Predictive Medical Data Mining: Case Study

Jiří Kléma, Olga Štěpánková, Lenka Nováková

Department of Cybernetics, CTU Prague,
Technicka 2, 166 27 Prague 6, Czech Republic,
{klema,step,novakl}@labe.felk.cvut.cz

Abstract. This paper presents a case study concerning scheduling and resource allocation issues at a spa. The paper is data-mining oriented. It discusses and describes how the history data can be used as a source for data-mining leading to discovery of rules or algorithms useful for prediction of resources requirements. In particular, we focused to identify groups of patients which appear frequently in the training set and which exhibit characteristic behavior or requirements of spa utilities. Then we predicted a set of health procedures to be passed for each member of such group. This approach resulted in a health procedure prediction algorithm satisfactory for early and convenient scheduling.

1 Introduction

Data mining (Fayyad et al. 1996) is concerned with finding patterns in data which are interesting (according to some user-defined measure of interestingness) and valid (according to some user defined measure of validity). Related research areas include database technology and data warehouses, statistics, machine learning, pattern recognition and soft computing, text and web mining and visualization. Numerous data mining methods exist, including predictive data mining methods, which typically result in that can be used for prediction and classification, and descriptive data mining methods that discover structure within data such as associations, clusters, etc. If the results of data mining are non black-box symbolic descriptions, the result of predictive data mining is usually a domain model, whereas the aim of descriptive data mining is the discovery of individual patterns (Lavrac, Grobelnik, 2003).

Modern medicine generates, almost daily, huge amounts of heterogeneous data. For example, medical data may contain images, signals like ECG, clinical information like temperature, cholesterol levels, etc., as well as the physician's interpretation. With the widespread use of medical information systems that include databases which have recently featured explosive growth in their sizes, physicians and medical researchers are faced with a problem of making use of the stored data. The traditional manual data analysis has become insufficient, and methods for efficient computer-assisted analysis indispensable, in particular those of data mining and other related techniques of knowledge discovery in databases and intelligent data analysis. This paper presents process of solution of a particular medical predictive task regarding early prediction of resource allocation at a spa.

1.1 Resource Allocation at a Spa

Spa facilities offer a set of health procedures to treat the medical problems of patients who attend a health farm for treatment in a given time period. Each patient obtains individual treatment – a set of procedures, assigned to the patient by the spa physician. The physician's recommendation is based on results of a careful examination of the patient upon his or her arrival. But this recommendation is not enough to ensure that each patient receives exactly those procedures required. To reach this goal all necessary resources of the spa (human resources – skilled personnel, and technical equipment – e.g., a bath tube or diathermia) have to be available in the required quantities for the patient.

The spa organization aims to provide appropriate individual treatment for each of its patients. It is vital for the spa administration to know in advance (before the arrival of a new group of patients) what the total amount of required procedures in the considered period will be. Such prediction of resource requirements cannot be based on data from patient's medical history, as it is not available to the spa administration before the patient enters the health farm. The only data about patients available for the prediction process is the data the spa administration receives as a part of the patient's application for treatment (e.g., sex, age, planned length of stay, type of disorder). Lauryn v.o.s., a company developing information systems for health farms (known here as the SPA company) and AI Group of CTU established a data mining project (abbreviated as the *SPA project* in the rest of the paper) with the following objectives:

– to study the possibility to predict a few weeks in advance the overall number of prescriptions of the specific health procedures within a specific time period (one week),
– to identify previously unknown groups of clients (e.g., women 50 to 60 years old having the same disorders) exhibiting characteristic behaviors or requirements for procedures. For each group, predict the corresponding treatment (the set of all procedures).

The above-mentioned objectives suggest predictive orientation of the SPA project. The main goal is early prediction of resource allocation connected with demand for specific health procedures. Nevertheless, search for interesting patterns (description of characteristic groups of clients) seems to be another project goal.

2 Data Preprocessing

Our intention is to predict spa resources requirements given all available information about the group of patients to be present at the spa in the considered week. There are three premises to such a data-mining exercise:

- Information available about each patient before his/her arrival is a significant factor in determining the schedule of procedures that will be prescribed by the spa physician to the considered person.
- Treatment schedules prescribed by different spa physicians are consistent.
- The full set of procedures offered by the spa is fixed, no procedures are added or removed.

If all the premises are true, the history data could be used to search for prediction rules. The data set was available as a relational database consisting of four relational tables – *Patients* (patient description, attributes as age, sex, disorder, cure type, …), *Procedures* (procedure description, name, price, type), *Prescriptions* (which procedure was prescribed to which patient and when) and *Forbidden_combinations* (procedures that cannot be scheduled at the same day). Our considered history data set, referred to as training data, is based on real life data about all 17 953 patients attending one specific spa resort during the years 1999, 2000 and their treatment schedules (protection of patient's personal data has been ensured by the administration of the spa). Data from the same facility covering the year 2001 are used as a test set.

In this project phase, we also used two vizualization tools: scatter-plot matrix and parallel coordinates. *Scatter-plot matrix* shows relation between two selected attributes. This method offers the first view on the considered data. We used this method to choose interesting attributes and to appoint interval for each attribute. A simple example of this type of vizualization is shown in Figure 1. It confirms that there is a linear dependency between attributes *age* and *born*. The attributes are redundant and one of them can only be used in farther project phases. *Parallel coordinates* method can be applied to find relation (dependency) between attributes.

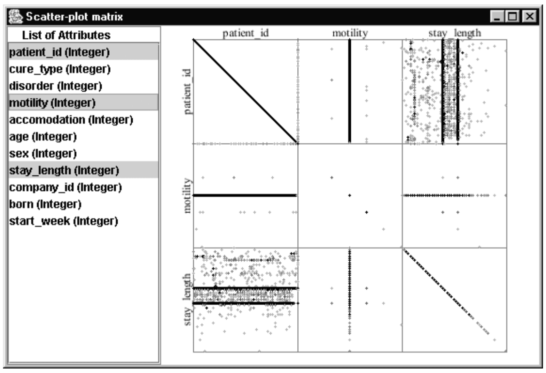
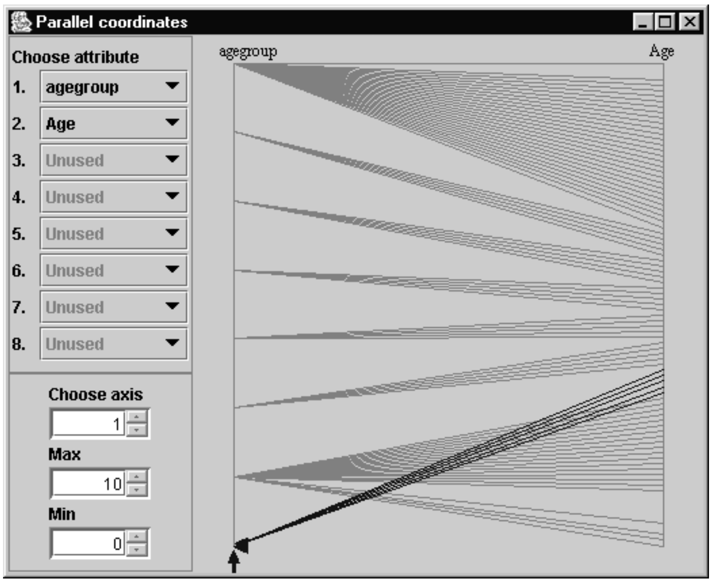


Fig. 1. The linear dependency between attributes *age* and *born*.



Choose attribute

1. agegroup
2. Age
3. Unused
4. Unused
5. Unused
6. Unused
7. Unused
8. Unused

Choose axis: 1

Max: 10

Min: 0

Fig. 2. An example of improper discretization of *age*.

2.1 Data Aggregation

It is interesting to note that there are two principal modeling directions. One direction starts by predicting all procedures to be prescribed to a single patient. The total for one week is obtained as a sum of predictions for individual patients actually present at the spa (the individual centered approach). The second modeling direction is based on the fact that the patient description is only approximate (the spa staff are provided with only basic information about clients). Consequently, all the patients can be split among characteristic groups containing patients with the same/similar description (the aggregated approach). The total for one week is obtained as the sum of predictions for the individual groups, and each week is represented entirely by a vector of representations of the patient groups.

A new aggregated table generated with a heavy support of SumatraTT consists of 81 attributes (Gr1, …, Gr81) corresponding to the upper mentioned groups and 35 additional attributes representing the procedures (Pr1, …, Pr40 – five of procedures are never prescribed). One record summarizes data concerning all patients present in the spa during a single week. Let us specify the contents of the table for the week *i*:

– Gr_{ik} is the number of days spent by patients belonging to the k-th group during the i-th week.
– Pr_{ij} is the total number of all prescriptions of the j-th procedure during the i-th week.

Aggregated representation has several advantages. It is very concise, with negligible information loss (considering information relevant for prediction), and it already pre-suggests candidates for characteristic groups with respect to the individual procedures.

3 Predictive Modeling

This section presents several different modeling approaches to the solved task. Majority of models deals with the aggregated representation, one of the models is based on the original data and follows the individual centered approach (Bayesian statistical model).

3.1 Simple Regression

A simple regression approach represents the most straightforward solution of the given predictive task in terms of the selected representation. The model is most simplified as it assigns all the patients to the same group while ignoring any patient specific information. It utilizes the overall number of patient-days in the predicted week only:

$$Pr_{ij}^{pred} = a_j \, GrAll_i \tag{1}$$

where $GrAll_i$ is the number of days spent by all the patients during the i-th week $(GrAll_i = Gr_{i1} + \ldots + Gr_{i81})$,

a_j is the regression coefficient learnt for the j-th procedure on the training data (average number of the j-th procedures prescribed per patient and day).

Apparently, this non-informed prediction represents the worst case prediction result and when compared with well-informed models it can give a basic outline of utility of patient description. When averaged over all the procedures, the simple regression model gives MAPE about 17.5%.

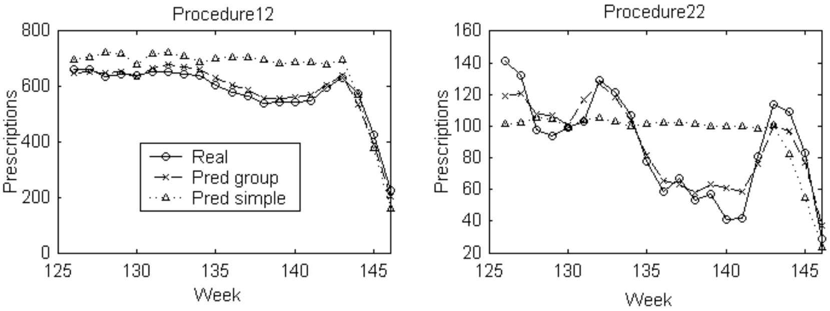
3.2 Regression by Patient Groups

The regression by patient groups tries to put in use differences among the individual patient groups. Instead of learning the single regression coefficient a_j it learns separate coefficients for all the considered groups. The total in the predicted week is then the sum of predictions obtained for the considered groups:

$$Pr_{ij}^{pred} = \sum_{k=1}^{81} a_{jk} \, Gr_{ik} \tag{2}$$

where a_{jk} is the regression coefficient learnt on the training data for the j-th procedure and the k-th group.

When averaged over all the procedures, the regression by patient groups gives MAPE about 14.5% (further denoted as the general MAPE), i.e., it brings general improvement as compared with the simple regression (17.5%). When regarding the individual health procedures, two different points of view have been considered: the above-mentioned MAPE and ability to follow the real trends. Considering these criteria, the regression by groups is significantly better for 6 procedures (see Figure 3), on the other hand it does not show any significant difference in the other 29 procedures (see Figure 4).



a) b)

Fig. 3. Examples of health procedures for which the regression by groups (Pred group) significantly outperforms the simple regression (Pred simple).

This diversity confirms that some procedures show little sensitivity to patient characteristics and they are prescribed with a nearly uniform distribution over the patient set. For these procedures, the simple regression either represents the competent solution immediately (see Figure 4a) or it can be a good solution having reduced a systematic prediction error, i.e., regarding long-term changes of a_j (see Figure 4b). This issue can be solved by a heuristic approach presented in Section 3.4.

The last minority of procedures might ask for a different group definition. These procedures can be predicted on bases of the procedure specific group definition that can be precisely tuned regarding the target procedure. We have applied the LISp-Miner system (Rauch, Simunek, 2000) to derive specific association rules describing the strong groups relevant to the critical procedures. The group segmentations can be surprisingly simple. For example, application of the single association rule resulting into two-group segmentation improves the prediction ability to follow the real trends of Pr37:

Cure_type(1, ..., 6) and Sex(Woman) → "Pr37 is likely to be prescribed"

The given rule splits the patient set between two almost equal sized sub-groups. For the first group, the frequency of Pr37 prescriptions is about twice higher than in the original set. On the contrary, the second group shows almost zero frequency of Pr37 prescriptions.

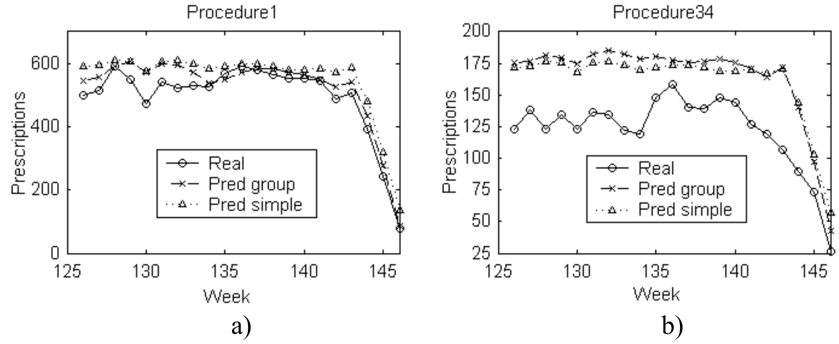


Fig. 4. Examples of health procedures for which both the regression approaches show similar performance, either working well (a) or having high MAPE (b).

3.3 Instance-Based Reasoning

Instance-based reasoning methods such as the nearest neighbor and locally weighted regression are conceptually straightforward approaches to approximating real-valued or discrete-valued target functions. Learning in these algorithms consists of simply storing the presented training data. When a new query instance is encountered, a set of similar related instances is retrieved from the memory and used to classify the new query instance. The key difference between this approach and other learning methods is that the IBR approach can construct a different approximation to the target function for each distinct query instance to be classified/predicted. Another advantage is that this learning approach leaves a large space to tailor the final solution to the given domain in terms of selection of the distance metric, kernel function, score function, feature-weighting or dealing with case memory (training instances). However, this adjustment requires many parameters to be set. Parameter setting is a non-trivial task that influences the knowledge model returned by the algorithm. Parameter values are usually set approximately, according to some characteristics of the target problem obtained in various ways. The usual way is to use background knowledge about the target problem (if available) and perform some testing experiments.

iBARET (Instance-Based REasoning Tool) is an implementation of the principal characteristics of IBR theory developed at the Gerstner Laboratory, CTU. It is a universal tool for modeling and predicting in domains described by a number of numeric or symbolic values with restricted or no background knowledge. It is suitable for applications both in classification and regression tasks. In order to provide an effective structure of the learning system, iBARET's architecture is split into two parts: a server and a client communicating via the Case Query Language (CQL). The server part is a database complemented by a search engine, which searches the case memory for the nearest neighbors, while the client corresponds to a modeling interface, which generates queries to the database, evaluates its responses and adjusts the model parameters.

iBARET incorporates various techniques and standards. For example, it supports the Predictive Model Markup Language (PMML) for model exchange and therefore it can be used in combination with other learning tools. The most straightforward cooperative utilization lies in its application to unclassified data after processing within a clustering tool. Since this paper focuses on model tuning, the further system description deals with the relevant system feature tuning. A more detailed general system description can be found in (Kléma, Palouš, 2001).

There are two principal approaches to parameter optimization in machine learning (Kohavi, 1997). The *filter approach* optimizes the model parameters in the framework of a preprocessing step. The optimization is based on dependencies revealed in the data, and it is completely independent from the induction algorithm. Working with iBARET, the filter approach corresponds to an external optimization with consecutive parameter loading into the system. In the *wrapper approach*, parameter optimization is conducted using the induction algorithm as a black box. iBARET is a batch method with performance bias, i.e. it uses feedback from the performance function during training. This function is calculated for each setting of the parameter vector. The system applies a genetic algorithm in order to search effectively through the parameter space. The following section discusses the various modifications of search algorithms implemented in iBARET.

Having aggregated representation, instance-based reasoning represents a natural way to answer both the questions of interest at once. The predictive task can be solved on the basis of similarity between weeks, while the identification of characteristic groups can arise from optimization of the predictive performance. In other words, a patient group can be denoted as characteristic for a procedure when it shows a significantly higher importance factor (its feature weight) than other groups in the framework of the predictive step.

A certain disadvantage of our representation is that it does not contain enough instances to start automated optimization in terms of the wrapper approach. If we restrict our optimization process to the feature (group) weights only, we still have 81 parameters to be optimized on 150 instances (number of weeks per three years). This ratio does not give any room for a reliable search through the parameter space, and it restricts us toward methods utilizing a heuristic parameter setting in terms of the filter approach.

The domain offers a natural way to estimate the group weights in accordance with the average number of prescriptions of the given procedure in the observed group (see that weights w_{jk} equal the group regression coefficients a_{jk} used in Eq. 2):

$$w_{jk} = \frac{1}{n} \sum_{i=1}^{n} \frac{Pr_{ijk}}{Gr_{ik}} \qquad (3)$$

where w_{jk} is the group weight for the j-th procedure and the k-th group,

Pr_{ijk} is the real number of prescriptions of the j-th procedure to the patients belonging to the k-th group in the i-th week,

n the number of instances used for learning the weights (number of weeks).

Table 1. Comparison of mean absolute relative errors reached for uniform feature weights (Uniform) and optimized weights (Weighted). When one method significantly outperforms the other (paired difference t-test, α=0.05), the mean absolute relative error is shown in bold.

Procedure	Uniform [%]	Weighted [%]	Procedure	Uniform [%]	Weighted [%]
Pr1	13.9	**8.0**	Pr21	13.9	**7.1**
Pr3	17.2	15.5	Pr22	31.9	**13.4**
Pr4	19.7	6.2	Pr23	24.4	23.2
Pr5	15.4	**6.9**	Pr26	17.3	8.3
Pr6	19.9	18.8	Pr28	13.6	7.4
Pr7	36.5	33.2	Pr29	17.4	**9.6**
Pr8	13.8	9.1	Pr30	23.9	21.9
Pr9	11.9	10.0	Pr31	25.9	**22.5**
Pr10	15.9	**7.5**	Pr32	29.4	31.1
Pr11	17.1	**5.3**	Pr33	11.8	**3.6**
Pr12	19.9	**4.7**	Pr34	26.5	20.5
Pr13	35.6	38.8	Pr35	13.4	**8.1**
Pr14	15.5	10.7	Pr36	31.5	24.6
Pr15	24.3	**18.7**	Pr37	18.3	20.7
Pr16	18.6	**14.5**	Pr38	12.4	**6.3**
Pr18	22.9	**15.9**	Pr39	**42.9**	50.9
Pr19	22.8	**13.9**	Pr40	22.2	**17.8**
Pr20	13.4	**6.5**	**AVG**	20.9	**15.5**

These weights can be directly applied within iBARET and used to predict the individual numbers of prescriptions. **Table 1** compares the results reached by the weighted and the uniform predictive algorithms. The results suggest that the utilization of group weights brings a significant improvement for 18 out of 35 health procedures, while uniform weights are better for 1 procedure (paired difference t-test, α=0.05). A stronger paired difference t-test (α=0.01) still claims 11 significant improvements when using calculated weights. Consequently, the procedures can be divided into two principal clusters: the first contains procedures that are prescribed evenly over the patient space, while the second consists of the special procedures relevant to patients with specific disorders. For the second cluster it makes sense to

identify the characteristic patient groups. In this sense, the results confirm the expert's prior expectation. An example of a procedure belonging to the second cluster is shown in Figure 5. The diagram shows the improvement in the peat bog bath prediction when considering different importance of the individual patient groups. At the same time, the prediction process reveals the patient groups characteristic for this procedure, e.g., patients aged between 20 and 60 having disorders 5, 9 or 13 are more likely to have this procedure prescribed.
The applicability of this solution is approved by a comparison of these results with the results provided by the individual centered approach. Although the second approach utilizes all the available information (and thus it is much more time and memory demanding) it does not outperform the aggregated method. This confirms the suitability of the suggested simplification.

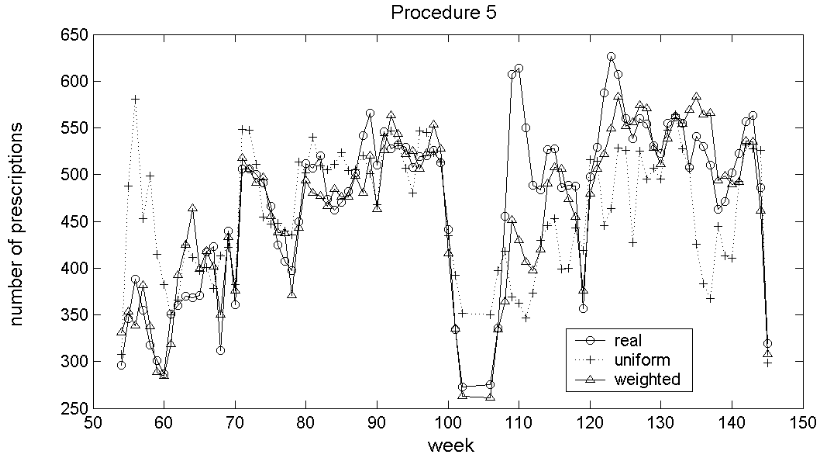


Fig. 5. Procedure 5 – peat bog – a comparison between the real number of prescriptions and the numbers predicted by the uniform and the weighted method (Pr5 line in Table 1).

3.4 Heuristic Approaches

Another simple predictive algorithm is based on a pure copy of the number of prescriptions in the last week. The simple regression is used if and only if *GrAll*$_i$ changes rapidly from one week to the other one. This approach gives the general MAPE 12%, i.e., it clearly outperforms more sophisticated methods. However, the predictions have to be often available several weeks in advance. This demand asks for utilization of the precedent weeks which decreases the prediction accuracy – 15.3% / 1 (when available 1 week in advance), 17.3% / 2, 18.7% / 3. Utility of the previous weeks is obvious, the major part of patients stays for 3 or 4 weeks. It follows that not more that one third of the patients changes each week.

Although this history approach cannot be applied for longer-term predictions directly, it brings forward time equability of most procedures. It can be utilized in correction of the systematic error observed in both regression approaches. For certain part of procedures, the prediction can be improved by subtracting the error of the same predictor taken from the last evaluated week.

3.5 Bayesian Statistical Prediction

Bayesian statistical inference offers another straightforward way to predict weekly number of health procedure applications. The prediction is based on the original task representation, i.e., it follows the individual centered approach.

$$Pr_{ij}^{pred} = \sum_{p=1}^{pat\_in\_week\_i} P\left(Pr_j | a_{1p} \cdots a_{xp}\right) Pat_{ip} \qquad (4)$$

where P($Pr_j|a_{1p}\ldots a_{xp}$) denotes the conditional probability that procedure Pr_j will be prescribed to patient p described by the vector of attributes $a_{1p}\ldots a_{xp}$ on one day of his stay,

pat_in_wek_i is the number of patients presented at the spa during i-th week,

Pat_{ip} is the number of days spent at the spa by patient p in the i-th week.

The conditional probability P($Pr_j|a_{1p}\ldots a_{xp}$) is estimated on basis of historical data using well-known Bayes' theorem. The attributes are assumed to be independent given the health procedure (Naïve Bayes):

$$P\left(Pr_j | a_{1p} \cdots a_{xp}\right) = \frac{P\left(Pr_j\right) \prod_i P\left(a_{ip} | Pr_j\right)}{P\left(Pr_j\right) \prod_i P\left(a_{ip} | Pr_j\right) + P\left(\neg Pr_j\right) \prod_i P\left(a_{ip} | \neg Pr_j\right)} \qquad (5)$$

4 Conclusion

The approach described in this paper results in the general mean absolute prediction error which is approximately 12%. This error is reached by the instance-based reasoning approach or the regression by groups with the heuristic correction. Most of procedures (32) are predicted using the aggregated dataset introduced in Section 2.1. The prediction of the remaining procedures (3) is based on the specific groups derived by the LispMiner.

The SPA company finds the project outcome very valuable. The suggested methodology can be incorporated into their information system for spa L-BIS. They plan to describe the reached results and consequent service improvements in their customer journal. In particular, the availability of reasonably accurate prediction has significant impact on the following activities of the spa complex:

- The full capacity of qualified workers operating the balneo services is utilized as the optimal staff structure for the considered week (or longer period) can be estimated from the predicted needs (some members of staff can be moved to the overloaded procedures, offered vacations or send for training elsewhere, etc.).

- The operating regime of various balneo services is tuned according to the actual needs. It results in the operating cost savings (electricity, water, etc.). The extra capacity can be offered to the general public well in advance.
- Consequently, the overall quality of the provided care is improved – the client always gets those treatments that his condition requires.

In future they plan to use the prediction for scheduling of energy consumption needed for warming-up the bath-tube and pool water. The developed method is easily re-usable for other similar facilities. According to the domain experts, every spa facility has a different structure of patients, even if they offer almost the same procedures. It means that a new segmentation of patients has to be designed for every spa facility. Currently, it is the only step where SumatraTT cannot help. This will be improved when current development of a new statistical template for SumatraTT is finished. On the other hand, the other steps of data pre-processing and analysis remain the same. In this context, SumatraTT proves to be an indispensable tool for the considered DM tasks as it can replace lot of tedious and demanding data processing. There have been developed appropriate data processing templates to do the job. Each template takes groups' description stored in a table and generates and executes SQL commands that calculate aggregated values. Finally, the data is exported into a text file. There is being prepared a script ensuring export of the data into the WEKA format. This opens possibility to apply any algorithm provided by the rich WEKA ML package including the regression, too.

References

1. Aubrecht, P., Kouba, Z.: Meta-Data Driven Data Transformation. Proc. of the 5th World Muti-conference on Systemics, Cybernetics and Informatics, 2001.
2. Blockeel, H., Moyle, S.: Collaborative data mining needs centralized model evaluation, submitted to DMLL-2002 at ICML-2002
3. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R.: CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM consortium, 2000.
4. Fayyad, U., Piatetski-Shapiro, G., Smith, P: From Data Mining to Knowledge Discovery: An overview. In Fayyad, U., Piatetski-Shapiro, G., Smith, P., Uthurusamy R. (eds.): Advances in Knowledge Discovery and Data Mining, 1-34. MIT Press, Cambridge, MA, 1996.
5. Fayyad, U., Piatetski-Shapiro, G., Smith, P., Uthurusamy R. (eds.): Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA, 1996.
6. Kléma, J., Palouš, J.: iBARET - Instance-Based Reasoning Tool. In: Final Programme & Proceedings. Aachen : Verlag Mainz, vol. 1, p. 55. 2001.
7. Kohavi, R., John, G. H.: Wrappers for Feature Subset Selection. Artificial Intelligence, 15 97, 273-324, 1997.
8. Lavrac, N., Grobelnik, M.: Data Mining. To appear in: Data Mining and Decision Support: Integration and Collaboration, Kluwer Publishers, 2003.
9. Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann, California, 1999.

10. Rauch, J., Simunek, M.: Mining for 4ft Association Rules. In Discovery Science 2000. Red. Arikawa, S. – Morishita S. Springer Verlag 2000, pp. 268 – 272.
11. Stepankova, O., Klema, J., Miksovsky, P.: Collaborative Data Mining and Data Exchange: A Case Study. In: IDDM Workshop Proceedings, ECML/PKDD 2002. Helsinki University of Technology, 2002, vol. 1, p. 135-140.
12. Stepankova, O., Klema, J., Lauryn, S., Miksovsky, P. and Novakova, L.: Data Mining for Resource Allocation: A Case Study. In proc. of the 5th IEEE/IFIP Int. Conf. on Information Technology for BALANCED AUTOMATION SYSTEMS (BASYS 2002), Cancún, Mexico, Kluwer Academic Publishers, ISBN 1-4020-7211-2, pp.477-484, 2002.
13. Witten, I., Frank, E.: Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.