ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 9, November 2012

A Survey on Temporal Data Clustering

M. Yasodha[1], Dr. P. Ponmuthuramalingam[2]

Assistant Professor, Dr NGP Arts and Science College, Coimbatore, India[1]

Associate Professor, Government Arts College (autonomous), Coimbatore, India[2]

ABSTRACT: **Temporal data clustering provides underpinning techniques for discovering the intrinsic structure and condensing information over temporal data. To classify data mining problems and algorithms used two dimensions: data type and type of mining operations. One of the main issue that arise during the data mining process is treating data that contains temporal information. Temporal data representations are generally classified into two categories: piecewise and global representations the area of temporal data mining has very much attention in the last decade because from the time related feature of the data, one can extract much significant information which cannot be extracted by the general methods of data mining. Many interesting techniques of temporal data clustering were proposed and shown to be useful in many applications. Since temporal data clustering brings together techniques from different fields such as databases, statistics and machine learning the literature is scattered among many different sources. In this paper, present a survey on temporal data clustering.**

Keywords: Temporal Data, Data Mining, Clustering

I. INTRODUCTION

Temporal data are ubiquitous in the real world and there are many application areas ranging from multimedia information processing to temporal data mining. Unlike static data, there is a high amount of dependency among temporal data and the proper treatment of data dependency or correlation becomes critical in temporal data processing. Temporal clustering analysis provides an effective way to discover the intrinsic structure and condense information over temporal data by exploring dynamic regularities underlying temporal data in an unsupervised learning way. Its ultimate objective is to partition an unlabeled temporal data set into clusters so that sequences grouped in the same cluster are coherent. In general, there are two core problems in clustering analysis, i.e., model selection and grouping. The former seeks a solution that uncovers the number of intrinsic clusters underlying a temporal data set, while the latter demands a proper grouping rule that groups coherent sequences together to form a cluster matching an underlying distribution. Clustering analysis is an extremely difficult unsupervised learning task. It is inherently an ill-posed problem and its solution often violates some common assumptions [1]. In particular, recent empirical studies [2] reveal that temporal data clustering poses a real challenge in temporal data mining due to the high dimensionality and complex temporal correlation. In the context of the data dependency treatment, classify existing temporal data clustering algorithms as three categories: temporal-proximity-based, model-based, and representation-based clustering algorithms.

II. RELATED WORK

Representing spatio-temporal data in a concise manner can be done by converting it into a trajectory form. In Hwang et al. [3] 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 derivative of each linear function yields the direction and the speed in the associated time interval. A trajectory is a disjunction of all its linear pieces. For example, a trajectory of the user moving on a 2-D space may consist of the following two linear pieces:

$[(x=t-3)\Lambda(y=t+3)\ \Lambda(0<t<2)]U[(x=6)\ \Lambda(y=-t)\Lambda(3<t<5)]$

In Pfoser [4], a linear interpolation is used. The sampled positions then become the endpoints of line segments of polylines and the movement of an object is represented by an entire polyline in 3D space. A trajectory T is a sequence $<(x1,y1,t1),(x2,y2,t2),\ldots,(xk,yk,tk)>$. Objects are assumed to move straight between the observed points with a constant speed. The linear interpolation seems to yield a good tradeoff between flexibility and simplicity. Anagnostopoulos et al. [5] summarize spatiotemporal data. They propose a distance-based segmentation criterion in an attempt to create minimal bounding rectangles (MBRs) that bound close data points into rectangular intervals in such a way that the original pairwise distances between all trajectories are preserved as much as possible. A variance-based hybrid variation is presented as a compromise between running time and approximation quality.

Several similarity measures are used in the literature. Anagnostopoulos et al. [5] define the distance between two trajectory segmentations at time t as the distance between the rectangles at time t, and the distance between two segmentations is the sum of the distances between them at every time instant. The distance between the trajectory MBRs is a lower bound of the original distance between the raw data, which is an essential property for guaranteeing correctness of results for most mining tasks.

In D'Auria, et al. [6] the similarity of trajectories along time is computed by analyzing the way the distance between the trajectories varies. More precisely, for each time instant they compare the positions of moving objects at that moment, thus aggregating the set of distance values. The distance between trajectories is computed as the average distance between moving objects.

In Li, et al.[7] the similarity of objects within a moving micro cluster is measured by distance on profiles of objects. Similar objects are expected to have similar initial locations and velocities.

Unsupervised classification, or clustering, derives structure from data by using objective criteria to partition the data into homogeneous groups so that the within group object similarity and the between group object dissimilarity are optimized simultaneously (Jain & Dubes 1988) [8]. Categorization and interpretation of structure are achieved by analyzing the models constructed in terms of the feature value distributions within each group. In the past, cluster analysis techniques have focused on data described by static features, i.e., value of the feature does not change, or the change is negligible, over time. In many real world applications, that cover diverse domains, such as engineering systems, human physiology, and economic and social systems, the dynamic characteristics, i.e., how a system interacts with the environment and evolves over time, are of interest. Dynamic characteristic of these systems is best described by temporal features whose values change significantly during the observation period. Clustering temporal data is inherently more complex than clustering static data because:
(i) the dimensionality of data is significantly larger, and
(ii)the model structures that describe individual cluster structures and more complex and harder to analyze and interpret.

The author proposed, assume the temporal data sequences that define the dynamic characteristics of the phenomenon under study satisfy the Markov property, and the data generation may be viewed as a probabilistic walk through a fixed set of states. When state definitions directly correspond to feature values, a Markov chain model representation of the data may be appropriate (Sebastiani et al. 1999) [9]. When the state definitions are not directly observable, or it is not feasible to define states by exhaustive enumeration of feature values (e.g., when the data is described with multiple continuous valued temporal features), a Hidden Markov model (HMM) representation is more appropriate. In this paper, the HMM methodology is applied for unsupervised learning of temporal data.

Consider the example of continuous speech recognition. Techniques have been developed (Rabiner & Jung 1993) [10] for analyzing the speech spectrum and extracting temporal features from a continuous speech signal. HMMs (usually 2-10 states per HMM) define feature sequences that correspond to subword units. Lexical and semantic rules are then employed to recognize full sentences. A lot of background knowledge in the form of phonemes and syllable for isolated word recognition, phone-like units (PLUs) for sub-word recognition and syllable and acoustic units for sentence-level recognition have been developed through years of rigorous analysis. Therefore, the HMM structures used for recognition are well-defined, and the speech recognition task is reduced to learning model parameters for efficient and robust recognition. On the other hand, when one considers problems, such as therapeutic effects of a new medical procedure or analysis of stock market data, not much is known about models that can be employed to structure, analyze, and interpret available data. To address these problems, need to develop methodologies that enable us to derive dynamic models from temporal data.

Our goal is to develop through the HMM models derived from available data, an accurate and explainable representation of system dynamics in a given domain. It is important for our clustering system to determine the best partition, i.e., number of clusters in the data, and the best model structure, i.e., the number of states in a model. The first step allows us to break up the data into homogeneous groups, and the second step provides an accurate state-based representation of the dynamic phenomena corresponding to each group. The tasks of this approach is by [11]

(i) developing an explicit HMM model size selection procedure that dynamically modifies the size of the HMMs during the clustering process, and
(ii) casting the HMM model size selection and partition selection problems in the Bayesian model selection framework. Illustrate the clustering and model interpretation process on an ecological data set in the last section of this paper.

Temporal-proximity-based [12], [13], [14] and model-based clustering algorithms [15], [16], [17] directly work on temporal data. Therefore, temporal correlation is dealt with directly during clustering analysis by means of temporal similarity measures [12], [13], [14], e.g., dynamic time warping, or dynamic models [15], [16], [17], e.g., hidden Markov model. In contrast, a representation-based algorithm converts temporal data clustering into static data clustering via a parsimonious representation that tends to capture the data dependency. Based on a temporal data representation of fixed yet lower dimensionality, any existing clustering algorithm is applicable to temporal data clustering, which is efficient in computation.

Various temporal data representations have been proposed [18], [19], [20], [21], [22], [23], [24], [25] from different perspectives. To our knowledge, there is no universal representation that perfectly characterizes all kinds of temporal data; one single representation tends to encode only those features well presented in its own representation space and inevitably incurs useful information loss. Furthermore, it is difficult to select a representation to present a given temporal data set properly without prior knowledge and a careful analysis. These problems often hinder a representation-based approach from achieving the satisfactory performance.

As an emerging area in machine learning, clustering ensemble algorithms have been recently studied from different perspective, e.g., clustering ensembles with graph partitioning [26], [28], evidence aggregation [27], [29], [30], [31], and optimization via semi definite programming [32]. The basic idea behind clustering ensemble is combining multiple partitions on the same data set to produce a consensus partition expected to be superior to that of given input partitions. Although there are few studies in theoretical justification on the clustering ensemble methodology, growing empirical evidences support such an idea, and indicate that the clustering ensemble is capable of detecting novel cluster structures [26], [27], [28], [29], [30], [31], [32]. In addition, a formal analysis on clustering ensemble reveals that under certain conditions, a proper consensus solution uncovers the intrinsic structure underlying a given data set [33]. Thus, clustering ensemble provides a generic enabling technique to use different representations jointly for temporal data clustering. Motivated by recent clustering ensemble studies [26], [27], [28] ,[29], [30], [31], [32], [33] and our success in the use of different representations to deal with difficult pattern classification tasks [34], [35], [36], [37], [38], present an approach to temporal data clustering with different representations to overcome the fundamental weakness of the representation-based temporal data clustering analysis. Our approach consists of initial clustering analysis on different representations to produce multiple partitions and clustering ensemble construction to produce a final partition by combining those partitions achieved in initial clustering analysis. While initial clustering analysis can be done by any existing clustering algorithms, propose a novel weighted clustering ensemble algorithm of a two stage reconciliation process. In our proposed algorithm, a weighting consensus function reconciles input partitions to candidate consensus partitions according to various clustering validation criteria. Then, an agreement function further reconciles those candidate consensus partitions to yield a final partition.
Previous methods on clustering spatio-temporal data have focused on grouping trajectories of similar shape. The one-dimensional version of this problem is equivalent to clustering time-series that exhibit similar movements. [39] Formalized a LCSS (Least Common Subsequence) distance, which assists the application of traditional clustering algorithms (e.g., partitioning, hierarchical, etc.) on object trajectories. In [40], regression models are used for clustering similar trajectories. Finally, [41] use traditional clustering algorithms on features of segmented time series. The problem of clustering similar trajectories or time-series is essentially different to that of finding moving clusters. The key difference is that a trajectory cluster has a constant set of objects throughout its lifetime, while the contents of a moving cluster may change over time. Another difference is that the input to a moving cluster discovery problem does not necessarily include trajectories that span the same lifetime. Finally, require the segments of trajectories that participate in a moving cluster to move similarly and to be close to each other in space.

A similar problem to the discovery of moving clusters is the identification of areas that remain dense in a long period of time. [42] proposed methods for discovering such regions in the future, given the locations and velocities of currently moving objects. This problem is different to moving clusters discovery in several aspects. First, it deals with the identification of static, as opposed to moving, dense regions. Second, a sequence of such static dense regions at consecutive timestamps does not necessarily correspond to a moving cluster, since there is no guarantee that there are common objects between regions in the sequence. Third, the problem refers to predicting dense regions in the future, as opposed to discovering them in a history of trajectories. The proposed work is also related to the incremental maintenance of clusters in data warehouses. Many researchers have studied the incremental updating of association rules for data mining. Closer to our problem are the incremental implementations of DBSCAN [43]

The author proposed that, [44] first, develop a practical temporal data clustering model by different representations via clustering ensemble learning to overcome the fundamental weakness in the representation-based temporal data clustering analysis. Next, we propose a novel weighted clustering ensemble algorithm, which not only provides an enabling technique to support our model but also can be used to combine any input partitions. Formal analysis has also been done. Finally, demonstrate the effectiveness and the efficiency of our model for a variety of temporal data clustering tasks as well as its easy-to-use nature as all internal parameters are fixed in our simulations.

III. CONCLUSION

In this paper, presented a survey on temporal data clustering approach on different representations and further propose a useful measure to understand clustering ensemble algorithms based on a formal clustering ensemble analysis [33]. In addition to that the approach does not suffer from a tedious parameter tuning process and a high computational complexity. Hence the temporal data clustering provides a promising yet easy-to-use technique for real world applications.

REFERENCES

[1] J. Kleinberg, “An Impossible Theorem for Clustering,” Advances in Neural Information Processing Systems, vol. 15, 2002.

[2] [2] E. Keogh and S. Kasetty, “On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Study,” Knowledge and Data Discovery, vol. 6, pp. 102-111, 2002.

[3] S.Y. Hwang, Y.H. Liu, J.K. Chiu, F.P. Lim (2005) "Mining Mobile Group Patterns: A Trajectory-based Approach". *Lecture Notes in Artificial Intelligence, PAKDD* 2005, v. 3518, , p 713-718

[4] D. Pfoser, "Indexing the Trajectories of Moving Objects". *IEEE Data Engineering Bulletin, 2002*

[5] 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.

[6] M. D'Auria, M. Nanni D., Pedreschi "Time-focused density-based clustering of trajectories of moving objects" To appear in JIIS Special Issue on "*Mining Spatio-Temporal Data"*, 2006

[7] Y. Li, J. Han, J. Yang, "Clustering Moving Objects". *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, p 617-622.

[8] Jain, A. K., and Dubes, D. C. 1988. *Algorithms for clustering data*. Prentice Hall.

[9] Sebastiani, P.; Ramoni, M.; Cohen, P.; Warwick, J.; and Davis, J. 1999. Discovering dynamics using bayesian clustering. In *Proceedings of the 3rd International Symposium on Intelligent Data Analysis*.

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

[11] Cen Li and Gautam Biswas,” Applying the Hidden Markov Model Methodology for Unsupervised Learning of Temporal Data

[12] E. Keogh and S. Kasetty, “On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Study,” Knowledge and Data Discovery, vol. 6, pp. 102-111, 2002.

[13] A. Jain, M. Murthy, and P. Flynn, “Data Clustering: A Review,” ACM Computing Surveys, vol. 31, pp. 264-323, 1999.

[14] R. Xu and D. Wunsch, II, “Survey of Clustering Algorithms,” IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645-678, May 2005.

[15] P. Smyth, “Probabilistic Model-Based Clustering of Multivariate and Sequential Data,” Proc. Int’l Workshop Artificial Intelligence and Statistics, pp. 299-304, 1999.

[16] K. Murphy, “Dynamic Bayesian Networks: Representation, Inference and Learning,” PhD thesis, Dept. of Computer Science, Univ. of California, Berkeley, 2002.

[17] Y. Xiong and D. Yeung, “Mixtures of ARMA Models for Model-Based Time Series Clustering,” Proc. IEEE Int’l Conf. Data Mining, pp. 717-720, 2002.

[18] N. Dimitova and F. Golshani, “Motion Recovery for Video Content Classification,” ACM Trans. Information Systems, vol. 13, pp. 408-439, 1995.

[19] W. Chen and S. Chang, “Motion Trajectory Matching of Video Objects,” Proc. SPIE/IS&T Conf. Storage and Retrieval for Media Database, 2000.

[20] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, “Fast Subsequence Matching in Time-Series Databases,” Proc. ACM SIGMOD, pp. 419-429, 1994.

[21] E. Sahouria and A. Zakhor, “Motion Indexing of Video,” Proc. IEEE Int’l Conf. Image Processing, vol. 2, pp. 526-529, 1997.

[22] C. Cheong, W. Lee, and N. Yahaya, “Wavelet-Based Temporal Clustering Analysis on Stock Time Series,” Proc. Int’l Conf. Quantitative Sciences and Its Applications, 2005.

[23] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrota, “Locally Adaptive Dimensionality Reduction for Indexing Large Scale Time Series Databases,” Proc. ACM SIGMOD, pp. 151-162, 2001.

[24] F. Bashir, “MotionSearch: Object Motion Trajectory-Based Video Database System—Index, Retrieval, Classification and Recognition,” PhD thesis, Dept. of Electrical Eng., Univ. of Illinois, Chicago, 2005.

[25] E. Keogh and M. Pazzani, “A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases,” Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, pp. 122-133, 2001.

[26] A. Strehl and J. Ghosh, “Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions,” J. Machine Learning Research, vol. 3, pp. 583-617, 2002.

[27] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, “Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data,” Machine Learning, vol. 52, pp. 91-118, 2003.

[28] X. Fern and C. Brodley, “Solving Cluster Ensemble Problem by Bipartite Graph Partitioning,” Proc. Int’l Conf. Machine Learning, pp. 36-43, 2004.

[29] A. Fred and A. Jain, “Combining Multiple Clusterings Using Evidence Accumulation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6 pp. 835-850, June 2005.

[30] N. Ailon, M. Charikar, and A. Newman, “Aggregating Inconsistent Information Ranking and Clustering,” Proc. ACM Symp. Theory of Computing (STOC ’05), pp. 684-693, 2005.

[31] A. Gionis, H. Mannila, and P. Tsaparas, “Clustering Aggregation,” ACM Trans. Knowledge Discovery from Data, vol. 1, no. 1, article no. 4, Mar. 2007.

[32] V. Singh, L. Mukerjee, J. Peng, and J. Xu, “Ensemble Clustering Using Semidefinite Programming,” Advances in Neural Information Processing Systems, pp. 1353-1360, 2007.

[33] A. Topchy, M. Law, A. Jain, and A. Fred, “Analysis of Consensus Partition in Cluster Ensemble,” Proc. IEEE Int’l Conf. Data Mining, pp. 225-232, 2004.

[34] K. Chen, L. Wang, and H. Chi, “Methods of Combining Multiple Classifiers with Different Feature Sets and Their Applications to Text-Independent Speaker Identification,” Int’l J. Pattern Recognition and Artificial Intelligence, vol. 11, pp. 417-445, 1997.

[35] K. Chen, “A Connectionist Method for Pattern Classification on Diverse Feature Sets,” Pattern Recognition Letters, vol. 19, pp. 545- 558, 1998.

[36] K. Chen and H. Chi, “A Method of Combining Multiple Probabilistic Classifiers through Soft Competition on Different Feature Sets,” Neurocomputing, vol. 20, pp. 227-252, 1998.

[37] K. Chen, “On the Use of Different Speech Representations for Speaker Modeling,” IEEE Trans. Systems, Man, and Cybernetics (Part C), vol. 35, no. 3, pp. 301-314, Aug. 2005.

[38] S. Wang and K. Chen, “Ensemble Learning with Active Data Selection for Semi-Supervised Pattern Classification,” Proc. Int’l Joint Conf. Neural Networks, 2007.

[39] Vlachos, M., Kollios, G., Gunopulos, D.: Discovering similar multidimensional trajectories In: Proc. of ICDE. (2002) 673–684

[40] Gaffney, S., Smyth, P.: Trajectory clustering with mixtures of regression models. In: Proc. of ICDM. (1999) 63–72

[41] Das, G., Lin, K.I., Mannila, H., Renganathan, G., Smyth, P.: Rule discovery from time series. In: Proc. of KDD. (1998) 16–22

[42] Hadjieleftheriou, M., Kollios, G., Gunopulos, D., Tsotras, V.J.: On-line discovery of dense areas in spatio-temporal databases. In: Proc. of SSTD. (2003)

[43] Kriegel, H.P., Kr¨ooger, P., Gotlibovich, I.: Incremental OPTICS: Efficient computation of updates in a hierarchical cluster ordering. In: Proc. of DaWaK. (2003) 224–233

[44] Yun Yang and Ke Chen,” Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 2, FEBRUARY 2011.

Biography

M. Yasodha is working as an Assistant Professor in the Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore and doing Ph.D., in Bharathiar University, Coimbatore. She has done her M.Phil., in the area of Data Mining in Bharathiar University, Coimbatore. She has done her post graduate degree MCA in Bharathiar University, Coimbatore. She has presented and published a number of papers in reputed journals. She has four years of teaching and research experience and her research interests include Data Mining, Web mining, Temporal Data mining and Text mining.

DR. P. **Ponmuthuramalingam** received his Masters Degree in Computer Science from Alagappa University, Karaikudi in 1988 and the Ph.D. in Computer Science from Bharathiar University, Coimbatore. He is working as Associate Professor and Head in Department of Computer Science, Government Arts College(Autonomous), Coimbatore. His research interest includes Text mining, Semantic Web, Network Security , Parallel Algorithms and Temporal Data Mining.