OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

CLASSIFICATION METHODS FOR DATA STREAM MINING

Rutuja Jadhav[1], Neha Sharma[2]
Assistant Professor, Computer Engineering Department, KKWIEER, Nashik, India[1]
Assistant Professor, Computer Engineering Department, KKWIEER, Nashik, India[2]
rhjadhav@kkwagh.edu.in[1], ngsharma@kkwagh.edu.in[2]

Abstract: Data Stream Mining is the process of extracting knowledge structures from continuous and rapid data records arriving at high speed. Stream mining is one of the emerging fields of research in Data Mining. With the growing use of Internet in this digital era, tremendous amount of data is generating exponentially which needs to be analysed. This data is continuous, very large in size and cannot be stored for a long time. So there is a need to processes the data as soon as it becomes available. Various algorithms are available for mining data from streams, which requires single or fewer number of scans. With the recent advancement in Internet of Things (IOT), huge data streams are generated, thus making stream mining one of the most promising area of research. This paper is a review of different Classification methods used for data stream mining.

Keywords: *Data Mining, Data streams, Classification*

I INTRODUCTION

Variety of information is collected in digital form in databases and in flat files such as Business Transactions, Scientific, Medical and Surveillance Data, Satellite Sensing, Digital Media and World Wide Web Repositories. Data mining is the process of discovering patterns in large data sets as mentioned above. The aim of data mining is to study and analyze the data, summarize useful information from it and identify relationship among them [1]. Various interesting patterns can be discovered from these relationships after performing some mathematical and statistical operations [2]. Data mining can be applied to different forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW).

Data Stream Mining is the process of extracting knowledge structures from continuous and rapid data records arriving at high speed. Stream mining is one of the emerging fields of research in Data Mining. With the growing use of Internet in this digital era, tremendous amount of data is generating exponentially which needs to be analysed. This data is continuous, very large in size and cannot be stored for a long time. So there is a need to processes the data as soon as it becomes available.

The stream data highlights many challenges to the mining process due to the presence of various attributes like: (i) temporarily ordered, (ii) fast changing, (iii) infinite in size [3]. Various methods are used for mining data from streams, such as Classification, Clustering and Outlier Analysis. In this paper we have focused on various classification methods used in stream data mining.

II CHALLENGES IN DATA STREAM MINING

This section explores different challenges while dealing with data streams

2.1 Challenges:

1. High speed of arrival & Infinite size of data

Memory management issues arise due to the huge length of data stream and high speed of data stream occurrence. This data is continuous, very large in size and cannot be stored for a long time. So there is a need to processes the data as soon as it becomes available. Summarizing techniques must be used to deal with the above problem [4].

2. Dynamic Nature

The Data streams change rapidly overtime. This continuous updation of data makes past data irrelevant for mining. So there is a need to develop models which can handle dynamic behavior of data streams [4].

3. Visualization Of Results

It becomes difficult for user to take decisions quickly, if the system generates results which are difficult to comprehend thereby making it complex for analysis. Intelligent monitoring can be one of the methods to deal with it [4].

4. Outliers Detection

Behaviour of outliers is unpredictable, as sometimes normal object may behave like outlier and vice versa with change in the data over time. Thus outlier handing is critical, as it may deform the entire clustering structure.

5. Multi-dimensional data streams

Working with multidimensionality of data is difficult, as all pairs of points may seem to be almost equidistant from one another [4].

6. Deciding Parameters

Identification of parameters to be used for data stream mining is exigent task, as it requires vast knowledge of the domain.

III FEATURES OF DATA STREAM MINING

- Incremental nature [4]
- Single scan functionality [5]
- Low time and space complexity [4]
- Concept evolution handling [5]
- Anytime result generation model [4]
- Robustness to outliers [5]

IV CLASSIFICATION TECHNIQUES FOR DATA STREAM MINING

Classification is a data mining function that assigns items in a collection to target categories or classes. Classification is a technique in which incoming data stream objects are studied or analyzed to decide which class they belong to. New classes can also be formed if an object does not belong to any available class [5].The goal of classification is to accurately predict the target class for each case in the data.

4.1 Generic model maintenance algorithm (GEMM) and FOCUS

The GEMM model algorithm is used for maintaining time varying subsets of a systematically evolving database. It is a generic algorithm which can maintain any class of models. It mainly deals operations where insertion of records is much cheaper as compared to the deletion operation [6].

FOCUS

It is a framework which measures the deviation between the underlying data sets [6]. It is used to estimate the amount of change in data characteristics.

- *Mining Method: - Classification is done using decision trees and algorithms for mining frequent item sets.*
- *Advantage:* - It manages the problem of concept drift.
- *Disadvantage: - Time complexity of the algorithm is very high.*

4.2 Very fast decision tree (VFDT) and Concept adapting very fast decision tree (CVFDT)

VFDT technique for classification is based on Hoeffding decision tree algorithm, in which splitting of the current best attributes is performed based on threshold value specified by the user. Only promising nodes are considered in VFDT and rescanning of the database is done as and when time is available [7]

The CVFDT algorithm provides better speed and accuracy as compared to VFDT. As it is flexible in nature it has the ability to identify and respond to the changes on the example generating process. In CVFDT consistency of model is maintained by using a sliding window of examples and performs operations over it. It ensures that decisions made are updated from time to time based on monitoring quality of prior decisions on the recent data [7].

- *Mining Method:* - Classification is done using decision tree approach.
- *Advantage:*- Operates at high speed and requires less memory
- *Disadvantage:*- Does not manage the problem of concept drift and is a very prolonged process.

4.3 Context distanced measure (CDM)

CDM is a bottom up technique. It is a classification method that at each level of hierarchy. It aggregates the distance of the entities at the lower level [8]

- *Mining Method:- It* uses 2 classification methods i.e decision tree algorithm and bayes network model.
- *Advantage:* - Appropriate factors are used for distance measurement between events.
- *Disadvantage:* - *User* defined information is difficult to handle.

4.4 On-Demand Stream Classification (ODSC)

On-Demand Stream is a classification method which trains the model so that it can adapt rapidly with the changing data stream. It dynamically chooses the most suitable window of precedent training data to build the classifier. With the evolving data stream, system offers an efficient solution maintaining high accuracy. [9].

- *Mining Method:* - Classification is based on concept of micro-cluster, which is linked with class label.
- *Advantage:*- Works at immense speed and requires less memory.

- *Disadvantages:* - It is a very prolonged process.

4.5 Ensemble-based Classification (EBC)

Ensemble-based Classification uses amalgamation of multiple classification algorithms which makes it the most recent method used for the classification of the data stream. In this method the arriving data stream is separated into dissimilar chunks. Different classifier algorithm is applied on each chunk. Once the results of each chunk is obtained , the best possible results are selected for obtaining desired knowledge. As compared to single classifier algorithms, Ensemble-based Classification method serves better as it is very effective for most of the classification models [10].

- *Mining Method:- It* uses amalgamation of multiple classification algorithms.
- *Advantage:* - It manages the concept drift problem performs well in single pass operations and offers high accuracy.
- *Disadvantage:* - Operates at low speed and memory requirement is high.

V COMPARATIVE STUDY OF CLASSIFICATION TECHNIQUES

Following table gives a comparative study of various classification techniques for data stream mining based on parameters namely a) Mining Method b) Concept Drift c) Processing Speed and d) Memory Required.

Table 1: Comparative Study of Various Classification Techniques

Classification Technique	Mining Method	Concept Drift	Processing Speed	Memory
GEMM & FOCUS	Decision Tree & frequent item sets	Resolves	Low	Low
VFDT & CVFDT	Decision Tree	Does not Resolve	High	Low
CDM	Decision Tree & Bayes netwok model.	May Resolve	High	Low
ODSC	Micro Cluster	May Resolve	High	Low
EBC	Combination of multiple classifiers	Resolve	Low	High

VI DATA STREAMING TOOLS

6.1 Weka

Weka is a Java based free and open source software licensed under the GNU GPL and available for use on Linux, Mac OS X and Windows. It comprises a collection of machine learning algorithms for data mining. It packages tools for data pre-processing, classification, regression, clustering, association rules and visualisation.

6.2 Rapid Miner

Rapid Miner is available in both FOSS and commercial editions and is a leading predictive analytic platform. Besides the standard data mining features like data cleansing, filtering, clustering, etc, the software also features built-in templates, repeatable work flows, a professional visualisation environment, and seamless integration with languages like Python and R into work flows that aid in rapid prototyping.

6.3 Orange

It is a Python library that powers Python scripts with its rich compilation of mining and machine learning algorithms for data pre-processing, classification, modelling, regression, clustering and other miscellaneous functions.

6.4 Knime

Knime is one of the leading open source analytic, integration and reporting platforms that comes as free software and as well as a commercial version. Written in Java and built upon Eclipse, its access is through a GUI that provides options to create the data flow and conduct data pre-processing, collection, analysis, modelling and reporting.

6.5 DataMelt

DataMelt or DMelt does much more than just data mining. It is a computational platform, offering statistics, numeric and symbolic computations, scientific visualisation, etc. DMelt provides data mining features like linear regression, curve fitting, cluster analysis, neural networks, fuzzy algorithms, analytic calculations and interactive visualisations using 2D/3D plots and histograms.

6.6 Apache Mahout

Mahout is primarily a library of machine learning algorithms that can help in clustering, classification and frequent pattern mining. It can be used in a distributed mode that helps easy integration with Hadoop.

6.7 ELKI

ELKI is open source software written in Java and licensed under AGPLv3. This software focuses especially on cluster analysis and outlier detection with a compilation of numerous algorithms from both these domains.

6.8 MOA

Massive Online Analysis (MOA), as the name suggests, is primarily data stream mining software that is well suited for applications that need to handle volumes of real-time data streams at a high speed.

6.9 KEEL

KEEL (Knowledge Extraction for Evolutionary Learning) is a Java based open source tool distributed under GPLv3. It is powered by a well-organised GUI that lets you

manage (import, export, edit and visualise) data with different file formats, and to experiment with the data (through its data pre-processing, statistical libraries and some standard data mining and evolutionary learning algorithms).

6.10Rattle

It is expanded to 'R Analytical Tool To Learn Easily', has been developed using the R statistical programming language. The software can run on Linux, Mac OS and Windows, and features statistics, clustering, modelling and visualisation with the computing power of R.

VII CONCLUSION

In this paper we have discussed about data stream mining, challenges and features in data stream mining. Further we have explored various classification techniques in data stream mining. Insight of this study provides a better understanding of the domain to carry out further research. From the above comparative study, Ensemble based classifier, being the most recent classification technique, has a greater scope of enhancement in time and space consumption. Also a study of different data stream mining tools is presented. Since the data stream is being generated continuously at a fast pace, there is always a need to achieve better results and determine the optimal solution.

REFERENCES

[1] Almasoud, A. M., Al-Khalifa, H. S., & Al-Salman, A. (2015, October). Recent developments in data mining applications and techniques. In *Digital Information Management (ICDIM), 2015 Tenth International Conference on* (pp. 36-42). IEEE

[2] Shukla, M., & Kosta, Y. P. (2016, August). Empirical analysis and improvement of density based clustering algorithm in data streams. In *Inventive Computation Technologies (ICICT), International Conference on* (Vol. 1, pp. 1-4). IEEE

[3] Kamber, M., & Han, J. (2002). Data mining: concepts and techniques. *Newsletter ACM SIGMOD, 31*(2), (pp. 66-68).

[4] Toshniwal, D. (2013, February). Clustering techniques for streaming data-a survey. In *Advance Computing Conference (IACC), 2013 IEEE 3rd International* (pp. 951-956). IEEE

[5] Ashish Kumar & Ajmer Singh, Stream mining a review: tools and techniques. International Conference on Electronics,Communication and Aerospace Technology ICECA 2017 IEEE (pp. 27-32)

[6] Ganti, V., Gehrke, J., & Ramakrishnan, R. (2002). Mining data streams under block evolution. *ACM SIGKDD Explorations Newsletter*, *3*(2), 1-10.

[7] Chi, Y., Wang, H., & Yu, P. S. (2005, August). Loadstar: load shedding in data stream mining. In *Proceedings of the 31st international conference on Very large data bases* (pp. 1302-1305). VLDB Endowment.

[8] Kwon, Y., Lee, W. Y., Balazinska, M., & Xu, G. (2008, December). Clustering events on streams using complex context information. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on* (pp. 238-247). IEEE.

[9] Aggarwal, C. C., Han, J., Wang, J., & Yu, P. S. (2004, August). On demand classification of data streams. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 503-508). ACM.

[10] Wang, H., Fan, W., Yu, P. S., & Han, J. (2003, August). Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 226-235). ACM.