Data Analysis – a Rough Set View

Zdzisław Pawlak
Institute of Theoretical and Applied Informatics
Polish Academy of Sciences
ul. Bałtycka 5, 44 000 Gliwice, Poland
e-mail:zpw@ii.pw.edu.pl

1 Introduction

Rough set theory is a new mathematical approach to data analysis. Basic idea of this method is hinges on classification of objects of interest into similarity classes (clusters) containing objects which are indiscernible with respects to some features, e.g., colour, temperature etc., which form basic building blocks of knowledge about reality, and are employed next to find out hidden patterns in data. Basis of rough set theory can be found in [29, 32, 38, 46].

Rough set theory has some overlaps with other methods of data analysis, e.g., statistics, cluster analysis, fuzzy sets, evidence theory and other but it can be viewed in its own rights as an independent discipline.

The rough set approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, inductive reasoning and pattern recognition. It seems of particular importance to decision support systems and data mining.

Rough set theory has been successfully applied in many real-life problems in medicine, pharmacology, engineering, banking, financial and market analysis and others. More about applications of rough set theory can be found in [15, 16, 37, 45, 49, 53] and others.

Very promising new areas of application of the rough set concept seems to emerge in the near future. They include rough control, rough data bases, rough information retrieval, rough neural network and others.

2 An example

Starting point of rough set theory is a set of data (information) about some objects of interest. Data are usually organized in a form of a table called *information system* or *information table.*

A very simple, fictitious example of an information table is shown in Table 1. The table describes six cars in terms of their (attributes) features such as *fuel consumption* (F), *perceived quality* (Q), *selling price* (P) and *marketability* (M).

Table 1: An example of information system

Car	*F*	*Q*	*P*	*M*
1	*high*	*fair*	*med.*	*poor*
2	*v. high*	*good*	*med.*	*poor*
3	*high*	*good*	*low*	*poor*
4	*med.*	*fair*	*med.*	*good*
5	*v. high*	*fair*	*low*	*poor*
6	*high*	*good*	*low*	*good*

Our main problem can be characterized as determining the nature of the relationship between selected features of the cars and their marketability. In particular, we would like to identify the main factors affecting the market acceptance of the cars.

Information systems with distinguished decision and condition attributes are called *decison tables*.

Each row of a decision table determines a *decision rule*, which specifies *decisions* (*actions*) that should be taken when conditions pointed out by *condition* attributes are satisfied. For example in Table 1 the condition (*F,high*), (*Q,fair*), (*P,med*) determines uniquely the decision (*M,poor*). Decision rules 3) and 6) in Table 1 have the same conditions but different decisions. Such rules are called *inconsistent* (*nondeterministic, conflicting, possible*); otherwise the rules are referred to as *consistent* (*certain, deterministic, nonconflicting, sure*). Decision tables containing inconsistent decision rules are called *inconsistent* (*nondeterministic, etc*); otherwise the table is *consistent* (*deterministic, etc*).

The number of consistent rules to all rules in a decision table can be used as *consistency factor* of the decision table, and will be denoted by $\gamma(C, D)$, where C and D are condition and decision attributes respectively. Thus if $\gamma(C, D) < 1$ the decision table is consistent and if $\gamma(C, D) \neq 1$ the decision table is inconsistent. For example for Table 1 $\gamma(C, D) = 4/6$.

In what follows information systems will be denoted by $S = (U, A)$, where U – is *universe*, A is a set *attributes*, such that for every $x \in U$ and $a \in A$, $a(x) \in V_a$, and V_a is the domain (set of values of a) of a.

3 Approximations of sets

Now the indiscernibility relation will be used to define basic operations in rough set theory, which are defined as follows:

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

$$B^*(X) = \{x \in U : B(x) \cap X \neq \emptyset\},$$

assigning to every $X \subseteq U$ two sets $B_*(X)$ and $B^*(X)$ called the *B-lower* and the *B-upper approximation* of X, respectively.

Hence, the B-lower approximation of a concept is the union of all B-granules that are included in the concept, whereas the *B-upper* approximation of a concept is the union of

all B-granules that have a nonempty intersection with the concept. The set

$$BN_B(X) = B^*(X) - B_*(X)$$

will be referred to as the *B-boundary region* of X.

If the boundary region of X is the empty set, i.e., $BN_B(X) = \emptyset$, then X is *crisp* (*exact*) with respect to B; in the opposite case, i.e., if $BN_B(X) \neq \emptyset$, X is referred to as *rough* (*inexact*) with respect to B.

Rough sets can be also defined using a *rough membership function* [25], defined as

$$\mu_X^B(x) = \frac{card(B(x) \cap X)}{card(B(x))}.$$

Obviously

$$0 \leq \mu_X^B(x) \leq 1.$$

Value of the membership function $\mu_X^B(x)$ is a conditional probability $\pi(X|B(x))$, and can be interpreted as a degree of *certainty* to which x belongs to X (or $1 - \mu_X^B(x)$, as a degree of *uncertainty*).

The rough membership function can be generalized as follows [31]:

$$\mu(X, Y) = \frac{card(X \cap Y)}{card\, X},$$

where $X, Y \subseteq U, X \neq \emptyset$.

The function $\mu(X, Y)$ expresses the *degree to which X is included in Y*. Obviously, if $\mu(X, Y) = 1$, then $X \subseteq Y$.

If X is included to a degree k in X we will write $X \subseteq_k Y$, and say that X is *roughly included* in Y.

The rough inclusion can be also used to more general definition of approximations, which are defined below:

$$B_k(X) = \bigcup_{k \leq l \leq 1} \{x \in U : B(x) \subseteq_l X\},$$

$$B^k(X) = \bigcup_{0 < l \leq 1} \{x \in U : B(x) \subseteq_l X\},$$

where k $(0 < k \leq 1)$ is a treshold – and are called *k-lower* and *k-upper B-approximation* of X, respectively.

The *k-boundary B-region* of X is defined as

$$BN_B^k(X) = B^k(X) - B_k(X) = \bigcup_{0 < l \leq k} \{x \in U : B(x) \subseteq_l X\},$$

For $k = 1$ we obtain the previous definitions.

This generalization is a variety of variable precision rough set model proposed by Ziarko [68].

4 Dependency of attributes

Another important issue in data analysis is discovering dependencies between attributes. Suppose that the set of attributes A in a database $S = (U, A)$ is divided into two subsets C and D, called *condition* and *decision* attributes respectively, such that $C \cup D = A$ and $C \cap D = \emptyset$. Such databases are called *decision tables.*

Intuitively, a set of attributes D *depends totally* on a set of attributes C, denoted $C \Rightarrow D$, if all values of attributes from D are uniquely determined by values of attributes from C. In other words, D depends totally on C, if there exists a functional dependency between values of D and C.

We would need also a more general concept of dependency, called a *partial dependency* of attributes. Intuitively, the partial dependency means that only some values of D are determined by values of C.

Formally dependency can be defined in the following way. Let D and C be subsets of A.

We will say that D *depends on* C to a *degree* k $(0 \leq k \leq 1)$, denoted $C \Rightarrow_k D$, if

$$k = \gamma(C, D) = \frac{card(POS_C(D))}{card\,(U)},$$

where

$$POS_C(D) = \bigcup_{X \in U/D} C_*(X),$$

called a *positive region* of the partition U/D with respect to C, is the set of all elements of U that can be uniquely classified to blocks of the partition U/D, by means of C.

Obviously

$$\gamma(C, D) = \sum_{X \in U/D} \frac{card(C_*(X))}{card\,(U)}.$$

If $k = 1$ we say that D *depends totally* on C, and if $k < 1$, we say that D *depends partially* (to a *degree* k) on C, and if $k = 0$, D *does not depend on* C.

The coefficient k expresses the ratio of all elements of the universe, which can be properly classified to blocks of the partition U/D, employing attributes C and will be called the *degree of the dependency.*

For example in Table 1 the degree of dependency between the attribute P and the set of attributes $\{E, Q, L\}$ is $2/3$.

Obviously if D depends totally on C then $I(C) \subseteq I(D)$. That means that the partition generated by C is finer than the partition generated by D.

The function $\gamma(C, D)$ can be regarded as a generalization of the rough inclusion function $\mu(X, Y)$, for it expresses to what degree partition generated by C, i.e., U/C is included in the partition generated by D, i.e., U/D.

5 Reduction of attributes

A reduct is a minimal set of condition attributes that preserves the degree of dependency. It means that a reduct is a minimal subset of condition attributes that enables to make the same decisions as the whole set of condition attributes.

Formally if $C \Rightarrow_k D$ then a minimal subset C' of C, such that $\gamma(C,D) = \gamma(C',D)$ is called a *D-reduct* of C.

For example, in Table 1 we have two reducts $\{E,Q\}$ and $\{E,L\}$ of condition attributes $\{E,Q,L\}$. This means that Table 1 can be replaced either by Table 2 or Table 3.

Store	E	Q	P
1	*high*	*good*	*profit*
2	*med.*	*good*	*loss*
3	*med.*	*good*	*profit*
4	*no*	*avg.*	*loss*
5	*med.*	*avg.*	*loss*
6	*high*	*avg.*	*profit*

Table 2

Store	E	L	P
1	*high*	*no*	*profit*
2	*med.*	*no*	*loss*
3	*med.*	*no*	*profit*
4	*no*	*no*	*loss*
5	*med.*	*yes*	*loss*
6	*high*	*yes*	*profit*

Table 3

It is easy to check that both Table 2 and Table 3 preserve degree of dependency between attributes P and E, Q, L.

Reduction of attributes is the fundamental issue in rough set theory.

6 Decision rules

Let S be a decision table and let C and D be condition and decision attributes, respectively.

By Φ, Ψ etc. we will denote logicals formulas built from attributes, attribute-values and logical connectives (*and, or, not*) in a standard way. We will denote by $|\Phi|_S$ the set of all object $x \in U$ satisfying Φ and refer to as the *meaning* of Φ in S.

The expression $\pi_S(\Phi) = \frac{card(|\Phi|_S)}{card(U)}$ will denote the probability that the formula Φ is true in S.

A *decision rule* is an expression in the form *"if...then..."*, written $\Phi \rightarrow \Psi$; Φ and Ψ are refered to as *condition* and *decision* of the rule respectively.

A decision rule $\Phi \rightarrow \Psi$ is *admissible* in S if $|\Phi|_S$ is the union of some C-elementary sets, $|\Psi|_S$ is the union of some D-elementary sets and $|\Phi \wedge \Psi|_S \neq \emptyset$. In what follows we will consider admissible decision rules only.

Examples of decision rules admissible in Table 1 are given below:

1) *if* (*E, high*) *and* (*Q, good*) *and* (*L, no*) *then* (*P, loss*)

2) *if* (*E, med.*) *and* (*Q, avg.*) *then* (*P, loss*)

3) *if* (*Q, avg.*) *then* (*P, loss*)

With every decision rule $\Phi \rightarrow \Psi$ we associate a *certainty factor*

$$\pi_S(\Psi|\Phi) = \frac{card(|\Phi \wedge \Psi|_S)}{card(|\Phi|_S)},$$

which is the conditional probability that Ψ is true in S given Φ is true in S with the probability $\pi_S(\Phi)$.

Besides, we will also need a *coverage factor* [60]

$$\pi_S(\Phi|\Psi) = \frac{card(|\Phi \wedge \Psi|_S)}{card(|\Psi|_S)},$$

which is the conditional probability that Φ is true in S given Ψ is true in S with the probability $\pi_S(\Psi)$.

Let $\{\Phi_i \rightarrow \Psi\}_n$ be a set of decision rules such that all conditions Φ_i are pairwise mutally exclusive, i.e., $|\Phi_i \wedge \Phi_j|_S = \emptyset$, for any $1 \leq i, j \leq n$, $i \neq j$, and

$$\sum_{i=1}^{n} \pi_S(\Phi_i|\Psi) = 1.$$

Then the following property holds:

$$\pi_S(\Psi) = \sum_{i=1}^{n} \pi_S(\Psi|\Phi_i) \cdot \pi_S(\Phi_i). \qquad (*)$$

For any decision rule $\Phi \rightarrow \Psi$ the following property is true:

$$\pi_S(\Phi|\Psi) = \frac{\pi_S(\Psi|\Phi) \cdot \pi_S(\Phi)}{\sum_{i=1}^{n} \pi_S(\Psi|\Phi_i) \cdot \pi_S(\Phi_i)}. \qquad (**)$$

This relationship first was observed by Łukasiewicz [1, 21]. It can be easily seen that the relationship between the certainty factor and the coverage factor, expressed by the formula (**) is the Bayes' Theorem. However, the meaning of Bayes' Theorem in this case differs from that postulated in statistical inference. In statistical data analysis based on Bayes' Theorem, we assume that prior probability about some parameters without knowledge about the data is given. The posterior probability is computed next, which tells us what can be said about prior probability in view of the data. In the rough set approach the meaning of Bayes' Theorem is unlike. It reveals some relationships in the database, without referring to prior and posterior probabilities, and it can be used to reason about data in terms of approximate (rough) implications. Thus, the proposed approach can be seen as a new model for Bayes' Theorem, and offers a new approach to data analysis.

References

[1] An, A. Chan, C., Shan, N., Cercone, N., Ziarko, W.: Applying knowledge discovery to predict water-supply consumption. IEEE Expert 12/4 (1997) 72-78

[2] Arciszewski, T., Ziarko, W.: Inductive learning in civil engineering: rough sets approach. Microcomputers and Civil Engineering 5/1 (1990)

[3] Czogała, E., Mrózek, A., Pawlak, Z.: The idea of rough-fuzzy controller. International Journal of Fuzzy Sets and Systems 72 (1995) 61-63

[4] Jackson, A. G., Ohmer, M., Al-Kamhawi, H.: Rough sets analysis of chalcopyrite semiconductor band gap data. In: T. Y. Lin (ed.), The Third International Workshop on Rough Sets and Soft Computing Proceedings (RSSC'94), November 10-12, San Jose State University, San Jose, California, USA (1994) 408-417

[5] Jackson, A. G., Leclair, S. R., Ohmer, M. C., Ziarko. W., Al-Kamhwi, H.: Rough sets applied to material data. Acta Metallurgica et Materialia (1996) 4475

[6] Jackson, A. G., Pawlak, Z., Leclair, S. R.: Rough set and discovery of new materials. Journal of Alloys and Compounds (to appear)

[7] Kowalczyk, W.: Analyzing temporal patterns with rough sets. In: EUFIT-96: The fourth European Congress on Intelligent Techniques and Soft Computing, September 2-5, Aachen (1996) 139-143

[8] Lin, T. Y.: Fuzzy controllers: an integrated approach based on fuzzy logic, rough sets, and evolutionary computing. In: T. Y. Lin and N. Cercone (eds.), Rough Sets and Data Mining. Analysis for Imprecise Data, Kluwer Academic Publishers, Boston, London, Dordrecht (1997) 123-138

[9] Lin, T. Y., Cercone, N.,(eds.): Rough Sets and Data Mining (Analysis of Imperfect Data, Kluwer Academic Publishers, Boston, London, Dordrecht (1997) 430 Lingras, P.: Rough neural networks. Sixth International Conferences, Information Procesing and Management of Uncertainty in Knowledge-Based Systems, Proceedings (IPMU'96), Volume II, July 1-5, Grenada (1996) 1445-1450

[10] Mrózek, A.: Rough sets in computer implementation of rule-based control of industrial processes. In: R. Sowiski (ed.), Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory. Kluwer Academic Publishers, Boston, London, Dordrecht (1992) 19-31

[11] Munakata, T.: Rough control: a perspective. In: T. Y. Lin and N. Cercone (eds.), Rough Sets and Data Mining. Analysis for Imprecise Data. Kluwer Academic Publishers, Boston, London, Dordrecht (1997) 77-88

[12] Nowicki R., Słowiński R., Stefanowski J.: Possibilities of appling the rough sets theory to technical diagnostics. In: Proceedings of the IXth National Symposium on Vibration Techniques and Vibroacoustics, December 12-14, AGH University Press, Kraków (1990) 149-152

[13] Nowicki, R., Słowiński, R., Stefanowski, J.: Rough sets analysis of diagnostic capacity of vibroacoustic symptoms. Journal of Computers and Mathematics with Applications 24/2 (1992) 109-123

[14] Nowicki, R., Słowiński, R., Stefanowski, J.: Evaluation of vibroacoustic diagnostic symptoms by means of the rough sets theory. Journal of Computers in Industry 20 (1992) 141-152

[15] Oehrn, A.: Rough logic control. In: (Project), Technical Report. Knowledge Systems Group, The Norwegian University of Science and Technology, Trondheim, Norway (1993)

[16] E. Orłowska (ed.): Incomplete Information: Rough Set Analysis. Physica-Verlag, Heidelberg (1997)

[17] S.K. Pal, A. Skowron (eds.): Fuzzy Sets, Rough Sets and Decision Making Processes.Springer–Verlag, Singapore (in preparation)

[18] Pawlak, Z.: Rough Sets - Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Boston, London, Dordrecht (1991) 229.

[19] Pawlak, Z.: Rough set theory and its applications to data analysis. Cybernetics and Systems (to appear)

[20] Pawlak, Z., Grzymała-Busse, J. W., Słowiński, R., Ziarko, W.: Rough sets. Communication of the ACM 38 (1995) 88-95

[21] Pawlak, Z., Munakata, T.: Rough Control Application of rough set theory to control. Fourth European Congress on Intelligent Techniques and Soft Computing, Proceedings EUFIT'96, Volume I, September 2-5, Germany (1996) 209-218

[22] Pawlak, Z.: Reasoning about data (a rough set perspective. In: L. Polkowski, A. Skowron (eds.), Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence, 1424 Springer, First International Conference, RSCTC'98, Warsaw, Poland, June, Proceedings, (1998) 25-34

[23] L. Polkowski, A. Skowron (eds.): Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence, 1424 Springer, First International Conference, RSCTC'98, Warsaw, Poland, June, Proceedings, (1998)

[24] Słowiński, R.: Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory. Kluwer Academic Publishers, Boston, London, Dordrecht (1992)

[25] Sowiski, R.: Rough set approach to decision analysis. AI Expert 10 (1995) 18-25

[26] Sowiski, R.: Rough set theory and its applications to decision aid. Belgian Journal of Operation Research, Special Issue Francoro 35/3-4 (1995) 81-90

[27] Słowiński, R., Stefanowski, J., Susmaga, R.: Rough set analysis of attribute dependencies in technical diagnostics. In: S. Tsumoto, S. Kobayashi, T. Yokomori, H. Tanaka and A. Nakamura (eds.), The fourth International Workshop on Rough Sets, Fuzzy Sets, and Machnine Discovery, Proceedings (RS96FD), November 6-8, The University of Tokyo (1996) 284-291

[28] Stefanowski, J., Słowiński, R., Nowicki, R.: The rough sets approach to knowledge analysis for classification support in technical diagnostics of mechanical objects. In: F. Belli and F. J. Radermacher (eds.), Industrial & Engineering Applications of Artificial Intelligence and Expert Systems. Lecture Notes in Economics and Mathematical Systems 604, Springer-Verlag, Berlin (1992) 324-334

[29] Szladow, A., Ziarko W.: Adaptive process control using rough sets. Proceedings of the International Conference of Instrument Society of America, ISA/93, Chicago (1993) 1421-1430

[30] Szladow, A., Ziarko W.: Application of rough sets theory to process control. Proceedings of Calgary 93 Symposium of Instrument Society of America, Calgary (1993)

[31] Tsumoto, S., Kobayashi, S., Yokomori, T., Tanaka, H., Nakamura, A., (eds.): The Fourth Internal Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, PROCEEDINGS. The University of Tokyo (1996)

[32] Wang, P. P., (ed.): Second Annual Joint Conference on Information Sciences, PROCEEDINGS. Wrightsville Beach, North Carolina, USA (1995)

[33] Wang, P., (ed.): Joint Conference of Information Sciences, Vol. 3. Rough Sets and Computer Sciences, Duke University (1997)

[34] Ziarko, W.: Acquisition of control algorithms from operation data. In: R. Słowiński (ed.), Intelligent Decision Support, Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, Boston, London, Dordrecht (1992) 61-75

[35] Ziarko, W., (ed.): Rough Sets, Fuzzy Sets and Knowledge Discovery. Proceedings of the International Workshop on Rough Sets and Knowledge Discovery (RSKD'93), Banff, Alberta, Canada, October 12–15, Springer-Verlag, Berlin (1993)

[36] Ziarko, W., Katzberg, J.: Control algorithms acquisition, analysis and reduction: machine learning approach. In: Knowledge-Based Systems Diagnosis, Supervision and Control, Plenum Press, Oxford (1989) 167-178

[37] Ziarko, W., Katzberg, J.: Rough sets approach to system modelling and control algorithm acquisition. Proceedings of IEEE WESCANEX 93 Conference, Saskatoon (1993) 154-163

[38] Zak J., Stefanowski J.: Determining maintenance activities of motor vehicles using rough sets approach. In: Proceedings of Euromaintenance'94 Conference, Amsterdam (1994) 39-42