An Analytical Framework for Data Stream Mining Methodologies Considering Issues and Necessity

Shabina Sayed[1] Shoeb Ahmed Ansari[2] Rakesh Poonia[3]
[1,2]Department of Computer Science [3]Department of Computer Applications
[1]Jodhpur National University, Jodhpur, Rajasthan, India [2]Shri. J. Jagadish Prasad Jabnormal Tibrewala University, Jhunjhunu, Rajasthan, India [3]Govt. Engineering College, Bikaner, Rajasthan, India

Abstract— In today's information society, computers are used to gather and share data anytime and anywhere. This concerns applications such as social networking, banking, telecommunication, healthcare, research, and entertainment, among others. As a result, a huge amount of data related to all human activity is gathered for storage and processing purposes. These data sets may contain interesting and useful knowledge represented by hidden patterns, but due to the volume of the gathered data, it is impossible to manually extract that knowledge. That is why data mining and knowledge discovery methods have been proposed to automatically acquire interesting, non-trivial, previously unknown and ultimately understandable patterns from very large data sets. A new class of emerging applications generates data at very high rates in the form of transient data streams. Due to their speed and size, it is impossible to store them permanently. Applications of data stream analysis can vary from critical scientific and astronomical applications to important business and financial ones. Algorithms, systems, and frameworks that address streaming challenges have been developed over the past 10 years. In this paper, we review the theoretical foundations of data stream analysis, mining data stream systems. Finally, we outline and discuss research problems in stream mining field of study.

Key words: Data Stream, Knowledge, Discovery, Concept Drift

I. INTRODUCTION

Data stream application domains include network mon-toring, security, telecommunication data management web applications, and sensor networks. The introduction of this new class of applications has opened an interesting line of research problems including novel approaches to knowledge discovery called data stream mining [1]. Current research in data mining is mainly devoted to static environments, where patterns hidden in the data are fixed and each data tuple can be accessed more than once. The most popular data mining task is classification, defined as generalizing a known structure to apply it to new data [2]. Traditional classification techniques give great results in static environments, however, they fail to successfully process data streams because of two factors: their overwhelming volume and their distinctive feature - concept drift. Concept drift is a term used to describe changes in the learned structure that occur over time.

These changes mainly involve substitutions of one classification task with another but also include steady trends and minor fluctuations of the underlying probability distributions [4, 5]. For most traditional classifiers the occurrence of concept drift leads to a drastic drop in classification accuracy. That is why recently; new classification algorithms dedicated to data streams have been proposed. These research issues should be addressed in order to realize robust systems that are capable of fulfilling the needs of data stream mining applications. In this review paper, we present the state of the art in this growing vital field. The paper is organized as follows. Section 2 presents the theoretical background of data stream analysis. In sections 3 and 4 mining data stream techniques and systems are reviewed respectively. Open and addressed research issues in this growing field are discussed in the section 5. Finally, section 6 summarizes this review paper. Sections 7 enlist the references.

II. THEORETICAL FOUNDATIONS

A large number of applications generate data streams [7, 8] for e.g. Telecommunication (call records), System management (network events), Surveillance (sensor network, au- audio/video), Financial market (stock exchange), Day to day business (credit card, ATM transactions, Etc.).The main Task of these applications are:

Real-time query answering, statistics maintenance, and pattern discovery on data streams.

- Structured low-volume: Wire services, Phone call connection reports, Phone and organization directory, Badge access tracking, Customer Lists, Account History, Personal address book, Personal records, Payroll databases, Expense reports, Logs of tunnel activities, Purchasing logs, Supplier relationships, Work logs/project history, Temperature in machine room for IS reliability, Active monitoring remote copy to disaster site, Disaster site monitoring, Credit reports, Biometric access control.
- Structured high-volume: Stock Exchange Transactions, Web pages for news/weather, audit records, CRM Databases [3], Web access logs and network logs, Company Website, Mutual fund valuation and transactions, "Financial product" sales, Credit/Debit card transactions, RFID Tracking logs, analyze signatures.
- Unstructured low-volume: Email, Trading floor sound, Chat, Instant Messages, Reports Internal, Printed reports, Handphone logs, Courier records, Call Center Data & Logs, Pager, External proprietary reports and data, Customer enquiries, Customer complaints, Public records, Patents, FAX, Scanned checks, RF Monitoring (look for rogue hubs), Print stream monitoring, Calendars.

Technology breakthroughs are needed to manage and analyze continuous streams for knowledge extraction to adapt system management rapidly based on changes of the data and the environment. There are various application need to make numerous real-time decisions about priorities of what inputs to examine, what analyses to execute, etc. Some application needs to operate over physically distributed sites. Hence the stream mining system should be highly secure and

provide support protection of private information and also it should be scalable in many dimensions.

Data stream mining application require Computation [10] at different granularity i.e. low, medium and high. Network link monitoring system require less computation. For such system, we can use algorithmic techniques such as Sampling, sketch maintenance etc. If the application requires query processing, data mining, knowledge discovery, we can process the data stream using Classification, clustering and load shedding. But there are other applications which consists of Evolving concepts (concept drift) which cannot be processed using conventional methodology. They require high [8, 9] granularity of computation e.g. stream management system. So for such application, we should use semantics like ontology, description logics etc.

III. Existing Data Stream Mining Techniques

Research problems and challenges that have been arisen in mining data streams have its solutions using well established statistical and computational approaches. We can categorize these solutions to data-based and task-based ones. In data-based solutions, the idea is to examine only a subset of the whole dataset or to transform the data vertically or horizontally to an approximately smaller size data representation. At the other hand, in task-based solutions, techniques from computational theory have been adopted to achieve time and space efficient solutions. In this section, we review these theoretical foundations.

A. Data based technique

Data-based techniques refer to summarizing the whole dataset or choosing a subset of the incoming stream to be analyzed. Sampling, load shedding, and sketching techniques represent the former one. Synopsis data structures and aggregation represent the later one. Here is an outline of the basics of these techniques with pointers to its applications in the context of data stream analysis.

1) Sampling

Sampling refers to the process of probabilistic choice of a data item to be processed or not. Sampling is an old statistical technique that has been used for a long time. Boundaries of the error rate of the computation are given as a function of the sampling rate. Very Fast Machine Learning techniques [5] have used Hoeffding bound to measure the sample size according to some derived loss functions. The problem with using sampling in the context of data stream analysis is the unknown dataset size. Thus the treatment of data stream should follow a special analysis to find the error bounds. Another problem with sampling is that it would be important to check for anomalies for surveillance analysis as an application in mining data streams. Sampling may not be the right choice for such an application. Sampling also does not address the problem of fluctuating data rates. It would be worth investigating the relationship among the three parameters: data rate, sampling rate and error bounds.

2) Load Shedding

Load shedding refers to the process of dropping a sequence of data streams. Load shedding has been used successfully in querying data streams. It has the same problems of sampling. Load shedding is difficult to be used with mining algorithms because it drops chunks of data streams that could be used in the structuring of the generated models or it might represent a pattern of interest in time series analysis.

3) Sketching

Sketching is the process of randomly project a subset of the features. It is the process of vertically sample the incoming stream. Sketching has been applied in comparing different data streams and in aggregate queries. The major drawback of sketching is that of accuracy. It is hard to use it in the context of data stream mining. Principal Component Analysis (PCA) would be a better solution that has been applied in streaming applications.

4) Synopsis Data Structures

Creating synopsis of data refers to the process of applying summarization techniques that are capable of summarizing the incoming stream for further analysis. Wavelet analysis, histograms, quantiles and frequency moments have been proposed as synopsis data structures. Since synopsis of data does not represent all the characteristics of the dataset, approximate answers are produced when using such data structures.

5) Aggregation

Aggregation is the process of computing statistical measures such as means and variance that summarize the incoming stream. Using this aggregated data could be used by the mining algorithm. The problem with aggregation is that it does not perform well with highly fluctuating data distributions.

B. Task-based Techniques

Task-based techniques are those methods that modify existing techniques or invent new ones in order to address the computational challenges of data stream processing. Approximation algorithms, sliding window, and algorithm output granularity represent this category.

In the following subsections, we examine each of these techniques and its application in the context of data stream analysis.

1) Approximation Algorithms

Approximation algorithms have their roots in algorithm design. It is concerned with design algorithms for computationally hard problems. These algorithms can result in an approximate solution with error bounds. The idea is that mining algorithms are considered hard computational problems given its features of continueality [11,12] and speed and the generating environment that is featured by being resource constrained. Approximation algorithms have attracted researchers as a direct solution to data stream mining problems. However, the problem of data rates with regard to the available resources could not be solved using approximation algorithms. Other tools should be used along with these algorithms in order to adapt to the available resources.

2) Sliding Window

The inspiration behind the sliding window is that the user is more concerned with the analysis of most recent data streams. Thus the detailed analysis is done over the most recent data items and summarized versions of the old ones. This idea has been adopted in many techniques in the undergoing comprehensive data stream mining system.

3) Algorithm Output Granularity

The algorithm output granularity (AOG) [6] introduces the first resource-aware data analysis approach that can cope with

fluctuating very high data rates according to the available memory and the processing speed represented in time constraints. The AOG performs the local data analysis on a resource-constrained device that generates or receive streams of information.

4) Classification of stream challenges

There are different challenges [13] in data stream mining that cause many research issues in this field. Regarding data stream requirements, developing stream mining algorithms is needed more studying than traditional mining methods.

IV. Research issue of DSM system

In recent years with the emerging modern equipment in which smart devices play an important role, the previously defined classification task becomes obsolete. The size of the training data is potentially unbounded and continuously incoming, as it was presented in the previous section. It is clear that it becomes infeasible to use all the information and the distribution is likely to be non-stationary. With these characteristics in mind, Domingos and Hulten (2003) [14, 15, 17] formulated the following properties that a classifier for data streams should possess in order to effectively work with streaming data.

- It must require small constant time per record, otherwise, it will inevitably fall behind the data, sooner or later.
- It must use only a fixed amount of main memory, irrespective of the total number of records it has seen.
- It must be able to build a model using at most one scan of the data, since it may not revisit old records, and the data may not even be available in secondary storage at future point in time.
- It must make a usable model available at any point in time, not only when it is done processing the data, since it may never be done the processing.
- It should ideally produce a model that is equivalent (or nearly identical) to the one that would be obtained by corresponding ordinary database mining algorithm, operating without the above constraints.
- When the data-generating phenomenon is changing over time (e.g., when the concept drift is present), the model at any time should be up-to-date, but also include all the information from the past that has not become outdated. Few of the algorithms applied in the traditional machine learning are inherently incremental, satisfying most of the above stated characteristics and therefore directly applicable to data stream mining. The best example is a Naive Bayes classifier [18] which requires storing only basic statistics. Other techniques need to be altered in order to meet the demands of data streams such as decision trees.

Data stream mining is a stimulating field of study that has raised challenges and research issues to be addressed by the database and data mining communities. Following is a discussion of both addressed and open research issues. The following is a brief discussion of previously addressed issues is as follows:

1) Unbounded memory requirements due to the continuous flow of data streams.
2) Required result curacy.
3) Transferring data mining results over a wireless network with a limited bandwidth.
4) Modelling changes of mining results over time.
5) Developing algorithms for mining results changes.
6) Visualization of data mining results on small screens of mobile devices.
7) Interactive mining environment to satisfy user requirements.
8) The integration between data stream management systems and the ubiquitous data stream.
9) Mining approaches.
10) The needs of real world applications.
11) Data stream pre-processing.
12) Model overfitting.
13) Data stream mining technology.
14) The formalization of real-time accuracy evaluation.

V. The proposed analytical framework

This research ends in an analytical framework which is shown in Table 2. This framework tries to show the efficiency of data mining applications in developing the novel data stream mining algorithms. These algorithms are classified base on the data mining tasks. We described the details of these algorithms based on preprocessing steps and the following steps. In addition, this framework can direct future works in this field.

Some of the most important results that have been reached during this research are:

- Mining data streams have raised a number of research challenges for the data mining community. Due to the resource and time constraints, many summarization and approximation techniques have been adopted from the fields of statistics and computational theory.
- There are many open issues that need to be addressed. The development of systems that will fully address these issues is crucial for accelerating the science discovery in the fields of physics and astronomy, as well as in business and financial applications.

VI. Conclusion

In this paper, we reviewed and analyzed data mining applications for solving data stream mining challenges. At first, we presented a comprehensive classification for data stream mining algorithms based on data mining applications. In this classification, we separate algorithms with preprocessing from those without preprocessing. In addition, we classify preprocessing techniques in a distinct classification. In the following, the layered architecture of the classification represents almost all of the challenges that are mentioned in various researchers. Then we discussed the application of data mining techniques for addressing the challenges of data stream mining, and then we presented an analytical framework regarding these applications. Results are shown that it is necessary to adopt many summarization and approximation techniques from the fields of statistics and computational theory, besides crucial changes that are needed in common data mining techniques. In spite of the researches that have been done on data mining application in data stream mining so far, there are still wide areas for further researches.

References

[1] Max Bramer. Principles of Data Mining. Springer, 2007.

[2] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic S- myth. From data mining to knowledge discovery: An overview. In Advances in Knowledge Discovery and Data mining, pages 134. American Association for Artificial Intelligence, 1996.
[3] John F. Gantz, David Reinsel, Christopeher Chute, Wolfgang Schlichting, Stephen Minton, Anna Toncheva, and Alex Man- frediz. The expanding digital universe: An updated forecast of worldwide information growth through 2011. Technical report, IDC Information and Data, 2008.
[4] Elena Ikonomovska, Suzana Loskovska, and Dejan Gjorgjevik.A survey of stream data mining, 2005.
[5] Ludmila I. Kuncheva. Classifier ensembles for changing environments. In Fabio Roli, JosefKittler, and Terry Windeatt, edi- tors, Multiple Classifier Systems, volume 3077 of LectureNotes in Computer Science, pages 115. Springer, 2004.
[6] Albert Bifet and Richard Kirkby. Data stream mining: a practical approach. Technical report, the University of Waikato, August 2009.
[7] Arvind Arasu, Gurmeet Singh Manku. Approximate Counts and Quantiles over Sliding Windows. In the ACM Symposium on Principles of Database Systems (PODS) 2004.
[8] Brian Babcck, Chris Olston. Distributed Top-k Monitoring. In the ACM International Conference on Management of Data (SIGMOD) 2003.
[9] Brian Babcock, Mayur Datar, Rajeev Motwani, LiadanO O'Callaghan.Maintaining Variance and k-Medians over Data Stream Windows. In the ACM Symposium on Principles of Database Systems (PODS) 2003.
[10] Yunyue Zhu, Dennis Shasha. StatStream: Statistical Moni- toring of Thousands of Data Streams in Real Time. In the International Conference on Very Large Data Bases (VLDB) 2002.
[11] Golab, L. and zsu, M.T. (2003): Issues in Data Stream Management. ACM SIGMOD Record, 32 (2): pp. 5 14.
[12] Gaber, M.M., Krishnaswamy, S., and Zaslavsky, A., (2003): Adaptive Mining Techniques for Data Streams Using Algorith- m Output Granularity. In Proceedings of the Australasian Data Mining Workshop.
[13] Muthukrishnan, S., (2003): Data streams: algorithms and applications. In Proceedings of the fourteenth annual ACM- SIAM symposium on discrete algorithms
[14] Gaber, M.M., Zaslavsky, A., and Krishnaswamy, S., (2004): Resource- Aware Knowledge Discovery in Data Streams. In Proceedings of First International Workshop on Knowledge Discovery in Data Streams. Pisa, Italy.
[15] Teng, W., Chen, M., and Yu, P.S., (2004): Resource-Aware Mining with Variable Granularities in Data Streams. In Pro- ceedings of the 4th SIAM International Conference on Data Mining. Lake Buena Vista, USA, pp. 527-531.
[16] Kargupta, H., Park, B., and Sarkar, K. (2002): MobiMine: Monitoring the Stock Market from a PDA. ACM SIGKDD Explorations Newsletter, 3 (2): pp. 37-46
[17] Haixun Wang, Wei Fan, Philip S. Yu, and Jiawei Han. Mining concept-drifting data streams using ensemble classifiers. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, KDD, pages 226235. ACM, 2003.
[18] Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis, Volume 95. Wiley

