Mining biomedical time series by combining structural analysis and temporal abstractions

R. Bellazzi, P. Magni, C. Larizza, G. De Nicolao, A. Riva and M. Stefanelli
Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy

This paper describes the combination of Structural Time Series analysis and Temporal Abstractions for the interpretation of data coming from home monitoring of diabetic patients. Blood Glucose data are analyzed by a novel Bayesian technique for time series analysis. The results obtained are post-processed using Temporal Abstractions in order to extract knowledge that can be exploited "at the point of use" from physicians. The proposed data analysis procedure can be viewed as a Knowledge Discovery in Data Base process that is applied to time-varying data. The work here described is part of a Web-based telemedicine system for the management of Insulin Dependent Diabetes Mellitus patients, called T-IDDM.

INTRODUCTION

The extraction of knowledge from a biomedical data base requires the completion of the so-called Knowledge Discovery in Data Base (KDD) process[1]: data are pre-processed and then analyzed by Data Mining (DM) algorithms, whose results must be subsequently interpreted and visualized. This process can be iterated until *useful* information is taken out. This paper deals with the problem of understanding the evolution of patients undergoing a therapy over a (relatively long) time. With respect to classical KDD problems, the number of data involved in this analysis is low, but the time dimension adds an additional source of complexity, so that it is crucial to combine different techniques for producing useful knowledge, that can be exploited by final users. The capability of automatically interpreting the results of DM algorithms is essential for moving from a research oriented to a user oriented approach to data analysis.

In our work we were interested in providing physicians with the most recent time series (TS) analysis techniques, that can be exploited to evaluate the metabolic control achieved by diabetic patients during home monitoring[2]. The real understanding of a TS analysis is sometimes difficult even for statisticians; as a matter of fact, the typical result is often a new time series (or a collection of time series) that represents a smoothed version of the original data set. For this reasons we have decided to interpret the TS analysis result by resorting to Temporal Abstraction (TA) techniques[3], that can provide a concise description of the time course of a certain variable. In this way, TS analysis is used to generate a collection of smoothed time series that are then summarized in an abstracted and comprehensible view for the user. The KDD process has hence been implemented through a four-step procedure: i) the data are pre-processed in order to test the applicability of the proposed algorithms; ii) a structural Bayesian analysis is performed on the data; iii) the results of the second step are further elaborated through TA techniques; iv) the output of the TA analysis is shown to physician for metabolic control evaluation. In this paper we will describe each of these steps.

This work is part of a EU funded telemedicine project, called T-IDDM (Telematic Management of Insulin Dependent Diabetes Mellitus), devoted to provide patients and physicians with an Information Technology infrastructure for a better management of type I diabetes (IDDM). In this project, physician relies on a set of distributed Web services, provided by a Medical Workstation. The solutions described in this paper are part of the data analysis and visualization sub-systems, that are linked with the data-management and decision support tools of the architecture. For further details see Riva et al[4].

INTERPRETING DIABETIC PATIENTS TIME SERIES

The complexity of analyzing data coming from home monitoring of IDDM patients is well known and widely described in the literature[2,5,6,7,8]. Physicians must evaluate the status of the patient's glucose metabolic control every 2/4 months by analyzing the data coming from home monitoring, that usually comprises Blood Glucose Levels (BGL), insulin dosages, meal intakes, physical exercise and occurrence of events that may affect glucose metabolism (e.g. fever). These data are then combined with mid-term control variables, like glycosylated hemoglobin, to revise the insulin therapy. In real clinical practice, often the only

1091—8280/98/$5.00 © 1998 AMIA, Inc.

available data are the BGL measurements, that may be automatically down-loaded from blood glucose reflectometers. This practical limitation has lead to the definition of decision support tools that are mainly based on the BGL TS analysis[9].

In particular, a way to judge the outcome of a certain therapy scheme is to check if the BGL measurements follow a cyclo-stationary behavior, i.e. if the daily course of glycemia is approximately the same over the monitoring time. The result of this analysis is the patient's *modal day*, a characteristic daily BGL pattern that summarizes the typical patient's response to the therapy in a specific monitoring period. It may be easily derived by the frequency histograms of BGL measurements in the different times of the day[2,6,7,10]. Of course, in addition to the modal day, it is helpful to know if the BGL TS followed a certain trend during the monitoring.

An interesting way of coping with the trend/cycle pattern detection has been presented by T. Deutsch et al.[2] and has been implemented in the UTOPIA system[9]: the BGL TS is analyzed by applying a technique widely used in econometrics: the Structural Time Series analysis. In our work we moved from the UTOPIA results to two new directions: i) the Structural TS analysis has been implemented in a Bayesian context for obtaining time varying results over a certain monitoring period (see next section); ii) the results of the DM algorithm of step i) have been post-processed by a TA technique for further interpretation (see section 4).

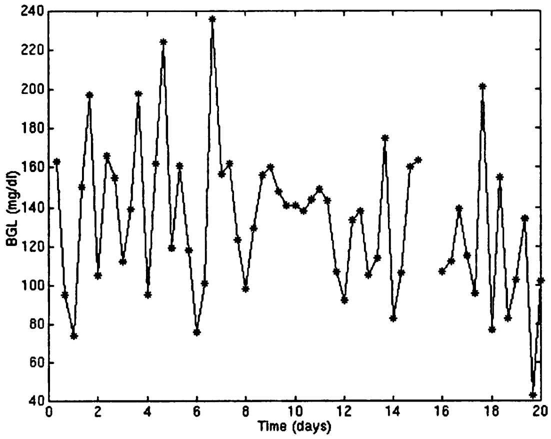


Figure 1: Measurements of BGL collected over 20 home monitoring days on a 12 year old pediatric patient.

A TRAINING EXAMPLE

In order to explain each step of our proposed method, we refer to an example taken from the data-set collected in the Policlinico S. Matteo Hospital of Pavia. Figure 1 shows the BGL measurements collected over twenty days by a 12 year old patient. She measured BGL three times a day. Our goal is to extract *trends* and the *daily cycles* from that data. This goal seems quite complex at a first glance.

STRUCTURAL TIME SERIES ANALYSIS: A BAYESIAN APPROACH

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 (ε), so that, for each measurement BGL_i (see Deutsch et al.[2]):

$$\mathrm{BGL}_i = \mathrm{T}_i + \mathrm{C}_i + \varepsilon_i \qquad (1)$$

The goal of the TS Analysis is then, starting from BGL_i, to extract T_i and C_i. This task can be performed by resorting to a variety of approaches, comprising Kalman *filtering*. Apart from the different technical choices, a fundamental issue must be decided: if, given a certain monitoring period, it is necessary to extract the best trend and the best cyclic components, or if it is important to detect local trend and local cycles. In our example, the first choice will lead to select the best linear regression ($\mathrm{BGL} = \mathrm{BGL}_0 + c \times$ time), and the most probable BGL daily pattern (e.g. high BGL at breakfast and low BGL at dinner); the second choice will instead allow the user to detect different trends within the monitoring period as well as different daily behaviors (e.g high BGL at breakfast and low BGL at dinner until day 10 and then high BGL at breakfast and dinner).

In our work we chose the second approach, that provides the physician with (at the end of the KDD process) a deep interpretation of the original TS. In particular we have exploited a Bayesian approach for signal reconstruction presented by Bellazzi and Magni[11].

The Trend dynamics is described by introducing an additional variable (S_i) that represents the variation of the Trend component of one measurement to the next one, so that $T_{i+1}-T_i=S_i$. If we assume that the S_i time course is described by a Markov Process, the time evolution of the trend component can be specified by the probability distribution $P(S_i \mid S_{i-1})$.

The cycle dynamics requires a more complex model[12]. At each measurement time, the cycle C_i is seen as a linear composition of a sine and a cosine wave, with frequency determined by the daily measurement time (e.g. if there are three measurements per day, the frequency (f) is 1/3), so that

$$\begin{aligned} C_{i+1} &= C_i \cos(2\pi f) + R_i \sin(2\pi f) \\ R_{i+1} &= -C_i \sin(2\pi f) + R_i \cos(2\pi f) \end{aligned} \qquad (2)$$

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

A Dynamic Bayesian Network[13,14] can' easily represent this model, as shown in Figure 2.

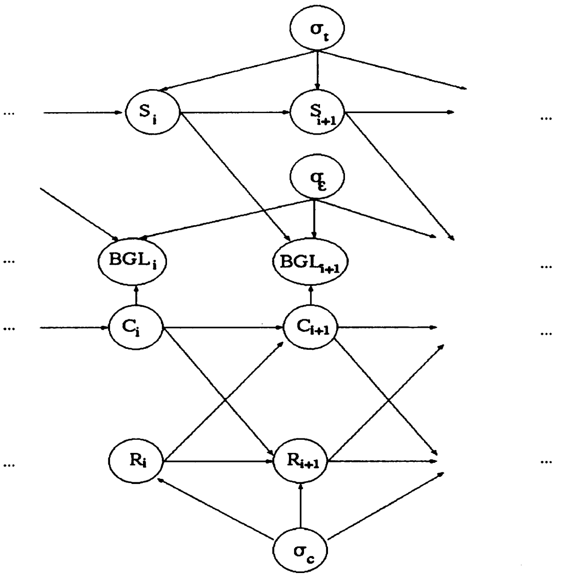


Figure 2. The Bayesian Network representation of the Structural Time Series Analysis

By assuming that:

$$\begin{aligned} &P(S_{i+1} \mid S_i) = N(S_i, \sigma_t^2) \\ &P(R_{i+1} \mid R_i, C_i) = N(-C_i \sin(2\pi f) + R_i \cos(2\pi f), \sigma_c^2) \\ &P(BGL_i) = N(T_i + C_i, \sigma_\varepsilon^2) \end{aligned}$$

where N(.,.) denotes the Normal distribution, it is possible to estimate T_i and C_i given BGL_i by resorting to a Markov Chain Monte Carlo method, that is able to work also in presence of unknown variances $(\sigma_t^2, \sigma_c^2, \sigma_\varepsilon^2)$[11].

The final outcome of the BN machinery presented above is hence the extraction of two new TS (T and C), from the TS of BGL. Such TS express, at each measurement time, the local trend and cycle components.

Figure 3 shows the results obtained on the training example. T is smoother than the original TS, and presents an increasing trend during the first week, and then a decreasing trend. C shows two different cycles, the first one until day 10 and the second one in the remaining period.

It is possible to note that at lunch (circles) BGL measurements are usually higher than in the rest of the day. The interpretation of the results is not easy for the user, in particular for what concerns the analysis of the cycle component. So, to transform the above presented DM method into knowledge, we need a further step.

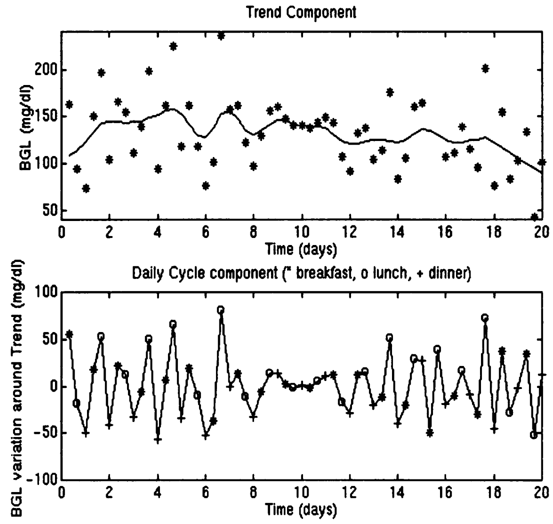


Figure 3: *The results of the structural Time Series analysis. The upper panel shows the trend component and the original data, while the lower panel shows the cycle component; the different daily measurements are highlighted (breakfast measurements are indicated with stars, lunch with circles and dinner with crosses)*

TEMPORAL ABSTRACTIONS FOR END-USER INTERPRETATION

TAs are methods used to obtain an abstract description of the course of multi-dimensional TS by extracting their most relevant features. Hence, in patient monitoring, TAs provide a useful instrument to transform the fragmentary representation of the patient's history into a more compact one. 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*). A formalization of the method is described in the work of Shahar[3], and its application in the diabetes domain can be found in several papers[3,6,7].

In our application we apply TA mechanisms that extract *trends* (increase, decrease or stationary patterns), and *states* (e.g. low, normal, high values) from an uni-dimensional time series. These mechanisms are exploited as means for interpreting the results obtained by the method described in the previous section. In particular, we have applied the following analyses:

a) The T component (see previous section) expresses the BGL local trend. Since daily cycles are cleaned out, the T TS is smoother than the original one and it can be easily analyzed by applying the trend TA mechanism. The final results are the intervals corresponding to the periods of significant BGL increase or decrease.

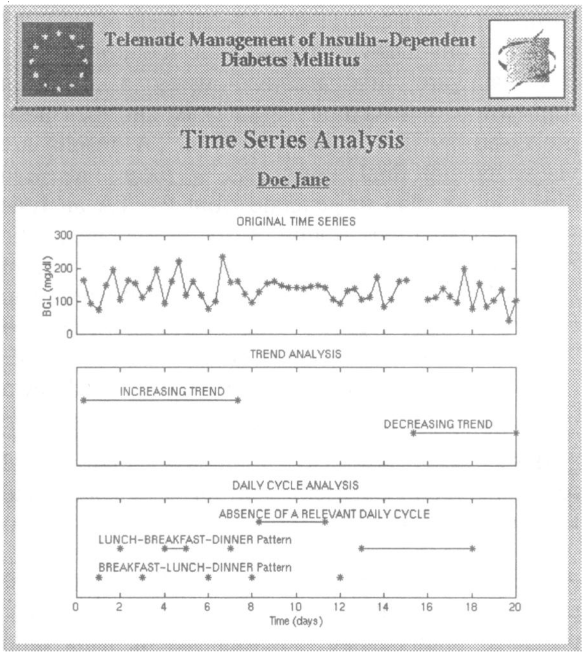
b) The C component interpretation needs the following post-processing procedure:

i) The monitoring period is analyzed to select the intervals where the C is a significant component of the original TS. This can be done by a TA that checks the amplitude of oscillations of C, e.g. if the difference between two consecutive C values is lower than a threshold, the corresponding period is discharged. The remaining analysis is performed on the intervals selected at this step. ii) A BGL pattern is extracted for each day. It is represented as a list of measuring times ordered by the corresponding BGL (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>*); iii) The days with the same pattern in the C TS are searched and aggregated with a state TA mechanism.

The output of this phase is hence a collection of episodes that express: a) the local trends during the monitoring period; b) the presence or absence of the possible cyclic patterns.

VISUALIZATION

As mentioned in the introduction, the KDD process is ended when the results arrive "at the point of use". In the system we are developing, the final user (physician) is provided with a Web-based interface, that allows him/her to have access to all the services needed for IDDM patient management. Also the TS analysis is embedded into this framework, and the TA results are visualized as shown in Figure 4.



Telematic Management of Insulin-Dependent Diabetes Mellitus

Time Series Analysis

Doe Jane

Figure 4: *The visualization of the TS analysis after applying TAs.*

In the upper panel the original TS is shown, while the trend analysis is depicted in the middle panel and the cycle analysis in the lower one. In the trend analysis picture, the increasing and decreasing episodes are labeled and their time spans and locations are graphically shown through bars that connect the start and end time of each episode. In the example, a increasing trend episode starts at day 1 and ends at day 7, while a decreasing trend episode starts at day 15 and ends at day 20. In the cycle analysis, the most significant daily patterns are selected (the patterns that span more than 50% of the total monitoring time), and, the period with an absence of a relevant daily pattern is highlighted. Again, the time spans and locations are graphically shown through bars that connect the initial and end time of each episode. In the example, it is possible to notice that in the days 10 and 11 there was an absence of a relevant daily cycle, while during the last monitoring period the patient had a persistent cycle with minima at dinner and maxima at lunch (*<lunch-breakfast-dinner>* pattern).

By combining the information coming from the trend and cycle analysis, therapeutic suggestions may be straightforward.

DISCUSSION AND FUTURE RESEARCH EFFORTS

In this paper we have described a process of knowledge extraction from a biomedical TS. In particular we concentrated on the analysis of BGL TS coming from IDDM patients home monitoring. The raw data are first analyzed using a Bayesian framework for structural TS analysis, and then post-processed by TA techniques. The TA results are visualized to final users. As a future research direction we plan to generate from the TA analysis results a textual description of the patients metabolic response. The approach herein presented differs from standard statistical techniques for TS smoothing in two ways: a) the degree of smoothness of the derived curves is automatically estimated from the data (through the estimate of the unknown variances $\sigma_t^2, \sigma_c^2, \sigma_\varepsilon^2$); b) the results of the TS analysis are abstracted into a high level representation that is useful both for visualization and for decision support

As mentioned in the introduction, the presented KDD process is integrated into a larger framework, that provides telemedicine services to patients and physicians for IDDM management. Although the analysis system herein presented is powerful, its applicability is limited to periods in which the number of daily BGL measurements is nearly constant, so that missing data can be considered "missing at random"[15]. This DM algorithm is not able to cope with more complex situations, in which the patient's life style changes abruptly; in this case more "weak" techniques should be applied, as described in Bellazzi et al.[7]. To provide physicians with an even more "intelligent" support for data analysis, we plan to pre-process the data in order to detect automatically *what* technique should be used in *each* period.

Acknowledgments

Dr. Stefano Fiocchi is gratefully acknowledged for having provided the data and for having helped us in formalizing medical knowledge. We thank Emanuela Piccolini for her contribution in the system implementation. This work is part of the EC-funded T-IDDM project (HC 1047).

References

1 Fayyad U, Uthurusamy R. Data Mining and Knowledge Discovery in Databases. Communications of the ACM. 1996; 39: 24-26.

2 Deutsch T, Lehmann ED, Carson ER, Roudsari AV, Hopkins KD, Sönksen P. Time series analysis and control of blood glucose levels in diabetic patients. Computer Methods and Programs in Biomedicine. 1994; 41:167-182.

3 Shahar Y. A Framework for Knowledge-Based Temporal Abstraction. Artificial Intelligence. 1997; 90: 79-133.

4 Riva A, Bellazzi R, Stefanelli M. A Web-Based System for the Intelligent Management of Diabetic Patients. MD Computing. 1997; 14:360-364.

5 Lehmann ED. Application of computers in clinical diabetes care. Diab. Nutr. Metab. 1997; 10: 45-59.

6 Kahn MG, Abrams CA, Orland, MJ et al. Intelligent computer-based interpretation and graphical presentation of self-monitored blood glucose and insulin data. Diab. Nutr. Metab. 1991; 4: 99-107.

7 Bellazzi R, Larizza C, Riva A. Temporal abstractions for interpreting chronic patients monitoring data. Intelligent Data Analysis - an International journal (accepted for publication). http://www.elsevier.com/locate/ida.

8 Andreassen S, Benn J, Hovorka R, Olesen KG, Carson ER. A probabilistic approach to glucose prediction and insulin dose adjustment: description of metabolic model and pilot evaluation study. Computer Methods and Programs in Biomedicine. 1994; 41:153-165.

9 Deutsch T, Roudsari AV, Leicester HJ, Theodorou T, Carson ER, Sönksen PH. UTOPIA: a consultation system for visit-by-visit diabetes management. Medical Informatics. 1996;21: 345-358.

10 Shahar Y, Musen MA. Knowledge-Based Temporal Abstraction in Clinical Domains. Artificial Intelligence in Medicine. 1996; 8: 267-298.

11 Bellazzi R, Magni P, De Nicolao G. Dynamic Probabilistic Networks for Modelling and Identifying Dynamic Systems: a MCMC Approach. Intelligent Data Analysis: an International Journal. 1997; 1 (4) http://www.elsevier.com/locate/ida.

12 Harvey A. Structural Time Series Model and the Kalman Filter. Cambridge: Cambridge University Press. 1990.

13 Dagum P, Galper A. Time Series prediction using belief network models. Int. J. Human-Computer Studies. 1995; 42: 617-632.

14 Aliferis CF, Cooper GF, Pollack ME, Buchanan BG, Wagner MM. Representing and developing temporally abstracted knowledge as a means towards facilitating time modeling in medical decision support systems. Computers in Biology and Medicine. 1997; 27: 411-434.

15 Ramoni M, Sebastiani P. The use of exogenous knowledge to Learn Bayesian Networks from Incomplete Databases. Advances in Intelligent Data Analysis. Lecture Notes in Computers Science 1280. Berlin: Springer. 1997; 537-549.