ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-606-5-1117

Mining Physiological Data for Discovering Temporal Patterns on Disease Stages

Corrado Loglisci and **Michelangelo Ceci** and **Donato Malerba**[1]

Abstract. Analyzing physiological data can be of great importance in unearthing information on the course of a disease. In this paper we propose a data mining approach to analyze these data and acquire knowledge, in the form of temporal patterns, on the physiological events which can frequently trigger particular stages of disease. The application to the sleep sickness scenario is addressed to discover patterns, expressed in terms of breathing and cardiovascular system time-annotated disorders, which may trigger particular sleep stages.

1 Introduction

Physiological data consist of the measurements over time of some parameters which describe the course of diseases. The vast quantities and the complexity of such data make the activity of interpretation so arduous that resorting to automatic techniques of analysis becomes necessary. Although of great usefulness in many scenarios, most of analysis approaches proposed in the literature still present two limitations. First, the temporal dimension is a valuable source of information and inferring time-based information, such as timing of heart failure, duration of apnoea, can help to derive meaningful conclusion on the course of diseases. Nevertheless, existing approaches take into account this dimension, but only few attempts have been done to automatically infer temporal knowledge in the context under investigation. Second, existing approaches are mostly based on the apriori defined models of disease that become inapplicable when domain knowledge used to define these models is not promptly available or it has not been previously acquired, as in the case of new pathologies.

These considerations motivate the current work whose main peculiarity is that of analysing physiological data without (necessarily) relying on domain medical information. In particular, we propose a temporal data mining approach that mines time-varying physiological-data and discovers patterns of time-annotated *physiological events* which can trigger particular *stages* of disease. A physiological event is associated to a physiological variation while a stage corresponds to a specific state of disease which holds in a period of time. Therefore, two consecutive stages denote a transition in disease while the events which occur in the first stage and do not occur in the second one can be responsible of the transition of disease towards the second stage. Patterns are discovered from the events detected on a collection of pairwise stages of interest. Such a collection is properly created in order to consider only pairs of stages which depict similar transitions. The usage of pattern discovery is therefore addressed to find out the most frequent (and maybe significant) events which can determine similar transitions and, thus, can trigger analogous stages.

[1] Department of Computer Science, University of Bari "Aldo Moro", Italy, email: {loglisci, ceci, malerba}@di.uniba.it

2 Problem Formulation

Before describing the computational solution to the scientific problem investigated in this work, here we introduce some necessary concepts. Let $P : \{a_1, \ldots, a_m\}$ be the finite set of real-valued physiological parameters. Physiological data consists of a collection Mp of time-ordered measurements of the set P.

A disease stage S_j is a 4-tuple $S_j : \langle ts_j, te_j, C_j, SV_j \rangle$, where $[ts_j..te_j]$ $(ts_j, te_j \in \tau, ts_j \leq te_j)$[2] represents the time-period of the stage, while $C_j : \{f_1, f_2, \ldots\}$ is a finite set of *fluents*, namely facts in terms of parameters P that are true during the time-period $[ts_j..te_j]$. SV_j is the set $\{sv_1, \ldots, sv_k, \ldots, sv_m\}$ containing high-level descriptions of parameters P during $[ts_j..te_j]$. A physiological event e is a signature $e : \langle t_F, t_L, Ea, IEa, SEa \rangle$, where $[t_F..t_L]$ is the time-interval during which the event e occurs $(t_F, t_L \in \tau)$, $Ea : \{ea_1, \ldots, ea_k, \ldots, ea_{m'}\}$ is a subset of P and contains m' distinct parameters which take values in the intervals $IE_a : [inf_1, sup_1], \ldots, [inf_k, sup_k], \ldots [inf_{m'}, sup_{m'}]$ respectively during $[t_F..t_L]$. Finally, $SE_a : \{sv_1, \ldots, sv_k, \ldots, sv_{m'}\}$ is a set of m' symbolic values associated to Ea. In particular, IE_a is a quantitative description of the event, while SE_a is a qualitative representation of the trend of values taken by each ea_k during $[t_F..t_L]$.

To naturally model the complexity of physiological events we resort to the approaches in Inductive Logic Programming (ILP) [6] which permits us to suitably represent patterns of such events. A temporal pattern T_P is a set of atoms $p_0(t_0^1)$, $p_1(t_1^1, t_1^2)$, $p_2(t_2^1, t_2^2)$, …, $p_r(t_r^1, t_r^2)$, where p_0 is the *key* predicate, p_i, $i = 1, \ldots, r$, is either a *structural* predicate or a *property* predicate or an *is_a* predicate or a temporal predicate, while t_i^j are either constants, which correspond to values of predicate arguments, or variables, which identify events or physiological parameters. The key, structural, is_a, property predicates define the content and temporal information of the events, while temporal patterns define the relationships between two interval-based events in Allen temporal logic [1].

The problem of discovering temporal patterns on physiological data is in this work solved through a four-stepped computational solution described in the following section.

3 Temporal Pattern Discovery

A stage can be seen as one of the steps of disease characterized by numerical (C_j), symbolic (SV_j) and temporal $([ts_j..te_j])$ components, and therefore, corresponds to one of the distinct segments of Mp. To determine these components we resort to the method proposed in [3]. More precisely, the periods of time $[ts_j..\, te_j]$ are obtained with a technique of temporal segmentation which splits Mp in a succession of

[2] τ is a finite totally ordered set of time-points.

multi-variate segments. This produces a sequence of segments of Mp different from each other such that two consecutive segments have different fluents. The fluents C_j are generated with an ILP system and correspond to numerical interval-values formulae which characterize the measurements included in $[ts_j .. te_j]$ and discriminate these from the measurements in $[ts_{j-1} .. te_{j-1}]$ and $[ts_{j+1} .. te_{j+1}]$. Finally, the components SV_j are determined as a representation of the slope of the regression line built on the values taken by each $a_k \in P$ in the time interval $[ts_j..te_j]$.

A collection R of pairwise stages is properly created to discover patterns from similar transitions. Pairwise stages appropriate for R are identified on the basis of a similarity value: pairs whose first stages and second stages have similarity value which exceeds a user-defined numerical threshold CS ($CS \in [0, 100]$) are considered. For instance, two pairs (S_j, S_{j+1}), (S_k, S_{k+1}) are collected in R if the similarity between S_j and S_k, and the similarity between S_{j+1} and S_{k+1} exceeds CS. In this work the similarity between two stages S_j and S_k corresponds to the similarity between their fluents C_j, C_k under the assumption that the symbolic values SV_j, SV_k are identical. Since the fluents are sets of interval-valued formulae, the similarity between C_j and C_k is so computed: $Sim(C_j, C_k) = (\sum_{f_j \in C_j, f_k \in C_k} (1 - Diss(f_j, f_k))) * 100/(|C_j| * |C_k|)$ where $f_j(f_k)$ is a single interval-valued formula of C_j (C_k). To compute $Diss(f_j, f_k)$ we resort to dissimilarity functions specific for interval-valued data. In particular, we consider the Gowda and Diday's [2] dissimilarity measure defined as

$$Diss(f_j, f_k) = \sum_{h=1...|P|} \delta(f_{j_h}, f_{k_h})$$

where, f_{j_h}, f_{k_h} are the intervals assumed by the parameter a_h, $|P|$ is number of intervals (parameters), and $\delta(f_{j_h}, f_{k_h})$ is obtained considering three types of dissimilarity measures incorporating different aspects of similarity. It should be noted that several collections of similar transitions can be actually created from the pairs of stages: the final collection R is selected by the user in the set of possibly overlapping collections.

Once the collection R of pairs of stages has been identified, for each pair (S_j, S_{j+1}) we look for events which may trigger the transition from S_j to S_{j+1} by exploiting the approach we proposed in [4]. The basic idea is that of mining candidate events and, then, selecting from the candidates events those which are most statistically interesting. The candidates are identified as variations in the measurements of two consecutive time-windows which slide back in time over the stages S_j and S_{j+1}. Once the candidates for a single pair (S_j, S_{j+1}) are generated, the sequence with the most statistically interesting events is identified by selecting the *most supported* ones. An event e_u is called *most supported* if it is the only one event which meets the following condition: there exists a set of candidate events $E : \{e_1, e_2, \ldots, e_t\}$ such that the set of parameters Ea and the time interval $[t_F..t_L]$ associated to e_u are included in the set of parameters and in the time intervals of all candidates in the set E. Moreover, the numerical interval IEa of each parameter in e_u is included in the corresponding numerical interval IEa of all candidates in the set E while the symbolic values SEa associated to the parameters of e_u coincide with the symbolic values of all candidates in the set E. The sequence of the most supported events for each pair of disease stages $(S_j, S_{j+1}) \in R$ forms the set ES of sequences of events from which temporal patterns will be discovered.

Discovery of temporal patterns from ES is performed by resorting to the ILP method for frequent patterns mining implemented in SPADA [5]. The sequences of physiological events are described in Datalog logic language and stored as sets of ground atoms in the extensional part D_E of a deductive database D. Ground atoms are logic predicates (in this work key, structural, properties and is_a predicates) which have only constant terms. The intensional part D_I of the database D is rather defined with the predicates based on Allen temporal logic [1]. D_I represents background knowledge on the problem, such as the precedence relationship between two events through the predicate $before()$. The discovery process performs a breadth-first search in the space of the patterns, from the most general to the most specific patterns, and prunes portions of the search space which contain only infrequent patterns. Infrequent (frequent) patterns are those patterns whose support[3] is less than (greater than or equal to) a minimum threshold $minF$. Discovered frequent patterns correspond to the temporal patterns of interest in this work.

4 Application to Sleep Disorders

The proposed approach was applied to the scenario of Sleep Disorders to discover patterns of disorders (i.e., events) of the cardiovascular and the respiratory systems which may trigger particular stages of the central nervous system during sleep. Dataset consists of physiological measurements collected at 1 second during sleep for only one patient[4]. Physiological parameters are *ecg, airflow, thorex, abdoex, pr, saO2*, which describe the cardiovascular and respiratory systems, while *eeg, leog, reog, emg* are used to describe the central nervous system. Temporal patterns were discovered at different experimental settings, and, more precisely, by tuning CS to 60%, 70%, 80% and the minimal duration of stages (afterwards, $minSD$) at 60, 120, 180 seconds.

Patterns with higher support were discovered at lowest values of CS and $minSD$, for instance the pattern:

$sequence(S), event(E1, S), event(E2, S), event(E3, S), before(E1, E2), before(E2, E3),$
$parameter\_of(E1, P1), is\_a(P1, abdoex), value\_interval(P1,' [-1.412, 0.722]'),$
$symbolic\_value(P1,' STRONG\_INCREASE'), parameter\_of(E2, P2), is\_a(P2, airflow),$
$value\_interval(P2,' [-2.322, 3.482]'), symbolic\_value(P2,' STRONG\_DECREASE'),$
$parameter\_of(E3, P3), is\_a(P3, saO2), value\_interval(P3,' [94.013, 95.012]'),$
$symbolic\_value(P3,' DECREASE')$ $[support = 21.4\%]$

is supported by a percentage of 21.4% of the sequences of events. Patterns with lower support but with more predicates are rather discovered at higher values of CS and $minSD$. Indeed, a larger $minSD$ value leads to the generation of wider time-windows and to a numerous set of distinct events which results in reducing the support of discovered patterns.

Acknowledgment. This work is partial fulfillment of the research objective of the ATENEO-2009 project "Modelli e metodi computazionali per la scoperta di conoscenza in dati biomedici".

REFERENCES

[1] J. F. Allen, 'Maintaining knowledge about temporal intervals', *Commun. ACM*, **26**(11), 832–843, (1983).
[2] E. Diday and F. Esposito, 'An introduction to symbolic data analysis and the sodas software', *Intell. Data Anal.*, **7**(6), 583–601, (2003).
[3] C. Loglisci and M. Berardi, 'Segmentation of evolving complex data and generation of models', in *ICDM Workshops*, pp. 269–273, (2006).
[4] C. Loglisci and D. Malerba, 'Discovering triggering events from longitudinal data', in *ICDM Workshops*, pp. 248–256, (2008).
[5] D. Malerba and F. A. Lisi, 'An ILP method for spatial association rule mining', in *Workshop on Multi-Relat. Data Mining*, pp. 18–29, (2001).
[6] S. Muggleton, *Inductive Logic Programming*, Academic Press, 1992.

[3] The support of a pattern P is the percentage of sequences which covers P.
[4] accessible at PhysioBank site http://www.physionet.org/physiobank/