Rough Sets
and Their Applications

by
Zdzislaw Pawlak

ICS Research Report 18/92

Warsaw University of Technology

Warsaw, September 1992
ROUGH SETS
AND THEIR APPLICATIONS

Zdzislaw Pawlak

Abstract

The rough set theory is a new mathematical tool to reason about uncertainty and vagueness. The proposed approach has been implemented on minicomputers and workstations in Poland and abroad and has found many interesting, real life applications. The article presents the philosophy underlying the rough sets theory and its applications are outlined.

Key words

Approximations, decision rules, expert systems, fuzzy sets, imprecision, information systems, knowledge, rough sets, uncertainty, vagueness.

1. Introduction

The rough sets theory was proposed by the author in 1982 [24] as a new mathematical tool to reason about vagueness and uncertainty. After ten years of pursuing the rough set theory it turned out that it is of substantial importance to AI and cognitive sciences, in particular expert systems, decision support systems, machine learning, machine discovery, inductive reasoning, pattern recognition, decision tables and the like.

By now several hundred articles concerning rough sets and their applications have been published. The rough sets approach has proved to be a very effective tool, with many successful, real-life applications to its credit [1,8,12-14,16-18,20-23,30,32-34,39,49] and several computer systems based on this idea were implemented on personal computers and work stations in Poland and abroad [10,35,42].

The rough set concept overlaps with other theories of uncertainty and vagueness, in particular with fuzzy sets [3-7,27,38,45], evidence theory [35], statistics [15], however it can be considered in its own rights.

More about rough sets theory and its applications can be found in [25] and [33].

2. What are Rough Sets?

The rough sets philosophy bears on the idea of classification. Our claim is that the ability to classify is fundamental feature of any living organism or robot, (agents) who in order to behave rationally in the outer realm, must constantly classify real or abstract entities, events, processes, signals etc., called in what follows objects. In order to classify one has to postpone small differences between objects, thus forming classes of objects which are not noticeably different. These indiscernibility classes can be viewed as elementary concepts used by the agent to build up his knowledge about reality. For example,
if objects are classified according to color, then the class of all objects classified as red form the concept of *redness*. Hence any agent, equipped with mechanisms of various classification patterns, forms a variety of elementary concepts, which are elementary building blocks (*granules, atoms*) of his knowledge about the world and himself. Therefore, in what follows we will call the family of classifications associated with an agent - his *knowledge*.

The granularity of knowledge causes that some notions cannot be expressed precisely within the available knowledge and can be defined approximately only. More exactly, a concept which can be expressed in terms of elementary concepts associated with the considered knowledge is *precise*, otherwise the concept is *rough* (*imprecise, vague*), and can not be expressed employing the knowledge.

In the rough sets theory each vague concept is replaced by a pair of precise concepts called its *lower* and *upper approximation*; the lower approximation of a concept consists of all objects which *surely* belong to the concept, whereas the upper approximation of the concept consists of all objects which *possibly* belong to the concept. For example, the concept of an *odd* (even) *number* is precise, because for each number it can be decided whether it is odd (even) or not - whereas the concept of a *beautiful women* is vague, because for some women it can not be decided whether they are beautiful or not, (there are boundary-line cases). Difference between the upper and the lower approximation is a *boundary region* of the concept, and it consists of all objects which cannot be classified with certainty to the concept or its complement employing available knowledge. The greater the boundary region, the more vague is the concept; if the boundary region is empty the concept is precise.

The idea of approximations is the basic tool in the rough set philosophy and will be defined more precisely in the next section.

4. Formal Definition

In order to present the above ideas formally we need a suitable method of representing classifications, which, as mentioned before, are the starting point of the rough set theory. To this end we will use the concept of an *information systems*, known also as an *attribute-value systems* or an *knowledge representation systems*.

Information system is a finite table with rows labelled by *objects*, columns are labelled by *attributes*, moreover with each attribute a finite set of its *values*, called *domain* of the attribute, is associated. To each object and an attribute a value of the attribute is associated. For example if the object were an *apple* and the attribute - color, then the corresponding entry in the table could be *red*.

Simple example of such table, which characterizes six stores in terms of some factors is shown below.

<table>
<thead>
<tr>
<th>Store</th>
<th>E</th>
<th>Q</th>
<th>S</th>
<th>R</th>
<th>L</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>high</td>
<td>good</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>500</td>
</tr>
<tr>
<td>2</td>
<td>high</td>
<td>good</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>-100</td>
</tr>
<tr>
<td>3</td>
<td>med.</td>
<td>good</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>200</td>
</tr>
<tr>
<td>4</td>
<td>low</td>
<td>avg.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>70</td>
</tr>
<tr>
<td>5</td>
<td>low</td>
<td>good</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>high</td>
<td>avg.</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>-20</td>
</tr>
</tbody>
</table>
Objects in the table are stores numbered from one to six and attributes are the following factors:

- \( E \) - empowerment of sales personnel
- \( Q \) - perceived quality of merchandise
- \( S \) - segmented customer base
- \( R \) - good refund policy
- \( L \) - high traffic location
- \( P \) - store profit or loss (in millions of US dollars)

Attribute \( E \) has the values high, medium and low; attribute \( S \) has values good and average; attributes \( R, L \) and \( P \) have attribute values yes and no, whereas values of attribute \( P \) are integers.

It is easily seen that each attribute in the table defines a partition of objects, i.e. an equivalence relation, such that two objects belong to the same equivalence class if they have the same attribute values. Thus attributes in the information system represent various classification patterns and the whole table can be regarded as a simple way of notation for families of classifications, or what is the same - families of equivalence relations.

Formally an information system is a pair \( S = (U,A) \), where \( U \) is a non-empty finite set of objects called the universe and \( A \) is a finite set of attributes. With every attribute \( a \) a set of its values, called the domain of \( a \), and denoted \( V^a \) is associated. Every attribute \( a \in A \), is a function \( a: U \rightarrow V^a \), which to each object \( x \in U \) uniquely associates an attribute value from \( V^a \). Objects can be anything we can think of, for example states, processes, moments of time, physical or abstract entities etc.

Every subset of attributes \( B \subseteq A \) defines uniquely an equivalence relation

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

As usually \( U/IND(B) \) denotes the family of all equivalence classes of the equivalence relation \( IND(B) \), i.e. the classification corresponding to \( IND(B) \).

The lower approximation of \( X \subseteq U \) by \( B \) is the union of equivalence classes of \( IND(B) \) which are included in \( X \), or formally

\[
BX = \bigcup \{Y \in U/IND(B) : Y \subseteq X\}
\]

The upper approximation of \( X \subseteq U \) by \( B \) is the union of all equivalence classes of \( IND(B) \) which have not-empty intersection with \( X \), i.e.

\[
\overline{BX} = \bigcup \{Y \in U/IND(B) : Y \cap X \neq \emptyset\}
\]

The boundary-line region is of course defined as \( B_{\overline{N}}(X) = \overline{BX} - BX \) and will be called the \( B \)-boundary of \( X \).

Set \( BX \) consists of all elements of \( U \) which can be with certainty classified as
elements of $X$ employing knowledge $B$; Set $\bar{B}X$ is the set of all elements of $U$ which can be possibly classified as elements of $X$ using set of attributes $B$; set $BN_B(X)$ is the set of all elements which cannot be classified either to $X$ or to $\neg X$ by means of attributes from $B$.

Now we are able to give the definition of the rough set. A set $X \subseteq U$ is rough with respect to $B$, if $\bar{B}X \neq \neg X$, otherwise the set $X$ is exact (with respect to $B$).

Thus a set is rough if it does not have sharp defined boundary, i.e. it can not be uniquely defined employing available knowledge.

For practical applications we need numerical characterization of vagueness, which will be defined as follows:

$$\alpha_B(X) = \frac{\text{card } B\overline{X}}{\text{card } \bar{B}X}$$

where $X \neq \emptyset$, called the accuracy measure.

The accuracy measure $\alpha_B(X)$ is intended to capture the degree of completeness of our knowledge about the set (concept) $X$.

Obviously $0 \leq \alpha_B(X) \leq 1$, for every $B$ and $X \subseteq U$; if $\alpha_B(X) = 1$ the $R$-boundary region of $X$ is empty and the set $X$ is definable in knowledge $B$; if $\alpha_B(X) < 1$ the set $X$ has some non-empty $B$-boundary region and consequently is undefinable in knowledge $B$.

The idea of approximation of sets is the basic tool in the rough set approach and is used to approximate description of some concepts (subsets of the universe) by means of attributes. For example, we might be interested whether there are factors characteristic for stores having high (above 100 Millions dollars) profit, and if not - to find the lower and the upper characteristic of these stores. The reader is advised to answer this question using the above given definitions.

Starting from the concept of classification we can also define a variety of other notions fundamental to rough sets philosophy and applications - needed to discover various relations between attributes, and objects. The most important ones are the dependency of attributes (cause-effect relations), redundancy of attributes and decision rule generation.

For example we may be interested whether the factor $P$ (store profit or loss) depends, exactly or approximately, on the remaining five factors, i.e. whether values of factor $P$ are determined by values of factors $E,Q,S,R$ and $L$ (dependency of attributes). If so, then the question arises if all the factors really influence the factor $P$ (redundancy of attributes), and if not, which are the ones which matters. The most important problem is to find a set of decision rules (exact or approximate) which determine the stories performance.

All these problems can be easily defined and investigated within the rough set theory, however we will drop these considerations here and an interested reader is referred to the book [25].
5. Applications of Rough Sets

The rough sets theory has proved to be very useful in practice. Many real life applications in medicine, pharmacology, industry, engineering, control, social sciences, earth sciences and other have been successfully implemented. Some of them are listed in the references [1,8,12-14,16-18,20-23,30,32-34,39,49]. Besides, the book edited by professor Roman Slowinski [33] can be used as a reference book on applications of the rough sets theory.

By now rough sets have been mainly used to data analysis. Data are very often imprecise. For example in medicine body temperature, blood pressure etc. have usually not exact numerical values but are rather expressed as qualitatively values, like normal, above normal or below normal etc.

Rough set theory is mainly used to vague data analysis. Main problems which can be solved using rough set theory in data analysis are data reduction, (elimination of superfluous data), discovering of data dependencies, data significance, decision (control) algorithms generation from data, approximate classification of data, discovering similarities or differences in data, discovering patterns in data and the like.

Machine learning is another important area where rough sets can be use. There is a variety of approaches to machine learning, however by now no commonly accepted theoretical foundations have been developed. It seems that the rough set approach, can be used as a theoretical basis for some problems in machine learning. Some ideas concerning the application of rough sets in this area can be found in [2,9,10,28,29,41,43,44,46].

Rough sets approach offers alternative methods to switching circuits synthesis and minimization, fault diagnosis and others. This is closely connected with boolean reasoning methods [19].

Image processing is also a promising field of the rough sets theory applications. Using basic concepts of the rough sets theory one can easily develop many basic algorithms for image processing and character recognition like, for example thinning algorithms, contour tracing etc.

In all the above mentioned classes of applications the same mathematical background is employed but different inference mechanisms as well as data structures are involved.

6. Problems

The concept of the rough set has inspired variety of research of both theoretical and practical nature. We will briefly outline some of them stressing the more practical ones.

Complexity and practical efficiency of the basic algorithms, are the most important problems to be studied more exactly in the near future. Besides, comparison to other theories (e.g. like fuzzy sets, theory of evidence, neural networks, mathematical morphology and others) are also of great importance and are by now extensively investigated. Also more practical questions need appropriate attention. In particular problems related to incomplete, and distributed data seem of primary importance, for very little has been done in these areas. The developed algorithms based on the rough sets approach are very well suited to parallel
processing, especially when appropriate hardware could be developed. Finally computing machine based on the rough sets concept, in which decision rules would play the role of elementary instructions is worthy consideration. Decision support systems would gain momentum having such tools.

References


[27] Z. Pawlak and A. Skowron. "Rough Membership Functions", Fuzzy Logic for


[40] A. Wasilewska." Syntactic Decision Procedures in Information Systems."


Recently published Research Reports of the Institute of Computer Science, W.U.T.


