Modelling of Rough-Fuzzy Classifier

KŘUPKA JIŘÍ, JIRAVA PAVEL
Institute of System Engineering and Informatics
Faculty of Economics and Administration, University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC
jiri.krupka@upce.cz, pavel.jirava@upce.cz

Abstract: The paper reflects the trend of the past years which is based on the diffusion of various traditional approaches and methods to the way of tackling new problems, in this case to the classification. Two components of the computational intelligence are applied in a classification model. It means rough and fuzzy sets on the basis of which the data classification hybrid model is proposed. In the second part the current knowledge in the investigated field are summarized and briefly explained. The algorithm for uncertain data operations and conditioned rules generation is introduced too. We have brought in an original toolbox called RSTbox for data processing and automatic rules generation. Proposed model and toolbox are carried out in MATLAB, tested on more data files, and compared to others, already known classification methods.

Key-Words: Classification, rough sets theory, fuzzy sets theory, rules generation, hybrid model, evaluation

1 Introduction

A role of classification is to classify objects, events and real-life situations into classes. Each of the reviewed objects is unique, original and its classification means a certain degree of generalization. Let's define a system for the particular objects i.e. input and output variables, elements (objects) and their mutual relations. Defining and collecting the data of input/output variables cannot be generalized, even though this stage influences the classification result. An application of classification methods based on the computational intelligence (CI) represents an effective tool for realization of a classification model.

Areas of CI (fuzzy sets, neural networks, genetic algorithms, rough sets etc.) belong to a fast developing field in the applied research. It is composed of several theories and approaches which, despite being different from one another, have two common denominators which are the non-symbolic representation of pieces of knowledge [2] and „bottom-up" architecture where the structures and paradigms appear from an unordered beginning [2,5,45]. On the basis of achieved classification results it seems to be effective and up-to-date to tackle the classification problem using a hybrid approach combining rough sets and fuzzy sets (FSs), both belonging to the field of the CI research.

The rough sets theory (RST) [21,27,30,31], due to prof. Pawlak, is based on the research of information system logical properties, and uncertainty in it is expressed by a boundary region (BR). Every investigated object is related to a specific piece of information, to specific data. The objects which are characterized by the same pieces of information are mutually undistinguishable from the point of view of the accessible pieces of information. This is expressed in RST by the indiscernibity relations.

The theory of FSs, due to prof. Zadeh, is a relatively new approach to uncertainty. In this theory an element belongs to a set according to the membership degree (membership function values) [44,45,46], i.e. in a closed interval. It is an enlargement of the traditional sets theory in which an element either is or is not a set member. If we endeavour to describe and model a particular reality problem we encounter a certain discrepancy. On one hand, there is the accuracy of mathematical methods by which a specific problem is described and, on the other hand, there is a very complicated reality necessitating a range of simplifications and the consequent inaccuracy, infidelity of the model arising from them.

RST and FSs are applied in a classifier modelling [19]. Our case deals with a hybrid classifier; it means a rough-fuzzy classifier (RFC). RST were used for a definition of IF-THEN rules and FSs were applied in RFC as a fuzzy inference system (FIS). FIS have been successfully applied in fields such as modelling of municipal creditworthiness, automatic control, decision analysis, data analysis, decision systems or expert system [4,7].

Goals of this paper is to create, verify and analyse a hybrid data classifier model. We applied a

rough set toolbox (RSTbox) for a rules generation. If-then rules were then used in Mamdani type of FIS which represents a kernel of RFC.

2 Problem Formulation

A definition of RST is connected with a term "an information system". From the view of RST is an information system (IS) can be defined as an information table [7,27,32] which represents a data set where: every column represents an attribute that can be measured for each object. A human expert or user may also supply the attribute. Each row represents a case or generally an object. More formally [7], IS is the 4-tuple:

$$IS=(U, A, V_a, f_a) \text{ for } \forall\ a_i \in A,\ i=1,2,\ldots,n\ , \qquad (1)$$

where: $U=\{x_1, x_2\ldots, x_m\}$ is a finite sets of objectives (universe), $A=\{a_1, a_2\ldots, a_n\}$ is a finite set of attributes, V_a is the domain of the attributes, f_a: $U \rightarrow V_a$ is a information function such that $f(x,a) \in V_a$ for each $a \in A$, $x \in U$ [7].

It is possible to express IS [7,21] as a decision table (see the Table 1) where: a_i is i-th attribute; x_j is j-th object; v_{ji} is an attribute value from its domain and d is a decision attribute with value h_r; for j=1,2,…,m and r=1,2,…,q.

Table 1 Decision table

Objects	Attributes					Decision attribute
	a_1	a_2	a_3	…	a_n	d
x_1	v_{11}	v_{12}	v_{13}	…	v_{1n}	h_1
x_2	v_{21}	v_{22}	v_{23}	…	v_{2n}	h_2
x_3	v_{31}	v_{32}	v_{33}	…	v_{3n}	h_3
…	…	…	…	…	…	…
x_m	v_{m1}	v_{m2}	v_{m3}	…	v_{mn}	h_q

The real life data set is represented as a table, too. For each object-attribute pair there is a known descriptor (a specific and precise value of an attribute).

A limited discernibility of objects by means of the attribute values generally prevents their precise classification [40]. In practice, the input data presented as decision tables, can have the missing attribute and decision values, i.e., decision tables are incompletely specified. The attribute values can be uncertain because of many reasons.

In practice, the input data presented in Table 1 can have missing attribute and decision values, i.e., decision tables are incompletely specified. The attribute values can be missing because of many reasons. The missing data are common part of data analysis and we need to bear their existence in mind [31]. It is necessary to define and divide the cases that can occur in reality. According to [35,36] we calculate with three algorithms of missing data operating.

The first is called „Missing Completely at Random" (MCAR), and the specialized technical literature also calls it Rubin´s condition of MCAR. Its prerequisites are that the incomplete nature of an observation is independent on what the particular observation contains, or what it would contain supposing it was complete [26]. To put it in other words, by leaving out the missing values we suppose that the complete observations form a random selection from the original data (complete) file. The second algorithm is known under the name „Missing at Random" (MAR). It says that if the data are given, the missing data algorithm is not dependent on the fact whether the data were not observed. The last algorithm is called „Non-ignorable" or „Missing Not at Random", and in this case we can not ignore the missing data algorithm [26].

For our convenience in this paper four types of uncertainty [41] need to be distinguished: discretization of quantitative attributes; imprecise values of quantitative attribute; multiple values of attribute and unknown or missing values of attribute [11,12,40].

Uncertainty coming from unknown or missing attributes occurs when the attribute value is unknown. There are two main reasons why an attribute value is missing: either the value was lost (e.g., was erased) or the value was not important. In the first case, the attribute value was useful but currently we have no access to it. One of RST approaches to data mining [43] is system "Learning from Examples using Rough Sets" (LERS). LERS uses two algorithms: Learning from Examples Module version 1 (LEM1) and Learning from Examples Module version 2 (LEM2) [10,11,13]. The approach to the missing attribute values [22,37,41] when all missing values were lost, was presented in an adapted LEM2 algorithm [13]. There are several areas of the use and modifications of LEM1 and LEM2 (MLEM2) algorithms. For example, in [13] the base rule induction via clustering decision classes is proposed. To induce the decision rules, LEM2 algorithm is used. In other pieces of scientific literature MLEM2 algorithm for decision rules generation in the area of data preparation for the cardio logical decision support was implemented. Other approaches to the

rule induction using LEM2 algorithm are described in [11,12].

2.1 Rough Sets

The assumption that objects can be seen only through the information available about them leads to the view that knowledge has a granular structure. Thus some objects appear as similar and undiscerned. Therefore in RST [30] we assume that any vague concept is replaced by a pair of precise concepts – the lower and upper approximation of the vague concept. The lower approximation (LA) consists of all objects which surely belong to the concept, and the upper approximation (UA) consists of all objects which possibly belong to the concept. And the difference between UA and LA is called BR.

The approximations are two basic operations in RST [31]. Suppose we are given two finite and non empty sets U and A, U is called the universe and A is a set of attributes. With attributes $a \in A$ we associate a set V_a (value set) called the domain of a. Any subset B of A determines a binary relation IND(B) on U which will be called an indiscernibility relation [21]:

$$IND(B)=\{(x,y)\in U \mid \forall\, a \in B \;\; a(x)=a(y)\}, \qquad (2)$$

where: IND(B) is an equivalence relation and is called B-indiscernibility relation. If $(x,y)\in IND(B)$, then x and y are B-indiscernible (indiscernible from each other by attributes from B). The equivalence classes of the B-indiscernibility relation will be denoted B(x).

The indiscernibility relation will be used now to define basic concept of RST. Let IS be defined by (1) and $B \subseteq A$, $X \subseteq U$. We can approximate X using only the information contained in B by constructing LA and UA of X in the following way:

$$\text{LA: } \underline{B}(X)=\{x\in U: B(x)\subseteq X\}, \qquad (3)$$

$$\text{UA: } \overline{B}(X)=\{x\in U: B(x)\cap X\neq\emptyset\}. \qquad (4)$$

The objects in LA can be with certainty classified as members of X on the basis of knowledge in B and the objects in UA are classified as possible members of X on the basis of knowledge in B. BR is the set of X and thus consists of those objects that we cannot definitely classify into X on the basis of knowledge B in the following way:

$$\text{BR: } BN_B(X)=\overline{B}(X)-\underline{B}(X). \qquad (5)$$

If the boundary region is empty, then set X is with respect to B. If the boundary region is not empty, set X is rough with respect to B.

The rough sets are defined by approximations and have properties defined in [21,30,31,32]. The RST is used in abundance applications (see more in [7,21,30,38]).

2.2 Fuzzy Sets

The theory of FSs is an approach to uncertainty [44,45,46]. In this theory an element belongs to a set according to the membership degree (membership function values) [45], i.e. to closed interval [0,1]. It is an enlargement of the traditional sets theory in which an element either is or is not a set member. During the process of a real system definition the effort to maximize the accuracy of a system description leads to the disproportionate rise of the number of definitions and conditions. In [45] the principal of incompatibility is formulated: "*If the complexity of a system rises, our ability to formulate accurate and significant judgements about its behaviour decreases, and the border is reached behind, the accuracy and relevance of which are practically mutually exclusive characteristics.*"

It is an enlargement of the traditional sets theory in which an element either is or is not a set member. If we endeavour to describe and model a particular reality problem we encounter a certain discrepancy. On one hand, there is accuracy of mathematical methods by which a specific problem is described and, on the other hand, there is a very complicated reality extorting a range of simplifications and the consequent inaccuracy, infidelity of the model arising from them.

Let U be a set we call universe. Let X be a variable which takes values from set U. Further, let real number N be allocated to every element $u\in U$ where $N(u)\in[0,1]$. Number N(u) indicates the possibility degree that variable X takes just value u. In the theory of FSs, FS on universe U is defined by membership function (MF) $\mu(x)$.

If $\mu_N(x)=0$ then x does not belong to FS N, if $\mu_N(x)=1$ then x belongs to FS N, if $\mu_N(x) \in [0,1]$ then x partially belongs to FS N, in other words it is not possible to certainly identify if X belongs to FS N [45,46].

The characteristic of a natural language given by the use of linguistic description of relations among parameters is characterized by the vagueness and uncertainty of semantics. There are several approaches solving this problem [3,9,20,42] and one of them is FSs.

A linguistic description uses a linguistic variable. It differs from a numerical variable where its values are not numbers but words or sentences in a natural or artificial language. In the field of the artificial intelligence (machine intelligence) there are various ways to represent the knowledge. Perhaps the most common way to represent the human knowledge is to form it into the natural language expressions of the type:

IF premise (antecedent)
THEN conclusion (consequent). (6)

This expression is referred to as IF-THEN rule-based form [34]. It typically expresses an inference such that we know a fact (premise, hypothesis, antecedent), then we can infer or derive another fact called a conclusion (consequent). This form of knowledge representation is approximate in the context of linguistics because it expresses human empirical and heuristic knowledge in our own language of communication.

The fuzzy rule-based system is useful in modelling some complex systems that can be observed by humans because it makes a use of linguistic variables as its antecedents and consequents. Linguistic variables can be naturally represented by FSs and logical connectivity of these sets.

In general, we can use several techniques for the decomposition of linguistic rules with multiple antecedents, it means conjunctive or disjunctive, based on (6). Most rule-based systems involve more than a rule. The process of obtaining the overall conclusions (consequents) from the individual conclusions contributed by each rules is known as an aggregation of rules and we use two typically aggregation strategy, conjunctive or disjunctive, too.

A typical if-then rule in a rule-based system is used to determine whether an antecedent (cause or action) infers a consequent (effect or reaction). We can use four mathematical procedures (methods) to conduct the inference of IF-THEN (fuzzy) rules for fuzzy systems based on linguistic rules [23,28,33]: Mamdani (or Mamdani and Asilian), Larsen, Sugeno (or Takagi, Sugeno and Kang), and Tsukamoto. These fuzzy inference methods have had several variations (for example, see more in [23] and [33]).

Through a fuzzy inference engine the rules contained in the rule base define the connection between the input and output fuzzy variables, in other words, the fuzzy inference is the process which uses FSs for formulating the mapping from a given input to an output. The general scheme of FIS involves inputs, fuzzification process, input and output MFs, base rules design, fuzzy logic operators, implication and aggregation, defuzzification and output. FIS have been successfully applied in the fields such as the modelling of classification and prediction systems in areas of municipal creditworthiness, automatic control, decision analysis, data analysis, decision systems, expert system etc. [23,33]. Because of its popularity and multi-disciplinary nature, FISs and its modifications are associated with a number of names, such as fuzzy logic neural networks, fuzzy-rule-based systems, fuzzy-genetics algorithms, fuzzy expert systems, rough-fuzzy modelling, rough-fuzzy hybridization, fuzzy logic controllers etc.

The fuzzy rule-base and MFs play an important role in analyzing and synthesizing FIS. MFs characterise the fuzziness in a fuzzy set. There are possibly more ways to assign the membership values or functions to fuzzy variables. For example, we can work with methods [33]: intuition, inference, rank ordering, neural networks, genetic algorithms and inductive reasoning. A fuzzy rule-base connects outputs with given inputs. We can present some automated methods for a proposal of fuzzy rule-base. These methods develop a rule-base or use a predetermined rule-base of the system such as the batch least squares, recursive least square, gradient, clustering, learning from example and modified learning from example (see more detail in [33]).

2.3 Rough-Fuzzy Approach

The theory FSs and RST are now attracting attention among researchers due to the representation of the knowledge processing. These two theories complement each other and as such they constitute important components of CI. There are various extensions of these two theories for processing.

The developments of rough and fuzzy extensions to the data processing make the hybrid approaches potentially rewarding research opportunities as well. A rough-fuzzy approach [25] has two main lines of thought in a hybridization of fuzzy and rough sets, the constructive approach and the axiomatic approach. The first one, generalized LA and UA are defined based on fuzzy relations that are called fuzzy-rough sets. The second [25] approach introduces the definitions for generalized fuzzy LA and fuzzy UA operators determined by a residual. The assumptions are found that allow a given fuzzy set-theoretic operator to represent LA or UA from a fuzzy relation. Different types of fuzzy relations produce different classes of fuzzy-rough set algebras. In addition to the previous approaches to

hybridization, other generalizations are possible (see more in [25]).

For example, in [16] a hybrid scheme that combines the advantages of fuzzy sets and rough sets in conjunction with statistical feature extraction techniques is introduced. The rough sets approach for generation of all reducts that contain minimal number of attributes and rules is introduced. FSs are applied to the fuzzy pre-processing of input data. In [34] a concept of fuzzy discretization of feature space for a rough set theoretic classifier is explained. The fuzzy discretization is characterised by a membership value, group number and affinity corresponding to an attribute value, in contrast to the crisp discretization which is characterised only by the group number. The merit of this approach over the crisp discretization in terms of classification accuracy, is demonstrated experimentally when overlapping data sets are used as an input to a rough set classifier. The generation [39] of effective feature pattern-based classification rules is essential to the development of any intelligent classifier which is readily comprehensible to the user. It means that an approach integrates a potentially powerful fuzzy rule induction algorithm with a rough set-assisted feature reduction method. In [29] the rough-fuzzy approach is used in case-based reasoning for generating cases, the linguistic representation of patterns is used to obtain a fuzzy granulation of feature space. RST is used to generate dependency rules corresponding to the information regions in the granulated feature space. The fuzzy MF corresponding to the informative regions are stored in cases.

3 Model Definition and Verification

On the basis of achieved classification results it seems to be effective and up-to-date to tackle the classification problem using a hybrid approach combining rough sets and FSs. The application of the classification methods based on CI represents an effective tool for the classification model implementation [5,9]. For example, we can speak about a probabilistic rough classifiers [8], a fuzzy classifiers [20,23] etc. The probabilistic rough classifier combines all positive aspects of rule induction systems with the flexibility of statistical techniques for classification.

Two natural approaches [23] to design a classifier are: to ask an expert how they solve the problem and try to encapsulate the knowledge in a fuzzy-base classifier; to collect input-output data (i.e. a labelled data set) and extract the classifier parameters from the data. The first model is said to be transparent (is interpretable in the domain context) and the second based on data may or may not be interpretable. Fuzzy classifier models are deemed to be able to integrate both approaches: expert (human) and data sources.

3.1 Definition of RFC model

Our case deals with a hybrid RFC [17,18] model. The kernel of our model is given by the following structure (see Fig.1). A whole range of scientific papers was dealing with the rules generation from analysed data and a lot of various methods and procedures using CI [14,22,37,41]. It means that RST were used to define IF-THEN rules (conditioned rules) and FSs were applied as Mamdani FIS.

The set of all deployment decision can be described approximately as noted below for the rule X_m:

$$\begin{aligned} &\text{IF } a_1 \text{ is } v_{j1} \text{ AND } a_2 \text{ is } v_{j2} \text{ AND } \dots \\ &\qquad \dots \text{ AND } (a_n \text{ is } v_{jn}) \text{ THEN } (d \text{ is } h_r), \end{aligned} \tag{7}$$

where: a_1,a_2 and a_n are attributes; v_{j1},v_{j2} and v_{jn} are values of attributes; d is decision and h_r is value of decision (see the Table 1).

In this case LEM1 algorithm was modified and used for a creation of rules set. For conciseness, this algorithm is summarised in a pseudo code (Fig.2).

We have implemented the algorithm as a graphic user interface (GUI) “RSTbox” [17,18,19], it means as environment functioning for automatic rules generation (see Fig.3). This tool is further applied to verify the proposed algorithms for partial calculations with real data. The input of GUI is a knowledge representation system IS=(U,A,V_a) that is formatted as a tabulator delimited table.

GUI output is s sorted reduced decision table with computed approximations. Every row in the table is associated with one decision rule. The decision table may include inconsistent rules.

The problem of the classification in our model consists of three phases. The first is the pre-processing of real data that have been pre-processed and modified into a suitable format. The histograms were created for them from which linguistic variables were derived. The whole data set was divided pursuant to “hold-out” [15] method into training and testing sets.

The second phase is the classification divided into RSTbox rules generation and FIS optimization, and the third is the output and classes interpretation as we can see in the following Fig.4.

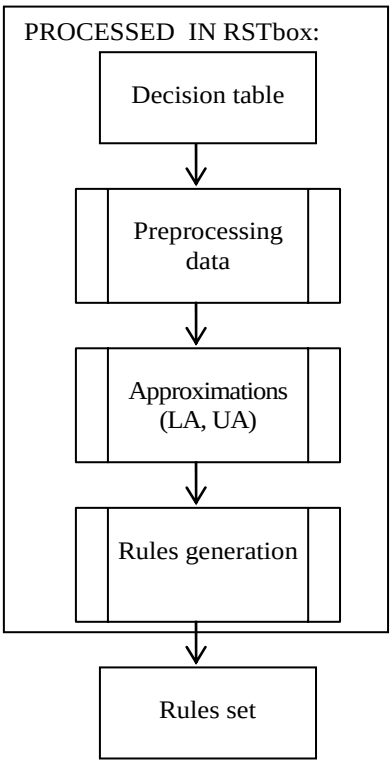

Fig.1 Structure of rules generation

% **algorithm procedure**
% **input**: IS as a decision table T = (U,A,D,f)
% where U= $x_1,x_2,\ldots,x_m$, A= $a_1,a_2,\ldots,a_n$,
% D= $h_1,h_2,\ldots h_q$, f is an information function
% **output**: NO Rules – set of IF-THEN rules for T;
begin
Create matrix S ,size m × (n+1), from table T,
S={$s_1,s_2,\ldots,s_{m*(n+1)}$}
if some object s_x = Ø **then** % x=1,2,…, m * (n+1)
for every object s_x **do** replace s_x by -1
if some vector X=[$x_{1,\ldots,}x_i$] contain -1
then delete x_i % i=1,2,…,m
end {if}
end {for}
end {if}
for reduced table T **do** compute IND(A)
% IND(A) is indiscernibility relations
if IND(A) contain redundant values
then delete redundant values
end {for}
for T, IND(A) compute lower approximation $\underline{A}(X)$
if $x_i \in \underline{A}(X)$
then create rule and insert it to NO Rules
end {if}
end {for}
end {algorithm}

Fig.2 Pseudo code of algorithm procedure

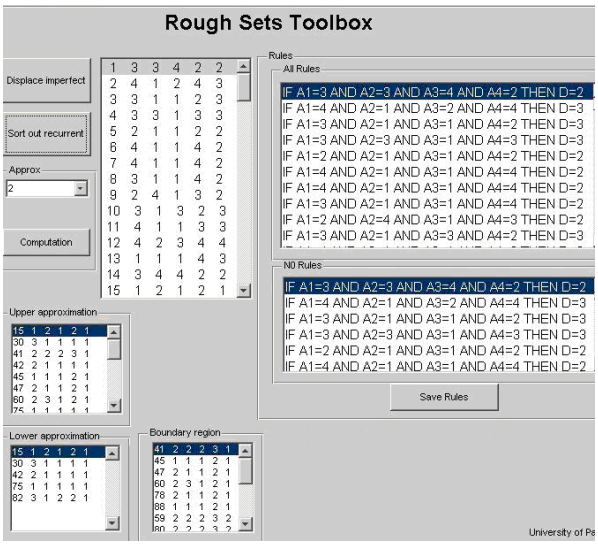

Fig.3 GUI of the RSTbox

The learned knowledge is presented in the form of a set of decision rules that can easily be explained and understood by users. Rough sets approach is applied in RSTbox for the generation of minimal fuzzy rule base for FIS in RFC_i= ={$RFC_1,RFC_2,\ldots,RFC_k$}. These sets of RFC_i use a various type of input MFs. The shape of these MFs is optimized by a real data histogram, and particular rules stresses adjustments were made. The RFC_i can be described as MISO system where the inputs are the attributes of a real data set, and the output is the decision about the classification.

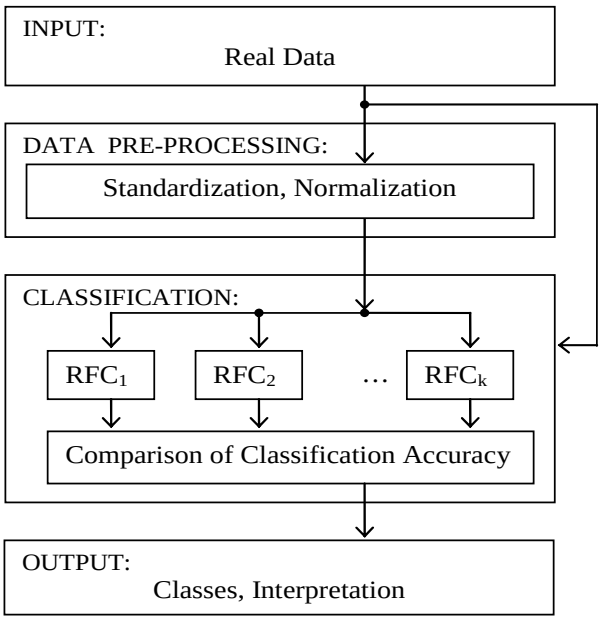

Fig.4 Model of rough-fuzzy classification

Finally, the accuracy of RFC_i classification is used for the best of them to be chosen. The models of the system were created and tested in

MATLAB\Simulink and the results were collectively evaluated.

The goal of the selected data experiments is to verify the correctness of the proposed RFC procedure (see Fig.5), to reach the high testing data classification accuracy even in comparison with the algorithms hitherto known. The real data set represents the input of our procedure. These pre-processed data are then used for the histograms computation. On their basis FIS and RSTbox, in which the pre-processed data are utilized, are subsequently modified. The classes are the outputs of the whole procedure.

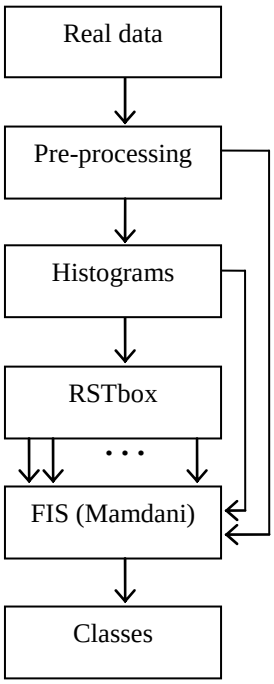

Fig.5 Rough-fuzzy classifier model

It is supposed that a classifier is a unit (algorithm, model) executing a classification, a classifier input is a set of attributes and a classifier output is a class allocation. The stated RFC model is based on the following assumptions:

- Let's specify a set by attributes $A=\{a_1, a_2, ..., a_n\}$ and $A_r=[v_{r1}, v_{r2}, ..., v_{rn}]$ is *n*-dimensional vector of attributes values where $r = 1,2,\ldots,m$.
- Let's suppose the classification into *R* classes be called $h_1, h_2, \ldots, h_R$. Let's mark *N*- dimensional attributes space by II.
- A class indicator $d\in\{h_1, h_2, ..., h_n\}$ is assigned to every $A_r\in\prod$. Function $d=f(A_r)$ is the rule defining this assignment.

A very important step is the explication and evaluation of the results obtained. It is not only that they are new and interesting but it is also the numeric parameters values applied on a selected model that is the thing. We speak about accuracy measure and error measure.

3.2 Verification of RFC model

The most important goal [23] in designing a classifier is to achieve the highest possible classification accuracy or the lowest possible error rate.

The classification accuracy is the ratio of correctly classified objects to the total amount of objects *x* in a set, expressed in percent (here denoted P_x). The parameter of the total classification error of a classifier model, obtained as the difference $100\text{-}P_x$, is frequently used, as well. The next well known numeric parameter is the resubstitution error [42] which is obtained as the ratio of correctly classified objects to the total amount of training data objects in a set, expressed in percent.

The methods used for the classifier accuracy evaluation according to [15,42] are:

- testing on the whole training data;
- k-fold validation;
- leave-one-out;
- testing on the testing data by „holdout" method;
- bootstrap.

In our experiments we used the testing on the whole training data, which is based on using one data set for both training and testing. This method is applicable, however, it bears the highest threat of overfitting, decreasing the testifying parameter abilities, and it is affected by the resubstitution error.

The other used method was the "holdout". The holdout method means, firstly, an accidental data division into two independent sets – training and testing. The usual division proportion is from 2/3 to 1/3 up to 4/5 to 1/5. The training set then serves for the model (classifier) creation and derivation, and the testing set for the classification accuracy determination. This method gives a more pessimistic P_x.

Once more, the goal of the selected data experiments is to verify the correctness of the proposed RFC procedure (see Fig.4), to reach a high testing data classification accuracy even in comparison with the algorithms hitherto known [14,41].

For the first part of the experiments IRIS-called data were used [1]. This is often cited and maybe the best known database to be found in the pattern recognition area. The database contains 150 records of iris flowers size measurements. The length and width of sepal and petal were measured. Three kinds

of iris were investigated - setosa, versicolor and virginica, where each iris plant refers to a class. One class is linearly separable from the other two, the latter being not linearly separable from each other. The second series of experiments was carried out with „WINE"- called database (wine recognition data) [1]. These data came into existence as the results of a chemical analysis of Italian-region-grown wines of three different kinds and they contain chemical elements values from 178 samples altogether. The analysis determined the quantities of 13 constituents found in each of the three types of wines. All attributes are continuous.

The experiments run according to Fig.4 and Fig.5. Firstly, the data were pre-processed and converted in a suitable format. Consequently, the histograms for the pre-processed data have been calculated. In the Fig.6 is the histogram of petal width parameter of iris plant (setosa, vesricolor and virginica) for IRIS data.

To proceed in „Holdout", the IRIS data were then divided into training (120 objects) and testing (30 objects). The second part proceeded concurrently with the whole data set. By using various MFs shapes (symmetric, non-symmetric) and MFs types: triangular (trimfbyexpert.fis), bell-shaped (gbellmfbyexpert.fis) and bell-shaped-gauss (gaussmfbyexpert.fis) the outputs were compared. A 30- and 150-object set was used for testing. The systems created in this way then were tested in Simulink-created models and the results collectively evaluated (see Fig.7).

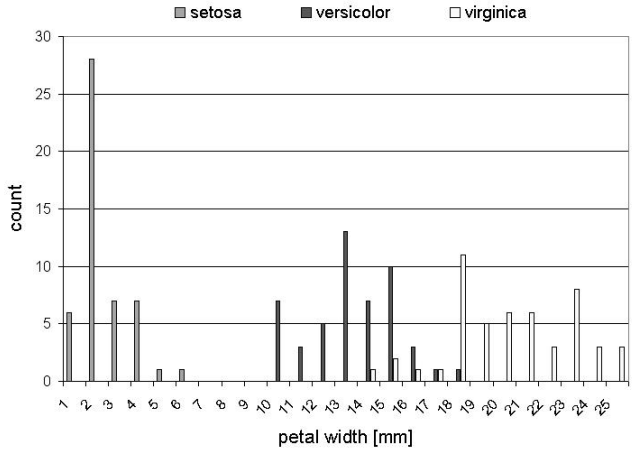

Fig.6 Histogram of PW parameter for IRIS data

Such MF (non-symmetric, triangular) for the petal-width (PW) parameter is presented in Fig.8.

We can see the example of the notation of FIS type Mamdani with non-symmetric triangular MF that MF based on the histogram in Fig.9.

In the Fig.10 are objects (for the first IRIS database 150 objects) on the axis x. They were ordered according to classes during the experiment (due to plasticity). Axis y represents classes 1, 2 and 3. Objects from interval (0,50] belong to class 1, objects from interval (50,100] belong to class 2 and objects from interval (100,150] belong to class 3. Objects outside from these interval represents incorrect classification.

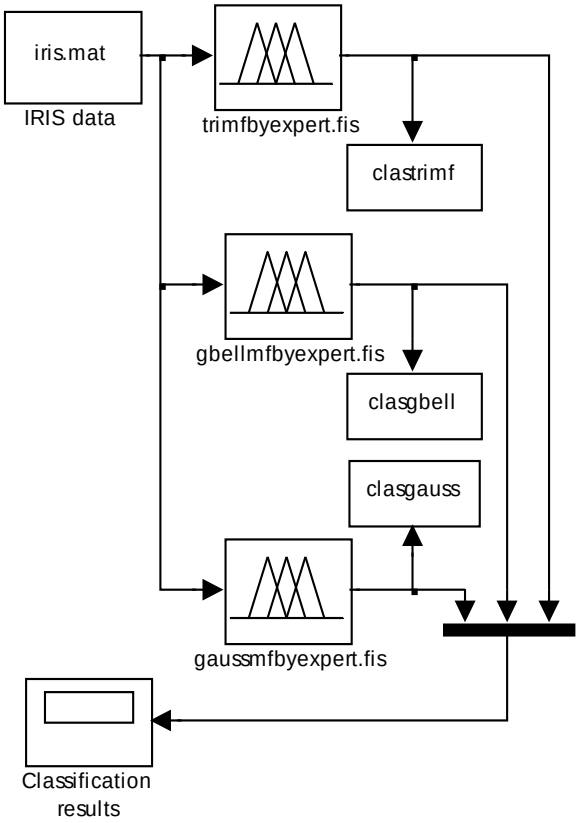

Fig.7 Simulation model of RFC

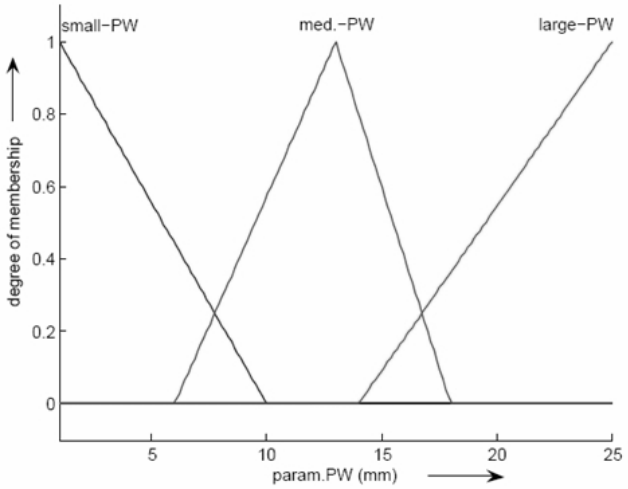

Fig.8 MFs of PW parameters for IRIS data

In the Fig.11 are objects (IRIS database testing data) on the axis x. They were ordered according to classes during the experiment. Axis y represents classes 1, 2 and 3. Objects from interval (0,10] belong to class 1, objects from interval (10,18] belong to class 2 and objects from interval (18,30] belong to class 3. Objects outside from these intervals represent, like in previous figure, incorrect classification.

```
[System]
Name='trimfbyexpert'
Type='mamdani'
…
[Rules]
1 2 1 1, 1 (0.085271) : 1
1 3 1 1, 1 (0.24806) : 1
1 1 1 1, 1 (0.00775) : 1
1 1 2 2, 2 (0.03876) : 1
1 1 2 3, 3 (0.007752) : 1
3 3 3 3, 3 (0.0386) : 1
2 3 1 1, 1 (0.03876) : 1
2 1 2 2, 2 (0.093) : 1
2 2 3 3, 3 (0.0155) : 1
2 3 2 3, 2 (0.007752) : 1
3 3 2 2, 2 (0.10078) : 1
3 2 2 3, 3 (0.0155) : 1
3 1 2 2, 2 (0.023256) : 1
3 2 3 2, 3 (0.0155) : 1
3 1 2 3, 3 (0.007752) : 1
1 3 1 2, 1 (0.0235) : 1
```

Fig.9 Part of Mamdani FIS algorithm "trimfbyexpert.fis"

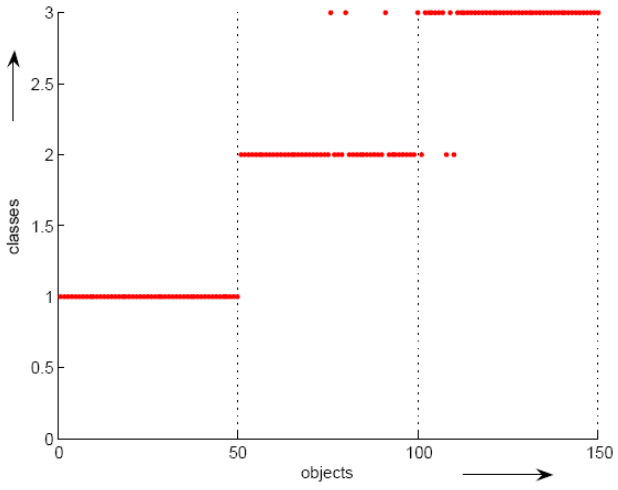

Fig.10 Graphical output for FIS „trimfbyexpert.fis" (data $IRIS_{150}$)

The second series of experiments was carried out with „WINE"-called database (wine recognition data) [1]. These data came into existence as chemical analysis results of Italian-region-grown wines of three different kinds and they contain chemical elements values from 178 samples altogether. Using „hold-out" method the data have been divided into training objects and testing objects. In the Fig.12 are objects (WINE database 178 objects) on the axis x. They were ordered according to classes during the experiment. Axis y represents classes 1, 2 and 3. Objects from interval (0,59] belong to class 1, objects from interval (59,130] belong to class 2 and objects from interval (130,178] belong to class 3. Objects outside from these intervals represent, like in previous figures, incorrect classification.

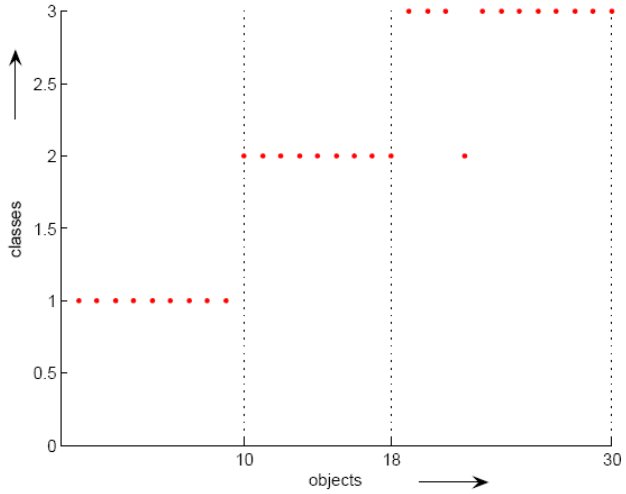

Fig.11 Graphical output for FIS „trimfbyexpert.fis" (testing data $IRIS_{30}$)

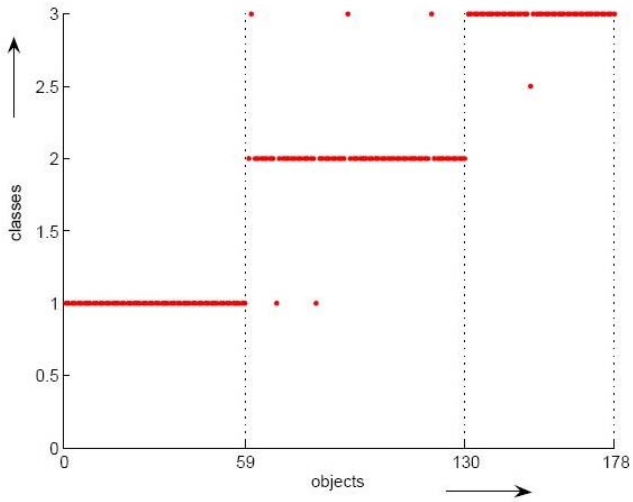

Fig.12 Graphical output for FIS „trimfbyexpert.fis" (data $WINE_{178}$)

In the last Fig.13 are objects (Wine database testing data) on the axis x. They were ordered according to classes during the experiment. Axis y represents classes 1, 2 and 3. Objects from interval (0,18] belong to class 1, objects from interval (18,34] belong to class 2 and objects from interval

(34,41] belong to class 3. Objects outside these intervals represent incorrect classification, too.

The outputs are demonstrated in following tables (Table 2 to 7).

The same procedure was used for the wine data, the data set in this case was divided into training (138 objects) and testing (40 objects).

The resulting classification accuracy [15] denoted P_x is the ratio of correctly classified objects to the total amount of objects x in a set, expressed in percent, as we can see in Table 2 and 5.

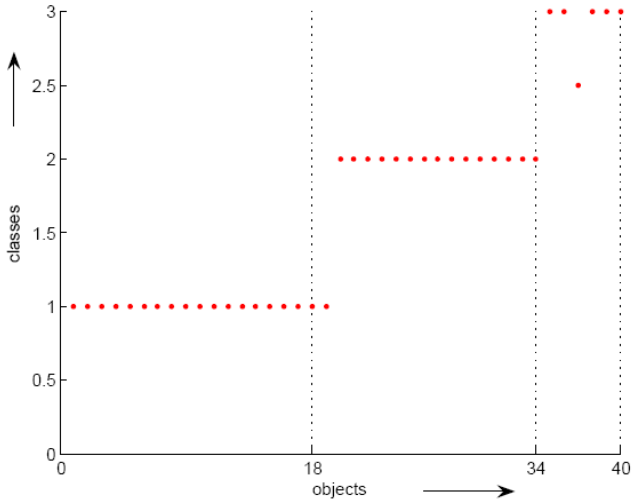

Fig.13 Graphical output for system „trimfbyexpert“, testing data WINE40

Resulting classification accuracy P_x (P_{IRIS}, P_{WINE}) is the ratio of correctly classified objects to the total amount of objects x in a set, expressed in percent, how we can see in Table 2, 3 and 4 where T-fis is trimfbyexpert.fis, G-fis is gaussmfbyexpert.fis, B-fis is bellmfbyexpert.fis; 150 and 30 are numbers of element into testing sets.

Table 2 The best results for $test_{150}$ and $test_{30}$ datasets

	T-fis ($test_{150}$) „optimistic“	T-fis ($test_{30}$) „pessimistic“
P_{IRIS}	95,33%	93,33%

Table 3 Comparison of classification accuracy for different MFs

$Test_{150}$	Symetric MF		
	T-fis	G-fis	B-fis
P_{IRIS}	73,33%	71,33%	79,33%

The second series of experiments was carried out with „WINE“-called database (wine recognition data) [1]. These data came into existence as chemical analysis results of Italian-region-grown wines of three different kinds and they contain chemical elements values from 178 samples altogether. Using „hold-out“ method the data have been divided into training (138 objects) and testing (178 objects) with the whole data set being concurrently operated in one part of the experiment and the training and testing data in the second. The procedure corresponded to „IRIS“ database experiment. The outputs are demonstrated in Table 5, 6 and 7.

Table 4 Comparison of classification accuracy for different MFs

$Test_{150}$	Modified MF		
	T-fis	G-fis	B-fis
P_{IRIS}	95,33%	90%	87,33%

Table 5 Best results for $test_{178}$ and $test_{40}$ datasets

	T-fis ($test_{178}$) „optimistic“	T-fis ($test_{40}$) „pessimistic“
P_{WINE}	96,6%	95%

Table 6 Comparison of classification accuracy for different MFs

$Test_{178}$	Symetric MF		
	T-fis	G-fis	B-fis
P_{WINE}	87,6%	83,2%	78,6%

Table 7 Comparison of classification accuracy for different MFs

$Test_{178}$	Modified MF		
	T-fis	G-fis	B-fis
P_{WINE}	96,6%	84,2%	83,6%

The classification results have been compared with methods published in [14,41], as we can see in Table 8 and 9.

Table 8 Comparison of classification accuracy for various methods

	Other methods [14,41]				
	ID3	Hong-and-Chen's	C5rules	EFUNN	PRISM
P_{IRIS}	90,7%	96,67%	92%	96%	90%
P_{IRIS}	Presented approach rough-fuzzy				
	95,33% (optimistic) / 93,33% (pessimistic)				

Table 9 Comparison of classification accuracy for various methods

	Other methods [14]			
	LDA	C5rules	1NN	kNN, Euclidean, k=1
P_{WINE}	98,9%	92,1%	96,1%	95,5%
	Presented approach rough-fuzzy			
P_{WINE}	96,6% (optimistic) / 95% (pessimistic)			

4 Conclusion

This paper dealt with the problem of data classification. When describing real systems it is possible to express their description by using a natural language. This description is uncertainty-loaded. To operate with uncertainty it is suitable to use RST and FSs theories, or possibly their combinations. For this reason the introduction section summarizes the basic ideas of the presented theories. The following sub-chapter is devoted to the rules base generation on the basis of RST by the means of a modified algorithm which results from LEM1 algorithm. This algorithm became the basis for RSTbox which was used for proposing the RFC. The following experiments referred to the problem of data classification using a hybrid approach.

For the experiments the data (IRIS, WINE) from [1], which are generally known and used as benchmark data, were used. The experiments verified the proposed model for the data classification, and the results were compared with other available classification methods presented in [14,41], and were applied to the same data.

The presented RFC turned out to appear suitable. The classification accuracy for IRIS data reached 93,33% (see the Table 2). The proposed procedure allowed to reach 95% (see the Table 5) of the classification accuracy for WINE data. On the basis of the above stated facts it can be claimed that the proposed RFC model is functional, relatively successful compared with other methods, and can be used to carry out various databases classification.

The areas where future investigations will be directed can be divided into two groups. First, it will be the investigation of theoretical context and the possibilities to use the proposed procedures. Secondly, it is further data analysis tools development (expansion of the implemented RST box).

This article showed that data classification hybrid approach combining RST and FSs is suitable and offers, compared to other approaches, very good results. Further investigations could focus on one of the proposed model part, namely rules generating. It is possible to assume that the number of conditioned rules will be possible to reduce when keeping input values accuracy. Thereby, model calculation demandingness would decrease to a considerable extent.

Further research can also focus on the area of using an algorithm "Missing Not at Random" for a rules generation. In the field of RFC it is feasible to proceed from the supervised learning technique to the combined approach by using pre-processed data in the first phase, e.g. "a self-organization map".

In the field of SW tool development (RSTbox) several properties that could be improved and extended in this toolbox present themselves. That is, e.g. the expansion of possibilities to import and export data used in various formats. Further, the generated rules could be directly imported into the conditioned rules base. Also, further tool optimization (applied algorithms) to reduce time and HW demandingness for operating with very large databases would be suitable. Similarly, graphic user interface would be marked with changes.

5 Acknowledgement

The paper is supported by the National Science Foundation of the Czech Republic, Grant No. 402/08/0849.

References:

[1] Asuncion, A., Newman, D.J., *UCI Repository Of Machine Learning Databases and Domain Theories* [online], Irwine, USA. Accessible from URL: http://www.ics.uci.edu/ MLRepository. html> [cit.2007-06-03].

[2] Bezdek, J.C., What is Computational Intelligence?, *Computational Intelligence: Imitating Live*. Piscataway: IEEE Press, 1994, pp. 1-12.

[3] Boixoixader, D., Jacas, J., Recasens, J., Upper and Lower Approximations of Fuzzy Sets, *International Journal of General System*, Vol.29, pp. 555-568.

[4] Brown, D. G., Classification and Boundary Vagueness in Mapping Resettlement Forest Types, *International Journal of Geographical Information Science*, Vol.12, 1998, pp. 105-129.

[5] Cpalka, K., Rutkowski, L., A New Method for Compexity Reduction of Neuro-Fuzzy

Systems, *WSEAS Transaction on Systems*, Vol.5, No.11, 2006, pp. 2514-2521.

[6] Dubois, D., Prade, H., *Fuzzy Information Engineering and Soft Computing: A guided tour of Applications*, New York: John Wiley & Sons, 1997.

[7] Düntsch, I., Gediga, G., *Rough Set Data Analysis - A Road to Non-invasive Knowledge Discovery*, Angor: Methodos, 2000.

[8] Gorzalczany, M. B., Piasta, Z., Neuro-Fuzzy Approach versus Rough-Set Inspired Methodology for Intelligent Decision Support, *Information Sciences*, Vol.120, No.1-4, pp. 45-68.

[9] Greco, S., Matarazzo, B., Slowinski, R., The Use of Rough Sets and Fuzzy Sets in MCDM, *Multicriteria decision making: Advances in MCDM models, algorithms, theory, and applications*. Dordrech: Kluwer Academic Publisher, 1999, pp. 14-59.

[10] Grzymała-Busse, J. W., Goodwin, L. K., Predicting Preterm Birth Risk Using Machine Learning from Data with Missing Values, *Bulletin of International Rough Set Society*, 1997, pp. 17-21.

[11] Grzymała-Busse, J. W., Rough Set Strategies to Data with Missing Attribute Values, *Proc. of the Wworkshop on Foundations and New Directions in Data Mining*. Melbourne, 19-22 November 2003, pp. 56-63.

[12] Grzymała-Busse, J. W, Siddhaye, S., Rough Set Approaches to Rule Induction from Incomplete Data, *Proc. of the IPMU2004*. Perugia, Italy, 4-9 July 2004, Vol.2, pp. 923-930.

[13] Grzymała-Busse, J. W., Wang, A. Z., Modified Algorithms LEM1 and LEM2 for Rule Induction from Data with Missing Attribute Values, *Proc. of the fifth International Workshop on Rough Sets and Soft Computing*, Research Triangle Park, NC, 2–5 March 1997, pp. 69-72.

[14] Guo, G. etc., Similarity-Based Data Reduction Techniques, *Journal of Research and Practice in Information Technology*. Vol.2, No.37, Australian Computer Society, 2005, pp. 211-232.

[15] Han, J., Kamber, M., *Data mining*, San Francisco: Elsevier, 2nd edition, 2006.

[16] Hassanien, A. E., Fuzzy Rough Sets Hybrid Scheme for Breast Cancer Detection, *Image and Vision Computing*, Vol.25, No.2, 2007, pp. 172-183.

[17] Jirava, P., *Information System Analysis Based on Rough Sets*, Theses of the Dissertation. Pardubice: University of Pardubice, 2007.

[18] Jirava, P., Křupka, J., Generation of Decision Rules from Nondeterministic Decision Table based on Rough Sets Theory, *Proc. of the 4th International Conference on Information Systems and Technology Management*, Sao Paolo, Brasil, 2007, pp. 566-573.

[19] Jirava, P., Křupka, J., Classification Model based on Rough and Fuzzy Sets Theory, *Proc. of the 6th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetic*. WSEAS Press, 2007.

[20] Křupka, J., Kašparová, M., Modelling of Internal Human Population Migration Classifiers by Fuzzy Inference System and Its Hierarchical Sstructure, *WSEAS Transaction on Systems*, Vol.6, No.3, 2007, pp. 461-467.

[21] Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A., Rough sets: A tutorial. In: S.K. Pal and A. Skowron (Eds.), *Rough-Fuzzy Hybridization: A New Trend in Decision-Making*. Singapur: Springer-Verlag. 1998, pp. 3-98.

[22] Kudo, Y., Murai, T., A Method of Generating Decision Rules in Object Oriented Rough Set Models, *Rough Sets and Current Trends in Computing*, Kobe, Japan, 6-8 October 2006.

[23] Kuncheva, L. I., *Fuzzy Classifier Design*, New York: Physica-Verlag, 2000.

[24] Lingras, P., Yao, Y. Y., Time Complexity of Rough Clustering: Gas versus k-Means, *Rough Sets and Current Trends in Computing 2475*, 2002.

[25] Lingras, P., Jensen, R., Survey of Rough and Fuzzy Hybridization, *Fuzzy Systems Conference*, 23-26 July, 2007, pp. 1–6.

[26] Little, R. J. A., Rubin, D. B., *Statistical Analysis with Missing Data*, 2nd edition, New York: Wiley, 2005.

[27] Montazer, Gh., Sabzevari, R., Pour Khatir, H. Gh., Intelligent Parameter Reduction Using Rough Sets Theory and Sensitivity Analysis, *WSEAS Transaction on Systems*, Vol.6, No.3, 2007, pp. 623-630.

[28] Nguen, H. T. etc., *First Course in Fuzzy and Neural Control*. Boca Raton: Chapman and Hall/CRC, 2003.

[29] Pal, S. K., Soft Data Mining, Computational Theory of Perceptions, and Rough-Fuzzy Approach, *Information Science*, Vol.163, 2004, pp. 5-12.

[30] Pawlak, Z., Rough Sets, *Int. J. of Information and Computer Sciences*, Vol.11, No.5, 1982, pp. 341-356,

[31] Pawlak, Z., A Primer on Rough Sets: A New Approach to Drawing Conclusions from Data, *Cardozo Law Review*, Vol.22, No.5-6, 2001, pp. 1407-1415.

[32] Polkowski, L., *Rough Sets, Mathematical Foundations, Advances in Soft Computing*. Physica-Verlag, 2002.

[33] Ross, T. J., *Fuzzy Logic with Engineering applications*, 2nd edition, West Sussex: Wiley, 2004.

[34] Roy, A., Pal, S. K., Fuzzy Discretization of Feature Space for a Rough Set Classifier, *Pattern Recognition Letters*, Vol.24, No.6, 2003, pp. 895-902.

[35] Rubin, D. B., Inference and Missing Data. *Biometrika*, Vol.63, No.3, 1976, pp. 581-592.

[36] Rubin, D. B., Little, R. J., *Statistical Analysis with Missing Data*. New York: Willey, 1987.

[37] Sakai, H., Nakata, M., On Possible Rules and Apriori Algorithm in Non-deterministic Information Systems. *Rough Sets and Current Trends in Computing*, Kobe, Japan, 6-8 October 2006.

[38] Sarkar, M., Fuzzy-Rough Nearest Neighbor Algorithms in Classification. *Fuzzy Sets and Systems Archive*, Vol.158, No.19, 2007, pp. 2134-2152.

[39] Shen, Q., Chouchoulas, A., A Rough-Fuzzy Approach for Generating Classification Rules, *Pattern Recognition*, Vol.35, 2002, pp. 2245-2438.

[40] Stefanowski, J., Słowiński, R., Rough Set Reasoning about Uncertain Data. *Fundamenta Informaticae 27*, IOS Press, 1996, pp. 229-243.

[41] Stefanowski, J., On Rough Set Based Approaches to Induction of Decision Rules. *Rough Sets in Knowledge Discovery*. Vol.1, Heidelberg: Physica Verlag, 1998, pp. 500-529.

[42] Witten, I. H., Eibe, F., *Data Mining: Practical Machine Learning Tools and Techniques.* 2nd edition, San Francisco: Morgan Kaufmann, 2005.

[43] Paharia, A. K., Bhawsar, Y., Singh, D., Data Mining: As an Imperative Tool for Discovering Knowledge, *Proc. of the 6th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetic*, WSEAS Press, 2007, pp. 376-379.

[44] Zadeh, L. A., Fuzzy Sets. *Information and Control*, Vol.8, 1965, pp. 338-353.

[45] Zadeh, L. A., Outline of a New Approach to the Analysis of Complex Systems and Decission Processes. *IEEE Trans .S.M.C.*, Vol.3, 1973, pp. 28-44.

[46] Zadeh, L. A., The Roles of Fuzzy Logic and Soft Computing in the Conception, Design and Deployment of Intelligent Systems. *Software Agents and Soft Computing*, 1997, pp. 183-190.