Temporal Data Mining: An Overview

Mohd. Shahnawaz, Ashish Ranjan, Mohd Danish

Abstract— To classify data mining problems and algorithms we 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. 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 can not be extracted by the general methods of data mining. Many interesting techniques of temporal data mining were proposed and shown to be useful in many applications. Since temporal data mining brings together techniques from different fields such as databases, statistics and machine learning the literature is scattered among many different sources. In this paper, we present a survey on techniques of temporal data mining.

Index Terms— Temporal Data; Temporal Data Mining; TDM Task; Temporal Sequence Mining.

I. INTRODUCTION

Due to rapid increase in storage of data, the interest in the discovery of hidden information in databases has exploded in the last decade. This discovery has mainly been focused on association rule mining, data classification and data clustering. One major problem that arises during the mining process is treating data with temporal feature i.e. the attributes related with the temporal information present in the database. This temporal attribute require a different procedure from other kinds of attributes. However, most of the data mining techniques tend to treat temporal data as an unordered collection of events, ignoring its temporal information. The aim of this paper is to present an overview of the techniques proposed to deal specifically with temporal data mining. Our objective is not only to enumerate the techniques proposed so far, but also to classify and organize them in a way that may be of help for reader looking for solutions to a specific problem. Other overviews of this area have appeared in the literature [1], but is not widely available, while the second one uses a significantly different approach to classify the existing work and is considerably less comprehensive than the present work.

Temporal data mining is an important extension of the data mining & it is non-trivial extraction of implicit, potentially useful and previously unrecorded information with an implicit or explicit temporal content, from large database. This paper is organized as Section 1 contains introduction to temporal data mining, section 2 describes different types of temporal data. Section 3 describes about the temporal data mining task. Section 4 describes mining temporal sequence. Section 6 presents paper conclusions.

Manuscript received October 06, 2011.

Mohd. Shahnawaz, Department of Computer Science & Engineering, Infinity Management & Engineering College, Sagar, India, +91-9755624180, (e-mail: shahnawaznbd@gmail.com).

Ashish Ranjan, Department of IT, OICL, New Delhi, India, +91-9212536698, (e-mail: ashishkumar.ranjan@gmail.com).

Mohd Danish, Department of IT, NIC, New Delhi, India, +91-9990261541, (e-mail:danishnbdr@gmail.com).

II. TEMPORAL DATA

A. Static Data

This type of temporal data has Zero temporality i.e. static data are free from any temporal reference. Inferences that can be derived from this data are free from any temporality.

B. Sequences

Sequences are ordered sequence of the events or transaction. Though there may not be any explicit reference to time, there exists a sort of qualitative temporal relationship between data items. In any transactional data if a transaction appears before another, it implies that the former transaction has occurred before the latter. In this type of temporal data have the temporal relationship like before, after, during, meet and overlap etc. Such relationships are called qualitative relationship between time events.

C. Time Stamped

This category of the temporal data has explicit time related information. Relationship can be quantitative i.e. we can find the exact temporal distance between data element. The inference made through this type of data may be temporal or non temporal.

D. Time Series

Time series data is special case of the time stamped data. In time series data events have uniform distance on the time scale.

E. Fully Temporal

Data of this category is fully time dependent. The inferences are also strictly temporal.

III. TEMPORAL DATA MINING TASK

The possible objectives of data mining, which are often called *tasks* of data mining [1] can be classified into some broad groups. For the case of temporal data mining, these tasks may be grouped as Association, Prediction, Classification, Clustering, Characterization, Search & retrieval, Pattern discovery, Trend Analysis and Sequence Analysis

This categorization is neither unique nor exhaustive, the only objective being to facilitate an easy discussion of the numerous techniques in the field.

First four have been investigated extensively in traditional time series analysis and pattern recognition. Algorithms for pattern discovery in large databases however are of more recent origin and are mostly discussed only in data mining literature. In this section, we provide an overview of first eight temporal data mining techniques. In the next section, we provide a detailed description of knowledge discovery techniques for sequential data.

A. Association

The discovery of relevant *association rules* is one of the most important methods used to perform data mining on transactional databases. An effective algorithm to discover association rules is the *apriori* [5]. Association rule discovery is an important task in data mining in which we extract the relation among the attribute on the basis of support and confidence. The association rule discovery can be extended to temporal association. One of the most common approaches to mining frequent patterns is the *apriori* method and when a transactional database represented as a set of sequences of transactions performed by one entity is used, the manipulation of temporal sequences requires that some adaptations be made to the *apriori* algorithm. The most important modification is on the notion of *support*: support is now the fraction of entities, which had consumed the *itemsets* in any of their possible transactions, i.e. an entity could only contribute one time to increment the support of each *itemset*, beside it could had consumed that *itemset* several times. After identifying the large *itemsets*, the *itemsets* with support greater than the minimum support allowed, they are translated to an integer, and each sequence is transformed in a new sequence, whose elements are the large *itemsets* of the previous one. The next step is to find the large sequences. For achieve this, the algorithm acts iteratively as *apriori*: first it generates the candidate sequences and then it chooses the large sequences from the candidate ones, until there are no candidates.

Candidate generation is the most costly operations in apriori-based approaches. A proposal to frequent pattern mining states that it is possible to find frequent patterns avoiding the candidate generation test [10].

In general association rule discovery the notion of a typical rule X =>Y (which states if X occurs then Y occurs),it can be extended to a rule with a new meaning: X $=>^{T}$ Y (which states: if X occurs then Y will occur within time T) [4]. Stating a rule in this new form, enable us to controlling the impact of the occurrence of an event to the other event occurrence, within a specific time interval.

A *cyclic rule* is one that occurs at regular time intervals, for example at every second Saturday of a month govt offices remain closed. In order to discover this rules, An efficient approach to discover cyclic rules consists the process: first discover the cyclic large *itemsets* and then generate the rules. Extension to this method consists in allowing the existence of different time units, such as days, weeks or months, and is achieved by defining calendar algebra to define and manipulate groups of time intervals. Rules discovered of such type are called *calendared association rules.*

Weighted temporal association rules can be extracted, it is seen that data have different importance on different time i.e. if we have a database of 50 years, it may be the data of last decade reflect more importance in comparison to the first decade. To extract the association rule in this situation we give more weightage to the data that is more important, TWM algorithm for weighted temporal association rule is given in [2]. Time weighted Miner show the better result in respect of accuracy. Algorithm for the time weighted mining is given in [2], here the whole algorithm is divided into 4 procedures. Each procedure perform specific task.

B. Prediction

Prediction has a versatile significance in the data mining. It is the forecasting for future on the basis of past. The task of time-series prediction has to do with forecasting (typically) future values of the time series based on its past samples. For this purpose, we need to build a predictive model for the data. Probably the earliest example of such a model is due to Yule way back in [9]. The autoregressive family of models can be used to predict a future value as a linear combination of earlier sample values, provided the time series is stationary. Linear non stationary models like ARMA models have also been found useful in many economic and industrial applications where some suitable variant of the process can be assumed to be stationary. Another popular work-around for non stationary is to assume that the time series is piece-wise stationary. The series is then broken down into smaller pieces called as “frames” within each of which, the stationary condition can be assumed to hold and then separate models are learnt for each frame. In addition to this standard ARMA family of models, there are many nonlinear models for time series prediction e. g., neural networks have good for nonlinear modeling of time series data [4]. The prediction problem for symbolic sequences has been addressed in Artificial Intelligence research. Consider various rule models such as disjunctive normal form model, periodic rule model etc. Based on these models sequence-generating rules are obtained that state some properties that constrain which symbol can appear next in the sequence.

In many cases, prediction may be formulated as classification, association rule finding or clustering problems. Generative models can also be used effectively to predict the evolution of time series. In spite of prediction problems have some specific characteristics that differentiate them from other problems. A vast literature exists on prediction of time series, in a variety of domains [11]. But we have failed to find in the data mining literature significant applications that involve prediction of time series and that do not fall into any of the previously described categories. Granted, several authors have presented work that aims specifically obtaining algorithms that can be used to predict the evolution of time series. In the particular domain of prediction, care must be taken with the domain where prediction is to be applied [11]. Prediction gains the importance in various fields like medical, finance, environmental & engineering with an exponential rate.

C. Classification

In classification one classifies the unknown set of attributes in any one of the predefined class [1].

In temporal classification, each temporal sequence presented in the database is assumed to belong to one of the predefined classes or categories and our goal is to automatically determine the corresponding category/class for the given input temporal set of attributes. There are many examples of sequence classification applications, like Handwriting recognition speech recognition, gesture recognition, demarcating gene and non-gene regions in a genome sequence, on-line signature verification, etc. The task of a speech recognition system is to transcribe speech signals into their corresponding textual representations.

There are some pattern recognition applications in which even images are viewed as sequences. For example, images of handwritten words are sometimes regarded as a sequence of pixel columns or segments proceeding from left to right in the image. Recognizing the words in such sequences is another interesting sequence classification application. In on-line handwritten word recognition and signature verification applications, the input is a sequence of pixel coordinates drawn by the user on a digitized tablet and the task is to assign a pattern label to each sequence. As is the case with any standard pattern recognition framework, in these applications also, there is a feature extraction step that precedes the classification step. For example, in speech recognition, the standard analysis method is to divide the speech pattern into "frames" and apply a feature extraction method on each frame. In gesture recognition, motion trajectories and other object-related image features are obtained from the video sequence. The feature extraction step in sequence recognition applications typically generates, for each pattern such as a video sequence or speech utterance, a sequence of feature vectors that must then be subjected to a classification step. Over the years, sequence classification applications have seen the use of both pattern based as well as model based methods. In a typical pattern based method, prototype feature sequences are available for each class. The classifier then searches over the space of all prototypes, for the one that is closest or most similar to the feature sequence of the new pattern. Typically, the prototypes and the given features vector sequences are of different lengths. Thus, in order to score each prototype sequence against the given pattern, sequence aligning methods like Dynamic Time Warping are needed. Another popular class of sequence recognition techniques is a model based method that use Hidden Markov Models (HMMs) [4].

D. Clustering

Clustering techniques are used to divide the data in the groups on the basis of similarity measure. There exist several clustering technique algorithm like K-means, K- medoids etc [5]. Clustering of sequences or time series is concerned with grouping a collection of time series or sequences based on their similarity. Clustering is of particular interest in temporal data mining since it provides an attractive mechanism to automatically find some structure in large data sets that would be otherwise difficult to summarize or visualize. There are many applications where a time series clustering activity is relevant e.g. web activity logs, clusters can indicate navigation patterns of different user groups. Another example could be clustering of biological sequences like proteins or nucleic acids so that sequences within a group have similar functional properties. There are a variety of methods for clustering sequences. At one end of the spectrum, we have model based sequence clustering methods. Learning mixture models, for example, constitute a big class of model based clustering methods. In case of time series clustering, mixtures of, e. g., ARMA models or Hidden Markov Models are in popular use. The other broad class in sequence clustering uses pattern alignment-based scoring or similarity measures to compare sequences. Some techniques use both model based as well as alignment based methods.

E. Characterization

Characterization is a summarization of the general characteristics or features of a target class of data. Characterization can be extended to temporal data. An interesting experiment would be extending to the concept of decision tree construction on temporal attributes e.g. a rule could be: The first case of filarial is normally reported after the first pre-monsoon rain and during the month of May-August. The output of characterization can be represented in various forms e.g. pie charts, bar charts, curves, multidimensional data cubes [1].

F. Search and Retrieval

The problem of searching is concerned with efficiently locating subsequences often referred to as queries in large archives of sequences or sometimes in a single long sequence. Query-based searches have been extensively studied in language and automata theory [10]. However the problem of efficiently locating exact matches of substrings is well solved, the situation is different when looking for *approximate* matches. In typical data mining applications like content-based retrieval, it is approx matching that we are more interested in. In content-based retrieval, a query is presented to the system in the form of a sequence. The task is to search a typically large data base of sequential data and retrieve from it sequences or subsequences *similar* to the query sequence e.g in large music database the user could a query and the system should retrieve tracks that resemble it. In all such problems there is a need to quantify the extent of similarity between any two sequences. Given two sequences of equal length we can define a measure of similarity by considering distances between corresponding elements of the two sequences. When the sequence elements are feature vectors standard metrics such as Euclidean distance may be used for measuring similarity between two elements. However, sometimes the Euclidean norm is unable to capture subjective similarities effectively. For example, in speech or audio signals, similar sounding patterns may give feature vectors that have large Euclidean distances and vice versa [7]. Idea in these measures is to perform the comparison in spectral domain by emphasizing differences in those spectral components that are perceptually more relevant. Similarity measures based on other transforms have been discovered as well e.g. in [6] present a comparison of DFT and DWT-based similarity searches. When the sequences consist of symbolic data we have to define dissimilarity between every pair of symbols which in general is determined by the application. Choice of similarity or distortion measure is only one aspect of the sequence matching problem.

In problem of determining similarity between pairs of sequences, the sequences may be of different lengths or size. In such cases, it is not possible to simply calculate distances between corresponding elements of the sequences. This brings us to the second aspect of sequence matching, namely, sequence alignment. Essentially we need to properly insert 'gaps' in the two sequences or decide which should be corresponding elements in the two sequences. Time warping methods have been used for sequence classification and matching for many years. In speech applications, Dynamic Time Warping (DTW) is a systematic and efficient method based on dynamic programming that identifies which correspondence among feature vectors of two sequences is best when scoring the similarity between them. In recent times, there are many situations in which such symbolic sequence matching problems find applications. For example, many biological sequences such as genes, proteins, etc.

G. Pattern Discovery

Unlike in search and retrieval applications, in pattern discovery there is no specific query in hand with which to search the database [5]. The objective is simply to unearth all *patterns of interest*. It is worthwhile to note at this point that whereas the other temporal data mining tasks discussed earlier in (i. e. sequence prediction, classification, clustering and matching) had their origins in other disciplines like estimation theory, machine learning or pattern recognition; the pattern discovery task has its origins in data mining itself. In that sense, pattern discovery, with its exploratory and unsupervised nature of operation, is something of a sole preserve of data mining. For this reason, this review lays particular emphasis on the temporal data mining task of pattern discovery.

H. Trend Analysis

Trend analysis in temporal database referred to change in attribute due to the change in time [1]. The analysis of one or more time series of continuous data may show similar trends, i.e. similar shape across the time axis. For example," the deployment of the data mining system is increasingly becoming popular in the banking industry. These types of analysis are of higher level than the earlier ones. Here we are trying to find the relationships of change in one or more static attributes, with respect to changes in the temporal attributes. Trend analysis is very useful for the decision support system & decision maker

IV. Mining Temporal Sequences

The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequence of events. An efficient approach to mining casual relations is sequence mining. The discovery of relations between sequences of events involves mainly three steps [11]:

Representation and modeling: In this sequence of the temporal data are transformed into a suitable form.

1) Similarity Measure: Definition of similarity measures between sequences.
2) Mining Operation: application of models and representations to the actual mining problems.

Other authors have used a different approach to classify data mining problems and algorithms they used three dimensions: data type, mining operations and type of timing information. Although both approaches are equally valid, we preferred to use representation, similarity and operations, since it provided a more comprehensive and novel view of the field depending on the nature of the event sequence, the approaches to solve the problem may be quite different. A sequence composed by a series of nominal symbols from a particular alphabet is usually called a *temporal sequence* and a sequence of continuous, real-valued elements, is known as a *time series*. Time series or, more generally, temporal sequences, appear naturally in a variety of different domains engineering, scientific research, finance and medicine. In engineering matters, they usually arise with either sensor-based monitoring, such as telecommunications control or log-based systems monitoring. In finance, applications on the analysis of product sales or inventory consumptions are of great importance to business planning see for instance [10]. Another very common application in finance is the prediction of the evolution of financial data (e.g. [1], [5], [8], [11]). In healthcare, temporal sequences are a reality for decades; with data originated by complex data acquisition systems like ECG's [4] or even with simple ones like measuring the patient temperature or treatments effectiveness [12]. In the last years development of medical informatics, the amount of data has increased considerably and more than ever, the need to react in real-time to any change in the patient behavior is crucial.

Sequence mining problem: The most general form of the sequence mining problem [5] can be stated as, Let $\sum=\{i_1, i_2, \ldots\ldots\ldots i_m\}$ be a set of distinct items comprising the alphabet. An event is a non-empty, disordered collection of item. Without any loss of generality, we write the items in an event in some predefined order. An event is denoted $(i_1, i_2, \ldots\ldots\ldots i_k)$ where i_j is an item in $\sum$. Any event that is given as input will also be called a transaction is known to us prior to the process and an event is generated during the algorithm, we can use both terms interchangeably if there is no ambiguity. Following algorithms has been discovered for sequence mining. In the next section a brief description is given.

A. GSP Algorithm

As mostly algorithms for sequence mining are based on Apriori algorithm, GSP algorithm is also based on apriori algorithm. One way to use the level wise paradigm is to first discover all frequent items in a level wise fashion. GSP Algorithm makes multiple passes over the database. In the first pass, all single items (1- sequences) are counted. From the frequent items a set of candidate 2 sequences are formed, and another pass is made to gather their support. The frequent 2 sequence are used to generate the candidate 3-Sequences, and this process is repeated until no more frequent sequences are found. There are two main steps in the algorithm

GSP Algorithm

F_1=the set of frequent 1-sequence

K=2,

do while $F_{k-1} \neq \emptyset$;

generate candidate sets C_k

for all input sequences s in the database D do

increment count of all a in ck if s support a

F_k={ *a* ε C_k *i.e. its frequency exceeds the threshold* }

K=k+1

Set of all frequent sequences is the union of all F_k*s*

End do

Candidate generation: By joining frequent (k-1) frequent sequence with itself. A pruning phase eliminates any sequence, at least one of whose sequences is not frequent.

Support Counting: Normally, a hash tree based search is employed for efficient support counting. Finally non-maximal frequent sequences are removed.

B. SPADE

SPADE is based on the level wise paradigm. In the year 2000, Zaki proposed SPADE (Sequential Pattern Discovery using Equivalance Claasses) for discovering the set a of all frequent sequences [5]. SPADE draws its motivational from the level wise algorithm (GSP) and vertical mining. Initially, SPADE uses a vertically structured database. In this structure a list is associated with each item. This list include the sequences-id and the transaction-id containing the item. Using this structure, all frequent sequences can be enumerated via simple temporal joins. The SPADE starts from the bottom-most (the least element) of the lattice and gradually works in a bottom-up way to generate all frequent sequences. It maintains the vertical structure as it proceeds from the least element to the maximal elements.

C. SPIRIT

Garofarakasi, Rajeev Rastogi and K Shim, in 1999, design SPIRIT (Sequential Pattern Mining with Regular Expression Constraints) is a sequence mining algorithm [3] that allows the user to specify the constraints in the form of regular expressions. Constrained sequence mining problem is used to discover frequent sequences from D, such that each of these frequent subsequences satisfy the user specified constraint set C.

D. WUM

Web Utilization Miner is also a constrained sequences miner. Its primary purpose is to analyze the navigational behavior of users on the web, but it is appropriate for sequential pattern discovery in any type of log. The main idea is to apply mining on disk-resident condensed data instead of the original log events.

V. CONCLUSION

Knowledge discovery in databases (KDD) is a process of extraction meaningful information from the databases. Temporal data mining task is so crucial compare to the ordinary mining, cause is handling time related features of the dataset. We have presented a comprehensive overview of techniques for the mining of temporal sequences. This domain has relationships with many other areas of knowledge, and an exhaustive survey that includes all relevant contributions. We have surveyed and classified a significant temporal data mining techniques, taking into consideration the representations they utilize, the similarity measures they propose and the applications they have been applied to. This survey has shown that a significant number of algorithms exist for the TDM. We can believe that the field will be significantly enriched if knowledge from other sources is incorporated into data mining algorithms and applications.

REFERENCES

1. Han J, Kamber M 2001 Data mining: Concepts and techniques (San Fransisco, CA: Morgan Kauffmann)
2. Kanak Saxena, Efficient Mining Weighted Temporal Association Rules. 2009 World Congress on Computer Science and Information Engineering pp 421-425 IEEE Computer Society.
3. Garofarakasi, Rajeev Rastogi and K Shim, in 1999, design SPIRIT: Sequential Pattern Mining with Regular Expression Constraints. Bell Labs Tech. Memorndum BL0112370-990223-03TM, February 1999.
4. SRIVATSAN LAXMAN and P S SASTRY, A survey of temporal data mining, Sadhana Vol. 31, Part 2, April 2006, pp. 173–198.
5. A.K. Pujari 2007, Data Mining Techniques, University Press ISBN 8173713804.
6. WuY-L, AgrawalD, AbbadiAE 2000Acomparison of DFT and DWT based similarity search in time series databases. In Proc. Ninth Int. Conf. on Information and Knowledge Management, McLean,VA, pp 488–495
7. Gray R M, Buzo A, Gray Jr. A H, Matsuyama Y 1980 Distortion measures for speech processing. IEEE Trans. Acoust., Speech Signal Process. 28: 367–376
8. Haselsteiner E, Pfurtscheller G 2000 Using time-dependent neural networks for EEG classification. IEEE Trans. Rehab. Eng. 8: 457–463.
9. Yule G 1927 On a method of investigating periodicity in distributed series with special reference to Wolfer's sunspot numbers. Philos. Trans. R. Soc. London A226
10. Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns without Candidate Generation. ACM SIGMOD Int. Conf. on Management of Data (2000) 1-12.
11. Claudia m M. Antunes and Arlindo L. Oliveira: Temporal Data Mining: an overview. Lecture Notes in Computer Science.
12. Coiera, E.: The Role of Knowledge Based Systems in Clinical Practice. In Barahona, P., Christensen, J.: Knowledge and Decisions in Health Telematics – The Next Decada. IOS Press Amsterdam (1994) 199-203

AUTHORS PROFILE

Mohd. Shahnawaz receive his bachelor degree in Engineering and Technology in computer science discipline after that he receive master degree as M.Tech from Rajive Gandhi Proudhogiki Vishwavidyalya Bhopal, He is currently working as Assistant Professor in the department of computer science & engineering at Infinity Management & Engineering College, Sagar M.P.

Ashish Ranjan receive his bachelor degree in Engineering and Technology in computer science discipline after that he receive master degree as M.Tech from Rajive Gandhi Proudhogiki Vishwavidyalya Bhopal, He is currently working as Administrative Officer (IT) in Oriental Insurance Corporation of India.

Mohd. Danish receive his bachelor degree in computer science discipline after that he receive master degree as M.C.A from Jamia Humdard University, New Delhi. Currently He is working in National Informatics Centre, New Delhi.