Intelligent analysis of clinical time series: an application in the diabetes mellitus domain


Dipartimento di Informatica e Sistemistica, Università di Pavia, via Ferrata 1, 27100 Pavia, Italy

Received 3 December 1999; received in revised form 25 February 2000; accepted 27 March 2000

Abstract

This paper describes the application of a method for the intelligent analysis of clinical time series in the diabetes mellitus domain. Such a method is based on temporal abstractions and relies on the following steps: (i) ‘pre-processing’ of raw data through the application of suitable filtering techniques; (ii) ‘extraction’ from the pre-processed data of a set of abstract episodes (temporal abstractions); and (iii) ‘post-processing’ of temporal abstractions; the post-processing phase results in a new set of features that embeds high level information on the patient dynamics. The derived features set is used to obtain new knowledge through the application of machine learning algorithms. The paper describes in detail the application of this methodology and presents some results obtained on simulated data and on a data-set of four diabetic patients monitored for > 1 year. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Intelligent data analysis; Diabetes mellitus; Temporal abstractions; Time series analysis

1. Introduction

Intelligent data analysis (IDA) is a new research field, mainly related to developing and applying methods that automatically transform data into information through the exploitation of the background knowledge available on the domain [3,22]. This approach seems particularly useful in medicine, where the association of a precise meaning to the data is often related to the recognition of the ‘context’ in
which the data have been collected. The use of knowledge is also important to properly analyze data in situations in which uncertainty plays a major role and/or when the number of available data is low, due to ethical or cost reasons; the exploitation of knowledge allows the convenient extraction of useful information from each single datum. IDA, therefore, can be viewed as a fundamental step in the knowledge management process within hospital information systems — it provides methods for exploiting existing explicit knowledge to transform data into information. As a natural consequence, IDA collates methodological contributions that come from several disciplines — from AI to Bayesian statistics and from cognitive science to mathematical modeling.

In this paper, we will describe the application of IDA techniques to the problem of analyzing and interpreting time series (TS) coming from the long-term monitoring of chronic patients. The analysis of multi-variate TS is a ubiquitous problem in science and represents a crucial challenge in biomecine applications, such as clinical monitoring, where several parameters must be contemporaneously examined to understand the patient’s overall situation. This rather complex task has traditionally been faced with descriptive and inferential statistical techniques [13]; within the IDA context, an AI-based methodology, known as temporal abstractions (TAs), has been proposed and successfully exploited in several application domains [4,16,21,28]. The principle of TA methods is to move from a time-point to an interval-based representation of the monitoring data: the time-stamped raw data are aggregated into intervals on the basis of a certain number of conditions, correspondent to the definition of a particular ‘abstract’ episode.

Therefore, if we look for different episodes, we obtain an ‘abstract description’ of (multi-variate) time-stamped data, that contains the patterns considered useful for a correct interpretation of the dynamics of the system under observation. A detailed presentation of this methodology can be found in Ref. [28].

Once a collection of TAs have been obtained, they provide a powerful high level description of the patient’s behavior. TAs can be used to ‘mine’ data collected over time, performing analysis at different levels of abstraction, aggregation and granularity.

In many application domains, a robust application of TAs may be hampered by the presence of noise on the data and in these situations, the derived abstractions can be highly dependent on the values assumed by the parameters that define the episodes.

Several authors have, therefore, found it useful to ‘pre-process’ the data in order to obtain more robust abstract episode calculations. For example, the idea of applying noise reduction techniques to the original TS in combination with TAs has been applied in the monitoring of intensive care unit patients [16]. Of course, the pre-processing techniques can be correspondent to classical data validation and outliers detection processes, but may also include more complex goals, such as the extraction of trend or periodic patterns from raw data.

Statistics may be also used to ‘post-process’ TAs, in order to obtain high level summaries of the patient dynamics over a certain monitoring period. For example, it would be easy to know how many ‘increasing’ episodes a certain variable had
during a patient’s follow-up, their average duration and the percentage of duration in the overall period. Under this perspective, the post-processing of TAs can be viewed as the computation of abstracted descriptive statistics: the number, duration and type of TA episodes can be considered as a summary of the time series at an abstract level. An interesting research direction, therefore, is to investigate whether these summaries could also be used to automatically learn some characteristics of the dynamic behavior of the patient under study. The description of the patient characteristics, at an abstract level, should allow the extraction of regularities and understand similarities that could be difficult or nearly impossible to derive from raw data.

On the basis of the above mentioned considerations, we may summarize the basic steps of the TA-based analysis in chronic patients monitoring through the general scheme shown in Fig. 1.

Summarizing such a scheme, from raw data, we obtain a pre-processed data set through the application of suitable filtering techniques; by applying TA mechanisms, we obtain a new ‘abstract level’ of episodes, that, thanks to proper presentation and visualization techniques, can conveniently help the user in ‘interpreting’ the patient’s behavior. The TA-based post-processing allows new features to be obtained that embed high level information on the patients dynamics; such features can be used to derive ‘new’ knowledge through the application of machine learning algorithms.

The aim of this paper is to present the application of the above presented scheme to the problem of analyzing data coming from home monitoring of type I diabetic patients. In this paper, we will describe the methods used in each step and we will present some results. In particular, we will report an assessment study done on a simulated patient and we will show the results obtained on four real patients, who have been monitored for > 1 year at the Policlinico S. Matteo Hospital of Pavia. Such patients have been enrolled within the telemedicine project — Telematic Management of Insulin Dependent Diabetes Mellitus’ (T-IDDM), funded by the European Commission. T-IDDM has been devoted to providing patients and

![Fig. 1. The TA-based analysis scheme.](image-url)
physicians with an information technology infrastructure for better diabetes management. In this project, the physician relies on a set of distributed web services, provided by a medical workstation. The approach described in this paper is part of the data analysis and visualization tools that are linked with the data-management and decision support modules of the whole system. For further details see Ref. [5].

2. Background

2.1. The application domain: a short summary

Diabetes mellitus is one of the major chronic diseases in industrialized countries. Its relevance (≈ 5%) in the European population and its related costs, force the health care institution towards the improvement of the treatment quality; rather interestingly, information technology has been recognized as one of the potential means for obtaining such improvement [23]. In particular, insulin-dependent (IDDM) patients (≈ 10% of the total diabetic population) are required to undergo intensive treatment to increase their life-expectancy [31]. This treatment is composed of several (from three to four) insulin injections per day and a careful blood glucose level (BGL) self-monitoring before (and sometimes after) each meal. Patients are required to collect BGL, insulin dosages, meals intakes, physical exercise and occurrence of events that may affect glucose metabolism (e.g. fever). All these data are evaluated by physicians every 2–4 months in order to assess the status of the patient’s glucose metabolism and to revise the insulin therapy.

In particular, diabetologists judge a certain therapy scheme by evaluating the number of potentially dangerous situations, such as hypoglycemias (very low BGL values) and hyperglycemias (very high BGL values). Moreover, they try to understand if there is a regularity in the occurrence of these episodes at a certain time of the day, in order to highlight potential mistakes in the insulin dose distribution. Finally, they use a laboratory test, the glycated hemoglobin, that reflects the average BGL over the last 2 months, for assessing the general control quality. Physician expertise and knowledge about the patient habits are sometimes used to forecast variations in the metabolic behavior: for example, if it is known that a certain patient, during summer-time, has a higher risk of hypoglycemias related to an increased physical activity, the overall amount of insulin is conveniently reduced in advance.

Since the early 1980s, several systems have been proposed to help patients and physicians in data collection, data analysis, decision support and, more recently, in a telematic management of the disease [5]. Nevertheless, the analysis of data coming from home monitoring of IDDM patients still remains a rather complex task. A wide spectrum of approaches have been proposed in the literature [2,13,20,23]. One of the main difficulties is related to the problem that, in real clinical practice, often the only available data are the BGL measurements that may be automatically down-loaded from blood glucose reflectometers. This practical limitation has led to the definition of decision support tools that are mainly based on the BGL TS
analysis [12]. For these reasons, in the following we will describe some techniques based on TAs for analyzing the two most common TS available in the diabetes domain: BGL and glycosuria (glucose in the urine).

2.2. Temporal abstraction

The problem solving method underlying TAs is based on an explicit ontology and a model of time adapted from Ref. [28] and described in detail in Ref. [4]. The principle of the TA method is to move from a time-point to an interval-based representation of longitudinal data. In our data model, all clinical data (measures of clinical parameters, e.g. BGL or glycosuria and actions, e.g. insulin injection) are time-stamped entities, called events, while TAs, which refer to situations persistent over time periods, are represented with intervals, called episodes.

The TA task is decomposed into two subtasks, each one solved by a specific mechanism: basic TA — solved by mechanisms that abstract time-stamped data into intervals (input data are events and outputs are episodes); complex TA — solved by mechanisms that abstract intervals into other intervals (input and output data are episodes).

Basic TAs aggregate events (time-stamped data) into episodes (intervals) by detecting clusters of adjacent observations falling within a specific set of qualitative levels or showing definite patterns. In particular, state TAs are defined to detect episodes associated with qualitative levels of time-varying variables, like hypoglycemia or hyperglycemia, while trend TAs detect patterns like increase, decrease and stationarity in a numerical time series. Each TA mechanism requires the setting up of several parameters to give a complete specification of the episode in dependence of the characteristics of the application [4]. Some parameters are common to all TAs, while others are specific for the particular TA. Examples of the parameters that have to be specified for all TAs are the ‘granularity’, that defines the maximum gap that allows two measurements to be aggregated in the same episode and the ‘minimal extent’, that specifies the minimal time-span of an episode to be considered relevant in that context. An example of a TA specific parameter is represented by the ‘slope’ of the trend abstractions, which defines the minimum increase or decrease rate to be detected.

Complex TAs aggregate episodes into more abstract ones through the application of the Allen temporal logic operators [1]. Within this paper, we exploit complex abstractions to detect two different metabolic events: the ‘Somogyi effect’ and the ‘metabolic instability’. The Somogyi effect, defined as a response to hypoglycemia while asleep with counter-regulatory hormones causing morning hyperglycemia, is detected by looking for ‘hyperglycemia at breakfast with absence of glycosuria’. A complex abstraction, called ‘suspected Somogyi effect’, is therefore defined as a time interval in which an episode of ‘hyperglycemia’ overlaps an episode of ‘glycosuria absent’ at breakfast. The overlapping is detected by applying the corresponding Allen’s temporal operator. ‘Metabolic instability’ occurs when high oscillations are present in the BGL time series. The metabolic instability TA is, therefore, defined as a sequence of ‘increasing’ and ‘decreasing’ BGL trends. Such
3. Intelligent data analysis in diabetes

As stated in the introduction, the goal of this paper is to show how the overall process of TA-based data analysis works when applied to a real problem. As shown in Ref. [4], several TAs can be derived for the analysis of IDDM patients data; however, in this paper we will describe in detail only the TAs useful to obtain the high level summaries that we exploited in the knowledge extraction step (Fig. 1).

We will start describing the pre-processing step, performed in our setting by relying on a new data filtering technique.

3.1. Pre-processing with structural filtering

An interesting way to judge the outcome of a certain therapy scheme starting from the analysis of BGL TS, is to check if it follows a cyclo-stationarity behavior, i.e. if the daily course of glycemia is approximately the same over the monitoring time. A cyclo-stationarity behavior is therefore characterized by the absence of significant trends (stationarity) and by a periodic (with period equal to 1 day) course of BGL. The characteristic daily BGL pattern that summarizes the typical patient’s response to the therapy is called ‘modal day’ and is usually derived by the frequency histograms of BGL measurements in the different times of the day (see Refs. [4,13,20,29] for a detailed discussion).

Looking for modal days and trends can be viewed as a search for a prototypical structure in the data and can be faced with a TS technique known as structural analysis. Structural filtering has been proposed in the diabetes field by Deutsch et al. [13] and their experience motivated our work in this field. By itself, structural analysis is able to provide a collection of TS’s that express the components of the original one. It, therefore, performs a particular kind of pre-processing.

The basic assumption of structural TS analysis is that each measurement of the predicted variable can be expressed as a sum of separate components, that represent its underlying structure.

In the case of BGL TS, the structure can be chosen as a composition of a trend component \(T\), a cyclic component \(C\) and a stochastic component \(v\), so that, for each measurement \(BGL_i\):

\[
BGL_i = T_i + C_i + v_i
\]  

(1)

The goal of the TS analysis is then, starting from \(BGL_i\), to extract \(T_i\) and \(C_i\). This filtering operation can be done by resorting to a variety of approaches, comprising Kalman filtering and least squares fitting.
In our approach we were interested in extracting new (smoothed) time series to be analyzed by TAs. Therefore, we looked for algorithms able to detect local trend and local cycles, to further interpret. In particular, we have exploited a general approach for Bayesian signal reconstruction described in Ref. [6].

In order to detect local trends, the $T$ dynamics is described by introducing an additional variable ($S$) that represents the random variation of $T$ from one measurement to the next, so that $T_{i+1} - T_i = S_i$. If we assume that the $S_i$ time course is described by a Markov chain, the time evolution of the $T$ component can be specified by the probability distribution $P(S_i|S_{i-1})$.

The $C$ dynamics requires a more complex model [15]. At each measurement time, $C_i$ is seen as a linear composition of a sine and cosine wave, with period 1 day, so that, if for example, there are three measurements per day, the frequency ($f$) is 1/3. The model for $C$ is hence:

$$C_{i+1} = C_i \cos(2\pi f) + R_i \sin(2\pi f)$$
$$R_{i+1} = -C_i \sin(2\pi f) + R_i \cos(2\pi f)$$  \hspace{1cm} (2)

The randomness of such model can be introduced by supposing that the $R_i$ component is a stochastic variable. Given Eq. (2), the system evolution is described by the probability distribution $P(R_{i+1}|R_i, C_i)$.

By assuming that:

$$P(S_i|S_{i-1}) = N(S_{i-1}, \sigma_s^2)$$
$$P(R_{i+1}|R_i, C_i) = N(-C_i \sin(2\pi f) + R_i \cos(2\pi f), \sigma_r^2)$$
$$P(BGL_i|T_i, C_i) = N(T_i + C_i, \sigma_B^2)$$

where, $N(\cdot, \cdot)$ denotes the normal distribution, the problem stands in estimating the couples $S_i, T_i$ and $R_i, C_i$, together with the process variances $\{\sigma_s^2, \sigma_r^2, \sigma_B^2\}$ given the observations BGL. Such a problem cannot be solved in closed form, so it is necessary to resort to an iterative estimation method, known as Markov chain Monte Carlo (MCMC). MCMC methods are based on two steps: a Markov chain and a Monte Carlo integration. By sampling from suitable probability distributions, a Markov chain, that converges (in distribution) to the target distribution, i.e. the distribution to be integrated, is generated. Then, the expectation is calculated through Monte Carlo integration over the obtained samples. As sampling strategy, in our work we used the Gibbs sampling scheme, proposed originally in Ref. [14]. This method allows the derivation, together with the point estimates and the confidence intervals of the variables; moreover, with respect to standard Kalman filtering, the method allows the estimation of the ‘process’ prior statistics, namely $\sigma_s^2, \sigma_r^2$. For a more detailed discussion see Refs. [7] and [9] for technical details.

The final outcome of the Bayesian machinery presented above is hence the extraction of two new TS ($T$ and $C$), from the BGL TS. Such TS express, at each measurement time, the trend and cycle components. Rather interestingly, this model can be easily represented and solved by using a dynamic Bayesian network [11], as described in Ref. [6].
3.2. Performing temporal abstractions

As mentioned before, in the analysis of diabetic patients data we extracted TA episodes both from original data and from the TS obtained through the pre-processing phase. A detailed description of the analysis performed on raw data can be found in Ref. [4].

For the purpose of TA-based summary extraction, we considered only the ‘state’ abstractions performed on the raw BGL, resulting in hyper and hypoglycemias episodes and the ‘complex’ abstraction related to the detection of the Somogyi effect.

Moreover, we exploited some basic TAs on the pre-processed data, i.e. the two structural components of the BGL TS. In particular, we applied the following analyses:

1. $T$ has been analyzed by applying the trend TA mechanism. The final results of this step of the analysis are the intervals corresponding to the periods of relevant BGL increase or decrease.

2. The $C$ component interpretation deserved the following procedure:
   2.1. The monitoring period has been analyzed to select the intervals where $C$ can be considered a significant component of the original TS. This task was performed by searching stationarity patterns in the $C$ TS. The related TA mechanism aggregates adjacent observations giving rise to oscillations with amplitude lower than a threshold, determined on the basis of the available physiological knowledge (in our case 20 mg/dl). The episodes, so extracted, can be interpreted as periods without relevant BGL cyclic patterns. The method discharges these episodes from the successive processing and focuses further analysis on the remaining intervals.
   2.2. Over the selected periods, a BGL cyclic pattern was extracted for each day. It was derived as the list of daily time measurements arranged so that the corresponding BGL level is in decreasing order (e.g. if, given three measurements per day, the maximum BGL measurement is at lunch and the minimum is at breakfast, the pattern is < lunch, dinner, breakfast >).
   2.3. The days with the same BGL cyclic patterns were searched and aggregated with a state TA mechanism in order to check the persistence of each pattern.

The outputs of this phase are the collection of episodes that express the local trends during the monitoring period and the intervals corresponding to the occurrence of all the possible cyclic patterns.

Since the trend TS expresses more reliably the metabolic trend of BGLs, we applied the ‘metabolic instability’ TA to $T$ (Section 2.2).

3.3. Post-processing techniques

From the TA analysis we obtain a large number of episodes that can be used to characterize the patient behavior. Referring to Fig. 1, once the data have been pre-processed and the TAs are obtained, useful information is made available to
final users. The first way to reach this goal is to show, with proper visualization techniques, the results to physicians [17,30]. Another effective way is to produce text summaries that report the results obtained and some conclusions in textual form (e.g. ‘the patient had an unstable BGL control from day x to day y with an increasing trend, absence of relevant cycles and a suspected Somogyi effect at day z’) as presented in Ref. [8].

As a further step, it is also possible to post-process the TA analysis results to derive a set of summaries that can be used to acquire some new knowledge on the intra-patient variability. In particular, we exploit the TA analysis results as a set of features that are able to characterize a ‘single control period’, defined as a period with the ‘same insulin protocol’ and to compare this period with past control periods, to obtain clusters of similar dynamic behaviors. In such a way, we should be able to derive a picture of the physician actions and the patient reaction looking at the past history of the patient.

More in detail, the post-processing step has been subdivided in two distinct phases: (i) feature selection and TA-based summaries calculation; and (ii) clustering of different control periods.

3.3.1. Feature selection and TA-based summaries calculation

In the current stage of our work, the decision as to what TA should be used to characterize the control period has been taken relying on medical knowledge. However, we plan to evaluate the feature set on the basis of the results obtained and the experts suggestions. The TA we used in our first experience were the following:

Hyperglycemia Hypoglycemia Metabolic instability Somogyi
(hyper) (hypo) (inst) (som)

The TA-based summaries for the hyper, hypo and inst TAs were derived by simply counting the total time span of the different episodes and by dividing such extent by the total time span of the considered control period. More formally, we defined the normalized TA (n.TA<sub>ta</sub>) of the generic temporal abstraction ta as:

\[
n.TA_{ta} = \frac{\sum_{i=1}^{n} \text{DUR}_i(ta)}{\text{LEN}}
\]

where DUR<sub>i</sub>(ta) is the duration of the i-th episode of ta and LEN is the total time span of the control period. The parameter LEN may assume a different value in dependence on the granularity of the TA; in particular, for the hyper, hypo and inst TAs, LEN is equal to the number of measurements (N.BGL).

For the som TA, it was decided to exploit the absolute value of the episode counts — such choice was motivated by the need for emphasizing each Somogyi effect occurrence, even disregarding its persistence.

Clearly, while this summary is able to provide an aggregated picture of the period, it does not account for high variability in the episodes duration and, therefore, can be considered as an average value for the TA.
To complete the set of features selected, we also included the average insulin requirement (IR)\(^1\) of the patient during the monitoring period.

Therefore, the following is the final feature set:

\[ n.TA_{\text{hyper}} \quad n.TA_{\text{hypo}} \quad n.TA_{\text{inst}} \quad \text{som} \quad \text{N.BGL} \quad \text{N.DAYS} \quad \text{IR} \]

It can easily be argued that the selected feature set is not unique, even using the mild criterion of patho-physiological significance. However, we preferred to select a sort of minimal set (in the opinion of experts) in order to be able to more easily assess the algorithm performance.

In summary, each period has been characterized by a collection of ‘summary features’. Such features describe the insulin therapy (IR), other features the overall BGL values (hyper and hypo), while the remaining features (som and inst) summarize the dynamics behavior provided by the pre-processed data.

### 3.3.2. Clustering

Since the summary features derived in the previous section represent a picture of the patient’s metabolic control over a certain monitoring period, it is of interest to aggregate periods with similar characteristics. Cluster analysis has three important goals: (1) it is possible to check if the different periods are time-dependent by inspecting if the clusters contain consecutive periods; (2) it is possible to show to physicians periods with similar overall metabolic control; and (3) it allows the classification of the last (current) period as belonging to a certain cluster.

From a technical viewpoint, for clustering we applied the \(k\)-means algorithm [26], since all the features used were continuous ones. The number of clusters was automatically set in order to have approximately three periods per cluster.

### 4. Assessment

To assess the TA-based analysis process described in Fig. 1 we performed three different evaluations:

1. Evaluation of the utility of pre-processing to extract trends and cycles on a (real) test case. With this first evaluation, we aimed at comparing the performance of the combined approach, here presented with single-method approaches (pre-processing or TAs only).

2. Evaluation of the post-processing techniques on simulated data. Since the post-processing results cannot be compared with gold standard ones on real patients, we simulated the response of a patient over ten control periods, characterized by the alternation of three different metabolic behaviors. In this way, it was possible to assess the capability of our approach to assign the different periods to the proper clusters.

---

\(^1\) The insulin requirement is the total daily insulin dose per kg.
3. Application of the cluster analysis on the data coming from the home monitoring of four pediatric patients affected by IDDM.

4.1. Pre-processing

In this section, we will show how the application of the pre-processing methodology previously proposed may enhance the capability of TA methods to extract trends and cyclical components from an original time series.

We used, as test set, the data coming from the home monitoring of a 14-year-old male IDDM patient. Such data have been collected during the verification phase of the T-IDDM project in Pavia, at the Department of Pediatrics of the Policlinico S. Matteo Hospital of Pavia.

Fig. 2 shows the data under analysis, corresponding to 33 monitoring days, during which the insulin protocol had not been changed by physicians. The data reflect a high variability and some missing data are present. It is difficult to extract trends and/or cycles from visual inspection and the histogram analysis allows the detection of only some of the hyperglycemia problems at breakfast and dinner.

When the structural TS analysis is applied, the patient’s behavior is more clearly identifiable, as shown in Fig. 3. The algorithm, together with the structure decomposition of the signal, allows the estimating of the value of the missing measurements.
Let us note, for example, at the end of the monitoring time, the presence of relevant increasing and decreasing trends corresponding to a small amplitude of the cycle component. Finally, the analysis of the cycles confirms that the lunch measurements (see Fig. 3 bottom) are usually the lowest in each day.

Fig. 4 shows the result obtained after TA processing. The lack of cyclo-stationarity is highlighted by the 'absence of relevant daily cycle' episode, while the extent of cyclical patterns are clearly depicted.

It is of interest to test if the proposed machinery is really useful, or if it was possible to obtain similar results by simply applying TAs on the original TS. Fig. 5 shows the TA obtained for the cycles extraction. This analysis was unable to derive meaningful patterns, so that each day is seen as it possesses a separate behavior.

Fig. 3. The trend (left figure) and cycle (right figure). The trend is superimposed with the original data, while in the cycle breakfast (stars), lunch (crosses) and dinner (circles) are highlighted.

Fig. 4. Temporal abstractions applied over the trend and cycle components. It is possible to note that when the trend oscillations become higher, the cycle component becomes irrelevant.
The results obtained can be finally transformed in textual reports, as discussed in Ref. [8].

4.2. Evaluation of clustering on simulated data

To evaluate the reliability of the TA-based summary methods, we have simulated the response of a patient over ten control periods.

The periods were characterized by nearly similar therapeutic protocols and the overall BGL response was taken as the sum of the output from a mathematical model of BGL/insulin dynamics [10], a known external disturbance and a 15% additive Gaussian noise. The known external disturbance was introduced to simulate unexpected variations in BGL behavior due to change in patient life-style or to unmodelled physiological effects (such as the adrenaline effect).

In particular, we have defined three different types of response:

1. Stabilized control: the patient underwent therapy with 5 U of fast acting (regular) insulin before breakfast, 10 U of regular insulin before lunch and a mixed insulin injection at dinner time of 4 U of regular insulin plus 8 U of intermediate action (NPH) insulin. Meal intakes were taken as 50, 100 and 100 g (glucose equivalents) at breakfast, lunch and dinner, respectively.

2. Poor meal at dinner time: the patient underwent therapy with 5 U of regular insulin before breakfast, 10 U of regular insulin before lunch and a mixed insulin injection at dinner time of 4 U of regular insulin plus 9 U of NPH insulin. Meal intakes were taken as 50, 100 and 50 g (glucose equivalents) at...
Table 1

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>IR</th>
<th>N.DAYS</th>
<th>N.GBL</th>
<th>n.(T_A_{\text{hypo}})</th>
<th>n.(T_A_{\text{hyper}})</th>
<th>som</th>
<th>n.(T_A_{\text{inout}})</th>
<th>Cl (guess)</th>
<th>Cl (true)</th>
</tr>
</thead>
<tbody>
<tr>
<td>28/12/98</td>
<td>18/01/99</td>
<td>0.675</td>
<td>22</td>
<td>88</td>
<td>0</td>
<td>0.420</td>
<td>2</td>
<td>0.000</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18/01/99</td>
<td>09/02/99</td>
<td>0.700</td>
<td>23</td>
<td>84</td>
<td>0.012</td>
<td>0.679</td>
<td>2</td>
<td>0.679</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>09/02/99</td>
<td>02/03/99</td>
<td>0.700</td>
<td>22</td>
<td>84</td>
<td>0.024</td>
<td>0.619</td>
<td>1</td>
<td>0.610</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>02/03/99</td>
<td>22/03/99</td>
<td>0.700</td>
<td>21</td>
<td>84</td>
<td>0</td>
<td>0.250</td>
<td>1</td>
<td>0.250</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>22/03/99</td>
<td>12/04/99</td>
<td>0.675</td>
<td>21</td>
<td>88</td>
<td>0</td>
<td>0.432</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12/04/99</td>
<td>03/05/99</td>
<td>0.700</td>
<td>22</td>
<td>88</td>
<td>0.023</td>
<td>0.602</td>
<td>2</td>
<td>0.602</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>03/05/99</td>
<td>25/05/99</td>
<td>0.675</td>
<td>23</td>
<td>88</td>
<td>0</td>
<td>0.443</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>25/05/99</td>
<td>14/06/99</td>
<td>0.675</td>
<td>21</td>
<td>84</td>
<td>0</td>
<td>0.405</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14/06/99</td>
<td>06/07/99</td>
<td>0.700</td>
<td>23</td>
<td>88</td>
<td>0</td>
<td>0.239</td>
<td>1</td>
<td>0.239</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>06/07/99</td>
<td>26/07/99</td>
<td>0.700</td>
<td>21</td>
<td>84</td>
<td>0.012</td>
<td>0.655</td>
<td>2</td>
<td>0.571</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
breakfast, lunch and dinner, respectively. This case was characterized by a sawtooth external disturbance that was related to an increase in the oscillations.

3. Rich meal at lunch time: The patient underwent therapy with 5 U of regular insulin before breakfast, 10 U of regular insulin before lunch and a mixed insulin injection at dinner time of 5 U of regular insulin plus 8 U of NPH insulin. Meal intakes were taken as 50, 120 and 100 g (glucose equivalents) at breakfast, lunch and dinner, respectively. Also, this case was affected by a sawtooth external disturbance, but only for 1 week.

The TA-based summaries and the clustering results are reported in Table 1. The derived clusters are denoted with 'cl'.

The clustering algorithm that exploited the TA-based summaries was able, in eight out ten cases, to correctly cluster the control periods (see columns cl-guess and cl-true). Problems occurred in discriminating cluster 3 from cluster 1 in the control periods with no hypoglycemic episodes. Although such a test was performed in unfavorable conditions — due to the relative lack of ‘a posteriori’ distinction in the multi-dimensional feature space — it revealed the capability of the TA-summaries to highlight instability periods and hence, to drive the clustering towards the recognition of the underlying behaviors.

4.3. Application on a real data set

We applied the methodology for data analysis and TA-based clustering, described in the previous section, on four patients monitored in the periods reported in Table 2. For the sake of completeness, let us note that patient 1 is the same patient exploited in Section 4.1.

For all patients, data were first pre-processed using structural filtering and then the TA-based summaries were derived, as explained in Section 3.3.

Although an extensive comment is out of the scope of this paper, we report some conclusions we drew together with the diabetologist responsible for such patients. Such physician was not involved in the system definition.

- Patient 1. The TA summaries and the corresponding clusters are shown in Table 3. Control periods of < 1 week have been discarded. It is apparent that cluster 3 groups periods with the presence of Somogyi effects, relatively high oscillations in BGL and a low percentage of hypoglycemias. Cluster 1, on the contrary, is characterized by a relatively low percentage of hyperglycemia and a relatively
Table 3

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>IR</th>
<th>N.DAYS</th>
<th>N.BGL</th>
<th>n.TA_{hypo}</th>
<th>n.TA_{hyper}</th>
<th>som</th>
<th>n.TA_{inst}</th>
<th>Cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>13/02/98</td>
<td>16/03/98</td>
<td>1.170</td>
<td>32</td>
<td>45</td>
<td>0.044</td>
<td>0.400</td>
<td>0</td>
<td>0.422</td>
<td>2</td>
</tr>
<tr>
<td>16/03/98</td>
<td>24/04/98</td>
<td>1.226</td>
<td>39</td>
<td>101</td>
<td>0.030</td>
<td>0.515</td>
<td>0</td>
<td>0.743</td>
<td>2</td>
</tr>
<tr>
<td>24/04/98</td>
<td>29/06/98</td>
<td>1.236</td>
<td>67</td>
<td>182</td>
<td>0.055</td>
<td>0.462</td>
<td>1</td>
<td>0.643</td>
<td>3</td>
</tr>
<tr>
<td>29/06/98</td>
<td>18/08/98</td>
<td>1.200</td>
<td>51</td>
<td>152</td>
<td>0.164</td>
<td>0.349</td>
<td>0</td>
<td>0.579</td>
<td>1</td>
</tr>
<tr>
<td>18/08/98</td>
<td>18/09/98</td>
<td>1.186</td>
<td>32</td>
<td>92</td>
<td>0.185</td>
<td>0.293</td>
<td>1</td>
<td>0.500</td>
<td>1</td>
</tr>
<tr>
<td>18/09/98</td>
<td>10/11/98</td>
<td>1.179</td>
<td>54</td>
<td>159</td>
<td>0.170</td>
<td>0.201</td>
<td>0</td>
<td>0.667</td>
<td>1</td>
</tr>
<tr>
<td>18/11/98</td>
<td>03/12/98</td>
<td>1.105</td>
<td>16</td>
<td>48</td>
<td>0.062</td>
<td>0.500</td>
<td>0</td>
<td>0.604</td>
<td>2</td>
</tr>
<tr>
<td>03/12/98</td>
<td>20/01/99</td>
<td>1.158</td>
<td>49</td>
<td>148</td>
<td>0.054</td>
<td>0.412</td>
<td>0</td>
<td>0.527</td>
<td>2</td>
</tr>
<tr>
<td>20/01/99</td>
<td>02/02/99</td>
<td>1.263</td>
<td>14</td>
<td>43</td>
<td>0.093</td>
<td>0.279</td>
<td>0</td>
<td>0.744</td>
<td>1</td>
</tr>
<tr>
<td>02/02/99</td>
<td>06/05/99</td>
<td>1.105</td>
<td>93</td>
<td>290</td>
<td>0.076</td>
<td>0.452</td>
<td>3</td>
<td>0.728</td>
<td>3</td>
</tr>
<tr>
<td>06/05/99</td>
<td>01/06/99</td>
<td>1.228</td>
<td>27</td>
<td>83</td>
<td>0.133</td>
<td>0.301</td>
<td>2</td>
<td>0.578</td>
<td>3</td>
</tr>
<tr>
<td>01/06/99</td>
<td>03/08/99</td>
<td>1.070</td>
<td>64</td>
<td>193</td>
<td>0.166</td>
<td>0.244</td>
<td>0</td>
<td>0.446</td>
<td>1</td>
</tr>
<tr>
<td>03/08/99</td>
<td>19/11/99</td>
<td>0.948</td>
<td>109</td>
<td>235</td>
<td>0.064</td>
<td>0.464</td>
<td>3</td>
<td>0.515</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>IR</th>
<th>N.DAYS</th>
<th>N.BGL</th>
<th>n.TA_{hypo}</th>
<th>n.TA_{hyper}</th>
<th>som</th>
<th>n.TA_{inst}</th>
<th>Cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>23/11/98</td>
<td>19/12/98</td>
<td>1</td>
<td>27</td>
<td>76</td>
<td>0.132</td>
<td>0.474</td>
<td>1</td>
<td>0.789</td>
<td>1</td>
</tr>
<tr>
<td>19/12/98</td>
<td>22/03/99</td>
<td>1</td>
<td>94</td>
<td>218</td>
<td>0.161</td>
<td>0.422</td>
<td>8</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>22/03/99</td>
<td>31/05/99</td>
<td>1.040</td>
<td>70</td>
<td>149</td>
<td>0.181</td>
<td>0.376</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>31/05/99</td>
<td>11/06/99</td>
<td>1</td>
<td>12</td>
<td>25</td>
<td>0.240</td>
<td>0.440</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>11/06/99</td>
<td>25/06/99</td>
<td>0.880</td>
<td>15</td>
<td>36</td>
<td>0.083</td>
<td>0.417</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>25/06/99</td>
<td>27/08/99</td>
<td>0.800</td>
<td>64</td>
<td>143</td>
<td>0.056</td>
<td>0.545</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>27/08/99</td>
<td>12/11/99</td>
<td>0.962</td>
<td>78</td>
<td>152</td>
<td>0.132</td>
<td>0.421</td>
<td>12</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

... continues...

... continues...

high percentage of hyperglycemia. It is interesting to note that cluster 1 is more likely to occur during summertime. This was explained by the increased physical activity of the 15-year-old male during the summer. Since the last period (3/08–19/11) seems to present the same problems as in the period 2/2–1/6 (grouped in cluster 3), characterized by a high frequency of Somogyi effects, it is possible to draw the conclusion that the patient is likely to take too much insulin at bed-time.

• Patient 2. The TA summaries and the corresponding clusters are shown in Table 4. In this case, cluster 1 detects the only period with instability, while the remaining two clusters are not well separated, as in the previous case. It is, however, possible to conclude that cluster 2 groups periods with a relatively high number of Somogyi effects and the contemporaneous presence of a relevant number of hyperglycemias. It is also important to note that the period from June 11th to June 25th, is a relatively short period, and hence it is difficult to meaningfully aggregate them.
• Patient 3. The TA summaries and the corresponding clusters are shown in Table 5. Such patient is characterized by relatively short control periods and high instability; cluster 1 groups periods with a high number of Somogyi effects, while clusters 2 contain periods with a high number of hypoglycemias. Cluster 3 presents periods with a relatively high number of hyperglycemias and low number of hypoglycemias. Rather interestingly, the method allows the highlighting of the persistence in the same cluster during the late summer period (27/07–30/09).

• Patient 4. The TA summaries and the corresponding clusters are shown in Table 6. Such results present, as for patient 2, a first cluster with the only period with high instability. Then, cluster 3 collects all periods with Somogyi effects, revealing a progressively worrying condition on the patient’s behavior.

5. Comparison with related approaches

As mentioned in Section 1, a high number of approaches have been presented for the analysis of data coming from diabetic patients’ home monitoring. A number of such approaches have been devoted to the prediction of BGL time series [2, 23, 24], while a few were oriented to an overall interpretation of the patient’s behavior.

Table 5

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>IR</th>
<th>N.DAYS</th>
<th>N.BGL</th>
<th>n.TA_{hyper}</th>
<th>n.TA_{hypo}</th>
<th>som</th>
<th>n.TA_{inst}</th>
<th>Cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>13/01/99</td>
<td>10/04/99</td>
<td>1.231</td>
<td>87</td>
<td>170</td>
<td>0.076</td>
<td>0.318</td>
<td>5</td>
<td>0.576</td>
<td>1</td>
</tr>
<tr>
<td>10/04/99</td>
<td>30/04/99</td>
<td>1.192</td>
<td>21</td>
<td>40</td>
<td>0.175</td>
<td>0.275</td>
<td>2</td>
<td>0.625</td>
<td>2</td>
</tr>
<tr>
<td>30/04/99</td>
<td>31/05/99</td>
<td>1.135</td>
<td>32</td>
<td>93</td>
<td>0.097</td>
<td>0.344</td>
<td>3</td>
<td>0.634</td>
<td>2</td>
</tr>
<tr>
<td>31/05/99</td>
<td>21/06/99</td>
<td>1.077</td>
<td>22</td>
<td>66</td>
<td>0.091</td>
<td>0.515</td>
<td>3</td>
<td>0.773</td>
<td>3</td>
</tr>
<tr>
<td>21/06/99</td>
<td>20/07/99</td>
<td>1.192</td>
<td>30</td>
<td>71</td>
<td>0.169</td>
<td>0.366</td>
<td>3</td>
<td>0.662</td>
<td>2</td>
</tr>
<tr>
<td>27/07/99</td>
<td>28/08/99</td>
<td>1.094</td>
<td>33</td>
<td>71</td>
<td>0.042</td>
<td>0.563</td>
<td>3</td>
<td>0.662</td>
<td>3</td>
</tr>
<tr>
<td>28/08/99</td>
<td>08/09/99</td>
<td>1.057</td>
<td>12</td>
<td>30</td>
<td>0.033</td>
<td>0.567</td>
<td>1</td>
<td>0.467</td>
<td>3</td>
</tr>
<tr>
<td>08/09/99</td>
<td>30/09/99</td>
<td>1.057</td>
<td>23</td>
<td>46</td>
<td>0.043</td>
<td>0.543</td>
<td>3</td>
<td>0.891</td>
<td>3</td>
</tr>
<tr>
<td>06/10/99</td>
<td>29/11/99</td>
<td>1.132</td>
<td>55</td>
<td>156</td>
<td>0.115</td>
<td>0.436</td>
<td>11</td>
<td>0.885</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>IR</th>
<th>N.DAYS</th>
<th>N.BGL</th>
<th>n.TA_{hyper}</th>
<th>n.TA_{hypo}</th>
<th>som</th>
<th>n.TA_{inst}</th>
<th>Cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/01/98</td>
<td>30/03/98</td>
<td>0.719</td>
<td>78</td>
<td>191</td>
<td>0.084</td>
<td>0.335</td>
<td>0</td>
<td>0.607</td>
<td>1</td>
</tr>
<tr>
<td>30/03/98</td>
<td>03/03/99</td>
<td>0.703</td>
<td>339</td>
<td>496</td>
<td>0.036</td>
<td>0.456</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>03/03/99</td>
<td>30/04/99</td>
<td>0.719</td>
<td>58</td>
<td>135</td>
<td>0.022</td>
<td>0.570</td>
<td>6</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>30/04/99</td>
<td>20/05/99</td>
<td>0.812</td>
<td>21</td>
<td>47</td>
<td>0.043</td>
<td>0.553</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>20/05/99</td>
<td>25/06/99</td>
<td>0.703</td>
<td>37</td>
<td>80</td>
<td>0.013</td>
<td>0.562</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>25/06/99</td>
<td>14/11/99</td>
<td>0.625</td>
<td>143</td>
<td>321</td>
<td>0.053</td>
<td>0.592</td>
<td>19</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
[12,20,29], including some commercial products, like Camit-Pro™ or Euro-touch™. While the difference in our approach with respect to the former class of such system is quite clear, a brief comment should be given to highlight the differences and similarities of our approach with respect to the latter. As in Ref. [29], the method herein presented can be considered as a general framework that may exploit all the variables collected during the home monitoring; thanks to the TA methodology, the knowledge about multi-variate patterns can be completely expressed and used in the data analysis. However, unlike all other approaches in the diabetes field, TA results are post-processed to search for similarities in the dynamic behaviors collected during a single case history. This idea has been inspired by some recent data mining papers, in which clustering has been applied to capture similar dynamics of Markov chains that describe a military scenario [27] and in which a supervised learning method has been applied to derive comprehensible descriptions of multivariate time series with applications to speech recognition [19]. Neither of these two approaches uses TAs as the central scheme of the analysis.

Finally, with respect to other applications of IDA to biomedical time series analysis (see Refs. [16,18,25]) we presented a sound pre-processing technique based on statistical method. Such a technique allows the estimation of trends and periodic components of a time series, together with the confidence intervals of the estimate. In addition, thanks to the capability of also estimating the process variances (see Section 3.1), the method automatically derives the 'smoothness' of the trend TS, and the change points between increasing and decreasing patterns. The cons of such a method are, however, represented by the computation time that hampers its use in real-time applications (such as ICU monitoring).

6. Conclusions — strengths and limitations

In Section 1 of this paper, we have presented a general multi-step methodology for exploiting TAs in the context of chronic patients’ monitoring. Such an approach has been applied to data coming from the home monitoring of diabetic patients.

The overall approach presents several novelty aspects:

- TAs are seen as a core tool for performing IDA on TS data. They are used not only for extracting information from the data, but also to generate high level summaries of the patients dynamic behavior. Such summaries are then used to find 'clusters' of behaviors over the monitoring time. The clusters reveal the presence of recurrence or persistence in the metabolic behavior of the patients.
- The methodology herein proposed relies on the exploitation of different methods in the different phases of the analysis: (i) a pre-processing phase applies signal processing techniques on the raw data and extracts structural components of the TS; (ii) thanks to the pre-processing phase a robust application of the TA methods is performed: visualization or text report generation techniques are then used to present the results obtained; (iii) simple statistics allows one to summa-
rize TAs and to extract the TA-based summaries; and (iv) unsupervised machine learning algorithms are applied to derive the clusters.

- Domain knowledge is applied also in the pre-processing phase, through the adoption of ‘structural’ assumptions on the nature of the signals at hand (i.e. BGL values).
- Rather than trying to learn the characteristics of a population of patients, the application of the clustering technique aims at understanding the overall behavior of a single patient. Finally, to the knowledge of the authors, this is one of the first attempts to merge TAs with machine learning analysis within the general IDA process.
- The search for regularities in the data of a very long monitoring period can also be performed, relying only on data processing techniques [12]. Although this approach does not have all the ‘by-products’ of the TA-based analysis, a comparison could be performed.
- The features selected in this paper for the particular application domain could be modified or improved. In particular, they seem insufficient to take a detailed picture of the patient situation. For example, the variability in the episode duration is not taken into account and the n.TA inst feature for describing trends variation does not allow highlighting of increasing or decreasing only situations. Moreover, the features corresponding to hypoglycemias and hyperglycemias can be expanded to report the results obtained in each time of the day (number of hyperglycemias and hypoglycemias at breakfast, lunch, dinner etc.). A trade-off between understanding of the results and completeness is clearly evident in the application of clustering techniques.
- It is apparent that TA-based summaries can be considered as features for a case-based retrieval tool. Such retrieval can be more efficiently performed if the selected features are divided on the basis of different ‘dimensions’. In the case of diabetes, such dimensions can be considered as the single aspects that characterize the metabolic control. For example, it might be interesting to select and retrieve past periods that are similar to a current case for the therapy only, or to select only similar BGL behavior disregarding the other aspects. At the present stage of our work, we are performing some tests on three different dimensions: the therapy, the outcome and the metabolic control.

Acknowledgements

We gratefully acknowledge Dr Giuseppe d’Annunzio and Dr Stefano Fiocchi for providing their support and medical knowledge. We sincerely thank Alberto Riva, without whom this work could not have been done. Finally, we thank the anonymous reviewers for their help in improving the paper. This work is part of the project T-IDDM (HC 1047), funded by the European Commission.
References


