Knowledge Discovery from Static Datasets to Evolving Data Streams and Challenges

V.Sidda Reddy
Professor
Sagar Institute of Technology
Hyderabad, A.P, India

M.Narendra
Assistant Professor
CMR Engineering College
Hyderabad, A.P, India

K.Helini
Assistant Professor
SCET
Hyderabad, A.P, India

ABSTRACT

Mining data streams has recently become an important active research work and more widespread in several fields of computer science and engineering. It has proven successfully in many domains such as wireless sensor networks, ATM transactions, search engines, web analysis and weather monitoring. Data steams can be considered a subfield of machine learning, data mining and knowledge discovery. Data Mining is a step in the process of knowledge discovery from large amount of data. Traditional data mining techniques can not be easily applied to the data stream mining due to unique characteristics of data streams. In this research work, we will survey the main techniques and applications of data mining and data stream mining. We then study, the computational and miming challenges in particular, on-line mining of continuous, high-speed massive data streams.

Keywords

Knowledge Discovery, Data Mining, Data Streams, Data Stream Mining

1. INTRODUCTION

Now a day, Knowledge Discovery (KD) and Data Mining (DM) is becoming most challenging research work for database and data miming community. Knowledge Discovery and Data Mining is an interdisciplinary area for extracting useful knowledge from large amount of data. Traditionally, Knowledge workers and data mining researches have primarily concentrated on data querying and data analysis of static datasets (data warehouse) for decision-support and business-intelligence activities [1]. In data mining applications, data arrives from static datasets then data mining techniques are applied for knowledge discovery. Static databases are rich in information that can be used in the data mining. Data mining has traditionally have been performed over static databases, where data mining methods can afford to read the data randomly one or more times if needed. In recent years, advances in both hardware and software technologies coupled with high-speed data generation has led to data streams and data stream mining. Over the past few years, a considerable number of studies have been made on data stream mining. We can categorize the literature into two big research directions first one is Data Stream Management Mining (DSMS) techniques and second one is Data Stream Mining (DSM) extended from data mining techniques.

We organized the rest of the paper as follows: Section II data mining techniques, applications and challenges. Section III describes the data stream characteristics and mining requirements, data stream mining models and advantages. Section IV data stream mining challenges. Section V conclusion and feature work.

2. DATA MINING TECHNIQUES

Knowledge Discovery in Databases (KDD) is the algorithmic and statistical data analysis to extract knowledge and useful patterns from gigantic data. Data Mining is a step in the process of knowledge discovery in databases [2]. Different data mining methods have been proposed for extracting knowledge such as classification, prediction, association, and clustering.

2.1 Classification

Classification predicts categorical class labels of unknown data objects. It involves two steps, in the first step a model is constructed describing a predetermined set of data classes by using training data. In the second step, the constructed model is used to classify unknown data objects. Classification methods include decision tree induction, Bayesian classification, Rule-based classification, Backpropagation and Associative classification. Classification methods mainly used applications such as target marketing, credit approval, medical diagnosis and fraud detection.

Major issues in classification as follows:

- Data cleaning
- Relevance analysis
- Data transformation

2.2 Prediction

Prediction is data analysis technique, that models continuous-valued functions i.e., predict unknown or missing values. It is similar to classification, classification predict categorical class labels of unknown data objects where as prediction models continuous-valued functions. Prediction can be modeled by statistical technique regression. Regression analysis describes the relationship between two (or more) variables. Regression analysis used to predict a dependent variable, based on the value of at least one independent variable. Regression methods include linear and multiple regressions, non-linear regression, poison regression, log-linear model and regression trees. Prediction commonly used applications such as time-series data analysis, weather forecast and sales analysis. A common issue in predication includes:

- Accuracy
- Robustness
- Scalability
- Interpretability
- Training time

2.3 Association

Association rule is an important and well researched method for finding frequent patterns or associations among set of objects in large database. It was first introduced by Agrawal in 1993 for discovering regularities between products in transaction data of super market [3-4]. Association rule mining widely used applications such as market basket analysis, cross-marketing, DNA sequence analysis, classification and clustering. Major challenges of association rule mining includes

- Normal association rule mining does not have any target item(s)
- Mining data on the hard disk (not in main memory)
- Produce a huge number of rules (thousand, tens of thousands, millions)

2.4 Clustering

Clustering is an important unsupervised learning technique where as classification is supervised learning technique. Clustering is a process of organizing objects into groups called clusters or classes; objects with in the same cluster are similar and dissimilar to the objects belonging to other cluster. Clustering methods includes partitioning algorithm, hierarchy algorithm, density-based algorithm, grid-based method and model-based method. Cluster analysis has wide applications including web analysis, pattern reorganization, medical diagnostic, and text data analysis. Challenges in cluster analysis as follows

- Scalability
- Ability to deal with different type of attributes
- Domain knowledge to determine input parameters
- Able to deal with noise and outliers.

Data mining techniques are suitable for simple and structured databases such as transaction databases, relational databases, object oriented databases and data warehouses [5]. Traditional data mining techniques are infeasible for mining data streams due to unique characteristics such as continues flow, high-speed, infinite length, ordered sequence and fast changing. The challenges that the traditional data mining techniques associated for mining data streams as follows:

- Traditional data mining systems, data store in static database for later analysis, where as in the applications of data streams, the fact that massive amount data arrive at high-speed makes data mining methods are infeasible.
- The data mining system is feasible, when input data is large and arrives from static datasets or data warehouse for the reason that data in static dataset is stable and accessed randomly. Still data mining system does not fusible to mine dynamic datasets or real-time datasets called data stream for the reason that data streams are continuous flow of data generated at high-speed in real-time applications.

3. DATA STREAM MINING

In recent years, which the advances in processing and communication techniques for the collection of data continuously with high-speed, there are many emerging applications to deal with the rapid growth of data sets which are referred to data streams

3.1 Data streams

A *data stream* is a continuous, high-speed and ordered sequence of data items. High-speed refers to phenomenon that the data rate is high relatively to the computational power [6]. Data streams have unique characteristics as follows:

- *Continuous flow*: Data streams are a continuous flow that the data stream elements flow one after another continuously, for example network traffic, sensor data and call center records.
- *Infinite length*: Data streams are huge volumes of data, possibly infinite length, for example web logs and web page click streams.
- *Fast changing*: Data stream objects values change frequently for example stock exchange.
- *High-speed*: Data streams are generated rapidly and flow at high-speed for example data, surveillance and video streams
- *Ordered sequence*: Data stream is an ordered sequence of data items that arrives in timely order for example, ATM transactions and credit card transaction flows.

3.2 Mining Requirements

From data stream characteristics we come to know that evolving data streams must meet several requirements for knowledge processing as follows [7]:

- process an pattern at a time, and examine it only once
- use a limited amount of memory
- working in limited amount of time
- be ready to predict at any time
- knowledge discovery from evolving data streams or dynamic datasets

3.3 Data stream models

The fast few 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. Mining data streams are real-time process of extracting interesting patterns from continuous and high-speed data items. Traditional data mining models are infeasible for mining data streams due their unique features. The first data stream model was proposed by Henzinger, Raghava, and Rajgopalan [8].The basic data stream processing model as follows:

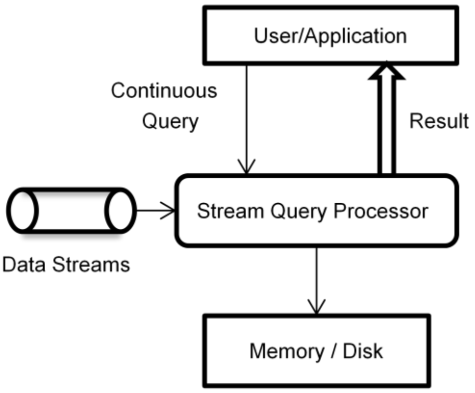

Fig 1: Basic Data Stream Processing Model

The research of data stream mining in general can be classified into three categories according to the stream processing model as follows [9-11]:

3.3.1 Landmark Window Model

In a landmark window model, all stream objects that have been observed up to the current time are contained in the window. All objects contribute with equal weight to the result, regardless of their time stamp. A landmark window increases over time as a new stream object arrives. The benefit of landmark window model is stream objects are simply accumulated over time but no need to remove objects form the window.

3.3.2 Damped Window Model

In a damped window model, the influence of stream objects on the stream mining result fades over time according to a user-defined fading function, which is monotonically decreasing. Objects in the damped window contribute to the mining result with decreasing weight as their age. Older data samples contribute less weight towards the pattern emerging in recent data. The benefit of damped window model is that the model deems different weights for new and old data samples.

3.3.3 Sliding Window Model

In sliding window model, the range of mining is confined to the stream objects contained in a window. The window always covers a certain number of most recent stream objects and the mining task focus on there objects at any point. The benefit of a sliding window is that because only recent object are considered when computing a data mining result, changes in the properties of the objects will be reflected in the mining results much faster than landmark window model. The main issue in sliding window model is that as a new basic window arrives oldest window must be removed from the memory as well their contribution also discarded from the memory, this will affect mining result.

These three data stream process models have been using in recent researches on data stream mining. Choosing which kind of model to utilize mainly depends on the specific application and data stream nature.

4. DATA STREAM MINING CHALLENGES

A data stream refers to huge volume of data generated by rapidly in real-time applications. When the underlying data is very large, continuous flow and high-speed, it leads to number of computational and mining challenges listed [12-15]:

- The key property of data stream is that continuous flow i.e., data flow continuously in sequence order at high-speed. Stream data can not access random. Every data item must be accepted as it arrives in the order that it arrives. Once inspected or ignored or discarded with no ability to retrieve it again, this will affect efficiency and result of stream mining.
- Hence continuous data flow is the one of the factor that challenges mining data stream.
- Data stream applications such as web logs and web page click streams produce huge volumes possibly infinite length data streams. So data stream mining allows processing of the data is many times larger than available working memory. Due to memory limitations mining infinite length data streams are most challenging task.
- Data streams are high-speed and generated rapidly by real-time applications such as stock market, bio-informatics and video streams. So data stream mining system must process high-speed and gigantic data with in time limitations. Due to limited resource and strict time constraints, obtaining the exact results will not be possible.
- Data streams are flow in sequence order due this property multiple and random access of such data streams is expensive rather almost impossible.
- Data stream such as spatial data streams and web click streams arrive in multiple-dimensional or low level. Where as stream mining queries are often complex need multi-level and multi-dimensional processing so the mining such data streams are most challenging task.
- Data stream elements change rapidly overtime. Thus, data from the past may become irrelevant for the mining.

5. CONCLUSION

Traditional data mining methods can not be easily applied to the data stream domain due to the unique characteristics of data streams. In this research paper we have addressed the data mining techniques, data stream mining models and challenges of data stream mining. Different models are proposed for data mining and data stream mining, the proposed models have advantages as well issues addressed in this review paper. We have also discussed data stream mining open challenges in knowledge discovery process. Mining data streams is still in its infancy state. Finally, we are concluding that due to unique characteristics of data streams, still research will be carrying considerable.

6. ACKNOWLEDGMENTS

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.

7. REFERENCES

[1] Carlo Zaniolo, and Hetal Thakkar: Mining Data Bases and Data Streams. University of California, Los Angeles.

[2] J. Han and M. Kamber*:* Data Mining, Concepts and Techniques. Morgan Kaufman, 2001

[3] R. Agrawal, T. Imielinski, and A. Swami: Mining Association Rules between Sets of Items in Large Databases. In Proceedings of 1993Internatinal Conference on Management of Data

[4] R.Agrawal and R.Srikan: Fast Algorithms for Mining Association Rules. Proceeding on the 20th VLDB conference 1994.

[5] Mahnoosh Kholghi, Hamed Hassanzadeh, and MohammadReza Keyvanpour: Clssification and Evaluation of Data Mining Techniques for Data Stream Requirements. ISCCCA, 2010.

[6] Mohamed Medhat Gaber, Shonali Krishnaswamy and Arkady Zaslavsky: Cost-Efficient Mining Techniques for Data Streams. Conferences in Research and Practice in Information Technology (DMWI 2004), Vol. 32.

[7] Albert Bifet and Richard Kirkby: Data Stream Mining A Practical Approach. The University of WAIKATO, 2009.

[8] M.Henziger, P. Raghavan, and S.Rajagopalan: Computing on data streams. In TR1998-001 Compag System Research. 1996.

[9] Aggarwal. C.C.: Data Streams: Models and Algorithms Springer Berlin Heidelberg, 2007.

[10] Chang, J. and Lee, W. : A sliding window method for finding recently frequent itemsets over online data streams ,JISE ,Vol. 20, 2004.

[11] J. Chang and W. Lee : A Sliding Window Method for Finding Recently Frequent Itemsets over Online Data Streams, JISE, Vol.20, 2004

[12] Mahnoosh Kholghi, and MohammadReza Keyvanpour: An Analytical Framework for Data Stream Mining Techniques Based on Challenges and Requirements, IJEST, 2011.

[13] Gaber M.M., Zaslavsky A, and, Krishnaswamy S: Mining Data Streams: A Review, SIGOD, 2005.

[14] Gaber M.M., Zaslavsky A, and, Krishnaswamy S: Resource-Aware Knowledge Discovery in Data Streams, In Proceedings of FIWKDDS, 2004.

[15] Babcock, B., Babu, S., Datar, M., Motwani, R. and Widom, J. : Models and issues in data stream systems, In proceedings of t SIGMOD-SIGACT-SIGART Symposium on Principles of database systems (PODS), New York , 2002.