SDS-Rules and Classification on PKDD 2003 Discovery Challenge

Tomáš Karban

Department of Software Engineering, Faculty of Mathematics and Physics,
Charles University, Malostranské nám. 25, 118 01 Prague, Czech Republic
tomas.karban@matfyz.cz

Abstract. This article describes the effort of knowledge discovery in database of STULONG project of extensive epidemiological study of atherosclerosis primary prevention, which was developed in Czechoslovakia in 1976 and lasted approx. 20 years. Data of this project is a subject of a discovery challenge of ECML/PKDD 2003 conference. For the purpose of studying differences between couples of sets, SDS-rules are described in the first part of this article. Second part covers data preprocessing and preparation, third part presents knowledge discovered using the SDS-Miner tool (for SDS-rules) and Weka (for classification).

1. Introduction

In the early seventies of the 20th century, a project of extensive epidemiological study of atherosclerosis primary prevention was developed in Czechoslovakia. It was entitled "National Preventive Multifactor Study of Heart Attacks and Strokes". The group of 1417 men born in the years 1926–1937 was studied in details concerning social characteristics, physical activity, smoking, alcohol consumption, personal anamnesis, physical and biochemical examination. After an entry examination (performed in the years 1976–1979), the men were invited for control examinations during the following 20 years. Control examinations covered the changes in patient employment, physical activity, diet, smoking, and personal anamnesis, as well as detailed physical and biochemical examination. Up to the year 2001, 389 patients died and the date and cause of death was recorded. For a group of 403 patients, details similar to control examination were gathered through a letter questionnaire.

The main study aims were:

- to identify atherosclerosis risk factors prevalence,
- to follow their development and impact on the examined men health,
- to study the impact of the risk factors intervention,
- after 10–12 years, compare risk factors profile and health between the groups of patients, who originally showed / did not show atherosclerosis risk factors.

The risk factors are as follows: arterial hypertension, hypercholesterolemia, hypertriglyceridemia, smoking, overweight, and positive family case history. In control examinations, hyperglycemia, HDL cholesterol, LDL cholesterol and uric acid were taken into account as well.

The patients were divided into 3 basic groups:

- normal group – no risk factors present,
- risk group – some risk factors present,
- pathologic group – manifestation of a cardiovascular or another serious disease.

In this paper, the following analytic questions are studied:

Q1. Are there any strong relations concerning entry examination and death?

Q2. Are there differences in the entry examination between men of the risk group, who came down with observed cardiovascular disease (during control examinations) and those who stayed healthy?

2. SDS-rules

The idea of SDS-rules came up from the STULONG project and the need to study couples of sets with respect to some chosen property or quality. Certainly, it would be useful to get results about prevalence of diseases between different groups of patients, to study differences between normal and risk group of patients, to compare groups of patients with different physical, social, family or biochemical background or to search for couples of sets that differ markedly in the selected property.

Let's note that the abbreviation SDS stands for "Set Differs from Set".

In the STULONG project, the mining of association rules was already done (in the sense of a GUHA method, see [1] and [2]). Association rules can describe among others relations of implication or equivalence of two properties (denoted as antecedent and succedent). Specifically, the quantifier of founded implication can be understood as a characterization of a property given by succedent on the set of objects given by antecedent.

This notion can be extended to describe some property on two disjoint sets of objects. The SDS-rule is the expression of the form $\blacktriangle(\alpha,\beta,\psi)$, where α and β define the two disjoint sets A and B, ψ denotes the studied property and operator $\blacktriangle$ stands for an SDS-quantifier. As in the case of association rules, α, β and ψ are derived Boolean attributes in the form of literal conjunctions.

SDS-rules are verified on the basis of six-fold table (see the Fig. 1).

$\mathcal{M}$	ψ	$\neg\psi$	
α	a	b	r
β	c	d	s
$\neg(\alpha \vee \beta)$	e	f	t
	k	l	n

Fig. 1. Six-fold table $\mathrm{SDS}(\alpha,\beta,\psi,\mathcal{M})$ of α, β and ψ in the matrix $\mathcal{M}$

This six-fold table is a contingency table with frequencies of objects (not-)satisfying α, β and ψ. Based on this table, various quantifiers can be defined. Quantifier is in fact a truth-condition concerning frequencies $a...f$. In this data-mining effort, the quantifier of symmetric difference (in additive form) was used. It is defined as follows:

$$\Delta^{+}_{k,Base}(\alpha,\beta,\psi) = \left| \frac{a}{a+b} - \frac{c}{c+d} \right| \geq k \wedge (a+b) > Base \wedge (c+d) > Base$$

The first part of the conjunction describes the minimum amount of difference required (in absolute value). The other two parts enforce the minimum size of both sets *A* and *B*.

It is possible to define the multiplicative form of this quantifier by replacing the difference part with multiplicative expression:

$$\Delta^{*}_{k,Base}(\alpha,\beta,\psi) = \left(\frac{a}{a+b} \geq k \cdot \frac{c}{c+d} \vee \frac{a}{a+b} \leq \frac{1}{k} \cdot \frac{c}{c+d} \right) \wedge \ldots$$

This multiplicative form of symmetric difference quantifier was not used in this paper, because it is too sensitive about very low incidence of the property ψ in the sets *A* and *B*. For instance for $k = 2$, the 1% and 2% representation of the property ψ in the sets *A* and *B* respectively is considered a valid difference, but these values denote "very low presence" in both sets and thus should not be considered different. The additive form does not suffer with this weakness; on the other hand it considers the representations (0%, 10%) and (45%, 55%) the same importance, although the first one might be more important for its 0% value. From this perspective, the usage of some statistical difference test would overcome both pitfalls.

More details about SDS-rules and their affinity with association rules can be obtained in [5].

3. Data preparation

The input data is stored in 4 tables (data matrices): Entry, Control, Letter and Death. For the purpose of SDS-rules mining and classification, analyzed data must be represented in a single data matrix. Furthermore, the SDS-rules cannot handle numeric values, only categorized ones.

In the case of analytical question Q1, the tables Entry and Death must be joined. In the case of Q2, the table Entry must be joined with the one-bit information from Control table: "patient stayed healthy / patient came down with some cardiovascular disease". This one bit information is given by attributes HODNx in the Control table; patient is considered healthy iff all values of HODNx attributes indicate "disease not found".

In the Entry table, the following changes were made:

STAV, VZDELANI, ZODPOV, TELAKTZA, AKTPOZAM, DOPRAVA, DOPRATRV, ALKOHOL, BOLHR, BOLDK, DUSNOST, MOC	The value "not stated" replaced with NULL
SMOKING	The value 13 "not stated" was replaced by the most probable value 4 "15–20 cigs./day", all of them were former smokers
PIVO7, PIVO10, PIVO12	Dropped, type of beer considered unimportant

VINO, LIHOV	Information is implicitly contained in VINOMN and LIHMN
IML, HTD, HTL, ICT, ICTL, DIABD, DIABL, HYPLD, HYPLL	Info about diet and medicines for personal anamnesis attributes dropped (sporadic values)
IMTRV, HTTRV, ICTTRV, DIABTRV, HYPLTRV	Info, how long before entry examination the disease appeared, not considered
SYST1, SYST2	Replaced with an arithmetic mean SYST
DIAST1, DIAST2	Replaced with an arithmetic mean DIAST

4. Classification Results

In the field of classification, two tasks were solved:

- classification of the death cause, based on entry table,
- classification of the "healthy" attribute derived from attributes HODNx of the Control table (A2 questionnaire), based on Entry table.

For classification, Weka 3.2.3 was used (see [4]). Only two algorithms were chosen for the sake of simplicity – C4.5 decision trees (J48) and Neural Network classifier. Both algorithms can deal with numeric attributes directly without categorization.

The attribute PRICUMR (cause of death) value 10 (cause unknown) was replaced by NULL value, thus reducing the number of dead patients to 381. Distribution of the cause is following:

Myocardial infarction	80	21,00%
Coronary heart disease	33	8,66%
Stroke	30	7,87%
Other causes	79	20,73%
Sudden death	23	6,04%
Tumorous disease	114	29,92%
General atherosclerosis	22	5,77%

Let's consider very simple degenerate classifier, which classifies all objects into the biggest target class – tumorous disease in this case. Such a classifier would classify correctly 29.92% of cases (objects). The experiments with parameter settings for the classifiers prove, that the cause of death cannot be successfully predicated from the Entry data matrix, because all models were in their successfulness very close to 30%. See the table of results. Based on the parameter settings, all models tended to classify most of the objects to the class of tumorous disease.

It is important to note at this point, that the successfulness was measured by cross-validation method with the number of 20 folds. This method works by building a classification model over a $^{19}/_{20}$ of all patients and then testing the model on the $^{1}/_{20}$ of patients not used or building the model. This is repeated 20 times with all patients gradually used for testing.

J48 PART –C 0.1 –M 5	32.02%	Variant of C4.5 algorithm, which transforms the decision tree to the list of decision rules
Stack of some J48 with different param settings + Bayes with NN on top	30.18%	Various classifiers are stacked together, there is a neural network on top of the stack doing the final decision
Neural Network	29.56%	Back propagation learning method
J48 –N 20 –M 5	29.40%	C4.5 algorithm

The second classification to two classes of "healthy" attribute was the same failure. Only the risk group of patients was taken into account for this classification. The best result was achieved by the J48 classifier (–C 0.25 –M 5) producing 65.54% successful classifications, while the major group (those who stayed healthy) has 65.79% of the risk group patients.

Based on these results, we can draw the conclusion, that the set of attributes used to classify the patients is not sufficient to cover (sense) the influences leading to the cause of death or the presence of some cardiovascular disease. This is true for the classification of the whole set of studied patients. Different results might be achieved for some important subsets based on known attributes; unfortunately Weka does not easily allow generating subsets and building models on them.

5. SDS-rules Results for Q1

As noted in the section 2, SDS-rules can describe differences between two selected groups of patients in some studied property. This paper concerns only with symmetric difference quantifier and hence the studied property can be simply called "the difference".

Results are presented in tables. The first two columns describe the first and the second set of patients by a conjunction of literals (attributes and their respective values). Difference is listed in the third column in the same form. The next six columns list the frequencies in the SDS-table. The last column quotes the value of difference in the percentage form. The explanation for attribute names and values can be found on the PKDD Discovery Challenge 2003 homepage [6].

Let's analyze the following example:

first_set	second_set	difference	a	b	c	d	e	f	Diff
stav(2) & vzdelani(2) & zodpov(3)	zodpov(5)	pricumr(8)	7	6	0	11	70	279	53.8%

In this example, the difference in the property PRICUMR(8) was found (PRICUMR denotes the cause of death, value 8 means "other causes"). The first set is given by the conjunction STAV(2) & VZDELANI(2) & ZODPOV(3), which contains divorced patients with apprentice school education and "other" job responsibility. The second set is given by ZODPOV(5) containing the pensioners for "other reasons". The first set contains more deaths for other causes than the second by 53.8%. Whether more or less must be find out by comparing the values $a/(a+b)$ and $c/(c+d)$.

For every case, only the first few SDS-rules are listed. In the note behind every table there are parameters of quantifier with total number of SDS-rules discovered with that setting.

The first group of results is for the Q1, where the strong differences in the cause of death were studied for various groups of patients. The second group of results (for Q2) has the groups of patients fixed to the risk group. First set contains 273 patients, who came down with some cardiovascular disease during the control examinations; the second set contains 525 patients, who stayed healthy. Note that the precise number of patients in groups might be slightly lower if the groups are compared in a property with some missing values.

It is necessary to present the categorization of numeric attributes. The categorization always splits possible values into suitable intervals, which have meaningful number of objects inside and are easy to interpret. The numbers (categories) assigned to intervals respect the order of intervals going from zero to the number of intervals minus one. Let's set an example on VYSKA (the patients' height) attribute:

interval	category
< 160cm	0
< 165cm	1
< 170cm	2
< 175cm	3
< 180cm	4
< 185cm	5
< 190cm	6
>= 190cm	7

The other numeric attributes (VAHA 13, BMI 14, SYST 9, DIAST 7, TRIC 8, CHLST 5 and TRIGL 8) are processed in a similar way; the numbers by attributes state the number of created categories.

Q1a. Social → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
stav(2) & vzdelani(2) & zodpov(3)	zodpov(5)	pricumr(8)	7	6	0	11	70	279	53.8%
stav(2) & vzdelani(2) & zodpov(3)	vzdelani(3) & zodpov(2)	pricumr(8)	7	6	1	36	69	255	51.1%
stav(1) & vzdelani(3) & zodpov(2)	stav(2) & vzdelani(2) & zodpov(3)	pricumr(8)	1	30	7	6	70	262	50.6%
stav(2) & zodpov(3)	zodpov(5)	pricumr(8)	13	13	0	11	64	272	50.0%

Total 768 rules for quantifier setting k = 0.1 and Base = 10.

Q1b. Physical Activity → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
telaktza(3) & doprava(4)	telaktza(4) & doprava(3)	pricumr(16)	7	7	2	15	104	219	38,2%
telaktza(3) & aktpozam(1)	aktpozam(2) & doprava(4)	pricumr(16)	5	19	15	14	89	209	30,9%
aktpozam(1) & dopratrv(6)	aktpozam(3) & dopratrv(5)	pricumr(8)	8	14	1	15	67	264	30,1%
aktpozam(2) & doprava(4)	doprava(3) & dopratrv(6)	pricumr(16)	15	14	18	64	74	160	29,8%
telaktza(2) & doprava(3)	telaktza(3) & doprava(4)	pricumr(16)	6	23	7	7	100	214	29,3%
telaktza(3) & doprava(4)	doprava(3) & dopratrv(6)	pricumr(16)	7	7	18	64	82	167	28,0%
telaktza(3) & aktpozam(1)	aktpozam(3) & dopratrv(5)	pricumr(8)	8	16	1	15	69	263	27,1%
aktpozam(2) & doprava(4)	aktpozam(3) & dopratrv(6)	pricumr(16)	15	14	3	9	89	225	26,7%

Total 575 rules for quantifier setting k = 0.1 and Base = 10.

Q1c. Smoking → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
koureni(5) & dobakour(9)	dobakour(7,8)	pricumr(16)	3	16	6	5	105	246	38,8%
koureni(5,6) & dobakour(9)	dobakour(7,8)	pricumr(16)	4	18	6	5	104	244	36,4%
koureni(4,5) & dobakour(9)	dobakour(7,8)	pricumr(16)	12	41	6	5	96	221	31,9%
koureni(4) & dobakour(10)	dobakour(7,8)	pricumr(16)	20	67	6	5	88	195	31,6%
dobakour(9)	dobakour(7,8)	pricumr(16)	17	56	6	5	91	206	31,3%
koureni(2,3) & dobakour(9)	dobakour(7,8)	pricumr(16)	4	13	6	5	104	249	31,0%
koureni(4) & dobakour(9,10)	dobakour(7,8)	pricumr(16)	29	92	6	5	79	170	30,6%

Total 803 rules for quantifier setting k = 0.1 and Base = 10.

Q1d. Alcohol → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
alkohol(2) & pivomn(3)	alkohol(3) & lihmn(9)	pricumr(8)	1	13	4	7	74	281	29,2%
alkohol(2) & pivomn(3)	alkohol(3) & vinomn(5)	pricumr(9)	4	10	0	65	19	282	28,6%
alkohol(2) & pivomn(3)	alkohol(3) & lihmn(8)	pricumr(9)	4	10	1	57	18	290	26,8%
alkohol(2) & pivomn(3)	pivomn(1) & vinomn(5)	pricumr(9)	4	10	1	43	18	304	26,3%
alkohol(2) & pivomn(3)	pivomn(1) & lihmn(8)	pricumr(9)	4	10	1	30	18	317	25,3%
alkohol(2) & pivomn(3)	pivomn(2) & lihmn(7)	pricumr(9)	4	10	3	86	16	261	25,2%
alkohol(2) & pivomn(3)	alkohol(3) & pivomn(2)	pricumr(9)	4	10	3	83	16	261	25,1%

Total 374 rules for quantifier setting k = 0.1 and Base = 10.

Q1e. Sugar, Coffee, Tea → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
kava(2) & cukr(7,8,9)	kava(3) & cukr(8,9,10)	pricumr(8)	0	17	9	13	65	255	40,9%
kava(2) & cukr(7,8,9)	kava(3) & cukr(7,8,9)	pricumr(8)	0	17	6	9	68	259	40,0%
kava(2) & cukr(8,9,10)	kava(3) & cukr(8,9,10)	pricumr(8)	1	21	9	13	64	251	36,4%
kava(2) & cukr(8,9,10)	kava(3) & cukr(7,8,9)	pricumr(8)	1	21	6	9	67	255	35,5%
kava(2) & caj(6)	kava(3) & cukr(4,5,6)	pricumr(16)	2	9	12	11	94	236	34,0%
kava(1) & cukr(5,6,7)	kava(3) & cukr(4,5,6)	pricumr(16)	6	24	12	11	89	222	32,2%
kava(2) & cukr(7,8,9)	kava(3) & cukr(9,10,11)	pricumr(8)	0	17	4	9	70	259	30,8%
kava(1) & caj(5)	kava(3) & cukr(4,5,6)	pricumr(16)	14	51	12	11	82	196	30,6%

Total 1652 rules for quantifier setting k = 0.1 and Base = 10.

Q1f. Personal Anamnesis → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
diabet(1)	diabet(2) & hyplip(1)	pricumr(5)	2	9	8	9	39	164	28,9%
ht(2) & hyplip(1)	ht(2) & hyplip(2)	pricumr(5)	5	6	28	135	18	58	28,3%
im(1)	im(2) & diabet(1)	pricumr(8)	1	11	4	7	73	281	28,0%
im(1) & diabet(2)	diabet(1)	pricumr(8)	1	11	4	7	73	281	28,0%
im(1) & diabet(2)	im(2) & diabet(1)	pricumr(8)	1	11	4	7	73	281	28,0%
ht(2) & hyplip(1)	hyplip(2)	pricumr(5)	5	6	39	169	3	5	26,7%

Total 70 rules for quantifier setting k = 0.1 and Base = 10.

Q1g. Questionnaire A_2 → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
bolhr(2) & dusnost(1)	bolhr(3) & boldk(1)	pricumr(5)	6	46	9	8	64	242	41,4%
bolhr(2) & dusnost(1)	bolhr(3)	pricumr(5)	6	46	10	9	64	241	41,1%
bolhr(3) & boldk(1)	boldk(2) & dusnost(1)	pricumr(5)	9	8	2	13	67	275	39,6%
bolhr(2) & boldk(1)	bolhr(3) & boldk(1)	pricumr(5)	9	52	9	8	61	236	38,2%
bolhr(2) & boldk(1)	bolhr(3)	pricumr(5)	9	52	10	9	61	235	37,9%
bolhr(2)	bolhr(3) & boldk(1)	pricumr(5)	13	60	9	8	57	228	35,1%
bolhr(2)	bolhr(3)	pricumr(5)	13	60	10	9	57	227	34,8%
bolhr(1) & dusnost(2)	bolhr(3) & boldk(1)	pricumr(5)	6	27	9	8	64	260	34,8%
bolhr(1) & dusnost(2)	bolhr(3)	pricumr(5)	6	27	10	9	64	259	34,4%
bolhr(1) & boldk(2)	bolhr(3) & boldk(1)	pricumr(5)	3	13	9	8	66	275	34,2%
bolhr(1) & boldk(2)	bolhr(3)	pricumr(5)	3	13	10	9	66	274	33,9%

Total 164 rules for quantifier setting k = 0.1 and Base = 10.

Q1h. Physical Examination → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
vaha_cat(6,7,8) & bmi_cat2(5,6,7)	bmi_cat2(8,9,10) & subsc_cat(5,6,7)	pricumr(16)	15	13	2	20	97	234	44,5%
vaha_cat(6,7,8) & bmi_cat2(5,6,7)	bmi_cat2(9,10,11) & subsc_cat(5,6,7)	pricumr(16)	15	13	2	19	97	235	44,0%
vaha_cat(6,7,8) & bmi_cat2(5,6,7)	bmi_cat2(11,12,13) & syst_cat2(6,7,8)	pricumr(16)	15	13	3	23	96	231	42,0%
vaha_cat(2,3,4) & diast_cat2(4,5,6)	vaha_cat(6,7,8) & subsc_cat(5,6,7)	pricumr(8)	1	20	13	15	65	267	41,7%
vaha_cat(6,7,8) & bmi_cat2(5,6,7)	bmi_cat2(10,11,12) & subsc_cat(3,4,5)	pricumr(16)	15	13	3	22	96	232	41,6%
bmi_cat2(5,6,7) & tric_cat(3,4)	bmi_cat2(8,9,10) & subsc_cat(5,6,7)	pricumr(16)	14	14	2	20	98	233	40,9%
vaha_cat(5,6,7) & bmi_cat2(10,11,12)	vaha_cat(6,7,8) & bmi_cat2(5,6,7)	pricumr(16)	4	27	15	13	95	227	40,7%
vaha_cat(6,7,8) & bmi_cat2(5,6,7)	vaha_cat(6,7,8) & bmi_cat2(11,12,13)	pricumr(16)	15	13	3	20	96	234	40,5%
bmi_cat2(5,6,7) & tric_cat(3,4)	bmi_cat2(9,10,11) & subsc_cat(5,6,7)	pricumr(16)	14	14	2	19	98	234	40,5%

Total 411 rules for quantifier setting k = 0.1 and Base = 10.

Q1i. Biochemical Examination → Death

first_set	second_set	difference	a	b	c	d	e	f	Diff
trigl_cat(4,5) & moc(1)	moc(2)	pricumr(16)	10	11	2	14	101	236	35,1%
chlst_cat(4) & trigl_cat(2,3)	chlst_cat(0,1) & trigl_cat(6,7)	pricumr(5)	5	8	1	18	74	275	33,2%
chlst_cat(1) & trigl_cat(6,7)	chlst_cat(4) & trigl_cat(2,3)	pricumr(5)	1	15	5	8	74	278	32,2%
chlst_cat(3) & trigl_cat(3,4)	chlst_cat(4) & trigl_cat(2,3)	pricumr(5)	1	15	5	8	74	278	32,2%
chlst_cat(1) & trigl_cat(2,3)	chlst_cat(3) & trigl_cat(3,4)	pricumr(16)	3	13	8	8	103	246	31,3%
chlst_cat(1) & trigl_cat(6,7)	chlst_cat(3) & trigl_cat(3,4)	pricumr(8)	6	10	1	15	72	277	31,3%
chlst_cat(3) & trigl_cat(3,4)	chlst_cat(0,1) & trigl_cat(2,3)	pricumr(16)	8	8	3	13	103	246	31,3%
chlst_cat(3) & trigl_cat(1,2)	chlst_cat(0,1) & trigl_cat(6,7)	pricumr(5)	18	32	1	18	61	251	30,7%
trigl_cat(3,4) & moc(1)	moc(2)	pricumr(16)	21	28	2	14	90	219	30,4%
chlst_cat(0,1) & trigl_cat(6,7)	chlst_cat(3,4) & trigl_cat(1,2)	pricumr(5)	1	18	22	40	57	243	30,2%
chlst_cat(4) & trigl_cat(2,3)	chlst_cat(2,3) & moc(2)	pricumr(5)	5	8	1	11	73	277	30,1%
chlst_cat(3) & trigl_cat(1,2)	chlst_cat(3) & trigl_cat(3,4)	pricumr(16)	10	40	8	8	96	219	30,0%

Total 1681 rules for quantifier setting k = 0.1 and Base = 10.

7. SDS-rules Results for Q2

Q2a. Healthy → Social

first_set	second_set	difference	a	b	c	d	e	f	Diff
healthy(0)	healthy(1)	zodpov(1)	53	213	124	395	0	0	3,97%
healthy(0)	healthy(1)	vzdelani(3) & zodpov(3)	32	238	46	475	0	0	3,02%
healthy(0)	healthy(1)	zodpov(3)	126	140	232	287	0	0	2,67%
healthy(0)	healthy(1)	vzdelani(3) & zodpov(1)	18	252	47	474	0	0	2,35%
healthy(0)	healthy(1)	vzdelani(2) & zodpov(3)	54	218	116	407	0	0	2,33%
healthy(0)	healthy(1)	stav(3)	15	258	41	484	0	0	2,32%
healthy(0)	healthy(1)	stav(1) & vzdelani(3) & zodpov(3)	25	246	37	484	0	0	2,12%

Total 25 rules for quantifier setting k = 0.01 and Base = 10.

Q2b. Healthy → Physical Activity

first_set	second_set	difference	a	b	c	d	e	f	Diff
healthy(0)	healthy(1)	aktpozam(2) & dopratrv(5)	126	132	273	226	0	0	5,87%
healthy(0)	healthy(1)	aktpozam(2) & doprava(1)	40	222	103	395	0	0	5,42%
healthy(0)	healthy(1)	aktpozam(2) & doprava(1) & dopratrv(5)	35	225	93	405	0	0	5,21%
healthy(0)	healthy(1)	aktpozam(2)	192	81	390	134	0	0	4,10%
healthy(0)	healthy(1)	doprava(1)	59	196	134	359	0	0	4,04%
healthy(0)	healthy(1)	doprava(1) & dopratrv(5)	53	200	122	371	0	0	3,80%
healthy(0)	healthy(1)	telaktza(1) & aktpozam(2) & doprava(1)	24	243	64	443	0	0	3,63%
healthy(0)	healthy(1)	telaktza(1) & doprava(1)	37	226	88	416	0	0	3,39%
healthy(0)	healthy(1)	telaktza(1) & aktpozam(2) & doprava(1) & dopratrv(5)	21	245	57	450	0	0	3,35%
healthy(0)	healthy(1)	aktpozam(1)	59	214	96	428	0	0	3,29%
healthy(0)	healthy(1)	aktpozam(1) & doprava(4) & dopratrv(5)	14	254	11	508	0	0	3,10%

Total 75 rules for quantifier setting k = 0.01 and Base = 10.

Q2c. Healthy → Smoking

first_set	second_set	difference	a	b	c	d	e	f	Diff
healthy(0)	healthy(1)	koureni(4)	113	160	192	333	0	0	4,82%
healthy(0)	healthy(1)	koureni(1)	42	231	105	420	0	0	4,62%
healthy(0)	healthy(1)	koureni(1) & dobakour(0)	42	231	105	420	0	0	4,62%
healthy(0)	healthy(1)	koureni(1) & dobakour(0) & byvkurak(0)	42	231	105	420	0	0	4,62%
healthy(0)	healthy(1)	koureni(1) & byvkurak(0)	42	231	105	420	0	0	4,62%
healthy(0)	healthy(1)	dobakour(0)	42	231	105	420	0	0	4,62%
healthy(0)	healthy(1)	dobakour(0) & byvkurak(0)	42	231	105	420	0	0	4,62%
healthy(0)	healthy(1)	koureni(4) & dobakour(10)	81	192	133	392	0	0	4,34%
healthy(0)	healthy(1)	dobakour(10) & byvkurak(0)	157	116	280	245	0	0	4,18%

Total 28 rules for quantifier setting k = 0.01 and Base = 10.

Q2d. Healthy → Alcohol

first_set	second_set	difference	a	b	c	d	e	f	Diff
healthy(0)	healthy(1)	alkohol(2) & pivomn(2)	78	194	192	333	0	0	7,89%
healthy(0)	healthy(1)	alkohol(3) & pivomn(2)	77	195	110	415	0	0	7,36%
healthy(0)	healthy(1)	alkohol(2)	137	130	295	218	0	0	6,19%
healthy(0)	healthy(1)	pivomn(1) & lihmn(7)	62	211	89	436	0	0	5,76%
healthy(0)	healthy(1)	pivomn(1)	89	184	143	382	0	0	5,36%
healthy(0)	healthy(1)	alkohol(3) & pivomn(3)	16	257	58	467	0	0	5,19%
healthy(0)	healthy(1)	alkohol(2) & pivomn(2) & lihmn(8)	37	236	98	427	0	0	5,11%

Total 80 rules for quantifier setting k = 0.01 and Base = 10.

Q2e. Healthy → BMI

first_set	second_set	difference	a	b	c	d	e	f	Diff
healthy(0)	healthy(1)	bmi_cat2(9,10,11,12)	79	194	87	438	0	0	12,37%
healthy(0)	healthy(1)	bmi_cat2(8,9,10,11,12)	106	167	141	384	0	0	11,97%
healthy(0)	healthy(1)	bmi_cat2(9,10,11,12,13)	86	187	107	418	0	0	11,12%
healthy(0)	healthy(1)	bmi_cat2(9,10,11)	64	209	72	453	0	0	9,73%
healthy(0)	healthy(1)	bmi_cat2(2,3,4,5)	78	195	199	326	0	0	9,33%

Total 24 rules for quantifier setting k = 0.01 and Base = 10.

Q2f. Healthy → Blood Pressure

first_set	second_set	difference	a	b	c	d	e	f	Diff
healthy(0)	healthy(1)	diast_cat(1)	117	156	169	356	0	0	10,67%
healthy(0)	healthy(1)	syst_cat(0) & diast_cat(0)	32	241	110	415	0	0	9,23%
healthy(0)	healthy(1)	syst_cat(0)	43	230	130	395	0	0	9,01%
healthy(0)	healthy(1)	syst_cat(1)	134	139	211	314	0	0	8,89%
healthy(0)	healthy(1)	syst_cat(1) & diast_cat(1)	83	190	119	406	0	0	7,74%
healthy(0)	healthy(1)	diast_cat(0)	63	210	160	365	0	0	7,40%

Total 13 rules for quantifier setting k = 0.01 and Base = 10.

Q2g. Healthy → Cholesterol

first_set	second_set	difference	a	b	c	d	e	f	Diff
healthy(0)	healthy(1)	cholesterol(2)	147	126	184	341	0	0	18,80%
healthy(0)	healthy(1)	cholesterol(1)	90	183	232	293	0	0	11,22%
healthy(0)	healthy(1)	cholesterol(0)	36	237	109	416	0	0	7,58%

Total 3 rules for quantifier setting k = 0.01 and Base = 10.

Q2x. Entry → Healthy

This task was extraordinary. It reversed the approach for studying the differences between the patients, who stayed healthy during the control examinations and those who didn't. In this case, the HEALTHY attribute was used in the role of difference attribute, while the groups of patients were generated out of Entry table attributes (18 of them). The group could be defined as a conjunction of 2 literals max (for the task to be feasible, 3 is just too much).

The idea behind this task was to prove, that there are groups, which differ strongly in the HEALTHY attribute. From the tasks Q2a through Q2g you might get suspicion, that there are no strong differences in there, for no task showed any. Note that reversing the sets definition and difference definition has one important consequence: when defining sets by HEALTHY attribute, the whole set of patients was always studied (because there are only 2 possible values defining two sets); on the contrary – defining the groups by conjunctions of Entry table attributes, only subsets are studied (at least 50 patients in every set forced).

first_set	second_set	difference	a	b	c	d	e	f	Diff
pivomn(1) & bmi_cat(3)	pivomn(3) & cholesterol(1)	healthy(0)	36	34	8	44	229	447	36,0%
alkohol(3) & cholesterol(2)	pivomn(3) & cholesterol(1)	healthy(0)	55	56	8	44	207	420	34,2%
telaktza(3) & pivomn(1)	pivomn(3) & cholesterol(1)	healthy(0)	30	31	8	44	234	448	33,8%
pivomn(1) & cholesterol(2)	pivomn(3) & cholesterol(1)	healthy(0)	48	50	8	44	217	431	33,6%
stav(1) & alkohol(1)	alkohol(2) & cholesterol(0)	healthy(0)	28	27	15	64	225	423	31,9%
zodpov(3) & bmi_cat(0)	bmi_cat(3) & cholesterol(2)	healthy(0)	13	46	66	57	193	421	31,6%
xtlak_cat(1) & cholesterol(1)	xtlak_cat(2) & cholesterol(2)	healthy(0)	10	50	70	75	193	400	31,6%
koureni(4) & xtlak_cat(2)	xtlak_cat(1) & cholesterol(1)	healthy(0)	65	72	10	50	198	403	30,8%
telaktza(1) & bmi_cat(0)	bmi_cat(3) & cholesterol(2)	healthy(0)	12	40	66	57	193	428	30,6%
alkohol(2) & cholesterol(0)	alkohol(3) & cholesterol(2)	healthy(0)	15	64	55	56	200	399	30,6%
alkohol(2) & bmi_cat(0)	bmi_cat(3) & cholesterol(2)	healthy(0)	13	43	66	57	193	422	30,4%
alkohol(3) & bmi_cat(1)	bmi_cat(3) & cholesterol(2)	healthy(0)	13	43	66	57	193	424	30,4%
dobakour(9) & pivomn(1)	pivomn(3) & cholesterol(1)	healthy(0)	25	30	8	44	240	451	30,1%
koureni(4) & bmi_cat(0)	bmi_cat(3) & cholesterol(2)	healthy(0)	13	42	66	57	194	426	30,0%
alkohol(2) & cholesterol(0)	pivomn(1) & cholesterol(2)	healthy(0)	15	64	48	50	210	410	30,0%

Total 587 rules for quantifier setting k = 0.2 and Base = 50.

8. Conclusion, problems, unanswered questions

In this data-mining effort in STULONG project, I have learned the following lessons:

- It is impossible to present all discovered knowledge briefly in one article. There are virtually thousands of SDS-rules; the number of them is adjusted by the quantifier parameter setting and it is almost impossible to guess the right values. Furthermore, the importance of discovered knowledge is on expert judgment, therefore it is necessary to mine and review more than say top 10 rules.
- In the case of complicated database (like this one), it is an essential practice to structure the individual data-mining task into groups and hierarchies from the most general to more specific, as it was suggested in [3].
- Complicated database leads to exponential number of tasks provided that every combination of attributes is used. Total number of tasks is markedly lowered by allowing only the combinations from one "logical group" of attributes – this approach was used in this paper. While making the data-mining feasible, this approach a priori disables studying of interesting combinations that may arise from unrelated attributes, although they may be important for expert. Let's exemplify the combination "heavy smoking and drinking of alcohol leading to some cause of death", which was not studied. In the interaction with expert, such unordinary combinations should be discussed.
- In the data preparation step, it is necessary to pay attention to unknown/missing values in database. Those cannot be confused with specific attribute values. In the source data, particular attribute values have the meaning of "unknown/not stated" and vice versa – NULL values have sometimes the known value "no/none".
- It is difficult to work with blood pressure as it is a pair of mutually dependent numeric values (systolic and diastolic BP). It needs a discussion with an expert, whether and how it is possible to represent it either with one numeric value or one categorized value.
- Reviewing of hundreds of SDS-rules is impractical; it is necessary to further develop effective sorting and filtering tool for expert to get simple access to desired knowledge. It is also practical (although not necessary) to provide attribute values legend simultaneously.
- Most SDS-rules produce further questions. When the expert sees the difference between two groups of patients, the questions for the "SDS-rule neighborhood" might arise: what about other values of difference attributes for the same groups, are they important for difference? What if we drop some literals from sets definition, will the difference be influenced heavily? The tool for result reviewing should provide such a "neighborhood browsing" capability.

References

1. Hájek, P. – Havránek, T.: Mechanizing Hypothesis Formation – Mathematical Foundations for a General Theory, Springer-Verlag, 1978
2. Rauch, J. – Šimůnek, M.: Alternative Approach to Mining Association Rules (in FDM 2002, The Foundation of Data Mining and Knowledge Discovery, The Proceedings of the Workshop of ICDM02, ISBN: 4-947717-02-6, pages 157-162), Japan, December 2002
3. Dolejší, P. – Lín, V. – Rauch, J. – Šebek, M.: System of KDD Tasks and Results within the STULONG Project, PKDD 2002 Discovery Challenge
4. Weka 3: Machine Learning Software in Java
   http://www.cs.waikato.ac.nz/ml/weka/
5. Karban, T. – Rauch, J. – Šimůnek, M.: SDS-Rules and Association Rules, Proceedings of the 12th Annual Conference of Doctoral Students – WDS 2003, Prague
6. The homepage of PKDD Discovery Challenge 2003
   http://euromise.vse.cz/challenge2003/