Systems Modelling on the basis of Rough and Rough-Fuzzy Approach

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

Abstract: In this paper the modelling of the information , economic and social systems is presented. The models are based on the rough sets theory, and the fuzzy and rough sets theory. These models have represented two real information systems, and a system of an internal human population migration. The information systems are represented as a table where every column represents an attribute (a variable, a property). This attribute can be measured or may also be supplied by a human expert. To obtain the necessary data questionnaires were use. To the migration model selected socioeconomic data, indicators, are applied. Economic and demographic indicators that affect size of migration for districts in the Czech Republic are defined. In data pre-processing we focused on different processing of data inputs. It means for all indicators we used selected data discretization techniques. In the migration models creation phase we deal with a new design of membership function shapes and rule base definition. The classifier models were carried out in MATLAB. Performed experiments have proven the accuracy of the proposed approach.

Key-words: Modelling, rough sets theory, fuzzy sets, information system, internal human migration, evaluation, classification

1 Introduction

In this article we will review the basic concepts and definitions of rough sets, fuzzy and rough sets, and approaches related to an evaluation / classification of systems. Systems can be usually described and defined by humans [13, 31].

A model of system [5] is an idealized representation , an abstract and simplified description, of a real world situation that is to be studied and/or analyzed.

We will deal with a problem of information systems (ISs) evaluation/classification. We evaluated two computer ISs (IS STAG and IS OPAC), which run on the intranet at our university, by gaining the information / data from their users. The first of the two ISs is STAG. Its main goal is to provide an organizational and administrative support for: students (on-line registration to courses and exams, access to timetable, curricula, syllabi, and further information), faculty (student records, computerized check of fulfilment of the requirements by every individual student, and further standard outputs), departments (offer and advertising of courses, information about students enrolled to courses, agenda of exams, and of registration to exams), persons responsible for timetabling (support for creating the timetable by providing free room search and collision checking), internal accreditation board as a part of the internal quality assurance system, and all individual faculty members (timetable information, lists of students registered to courses, search for free rooms, etc.). The second of the two ISs is the library information system OPAC. The main goal of this information system is: online book reservation, retrieval catalogue and library services, quick search in library database, and readers monitoring accounts.

In the next group of experiments we will work with an internal demography model. Demography is the statistical study of all populations. It can be a very general science that can be applied to any kind of dynamic population, that is, one that changes over time or space (see population dynamics). Demography encompasses the study of the size, structure, distribution of populations, and the way populations change over the time due to births, deaths, migration and ageing. Changes in the population number and population increase are basic topics of demography. A natality, mortality and a spatial mobility – migration influence the status of the population number directly. It teams up with population geography that deals with migrations and population distribution. Population evolution is a result of natural reproduction population (births, deaths) but also of the migration result. Many factors influence the size of the population migration. They are job opportunities, an environment, etc. (more in [30, 37]).

The work reflects the past years trend, which is based on the diffusion of various traditional methods and approaches to the way of tackling new problems. Methods of computational intelligence (CI) are used for

modelling of systems [2]. 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, 31]. They have been used in many uncertainty information processing systems successfully.

1.1 Rough Sets Approach

The rough sets theory (RST) [26, 27, 29] is based on the research of information system's logical properties, and broadly speaking uncertainty in it is expressed by a boundary region. 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.

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.

The approximations are two basic operations in RST [27]. 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 [17]:

$$IND(B)=\{(x,y)\in U \mid \forall\, a \in B \;\; a(x)=a(y)\}, \tag{1}$$

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 $B \subseteq A$ and $X \subseteq U$. We can approximate X by using only the information contained in B by constructing lower approximation (2) and upper approximation (3) of X in the following way:

$$\underline{B}(X)=\{x\in U: B(x)\subseteq X\} \text{ and} \tag{2}$$

$$\overline{B}(X)=\{x\in U: B(x)\cap X\neq\varnothing\}. \tag{3}$$

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

$$BN_B(X)=\overline{B}(X)-\underline{B}(X), \tag{4}$$

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, then set X is rough with respect to B. Rough sets are defined by approximations and have properties defined in [17, 26, 27, 29].

1.2 Fuzzy Sets Approach

The theory of fuzzy sets (FSs) is an approach, instrument specifying how well an object satisfied a vague description. In this theory an element belongs to a set according to the membership degree (membership function values) [40, 41, 42], 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.

Let S 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 S takes just value u. In the theory of FSs, FS on universe U is defined by membership function (MF) $\mu(s)$. If $\mu_N(s)=0$ then s does not belong to FS N, if $\mu_N(s)=1$ then s belongs to FS N, if $\mu_N(s) \in [0,1]$ then s partially belongs to FS N, in other words it is not possible to certainly identify if s belongs to FS N [41, 42].

Let´s suppose it is defined a fuzzy system with input and output fuzzy variables. The connection betweens the inputs and outputs is performed by fuzzy inference process which uses FSs for formulating the mapping from a given input to an output. This process uses fuzzy rules / IF-THEN rules in the following way:

IF premise (antecedent)
THEN conclusion (consequent). (5)

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 [22, 25, 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 [22] and [33]).

1.3 Rough-Fuzzy Approach

The theory FSs and RST are 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 [24] 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 [24] 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 [24]).

For example, in [9] 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 [36] 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 [32] 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.

RST and FSs are applied in a classifier modelling [12, 19]. This case [12, 19] 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 [3, 4].

In this paper, RST is applied to an information systems (ISs) evaluation. Its goals are:

- to collect real data about ISs
- to suggest and realize a tool based on RST for generating conditioned rules
- to use RST box for ISs evaluation.

In the second group of experiments RST and FSs are used in classification models of human internal migration (it means migration rate (MR) in the Czech Republic. Its goals are:

- to use a different way of the data pre-processing; it means on selected data discretization techniques [8, 39]
- to apply RST box for generation of IF-THEN rules for MR fuzzy (Mamdani's FIS)
- to optimize MF shapes in Mamdani's FIS on the basis of selected data discretization techniques (histogram analysis and binning).

Both parts should validate usability of rough and rough-fuzzy approach in research areas.

2 Evaluation of Information Systems

The next step of the evaluation was to collect real data about these ISs (IS STAG and IS OPAC). Both of them are frequently used by university students. Therefore we can evaluate these systems on the basis of information gained from a student questionnaire. They classified attributes of ISs by linguistic descriptors. The model of the ISs evaluation is in the Fig.1.

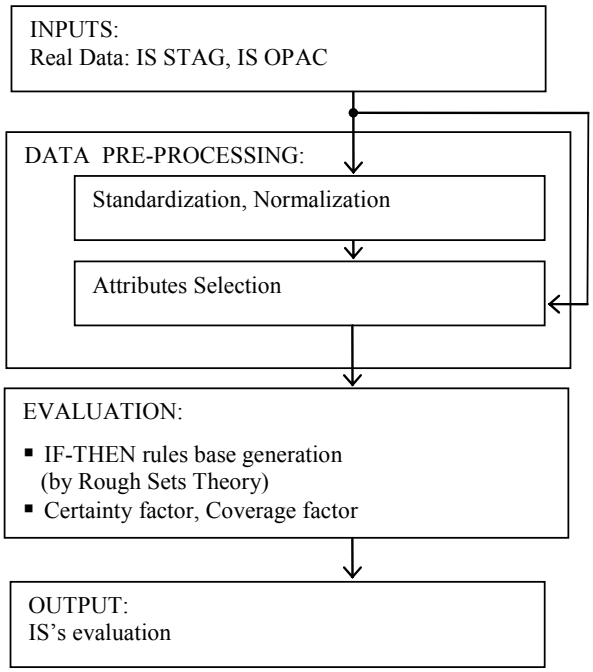

Fig.1 Model of ISs evaluation

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, computational intelligence, soft computing) 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 of the rule. This expression is referred to as IF-THEN rule-based form [31]. 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). Linguistic variables (small, high, …) are transformed into FSs. 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 use of rules of this form is the basic frame for fuzzy control.

From a process evaluation point of view, IS can be defined as a decision table [17] which represents a data set where every column represents an attribute that can be measured for each object (see the Table 1). The attribute a_i is i-th attribute; x_j is j-th object; v_{ji} is attribute value and d is decision attribute with value h_r for i=1,2,…,n; j=1,2,…,m and r = 1,2,…, q.

2.1 IF-THEN Rules Generation

First problem, we tried to resolve the evaluation on the basis of IF-THEN rules generation from the data. A whole range of scientific papers that deal with rules generation from analysed data and a lot of various methods and procedures using can be found in [21, 31, 35, 38].

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

For the generation of IF-THEN rules we have modified LEM1 algorithm and implemented the procedure as a tool [11, 12]. This tool is further applied to verify the proposed algorithms for partial calculations with real data. For conciseness, this modified algorithm is summarised in a pseudo code (Fig.2).

```
% algorithm procedure
% input: IS as a decision table  T = (U,A,D,f)
% where U= x1,x2 ,…,xm , A= a1,a2,…,an ,
% D= h1,h2,…hq , 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={s1,s2,…,sm*(n+1)}
    if  any object sx = Ø then //(x=1,2,…, m *
    (n+1))
          for every object sx do replace sx by -1
                    if any vector X=[x1,…,xi]contain -1 then
                        //i=1,2,…m
                        delete xi
                    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 A(X)
          if xi ∈ A(X) then
          create rule and insert it to NO Rules
          end  {if}
     end  {for}
end  {algorithm}
```

Fig.2 Algorithm procedure

This algorithm is the kernet of a toolbox [11, 12] called Rough Sets Toolbox (RSTbox) functioning for automatic rules generation. We created this toolbox in MATLAB environment. RSTbox is able to work with data in plain text format. It consists of four modules and among others supports data pre-processing, input data

selection, if-then rules generation and computation of approximations. This instrument is further applied for verifying the proposed algorithms for partial calculations with collected data. Entering screen from this toolbox (with loaded data) is in the Fig.3.

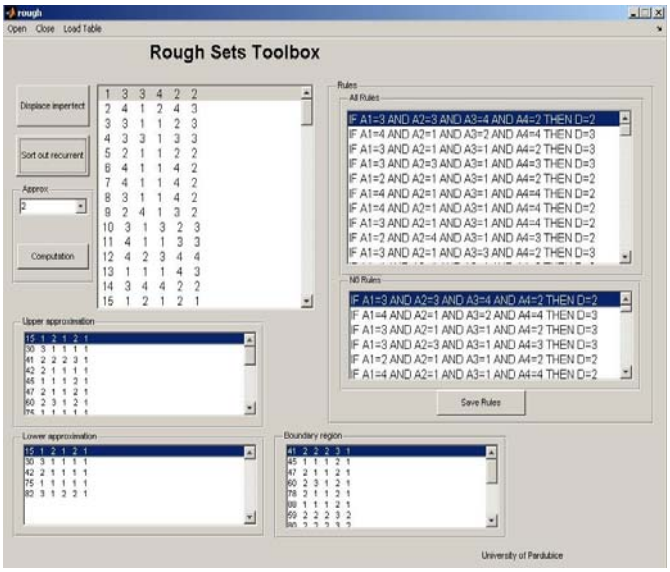

Fig.3 Window of Rough Sets Toolbox in MATLAB

Outputs of RSTbox are:

- a set of IF-THEN rules generated in *.txt format (All Rules)
- computed approximations for analyzed data (lower and upper approximation and boundary region)
- a set of certain rules (NO Rules) for analyzed data

2.2 Modelling of ISs Evaluation

We defined through the process of ISs (IS STAG and IS OPAC) evaluation questionnaires for a retrieval of real data. To obtain the necessary data, questionnaires with thirteen questions were used. Twelve questions were bipolar [16] and one question open – ended, all representing the attributes of ISs.

On the basis of the questionnaires and discussions among university ISs users three questions from the questionnaires for the evaluation of ISs were chosen: "What amount of financial resources should the organization invest in IS/IT every year?"; "Is the graphical interface user-friendly?" and "Do you agree with deployment of IS?" The first question represents the cost attribute cost with a low, middle, and high scope, the second describes the graphical interface attribute with yes or no scope, and the third is the decision deployment attribute with yes or no values.

We can define the rule form R1, R2,…, R12 for ISs evaluation in the Table 2.

Table 2 Rule form for ISs evaluation

Rule	Syntax of rules
R1	IF (costs is low) AND (graphical interface is friendly) THEN (deployment is yes)
R2	IF (costs is low) AND (graphical interface is not friendly) THEN (deployment is yes)
R3	IF (costs is middle) AND (graphical interface is friendly) THEN (deployment is yes)
R4	IF (costs is middle) AND (graphical interface is not friendly) THEN (deployment is yes)
R5	IF (costs is high) AND (graphical interface is friendly) THEN (deployment is yes)
R6	IF (costs is high) AND (graphical interface is not friendly) THEN (deployment is yes)
R7	IF (costs is low) AND (graphical interface is friendly) THEN (deployment is no)
R8	IF (costs is low) AND (graphical interface is not friendly) THEN (deployment is no)
R9	IF (costs is middle) AND (graphical interface is friendly) THEN (deployment is no)
R10	IF (costs is middle) AND (graphical interface is not friendly) THEN (deployment is no)
R11	IF (costs is high) AND (graphical interface is friendly) THEN (deployment is no)
R12	IF (costs is high) AND (graphical interface is not friendly) THEN (deployment is no)

If the prerequisites for IS STAG are as follows:

- on the basis of rules Rule1,2,3 and 5, the set {1,2,3,5} can be defined. It is the set of cases, denoted X_1
- answers (costs and graphical interface) are the set of attributes, denoted B (subset of all attributes)

then it can be defined : the lower approximation (2) is Ø; the upper approximation (3) is the set {1,2,3,5,7,8,9,11}; the boundary region (4) is the set {1,2,3,5,7,8,9,11} and the set is internally *B*-undefinable.

If the prerequisites for IS OPAC are follows:

- set {1,3,4,5,6} is the set of cases, denoted X_2
- answers (costs and graphical interface) are the set of attributes, denoted B (subset of all attributes)

then can it can be defined: the lower approximation (2) is the set {5}; the upper approximation (3) is the set {1,3,4,5,6,7,9,10,12}; the boundary region (4) is the set {1,3,4,6,7,9,10,12} and the set is roughly *B*-definable.

These can be seen in the Fig.4.

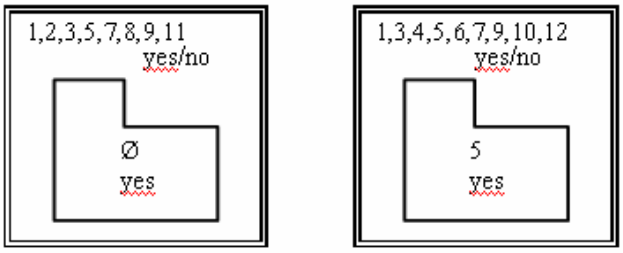

Fig.4 The RST interpretation of IS STAG and IS OPAC

The interview's results were adapted into a decision table. For the next computation, various combinations of attributes contained in a decision table were chosen. From this table lower and upper approximation, certainty *CeF* [27], and coverage *CoF* [27] factors were computed:

$$CeF = \frac{number\ of\ all\ cases\ satisfying\ \Phi\ and\ \Psi}{number\ of\ all\ cases\ satisfying\ \Phi}, \quad (6)$$

$$CoF = \frac{number\ of\ all\ cases\ satisfying\ \Phi\ and\ \Psi}{number\ of\ all\ cases\ satisfying\ \Psi}. \quad (7)$$

Every decision rule is an implication if Φ then Ψ, where Φ is a condition and Ψ is a decision; Φ and Ψ are logical formulas created from the attributes values and describing some properties of facts.

With every decision rule we associate two factors (*CeF* and *CoF*) [27] (see in the Table 3), where: if *CoF* = 1 then the rule is called "certain" and if 0 < *CoF* < 1 then the rule is called "uncertain".

Table 3 Certainty and coverage factors of IS STAG and IS OPAC

Appropriate rule	IS STAG		IS OPAC	
	CeF	*CoF*	*CeF*	*CoF*
R1	0.5	0.416667	0.84615	0.578947
R2	0.2	0.083333	0	0
R3	0.28571	0.166667	0.66666	0.210526
R4	0	0	0.5	0.052632
R5	0.8	0.333333	1	0.052632
R6	0	0	0.66666	0.105263
R7	0.5	0.238095	0.15384	0.222222
R8	0.8	0.190476	1	0.333333
R9	0.71428	0.238095	0.33333	0.222222
R10	1	0.142857	0.5	0.111111
R11	0.2	0.047619	0	0
R12	1	0.142857	0.33333	0.111111

From the computed factors we could induce the following statements for particular IS. Firstly, *CeF* and decision rules (R1 to R12) for IS STAG were used. For R1 and R7 the valid statement is: low costs and friendly graphical interface caused a positive decision (deployment = yes) in 50 % of the cases and negative decision in 50% of the cases. For R2 and R8 the valid statement is: middle costs and unfriendly graphical interface caused a positive decision (deployment = yes) in 20 % of the cases and negative decision in 80% of the cases etc. Secondly, *CoF* and decision rules (R1 to R12) were used. For R1 till R6 the valid statements are: 42 % positive decisions occurred when costs are low and graphical interface is friendly, 8 % positive decisions occurred when costs are low and graphical interface is unfriendly etc. Analogically, the statements have been constructed for IS OPAC. Based on this statement we can propose the conclusion that "graphical interface" attribute is crucial (for positive decision). In the first case (IS STAG) 92%, and in the second case (IS OPAC) 84% positive decisions occurred if graphical interface is friendly.

3 Classification of Migration Rates

In previous papers [14, 20] we created classification models MR in the 76 districts of the Czech Republic that expresses a number of migrants per 1 000 people to date (July, 1) in the year t) on the basis of a system approach. To define factors (indicators) that affect the MR was the part of the model creation. Basic demographic indicators and selected economic indicators detailed description has been published in [14]. There are: a crude marriage rate (CMR) - it is a number of marriages per 1 000 people to the date in the year t; a crude birth rate (CBR) - it is a number of live births per 1 000 people to the date in the year t); a crude abortion rate (CAR) - it is a number of abortions per 1 000 people to the date in the year t); a crude death rate (CDR) - it is a number of deaths per 1 000 people to the date in the year t); a crude divorce rate (CDiR) - it is a number of divorces per 1 000 people to the date in the year t); an unemployment rate (UR), and a gross average monthly wage (W).

In data pre-processing phase we focused on the correlation analysis and on the creation of classes of MR by an expert evaluation and several hierarchical and non-hierarchical cluster methods. We defined the centre of gravity for these clusters and extracted lexical variables (values) for the classification classes of MR on the basis of the centre of gravity. Then we concentrated on the classification by decision trees and neural networks. A lot of classification models were created and the best result was achieved by RBFNN in [14].

In order to an achieving the better results we used a fuzzy set in next models creation. In [20] we can see a modelling of classifiers based on FIS and its hierarchical structure for a model of an economic and social system. In data pre-processing were used selected cluster and fuzzy cluster methods. An optimization of FIS classifiers and a realization of the hierarchical structure classifier have been suggested and compared.

3.1 Data Pre-processing

Because we work by basic demographic and selected economic indicator, we can state we design a socio-economic system. Such system we can determine as a system based on economic and social indicators, socioeconomic environment and interactive relations. In

dependence on the field of our interests we work with varied data sources. It means we select specific economic and social indicators.

The sources of data may be classified in several ways, including the methods of collection and compiling the data, the ways of accessing the data, the agencies that collect and compile the data etc. The basic methods of collecting demographic data may be listed as enumeration, registration, and maintenance of administrative records [37]. In enumeration, the collecting authority initiates contact with the members of a population in order to collect and compile data about the status of the population at a specified date (e. g. sample survey). In the registration method, the member of the population report an event to the collection authority, e. g. birth and death registration. In the Czech Republic municipal authorities carry out these activities. The administrative records method first calls for a general registration/enumeration to initiate a file of individuals eligible or obligated to register, and then for a continuing registration by individuals as they become eligible for registration. Administrative records are usually established as an administrative file to collect nondemographic data. In spite of it's having features in common with the other methods, this method is distinctive enough to be listed separately (e.g. Medicare records, drivers license registration).

The Czech Statistical Office belongs to the most important agencies for data collection in the Czech Republic. It purveys general statistical information about this country and varied data about Czech municipalities and regions. As well there we can find European data and international comparison, too and we can count it as a basic source of demographical and socioeconomic data.

In design of classification model we work with districts. Every one is an object $\mathbf{o}_b$ and is described by p indicators (characteristics). A vector of measurement $\mathbf{o}_b$ contains values z_{bd} of p characteristics in formula (8) that it is the following:

$$\mathbf{O}_b = \{z_{b1}, z_{b2}, \ldots, z_{bd}\}, \quad (8)$$

for b-th object $\mathbf{o}_b$, (b= 1, 2, …, e). The input set of the objects which are determined for the clustering, can be expressed by a formula of objects matrix $\mathbf{O}$(e×d), where e is a number of objects, for b =1, 2, ..., e and d is a number of characteristics, for c = 1, 2, ..., d.

After determination a factors intensity on the migration presented in [14] we focused on selected data discretization techniques. It means for all indicators we used the histogram analysis and binning methods.

Histogram analysis [8] is an unsupervised discretization technique because it does not use class information. Histograms use binning to approximate data distributions and belong to a popular form of data reduction. They are also often used to explore the data before manipulations and model building and are frequently used to reveal imbalances in the data. By using of value of modes (it is such value z_{mod} of random variable Z, in that is frequency function of random variable Z locally maximum value, it is:

$$f(z_{mod}) = \max (f(z)), \quad (9)$$

and showing of histograms we determined intervals of bins.

In this analyse we used two types of partitioning rules. There are equal-width (M2) equal-frequency (M3) and an expert evaluation approach by modes of histograms of each indicator (M1).

In an equal-with histogram, the width of each bucket range is uniform and in an equal-width histogram, the buckets are created by the frequency of each bucket. Frequency in each bucket is constant.

Binning [8, 39] is a top-down splitting technique based on a specified number of bins. These methods are also used as discretization methods for numerosity reduction and concept hierarchy generation. Attribute values can be discretized by applying equal-width or equal-frequency binning, and then replacing each bin value by the bin mean or median. Binning does not use class information and is therefore an unsupervised discretization technique, too. It is sensitive to the user-specified number of bins, as well as the presence of outliers.

In our case we applied these binning methods for data smoothing [there are the smoothing by bin means (M4), and smoothing by bin boundaries (M5).

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. Smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value [8]). For example the results of the CBR indicator splitting are in the Table 4.

Table 4 Results of the CBR indicator splitting

Method	Bin 1	Bin 2	Bin3
M1	(8.16; 8.74)	(8.75; 10.04)	(10.05;11.11)
M2	(8.16;9.14)	(9.15;10.12)	(10.13;11.11)
M3	(8.16;9.23)	(9.24;9.79)	(9.8; 11.11)
M4	(8.93;8.93)	(9.51;9.51)	(10.31;10.31)
M5	(8.16;9.24)	(9.24;9.78)	(9.78;11.11)

In the next part of this paper the M1, M2 and M3 methods are only used and classification model creation is described. These models use the demographic and economic indicators values for the year of 2004. In these models inputs: UR, W, CBR, and CDR were used.

3.2 Modelling of Migration Rate Classifiers

The basic scheme of this problem is depicted in the Fig.5. Because formerly designed classification models based on the non-hierarchical cluster analysis, neural networks and regression trees are described in [39] and fuzzy models in [22, 23, 28] we focused on design of new approach for the definition of membership function shapes and base of rules for FIS.

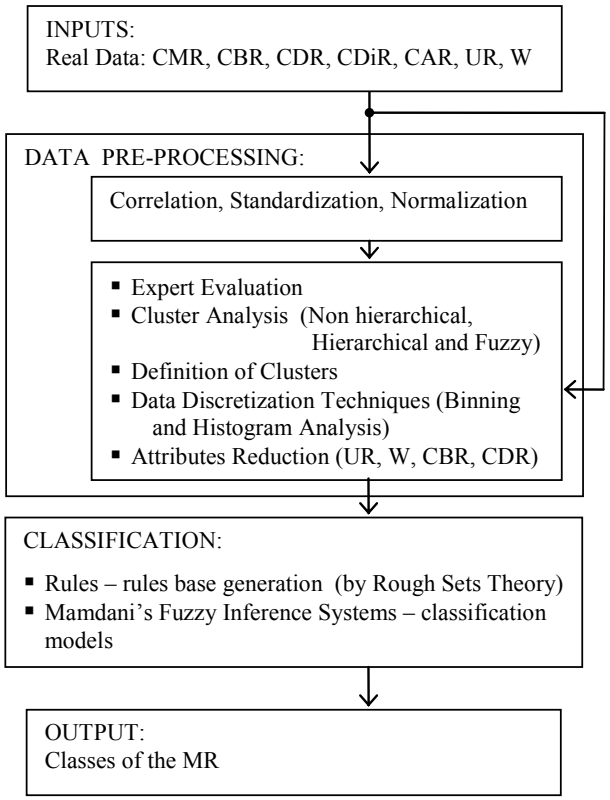

Fig.5 Model of hybrid classifier

Parameters for a definition model of MR evaluation can be expressed by an incompleteness and disproportion. Classification deals with knowledge and data characterized by an uncertainty. This was realized by means of FIS [22, 33].

The FIS can be defined as a multiple inputs and single output system. It is more described in [33, 42]. A disadvantage of this approach to the design of FIS [31] is an exponential growth of the number of the fuzzy rules (FRs) in the base of FRs (BFRs) and the FIS can be realized ineffectively. This problem can be removed by a hierarchical structure of FIS [6, 23, 28]. In the hierarchical structure of FIS it is necessary to determine the number of FRs for the first and other levels, see more in [6, 20].

The heuristic approach for the creation of FIS (it means the shape and number of membership function of inputs and output variables, and BFRs) was used because an exact general method for definition of their number does not exist [33]. Conditioned IF-THEN FRs were generated by rough sets theory based algorithm [7, 10, 18, 21, 27, 29] by the lower approximation. This procedure was implemented as a toolbox called Rough Sets Toolbox (RST) [10] in MATLAB environment functioning for minimal automatic rules generation.

Generally we can say, that our classification model uses Mamdani FIS and it was defined for 4 input variables (UR, W, CBR and CDR) and 1 output variable MR. It contains the fuzzification process, the inference mechanism and the defuzzification process [22, 33, 42]. For inputs (output) variables there were defined membership functions (MFs) of fuzzy sets (MF1, MF2 and MF3) where FRs are written in form IF antecedent THEN consequent. Consequently on the basis of a lot of simulations we have selected the centre of gravity and mean of maximum defuzzification method. For methods (M1, M2 and M3) we designed 8 types of FIS.

We used min and max values from these variables for a definition of the universe. The three triangle MFs of fuzzy sets for each variable were designed by expert for different tested models (these input MFs were symmetric and non-symmetric). The values of non-symetric MFs of the CBR indicator are defined in the Table 5.

Table 5 MF for CBR indicator

Method	MF1	MF2	MF3
M1	(8.16; 8.2; 8.7)	(8.4; 9.5; 10.0)	(9.5; 10.5; 11.1)
M2	(8.16; 8.84; 9.1)	(8.84; 9.29; 10.2)	(9.52; 10.43; 11.1)
M3	(8.16; 8.7; 9.24)	(8.7; 9.47; 9.7)	(9.47; 0.45; 11.11)

Testing on the whole training data were used in experiments. This approach is based on using one data set for both training and testing. This method is applicable, however, bears the highest thread of overfitting and decreasing the testifying parameter abilities. In this case if we used training set for testing, we can only determine the resubstitution error [8, 39]. It is the error rate in the training data set. It is calculated by resubstituting the training instances into a classifier that was constructed from them. Although it is not a reliable predictor of the true error rate on new data, it is nevertheless often useful to know. On the basis of the small amount of data we used this type of testing, only.
The best result of MR classification was achieved for model of the M2 method. It uses symmetric inputs and outputs MFs, the mean of maximum defuzzification method and 45 FRs generated and optimized by RST. The accuracy of classification is 62 % for this model.

4 Conclusion

A lot of approaches for generating conditioned rules from decision tables and classification are proposed up

to the present day. We have presented, in this article, an approach based on rough set theory.

Application RSTbox, which is based on rough sets, was designed and created in MATLAB. This tool was used in experiments with data collected at the University of Pardubice. The experiments proved the accuracy (correctness) of the RSTbox and possibility of drawing rules and conclusions from data with this tool.

From the computed values we can see that for IS users graphical interface is a crucial attribute. In the first case (IS STAG) 92% and in second case (IS OPAC) 84% positive decisions occurred, when graphical interface is friendly. If we compare both information systems on the basis of acquired information summarized in the Table 3, we can tell IS OPAC is perceived better than IS STAG.

Demographic and economic indicators that influence the size of MR were defined too in the paper. A comparing of classification results with the final classes (classification model of MR on the basis of hybrid FIS) with FIS [20] and hierarchical FIS [20] (31%) achieved better classification results (62%).

For testing we used the training data set, although it will be possible to use new testing dataset for finding of real quality of classifier. The hold-out method [8] is possible to use there. In this case it is the question:

Is this achieved quality of designed classifier acceptable for a socio-economic system classifier?

On the other hand for an objectification of the classifier operation, we tested this hybrid FIS classifier also (HFISC) [10] by known databases [1].

Achieving of better results is conditioned by defining other factors. These are e.g. a description of districts from the point of view of an environment, an area topology, a structure of the population education, job opportunities etc. The usage of a fuzzy set appears convenient for district rating by these factors.

5 Acknowledgement

The work was supported by the National Science Foundation of the Czech Republic under Grant No. 402/08/0849 with title Model of Sustainable Regional Development Management.

References

[1] Asuncion A, Newman DJ, *UCI repository of machine learning databases and domain theories*. URL: <ww:http://www.ics.uci.edu/MLRepository.html> , 1999.

[2] Bezdek JC, What is computational intelligence? *Computational Intelligence: Imitating Live*, 1994, pp.1-12.

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

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

[5] Gass SI, Harris CM, *Encyclopedia of operations research and management science* (Kluwer Academic Publishers, Boston, 2004).

[6] Gegov AE, Frank PM, Hierarchical Fuzzy Control Multivariable Systems. *Fuzzy Sets and Systems*, Vol.72, 1995, pp.299-310.

[7] Grzymała-Busse JW, Siddhaye S, Rough set approaches to rule induction from incomplete data. *Proc. of the IPMU2004* , Vol.2, 2004, pp.923-930.

[8] Han J, Kamber M, *Data mining: concepts and techniques,* San Francisco: Morgan Kaufmann Press, 2001.

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

[10] Jirava P, *Information system analysis based on rough sets: theses of the dissertation,* Pardubice: University of Pardubice, 2007.

[11]] 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 CONTECSI 2007.* Sao Paolo, Brasil, 2007, pp.566-573.

[12] 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.

[13] Kašparová M, Prediction Model Analysis of Financing of Basic Transport Services, W*SEAS Transaction on Systems*, WSEAS Press, Vol.5, 2007, pp.211-218.

[14] Kašparová M, Křupka J, Classification and prediction models for internal population migration in distrists, *WSEAS Transaction on Systems,* Vol.5*,* 2006, pp.1540-1547.

[15] Kašparová M, Jirava P, Křupka J, Hybrid Approach For Modelling Of Internal Human Population Migration Classifiers. *Proc. of the 12th IASTED International Conference Artificial Intelligence and Soft Computing (ASC 2008)*, Ed. A.P. Del Pobil, September 1-3, 2008 Palma de Mallorca, Spain, ACTA Press: Anaheim, Calgary, Zurich, 2008, pp.50-54.

[16] Kendall KE, Kendall JE, *System Analysis and Design*. Pearson Education, 2004.

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

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

[19] Křupka J, Jirava P, Modelling of Rough-Fuzzy Classifier. W*SEAS Transaction on Systems*, WSEAS Press, Vol.7, 2008, pp.251-263.

[20] Křupka J, Kašparová M, Modelling of internal human population migration classifiers by fuzzy inference system and its hierarchical structure, *WSEAS Transaction on Systems*, Vol.6*,* 2007, pp.461-466.

[21] 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, Japan, October 6-8, 2006.

[22] Kuncheva LI, *Fuzzy Classifier Design,* New York: Physica-Verlag, 2000.

[23] Lee ChCh, Fuzzy logic in control systems: fuzzy logic controller - part I and II, *IEEE Transaction on Systems, Man, and Cybernetics*, Vol.20, 1990, pp.404-433.

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

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

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

[27] 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.

[28] Pedrycz W, *Fuzzy control and fuzzy systems,* London: Research Studies Press Ltd., 1993.

[29] Polkowski L, *Rough Sets, Mathematical Foundations, Advances in Soft Computing*. Heidelberg: Springer Verlag, 2002.

[30] Preston SH, Heuveline P, Guillot M, *Demography: measuring and modeling population processes,* Malden: Blackwell Publishing, 2006).

[31] Olej V, Křupka J, *Analysis of Decision Processes of Automation Control Systems with Uncertainty*. University Press Elfa, Košice, 1996.

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

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

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

[35] 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.

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

[37] Siegel JS, *Applied demography: applications to bussines*, g*overnment, law, and public policy,* San Diego: Academic Press, 2002.

[38] Shavlik JW, Dietterich GT, *Readings in Machine Learning*. Morgan Kaufman, 1990.

[39] Witten IH, Frank E, *Data mining: practical machine learning tools and techniques,* Amsterdam: Morgan Kaufmann Publishers, 2005.

[40] Zadeh LA, Fuzzy Sets. *Information and Control*, Vol.8, 1965, pp.338-353.

[41] Zadeh LA, 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.

[42] Zadeh LA, 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