USING CASE-BASED REASONING TO EXPLAIN EXCEPTIONAL CASES

Rainer Schmidt[1] and Olga Vorobieva[1,2]
[1] *Institute for Medical Informatics and Biometry, University of Rostock, Rostock, Germany*
[2] *Sechenov Institute of Evolutionary Physiology and Biochemistry, St.Petersburg, Russia*

Keywords: Case-Based Reasoning, Medicine, Exceptional Cases.

Abstract: In medicine many exceptions occur. In medical practice and in knowledge-based systems too, it is necessary to consider them and to deal with them appropriately. In medical studies and in research, exceptions shall be explained. We present a system that helps to explain cases that do not fit into a theoretical hypothesis. Our starting points are situations where neither a well-developed theory nor reliable knowledge nor a case base is available at the beginning. So, instead of reliable theoretical knowledge and intelligent experience, we have just some theoretical hypothesis and a set of measurements. In this paper, we propose to combine Case-Based Reasoning with a statistical model. We use Case-Based Reasoning to explain those cases that do not fit the model. The case base has to be set up incrementally, it contains the exceptional cases, and their explanations are the solutions, which can be used to help to explain further exceptional cases.

1 INTRODUCTION

In medical studies and in research, exceptions shall be explained. We have developed ISOR, a case-based dialogue system that helps doctors to explain exceptional cases. ISOR deals with situations where neither a well-developed theory nor reliable knowledge nor a proper case base is available. So, instead of reliable theoretical knowledge and intelligent experience, we now have just some theoretical hypothesis and a set of measurements. In such situations the usual question is, how do measured data fit to theoretical hypotheses. To statistically confirm a hypothesis it is necessary that the majority of cases fit the hypothesis. Mathematical statistics determines the exact quantity of necessary confirmation (Kendall and Stuart, 1979). However, usually a few cases do not satisfy the hypothesis. We examine these cases to find out why they do not satisfy the hypothesis. ISOR offers a dialogue to guide the search for possible reasons in all components of the data system. The exceptional cases belong to the case base. This approach is justified by a certain mistrust of statistical models by doctors, because modelling results are usually unspecific and "average oriented" (Hai, 2002), which means a lack of attention to individual "imperceptible" features of concrete patients.

The usual case-based reasoning (CBR) assumption is that a case base with complete solutions is available. Our approach starts in a situation where such a case base is not available but has to be set up incrementally. So, we must

1. Construct a model,
2. Point out the exceptions,
3. Find causes why the exceptional cases do not fit the model, and
4. Develop a case base.

So, we combine case-based reasoning with a model, in this specific situation with a statistical one. The idea to combine CBR with other methods is not new. For example Care-Partner resorts to a multi-modal reasoning framework for the co-operation of CBR and rule-based reasoning (RBR) (Bichindaritz et al, 1998). Another way of combining hybrid rule bases with CBR is discussed by Prentzas and Hatzilgeroudis (Prentzas and Hatzilgeroudis, 2002). The combination of CBR and model-based reasoning is discussed in (Shuguang et al, 2000). Statistical methods are used within CBR mainly for retrieval and retention (e.g. Corchado et al, 2003; Rezvani and Prasad, 2003). Arshadi proposes a method that combines CBR with statistical methods like clustering and logistic regression (Arshadi and Jurisica, 2005).

Schmidt R. and Vorobieva O. (2008).
USING CASE-BASED REASONING TO EXPLAIN EXCEPTIONAL CASES.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - AIDSS*, pages 119-124
DOI: 10.5220/0001669501190124

1.1 Dialyse and Fitness

Hemodialysis means stress for a patient's organism and has significant adverse effects. Fitness is the most available and a relative cheap way of support. It is meant to improve a physiological condition of a patient and to compensate negative dialysis effects. One of the intended goals of this research is to convince the patients of the positive effects of fitness and to encourage them to make efforts and to go in for sports actively. This is important because dialysis patients usually feel sick, they are physically weak, and they do not want any additional physical load (Davidson et al, 2005).
At our University clinic in St. Petersburg, a specially developed complex of physiotherapy exercises including simulators, walking, swimming etc. was offered to all dialysis patients but only some of them actively participated, whereas some others participated but were not really active. The purpose of this fitness offer was to improve the physical conditions of the patients and to increase the quality of their lives.

2 EXPLANATION MODEL

For each patient a set of physiological parameters is measured. These parameters contain information about burned calories, maximal power achieved by the patient, his oxygen uptake, his oxygen pulse (volume of oxygen consumption per heart beat), lung ventilation and others. There are also biochemical parameters like haemoglobin and other laboratory measurements. More than 100 parameters were planned for every patient. But not all of them were really measured.
Parameters are supposed to be measured four times during the first year of participating in the fitness program. There is an initial measurement followed by a next one after three months, then after six months and finally after a year. Unfortunately, since some measurements did not happen, many data are missing. Therefore the records of the patients often contain different sets of measured parameters.
It is necessary to note that parameter values of dialysis patients essentially differ from those of non-dialysis patients, especially of healthy people, because dialysis interferes with the natural, physiological processes in an organism. In fact, for dialysis patients all physiological processes behave abnormally. Therefore, the correlation between parameters differs too.
For statistics, this means difficulties in applying statistical methods based on correlation and it limits the usage of a knowledge base developed for normal people. Non-homogeneity of observed data, many missing values, many parameters for a relatively small sample size, all this makes our data set practically impossible for usual statistical analysis.
Our data set is incomplete therefore we must find additional or substitutional information in other available data sources. They are databases – the already existent Individual Base and the sequentially created Case Base and the medical expert as a special source of information.

2.1 Setting up a Model

We start with a medical problem that has to be solved based on given data. In our example it is: "Does special fitness improve the physiological condition of dialysis patients?" More formally, we have to compare physical conditions of active and non-active patients. Patients are divided into two groups, depending on their activity, active patients and non-active ones.
According to our assumption, active patients should feel better after some months of fitness, whereas non-active ones should feel rather worse. We have to define the meaning of "feeling better" and "feeling worse" in our context. A medical expert selects appropriate factors from ISOR's menu. It contains the list of field names from the observed database.
The expert selects the following main factors

- F1: O2PT - Oxygen pulse by training
- F2: MUO2T - Maximal Uptake of Oxygen by training
- F3: WorkJ – performed Work (Joules) during control training

Subsequently the "research time period" has to be determined. Initially, this period was planned to be twelve months, but after a while the patients tend to give up the fitness program. This means, the longer the time period, the more data are missing. Therefore, we had to make a compromise between time period and sample size. A period of six months was chosen.
The next question is whether the model shall be quantitative or qualitative? The observed data are mostly quantitative measurements. The selected factors are of quantitative nature too. On the other side, the goal of our research is to find out whether physical training improves or worsens the physical condition of the dialysis patients.
We do not have to compare one patient with another patient. Instead, we compare every patient with his own situation some months ago, namely just before the start of the fitness program. The success shall not be measured in absolute values, because the health statuses of patients are very different. Thus, even a

modest improvement for one patient may be as important as a great improvement of another. Therefore, we simply classify the development in two categories: "better" and "worse". Since the usual tendency for dialysis patients is to worsen in time, we added those few patients where no changes could be observed to the category" better".
The three main factors are supposed to describe the changes of the physical conditions of the patients. The changes are assessed depending on the number of improved factors:

- Weak version of the model: at least one factor has improved
- Medium version of the model: at least two factors have improved
- Strong version of the model: all three factors have improved

The final step means to define the type of model. Popular statistical programs offer a large variety of statistical models. Some of them deal with categorical data. The easiest model is a 2x2 frequency table. Our "Better/ Worse" concept fits this simple model very well. So the 2x2 frequency table is accepted. The results are presented in Table 1.

Table 1. Results of Fisher's Exact Test, performed with an interactive Web-program: http://www.matforsk.noIola/fisher.htm. The cases printed in bold have to be explained.

Improve-ment mode	Patient's physical condition	Active	Non-active	Fisher Exact p
Strong	Better	28	2	< 0.0001
	Worse	**22**	21	
Medium	Better	40	10	< 0.005
	Worse	**10**	12	
Weak	Better	47	16	< 0.02
	Worse	**3**	6	

According to our assumption after six months of active fitness the conditions of the patients should be better.
Statistical analysis shows a significant dependence between the patient's activity and improvement of their physical condition. Unfortunately, the most popular Pearson Chi-square test is not applicable here because of the small values "2" and "3" in Table 1. But Fisher's exact test (Kendall and Stuart, 1979) can be used. In the three versions shown in Table 1 a very strong significance can be observed. The smaller the value of p is, the more significant the dependency.

Exceptions. So, the performed Fisher test confirms the hypothesis that patients doing active fitness achieve better physical conditions than non-active ones. However, there are exceptions, namely active patients whose health conditions did not improve.
Exceptions should be explained. Explained exceptions build the case base. According to Table 1, the stronger the model, the more exceptions can be observed and have to be explained. Every exception is associated with at least two problems. The first one is "Why did the patient's condition get worse?" Of course, "worse" is meant in terms of the chosen model. Since there may be some factors that are not included in the model but have changed positively, the second problem is "What has improved in the patient's condition?" To solve this problem we look for significant factors where the values improved.
In the following section we explain the set-up of a case base on the strongest model version.

2.2 Setting up a Case Base

We intend to solve both problems (mentioned above) by means of CBR. So we begin to set up the case base up sequentially. That means, as soon as an exception is explained, it is incorporated into the case base and can be used to help explaining further exceptional cases. We chose a random order for the exceptional cases. In fact, we took them in alphabetical order.
The retrieval of already explained cases is performed by keywords. The main keywords are "problem code", "diagnosis", and "therapy". In the situation of explaining exceptions for dialysis patients the instantiations of these keywords are "adverse effects of dialysis" (diagnosis), "fitness" (therapy), and two specific problem codes. Besides the main keywords additional problem specific ones are used. Here the additional key is the number of worsened factors. Further keywords are optional. They are just used when the case base becomes bigger and retrieval is not simple any longer.
However, ISOR does not only use the case base as knowledge source but further sources are involved, namely the patient's individual base (his medical history) and observed data (partly gained by dialogue with medical experts). Since in the domain of kidney disease and dialysis the medical knowledge is very detailed and much investigated but still incomplete, it is unreasonable to attempt to create an adequate knowledge base. Therefore, a medical expert, observed data, and just a few rules serve as medical knowledge sources.

2.2.1 Expert Knowledge and Artificial Cases

Expert's knowledge can be used in many different ways. Firstly, we use it to acquire rules, and secondly, it can be used to select appropriate items from the list of retrieved solutions, to propose new solutions and last but not least – to create artificial cases.
Initially, artificial cases are created by an expert, afterwards they can be used in the same way as real cases. They are created in the following situation. An expert points out a factor F as a possible solution for a query patient. Since many values are missing, it can happen that just for the query patient values of factor F are missing. The doctor's knowledge in this case can not be applied, but it is sensible to save it anyway. Principally there are two different ways to do this. The first one means to generate a correspondent rule and to insert it into ISOR's algorithms. Unfortunately, this is very complicated, especially to find an appropriate way for inserting such a rule. The alternative is to create an artificial case. Instead of a patient's name an artificial case number is generated. The other attributes are either inherited from the query case or declared as missing. The retrieval attributes are inherited. This can be done by a short dialogue (Figure 1) and ISOR's algorithms remain intact. Artificial cases can be treated in the same way as real cases, they can be revised, deleted, generalised etc.

2.2.2 Solving the Problem "Why Did some Patients Conditions Became Worse?"

As results we obtain a set of solutions of different origin and different nature. There are three categories of solution: additional factor, model failure, and wrong data.

Additional Factor. The most important and most frequent solution is the influence of an additional factor. Only three main factors are obviously not enough to describe all medical cases. Unfortunately, for different patients different additional factors are important. When ISOR has discovered an additional factor as explanation for an exceptional case, the factor has to be confirmed by a medical expert before it can be accepted as a solution. One of these factors is Parathyroid Hormone (PTH). An increased PTH level sometimes can explain a worsened condition of a patient (Davidson et al, 2005). PTH is a significant factor, but unfortunately it was measured only for some patients.
Some exceptions can be explained by indirect indications. One of them is a very long time of dialysis (more than 60 months) before a patient began with the training program.
Another solution was a phosphorus blood level. We used the principle of artificial cases to introduce the factor phosphorus as a new solution. One patient's record contained many missing data. The retrieved solution meant high PTH, but PTH data in the current patient's record was missing too. The expert proposed an increased phosphorus level as a possible solution. Since data about phosphorus data was missing too, an artificial case was created, that inherited all retrieval attributes of the query case while the other attributes were recorded as missing. According to the expert high phosphorus can explain the solution. Therefore it is accepted as an artificial solution or a solution of an artificial case.

Model Failure. We regard two types of model failures. One of them is deliberately neglected data. Some data had been neglected. As a compromise we just considered data of six months and further data of a patient might be important. In fact, three of the patients did not show an improvement in the considered six month but in the following six months. So, they were wrongly classified and should really belong to the "better" category. The second type of model failure is based on the fact that the two-category model was not precise enough. Some exceptions could be explained by a tiny and not really significant change in one of the main factors. Wrong data are usually due to a technical mistake or to not really proved data. For example, one patient was reported as actively participating in the fitness program but really was not.

2.3 Illustration of the Program Flow

Figure 1 shows the main dialogue of ISOR where the user at first sets up a model (steps one to four), subsequently gets the result and an analysis of the model (steps five to eight), and then attempts to find explanations for the exceptions (steps nine and ten). Finally the case base is updated (steps eleven and twelve). On the menu (Figure 1) we have numbered the steps and explain them in detail.
At first the user has to set up a model. To do this he has to select a grouping variable. In this example CODACT was chosen. It stands for "activity code" and means that active and none active patients are to be compared. Provided alternatives are the sex and the beginning of the fitness program (within the first year of dialysis or later). In another menu the user can define further alternatives. Furthermore, the user has to select a model type (alternatives are "strong", "medium", and "weak"), the length of time that should be considered (3, 6 or 12 months), and main factors have to be selected. The list contains the factors from the observed database. In the example

three factors are chosen: O2PT (oxygen pulse by training), MUO2T (maximal oxygen uptake by training), and WorkJ (work in joules during the test training). In the menu list, the first two factors have alternatives: “R” instead of “T”, where “R” stands for state of rest.
When the user has selected these items, the program calculated the table. “Better” and “worse” are meant in the sense of the chosen model, in the example of the strong model. ISOR does not only calculate the table but additionally extracts the exceptional patients from the observed database. In the menu, the list of exceptions shows the code names of the patients. In the example patient “D5” is selected” and all further data belong to this patient.

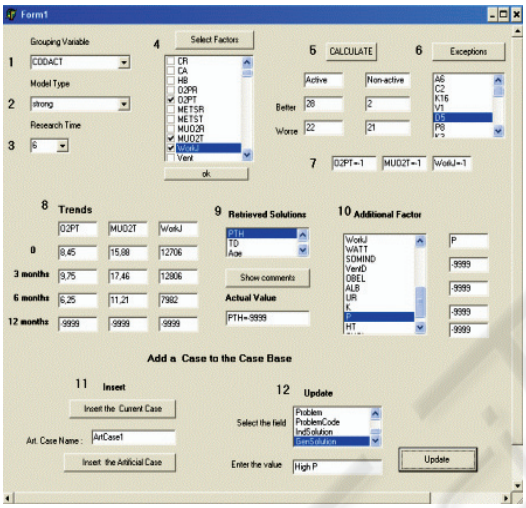

Figure 1: ISOR’s program flow.

The goal is to find an explanation for the exceptional case “D5”. In point seven of the menu it is shown that all selected factors worsened (-1), and in point eight the factor values according to different time intervals are depicted. All data for twelve months are missing (-9999).
The next step means creating an explanation for the selected patient “D5”. From the case base ISOR retrieves general solutions. The first retrieved one in this example, the PTH factor, denotes that the increased Parathyroid hormone blood level may explain the failure. Further theoretical information (e.g. normal values) about a selected item can be received by pressing the button “show comments”. The PTH value of patient “D5” is missing (-9999). From menu point ten the expert user can select further probable solutions. In the example an increased phosphorus level (P) is suggested. Unfortunately, phosphorus data are missing too. However, the idea of an increased phosphorus level as a possible solution shall not be lost. So, an artificial case has to be generated.
The final step means inserting new cases into the case base. There are two sorts of cases, query cases and artificial cases. Query cases are stored records of real patients from the observed database. These records contain a lot of data but they are not structured. The problem and its solution transform them into cases and they get a place in the case base. Artificial cases inherit the key attributes from the query cases (point seven in the menu). Other data may be declared as missing. By the update function the missing data can be inserted later on. In the example of the menu, the generalised solution “High P” is inherited, it may be retrieved as a possible solution (point 9 of the menu) for future cases.

3 EXAMPLE

By an example we demonstrate how ISOR attempts to find explanations for exceptional cases. Because of data protection we cannot use a real patient. It is an artificial case but it is a typical situation.

Query Patient. a 34-year old woman started with fitness after five months of dialyse. Two factors worsened Oxygen pulse and Oxygen uptake, and consequently the condition of the patient was assessed as worsened too.

Problem. Why the patient’s condition deteriorated after six months of physical training.

Retrieval. The number of worsened factors is used as an additional keyword in order to retrieve all cases with at least two worsened factors.

Case Base. It does not only contain cases but more importantly a list of general solutions. For each of the general solutions there exists a list that contains the concrete solutions based on the cases in the case base.

The list of general solutions contains five items:

1.) Concentration of Parathyroid Hormone
2.) Period of dialyse is too long.
3.) An additional disease
4.) A patient was not very active during the fitness program.
5.) A patient is very old.

Individual Base. The patient suffers from a chronic disease, namely from asthma.

Adaptation. Since the patient started with fitness already after five months of dialyse, the second general solution can be excluded. The first general solution might be possible, though the individual base does not contain any information about PTH. Further lab tests showed PTH = 870. So, PTH is a solution.
Since an additional disease, bronchial asthma, is found in the individual base, this solution is checked. Asthma is not contained as solution in the case base, but the expert concludes that asthma can be considered as a solution. Concerning the remaining general solutions, the patient is not too old and she proclaims that she was active at fitness.

Adapted Case. The solution consists of a combination of two factors, namely a high PTH concentration and an additional disease, asthma.

4 CONCLUSIONS

In this paper, we have proposed to use CBR in ISOR to explain cases that do not fit a statistical model. Here we presented one of the simplest statistical models. However, it is relatively effective, because it demonstrates statistically significant dependencies, in our example between fitness activity and health improvement of dialysis patients, where the model covers about two thirds of the patients, whereas the other third can be explained by applying CBR. Since we have chosen qualitative assessments (better or worse), very small changes appear to be the same as very large ones. We intend to define these concepts more precisely, especially to introduce more assessments. The presented method makes use of different sources of knowledge and information, including medical experts. It seems to be a very promising method to deal with a poorly structured database, with many missing data, and with situations where cases contain different sets of attributes.

ACKNOWLEDGEMENTS

We thank Professor Alexander Rumyantsev from the Pavlov State Medical University for his close co-operation. Furthermore we thank Professor Aleksey Smirnov, director of the Institute for Nephrology of St-Petersburg Medical University and Natalia Korosteleva, researcher at the same Institute for collecting and managing the data.

REFERENCES

Arshadi, N., Jurisica, I., 2005. Data Mining for Case-based Reasoning in high-dimensional biological domains. *IEEE Transactions on Knowledge and Data Engineering* 17 (8). 1127-1137

Bichindaritz, I., Kansu, E., Sullivan, K.M., 1998. Case-based Reasoning in Care-Partner. In: *EWCBR-98, European Workshop on Case-Based Reasoning*. Springer, Berlin, 334-345

Corchado, J.M., Corchado, E.S., Aiken, J., Fife, C., Fernandez, F., Gonzalez, M., 2003. Maximum likelihood Hebbian learning based retrieval method for CBR systems. In: *ICCBR-2003, International Conference on Case-Based Reasoning*. Springer, Berlin, 107-121

Davidson, A.M., Cameron, J.S., Grünfeld, J.-P. (eds.), 2005. *Oxford Textbook of Nephrology*, Volume 3. Oxford University Press

Hai, G.A., 2002. *Logic of diagnostic and decision making in clinical medicine*. Politheknica publishing, St. Petersburg

Kendall, M.G., Stuart, A., 1979. *The advanced theory of statistics*. Macmillan publishing, New York

Prentzas, J., Hatzilgeroudis, I., 2002. Integrating Hybrid Rule-Based with Case-Based Reasoning. In *ECCBR-2002, European Conference on Case-Based Reasoning*. Springer, Berlin 336-349

Rezvani, S., Prasad, G., 2003. A hybrid system with multivariate data validation and Case-based Reasoning for an efficient and realistic product formulation. In *ICCBR-2003, International Conference on Case-Based Reasoning*. Springer, Berlin (2003) 465-478

Shuguang, L., Qing, J., George, C., 2000: Combining case-based and model-based reasoning: a formal specification. In *APSEC'00, Asia Pacific Software Engineering Conference*, 416