

A Framework for Handling Evolving Data Streams

Tasneem Hasan

Professor

Department of CSE

ITM College of Engineering Nagpur, INDIA

Corresponding Author's Email Id: *tasneemh@itmnapur.ac.in*

Abstract

With recent advancement in technology need for analysis of such unbounded streams is increasing day by day. Data mining process helps to excavate useful knowledge from rapidly generated raw data streams. In context with the continuously generated data, mining data streams is emerging challenging task in which several issues like limited space, limited time, accuracy, handling evolving data need to be considered. In this paper the main method of research is clustering which is focused to handle evolving data streams. Most of the previously proposed methods inherit the drawbacks of k means method and fail to handle the issues. A hybrid data mining approach encompassing windowing, grid and density clustering and divide and merge method is proposed in this paper. A dynamic data stream clustering algorithm (DDS) is used in which a dynamic density threshold is designed to accommodate the changing density of grids with time in data stream. At last divide and merge approach is used to handle varying data points and further refine the quality of result obtained.

Keywords: *Data streams; Data mining; Concept drift; Clustering; Threshold*

INTRODUCTION

With the emerging applications involving massive data e.g. customer click, retail chain transaction, stock trading, emails etc leads to

the generation of continuous huge bulk of information streams. These data streams contains continuous fast and varying data points on which various data mining method

like clustering ,classification etc could be applied. The traditional method fails to handle such huge amount of dynamic data. Huge size of continuously flowing data has put forward a number of challenges in data stream analysis. The application of various data mining methods on these uncertain data streams helps to excavate hidden information which can be used in processing of data streams.

A data stream mining is an approach to extract meaningful information (knowledge) from raw data streams and find changes and evolution of stream overtime. A general form of data mining technique involves clustering which is important and difficult research topic in data mining, artificial intelligence, machine learning involving wide range of application such as website analysis, stock market analysis, astronomers, news organization etc. Clustering aims to identify and partition them into clusters such that object belonging to same cluster are similar and help user understand natural grouping of attributes. Various clustering algorithms are available for processing data streams but majority of them work on static data streams. Many issues can be envisioned

with handling of data streams which are summarized as follows: [1]

- **Large space:** Data streams have enormous volumes of continuously incoming data.
- **Dynamic data:** Data streams are fast, changing, uncertain and require fast response to incorporate changes in data and reflect it in output.
- **Noise:** Any approach applied to data streams should be able to deal with noise and outliers.
- **Single scan:** Since data streams have infinite volume of information which is fast and changing, hence stream data should be read only once.
- **Light weight:** Techniques applied to vast data streams should process stream in less time and memory to should provide an optimal output.

Data streams have an important characteristic that it is unpredictable, dynamic and uncertain in nature. A difficult problem with learning in many real-world scenarios is that the concepts are often not

stable but change with time. Concept of interest may depend on some hidden context, not given explicitly.

A typical example is weather prediction rules that may vary radically with the season. Often the cause of changes in continuously flowing data streams is hidden, not known a priori, making the learning task more complicated. Knowledge discovery and data mining from time-changing data streams and efficient concept drift handling on data streams have become important topics in the machine learning community and have gained increased focus recently [2].

Data mining is process of efficiently digging out useful information from underlying raw data. These are combination of techniques that make use of computer programs to automatically excavate model that represents useful patterns [3]. Due to recent technological advances and availability of streaming data in diverse range of fields increases importance of streaming data analysis [4]. Data mining techniques for streams are broadly classified in supervised and unsupervised techniques.

Supervised approach includes classification which is constrained to work only on labeled data set. Unsupervised approach encompasses clustering in which data set is unlabelled. As mentioned earlier data streams are uncertain so we cannot predict whether incoming data points would be labeled or unlabelled. Clustering approach is assumed to be more suitable for data streams because since it works on unlabelled data sets, it would also produce exceptionally good results if applied to labeled datasets.

Cluster analysis or clustering is a primary data mining tool which was first used by Tryon in 1939. By partitioning data streams into homogenous clusters in which objects have high intra class similarity and low inter class similarity, data miners can interpret about data properties which can then be utilized to develop efficient classification systems for new data or predictive model for hidden unforeseen events [12]. Due to versatility of clustering present section intends to provide broad overview on various commonly used clustering approaches.

Guha et. al. [5] adapted k means approach for clustering in which data is divided into

chunks, each of which is then separately clustered in weighted centers. At final stage centers are again clustered to give final clusters. This approach was again improved to produce STREAM [6] which uses moving window and produces centers at each particular stage.

This approach does not need to specify in advance the number of clusters expected at the output and instead evaluates the performance by using a combination of sum of squared distances and the number of centers used.

Online phase of CLUSTREAM [7] generates micro clusters which absorbs incoming data points and cluster them based on time window. In offline phase we get macro clusters which are generated based on cluster number, duration, etc using k means approach.

Divide and conquer technique was used by Edwin Lughofer [8] to devise dynamic split and merge method which would be able to handle dynamic nature of data. Merging is performed using weighted averaging of cluster centers based on homogeneity condition between two clusters. Splitting is

done when clusters are shrinking inside one large cluster to form separate clusters using Bayesian information criterion. It requires more space and processing time. It copes up with dynamic data to provide high purity and cleaner cluster.

Density based approach for clustering was utilized in [9] to present DENSTREAM which is extension of DBSCAN and is similar to CLUSTREAM. It defines structure of potential micro cluster and outlier micro cluster. When new data points arrives it checks if it can be fused with potential micro cluster, if not then try fusing it with outlier micro cluster else create new outlier micro cluster. When request for clustering occurs it uses DBSCAN to generate final output clusters. Based on some predefined criteria some outlier micro clusters are transformed into potential micro cluster rest are discarded which leads to loss of knowledge points.

rDENSTREAM algorithm [10] was proposed to overcome the drawback of DENSTREAM. It adds retrospect phase in above two steps on DENSTREAM. Here result of macro clustering of potential micro cluster is regarded as classifier, outlier micro cluster are not discarded but stored in buffer

which is later used to relearn the classifier to provide improved clustering accuracy and limit the loss of knowledge point. Although it overcomes the drawback of DENSTREAM by not discarding data points but requires additional space for this purpose.

A combination of grid and density based approach was used to present novel DD stream method [11]. The key idea is that incoming infinite data points are represented in form of finite number of cells to form grid structure on which clustering is done based on density of grids. DD stream [11] overcomes drawbacks of loss of boundary points by making use of DCQ means algorithm. DD stream is two phase approach that aims to retain boundary points and hereby produce good quality cluster.

In first phase incoming data points are read and mapped in grid cells. It associates density decay factor by using eigen vector to determine distance between data points and neighboring grid cell. In next phase DCQ means is used to calculate distance of data points on boundary with grid cluster centers and adds that point to the minimum distance grid.

In the light of these issues this paper presents a hybrid approach which includes windowing, dynamic data stream algorithm (DDS) for clustering which is based on grid and density approach. In order to further increase the efficiency and accuracy of the algorithm and to handle varying nature of data stream, divide and merge approach is also appended. The rest of the paper is organized as follows. Section II presents proposed system architecture .Section III provides a brief discussion on drawbacks of previous methods and advantages of proposed method. Finally conclusion and future work appears in section IV.

PROPOSED WORK

The grid-based clustering method uses a multi resolution grid data structure. It forms the grid data structure by dividing the data space into a number of cells and perform the clustering on the grids. Clustering depends on the number of grid cells and independent of the number of data objects .Grid-based method could be natural choice for data stream in which the infinite data streams map to finite grid cells. The synopsis information for data streams is contained in the grid cells. Grid can be defined as

division of data space into finite number of cells that form a grid structure.

Grid cell can be dense, sparse or empty. If a grid cell does not contain even a single data point then such grids are called empty grids and are discarded. Threshold specifies the minimum number of points to classify whether the grid is dense or sparse. If a grid cell contains density value greater than or equal to the threshold then that cell is considered as dense grid. If the density value of grid cell is less than threshold then that grid would be considered as sparse grid.

The density of grid is calculated using its dynamic weight. It is suitable for updating the grid cell feature with negligible time and storage overhead.

The figure below depicts the proposed system architecture.

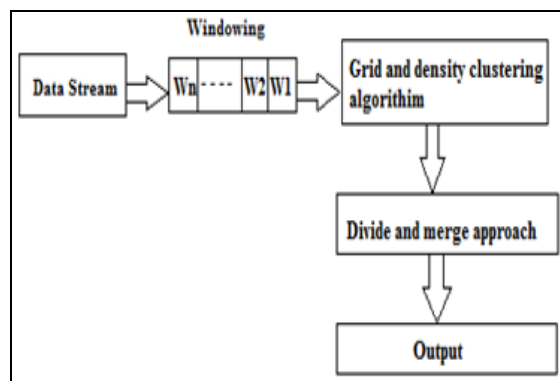


Figure 1 Proposed system architecture

As shown in figure. 1 there are following three modules. The first module is windowing which takes raw data streams as input and applies sliding window technique to form windowed stream. Huge bulk of information cannot be given directly to clustering algorithm as it would lead to space overhead and consume more time. Windowing simply partitions the huge data stream into blocks or windows. This windowed stream would be provided input to next module. The next module is grid and density based clustering where the windows obtained from previous stage are given one by one to the clustering algorithm.

A dynamic data stream algorithm (DDS) which aim to cluster uncertain data points is used here. Here first grid mapping is done where infinite data stream is mapped onto a finite grid cells. For every data point that has been located on grid cells, calculate its dynamic weight which depends on rate of processing, arrival time of data point and existence probability of data point. Using dynamic weight, dynamic density is calculated. For every non empty grid cell dynamic density can be defined as

summation of dynamic weight of all data points that belong to that grid cell.

Using the dynamic density and threshold classification of grid cells into dense and sparse grid is done. By comparing the value of dynamic density with threshold each cell can be stamped as either dense or sparse. If the value of dynamic density of a cell in grid is greater than or equal to threshold then classify that cell as dense. If the value of dynamic density of a cell in grid is less than threshold then it is considered as sparse.

Finally clustering is done based on density of grid cells. Every empty grid cells are discarded. For every dense grid if its neighbor is also dense then include them in one cluster and finds its neighbor. If neighbor is sparse then consider it boundary point for cluster. For every sparse grid if its neighbor is also sparse then consider it as noise point.

To handle the dynamic and unpredictable nature of data streams re clustering is done on the output obtained from grid and density clustering method. The intermediate clusters obtained from previous step would be provided input to divide and merge

technique to produce final output. A specific problem in evolving cluster models arise whenever two or more local data clouds (each one modeled by a cluster) are moving together or are delaminating inside one cluster. In such cases, where concept drift occurs clusters should be dynamically merged or split in order to account for correct cluster partitions.

DISCUSSION

The algorithms used in the proposed method aims to handle large volumes of data to yield meaningful patterns as output. As mentioned before, the major challenging issues in data stream mining are represented by limited time, limited space, accuracy of output, handling of concept drift. A good data mining technique should aim to satisfy these issues for efficient handling of data streams.

This section discusses the drawbacks of previous methods used and analyzes the methods used in proposed approach and lists the advantages of these methods.

STREAM [6] has various drawbacks which include failure to handle dynamic data,

requires more time and space. It has poor accuracy and it inherits all drawbacks of K means. However it is relatively simple and easy to implement.

CLUSTREAM [7] fails to handle changing data, leads to loss of knowledge points, and suffers from drawbacks of K means. It also requires more space and time, but has slightly more accuracy than STREAM. Dynamic split and merge [8] copes up with dynamic data to provide high purity and cleaner cluster. However it should be used in combination of some other technique to refine the results. As a standalone system it fails to produce better results.

DENSTREAM [9] leads to loss of data points due to which less accurate clusters are obtained. However it requires less processing time and space and can also handle concept drift. rDENSTREAM algorithm [10] overcomes the drawback of DENSTREAM by not discarding data points but requires additional space for this purpose. It requires more time but has better accuracy than DENSTREAM and can also handle concept drift.

DD stream method [11] overcomes the weakness of various existing methods by retaining boundary points, efficiently handles changing nature of data and has good accuracy and scalability. However it has more space requirement and further more improvements can be done to increase its accuracy. Based on this discussion it can be concluded that none of these approaches handle all the issues simultaneously.

Hence more stress should be laid to devise a hybrid approach that satisfies all the constraints efficiently which would be a major landmark in research area of data stream mining.

With the aim to handle all issues simultaneously the proposed system uses combination of three different methods. As mentioned earlier data streams have huge volumes of information and require more space hence windowing is used here as it requires less time, less memory and can be efficiently used to keep only the statistics of data.

One of the issues envisioned with efficient handling of data streams is that the technique applied should light weight and

produce accurate output. Grid and density based clustering can handle large data sets, fast processing, takes less time ,insensitive to parameters and input order, maps infinite data streams into finite data streams , comparatively more accurate than other clustering approach. The DDS algorithm used is not sensitive to anomalous data points, can efficiently handle uncertain data streams and has high accuracy.

To handle concept drifting data streams the results are again re clustered using divide and merge. A specific problem in evolving cluster models arise whenever two or more local data clouds (each one modeled by a cluster) are moving together or are delaminating inside one cluster. In such cases, where concept drift occurs clusters should be dynamically merged or split in order to account for correct cluster partitions. Divide and merge improves cluster quality and contributes in handling evolving data streams.

CONCLUSION

Data mining, also called Knowledge-Discovery is one of the hot topics in the field of knowledge extraction. With the

growing use of technology efficient handling of data streams is needed.

On the basis of the previous clustering method it can be concluded that none of the clustering methods accurately handle varying data streams. Previous clustering method fails to handle all the issues of data streams simultaneously. Hence a hybrid data mining approach was proposed in this paper that combines the advantages of individual methods to form a more robust system. The main research methodology used here for handling evolving data streams is clustering because clustering can handle both labeled as well as unlabelled data streams efficiently. Windowing helps in conserving memory, DDS clustering algorithm helps in fast processing of data streams and divide and merge further increases the accuracy by handling varying data points by re clustering them.

The future work encompasses practical implementation of this approach and its performance comparison with existing methods.

REFERENCES

1. Yi-Hong Lu, Yan Huang, "Mining data streams using clustering", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, IEEE, Guangzhou, August 2005, pp 2079-2083.
2. Kapil Wankhade, Vahida Attar, Pradeep Sinha, "A fast and light classifier for data streams", Springer, 2010, pp 200-207.
3. K.Prasanna Lakshmi, Dr.C.R.K.Reddy, "A Survey on different trends in Data Streams", Proceeding of International Conference on Networking and Information Technology, IEEE, 2010, pp 451-455.
4. Amr Magdy, Noha A. Yousri, and Nagwa M. El-Makky, "Discovering Clusters with Arbitrary Shapes and Densities in Data Streams ", In proceedings of 10th International Conference on Machine Learning and Applications, IEEE, 2011, pp 279-282.
5. S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams," Foundations of Computer Science, Annual IEEE Symposium, 2000, pp 359-366.
6. L. O'callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, " Streaming-data algorithms for high-quality clustering", Proceedings of the 18th International Conference on Data Engineering (ICDE.02), IEEE, 2002. pp 685–694.
7. C. Aggarwal, J. Han, J. Wang, and P. Yu, "A framework for clustering evolving data streams", In Proceedings of the 29th international conference on Very large data bases, 2003, Volume 29, pp 81–92.
8. Edwin Lughofer, "Dynamic Evolving Cluster Models using On-line Split-and-Merge Operations", In proceedings of 10th International Conference on Machine Learning and Applications, IEEE, 2011, pp 20-26.

9. Cao F, Ester M, Qian W and Zhou A, “Density-Based Clustering over an Evolving Data Stream with Noise”, Proceedings of the SIAM Conference on Data Mining, 2006, pp 326-337.
10. Liu Li-xiong, Huang Hai, Guo Yun-fei, Chen Fu-cai, “rDenStream, A Clustering Algorithm over an Evolving Data Stream”, IEEE, 2009.
11. Chen Jia, Cheng Yu Tan, Ai Yong, “A Grid and Density-based Clustering Algorithm for Processing Data Stream”, In proceedings of Second International Conference on Genetic and Evolutionary Computing, IEEE, 2008, pp 517-521.
12. T. Soni Madhulatha, “OVERVIEW OF STREAMING-DATA ALGORITHMS”, In journal of Advanced Computing: An International Journal (ACIJ), 2011, Vol.2, No.6.