# Analysis of STULONG Data by Rough Set Exploration System (RSES)

Sinh Hoa Nguyen, Hung Son Nguyen

[1] Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008, Warsaw, Poland
[2] Institute of Mathematics, Warsaw University
Banacha 2, 02-097, Warsaw, Poland
hoa@mimuw.edu.pl son@mimuw.edu.pl

**Abstract.** Rough Set Exploration System (RSES)[1] is a software tool designed and implemented at Warsaw University. It is a library of methods and a graphical user interface supporting variety of rough-set-based computations. We demonstrate features and abilities of RSES on the PKDD/ECML challenge over STULONG data.

## 1 Introduction

Rough Set Exploration System has been created to enable multi-directional practical investigations and experimental verification of research in decision support systems and classification algorithms, in particular of those with application of rough set theory.

First version of RSES and the library RSESlib was released several years ago. After modifications, improvements and removal of detected bugs it was used in many applications. The RSESlib was also used in construction of computational kernel of ROSETTA — an advanced system for data analysis (see [7]).

The first version of Rough Set Exploration System (RSES v. 1.0) in its current incarnation was introduced approximately two years ago (see [3]). Present version (2.0) introduces several changes, improvements as well as new algorithms - the result of selected, recent research developments in the area of rough-set-based data analysis.

Another important change in terms of technical development is the re-implementation of the RSES core classes in Java$^{TM}$ 2. Most of the computational procedures are now written in Java using its object-oriented paradigms. The migration to Java simplifies some development operations and, ultimately, leads to improved flexibility of the product permitting future migration of RSES software to operating systems other than Windows.

The main subject of our work is applying RSES to answer to challenge questions related to STULONG data. In Section 2, we introduce some basic notions related to rough set based methods. The data preparation process is described in Section 3. We emphasise the pivotal role of decision rules in analysis of decision

[1] RSES homepage: http://logic.mimuw.edu.pl/∼rses/

tables and some useful methods of decision rule extraction and improvement in Section 4. The results of experiments related to challenge question nr 6 are presented in Section 5. The paper ends with some concluding remarks in Section 6.

## 2 Basic notions

The notion of concept approximation is a focal point of many approaches to data analysis based on rough set theory. In the majority of rough set applications the approximations are used only at some stage of inductive learning. Most of existing solutions (see [4, 7, 11]) make use of decision rules derived from data. The structure of data that is central point of our work is represented in the form of *information system* or, more precisely, the special case of information system called *decision table*.

An *information system* [8] is a pair $\mathbb{S} = (U, A)$, where $U$ is a non-empty, finite set of *objects* and $A$ is a non-empty, finite set, of *attributes*. Each $a \in A$ corresponds to the function $a : U \to V_a$ called *evaluation function*, where $V_a$ is called the *value set* of $a$. Elements of $U$ could be interpreted as cases, states, patients, observations etc.

The above formal definition of information systems is very general and it covers many different "real information systems", e.g. data base systems. For simplification, we will use the simplest form of information systems called *information tables*, which can be implemented as two–dimensional arrays (matrixes) in every programming language. In information table, we usually associate its rows to objects, its columns to attributes and its cells to values of attributes on objects.

In supervised learning problems, objects from training set are pre-classified into some *categories* or *classes*. To manipulate this type of data we use a special case of information systems called *decision systems* which are information systems of the form $\mathbb{S} = (U, A \cup \{dec\})$, where $dec \notin A$ is a distinguished attribute called *decision*. The elements of attribute set $A$ are called *conditions*. Without loss of generality one can assume that the domain $V_{dec}$ of the decision $dec$ is equal to $\{1, \ldots, d\}$. The decision $dec$ determines a partition

$$U = CLASS_1 \cup \ldots \cup CLASS_d$$

of the universe $U$, where $CLASS_k = \{x \in U : dec(x) = k\}$ is called the $k^{th}$ *decision class of* $\mathbb{S}$ for $1 \leq k \leq d$.

## 3 Preparation of STULONG data for RSES

We concentrate on the challenge question nr 6 related to differences on entry examination between men from risk group, who came down with some cardiovascular diseases and those who stay healthy.

To answer to this question, we create a new decision table as a input data set for RSES.

1. **Objects:** objects of this table are related to patients belonging to risk group (KONSKUP = 3,4). There are 859 such objects.
2. **Attributes:** attributes of our decision table are defined from Entry table. We collect 24 attributes to the decision table. Some of them are original attributes from Entry table: STAV, VZDELANI, ZODPOV, TELAKTZA, AKTPOZAM, DOPRAVA, DOPRATRV, KOURENI, DOBAKOUR, ALKOHOL, VINO, LIHOV, PIVOMN, VINOMN, LIHMN, SYST1, DIAST1, SYST2, DIAST2. We also created some new attributes from those existing in Entry table:
   - WIEK: this attribute describes ages of patients. Its values are computed by ROKNAR and ROKVSTUP attributes.
   - BMI: describes whether the patient is overweight or not. This attribute is computed by HEIGHT and WEIGHT attributes using formula:

$$\text{BMI} = 1 \Leftrightarrow \text{weight in kg} \;/\; (\text{height in m})^2 \geq 25$$

   - PIVO: describes whether the patient drinks beers or not. This attribute is computed by merging 3 attributes PIVO7, PIVO10 and PIVO12 from alcohol attribute group.
   - CHLST, TRIGL: have been discretized according to the instruction in the STULONG Discovery Challenge 2003 homepage.
3. **Decision:** We create a new attribute related to the future history of patients, and it will be used for prediction purpose . This attribute, called "HDISEASE", describes the susceptibility of patients to heart diseases. The values of this attribute are computed from Control and Death tables as follow:
   - For a given patient, the value of HDISEASE attribute is equal to 1 if either he has heart disease in his history (i.e., he has at least one positive value on attributes HODN2, HODN3, HODN13, HODN14), or he dies by heart disease (i.e., he is presented in the Death table with PRICUMR attribute equal to 05, 06, 07, 17);
   - The value of HDISEASE attribute is equal to 2 if either he has other disease in his history (i.e., he has at least one positive value on attributes HODN1, HODN4, HODN11, HODN12, HODN15, HODN21, HODN23), or he dies by other disease (i.e., he is presented in the Death table with PRICUMR attribute equal to 08, 09, 10,16);
   - In other cases the value of HDISEASE attribute is equal to 0;

The frequency of HDISEASE attribute values within risk group is presented in the Table 1.

## 4 Rough set and decision rule extraction from data

One of the main topics of rough set theory is extraction of optimal, consistent decision rules from decision tables. Using those decision rules, one can define lower and upper approximation of concepts [1].

| HDISEASE - susceptibility to heart diseases | | |
|---|---|---|
| **Code** | **Meaning** | **Number of patients** |
| 0 | he will stay healthy | 502 |
| 1 | he will have heart disease or will die by heart disease | 221 |
| 2 | he will have other disease (or die by other disease) | 136 |

**Table 1.** Frequency of HDISEASE attribute values in risk group

Let $\mathbb{S} = (U, A \cup \{dec\})$ be a given decision table. *Decision rule* is a formula of the form

$$(a_{i_1} = v_1) \wedge ... \wedge (a_{i_m} = v_m) \Rightarrow (dec = k) \tag{1}$$

where $1 \leq i_1 < ... < i_m \leq |A|$, $v_i \in V_{a_i}$. Atomic formulas $(a_{i_j} = v_j)$ are called *descriptors*. We say that rule $r$ is *applicable* to the object, or alternatively, the object *matches* rule, if its attribute values satisfy the premise of the rule. With the rule we can connect some numerical characteristics such as *matching* and *support* (see [2]) as follows:

- $length(\mathbf{r})$ = the number of descriptors in the premise of $\mathbf{r}$ (i.e. the left hand side of implication).
- $[\mathbf{r}]$ = *carrier of* $\mathbf{r}$, i.e. the set of objects satisfying the premise of $\mathbf{r}$.
- $support(\mathbf{r}) = card([\mathbf{r}])$ = the number of objects satisfying the premise of $\mathbf{r}$;
- $confidence(\mathbf{r})$ = the measure of truth of the decision rule:

$$confidence(\mathbf{r}) = \frac{card([\mathbf{r}] \cap CLASS_k)}{card([\mathbf{r}])}$$

- The decision rule $\mathbf{r}$ is called *consistent* with $\mathbb{A}$ if $confidence(\mathbf{r}) = 1$.

Rough set theory offers many solutions for KDD tasks []. The existing methods based on rough set theory usually use decision rules to resolve both classification and description tasks. In data mining philosophy, we are interested on *short, strong* decision rules with *high confidence*. The linguistic features like “short”, ”strong” or “high confidence” of decision rules can be formulated by term of their length, support and confidence. RSES owns many methods related to decision rule extraction as well as decision rule improving. Let us mention the most important methods, which have been implemented in RSES.

### 4.1 Rule extraction by RSES

Given a decision table $\mathbb{S} = (U, A \cup \{dec\})$, the decision rule:

$$\mathbf{r} =_{def} (a_{i_1} = v_1) \wedge ... \wedge (a_{i_m} = v_m) \Rightarrow (dec = k)$$

is called minimal consistent decision rule if it is consistent with $\mathbb{S}$ and any decision rule $\mathbf{r}'$ created from $\mathbf{r}$ by removing one of descriptors from left hand side of $\mathbf{r}$ is not consistent with $\mathbb{S}$.

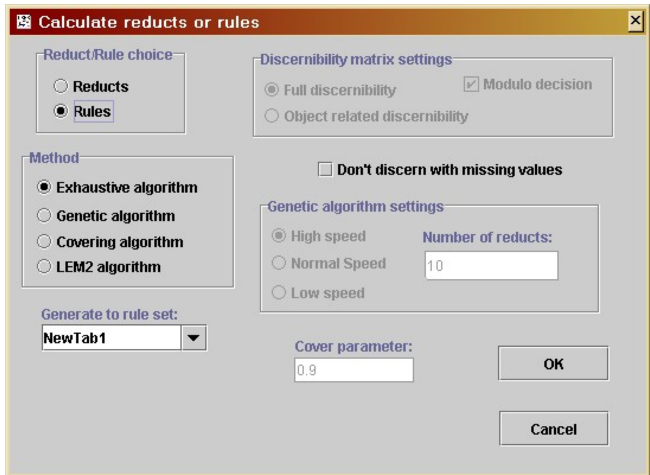


**Fig. 1.** Rule extraction menu in RSES

Unfortunately, the number of all minimal consistent decision rules for a given decision table can be exponential with respect to the size of decision table. Four heuristics have been implemented in RSES:

1. Exhaustive Algorithm: this algorithm realizes the computation of *object oriented reducts* (or local reducts). It has been shown that any minimal consistent decision rules for a given decision table $\mathbb{S}$ can be obtained from objects by reduction of redundant descriptors. The method is based on Boolean reasoning approach.
2. Genetic Algorithm: using genetic algorithm with permutation encoding and special crossover operator [2], one can compute a predefined number of minimal consistent rules.
3. Covering Algorithm: this algorithm searches for minimal (or very close to minimal) set of rules which cover the whole set of objects.
4. LEM2 Algorithm: this is a realisation of LEM2 algorithm, which is an other kind of covering algorithm (see [4]).

### 4.2 Rule improvement by RSES

After calculation of decision rules, one can select some most interesting rules for further usage. Some time we can not do it because of the low support of calculated rules. In RSES, the supports of decision rules can be increased by the following ways:

1. Discretization of real value attributes: two discretization methods called “local” and “global” methods [2] have been implemented in RSES. Both methods are based on Boolean reasoning approach. The local method is build on decision tree and it usually products more cuts than global method.

2. Rule shortening: Removing some descriptors from a given decision rule can increase its support, but it also decreases its confidence. In RSES, we can determine the "shortening ratio", which is a minimal acceptable threshold for confidences of decision rules in shortening process.
3. Generalization of rules: By generalized decision rules we denote the implications of the form:

$$\mathbf{r} =_{def} (a_{i_1} \in S_1) \wedge ... \wedge (a_{i_m} \in S_m) \Rightarrow (dec = k)$$

where $S_j \subset V_{a_{i_j}}$. RSES can construct generalized decision rules by merging those rules containing some common descriptors.

## 5 Experimental results

In this section, we present the results of experiments with RSES on STULONG data. We used RSES to search for decision rules which differ patients from risk group with respect to the heart disease.

| Class HDISEASE=0. Minimal confidence = 0.8 | |
|---|---|
| **Rules** | **Support** |
| (TELAKTZA=1) & (TRIGL=0) & (ALKOHOL=2) & (ZODPOV=2) | 48 |
| (PIVOMN=(2.0,3.0)) & (ALKOHOL=2) & (TRIGL=0) & (VZDELANI=3) | 42 |
| (TELAKTZA=1) & (VINOMN=(5.0,6.0)) & (ZODPOV=2) & (DOPRAVA=3) | 41 |
| (PIVO=1) & (LIHMN=(8.0,Inf)) & (TELAKTZA=1) & (BMI=0) | 38 |
| (STAV=1) & (TRIGL=0) & (DIAST1=(63.0,73.0)) | 36 |
| (PIVOMN=(2.0,3.0)) & (LIHMN=(8.0,Inf)) & (TELAKTZA=1) & (BMI=0) | 35 |
| (PIVOMN=(2.0,3.0)) & (TELAKTZA=1) & (BMI=0) & (LIHOV=12) | 35 |
| (LIHMN=(8.0,Inf)) & (ALKOHOL=2) & (VINOMN=(5.0,6.0)) & (CHLST=0) | 32 |
| (PIVOMN=(2.0,3.0)) & (ALKOHOL=2) & (ZODPOV=2) & (VZDELANI=3) | 31 |
| (VINOMN=(5.0,6.0)) & (ALKOHOL=2) & (PIVOMN=(2.0,3.0)) & (LIHMN=(8.0,Inf)) & (ZODPOV=2) | 31 |
| (PIVOMN=(2.0,3.0)) & (ALKOHOL=2) & (ZODPOV=3) & (VZDELANI=3) | 26 |
| (BMI=0) & (ZODPOV=2) & (DOPRAVA=1) | 24 |

**Table 2.** Results of Scenario 1: the strongest decision rules for the class HDISEASE=0, i.e., group of patients who will stay healthy.

### 5.1 Experiment 1

The first experiment was done on the whole decision table. We have tested many set of methods and corresponding parameters. After evaluation of results, we selected the best scenarios which return the most interesting decision rules.
**Scenario 1:** (local discretization + LEM2 + shortening)

In this scenario, the input decision table is performed with the following operations:

1. Discretization of real value attributes by local method;
2. Computation of decision rules using LEM2 method (covering ratio = 0.95);
3. Shortening of rules using shortening ratio = 0.8;

We have obtained 185 decision rules, which are supported by at least 5 objects (patients). In Table 2 and Table 3 we present the extracted rules which are supported by largest number of objects for classes HDISEASE=0 and HDISEASE=1. All these rules have confidence better than 0.8.

| Class HDISEASE=1. Minimal confidence = 0.8 | |
|---|---|
| **Rules** | **Support** |
| (STAV=1) & (CHLST=1) & (VINOMN=(-Inf,5.0)) & (PIVOMN=(2.0,3.0)) & (ALKOHOL=3) & (TRIGL=0) & (VZDELANI=2) | 9 |
| (TELAKTZA=3) & (AKTPOZAM=1) & (VZDELANI=3) | 9 |
| (LIHMN=(-Inf,8.0)) & (DOPRAVA=3) & (DOPRATRV=(6.0,8.0)) & (KOURENI=(5.0,6.0)) & (ZODPOV=3) & (BMI=0) | 8 |
| (STAV=1) & (AKTPOZAM=1) & (ALKOHOL=2) & (DOPRAVA=4) | 7 |
| (TELAKTZA=3) & (DOPRATRV=(-Inf,6.0)) & (ALKOHOL=3) & (WIEK=(49.0,50.0)) | 7 |
| (CHLST=1) & (PIVOMN=(-Inf,1.0)) & (TELAKTZA=1) & (DOBAKOUR=(9.0,10.0)) | 6 |
| (STAV=1) & (DOPRATRV=(-Inf,6.0)) & (VINOMN=(-Inf,5.0)) & (DOPRAVA=3) & (VZDELANI=2) & (BMI=1) & (TELAKTZA=3) | 6 |
| (LIHMN=(-Inf,8.0)) & (KOURENI=(5.0,6.0)) & (CHLST=0) & (VINOMN=(5.0,6.0)) & (ALKOHOL=2) | 6 |
| (PIVOMN=(2.0,3.0)) & (LIHMN=(8.0,Inf)) & (ZODPOV=3) & (WIEK=(51.0,52.0)) | 6 |
| (LIHMN=(8.0,Inf)) & (VZDELANI=2) & (DOPRATRV=(-Inf,6.0)) & (TRIGL=0) & (TELAKTZA=4) & (DOBAKOUR=(10.0,Inf)) | 6 |

**Table 3.** Results of Scenario 1: the strongest decision rules for the class HDISEASE=1, i.e., group of patients who will have problem with heart disease.

**Scenario 2:**(global discretization + LEM2 + shortening):

1. Discretization of real value attributes by global method;
2. Computation of decision rules using LEM2 method;
3. Shortening of rules using shortening ratio = 0.8;

This time, we have obtained 196 decision rules, which are supported by at least 5 objects (patients). In Table 4 and Table 5 we present the extracted rules which are supported by largest number of objects for classes HDISEASE=0 and HDISEASE=1.

| Class HDISEASE=0. Minimal confidence = 0.8 | |
|---|---|
| **Rules** | **Support** |
| (ALKOHOL=2) & (ZODPOV=2) & (TRIGL=0) & (TELAKTZA=1) | 48 |
| (DOBAKOUR=MISSING) & (SYST2=(125.5,142.5)) | 44 |
| (LIHMN=(7.5,Inf)) & (DIAST1=(-Inf,77.5)) & (BMI=0) | 37 |
| (ALKOHOL=2) & (ZODPOV=2) & (KOURENI=(-Inf,3.5)) | 34 |
| (DOPRATRV=(-Inf,5.5)) & (TRIGL=0) & (BMI=0) & (VZDELANI=4) | 34 |
| (LIHMN=(7.5,Inf)) & (TELAKTZA=1) & (ALKOHOL=2) & (DOBAKOUR=(9.5,Inf)) & (SYST2=(107.5,125.5)) | 32 |
| (PIVOMN=(1.5,2.5)) & (ALKOHOL=2) & (VZDELANI=3) & (ZODPOV=2) | 31 |
| (PIVOMN=(1.5,2.5)) & (DOBAKOUR=(-Inf,9.5)) & (VZDELANI=4) | 31 |
| (WIEK=(47.5,49.5)) & (ZODPOV=2) & (TELAKTZA=1) | 31 |
| (LIHMN=(7.5,Inf)) & (SYST2=(107.5,125.5)) & (CHLST=0) | 30 |
| (ALKOHOL=2) & (ZODPOV=2) & (WIEK=(47.5,49.5)) | 28 |
| (PIVOMN=(1.5,2.5)) & (ZODPOV=1) & (SYST2=(125.5,142.5)) | 27 |
| (PIVOMN=(1.5,2.5)) & (ZODPOV=1) & (CHLST=0) | 25 |
| (BMI=0) & (ZODPOV=2) & (DOPRAVA=1) | 24 |
| (ZODPOV=2) & (SYST1=(-Inf,111.0)) | 24 |
| (CHLST=1) & (LIHMN=(-Inf,7.5)) & (DOPRATRV=(-Inf,5.5)) & (VZDELANI=4) & (TRIGL=0) | 24 |

**Table 4.** Results of Scenario 2: the strongest decision rules for the class HDISEASE=0.

**Comments on Experiment Results:** The results of experiments on STULONG data, obtained by using RSES system, can be used for many goals.

- Decision rules should be verified by experts to check whether strong rules are new knowledge or just by accident. We hope the medical experts will find some interesting rules among those, which have been extracted by RSES.
- We can have some conclusions related to the analytic question 6:
  1. The class HDISEASE=0 (i.e., the group of patients who will stay healthy) is more regular than other classes. All rule extraction scenarios confirm the observation, that there are more strong rules for this class (see Table 6).
  2. Building classifiers using decision rules: RSES can classify new cases using rules extracted from data. Again, the class HDISEASE=0 is easier for prediction task. We used the decision rules (extracted from risk group) to classify patients from Normal group and Pathological group (HODSKUP= 1,2,5,6). The accuracy of classification on the class HDISEASE=0 was above 90%, whereas the accuracy of classification on the class HDISEASE=1 was only 36%.

| Class HDISEASE=1. Minimal confidence = 0.8 | |
|---|---|
| **Rules** | **Support** |
| (LIHMN=(-Inf,7.5)) & (DOPRAVA=3) & (DOPRATRV=(5.5,6.5)) & (KOURENI=(4.5,5.5)) & (ZODPOV=3) & (BMI=0) | 8 |
| (STAV=1) & (PIVOMN=(1.5,2.5)) & (AKTPOZAM=1) & (DOPRAVA=4) | 8 |
| (BMI=1) & (KOURENI=(4.5,5.5)) & (ALKOHOL=3) & (SYST1=(131.0,161.0)) & (TRIGL=0) | 8 |
| (ALKOHOL=3) & (DIAST1=(77.5,81.0)) & (DIAST2=(72.5,84.5)) & (WIEK=(47.5,49.5)) | 8 |
| (CHLST=1) & (LIHMN=(-Inf,7.5)) & (PIVOMN=(1.5,2.5)) & (TELAKTZA=1) & (WIEK=(50.5,Inf)) | 8 |
| (DOPRATRV=(-Inf,5.5)) & (TELAKTZA=3) & (AKTPOZAM=1) & (PIVOMN=(-Inf,1.5)) | 8 |
| (STAV=1) & (AKTPOZAM=1) & (ALKOHOL=2) & (DOPRAVA=4) | 7 |
| (DOBAKOUR=(9.5,Inf)) & (STAV=1) & (VZDELANI=2) & (WIEK=(47.5,49.5)) & (ALKOHOL=3) | 7 |
| (ALKOHOL=3) & (KOURENI=(3.5,4.5)) & (LIHMN=(7.5,Inf)) & (WIEK=(50.5,Inf)) | 6 |
| (AKTPOZAM=2) & (DOPRATRV=(-Inf,5.5)) & (VINOMN=(4.5,Inf)) & (WIEK=(47.5,49.5)) & (DIAST1=(77.5,81.0)) & (DIAST2=(72.5,84.5)) | 6 |

**Table 5.** Results of Scenario 2: the strongest decision rules for the class HDISEASE=1.

| | HDISEASE=0 | HDISEASE=1 | HDISEASE=0 |
|---|---|---|---|
| $p = 5\%$ | 11 | 0 | 0 |
| $p = 4\%$ | 16 | 2 | 1 |
| $p = 3\%$ | 39 | 5 | 4 |
| $p = 2\%$ | 65 | 19 | 29 |

**Table 6.** Number of decision rules (extracted by Scenario 1) supported by at least $p\%$ of objects from a given class

### 5.2 Experiment 2

In this experiment we applied the scenario 2 to some groups of attributes only. We took under consideration 5 groups of attributes:

**G1** : Social factors
**G2** : Physical activity
**G3** : Smoking
**G4** : Alcohol
**G5** : BMI + Blood pressure + Cholesterol level

Single groups **G1, G2, G3, G4** of attributes did not return good rules, because the decision tables restricted to attributes from those groups are not consistent. We repeated the scenario 2 to the following decision tables:

**Social_Act_HDISEASE:** this table was created by merging attributes from **G1** and **G2**. The strongest decision rules are presented in Table 7.

**Sm_Alc_HDISEASE:** this table was created by merging attributes from **G3** and **G4**. The strongest decision rules are presented in Table 8.

**BMI_Blood_Chlst_HDISEASE:** this table was created by merging attributes from **G5**. The strongest decision rules are presented in Table 9.

| Class HDISEASE=0. Minimal confidence = 0.8 | |
|---|---|
| **Rules** | **Support** |
| (TELAKTZA=1) & (DOPRATRV=(8.5,Inf)) | 22 |
| (TELAKTZA=2) & (WIEK=(38.5,41.5)) | 18 |
| (STAV=1) & (DOPRAVA=3) & (ZODPOV=3) & (TELAKTZA=4) | 18 |
| (TELAKTZA=1) & (ZODPOV=2) & (WIEK=(48.5,49.5)) | 17 |
| (TELAKTZA=1) & (WIEK=(41.5,42.5)) | 16 |
| (VZDELANI=4) & (ZODPOV=3) & (WIEK=(45.5,48.5)) | 15 |
| (VZDELANI=3) & (ZODPOV=2) & (WIEK=(48.5,49.5)) | 14 |
| (STAV=1) & (DOPRATRV=(5.5,6.5)) & (TELAKTZA=4) | 14 |
| (VZDELANI=3) & (ZODPOV=2) & (WIEK=(38.5,41.5)) | 14 |
| Class HDISEASE=1. Minimal confidence = 0.8 | |
| **Rules** | **Support** |
| (ZODPOV=5) | 11 |
| (VZDELANI=3) & (TELAKTZA=3) & (AKTPOZAM=1) | 9 |
| (VZDELANI=2) & (WIEK=(49.5,50.5)) & (ZODPOV=1) | 6 |
| (DOPRATRV=(5.5,6.5)) & (VZDELANI=3) & (TELAKTZA=3) & (WIEK=(45.5,48.5)) | 5 |

**Table 7.** Results of Scenario 2 to the Social_Act_HDISEASE table: the strongest decision rules for the classes HDISEASE=0 and HDISEASE=1.

**Comments on Experiment Results:** One can see that rules extracted by experiment 2 are more compact, but their supports are lower in general.

Attributes from social factors and physical activity groups seem to be best in description of susceptibility to heart disease.

Attributes from smoking and alcohol groups seem to be best in description of patients who will stay healthy.

## 6 Conclusions

We presented some experiments of knowledge discovery process using RSES system. In our opinion, first results are quite promising. We plan to continue our research on STULONG data using more advanced tools in RSES like classification, decomposition, new feature extraction e.t.c.

We also plan to include some new methods to RSES related to risk analysis and unbalanced decision tables. These topics are intensively studied by our research group but are not implemented yet.

| Class HDISEASE=0. Minimal confidence = 0.75 | |
|---|---|
| **Rules** | **Support** |
| (LIHOV=12 & (DOBAKOUR=MISSING | 53 |
| (LIHMN=(7.5,8.5) & (DOBAKOUR=MISSING | 47 |
| (LIHMN=(7.5,8.5) & (KOURENI=(-Inf,1.5) | 47 |
| (LIHMN=(-Inf,7.5) & (DOBAKOUR=(8.5,9.5) & (PIVOMN=(1.5,2.5) | 29 |
| (VINOMN=(4.5,5.5) & (KOURENI=(2.5,3.5) | 27 |
| (ALKOHOL=3 & (VINO=11 & (DOBAKOUR=MISSING | 23 |
| (PIVOMN=(1.5,2.5) & (VINOMN=(4.5,5.5) & (LIHOV=12 & (ALKOHOL=2 & (DOBAKOUR=(9.5,Inf) & (KOURENI=(3.5,4.5) | 22 |
| (VINOMN=(-Inf,4.5) & (ALKOHOL=2 & (KOURENI=(-Inf,1.5) | 21 |

**Table 8.** Results of Scenario 2 to the Sm_Alc_HDISEASE table: the strongest decision rules for the classes HDISEASE=0, there are no interesting rules for the class HDISEASE=1.

| Class HDISEASE=0. Minimal confidence = 0.8 | |
|---|---|
| **Rules** | **Support** |
| (BMI=0) & (SYST2=(107.5,112.5)) | 33 |
| (DIAST1=(52.5,67.5)) | 30 |
| (BMI=0) & (DIAST1=(72.5,77.5)) | 24 |
| (DIAST1=(77.5,83.5)) & (SYST2=(107.5,112.5)) | 21 |
| (TRIGL=0) & (SYST1=(107.5,111.0)) | 21 |
| (BMI=0) & (SYST1=(107.5,111.0)) | 21 |
| (BMI=0) & (TRIGL=0) & (DIAST2=(86.5,91.0)) | 18 |
| (BMI=0) & (SYST1=(-Inf,107.5)) | 18 |
| Class HDISEASE=1. Minimal confidence = 0.8 | |
| **Rules** | **Support** |
| (DIAST1=(122.5,Inf)) | 6 |
| (SYST2=(162.5,192.5)) & (DIAST1=(102.5,109.0)) | 5 |
| (DIAST2=(76.5,81.0)) & (SYST1=(131.0,136.5)) & (DIAST1=(77.5,83.5)) | 5 |
| (BMI=0) & (SYST1=(131.0,136.5)) & (DIAST1=(77.5,83.5)) | 5 |

**Table 9.** Results of Scenario 2 to the BMI_Blood_Chlst_HDISEASE table: the strongest decision rules for the classes HDISEASE=0 and HDISEASE=1.

## References


1. Bazan, J.G, Nguyen H. S., Skowron A., Szczuka M.: A View on Rough Set Concept Approximations. In G. Wang, Q. Liu, Y. Yao, A. Skowron (Eds.): Proceedings of RSFDGrC 2003, Chongqing, China, Lecture Notes in Computer Science 2639 Springer 2003, pp. 181-188
2. Bazan, J.G., Nguyen, H.S., Nguyen, S.H, Synak, P., Wróblewski, J.: Rough Set Algorithms in Classification Problem. In: Polkowski, L., Tsumoto, S., Lin, T.Y. (eds), Rough Set Methods and Applications, Physica-Verlag, 2000 pp. 49–88.
3. Bazan J., Szczuka M. RSES and RSESlib - A Collection of Tools for Rough Set Computations, Proc. of RSCTC'2000, LNAI 2005, Springer Verlag, Berlin, 2001
4. Grzymała-Busse J., A New Version of the Rule Induction System LERS Fundamenta Informaticae, Vol. 31(1), 1997, pp. 27–39

5. Nguyen S.H., Nguyen H.S., Discretization Methods in Data Mining, In [9] vol.1, pp. 451-482
6. Nguyen S.H., Skowron A. and Synak P., Discovery of Data Patterns with Applications to Decomposition and Classfification Problems. In [9] vol.2, pp. 55-97.
7. Ørn A., The ROSETTA Homepage, `http://www.idi.ntnu.no/∼aleks/rosetta`
8. Pawlak Z.: *Rough sets: Theoretical aspects of reasoning about data*, Kluwer Dordrecht, 1991.
9. Polkowski L., Skowron A.(ed.), Rough Sets in Knowledge Discovery vol. 1 and 2, Physica Verlag, Heidelberg, 1998
10. Wróblewski J., Covering with Reducts - A Fast Algorithm for Rule Generation, Proceeding of RSCTC'98, LNAI 1424, Springer Verlag, Berlin, 1998, pp. 402-407
11. The RSES Homepage, `http://logic.mimuw.edu.pl/∼rses`