International Journal of Research in Advent Technology, Vol.6, No.1, January 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

Utilizing Streaming Algorithms for Mining Large Databases

A.Poongodi
Research Scholar, Bharathiar University,
Assistant Professor, KG College of Arts and Science, Coimbatore.

Abstract - With the advance in both hardware and software technologies, automated generation data and storage has become faster than ever. Streaming data processing is beneficial in most scenarios where new, dynamic data is generated on a continual basis. It applies to most of the industry segments and big data use cases. The 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 data stream mining techniques are new data mining techniques or modifying existing ones to mine high dimensional, high-speed and fast changing data of large databases. The main challenges include that the data stream mining needs to handle data distribution and concept drifting. This paper analyzes the uses of streaming algorithms for mining large databases and 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, Pharmacovigilance, ADRS, Concept Drifting, Data Stream Mining, Hoeffding Tree.

I. 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 technique 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 large databases using streaming algorithms 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 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 large databases 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 general data stream mining approach and FDA for sample large database. 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 with classification techniques. VI concludes the paper.

II. MINING DATA STREAMS

Data stream is a high-speed continuous flow of data from diverse resources. 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.

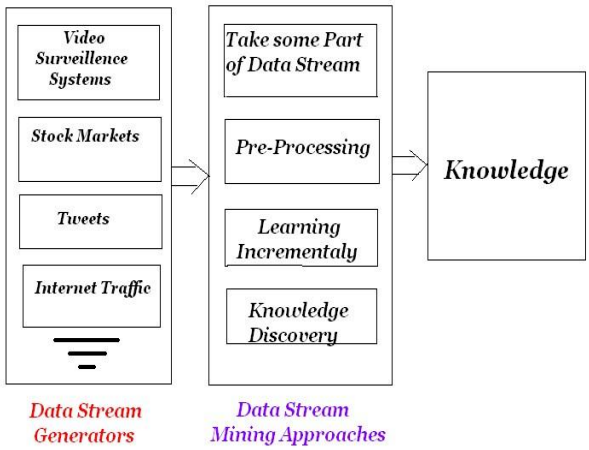


Figure 1. General data stream mining procedure

From the above Figure 1, The generated data 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].

A. FAERS

Medicines are designed to cure, treat, or prevent diseases; however, there are also risks in taking any medicine - particularly short term or long term adverse drug reactions (ADRs) can cause serious harm to patients. Pharmacovigilance (PhV) is the science that concerns with the detection, assessment, understanding and prevention of ADRs[36]. Food and Drug Administration (FDA) is the Federal public health agency that has regulatory responsibility for ensuring the safety of all marketed medical products, including pharmaceuticals (drugs and biologics). The availability of safe and effective pharmaceutical products depends on reporting of ADRs by all the parties involved i.e. the consumers or the patients, the healthcare providers and the drug manufacturers. The manufacturers have to compulsorily report ADRs.

All unsolicited reports from health care professionals or consumers, received by the FDA via either voluntary or mandatory route, are called spontaneous reports. Spontaneous reports are a part of a clinical observation that originates outside of a formal study. The individual spontaneous reports of ADRs, medication errors, and product quality problems, sent directly to the FDA through the Med Watch program or to the manufacturer and then indirectly from the manufacture to the FDA, combined with data from formal clinical studies and from the medical and scientific literature, comprise the primary data source upon which post marketing surveillance depends. The FDA continuously strives to implement newer surveillance techniques for detecting, reporting and evaluating adverse events. Another approach that might be used by FDA to detect adverse events is analyzing claim databases with a large sample size [36].

The safety profile of medicinal product may change in the post marketing environment. Regulatory agencies monitor product safety through a variety of mechanisms including signal detection of the adverse experience safety reports in databases and by requiring and monitoring risk management plans, periodic safety update reports and post authorization safety studies.

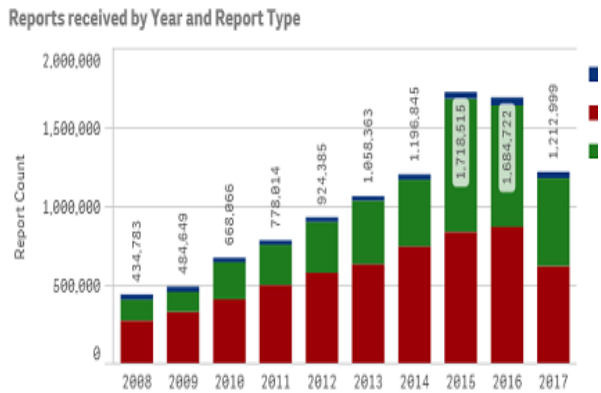
B. POSTMARKETTING REPORTING OF ADVERSE EXPERIENCES.

The main source of safety information for newly approved drugs is the routine post marketing surveillance of adverse experiences. Four major data- bases are used. The World Health Organization for International Drug Monitoring (Uppsala Monitoring Centre) has VigiBase™, which was started in 1968 and contains over 7 million individual case safety reports from 144 member

countries [WHO 2012]. The United States Food and Drug Administration (FDA) has the Adverse Event Reporting System (AERS), which was started in 1969 and contains over 4 million reports [FDA 2011a], and the Vaccine Adverse Event Reporting System (VAERS) [VAERS 2013], which was started in 1990 and contains over 400,000 reports. The European Medicines Agency (EMA) has Eudra Vigilance, which was started in 2001[37] FDA currently receives approximately two million adverse events, use error, and product complaint reports each year from consumers, health care professionals, manufacturers, and others. These reports are entered into various databases maintained by the FDA for subsequent analyses to identify potential safety issues and enhance the understanding of those issues, Since the 1990s [36].

	Total Reports	Expedited	Non-Expedited	Direct
2017	**1,212,999**	615,558	557,058	40,383
2016	**1,684,722**	864,389	769,534	50,879
2015	**1,718,515**	831,924	845,052	41,539
2016	**1,684,722**	864,389	769,534	50,879
2015	**1,718,515**	831,924	845,052	41,539
2014	**1,196,845**	739,581	423,150	34,114
2013	**1,058,363**	626,692	403,368	28,303

Table 1. Reports received by Year wise FDA dataset



Graph1. Reports received by Year wise FDA dataset

The table 1 and graph 1 shows FDA dataset in year wise [39] These databases represents high speed and fast changing data. The streaming algorithms are used to handle these databases with the feature of data distribution and concept drift.

III. 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, ADR reports in regulatory agencies where nodes are resources constrained, it is not possible to have algorithms that consume huge amount of memory 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.

IV. 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 Ensemble–Based Classification, Very Fast Decision Tree (VFDT) and CVFDT, On-Demand Classification, On-Line Information Network (OLIN), Lightweight Classification (LWClass), Scalable Classification Algorithm by Learning Decision Patterns (SCALLOP) and Adaptive Nearest Neighbor Classification for Data-Streams(ANNCAD). 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. The selection of techniques is based on the soundness of the techniques and how well the techniques address important research challenges.

V. EVALUATION OF STREAMING CLASSIFICATION ALGORITHMS

Classification is the process of predicting the class label of an unknown data instance based on model constructed from leaning on training instances [1]. Various classification algorithms for streaming data are available. Some of the available streaming data classification algorithms along with their key features are chronologically listed in table[38]. Distributed data mining is mining of data from different sources.[40].

Distributed also consists large and high volume of data processing, and implemented with big data concepts.

Streaming Classification Algorithm	Year	Key Features
Ensemble Algorithms [26][27]	2001	Provides robustness and handles concept drift but needs to be carefully used for high speed data streams.
VFDT[35][36]	2000	Require lesser memory and does prediction at any moment of time during training. It uses Hoeffding bound to assess the number of Minimum instances required to grow the decision tree.
CVFDT[25]-[26]	2001	Advancement of VFDT that enables Concept Adaptation.
On-Demand Classifier[28]	2004	Based on micro clustering, dynamically adapts and/or selects sliding window size for better performance and concept Adaptation.
OLIN[29]	2002	Requires lesser memory and uses the info-fuzzy network(IFN) for concept adaptation
Weighted Classifier Ensemble[28]	2003	Deals well with concept drifts by using ensemble of weighted classifiers on chunks of data instances from data streams rather revising the model(which is time taking process)
ANNCAD[29]	2005	An incremental algorithm that adaptively searches for nearest neighbors by multi resolution representation of data. It facilitates low model update cost.
Random Forest Based Classification Algorithm[30]	2011	Handles evolving data streams even with intermittent labeled data instances arrival in one pass.
Vertical Hoeffding Tree(VHT)[31]	2013	A variation of VFDT that performs distributed parallel computation by vertically partitioning (attribute based) data sets.
Similarity-based data stream Classifier (SimC)[32]	2014	Uses new insertion /removal approach for quickly capturing and representing changes in data to improve performance. Also, Incorporates new class labels and discards obsolete class labels during the execution.
Online Stream Classifier with incremental semi-supervised learning[33]	2015	Utilizes the selective self-training based semi supervised learning approach to achieve at the par classification accuracy even with availability of only 1% labeled data.
Distance-Based Ensemble Online Classifier with kernel Clustering.[34]	2015	An ensemble of classifier is constructed on the basis of portfolio of distance measures.

some approach uses ensemble techniques[26,27,30] to adapt to concept drifts. Obviously, ensemble based approaches are robust but require to maintain several models at a time, hence require more memory to retain those models. In terms of processing time, there is a tradeoff between the ensemble approach and model update frequency of other approaches.[38]

Performance Evaluation Measures

One of the challenges of stream data mining tasks is how to evaluate the performance of the mining tasks. Performance measure of stream data mining classifications are listed in table[38].

Task	Evaluation Measure	Major Purpose	Value Significance
Classification	Kappa statistics[6]	Assess performance In imbalance data stream case	Higher value means better performance
	Temporal Negative –Kappa statistics[6]	Assess performance in case of temporal dependent data streams	Negative Values means worse performance.

Table 3. Performance Evaluators of streaming using classification[38]

VI. CONCLUSION

The major objective of this article is to analyze and clarify the various data stream mining techniques and 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. An evaluation 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; we are concluding that due to unique characteristics of data streams, Many works were done on mining large databases using classification[1], catlett[2] presented a decision-tree based learner that can handle thousands of datasets. Hoeffding trees [7] can access data sequentially and just required one scan. In HT new samples can be added any time and is incremental in nature. Hoeffding Tree and its extensions helps high speed learning of large volume samples with less memory utilization.

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: conceptsand 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 Intelligence.

[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:loadshedding 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] Gama, J., Fernandes, R., Rocha, R: Decision Trees for mining Data Streams. Intelligent Data Analysis. 10,23

[26] Stree, W. Nick and YongSeog Kim. "A streaming ensemble algorithm (SEA) for large-scale classification. "In Proceedings of seventh ACM SIGKDD international conference on knowledge discovery and data mining, pp 377-382. ACM, 2001.

[27] Bifet, A Holmes, G Kirkby, R., and P Fahringer, B.2011. In MOA: DATA STREAM MINING –A Practical approach, The University of Waikato, 107-139.

[28] Wang, Haixum, Wei Fan, Philip S. Yu, and Jiawei Han. "Mining concept-drifting data streams using ensemble classifiers". In Proceedings of ninth ACM SIGKDD international conference on knowledge discovery and data mining, pp . 226-235, ACM, 2003.

[29] La, Yan-Nei, and Carlo Zaniolo. "An adaptive nearest neighbor classification algorithm for streams "In knowledge Discovery in dartabases: PKDD 2005, pp 108-120. Speinger Berlin Heidelberg. 2005

[30] Abdulsalam, H: Skillicorn, D.B :, Martin , P., "Classification Using Streaming Random Forests, "Knowledge and Data Engineering, IEEE Transactions on, Vol 123, no.1, pp. 22-36, Jan 2011.

[31] Prasad, B.R., and Aagarwal , S. " Critical Parameter analysis for Vertical Hoeffding Tree for optimized performance using SAMOA, " Int J. Mach. Learning &

[32] Loo H. and Marsano M. N. Online data stream classification with incremental semi-supervised learning In Proceedings of the second ACM IKDD Conference on Data Sciences, ACM, 2015.

[33] Jedrejowicz J., amd Piotr J Distance-Based Ensemble Online Classifier with Kernel Clustering , Intelligent Decision Technologies, Springer International Publishing, 279-289, 2015.

[34] J.Han, M,Kamber and J. Pei, Data Mining Concepts and Techniques, 3rd Edition, Morgan Kaufmann,(2011).

[35] J.A.Danel, Aurora: A New Model and Architecture for Data Stream Management, The VLDB Journal- The International Journal on Very Large Data Bases, Vol. 12, no.2. 2003, pp 120-139.

[36] Mei Liu, Michael E.Matheny, ACM SIGKDD Explorations Newsletter, Volume 14, Issue 1, June 2012, Pages 35-42

[37] Robert G Sharrar and Gretchen S Dieck, "Monitoring product safety in the postmarketing environment" Therapeutic Advances in Drug Safety, (2013) 4(5) 211–219.

[38] Bakshi Rohit Prasad and Sonali Agarwal, Stream Data Mining:Platforms, Algorithms, Performance Evaluaters and Research Trends, International Journal of Database Theory and Application, Vol.9.No.9(2016), PP.201-218

[39] https://www.fda.gov

[40] Ajitha.P, Dr.E.chandra, "A Survey on Outliers Detection in Distributed Data mining for Big Data", Journal of Basic and Applied Scienific Research 2015, pp:31-38.