Learning Rules with Complex Temporal Patterns in Biomedical Domains

Lucia Sacchi, Riccardo Bellazzi, Cristiana Larizza,
Riccardo Porreca, and Paolo Magni

Dipartimento di Informatica e Sistemistica, University of Pavia, via Ferrata 1,
27100 Pavia, Italy
lucia@aim.unipv.it
{riccardo.bellazzi, cristiana.larizza, paolo.magni}@unipv.it

Abstract. This paper presents a novel algorithm for extracting rules expressing complex patterns from temporal data. Typically, a temporal rule describes a temporal relationship between the antecedent and the consequent, which are often time-stamped events. In this paper we introduce a new method to learn rules with complex temporal patterns in both the antecedent and the consequent, which can be applied in a variety of biomedical domains. Within the proposed approach, the user defines a set of complex interesting patterns that will constitute the basis for the construction of the temporal rules. Such complex patterns are represented with a Temporal Abstraction formalism. An APRIORI-like algorithm then extracts precedence temporal relationships between the complex patterns. The paper presents the results obtained by the rule extraction algorithm in two different biomedical applications. The first domain is the analysis of time series coming from the monitoring of hemodialysis sessions, while the other deals with the biological problem of inferring regulatory networks from gene expression data.

1 Introduction

The application of data mining techniques to the medical and biological domain has gained great interest in the last few years, also thanks to the encouraging results that have been achieved in many fields [1,2]. One issue of particular interest in this area is represented by the analysis of temporal data, usually referred to as Temporal Data Mining (TDM) [3,4,5]. Within TDM, research usually focuses on the analysis of time series, collected measuring clinical or biological variables at different points in time. The explicit handling of time in the data mining process is extremely attractive, as it gives the possibility of deepening the insight into the temporal behavior of complex processes, and may help to forecast the future evolution of a variable or to extract causal relationships between the variables at hand.

An increasing number of TDM approaches is currently applied to the analysis of biomedical time series; in functional genomics, for example, clustering techniques have been largely exploited to analyze gene expression time series, in order to assess the function of unknown genes, relying on the assumption that genes with similar profiles may share similar function [6,7,8]. TDM has also been successfully used to

S. Miksch et al. (Eds.): AIME 2005, LNAI 3581, pp. 23 – 32, 2005.

study gene expression time series of particular cell lines which are crucial for understanding key molecular processes of clinical interest, such as the insulin actions in muscles [9] and the cell cycle in normal and tumor cells [10]. Several works have been proposed also for what concerns the representation and processing of time series coming from the monitoring of clinical parameters, collected for example during an ICU staying [11,12].

In this paper, we are interested into one of the most attractive applications of AI-based TDM: the extraction of temporal rules from data. Unlike association rules, temporal rules are characterized by the fact that the consequent is related to the antecedent of the rule by some kind of temporal relationship [13]; moreover, a temporal rule typically suggests a cause-effect association between the antecedent and the consequent of the rule itself. When applied to the biomedical domain, this could be of particular interest, for example in reconstructing gene regulatory networks or in discovering knowledge about the causes of a target event [4].

An interesting approach to the problem of extracting temporal rules has been presented in [14,15] where the authors, exploiting the ideas of Hoppner [13] and the well-known APRIORI algorithm [16], have defined a method for the discovery of both association and temporal rules to get an insight into the possible causes of non-adherence to therapeutic protocols in hemodialysis, through the analysis of a set of monitoring variables. The TDM approach relied basically on two phases, the first one concerning the time series representation while the second dealing with rule extraction. In particular, the time series are first summarized through qualitative patterns extracted with the technique of Temporal Abstractions; then, possible associations between those patterns and the non-adherence events are searched with an APRIORI-like procedure. The mentioned method, however, only treats rules with antecedents composed by the conjunction of simple patterns (i.e. patterns of the kind "increasing", "decreasing", …), where the conjunction is interpreted as a co-occurrence relationship (i.e. "variable A increasing" occurs at the same time of "variable B decreasing"). If this conjunction temporally precedes another simple pattern, say "variable C increasing", sufficiently often, a rule of the kind "variable A increasing and variable B decreasing precedes variable C increasing" is generated.

In this paper, we propose an extension of the method described in [14,15] in order to extract rules with arbitrarily complex patterns as members of both the rule antecedents and consequents. The data miner can define in advance such patterns, or they might be automatically generated by a complex pattern extractor. This extension is able to deal with the search of relationships between complex behaviors, which can be particularly interesting in biomedical applications. For example, a drug is first absorbed and then utilized, so that its plasma distribution precedes its effect in the target tissue. In this case, it would be important to look for complex episodes of "up and down" type in the drug plasma concentration, to automatically extract temporal knowledge in the data. Therefore, the method that we propose in this paper enables the user to define episodes of interest, thus synthesizing the domain knowledge about a specific process, and to efficiently look for the specific temporal interactions between such complex episodes.

The paper is structured as follows: we first describe the new method for the extraction of complex temporal patterns from data; then, we introduce two different

biomedical applications where the method is to provide interesting results on real data sets. Finally we discuss pros e cons of the proposed approach.

2 The Complex Temporal Rules Extraction Method

As shown in Figure 1, the method proposed in this paper develops following different steps that, starting from the raw time series and passing through different stages of representation, leads to the construction of a set of temporal rules, where both the antecedent and the consequent are made up of complex patterns.

2.1 A Formalism for Time Series Representation: The Technique of Temporal Abstractions

To be able to extract temporal rules from the data, we need first of all a suitable representation of the time series [13]. A convenient technique to extract a compact and meaningful representation of temporal data is to resort to Temporal Abstractions (TAs) [17].

TAs are an AI methodology characterized by the move from a time-stamped (quantitative) representation of temporal data to an interval-based qualitative one. In each interval, a specific pattern is verified in the data; such patterns represent a meaningful summary of the original data, and can be used to derive features that characterize the dynamics of the system under observation. Algorithms that, taking as input a time series, generate an interval series, with intervals corresponding to the time periods in which a specific pattern is present in the input, are referred to as TA *mechanisms*. TA mechanisms represent the fundamental step of TA-based analysis.

Within TAs, we can distinguish between two main categories: basic and complex abstractions. Basic TAs are used to detect simple patterns in numerical or symbolic time series. More precisely, we can extract *Trend* Temporal Abstractions, to capture increasing, decreasing or stationary courses in a numerical time series, and *State* TAs, to detect qualitative patterns corresponding to low, high or normal values in a numerical or symbolic time series. Complex TAs, on the other hand, correspond to intervals in which specific temporal relationships between basic or other complex TAs hold. These relationships are typically identified by the temporal operators defined in Allen algebra [18].

In our approach, we aim at representing the time series through a set of complex TAs, that will be denoted as *complex temporal events*; to obtain such representation, data are processed following two consecutive steps.

1. Raw time series are initially processed with a suitable TA mechanism [14] to describe the time course of a variable as a set of consecutive Basic *trend* TAs. The resulting representation is a label made up of simple qualitative elements of the kind [Increasing], [Decreasing], [Steady] (Figure 1 a) and b)).

2. The Basic TA description represents the basis for the creation of the final complex TA representation, that will be then used in the rule extraction algorithm. This representation is based on the definition of *complex temporal events*, that are complex abstractions defined as the time intervals in which specific interesting

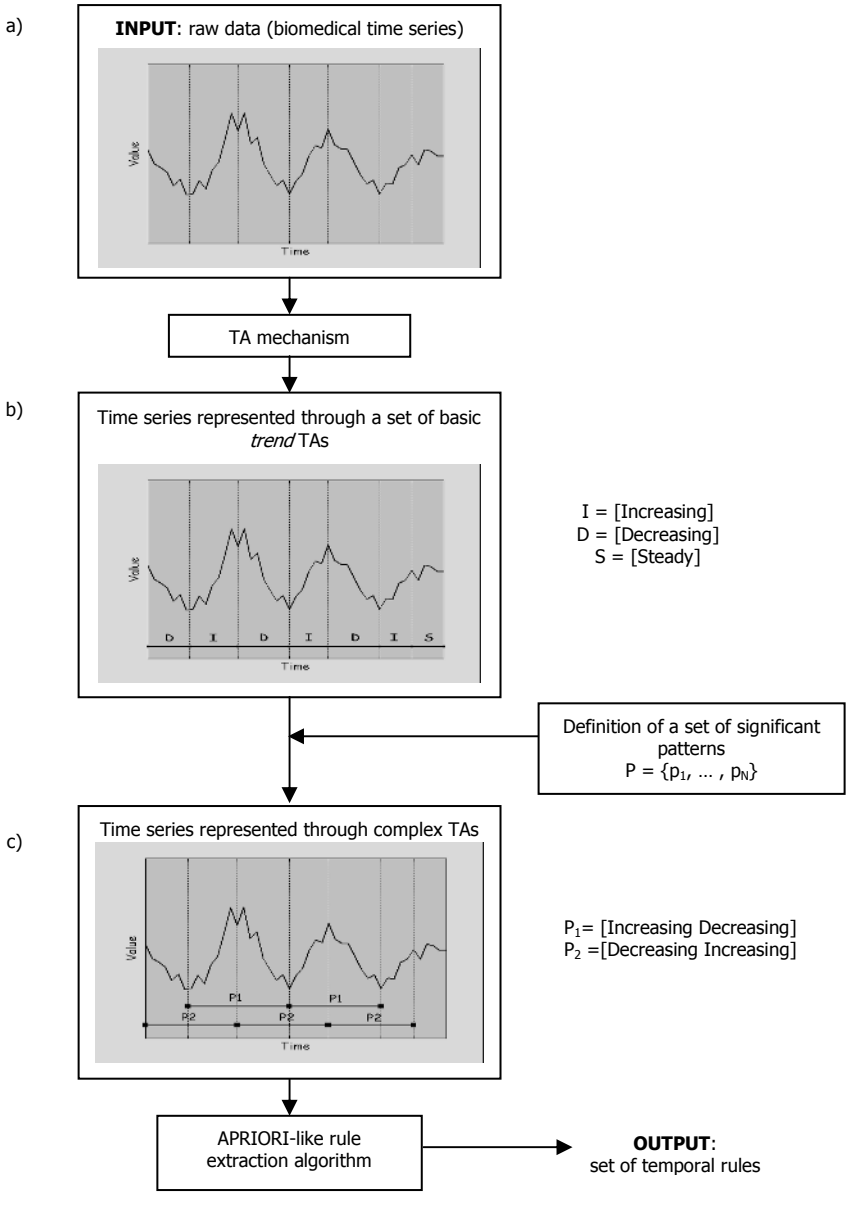


Fig. 1. The steps of the algorithm for temporal rules extraction

patterns occur in the input data. One of the core aspects of this phase is the definition of a set of significant patterns $P = \{p_1,\ldots p_n\}$, where each p_i is made up by the composition of simple labels like [Increasing], [Decreasing], [Steady]. In general, the

set P may be both user-defined or automatically suggested to the researcher after a pre-processing of the initial qualitative representation of the variables performed with an appropriate strategy, such as the one presented in [19]. Moreover, the definition of P clearly relies on the clinical/biological knowledge on the kind of relationships that may be useful to explore with respect to the analysis purposes. As an example, let us consider a situation in which there is interest to investigate if a particular peak in the dynamics of a variable V_1 is often related to an opposite peak of another variable V_2. We can formalize this problem by defining P as P={[Increasing Decreasing], [Decreasing Increasing]}. The rule extraction will then be performed only on those series that present the first pattern in V_1 and the second pattern in V_2; such rules will look for a temporal significant relationship between those two patterns.

The steps that lead from the raw data to the representation through *complex temporal events* are depicted in Figure 1 a), b) and c). The example reveals also that, in general, it is possible to find more than one example of the same pattern within a single time series and that the intervals that correspond to different patterns may overlap when dealing with the same profile.

2.2 Temporal Rules Extraction

Once we have derived a representation of the temporal profiles based on the definition of a set of complex TAs, the method for temporal rules extraction develops through a strategy which looks for both the antecedent and the consequent of the rule coming from the set of complex TAs that represent the time series. The rules extraction strategy will then look for rules in which a set of contemporaneous TAs (the antecedent) has a precedence temporal relationship with another TA (the consequent). Notice that, since temporal rules are derived through the combination of complex temporal abstractions on the basis of a temporal relationship, they can be considered themselves as complex TAs.

More formally, we consider temporal relationships expressed by the temporal operator PRECEDES, defined as follows: given two episodes, A and C, with time intervals $[a_1, a_2]$ and $[c_1, c_2]$, we say that A PRECEDES C if $a_1 \leq c_1$ and $a_2 \leq c_2$. Note that PRECEDES includes the Allen's temporal operators OVERLAPS, FINISHED-BY, MEETS, BEFORE, EQUALS and STARTS. Moreover, the PRECEDES relationship may be constrained by some parameters, that set some restrictions on the mutual position of the intervals involved [15]. These parameters are: the right shift (RS), defined as the maximum allowed distance between c_2 and a_2, the left shift (LS), defined as the maximum allowed distance between c_1 and a_1 and the gap, defined as the maximum distance between a_2 and c_1, when $c_1 > a_2$.

The search procedure is then aimed at defining rules of the kind $A \rightarrow_p C_i$, where A is a conjunction of complex TAs and constitutes the antecedent, while C_i is the consequent of the rule. The notation $\rightarrow_p$ defines the PRECEDES temporal relationship between A and C_i.

The search procedure develops following the ideas of [14,15], after having properly modified the definitions of the set from which to extract the antecedent and the consequent episodes. In more detail, in order to define confidence and support, we need first to introduce some quantities:

- TS: the time span, i.e. the total duration of the observation period in which the rule is derived;
- RTS: the rule time span, i.e. the time span corresponding to the union of the episodes in which both the antecedent and the consequent of the rule occur;
- NAT: the number of times (episodes) in which the antecedent occurs during the TS;
- NARTS : the number of times (episodes) in which the antecedent occurs during the RTS.

We can therefore define:

- Support (Sup) = RTS / TS[1];
- Confidence (Conf) = NARTS / NAT.

The strategy to extract temporal rules develops then as follows:

1. Put all the *complex temporal events* that represent the time series in the set A_0;
2. Fix the consequent as an episode $c_i \in A_0$;
3. Apply the PRECEDES operator between each a_i and c_i, where a_i is such that $a_i \in A_0 - \{c_i\}$ and a_i doesn't refer to the same time series as c_i. Put the results in the new set A_1 and set the counter k to 1. In general A_1 is formed by those rules that show a support greater then a fixed threshold;
4. Repeat:
   - Set k=k+1;
   - Generate the set A_k from A_{k-1} such that each rule in A_k has cardinality k (conjunction of k TAs in the antecedent) and verifies the PRECEDES relationship. Even in this case, it is possible to state a restriction for A_k based on a threshold on the support;

   Until: A_k is empty;
5. Put $A = A_{k-1}$ and repeat from step 2 for another consequent $c_j \in A_0$.

3 Results

In this section we show the results obtained by applying the method to two different problems, the first in the clinical domain, while the other one concerning the biological problem of inferring gene regulatory networks from data. These examples allow to understand the wide spectrum of applicability of the proposed solutions.

3.1 Analysis of Time Series Coming from Haemodialysis Sessions Monitoring

The first application we introduce is about the use of the rule extraction algorithm to analyze time series coming from the monitoring of several variables during different dialysis sessions[2]. In particular, we have considered a single patient undergoing to 86

[1] Several definitions of support can be considered; in our case, we chose to consider the real time span of the episodes, in order to take into account low frequency episodes with long TS.

[2] The data have been made available by courtesy of the Dialysis Unit of the A.O. of Vigevano, Italy.

dialysis cycles, looking for the relationships occurring between arterial pressure and heart rate. In our study, we have considered systolic pressure (SP), diastolic pressure (DP) and heart frequency (HR). From a clinical point of view, it is interesting to look for temporal relationships that highlight a negative correlation between the pressure variables on one hand and the heart frequency on the other. Such relationships may be related to hypertension or hypotension episodes. Relying on this assumption, we have identified the set P as P = {[Increasing Steady Decreasing], [Decreasing Steady Increasing]}. The rule extraction algorithm searches for rules that satisfy the PRECEDES operator between opposite *complex temporal events* in the antecedents with respect to the consequents (e.g. [Increasing Steady Decreasing] vs [Decreasing Steady Increasing]). Table 1 shows the results we obtained fixing a threshold for the confidence, Conf ≥ 0.7, and for the support, Sup ≥ 0.1.We have been able to derive rules with complex patterns involving one or more variables. The first rule extracts a contemporaneous pattern for Systolic and Diastolic pressure, in which an up and down pattern is followed by a down and up pattern of the Heart Rate; the rule has confidence 0.7 and is verified in 10 dialysis over 86. Other two similar rules, which relate Heart Rate with Systolic and Diastolic pressures are also found. In both cases the number of dialysis in which the rule is verified is 24. These episodes are clinically relevant, since they correspond to the patient response to hypertension, probably due to vasoconstriction.

Table 1. The rules derived from the analysis of the haemodialysis data ([ISD]= [Increasing Steady Decreasing], [DSI] = [Decreasing Steady Increasing])

RULE (OPERATOR: PRECEDES) P={[Increasing Steady Decreasing], [Decreasing Steady Increasing]}					
Antecedent		Consequent		Confidence	Support
Variable	Pattern	Variable	Pattern		
SP DP	[ISD] [ISD]	HR	[DSI]	0.706	0.156
HR	[DSI]	DP	[ISD]	0.755	0.398
HR	[DSI]	SP	[ISD]	0.8	0.407

3.2 Analysis of Gene Regulatory Networks Through Gene Expression Data

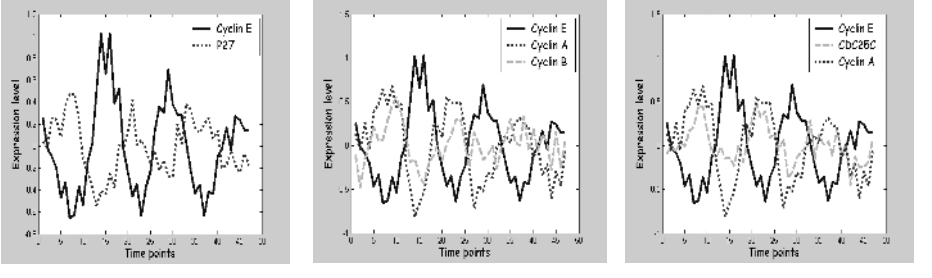
The second study regards the attractive biological problem of inferring genetic regulatory networks starting from gene expression data. In this domain, our algorithm could be particularly suited since it allows to describe patterns of synchronization and precedence in gene expressions; such patterns might be the evidence of a close relationships between genes. Moreover, by highlighting the relationships between synchronized gene sets, we can gain insight into the temporal sequence of macro-processes, potentially suggesting cause-effect relationships between the involved genes.

In this paper we have analysed the data coming from DNA microarray experiments on human cell cycle, presented in [10] and available at http://genome-www.stanford.edu/Human-CellCycle/Hela/. From the whole dataset, we extracted 5

time series of 47 samples that correspond to some of the human genes which are known to regulate cell cycle [20]. We have considered the rules characterized by pattern P={[Increasing Decreasing], [Decreasing Increasing]}. This pattern is useful to highlight the synchronization and phase shifts between genes during the cell cycle. The rules have been derived with confidence Conf =1 and support Sup ≥ 0.7. Rather interestingly, the most important known relationships between genes are automatically derived by the algorithm. Table 2 and Figure 2 show some examples related to the gene for Cyclin E, the protein which regulates the transition from the phase G1 to S. Protein P27, a Cyclin E repressor, is always in opposition to Cyclin E; this is expressed by a precedence relationship, where a peak of one gene always precedes the peak of the other. Moreover, Cyclin A and B regulate transition from phase S to G2 and M, and CDC25 is a protein which favors the transition from G2 to M. Such genes are always find to be synchronized, and their complex pattern periodically precedes Cyclin E.

Table 2. Examples of the Rules extracted from the analysis of gene expression data of human cell cycle ([ID] = [Increasing Decreasing], [DI] = [Decreasing Increasing])

RULE (OPERATOR: PRECEDES) P = {[Increasing Decreasing] [Decreasing Increasing]}					
Antecedent		Consequent		Confidence	Support
Gene	Pattern	Gene	Pattern		
P27	[ID]	Cyclin E	[DI]	1	0.915
Cyclin A Cyclin B	[ID] [ID]	Cyclin E	[ID]	1	0.745
CDC25C Cyclin A	[ID] [ID]	Cyclin E	[ID]	1	0.745



a) b) c)

Fig. 2. The time series corresponding to the temporal rules reported in Table 2. Protein P27 is always in opposition to Cyclin E (a); Cyclin A, Cyclin B and CDC25 are synchronized; their complex pattern periodically precedes Cyclin E (b,c)

4 Discussion

In this paper we have presented a new method for the automated generation of temporal rules which involve complex patterns in both the antecedent and the consequent. This algorithm is particularly suited for exploring the temporal relationships between the variables collected in different kind of biomedical data bases. It is important to note that the method performs a knowledge-based search in the data set. Knowledge is required to the domain expert to specify the parameters of the temporal abstraction mechanisms, such as the minimum slope of the trends used to define the simple patterns; moreover, additional knowledge is needed to select the interesting temporal complex patterns and to completely specify the precede temporal relationship. Finally, the rules are extracted after the choice of support and confidence. Such a knowledge-intensive procedure can be considered both a strength and a weakness of the approach. It is a strength, since it allows the extraction of results which are driven by background knowledge in an explicit way. This facilitates explanation and user control on the output. It is a weakness, since the degrees of freedom of the users are high, so that the results may be difficult to reproduce if all the parameters are not clearly reported in the data analysis process; moreover it requires a workload to the user in knowledge elicitation. In our opinion, it can be used as a useful instrument to complement data-driven approaches to gain insight in complex temporal behaviours, which are common in biomedical domains.

Given the great flexibility that characterizes the algorithm, immediate extensions may lead to consider both Trend and State TAs for the representation of the temporal profiles; moreover also different relationships between the events may be taken into account. As a future development we are including in the algorithm a strategy to automatically propose to the user the most relevant complex patterns in the data; moreover, we are working on a technique for result visualization through a semantic network of temporal relationships.

References

1. Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., Tockman, M., Clark, R.A.: Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med 32 (2004) 71-83.
2. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286 (1999) 531-537
3. Augusto, J.C.: Temporal reasoning for decision support in medicine. Artif Intell Med. 33 (2005) 1-24
4. Roddick, J.F., Spiliopoulou, M.: A survey of temporal knowledge discovery paradigms and methods. IEEE Transactions on Knowledge and Data Engineering 14 (2002) 750-767
5. Lin, W., Orgun, M.A., Williams, G.J.: An overview of temporal data mining. In: Simeoff, S., Williams, G., Hegland, M. (eds.): Proceedings of the 1st Australian Data Mining Workshop (ADM02), (2002) 83-90
6. Eisen, M., Spellman, P.T., Botstein, D., Brown, P.O.: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95 (1998)14863-14868

7. Ramoni, M., Sebastiani, P., Cohen, P.: Bayesian clustering by dynamics. Machine Learning 47 (2002) 91-121
8. Hvidsten, T.R., Lægreid, A., Komorowski, J.: Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. Bioinformatics 19 (2003)1116-1123
9. Di Camillo, B., Sreekumar, R., Greenlund, L.J., Toffolo, G., Cobelli, C., Nair, S.K.: Selection of insulin regulated genes based on array measurement error. In: Proceedings of Genomics of Diabetes and Associated Diseases in the Post Genome Era, (2003) 113
10. Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., Botstein, D.: Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13 (2002) 1977-2000
11. Salatian, A., Hunter, J.R.W.: Deriving Trends in Historical and Real-Time Continuously Sampled Medical Data. Journal of Intelligent Information Systems 13 (1999) 47-71
12. Salatian, A.: Interpreting Historical ICU Data Using Associational and Temporal Reasoning. In: ICTAI (2003) 442-450
13. Höppner, F.: Discovery of temporal patterns-learning rules about the qualitative behaviour of time series. In: Proceedings of the Fifth PPKDD, LNAI (2001) 192-203
14. Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R.: Quality assessment of hemodialysis services through temporal data mining. In: Dojat, M., Keravnou, E., Barahona, P. (eds.): Artificial Intelligence in Medicine. Ninth Conference on Artificial Intelligence in Medicine in Europe, LNAI 2780. Springer-Verlag (2003) 11-20
15. Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R.: Temporal data mining for the quality assessment of hemodialysis services. Artificial Intelligence in Medicine, May 2005.
16. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the International Conference on Very Large Databases. Morgan Kaufmann (1994) 478-499
17. Shahar, Y.: A framework for knowledge-based temporal abstraction. Artificial Intelligence 90 (1997) 79-133
18. Allen, J.F.: Towards a general theory of action and time. Artificial Intelligence 23 (1984) 123-154
19. Sacchi, L., Bellazzi, R., Larizza, C., Magni, P., Curk, T., Petrovic, U., Zupan, B.: Clustering gene expression data with temporal abstractions. In: Medinfo. (2004) 798-802
20. Tyson, J.J., Chen, K., Novak, B.: Network dynamics and cell physiology. Nat Rev Mol Cell Biol 2 (2001) 908-916