Data Mining Support for Scheduling and Resource Allocation:

A Case Study

Olga Štěpánková[1], Štěpán Lauryn[2],
Petr Aubrecht[1], Jiří Klema[1], Petr Mikšovský[1], Lenka Nováková[1], Jiří Palouš[1]

1: Gerstnerova Laboratoř, Katedra kybernetiky FEL ČVUT, Technická 2, 166 27 Praha 6
{step, aubrech, klema, miksovsp, palous@labe.felk.cvut.cz}

2: Lauryn,v.s.o., Pražská 255, 530 06 Pardubice 6 {stepanl@lauryn.cz}

1 Introduction

In industry, everybody understands that efficiency of any serious industrial activity is significantly influenced by scheduling. 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 SKLAD. But scheduling is as 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 is a well known case which is 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 gets an individual treatment, i.e. a set of procedures assigned to the patient by the spa physician, who makes his choice 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 keeping in mind 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 person cannot exercise all the day a single very demanding job. On the other hand he/she cannot switch among his/her skills every now and than 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 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 extra help will be needed for the limited period of time or that a long day will have to be introduced for certain 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

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). The structure of the records is reviewed at Table 1. 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, their domain and its size is given in Table 1.

Attribute name	Domain	Size of the domain	Most frequent values and their cover (in descending order)	Size of the restricted domain
Sex	M, F	2		2
Cure_type	1,..,8	8	5 (59%)	4
Disorder	2,..,13	12	13 (24,45 %), 5 (22 %), 3 (16,6 %)	8
Motility	2,..,9	8	4 (98,5%)	1
Stay_lenth	3,..,35	33	21 days (60,22%), 28 days (27,6%)	2
Accomodation	1,..,5	5	Each value is covered by 24,6 % to 11 % patients	5
Age	18,..,101	84	50 – 80 years (57%)	8

Suppose each value of each attribute has the same weight from the point of view of the considered prediction task. How many different type 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. That is why Table 1 is complemented by the relevant information which is specific for the training data.
The most frequent values have to be represented even in the restricted domain (the original domain has finer granularity than the restricted one). The size of restricted domains has to be at least that given in the last column of the Table 1. Now it seems we can afford to consider age in decades. Under these conditions we have to distinguish 2 x 4 x 8 x + 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 in both in the original and in the simplified case failed. That is why the goal for the data-mining was specified more precisely 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.

2.1 Data pre-processing and search for restricted attribute domains

Data exploration mentioned above used SQL and it identified some mistakes or misprints. Consequently appropriate cleaning (standardization of considered cases) was designed. To simplify the prediction task, it is necessary to restrict the size of domains of some these attributes. The restriction has to reflect the structure of the treated data – the minor and rare cases can be neglected. The attributes (modified by a change in the domain of values) can be identified easily by the prefix CTU:

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).

2.2 New Table CTU_GWEEKS

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. Instead of characterizing a patient with respect to all the attributes originally used, the set of considered attributes has been restricted to the following ones (the size of the corresponding domain is given in brackets): Sex (2), CTU_Gdisorder (6), CTU_Gage (6) and CTU_Gcompany (2) described in 2.1.
Each patient can be uniquely assigned to one of 144 = 2x6x6x2 different groups corresponding to the Cartesian product of the relevant domains. Following the expert's advise we have decided to focus only to the "regular patients", i.e. taken from the table *patient_regul*. Among these patients there appear 128 groups only.

A new table CTU_GWEEKS consists of 128 attributes (Gr1..Gr128) corresponding to the upper mentioned groups and 38 attributes representing the procedures (Proc1..Proc40). One record sumarizes data concerning all patients present in the spa during a single week. Let as specify the contents of the table for the week *n*:

- Gr1[*n*] is the number of days spent by patients of belonging to the group Gr1 during the week *n*. The same applies to Gr2, etc.
- Proc1[*n*] is the total number of all prescriptions of procedure Proc1 during the week *n*. The same applies to Proc3, etc.

The final table CTU_GWEEKS contains 125 records corresponding to all the weeks in the period 1999-2000.

3 Methods for Procedure Prediction

The aim is to predict the total number of prescriptions of different procedures per week (denoted bellow **TNP_W**) from some characteristics of the considered group of patients present in the spa that specific week. How difficult is this task? This can be estimated from the variability of the number of procedures prescribed to a patient per day – see Tab 2a, 2b.

In the worst case, using the average value per patient and a day we can make a mistake about 3%. How does it work when applied to the group as a whole? This is studied in 3.1. The later sections work with more sophisticated methods.

3.1 Prediction and average number of procedures' prescriptions per patient and day

3.1.1 Naive statistical prediction NSP

Suppose, the only characteristics of the considered group of patients is the size of the group, namely number of patient-days the considered group is going to occupy the spa. The statistical evaluation of the available data denoted as **NSP** applies the steps a) - c) of the following algorithm:

a) Choose the evaluated period providing the statistical data.
b) For each procedure and week *i* in the training data calculate the average number of applications of the considered procedure prescribed to any patient per day that week **TNP_P_W** (*i*)= TNP_W(*i*)/ Patient-days(*i*).
c) Find the mean value and deviation of **TNP_P_W** (*i*) in the evaluated period of weeks. This mean value serves as a rough estimate of the number of applications of the procedure *i* per patient-day. This estimate can be used to predict the total number of prescriptions of procedure *i* given number of considered patient-days. Linear regression could be applied for the same purpose, too.

The **NSP solution** has been verified [Nováková] on the test data and the resulting MAPE error, defined bellow, is about 30 % (see Tab. 3). This is not an acceptable solution. Surprisingly, the worst prediction accuracy (resulting error about 40 %) is obtained for the procedures 7 and 31, both of which have not the highest deviation

in Tab. 2. Slightly better predicted are the procedures 16, 20, 22, 30, 32, 36 and 37 with the error rate 30 % approximately. Obviously, to reach better results finer structure of the group of patients has to be taken into

3.1.2 Regression

There are several ways how to apply regression for available data:

- First, we will generate a simple regression model (**Simple REG**) that deals purely with an overall number of patients present in spa the considered week ($Proc_j\_num_i\_pred = a_1 Gr\_all_i + a_0$, where Gr_all_i is an overall patient number in the i-th week).
- Later on, this model (denoted as **Simple REG – corr**.) was improved by subtracting the error of the same predictor known from the last week.
- Another simple approach (denoted as **Last week**) is based on pure utilization of the result reached by predicting the same amount of the procedures as prescribed in the last week. The first linear approach is used if and only if overall number of patients changes rapidly from one week to the other one.

The results reached on unseen testing data (21 weeks) are presented in Tab 4. These results are evaluated in terms of mean absolute error (MAE), mean absolute percentage error (MAPE) and their standard deviations (STDEV). These error measures are defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} \left| x_i^{prel} - x_i^{real} \right|}{n}, \qquad MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{\left| x_i^{prel} - x_i^{real} \right|}{x_i^{real}},$$

3.2 Decision tree

We have tried to generate a **decision tree** form the considered data [Nováková] in order to answer the question „will this specific procedure be prescribed for the given patient or not?“ This did not bring much better results than the pure statistical solution (prediction accuracy about 70 %). All over the fact that we have not obtained any satisfiable results there is an observation which is worth of mentioning: „The first attribute used for splitting the root in all constructed trees was ‚type of disorder', ever."

A simple table (see Table 5) provides a preliminary analysis of the relation between the type of the patient's disorder and the frequency of prescriptions of a specific procedure. It gives a positive indication of their mutual influence. This influence should be reflected in characterization of different groups of patients. The first step towards this goal has been the introduction of the cumulative attributes in 2.4. The table CTU_GWEEKS serves as an input for an IBR system iBARET.

3.3 iBARET Results (Prediction based on CTU_GWEEKS table)

The iBARET algorithm [Kléma] belongs to the group of instance-based learning (IBL) algorithms, it can be considered as a derivative of kNN method. This chapter presents experiments carried on with its two basic input settings based on CTU_GWEEKS table. First, it was proved that number of features must be reduced. Obviously, learning 128 feature weights on 125 examples gives big space for over-training. That is why, we have tried to identify patient groups that are principal for the given procedure and thus reduce the number of features. It was decided to use 10 features only, the groups with the largest feature weights were used (Gr61, Gr63, Gr128, Gr64, Gr66, Gr44, Gr76, Gr11, Gr38, Gr_All). The most relevant features (patient groups) were identified in an iterative manner – 128 feature weights were learnt first and then there were selected just 10 features having the largest weights for re-training. This model is denoted as iBARET – reduced. The second iBARET model deals with 128 features but applies sort of windowing. It means that the case memory does not contain the training set only but it always contains all the records that precede the currently predicted testing record.

Model	MAE	STDEV	MAPE [%]	STDEV [%]
iBARET - reduced	52	66	16	42
iBARET – win.	81	78	19	18
Simple REG – corr.	49	44	13	25
Last week	37	38	9	18

Tab. 4. Procedure 6 – evaluation on testing data (21 weeks)

Generally, it can be concluded that we have constructed predictive models that are likely to predict with MAPE slightly higher than 10%. Relative success of the trivial models with respect to the more sophisticated models can be explained as follows:

- the model that uses the last week data probably utilizes the fact that majority of patients stay from one week to the other one and that is why the similar amount of procedures of the given type is expected,
- weak relationship between the patient structure and number of prescribed procedures (or at least our inability to utilize it) suggests influence of other factors on procedure application (spa capacity, constraints on different procedures, etc.).

3.4 Refinement of the Statistical Approach

All over the fact that the results obtained by iBARET are better than those obtained in 3.1 and 3.2, we are not fully satisfied with them. One of the reasons for this failure is the size of the training data: the relation between number of available cases and complexity of their description (number of used attributes) is too low. This is due to the fact that the structure of groups designed in 2.1 and 2.5 is too delicate. Which groups are really necessary to solve the problem?

3.4.1 Impact of various attributes on prescription of a procedure

Let us analyze the relation between prescriptions of the procedure and various available attributes. The relation between **prescriptions and the type of the patient's disorder** is shown on Tab. 5. Here, it is clear that the patient's type of disorder has a significant influence on procedures prescribed to the patient. This is obvious in the case of some procedures, namely:

Procedure	D2	D3	D5	D6	D7	D9	D13	others
1	12,3%	24,6%	16,8%	4,4%	3,4%	3,9%	27,4%	7,1%
3	11,3%	15,3%	21,8%	2,1%	0,1%	13,0%	29,8%	6,6%
4	1,2%	4,7%	2,1%	14,0%	**59,4%**	0,6%	11,9%	6,1%
...								
39	11,6%	18,5%	4,0%	7,8%	**44,9%**	2,2%	9,0%	2,1%

Tab. 5. Disorder - Procedure relation – few examples

- „Patients with disorder D2, D3, D5 or D9 are being prescribed the procedure **4** very rarely (less than 5 %)."
- „Patients with disorder D7 are getting 45 % of all prescriptions of the procedure **39**."
- „Patients with disorder D5 will probably not be prescribed the procedure **39**."

Do the other original attributes influence the amount of procedures prescribed? Each of the considered attributes is discrete and its single value identifies a specific group of patients. Let us consider a fixed procedure *n* and a fixed attribute At. To find out the influence of any specific value *a* of At on the frequency with which the procedure *n* is prescribed we introduce a coefficient of **subgroup significance** [Nováková] denoted as ***sgs(a):***

> Let Any_ A_P(*n*) be the number the procedure *n* is prescribed to a patient per day on average and A_P(*n*,*a*) be the corresponding number for the subgroup of patients with the value of the attribute At equal to *a*. More precisely, A_P(*n*,*a*) is the number the procedure *n* is prescribed per day to a patient belonging to the considered subgroup (with the At value equal to *a*). Coefficient ***sgs(a)*** equals to the relation A_P(*n*,*a*)/ Any_A_P(*n*).

Obviously, if the ***sgs*** coefficient of a certain group is close to 1, this group does not exhibit special influence on the number of procedures prescribed (this group does not diverge from the average). On the other hand if ***sgs*** of a subgroup differs significantly from 1 (it is e.g. 2 or 0) the considered subgroup has important impact on the amount of procedures prescribed.

The coefficient of subgroup significance is calculated in Table 6 a, b for

- Age group of patient
- Sex of the patient

Procedure	Age A	Age B	Age C	Age D	Age E
1	0,49	0,59	0,84	1,29	1,71
3	1,34	1,27	1,16	0,78	0,47
4	1,16	0,95	0,90	1,03	1,10
5	1,61	1,66	1,27	0,51	0,08
...					
39	0,58	0,55	0,76	1,21	2,09
40	0,07	0,25	0,67	1,45	2,59

Procedure	Men	Women	Difference
1	1,06	0,97	0,09
3	1,54	0,75	0,78
4	1,01	0,99	0,02
5	1,15	0,93	0,22
...			
39	0,63	1,17	-0,54
40	1,00	1,00	-0,01

Tab. 6 a,b. Coefficient **sgs** for the considered groups of patients

The relations in the Tables 6 a, b support the hypothesis that both the value of the attribute sex and age group exhibit significant influence on the amount of procedures prescribed. This is obvious e.g. in the case of procedures printed in bold in the table 6.

3.4.2 New domains for the used attributes

Grouping of patents applied in the table CTU_GWEEKS did not consider the original attribute Cure_type, which can have 9 different values. Instead it applied the attribute Company. Surprisingly, in the year 2000 there appears a very limited amount of patients from groups generated as combination of the chosen disorder and Cure_type (see Tab. 7). The most frequent combinations will become a basis for introducing the new characteristics of the group of patients.

Procedure	D2C5	D3C5	D5C5	D5C6	D6C5	D7C5	D9C5	D13C5	Others
1	13%	25%	4%	8%	4%	1%	4%	28%	12%
3	12%	15%	7%	11%	2%	0%	12%	29%	12%
4	1%	5%	1%	1%	14%	57%	1%	12%	7%
5	17%	11%	7%	10%	0%	0%	12%	29%	12%
39	11%	15%	1%	2%	8%	44%	2%	9%	7%
40	12%	28%	3%	9%	1%	4%	3%	28%	13%

Tab. 8. Dependency between disorder X cure_type combinations and the amount of procedures prescribed

We will consider groups of patients corresponding to the combinations D2C5, D3C5, D5C5C6, D6C5, D7C5, D9C5, D13C5 and Others (any other combination of Disorder and CureType). Now, we will try to find out the influence of this combined attribute on the amount of procedures prescribed (see Tab. 8.). Even here, there is no doubt that there is a relation between both arguments (compare e.g. the procedures 21, 22, 23).

		Disorder												
		-	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
	-	682	190	456	28	551	170	550	6	118	12	2	7	728
Cure type	**1**	0	2	36	15	349	13	2	0	2	1	0	1	42
	2	0	0	5	0	28	0	0	0	0	0	0	0	1
	5	1	**1234**	**2146**	93	**639**	**560**	**1642**	22	**622**	117	7	32	**3277**
	6	0	10	179	46	**950**	23	40	0	1	2	1	8	159
	7	0	1	17	2	138	1	4	0	1	0	0	1	25
	8	0	0	0	0	1	0	0	0	0	0	0	0	0

Tab. 7. Number of patients corresponding to combinations of Disorder and Cure_type attributes

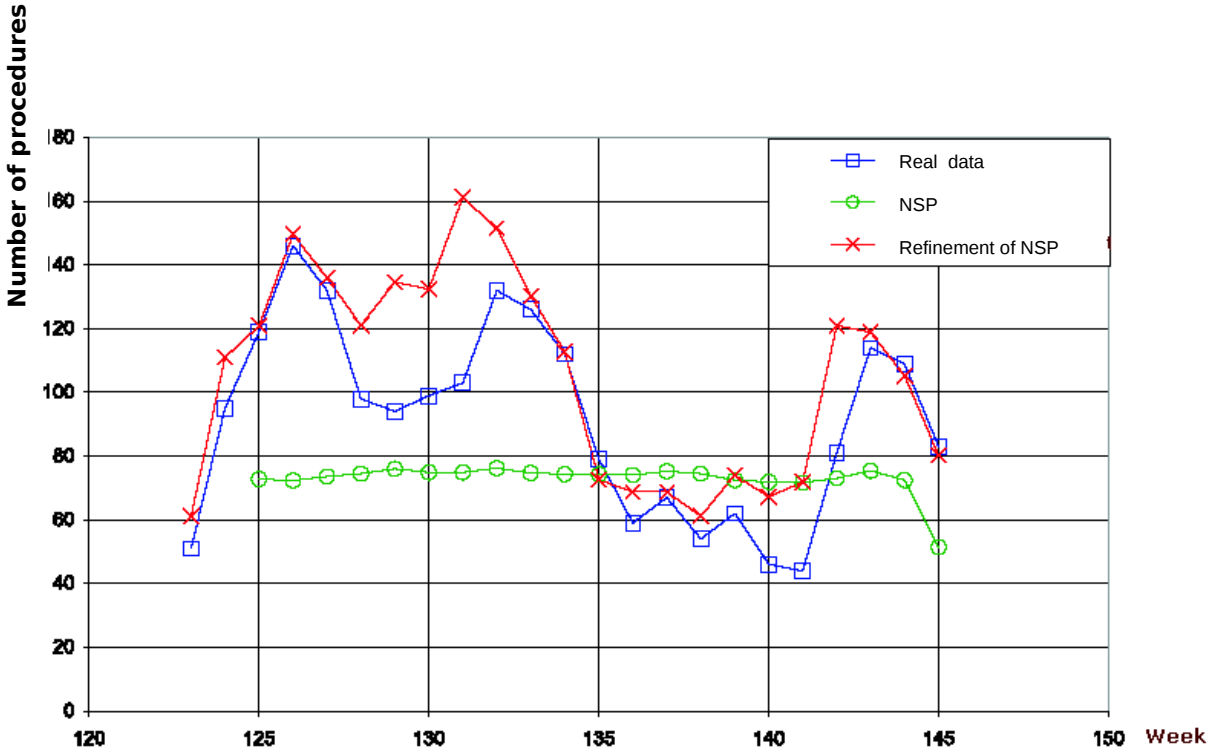
3.4.3 Introduction of the new patient groups

The relations verified by upper mentioned experiments give a means to define new types of groups in a way similar to that described in 2.5. This time the Cartesian product will be the result of combining the attributes SEX (2 values), AGE_DIS (5 values), Disorder_CureType (8 possible values). This leads to the introduction of 2*5*8 = 80 disjunctive groups.

3.4.4 Prediction and the new patient groups

For each of these groups independently we will use the regression as a means for prediction of the number of prescribed procedures. The total in the predicted week is then the sum of predictions obtained for the considered groups.

The results for the procedure No. 22 are depicted on Fig. 3 [Nováková] together with the results obtained by the simple method NSP (see 3.1) and the real values. The MAPE of the introduced refinement of statistical method



is 18 %,

Fig. 3. Comparison of methods used (procedure 22)

while it is 32% for NSP.

4 Lessons learnt

The spa task treats time-series - consequently significant preprocessing was necessary. Here SumatraTT proved to be very useful for most DM tasks, namely:

- Data aggregations – counting the total of procedures wrt. patient, week, etc.
- Data transformation – new data set 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 various specific formats required by the used DM tools.

After each transformation or aggregation step it is necessary to verify that the resulting data are generated correctly. The best way is to find appropriate cross-checking sums which have to fit both in the old and in the new dataset.

4.1 Spa DM process: review of important steps taken

Let us review the most important steps which have been taken in the spa DM process and which seem to lead to applicable results:

1. Data exploration used SQL and suggested some cleaning (standardization of considered cases) - see 2.1, 2.2.
2. Preliminary statement of the DM goal helped to design new attributes, namely the week number (see 2.3).

3. The training dataset restriction as a result of analysis of the total number of patients - see Fig. 1.
4. Identification of the attributes, which do influence the number of procedures prescribed to a patient using the ***sgs*** coefficient – see 3.4.
5. The attributes identified in the former step have to be carefully discretized to shrink their domain. The new domains serve for the definition of the significant subgroups as the Cartesian product of the corresponding domains– see 3.4.2, 3.4.3.
6. Finding parameters necessary for prediction on unseen data using regression or iBaret for each of the procedures individually.
7. Evaluation on unseen data and calculation of MAPE for all procedures.

Most of these steps require good deal of tedious work. It was mentioned earlier that the data preprocessing phases (1. - 4.) work extensively with SQL and that they are supported by SumatraTT. What about the step 6? To find the parameters for prediction it is necessary to summarize available data for each procedure independently w.r.t. the subgroups defined in the step 5 in a way analogous to the creation of the new table in 2.5. To apply iBaret there will be needed the new table **NG_Week** consisting of records summarizing data concerning all patients present in the spa during a single week. Each record has 80 attributes bound to the subgroups introduced in the step 5 (NGr1..NGr80) and 38 attributes representing the procedures (Proc1..Proc40). Let as specify the contents of the table **NG_Week** for the week *n*:

- NGr1[*n*] is the number of days spent in spa by the patients of belonging to the group NGr1 during the week *n*. The same applies to Gr2, etc.
- Proc1[*n*] is the total number of all prescriptions of procedure Proc1 during the week *n*. The same applies to Proc3, etc.

The final table **NG_Week** contains 125 records corresponding to all the weeks in the period 1999-2000.

To use regression, there will be required additional information on individual procedures w.r.t. the defined subgroups. Let ***i*** be a fixed procedure. The new table **P_G_Week** ^{i} consists of records summarizing data concerning procedure ***i*** for all patients present in the spa during a single week. Each record has 160 attributes bound to the subgroups introduced in the step 5, the corresponding structure is (PNGr1 ^{i}.. PNGr80 ^{i}, ANGr1^{i}.. ANGr80 ^{i}). Let as specify the contents of the table **P_G_Week** ^{i} for week *n*:

- PNGr1 ^{i} [*n*] is the total number of prescriptions of the procedure ***i*** to the patients belonging to the group NGr1 during the week *n*. The same applies to PNGr2, etc.
- ANGr1 ^{i} [*n*] is the average number of prescriptions per day of the procedure ***i*** to the patients belonging to the group NGr1 during the week *n*, i.e. ANGr1 ^{i} [*n*] = PNGr1 ^{i} [*n*] / NGr1[*n*]. The same applies to ANGr2 ^{i} [*n*], etc.

The final table CTU_GWEEKS contains 125 records corresponding to all the weeks in the period 1999-2000. Regression is applied on each of the columns (ANGr1^{i}.. ANGr80 ^{i}) independently.

Lot of effort would be necessary to create several dozens of tables with rather complex relation to the original data. Fortunately, we can rely here on the support of SumatraTT and its GUI. It is ready to develop a transformation of the input data into the structure described as the table **NG_Week**. This tool will be used to extract the relevant information from the test data-set. Moreover, another SumatraTT transformation with a parameter ***i*** can be designed to calculate the content of **P_G_Week** from the input data.

5 Results and Conclusions

The approach described in this paper results in the MAPE prediction error which is approximately 12%. The sapa administration requested 20% precision only. 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.

Acknowledgements

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*

References

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/

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

Klema Jiri and Palous Jiri: **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

Klema Jiri, Lhotska Lenka, Stepankova Olga and Palous Jiri: 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

Nováková, L.: Praktické aplikace metod strojového učení. Diplomová práce K333, FEL ČVUT, Praha leden 2002

WEKA: http://www.cs.waikato.ac.nz/ml/weka/index.html