ROUGH SETS AND DATA MINING

Zdzisław Pawlak

Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Baltycka 5, 44 100 Gliwice, Poland

ABSTRACT

The paper gives basic ideas of rough set theory - a new approach to vague data analysis. The lower and the upper approximation of a set the basic operations of the theory, are intuitively explained and formally defined. Some applications of rough set theory are briefly outline and some future problems pointed out.

INTRODUCTION

Rough set theory [7] is a new mathematical approach to data analysis and data mining. After 15 year of pursuing rough set theory and its application the theory has reached a certain degree of maturity. In recent years we witnessed a rapid grow of interest in rough set theory and its application, world wide. Many international workshops, conferences and seminars included rough sets in their programs. A large number of high quality papers have been published recently on various aspects of rough sets.

The connection of rough set theory and many other theories has been clarified. Particularly interesting is the relationship between fuzzy set theory and Dempster-Shafer theory of evidence. The concepts of rough set and fuzzy set are different since they refer to various aspects of imprecision [10], whereas the connection with theory of evidence is more substantial [12]. Besides, rough set theory is related to discriminant analysis [4], Boolean reasoning methods [13] and others. The relationship between rough set theory and decision analysis is presented in [11,15]. Several extension of the "basic" model of rough set have been proposed and investigated.

Various real life-applications of rough set theory have shown its usefulness in many domains. 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. No doubt that rough set theory can contribute essentially to material sciences, a subject of special interest to this conference.

BASIC CONCEPTS

Rough set philosophy is founded on the assumption that with every object of the universe of discourse we associate some information (data, knowledge). Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory.

Any set of all indiscernible (similar) objects is called an elementary set, and form a basic granule (atom) of knowledge about the universe. Any union of some elementary sets is referred to as a crisp (precise) set - otherwise the set is rough (imprecise, vague).

Each rough set has boundary-line cases, i.e., objects which cannot be with certainty classified, by employing the available knowledge, as members of the set or its complement.

Obviously rough sets, in contrast to precise sets, cannot be characterized in terms of information about their elements. In the proposed approach with any rough set a pair of precise sets - called the lower and the upper approximation of the rough set is associated. The lower approximation consists of all objects which surely belong to the set and the upper approximation contains all objects which possible belong to the set. The difference between the upper and the lower approximation constitutes the boundary region of the rough set. Approximations are two basic operations used in rough set theory.

Data are often presented as a table, columns of which are labeled by *attributes*, rows by *objects* of interest and entries of the table are *attribute values*. Such tables are known as *information systems, attribute-value tables*, *data tables* or *information tables*.

Usually we distinguish in information tables two kinds of attributes, called *condition* and *decision* attributes. Such tables are known as *decision tables.* Rows of a decision table are referred to as "if...then..." *decision rules*, which give conditions necessary to make decisions specified by the decision attributes. An example of a decision table is shown in Table 1.

Pipe	*C*	*S*	*P*	*Cracks*
1	*high*	*high*	*low*	*yes*
2	*avg.*	*high*	*low*	*no*
3	*avg.*	*high*	*low*	*yes*
4	*low*	*low*	*low*	*no*
5	*avg.*	*low*	*high*	*no*
6	*high*	*low*	*high*	*yes*

Table 1

The table contains data concerning six cast iron pipes exposed to high pressure endurance test. In the table *C*, *S* and *P* are condition attributes, displaying the percentage content in the pig-iron of coal, sulfur and phosphorus respectively, whereas the attribute *Cracks* revels the result of the test. The values of condition attributes are as follows (C, high) > 3.6%, 3.5% ≤ (C, avg.) ≤ 3.6%, (C, low) < 3.5%, (S, high) ≥ 0.1%, (S, low) < 0.1%, (P, high) ≥ 0.3%, (P, low) < 0.3%.

Main problem we are interested in is how the endurance of the pipes depend on the compounds *C*, *S* and *P* comprised in the pig-iron, or in other words, if there is a functional dependency between the decision attribute *Cracks* and the condition attributes *C, S* and *P.* In the rough set theory language this boils down to the question, if the set {2,4,5}of all pipes having no cracks after the test (or the set {1,3,6}of pipes having cracks), can be uniquely defined in terms of condition attributes values.

It can be easily seen that this is impossible, since pipes 2 and 3 display the same features in terms of attributes *C*, *S* and *P*, but they have different values of the attribute *Cracks*. Thus information given in Table 1 is not sufficient to solve our problem. However we can give a partial solution. Let us observe that if the attribute *C* has the value *high* for a certain pipe, then the pipe have cracks, whereas if the value of the attribute *C* is *low*, then the pipe has no cracks. Hance employing attributes *C*, *S* and *P*, we can say that pipes 1 and 6 *surly* are good, i.e., *surely* belong to the set {1, 3, 6}, whereas pipes 1, 2, 3 and 6 *possible* are good, i.e., possible belong to the set {1, 3, 6}.Thus the sets {1, 6}, {1, 2, 3, 6} and {2, 3} are the lower, the upper approximation and the boundary region of the set {1, 3, 6}.

This means that the quality of pipes cannot be determined exactly by the content of coal, sulfur and phosphorus in the pig-iron, but can be determined only with some approximation.

In fact approximations determine the dependency (total or partial) between condition and decision attributes, i.e., express functional relationship between values of condition and decision attributes.

The degree of dependency between condition and decision attributes can be defined as a *consistency factor* of the decision table, which is the number of conflicting decision rules to all decision

rules in the table. By conflicting decision rules we mean rules having the same conditions but different decisions. For example, the consistency factor for Table 1 is 4/6 = 2/3, hence the degree of dependency between cracks and the composition of the pig-iron is 2/3. That means that four out of six (ca. 60%) pipes can be properly classified as good or not good on the basis of their composition.

We might be also interested in reducing some of the condition attributes, i.e. to know whether all conditions are necessary to make decisions specified in a table. To this end we will employ the notion of a *reduct* (of condition attributes). By a reduct we understand a minimal subset of condition attributes which preserves the consistency factor of the table. It is easy to compute that in Table 1 we have two reducts $\{C, S\}$ and $\{C, P\}$. Intersection of all reducts is called the *core.* In our example the core is the attribute *C*.

That means that in view of the data coal is the most important factor causing cracks and cannot be eliminated from our considerations, whereas sulfur and phosphorus play a minor role and can be mutually exchanged as factors causing cracks.

Now we present the basic concepts more formally.

Suppose we are given two finite, non-empty sets *U* and *A*, where *U* is the *universe*, and *A* – a set *attributes*. 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:

> $xI(B)y$ 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*.

Obviously *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*) belong 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. In the rough set approach the elementary sets are the basic building blocks of our knowledge about reality.

The indiscernibility relation will be used next to define basic concepts of rough set theory. Let us define now the following two operations on sets

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

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

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

Rough set can be also characterized numerically by the following coefficient

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

called *accuracy* of *approximation*, where $|X|$ denotes the cardinality of X. Obviously $0 \le \alpha_B(X) \le 1$. If $\alpha_B(X) = 1$, X is crisp with respect to B (X is precise with respect to B), and otherwise, if $\alpha_B(X) < 1$, X is rough with respect to B.

Approximation can be employed to define dependencies (total or partial) between attributes, reduction of attributes, decision rule generation and others, but will not discuss these issues here. For details we refer the reader to references.

APPLICATIONS

Rough set theory has found many interesting applications. 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 main advantage of rough set theory is that it does not need any preliminary or additional information about data - like probability in statistics, or basic probability assignment in Dempster-Shafer theory and grade of membership or the value of possibility in fuzzy set theory.

The rough set theory has been successfully applied in many real-life problems in medicine, pharmacology, engineering, banking, financial and market analysis and others. Some exemplary applications are listed below.

There are many applications in medicine. In pharmacology the analysis of relationships between the chemical structure and the antimicrobial activity of drugs has been successfully investigated. Banking applications include evaluation of a bankruptcy risk and market research. Very interesting results have been also obtained in speaker independent speech recognition and acoustics. The rough set approach seems also important for various engineering applications, like diagnosis of machines using vibroacoustics symptoms (noise, vibrations) and process control. Application in linguistics, environment and databases are other important domains.

First application of rough sets to material sciences, particularly interesting to this community, can be found in [2, 3]. Rough set approach to materials science provides new algorithmic method for predicting and understanding material properties and behaviour, which can be very useful in creating new materials [3].

More about applications of rough set theory can be found in [5,6,14,19,20,21,22].

Application of rough sets requires a suitable software. Many software systems for workstations and personal computers based on rough set theory have been developed. The most known include LERS [1], Rough DAS and Rough Class [16] and DATALOGIC [17]. Some of them are available commercially.

CONCLUSION

Rough set approach to data analysis has many important advantages. Some of them are listed below.

- Provides efficient algorithms for finding hidden patterns in data.
- Identifies relationships that would not be found using statistical methods.
- Allows both qualitative and quantitative data.

- Finds minimal sets of data (data reduction).
- Evaluates significance of data.
- Generates sets of decision rules from data.
- It is easy to understand.
- Offers straightforward interpretation of obtained results.
- Most algorithms based on the rough set theory are particularly suited for parallel processing, but in order to exploit this feature fully, a new computer organization based on rough set theory is necessary.

Although rough set theory has many achievements to its credit, nevertheless several theoretical and practical problems require further attention.
Especially important is widely accessible efficient software development for rough set based data analysis, particularly for large collections of data.
Despite of many valuable methods of efficient, optimal decision rule generation methods from data, developed in recent years based on rough set theory - more research here is needed, particularly, when quantitative attributes are involved. In this context also new discretization methods for quantitative attribute values are badly needed. Also an extensive study of a new approach to missing data is very important. Comparison to other similar methods still requires due attention, although important results have been obtained in this area. Particularly interesting seems to be a study of the relationship between neural network and rough set approach to feature extraction from data.
Last but not least, rough set computer is badly needed for more serious applications. Some research in this area is already in progress.
For basic ideas of rough set theory the reader is referred to [8,9,15,18].

ACKNOWLEDGMENTS

The author gratefully acknowledge the support of the Air Force Contract F61708-97-WO196.

REFERENCES

1. J.W. Grzyma³a-Busse, in Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, (R. S³owiñski ed.), Kluwer Academic Publishers, Boston, London, Dordrecht, 1992, P.471.
2. A.G. Jackson, M. Ohmer and H. Al-Kamhawi, in The Third International Workshop on Rough Sets and Soft Computing Proceedings (RSSC'94), (T.Y. Lin ed.), San Jose State University, San Jose, California, USA, 1994.
3. A.G. Jackson, S.R. LeClair, M.C. Ohmer, W. Ziarko and H. Al-Kamhwi, Acta Metallurgica et Materialia, 1996, p.4475.
4. E. Krusiñska, R. S³owiñski and J. Stefanowski, Applied Stochastic Models and Data Analysis, 8, 1992, p.43.
5. T.Y. Lin and N. Cercone, Rough Sets and Data Mining - Analysis of Imperfect Data, Kluwer Academic Publishers, Boston, London, Dordrecht, 1997, P.430.
6. T.Y. Lin and A.M.Wildberger, The Third International Workshop on Rough Sets and Soft Computing Proceedings RSSC'94), San Jose State University, San Jose, California, USA, 1995.
7. Z. Pawlak, International Journal of Computer and Information Sciences, 11, 1982, p.341.
8. Z. Pawlak, Rough Sets - Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Boston, London, Dordrecht, 1991, P.229.

9. Z. Pawlak, J.W. Grzyma³a-Busse, R. S³owiñski and W. Ziarko, Communication of the ACM, 38, 1995, p.88.
10. Z. Pawlak and A. Skowron, in Advances in the Dempster Shafer Theory of Evidence, (R.R Yaeger, M. Fedrizzi and J. Kacprzyk eds.), John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, 1994.
11. Z. Pawlak and R. S³owiñski, European Journal of Operational Research, 72, 1994, p.443.
12. A. Skowron and J.W. Grzyma³a-Busse, in Advances in the Dempster-Shafer Theory of Evidence, (R.R, Yaeger, M. Fedrizzi and J. Kacprzyk eds.), John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, 1994.
13. A. Skowron and C. Rauszer, in Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, (R. S³owiñski ed.), Kluwer Academic Publishers, Boston, London, Dordrecht, 1992, P.471.
14. R. S³owiñski, Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, Boston, London, Dordrecht, 1992, P.471.
15. R. Słowiński, AI Expert, 10, 1995, p.18.
16. R. S³owiñski and J. Stefanowski, J., (1992), in Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, (R. S³owiñski ed.), Kluwer Academic Publishers, Boston, London, Dordrecht, 1992, P.471.
17. A. Szladow, PC AI , 7/1, 1993, p.40.
18. A. Szladow and W. Ziarko, AI Expert, 7,1993, p.36.
19. S. Tsumoto, S. Kobayashi, T. Yokomori, H. Tanaka and A. Nakamura, The Fourth Internal Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, PROCEEDINGS, The University of Tokyo, 1996, P.465.
20. P.P. Wang, Second Annual Joint Conference on Information Sciences, PROCEEDINGS, Wrightsville Beach, North Carolina, USA, 1995.
21. P. Wang, Joint Conference of Information Sciences, Vol. 3. Rough Sets and Computer Sciences, Duke University, 1997, P.449.
22. W. Ziarko, 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, P.476.