ROUGH SETS AND DATA ANALYSIS

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

Abstract

1 Basic Philosophy

In this talk we are going to present basic concepts of a new approach to data analysis, called rough set theory[13]. The theory has attracted attention of many researchers and practitioners all over the world, who contributed essentially to its development and applications.

Rough set theory overlaps with many other theories, especially with fuzzy set theory[2,15], evidence theory[19] and Boolean reasoning methods[18], discriminant analysis[5] - nevertheless it can be viewed in its own rights, as an independent, complementary, and not competing discipline.

Rough set theory is based on classification. Consider, for example, a group of patients suffering from a certain disease. With every patient a data file is associated containing information like, e.g. body temperature, blood pressure, name, age, address and others. All patients revealing the same symptoms are indiscernible (similar) in view of the available information and can be classified in blocks, which can be understood as elementary granules of knowledge about patients (or types of patients). These granules are called elementary sets or concepts, and can be considered as elementary building blocks of knowledge about patients. Elementary concepts can be combined into compound concepts, i.e. concepts that are uniquely defined in terms of elementary concepts. Any union of elementary sets is called a crisp set, and any other sets are referred to as rough (vague, imprecise). With every set X we can associate two crisp sets, called the lower and the upper approximation of X. The lower approximation of X is the union of all elementary set which are included in X, whereas the upper approximation of X is the union of all elementary set which have non-empty intersection with X. In other words the lower approximation of a set is the set of all elements that surely belongs to X, whereas the upper approximation of X is the set of all elements that possibly belong to X. The difference of the upper and the lower approximation of X is its boundary region.

0-7803-3687-9/96©1996 IEEE

Obviously a set is rough if it has non empty boundary region; otherwise the set is crisp. Elements of the boundary region cannot be classified, employing the available knowledge, either to the set or its complement. Approximations of sets are basic operation in rough set theory.

Basics of rough set theory can be found in (Grzymała-Busse, 1988, Grzymała-Busse, 1995, Pawlak, 1991, Pawlak, et al 1995, Słowiński, 1995, Szladow and Ziarko, 1993).

2 Information Tables and Rough Sets

Information is often available in a form of data tables, known as information systems, attribute-value tables or information tables. Columns of an information table are labeled by attributes, rows - by objects and entries of the table are attribute values. Objects having the same attribute values are indiscernible with respect to these attributes and belong to the same block of the partition (classification) determined by the set of attributes. Basic problems in data analysis which can be tackled employing the rough set approach are the following:

- Characterization of set of objects in terms of attribute values.
- Finding dependencies (total or partial) between attributes.
- Reduction of superfluous attributes (data).
- Finding the most significance attributes.
- Decision rule generation

Rough set theory offers simple algorithms to answer the above questions and enables straightforward interpretation of obtained results.

3 Applications and Advantages

The rough set methodology has found many real-life applications in medical data analysis, finance, banking, engineering, voice recognition, image processing and others. The proposed method has many important advantages. Some of them 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 method is particularly suited for parallel processing, but in order to exploit this feature fully a new hardware solutions are necessary.

Some information about application of rough set theory are reported in (Lin and Wilderberg, 1994, Lin, 1995, Słowiński, 1992, Ziarko, 1993).

4 Further Research

More than 1000 papers have been published on rough set theory and its applications till now. Despite many important theoretical contributions and extensions of the original model some essential research problems requires extensive research. Some of them are listed below:

- Rough logic, based on the concept rough truth seems to be a very important issue.
- Theory of rough relation and especially rough function is necessary in many applications.
- Comparison with many other approaches to data analysis, in particular neural networks, genetics algorithms and others.

Besides, some practical problems related with application of rough sets in many domains are of great importance.

- Discretization methods for continuous data.
- Reduction methods for a large databases.
- Decomposition of large information tables.
- Efficient and widely assessable software is necessary to further development of various applications.
- Development of rough set computer seems to be badly needed in order to exploit fully the advantages of the rough set approach to data analysis.

Besides, rough control, the application of rough set theory to control, seems to be a very promising area of application are of the rough set concept and need due attention [1,7,8,10,11,22,23,25,27].

References

1. Czogała, E, Mrózek, A., Pawlak, Z., (1995). The Idea of Rough-Fuzzy Controller. International Journal of Fuzzy Sets and Systems, 72, 61-73.
2. Dubois, D, Prade, H., (1992).Putting Rough Sets and Fuzzy Sets Together. In Intelligent Decision Support - Handbook of Advances and Applications of the Rough Set Theory, R. Slowinski ed., Kluwer Academic Publishers, Dordrecht, Boston, London, 203-232.
3. Grzymała-Busse, J., (1988). Knowledge Acquisition under Uncertainty - a Rough Set Approach, Journal of Intelligent and Robotics Systems, 1, 3-16.
4. Grzymała-Busse, J., (1995). Rough Sets , Advances in Imaging and Electrons Physics, 94, to appear.
5. Krusińska, E., Słowiński, R., Stefanowski, J., (1992). Discriminant versus Rough Set Approach to Vague Data Analysis, Journal of Applied Statistics and Data Analysis, 8, 43-56.
6. Lin, T.Y., Wilderberg, A.M. eds., (1994). Soft Computing, Proceedings of the Third International Workshop on Rough Sets and Soft Computing (RSSC!94), November 10-12, 1994, San Jose State University, San Jose, California, USA.
7. Lin, T.Y., (1995a). Fuzzy Logic Controller and Rough Logic, Soft Computing: Rough Sets, Fuzzy Logic, Neural Network, Uncertainty Management and Knowledge Discovery, Society of Computer Simulation, Society of Computer Simulation, 125-129.
8. Lin, T.Y. (1995b). Rough-Fuzzy Controller for Complex Systems, The Fourth Annual International Conference on Fuzzy Theory and Technology, Proceedings of Second Annual Joint Conference on Information Science Wrightsville Beach, North Carolina, Sept. 28-Oct. 1, 1995, 18-21.
9. Lin, T.Y., ed., (1995c). CSC-95, 23rd Annual Computer Science Conference on Rough Sets and Database Mining, Conference Proceedings, March 2, San Jose State University, San Jose, California, USA.
10. Mrózek, A. (1986). Use of rough sets and decision tables for implementing rule-based control of industrial processes, Bulletin of the Polish Academy of Sciences, T. Sc., 34/5-6, 357-371.
11. Munakata, T., (1995a). Rough Control: A perspective. In CSC-95, 23rd Annual Computer Science Conference on Rough Sets and Database Mining, Conference Proceedings, March 2, San Jose State University, San Jose, California, USA.
12. Munakata, T. (1995b). Rough control: Basic ideas and applications, the Proceeding of the Second Annual Joint Conference on Information

Sciences, Wrightsville Beach, NC, Sept. 28 - Oct. 1, 1995, 340-343.
13. Pawlak, Z., (1982). Rough Sets, International Journal of Computer and Information Sciences, 11, 341-356.
14. Pawlak, Z., (1991). Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, Boston, London.
15. Pawlak, Z., Skowron , A., (1994). Rough Membership Functions. In Advances in the Dempster-Shafer Theory of Evidence, R.R. Yeager, M. Fedrizzi, J. Kcprzyk, eds., John Wiley and Sons, New York, 251-271.
16. Pawlak, Z., Słowiński, R., (1994). Decision Analysis using Rough Sets, International Transactions on Operational Research, 1, 107-104.
17. Pawlak, Z., Grzymała-Busse, J., Słowiński, R., Ziarko, W., (1995). Rough Sets, Communication of the ACM, 38, 88-95.
18. Skowron, A., Rauszer, C., (1992). The Discernibility Matrices and Functions in Information Systems. In Intelligent Decision Support - Handbook of Advances and Applications of the Rough Set Theory, R. Słowiński, ed., Kluwer Academic Publishers, Dordrecht, Boston, London, 311-369.
19. Skowron, A., Grzymała-Busse, J., (1994). From Rough Set Theory to Evidence Theory. In Advances in the Dempster-Shafer Theory of Evidence, R.R. Yeager, M. Fedrizzi, J. Kcprzyk, eds., John Wiley and Sons, New York, 251-271, 193-235.
20. Słowiń ski, R. ed., (1992). Intelligent Decision Support - Handbook of Advances and Applications of the Rough Set Theory, Kluwer Academic Publishers, Dordrecht, Boston, London.
21. Słowiński, R., (1995). Rough Set Approach to Decision Analysis, AI Expert, 10, 18-25.
22. Szladow, A. Ziarko, W. (1992). Knowledge-based process control using rough sets. In Słowiński, R., ed. Decision Support by Experience: Rough Sets Approach. Kluver Academic Publishers, 49-60.
23. Szladow A., Ziarko, W., (1993a). Adaptive process control using rough sets, Paper # 93-384, Instrument Society of America, 1993, 1421-1430.
24. Szladow, A., Ziarko, W., (1993b). Rough Sets: Working with Imperfect Data, AI Expert, 8, 36-41.
25. Ziarko, W., (1992). Generation of control algorithms for computerised controllers by operator supervised training, In Modelling, Identification and Control, Proceedings of the Eleventh IASTED International Conference, M.H. Hamza, ed., Innsbruck, Austria, Feb., 1992, 510-513.
26. Ziarko, W., ed., (1993). Rough Sets, Fuzzy Sets and Knowledge Discovery. Proceeding of the International Workshop on Rough sets and Knowledge Discovery (RSKD-93), Banff, Alberta, Canada, Springer Ver-

lag, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.
27. Ziarko, W., Katzberg, J.D., (1989) Control algorithm acquisition, analysis and reduction: A machine learning approach, In Knowledge-based System Diagnosis, Supervision, and Control, S. G. Tzafestas, ed., Plenum Press, 167-180.