ON ROUGH SETS

Zdzisław Pawlak

Institute of Computer Science
Polish Academy of Sciences
P.O. Box 22
00-901 Warszawa, PKiN

1. Introduction

We present in this note a new approach to data analysis. The method arose from studying some problems of artificial intelligence, such as expert systems, inductive reasoning, learning by examples, cluster analysis and others.

In many problems in those areas we face the following situation: we are given set of objects (or states of an object), and we are unable to distinguish some objects (states) by the available means of observation or description. To cope with such situations we introduce an indiscernibility relation which expresses our ability to discern objects (or states) under consideration. It is assumed that the indiscernibility relation is an equivalence relation. Thus, the indiscernibility relation express in a certain sense our accuracy of observation, and equivalence classes of the relation, called here atoms, are the least subsets we are able to observe "through" the indiscernibility relation.

Clearly all observable sets by given indiscernibility relation are only union of atoms. Others sets are not observable. Thus if we want to deal with any subsubset of a given set, we can only do that with some approximation. To express this precisely we introduce two operations on sets; lower and upper approximation of a set. Lower approximation of set X is the greatest union of atoms included in set X, and upper approximation of X is the least union of atoms including set X. Thus lower approximation of a set contains all elements which definitely belong to set X, and upper approximation of set X contains all elements which possibly belong to X. The difference between upper and lower appro-

ximation is a borderline region of set X, which determine limits of tolerance for deciding whether given elements belong to X or not.

It turns out that lower and upper approximations coincide with interior and closure operation respectively in a certain topological space, generated by the indiscernibility relation, so one can use standard topological methods to investigate the problem of uncertain membership of elements to set X.

Eventually this leads to a new concept of a set with no clearly defined boundaries, called here rough set. Rough set can be viewed as a pair of sets (lower and upper approximations) or a family of set having the same lower and upper approximations.

We shall not consider in this note the idea of rough set itself, but we restrict our considerations to approximation operations only. The idea of a rough set has been first published in Pawlak (1982a).

Let us notice that the concept of rough set cannot be expressed in terms of fuzzy sets, for there are essential differences between those two concepts (see Pawlak (1984e)).

It seems however that this idea has some connections with Alternative Set Theory (Vopenka (1979)) and Non-Standard Analysis (Robinson (1966)), but these problems have been not studied intensively as yet.

The proposed approach needs deeper mathematical insight, and some investigations are carried out to meet this requirement, but there are not concrete results until now.

The rough set idea has also some logical flavour, and many results have been obtained in this direction by Orłowska (see the enclosed references).

The concept of the rough set although for from being fully explored found some interesting applications.

From practical point of view the most interesting one seems to be a new approach to expert systems design. Several systems to support medical data analysis have been implemented (see Pawlak (1984d), Doroszewski at el. (1984)) which show some advantages in comparison to traditional methods. Also the application of rough set approach to support industry processes control has been developed by Mrożek (see Mrożek (1984)).

The rough set approach seems to be also a good departure point to study foundations of knowledge representation (see Orłowska, Pawlak (1984a)). There are also trials to apply this idea to formal languages theory (Kierczak (1984)), approximate concept learning (Konrad et al. (1981), Tu-Hue Le (1982)), probability (Pawlak (1984), mechanics (Woźniak (1983)) measurement theory (see Orłowska, Pawlak (1984d)) and others.

2. Basic concepts

2.1. Indiscernibility

Let U be a set called universe and let R be a binary relation over U, called indiscernibility relation. We assume through this paper that R is an equivalence relation. If $(x,y)\in R$ we say that x and y are R-indiscernible. Equivalence classes of the relation R are called R-elementary sets or R-atoms. An equivalence class of a relation R containing element x will be denoted by $[x]_R$. Any finite union of R-atoms will be called R-definable set. We assume that the empty set is R-definable for every R. The family of all R-definable sets will be denoted by Def(R). We shall ommit the letter R if the indiscernibility relation is understood, writing for example atoms, elementary sets - instead R-atoms or R-elementary sets,

Given an universe U and an indiscernibility relation R, then the ordered pair $A = (U,R)$ will be called an approximation space.

Speaking informally, the approximation space $A = (U,R)$ says with what accuracy we can distinguish elements of the universe U, or in other words, that we observe elements of the universe U over the indiscernibility relation R. That is to mean that elements of the universe U which are indistinguishable by given means of observation are glued together forming atoms (equivalence classes of the indiscernibility relation), and each atom can be observed only as a whole. Consequently we are able to observe only definable sets. Not every subset of the universe U is thus observable; some sets can be observed only with some approximation. To express this fact precisely we employ the notion of an approximation of a set in a given approximation space.

2.2. Approximation of a set

Let $A = (U,R)$ be an approximation space, and let $X \subset U$. be a certain subset of U.

An upper R-approximation of X, denoted $\overline{R}X$ is defined as

$$\overline{R}X = \left\{x \in U : [x]_R \cap X \neq \emptyset \right\},$$

i.e. $\overline{R}X$ is the least R-definable set including X.

A lower R-approximation of X, denoted $\underline{R}X$ is defined as

$$\underline{R}X = \left\{x \in U : [x]_R \subset X \right\},$$

i.e. $\underline{R}X$ is the greatest R-definable set included in X.

Set $Bn_R X = \overline{R}X - \underline{R}X$ will be called R-boundary of X.

By means of approximations we can define positive, negative and borderline region of a set, i.e.

$\underline{R}X$ - R-positive region of set X

$U-\overline{R}X$ - R-negative region of set X

$Bn_R X$ - R-borderline region of set X.

In R-positive region of set x there are only elements which definitely belong to set X; in R-negative region of set X there are only elements which definitely do not belong to set X, and the R-borderline region of set X is a doubtful area, consisting of elements which membership to set X cannot be decided by the given "accuracy" of observation, expressed by the indiscernibility relation R.

To express facts mentioned above we can also introduce two membership functions $\underline{\in}_R$, $\overline{\in}_R$ defined as follows:

$$x \underline{\in}_R X \text{ iff } x \in \underline{R}X$$
$$x \overline{\in}_R X \text{ iff } x \in \overline{R}X$$

and which read "x surely belongs to X", and "x possibly belongs to X" respectively. Thus we may interpret approximations as counterparts of necessity and possibility in model logic.

It would seem that the concept of approximations of a set, can be expressed in terms of fuzzy set theory (see Zadeh (1965)) assuming the following fuzzy membership function

$$\mu'_X(x) = \begin{cases} 1 & \text{iff } x \in \underline{R}X \\ 1/2 & \text{iff } x \in Bn_R X \\ 0 & \text{iff } x \in U - \overline{R}X, \end{cases}$$

however this membership function cannot be extended for union and intersection of sets (for details see Pawlak (1984d)).

Thus the concept of approximation of a set cannot be expressed in fuzzy sets theory.

2.3. Properties of approximations

Approximation space $A = (U,R)$ defines uniquely the topological space $T_A = (U, Def(R))$, where $Def(R)$ is topology for U, and it is the family of open and closed sets in T_A. The family of all R-elementary sets in A is a base for T_A, and lower and upper approximations of X are interior and closure respectively in the topological space T_A. Thus $\underline{R}X$ and $\overline{R}X$ have the following properties:

(A1) $\underline{R}X \subset X \subset \overline{R}X$

(A2) $\underline{R}U = \overline{R}U = U$

(A3) $\underline{R}\emptyset = \overline{R}\emptyset = \emptyset$

(A4) $\overline{R}(X \cup Y) = \overline{R}X \cup \overline{R}Y$

(A5) $\underline{R}(X \cup Y) \supset \underline{R}X \cup \underline{R}Y$

(A6) $\overline{R}(X \cap Y) \subset \overline{R}X \cap \overline{R}Y$

(A7) $\underline{R}(X \cap Y) = \underline{R}X \cap \underline{R}Y$

(A8) $\overline{R}(-X) = -\underline{R}X$

(A9) $\underline{R}(-X) = -\overline{R}X$

Moreover approximations obey the following properties:

(A10) $\underline{R}\underline{R}X = \overline{R}\underline{R}X = \underline{R}X$

(A11) $\overline{R}\overline{R}X = \underline{R}\overline{R}X = \overline{R}X$.

2.4. Undefinable sets

It is obvious that set X is R-definable iff $\underline{R}X = \overline{R}X$; otherwise set X is not R-definable, i.e. X is R-undefinable. One can split undefinable sets into four following classes:

(B1) If $\underline{R}X \neq \emptyset$ and $\overline{R}X \neq U$, X will be called roughly R-definable

(B2) If $\underline{R}X = \emptyset$, X will be called internally R-undefinable

(B3) If $\overline{R}X = U$, X will be called externally R-undefinable

(B4) If $\underline{R}X = \emptyset$ and $\overline{R}X = U$, X will be called totally R-undefinable.

Thus "looking" at the subset $X \subset U$ with "precision" determined by the indiscernibility relation R, we can face the following five situations:

1) Set X is R-definable, i.e. for every element $x \in U$ we can decide whether x belongs to X or not,
2) Set X is roughly R-definable, i.e. for some elements $x \in U$ we can decide whether x belongs to X or not, but there are elements (belonging to R-borderline region of set X), for which we are unable to decide their membership to set X.
3) Set X is internally R-undefinable, i.e. for any element $x \in U$ we cannot decide whether x definitely belongs to set X.
4) Set X is externally R-undefinable, i.e. for any element $x \in U$ we cannot exclude x being member of X.
5) Set X is totally R-undefinable, i.e. for any element $x \in U$ we cannot decide whether x belongs to X or not.

The above five cases are depicted on figure 1.

2.5. Accuracy of approximation

In order to express more exactly concept of approximation of a set we introduce the notion of accuracy coefficient $\alpha_R(X)$ of a set X in the approximation space $A = (U,R)$, defined as follows:

$$\alpha_R(X) = \frac{\mu_R(X)}{\mu_R(X)} = \frac{\mu(\underline{R}X)}{\mu(\overline{R}X)},$$

where μ is Jordan measure of set X. In case when X is finite we assume $\mu(\underline{R}X) = card(\underline{R}X)$ and $\mu(\overline{R}X) = card(\overline{R}X)$.

Obviously

$$0 \leq \alpha_R(X) \leq 1$$

and $\alpha_R(X) = 1$ iff X is R-definable and $\alpha_R(X) = 0$ iff X is R-undefinable.

The accuracy coefficient express ratio of elements surely belonging to set X, to those possibly belonging to set X.

3. Example of application

3.1. Introduction

In this section we give an application of ideas introduced in previous section to some artificial intelligence problems.

In inductive reasoning or learning by examples we characterize objects in terms of attributes like, colour, size etc. Each attribute assume values from given set, for example colour may have value, blue, green, etc.

The question arises whether we can define uniquely any subset of object in terms of their attributes? In other words: we are given set of examples of a certain concept, and we ask whether this concept can be characterized in terms of features (attributes) of examples?

In order to formulate this problem precisely we introduce first some necessary notions.

3.2. Knowledge Representation System

Let U be set of objects and let Q be set of attributes of objects belonging to U. With every $q \in Q$ we associate set V_q values of q, called domain of q. We describe each object from U by determining all its features, i.e. by giving values of all attributes associated with it. This description of objects in terms of their attributes, can be understood as knowledge about objects.

To express above considerations more precisely we introduce the notion of Knowledge Representation System.

By a Knowledge Representation System S we mean an ordered quadruple (see Pawlak (1981b))

$$S = (U, Q, V, \varrho)$$

where

U - is set of objects

Q - is set of attributes

$V = \bigcup_{q \in Q} V_q$, V_q - is domain of q

$\varrho : U \times Q \to V$ is an information function

We assume that

$$\varrho(x,q) \in V_q \text{ for every } q \in Q.$$

Function $\varrho_x(q) = \varrho(x,q)$ for every $x \in U$ and $q \in Q$ will be called information about x in S.

Thus information about any object in a given knowledge representation system is a description of the object in terms of its features available in the system.

3.3. Indiscernibility relations generated by knowledge representation System

Let $S = (U, Q, V, \varrho)$ be a knowledge representation system and let $P \subseteq Q$ be a subset of attributes.

We say that objects $x, y \in U$ are P-indiscernible in S, $x \tilde{\tilde{P}} y$, iff

$$\varrho_x(p) = \varrho_y(p) \text{ for every } p \in P.$$

Obviously P is an equivalence relation. So each knowledge Representation System generates a family of approximation spaces

$$A_P = (U, \tilde{\tilde{P}})$$

where $P \subseteq Q$.

In other words having a knowledge representation system $S = (U, Q, V, \varrho)$ we can employ the notion of approximations of sets by means of subset $P \subseteq Q$ of attributes in the system. Thus

$\underset{\sim}{P}X, \tilde{P}X$, denote lower, and upper P-approximation of set X, respectively.

For the sake of simplicity we shall write P instead of $\sim P$, for example $\underset{\sim}{P}X(\tilde{P}X)$ instead of $\underset{\sim}{P}X(\tilde{P}X)$ etc.

Suppose we are given knowledge representation system $S = (U,Q,V,\varrho)$, $X \subset U$, $P \subset Q$, and we want characterise set X in terms of attributes P. Then we have the following possibilities:

1) X is P-definable
2) X is roughly P-definable
3) X is internally P-undefinable
4) X is externally P-undefinable
5) X is totally P-undefinable.

That is to mean that if we want to learn a certain concept by giving examples of that concept (set X), and we want characterize the concept in terms of features (set P of attributes) of examples which represent the concept we may face one of the following situations:

1) The concept can be learned if X is P-definable
2) The concept can be learned roughly, if X is roughly P-definable
3) The concept cannot be learned (one can learn only counter example of the concept) if X is internally P-undefinable
4) The concept cannot be learned fully (one cannot learn counter example), if X is externally P-undefinable
5) The concept cannot be grasped by given examplex, if set X is totally P-undefinable.

The idea mentioned above is of primary importance for many branches of artificial intelligence, like pattern recognition, inductive inference, expert systems and others.

3.4. Application to medical diagnosis

The proposed approach has been applied to medical data analysis. A file of 150 patients suffering from heart disease seen in a hospital in Warsaw was used as a data base. Every patient were characterized by 7 symptoms (attributes), and all patients have been divided by experts into six classes corresponding to the grade of the disease advance.
The question was whether by means of assumed symptoms one can determine the grade of disease advance.

Results of computation are given in the table below.

Class Number	Number of Patients	Lower Approx.	Upper Approx	Accuracy
1	10	11	15	0,47
2	46	33	51	0,72
3	42	39	45	0,87
4	33	30	36	0,83
5	15	15	15	1,00
6	4	4	4	1,00

Table 1

Higher class numbers in the table represent higher disease advance.

One can see from the table that the assumed symptoms are fully characteristic for classes 5 and 6 only representing the highest degree of disease advance, i.e. these two classes are definable by the assumed symptoms. The remaining classes are roughly definable by the considered attributes and the

accuracy of approximation is very low (0,27) for the initial stage of the disease, what is to mean that the symptoms are not enough characteristic in early stages of the disease.

For more details concerning this example see Pawlak (1984d).

Several other medical data analysis examples have been computed using this method, and the results show, that the presented approach can be of practical value, especially when data about patients are vague.

4. Conclusion

In this note we presented only part of the research activity concerning the rough sets approach to data analysis. It is to early to state firmly whether the proposed approach gives really new, valuable tools for data analysis or not, however results obtained so far seem to convince that the area is worth of investigating.

ACKNOWLEDGMENT

Thanks are due to prof. G. Rozenberg for the proposal of writing the paper.

BIBLIOGRAPHY

Doroszewski, J., Janik, D., Tyrcha, J., (1984) On application of rough sets to medical diagnosis (manuscript)

Farinas del Cerro, L., Orłowska. E., (1983) DAL - a logic for data analysis. Languages et Systemmes Informatiques, Rep. 183.

Kierczak, K., (1984) Rough grammars. Fundamenta Informaticae (to appear)

Konrad, E., Orłowska, E., Pawlak, Z., (1981) Knowledge Representation Systems. ICS PAS Reports No 433.

Konrad, E., Orłowska. E., Pawlak, Z., (1982) On approximate concept learning, 1982 European Conference on Artificial Intelligence, 12-14 July 1982, ORSAY, France, 17-19. (Discussion Papers) also Technische Universität Berlin, Bericht No 81-7 (1981)

Marek, W., Pawlak, Z., (1983) Rough sets and information systems. Fundamenta Informaticae

Mrózek, A., (1984) Rough classification in identification, analysis and estimation of human-operator inference model (in Polish). Podstawy sterowania, vol. 14, No 1 (to appear)

Novotny, M., Pawlak, Z., (1983) On a representation of rough sets by means of information systems, Fundamenta Informaticae, vol. VI, No 3-4, pp. 276-285

Orłowska, E., (1982a) Logic of vague concepts, ICS PAS Reports, No 474

Orłowska, E., (1982b) Languages of approximate information. ICS PAS Reports, No 479

Orłowska, E., (1983a) Representation of vague information. ICS PAS Reports, No 503

Orłowska, E., (1983b) Semantics of vague concepts. 7th International Congress of Logic, Methodology and Philosophy of Sciences, Salzburg, Austria, Vol. 2, pp. 127-130

Orłowska, E., (1994) Modal logic in the theory of information systems. Zeitschrift für Math. Logik und Grundlagen der Math. (to appear)

Orłowska, E., Pawlak, Z., (1984b) Foundation of knowledge representation. ICS PAS Reports, No 537

Orłowska, E., Pawlak, Z., (1984b) Representation of nondeterministic information. Theoretical Computer Sciences, (to appear)

Orłowska, E., Pawlak, Z., (1984c) Expressive Power of Knowledge Representation. International Journal of Man-Machine Studies (to appear)

Orłowska, E., Pawlak, Z., (1984d) Measurement and Indiscernibility. Bull. Polish Acad. Sci. (to appear)

Orłowska, E., Pawlak, Z., (1984e) Basis for Artificial Intelligence. Fifth Generation Computer Systems, North-Holland (to appear)

Pawlak, Z., (1981a) Rough relations. ICS PAS Reports No 435

Pawlak, Z., (1981b) Information Systems - Theoretical Foundations. Information Systems, Vol. 6, No 3, pp. 205-218

Pawlak, Z., (1982a) Rough Sets. International Journal of Information and Computer Sciences, vol. 11, No 5, pp. 341-356

Pawlak, Z., (1982b) Rough function. ICS PAS Reports No 456

Pawlak, Z., (1983) Information Systems. (The book in Polish), Warszawa

Pawlak, Z., (1984a) On Superfluous Attributes in Knowledge Representation Systems. Bull. Polish Academy of Sciences (to appear)

Pawlak, Z., (1984b) Rough Probability. Bull. Polish Academy of Sciences (to appear)

Pawlak, Z., (1984c) Discrimination Power of Attributes in Konwledge Representation Systems. Bull. Polish Academy of Sciences (to appear)

Pawlak, Z., (1984d) Rough classification. International Journal of Man-Machine Study (to appear)

Pawlak, Z., (1984e) Rough sets and fuzzy sets. ICS PAS Reports (to appear)

Robinson. A., (1966) Non-standard analysis. North-Holland Publishing Company, Amsterdam

Tu-Hue Le (1982) Approximative Mustererkennung, Technische Universität, Berlin

Vopenka, P., (1979) Mathematics in the Alternative Set Theory. Teubaer-Texte zur Mathematik, Leipzig

Woźniak, Cz., (1983) Manuscript

Żakowski, W., (1984) Approximation in the space (U,R). Demonstratio Mathematicae (to appear)

Zadeh, L.A., (1965) Fuzzy sets. Information and Control, vol. 8, pp. 338-353.

A Complete Axiomatization for full Join Dependencies in Relations

B. Thalheim
Technische Universität Dresden
Sektion Mathematik
8027 Dresden
Mommsenstr. 13
GDR

In the Relational model (see /2/), data are stored in tables. The central problem of relational database design is, how to choose these tables. The introduction of data dependencies (more than 52 types), such as functional, multivalued and join dependencies and subsequent formalizations of desirable schemata are provided partial solutions to this central problem. The concept of functional