Intra-patients learning by combining clustering and temporal abstractions

R. Bellazzi, C. Larizza, S. Montani, M. Stefanelli

Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy

Abstract

In this paper we present a method for the analysis and subsequent clustering of data coming from home monitoring of diabetic patients. Our method aims at characterising the patient behaviour over time in order to be able to cluster periods with similar dynamics, and to provide a mean to physicians for better learning the action/reaction pattern that a certain patient shows in response to Insulin therapy. The method is described and a case study is reported.

1. Introduction

The analysis of multi-variate time series is an ubiquitous problem in science, and represents a crucial challenge in biomedicine applications, such as clinical monitoring, where several parameters must be contemporaneously examined to understand the patient's overall situation.

This rather complex task has been traditionally faced with descriptive and inferential statistical techniques[Deutsch,1994]. More recently, an AI-based methodology, known as Temporal Abstractions (TAs) has been proposed and successfully exploited in several application domains [Shahar,1996, Bellazzi,1998, Horn,1997, Larizza,1992]. TAs are able to summarise the time course of multi-variate data through abstracted episodes which are valid over a certain time period. A detailed presentation of this methodology can be found in [Shahar,1996].

At a first glance, TAs can be viewed as the AI counterpart of descriptive statistics, that are able to summarise the data through some "sufficient statistics", i.e. mean and standard deviation in normal distributed observations. The number, duration and type of TA episodes can be considered as a summary of the time series at an abstract level. From both a research and an application viewpoint, it is interesting to investigate whether these summaries could be also used to automatically learn some characteristics of the dynamic behaviours under study.

The aim of this paper is to present an approach and to show some results that we have obtained in using TAs for clustering "similar" periods of home monitoring of diabetic patients. With respect to standard Machine Learning approaches, that usually exploit data coming from several subjects to classify the current one, in this paper we will use data coming from the same subject to highlight similar periods in the history of past collected data.

2. Problem Statement

As Diabetes Mellitus is one of the chronic diseases with highest incidence in western countries, there is a large body of research on the use of Information Technology for improving the quality of patient's care [Lehmann,1995]. One important aspect is related to the optimisation of Insulin therapy in Insulin-treated patients through a careful analysis of the parameters that are collected during patients' self-monitoring. Such parameters comprise the Blood Glucose concentration Level (BGL), the presence of Glucose and Ketons in the urine (Glycosuria and Ketonuria) and the occurrence of events or activities that may affect glucose metabolism, such as physical exercise or concurrent diseases. In this paper we will derive some features that summarise the BGL dynamics and the patients' behaviour over a time period characterised by a given insulin protocol scheme. By using these features, we will try to understand if the history of Blood Glucose control of a single patient presents some regularities or recurrence, so allowing, if such regularities exist, to understand if they are caused by the physicians decisions or by other factors, such as the different seasons or periods of the year.

3. Temporal Features

A high number of features have been proposed in the literature to summarise the metabolic behaviour of a patient under Insulin treatment [Bellazzi,1998, Deutsch,1994, Lehmann,1995]. Within the possible choices, we decided to apply a two steps approach for extracting our feature set. Such approach is briefly summarised below:

1. **Data processing and filtering**. The first step is the pre-processing of raw data. This pre-processing was performed by applying both TAs and signal processing techniques.
    1.1 *Temporal abstractions on raw data*. To summarise the BGL daily distribution we classify the BGL measurements collected at the same day time in qualitative levels (low, normal, high, very high, ...) and then we join together consecutive periods with the same qualitative level. Such summaries are computed through methods called State Temporal Abstractions. At the same time, we apply complex temporal abstraction mechanisms; for example, we look for Somogy effect (fasting high level of Blood Glucose in response to a nocturnal hypoglycemia): this effect is detected through a combination of high BGL at breakfast time with absence of glycosuria (see [Bellazzi,1998] for a detailed explanation).
    1.2 *Structural filtering and Temporal Abstractions*. Since it is well known that the time course of Blood Glucose in normal subjects follows a cyclo-stationary behaviour, an interesting analysis is the extraction of trends and daily cycles from the time series of pathologic subjects. In this way it is possible to understand if persistent cyclical patterns exists, and if there are significant trends buried in the (often highly unstable) original data sets [Deutsch, 1994]. The two components of the BGL time series are detected resorting to signal processing methods [Bellazzi, 1998 (2)]. Once the two series are available, it is possible to apply TA techniques for deriving episodes of increase or decrease in the trend time series, as well as for counting the time periods characterised by the same cyclical (daily) pattern (see [Bellazzi, 1999] for details).
2. **Feature structuring**. After step one, from the raw data we obtain a large number of data analysis results, that can be used as features for learning about the patients behaviour. The first way to reach this goal is to show with proper visualisation techniques the results to physicians [Shahar, 1999]. 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 Somogy effect at day z"). A third (new) way of exploiting data is to use the data analysis results as a set of features that are able to characterise a single *control period*, defined as *a period with the same Insulin protocol*, and to compare this period with past periods, to obtain clusters of similar dynamic behaviours. In such a way we are able to derive a picture of the physician actions and the patient reaction looking at the past history of the patient. More in detail, we have subdivided the derived features into two layers: a summary layer and a detailed layer. The *summary layer* reports features that with *one numeric value* characterise a *single aspect* of the patient dynamic behaviour over a control period. The *detailed layer* reports features that detail with *a collection of numbers* the same *single aspects* of the patient dynamic behaviour over a control period. For example, a summary feature reports the percentage of hyperglycemic episodes occurred in a period, while a detailed feature reports the minimum duration, the maximum duration and the average level of hyperglycemic episodes in a single time period.

4. Clustering and retrieval

Having classified all the extracted features in the two layers, it is possible to define a strategy for characterising the past monitoring periods and for comparing the current period (e.g. at visit time) with the past behaviours. Again, our proposed strategy is composed by two steps:

- the first step relies on summary features and derives a certain number of clusters from the periods at hand. Then, the past periods belonging to the cluster of the current one are retrieved.
- the second step relies on the detailed features and orders the retrieved periods in dependence to their relative distance with the current one.

Clustering. Each period is characterised by a collection of summary features. Some features describe the insulin therapy, other features the BGL, while some others summarise the trend and cycle components; finally a Somogy effect feature is included. To cluster periods a modification of the k-means algorithm is applied [Mitchell, 1997]; in particular, we use the Heterogeneous Euclidean-Overlap metric (HEOM) as the distance measure for clustering [Wilson,1997]. HEOM is able to account for discrete (nominal) and continuous variables. The number of clusters is automatically set in order to have approximately three periods per cluster.

The cluster analysis has three important side-effects: first, it is possible to check if the different periods are time-dependent, by inspecting if the clusters contain consecutive periods; second it is possible to show to physicians periods with similar overall metabolic control; third, it allows to classify the current period as belonging to a certain cluster.

Retrieval and period ordering. The classification allows us to retrieve all periods belonging to a certain cluster. At this second stage, the HEOM is computed taking into account only the detailed features; this allows to order the different periods and to show the closest period at the detailed level.

Data Mining over different dimensions. It is apparent that the clustering and retrieval steps can be performed not only by considering all the features, but also by selecting the features that characterise a single aspect of the metabolic control. For example, it might be interesting to select and retrieve past periods that are similar for the therapy only, or to select only similar BGL behaviour disregarding other aspects. By using the proposed approach, it is possible to mine the data following different dimensions, and to obtain different views that might be related to different evaluation purposes. To facilitate this end, we also sub-divided all the features in dimensions, so that the user may select one dimension to start the overall process. At the present stage of our work, we exploited three dimensions: the therapy, the outcome and the metabolic control.

5. An example

For better clarifying the approach previously presented, let us consider an example. We consider a 14 years old male diabetic patient followed from March 3rd 1998 to May 30th 1999. During this period he collected 1351 BGL measurements (around three times a day), 1035 Glycosuria measurements and he registered 1649 Insulin dosages. Thirteen different therapeutic protocols were suggested by the physician, so that we can compare the analysis performed on the data collected in the last period with the one obtained considering the previous 12. For the all monitoring periods the analysis summarised in Section 4 was performed.

An example of the data collected in one of the monitoring periods is reported in Figure 1, while Figure 2 shows the result of the Temporal Abstractions and structural filtering algorithms on the same raw data.

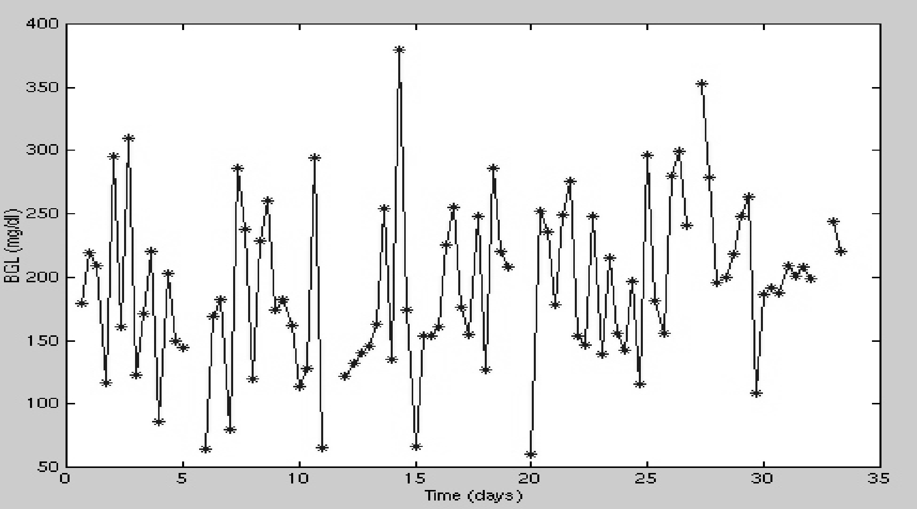


Figure 1. The BGL data collected over one month

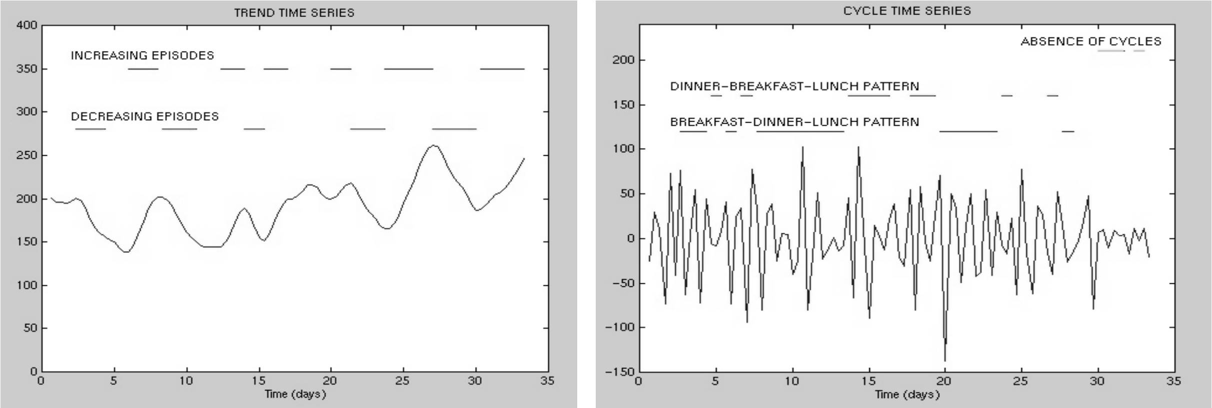
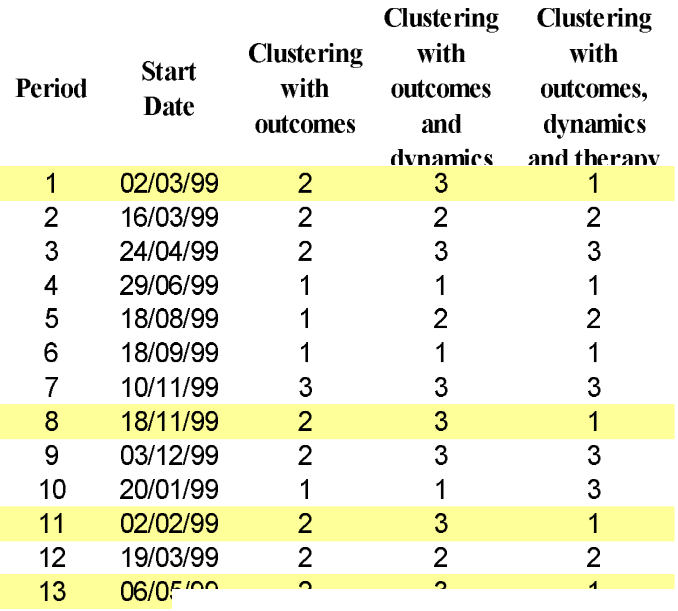


Figure 2. The results of structural analysis. Trends and cycle components are shown; the duration of increasing and decreasing episodes are depicted (left); moreover, the different daily patterns extracted from data are reported (right): a breakfast-dinner-lunch pattern is a daily pattern in which the lowest BGL value in one day is at lunch and the highest is at dinner time.

				BGL metabolism		*OUTCOME*			*Therapy*
Period	**start date**	**period length**	**# BGLs**	**Cycle bandwidth**	**Cutoff frequency of trend spectrum (x10-5)**	**% Hypo**	**% Hyper**	**Somogy**	**Daily Insulin dosage**
1	02/03/99	16	48	0,1591	*.8*	6%	33%	0	62
2	03/03/99	37	111	0,1720	*.4*	2%	44%	0	65
3	04/03/99	64	192	0,1718	*.9*	5%	41%	1	68
4	05/03/99	52	156	0,1669	*.5*	15%	29%	0	68
5	06/03/99	32	96	0,1700	*.25*	16%	27%	0	67
6	07/03/99	54	162	0,1693	*.8*	15%	20%	0	67
7	08/03/99	8	24	0,1591	*.75*	21%	46%	0	66
8	09/03/99	16	48	0,1664	*.8*	8%	44%	0	64
9	10/03/99	49	147	0,1583	*.9*	8%	39%	0	67
10	11/03/99	13	39	0,1603	*.85*	13%	26%	0	73
11	12/03/99	46	138	0,1711	*.9*	8%	36%	0	64
12	13/03/99	49	147	0,1724	*.4*	7%	51%	3	67
13	14/03/99	24	72	0,1730	*1.01*	8%	32%	2	70

Table 1. A portion of the summary features. The features are grouped in the three different dimensions

Once all the analyses have been performed, the features of the summary layer are used to cluster periods with similar behaviours. Table 1 reports a subset of the summary features of the periods.
As it is apparent from the Table, while the features related to therapy (e.g the daily insulin dosage) and to the outcome are quite easy to understand, the features related to the BGL behaviour need a further explanation. The cutoff frequency of trend spectrum is related with the energy associated with the trend component: the highest is this frequency, the more unstable is the trend time series; the cycle bandwidth is associated with the distribution of the cycle time series around a single pattern: the largest is the band, the more frequent the patient changes its daily pattern. The table highlights the features of the current case.
Table 2 shows the results of the cluster analysis. Rather interestingly, the clusters obtained considering all features are not consecutive. Moreover, the current period is clustered with the first one and with all the periods that showed a nearly stable control (see Table 1).
An interesting result of the analysis is that, after a period of increased hypo and hyperglycemias, the physician succeeded in a normalisation of the BGL level, but had to "pay the price" of a high increase in the insulin requirement (see Table 1). Finally it is possible to see that there is a substantial similar clustering across the different dimensions.



Period	Start Date	Clustering with outcomes	Clustering with outcomes and dynamics	Clustering with outcomes, dynamics and therapy
1	02/03/99	2	3	1
2	16/03/99	2	2	2
3	24/04/99	2	3	3
4	29/06/99	1	1	1
5	18/08/99	1	2	2
6	18/09/99	1	1	1
7	10/11/99	3	3	3
8	18/11/99	2	3	1
9	03/12/99	2	3	3
10	20/01/99	1	1	3
11	02/02/99	2	3	1
12	19/03/99	2	2	2
13	06/05/99	2	3	1

Table 2. The result of clustering by using the summary features over the different dimensions.

Once the periods belonging to the closest cluster are selected, it is possible to retrieve the closest ones by considering the detailed features list. In our example, the closest period is number 1, and a detailed features comparison for the trend abstraction features is shown in Table 3.

Period	total length	trend	min length	max length	mean length	% length
1	48	decrease	5	7	6,00	38%
1	48	increase	4	6	5,00	31%
13	72	decrease	3	6	4,22	53%
13	72	increase	3	6	4,00	28%

Table 3. Comparison of the detailed features of the trend abstractions between period 1 and 13

The similarities between period 13 and 1 are highlighted in the case of "increasing trend abstractions": the same maximum episodes duration and a nearly similar percentage length of the episodes is found in both periods.

6. Conclusions

In this paper we have shown a method for exploiting a number of high level features extracted from a patient monitoring data to cluster periods with the same overall dynamic behaviour. In particular, we have applied a two-step methodology that allows to progressively refine the search for similar periods in the patients history by the use of a set of features at different levels of detail. One key issue of our proposed method is the extraction of the features used in the clustering and retrieval steps by resorting to a combination of statistical techniques and Temporal Abstractions. A second important issue is the capability of providing a mean to the user for mining temporal data through different search dimensions, as well as with different levels of detail.

The piece of research herein presented has some recent related works. In particular, a work of Sebastiani and Ramoni [Sebastiani,1999] presented a method for characterising a time series through a Markov Chain, and therefore to derive clusters of similar dynamic behaviours through a similarity index calculated on the probability transition matrices. Such approach cannot be exploited in our context since the dimension of the feature space, that reflects the dimension of the space variables in the Markov model, may be rather large. Another interesting method for dealing with learning the dynamics was recently presented by Kadous, that used a particular kind of Temporal Abstractions to derive a high level feature description of the system dynamics from multi-variate data [Kadous,1999]. Such method also uses clustering for recognising similar dynamic behaviour. The difference with the approach presented in this paper is related to technique used for data processing and for feature extraction.

As a further step, our plan is to investigate the crucial problem of feature selection, in order to reduce the feature space in both summary and detailed layer; we plan to define a feature selection strategy that is dependent on the search dimension in the overall search space.

From the application viewpoint, we plan to evaluate the approach herein presented by assessing the perceived subjective utility of this method from physician, and by including a tool in our system for tele-care of Diabetic Mellitus patients, defined within the T-IDDM project (HC 1047) funded by the European Commission.

Acknowledgements. This work is part of the project T-IDDM (HC 1047), funded by the European Commission.

References

[Bellazzi,1998] Bellazzi, R., Larizza, C., Riva, A., Temporal Abstractions for Interpreting Diabetic patients monitoring data, Intelligent Data Analysis, 2 (1998) 97-122.

[Bellazzi, 1998 (2)] Bellazzi, R., Magni, P., Larizza, C., De Nicolao, G., Riva A., and Stefanelli, M., Mining biomedical time series by combining structural analysis and temporal abstractions, JAMIA, Symposium supplement 1998, 160--164.

[Bellazzi, 1999] Bellazzi, R., Larizza, C., Magni, P., Montani S. and De Nicolao, G., Intelligent Analysis of Clinical Time Series by combining Structural Filtering and Temporal Abstractions, Lecture Notes In Artificial Intelligence 1620, W. Horn, Y. Shahar, G. Lindberg, S. Andreassen, J. Wyatt eds, pp. 261-270, 1999.

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

[Horn,1997] Horn, W., Miksch, S., Egghart, G., Popow, C., Paky, F., Effective data validation of high frequency data: time-point-, time-interval- and trend-based- methods. Computers in Biology and Medicine, 25 (1997) 389-409.

[Kadous,1999] Kadous, M.W., Learning Comprehensible Descriptions of Multivariate Time Series, Proceedings of ICML 99, Ljubljana, 1999.

[Larizza,1992] Larizza, C., Moglia, A., and Stefanelli, M., M-HTP: A system for monitoring heart transplant patients, Artificial Intelligence in Medicine. 4, 111-126, 1992.

[Lehmann,1995] Lehmann, E.D and Deutsch, T., Application of computers in diabetes care - a review. Part I and II., *Medical Informatics*. 20, 281-302, 1995.

[Mitchell,1997] Mitchell, T., Machine Learning, Mc-Graw Hill, 1997.

[Sebastiani,1999] Sebastiani, P., Ramoni, M., Discovering Dynamics using Bayesian Clustering, Proceedings of IDA 99, Amsterdam, 1999.

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

[Shahar,1997] Shahar, Y., A Framework for Knowledge-Based Temporal Abstraction, Artificial Intelligence. 90(1-2), 79-133, 1997.

[Shahar,1999] Shahar, Y., Timing is everything: Temporal Reasoning and Temporal Data Maintenance in Medicine, Lecture Notes In Artificial Intelligence 1620, W. Horn, Y. Shahar, G. Lindberg, S. Andreassen, J. Wyatt eds, pp. 30-46, 1999.

[Wilson,1997] Wilson, D.R., Martinez, T.R., Improved heterogeneous distance functions, Journal of Artificial Intelligence Research, 6 (1997) 1-34.