] is a density based clustering algorithm which exhibits high quality and efficiency. It can detect concept drifts in real time. However, it is highly complex in nature. AWSOM [30] is prediction based algorithm which can detect patterns efficiently, consumes less memory space, and completes clustering in a single pass.

With respect to classification techniques, LWClass [18] uses classification based on class weights. It exhibits high speed and consumes less memory. Its drawbacks are time consuming and can't be adapted to concept drifts. On-demand stream classification [10] uses micro-clustering approach. It exhibits dynamic updates, high speed, and consumes less memory space. VFDT and CVFDT algorithms produce decision trees. They are high speed and consume less memory space. However, they can't be adapted to concept drifts, and they are time consuming and costly learning. GEMM and FOCUS [15] are meant for generating decision trees and frequent item sets respectively. They follow incremental mining approach and also detect concept drifts. However, they are very time consuming and costly learning. OLIN [16] makes use of info-fuzzy techniques for building tree like result. It exhibits a dynamic update. Its drawbacks include low speed, causes memory problem besides time consuming and costly learning. CDM [19] produces decision trees. It is very suitable for distance measurement between events. It has drawback such as user-defined information complexity. Ensemble-based classification [21] uses combination of various classifiers. Its qualities include a single pass, dynamic update, ability to detect drifts, and highly accurate. However, it suffers from low speed, storage problem, and time consuming. ANNCAD [22] uses incremental classification and exhibits a dynamic update. Its drawbacks are same as that of Ensemble-based classification. SCALLOP [24] is suitable for scalable classification of numerical data streams. It exhibits dynamic update. It also suffers drawbacks same as that of ANNCAD.

With respect to techniques pertaining to frequency counting and time series analysis, approximate frequent counts [32] generates frequent item sets. It exhibits an incremental update, simplicity, consume less memory, and complete processing in a single pass. However, it generates approximate output with more error range possibility. FPStream [33] also generates frequent item sets.

## 6. CONCLUSION

The major objective of this article is to analyze and clarify the various data mining techniques and data stream mining challenges in real time applications. The data mining techniques that act on data streams are classified into clustering, classification, frequency counting and time series analysis. A survey on these techniques reveal the facts that from the classification techniques VFDT, CVFDT, CDM, on demand stream classification, ensemble-based classification, and ANNCAD are applicable and feasible for mining data streams while GEMM, FOCUS, OLIN, SCALLOP are not feasible; with respect to clustering techniques VFKM, CluStream, AWSOM, and HPStream are applicable while stream and locale search, and D-Stream are partially feasible; with respect to frequency counting and time series analysis both FPStream and Approximate Frequent Counts techniques are applicable to mining data streams. Finally, we are concluding that due to unique characteristics of data streams, still research will be carrying considerable.


## AKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments. We also thank the authors of all references for helping us setup the paper.


## REFERENCES


[1] Shearer C. (2000). The CRISP-DM model: the new blueprint for data mining, *Journal of Data Warehousing*, Vol. 5, No. 4, pp. 4-15.

[2] Gaber M.M., Zaslavsky A., Krishnaswamy S. (2005). Mining data stream: *a review. SIGMOD Record*. Vol. 34, No. 2, pp. 18-26.

[3] Kholghi M., Hassanzadeh H., Keyvanpour M. (2010). Classification and evaluation of data mining techniques for data stream requirements, *International Symposium on Computer, Communication, Control and Automation (3CA)*, pp. 474-478.

[4] Yang H., Fong S. (2010). An experimental comparison of decision trees in traditional data mining and data stream mining, *IEEE Xplore 2010 international Conference*, pp. 442-447.

[5] Han J., Kamber M. (2006). Data mining: concepts and techniques, second edition, *The Morgan Kaufmann Series in Data Management Systems: Elsevier*.

[6] Aggrawal C.C. (2007). Data Streams: Models and Algorithms: Springer.

[7] Chu F. (2005). Mining techniques for data streams and sequences, *Doctor of Philosophy Thesis*: University of California.

[8] Gama J., Rodrigues P.P. (2009). An overview on mining data streams, *Studies Computational Intelligence*. Springer Berlin/Heidelberg, pp. 29–45.

[9] Khan. (2000). Data stream mining: challenges and techniques, *Proceedings of 22th International Conference on Tools with Artificial Intelligen*ce.

[10] Muthukrishnan S. (2003). Data streams: algorithms and applications, *Proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms.*

[11] Golab L., Özsu M.T. (2003). Issues in data stream management, *ACM SIGMOD Record*, Vol. 32, No. 2, pp. 5-14.

[12] Chi Y., Wang H., Yu P.S. (2005). Loadstar: load shedding in data stream mining, *Proceedings of the 31st VLDB Conference*, Trondheim, Norway. pp. 1302-1305.

[13] Gaber M.M., Krishnaswamy S., Zaslavsky A. (2003). Adaptive mining techniques for data streams using algorithm output granularity, *The Australasian Data Mining Workshop.*

[14] Teng W., Chen M., Yu P.S. (2004). Resource-aware mining with variable granularities in data streams, *Proceedings of the 4th SIAM International Conference on Data Mining*, Lake Buena Vista, USA. pp. 527-53.

[15] Ganti V., Gehrke J., Ramakrishnan R. (2002). Data streams under block evolution, *ACM SIGKDD Explorations Newsletter*, Vol. 3, No. 2, pp. 1-10.

[16] Last M. (2002). Online classification of nonstationary data streams, *Intelligent Data Analysis*, Vol. 6, No. 2, pp. 129-147.

[17] Chi Y., Wang H., Yu P.S. (2005). Loadstar: load shedding in data stream mining, *Proceedings of the 31th VLDB Conference*, Trondheim, Norway. pp. 1302-1305.

[18] Gaber M.M., Krishnaswamy S., Zaslavsky A. (2006). On-board mining of data streams in sensor networks advanced, *Methods of Knowledge Discovery from Complex Data*, Springer, pp.307-335.

[19] Kwon Y., Lee W.Y., Balazinska M., Xu G. (2008). Clustering events on streams using complex context information, *Proceedings of the IEEE International Conference on Data Mining Workshop*. pp. 238-247.

[20] Wang H., Fan W., Yu P., Han J. (2003). Mining concept-drifting data streams using ensemble classifiers, *Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining*, Washington DC, USA.

[21] Law Y., Zaniolo C. (2005). An adaptive nearest neighbor classification algorithm for data streams, *Proceedings of the 9th European Conference on the Principals and Practice of Knowledge Discovery in Databases*, Verlag, Springer.

[22] Law Y., Zaniolo C. (2005). An adaptive nearest neighbor classification algorithm for data streams, *Proceedings of the 9th European Conference on the Principals and Practice of Knowledge Discovery in Databases*, Verlag, Springer.

[23] Ferrer-Troyano F.J., Aguilar-Ruiz J.S., Riquelme J.C. (2004). Discovering decision rules from numerical data streams, *Proceedings of the 2004 ACM symposium on Applied computing*, Nicosia, Cyprus. pp. 649-653.

[24] O'Callaghan L., Mishra N., Meyerson A., Guha S., Motwani R. (2002). Streaming-data algorithms for high-quality clustering, *Proceedings of IEEE International Conference on Data Engineering*.

[25] Domingos P., Hutten G. (2002). Mining high-speed data streams, *Proceedings of the Association for Computing Machinery 6th International Conference on Knowledge Discovery and Data Mining*.

[26] Domingos P., Hulten G. (2001). A general method for scaling up machine learning algorithms and its application to clustering, *Proceedings of the Eighteenth International Conference on Machine Learning*, Morgan Kaufmann. pp. 106-113.

[27] Hulten G., Spencer L., Domingos P. (2001). Mining time-changing data streams, *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, San Francisco, California. pp. 97-106.

[28] Aggarwal C., Han J., Wang J., Yu P.S. (2003). A framework for clustering evolving data streams, *Proceedings of the 29th VLDB Conference*, Berlin, Germany.

[29] Chen Y., Tu L. (2007). Density-based clustering for real-time stream data, *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Jose, California, USA, pp. 133-142.

[30] Papadimitriou S., Faloutsos C., Brockwell A. (2003). Adaptive, hands-off stream mining, *Proceedings of the 29th International Conference on Very Large Data Bases VLDB.*

[31] Aggarwal C.C., Han J., Wang J., Yu P.S. (2004). A framework for projected clustering of high dimensional data streams, *Proceedings of the 30th Conference VLDB*, Toronto, Canada.

[32] Manku G.S., Motwani R. (2002). Approximate frequency counts over data streams, *Proceedings of the 28th International Conference on VLDS*, Hong Kong, China.

[33] Giannella C., Han J., Pei J., Yan X., Yu P.S. (2003) Mining frequent patterns in data streams at multiple time granularities, *Data Mining: next generation challenges and future directions*, MIT/AAAI Press, pp. 191-212.