V.S. Alagar, S. Bergler and F.Q. Dong (Eds.)

Incompleteness and Uncertainty in Information Systems

Proceedings of the SOFTEKS Workshop on Incompleteness and Uncertainty in Information Systems, Concordia University, Montreal, Canada, 8–9 October 1993

Published in collaboration with the British Computer Society

Springer-Verlag
London Berlin Heidelberg New York
Paris Tokyo Hong Kong
Barcelona Budapest

Knowledge and Uncertainty a Rough Set Approach

Zdzisław Pawlak
Institute of Computer Science
Warsaw Technical University
zpw@ii.pw.edu.pl

1 Introduction

The idea of a rough set has been proposed by the author as a new mathematical tool to deal with vagueness and uncertainty. It seems to be of fundamental importance to AI and cognitive sciences, in particular expert systems, decision support systems, machine learning, machine discovery, inductive reasoning pattern recognition, decision tables and others.

The rough set theory, besides its methodological significance, turn out to be very useful in practice and its importance to data analysis seems to be unquestionable. Main advantage of the rough set theory in applications is in discovering patterns in data, data reduction, discovering of data dependencies, data significance, decision algorithms generation from data, approximate classification of data, discovering similarities or differences in data, and others. Many real life applications of this concept have been implemented. More about the application of the rough set theory can be found in Slowinski [38].

In the lecture basic concepts of the rough set theory will be outlined and its philosophical background briefly presented. For more detailed exposition the reader is referred to Pawlak [26].

2 Knowledge and classification

Theory of knowledge for a long time has been subject of interest of philosophers and logicians [17,18,20,29]. Recently new momentum to this area of research have been given by AI researchers ([1,4,5,6,7,8,9,13,14,15,16,19,22,23,24,25,30, 31,32,33,39,40] and others). There is no till now, however, widely shared body of opinion, as to how understand, represent and manipulate knowledge.

Intuitively, the idea of knowledge seems to be best expressed by Russell, who says that "knowing" is a kind of relationship between an organism and the environment [35], which can be perceived as a body of information about some parts of reality that is needed to behave rationally in the real world.

It seems natural that this kind of understanding of knowledge must be based on the ability of an organism to classify various states of the real world and the organism states itself, and consequently that each organism must be equipped with a variety of classification skills, on concrete and abstract level.

For example, knowledge of any organism about its environment must be based on the ability to classify variety of situations (e.g. safe-unsafe, dark-bright, etc.). On more specific level classification of sensory signals, like color,

"Apart from the known and the unknown, what else is there?"

Harold Pinter in *The Homecoming*

temperature etc., seems to be of fundamental significance to acquire knowledge needed by any organism.

Therefore we will assume, that knowledge is based on the ability to classify objects, and by object we understand anything we can think of, for example entities, states, processes, abstract concepts, signals, events etc.

Classification requires that small differences between objects being classified must be postponed, i.e. object being in the same class are *indiscernible*. For example when classifying some objects according to color, in order to form a class of red objects, we have to postpone small differences between various shades of red, so that all objects in the same class are indiscernible. Therefore the indiscernibility is a fundamental concept in the presented approach.

3 Knowledge and information systems

As said before we assume that knowledge is manifested by the ability to classify. Therefore we will define formally knowledge as a family of partitions over a fixed, finite universe. For mathematical reasons instated of partitions we may also use the corresponding equivalence relations. Thus knowledge can be also defined as family of equivalence relations over the universe. Because we need a "language" to represent various partitions, we will employ to this end so called information systems, called also attribute-value tables.

An *information system* is a finite table rows of which are labelled by objects of the universe U, whereas columns - are labelled by attributes from a fixed set A and each entry of the table corresponding to object x and attribute a is an *attribute value*. The set of all values of an attribute a is called the domain of a and is denoted V_a . For example, if the attribute is COLOR then its domain may be *red, green, blue*, etc. An example of an information system is given in the table below.

U	COLOR	SHAPE	SIZE
1	*red*	*tri.*	*small*
2	*green*	*square*	*small*
3	*blue*	*round*	*large*
4	*red*	*square*	*small*
5	*green*	*round*	*large*

Table 1

The table contains data about five children toy blocks, described by three features (attributes) *COLOR, SHAPE* and *SIZE*. The domains of the corresponding attributes are {*red, blue, green*}, {*triangular, square*}, round and {*small, large*}.

It is easily seen that each subset of attributes determines partition of objects of the universe into blocks having the same features, i.e. being indiscernible by this features.

"Knowing is a relation of the organism to something else or to a part of itself"

Bertrand Russel in *An Inquiry into Meaning and Truth*

Formally an information system can be defined as a pair $S = (U, A)$, where U - is the universe and A - is the set of attributes. Each attribute a can be understood as a total function $a : U \rightarrow V_a$, which to every objects associates the attribute value.

With every subset of attributes $B \subseteq A$, we associate a binary relation $IND(B)$, called an *indiscernibility relation* and defined thus:

$$IND(B) = \{(x, y) \in U^2 : \text{ for every } a \in B, a(x) = a(y)\}.$$

Obviously *IND(B)* is an equivalence relation and

$$IND(B) = \bigcap_{a \in B} IND(a).$$

An equivalence class of the relation *IND(B)* containing the object x will be denoted $[x]_B$, and the partition generated by *IND(B)*, i.e. family of all equivalence classes of $IND(B)$, is denoted as $U/IND(B)$, or in short U/B.

Any subset of the universe will be called concept or category in $S = (U, A)$; in particular equivalence classes of any relation $IND(a)$ will be referred to as *primitive* concepts of a (in S), where as equivalence classes of any relation $IND(B)$ will be called basic concepts of B (in S), provided card $(B) > 1$.

An attribute value $a(x)$ can be viewed as a name (description, label)) of the it primitive category of a containing x (i.e.- the name of $[x]_a$), whereas the set $a(x)_{a \in B}$, can be considered as a name of the basic category $[x]_B$.

Remark. In fact we should distinguish between the name (intention), of a concept, e.g. *red*, and its meaning (extension), e.g. the set of all red objects, but for the sake of simplicity we will make not this distinction, whenever it will make no confusion.

For example if *U/COLOR* is a partition classifying objects according to color, than *red, green, blue*, etc., i.e. sets of red, green or blue objects, are primitive concepts of our knowledge; if elements of the universe are classified according to *COLOR (red, green, blue)*, and *SHAPE (triangle, square, round,)* then the corresponding basic categories would be *green and square* (green square) *red and triangular* (red triangle), etc. Thus basic categories are fundamental building blocks, or basic properties of the universe which can be expressed employing this knowledge. Hence an information system contains descriptions of all basic categories available in considered knowledge.

Knowledge in the presented approach can be viewed as a family of basic concepts or categories, which form elementary granules (atoms) of knowledge, i.e. having an information system $S = (U, A)$ knowledge determined by S can be defined as $K = U/A$. Many important problems can be formulated and solved in the proposed framework. For example we can easily define some useful notions. If $S = (U, B)$, $S' = (U, B')$ and $U/B = U/B'$ we will say that S is *equivalent* to S', symbolically $S \simeq S'$. If $IND(B') \subseteq IND(B)$, then and S will be called *finer* then S', or S' - *coarser* then S, symbolically $S' \leq S$.

"Reality, or the world we all know, is only a description"

Carlos Castaneda in *Journey to Ixtlan: The Lesson of Don Juan*

Often the question arises whether all primitive concepts in S are necessary in order to define all basic concepts in S. This problem arises in many practical applications—and will be referred to as ***knowledge reduction.*** Formally this can be formulated in the discussed framework as follows. Suppose we are given $S = (U, A)$ and $B \subseteq A$. The question is whether $U/B = U/A$?

To answer this question we need some auxiliary notions.

- Let $a \in B$. We will say that a is *superfluous* in B if $IND(B) = IND(B - \{a\})$, otherwise a is *indispensable* in B.
- The set of attributes B is *independent* if all its attributes are indispensable.
- The set B' is a *reduct* of B if
  - B is independent, and
  - $IND(B') = IND(B)$.

Thus a reduct of B is the minimal subset of B such that $U/B' = U/B$, i.e. B' determines the same family of basic concepts as the set B. In other words reduction of knowledge boils down to the elimination of superfluous attributes in the information system.

The set of all independent attributes in B is referred to as the *core* of B, and is denoted *CORE(B)*. The following interesting property is valid:

$$CORE(B) = \cup RED(B), \tag{1}$$

where *RED(B)* is the family of all reducts of B.

Remark. The problem of reduction of knowledge is related to the general idea of independence discussed in mathematics as formulated by Marczewski [21] see also an overview paper by Glazek [12].

More about reduction of attributes can be found in [37].

4 Uncertainty, vagueness and rough sets

In this section we would like to discuss the central problem of our approach, the problem of vagueness and uncertainty. There are many conceptions of vagueness and uncertainty in logical and philosophical literature [2,3,10,34]. We present here so called "boundary-line" view, which is due to Frege, who writes:

The concept must have a sharp boundary. To the concept without a sharp boundary there would correspond an area that had not a sharp boundary-line all around. ([11]).

"We must distinguish between truth, which is objective, and certainty, which is subjective"

Karl R. Popper (1992)

Thus Frege's idea of vagueness is based on the boundary-line cases. i.e. if a concept is precise every object can be classified as belonging to this concept or not, whereas for vague concepts this is not the case and some object cannot be classified to the concept or its complement, forming thus the boundary line cases. For example the concept of an *odd (even) number* is precise, because every number is either odd or even - whereas the concept of a *beautiful women* is vague, because for some women it cannot be decided whether they are beautiful or not, (there are boundary-line cases). Thus if a concept is vague we are uncertain whether some objects (the boundary-line cases) belong to the concept or not. Hence vagueness is a property of concepts (sets), whereas uncertainty is a property of objects (elements), i.e. if a concept is vague its extension is uncertain.

The ideas considered in the previous sections can be easily employed to express these considerations more precisely.

In the presented approach concept is a subset of the universe. Suppose we are given an information system $S = (U, A)$ and let $X \subseteq U$. The concept (set) X will be said to be precise, if it is an union of some basic concepts (sets) of S, otherwise the concept (set) is vague (rough). Thus precise concepts can be defined in terms of basic concepts in S, whereas this is not the case for vague concepts. (Let us note that some concepts can be precise in one information system but vague in another one).

Basic idea of our approach to vagueness consists in replacing vague concept by a pair of precise concepts, called its *lower* and *upper approximations*. The difference between the upper and the lower approximation is the boundary region. For example the lower approximation of the concept of a beautiful women contains all women which are beautiful with certainty, whereas the upper approximation of this concept contains all women which are possibly beautiful, and the boundary region of this concept is formed by all women which can not be classified with certainty as beautiful or not beautiful. The "size" of the boundary region can be used as a *measure* of vagueness of the vague concept. The greater the boundary region, the "more" vague is the concept; precise concepts do not have the boundary region at all.

Formally the above considerations can be presented as follows.

Let $S = (U, A)$ be an information system, $X \subseteq U$ and $B \subseteq A$. With each subset $X \subseteq U$ and the set of attributes B we associate two subsets:

$$\underline{B}X = \cup\{Y \in U/B : Y \subseteq X\}$$

$$\overline{B}X = \cup\{Y \in U/B : Y \cap X \neg \emptyset\}$$

called the $B-$*lower* and the $B-$*upper* approximation of X (in S) respectively.

Set $BN_B(X) = \overline{B}X - \underline{B}X$ will be called the $B-$*boundary* of X. The set $\underline{B}X$ is the set of all elements of U which can be with *certainty* classified as elements of X, employing set of attributes B; the set $\overline{B}X$ is the set of elements of U which can be *possibly* classified as elements of X, employing the set of

attributes B; the set $BN_B(X)$ is the set of elements which cannot be classified either to X or to $-X$ using B.

The boundary region is the undecidable area of the concept X, and none of the objects belonging to the boundary region can be classified with certainty to X or $-X$ by using the set of attributes B.

Obviously a concept X is vague (rough) with respect to B, if and only if $\overline{B}X \neq \underline{B}X$, otherwise the concept X is precise.

In order to express how "vague" is a concept we can use a numerical evaluation of vagueness by defining the accuracy measure

$$\alpha_B(X) = card\underline{B}/card\overline{B}$$

where $X \neq \emptyset$.

Obviously $0 \leq \alpha_B(X) \leq 1$, for every B and $X \subseteq U$; if $\alpha_B(X) = 1$ the boundary region of X is empty and the set X is precise with respect to B; if $\alpha_B(X) < 1$, the set X has some non-empty R-boundary region and consequently is vague with respect to B.

Besides characterization of vague concept by means of numerical values one can also define qualitative characterization of vagueness, showing that there are four basic classes of vagueness, as defined below.

a) If $\underline{B}X \neq \emptyset$ and $\overline{B}X \neq U$, then we say that X is *roughly B-definable*,

b) If $\underline{B}X = \emptyset$ and $\overline{B}X \neq U$, then we say that X is *internally B-undefinable*,

c) If $\underline{B}X \neq \emptyset$ and $\overline{B}X = U$, then we say that X is *externally B-undefinable*,

d) If $\underline{B}X = \emptyset$ and $\overline{B}X = U$,then we say that X is *totally B-undefinable*.

The intuitive meaning of this classification is the following:

If set X is roughly B-definable, this means that we are able to decide for some elements of U whether they belong to X or $-X$.

If X is internally B-undefinable, this means that we are able to decide whether some elements of U belong to $-X$, but we are unable to decide for any element of U, whether it belongs to X or not.

If X is externally B-undefinable, this means that we are able to decide for some elements of U whether they belong to X, but we are unable to decide, for any element of U whether it belongs to $-X$ or not.

If X is totally B-undefinable, we are unable to decide for any element of U whether it belongs to X or $-X$.

That means, that the set X is roughly definable if there are some objects in the universe which can be positively classified, to the set X employing the set of attribute B. This definition also implies that there are some other objects which can be classified without any ambiguity as being outside the set X.

External B-undefinability of a set refers to the situation when positive classification is possible for some objects, but it is impossible to determine that an object does not belong to X on the basis of its features expressed by the set of attributes B.

Having defined the vagueness we are now in a position to define uncertainty. As mentioned before uncertainty is related to elements of the universe and

expresses how "strongly" an element belong to a concept. This idea can be expressed by the formula [27].

$$\mu_{X,B}(x) = card([x]_B \cap X)/cardX.$$

The intuitive meaning of the above formula is obvious. It is interesting to compare this formula with the membership function in the fuzzy set theory, but we will not discuss this problem here. More about it can be found in [27,28,36].

5 Conclusion

The rough set theory besides, its importance to data analysis, contributed also to understanding better vagueness and uncertainty.

References

[1] Aikins JS. Prototypic a knowledge for expert systems. Artificial Intelligence 1983; 20:163–210

[2] Black M. Vagueness. The Philosophy of Sciences 1937; 427–455

[3] Black M. Reasoning with loose concepts. Dialog 1963; 2:1–12

[4] Bobrow DG. A panel on knowledge representation. Proc Fifth Int'l Joint Conference on Artificial Intelligence, 1977, Carnegie-Melon University, Pittsburgh, PA

[5] Bobrow DG, Winograd T. An overview of KRL: a knowledge representation language. Journal of Cognitive Sciences 1977; 1:3–46

[6] Brachman RJ, Smith BC. Special issue of knowledge representation. SIGART Newsletter 1980; 70:1–138

[7] Brachman RJ, Levesque HJ. (eds) Readings in knowledge representation. Morgan Kaufmann Publishers Inc, 1986

[8] Buchanan B, Shortliffe E. Rule based expert systems. Addison-Wesley, Reading, Mass, 1984

[9] Davis R, Lenat D. Knowledge-based systems in artificial intelligence. McGraw-Hill, 1982

[10] Fine K. Vagueness, truth and logic. Synthese 1975; 30:265–300

[11] Frege G. Grundgesetze der arithmentik. 1903;2. Geach, Black (eds) In: Selections from the philosophical writings of Gotlob Frege, Blackweil, Oxford, 1970

[12] Glazek K. Some old and new problems in the independence theory. Colloquium Mathematicum 1979; 17:127–189

[13] Grzymala-Busse J. On the reduction of knowledge representation. Systems Proc of the 6th Int'l Workshop on Expert Systems and their Applications, Avignon, France, 1986; pp 463–478

[14] Grzymala-Busse J. Knowledge acquisition under uncertainty - a rough set approach. Journal of Intelligent and Robotics Systems 1988; 1:3–16

[15] Halpern J. (ed) Theoretical aspects of reasoning about knowledge. Proc of the 1986 Conference, Morgan Kaufman, Los Altos, CA 1986

[16] Hayes-Roth B, McDermott J. An inference matching for inducing abstraction. Communication of the ACM 1978; 21:401–410

[17] Hempel CG. Fundamental of concept formation in empirical sciences. University of Chicago Press, Chicago, 1952

[18] Hintika J. Knowledge and belief. Cornell University Press, Chicago, 1962

[19] Holland JH, Holyoak KJ, Nisbett RE, Thagard PR. Induction: processes of inference, learning, and discovery, MIT Press, 1986

[20] Hunt EB. Concept formation. John Wiley and Sons, New York, 1974

[21] Marczewski E. A general scheme of independence in mathematics, BAPS 1958; 731–736

[22] McDermott D. The last survey of representation of knowledge. Proc of the AISB/GI Conference on AI, Hamburg, 1978, pp 286–221

[23] Minski M. A framework for representation knowledge. In: Winston P (ed) The psychology of computer vision, McGraw-Hill, New York, 1975, pp 211–277

[24] Newell A. The knowledge level. Artificial Intelligence 1982; 18:87–127

[25] Orlowska E. Logic for reasoning about knowledge. Zeitshrift fur Math Logik und Grundlagen der Math 1989; 35:559–572

[26] Pawlak Z. Rough sets—theoretical aspects of reasoning about data. Kluwer Academic Publishers, 1991

[27] Pawlak Z, Skowron A. From the rough set theory to evidence theory. In: Fedrizzi M, Kacprzyk J, Yager RR (eds) Advances in the Dempster-Shafer theory of evidence, John Wiley and Sons, 1992 (to appear)

[28] Pawlak Z, Skowron A. Rough membership functions: a tool for reasoning with uncertainty. Algebraic Methods in Logic and Computer Science, Banach Center Publications, Institute of Mathematics, Polish Academy of Sciences, Warsaw, 1993; 28:135–150

[29] Popper K. The logic of scientific discovery. Hutchinson, London, 1959

[30] Rauszer C. Logic for information systems. Fundamenta Informaticae 1992 (to appear)

[31] Rauszer C. Knowledge representation for group of agents. In: Wolenski J (ed) Philosophical logic in Poland, Kluwer Academic Publishers, 1992 (to appear)

[32] Rauszer C. Rough logic for multi agent systems. Proc of the Conference Logic at Work, Amsterdam, 1992 (to appear)

[33] Rauszer C. Approximate methods for knowledge systems. Proc of the 7th Int'l Symposium on Methodologies for Intelligent Systems, Trondheim, 1993, pp 326–337

[34] Russell B. Vagueness. Australian Journal of Philosophy 1923; 1:84-92

[35] Russell B. An inquiry into meaning and truth. George Allen and Unwin, London, 1950

[36] Skowron A, Grzymala-Busse J. From the rough set theory to evidence theory. In: Fedrizzi M, Kacprzyk J, Yager RR (eds) Advances in the Dempster-Shafer theory of evidence, John Wiley and Sons, 1991 (to appear)

[37] Skowron A, Rauszer C. The discernibility matrices and functions in information systems, In: Slowinski R (ed) Intelligent decision support. Handbook of advances and applications of the rough set theory, Kluwer Academic Publishers, 1992, pp 311-362

[38] Slowinski R (ed) Intelligent decision support. In: Handbook of advances and applications of the rough set theory, Kluwer Academic Publishers, 1992

[39] Ziarko W. On reduction of knowledge representation. Proc 2nd Int'l Symp on Methodologies of Intelligent Systems, Charlotte, NC, 1987, pp 99–113

[40] Ziarko W. Acquisition of design knowledge from examples, Math Comput Modeling 1988; 10:551–554