Data, Decision Rules and Rough Sets

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. 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. The theory has found many applications.

In the paper basic concepts of rough set theory will be presented and discussed. More about rough set theory and its applications can be found in the enclosed references.

2 The philosophy of rough sets

The starting point of rough set theory is the indiscernibility relation, first formulated by Leibniz and known as the *principle of the identity of indiscernibles*, which is the following:

"It all properties of objects are the same then the objects are identical".

This principle leads to clustering of elements of interest into *granules*, of indiscernible (similar) objects. In rough set theory these granules, called *elementary sets* (*concepts*) form basic building blocks of knowledge about the universe.

Every union of elementary concepts is referred to as a *crisp* or *precise* concept (set); otherwise a concept (set) is called *rough, vague* or *imprecise*. Thus rough concepts cannot be expressed in terms of elementary concepts. However, they can be expressed approximately by means of elementary concepts by employing the idea of the *lower* and the *upper approximation* of a concept. The lower approximation of a concept is the union of all elementary concepts which are included in the concept, whereas the upper approximation is the union of all elementary concepts which have nonempty intersection with the concept. In other words, the lower and the upper approximation of a concepts are the union of all elementary concepts which are surely and possibly included in the concept respectively. The difference between the lower and the upper approximation of the concept is its *boundary region*. Now it can easily be seen that a concept is rough if it has nonempty boundary region, i.e., its lower and upper approximation are nonidentical. Obviously, if the lower and the upper approximations of the concept are the same, i.e., its boundary region is empty – the concept is crisp.

3 Database

The basic ideas of rough set theory can be explained in general terms but in order to give more intuitive insight into the theory we will start our consideration from a database. Let us first give a simple example of database presented in Table 1.

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

Table 1

In the database six stores are characterized by four attributes, shown below:

E – empowerment of sales personnel,
Q – perceived quality of merchandise,
L – high traffic location,
P – store profit or loss.

Let us observe that each store has different description in terms of attributes E, Q, L and P, thus all stores may be distinguished (discerned) employing information provided by all attributes. However, stores 2 and 3 are indiscernible in terms of attributes E, Q and L, since they have the same values of these attributes. Similarly, stores 1, 2 and 3 are indiscernible with respect to attributes Q and L, etc.

Each subset of attributes determines a partition (classification) of all objects into classes having the same description in terms of these attributes. For example, attributes Q and L aggregate all stores into the following classes $\{1, 2, 3\}, \{4\}, \{5, 6\}$. Thus, each database determines a family of classification patterns which are used as a basis of further considerations.

Suppose we are interested in the following problem: what are the characteristic features of stores having profit (or loss) in view of information available in Table 1. In other words, the question is whether we are able to describe set (concept) $\{1, 3, 6\}$ (or $\{2, 4, 5\}$) in terms of attributes E, Q and L. Obviously this question cannot be answered uniquely in our case since stores 2 and 3 display the same features in terms of attributes E, Q and L, but store 2 makes a profit, whereas store 3 has a loss. Thus information given in Table 1 is not sufficient to answer this question. However, we can give a partial answer to this question. Let us observe that if the attribute E has the value *high* for a certain store, then the store makes a *profit*, whereas if the value of the attribute E is *low*, then the store has a loss. Thus, in view of information contained in Table 1, we can say for sure that stores 1 and 6 make a profit, stores 4 and 5 have a loss, whereas stores 2 and 3 cannot be classified as making a profit or having a loss. Therefore we can give approximate answers only. Employing attributes E, Q and L, we can say that stores 1 and 6 *surely* make a profit, i.e., *surely* belong to the set $\{1, 3, 6\}$, whereas stores 1,2,3 and 6 *possibly* make a profit, i.e., possibly belong to the set $\{1, 3, 6\}$. We will say that the set $\{1, 6\}$ is the *lower approximation* of the set (concept) $\{1, 3, 6\}$, and the set $\{1, 2, 3, 6\}$ – is the *upper approximation* of the set $\{1, 3, 6\}$. The set $\{2, 3\}$, being the difference between the upper approximation and the lower approximation is referred to as the *boundary region* of the set $\{1, 3, 6\}$. New let us give same formal notations and definitions.

By a *database* we will understand a pair $S = (U, A)$, where U and A, are finite, nonempty sets called the *universe*, and a set of *attributes* respectively. With every attribute $a \in A$ we associate a set V_a, of its *values*, 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*, and is defined as follows:

$(x, y) \in I(B)$ if and only if $a(x) = a(y)$ for every $a \in A$, where $a(x)$ denotes the value of attribute a for element x.

It can easily be seen that $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e. partition determined by B, will be denoted by $U/I(B)$, or simple U/B; an equivalence class of $I(B)$, i.e. block of the partition U/B, containing x will be denoted by $B(x)$.

If (x, y) belongs to $I(B)$ we will say that x and y are *B-indiscernible*. Equivalence classes of the relation $I(B)$ (or blocks of the partition U/B) are refereed to as *B-elementary sets*.

4 Approximations

Now we can define two basic operations in rough set theory:

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

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

which assign to every subset X of the universe U two sets $B_*(X)$ and $B^*(X)$ called the *B-lower* and the *B-upper approximation* of X, respectively.

The set

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

will be called the *B-boundary region* of X.

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

Rough sets can also be defined by using a *rough membership function*, defined as

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

Obviously

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

The value of the membership function $\mu_X^B(x)$ is a kind of conditional probability, and can be interpreted as a degree of *certainty* that x can be classified as X, employing set of attributes B.

5 Dependency of attributes

Another important issue in rough set theory is discovering dependencies between attributes in a database. 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. In Table 1 there are not total dependencies whatsoever.

We would need also a more general concept of dependency of attributes, called a *partial dependency* of attributes. The partial dependency means that only some values of D are determined by values of C. For example, in Table 1 the attribute P depends partially on attributes E, Q and L.

Formally, the above idea can be formulated as follows.

Let D and C be subsets of A. We say that D *depends to a degree* $k, 0 \leq k \leq 1$, on C, denoted $C \Rightarrow_k D$, if

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

where

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

For the dependency $\{E, Q, L\} \Rightarrow_k \{P\}$ in Table 1 we have $k = 4/6 = 2/3$.

If $k = 1$, then D depends on C totally; $0 < k < 1$ then depends partially on C, and if $k = 0$, then D does not depend on C.

6 Reduction of attributes

A reduct is a subset 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 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. 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.

7 Knowledge base

Each dependency $C \Rightarrow_k D$ in a database induces a set of decision rules of the form *"if... then..."*, called a *knowledge base.*

Decision rules are implications defined in the standard way from attributes, attribute values and propositional connectives.

Given $x \in U$ and $B \subseteq A$ by $\Phi_x^B = \bigwedge_{a \in B}(a, v)$ we mean a formula such that $a(x) = v$ and $v \in V_a$.

Every dependency $C \Rightarrow_k D$ determines a set of *decision rules*

$$\{\Phi_x^C \to \Phi_x^D\}_{x \in U},$$

called *knowledge base.* For example for Table 1 we get the following decision rules (knowledge base):

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

2) *if* (E, *med.*) *and* (Q, *good*) *and* (L, *no*) *then* (P, *loss*),

3) *if* (E, *med.*) *and* (Q, *good*) *and* (L, *no*) *then* (P, *profit*),

4) *if* (E, *no*) *and* (Q, *avg.*) *and* (L, *no*) *then* (P, *loss*),

5) *if* (E, *med.*) *and* (Q, *avg.*) *and* (L, *yes*) *then* (P, *loss*),

6) *if* (E, *high*) *and* (Q, *avg.*) *and* (L, *yes*) *then* (P, *profit*).

This set of decision rules describes some relationships hidden in Table 1 (the database), i.e., the partial dependency between condition and decision attributes. It can be also understood as a description of approximations of decisions in terms of conditions. Thus a knowledge base can be seen as a language used to express our knowledge about the universe, hidden in the database.

8 Credibility factor of a decision rule

With every decision rule $\Phi_x^C \to \Phi_y^C$ we associate a *credibility factor* defined below:

$$\mu(\Phi_x^C, \Psi_x^D) = \frac{card(|\Phi_x^C \wedge \Psi_x^D|_S)}{card\,|\Phi_x^C|_S}.$$

There exists an interesting relationship expressed by the following formula:

$$\pi_S(\Psi_x^D) = \sum_{y \in \hat{D}(x)} (\pi_S(\Phi_y^C) \cdot \mu_S(\Phi_y^C, \Psi_y^D)) = \sum_{y \in \hat{D}(x)} \pi_S(\Phi_y^C \wedge \Psi_y^D) \qquad (*)$$

where $\hat{D}(x)$ is the set of representatives of $D(x)/I(C)$, and $\pi_S(\Phi_x^C) = \frac{card(|\Phi_x^C|_S)}{card(U)}$ and $\pi_S(\Phi_x^C)$ is a probability that the formula Φ_x^C is satisfied in S.

Thus formula (*) shows the relationship between the probability of conditions, certainty factor of the decision rule and the probability of decisions.

Let us notice that the credibility factor can be viewed as a counterpart of a truth value of an implication in classical logic and thus any decision rule is in a sense an inference rule used to reason about properties of data.

9 Conclusions

Basic concepts of rough set theory are approximations, dependencies and decision rules. In fact approximations and dependencies are mutually exchangeable concepts used to express our imperfect knowledge about reality. Approximations express local properties of a database, i.e., they describe particular decisions in terms of conditions - whereas dependencies (partial or total) are used to describe global properties of a database, i.e., general relationships in the database. Decision rules form a description language for both approximations as well as dependencies.

Many real life-application of rough set theory have shown its usefulness in many domains. Medicine, pharmacology, banking, market research, engineering, speech recognition, material sciences, control, linguistic, databases are just some exemplary fields of application, where rough set theory has demonstrated its usefulness.

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

Despite of many achievements a wide variety of problems still requires consideration and extensive research. Particularly important seems to be hardware implementation of rough computer, which is badly needed to pursue many applications in the years to come.

About two thousand papers have been published so far about rough set theory and its application. Some information about the present state of rough set theory and some applications can be found in the references.

References

[1] Grzymała-Busse, J.: On the reduction of knowledge representation systems. In: Proc. of the 6th International Workshop on Expert Systems and their Applications **1**, Avignon, France, April 28–30 (1986) 463–478

[2] Grzymała-Busse, J.: Knowledge acquisition under uncertainty – a rough set approach. Journal of Intelligent and Robotics Systems **1** (1988) 3–16

[3] Hu, X., Cercone, N.: Mining knowledge rules from databases: A rough set approach. In: Proceedings of the 12th International Conference on Data Engineering, New Orleans (1995) 96–105

[4] Hu, X., Cercone, N., Ziarko, W.: Generation of multiple knowledge from databases based on rough set theory. In: T.Y. Lin, N. Cercone (eds.): Rough Sets and Data Mining. Analysis of Imprecise Data. Kluwer Academic Publishers, Boston, Dordrecht (1997) 109–121

[5] Hu, X., Shan, N., Cercone, N., Ziarko, W.: DBROUGH: A rough set based knowledge discovery system. In: Z.W. Ras, M. Zemankova (eds.), Proceedings of the Eighth International Symposium on Methodologies for Intelligent Systems (ISMIS'94), Charlotte, NC, October 16–19, 1994, Lecture Notes in Artificial Intelligence **869**, Springer-Verlag (1994) 386–395

[6] Komorowski, J., Polkowski, L., Skowron, A.: Rough sets for data mining and knowledge discovery (abstract of tutorial). In: J. Komorowski, J. Żytkow (eds.), First European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'97), June 25–27, Trondheim, Norway, Lecture Notes in Artificial Intelligence **1263**, Springer-Verlag, Berlin (1997) 393–393

[7] Kowalczyk, W., Piasta, Z.: Rough sets–inspired approach to knowledge discovery in business databases. In: The Second Pacific–Asian Conference on Knowledge Discovery and Data Mining, (PAKDD'98), Melbourne, Australia, April 15–17 (1998) (accepted)

[8] T.Y. Lin (ed.): Proceedings of the Third International Workshop on Rough Sets and Soft Computing (RSSC'94). San Jose State University, San Jose, California, USA, November 10–12 (1994)

[9] T.Y. Lin (ed.): Proceedings of the Workshop on Rough Sets and Data Mining at 23rd Annual Computer Science Conference. Nashville, Tennessee, March 2 (1995)

[10] T.Y. Lin (ed.): Journal of the Intelligent Automation and Soft Computing **2/2** (1996) (special issue)

[11] T.Y. Lin (ed.): International Journal of Approximate Reasoning **15/4** (1996) (special issue)

[12] T.Y. Lin, N. Cercone (eds.): Rough Sets and Data Mining. Analysis of Imprecise Data. Kluwer Academic Publishers, Boston, Dordrecht (1997)

[13] T.Y. Lin, A.M. Wildberger (eds.): Soft Computing: Rough Sets, Fuzzy Logic, Neural Networks, Uncertainty Management, Knowledge Discovery. Simulation Councils, Inc., San Diego, CA (1995)

[14] Mrózek, A., Płonka, L.: Knowledge representation in fuzzy and rough controllers. In: M. Dąbrowski, M. Michalewicz, and Z.W. Ras (eds.), Proceedings of the Third

International Workshop on Intelligent Information Systems, Wigry, Poland, June 6–10, 1994, Institute of Computer Science Polish Academy of Sciences, Warsaw (1994) 324–337

[15] Munakata, T.: Rough control: A perspective. In: T.Y. Lin, N. Cercone (eds.): Rough Sets and Data Mining. Analysis of Imprecise Data. Kluwer Academic Publishers, Boston, Dordrecht (1997) 77–88

[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, Dordrecht (1991)

[19] Pawlak, Z., Skowron, A.: Rough membership functions. In: R.R. Yaeger, M. Fedrizzi, and J. Kacprzyk (eds.), Advances in the Dempster Shafer Theory of Evidence, John Wiley & Sons, Inc., New York (1994) 251–271

[20] 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

[21] Płonka, L., Mrózek, A., Winiarczyk, R., Maitan, J.: Implementing rule–oriented knowledge bases on smart networks. In: Proceedings of the Fourth International Workshop on Intelligent Information Systems, Augustów, Poland, June 5–9, 1995, Institute od Computer Science, Polish Academy of Sciences, Warsaw (1995)

[22] 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)

[23] R. Słowiński (ed.): Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Boston, Dordrecht (1992)

[24] R. Słowiński, J. Stefanowski (eds.), Foundations of Computing and Decision Sciences **18/3–4** (1993) 155–396 (special issue)

[25] Swiniarski, R., Berzins, A.: Rough sets for intelligent data mining, knowledge discovering and designing of an expert systems for on–line prediction of volleyball game progress. In: S. Tsumoto, S. Kobayashi, T. Yokomori, H. Tanaka, and A. Nakamura (eds.): Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD'96). The University of Tokyo, November 6–8 (1996) 413–418

[26] Szladow, A.J., Ziarko, W.: Knowledge–based process control using rough sets. In: R. Słowiński (ed.): Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Boston, Dordrecht (1992) 49–60

[27] S. Tsumoto (ed.): Bulletin of International Rough Set Society **1/1** (1996)

[28] S. Tsumoto (ed.): Bulletin of International Rough Set Society **1/2** (1997)

[29] S. Tsumoto (ed.): Bulletin of International Rough Set Society **2/1** (1998)

[30] S. Tsumoto, S. Kobayashi, T. Yokomori, H. Tanaka, and A. Nakamura (eds.): Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD'96). The University of Tokyo, November 6–8 (1996)

[31] Tsumoto, S., Tanaka, H.: Extraction of medical diagnostic knowledge based on rough set based model selection and rule induction. In: S. Tsumoto, S. Kobayashi, T. Yokomori, H. Tanaka, and A. Nakamura (eds.): Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD'96). The University of Tokyo, November 6–8 (1996) 426–436; see also: Y.-Y. Chen, K. Hirota, and J.-Y. Yen (eds.), Proceedings of 1996 ASIAN FUZZY SYSTEMS SYMPOSIUM – Soft Computing in Intelligent Systems and Information Processing, December 11–14, Kenting, Taiwan, ROC. (1996) 145–151

[32] Tsumoto, S., Tanaka H.: Domain knowledge from clinical databases based on rough set model. In: P. Borne, G. Dauphin–Tanguy, C. Sueur, and S. El Khattabi (eds.), Proceedings of IMACS Multiconference: Computational Engineering in Systems Applications (CESA'96) **3/4** July 9–12, Lille, France, Gerf EC Lille – Cite Scientifique (1996) 742–747

[33] Tsumoto, S., Ziarko, W., Shan, N., Tanaka, H.: Knowledge discovery in clinical databases based on variable precision rough sets model. In: Proceedings of the 19th Annual Symposium on Computer Applications in Medical Care, New Orleans, 1995, Journal of American Medical Informatics Association Supplement (1995) 270–274

[34] P.P. Wang (ed.): Proceedings of the International Workshop on Rough Sets and Soft Computing at Second Annual Joint Conference on Information Sciences (JCIS'95), Wrightsville Beach, North Carolina, 28 September - 1 October (1995)

[35] P.P. Wang (ed.): Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at Third Annual Joint Conference on Information Sciences (JCIS'97). Duke University, Durham, NC, USA, Rough Set & Computer Science **3**, March 1–5 (1997)

[36] Ziarko, W.: On reduction of knowledge representation. In: Proc. 2nd International Symp. on Methodologies of Intelligent Systems, Charlotte, NC, North Holland (1987) 99–113

[37] Ziarko, W.: Acquisition of design knowledge from examples. Math. Comput. Modeling **10** (1988) 551–554

[38] Ziarko, W.: Determination of locally optimal set of features for representation of implicit knowledge. In: Proceedings of International Conference on Computing and Information, Toronto, North Holland (1989) 433–438

[39] Ziarko, W.: Rough sets and knowledge discovery: An overview. In: W. Ziarko (ed.): Rough Sets, Fuzzy Sets and Knowledge Discovery (RSKD'93). Workshops in Computing, Springer-Verlag & British Computer Society, London, Berlin (1994) 11–15

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

[41] W. Ziarko (ed.): Rough Sets, Fuzzy Sets and Knowledge Discovery (RSKD'93). Workshops in Computing, Springer-Verlag & British Computer Society, London, Berlin (1994)

[42] Ziarko, W.: Introduction to the special issue on rough sets and knowledge discovery. In: W. Ziarko (ed.): Computational Intelligence: An International Journal **11/2** (1995) (special issue) 223–226

[43] Ziarko, W., Shan, N.: Knowledge discovery as a search for classifications. In: T.Y. Lin (ed.): Proceedings of the Workshop on Rough Sets and Data Mining at 23rd Annual Computer Science Conference. Nashville, Tennessee, March 2 (1995) 23–29

[44] Ziarko, W., Shan, N.: Rough sets and knowledge discovery. Encyclopedia of Computer Science and Technology, Marcell Dekker Inc. **35**, supplement **20** (1996) 369–379