Categorizing and Analyzing the Data Stream Requirements in Distributed System

[1] G.Surekha, [2]P. Lakshmi Priya
[1]Assistant Professor, [2]Assistant Professor
Computer Science and Engineering
Vidya Jyothi Institute of Technology, Hyderabad, India

Abstract: Data mining refers to extracting or mining knowledge from large amounts of data. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks are classified into two categories descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. Stream data flow in and out of a computer system continuously and with varying update rates. They are temporally ordered, fast changing, massive, and potentially infinite. It may be impossible to store an entire data stream or to scan through it multiple times due to its tremendous volume. In distributed systems, database systems and data mining the managing and processing of data stream is an active research area. Data Streams are highly dynamic nature due to high data distribution. Data streams are applied in network monitoring, tele communication systems and sensor networks. In online monitoring time and space utilization should be very efficient. My quest analyses and classifies application of diverse data mining techniques in efficient aspects of data stream mining. A solution is designed to map data mining techniques to data stream mining challenges and requirements.

Keywords: **Data mining, Data stream, Classification, Distributed System, Sensor networks.**

I. INTRODUCTION

Data mining techniques are suitable for simple and structured data sets like relational databases, transactional databases and data warehouses. Fast and continuous development of advanced database systems, data collection technologies, and the World Wide Web, makes data grow rapidly in various and complex forms such as semi structured and non-structured data, spatial and temporal data, and hypertext and multimedia data. Therefore, mining of such complex data becomes an important task in data mining realm. In recent years different approaches are proposed to overcome the challenges of storing and processing of fast and continuous streams of data.

Data stream can be conceived as a continuous and changing sequence of data that continuously arrive at a system to store or process . Imagine a satellite-mounted remote sensor that is constantly generating data. The data are massive (e.g., terabytes in volume), temporally ordered, fast changing, and potentially infinite. These features cause challenging problems in data streams field. Traditional OLAP and data mining methods typically require multiple scans of the data and are therefore infeasible for stream data applications. Whereby data streams can be produced in many fields, it is crucial to modify mining techniques to fit data streams.

Data stream mining has many applications and is a hot research area. With recent progress in hardware and software technologies, different measurement can be done in various fields. These measurements are continuously feasible for data with high changing ratio. Common applications which require mining of large amount of data to find new patterns are sensor networks, store and search of web events, and computer networks traffic. These patterns are valuable for decision makings. Data Stream mining refers to informational structure extraction as models and patterns from continuous data streams. Data Streams have different challenges in many aspects, such as computational, storage, querying and mining. Data stream mining is the extraction of structures of knowledge that are represented in the case of models and patterns of infinite streams of information.

For extracting knowledge or patterns from data streams, it is crucial to develop methods that analyze and process streams of data in multidimensional, multi-level, single pass and online manner. These methods should not be limited to data streams only, because they are also needed when we have large volume of data. Moreover, because of the limitation of data streams, the proposed methods are based on statistic, calculation and complexity theories. For example, by using summarization techniques that are derived from statistic science, we can confront with memory limitation. In addition, some of the techniques in computation theory can be used for implementing time and space efficient algorithms. By using these techniques we can also use common data mining approaches by enforcing some changes in data streams.

II. LITERATURE SURVEY

The problem of data stream classification is the arrival of data in an abstractly immeasurable stream and the chance to determine each record is briefed [1]. An online stream classification algorithm executing in amortized O (1) time controls the irregular appearance of labeled records. The algorithm judge internally on the basis of quality of updates models. The models are

updated from the unlabelled records to identify the inclusion of labeled records. The stream-classification algorithm handles multiple target classes. A new technique for temporal data mining based on classification rules with human domain experts easy understating is enabled [5]. Generally, time series are disintegrated into short segments, and short-term trends of the time series within the segments. The disintegrated time series short segments like average, slope and curvature are segmented by polynomial models. The classifiers determine short sequences of flow in consecutive segments with their rule premises. The resultants slowly allot an input to a class. The time series are assigned to originate with the productive time series model as classifier detects the anomalies. The time series pass provides a faster modeling in segmenting and piece wising polynomial models. The method is appropriate to problems with harsh timing constraints.

A large data set source classification is a limitation. A test chain classification is enabled to overcome the limitation [7]. The chain classification for large dataset classifies data based on the knowledge in the extending areas of adjacent scenes. The basic idea was to classify one dataset initially where the best truth data is present. Then to classify the adjacent dataset using classification of the initial overlap set as a training data. The arrival of novel class aspects are handled in the data stream classification method [2]. The data stream classification technique binds a novel class detection mechanism into traditional classifiers. The integration implements automatic detection of novel classes before the occurrence of true labels novel class instances. More time consumed for recognizing similarity instances in a class model. In addition, most existing stream classification techniques access the true label of a data point speedily after the data point is classified. ‘ The ranking and classification are combined to provide more accurate determination of a heterogeneous information network [6]. Highly ranked objects within a class play a vital role in classification. Similarly a class membership detail is essential for evaluating a quality ranking over a dataset.

The efficiency of collective classification models is improved depending upon the amount of class labels information present [5]. An availability change is noticed with the increase in test set labels for the relative performance of statistical relational models. The Data partitioning methods like bagging and boosting are greatly utilized in multiple classifier systems. The classification accuracy is highly improved based on data partitioning methods. The analysis of training data distribution and its effects on the activities of multiple classifier systems is studied [3]. Various feature-based and class-based measures are used to determine the statistical functioning of the training partitions. The measures assess the importance of different types of training partitions.

REFERENCES

[1] Hanady Abdulsalam, David B. Skillicorn and Patrick Martin, “Classification Using Streaming Random Forests”, IEEE Transactions On Knowledge And Data Engineering, January 2011.

[2] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, “Classification and Novel Class Detection in ConceptDrifting Data Streams under Time Constraints”, IEEE Transactions On Knowledge And Data Engineering, June 2011.

[3] Rozita A. Dara, Masoud Makrehchi and Mohamed S. Kamel, “Filter-Based Data Partitioning for Training Multiple Classifier Systems”, IEEE Transactions On Knowledge And Data Engineering, April 2010.

[4] Ahmed Abbasi, Stephen France, Zhu Zhang, and Hsinchun Chen, “Selecting Attributes for Sentiment Classification Using Feature Relation Networks”, IEEE Transactions On Knowledge and Data Engineering, March 2011.

[5] Dominik Fisch, Thiemo Gruber, and Bernhard Sick, “Swift Rule: Mining Comprehensible Classification Rules for Time Series Analysis”, IEEE Transactions On Knowledge And Data Engineering, May 2011.

[6] E.W.T. Ngai, Li Xiu and D.C.K. Chau, “Application of data mining techniques in customer relationship management: A literature review and classification”, ELSEVIER Science Direct 2009.

[7] Carlos Ordonez and Sasi K. Pitchaimalai, “Bayesian Classifiers Programmed in SQL”, IEEE Transactions on Knowledge and Data Engineering, January 2010.

[8] B. Babcock, M. Datar, and R. Motwani. “Sampling from a Moving Window over Streaming Data.” In Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2002, pages 633–634.

[9] B. Babcock, M. Datar, and R. Motwani. Load Shedding Techniques for Data Stream Systems (short paper) In Proc. of the 2003 Workshop on Management and Processing of Data Streams, June 2003.