31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Methods and Tools for Mining Multivariate Temporal Data in Clinical and Biomedical Applications

Riccardo Bellazzi, Lucia Sacchi, and Stefano Concaro

Abstract—**Temporal data mining is becoming an important tool for health care providers and decision makers. The capability of handling and analyzing complex multivariate data may allow to extract useful information coming from the day-by-day activity of health care organizations as well as from patients monitoring. In this paper we review the main approaches presented in the literature to mine biomedical time sequences and we present a novel approach able to deal with "point-like" and "interval-like" events. The methods is described and the results obtained on two clinical data sets are shown.**

I. Introduction

Health care organizations (HCO) have nowadays evolved into data-rich, complex enterprises. The large amounts of data collected during the HCO day-by-day activities may be of great utility to health care providers and decision makers, provided that suitable data analysis methods and tools are available to them. An example is given by the retrospective analysis of the administrative records related to healthcare "events", such as patients' hospital admissions and discharges, drug prescriptions and lab exams. In order to analyze such data, it is often necessary to implement large data warehouses and to design ad-hoc data mining algorithms. When this happens, though, this kind of data become an invaluable source of information to assess health care activities, to implement organizational learning strategies and to design personalized-medicine interventions [1].

One of the most important needs of data mining algorithms to analyze HCO data is to take into account the temporal dimension. As a matter of fact, clinical monitoring data, administrative information, molecular medicine experiments very often deal with time. For this reason, in this paper we are interested to the algorithms which have been designed to extract temporal information from time series and time sequences of clinical and biomedical data. The research field aimed at extending traditional Data Mining methods to explicitly handle temporal reasoning and to incorporate the recognition of temporal features is called Temporal Data Mining (TDM) [2,3].

The first class of methods that have been proposed in the literature are know as *sequential pattern mining*. They extract *frequent* sequential patterns that occur in databases of time-stamped data. Following the seminal algorithm reported in [4], several papers dealt with the improvement of the sequential pattern search strategy in order to maximize the computational performance [5,6,7,8]. Usually, sequential pattern mining is related to events that can be represented by single time points, such lab exams or economic transactions [9,10,11]. For this reason, several methods have been proposed to take into account "events with duration", in which sequential patterns are derived from interval-based data [12,13,14]. Health care, however, is characterized by events of heterogeneous temporal nature with respect to the temporal granularity used to collect the data [15]. For example, considering a temporal granularity of one day, we may collect events with duration (interval-like), such as hospitalizations and drug delivery and "instantaneous" events (point-like), such as day-hospital admissions and ambulatory visits.

In this paper we present the results obtained by an original method proposed by the authors, which is able to deal with both interval-like and point-like events. The method extracts a set of Temporal Association Rules and is based on the knowledge-based temporal abstraction paradigm. The proposed approach is an example of a general methodology that can be conveniently applied to a variety of health care, biomedical and clinical problems, including patient's monitoring, health care administrative data management and molecular biology.

II. Methods

A. Patterns and temporal abstractions

The method proposed in this paper is aimed at finding "useful patterns" in the temporal data. Those patterns will be then mined in terms of Temporal Association Rules, as described in subsection II.B.

The notion of *pattern* is intuitively related to the representation of a property or a behavior of interest that may occur in the data; such an abstract property is in general described in qualitative terms. A simple pattern may be for example a decreasing trend of a variable, while a more complex one might be a saturation behavior. Usually a pattern has a duration and is therefore associated with a time interval.

The framework of Temporal Abstractions (TAs) [12,13,14,16] is a suitable way to formalize the intuitive

This work was supported in part by the FIRB Project ITALBIONET, Italian Bioinformatics Network, funded by the Italian Ministry of Education, University and Research. SC is supported by the Azienda Sanitaria Locale of Pavia, Italy.

The authors are with the Dipartimento di Informatica e Sistemistica, University of Pavia, via Ferrata 1, 27100 Pavia.

RB is the corresponding author: tel: +39-0382-985720; fax: +39-0382-985720; e-mail: riccardo.bellazzi@unipv.it.

definition of pattern given above. TAs represent time series through time intervals. Following the data model proposed in [16], raw temporal data are represented as time-stamped entities, called *events*, while their abstract representation is given by TAs as a sequence of intervals, called *episodes*. we will denote a generic episode as e ≡ (e.start,e.end), where e.start and e.end are respectively the starting and the ending point of the interval. A qualitative label, corresponding to a specific behavior of interest, is then used to characterize each episode. The algorithms which are devoted to the generation of episodes from events (or from other episodes) are known as TA *mechanisms*.

TAs are usually classified into two main categories:

- *Basic* TAs, searched by mechanisms that abstract time-stamped data into intervals (input data are events and outputs are episodes);
- *Complex* TAs, searched by mechanisms that abstract intervals into other intervals (input and output data are episodes).

Among Basic TAs, we will herein deal with *state* TAs, which are used to detect qualitative patterns corresponding for example to low, high or normal values in numerical or symbolic time series, and *trend* TAs, used to capture increasing, decreasing or stationary courses in numerical time series.

Complex TAs correspond to intervals over which specific temporal relationships between basic or other complex TAs hold; such temporal relationships are usually identified through Allen's temporal operators [17]. TAs therefore allow to associate a qualitative property of interest, which we will refer to as the *pattern*, to the set of episodes where this behavior is verified in the data.

TA mechanisms are algorithms which depend on a set of parameters specified by the user or by the data analyst according to the application of interest. Examples of these parameters are the minimum slope to trigger the detection of a trend TA, or the threshold values needed to map the quantitative values of a variable to a set of qualitative state labels (e.g. high or low).

In case the data are simply "point-like" events, they may be treated as if they are TAs with no duration.

B. *Temporal Association Rules*

Once the data are represented through TAs, it is possible to apply algorithms to derive temporal relationships between them. In particular we want to extract frequent temporal relationships in a database reporting the sequences of TAs coming from different subjects.

To this end, we have defined a new method to find Temporal Association Rules (TAR) in a set of TA transformed data. A TAR is a relationship where a temporal operator holds between an antecedent, made of one or more patterns, and a consequent, made of a single pattern. The available relationships can be defined through the PRECEDES temporal operator [18]. Given two episodes *e1* and *e2*, the PRECEDES relationship holds (*e1 PRECEDES e2*) if *e1.start* ≤ *e2.start* and *e1.end* ≤ *e2.end*. The PRECEDES operator extends the Allen's temporal algebra operators including also "point-like" events [19]. The PRECEDES operator is constrained by three design parameters: the *left shift* (LS), defined as the maximum allowed distance between e1.start and e2.start, the *gap* (G), defined as the maximum allowed distance between e1.end and e2.start, and the *right shift* (RS), defined as the maximum allowed distance between e1.end and e2.end. All the operators were implemented in order to handle both "interval-like" and "point-like" events.

The method for TAR extraction is based on an Apriori-like strategy, and it is made up of three steps:

1. iterative selection of a variable as consequent of the rule;
2. extraction of the *basic-set* of rules, that is the whole set of rules with single cardinality in the antecedent;
3. extraction of *complex rules*, defined as rules with antecedent of multiple cardinality K obtained through the intersection of the episodes of the antecedents of the rules of cardinality K-1.

In addition to the implementation presented in [14], the current version of the algorithm offers the optional opportunity to select specific target rule types, defining the classes of the variables allowed for the antecedent and the consequent selection, respectively.

The TAR extraction algorithm looks for frequent associations, where the frequency is computed through the *support* of a rule. The support is defined as the proportion of subjects for which the rule is verified over the total number of subjects involved in the study. Those who verify the rule are the subjects in which there are episodes with low frequency but long lasting in time, or short episodes (as for events without a duration) but with high frequency.

The quality of the rules is evaluated through its *confidence*. The confidence is defined as the proportion of subjects satisfying the rule over the subjects satisfying the antecedent. According to this definition, the confidence represents the probability to find one episode of the consequent satisfying the temporal relationship given that one episode of the antecedent occurs.

III. Applications and Results

A. *Analysis of Dialysis monitoring data*

As a first example of the application of the TAR extraction, we show the results obtained in the context of hemodialysis (HD) patient monitoring [14,18]. The problem was related to the retrospective evaluation of the HD data for quality assessment of patients' treatment. We analyzed 36 patients coming from the Limited Assistance Dialysis Center of Mede, remotely managed by the Dialysis Unit of the A.O. of Vigevano, Italy.

Among other variables, we studied the temporal behavior of the systolic pressure (*SP*), the diastolic pressure (*DP*) and the heart rate (*HR*). The measurements have been performed through a digital sphygmomanometer. In this case we have looked for patterns related to hypotension or hypertension episodes, which usually involve a counter-regulation of HR.

Trend TAs have been applied to extract increasing, decreasing, steady patterns; moreover, complex abstractions were used to detect increasing/decreasing-like behaviors.

Trends are detected through a sliding window algorithm, fixing a threshold of the 5% on the slope change to detect the patterns of increase and decrease. This choice is justified by the fact that a change in pressure values or in heart rate of 5 out of 100 units is clinically significant; such variation can be properly detected by considering the precision of the measurement instrument.

From a computational viewpoint, the rule extraction algorithm works by searching opposite patterns between SP, DP and HR. Considering an average treatment duration of four hours, with measurements taken every five minutes, we set *LS*=*RS*=40 and *G*=30.

Table I shows the results obtained fixing a threshold for the confidence, $Conf \geq 0.5$, and for the support, $Sup \geq 0.1$.

TABLE I
The temporal patterns extracted on a data set of 36 HD patients.

Antecedent		Consequent		Confidence	Support
Variable	Pattern	Variable	Pattern		
SP DP	╲_╱ ╲_╱	HR	╱‾╲	0.580	0.119
HR	╲_╱	DP	╱‾╲	0.623	0.192
HR	╲_╱	SP	╱‾╲	0.622	0.199

Interesting rules which describe relationships between complex patterns involving one or more variables were detected. The first rule extracts a contemporaneous pattern of *SP* and *DP*, in which a down and up pattern is followed by an up and down pattern of *HR*. From the clinical viewpoint, this rule highlights the occurrence of hypotension episodes taking place during dialysis treatments: when arterial blood pressure decreases, the organism reacts with an increase in the HR, which then goes back to normal values as soon as blood pressure increases. Other similar rules, which relate *HR* with *SP* and *DP* were also found. These episodes are clinically relevant, since they correspond to the patient's response to blood pressure instability.

Looking at the small values obtained for confidence and support in this case, an interesting clinical issue that may rise is whether it is possible to find a group of patients particularly prone to hypo or hypertensive episodes during HD. To tackle this problem, we ran the rule extraction algorithm on each patient separately and then evaluated confidence and support of the obtained rules. We thus identified a group of 10 patients showing an high number of precedence rules involving blood pressure and heart rate, in which the variables present an opposite pattern. The TAR mining algorithm was then run on those patients and the results are shown in Table II ($Conf \geq 0.7$, $Sup \geq 0.1$). In this table it is possible to observe a relevant improvement in the confidence and support of the first rule. In this application, the proposed approach clearly shows its capability of answering to clinically relevant questions, together with extracting useful information from the data.

TABLE II
The temporal patterns extracted on a subset of 10 HD patients showing frequent hypo- and hypertensive episodes.

Antecedent		Consequent		Confidence	Support
Variable	Pattern	Variable	Pattern		
SP DP	╲_╱ ╲_╱	HR	╱‾╲	0.728	0.154
HR	╲_╱	SP	╱‾╲	0.734	0.342

B. *Mining administrative and clinical data*

In the context of the Italian National Healthcare System the Regional Healthcare Agencies (ASL) have a central role in the coordination of the care delivery process to the assisted population. The ASL of Pavia has a large datawarehouse, which records all the main healthcare expenditures of the assisted population. Such administrative data refers to: hospital admissions, drug prescriptions (provided through the ATC[1] code) and ambulatory visits. In the period between January 2007 and October 2008 (22 months) the ASL was able to collect, together with administrative data, also the clinical information of a selected subgroup of about 1000 diabetic patients: a total of about 5000 patients visits were recorded.

In order to merge administrative and clinical data in a uniform representation, the clinical data underwent to a pre-processing procedure, where both basic state and trend TAs were exploited. *State detection* was applied to discretize continuous values in state intervals, determined on the basis of physiological thresholds suggested by an expert clinician. *Trend detection* was used to describe the change of the state with respect to the previous visit, thus allowing the definition of *Increasing*, *Steady* or *Decreasing* intervals for each variable.

In our analysis we selected a specific rule template, where the antecedent selection is limited to state abstractions and drug prescriptions, while the consequent selection is limited to trend abstractions. This allows us to evaluate whether a combination of drug prescriptions, in conjunction with a clinical status, frequently show a temporal association with a subsequent variation of the clinical conditions in the considered population [20,21].

We herein present the results obtained by setting the thresholds for support and confidence to $Sup \geq 0.01$ and $Conf \geq 0.3$. The selected temporal operator was BEFORE.

[1] Anatomical Therapeutic Chemical classification system

One of the TARs obtained after the application of the algorithm is reported below:

BMI overweight (between 25 and 30) *AND Glycaemia regular* (between 75 and 110 mg/dl) *AND HbA1c high* (between 7-8%) *AND Anti-Hypertensive Therapy: Yes*
PRECEDES
Glycaemia Increasing
(support: 0.013 confidence: 0.56, Gap: 1 visit).

The rule illustrates that a subject found to be overweight, under anti-hypertensive therapy, showing a normal glycemic value and high glycated hemoglobin has a probability of 56% to have an increase in glycaemia within the following visit.

In total we have extracted more than a hundred rules, which have been ranked on the basis of their confidence, support and clinical interest. After filtering, the rules analyzed by the clinical experts were 24. All of them were meaningful and interesting either for their clinical relevance or because they were potentially useful to plan health care interventions.

IV. Software Tools

Although temporal data mining is a ripe technology, very few software tools are available, especially in the biomedical and clinical fields. Some tools have been devoted to the statistical analysis of time series, with particular reference to clustering [22]. The development of sequential pattern mining algorithm was followed by the release of the corresponding tools [6,10,11], but such tools are limited to the analysis of temporal sequences ("point-like" events) and still require a data preparation step which is not straightforward.

As regards TAs, the main tools are represented by the KNAVE/IDAN systems, which are designed as temporal decision support tools [23]. A set of open-source java classes for temporal abstractions have been made available by the Tempo project [24]. Finally, we have implemented the methods presented in this paper in a Matlab-based tool, which will be soon available.

V. Conclusion

Temporal Data Mining is nowadays moving from pure research to applications. Algorithms are now almost ready for exploitation in health care and biomedical data analysis. To complete the transition, new user friendly tools should be designed and deployed; it is likely that in the near future such tools will start to be available.

Acknowledgment

The authors acknowledge Carlo Cerra, MD, Roberto Bellazzi, MD, Blaz Zupan, PhD, for their collaboration. We also thank the "Azienda Sanitaria Locale di Pavia" and the "Unità Operativa di Nefrologia e Dialisi di Vigevano" for their support.

References

[1] M. Stefanelli, "The socio-organizational age of artificial intelligence in medicine", *Artif Intell Med 23* (2001) 25-47.

[2] J.F. Roddick and M. Spiliopoulou, "A Survey of Temporal Knowledge Discovery Paradigms and Methods", *IEEE Transactions on Knowledge and Data Engineering* 14 (2002) 750-767.

[3] A.R. Post and J.H. Harrison, "Temporal data mining", *Clin Lab Med 28* (2008) 83-100.

[4] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", *20th International Conference on Very Large Data Bases* (1994) 487-499.

[5] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements", *5th International Conference on Extending Database Technology: Advances in Database Technology* (1996) 3-17.

[6] J. Pei, J. Han, B.M. Asl, H. Pinto, Q. Chen, U. Dayal and M. Hsu, "PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth", *17th International Conference on Data Engineering* (2001) 215-224.

[7] R.T. Ng, L.V.S. Lakshmanan, J. Han and A. Pang, "Exploratory mining and pruning optimizations of constrained associations rules", *In Proc. 1998 ACMSIGMOD Int. Conf. Management of Data (SIGMOD'98)* (1998) 13-24.

[8] R.J. Bayardo, R. Agrawal and D. Gunopulos, "Constraint-Based Rule Mining in Large, Dense Databases", *15th International Conference on Data Engineering* (1999) 188-197.

[9] R. Agrawal and R. Srikant, "Mining Sequential Patterns", *11th International Conference on Data Engineering* (1995) 3-14.

[10] M.J. Zaki, "SPADE: An Efficient Algorithm for Mining Frequent Sequences", *Machine Learning* 42 (2001) 31-60.

[11] J. Ayres, J. Flannick, J. Gehrke and T. Yiu, "Sequential PAttern mining using a bitmap representation", *8th ACM SIGKDD international conference on Knowledge discovery and data mining* (2002) 429-435.

[12] P.S. Kam and A.W.C. Fu, "Discovering Temporal Patterns for Interval-Based Events", *2nd International Conference on Data Warehousing and Knowledge Discovery* (2000) 317-326.

[13] F. Höppner and F. Klawonn, "Finding Informative Rules in Interval Sequences", *4th International Conference on Advances in Intelligent Data Analysis* (2001) 125-134.

[14] L. Sacchi, C. Larizza, C. Combi and R. Bellazzi, "Data mining with Temporal Abstractions: learning rules from time series", *Data Mining and Knowledge Discovery* 15 (2007) 217-247.

[15] C. Combi, M. Franceschet and A. Peron, "Representing and Reasoning about Temporal Granularities", *Journal of Logic and Computation* 14 (2004) 51-77.

[16] Y. Shahar, "A framework for knowledge-based temporal abstraction", *Artificial Intelligence* 90 (1997) 79-133.

[17] J.F. Allen, "Towards a general theory of action and time", *Artificial Intelligence* 23 (1984) 123-154.

[18] R. Bellazzi, C. Larizza, P. Magni and R. Bellazzi, "Temporal data mining for the quality assessment of hemodialysis services", *Artif Intell Med 34* (2005) 25-39.

[19] M.B. Vilain, "A system for reasoning about time", *2nd National Conference in Artificial Intelligence* (1982) 197-201.

[20] R. Raj, M.J. O'Connor and A.K. Das, "An ontology-driven method for hierarchical mining of temporal patterns: application to HIV drug resistance research", *AMIA Annu Symp Proc* (2007) 614-9.

[21] S. Concaro, L. Sacchi, C. Cerra, P. Fratino, R. Bellazzi. "Mining Healthcare data with temporal association rules: improvements and assessment for a practical use", I*n Proc. of AIME 2009, Verona* (accepted for publication).

[22] P. Magni, F. Ferrazzi, L. Sacchi, R. Bellazzi. "TimeClust: a clustering tool for gene expression time series", *Bioinformatics,* 24 (2008) 430-432

[23] S.B. Martins, Y. Shahar, D. Goren-Bar, M. Galperin, H. Kaizer, L.V. Basso, D. McNaughton, M.K. Goldstein. "Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data." *Artif Intell Med.* 43 (2008)17-34

[24] P. Ciccarese, C. Larizza. "A framework for temporal data processing and abstractions". *AMIA Annu Symp Proc.* (2006)146-50.