Discovering Temporal Patterns of Complex Events in Biosignal Data

Corrado Loglisci, Michelangelo Ceci Donato Malerba

Dipartimento di Informatica, Università degli Studi di Bari

Abstract. Analyzing biosignal data is an activity of great importance which can unearth information on the course of a disease. In this paper we propose a temporal 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 proposed approach is realized through a four-stepped computational solution: first, disease stages are determined, then a subset of stages of interest is identified, subsequently physiological time-annotated events which can trigger those stages are detected, finally, patterns are discovered from the extracted events. The application to the sleep sickness scenario is addressed to discover patterns of events, in terms of breathing and cardiovascular system time-annotated disorders, which may trigger particular sleep stages.

1 Introduction

Biosignal data is a particular kind of biomedical data which consist of the measurements over time of some parameters (e.g., blood oxygen, respiration rate, heart rate, etc.) which describe the health conditions of a patient. They can convey relevant information on the clinical picture of patients and, in particular, on the course of diseases. However, the vast quantities and the complexity of biosignal data make the interpretation of them so arduous that resorting to automatic techniques of analysis becomes necessary.

In the literature two main analysis approaches have been pursued. In the first one, the analysis process aims to acquire information about the course of disease only for a limited set of time-stamped measurements. The problem is usually faced by applying *trend detection* techniques aiming at either generating high-level descriptions of the parameters [15]. The second one deals with data which describe the complete course of disease over time, and the goal is tracking each stage of the disease with a pre-existing disease model [1]. Such a model is often represented in the form of deterministic automaton where each state corresponds to an expected stage of disease while a transition between two states denotes the change from one stage to another one. Although these approaches have shown to be of great usefulness in many scenarios, they still present two issues.

First, time dimension is a valuable source of information which can help to derive meaningful conclusions. Existing approaches take into account this dimension, but only few attempts have been done to automatically infer temporal knowledge in the context under investigation (e.g. [5]). A realistic analysis should rather take into account time dimension to discover temporal information, such as timing of heart failure or duration of apnoea. Indeed, temporal aspect is intrinsic in the nature of the biosignal data and it is of so vital importance that it is considered when making decisions in the medical tasks.

Second, most of existing approaches presents a strong dependence on the background knownledge: they are based on the apriori defined models of disease that become inapplicable when domain information used to define these models is not promptly available, as in the case of new pathologies.

These considerations motivate the current work. In particular, we propose a temporal data mining approach that aims at supporting the task of interpretation of the course of a disease. It mines time-varying biosignal data and discovers patterns of time-annotated *complex events* which can trigger particular *stages* of disease. A complex event is associated to a physiological variation while a stage corresponds to a specific state of disease which holds in a period of time. Two consecutive stages represent two different states and together they depict a transition in disease. So, given two consecutive stages, we assume that whatever happens in the first stage may affect the second one, and, since two consecutive stages are different each other, 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. Therefore, the transition can be ascribed to these events.

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) complex events which can determine similar transitions and, thus, can trigger analogous stages.

Another peculiarity of the approach we are going to present is that of analysing biosignal data without (necessarily) relying on domain medical information. This permits to have several advantages. First, the mining process is not limited by constraints imposed *apriori* by domain experts, and this facilitates the discovery of unexpected information. Second, clinicians can validate previously acquired knowledge with an empirical means, or also, they can find new comprehension to the conclusions drawn in evidence-based medicine. Third, the risk to discover trivial or uninteresting information, as largely believed when not exploiting prior knowledge, is mitigated by the statistical evidence of the mining results, which can be indicative of their relevance.

This work also extends our previous study [8], where we described an approach to discover physiological complex events potentially responsible to change the human physiology. The extension aims to: 1) find out regularities based on statistical evidence among complex events; 2) discover patterns of such events on similar transitions of disease.

The paper is organized as follows. In next section we define the problem in terms of four sub-problems. The computational solution for them is described in

Section 3. An application to the case of Sleep Slickness is presented in Section 4. Finally, some conclusions close this paper.

2 Problem Formulation

Before formally defining the scientific problem of interest 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 (e.g., $\{$*blood oxygen*, *heart rate* and *respiration rate* $\}$). A single biosignal consists of the time-ordered measurements of a parameter of P, while a set of biosignals is a collection Mp of time-ordered measurements of the set P.

Definition 1. *A disease stage S_j is a 4-tuple $S_j = \langle ts_j, te_j, C_j, SV_j\rangle$, where [$ts_j..te_j$] ($ts_j, te_j \in \tau$, $ts_j \leq te_j$)*[1] *represents the time-period of the stage, while $C_j : \{f_1, f_2, \ldots\}$ is a finite set of* fluents*, namely facts or properties 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 m symbolic values such that sv_k is a high-level description of the parameters $a_k \in P$ during $[ts_j..te_j]$.*

An example of state is $S_1 : \langle t_1, t_{10}, \{$ *blood oxygen* $\in$ *[6500;6700], heart rate* $\in$ *[69;71], respiration rate* $\in$ *[2300;5500]*$\}$, $\{$*blood oxygen is INCREASING, heart rate is STEADY, respiration rate is INCREASING*$\}$ $\rangle$ which can be interpreted as follows: S_1 is associated with the period of time $[t_1, t_{10}]$ and is characterized by the fact (fluent) that the parameters *blood oxygen*, *heart rate* and *respiration rate* range respectively in $[6500; 6700]$, $[69; 71]$, $[2300; 5500]$ and have increasing, steady and increasing trend respectively.

Definition 2. *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 when 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 $[t_F..t_L]$.

Two examples of events are $e_1 : \langle t_1, t_5, \{bloodoxygen\}, \{[6300; 6800]\}, \{DECREASING\}\rangle$ and $e_2 : \langle t_6, t_{10}, \{bloodoxygen\}, \{[6600; 7000]\}, \{INCREASING\}\rangle$ which can interpreted as follows: e_1 (e_2) is associated with the time-period $[t_1, t_5]$ ($[t_6, t_{10}]$) during which the parameter *blood oxygen* ranges in $[6300; 6800]$ ($[6600; 7000]$) and has a decreasing (increasing) trend.

Actually, events and sequences so defined have a description more complex than traditional one in Data Mining, where an event is represented by a symbol of a predefined alphabet and a sequence is a list of such symbols. Traditional

[1] τ is a finite totally ordered set of time-points. Henceforth, the corresponding order relation is denoted as $\leq$.

methods of pattern discovery become thus inapplicable to the Definitions 1, 2 and, moreover, possible transformation of complex events and sequences into traditional representations could imply loss of information and could not take into account the original structure of the events. To face this issue we resort to approaches synthesized in *Inductive Logic Programming* (ILP) which permit us to discover patterns from sequences of complex events given their peculiarity to naturally model complex data. This way, sequences and events of our interest (Definitions 1, 2) can be described in a logic formalism and represented as sets of *ground atoms* [3]. For instance, the sequence of the events above can be represented with the following set of atoms:

sequence(seq$_1$). event(seq$_1$,e$_1$). time_tF(e$_1$,1). time_tL(e$_1$,5). parameter_of(e$_1$,p$_1$). is_a(p$_1$,blood_oxygen). value_interval(p$_1$,'[6300;6800]'). symbolic_value(p$_1$,'DECREASE'). event(seq$_1$,e$_2$). time_tF (e$_2$,6). time_tL(e$_2$,10). parameter_of(e$_2$,p$_2$). is_a(p$_2$,blood_oxygen). value_interval(p$_2$,'[6600;7000]'). symbolic_value(p$_2$,'INCREASE').

where *sequence(seq$_1$)* is the atom which identifies the sequence seq$_1$ through the predicate *sequence*(); *event(seq$_1$, e$_1$)* is the atom which relates the sequence seq$_1$ to the event e$_1$ through *event*(); *time_tF (e$_1$, 1)* is the atom which assigns the specific value 1 to the attribute *time_tF* of e$_1$ through *time_tF*(); *parameter_of(e$_1$, p$_1$)* is the atom which relates the event e$_1$ to the parameter p$_1$ through *parameter_of*(); *is_a(p$_1$, blood_oxygen)* is the atom which assigns a specific value *blood_oxygen* to p$_1$ through *is_a*(), *value_interval(p$_1$,'[6300;6800]')* is the atom which assigns a specific interval of values [6300;6800] to p$_1$ through *value_interval*() and *symbolic_value(p$_1$,'DECREASE')* is the atom which assigns a specific symbolic value DECREASE to p$_1$ through *symbolic_value*(). Depending on their function, logic predicates used in the atoms can be classified as: 1)*property* predicates, which define the value taken by an attribute of an event or parameter (e.g., *time_tF*(), *symbolic_value*()); 2)*structural* predicates, which relate events, events with parameters and sequences with events (e.g., *event*(), *parameter_of*()); 3)*is_a* predicates, which identify specific events or parameters; 4)*key* predicate which identifies a specific sequence (i.e., *sequence*()).

After these preliminary concepts, it is possible formally define a temporal pattern:

Definition 3. *A temporal pattern T_P is a set of atoms $p_0(t_0^1)$, $p_1(t_1^1, t_1^2)$, $p_2(t_2^1, t_2^2)$, $\ldots$, $p_r(t_r^1, t_r^2)$, 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 property predicates, or variables, which identify sequences, events or physiological parameters.*

Temporal predicates in Definition 3 express possible temporal relationships between two events e$_1$, e$_2$ according to the Allen temporal logic[2]: *before(e$_1$, e$_2$), equal(e$_1$,e$_2$), meets(e$_1$,e$_2$), overlaps(e$_1$, e$_2$), during(e$_1$, e$_2$), starts(e$_1$, e$_2$), finishes(e$_1$, e$_2$).* For instance, the temporal pattern

Tp: sequence(Q), event(Q, E1), event(Q, E2), before(E1, E2), parameter_of(E1, P1), is_a(P1, blood_oxygen), value_interval(P1,'[6300;7000]'), symbolic_value(P1, steady), *is_a(P2, respiration_rate), value_interval(P2,'[2300;5500]'), symbolic_value(P2, strong_increase)*

expresses the fact that, for a subset of sequences, the event E_1 is followed by E_2, where in E_1 the blood oxygen has steady trend and ranges in [6300;7000] while in E_2 the respiration rate is strongly increasing with values in [2300;5500].

Considering the concepts thus far defined, the problem of discovering frequent temporal patterns of complex events in biosignals can be divided in four sub-problems formalised as follows:

1. *Given*: a set of biosignals $Mp : \{Mp_{t1}, Mp_{t2}, \ldots, Mp_{tn}\}$,
   Find: a finite set $S : \{S_1, S_2, \ldots\}$ of consecutive disease stages which represent distinct sub-sequences of *Mp*.
2. *Given*: a criterion CS to collect pairwise stages of interest from S,
   Find: a collection R of pairwise stages (S_j, S_{j+1}) which satisfy the criterion CS.
3. *Given*: the collection R,
   Find: a set ES of sequences $\langle e_1, e_2, \ldots\rangle$ of complex events for each pair (S_j, S_{j+1}) in R.
4. *Given*: the set *ES* and a user-defined threshold *minF*,
   Find: patterns in *ES* whose support exceeds the threshold *minF*.

A computational solution to each of these sub-problems is described in the following section.

3 Discovering Patterns in Biosignal Data

3.1 Determination of Disease Stages

As before introduced, a stage can be seen as one of the steps of disease characterized by numerical (C_j), symbolic ($\{sv_1,\ ...,\ sv_k,\ ...,\ sv_m\}$) and temporal ($[ts_j..te_j]$) properties. In other words, a stage corresponds to one of the distinct segments of *Mp*, and this provides us some hints on the approach to follow in order to determine the stages $S_j = \langle ts_j, te_j, C_j, SV_j\rangle$. The elements ts_j, te_j, C_j are obtained by resorting to the method we proposed in [6] which is here shortly described. It first identifies the periods of time $[ts_j..te_j]$ with a hybrid technique of segmentation of multi-variate time-series. This produces a sequence of segments which are different each other and it guarantees that two consecutive segments have different fluents: given three consecutive segments, $[ts_{j-1}..te_{j-1}]$, $[ts_j..te_j]$, $[ts_{j+1}..te_{j+1}]$, the fluents C_j associated to $[ts_j..te_j]$ are conditions which *characterize* the data included in $[ts_j..te_j]$ and *discriminate* those of the segments $[ts_{j-1}..te_{j-1}]$ and $[ts_{j+1}..te_{j+1}]$. In this sense the problem of determining C_j can be thus seen as a *Conceptual Inductive Learning* task [12] which permits to characterize each stage of disease and distinguish it from each other with a rigorous description. At this aim we exploit the inductive learning capabilities of ATRE learning system [9] which outputs a set of interval-valued atomic formulae. In the following, an example of set of conditions is reported:

$C_j : \{f_1 : \langle bloodoxygen \in [6500; 6700], heartrate \in [69; 71], respirationrate \in [2300; 5500]\rangle, f_2 : \langle bloodoxygen \in [6800; 7000], heartrate \in [55; 67], respirationrate \in [2500; 4000]\rangle, \ldots\}$

Finally, the values of the elements SV_j of S_j are derived by a temporal abstraction technique [14]. It is defined through a function $\Theta : \Pi \rightarrow \Lambda$ which provides an high-level representation $\lambda \in \Lambda$ of the most relevant features $\pi \in \Pi$ of data. In our case, Θ returns, for each parameter a_k, a representation of the slope of the regression line built on the values taken by a_k in the time interval $[ts_j..te_j]$: for instance, the slope values ranging in the interval (0.2..1] are described as INCREASING.

3.2 Collecting Pairwise Stages

As stated before, a collection R of pairwise stages is properly created in order to discover patterns from similar transitions: such patterns represent the events which more frequently trigger analogous stages. 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[2], and since the fluents are sets of interval-valued formulae for the subsection 3.1, the similarity between C_j and C_k is so computed:

$$Sim(C_j, C_k) = \frac{\sum\limits_{f_j \in C_j, f_k \in C_k} (1 - Diss(f_j, f_k))}{|C_j| + |C_k|} \tag{1}$$

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 [4] dissimilarity measure defined as:

$$Diss(f_j, f_k) = \sum_{h=1\ldots|P|} \delta(f_{j_h}, f_{k_h}) \tag{2}$$

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, namely $\delta(f_{j_h}, f_{k_h}) = \delta_\pi(f_{j_h}, f_{k_h}) + \delta_s(f_{j_h}, f_{k_h}) + \delta_c(f_{j_h}, f_{k_h})$, $(\delta_\pi, \delta_c,\ \delta_s \in [0, 1])$. In particular, δ_π indicates the relative position of f_{j_h}, f_{k_h} in their entire interval of values, δ_s indicates the relative sizes of f_{j_h}, f_{k_h} without referring to common parts between them, while δ_c is a measure of the non common parts between f_{j_h}, f_{k_h}.

[2] The notion of similarity between two stages does concern neither the time-periods $[ts_j, te_j]$ nor the sets SV_j, i.e., two stages can be similar although they are associated to different time-periods and symbolic values.

3.3 Detection of Complex Events

Once the collection R of pairwise 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}. The assumption that events that occur during the time interval $[ts_j..te_j]$ should not occur in $[ts_{j+1}..te_{j+1}]$ provides us some hints on the approach to follow in order to detect events. At this aim we integrate the algorithm we proposed in [7] which is here reported.

The basic idea is that of mining candidate events and, then, selecting from them the most *statistically interesting.* The algorithm for mining the set of candidate events $\{e \mid e = \langle t_F, t_L, Ea, IEa, SEa\rangle\}$ proceeds by iteratively scanning the physiological measurements included in the stages S_j (i.e., $\{Mp_{ts_j}, \ldots, Mp_{te_j}\}$) and S_{j+1} (i.e., $\{Mp_{ts_{j+1}}, \ldots, Mp_{te_{j+1}}\}$) with two adjacent time-windows which slide back in time (Figure 1). The candidates are identified by finding variations in the measurements between the windows w and w'. At the first iteration, the time-windows w, w' (w' immediately follows w) correspond to the last part of S_j and to the complete S_{j+1} respectively. If a candidate is found, then the next candidate is searched for the pair (w'',w), where the new time-window w'' has the same size of w (Figure 1a). Otherwise, the next candidate is searched for the pair (w'',w'), where w'' is strictly larger than w (Figure 1b). At the end of a single scan a sequence of candidates is obtained.

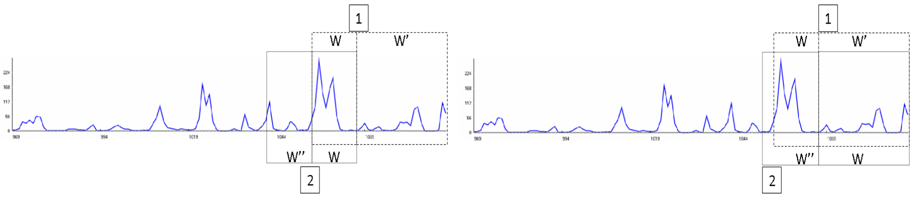
The intuition underlying the detection of candidate events for a given couple of windows (w, w') is that the intrinsic dependence of two parameters may change between the two adjacent time-windows. This idea is implemented in the following strategy: for each parameter a_i two multiple linear regression models are built on the remaining parameters in P by considering the distinct physiological measurements in w and w' respectively:

$$a_i = \beta'_0 + \beta'_1 a_1 + \ldots + \beta'_{i-1} a_{i-1} + \beta'_{i+1} a_{i+1} + \ldots + \beta'_m a_m,$$
$$a_i = \beta''_0 + \beta''_1 a_1 + \ldots + \beta''_{i-1} a_{i-1} + \beta''_{i+1} a_{i+1} + \ldots + \beta''_m a_m,$$

The couple of regression models which guarantees the lowest predictive information loss is selected. Let a_h be the parameter for which the lowest predictive information loss is obtained, the set of parameters $Ea = \{a_k \in P - \{a_h\} \mid \ |\beta'_k - \beta''_k| \leq \sigma_k\}$[3] is selected and associated with the time window $w : [t_F..t_L]$ to form the event $e : \langle t_F, t_L, Ea, IEa, SEa\rangle$.

The set Ea is further filtered in order to remove those parameters for which no interval of values which discriminates the measurements in w from those in w' can be generated: this permits also to determine the element IEa. In particular, for each $a_k \in Ea$ the interval $[inf_k, sup_k]$ is computed by taking the minimum (inf_k) and maximum (sup_k) value of a_k in w. If $[inf_k, sup_k]$ is weakly consistent with respect to values taken by a_k during the time window w' then a_k is kept, otherwise it is filtered out. Weak consistency is verified by computing the weighted average of the zero-one loss function on the measurements in w', where

[3] σ_k is automatically determined and is the standard deviation of the k-th coefficient of linear regression models computed on non-overlapping time-windows of size $t_L - t_F$ over (S_j, S_{j+1}).



(a) (b)

Fig. 1. Mining the candidate events: if the event is found for the pair (w,w'), then the next candidate is sought for (w'',w) (a), otherwise it is sought for the pair (w'',w') (b).

weights decrease proportionally with the time points in w'. Finally, the filtered set of m' parameters will be associated with a set of intervals $\{[inf_1, sup_1]$, …, $[inf_k, sup_k]$, …, $[inf_{m'}, sup_{m'}]\}$, which corresponds to the quantitative description IEa of the event $e : \langle t_F, t_L, Ea, IEa, SEa\rangle$.

The set SEa is determined through the same technique of temporal abstraction introduced in the subsection 3.1. It contains a symbolic value for each a_k that denotes the slope of the regression line built on the biosignal data in $[t_F..t_L]$.

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* events. An event e_u is called *most supported* if it meets the following two conditions: 1) there exists a set of candidates $\{e_1, e_2, \ldots, e_t\}$ which contains the same information of e_u, that is: $\forall e_q, q = 1, \ldots, t$, $e_q \neq e_u$, the set of parameters Ea associated to e_q includes the set of parameters associated to e_u, the time interval $[t_F, t_L]$ associated to e_q includes the time interval associated to e_u, and, finally, the set of symbolic values SEa and the intervals IEa associated to the parameters of e_q coincide; 2) no event e_v exists whose information is contained in a set of candidates $\{e_1, e_2, \ldots, e_{t'}\}$ with $|\{e_1, e_2, \ldots, e_{t'}\}| > |\{e_1, e_2, \ldots, e_t\}|$. The *support* of the event e_u is computed as follows: let $\{e_1, e_2, \ldots, e_z\}$ be the set of candidates such that the time interval associated to each of them contains that of e_u, and $\{e_1, e_2, \ldots, e_t\}$ be the set of candidates as described at the point 1), then the support of e_u is $supp(e_u) = (t + 1)/z$.

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.

3.4 Temporal Pattern Discovery

Discovery of temporal patterns from ES is performed by resorting to the ILP method for frequent patterns mining implemented in SPADA [10]. The sequences returned from the previous step are described in a logic formalism (Datalog language [3]) and stored as sets of ground atoms in the extensional part D_E of a deductive database D. Logic predicates used to describe ground atoms are those presented in Section 2: the predicates define the structure of the sequences $(sequence(), event())$, the content of events $(parameter\_of(), value\_interval()$, $symbolic\_value())$ and the associated temporal information $(time\_tf(), time\_tL())$.

The intensional part D_I of the database D is rather defined with the predicates based on Allen temporal logic [2]: D_I represents background knowledge on the problem (e.g., precedence relationships between two events through the predicate $before()$) and allows to entail additional atoms by applying these predicates to the extensional part.

For example: let D be a deductive database which contains sequences of events as defined in Section 2, the extensional part D_E includes the ground atoms:

sequence(seq1). sequence(seq2). event(seq1,e1). event(seq1,e2).event(seq2,e3).
event(s2,e4). time_tF(e1,10). time_tL(e1,25). time_tF(e2,22). time_tL(e2,23).
time_tF(e3,90). time_tL(e3,110).time_tF(e2,170). time_tL(e2,190).

where the constants $seq1$ and $seq2$ denote two distinct sequences, while the constants $e1$, $e2$, $e3$, $e4$ identify four events. The intensional part D_I is formulated as the logic program:

before(E1, E2) ← *event(S, E1),event(S, E2), E1* ≠ *E2, time_tL(E1,T1), time_tF(E2,T2), T1<T2, not(event(S, E3), E3*≠ *E1, E3*≠ *E2, time_tF(E3,T3F), time_tL(E3,T3L), T1<T3F, T3L<T2)*

by considering the atoms in D_E the ground atoms *before(e1, e2), before(e3, e4)* are entailed and added to D_E.

The process of discovery performs a search in the space of patterns by following the level-wise method proposed in [11]. In particular, it performs a breadth-first search in the space of pattern, evaluates them from the most general to the most specific and prunes portions of the search space which contain only infrequent patterns. Infrequent patterns are those patterns whose support is less than the threshold $minF$ (conversely, for frequent patterns), while the support of a pattern P is the percentage of sequences in D which covers the pattern P. The application of the level-wise method requires a generality ordering which is monotonic with respect to pattern support [10]. The generality ordering adopted by SPADA is based on the notion of θ-subsumption [13]. For instance, given three patterns

$$T_{P1} \equiv is\_a(P1, blood\_oxygen),$$
$$T_{P2} \equiv is\_a(P1, blood\_oxygen), symbolic\_value(P1,' INCREASE').$$
$$T_{P4} \equiv is\_a(P1, blood\_oxygen), symbolic\_value(P1,' INCREASE'),$$
$$value\_interval(P1,' [6300, 6800]').$$

T_{P1} is more general of T_{P2} which is more general of T_{P4} with respect to the considered generality order. The monotonicity property can be exploited in order to prune the search space. In fact, if T_{P1} is infrequent then the patterns more specific than T_{P1} are infrequent and thus they are pruned. A detailed description of SPADA system can be found in [10].

4 Application to Sleep Disorders

Sleep disorders are issues of great importance and widely investigated in medicine because some serious diseases are accompanied by typical sleep disturbances (e.g., daylight sleepiness), and this attracts the interest of the several scientific communities. The possible influence of physiological alterations on sleep disorders motivates our interest in this scenario. In particular, we apply the proposed approach to discover patterns of disorders (i.e., complex events) of the cardiovascular and the respiratory systems which may trigger particular stages of the central nervous system during sleep.

Dataset Description. Experiments concern the biosignal data of polysomnography (measurements of physiological parameters during sleep) of a patient observed from 21.30 p.m. to 6.30 a.m.. The dataset has been created by sampling measurements at 1 second and it is publicly accessible at PhysioBank site[4]. Physiological parameters are *eeg* (electroencephalogram), *leog*, *reog* (electrooculograms), *emg* (electromyogram), *ecg* (electrocardiogram), *airflow*, (nasal respiration), *thorex* (thoracic excursion), *abdoex* (abdominal excursion), *pr* (heart rate) and *saO2* (arterial oxygen saturation). Where, *ecg, airflow, thorex, abdoex, pr, saO2* describe the cardiovascular and respiratory systems, while *eeg, leog, reog, emg* describe the central nervous system.

Results. Different sets S of disease stages are obtained by tuning the minimal duration of the stages [6]. For each S different collections R are created by setting CS to 60, 70, 80. Pattern are discovered from these collections by setting the threshold $minF$ to 5% (Table 1). As we can see the number of discovered patterns (#patterns) is strongly dependent on the minimal duration of the stages, indeed the greater the stages, the higher the dissimilarity between the stages and the lower the number of similar pairwise stages (cardinality of R). This can be due to the fact that the fluents of stages with longer duration characterize and discriminate an higher number of physiological measurements. Therefore, they tend to be too specific for the set of data to characterize and very dissimilar from other fluents. In these cases, the cardinality of R is lower and this produces a set ES with a small number of sequences where it could be difficult to discover frequent patterns.

A first interesting result is produced when the minimal duration is set to 60s and CS to 60. In this case a set ES of nine sequences (as many the pairs of stages) of complex events is identified, while 579 frequent patterns are discovered, among them the most frequent one, which can trigger the transition depicted by the 9 pairs of stages, is so described:

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

This pattern involves temporal predicates ($before()$), structural predicates (e.g., $parameter\_of()$) and property predicates(e.g., $symbolic\_value()$) and it is supported by a percentage of 21.4% of the total sequences.

Patterns with more predicates but with lower support are rather discovered at higher values of the minimal duration. For instance, when the minimal duration

[4] http://www.physionet.org/physiobank/

Table 1. Results at different settings of the stages duration and CS ($minF$ =5%).

minimal duration(s)	$\|S\|$	CS	$\|R\|$	#pattern
60	139	60	9 pairwise stages	579
		70	3 pairwise stages	112
		80	0	0
120	126	60	6 pairwise stages	63
		70	3 pairwise stages	34
		80	0	0
300	31	60	3 pairwise stages	7
		70	1 pairwise stages	4
		80	0	0

is 120s and CS is 60, the set R of similar pairs and the set ES of sequences of events amount to 6. One of discovered patterns in this case is the following:

$\mathit{sequence}(S), \mathit{event}(E1, S), \mathit{event}(E2, S), \mathit{event}(E3, S), \mathit{before}(E1, E2), \mathit{before}(E2, E3), \mathit{before}(E3, E4), \mathit{parameter\_of}(E1, P11), \mathit{is\_a}(P11, \mathit{thorex}), \mathit{value\_interval}(P11, '[-3.984, 3.984]'), \mathit{symbolic\_value}(P11, '\mathit{INCREASE}'), \mathit{parameter\_of}(E2, P21), \mathit{is\_a}(P21, \mathit{abdoex}), \mathit{value\_interval}(P21, '[-1.757, 1.82]'), \mathit{symbolic\_value}(P21, '\mathit{STRONG\_INCREASE}'), \mathit{parameter\_of}(E2, P22), \mathit{is\_a}(P22, \mathit{thorex}), \mathit{value\_interval}(P22, '[-0.91, 2.071]'), \mathit{symbolic\_value}(P22, '\mathit{STRONG\_INCREASE}'), \mathit{parameter\_of}(E3, P3), \mathit{is\_a}(P3, \mathit{saO2}), \mathit{value\_interval}(P3, '[97.010, 98.009]'), \mathit{symbolic\_value}(P3, '\mathit{DECREASE}'), \mathit{parameter\_of}(E4, P4), \mathit{is\_a}(P3, \mathit{abdoex}), \mathit{value\_interval}(P3, '[-1.663, 1.443]'), \mathit{symbolic\_value}(P3, '\mathit{STEADY}')$ $[\mathit{support} = 7.14\%]$

Generally, a larger value of the minimal duration leads to the generation of wider time-windows and to a numerous set of different complex events that results in reducing the frequency of discovered patterns. This observation is also confirmed by the accuracy values (Table 2) of the event detection (subsection 3.3). Indeed, when the minimal duration is 120s (Table 2 at right) the number of true positive events (sensitivity) is lower while the number of false positives is higher, and this leads to avoid that the true positive events contribute to form the final set of frequent patterns.

Table 2. Accuracy of the event detection for the experiments in Table 1 when the minimal duration is 60s (left) and 120s (right)

specificity (%)	sensitivity (%)	$[t_F..t_L]$ width
44	71	10
46	68	15
43	64	20
48	70	25
48	71	30

specificity (%)	sensitivity (%)	$[t_F..t_L]$ width
39	67	20
42	62	30
40	59	40
36	64	50
41	66	60

5 Conclusions

In this work we have investigated some issues raised when analyzing time-ordered biosignal data. We proposed a temporal data mining approach guided only by data and which does not (necessarily) rely on domain medical knowledge. It can be addressed to support the initial investigations on disease and, more precisely, to discover patterns on the events which determine the progression of disease. In particular, patterns express sequences of time-annotated events which can frequently trigger particular stages of disease. Since frequency denotes regularity, discovered pattern can provide support for medical decision making on the basis on the evidence of regularities in the data. Finally, the application to the scenario of sleep sickness has led to the discovery of patterns already known in the medical literature, and others not previously known but, according to the domain experts, that have analysed them, worth of being further investigated.

References

1. K.-P. Adlassnig. Fuzzy systems in medicine. In J. M. Garibaldi and R. I. John, editors, *EUSFLAT Conf.*, pages 2–5. De Montfort University, Leicester, UK, 2001.
2. J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, 1983.
3. S. Ceri, G. Gottlob, and L. Tanca. *Logic Programming and Databases*. Springer, 1990.
4. E. Diday and F. Esposito. An introduction to symbolic data analysis and the sodas software. *Intell. Data Anal.*, 7(6):583–601, 2003.
5. T. Guyet, C. Garbay, and M. Dojat. Human/computer interaction to learn scenarios from icu multivariate time series. In S. Miksch, J. Hunter, and E. T. Keravnou, editors, *AIME*, volume 3581, pages 424–428. Springer, 2005.
6. C. Loglisci and M. Berardi. Segmentation of evolving complex data and generation of models. In *ICDM Workshops*, pages 269–273. IEEE Computer Society, 2006.
7. C. Loglisci and D. Malerba. Discovering triggering events from longitudinal data. In *ICDM Workshops*, pages 248–256. IEEE Computer Society, 2008.
8. C. Loglisci and D. Malerba. A temporal data mining approach for discovering knowledge on the changes of the patient's physiology. In *Proc. of 12th Conf. on Art. Int. in Medicine*, pages 26–35, 2009.
9. D. Malerba. Learning recursive theories in the normal ilp setting. *Fundam. Inform.*, 57(1):39–77, 2003.
10. D. Malerba and F. A. Lisi. An ILP method for spatial association rule mining. In *In Working notes of the First Workshop on Multi-Relational Data Mining*, pages 18–29, 2001.
11. H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.
12. S. R. Michalski, G. J. Carbonell, and M. T. Mitchell. *Machine learning an artificial intelligence approach volume II.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1986.
13. G. D. Plotkin. A note on inductive generalization. *Machine Intelligence*, 5:153–163, 1970.
14. Y. Shahar. A framework for knowledge-based temporal abstraction. *Artif. Intell.*, 90(1-2):79–133, 1997.
15. M. Stacey and C. McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial Intelligence in Medicine*, 39(1):1–24, 2007.