Data Mining – a Rough Set Perspective

Zdzisław Pawlak

Institute for Theoretical and Applied Informatics
Polish Academy of Sciences
Bałtycka 5, 44 000 Gliwice, Poland

1 Introduction

Data mining (DM) can be perceived as a methodology for discovering hidden patterns in data. DM is a relatively new area of research and applications, stretching, over many domains like statistics, machine learning, fuzzy sets, rough sets, cluster analysis, genetics algorithms, neural networks and others. Despite many various techniques employed in DM yet it can be seen as a distinct discipline with its own problems and aims.

Reasoning methods associated with discovering knowledge from data attracted attention of philosophers for many years. Particularly some ideas of B. Russell and K. Popper about data, induction and experimental knowledge can be viewed as precursory ones for DM.

Many valuable papers and books have been published on data mining recently. In this paper we will focus our attention on some problems pertinent to rough sets and DM [2, 3, 5, 6, 7, 8, 9, 11, 14, 15, 16, 19, 24, 32, 33, 36, 37].

Rough set theory has proved to be useful in DM, and it "... constitutes a sound basis for data mining applications" [4]. The theory offers mathematical tools to discover hidden patterns in data. It identifies partial or total dependencies (i.e. cause-effect relations) in databases, eliminates redundant data, gives approach to null values, missing data, dynamic data and others. The methods of data mining in large databases using rough sets have recently been proposed and investigated [5, 14, 16].

The theory is based on sound mathematical foundation. It can easily be understood and applied. Several software systems based on rough set theory have been implemented and many nontrivial applications of this methodology for knowledge discovery have been reported. More about rough sets and their applications can be found in [19].

The theory is not competitive but complementary to other methods and can also be often used jointly with other approaches (e.g. fuzzy sets, genetic algorithms, statistical methods, neural networks etc.).

The main objective of this talk is to give basic ideas of rough sets in the context of DM. The starting point of rough set theory is a data set. The theory can also be formulated in more general terms, however for the sake of intuition we will refrain from general formulation. Data are usually organized in the form of a table, columns of which are labeled by attributes, rows – by objects and entries of the table are attribute values. Such a table will be called a database. Next, basic operations on sets in rough set theory, the lower and the upper approximation of a set will be defined. These operations will be used to define the basic concepts of the theory (from the DM point of view) – total and partial dependency of attributes in a database. The concept of dependency of attributes is used to describe cause-effect relations hidden in the data. Further, a very important issue, reduction of data, will be introduced. Finally certain and possible decision rules determined by total and partial dependencies will be defined and analyzed. Besides, certainty and coverage factors of a decision rule will be defined and reasoning methods based on the idea outlined.

2 Database

An example of a simple database is presented in Table 1.

Table 1. An example of a database

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*

In the database six stores are characterized by four attributes:

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

Each store is described in terms of attributes E, Q, L and P.

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.

Formally a database will be defined as follows.

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 referred to as *B-elementary sets* or *B-granules*.

Equivalence relation as a basis for rough set theory for many applications is not sufficient. Therefore other relations e.g. a tolerance relation, an ordering relations and others, have been proposed, e.g. [21, 23, 31]. But for the sake of simplicity in this paper we will stick to the equivalence relation as a basis for rough set theory.

3 Approximations of Sets

First let us consider the following exmaple: what are the characteristic features of stores having profit (or loss) in view of information available in Table 1. It can easily be seen that this question cannot be answered uniquely 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. 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. 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}.

Approximations can be defined formally as operations 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 and defined as follows:

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

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

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.

"Roughness" of a set can be also characterized numerically as

$$\alpha_B(X) = \frac{card(B_*(X))}{card(B^*(X))},$$

where $0 \leq \alpha_B(X) \leq 1$ and if $\alpha_B(X) = 1$, X is crisp with respect to B, whereas if $\alpha_B(X) < 1$, X is rough with respect to B.

Rough sets can be also defined using a *rough membership function* [17], 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*).

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 C and D.

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.

Dependency is strictly related with approximations and is the basic issue in data mining, because it reveals relationships in a database.

Formally, dependency can be defined in the following way. Let C and D 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

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

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 $k = 0$, then 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.

5 Reduction of Attributes

A reduct is the 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 $\{E, Q, L\}$.

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

In large databases computation of reducts on the basis of the given definition is not a simple task and therefore many more effective methods have been proposed. For references see [19].

6 Significance of Attributes

The concept of a reduct enables us to remove some attributes in the database in such a way that the basic relationships in the database are preserved. Some attributes, however, cannot be removed from the database without changing their properties. To express this idea more precisely we will need the notion of *significance of an attribute*, which is defined next.

Let C and D be sets of condition and decision attributes respectively and let a be a condition attribute, i.e. $a \in C$. We can ask how the coefficient $\gamma(C, D)$ changes when removing the attribute a, i.e. what is the difference between $\gamma(C,D)$ and $\gamma(C-\{a\},D)$. We can normalize the difference and define the significance of an attribute a as:

$$\sigma_{(C,D)}(a) = \frac{(\gamma(C,D) - \gamma(C-\{a\},D))}{\gamma(C,D)} = 1 - \frac{\gamma(C-\{a\},D)}{\gamma(C,D)},$$

and denote simply by $\sigma(a)$, when C and D are understood.

Thus the coefficient $\sigma(a)$ can be understood as an error which occurs when attribute a is dropped. The significance coefficient can be extended to set of attributes as follows:

$$\sigma_{(C,D)}(B) = \frac{(\gamma(C,D) - \gamma(C-B,D))}{\gamma(C,D)} = 1 - \frac{\gamma(C-B,D)}{(\gamma C,D)},$$

denoted by $\sigma(B)$, if C and D are understood, where B is a subset of C.

If B is a reduct of C, then $\sigma(B) = 1$, i.e. after removing any reduct from the set of decision rules one cannot make sure decisions, whatsoever.

7 Decision Rules

Dependences between attributes are usually symbolized as a set of decision rules. For example, decision rules describing the dependency $\{E,Q\} \Rightarrow \{P\}$ in Table 1 are the following:

(*E*, *high*) *and* (*Q*, *good*) → (*P*, *profit*),
(*E*, *med.*) *and* (*Q*, *good*) → (*P*, *loss*),
(*E*, *med.*) *and* (*Q*, *good*) → (*P*, *profit*),
(*E*, *no*) *and* (*Q*, *avg.*) → (*P*, *loss*),
(*E*, *med.*) *and* (*Q*, *avg.*) → (*P*, *loss*),
(*E*, *high*) *and* (*Q*, *avg.*) → (*P*, *profit*).

A set of decision rules is usually referred as a *knowledge base*.

Usually we are interested in the optimal set of decision rules associated with the dependency, but we will not consider this issue here. Instead we will analyze some probabilistic properties of 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 logical formulas built up from attributes, attribute-values and logical connectives (*and, or, not*) in a standard way. We will denote by $|\Phi|_S$ the set of all objects $x \in U$ 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 referred to as *condition* and *decision* of the rule, respectively.

A decision rule $\Phi \to \Psi$ 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.

With every decision rule $\Phi \to \Psi$ 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* [26]

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

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

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

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

Then the following properties hold:

$$\pi_S(\Psi) = \sum_{i=1}^{n} \pi_S(\Psi|\Phi_i) \cdot \pi_S(\Phi_i), \tag{2}$$

$$\pi_S(\Phi|\Psi) = \frac{\pi_S(\Psi|\Phi) \cdot \pi_S(\Phi)}{\sum_{i=1}^{n} \pi_S(\Psi|\Phi_i) \cdot \pi_S(\Phi_i)}. \tag{3}$$

It can be easily seen that the relationship between the certainty factor and the coverage factor, expressed by the formula (3) is the Bayes' theorem [1]. The theorem enables us to discover relationships in the databases.

8 Conclusions

Data mining is the quest for knowledge in databases. Many methods have been proposed for knowledge discovery in databases. No doubt rough sets proved to be a valuable methodology for data mining. Some advantages of rough set theory in this context are listed below:

- provides efficient algorithms for finding hidden patterns in data
- finds minimal sets of data (data reduction)
- evaluates significance of data
- generates minimal sets of decision rules from data
- it is easy to understand and offers straightforward interpretation of results

The rough set approach to data mining is not competive to other methods but rather complementary and can be also used jointly with other approaches.

References

1. Borkowski L. (Ed.) (1970) Jan Łukasiewicz – Selected Works, North Holland Publishing Company, Amsterdam, London, Polish Scientific Publishers, Warszawa
2. Cercone N., Ziarko W. (1997) GRS: A Generalized Rough Set Model for Data Mining Applications. In: Wang P.P. (Ed.) Joint Conference of Information Sciences, March 1-5, Duke University 3, 181–186
3. Cios J., Pedrycz H., Swiniarski R. (1998) Data Mining for Knowledge Discovery, Kluwer Academic Publishers
4. Deogun J.S., Raghavan V.V., Sarkar A., Sever H. (1997) Data Mining: Trends in Research and Development. In: Lin T.Y., Cercone N. (Eds.) Rough Sets and Data Mining, Kluwer Academic Publishers, 9–45
5. Fernandes-Baizan M.C., Menasalvas Ruiz E., Castano M. (1996) Integrating RDMS and Data Mining Capabilities using Rough Sets. In: Proceedings of the Sixth International Conference, Information Processing and Management of Uncertainty in Knowledge Based Systems (IPMU'96) 2, July 1-5, Granada, Spain, 1439–1445.
6. Grzymała-Busse J. (1986) 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, 463–478
7. Hu X., Cercone N., Ziarko W. (1997) Generation of Multiple Knowledge from Databases Based on Rough Set Theory. In: Lin T.Y., Cercone N. (Eds.) Rough Sets and Data Mining. Analysis of Imprecise Data. Kluwer Academic Publishers, Boston, Dordrecht, 109–121
8. Hu X., Shan N., Cercone N., Ziarko W. (1994) DBROUGH: A Rough Set Based Knowledge Discovery System. In: Ras Z. W., Zemankova M. (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, 386–395
9. Kowalczyk W., Piasta Z. (1998) 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 (accepted)
10. Lin T. Y. (Ed.) (1994) 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
11. Lin T. Y. (Ed.) (1995) Proceedings of the Workshop on Rough Sets and Data Mining at 23rd Annual Computer Science Conference. Nashville, Tennessee, March 2
12. Lin T. Y. (Ed.) (1996) Journal of the Intelligent Automation and Soft Computing 2/2 (special issue)
13. Lin T. Y. (Ed.) (1996) International Journal of Approximate Reasoning 15/4(special issue)
14. Lin T. Y. (1996) Rough Set Theory in very Large Databases. In: Borne P., Dauphin-Tanguy G., Suer C., Khattabi S. El., (Eds.) Proceedings of IMACS Multiconference: Computational Engineering in Systems Applications (CESA'96) 3/4, July 9-12. Lille, France, Gerf EC Lille-Cite Scientifique, 936–941.

15. Lin T. Y., Cercone N. (Eds.) (1997) Rough Sets and Data Mining. Kluwer Academic Publishers
16. Millan M., Machuca F. (1997) Using Rough Set Theory to Exploit the Data Mining Potential in Relational Databases Systems. In: Wang P. P. (Ed.) Joint Conference of Information Sciences 3, March 1-5, Duke University, 344–347
17. Pawlak, Z., Skowron, A. (1994) Rough Membership Functions. In: Yaeger R. R., Fedrizzi M., Kacprzyk J. (Eds.) Advances in the Dempster Shafer Theory of Evidence, John Wiley & Sons, Inc., New York 251–271
18. Pawlak, Z. (1998) Reasoning about Data – a Rough Set Perspective. In: Polkowski L., Skowron A. (Eds.) Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence 1424 Springer, First International Conference, RSCTC'98, Warsaw, Poland, June, Proceedings, 25–34
19. Polkowski L., Skowron A., (Eds.) Rough Sets in Knowledge Discovery, Methodology and Applications, Vol. 1,2, Physica -Verlag, A Springer Verlag Company
20. Polkowski L., Skowron A. (Eds.) (1998) Rough Sets and Current Trends. In Computing, Lecture Notes in Artificial Intelligence 1424 Springer, First International Conference, RSCTC'98, Warsaw, Poland, June, Proceedings,
21. Skowron A., Stepaniuk J. (1994) Generalized Approximations Spaces. Proceedings of the Third International Workshop on Rough Sets and Soft Computing, San Jose, November 10–12, 156–163
22. Słowiński R. (Ed.) (1992) Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Boston, Dordrecht
23. Słowiński R. (1992) A generalization of the indiscernibility relation for rough set analysis of quantitative information. Rivista di Matematica per le Scienze Economiche e Sociali 15/1: 65–78
24. Swiniarski R., Berzins A. (1996) Rough Sets for Intelligent Data Mining, Knowledge Discovering and Designing of an Expert Systems for on–line Prediction of Volleyball Game Progress. In: Tsumoto S., Kobayashi S., Yokomori T., Tanaka H., Nakamura A. (Eds.) Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD'96). The University of Tokyo, November 6–8, 413–418
25. Tsumoto S., Kobayashi S., Yokomori T., Tanaka H., Nakamura A. (Eds.) (1996) Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD'96). The University of Tokyo, November 6–8
26. Tsumoto, S. (1998) Modelling Medical Diagnostic Rules Based on Rough Sets. In: Polkowski L., Skowron A. (Eds.) Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence 1424, Springer, First International Conference, RSCTC'98, Warsaw, Poland, June, Proceedings, 475–482
27. Wang P. P. (Ed.) (1995) 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
28. Wang P. P. (ed.) (1997) 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
29. Ziarko, W. (1987) On reduction of knowledge representation. In: Proc. 2nd International Symp. on Methodologies of Intelligent Systems, Charlotte, NC, North Holland, 99–113

30. Ziarko W. (Ed.) (1993) Proceedings of the Second International Workshop on Rough Sets and Knowledge Discovery (RSKD'93). Banff, Alberta, Canada, October 12–15
31. Ziarko W. (1993) Variable Precision Rough Set Model. Journal of Computer and System Sciences 46/1:39–59
32. Ziarko W. (1994) Rough Sets and Knowledge Discovery: An overview. In: Ziarko W. (Ed.) Rough Sets, Fuzzy Sets and Knowledge Discovery (RSKD'93). Workshops in Computing, Springer-Verlag & British Computer Society, London, Berlin, 11–15
33. Ziarko W. (Ed.) (1994) Rough Sets, Fuzzy Sets and Knowledge Discovery (RSKD'93). Workshops in Computing, Springer-Verlag & British Computer Society, London, Berlin
34. Ziarko W. (1995) Introduction to the Special Issue on Rough Sets and Knowledge Discovery. In: Ziarko W. (Ed.) Computational Intelligence: An International Journal 11/2:223–226 (special issue)
35. Ziarko W., Shan N. (1995) Knowledge Discovery as a Search for Classifications. In: Lin T. Y. (Ed.) Proceedings of the Workshop on Rough Sets and Data Mining at 23rd Annual Computer Science Conference. Nashville, Tennessee, March 2 23–29
36. Ziarko W., Shan N. (1996) Rough Sets and Knowledge Discovery. Encyclopedia of Computer Science and Technology, Marcell Dekker Inc. 35, supplement 20:369–379
37. Ziarko W. (1996) Review of the Basics of Rough Sets in the Context of Data Mining. The Fourth Intl. Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, Tokyo, 447–457