ELSEVIER

Computer Methods and Programs in Biomedicine 62 (2000) 205–218

Computer Methods and Programs in Biomedicine

www.elsevier.com/locate/cmpb

Diabetic patients management exploiting case-based reasoning techniques

Stefania Montani [a], Riccardo Bellazzi [a,*], Luigi Portinale [b], Giuseppe d'Annunzio [c], Stefano Fiocchi [c], Mario Stefanelli [a]

[a] *Dipartimento di Informatica e Sistemistica, Università di Pavia, via Ferrata 1, I-27100 Pavia, Italy*
[b] *Dipartimento di Informatica, Università di Torino, corso Svizzera 185, I-10149 Torino, Italy*
[c] *I.R.C.C.S. Policlinico S. Matteo, P.le. Golgi 2, I-27100 Pavia, Italy*

Received 8 December 1998; received in revised form 5 May 1999; accepted 12 May 1999

Abstract

In this paper we propose a case-based decision support tool, designed to help physicians in 1st type diabetes therapy revision through the intelligent retrieval of data related to past situations (or 'cases') similar to the current one. A case is defined as a set of variable values (or features) collected during a visit. We defined taxonomy of prototypical patients' conditions, or classes, to which each case should belong. For each input case, the system allows the physician to find similar past cases, both from the same patient and from different ones. We have implemented a two-steps procedure; (1) it finds the classes to which the input case could belong; (2) it lists the most similar cases from these classes, through a nearest neighbor technique, and provides some statistics useful for decision taking. The performance of the system has been tested on a data-base of 147 real cases, collected at the Policlinico S. Matteo Hospital of Pavia. The tool is fully integrated in the web-based architecture of the EU funded Telematic management of Insulin Dependent Diabetes Mellitus (T-IDDM) project.

Keywords: Case-based reasoning; Diabetes management; Decision support; Telemedicine

1. Introduction

After the publication of the DCCT study ([1]), intensive insulin therapy (IIT) has become mandatory for patients suffering from 1st type diabetes mellitus (DM-1). IIT, consisting in three to four insulin injections per day or the use of insulin pumps, is a data and knowledge intensive process, since it requires frequent blood glucose level (BGL) measurements, and the reporting of insulin injection amounts and other information on patients' diet and life-style. Moreover, the quantity and quality of the interactions between the patients and the health care providers is necessarily increased, in order to maintain a tight metabolic control and, at the same time, to improve the patient's self-consciousness and education on the disease. A natural goal of information technologies (IT) is, hence, to provide a valuable support to this disease management process [2].

* Corresponding author. Tel.: +39-0382-505511.
E-mail address: ric@aim.unipv.it (R. Bellazzi)

0169-2607/00/$ - see front matter
PII: S0169-2607(00)00068-7

Nowadays, the data collection is highly facilitated by the capability of commercial reflectometers of data storage and downloading, as well as by the increasing use of telemedicine systems [3]. It can also help diabetologists by providing them with a collection of tools for improving the quality of patient's care [4], from data-bases to simulation and education packages, and finally to decision-support systems. Anyway, managing all the data piling in the physicians' electronic desk, and extracting from them reliable information about the patients' status, often remains a problem difficult to be satisfactorily solved.

For these reasons, it is of interest to study methods that enable physicians in performing an intelligent consultation of the available databases, either for supporting their decisions during therapy revision, or for extracting useful information from the accumulated experience. It is also interesting to keep track of the 'problem/solution' patterns that occurred in the past, in order to manage and disseminate the expertise of the diabetologists. In the context of chronic diseases (and hence in particular in diabetes monitoring), all the aspects related to the management and maintenance of knowledge play a crucial role, at least for two different reasons: first, it is necessary to maintain the knowledge about a specific patient over time, so that, even in presence of changes in the physicians staff, the quality of the patient's care is not decreasing due to the lack of information; second, it is useful to manage the 'operative' knowledge of experts, in order to bring in surface their expertise and to keep it in the institution even when they move or retire.

To cope with the above mentioned problems, we have designed a new tool for data-base retrieval and knowledge management, based on the case-based reasoning (CBR) technology [5–7]. CBR is a problem-solving paradigm that utilizes the specific knowledge of previously experienced situations, called cases. Each case is usually described by a set of variable values, called features, and it is associated to a solution (or decision) and to an outcome. CBR basically consists in retrieving past cases that are similar to the current one and in reusing (by, if necessary, adapting) past successful solutions; the current case can be retained and put into the base of cases. CBR can, hence, be viewed as a methodology able to combine retrieval, reasoning and learning steps and to produce solutions to problems by taking into account past experience.

In current medical practice, CBR techniques can be limited to provide diabetologists with a tool to perform an intelligent retrieval of the data-base of past cases, in order to detect, during a periodical control visit, if the same metabolic behavior has already occurred to the same patient or to a similar one, and, in that case, in seeing what decision was taken in the past and what was the resulting outcome. Supporting physicians in this activity may be particularly interesting, considering that they may visit more than 100 patients every month. A crucial capability of the tool will be the possibility of analyzing the overall history related with the retrieved cases, showing the sequence of decision steps that precede and follow the retrieved situation. In addition, it will be possible to calculate from the selected sub-population some statistics that may be interesting for taking decisions, as well as for assessing the quality achieved in the treatment of the sub-population at hand [4].

The CBR tool, we have realized is fully integrated within the architecture of the EU funded Telematic management of Insulin Dependent Diabetes Mellitus (T-IDDM) project, devoted to the implementation of a web-based telemedicine system. In this project, physicians can exploit a medical workstation, that comprises a number of IT services useful for managing this kind of chronic patients, an information system for data handling, a data analysis and visualization system, a decision support system and a communication tool for data exchange with the patients' house; all of this sub-systems are integrated in a web-based service. Details on the T-IDDM project can be found in [8,9].

In this paper, we will describe the fundamental components of the CBR system, with examples and an evaluation of the retrieval tool performed on a set of 147 cases collected at the Department of Pediatrics of the Policlinico S. Matteo Hospital of Pavia, Italy.

2. The case-based reasoning paradigm

CBR is a reasoning paradigm that instead of relying on general rules or models, utilizes the specific knowledge contained into already solved instances of problems [6,7]. In the CBR model, problem-solving experience is explicitly taken into account by storing past solved problems and by suitably 'remembering' them when a new problem has to be tackled. A case is then a structured representation of past problems suitable for re-use.

Generally, a case consists of the following three basic information,

- the problem description, typically a set of ⟨feature, value⟩ pairs in terms of which the problem corresponding to the case may be characterized;
- the case solution representing the solution adopted for solving the corresponding problem;
- the case outcome representing the outcome of the applied solution.

For instance, in a medical domain the problem description may be the set of symptoms of the patient under examination, and of pathophysiological entities providing a compact description of the clinical time course; the solution may be the possible treatment, and the outcome may be the result of the treatment.

The use of CBR plays a significant role in many relevant tasks like diagnostic problem solving [10] or planning [11], since it can mimic (at some extent) the capability of human experts in solving a new case by retrieving similar cases solved in the past and by suitably adapting them to the situation at hand.

The suitability of CBR to solve complex problems has been widely discussed in the last few years and two basic possibilities emerged:

- Precedent CBR, where previous solutions to cases similar to the current one are used as a justification for the solution of the current case with almost no adaptation (e.g. legal reasoning).
- Case-based problem solving, where retrieved solutions to previous similar cases need to be adapted to fit the current situation (e.g. planning, design, diagnosis, etc.).

Case-based problem solving is, of course, the most general approach and can be summarized by the following four basic step known as the CBR cycle or the four 'Res' [7]:

1. RETRIEVE the most similar case(s) from the case library.
2. REUSE the case knowledge (typically the solution) to solve the new problem.
3. REVISE the proposed new solution.
4. RETAIN the relevant parts of this experience (typically the current case) for future problem solving.

The above CBR cycle puts emphasis on several aspects each CBR system has to deal with, namely how to represent cases, how to organize the case library, which kind of algorithm to use for retrieval, how to adapt a retrieved solution and when to add new cases or forget old ones.

In the whole cycle, some steps may be missing or they may be collapsed. For example, it is quite common to view the REUSE and REVISE steps as a single one or to avoid the RETAIN step if the current case is in some sense 'covered' by other cases already stored in the library.

Moreover, very often CBR systems are built essentially as tools for flexible retrieval, leaving to the user all the decisions concerning adaptation and re-use of retrieved solutions. Indeed, even if no capability for automated adaptation is provided by the CBR system, in many applications concerning decision support it is very useful to extract the knowledge concerning relevant past cases for further analysis. This is also the view taken by most commercial CBR tools [12]. As a consequence, devising an efficient retrieval process is fundamental for dealing with large case libraries, as can be the case in a medical setting.

3. Case-based retrieval for DM-1 management

As previously noticed, in the context of a periodical visit, CBR may have an important role in decision making, the retrieval of the therapy schemes adopted in the past, and of some indicators of the outcomes obtained on those occasions, could provide a first guideline on how to cope with the current problems. Since, we are inter-

ested in providing useful information to physicians, the data-base of past cases (called case memory) was structured by resorting to a taxonomy of prototypical classes, that express typical problems that may occur to patients. The retrieval process implemented in our tool is, hence, composed by two steps, a classification step, that proposes to the physician the class of cases to which the current case could belong, and a retrieval step, that effectively retrieves the 'closer' past cases. In the following, we will describe in detail the case memory structure and each one of the retrieval steps.

3.1. The case memory

3.1.1. Features

As shown in Section 2, a case is generally defined as a collection of features summarizing the problem, together with a solution and with the outcome obtained by applying the solution itself. More formally, a case C can be viewed as a triple

$$C = \{\langle F{:}f\rangle, \langle S{:}s\rangle, \langle O{:}o\rangle\}$$

where f being the vector of values for the set of descriptive features F, s the solution schemata selected from the solution space S and o the outcome of the solution selection in the space of the possible outcomes O.

In our application a case just coincides with a periodical visit, whose data, after having been collected and, when needed, discretized, represent the case features of set F. In more detail, we have defined 27 variables (see Table 1), of which 21 are nominal (discrete) and six are linear (continuous), extracted from three sources of information:

- General characterization, 11 features able to generically describe the patient, such as, sex, age, distance from diabetes onset.
- Mid-term information, 13 features actually collected during the visits, like weight and glycated hemoglobin (HbA1c) values.
- Short term (day-by-day) information, three features collected during home monitoring activity, i.e. the number of hypoglycemic episodes, the metabolic control and the physical activity.

Some features may be automatically obtained as abstractions of the raw data collected during the visit. For example, control trend and requirement trend of Table 1 are calculated as the variation of HbA1c and requirement, respectively, since the previous visit; metabolic control and hypoglycemias discretize and summarize the information coming from the day-by-day BGL data collection. By now, the day-by-day features of the cases stored in the data-base have been extracted from the patients' log-books, as we were working on retrospective data. A routinely use of the T-IDDM service, permitting the telematic transmission of the monitoring data from the patient's house to the medical workstation, will enable an automatic collection and elaboration of the metabolic control and life style information.

The solution s is represented by an array containing insulin types and doses, decided by the physician from an analysis of the data in the set F. The outcome o of the therapeutic decision is summarized by HbA1c and by the number of hypoglycemic events collected at the following visit.

3.1.2. Classes

As previously mentioned, the case memory was structured through the partitioning induced by a set of mutually exclusive classes. Such classes express the medical knowledge on the prototypical situation that may occur to DM-1 pediatric patients.

More precisely, we have defined a taxonomy, where the roots class (patient's problems) represents the most general class describing all the possible cases we may store into the case memory. Each remaining class in the hierarchy is a prototypical description of the set of situations it summarizes; the class/sub-class link represents the relation of further specialization (see Fig. 2). More precisely, leaf nodes represent the most detailed description of pathologies, or clinical course conditions taken into consideration; an inner node represents a class with certain properties that all the classes of its descending sub-tree have in common. For example, the inner node behavioral–puberal problems has four descendants: change life style; falsifier; no motivation; typical puberal problems.

The first three leaves are possible situations deriving from the patient's behavior, from problems related with a non-reported change in lifestyle to other psychological problems related with young patients, like data falsification or loss of motivation in following ITT. The fourth class identifies the typical alterations experienced by diabetic patients in puberty.

The prior knowledge associated to the inner node indicates a normal weight and a metabolic

Table 1
The features defining a case

Feature name	Type	Values
Characterization		
Sex	Nominal	[Male, female]
Height	Linear (continuous)	
Age	Linear (continuous)	
Neuropathies	Nominal	[Yes, no]
Other chronic diseases	Nominal	[Unrelated, related to hyperglycemia, related to hypoglycemia, absent]
Puberal stage	Nominal	[Infant, beginning puberal, puberal, adult]
Job	Nominal	[Not-sedentary-worker, sedentary-worker, not-sedentary-student, sedentary-student]
Retinopathies	Nominal	[Yes, no]
Anti insulin antibodies	Nominal	[Yes, no]
Nephropathies	Nominal	[Yes, no]
Distance from onset	Nominal	[Short, long]
Mid-term features		
Weight	Linear (continuous)	
Weight excess	Nominal	[Overweight, underweight, normal]
HbA1c	Linear (continuous)	
Other hormonal disorders	Nominal	[Yes, no]
Requirement trend	Nominal	[Increase, decrease, stationarity]
Control trend	Nominal	[Increase, decrease, stationarity]
Regular insulin	Nominal	[Regular, actrapid]
NPH insulin	Nominal	[Monotard, protaphane, intermediate]
Premixed insulin	Nominal	[Isophace, actraphane]
Premixed ratio	Nominal	[90/10, 80/20, 73,30, 60/40, 50/50]
Number of injections	Linear (continuos)	
Diet	Nominal	[Free, prescribed, controlled]
Requirement	Linear (continuous)	
Short-term features		
Metabolic control	Nominal	[Good, hypoglycemias, hyperglycemias, instable]
Hypoglycemias	Nominal	[None, some, many]
Physical activity	Nominal	[None, intensive-continuous, medium-continuous light-continuous, intensive-occasional medium-occasional, light-occasional]

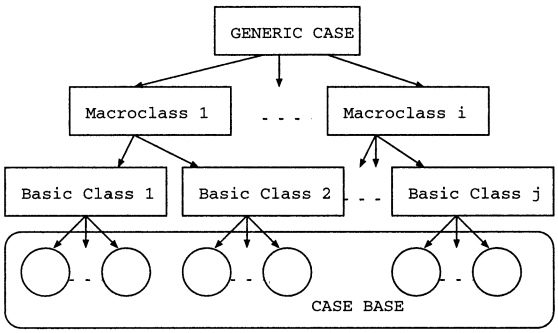

Fig. 1. Case memory organization.

control characterized by hyperglycemia. Each of the leaves is a specialization of the inner node, where additional feature values are specified. For example, for typical puberal problems, the feature puberal stage has to be puberal or beginning puberal, while this condition is not mandatory for the other three classes.

Leaves in the taxonomy are called basic classes, each case in the case memory is an instance of a unique basic class and can be retrieved through such a class. Classes at the upper level are denoted as macroclasses (see Fig. 1).

Of course, a revision of the taxonomy would be required for representing adult patients' problems, note for example, the presence of the puberal stage class and the absence of cardiovascular complications, that are frequent in adults.

3.2. Classification

Situation assessment and case search are strongly influenced by the organizational struc-

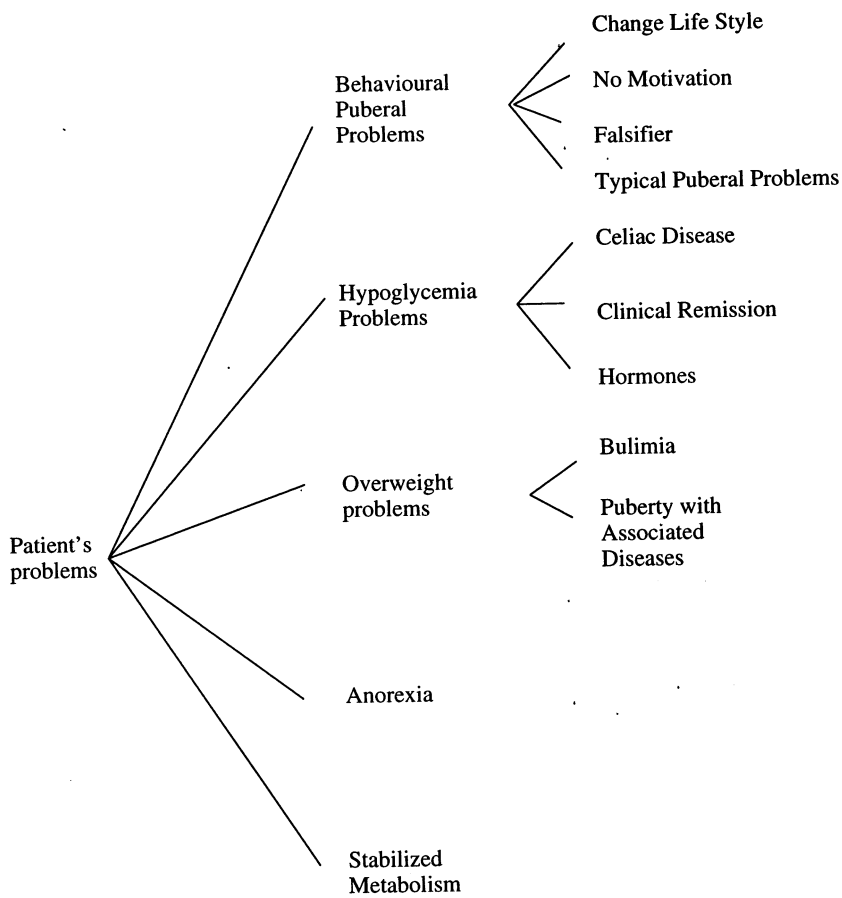

Fig. 2. Taxonomy of classes of prototypical situations that may happen during monitoring of pediatric DM-1 patients.

Table 2
Features chosen for the classification step

Sex	Job	Puberal stage	Other chronic diseases	Distance from onset	Weight excess	Diet	Control trend	Required trend	Metabolic control	Hypos	Physical activity

tures on which the case memory is based on. In our application, the definition of a taxonomy of prototypical situations has allowed us to implement a method to make case retrieval more flexible. The first step of the method is classification, the search space for similar cases is limited by identifying what is the context in which the current case should be interpreted, i.e. by finding what are the classes in the hierarchical structure that better represent the case itself. Classification may be performed on the leaves of the taxonomy tree, to find the most probable classes to which the input case could belong; but when several features in the case are missing, or when a less specific identification of the situation at hand is required, the classification step may be conducted just at the upper level of the tree, working on the more general macroclasses.

Classification is performed on a sub-set of the features, and in particular on the ones that physicians have considered more useful to discriminate the classes. Table 2 shows the features that have been chosen for this task.

Although, a high number of approaches were available for coping with the classification task [13], we found that a Bayesian classification method was the more suitable for our purpose. In fact, we were operating on a reduced training data set (147 cases), and the Bayesian approach let us explicitly consider the available prior knowledge. More precisely, we have implemented a Naive Bayes strategy [14], making the hypothesis of conditional independence among the features given a certain class. Although this assumption is quite strong, the method it is known to be robust in a variety of situations [14,15], even in presence of conditional dependencies among features. Some recent interesting research results on the Naive Bayes classifier may be found in [16].

For applying Naive Bayes, we calculate the probability that a case belongs to class c_i, given that the set of its features $f = \{f_1, \ldots, f_M\}$ is $\hat{f}$, through the following formula.

$$P(c_i | f = \hat{f}) \propto \prod_{j=1}^{M} p(c_i) p(f_j = \hat{f}_j | c_i)$$

The method classifies a case as belonging to the class that maximizes $P(c_i | f = \hat{f})$. The conditional probabilities $p(f_j = \hat{f}_j | c_i)$ are obtained through the Bayesian update formula for discrete distributions [17,18]; in particular, we use a re-parameterized version of the update formula known as m-estimate of probability [13], that modifies the prior knowledge with the information coming from the cases of the case memory as follows.

$$p(f_j = \hat{f}_j | c_i) = \frac{m\hat{p}_{ij} + \hat{N}_{ij}}{m + D_i}$$

where $\hat{N}_{ij}$ is the number of cases in the case memory of class i whose feature f_j assumes the value $\hat{f}_j$, while D_i is the total number of cases in class i. The medical knowledge is synthesized by the prior probability distribution ($\hat{p}_{ij}$), whose reliability is expressed by the implicit number of samples m. In other words, the larger is m, the larger is the confidence of the expert on the prior.

In our application, the prior probability value (p_{ij}) was derived from expert's opinion through a technique described in [19].

3.3. Retrieval

When classification has been completed, the physician may want to retrieve only past cases belonging to the most probable class found by the Naive Bayes classifier (intra-class retrieval), or may choose a whole set of classes to be considered (inter-class retrieval).

3.3.1. Intra-class retrieval

Looking into the portion of the case memory that store the most probable class, we can apply a nearest neighbor technique for retrieving cases, so that only the closest cases are shown to physicians. The distances of the cases from the current one is calculated by using a distance metric called Euclidean-overlap metric (HEOM) [20]

$$\mathrm{HEOM} = \sqrt{\sum_f d_f(x,y)^2}$$

where $d_f(x,y)=1$, if x or y are missing; $d_f(x,y)=$ overlap (x,y), if f is a symbolic feature, (i.e. 0 if $x=y$, 1 otherwise); $|x-y|/\mathrm{range}_f$; if f is a numeric and continuous feature.

3.3.2. Inter-class retrieval

For inter-class retrieval, we have implemented another nearest neighbor technique, again able to cope with missing data, and to take into account both numeric and symbolic features. Such technique is based on the heterogeneous value difference metric (HVDM) [20].

$$\mathrm{HVDM} = \sqrt{\sum_f d_f(x,y)^2}$$

where $d_f(x,y)=1$, if x or y is missing; $d_f(x,y)=$ $\mathrm{norm}_f(x,y)$ if f is a symbolic variable; $|x-y|/4\times\sigma_f$, if f is a numeric and continuous variable.

In more detail

$$\mathrm{norm}_f(x,y) = \sum_c \left| \frac{N_{fxc}}{N_{fx}} - \frac{N_{fyc}}{N_{fy}} \right|$$

where N_{fxc}, is the number of cases in which $f=x$ in class c, and N_{fx} is the number of cases in which $f=x$ in all the considered classes; the same applies to value y.

This nearest neighbor technique may be computationally inefficient when working with large data-bases. In fact, the complexity of HVDM is known to be $O(FnC)$ [20], where F is the number of features, n the number of cases and C is the number of classes (proportional to the number of cases).

For this reason, we have adapted also a non-exhaustive search procedure, implementing a pivoting strategy (see [21] for details).

The mechanism consists in:

- Finding the median case, i.e. the case with the minimum distance from all the other cases at hand, and computing the distance between the median case and all the other cases.
- Computing the distance between the median case and the input case
- Estimating the distance between the input case and all the remaining cases by using triangle inequality, thus finding a lower and an upper bound for the distance value.
- Applying an iterative procedure that progressively eliminates cases whose interval lower bound is higher than the minimum of all the upper bounds.

At the end of the retrieval step, all the cases belonging to the selected classes are ordered on the basis of their distance from the current case. The interface shows the first ten cases in the list, expected to be the most reliable; anyway, the user is allowed to inspect the remaining ones, and to have a general view of the information stored in the classes he is working on.

In the current implementation, it is also possible to retrieve all past patient's history, so verifying the outcomes of the therapeutic choice on the metabolic control, in both short and long periods. In fact, each case belonging to a certain patient is connected to the previous and to the following one by two chains of pointers. An intra-patient, upper-level retrieval can thus be performed, in order to learn strategies that, starting from a certain condition, led the patient to a target status (for example stabilized metabolism), through a series of class transitions (see Fig. 3). If a similar

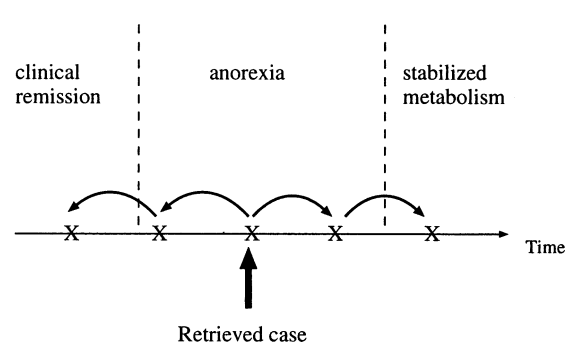

Fig. 3. An example of patient's history with class transitions.

background exists for the current patient, we may assume that therapeutic choices similar to the retrieved ones could lead to similar class transitions, and therefore to a similar conclusion after a certain time.

After the retrieval step, our tool makes some statistics on the basic decisional features of the retrieved cases; for example, it calculates the average variations in the injection number, and the proportion of cases in which the requirement trend has increased, limited to the cases with a positive outcome (i.e. with a reduced number of hypoglycemic episodes, and with a decreasing trend of HbA1c). This information is presented to the physician, who may then decide whether to rely on it, and to apply to the current protocol changes oriented in the average direction of the retrieved ones.

4. Results

Through the collaboration of the endocrinology unit of the Pediatric Clinic of Policlinico S. Matteo Hospital in Pavia, we collected 147 cases, coming from the clinical records of 29 pediatric patients. On such data, we have tested the system performances. We are aware that these results are very preliminary, due to the relatively small number of data collected; nevertheless, we believe it is interesting to present this information, as a start point for a future validation on a larger database.

Table 3
Cross-validation and validation error rates expressed in terms of ratio of incorrect classified cases and related percentages

Class	Error rate cross-validation on real data	Error rate validation on generated data
Stabilized metabolism	2/20 (10%)	69/1014 (7%)
Insulin resistant	0	0
Clinical remission	0	63/1011 (6%)
No motivation	0	0
Falsifier	4/18 (22%)	147/1009 (14%)
Change life style	12/15 (8%)	89/1011 (9%)
Celiac disease	0	72/1014 (7%)
Hormones	2/9 (22%)	0
Bulimia	0	0
Anorexia	0	0
Puberty with associated diseases	0	29/1003 (3%)
Typical puberal problems	0	98/1005 (10%)

The classification step has been tested through a leave-one-patient out cross-validation technique, both on the basic classes and on the macroclasses.

To classify a case c_j belonging to a certain subject s_i, first all cases of patient s_i were removed from the case base; then the case c_j was reinserted and classified. The posterior distributions of the class were calculated with the m-estimate formula applied to the cases belonging to the other patients. In this way, the cross-validation was performed only on cases independent from c_j.

When working on basic classes, a correct classification was obtained for the 83% of the whole case base, while in the 98% the correct class was one of the two most probable classes. A poor outcome was obtained for class falsifier, this is not surprising, since falsifiers are usually young patients that report wrong data on their diaries, to avoid complaints by parents and physicians about their nutritional and life habits. It is therefore, quite difficult, even for physicians themselves, to identify a falsifier, or to list a set of peculiar traits sufficient to describe him or her. Detailed results obtained performing the classification step are presented in the second column of Table 3.

The outcome was slightly improved when working on macroclasses, the correct class was obtained in 84% of the cases, and in 100% of the cases it was in the list of the most probable ones.

As a second step, we automatically generated more then 10 000 simulated cases, starting from the probability distribution derived from the feature occurrence in the case memory. On this expanded data-base, we performed a validation strategy, that resulted quite encouraging again, an error of less than 10% occurred in each class, excluding the class falsifier where the error rate reached 14%. Detailed results of validation experiment are shown in the third column of Table 3.

As a future verification step, we would like to conduct a prospective validation on real patients'

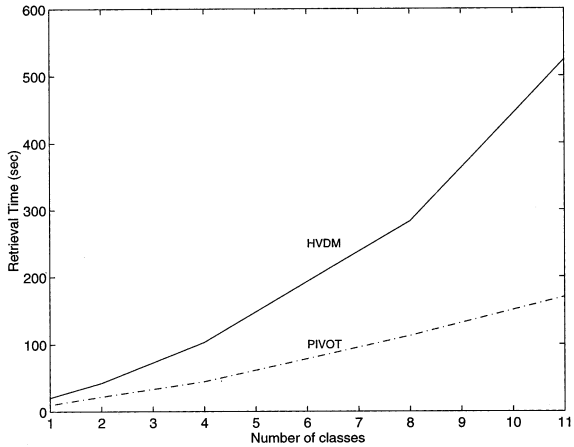

Fig. 4. The retrieval performances (in s) as a function of the number of classes.

cases, to compare the Bayesian classifier suggestions with a physician's opinion.

We tested the computational efficiency of the retrieval step on a Sun Sparc 10 machine; the classification time was of the order of milliseconds, while retrieval with the HEOM formula took 2 s in the 147 cases data-base, and up to 23 s in the 10 000 cases data-base.

Retrieval time with the HVDM algorithm ranged from 5 (on about 20 cases), to more than 500 s on the whole 10 000 cases data-base. The computation time with the pivoting algorithm grew linearly with the number of cases in the search space. In this situation retrieval time ranged from 3 (on about ten cases) to 170 s on the entire 10 000 cases data-base (see Fig. 4). Hence, by resorting to the most suitable algorithm, the system has a good performance (with respect to the application requirements) even in the presence of large data-bases.

5. Implementation

5.1. Details and integration in the T-IDDM architecture

The case-based retrieval tool described in the previous sections is one of the decision support tools of the medical workstation developed within the T-IDDM project. From an implementation point of view, it is fully integrated in a distributed, web-based environment, managed by lispweb, an extended, special-purpose web server, written in common lisp, that makes it possible to create more 'intelligent' and 'secure' applications while remaining in the context of web-based systems [9].

The components of the medical workstation application are:

- a knowledge base, composed of a structured description of the ontology of the domain under consideration;
- a relational data-base that collects all data as instances of the concepts defined at the ontological level;
- a data analysis tool, able to extract the patients' status from the collected data;
- a set of decision support tools (among which the CBR tool), that work on the objects contained in the knowledge base and on the analysis made on the data residing in the data-base;
- a user interface, providing data visualization and knowledge acquisition functionality;
- a telecommunication system, able to connect the medical workstation to the patients' houses.

In particular, the interaction between the CBR reasoning tool and the user (definition of a new case, classification and retrieval) takes place through a set of HTML pages, containing dynamically generated information, such as multicolumn tables and forms. All the cases are stored in an Oracle™ data-base, whose table structure mirrors the classes taxonomy; each leaf of the taxonomy tree matches a table, whose columns corresponds to the case features, and whose rows are instances of the class at hand.

The CBR tool does not require any additional effort from the user. Then, when the tool is invoked, it exploits all the knowledge saved in the information system, and provides the physician with a set of retrieved cases and with some statistics on their main features.

5.2. The system at work: an example of decision support

The CBR tool described in this paper is accessible through the T-IDDM user interface (an on-

line demo of the system can be found at http://aim.unipv.it/projects/tiddm). As a first step, the physician is allowed to get a summary of the previously stored patient's cases, to automatically generate a case from a periodical visit data set, or to edit a new case.

As an example, let us suppose that the physician decides to analyze a patient's visit data collected on May 21 1998 (we are using dummy names, but real patients' data). The physician is allowed to complete missing data, and to verify the correctness of the overall information; then he can start the classification procedure.

As shown in Fig. 5, as a first choice the system suggests Jane Doe to be a patient with some hormonal disorders (hypothyroidism); other probable alternatives are puberty with additional related diseases and data falsification.

By performing inter-class retrieval on the three most probable classes (hormones, puberty with associated diseases and falsifier), the system can rely on 30 cases. Nine of them are positive cases (due to the reduction of HbA1c at the following visit), and they are taken from the history of five different patients. Table 4 provides some statistical analysis on such cases.

If the physician concentrates only on the cases from the most probable class (i.e. hormones), he gets the list displayed in Fig. 6.

Table 5 shows the results of the statistical analysis performed on these nine retrieved cases, belonging to three different patients (patient 21, patient 40 and patient 44). Due to missing data about the outcome of cases number 6, 8 and 9, the system can just work on six cases, in three of which the selected therapy was not effective, because HbA1c did not decrease.

Cases number 2, 4 and 5 show a positive outcome (decreasing value of HbA1c after the application of the solution therapy, without an increase

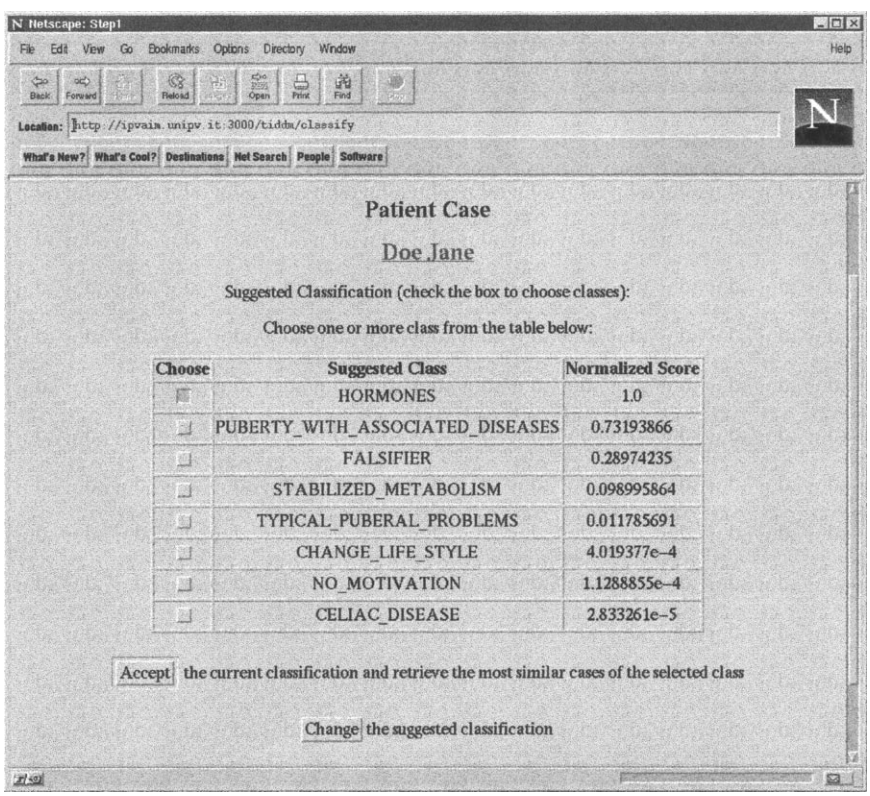

Fig. 5. Output of the classification step on the basic classes.

Table 4
Statistical analysis on the positive cases retrieved from classes hormones, puberty with associated diseases, and falsifier

Class	Positive case	Number of patients	Required trend	HbA1c	Hypos	Metabolic control
Hormones	3	1	2 Inc.,1 dec.	3 Inc.	3 Some	3 Instable
Puberty with A.D.	1	1	1 Inc.	1 Inc.	1 Lots	1 Instable
Falsifier	5	3	3 Inc., 1 stat., 1 dec.	1 Inc., 2 stat., 1 dec.	4 Some, 1 none	4 Instabel, 1 hyper

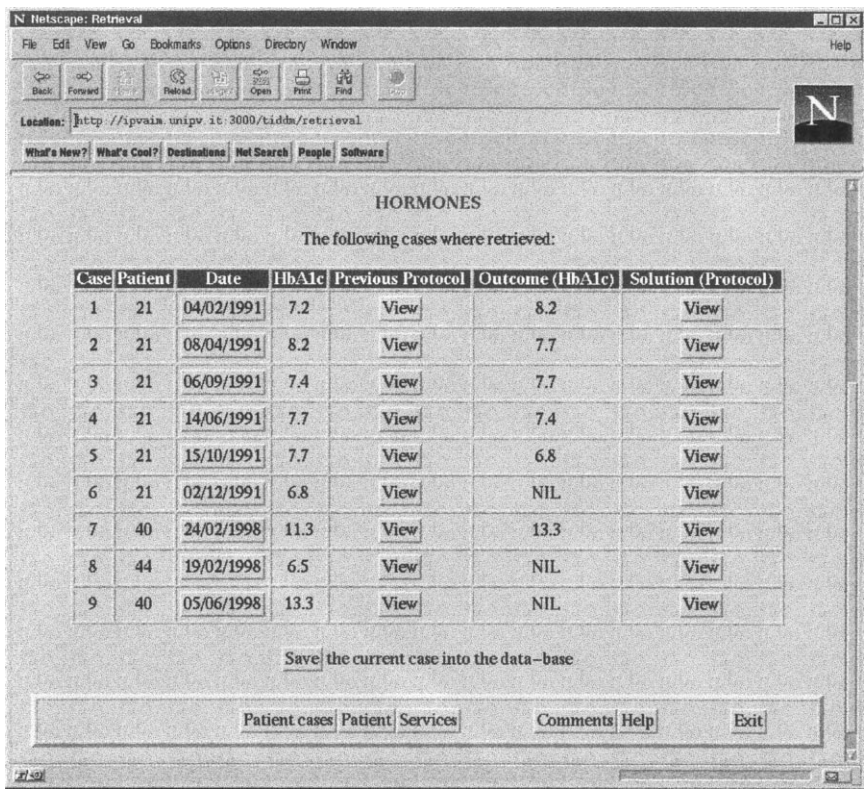

Fig. 6. Output of the intra-retrieval step.

Table 5
Detailed statistics on the cases retrieved from class hormones

Patient id	Sex	Age	Positive outcome cases	Negative outcome cases	Unreliable cases
21	Female	17	3	2	1
40	Female	16	0	1	1
44	Female	19	0	0	1

in the number of hypoglycemias). Table 6 summarizes some statistics on the three cases' features. From them the physician can observe a progressive increase in the value of insulin requirement, and may examine the current case to decide if a similar action could be suitable as well.

6. Conclusions and future research directions

In this paper, we proposed a case-based retrieval tool, able to support physicians during the revision of DM-1 patients therapy, by retrieving past cases similar to the current one from the available data-base. This system is fully integrated.with the web-based hospital information system defined in the T-IDDM project. Rather interestingly, CBR can be viewed as a system for knowledge management and behavioral learning, that does not require efforts to the health care providers other than collecting data during the periodical control visits. As a matter of fact, CBR is a 'lazy learning' paradigm, when new information (i.e. a new case) is collected, it is simply stored in the case memory, only at retrieval time learning takes place, as the system output is enriched by the presence of additional examples. This kind of paradigm is perfectly suitable for our application, as the physician, when using the system, just has to store in the data-base the information collected during a visit (but this work is routinely done, with a different purpose, which is the one of saving the visit data into the hospital information system). Moreover, the progressive collection of cases will automatically store data and results on new therapeutic solutions (e.g. the use of new insulin types, such as lispro), thus enriching the health care organization expertise, without requiring an explicit revision of the knowledge base.

We expect to obtain the first results about the clinical utility of the above presented tool through the evaluation studies that are currently carried on within the T-IDDM project. In particular the T-IDDM verification phase is taking place at the Policlinico S. Matteo, involving ten pediatric patients and three physicians, who will be using the system prototype during the next months. Patients and physicians from the other European medical centers of T-IDDM consortium will soon start with the demonstration phase as well, providing us with additional cases or patients.

From a methodological point of view, the proposed CBR system is the result of the integration of different intelligent data analysis techniques for decision support. Classification, nearest neighbor retrieval, qualitative and temporal abstractions are merged for the purpose of coping with the complex problem of DM-1 patients management. We plan to extend the work herein presented in several research directions. We do not believe that a classical CBR adaptation technique will be suitable in our application domain, automatically computing the insulin doses for the patient at hand starting from the therapies assigned to other retrieved subjects does not take into account the patient's characteristics and peculiar needs. Our tool currently calculates some statistics on the basic decisional features of the retrieved positive cases; as a future research direction, we plan to use our tool as a data mining tool, by extending its statistical analysis functionality, looking for significant correlation among patient's status parameters (HbA1c value and trend, hypoglycemias, weight excess, metabolic control) and insulin therapy descriptors (number of injections,

Table 6
Statistics on the reliable cases in class hormones[a]

Patient	Date	HbA1c	HbA1c trend	Required	Required trend	Metabolic control	Hypos	Injection number
21	08/04/1991	8.2	Increase	1.09	Decrease	Instable	Some	3
21	14/06/1998	7.7	Increase	1.1	Increase	Instable	Some	2
21	15/10/1998	7.7	Increase	1.11	Increase	Instable	Some	2

[a] Trends are calculated with respect to the previous visit.

requirement value and trend, single dose values and dose distributions over the day). The results will be analyzed to discover if it is possible to learn general strategies, typical of the different macro-classes or basic classes. In this case, classification would already provide an indication on the therapy adjustment directions to be applied in the current case. Moreover, additional strategies could be dynamically inferred from the retrieved cases, even if they do not belong to a single class. In this situation, instead of relying on precompiled indications, we would exploit the similarities between the current case and the retrieved ones; the presence in the case memory of additional cases would enhance the reliability of the results provided at this step. Finally, the most ambitious goal will be to integrate the CBR tool with other existing decision support tools, and in particular with a rule-based system, already implemented in the T-IDDM medical workstation. The search for general strategies provided by the CBR classification and retrieval would define a context in which the current patient's data should be interpreted, and would therefore reduce the search space of the rule-based system through the definition of a set of context-based meta-rules. The rule-based system would then provide a proper solution for the case at hand, by generating a personalized protocol, instead of just listing a series of different protocols, generically suitable for coping the patient's problems, as it currently does.

Acknowledgements

This paper is part of the EU TAP project T-IDDM HC 1047.

References

[1] The Diabetes Control and Complication Trial Research Group, The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus, New Engl. J. Med. 329 (1993) 977–986.

[2] E.D. Lehmann, Application of computers in clinical diabetes care, Diab. Nutr. Metab. 10 (1997) 45–59.

[3] A.M. Albisser, R.I. Harris, S. Sakkal, E. Pemberton, A. Rieflin, J.L. Nealon, Diabetes intervention in the information age, Med. Inform. 21 (1996) 297–316.

[4] K.R. Piwernetz, et al., Monitoring the targets of the Saint Vincent declaration and the implementation of quality management in diabetes care: the Diabcare initiative, Diabetic Med. 10 (1993) 371–377.

[5] D.B. Leake, J.L. Kolodner, A tutorial introduction to CBR, in: Case Based Reasoning: Experiences, Lessons and Future Directions, AAAI Press, 1996, pp. 31–65.

[6] J.L. Kolodner, Case-Based Reasoning, Morgan Kaufinann, Los Altos, CA, 1993.

[7] A. Aamodt, E. Plaza, Case-based reasoning: foundational issues, methodological variations and systems approaches, AI. Commun. 7 (1994) 39–59.

[8] R. Bellazzi, C. Cobelli, E. Gomez, M. Stefanelli, The T-IDDM Project: Telematic management of insulin dependent diabetes mellitus, in: M. Bracale, F. Denoth (Eds.), Health Telematics'95, 1995, pp. 271–276.

[9] A. Riva, R. Bellazzi, M. Stefanelli, A web-based system for the intelligent management of diabetic patients, MD. Computing 14 (1997) 360–364.

[10] J. Kolodner, R. Kolodner, Using experience in clinical problem solving: introduction and framework, IEEE Trans. Syst. Man Cybernetics 17 (1987) 420–431.

[11] K.J. Hammond, Case-Based Planning: Viewing Planning as a Memory Task, Academic Press, New York, 1989.

[12] I. Watson, Applying Case-Based Reasoning: Techniques for Enterprise Systems, Morgan Kaufinann, Los Altos, CA, 1997.

[13] T. Mitchell, Machine Learning, Mc Graw Hill, New York, 1997.

[14] I. Kononenko, Inductive and Bayesian learning in medical diagnosis, Appl. Artif. Intell. 7 (1993) 317–337.

[15] I. Zelic, I. Kononenko, N. Lavrač, V. Vuga, Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries, in: Proceedings of IDAMAP 97 workshop, IJCAI 97, Nagoya, Japan, 1997, pp. 61–67.

[16] S. Monti, G. Cooper, The impact of modeling the dependencies among patient findings on classification accuracy and calibration, JAMIA, Symp. Suppl. 1998, 592–596.

[17] D. Spiegelhalter, A. Dawid, S. Lauritzen, R. Cowell, Bayesian analysis in expert systems, Stat. Sci. 8 (1993) 219–283.

[18] A. Riva, R. Bellazzi, Learning temporal probabilistic causal models from longitudinal data, Artif. Intell. Med. 8 (1996) 217–234.

[19] S. Montani, R. Bellazzi, L. Portinale, S. Fiocchi, M. Stefanelli, A case-based retrieval system for diabetic patients therapy, In: Proceedings of IDAMAP 98 workshop, ECA1 98, Brighton, UK, 1998, pp. 64–70.

[20] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J. Artif. Intell. Res. 6 (1997) 1–34.

[21] L. Portinale, P. Torasso, D. Magro, Selecting most adaptable diagnostic solutions through pivoting-based retrieval, in: Lecture Notes in Artificial Intelligence, vol. 1266, Springer, Berlin, 1997, pp. 277–288.