ROUGH SETS THEORY AND UNCERTAINTY INTO INFORMATION SYSTEM

Pavel Jirava
Institute of System Engineering and Informatics
Faculty of Economics and Administration, University of Pardubice

***Abstract**: This article is focused on rough sets approach to expression of uncertainty into information system. We assume that the data are presented in the decision table and that some attribute values are lost. At first the theoretical background is described and after that, computations on real-life data are presented. In computation we work with uncertainty coming from missing attribute values.*

***Keywords**:* Information System, Rough Sets Theory, missing attribute value, characteristic relation.

Introduction

Information system can be defined by different manners. From the view of Rough Sets Theory is common following definition."A data set is represented as a table; 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. " This table is called an information table [3,6].

More formally, the information system *IS* is the 4-tuple

$$IS=(U, Q, V, f), \quad (1)$$

where: U is a finite sets of objectives (universe), $Q = \{q_1, q_2,\ldots, q_m\}$ is a finite set of attributes, V_q is the domain of the attribute q, $V = \bigcup_{q\in Q} V_q$ and $f: U \times Q \rightarrow V$ is a total function such that $f(x,q) \in V_q$ for each $q \in Q$, $x \in U$, called information function [4].

In practice input data, presented as decision tables, may have missing attribute and decision values, i.e., decision tables are incompletely specified. This is examined in the next part.

Rough Sets Theory and uncertainty

The main goal of the rough sets analysis is synthesize approximation of concepts from the acquired data [5]. Every object we explore we associate with some information (data). Objects characterized by the same data are indiscernible in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough sets theory [6].

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 rough set theory 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 rough sets theory [6]. Suppose we are given two finite and non empty sets U and A and/or Q. 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 $I(B)$ on U

which will be called an indiscernibility relation. Indiscernibility relation is defined [6] by following way:

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

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 rough sets theory.

Rough Sets Theory has been used to lot of branches of science and plenty of software systems that implement rough set methods were developed. The main goal of the rough set analysis is to synthesize approximation of concepts from the acquired data. A data set is represented as a table, where every column represents an attribute and each row represents a case, object. This table is usually called an information table or information system. 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. In practice attribute values are frequently uncertain. Rough Sets Theory deals with uncertainty problem outgoing from ambiguity of exact terms or with another kind of uncertainty outgoing from definition of attribute values [8].

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
- unknown or missing values of attribute

We will deal here only with the fourth case. Uncertainty coming from missing or unknown attributes occures 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. The first rough set approach to missing attribute values, when all missing values were lost, was described in 1997, where two algorithms for rule induction, LEM1 and LEM2, modified to deal with such missing attribute values, were presented. In 1999 this approach was extensively described together with a modification of the original idea in the form of a valued tolerance based on a fuzzy set approach [1].

In the second case the value does not matter, so such values are also called "do not care" conditions. The second rough set approach to missing attribute values, in which the missing attribute value is interpreted as a "do not care" condition, was used for the first time in 1991. Each missing attribute value was replaced by all possible values. This idea was further developed and furnished with theoretical properties in 1995 [1].

In practice input data may have missing attributes. So decision tables are incompletely specified. A characteristic relation describes these tables.

Characteristic relation is a generalization of the indiscernibility relation. For cases where missing attributes are lost will be characteristic relation denoted *LO(B)* in this article, where *B* is subset of set of all attributes. For $x, y \in U$ it is as follows:

$$(x, y) \in LO(B) \text{ if and only if } \rho(x, a) = \rho(y, a) \text{ for all } a \in B \text{ such that } \rho(x, a) \neq ?, \quad (3)$$

where: "ρ" is function, that maps the set of ordered pairs (case, attribute) into the set of all values and "?" represents lost values.

For any x the characteristic relation *LO(B)* may be presented by the characteristic set $I_B(x)$ in the following way:

$$I_B(x) = \{y \mid (x, y) \in LO(B)\}. \quad (4)$$

For cases where missing attributes are "do not care" will be characteristic relation denoted *DO(B)*. For $x, y \in U$ it is as follows:

$$(x, y) \in DO(B) \text{ if and only if } \rho(x, a) = \rho(y, a) \text{ or } \rho(x, a) = *$$
$$\text{or } \rho(y, a) = * \text{ for all } a \in B, \quad (5)$$

where: "*" represents do not care values.

For any *x* the characteristic relation *DO(B)* may be presented by the characteristic set $J_B(x)$ by the following way:

$$J_B(x) = \{y \mid (x, y) \in DO(B)\}. \quad (6)$$

Characteristic relation *LO(B)* is reflexive and characteristic relation *DO(B)* is reflexive and symmetric [1].

Every decision rule is an implication if Φ then Ψ , where Φ is condition and Ψ is decision; Φ and Ψ are logical formulas created from attributes values and described some properties of facts. With every decision rule we associate two conditional probabilities: the certainty factor (*CeF*) and coverage factor (*CoF*) [7], where:

$$CeF(\Psi I \Phi) = \frac{number\ of\ all\ cases\ satisfying\ \Phi\ and\ \Psi}{number\ of\ all\ cases\ satisfying\ \Phi}, \quad (7)$$

$$CoF(\Phi I \Psi) = \frac{number\ of\ all\ cases\ satisfying\ \Phi\ and\ \Psi}{number\ of\ all\ cases\ satisfying\ \Psi}. \quad (8)$$

If *CoF* = 1 then the rule is called "certain" and if 0 < *CoF* < 1 then the rule is called "uncertain".

Experimental part

We evaluated two information systems (the information system STAG and the library information system Daimon Opac1.6.0- OPAC), which runs on university intranet[1]. First of them is system STAG. Its main goal is to provide an organizational and administrative support to this system of study for students and staff. The main goals of the second system are online book reservation, retrieval catalogue and library services, quick searching in library database and monitoring reader's accounts. University students frequently use both of them. Therefore we can evaluate these systems on the basis of information gained from student questionnaire [2].

The results of interview are shown in following decision tables (Table 1) where were used three questions for computation: the 1st question (A1) "What amount of investment resources should organization every year invest in IS/IT?" with scope low, middle and high; the 2nd question (A2) "Is graphical interface user friendly?" with scope yes, no; the 3rd

[1] The University of Pardubice

question (D) “Would you choose the system for implementation in organization?” with scope yes, no. The number of analogous rules (cases) N or frequencies of answers were defined for information systems STAG and OPAC [2].

For example, it can be represented by the following decision rule, where decision rule express relationship between conditions and decisions:

R_x: *IF cost is low AND graphical interface is friendly THEN deployment is yes.*

There are calculations of certainty and coverage factors (Table 2) on the basis of formulae (7, 8).

Next, we can see these input data, with lost values, in the Table 3. This table is called “incompletely specified” and the lost values are denoted “?”. There are calculations of certainty and coverage factors (Table 4), too.

What the data from Table 4 tell us? From the decision rules and certainty factor for incompletely specified information system OPAC we can draw the following conclusion:

- values of certainty and coverage factors differ from the values in Table 2 with no missing attributes values. For example for information system STAG high costs and friendly graphical interface caused positive decision (deployment = yes) in 100 % of the causes (while in Table 2 only 80 %); 44 % positive decisions occurred when costs are high and graphical interface is friendly (while in Table 2 is 33 %); otherwise 89 % positive decisions occurred when graphical interface is friendly (while in Table 2 is 92 %);
- for is OPAC high costs and friendly graphical interface caused always positive decision (deployment = yes) (also in Table 2); 7 % positive decisions occurred when costs are high and graphical interface is friendly (while in Table 2 is 5 %); otherwise 87 % positive decisions occurred when graphical interface is friendly (while in Table 2 is 84 %).

Table 1. Decision table of information systems STAG and OPAC (no missing attributes values)

Appropriate rule	Attribute		Decision		
	A1 (cost)	A2 (graphical interface)	D (deployment)	N for IS STAG	N for IS OPAC
R1	Low	friendly	Yes	5	11
R2	Low	unfriendly	Yes	1	0
R3	Middle	friendly	Yes	2	4
R4	Middle	unfriendly	Yes	0	1
R5	High	friendly	Yes	4	1
R6	High	unfriendly	Yes	0	2
R7	Low	friendly	No	5	2
R8	Low	unfriendly	No	4	3
R9	Middle	friendly	No	5	2
R10	Middle	unfriendly	No	3	1
R11	High	friendly	No	1	0
R12	High	unfriendly	No	3	1

Table 2. Certainty and coverage factors of information systems STAG and OPAC				
Appropriate rule	IS STAG		IS OPAC	
	CeF	*CoF*	*CeF*	*CoF*
R1	0,5	0,416667	0,846153846	0,578947
R2	0,2	0,083333	0	0
R3	0,285714	0,166667	0,666666667	0,210526
R4	0	0	0,5	0,052632
R5	0,8	0,333333	1	0,052632
R6	0	0	0,666666667	0,105263
R7	0,5	0,238095	0,153846154	0,222222
R8	0,8	0,190476	1	0,333333
R9	0,714286	0,238095	0,333333333	0,222222
R10	1	0,142857	0,5	0,111111
R11	0,2	0,047619	0	0
R12	1	0,142857	0,333333333	0,111111

Table 3. Decision table of incompletely specified information systems STAG and OPAC (with missing attributes values)					
Appropriate rule	Attribute		Decision		
	A1 (cost)	A2 (graphical interface)	D (deployment)	N for IS STAG	N for IS OPAC
R_{LO1}	?	friendly	Yes	1	1
R_{LO2}	low	?	Yes	2	1
R_{LO3}	low	?	No	1	0
R_{LO4}	?	friendly	No	2	2
R_{LO5}	middle	?	Yes	-	2

Table 4. Certainty and coverage factors of incompletely specified information systems STAG and OPAC (missing attribute values "lost value")				
Appropriate rule	IS STAG		IS OPAC	
	CeF	*CoF*	*CeF*	*CoF*
R1	0,333333	0,222222	0,909090909	0,666667
R2	0,2	0,111111	0	0
R3	0,333333	0,222222	0,666666667	0,133333
R4	0	0	0	0
R5	1	0,444444	1	0,066667
R6	0	0	0,666666667	0,133333
R7	0,666667	0,222222	0,090909091	0,142857
R8	0,8	0,222222	1	0,428571
R9	0,666667	0,222222	0,333333333	0,142857
R10	1	0,166667	1	0,142857
R11	0	0	0	0
R12	1	0,166667	0,333333333	0,142857

Conclusion

Rough Sets represent a powerful tool for decision making about information systems, data mining and drawing conclusions from data, especially in those cases where some uncertainty exists in the data. This paper has discussed the data with missing attribute values.

There are two approaches to missing attribute values entitled “lost value” or “do not care” conditions.

We can see these input data, with lost values (Table 3). This table is called “incompletely specified” and the lost values are denoted “?”. Calculations of certainty and coverage factors were performed for both completely and incompletely specified tables (Table 2, 4) and conclusions for this tables were formulated. For computing the data acquired by the research at the University of Pardubice [2] were used.

Literature

1. GRZYMAŁA-BUSSE, J.W., SIDDHAYE, S. Rough Set Approaches to Rule Induction from Incomplete Data. In: *Proceedings of the IPMU2004*, Perugia , Italy, July 4-9 2004, Volume 2, pp.923-930.
2. JIRAVA, P., KŘUPKA, J. Rough Sets and Evaluation of Information System, In: *11th. International Conference on Soft Computing: Mendel 2005*, Brno, 2005, pp.178-183. ISBN: 80-214-2961-5.
3. KOMOROWSKI, J., POLKOWSKI, L., SKOWRON, A. Rough sets: a tutorial. In: *S.K. Pal and A. Skowron, editors, Rough-Fuzzy Hybridization: A New Method for Decision Making*, Springer-Verlag, Singapore, 1998.
4. KŘUPKA, J. Rough Sets Theory in Decision Analysis. In: *Scientific papers of the University of Pardubice, Series D 9/2004*. Pardubice: Univerzita Pardubice, 2004, pp.93-99. ISSN 1211-555X.
5. PAWLAK, Z. Rough sets. In: *Int. J. of Information and Computer Sciences, 11, 5*., 1982, pp.341-356.
6. PAWLAK, Z. Rough set elements. In: *Rough Sets in Knowledge Discovery I – Methodology and Applications*, Physica Verlag, Heidelberg, 1998, pp.10-31.
7. PAWLAK, Z. A Primer on Rough Sets: A New Approach to Drawing Conclusions from Data. In: *Cardozo Law Review*, Volume 22, Issue 5, 6th July 2001, pp.1407-1415.
8. SLOWINSKI, R., STEFANOWSKI, J. Rough-Set Reasoning about Uncertain Data. In: *Fundamenta Informaticae*, 27, 2/3, 1996, pp.229-243.

Contact Address:

Ing. Pavel Jirava
Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic
e-mail: pavel.jirava@upce.cz
tel: 466036001