ROUGH SETS AND DATA MINING

Analysis for Imprecise Data

T. Y. LIN
San Jose State University
San Jose, California, USA

N. CERCONE
University of Regina
Regina, Saskatchewan, CANADA

KLUWER ACADEMIC PUBLISHERS
Boston/London/Dordrecht

1

ROUGH SETS

Zdzisław Pawlak

Institute of Computer Science,
Warsaw University of Technolgy, Warsaw 00-665, Poland,
ul. Nowowiejska 15/19, zpw@ii.pw.edu.pl

The concept of the rough set is a new mathematical approach to imprecision, vagueness and uncertainty in data analysis.

The starting point of the rough set philosophy is the assumption that with every object of interest we associate some information (data, knowledge). E.g., if objects are patients suffering form a certain disease, symptoms of the disease form information about patients. Objects are similar or indiscrenible, if they are characterized by the same information. The indiscernibility relation generated thus is the mathematical basis of the rough set theory.

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

As a consequence of the above definition each rough set have boundary-line elements, i.e., elements which cannot be with certainty classified as members of the set or its complement. (Obviously crisp sets have no boundary-line elements at all). In other words boundary-line cases cannot be properly classified employing the available knowledge. Thuse rough sets can be viewed as a mathematical model of vague concepts.

In the rough set approach any vaque concept is characterized by pair of precise concepts – called the lower and the upper approximation of the vague concept. The lower approximation consists of all objects which surely belong to the concept and the upper approximation contain all objects which possible belong to the concept. Approximations constitute two basic operations in the rough set approach.

The above presented ideas can be illustrated by the following example. Suppose we are given data table – called also attribute-value table or information system – containing data about 6 patients, as shown below.

Patient	Headache	Muscle-pain	Temperature	Flu
p1	no	yes	high	yes
p2	yes	no	high	yes
p3	yes	yes	very high	yes
p4	no	yes	normal	no
p5	yes	no	high	no
p6	no	yes	very high	yes

Columns of the table are labelled by attributes (symptoms) and rows by objects (patients), whereas entries of the table are attribute values. Thus each row of the table can be seen as information about specific patient. For example patient p2 is characterized in the table by the following attribute-value set

{(Headache, yes), (Muscle-pain, no), (Temperature, high), (Flu, yes)},

which form information about the patient.

In the table patients p2, p3 and p5 are indiscernible wiht respect to the attribute Headache, patients p3 and p6 are indiscernible with respect to attributes Muscle-pain and Flu, and patients p2 and p5 are indiscernible with respect to attributes Headache, Muscle-pain and Temperature. Hance, for example, the attribute Headache generates two elementary sets {p2, p3, p5} and {p1, p4, p6}, whereas the attributes Headache and Muscle-pain form the following elementary sets, {p1, p4, p6}, {p2, p5} and {p3}. Similary one can define elementary set generated by any subset of attributes.

Because patient p2 has flu, whereas patient p5 does not, and they are indiscernible with respect to the attributes Headache, Muscle-pain and Temperature, thus flu cannot be characterized in terms of attributes Headache, Muscle-pain and Temperature. Hance p2 and p5 are the boundary-line cases, which connot be properly classified in view of the available knowledge. The remaining patients p1, p3 and p6 display symptoms which enable us to classify them with certainty as having flu, patients p1 and p5 cannot be excluded as having flu and patient p4 for sure has not flu, in view fo the displayed symptoms. Thus

the lower approximation for the set of patients having flu is the set {p1, p3, p6} and the upper approximation of this set is the set {{p1, p2, p3, p5, p6}. Similary p4 has not flu and p2, p5 can not be excludes as having flu, thus the lower approximation of this concept is the set {p4} whereas – the upper approximation is the set {p2, p4, p5}.

We may also ask whether all attributes in this table are necessary to define flu. One can easily see, for example that, if a patient has very high temperature, he has for sure flu, but if he has normal tempetrature he has not flu whatsoever.

In general basic problems which can be solved using the rough set approach are the following:

1) description of set of objects in terms of attribute values

2) dependencies (full or partial) between attributes

3) reduction of attributes

4) significance of attributes

5) decision rules generation

and others.

The rough set methodology has been applied in many real-life applications and it seems to be important to machne learning, decision analysis, knowledge discovery, expert systems, decision support systems, pattern recognition and others.

Some current research on rough controllers has pinted out a new very promising area of applications of the rough set theory.

The rough set concept coincided with many other mathematical models of vagueness and uncertainty – in particular fuzzy sets and evidence theory – but it can be viewed in its own rights.

REFERENCES

[1] Grzymała-Busse J.W., (1991), *Managing Uncertainty in Expert Systems.* Kluwer Academic Publishers, Dordrecht, Boston, London.

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

[3] Pawlak Z., (1982), "Rough sets". *International Journal of Computer and Information Sciences*, 11, 341–356.

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

[5] Pawlak Z., Grzymała-Busse J. W., Słowiński R., and Ziarko, W., (1995),"Rough sets", *Communication of the ACM*, 38, 88–95.

[6] Słowiński, R., (ed.), (1992), *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*, Kluwer Academic Publishers, Dordrecht.

[7] Ziarko, W., (ed.), (1993), *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.

Biographical Sketch

Zdzisław I. Pawlak is Professor of Computer Science and Member of the Polish Academy of Sciences. He is head of the Group for Algorithmic Method of Reasoning in the Institute of Theoretical and Applied Informatics, Polish Academy of Sciences and Director of the Institute of Computer Science, Warsaw University of Technology. Current research interests include intelligent systems and cognitive sciences, in particular, decision support systems, reasoning about knowledge, machine learning, inductive reasoning, vagueness, uncertainty and conflict analysis. **Author's Present Address:** Institute of Computer Science,

Warsaw University of Technolgy, Warsaw 00-665, Poland, ul. Nowowiejska 15/19, zpw@ii.pw.edu.pl