PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 133
Pozyskiwanie wiedzy i zarządzanie wiedzą 2010

Olga A. Siniavskaya, Boris A. Zhelezko
Belarus State Economic University, Minsk

KNOWLEDGE ACQUISITION BY MEANS OF ROUGH SETS THEORY

Summary: In this article possibilities of knowledge acquisition by means of rough sets theory are considered. The basic concepts of rough sets theory are given, including methods of calculation of basic indicators. Examples of rough sets theory application for calculation of probability of events' causes and consequences, estimations of results utility are described. The principle of economic information discretisation for its processing by means of rough sets theory methods is considered. Relation of rough sets theory with alternative decision-making theories and intellectual data analysis methods is shown.

Keywords: rough sets theory, information system, decision table, approximation, decision algorithm, knowledge acquisition.

1. Introduction

Rough sets theory is a new mathematical approach to intelligent analysis and data mining. The theory was suggested by polish scientist Z. Pawlak in 1982 [5]. Rough sets have many successful applications in practice, for example, in medicine, pharmaceutics, banking and finance, market analysis, environment control, etc. The goal of this paper is showing of the possibilities of rough sets theory for knowledge acquisition, and also determination of rough sets theory interrelation with alternative decision-making theories.

2. Basic concepts of rough sets theory

Rough sets philosophy is based on the assumption that any object may be associated with some information (data, knowledge). Objects characterized by the same information are *indiscernible* (*similar*) [6, p. 2]. Any subset of indiscernible objects is called an *elementary set* and forms a basic *granule* of knowledge domain. Any union of several elementary sets may be crisp (precise) or rough (imprecise) set. Objects which could not be precisely classified may be the rough sets elements. For

each rough set two precise sets correspond which called *the upper and the lower approximation*. *The lower* approximation includes all objects *surely* belonging to the set, and the upper approximation includes the objects *possibly* belonging to it. Rough set *boundary region* is a difference between the upper and the lower approximation.

The analytical base in the rough sets theory is an information system[1]. The notion of "information system" in rough sets theory terms essentially differs from the similar notions known in information science.

Information system is a data table with attributes in columns, investigated objects in rows and attributes values for the objects in cells [6].

Formally the information system can be presented in the form of $S = (U, A)$, where U – the finite nonempty set characterizing the problem area, and A – the set of attributes.

With each attribute $a \in A$ the set V_a of its values is associated, called the domain of a. Every subset B, included in A, determines a binary relation $I(B)$ on the set U. This binary relation is called "an indiscernibility relation" and defined as follows: $(x, y) \in I(B)$ if and only if $a(x) = a(y)$ for every $a \in B$, where $a(x)$ – the value of attribute a for an element x. Obviously, $I(B)$ is equivalence relation. The family of all equivalence classes $I(B)$, which is the part of subset B, is designated as $U/(IB)$ or simply U/B. Equivalence class $I(B)$, containing x, is designated as $B(x)$. If $(x, y) \in I(B)$, then x and y are called *B-indiscernible*. Equivalence classes $I(B)$ are called *B-elementary sets* or *B-granules* [6, p. 3].

If it is possible to select two attribute classes in information system: conditions attributes and decisions attributes, then the information system is called decision table, designated as $S = (U, C, D)$, where C and D are the sets of conditions and decisions attributed accordingly.

Table 1 is an example of decision table. Analogous example from environmental sphere was considered in the article [6].

Here we consider the economic situation example. In Table 1 data about sales centres and financial results of their activity are represented. The information is incomplete; given data do not allow receiving the unambiguous result of sales centre activity (many external, unknown for the analyst factors may influence on this result). Data *are inconsistent*; therefore the rules construction problem can be solved only approximately. In this example situations 1, 2 and 5 can be classified definitely as the reason of successful activity (profitability); situation 4 can be classified definitely as the unprofitability reason; situations 3 and 6 can be classified as the possible reasons or profitableness, or unprofitability.

[1] In the article [2, p. 249] "the information table" notion is used in this sense.

Table 1. Decision table example with data about sales centres

	conditions attributes			decision attribute	
Situation	Employees' qualification (EQ)	Range of goods (RG)	Location	Financial result (FR)	A number of sales centres
1	High	Various	City A	Profit	2
2	Very high	Various	City B	Profit	3
3	High	Narrow	City B	Profit	9
4	Low	Narrow	City A	Loss	12
5	Very high	Various	City C	Profit	7
6	High	Narrow	City A	Loss	3

Source: own elaboration based on [6, pp. 1-12].

Let us assume that in information system $S = (U, A)$, $X \subseteq U$, $B \subseteq A$ it is necessary to describe the set X by means of attribute values from the set B.

Lower approximation of the set X (the set of the all B-granules, included in the set X):

$$B_*(X) = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\}, \tag{1}$$

and upper approximation of the set X (the set of the all B-granules, which have nonempty intersection with the set X):

$$B^*(X) = \bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \varnothing\}. \tag{2}$$

Boundary region: $BN_B(X) = B^*(X) - B_*(X)$. If $BN_B(X) = 0$, then X is a precise set, otherwise X is a rough set.

In the example of Table 1: X_1={1, 2, 3, 5} is the set of profitability situations; $B^*(X_1)$ = {1, 2, 5} includes only *non-contradictory* situations; $B^*(X_1)$ = {1, 2, 3, 5, 6} also includes *contradictory* situations 3 and 6 in which there are different solution attribute values correspond to the same conditions attributes values.

Decision rules constitute the formal language for logical description of approximations. *Decision language* is the language of decision rules formal description. Decision rules are the expressions in the "if…then…" form: $\Phi \rightarrow \Psi$. Antecedent is denoted as Φ, consequent as Ψ, antecedent and consequent sets as *For(C)* and *For*(*D*) correspondingly, where "*For*" is a *formulae* set.

There is an example of decision rule:

(«EQ»=«High»)∧(«RG»=«Various»)→(«FR»=«profit»).

For decision rules evaluation and interpretation and conclusions obtaining on their base a number of quantity indicators are used. Let us consider their essence and formulas for their calculation.

1. *Support* of a rule is a maximum number of objects satisfying both antecedent and consequent:

$$supp_S(\Phi,\Psi) = card(\|\Phi \wedge \Psi\|_S). \tag{3}$$

2. *Certainty factor* is a frequency of occurrence of the objects having property Ψ in set of the objects having property Φ:

$$cer_S(\Phi,\Psi) = \frac{card(\|\Phi \wedge \Psi\|_S)}{card(\|\Phi\|_S)}. \tag{4}$$

This coefficient is widely used also in data mining and known in data mining terms as *confidence coefficient*.

3. *Coverage factor* is a frequency of occurrence of the objects having property Φ in set of the objects having property Ψ:

$$cov_S(\Phi,\Psi) = \frac{card(\|\Phi \wedge \Psi\|_S)}{card(\|\Psi\|_S)}. \tag{5}$$

4. *Strength* of a rule is the ratio of the *support of a rule* to a number of objects in the decision table:

$$\sigma_S(\Phi,\Psi) = \frac{supp_S(\Phi,\Psi)}{card(U)}. \tag{6}$$

Let us consider the following rule generated on the base of Table 1:

(«EQ»=«High»)∧(«RG»=«Narrow»)→(«FR»=«profit»).

For this rule:

supp = 9 (3rd situation took place 9 times).

$\sigma = \frac{9}{2+3+9+12+7+3} = 0.25$; *cer* = 9/(9 + 3) = 0.75;

cov = 9/(2 + 3 + 9 + 7) ≈ 0.43.

For *a certain* rule *cer* = 1, for *an uncertain* rule *cer* < 1. Certain rules correspond to the lower approximation, uncertain – to boundary region. Certainty and support are conditional probabilities which express precise degree of knowledge (data) about problem area.

Decision algorithm is a set of mutually excluding and exhaustive decision rules corresponding to the given decision table.

Inverse decision algorithm consists of rules in which antecedent and consequent interchange their position. Inverse algorithms are used for finding-out of the reasons which have caused those or other decisions.

For the considered example (Table 1), decision algorithm can be written as follows:

1. («EQ»=«High»)∧(«RG»=«Various»)→(«FR»=«profit»).

2. («EQ»=«Very high») → («FR»=«profit»).

3. («EQ»=«High»)∧(«RG»=«Narrow»)→(«FR»=«profit»).

4. («EQ»=«Low») → («FR»=«loss»).

5. («EQ»=«High»)∧(«RG»=«Narrow»)→(«FR»=«loss»).

And inverse decision algorithm can be written as follows:

1′. («FR»=«profit») → («EQ»=«High»)∧(«RG»=«Various»).

2′. («FR»=«profit») → («EQ»=«Very high»).

3′. («FR»=«profit») → («EQ»=«High»)∧(«RG»=«Narrow»).

4′. («FR»=«loss») → («EQ»=«Low»).

5′. («FR»=«loss») → («EQ»=«High»)∧(«RG»=«Narrow»).

In the above-mentioned example quantity indicators of decision evaluation have following values (Table 2).

Table 2. Quantity indicators of decision evaluation

Decision rules	Support	Strength	Certainty	Coverage
1	2	0.06	1	0.1
2	10	0.28	1	0.48
3	9	0.25	0.75	0.43
4	12	0.33	1	0.8
5	3	0.08	0.25	0.2

Source: own calculations.

Certainty factor values allow acquiring following knowledge:

– high employees' qualification and various range of goods or very high employees' qualification certainly maintain sales centres profit;
– low employees' qualification certainly causes sales centres unprofitability;
– high employees' qualification and narrow range of goods may cause:
- profit with probability equal to 0.75;
- loss with probability equal to 0.25.

On the other hand, the profit reasons may be the following:

– high employees' qualification and various range of goods with probability 0.1;
– very high employees' qualification with probability 0.48;
– high employees' qualification and narrow range of goods with probability 0.43.

The loss reasons may be the following:

– low employees' qualification with probability 0.8;
– high employees' qualification, but narrow range of goods with probability 0.2.

3. Data discretisation in decision tables

There are many situations in practice when descriptors are expressed by not equal, but close values. Let us consider the following example described in detail in article [1] (the fragment of decision table is presented in Table 3). For the enterprises bankruptcy diagnostic a number of indices are used, among them there are following indices: value of sales (SALES); the ratio of profit before tax to capital employed (ROCE); the ratio funds flow to total liabilities (FFTL); the ratio of current liabilities plus long-term debt to total assets (GEAR); the ratio of current liabilities to total assets (CLTA); the ratio of current assets to current liabilities (CACL); company age (AGE).

The indices are different for the firms and at first sight rough sets theory in this case is inapplicable. Nevertheless, within some time intervals of indices' values have identical quality (for example, high, average, low, etc.). The example of interval discretisation of indices' values is represented in Table 4.

After discretisation decision table will be such as Table 5 and become suitable to processing by means of rough sets theory. The received discrete values, in turn, may be associated with qualitative characteristics ("very high", "high", "average", "below average", etc.).

Table 3. Indices and results of firms' activity

Firm	SALES	ROCE	FFTL	GEAR	CLTA	CACL	AGE	Financial result
1	6 762	7.5364	0.1545	0.6233	0.6233	1.5489	74	profit
2	16 149	–1.0712	0.0271	1.2218	1.2218	0.6236	29	profit
3	8 086	15.2024	0.6163	0.3307	0.3307	2.3553	51	profit
4	7 646	31.2239	0.6312	0.5205	0.4829	1.6397	25	profit
5	11 528	1.3275	0.066	0.7124	0.6377	0.9967	40	loss
6	29 300	0.0745	0.0683	0.5977	0.4767	1.1994	25	loss
7	2 958	−9.4013	0.0145	1.4865	1.4865	0.4974	11	loss
8	2 978	8.5486	0.3285	0.3898	0.3883	2.0519	9	loss
…	…	…	…	…	…	…	…	…

Source: based on [1, pp. 561-576].

Table 4. Rules of quantitative values discretisation

Attribute	Interval 0	Interval 1	Interval 2	Interval 3
SALES	[2857, 5694)	[5694, 32683.5]	[32683.5, 167370]	
ROCE	[−37.3497, −9.31125)	[−9.31125, −6.3066)	[−6.3066, 1.71560)	[1.7156, 33.8451]
FFTL	[−0.3283, 0.029)	[0.029, 0.12095)	[0.12095, 0.21375)	[0.21375, 0.6312]
GEAR	[0.1212, 0.57495)	[0.57495, 0.793)	[0.793, 1.0985)	[1.0985, 3.5336]
CLTA	[0.1212, 0.465)	[0.465, 0.7026)	[0.7026, 1.4865]	
CACL	[0.4974, 1.16945)	[1.16945, 1.37075)	[1.37075, 4.4465]	
AGE	[2, 24.5)	[24.5, 90]		

Source: based on [1, pp. 561-576].

Table 5. Result discretisation of firm activity indices' values

Firm	SALES	ROCE	FFTL	GEAR	CLTA	CACL	AGE	Financial result
1	1	3	2	1	1	2	1	profit
2	1	2	0	3	2	0	1	profit
3	1	3	3	0	0	2	1	profit
4	1	3	3	0	1	2	1	profit
5	1	2	1	1	1	0	1	loss
6	1	2	1	1	1	1	1	loss
7	0	0	0	3	2	0	0	loss
8	0	3	3	0	0	2	0	loss
…	…	…	…	…	…	…	…	…

Source: based on [1, pp. 561-576].

4. Rough sets theory integration with alternative theories and methods

For the real problems decision, concerned with knowledge acquisition, in particular, for the problems of diagnostics and forecasting, it is quite often necessary to combine rough sets theory methods with other decision making theories and intellectual analysis methods, for example, genetic algorithms, decisions trees, probability theory, Dempster–Shafer theory of evidence, fuzzy sets theory, discriminant, statistical and economic analysis. Examples of rough sets theory use in integration with other theories and methods [1; 3; 4; 7-10], including practical examples, are briefly described in Table 6.

Table 6. Examples rough sets theory application integrated with alternative decision-making and intellectual analysis theories and methods

Authors, countries	Solving problem or theoretical research	Applied theories and methods	Description of initial data, tool maintenance, results
1	2	3	4
T.E. McKee, T. Lensberg (USA, Norway)	Corporation bankruptcy prediction	Rough sets theory, genetic algorithms	Data about 150 American companies were the base of decision algorithm generation. Data about 291 American companies from the period of 1991-1997 years were used for model validation. Classification quality on this sampling with the use of rough sets theory is equal to 67%. For the model in which rough sets theory and genetic algorithm were combined, data about 144 American companies were used, classification quality is equal to 80%.

1	2	3	4
L.-P. Khoo, L.-Y. Zhai (Singapore)	Diagnostics machines and mechanisms state	Rough sets theory, genetic algorithms	Special software *RClass-Plus* was applied, developed by authors for decision algorithm generation. Comparison of results with other programs (*ID3*, *LERS*, early version of *RClass*) was made.
Y.Y. Yao (Canada)	Theoretical research	Fuzzy and rough sets	Theoretical results
M. Quafafou (France)	Theoretical research	Fuzzy and rough sets theories combination, called α-RST	Theoretical results
P. Srinivasan, M.E. Ruiz, D.H. Kraft, J. Chen (USA)	Search systems (documents search on users queries)	Fuzzy and rough sets	The database of Medical Library containing 476313 concepts was used. Conformity of the document to user query was calculated, user received the documents whose conformity had been defined by the maximum value of the calculated coefficient.
L. Shen, F.E.H. Tay, L. Qu, Y. Shen (Singapore, China)	Diagnostics of the diesel engine unserviceability	Rough sets theory, method of quantity criteria values discretisation	Results of diesel engine tests were used. The probabilities of its unserviceability for various reasons were calculated.
M.J. Beynon, M.J. Peel (United Kingdom)	Corporation bankruptcy prediction	Rough sets theory modification, least squares method, FUSINTER-method of data discretisation	Information from "Financial analysis made easy (FAME)" database about more than 200 Britain corporations was used. From this database 30 bankrupts and 30 successful companies were selected. Data discretisation method was automated by means of Maple software.

Source: own elaboration.

5. Conclusions

Rough sets theory has many advantages, which make it useful and convenient tool for knowledge acquisition. In particular, preliminary information analysis allows obtaining a number of the important elements of knowledge about a problem situation: relations between attributes and/or criteria; information about their interaction on the base of approximation quality calculation of and its analysis by means of the fuzzy measures theory; the minimum subset of attributes or criteria (reduct), which includes all information necessary for decision-making; set of not reducing attributes and/or criteria (a core). The model of preferences is generated from the preliminary information, in the form of decision algorithm consisting of production rules.

Heterogeneous information (quantitative and qualitative, predetermined and not predetermined, real and fuzzy estimations, decision tables with missing values) may be processed by means of rough sets theory, and knowledge from such information may be acquired.

However, rough sets theory using is not recommended in the case of small samplings (less than 15 analyzed objects) because calculation of indices of causes and consequences probability may be incorrect;

References

[1] Beynon M.J., Peel M.J., *Variable precision rough sets theory and data discretisation*: *An application to corporate failure prediction*, "Omega" 2001, Vol. 29, pp. 561-576.
[2] Greco S., Matarazzo B., Slowinski R., *Rough sets methodology for sorting problems in presence of multiple attributes and criteria*, "European Journal of Operational Research" 2002, Vol. 138, pp. 247-259.
[3] Khoo L.-P., Zhai L.-Y., A *prototype genetic algorithm-enhanced rough set-based rule induction system*, "Computers in Industry" 2001, Vol. 46, pp. 95-106.
[4] McKee T.E., Lensberg T., *Genetic programming and rough sets: A hybrid approach to bancruptcy classification*, "European Journal of Operational Research" 2002, Vol. 138, pp. 436-451.
[5] Pawlak Z., *Rough sets*, "International Journal of Information & Computer Sciences" 1982, Vol. 11, pp. 341-356.
[6] Pawlak Z., *Rough sets and intelligent data analysis*, "Information Sciences" 2002, Vol. 147, pp. 1-12.
[7] Quafafou M., *α-RST: A generalization of rough sets theory*, "Information Sciences" 2000, Vol. 124, pp. 301-316.
[8] Shen L., Tay F.E.H., Qu L., Shen Y., *Fault diagnosis using Rough Sets Theory*, "Computers in Industry" 2000, Vol. 43, pp. 61-72.
[9] Srinivasan P., Ruiz M.E., Kraft D.H., Chen J., *Vocabulary mining for information retrieval: Rough sets and fuzzy sets*, "Information Processing and Management" 2001, Vol. 37, pp. 15-38.
[10] Yao Y.Y., *A comparative study of fuzzy sets and rough sets*, "Journal of Information Sciences" 1998, Vol. 109, pp. 227-242.

POZYSKIWANIE WIEDZY DZIĘKI TEORII ZBIORÓW PRZYBLIŻONYCH

Streszczenie: W pracy rozważa się możliwość pozyskiwania wiedzy dzięki teorii zbiorów przybliżonych. Zawiera ona podstawowe pojęcia z teorii zbiorów przybliżonych oraz metody obliczania podstawowych wskaźników stosowanych przez teorię. Opisany zostały przykłady obliczania prawdopodobieństwa przyczyn i skutków wydarzeń, oceny przydatności wyników. Praca określa zasady pobierania próbek danych do przygotowania się do ich przetwarzania metodami teorii zbiorów przybliżonych. Przedstawiono zależność pomiędzy teorią zbiorów przybliżonych a alternatywnymi metodami oraz teoriami eksploracji danych.