Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 14 (2020) Pp-146-149

BBRC
Bioscience Biotechnology Research Communications

Survey and Research Issues in Data Stream Mining

Lalit Agrawal[1*] and Dattatraya Adane[2]
[1,2]*Shri Ramdeobaba College of Engineering and Management, Nagpur, India*

ABSTRACT

These days, an immense assortment and volume of information is constantly getting created from heterogeneous sources accordingly prompting a huge enthusiasm for the rising field of information stream mining. Information stream mining is where information is extricated from the data accessible in the information streams. There exist numerous applications which require this information for impromptu creation and business needs. Therefore preparing information streams in a proficient way is investigated by the specialists. In this paper, we present a short survey of different strategies accessible for the information stream mining.

KEY WORDS: CLASSIFICATION, DATA MINING, DATA STREAMS MINING

INTRODUCTION

Conventional frameworks for dealing with information digging were reasonable for the essential and appropriately orchestrated sort of data. Along these lines this sort of framework takes a lot of time in assortment of information, stockpiling of the information and their preparation. However, with the adjustment in situation these days, the choice must be taken "on-the-fly". Before delving into the subtleties of the different difficulties and their answers accessible in the writing it is critical to talk about the idea of the information stream. The stream is ceaseless and perpetual in nature. As a result of this property it is hard to store the whole stream into a concentrated information base and afterward apply a calculation to separate information accessible inside it. Other than capacity, handling this huge volume of information additionally gives certain difficulties [Khalilian M., 2010]. To beat various issues, the analysts have planned an alternate sort of information digging calculations and approaches for capacity and investigation of information constantly getting created from information streams [Han J., 2011].

Exploration in the field of information stream is essentially inspired by numerous applications which include volume of information age from an assortment of sensor information, information from different Supermarket applications, phone logs, information from satellites and different sources. Conventional methodologies are insufficient to mine information in the present condition which requires a constant examination and fast activity to inquiries as the information recently was Static and changing intermittently yet these days it is persistent and quickly changing hence new calculations are required[Agrawal L., 2020]. This paper is organised as follows: Section 1 provides the introduction about the data streams. Section 2 provides various methodologies for processing the data stream. Section 3 explains various stream mining algorithms which are available and section 4 describes various researches that have been carried out in the field of data stream mining.

Methodologies of data stream processing: Information streams are tremendous in volume, accordingly it is exceptionally hard to store information locally before handling. Subsequently it is clear that there is a compromise between the precision and the storage space

ARTICLE INFORMATION

*Corresponding Author: agrawalls@rknec.edu
Received 17th Oct 2020 Accepted after revision 24th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

Clarivate Analytics crossref

NAAS Journal Score 2020 (4.31)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.14/35

needed. Synopsis [Aggarwal Charu C., 2007] is a kind of data structure that provides a summary of the data. This is smaller than the actual data sets but covers all the major aspects of the data. Hence the output generated considering the synopsis is approximately correct. We need efficiency both in terms of time complexity and space complexity. In the segment below we will discuss various such techniques which improve the accuracy of the prediction.

A. Random Sampling: Random sampling is considered as the simplest method for synopsis construction. In this type of method, data streams are sampled periodically. In this the specialisation of representation is not carried out instead multidimensional representation is generated related to the data points, hence the synopsis can be used with a variety of applications. Reservoir sampling is a technique which is utilized to choose arbitrary components which are fair with no substitution [Vitter J.S., 1985]. To choose tests of impartiality we should know the length of the information in advance yet since it is absurd on account of an information stream, this methodology is changed a bit. The primary concept in random sampling is selection of a reservoir of a sample of size[Gibbons P.B., 1998].

B. Sliding Windows: Instead of sampling data periodically, a new concept of sliding window can be used for analysis. The main motivation in this method is instead of computing a sample only the recent data is taken into consideration to make the decision which will replace the older data [Poosala V., 1999]. Window Size is 'a' and 'b' is considered as the arrival time of the new element. Considering this the expiry of the node can be calculated as 'a + b'. It also solves various problems of a memory requirement as only the data of the window size which is of a smaller size needs to be maintained.

C. Histograms: A synopsis data structure which measures the frequency distribution of values in a stream of information is known as histograms. It creates various extents by separating the information along the attributes to maintain the count of each bucket. The depth of histogram is further decided by the rule which is used to divide the data. Answers to the range queries can be effectively given using this strategy and in light of the fact that the main thing is to be determined is the bucket in which data falls, the query resolution can be additionally made proficient by deriving various methodologies from the available histogram [Poosala V., 1999]. The variant of this method is proposed as V-Optimal histograms [Jagadish H.V., 1998].

Data stream mining algorithms: The research in data stream mining has evolved continuously because of the enormous volume of data that is getting generated from a variety of applications and their business requirements. Various procedures have been proposed to extract meaningful information from the continuously evolving data streams.

A. Clustering: Various applications require the bifurcation of available information into different segments. Those segments are referred to as clusters. Various techniques are available to cluster the static information but it requires an additional overhead to cluster the dynamic, continuously evolving data streams as it has to be done in a single pass. Few methods for clustering data streams are mentioned below:

STREAM – Guha, Mishra, Motwani and O'Callaghan proposed a k-median based Stream Clustering Algorithm. It relies on the divide and conquer approach. In the initial stage this algorithm first fragments the incoming data stream into smaller fragments and after that it tries to identify smaller fragments by using the k-median algorithm. In the subsequent stage, weighted cluster centers are grouped in a small number of groups. This algorithm doesn't consider the concept evaluation concept in the stream.

CluStream [Aggarwal C.C., 2003] clustering methodology Works in two phases that are online and offline. In the online phase it stores the summaries of the data which is coming in the form of an information stream into the micro clusters. This idea is basically the advanced state of BIRCH [Poosala V., 1999]. Bigger segments are created by Offline components by applying the k-means grouping algorithm. ClusTree [Kranen A., 2009] Clustering methodology also works in two phases that are online and offline. In the Online phase it learns from the micro clusters. Any miscellaneous collection of algorithms can be utilized for offline components. This algorithm is termed as one of the best dynamic models.

HPStream [Aggarwal C.C., 2004] focuses on the grouping of multidimensional data streams. This technique gives more weight age to the more recent data whereas reducing the preference or the importance of the old data. It updates on incremental bases and it is different for every dimension. Hue Stream algorithm [Meesuksabai W., 2012] augments E-Stream algorithm [Udommanetanakit K., 2007], which has been discussed before this.. Probability distribution function is presented in this algorithm to support the vulnerability in various attributes functions. In this algorithm the proposed function is further utilised to join various incoming clusters or the new information which is approaching and it further decides the data where to put in by using the histogram methodology.

POD Clus[Rodrigues P.P., 2008] Is a popular model for grouping various incoming information streams. POD is known as probability and Distribution-based Clustering. This algorithm is better for two reasons: first it is appropriate for grouping by illustrations and it is also good by variable selections. To create the cluster data and to update it on the fly the data summary in the form of say means, standard deviation is utilised in this algorithm. This algorithm suits for the concept evolution because it allows the new groups to appear, the splitting of the existing groups and it also helps in merging of the two groups and removal of few groups.

B. Classification: Variety of classification strategies are available for stationary data. This is basically a two-step process where the new model is created by properly arranging the incoming data and then it is further utilised to predict the unknown class names from this new information. In traditional use the training information is stored in the database and it is available for screening multiple times but in the information streaming storing this continuously coming high speed data is impossible and not available for several times screening. We have mentioned below the few popular classification algorithms for the streaming data:

Hoeffding Tree Algorithm is proposed by Domingos and Hulten focuses on the concept of the splitting decision tree called the Hoeffding Tree. This name is derived from the Hoeffding bound.The core logic in this algorithm is that the hoeffding bound provides you with a certain level of assurance about the best attribute to divide further. The main advantage of this algorithm is that it gives high accuracy even if the data set is small in nature and in a single pass of screening. The main disadvantage of this algorithm is that it is not capable of handling the concept scenarios.

Fast Decision Trees [Hulten G., 2001] which is proposed by Domingos et al. made an serious effort to improve the rate and precision of classification. It divides the tree by identifying the current best attributes for splitting.The main advantage of this algorithm is that it gives better accuracy even if the information stream is small in nature and the main disadvantage of this algorithm is that it cannot handle the change in concept that is concept drift. To overcome this advantage this algorithm is further modified as Concept-adapting Very Fast Decision Tree(CVFDT). It uses the sliding window concept. Classification on Demand is based on the concept of Clustream [Aggarwal C.C., 2003]. The clustering process is further divided into two segments: in the first segment the analysis of data is carried out and in the later segment classification is carried out.

Research Issues: There are various research issues related to the extraction of data streams from heterogeneous sources and merging them to gain more accurate Insight. Few of these are mentioned in [Agrawal L.S., 2016].

- Handling the variety of input streams in real time
- Memory requirements to store and compute this information
- Privacy of the data
- After mining, the format of presentation is a research issue for heterogeneous data streams
- Data gets evolved continuously. So how to consider new knowledge and update the prediction is an important issue.
- Various tools are being developed by the researchers towards the analysis of data streams and features to include to cater all kind of data streams
- How the addition of new streams and removal of old streams will affect the current prediction and changes for future
- Genuine information may come at any time. How to differentiate between the outliers are common issues
- In data mining, data is discarded after it gets processed. Mechanism for data recovery if data is needed in future for reference
- Models to mine data streams should be smart enough to differentiate between change in concept and noise in data.

CONCLUSION

Many challenges are presented by the continuously evolving data streams and it has made researchers to focus and develop new ways to deal with such heterogeneous groups of data.In this survey paper, we highlighted the different issues raised by the data streams and review of a variety of clustering and classification approaches developed by the researchers to handle this information. One algorithm or technique is not suited to mine all kinds of data streams. The search for the best method to handle this data is still going on because of a variety of issues like accurate prediction, adaptability to the changing environment etc. We have presented a few popular algorithms in the field of clustering and classification of data streams.

From this survey we can conclude that the data streams are a huge volume of information, dynamic and contain many information needed for a variety of applications. So applying a static data mining algorithm is very difficult. The research in this field is still in its early phase and various issues mentioned in earlier sections can be solved by carrying out research in this domain. It is definite that data mining will play an important role in future business strategies.

REFERENCES

Aggarwal, Charu C., ed., "Data streams: models and algorithms", Springer Science & Business Media, Vol. 31, 2007.

Aggarwal, C.C., Philip, S.Y., Han, J. and Wang, J., 2003, January. A framework for clustering evolving data streams. In Proceedings 2003 VLDB conference (pp. 81-92). Morgan Kaufmann.

Aggarwal, C.C., Han, J., Wang, J. and Yu, P.S., 2004, August. A framework for projected clustering of high dimensional data streams. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 (pp. 852-863).

Agrawal, L.S. and Adane, D.S., 2016. Models and issues in data stream mining. International Journal Of Computer Science And Applications, 9(1), pp.6-10.

Agrawal, L. and Adane, D., 2020. Survey and Research Issues in Data Stream Mining. Helix, 10(04), pp.428-431.

Datar, M., Gionis, A., Indyk, P. and Motwani, R., 2002. Maintaining stream statistics over sliding windows.

SIAM journal on computing, 31(6), pp.1794-1813.

Gibbons, P.B. and Matias, Y., 1998, June. New sampling-based summary statistics for improving approximate query answers. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data (pp. 331-342).

Han, J., Pei, J. and Kamber, M., "Data mining: concepts and techniques", Elsevier, 2011.

Hulten, G., Spencer, L. and 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).

Jagadish, H.V., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K.C. and Suel, T., 1998, August. Optimal histograms with quality guarantees. In VLDB (Vol. 98, pp. 24-27).

Khalilian, M. and Mustapha, N., 2010. Data stream clustering: Challenges and issues. arXiv preprint arXiv:1006.5261.

Kranen, A., 2009. Self Adaptive any time clustering.

Meesuksabai, W., Kangkachit, T. and Waiyamai, K., 2012. Evolution-Based Clustering Technique for Data Streams with Uncertainty. Agriculture and Natural Resources, 46(4), pp.638-652.

Poosala, V., Ganti, V. and Ioannidis, Y.E., 1999. Approximate query answering using histograms. IEEE Data Eng. Bull., 22(4), pp.5-14.

Rodrigues, P.P., Gama, J. and Pedroso, J., 2008. Hierarchical clustering of time-series data streams. IEEE transactions on knowledge and data engineering, 20(5), pp.615-627.

Udommanetanakit, K., Rakthanmanon, T. and Waiyamai, K., 2007, August. E-stream: Evolution-based technique for stream clustering. In International conference on advanced data mining and applications (pp. 605-615). Springer, Berlin, Heidelberg.

Vitter, J.S., 1985. Random sampling with a reservoir. ACM Transactions on Mathematical Software (TOMS), 11(1), pp.37-57.