OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

CLUSTERING METHODS FOR DATA STREAM MINING

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

Abstract: **The process of finding patterns from huge data sets using various algorithms of machine learning, statistics, and database systems is known as Data mining. Data Stream mining is one of the promising areas of research in Data Mining. A data stream being an ordered sequence of instances, Stream Mining is the process of deriving knowledge structures from uninterrupted and speedy data records from these instances. Incredible amount of data is generated with the increasing use of Internet in this digital era, which has to be analysed. There is a need to process the data as soon as it becomes available as it is continuous and bulky in nature which cannot be stored for a long time. Various algorithms are available for data stream mining, which performs single or less 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 Clustering methods used for data stream mining.**

Keywords: *Data Mining, Data streams, Clustering*

I INTRODUCTION

Diverse amount of information is generated digitally in the form of 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 deals with the study and analysis of data, abridging useful information from it and discovering relationships among them [1]. A range of appealing patterns can be derived from the relationships obtained by performing various 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 [1]. Data Stream mining is one of the promising areas of research in Data Mining. A data stream being an ordered sequence of instances, Stream Mining is the process of deriving knowledge structures from uninterrupted and speedy data records from these instances. Incredible amount of data is generated with the increasing use of Internet in this digital era, which has to be analysed. There is a need to process the data as soon as it becomes available as it is continuous and bulky in nature which cannot be stored for a long time. Various algorithms are available for data stream mining, which performs single or less number of scans.

Due to the existence of various attributes like: (i) temporarily ordered, (ii) fast changing, (iii) infinite in size [3], the stream data brings along many challenges to the mining process Various methods are used for mining data from streams, such as Classification, Clustering and Outlier Analysis

In this paper we have focused on various Clustering methods used in stream data mining.

II FEATURES of DATA STREAM MINING

This section explores features of data stream mining

2.1 Incremental nature [4]

There is a need of flexible technique which updates the model as and when new incoming data arrives, thereby avoiding need to build a new model every time.

2.2 Single scan functionality [5]

Due to the continuous and rapid arrival of data, it cannot be stored for later usage because of enormous space requirement. Thus there is a need for single scan algorithms which will revise its model based on single access time of

streaming data points. In turn it will provide summaries of data for further analysis.

2.3 Low time and space complexity [4]

As huge amount of uninterrupted and speedy data records are generated, there is a need for time and space efficient algorithms working proficiently on summarized data.

2.4 Concept evolution handling [5]

Concept evolution occurs because of updating the features of data. In addition, it is a tendency in which fresh classes appear and the previous ones become obsolete over a period of time. So, an adaptive clustering model must be developed, this can be done by providing higher priority to fresh data & lesser to previous one.

2.5 Anytime result generation model [4]

Since data streams are boundless, it is difficult to wait for termination of the stream to produce the results. Thus, a model must be developed in such a way that it will be able to produce approximate results which will be available at any instance of time

2.6 Robustness to outliers [5]

A clustering technique must be developed in order to divide outlier from ordinary entity as it may alter the results.

III CHALLENGES IN DATA STREAM MINING

- *High speed of arrival & Infinite size of data[4]*
- *Dynamic Nature [4]*
- *Visualization Of Results [4]*
- *Outliers Detection*
- *Multi-dimensional data streams [4]*
- *Deciding Parameters*

IV CLUSTERING TECHNIQUES FOR DATA STREAM MINING

Clustering is a process of creating different groups called Clusters of input data. Elements within the same cluster have similar properties with a well defined frontier which is a dividing line between different clusters. Any number of clusters can be formed depending on the diversity of the data stream.

Following are the challenges while developing a clustering algorithm:

(i) Difficult to identify the cluster of objects that lie within the periphery of 2 or more clusters at same time.

(ii) Another issue is to determine what to do with objects that do not fall within the boundary of any cluster.

4.1 STREAM and LOCALSEARCH [6]

STREAM generates output whose cost is at most constant times the cost obtained by applying LSEARCH directly over entire stream[1]. Compared to LSEARCH, Algorithm STREAM is efficient in terms of space complexity as only O(*ik*) points need to be retained at the *ith* point of stream. So, retaining these O(*ik*) points may become excessively gigantic for very long streams. In this case there arises a need to recluster the weighted *ik* centres to preserve just k centres [6].

- *Mining Method:*- STREAM and LSEARCH uses K-Medians to form clusters.
- *Advantages:*- The incremental nature of the learning technique makes it flexible and cost efficient.
- *Disadvantages*:- Quality of clusters are degraded as data is generated at a very high speed.

4.2 VFKM [7]

VFKM stands for Very Fast K-means algorithm. Cluster formation in VFKM algorithm is entirely dependent on the requirements of the application. In this algorithm clustering is accomplished using two components Online component and Offline component.

*Online component is responsible for storing detailed summary of the statistics after regular intervals of time.

* Offline component on the other hand aids the analyst to identify the number of clusters to be produced or the time horizons to be considered.

Concept of pyramidal time frame is used in this algorithm in order to deal with issue of storage and to be able towork on fast data streams[7].

- *Mining Method:*- VFKM uses Method of K-Means for clustering.
- *Advantages:*- This algorithm is efficient in terms of time and space complexity.
- *Disadvantages:*- It is unable to carry out Multi-pass operations.

4.3 CluStream [8]

CluStream (Cluster Feature Vector-CFV) is a technique that uses cluster feature vectors to symbolize micro-clusters. The sum of square of all the tuples for each attribute dimension including the time attribute is preserved by CFV. This means that CFV calculates the mean and variance for each micro-cluster in each feature dimension and time[8].

- *Mining Method:*- CFV uses microclustering along with pyramidal time frame as a mining method.
- *Advantages:*- Acts great at concept drift detection and requires less time and space for operation.
- *Disadvantages:*- It does not allow Offline clustering.

4.4 D-Stream [9]

D-Stream (Density stream) – D-Stream is an algorithm that proposes a framework in which clustering of data stream is carried out using density based approach. It uses two components for the clustering process:

* Online component is responsible for structuring the input data in the form of a grid.

* Offline component then evaluates the density of the grid and forms clusters of the grids based on the calculated densities [9].

- *Mining Method:-* Cluster formation is based on the density of the grid.
- *Advantages:-* It identifies concept drift and produces quality output.
- *Disadvantages:-* It is not efficient.

4.5 WSOM (Arbitrary Window Stream modeling Method)

Interesting patterns are efficiently identified which are generated by sensors placed at remote locations with the help of this algorithm. It is beneficial to use WSOM as the operations are automated either before or during data gathering without any need of user interference. Updations are performed in a constant time using the logarithmic space. It requires limited number of resources to operate [10].

- *Mining Method:-* Operations are performed using Predictions.
- *Advantages :-* Less Memory required and single Pass updation of model is done dynamically
- *Disadvantages :-* High complexity.

V APPLICATIONS OF DATA STREAM MINING

Applications [4] of data stream mining are as follows:

5.1 Meteorological Research – Weather forecasters use clustering approach to identify the similarity in weather conditions at 2 different geographical locations on earth which are grouped in different clusters.

5.2 Stock Exchange – To handle fluctuations in the prices of stock market, clustering technique is used.

5.3 Detection of Abnormal Behaviour in Wireless Sensor Networks (WSN) – To determine the abnormality and intrusion in WSN, sensor nodes which are located at distinct geographical regions sends information about changes in its surrounding. So, data stream mining techniques are useful in indentifying unusual behaviour of nodes.

5.4 Supermarket – Clustering technique is used by the supermarket to keep track of the products they have sold, products frequently brought together to offer discounts to enhance the business.

VII CONCLUSION

In this paper we have discussed about data stream mining, challenges and features in data stream mining. Further we have explored various clustering techniques in data stream mining. Insight of this study provides a better understanding of the domain to carry out further research. Also applications of data stream mining are 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] Rutuja H. Jadhav & Neha G. Sharma. Clustering Methods for Data Stream Mining. Open Access International Journal of Science and Engineering(OAIJSE) Vol. 3, March 2018 pp. 94 – 97)

[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] O'callaghan, L., Mishra, N., Meyerson, A., Guha, S., & Motwani, R.(2002). Streaming-data algorithms for high-quality clustering. In *Data Engineering, 2002. Proceedings. 18th International Conference on* (pp.685-694). IEEE.

[7] Hulten, G., Spencer, L., & Domingos, P. (2001, August). Mining time changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp.97-106). ACM.

[8] Aggarwal, C. C., Han, J., Wang, J., & Yu, P. S. (2003, September). A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29* (pp.81-92). VLDB Endowment.

[9] Chen, Y., & Tu, L. (2007, August). Density-based clustering for realtime stream data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp.133-142). ACM.

[10] Papadimitriou, S., Brockwell, A., & Faloutsos, C. (2003, September). Adaptive, hands-off stream mining. In *Proceedings of the 29th international conference on Very large data bases-Volume 29* (pp. 560-571). VLDB Endowment.