DATA MINING AND RESOURCE ALLOCATION: A CASE STUDY

Olga Štěpánková[1], Jiří Kléma[1],
Štěpán Lauryn[2], Petr Mikšovský[1], Lenka Nováková[1]

[1]*Department of Cybernetics, CTU Prague,*
Technicka 2, 166 27 Prague 6, Czech Republic
e-mail: {step,klema,miksovsp}@labe.felk.cvut.cz

[2]*Lauryn,v.s.o., Pražská 255, 530 06 Pardubice 6,*
e-mail: {stepanl@lauryn.cz}

This paper presents a case study concerning scheduling and resource allocation issues in 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

In industry, everybody understands that scheduling influences efficiency of any serious industrial activity significantly. Due to good scheduling the customers can get their ordered goods in time and for reasonable price which is not increased due to extra fee for storage of the intermediate product. But industry is not the only domain where scheduling has clear economic impact. Scheduling is important in those types of complex services, which have to be ensured e.g. through cooperation of several persons using restricted number of resources. Such a situation appears often in medical environment. Nurses' rostering belongs to well-known cases frequently referred to when describing applications of various techniques for scheduling. But similar problems appear e.g. in a typical health farm or a spa.

Spa offers a set of various health procedures to heal medical problems of the patients who are arriving into the health farm for a restricted period. Obviously each patient obtains an individual treatment, i.e. a set of procedures assigned to the patient by the spa physician, who makes his recommendation after careful inspection of the patient upon his arrival. But recommendation of the spa physician is not enough to ensure that the patient gets those procedures he is supposed to get. To reach such a goal it is important to ensure that necessary resources are available in appropriate quantity. Two basic types of resources – human resources (appropriate skilled personal) and technical equipment (e.g. a bath tube or diathermia) - have to be combined while there has to be met a number of very diverse local constraints, e.g.

- Each member of the personal has several skills and can operate several types of equipment. On one hand, a single person cannot exercise all the day one type of a physically demanding job. On the other hand, he/she cannot switch among his/her skills every now and then as the adjustment can be time consuming, etc.
- Sometimes several pieces of equipment are situated in a single room but they cannot be used simultaneously.
- There is a minimal amount of patients for which certain type of equipment can be opened (sauna for 10 persons).

All over the fact that the groups of patients occupying the spa are changing frequently, the spa aims to provide the appropriate individual treatment for each of its patients. How can such a goal be achieved? It is vital for the spa administration to know in advance (before the group of patients arrives) what will be the total requirements for all procedures offered by the spa. This knowledge can point to the fact that special precautions should be taken for the considered limited period of time, e.g. extra day off can be offered to some professions

or on the contrary extra personal has to be hired or a long day introduced for certain wards ensuring specific procedures. Such decisions have to be planned several days or even weeks in advance. That is why timely prediction of resources requirements is vital for the spa administration. Can such a goal be achieved?

Administration gets the basic information (including rough anamnesis) about the patients to come several weeks in advance. Moreover, the administration owns all the data about the treatment of patients from the last years. Can the history data be used as a source for data-mining leading to discovery of rules or algorithms useful for prediction of resources requirements? In the rest of the paper we will describe a data-mining case-study providing a positive answer to the considered question. This case study is based on real life data.

2. CASE STUDY DESCRIPTION

2.1 Data Mining Goals, Available Data and Their Granularity

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

1. 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.
2. Treatment schedules prescribed by different spa physicians are consistent.
3. The full set of procedures offered by the health farm is fixed, no procedures are added or removed.

If all the premises are true, the history data from the last period (1 or 2 years) could be used to search for prediction rules. 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.

The method to solve the prediction task has to be chosen with respect to the complexity of the treated problem. First, what is known about a single patient before he/she arrives into the health farm? Each person is described using 7 discrete attributes (Sex, Cure_type, Disorder, Motility, Stay_length, Accommodation and Age). These attributes differ in their domain sizes and frequency of the individual attribute values (e.g., there are 8 possible values of Cure_type, the most often value covers 59% of patients, Stay_length can be from 3 to 35 days, but the patients mostly stay for 21 (60%) or 28 days (28%)).

Suppose each value of each attribute has the same weight from the point of view of the considered prediction task. How many different types of patients we would have to take into account? Let us forget about the attribute with the most extensive domain, the age. Even excluding the age we would have to distinguish 2 x 8 x 12 x 8 x 33 x 5 = 253 440 different types of patients. But our training data cover less then 1/10 of this amount only. It is clear we have to suggest appropriate simplification of the task.

The first step towards simplification is the change of granularity in the used domains – design of the restricted domains. Domains of some attributes are rather extensive (e.g. Stay_length), but what really counts is the frequency with which individual values appear in the considered training data. Obviously, the most frequent values have to be represented even in the restricted domain (the original domain has finer granularity than the restricted one). Now it seems we can afford to consider age in decades. Under these conditions we have to distinguish 2 x 4 x 8 x 2 x 5 x 8 = 5120 different types of patients. Even under this simplification, given data of less than 18 000 patients only we are not ready to learn to answer a question "Will the considered patient be prescribed procedure No. A?" The attempt to use ID3 for this purpose failed both in the original and in the simplified case. But we do not need to answer this very specific question. The spa administration does not want to replace their physicians by a SW system. What they need is an estimate of resources, which will be necessary in the coming weeks. That is why the final goal for the data-mining was rephrased as follows:

- Use the original attributes (with restricted domains) to identify groups of patients which appear frequently in the training set and which exhibit characteristic behavior or requirements of spa utilities.
- For each such group predict a set of procedures to be passed by a typical patient during a typical week. Consequently calculate sums of procedures for the specific week according to the actual number of patients in the spa and their distribution among the groups.

2.2 Data Aggregation

Each data entry in the original dataset describes one specific allocation of a single procedure for a specific patient. These entries have to be aggregated to make explicitly available necessary information about full week of a stay for each patient. This task was approached as a time-series problem leading to significant amount of

preprocessing. SumatraTT (Aubrecht, 2001a, 2001b) proved to be very useful for all the applied DM tools by ensuring the following tasks:

- Data aggregation – counting the total of procedures wrt. patient, week, etc.
- Data transformation – new dataset was generated in such a way that n-record of the original table were transformed into n-columns of one record in the new set (matrix transposition).
- Combined transformation.
- Export of the dataset into specific formats required by various DM tools applied.

2.3 New Table CTU_GWEEKS

Data exploration mentioned above used SQL and it identified some mistakes or misprints. Consequently appropriate cleaning (standardization of considered cases) was designed and the size of domains of some these attributes was restricted (e.g. the minor and rare cases are neglected). The new attributes (modified by a change in the domain of values) can be identified in the sequel easily by the prefix CTU. It does not seem difficult to define appropriate restriction in some cases, namely for the following attributes:

CTU_GDISORDER. Total number of 6 considered values covers 5 most frequent disorders, namely 2 (1465 patients), 3, 5, 7, 13 (4397 patients) and the rest is labeled as 0. Under this modification the number of patients with each considered disorder is higher than 1000.

CTU_GAGE. Total number of considered values is 6: A – age less than 45 (1461 patients), B – age within the interval <45, 55) covering 3226 patients, C - <55,65) covering 4439 patients, D - <65,75) with 4959 patients and E - 75 and more with 1895 patients.

CTU_GCOMPANY. Total number of considered values is 2, namely the original company value 29 (corresponding to more than 70% of all patients) and the others (labeled as 0).

Suppose the solution of the considered task is based only on the attributes Sex (2), CTU_Gdisorder (6), CTU_Gage (6) and CTU_Gcompany (2) - the size of the corresponding restricted domain is given in brackets. This restriction leads to introduction of 144 = 2 x 6 x 6 x 2 different groups corresponding to the Cartesian product of the relevant domains. Is it possible to neglect information e.g. about the Cure_type? What is the reasonable amount of groups to consider?

The time unit we are going to work with is a week. We know that there is about 500 of patients staying in the spa each week. In an extreme case, it can happen that all the considered groups of patients appear among the 500 spa visitors in a single week. Moreover, the domain expert claims that 1 or 2 patients more or less does not make a difference in spa resources requirements. Thus a group has to contain 5-10 patients at least to become significant for the considered prediction task. Consequently, it makes little sense to introduce more than 100 different patients' groups. These groups have to be derived from 144 upper mentioned groups, which are further split due to the value of the attribute Cure_type. Is this feasible at all? This seemingly intractable problem can be solved using similar type of data analysis as that used when restricting the domains of considered attributes. For the present task, there will be necessary to analyze the training data wrt. the size of groups of patients defined by combinations of values of considered attributes. This simple approach points to the fact that some combinations are rare or absent in the training data. This is most decisive for the combination Disorder x Cure_type and that is why this combination is replaced by a new combined attribute Disorder_CureType having 8 possible values only. Occurrence analysis in the training data entitles us to define new types of groups as the Cartesian product combining the attributes SEX (2 values), AGE_DIS (5 values), Disorder_CureType (8 possible values). Consequently, we have 2 x 5 x 8 = 80 disjunctive groups plus 1 additional one covering the rest of the patients.

A new table CTU_GWEEKS generated with a heavy support of SumatraTT consists of 81 attributes (Gr1, …, Gr81) corresponding to the upper mentioned groups and 35 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 *n*:

- 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.

The final table CTU_GWEEKS contains 147 records corresponding to all the weeks in the period 1999-2001 (126 records in the training set, 22 records in the test set).

3. PREDICTIVE MODELING

3.1 General Overview, Score Function

As defined above, for the effective resource allocation it is critical to know (predict) how many individual health procedures are going to be prescribed for the following time period. This chapter focuses on construction of

predictive models estimating the total number of prescriptions of different types of procedures in a particular week according to the actual number of patients in the spa and their distribution among the patient groups.

All the presented models deal with the datasets aggregated in the chapter 2. Of course, alternative models dealing with other task representations could be created. Let us mention a model dealing with a representation handling single patients rather than their groups. The model is based on a decision tree answering the question „will this specific procedure be prescribed for the given patient or not?“ The final weekly prediction was made as a sum of all the partial patient predictions. This approach did not bring satisfactory results. However, there was an observation which was worth of mentioning: „The first attribute used for splitting the root in all constructed trees was ‚type of disorder‘, ever.“ This observation was considered when defining cumulative attributes specifying significant patient groups. The other model handling single patients could be a probabilistic model (e.g., Naïve Bayes) answering question „what is the probability that a specific procedure will be prescribed for the given patient on a single day of his/her stay?“

All the presented models use the same scoring function frequently applied in regression tasks. The model fitting is evaluated in terms of mean absolute percentage error (MAPE) and its standard deviation (STDEV). This error measure is defined as follows:

$$S_{MAPE}(M_j) = \frac{100}{n} \sum_{i=1}^{n} \frac{\left| Pr_{ij}^{pred} - Pr_{ij}^{real} \right|}{Pr_{ij}^{real}} \quad [\%]$$

where M_j is the predictive model designed for the j-th procedure,
n is the number of predicted weeks,
Pr_{ij}^{real} is the real number of prescriptions of the j-th procedure in the i-th week,
Pr_{ij}^{pred} is the predicted number of prescriptions of the j-th procedure in the i-th week.

3.2 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 simplified in such a manner that it assigns all the patients to the same group. It means that it does not take advantage of division into 81 groups, does not consider any patient specific information and utilizes the overall number of patient-days in the predicted week only:

$$Pr_{ij}^{pred} = a_j GrAll_i$$

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%.

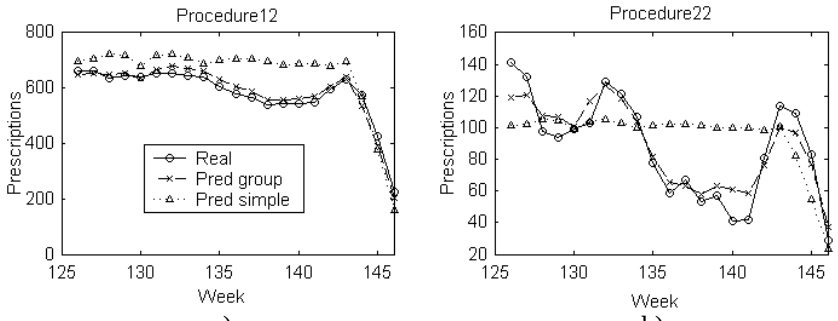
3.3 Regression by Patient Groups

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 (Novakova, 2002):

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

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 dented 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 1), on the other hand it does not show any significant difference in the other 29 procedures (see Figure 2).



a) b)

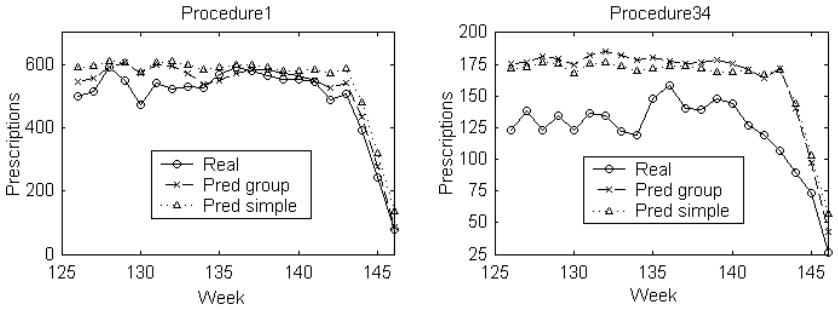
Figure 1 – Examples of health procedures for which the regression by groups (Pred group) significantly outperforms the simple regression (Pred simple).

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 2a) or it can be a good solution having reduced a systematic prediction error, i.e., regarding long-term changes of a_j (see Figure 2b). This issue can be solved by a heuristic approach presented in 3.5.

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.



a) b)

Figure 2 – Examples of health procedures for which both the regression approaches show similar performance, either working well (a) or having high MAPE (b).

3.4 Instance-Based Reasoning

Promising results have been obtained by application of the iBARET (Klema, Palous, 2001) system as well. This system utilizes algorithms belonging to the group of instance-based learning methods, it can be considered as derivative of kNN method. The prediction is based on searching through the case memory of the known week instances given by distribution of patients among groups (records of CTU_GWEEKS) and retrieving the most similar records to the current predicted record in the weighted and normalized Euclidean space.

The algorithm is provided with a ratio-based data. The absolute numbers of patient days are replaced by ratios, i.e., each Gr_{ik} is replaced by $Gr_{ik} / GrAll_i$. The motivation for transition between absolute and ratio-based data is straightforward – we do not want to retrieve weeks with the exactly same size of the individual groups, but with the same patient distribution among the groups. The final prediction is generated as follows:

$$Pr_{ij}^{pred} = \frac{1}{N} \sum_{n=1}^{N} Pr_{nj}^{real} \frac{GrAll_n}{GrAll_i}$$

where N is number of selected nearest neighbors,
Pr_{nj}^{real} is the real number of prescriptions of the j-th procedure in the week selected as the n-th nearest neighbor,
$GrAll_n$ is the number of days spent by all the patients during the week selected as the n-th nearest neighbor.

3.5 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.

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 simple regression or the regression by groups with the heuristic correction. Most of procedures (32) are predicted using the group definition given in 2.3. The prediction of the remaining procedures (3) is based on the specific groups derived by the LispMiner.

The spa administration requested 20% precision only. This request is satisfied for 31 of 35 predicted health procedures. It is hoped that the spa management can benefit from the prediction (information about the amount of necessary procedures available few weeks in advance). How can it be applied and what its main contributions? Prediction of reasonable accuracy can have significant impact on the activity of the spa complex in the following aspects:

- It will be possible to plan full use of capacity of workers operating the balneo services. Optimal staff structure for the considered week (or longer) period can be designed (some can be moved to the overloaded procedures, new people can be temporarily hired, planning of vacations, …).
- The operating regime of various balneo services will be tuned according to the actual needs (restriction of operating costs for electricity, water, …). Procedures, which are not necessary for patients staying in the spa, can be offered to the general public.
- Consequently the quality of the provided care will be improved – the client will ever get the procedures his health condition requires.

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 grouping of patients has to be designed for every spa facility. 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.

As it follows from the previous paragraphs, the developed method is easily re-usable for other similar facilities. The main difference can appear in grouping of patients. The grouping is carried out using quite simple statistical calculation. 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.

Acknowledgments

This research was supported by the EU project Sol-Eu-Net IST-1999-11495 *Data Mining and Decision Support for business competitiveness: A European virtual enterprise.*

5. REFERENCES

1. Aubrecht, P. (2001a). Specification of SumatraTT. Technical Report K333-2/01, CTU, Dept.of Cybernetics, Technická 2, 166 27 Prague 6, www: http://krizik.felk.cvut.cz:8080/SumatraReg/.
2. Aubrecht, P. and Kouba, Z. (2001b). Metadata Driven Data Transformation. In SCI 2001, volume I, pages 332-336. International Institute of Informatics and Systemics and IEEE Computer Society.
3. Klema, J., Palous, J. (2001): iBARET - Instance-Based Reasoning Tool**,** In ELITE Foundation, editor(s), European Symposium on Intelligent Technologies, Hybrid Systems and Their Implementation on Smart Adaptive Systems, 1, pages 55, 2001
4. Klema, J., Lhotska, L., Stepankova, O., Palous J. (2000): Instance-Based Modelling in Medical Systems, In Trappl R., editor(s), Cybernetics and Systems 2000, 2, pages 365-370, Vienna, Austria, April 2000. Austrian Society for Cybernetics Studies - ISBN 3-85206-151-2
5. Novakova, L.(2002): Prakticke aplikace metod strojového uceni. In Czech, Diploma thesis K333, FEE CTU, Prague, January 2002.
6. Rauch, J., Simunek, M. (2000): Mining for 4ft Association Rules. In Discovery Science 2000. Red. Arikawa, S. – Morishita S. Springer Verlag 2000, pp. 268 – 272.