IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 55-58

Rough Data Set Based Applications in Framing Decision Rules

Renu Vashist [1] & M.L Garg[2]

[1,2]School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, INDIA
Email: [1]vashist.renu@gmail.com, [2]ml.garg@smvdu.ac.in

ABSTRACT

Rough set theory is used for dealing with uncertainty in the hidden pattern of data. This paper outlines concepts of the rough set theory for lower and upper approximations, reduction of attributes and decision rules. It assumes that the information about the real world is given in the form of an information table which represents input data, gathered from any domain, such as, medicine, finance or the military. In this study we have acquired the data from the real estate domain and framed some possible and certain rules to make the decision related to the price of the house.

1. INTRODUCTION

Rough set theory, introduced by Zdzislaw Pawlak in the early 1980s is a new mathematical tool to deal with vagueness and uncertainty [2, 3,4]. Rough set theory can be seen as a mathematical approach to intelligent data analysis and data mining. This approach seems to be of fundamental importance to artificial intelligence (AI) and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, and knowledge discovery from databases, expert systems, decision support systems, inductive reasoning, and pattern recognition [1,5–9].

There are many important advantages of rough set approach to data analysis:

- Rough set theory(RST) provides efficient algorithms for finding hidden patterns in data;
- RST does not need any preliminary or additional information about data, such as probability distribution in statistics;
- RST evaluates significance of data;
- RST is used to generates sets of decision rules from data;
- RST offers straightforward interpretation of obtained results;
- RST is easy to understand..

Rough set theory can be considered as an extension of classical set theory. RST can be used for representing incomplete knowledge. The basic concept of the RST is the notion of approximation space; with every object of universe we associate some information i.e. Data and Knowledge.

Suppose we are given an information system $S = (U, A), X \subseteq U$ and $P \subseteq A$, where U and A, are finite, nonempty sets called the universe, and the set of attributes, respectively. Set A will contain two disjoint sets of attributes called condition and decision attributes and the system is donated by $S = (U, C, D)$ where C is called condition attribute and D is called decision attribute. With every attribute $a \in A$ we associate a set Va, of its values, called the domain of a.

Now we define two approximations $\underline{P(X)}$ and $\overline{P(X)}$ called the P-lower and the P-upper approximation of X respectively where

$$\underline{P(X)} = \bigcup x \in U \{P(x) : P(x) \subseteq X\} \text{ and}$$

$$\overline{P(X)} = \bigcup x \in U \{P(x) : P(x) \cap X \neq \phi\}.$$

Lower approximation will consist of all the members which surely belongs to the set and Upper approximation consist of all the members which possibly belongs to the set.

The boundary region is given by the set difference $\overline{P(X)} - \underline{P(X)}$ consists of those objects that can neither be ruled in nor ruled out as members of the target set X. If the boundary region is empty i.e $\overline{P(X)} = \underline{P(X)}$ then the set is crisp otherwise the set is rough.

Information Table in Rough Set

In this article we will assume that the information about the real world is given in the form of an *information table*. Thus, the information table represents input data, gathered from any domain, such as medicine, finance or the military. An example of such an information table is given in Table 1.

Rows of a table, labeled e1, e2, e3, e4, e5 and e6 in Table 1, are called *examples* (objects, entities). Properties

of examples are perceived through assigning values to some variables. We will distinguish between two kinds of variables: *attributes* (sometimes called condition attributes) and *decisions* (sometimes called decision attributes). For example, if the information table describes a hospital, the examples may be patients; the attributes, symptoms and tests and the decisions, diseases. Each patient is characterized by the results of tests and symptoms and is classified by the physicians (experts) as being on some level of disease severity. Here we are considering the data from the real estate. The attributes are the various parameters of a house and the decision is Price of house.

Table 1
Information Table
Condition Attributes

	Location	*Fireplace*	*Basement*
e1	bad	yes	yes
e2	good	yes	yes
e3	v_good	yes	yes
e4	bad	yes	no
e5	good	no	no
e6	v_good	yes	no

Indiscernibility Relation

The main concept of rough set theory is an *indiscernibility relation*, normally associated with a set of attributes for example the set consisting of attributes *Basement* and *fireplace* from Table1. Examples e1 and e2 are characterized by the same values of both attributes, for the attribute *Basement* the value is *yes* for e1 and e2 and for the attribute *fireplace* the value is *yes* for both e1 and e2. Moreover, example e3 is indiscernible from e1 and e2. Examples e4 and e6 are also indiscernible from each other. Obviously, the indiscernibility relation is an equivalence relation. Sets that are indiscernible are called *elementary sets*. Thus, the set of attributes *Basement* and *fireplace* defines the following elementary sets: {e1, e2, e3}, {e4,e6}, and {e5}. Any finite union of elementary sets is called a *definable set*. In our case, set {e1, e2, e3, e5} is definable by the attributes *basement* and *fireplace,* since we may define this set by saying that any member of it is characterized by the attribute *Basement* equal to yes and the attribute *fireplace* equal to yes or by the attribute *Basement* equal to no and the attribute *fireplace* equal to no.

With the help of indescernibility relation, we can define redundant or dispensable attributes. Table2 is called Decision Table because this table is having a Decision attribute. Usually decision is a single attribute. The decisions are the prices of houses according to the values of various condition attributes.

Decision Table 2

	Location	*Fireplace*	*Basement*	*Decision Price*
e1	bad	yes	yes	Low
e2	good	yes	yes	High
e3	v_good	yes	yes	High
e4	bad	yes	no	Low
e5	good	no	no	Low
e6	v_good	yes	no	High

If a set of attributes and its superset define the same indiscernibility relation (i.e. if elementary sets of both relations are identical), then any attribute that belongs to the superset and not to the set is redundant. In the example from Table 2, let the set of attributes be the set {*Location, Basement*} and its superset be the set of all three attributes, i.e. the set { *Location, Fireplace, Basement* }. Elementary sets of the indiscernibility relation defined by the set { *Location, Basement* } are singletons, i.e., sets {e1}, {e2}, {e3}, {e4}, {e5}, and {e6}, and elementary sets of the indiscernibility relation defined by the superset { *Location, Fireplace, Basement* } are also {e1}, {e2}, {e3}, {e4}, {e5}, and {e6} . Here the attribute *Fireplace* belongs to the superset and does't belong to the set. Thus, the attribute *Fireplace* is redundant. On the other hand, the set {*Location, Basement*} does not contain any redundant attribute, since elementary sets for attribute sets {*Location*} and {*Basement*} are not singletons. Such a set of attributes, with no redundant attribute, is called *minimal* (or independent).

Reduct or Covering

The set *P* of attributes is the *reduct* (or covering) of another set *Q* of attributes if *P* is minimal and the indiscernibility relations, defined by *P* and Q are same. In our example {*Location, Basement*} is a reduct of original set of attributes {*Location, Fireplace, Basement*}. Table3 represent a new decision table based on reduct.

Table 3
Reduced Decision

	Location	*Basement*	*Decision Price*
e1	bad	yes	Low
e2	good	yes	High
e3	v_good	yes	High
e4	bad	no	Low
e5	good	no	Low
e6	v_good	no	High

Concept

Now we will define decision attributes, we can define elementary set associated with the decision as subset of the set of all example with the same value of decision. Such subset are called concept. For table 2 and 3 the

concepts are {e1, e4, e5} and {e2, e3, e6}. The first concept corresponds to the set of all houses whose prices are low, the second one to the set of all houses whose prices are high. We can tell whether the prices of a house are low or high on the basis of attributes values of {*Location, Basement*}.

Now the data from Table3 is enhanced by two examples e7 and e8 as presented in Table4.

Table 4

	Location	*Basement*	*Decision Price*
e1	bad	yes	Low
e2	good	yes	High
e3	v_good	yes	High
e4	bad	no	Low
e5	good	no	Low
e6	v_good	no	High
e7	good	no	High
e8	v_good	no	Low

Elementary sets of indiscernibility relation defined by attributes *location* and *Basement* are {e1}, {e2},{e3}, {e4}, {e5, e7}, and {e6, e8}, while concepts defined by decision *price* are {e1, e4, e5, e8} and {e2, e3, e6, e7}. Obviously, in Table 4 the decision *price* does not depend on attributes *Location* and *Basement* since neither {e5, e7} nor {e6, e8} are subsets of any concept.In other words, neither concept is definable by the attribute set {*Location, Basement*}. We say that Table 4 is *inconsistent* because examples e5 and e7 are conflicting (or are inconsistent)—for both examples the value of any attribute is the same, yet the decision value is different. (Examples e6 and e8 are also conflicting).

2. LOWER AND UPPER APPROXIMATIONS

Rough set theory offers a tool to deal with inconsistencies[4]. For each concept X the greatest definable set contained in X and the least definable set containing X are computed. The former set is called a *lower approximation* of X the latter is called an *upper approximation* of X. In the case of Table 4, for the concept {e2, e3, e6, e7}, describing the houses having high price, the lower approximation is equal to the set {e2,e3}, and the upper approximation is equal to the set {e2, e3, e5, e6, e7, e8}, as depicted in Figure 1. Similarly, for the concept {e1, e4, e5, e8}, describing the houses having low prices, the lower approximation is {e1, e4} and the upper approximation is {e1, e4, e5, e6, e7, e8}. Either of these two concepts is an example of a *rough set*, a set that is undefinable by given attributes. The set {e5, e6, e7, e8}, containing elements from the upper approximation of X that are not members of the lower approximation of X, is called a *boundary region*. Elements of the boundary region cannot be classified as members of the set X. On the other hand, rough sets may also be defined as sets having nonempty boundary regions.

For any concept, rules induced from its lower approximation are certainly valid and such rules are called certain. While the rules induced from upper approximation of the concept are possibly valid and are called possible. The lower and upper approximation of two concepts is shown in Fig1.

Certain Rules

For Table 4 certain rules are:

(Location , bad) → (Price, low),

(Location , good) and (Basement , yes) → (Price , high),

(Location , v_good) and (Basement , yes) → (Price , high);

Possible Rules

For Table 4 possible rules are

(Basement , no) → (Price, low),

(Location , bad) → (Price , low),

(Location , good) → (Price , high),

(Location ,v_ good) → (Price , high).

A few measures of uncertainty were developed within rough set theory. The most frequently used are: a quality of lower approximation and a quality of upper approximation. For a given set X of examples, not necessarily definable by a set P of attributes, the quality of lower approximation is the ratio of the number of all elements in the lower approximation of X i.e 2 to the total number of examples i.e 8. Similarly, the quality of upper approximation is the ratio of the number of all elements in the upper approximation of X to the total number of examples. Thus, in the example from Table , for the concept X = {e1, e4, e5,e8}, the quality of lower approximation is 0.25 and the quality of upper approximation is 0.75. The quality of lower approximation may be interpreted as the ratio of the number of all certain classified examples by attributes from P as being in X to the number of all examples of the information table. It is a kind of relative frequency.

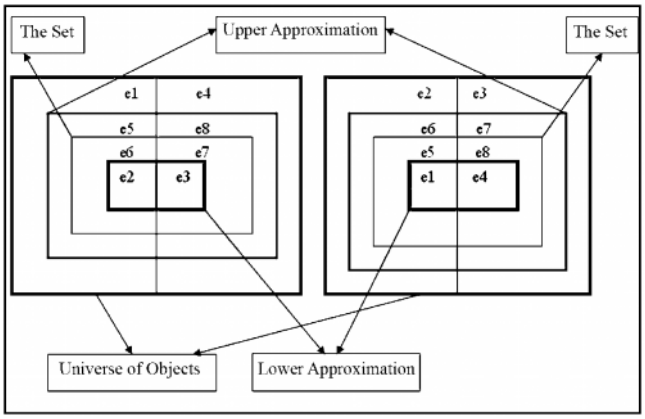

Fig 1: Lower and Upper Approximation of Two Concept

CONCLUSION

In the last two and a half decades the field of rough data set has taken a gigantic leap in terms of its applications in the growing number of disciplines like, economics, finance, medicine, business management, environmental engineering, software engineering, decision analysis, molecular biology and pharmacy.

Many types of improved rough set algorithms have been proposed for specific applications. An important and upcoming trend in the field of rough set is to develop hybrid methods, which incorporate rough sets and other computational methods such as fuzzy sets, neural networks etc. Although till date, it has a very limited application in the development of software based application tools. Therefore there are unlimited and interesting possibilities of development of integrated application software of rough sets as a future area of research.

REFERENCES

[1] S.K. Pal, A. Skowron (Eds.), *Rough Fuzzy Hybridization*, Springer, Berlin, 1999.

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

[3] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory, Knowledge Engineering and Problem Solving*, **9**, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

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

[5] Z. Pawlak, Decision Rules, Bayes_ rule and Rough Sets, in: N. Zhong, A. Skowron, S. Ohsuga (Eds.), *New Direction in Rough Sets, Data Mining, and Granular-Soft Computing, Springer*, Berlin, 1999, pp. 1–9.

[6] L. Polkowski, A. Skowron (Eds.), "Rough Sets and Current Trends in Computing", *Lecture Notes in Artificial Intelligence*, 1424, Springer, Berlin, 1998.

[7] L. Polkowski, A. Skowron (Eds.), "Rough Sets in Knowledge Discovery", **1–2**, Physica Verlag, A Springer Company, Berlin, 1998.

[8] L. Polkowski, S. Tsumoto, T.Y. Lin (Eds.), "Rough Set Methods and Applications–New Developments in Knowledge Discovery in Information Systems", Springer, Berlin, 2000, to Appear.

[9] N. Zhong, A. Skowron, S. Ohsuga (Eds.), *New Direction in Rough Sets Data Mining and Granular-Soft Computing*, Springer, Berlin, 1999.

[10] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory, Knowledge Engineering and Problem Solving*, **9**, Kluwer AcademicPublishers, Dordrecht, The Netherlands, 1991.[11] Z. Pawlak, Decision rules, Bayes' Rule and Rough Sets. In: Skowron et al. [280], pp. 1–9.