1 Introduction

Rough set theory is a new mathematical approach to data analysis. The rough set approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, inductive reasoning and pattern recognition. It seems of particular importance to decision support systems and data mining. The theory has found many applications.

In the paper basic concepts of rough set theory will be presented and discussed. More about rough set theory and its applications can be found in the enclosed references.

2 The philosophy of rough sets

The starting point of rough set theory is the indiscernibility relation, first formulated by Leibniz and known as the principle of the identity of indiscernibles, which is the following:

"It all properties of objects are the same then the objects are identical".

This principle leads to clustering of elements of interest into granules, of indiscernible (similar) objects. In rough set theory these granules, called elementary sets (concepts) form basic building blocks of knowledge about the universe.

Every union of elementary concepts is referred to as a crisp or precise concept (set); otherwise a concept (set) is called rough, vague or imprecise. Thus rough concepts cannot be expressed in terms of elementary concepts. However, they can be expressed approximately by means of elementary concepts by employing the idea of the lower and the upper approximation of a concept. The lower approximation of a concept is the union of all elementary concepts which are included in the concept, whereas the upper approximation is the union of all elementary concepts which have nonempty intersection with the concept. In other words, the lower and the upper approximation of a concepts are the union of all elementary concepts which are surely and possibly included in the concept respectively. The difference between the lower and the upper approximation of the concept is its boundary region. Now it can easily be seen that a concept is rough if it has nonempty boundary
region, i.e., its lower and upper approximation are nonidentical. Obviously, if the lower and the upper approximations of the concept are the same, i.e., its boundary region is empty – the concept is crisp.

3 Database

The basic ideas of rough set theory can be explained in general terms but in order to give more intuitive insight into the theory we will start our consideration from a database. Let us first give a simple example of database presented in Table 1.

<table>
<thead>
<tr>
<th>Store</th>
<th>$E$</th>
<th>$Q$</th>
<th>$L$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>high</td>
<td>good</td>
<td>no</td>
<td>profit</td>
</tr>
<tr>
<td>2</td>
<td>med.</td>
<td>good</td>
<td>no</td>
<td>loss</td>
</tr>
<tr>
<td>3</td>
<td>med.</td>
<td>good</td>
<td>no</td>
<td>profit</td>
</tr>
<tr>
<td>4</td>
<td>no</td>
<td>avg.</td>
<td>no</td>
<td>loss</td>
</tr>
<tr>
<td>5</td>
<td>med.</td>
<td>avg.</td>
<td>yes</td>
<td>loss</td>
</tr>
<tr>
<td>6</td>
<td>high</td>
<td>avg.</td>
<td>yes</td>
<td>profit</td>
</tr>
</tbody>
</table>

Table 1

In the database six stores are characterized by four attributes, shown below:

- $E$ – empowerment of sales personnel,
- $Q$ – perceived quality of merchandise,
- $L$ – high traffic location,
- $P$ – store profit or loss.

Let us observe that each store has different description in terms of attributes $E, Q, L$ and $P$, thus all stores may be distinguished (discerned) employing information provided by all attributes. However, stores 2 and 3 are indiscernible in terms of attributes $E, Q$ and $L$, since they have the same values of these attributes. Similarly, stores 1, 2 and 3 are indiscernible with respect to attributes $Q$ and $L$, etc.

Each subset of attributes determines a partition (classification) of all objects into classes having the same description in terms of these attributes. For example, attributes $Q$ and $L$ aggregate all stores into the following classes: \{1, 2, 3\}, \{4\}, \{5, 6\}. Thus, each database determines a family of classification patterns which are used as a basis of further considerations.

Suppose we are interested in the following problem: what are the characteristic features of stores having profit (or loss) in view of information available in Table 1. In other words, the question is whether we are able to describe set (concept) \{1, 3, 6\} (or \{2, 4, 5\}) in terms of attributes $E, Q$ and $L$. Obviously this question cannot be answered uniquely in our case since stores 2 and 3 display the same features in terms of attributes $E, Q$ and $L$, but store 2 makes a profit, whereas store 3 has a loss. Thus information given in Table 1 is not sufficient to answer this question. However, we can give a partial answer to this question.

Let us observe that if the attribute $E$ has the value $high$ for a certain store, then the store makes a profit, whereas if the value of the attribute $E$ is $low$, then the store has a loss. Thus in view of information contained in Table 1, we can say for sure that stores 1 and 6 make a profit, stores 4 and 5 have a loss, whereas stores 2 and 3 cannot
be classified as making a profit or having a loss. Therefore we can give approximate answers only. Employing attributes $E$, $Q$ and $L$, we can say that stores 1 and 6 surely make a profit, i.e., surely belong to the set \{1, 3, 6\}, whereas stores 1, 2, 3 and 6 possibly make a profit, i.e., possibly belong to the set \{1, 3, 6\}. We will say that the set \{1, 6\} is the lower approximation of the set (concept) \{1, 3, 6\}, and the set \{1, 2, 3, 6\} – is the upper approximation of the set \{1, 3, 6\}. The set \{2, 3\}, being the difference between the upper approximation and the lower approximation is referred to as the boundary region of the set \{1, 3, 6\}. New let us give same formal notations and definitions.

By a database we will understand a pair $S = (U, A)$, where $U$ and $A$, are finite, nonempty sets called the universe, and a set of attributes respectively. With every attribute $a \in A$ we associate a set $V_a$, of its values, called the domain of $a$. Any subset $B$ of $A$ determines a binary relation $I(B)$ on $U$, which will be called an indiscernibility relation, and is defined as follows:

$$(x, y) \in I(B) \text{ if and only if } a(x) = a(y) \text{ for every } a \in A,$$

where $a(x)$ denotes the value of attribute $a$ for element $x$.

It can easily be seen that $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e. partition determined by $B$, will be denoted by $U/I(B)$, or simple $U/B$; an equivalence class of $I(B)$, i.e. block of the partition $U/B$, containing $x$ will be denoted by $B(x)$.

If $(x, y)$ belongs to $I(B)$ we will say that $x$ and $y$ are $B$-indiscernible. Equivalence classes of the relation $I(B)$ (or blocks of the partition $U/B$) are refereed to as $B$-elementary sets.

## 4 Approximations

Now we can define two basic operations in rough set theory:

$$B_s(X) = \{x \in U : B(x) \subseteq X\},$$

$$B^*(X) = \{x \in U : B(x) \cap X \neq \emptyset\},$$

which assign to every subset $X$ of the universe $U$ two sets $B_s(X)$ and $B^*(X)$ called the $B$-lower and the $B$-upper approximation of $X$, respectively.

The set

$$BN_B(X) = B^*(X) - B_s(X)$$

will be called the $B$-boundary region of $X$.

If the boundary region of $X$ is the empty set, i.e. $BN_B(X) = \emptyset$, then the set $X$ is crisp (exact) with respect to $B$; in the opposite case, i.e. if $BN_B(X) \neq \emptyset$, then the set $X$ is rough (inexact) with respect to $B$.

Rough sets can also be defined by using a rough membership function, defined as

$$\mu^B_X(x) = \frac{\text{card}(B(x) \cap X)}{\text{card}(B(x))}.$$ 

Obviously

$$0 \leq \mu^B_X(x) \leq 1.$$
The value of the membership function \( \mu_B^X(x) \) is a kind of conditional probability, and can be interpreted as a degree of certainty that \( x \) can be classified as \( X \), employing set of attributes \( B \).

5 Dependency of attributes

Another important issue in rough set theory is discovering dependencies between attributes in a database. Intuitively, a set of attributes \( D \) depends totally on a set of attributes \( C \), denoted \( C \Rightarrow D \), if all values of attributes from \( D \) are uniquely determined by values of attributes from \( C \). In other words, \( D \) depends totally on \( C \), if there exists a functional dependency between values of \( D \) and \( C \). In Table 1 there are not total dependencies whatsoever.

We would need also a more general concept of dependency of attributes, called a partial dependency of attributes. The partial dependency means that only some values of \( D \) are determined by values of \( C \). For example, in Table 1 the attribute \( P \) depends partially on attributes \( E, Q \) and \( L \).

Formally, the above idea can be formulated as follows.

Let \( D \) and \( C \) be subsets of \( A \). We say that \( D \) depends to a degree \( k \), \( 0 \leq k \leq 1 \), on \( C \), denoted \( C \Rightarrow_k D \), if

\[
k = \gamma(C, D) = \frac{\text{card}(POS_C(D))}{\text{card}(U)} = \frac{\sum_{x \in U} \text{card}(C_*(X))}{\text{card}(U)},
\]

where

\[
POS_C(D) = \bigcup_{X \in U/I(D)} C_*(X).
\]

For the dependency \( \{E, Q, L\} \Rightarrow_k \{P\} \) in Table 1 we have \( k = \frac{4}{6} = \frac{2}{3} \).

If \( k = 1 \), then \( D \) depends on \( C \) totally; \( 0 < k < 1 \) then depends partially on \( C \), and if \( k = 0 \), then \( D \) does not depend on \( C \).

6 Reduction of attributes

A reduct is a subset of condition attributes that preserves the degree of dependency. It means that a reduct is a minimal subset of condition attributes that enables to make the same decisions as the whole set of condition attributes.

Formally if \( C \Rightarrow D \) then subset \( C' \) of \( C \), such that \( \gamma(C, D) = \gamma(C', D) \) is called a \( D \)-reduct of \( C \).

For example, in Table 1 we have two reducts \( \{E, Q\} \) and \( \{E, L\} \) of condition attributes. This means that Table 1 can be replaced either by Table 2 or Table 3.

<table>
<thead>
<tr>
<th>Store</th>
<th>( E )</th>
<th>( Q )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>high</td>
<td>good</td>
<td>profit</td>
</tr>
<tr>
<td>2</td>
<td>med.</td>
<td>good</td>
<td>loss</td>
</tr>
<tr>
<td>3</td>
<td>med.</td>
<td>good</td>
<td>profit</td>
</tr>
<tr>
<td>4</td>
<td>no</td>
<td>avg.</td>
<td>loss</td>
</tr>
<tr>
<td>5</td>
<td>med.</td>
<td>avg.</td>
<td>loss</td>
</tr>
<tr>
<td>6</td>
<td>high</td>
<td>avg.</td>
<td>profit</td>
</tr>
<tr>
<td>Store</td>
<td>E</td>
<td>L</td>
<td>P</td>
</tr>
<tr>
<td>-------</td>
<td>----</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>high</td>
<td>no</td>
<td>profit</td>
</tr>
<tr>
<td>2</td>
<td>med.</td>
<td>no</td>
<td>loss</td>
</tr>
<tr>
<td>3</td>
<td>med.</td>
<td>no</td>
<td>profit</td>
</tr>
<tr>
<td>4</td>
<td>no</td>
<td>no</td>
<td>loss</td>
</tr>
<tr>
<td>5</td>
<td>med.</td>
<td>yes</td>
<td>loss</td>
</tr>
<tr>
<td>6</td>
<td>high</td>
<td>yes</td>
<td>profit</td>
</tr>
</tbody>
</table>

Table 3

It is easy to check that both Table 2 and Table 3 preserve degree of dependency between attributes $P$ and $E$, $Q$, $L$. Reduction of attributes is the fundamental issue in rough set theory.

7 Knowledge base

Each dependency $C \Rightarrow_k D$ in a database induces a set of decision rules of the form "if... then...", called a knowledge base.

Decision rules are implications defined in the standard way from attributes, attribute values and propositional connectives.

Given $x \in U$ and $B \subseteq A$ by $\Phi^B_x = \wedge_{a \in B}(a, v)$ we mean a formula such that $a(x) = v$ and $v \in V_a$.

Every dependency $C \Rightarrow_k D$ determines a set of decision rules

$$\{\Phi^C_x \rightarrow \Phi^D_x\}_{x \in U},$$

called knowledge base. For example for Table 1 we get the following decision rules (knowledge base):

1) if (E, high) and (Q, good) and (L, no) then (P, profit),

2) if (E, med.) and (Q, good) and (L, no) then (P, loss),

3) if (E, med.) and (Q, good) and (L, no) then (P, profit),

4) if (E, no) and (Q, avg.) and (L, no) then (P, loss),

5) if (E, med.) and (Q, avg.) and (L, yes) then (P, loss),

6) if (E, high) and (Q, avg.) and (L, yes) then (P, profit).

This set of decision rules describes some relationships hidden in Table 1 (the database), i.e., the partial dependency between condition and decision attributes. It can be also understood as a description of approximations of decisions in terms of conditions. Thus a knowledge base can be seen as a language used to express our knowledge about the universe, hidden in the database.
8 Credibility factor of a decision rule

With every decision rule \( \Phi^C_x \rightarrow \Phi^C_y \) we associate a credibility factor defined below:

\[
\mu(\Phi^C_x, \Psi^D_x) = \frac{\text{card}(|\Phi^C_x \land \Psi^D_x|_S)}{\text{card}(|\Phi^C_x|_S)}.
\]

There exists an interesting relationship expressed by the following formula:

\[
\pi_S(\Psi^D_x) = \sum_{y \in \hat{D}(x)} (\pi_S(\Phi^C_y) \cdot \mu_S(\Phi^C_y, \Psi^D_y)) = \sum_{y \in \hat{D}(x)} \pi_S(\Phi^C_y \land \Psi^D_y) \quad (*)
\]

where \( \hat{D}(x) \) is the set of representatives of \( D(x)/I(C) \), and \( \pi_S(\Phi^C_x) = \frac{\text{card}(|\Phi^C_x|_S)}{\text{card}(U)} \) and \( \pi_S(\Phi^C_x) \) is a probability that the formula \( \Phi^C_x \) is satisfied in \( S \).

Thus formula (*) shows the relationship between the probability of conditions, certainty factor of the decision rule and the probability of decisions.

Let us notice that the credibility factor can be viewed as a counterpart of a truth value of an implication in classical logic and thus any decision rule is in a sense an inference rule used to reason about properties of data.

9 Conclusions

Basic concepts of rough set theory are approximations, dependencies and decision rules. In fact approximations and dependencies are mutually exchangeable concepts used to express our imperfect knowledge about reality. Approximations express local properties of a database, i.e., they describe particular decisions in terms of conditions - whereas dependencies (partial or total) are used to describe global properties of a database, i.e., general relationships in the database. Decision rules form a description language for both approximations as well as dependencies.

Many real life-application of rough set theory have shown its usefulness in many domains. Medicine, pharmacology, banking, market research, engineering, speech recognition, material sciences, control, linguistic, databases are just some exemplary fields of application, where rough set theory has demonstrated its usefulness.

Very promising new areas of application of the rough set concept seems to emerge in the near future, e.g. rough control, rough data bases, rough information retrieval and rough neural network and others.

Despite of many achievements a wide variety of problems still requires consideration and extensive research. Particularly important seems to be hardware implementation of rough computer, which is badly needed to pursue many applications in the years to come.

About two thousand papers have been published so far about rough set theory and its application. Some information about the present state of rough set theory and some applications can be found in the references.

References


