ISSN No. 0976-5697

Volume 5, No. 4, April 2014 (Special Issue)

International Journal of Advanced Research in Computer Science

CASE STUDY AND REPORT

Available Online at www.ijarcs.info

Clustering Techniques for Temporal Data:A Survey

Sweta C. Morajkar
Computer Engineering Department
Goa Engineering College Farmagudi, Ponda Goa
scmorajkar33@gmail.com

J.A Laxminarayana
Computer Engineering Department
Goa Engineering College Farmagudi, Ponda Goa
jal@gec.ac.in

Abstract: Temporal data Clustering is a process of grouping objects based on their temporal similarity. The field of temporal data mining has gained high popularity due to processing on large amount of data which is generated. Since an additional dimension is incorporated, it needs to be handled in an efficient way. As a consequence, large amounts of temporal data becomes available that needs to be analyzed. A complete understanding of the structure of data requires that it should be viewed as sequence of timestamped events. This paper presents a survey on the most significant techniques to deal with clustering of temporal data.

Keywords: Temporal Clustering; Data mining; Temporal data mining; Clustering; Stream Clustering

I. INTRODUCTION

Due to increase in number of technologies, large amount of data gets generated. Due to enormous size, analysis over such data becomes a difficult task. There is interest in the discovery of hidden information. The discovery deals with processes such as clustering, classification and pattern extraction. Due to high dependency among the structure in temporal sequences, proper treatment has to be provided for clustering process. Most of the existing techniques ignore the temporal information and assume data as an unordered collection of events.

Clustering is an approach to analyze temporal data at higher level of abstraction. It groups data according to some similarity measure. Representing and clustering two dimensional data is manageable but combining this result with time imposes number of challenges. If clustering is performed based on time points, more complex data types come into picture. This mostly occurs in time series data where two time series are compared in order to find similarity patterns.

Temporal clustering analysis plays an important role in discovering the intrinsic structure within the data elements. The main objective is to partition an unlabelled data into clusters. Model selection and grouping are most important factors considered in cluster analysis. Model selection is a task of selecting statistical model from a set of elements whereas grouping demands a proper grouping rule that groups sequences to form a cluster.

II. WHY CLUSTERING

Clustering can be considered the most important unsupervised learning technique. The method allows to find intrinsic structure among data. The patterns generated from clustering can be used for predicting some important concepts related to event. Examples of clustering include stock market data, behavioral clustering. Most of the clustering algorithms work on static data in that clustering has to be performed each time a new data point arrives. Due to its dynamic nature, evolutionary algorithms are required. Some data stream clustering algorithms have also been designed. The original algorithms such as K-means, DBSCAN have been modified in order to handle the evolving data. There is a need to discover hidden information in the stored data as large amount of data gets accumulated. The discovery in such data deals with classification, clustering, identifying interesting patterns etc. In order to find relationships among data elements, temporal dependencies have to be considered. Most of the data mining tasks do not consider temporal information and consider data as an unordered collection.

III. TEMPORAL DATA

Temporal data plays an important role in representing time evolving data. Temporal data includes valid time and transaction time. Valid time is the time period when a fact is true whereas transaction time is the time period when a fact is stored in database.

Examples of temporal data ranges from multimedia information processing to temporal data mining. Data is represented as a sequence of timestamped events. Timestamp associated with each record refers to some event related to particular entity. Analysis over such data allows us to find longitudinal patterns.

Before any data mining task is applied on the data, preprocessing needs to be applied. When we consider timestamped data, representation problem is important task since there are lots of dependencies between data elements. High dimensional data can be represented in multiple forms. One solution is to use data with transformation either keeping data in its original form or considering the time windows. A second solution is to represent data in discrete form.

IV. REPRESENTING TEMPORAL DATA

Temporal data is usually represented in the form of trajectory. A trajectory is the path that a moving object follows through space as a function of time [1].In order to represent moving objects; trajectory is divided into a set of functions with distinct time intervals. After obtaining the derivative of each linear function, we get speed.

A linear interpolation method is used in [2].The movement of an object is represented in polyline 3d space. Objects move straight within certain points with constant

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014

speed. In[3],a similarity measure is defined between segmentation of two trajectories at a particular time interval. This distance is defined as distance between the rectangles at that time. Minimum bounding rectangles are considered for finding the distance.

V. TECHNIQUES

A. *Bayesian Approach:*

Due to dynamic nature of real world systems, there is a need to build a model which will better represent the information in form of probabilities [4].In some well known domains such as recognizing speech; there exists sufficient knowledge to construct models. But in many other domains, this type of knowledge may not be available. So data driven approaches are used in order to analyze dynamic behavior. One of the most challenging tasks is to automatically build models and create structures using exploratory data analysis measures. Representations for temporal data are usually classified into two types: piecewise and global. In piecewise, representation is obtained by partitioning temporal data into segments based on certain criteria. The obtained set of partitions is then modeled into specific representation.

B. *Trajectory Clustering:*

Trajectories represent the data of moving objects over time. The data is usually represented in point form. Temporal data considered with moving patterns plays an important role in representing such evolving features. Each individual data is represented by a sequence of measurement given by function of temporal dimension. By using EM algorithm, objects that are likely to be generated from a trajectory are clubbed together. A model based technique is discussed in [5], where a cluster is denoted by a Markov model that estimates transitions between successive positions. Parameters are estimated by means of expectation maximization algorithm.

In a trajectory is a function that maps time to locations. To represent object movement, a trajectory is decomposed into a set of linear functions, one for each disjoint time interval. The output of each function yields direction and the speed in associated time interval.

A distance function is used in order to represent temporal characteristics of trajectories. The authors have focussed on improving the quality of trajectory clustering. If we consider two trajectories, they may be different with respect to whole time period. However, if only a sub interval is considered, trajectories are found to be similar.

C. *Stream Clustering:*

Stream clustering deals with clustering data in orderly fashion. It is an ordered sequence of points which can be accessed only once. Stream clustering process is motivated by many real world examples such as web log data, telephone call data etc. Based on the static clustering algorithms such as density based clustering, the algorithms for streaming clustering have been modified.

In case of stream clustering, concept of online and offline cluster components is used. Online component stores periodically detailed statistics. Offline clustering uses the saved summary of clusters and performs clustering. Two clustering methods perform well for stream data. The first process is known as microclustering which takes into account set of all points which are very close to each other. These microclusters are treated as individual unit and thus used for macroclustering.

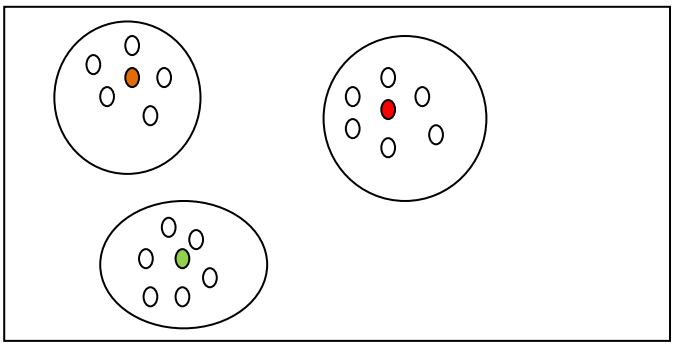

Figure 1. Microclusters

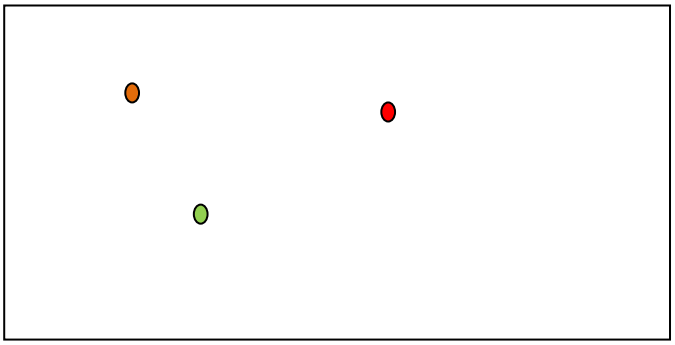

Figure 2. Macroclusters

Some of the challenges involved in streaming clustering are:

Stream data have massive volume. Thus it is difficult to store this data on disk. Data should be processed in single scan. By scanning the data once, summary information required for clustering is stored.

Data stream patterns continuously evolve over time [6].Due to evolutionary phenomenon of these streams, underlying clustering models require continous updation..A model structure must be available at any time.

StreamKM++: In order to choose first value for clustering, k-means++ algorithm is used [7]. This technique calculates weighted sample of data stream. The use of these coreset trees speeds up the time required for construction of coreset.

In CluStream[8], clustering process is divided into two components: offline and online. The online component creates a set of microclusters and offline component is used in order to computer clustering results based on snapshot data.

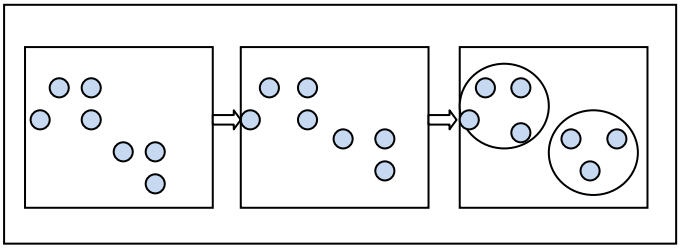

Figure 3. Snapshots at different time intervals

In [9], DenStream algorithm is used for evolving data streams. Existing DBScan method is enhanced in order to

handle streaming data. A fading function is used in order to reduce the weight in micro clusters over time. D-Stream is another density based streaming clustering algorithm. It divides the space into a set of grids. Each incoming point is mapped onto a corresponding grid. Similar to DenStream, a fading function is used in D-Stream to reduce weight over time period. Cobweb is an incremental clustering algorithm. It makes use of probabilistic approach. It uses classification tree. Each node in a tree is represented by concept. The probability value is associated with each node.

VI. CONCLUSION

The paper deals with temporal data clustering methods. It also explains different techniques to represent temporal data. The current clustering mechanisms require many parameters to be specified apriori, but clustering techniques for temporal data do not require any parameters. They are single pass clustering mechanisms. The paper also states the differences between all the techniques and also presents advantages and disadvantages associated with them.

Table I. Comparison of different stream clustering techniques

Sr. No.	***Stream Clustering Methodology***	***Clustering approach used***	***Parameters***	***Advantages***	***Disadvantages***
1	Stream Clustering	K medoids	Input as number of parameters	Incremental Learning	Low accuracy
2	StreamKM++	K means++	Number of clusters	Cluster quality is high	Time efficiency
3	Denstream	Density based approach	Cluster radius	High time efficiency	High Complexity
4	CluStream	Microclustering,Time frame clustering	Time window and number of clusters	High accuracy	Offline clustering
5	CobWeb	Incremental, classification	Probabilistic approach	Incremental Clustering	Probability distribution on separate attributes are independent of each other.

VII. REFERENCES

[1] S.Y. Hwang, Y.H. Liu, J.K. Chiu, F.P. Lim (2005) "Mining Mobile Group Patterns: A Trajectory-based Approach". T.B. Ho, D. Cheung, and H. Liu (Eds.): PAKDD 2005.

[2] Dieter Pfoser Christian S. Jensen Yannis Theodoridis "Novel Approaches to the Indexing of Moving Object Trajectories". Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000

[3] A. Anagnostopoulos, M. Vlachos , M. Hadjieleftheriou, E Keogh.,P.s. Yu, "Global Distance-Based Segmentation of Trajectories".KDD'06, Philadelphia, Pennsylvania, USA, August 20–23, 2006.

[4] Cen Li, Gautam Biswas "A Bayesian Approach to Temporal Data Clustering using Hidden Markov Models".

[5] Rabiner, L. R. 1989. "A tutorial on hidden markov models and selected applications in speech recognition". Proceedings of the IEEE 77(2):257–285.

[6] Daniel Barbara."Requirements for clustering Data Streams".SIGKDD Explorations,Volume 3.

[7] Marcel R. Ackermann, Marcus Martens, Christoph Raupach and Kamil Swierkot."StreamKM++: A Clustering Algorithm for Data Streams". ACM J. Exp. Algor. V, N, Article A January.

[8] Charu C. Aggarwal, Jiawei Han, Philip S. Yu."A Framework for Clustering Evolving Data Streams". Proceedings of the 29th VLDB Conference, Berlin.

[9] F. Cao,M. Ester,W. Qian, and A. Zhou. "Density-based clustering over evolving data stream with noise". Proceedings of the Sixth SIAM International Conference on Data Mining (SIAM 2006), pages 326–337, 2006.

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014