Temporal abstractions for pre-processing and interpreting diabetes monitoring time series *

Riccardo Bellazzi, Cristiana Larizza
Dipartimento di Informatica e Sistemistica
Università di Pavia
via Ferrata 1, 27100 Pavia, Italy

Alberto Riva
I.R.C.C.S. Policlinico S. Matteo
viale Golgi 1, 27100 Pavia, Italy

Abstract

In this paper we describe a number of intelligent data analysis techniques to pre-process and analyze data coming from home monitoring of diabetic patients. In particular, we show how the combination of temporal abstractions with statistical and probabilistic techniques may be applied to derive useful summaries of patients' behaviour over a certain monitoring period. Finally, we describe how Intelligent Data Analysis methods may be used to index past cases to perform a case-based retrieval in a data-base of past cases.

1 Introduction

Data pre-processing is an important step of data analysis, allowing to transform an original data set into a new, more meaningful and computationally manageable data set. More formally, following Fayad [Famili *et al.*, 1997], we can say that data pre-processing is "a transformation T that transforms the raw real world data vectors **X**, to a set of new data vectors **Y** such that: (i) **Y** preserves the information in **X**, (ii) **Y** eliminates at least one of the problems in **X** and (iii) **Y** is more useful than **X**". In this paper we will face the problem of pre-processing noisy biomedical time series by using Temporal Abstractions (TA)s. TAs are methods able to derive an abstract description of the course of longitudinal data by extracting their most relevant features [Kahn *et al.*, 1991; Shahar and Musen, 1996; Miksch *et al.*, 1995]. Although TAs have been traditionally used as the first step of automated reasoning, they can be easily applied for performing an intelligent data pre-processing. As a matter of fact, their use for solving this task is strictly related with the peculiarity of long-term monitoring of chronic patients. In this case, the monitoring data are often collected on an irregular time grid by patients themselves. The data are usually revised by the physician during periodic controls, that occur only few times in a year. During data revision it is very difficult to distinguish abnormal values from measurement errors. So, although a large amount of data is present, the information content is often quite low. In order to give a meaningful interpretation to the data, a suitable strategy is moving from the raw data to a more abstract representation of the patient's behaviour. Following this approach, we propose to extract from the original time series the relevant features of the monitoring period through TAs. In particular, we exploit TAs to pre-process the raw data and to obtain a new time series of abstract episodes, that will be further analyzed through quantitative methods for interpretation purposes.

In this paper we will discuss the application of TAs to the pre-processing of data coming from the home monitoring of patients affected by insulin-dependent diabetes Mellitus (IDDM). We will describe how to extract the episode time series and we will discuss how they can be used to derive in a coherent and sound fashion a number of indicators of the patient's behaviour. In particular we will introduce a new method for deriving well-established indicators of the degree of metabolic control, and we will propose a number of new analyses based on TAs. Finally, we will discuss how such TA-based indicators may be used to classify and intelligently retrieve past cases.

The work herein presented is part of the T-IDDM (Telematic Management of Insulin-Dependent Diabetes Mellitus patients) project, that proposes a distributed architecture for IDDM patient management [Bellazzi *et al.*, 1995]. This architecture is based on on two intelligent modules, a Patient Unit and a Medical Unit, connected by telecommunication services. The peculiarity of the applications of Intelligent Data Analysis methods in this context is described elsewhere [Bellazzi *et al.*, 1997]. In this paper we will concentrate on the methodologies applied to analyze the data at the medical unit site [Larizza *et al.*, 1997].

2 Basic features

2.1 The clinical problem

The application domain in which we are applying TA techniques is the interpretation of data coming from the home monitoring of Insulin-Dependent Diabetes Melli-

*This work is part of the EC project HC-1024, T-IDDM, Telematic Management of Insulin-Dependent Diabetes Mellitus.

tus (IDDM) patients. IDDM is caused by the destruction of the pancreatic cells producing insulin, the main hormone that regulates the glucose metabolism. IDDM patients control their glucose metabolism through the delivery of exogenous insulin several times a day. The onset and development of diabetic complications is strictly related to the degree of metabolic control that can be thus achieved [DCCT, 1993]. The patients perform self-monitoring of BGL and glycosuria (glucose in the urine) at home several times a day, usually before meals, and report the monitoring data and the therapy followed in a diary. Unfortunately, to reconstruct completely the BGL daily pattern, at least eight measurements per day would be required. This means that the BGL signal is under-sampled, and it is hence very difficult to distinguish abnormal values from measurement errors. So, although a large amount of data is present, often the information content is quite low. Moreover, the daily diary often does not contain relevant information, like diet modifications or physical exercises.

However, although the diary often contains only incomplete information, the physicians must use the few available data to assess therapeutic revisions during their routine clinical activity. Since the early 80's, several decision-support systems have been developed to assist patients and physician during the monitoring process. In particular, a number of tools have been proposed in the field of data interpretation. The methodologies used to this purpose range from simple statistical analysis and graphical representation of the raw data, to more complex techniques, like time-series analysis [Deutsch *et al.*, 1994], causal probabilistic networks [Andreassen *et al.*, 1994] and temporal abstractions [Shahar and Musen, 1996].

In this paper we explicitly cope with the problem of defining an effective system for helping data interpretation even when the information is very poor. We propose a novel approach based on the combination of TA methods with statistical and probabilistic techniques; our methodology is exploited to derive both well-known and new indicators in an integrated fashion, by soundly taking into account the incompleteness of the available data set.

2.2 Temporal Abstractions

TAs are methods that can be used to obtain an abstract description of the course of multi-dimensional time-series by extracting their most relevant features [Keravnou, 1995]. Hence, in patient monitoring, TA methods provide a useful instrument for summarizing the patient's history by transforming the fragmentary representation provided by the raw data into a more compact one based on series of abstract episodes.

The basic principle of TA methods is to move from a time-point to an interval-based representation of the data. Given a sequence of time stamped data (*events*), the adjacent observations which follow meaningful patterns are aggregated into intervals (*episodes*). In particular, to conceptualize the TA problem solving method, we defined an ontology which distinguishes two main classes of abstractions: BASIC abstractions for detecting predefined patterns in a time series, and COMPLEX abstractions, for investigating specific temporal relationships between intervals.

BASIC abstractions extract TRENDS (increase, decrease or stationarity patterns), or STATES (e.g. low, normal, high values) from a uni-dimensional time series.

COMPLEX abstractions search for specific temporal relationships between episodes which can be generated from a basic abstraction or from other complex abstractions. The relation between intervals can be any of the temporal relations defined by Allen [Allen, 1984]. This kind of TA can be exploited to extract multi-dimensional patterns or to detect uni-dimensional patterns of complex shapes.

Some examples of the use of both BASIC and COMPLEX abstractions are reported in [Bellazzi *et al.*, 1997].

3 TAs in the diabetes domain

In order to allow a proper interpretation of the data, we subdivided the 24-hour daily period into a set of consecutive non-overlapping time slices. This process is a first type of abstraction, that generates a qualitative time scale on the basis of the information about the patient's life-style, in particular the meal times. The relationships between actions (insulin intakes) and effects (BGL measurements) are derived using the concept of *competent time slice*: an action in a certain time slice will be competent for the BGL measurements in the time slices that it directly affects. For example, an intake of regular insulin will be competent for the time slices that cover the subsequent six hours.

The seven time slices used in this work are shown in the following:

Breakfast
Mid-Morning
Lunch
Afternoon
Dinner
Bed-Time
Night-Time

The measurements of BGL, glycosuria, and insulin dosages collected by the patient are always associated with a time-slice. Further information about the patient life-style, physical exercise and diet is very frequently missing. In our work, we have hence performed the analysis only on the time series of BGL, insulin and glycosuria. When detecting STATE patterns in time series of numerical variables a preliminary qualitative abstraction is carried out [Larizza *et al.*, 1995]. The mapping between the qualitative abstractions and the quantitative levels of each numerical variable depends on the time-slice and on the specific patient's characteristics. For example, the BGL normal range is wider in the morning

TA type	finding	Temporal Abstractions
STATE	BGL	Hypoglycemia Normoglycemia Hyperglycemia
	Insulin	Low Insulin Medium Insulin High Insulin
	Glycosuria	Glycosuria Absent/Traces Glycosuria Present
TREND	BGL	BGL Increase BGL Stationarity BGL Decrease
	Insulin	Insulin Increase Insulin Stationarity Insulin Decrease

Table 1: BASIC Temporal Abstractions for the *breakfast* time slice

than around lunch, and it is wider in pediatric patients than in adult ones.

Once the time slices have been generated, we have defined a set of BASIC and COMPLEX abstractions for each time slice. For example, the BASIC TAs defined for the *Breakfast* time slice are shown in Table 1.

Exploiting the BASIC abstractions and the above introduced definitions, we can characterize the patient behaviour through the concept of ABSTRACT STATE (AS), that corresponds to the combination of the TAs that are *true* in that period. We can construct the time series of the ASs along the daily time axis for each time-slice.

The general form of the abstract state in the i−th day for the j−th time-slice AS_{ij} is:

AS_{ij}={ BGL_{ij}, BGL-TREND_{ij}, COMPETENT INSULIN $(\text{C-I})_{ij}$, C-I TREND_{ij}, GLYCOSURIA_{ij}}

To explain the process of deriving the AS time series from the raw data, let's suppose to have a series of data records which are extracted from the daily diary of a patient. The general form of a record is the following:

day	BGL	C-I-dose	GLYCOSURIA

After choosing suitable parameters for the TAs (i.e. qualitative levels and corresponding ranges for STATE abstractions, minimum slope for TREND abstractions, etc.), the next step is to analyze the data and to detect the relevant episodes.

The result of this analysis is a collection of episodes for each TA of the form:

temporal abstraction	episodes list

where each episode is identified by the pair of values $[s, e]$ which represent the start and the end points of the episode detected.

Finally, from this collection of episodes, we derive the AS time series for each time slice. Each element of this series, representing an AS, is of the form:

day	BGL	BGL-TREND	C-I	C-I TREND	GLYCOSURIA

The above presented methodology proposes a two-step procedure for data pre-processing: i) the raw time-series is analyzed using TAs, moving from the original time scale to a new scale derived from the sequence of relevant patterns detected in the data; ii) a new time series of abstract episodes is obtained, going back from the new scale to the original one. The new time series contains more meaningful information than the raw one, incorporating temporal trends and, in general, it may include also multi-variable patterns.

4 Ta-based summaries in diabetes monitoring

Once that the AS time series is defined, a number of analyses can be performed. The more interesting ones are described in the following:

4.1 The Blood Glucose Modal Day

A major goal of monitoring diabetic patients is to verify that their glucose metabolism follows a daily *cyclo-stationary* pattern, modulated by meal intakes and by physical activity. To test this hypothesis, several authors suggest to extract the BGL *modal day* (BG-MD).

The BG-MD is a characteristic daily BGL pattern that summarizes the typical patient's response to the therapy in a specific monitoring period; it is usually derived as the collection of most probable BGL qualitative level in each time slice. The BG-MD is used to evaluate the protocol performance over the selected time interval, even when the information is poor (e.g. data on meals missing). Several approaches for extracting the BG-MD have been presented in the literature, from simple statistics to time series analysis [Kahn *et al.*, 1991; Deutsch *et al.*, 1994].

In our approach it is easy to derive the BG-MD by calculating the marginal probability distribution of the BGL from the AS time series. In particular, in this context it is possible to resort to a Bayesian technique described in [Ramoni and Sebastiani, 1996] that is able to deal with incomplete probabilistic models.

Let us suppose that we have a discrete variable with K states. Before starting data collection in a given period we may assign a prior probability to the occurrence of each state equal to $\frac{1}{K}$. After a certain monitoring period of N days, we collect D measurements, while $M = N-D$ data are missing. The posterior probability bounds of the occurrence of the l-th state can be calculated as:

$$p_{inf} = \frac{1+d_l}{K+N} \quad (1)$$

$$p_{sup} = \frac{1 + d_l + M}{K + N}$$

where d_l is the number of occurrences of the l-th state in the monitoring period.

We may hence define the *ignorance* in the monitoring period as:

$$\text{IG} = p_{sup} - p_{inf} = \frac{M}{K + N} \quad (2)$$

The IG is clearly proportional to the number of missing data in the monitoring period.

Since in our application we consider three BGL state TAs, *Hypoglycemia, Normoglycemia, Hyperglycemia*, K is equal to 3. As the monitoring process proceeds, the bounds on the probabilities [Riva and Bellazzi, 1996] are updated. At any time we obtain, for each time slice, an interval probability distribution over the BGL state abstractions. The modal day can be then extracted by taking the BGL states with the highest p_{inf} in each time slice.

By using the same procedure it is possible to extract the posterior distribution for each variable contained in the AS. For example, useful indicators are given by the *glycosuria modal day* and by the *control actions modal day*, calculated over the GLYCOSURIA and over the C-I, respectively.

4.2 The Blood Glucose Modal Daily Pattern

Although the BG-MD is a widely accepted instrument for summarizing the patient's metabolic control over a given monitoring period, it misses the information on the actual daily pattern, i.e. on the sequence of consecutive measurements in one day. As a matter of fact, the BG-MD only reflects the frequencies of measurements in each time-slice, and not the frequencies of patterns of consecutive measurements. Nevertheless the information derived on the BG-MD may be used as the basis to search for the more frequent daily pattern in the AS time series.

To achieve the above mentioned goal, we define the *blood glucose modal daily pattern* (BG-MDP) as the most frequent sequence of abstract states of the BGL variables in the different time slices of one day.

The BG-MDP can be derived exploiting the following steps:

- *Search parameter assessment.* The BG-MDP must be searched for in a discrete space whose dimension grows combinatorially with the number of time-slices and BGL states. In our application we have 343 combinations (7 time slices and 3 levels for BGL), so that an exhaustive search is clearly feasible. Nevertheless, there are two main reasons for avoiding such a brute-force approach:

  1) The frequency of a daily pattern is derived by joining the abstract states in each time-slice in one day.

  The joint abstract state of the i-th day is the set

  $$\text{AS}_i = \{\text{AS}_{i1}, \ldots, \text{AS}_{iJ}\} \quad (3)$$

  where J is the number of time-slices. The marginal probability of the BGL daily patterns in a monitoring period of N days is hence the joint probability of the BGL abstract levels over $\{\text{AS}_{i1}, \ldots, \text{AS}_{iJ}\}$, with $i = 1, \ldots, N$.

  The joint probability distribution can be derived using the same expression as in (1), thus providing upper and lower posterior probability bounds. If we consider all the time-slices, the missing data set will be the union of the missing data over the whole day. In other words, if we have three BGL measurements in a certain day, that day will be considered as "missing" since the complete BGL data set (seven measurements) is not available.

  Since the number of missing data will be clearly high, the joint probability estimate will be unreliable (i.e. affected by a large ignorance). This problem is crucial, since we must ensure that the joint probability distribution is not uniform over the possible daily patterns: this means that, if DP is the number of possible daily patterns, we must find at least one minimum probability bound greater than $\frac{1}{\text{DP}}$.

  2) Increasing the number of BGL states, the search space dimension rapidly increases. For example, using five levels the search space dimension is 16807.

  To overcome this limitations, we must define two parameters: the maximum allowed ignorance (MIG) in one time slice (i.e. the maximum number of allowed missing data) and the minimum probability bound for the joint probability distribution ($\bar{p_{inf}}$). In the following steps we will explain how it is possible to use these parameters to reduce the search space.

- *Time slices selection.* If we derive the BG-MD we calculate the ignorance level in each time slice; so, we include in the BG-MDP only the time slices that have an ignorance level lower than MIG.

- *Search.* Once the reduced set of daily time-slices has been defined, we can perform an exhaustive search of the daily patterns that have a lower probability bound higher than $\bar{p_{inf}}$.

- *Output list definition.* At the end of the search phase we can find a set of candidates for the BG-MDP. By definition, the BG-MDP is the most probable one; nevertheless it should be of interest to show to the physician the ordered list of candidates.

It is hence important to derive some criteria for choosing $\bar{p_{inf}}$ and MIG.
In order to ensure that the joint distribution is not uniform, a natural choice for $\bar{p_{inf}}$ is $\bar{p_{inf}} > \frac{1}{\text{DP}}$.

The number of missing data that affects the estimate of the joint probability distribution is the dimension of the set obtained by the union of the missing data of each time slice. If we denote this number as M^*, we can define the ignorance over the joint as:

$$\text{IG}^* = \frac{M^*}{\text{DP} + N} \tag{4}$$

where DP is the number of possible daily patterns.

Since MIG determines the maximum number of missing data in each time slice, we can derive the lower and upper bounds for IG^* as functions of MIG.

$$\begin{aligned} \text{IG}^*_{min} &= \text{MIG}\frac{K+N}{\text{DP}+N} \\ \text{IG}^*_{max} &= min(\frac{J\ \text{MIG}\ (K+N)}{\text{DP}+N};1) \end{aligned}$$

where J is the number of selected time-slices, K is the number of BGL states and DP has the same meaning as above.

Given the above defined bounds, it is possible to choose MIG in order to minimize the ignorance region over the joint probability distribution.

4.3 Exploiting TAs time-spans

Persistence Analysis

Interesting results can be obtained through the time span distribution of the episodes. If the time span of the BGL normoglycemic episodes have an exponential distribution, it is clear that the patient is not able to control its glucose metabolism for a long period. This kind of information is important for understanding the patient's response to a certain therapeutic protocol, as well as for classifying the quality of his/her self-management. In order to verify this hypothesis, we propose to apply some statistical methods on the time span distribution.

As a first step, we perform a non-linear least-squares fitting of the model:

$$f = \lambda e^{-\lambda t} \tag{5}$$

Once we have estimated λ, we test the hypothesis that the data follows an exponential law with λ parameter through the χ^2 statistics [Moller and Muliere, 1996]. If we discretize the time axis into h intervals, the frequencies of the episodes whose extent is within each interval can be reported in a data vector $\{f_1, \ldots, f_h\}$. We may hence calculate

$$\chi^2_{h-1} = N_e \sum_{i=1}^{h} \frac{(f_i - \theta_i)^2}{\theta_i} \tag{6}$$

where θ_i is the expected frequencies in the i-th interval under the hypothesis of exponential distribution, and N_e is the number of episodes.

The exponential distribution hypothesis is rejected with degree α if $\chi^2_{h-1} > \chi^2_{h-1}(1-\alpha)$.

Dynamic Behavior Interpretation

As shown in other papers [Bellazzi and Larizza, 1996; Larizza *et al.*, 1997], the AS time series can be used to interpret the cause-effect relationships underlying the patient's behaviour. In particular, since some data are missing, it is of interest to derive a probabilistic model expressed through the probability distribution of the next AS given the past history of abstract states [Riva and Bellazzi, 1996]. Such kind of models are able to explain the occurrence of critical situations detected during the monitoring period, and hence may be exploited to tailor the therapy to each patient. A typical example is regards the detection of the episodes related with a hypoglycemic event (e.g. a wrong therapy implementation).

Unfortunately, in the general case, the use of the AS in this context is quite complex, since the probabilistic model underlying the AS evolution is intrinsically non-Markovian: the state variables are represented with intervals whose different extents implicitly contain information on their past course. To obtain a Markov model it is necessary to resort to an augmented AS, denoted as AS^*, that includes also the information of the extent of each state variable calculated at that time.

We can hence obtain a new time series of AS^*: the AS^* of the i-th day for the j-th time slice can be written as:

$\text{AS}^*{}_{ij}$={ BGL_{ij}, BGL TIME-SPAN_{ij}, BGL-TREND_{ij} , BGL-TREND TIME-SPAN_{ij}, C-I_{ij}, C-I TIME-SPAN_{ij}, C-I TREND_{ij}, C-I TREND TIME-SPAN_{ij}, GLYCOSURIA_{ij}, GLYCOSURIA TIME-SPAN_{ij}}

where the time spans of the variables represent the current time extent of each episode.

In this setting, the problem of deriving the needed conditional probability distributions from the data is hampered by the "curse of dimensionality" problem, since the number of conditionals grows combinatorially with the number of variables involved in the model and with the number of their levels. It is hence necessary to derive a simplified model, that conjugates the need of forecasting the patient's behavior with the possibility of deriving the needed probability distributions from the available data. Moreover, in order to obtain a manageable model, also the time-span of the variables must be abstracted into a few qualitative levels, that can be automatically derived from the time-span distribution analysis.

Of course, also this solution requires a very high number of conditionals. So, it is necessary to resort to marginal models, able for example to interpret the next BGL given a subset of the extended AS^*.

A suitable solution for increasing the amount of data in the analysis is to consider the data coming from a homogeneous population of patients. In this case, it is possible to apply a Bayesian hierarchical model for conditional probability learning, able to preserve the patient's individuality while borrowing strength from population data (see [Riva and Bellazzi, 1996]).

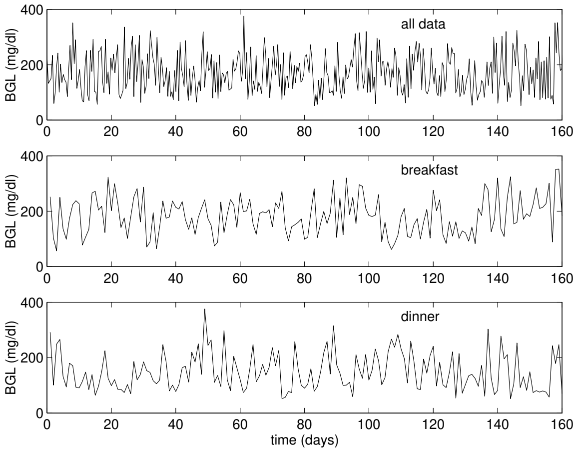

Figure 1: BGL monitoring data

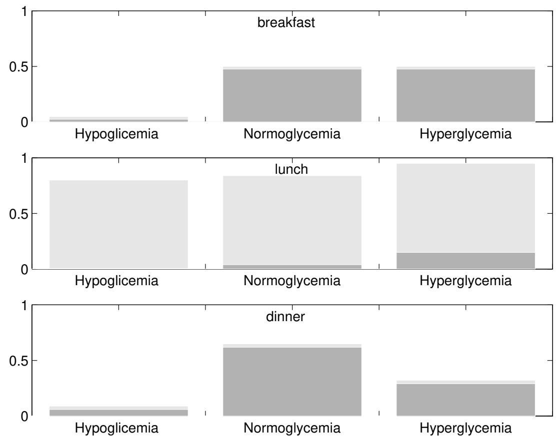

Figure 2: The BGL modal day

5 Data analysis example

In order to better explain the herein proposed methodologies we will consider a fourteen years-old female patient, monitored for a period of 165 days (about six months). The BGL data are shown in Fig. 1

The patient followed a therapeutic protocol with an insulin injection at breakfast and an insulin injection at dinner; so, the BGL time series mainly contains measurements in these two time-slices, while very few data are available for the lunch time slice.

More precisely, the data available over the monitoring period are reported in the following table:

Time slice	Breakfast	Lunch	Dinner
# of available data	161	30	160
# of missing data	4	135	5

It is now possible to derive the AS time series for each time slice. The AS is composed in each time slice as

DAY	BGL	BGL-TREND	NPH	NPH-TREND	GLYCOSURIA

where NPH is a type of insulin with a delayed effect. From this time series it is easy to derive the patient's BG-MD, as reported in Fig. 2.

The histograms of figure 2 show the minimum (heavy grey) and maximum (light grey) probability bounds for each BGL level. The lunch time slice is affected by a high ignorance, so that the derived BG-MD is clearly unreliable. On the contrary, the breakfast and dinner BG-MD provide a precise estimate of the patient's behaviour. Moreover, the dinner BG-MD has a single mode in the BGL distribution, while in the breakfast time-slice there are two states with the same bounds of probability; in this case the information provided by the BG-MD is ambiguous. This problem can be conveniently solved resorting to the BG-MDP analysis.

By fixing the MIG to 40%, the lunch time slice is excluded from the BC-MDP analysis. We now have eight possible patterns (two time slices and three levels); since DP $= 8$ the minimum probability bound is $p_{inf}^{-} = 0.125$.

In the following table the most probable BG-MDP (lower bounds) are reported in decreasing order:

pattern		probability
BREAKFAST	DINNER	
Hyperglycemia	Normoglycemia	0.35
Normoglycemia	Normoglycemia	0.29
Normoglycemia	Hyperglycemia	0.16
Hyperglycemia	Hyperglycemia	0.13

The ignorance of the estimates is 0.053, since $M^* = 9$, DP $= 8$, $N = 165$, so IG$^* = \frac{9}{8+165}$. With this method it is possible to identify with a high degree of accuracy the modal patient dynamics.

As mentioned in the previous section, it is important to verify whether the time span distribution of normoglycemic episodes is characterized by an exponential law. This index is useful for evaluating the stability of metabolic control. We will consider the episodes along the whole day.

Fig. 3 shows the time span distribution of normoglycemic episodes in the patient under consideration, as well as the derived exponential fit. The χ^2 statistics allow us to reject the hypothesis of exponential distribution (with a 99% degree of confidence), since there is a significant number of episodes with length of more than 4 days.

Finally, from the AS* (see previous section) we can derive a probabilistic model to describe the patient's behaviour given a certain TA state and its time span. In particular, for this example, we have derived a model in which we considered only the BGL state and its extent to explain the next BGL level. The time span of each BGL state has been abstracted into three levels (Low, Medium, High).

In the following table we reported the lower bound of the conditional probability tables learned from the patient's data in the dinner time slice.

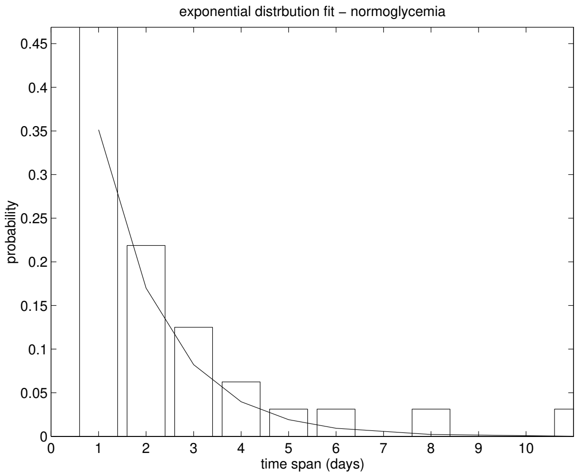

Figure 3: The time span distribution of normoglycemic episodes in the patient under consideration.

measured BGL = Hypoglycemia			
Next BGL Time Span	Hypo-glycemia	Normo-glycemia	Hyper-glycemia
Low	0.1818	0.3636	0.4545
Medium	0.2500	0.5000	0.2500
High	0.3333	0.3333	0.3333

measured BGL = Hypoglycemia			
Next BGL Time Span	Hypo-glycemia	Normo-glycemia	Hyper-glycemia
Low	0.0571	0.6000	0.3429
Medium	0.0435	0.5652	0.3913
High	0.0667	0.7333	0.2000

measured BGL = Hypoglycemia			
Next BGL Time Span	Hypo-glycemia	Normo-glycemia	Hyper-glycemia
Low	0.0882	0.5588	0.3529
Medium	0.0769	0.5385	0.3846
High	0.2857	0.4286	0.2857

An interesting information that can be derived from this model is the high probability of the Normoglycemic episodes in every table, reflecting the good quality of control achieved by the patient in the dinner time slice. Moreover, the time span of a Normoglycemic episode seems to be a good index that future measurements will be still Normal (the corresponding probability is 0.7333). Other applications of this technique can be found in [Bellazzi and Larizza, 1996].

6 Extensions: Case Indexing through TAS

The TA-based summaries introduced in the previous sections may help to define some indexes that can be used to classify patients and retrieve them in a case-based reasoning system [Slade, 1991]. Herein we propose a preliminary strategy for performing such task, composed by the following steps:

- *Case definition in the diabetes domain.*

  A *case* is a context in which the physician has taken a decision on the patient's therapy on the basis of the data collected over a certain monitoring period. The time evolution of a patient may hence generate several cases, that may significantly differ from each other. In a decision support system (not necessarily a case-based reasoner), it could be interesting to retrieve cases that are in some sense similar to the current one. TAs may be exploited to summarize the collected time-dependent data, and hence to provide an index of similarity between cases.

- *Indexing features.*

  We propose to index the cases on the basis of two classes of features:

  - *simple features.*
    The patient's identity, sex, age, weight, therapeutic protocol (as formally derived in [Riva and Bellazzi, 1995]), and glucosylated hemoglobin seem the natural indicators to be used. Each of such features can be codified with a single identifier (number, string).
  - *derived features.*
    In order to summarize the decision context, we propose to include the following TA-based summaries:
    BG-MD, BG-MDP, *Frequency of Hypoglicemias, Exponential distribution test.*
    The BG-MD and BG-MDP can be codified using arrays of dimension equal to the number of time-slices, each element being the state of the variable in the corresponding time slice. The *Exponential distribution test* is the result of the test introduced in Section 4.3 for verifying whether the time-span distribution of the Normoglycemic events follows an exponential law. This feature is hence only a yes/no value.

- *Case retrieval*

  Sex, age and therapeutic protocol may be used to select a first subclass of cases. The degree of similarity of the BG-MD and of the BG-MPD can be used to perform a further selection, exploiting a threshold over the similarity scale. For example, if we define a distance between the current case and a generic case in the j-th time slice as

  $$d_j = |\text{BG-MD}_{generic\ case}(j) - \text{BG-MD}_{current\ case}(j)|$$

  then the degree of similarity is

  $$\frac{1}{\sum_{j=1}^{J} d_j}$$

  A suitable choice for the threshold is 1/2, meaning that no more than two slices can be different for 1 level or that only one slice can differ for two levels. Clearly, we have supposed that the BGL state are codified with an ordered scale.

We can finally retrieve all the remaining cases; the case list is ordered on the basis of a similarity degree calculated over the remaining features. We are now testing several metrics for defining the relative distance of each feature, as well as for calculating the similarity degree, in order to properly weight cases that come from the current patient history.

7 Conclusions

In this paper we have described an approach to preprocess and interpret data coming from the home monitoring of diabetic patients using a combination of TAs and statistical and probabilistic techniques. In particular we have shown how it is possible to derive useful indicators of patients' metabolic control by moving from the original time-series to a new time series, built using TAs. Moreover, we have derived new indexes for the evaluation of the monitoring period, able to describe the typical patient day (the modal daily pattern); in this case we have applied a recently proposed Bayesian method for probability learning, able to explicitly take into account the incompleteness of the data set, through the notion of ignorance. Furthermore we have proposed a methodology for the analysis of the episodes time span, able to identify in a compact fashion the stability of the time series. Finally we faced the problem of exploiting causal relationships between episodes to derive suitable forecasting models. In the very last part of the paper, we have described a possible extension of the use of TA pre-processing, for defining an index of the similarity of past cases with the current one.

We believe that the proposed approach could be useful in every medical problem that involves long-term monitoring of chronic patients, in which a huge amount of data is present, with a low information content.

Acknowledgments. The authors sincerely thank Dr. Stefano Fiocchi for providing us the necessary medical knowledge and Prof. Mario Stefanelli for his methodological support.

References

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

[Andreassen *et al.*, 1994] S. Andreassen, J. Benn, R. Hovorka, K.G. Olesen, and E.R. Carson. A probabilistic approach to glucose prediction and insulin dose adjustment: description of metabolic model and pilot evaluation study. *Computer Methods and Programs in Biomedicine*, 41 (1994) 153-165.

[Bellazzi *et al.*, 1995] R. Bellazzi, C. Cobelli, E. Gomez, and M. Stefanelli. The T-IDDM project: Telematic management of Insulin Dependent Diabetes Mellitus. Health Telematics '95, M. Bracale, F. Denoth eds. (1995) 271–276.

[Bellazzi and Larizza, 1996] R. Bellazzi and C. Larizza. Interpreting longitudinal data through temporal abstractions: an application to diabetic patients monitoring. In: *Second International Symposium on Intelligent Data Analysis*, Birbeck College, London, 1997.

[Bellazzi *et al.*, 1997] R. Bellazzi, C. Larizza, and A. Riva. Cooperative Intelligent Data Analysis: an application to diabetic patients management, in: *Intelligent Data Analysis in Medicine and Pharmacology*, N.Lavrač, E. Keravnou, B. Zupan eds., Kluwer, 1997.

[DCCT, 1993] The Diabetes Control and Complication Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *The New England Journal of Medicine*, 14-329, (1993) 977–986

[Deutsch *et al.*, 1994] T. Deutsch, E.D. Lehmann, E.R. Carson, A.V. Roudsari, K.D. Hopkins, and P. Sönksen. Time series analysis and control of blood glucose levels in diabetic patients. *Computer Methods and Programs in Biomedicine*, 41 (1994) 167–182.

[Famili *et al.*, 1997] A. Famili, W. Shen, R. Weber, and E. Simoudis. Data Preprocessing and Intelligent Data Analysis. *Intelligent Data Analysis Journal*, Vol.1, 1997.

[Kahn *et al.*, 1991] M.G. Kahn, C.A. Abrams, M.J. Orland, J.C. Beard, S.B. Cousin, J.P. Miller, and J.V. Santiago, Intelligent computer-based interpretation and graphical presentation of self-monitored blood glucose and insulin data. *Diab. Nutr. Metab.*, 4 (1991) 99-107.

[Keravnou, 1995] E.T. Keravnou. Modelling medical concepts as time objects, *Lecture Notes in Medical Informatics*, P. Barahona, M. Stefanelli and J. Wyatt eds., Springer Verlag, Berlin (1995) 67–90.

[Larizza *et al.*, 1995] C. Larizza, G. Bernuzzi, and M. Stefanelli. A General Framework for Building Patient Monitoring Systems. *Lecture Notes in Artificial Intelligence*, P. Barahona, M. Stefanelli and J. Wyatt eds., Springer Verlag, Berlin (1995) 91–102.

[Larizza *et al.*, 1997] C. Larizza, R. Bellazzi, and A. Riva. Temporal abstractions for diabetic patients management. *Lecture Notes in Artificial Intelligence*, E. Keravnou, C. Garbay, R. Boud and J. Wyatt eds., Springer Verlag, Berlin (1997) 319–330.

[Miksch *et al.*, 1995] S. Miksch, W. Horn, C. Popow, and F. Paky. Therapy Planning Using Qualitative Trend Descriptions. In: *Lecture Notes in Artificial*

Intelligence, P. Barahona, M. Stefanelli and J. Wyatt eds., Springer Verlag, Berlin (1995) 197–208.

[Moller and Muliere, 1996] F. Moller, P. Muliere. Introduzione all'Inferenza Statistica. Gianni Luculiano Editore, Pavia, 1984.

[Ramoni and Sebastiani, 1996] M. Ramoni and P. Sebastiani. Robust Learning with missing data. *Knowledge Media Institute Report*, The Open University, KMI-TR-28, 1996.

[Riva and Bellazzi, 1995] A. Riva, R. Bellazzi. High level control strategies for diabetes therapy. In: *Lecture Notes in Artificial Intelligence*, 1995, P. Barahona, M. Stefanelli and J. Wyatt eds., Springer Verlag, Berlin, 185-196.

[Riva and Bellazzi, 1996] A. Riva and R. Bellazzi. Learning Temporal Probabilistic Causal Models from Longitudinal Data. *Artificial Intelligence in Medicine*, 8 (1996) 217–234.

[Shahar and Musen, 1996] Y. Shahar and M.A. Musen. Knowledge-Based Temporal Abstraction in Clinical Domains. *Artificial Intelligence in Medicine*, 8 (1996) 267–298.

[Slade, 1991] S. Slade. Case-Based Reasoning: a Research Paradigm. *AI Magazine*, 12 (1), Spring 1991, 42–55.