Classification Model based on Rough and Fuzzy Sets Theory

JIRAVA PAVEL, KŘUPKA JIŘÍ
Institute of System Engineering and Informatics
Faculty of Economics and Administration, University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC

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. 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. It even allows operating with uncertainty data. This model is carried out in MATLAB, and 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

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 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 [2,22]. 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) [14,16,17] is based on the research of information system logical properties, and uncertainty in it is expressed by a boundary region. Every investigated object is connected 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.

Our case deals with a hybrid 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 modeling of municipal creditworthiness, automatic control, decision analysis, data analysis, decision systems or expert system [4,6].

Goals of this paper are: to suggest and realize a toolbox for generating conditioned rules and to create and analyse a hybrid RFC data classifier model. The toolbox applies proposed algorithm of rules generation which exploits RST. These rules were used in Mamdaniho 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 [5,14] 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, 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, where $V_a=\{v_{11}, x_{12}, \ldots, v_{m1}, \ldots, v_{mn}\}$, f_a: $U \rightarrow V_a$ is a information function such that $f(x,a) \in V_a$ for each $a \in A$, $x \in U$ [5].

It is possible to express IS as a decision table (see the Table 1) where: a_i is i-th attribute (member of the set A);

x_j is j-th member of the set U, j=1,2,…,m; v_{ji} are attribute values and d = h_r is decision attribute for 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

Real live data set is represented as a table, where every column represents an attribute and each row represents a case, object. For each pair object-attribute there is known descriptor. Descriptor is specific and precise value of attribute. A limited discernibility of objects by means of the attribute values prevents generally their precise classification [20]. In practice input data, presented as decision tables, may have missing attribute and decision values, i.e., decision tables are incompletely specified. Values of attribute may be uncertain because of many reasons. Generally we can define four types of uncertainty: discretization of quantitative attributes; imprecise values of quantitative attribute; multiple values of attribute and unknown or missing values of attribute.

Uncertainty coming from unknown or missing attributes occurs when the value of an attribute is unknown. In practice input data presented as decision tables, may have missing attribute and decision values, i.e., decision tables are incompletely specified. 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 attribute value was useful but currently we have no access to it. RST approach to missing attribute values, when all missing values were lost, was presented in LEM2 algorithms [8,9,10].

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

The approximations are two basic operations in RST [17]. 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 [14]:

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

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 define (1) and let and $B\subseteq A$ and $X\subseteq U$. We can approximate X using only the information contained in B by constructing lower approximation and upper approximation of X on the following way:

$$\underline{B}(X)=\{x\in U\colon B(x)\subseteq X\} \text{ and} \tag{3}$$

$$\overline{B}(X)=\{x\in U\colon B(x)\cap X\neq\emptyset\}. \tag{4}$$

The objects in lower approximation can be with certainly classified as members of X on the basis of knowledge in B and the objects in upper approximation are classified as possible members of X on the basis of knowledge in B. The set

$$BN_B(X)=\overline{B}(X)-\underline{B}(X), \tag{5}$$

is called the boundary region of X and thus consists of those objects that we cannot decisively classify into X on the basis of knowledge B. If the boundary region is empty, then the set X is crisp with respect to B. If the boundary region is not empty, the set X is rough with respect to B. Rough sets are defined by approximations and have properties defined in [14,16,17,18].

The theory of FSs is an approach to uncertainty. In this theory an element belongs to a set according to the membership degree (membership function values) [22,23,24] that is to say in 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. 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.

The effort to maximize accuracy leads to the disproportionate rise of the number of definitions and conditions. In [23] 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 which accuracy and relevance are practically mutually exclusive characteristics.“

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 is defined on universe U is defined by membership function $\mu(x)$ in this theory.

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 [23,24].

Natural language (characteristic by use of linguistic description of relations among parameters) is characterized by vagueness and uncertainty of semantics. There are several approaches solving this problem [3,7,22] and one of them is FIS which uses theory of FSs for formulating the mapping from a given input to an output. General scheme of FIS involves inputs, fuzzification process, input membership functions, base rules design, fuzzy logic/FSs operators, implication and aggregation, defuzzification and output. The mapping then provides a basis from which decisions can be made, or patterns discerned. There are two commonly used types of FIS - Mamdani and Sugeno.

3 Modelling of Rough Fuzzy Classifier

The problem of classification in our model is composed of three phases: first is preprocessing, second is classification dividet into Rough Sets Toolbox rules generation and FIS and third is output and interpretation as we can see on folowing Fig. 1.

The kernel of our model is given by following algorithm. A whole range of scientific papers was dealing with rules generation from analysed data and a lot of various methods and procedures using CI [15,19,21]. The presented procedure uses RST mathematical apparatus for generating IF-THEN rules and the following algorithm was proposed for it (Fig. 2).

The presented algorithm was implemented as a toolbox [13] called Rough Sets Toolbox in MATLAB environment functioning for automatic rules generation. This instrument is further applied for verifying the proposed algorithms for partial calculations with machile learning repository data.

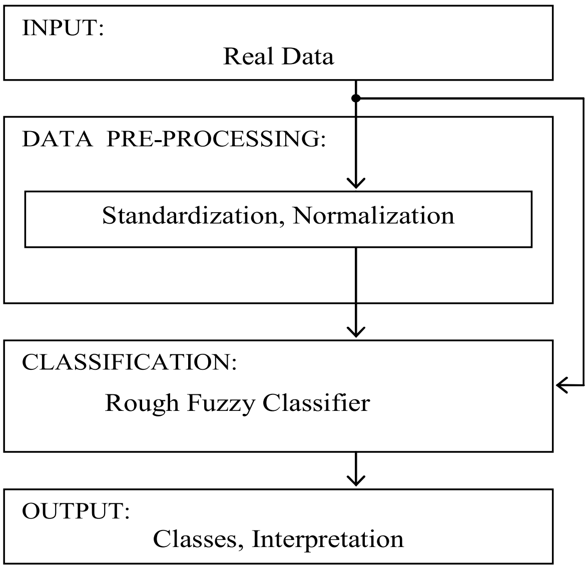

Fig. 1 Model scheme

```
%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 information function
%output: NO Rules – set of if-then rules for T;
begin
Create matrix S ,size m x (n+1), from table T,
S={$s_1,s_2,\ldots,s_{m*(n+1)}$}
   if any object $s_x$ = Ø then //(x=1,2,…, m * (n+1))
         for every object $s_x$ do replace $s_x$ by -1
                if any vector X=[$x_{1,\ldots,}x_i$]contain -1 then
                    //i=1,2,…m
                    delete $x_i$
                end {if}
         end {for}
   end {if}
   for reduced table T do compute I
   // I= indiscernibility relations IND(A)
          if IND(A) contain redundant values then
                   delete redundant values
   end {for}
   for T,I 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 Algorithm procedure

The data have been first pre-processed and modified into a suitable format (MS Excel software is used). Histograms have been created for them from which linguistic variables have been derived. The whole file has been divided pursuant to “hold-out” [12] method into training and testing set. Further, the data have been analysed using the toolbox and conditioned rules have been derived for them. They form the conditioned rules platform in FIS. Then, input and output membership functions in this system have been modified and particular rules stresses adjustments have been made. The systems created in this way then have been tested in Simulink-created models and the results collectively evaluated. The proposed classification procedure can be demonstrated as following:

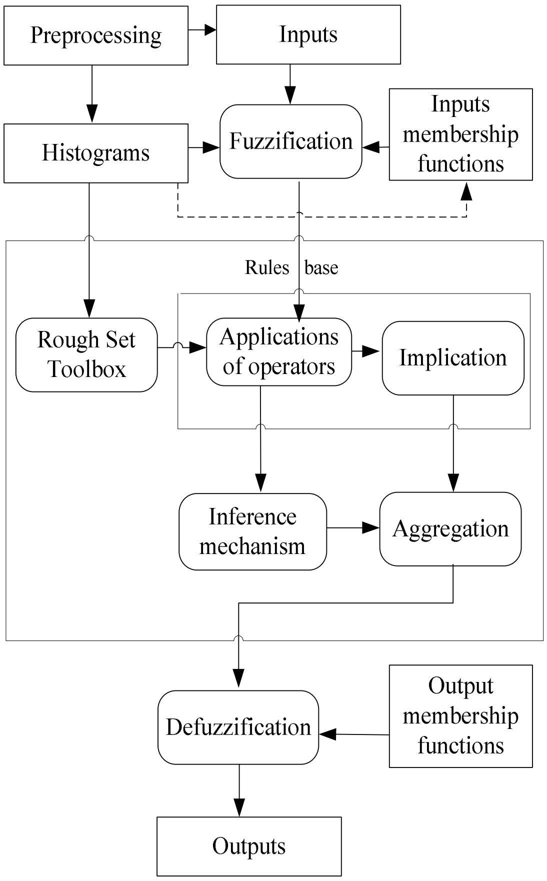

Fig. 3 Rough Fuzzy Classifier model

The goal of the selected data experiments is to verify the correctness of the proposed RFC procedure (see Fig. 3), to reach high testing data classification accuracy even in comparison with the algorithms hitherto known.

For the first part of the experiments IRIS-called data have been used [1]**Error! Reference source not found.** . The database contains 150 records of iris flowers size measurements. The length and width of sepal and petal have been measured. Three kinds of iris have been investigated - setosa, virginica and versicolo. 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 ojects.

Resulting classification accuracy denoted P_x 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 and 3.

Table 2. Resulting classification accuracy for IRIS data (%)

IRIS data		other classification methods [11,21]				
	RFC	C5	ID3	EFUNN	Hong-Chen´s	PRISM
P_{IRIS}	93,33	92,0	90,7	96,0	96,67	90,0

Table 3. Resulting classification accuracy for WINE data (%)

WINE data		other classification methods [11]			
	RFC	C5	LDA	FSM	kNN, k=1
P_{WINE}	95,0	92,1	98,9	96,1	95,5

4 Conclusion

In introduced paper was proposed new model for data classification. The kernel of this model is given by new algorithm. The proposed algorithms were used in RST toolbox which was successfully applied during the experiments.

The experiments proved the correctness of the proposed procedures and rough-fuzzy classifier functionality.

The databases from [1] that are generally known have been chosen for the experiments, and the results of the experiments were possible to be compared with the results of other classification methods applied on these databases. It turned out that the proposed rough-fuzzy classifier seems to be suitable and successful.

For the first database – „IRIS“ data database – classification accuracy 93.33% has been achieved. For „WINE“ data, which was the second applied database, classification accuracy reached by the presented procedure and methods was 95%.

Based on the above stated facts it can be claimed that the proposed rough-fuzzy classification model is functional, it is relatively successful (compared to other approaches) and using it, various databases classification can be carried out.

References:

[1] Asuncion, A., Newman, D.J. *UCI Repository Of Machine Learning Databases and Domain Theories* [online]. Irwine, USA. Accessible from

www:http://www.ics.uci.edu/MLRepository.html> [cit.2007-06-03].

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

[3] Brown, D. G., Classification and boundary vagueness in mapping resettlement forest types. In: *International Journal of Geographical Information Science*, 12/1998, pp.105-129.

[4] Dubois, D., Prade, H. *Fuzzy Information Engineering and Soft Computing: A guided tour of Applications*. New York : John Wiley & Sons, 1997, 712 p.

[5] Düntsch, I., Gediga, G. *Rough Set Data Analysis - A Road to Non-invasive Knowledge Discovery*. Angor: Methodos, 2000, 107 p.

[6] Greco, S., Matarazzo, B., Slowinski, R. The use of rough sets and fuzzy sets in MCDM. In: *Multicriteria decision making: Advances in MCDM models, algorithms, theory, and applications*. Dordrech: Kluwer Academic Publisher. 1999, pp.14-59.

[7] Grzymała-Busse, J.W., Goodwin, L.K. Predicting preterm birth risk using machine learning from data with missing values. In: *Bulletin of International Rough Set Society 1/2*. IRSS. 1997, pp.17-21.

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

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

[10] Grzymała-Busse, J.W., Wang, A.Z. Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. In: *Proc. of the fifth International Workshop on Rough Sets and Soft Computing* (RSSC'97) at the Third Joint Conference on Information Sciences(JCIS'97). Research Triangle Park, NC. 2–5 Maech 1997, pp.69-72.

[11] Guo, G. e.a. Similarity-Based Data Reduction Techniques. In: *Journal of Research and Practice in Information Technology*. No. 37/2. Australian Computer Society. 2005, pp.211-232.

[12] Han, J., Kamber, M. *Data mining*. San Francisco: Elsevier, second edition, 2006.

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

[14] 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.

[15] Kudo, Y., Murai, T. A method of Generating Decision Rules in Object Oriented Rough Set Models. In: *Rough Sets and Current Trends in Computing* RSCTC 2006. Kobe, Japonsko, October 6-8 2006.

[16] Pawlak, Z. Rough sets, In: *Int. J. of Information and Computer Sciences*, 11, 5, 1982, pp.341-356,.

[17] Pawlak, Z. A Primer on Rough Sets: A New Approach to Drawing Conclusions from Data. In: *Cardozo Law Review*, Volume 22, Issue 5-6, July 2001, pp.1407-1415.

[18] Polkowski, L. *Rough Sets, Mathematical Foundations, Advances in Soft Computing*. Physica – Verlag- A Springer-Verlag Company, 2002.

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

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

[21] Stefanowski, J. On rough set based approaches to induction of decision rules. In: *Rough Sets in Knowledge Discovery*. Vol 1, Heidelberg: Physica Verlag, 1998, pp.500-529.

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

[23] Zadeh, L.A. Outline of a new Approach to the Analysis of Complex Systems and Decission Processes. In: *IEEE Trans .S.M.C*. 3/1973, pp.28-44.

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