Knowledge Acquisition under Uncertainty through Rough Set

[1]Brojo Kishore Mishra, [2]Swarupananda Bissoyi and [3]Susanta Kumar Das

[1]Department of Computer Science, MITS, Rayagada.
mishra.bkm@gmail.com
[2]Development Lead, Samsung India S/w operation, Bangalore.
[3] Department of Computer Science, Berhampur University.
dr.dassusanta@yahoo.co.in

***Abstract* - Knowledge acquisition under uncertainty using rough set theory was first stated as a concept and was introduced by Z.Pawlak in1981. A collection of rules is acquired, on the basis of information stored in a data base-like system, called an information system. Uncertainty implies inconsistencies, which are taken into account, so that the produced are categorized into certain and possible with the help of rough set theory. The approach presented belongs to the class of methods of learning from examples. The taxonomy of all possible expert classifications, based on rough set theory, is also established. It is shown that some classifications are theoretically (and, therefore, in practice) forbidden. For a set of conditions of the information system, and a given action of an expert, lower and upper approximations of a classification, generated are computed in a straightforward way, using their simple definitions. Such approximations are the basis of rough set theory. From these approximations, certain and possible rules may be determined. Certain rules have been propagated separately during the inference process, producing new certain rules. Similarly, possible rules are likely to propagate in a parallel way. Example on the basic of knowledge Acquisition has been discussed in brief.**

***Key words* - Rough Set, Lower Approximation, Upper Approximation, Knowledge Acquisition and Rule Generation.**

I. INTRODUCTION

Rough Set Theory is a mathematical formalism for representing uncertainty that can be considered as an extension of the classical set theory. It has been used in many different research areas, including those related to inductive machine learning and reduction of knowledge in knowledge-based systems. We can observe the following about the rough set approach:

- Introduction of efficient algorithms for finding hidden patterns in data,
- Determination of optimal sets of data (data reduction),
- Evaluation of the significance of data,
- Generation of sets of decision rules from data,
- Easy-to-understand formulation,
- Straightforward interpretation of obtained results,
- Suitability of many of its algorithms for parallel processing.

Rough set theory, proposed by Pawlak in 1982, can be seen as a new mathematical approach to vagueness. The rough set philosophy is founded on the assumption that with every object of the universe of discourse we associate some information (data, knowledge). Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory. In rough set approach indiscernibility is defined relative to a given set of functional (attributes).

The basic assumption of rough set theory as put forth by Pawlak is that human knowledge about a universe depends upon their capability to classify its objects. Classifications of a universe and equivalence relations defined on it are known to be interchangeable notions. So, for mathematical reasons equivalence relations were considered by Pawlak to define rough sets. A pair of crisp sets, called the lower and upper approximations of the set, represents a rough set. The lower approximation of a rough set comprises of those elements of the universe, which can be said to belong to it definitely with the available knowledge. The upper approximation on the other hand comprises of those elements, which are possibly in the set with respect to the available information. The concept of rough sets was primarily concerned with the study of intelligent systems characterized by insufficient and incomplete information.

Any set of all indiscernible (similar) objects is called an elementary set and forms a basic granule of knowledge about the universe. Any union of some elementary sets is referred to as crisp set- otherwise the set is rough.

For algorithmic approach we divide the attributes into two types:

1. Conditions
2. Decisions (or Actions)

Objects are described by values of conditions, while classifications of experts are represented by values of decisions. For a set of conditions of the information system and a given action d of an expert, lower and upper approximations of a classification, generated by d, may be computed in a straight forward way, using their simple definition. Such approximations are the basis of rough set theory. From these approximations, certain and possible rules may be determined for action d, again in a straightforward way. Induced rules are categorized into certain and possible.

II. DEFINITIONS

Let *U* be an universe of discourse and *R* be an equivalence relation over *U*. By *U/R* we denote the family of all equivalence classes of *R*, referred to as categories or concepts of *R* and the equivalence class of an element $x \in U$ is denoted by $[x]_R$.

Definition 1 By a *knowledge base*, we understand a relational system $K = (U, \Re)$, where *U* is as above and $\Re$ is a family of equivalence relations over *U*.

Definition 2 For any subset $P(\neq \phi) \subset \Re$, the intersection of all equivalence relations in *P* is denoted by $IND(P)$ and is called the *indiscernibility relation over P*. We define

IND(K) = { IND(P): $P(\neq \phi) \subset \Re$ }.

Definition 3 Let $X \subseteq U$ and R be in IND(K).The sets $\underline{R}X$ and $\overline{R}X$ are called the *R-lower* and *R-upper approximation*s of X respectively and are defined as follows:

$\underline{R}X = \{x \in U: [x]_R \subseteq X\}$, $\overline{R}X = \{x \in U: [x]_R \cap X \neq \phi\}$.

Definition 4 For $X \subseteq U$, the *R–boundary of X* is denoted by $BN_R(X)$ and is defined as

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

Definition 5 A set $X \subseteq U$ is said to be *rough with respect to R* if and only if $\underline{R}X \neq \overline{R}X$; that is, $BN_R(X) \neq \phi$. X is said to be R-*definable* if and only if $\underline{R}X = \overline{R}X$ or $BN_R(X) = \phi$.

It may be noted that R-definable sets are crisp sets with respect to R. Many properties of the lower and upper approximations of rough sets, union of rough sets and intersection of rough sets have been obtained.

Definition 6 Let $x \in U$ and $X \subseteq U$. We say x *is certainly in* X with respect to R if and only if $x \in \underline{R}X$ and x *is possibly in* X if and only if $x \in \overline{R}X$.

III. APPROXIMATION OF CLASSIFICATIONS

Classifications of universes play central roles in basic rough set theory. We define below a classification formally.

Definition 1 Let $F = \{X_1, X_2, \ldots, X_n\}$ be a family of non empty sets defined over U. We say that F is a *classification of* U if and only if $X_i \cap X_j = \phi$ for $i \neq j$ and $\bigcup_{i=1}^{n} X_i = U$.

Definition 2 Let F be as above and R be an equivalence relation over U. Then $\underline{R}F$ and $\overline{R}F$ denote respectively the R-*lower* and R-*upper approximations* of the family F and are defined as

$\underline{R}F = \{\underline{R}X_1, \underline{R}X_2, \ldots\ldots\ldots\ldots, \underline{R}X_n\}$,

$\overline{R}F = \{\overline{R}X_1, \overline{R}X_2, \ldots\ldots\ldots\ldots \overline{R}X_n\}$.

We assume that F is a classification of U and R is an equivalence relation over U.

Grzymala-Busse has established some properties of approximation of classifications. These results are irreversible by nature. Pawlak has noted that these results of Busse establish that the two concepts, approximation of sets and approximation of families of sets (or classifications) are two different issues and that the equivalence classes of approximate classifications cannot be arbitrary sets. He has further stated that if we have positive example of each category in the approximate classification then we must have also negative examples of each category. In this article, we further analyze these aspects of theorems of Busse and provide physical interpretation of each one of them by taking a standard example.

One primary objective is to extend the results of Busse by obtaining necessary and sufficient type theorems and show how the results of Busse can be derived from them. The results of Busse we discuss here are in their slightly modified form as presented by Pawlak.

Theorems on approximation on classifications

In this section, we shall establish two theorems which have many corollaries generalizing the four

theorems established by Busse and presented in slightly modified forms by Pawlak. We shall also provide interpretations for most of these results including those of Busse and illustrate them through a simple example of toys.

We shall use the following notations for representational convenience:

$N_n = \{1, 2, ..., n\}$.

For any $I \subset N_n$, I^c is the complement of I in N_n.

Theorem 1 *For any* $I \subset N_n$, $\overline{R}(\bigcup_{i \in I} X_i) = U$ *if and only if* $\underline{R}(\bigcup_{j \in I^C} X_j) = \phi$.

Corollary 1 *Let* $F = \{X_1, X_2, ..., X_n\}$ *be a classification of U and let R be an equivalence relation on U and* $I \subset N_n$. *If* $\overline{R}(\bigcup_{i \in I} X_i) = U$ *then* $\underline{R}X_j = \phi$ *for each* $j \in I^c$.

Corollary 2 For *each* $i \in N_n$, $\overline{R}X_i = U$ *if and only if* $\underline{R}(\bigcup_{j \neq i} X_j) = \phi$.

Taking $I = \{i\}^c$, in Theorem 3.1 we get

Corollary 3 For *each* $i \in N_n$, $\underline{R}X_i = \phi$ *if and only if* $\overline{R}(\bigcup_{j \neq i} X_j) = U$.

Corollary.4 *If there exists* $i \in N_n$ *such that* $\overline{R}X_i = U$ *then for each* $j (\neq i) \in N_n$, $\underline{R}X_j = \phi$.

Corollary 5 *If* $\overline{R}X_i = U$ *for all* $i \in N_n$ *then* $\underline{R}X_i = \phi$ *for all* $i \in N_n$.

Theorem 2 For *any* $I \subset N_n$, $\underline{R}(\bigcup_{i \in I} X_i) \neq \phi$ *if and only if* $\bigcup_{j \in I^c} \overline{R}X_j \neq U$.

Corollary 6 For $I \subset N_n$, *if* $\underline{R}(\bigcup_{i \in I} X_i) \neq \phi$ *then* $\underline{R}X_j \neq U$ *for each* $j \in I^c$.

Corollary 7 For *each* $i \in N_n$, $\underline{R}X_i \neq \phi$ *if and only if* $\bigcup_{j \neq i} \overline{R}X_j \neq U$.

Taking $I = \{i\}^c$ in Theorem 3.2 we get

Corollary 8 For *all* $i, 1 \leq i \leq n$, $\overline{R}X_i \neq U$ *if and only if* $\underline{R}(\bigcup_{j \neq i} X_j) \neq \phi$.

By Corollary 3.7, $\underline{R}X_i \neq \phi \Rightarrow \bigcup_{j \neq i} \overline{R}X_j \neq U$ for each $j \neq i,\ 1 \leq j \leq n$.

Corollary 9 If there exist $i \in N_n$ *such that* $\underline{R}X_i \neq \phi$ *then for each* $j (\neq i) \in N_n$, $\overline{R}X_j \neq U$.

Corollary 10 If *for all* $i \in N_n$, $\underline{R}X_i \neq \phi$ *holds then* $\overline{R}X_i \neq U$ *for all* $i \in N_n$.

Some properties of classifications

In this section we shall establish some properties of measures of uncertainty and discuss in detail on properties of classifications with two elements and three elements.

III(A) MEASURES OF UNCERTAINTY

We denote the number of elements in a set A by card(A).

Definition 1 Let $F = \{X_1, X_2, ..., X_n\}$ be a classification of U and R be an equivalence relation on U. Then we denote the *accuracy of approximation of F by R* by $\alpha_R(F)$ and define it as

$$\alpha_R(F) = (\sum_{i=1}^{n} card(\underline{R}X_i)) / (\sum_{i=1}^{n} card(\overline{R}X_i)).$$

Definition 2 Let F and R be as above. Then we denote the *quality of approximation of F by R* is denoted by $\gamma_R(F)$ and define it as

$$\gamma_R(F) = (\sum_{i=1}^{n} card(\underline{R}X_i)) / card(U).$$

The accuracy of classification expresses the percentage of possible correct decision when classifying objects employing the knowledge of R. The quality of classification expresses the percentage of objects which can be correctly classified to classes of F employing knowledge of R.

Let R_1 and R_2 be any two equivalence relations on U. F_1 and F_2 be the classifications of U induced by R_1 and R_2 respectively.

Definition 3

(i) We say that R_2 *depends in degree k on* R_1 in U and denote it by $R_1 \xrightarrow{k} R_2$, if and only if $\gamma_{R_1}(F_2) = k$.

(ii) We say that R_2 *totally depends on* R_1 in U if and only if k =1.

(iii) We say that R_2 *roughly depends on* R_1 in U if and only if $0 < k < 1$.

(iv) We say that R_2 *is totally independent on* R_1 in U if and only if k = 0.

(v) We say F_2 *depends in degree k on* F_1 in U, written as $F_1 \xrightarrow{k} F_2$ if and only if $F_1 \xrightarrow{k} F_2$.

Property 1 *For any R-definable classification F in U,* $\beta_R(F) = \gamma_R(F) = 1.$

So, if a classification F is R-definable then it is totally independent on R.

Property2 *For any classification F in U and an equivalence relation R on U,* $0 \le \beta_R(F) \le \gamma_R(F) \le 1.$

IV ALGORITHM

Step 1: Take the universal set in a Two-D array and in which first column is universal set and second column will be the set in number to which the member will belong while entering knowledge set and third column will be the count of the member that is how many it has come in the sets .

Step 2: We will first set all the values of second and third column to 0 .

Step 3: After entering all the knowledge sets we will check the values of the table in second and third column if any of them is 0 or any value in third column is not one, the sets are declared as not classified and user is asked to enter the sets again .

Step 4: Now here one more 2-D array is used to find the lower and upper approximation of the decision set.

Step 5: The 2-D array will have the number of rows equal to the number of sets entered. So as soon the values of the decision set is entered the count column which is the first column in the second 2-D array gets incremented which denotes that how many members are from which set.

Step 6: Then we will check if the number of members in the decision set is equal to the number of members in original set then that set is included in Lower Approximation.

Step 7: If its number of members matched is not 0 then it will be included in Upper Approximation.

Step 8: Now we will just check and compare which all sets are there in lower and upper approximation if they are same then the set is Crisp else it is Rough.

Step 9: For finding types, two variables with initial value 0 for both are set, and we will check their values for finding types.

RULE GENERATION MODULE

As the development of a knowledge-based system (KBS) involves: identifying a real world problem solving task that is to be tackled, representing the key components of this task in the KBS, and implementing the inference process that produces solutions. Thus there are two key components involved in the knowledge engineering process. There is the task of producing a representation of the problem that captures the key features and the task of developing an inference mechanism.

So in this module our aim is to develop an inference mechanism using the representational model from the previous model and pattern recognition from that in the form or boundaries. The output of this module will be Rules which are the inferences from the knowledge databases.

In the previous module we were considering only one set for finding boundaries. Now we will be considering a classification and not a set because in the process of learning from examples, rules are derived from classifications generated by single decisions.

Classifications of universes play central roles in basic rough set theory. We define below a classification formally.

We assume that F is a classification of U and R is an equivalence relation over U.

Let F = $\{X_1, X_2, \ldots, X_n\}$ be a family of non empty sets defined over U

Module completes the following Tasks:

1. To input the data file and to store the conditions and decisions in separate arrays to represent them in rough set model.

2. To cross check the input file with the input given by user for conditions and decisions.

3. To store all the possible conditions in an array to find all the condition set possibilities and to proceed further with rough set model.

4. To represent the given input file in the Rough Set representation and check for Classifications using Representational Model developed in module 1 for both the condition sets and decision sets.

F is a *classification of* U if and only if $X_i \bigcap X_j = \phi$ for i ≠ j and $\bigcup_{i=1}^{n} X_i$ = U.

5. After finding the classifications we will find the Lower and Upper Approximation with several Decision sets.

Let F be as above and R be an equivalence relation over U. Then $\underline{R}$ F and $\overline{R}$ F denote respectively the R-*lower* and R-*upper approximations* of the family F and are defined as

$\underline{R}$ F={ $\underline{R}X_1, \underline{R}X_2, \ldots\ldots\ldots\ldots, \underline{R}X_n$ },

$\overline{R}$ F={ $\overline{R}X_1, \overline{R}X_2, \ldots\ldots\ldots\ldots \overline{R}X_n$ }.

6. To find measures of Uncertainty i.e.

Let $F = \{X_1, X_2, ..., X_n\}$ be a classification of U and R be an equivalence relation on U as above

1. Accuracy of Approximation.

$$\alpha_R(F) = (\sum_{i=1}^{n} card(\underline{R}X_i)) / (\sum_{i=1}^{n} card(\overline{R}X_i))$$

2. Quality of Approximation.

$$\gamma_R(F) = (\sum_{i=1}^{n} card(\underline{R}X_i)) / card(U)$$

We denote the number of elements in a set A by card(A).

7. To find how many rules will be certain and how many will be uncertain with conditions for which theory has been developed and presented below.

8. To find the final rules for decision sets.

It is easy to have the following observations regarding the existence of certain and possible rules from classifications:

V. OBSERVATION ON RULE GENERATION

Observation 1: For C-definable classifications, all the rules are certain rules.

Observation 2: For roughly C-definable strong and roughly C-definable weak classifications both certain and possible rules exist.

Observation 3: For totally C-definable, internally C-undefinable strong and internally C- undefinable weak classifications there are no certain rules.

Observation 4: For roughly C-definable strong classifications the number of certain rules is equal to the number of elements in the classification.

Observation 5: For all types of classifications other than C-definable classification has the property that there is at least one possible rule.

Observation 6: For roughly C-definable weak classifications there is at least one certain rule.

Observation 7: For totally C-undefinable classifications, there is no certain rule. The number of possible rules is equal to the number of elements in the classification.

Observation 8: For internally C-undefinable strong classifications, there is no certain rule. The number of possible rules is at most equal to the number of elements in the classification.

Observation 9: For internally C-undefinable weak classifications, there is no certain rule. There is no certainty about the existence of possible rules.

VI. ALGORITHM

Step 1: First of all the rough set model developed in Module 1 is used to form the input sets.

Step 2: Now as the decision sets entered are multiple, so to find Lower and Upper Approximation for the number of sets entered in decision the number of columns in both the 2-D array has been increased with the number of sets entered, so that it can store for all sets and then we can easily find the lower and upper approximation.

Step 3: Input is taken from the user for number of conditions and types of each and all combinations are made in a 2-D array named condition.

Step 4: Input is taken for the number of experts for decisions and types of decisions they take and store in an array named decision.

Step 5: Now the file is taken as the input and is cross checked first for the number of conditions and decisions given by the user and is accepted if

matched and if not then again asked by the user .

Step 6: File taken is stored in the form of Model given in module1 in Table array.

Step 7: Table array is checked for classification according to the module 1 .

Step 8: First the conditions given in table are combined into groups then with respect to number of test cases they are classified.

Step 9: Decisions in Table array are separately classified using Step 2.

Step 10: Lower and Upper Approximations for each decision set and for all experts are made individually Using Step2.

Step 11: Now first we will find the type of definability from the 11 cases we have developed.

Step 12: As we have the type of definability we will get the number of rules and inferences possible for the particular expert.

Step 13: Now we will find the value for Measures of Uncertainty:

1. The accuracy of approximation of decision set whose approximations are found: It is the ratio of sum of the numbers of all certainly classified objects of attributes from the attribute set, to the sum of the numbers of all possibly classified objects, by attributes from the same set.
2. The quality of approximation of decision set whose approximations are found: It is the ratio of the sum of the numbers of all certainly classified objects by attributes from the attribute set, to the number of all objects of the system.

Step 14: Using these two uncertainty parameters and the Lower and Upper Approximations Definability we will find the number of certain rules and uncertain rules and there dependencies for each expert with the theory developed.

Algorithm for finding Missing attributes:

Step 1: Capture the rules generated in module 2.

Step 2: Find the Case where attributes are missing.

Step 3: Find the Measures of Uncertainty, depending upon which missing attributes will be found.

Step 4: Find the appropriate rule from the rule table for missing attribute.

Step 5: Generate the attribute

VII. CONCLUSION

Most of the knowledge in real life is uncertain or imprecise in character. If one tries to do away with these then much of the desired knowledge is lost. Busse, for the first time developed theories to deal with such knowledge base through rough set approach. Also, he has taken inconsistency into account. Automation of his approaches has been carried out by us successfully in this project. It is worth noting that we have developed some of the algorithms ourselves and made enhancement of the theory whenever required.

A new approach to knowledge acquisition under uncertainty can be build based on rough set theory. The real world phenomena are consist of information system, where inconsistencies are included can be tested through this model. Such inconsistencies are present because of different actions of the same expert for different objects described by the same values of conditions. Different actions of different experts for the same object are another source of inconsistency. For a set of conditions of the information system, and a given action of an expert, lower and upper approximations of a classification, generated, may be computed in a straightforward way, using their simple definitions. Such approximations are the basis of rough set theory. From these approximations, certain and possible rules may be determined for action. Induced rules are categorized into certain and possible, because they are computed from lower and upper approximations, can be implemented. Certain rules may be propagated separately during the inference process, producing new certain rules to solve the problem of inconsistent and generate the output. Propagation of rules in both subsystems may occur concurrently. Thus, a new implementation of production systems, based on rough set theory, is expected to be more efficient in knowledge and data information dealing as compared to existing software in same direction. The build-up system will give immediate conclusions, useful for induction of rules, as in track of discussed examples.

The rule generation using rough sets has been studied by many authors and different types of efficient algorithms have been developed by them. These new methods can be implemented. However, it is worth noting that each algorithm has its specific area of application. So, one can develop a package of such rule generation algorithms in future. Also, reduction in the number of attributes in a knowledge base makes the storage and analysis more efficient. This can be added in future.

REFERENCE

1. Dubois, D. and Prade, H., Twofold fuzzy sets and rough sets, fuzzy sets and systems 17(1985).
2. J.W.Grzymala-Busse., On the reduction of knowledge representation systems, Proc. 6th international workshop

on expert systems and their applications, Aviguan, France, April 28-30 (1986), pp. 463-478 .

3. J.W.Grzymala-Busse., Knowledge acquisition under uncertainty – a rough set approach, journal of intelligent and robotic systems, (1988), pp. 3-16.
4. J.W.Grzymala-Busse., Rough set strategies to Data with missing attribute values, found and novel approaches in data mining, T.Y.L in, S.Ohsuga, C.J.Lian and X.hu (Eds). Studies in computational Intelligence, vol.9 .special edition, (2006), pp. 197-212.
5. Mishra, B,K., Knowledge Acquisition under Uncertainity – A Rough Set Approach, M.Tech – project Dissertation, Berhampur University, 2008.
6. Pawlak, Z., Rough sets, International journal of computer and information sciences ,II (1982),pp.341-356.
7. Pawlak, Z., Rough classification, International journal Man-Machine studies, 24 (1983), pp. 469-483.
8. Pawlak, Z., Rough sets. - Basic notions, Institute Comp.Sci. Acad. Sci. Rep. No. 431, Warsaw(1981).
9. Pawlak, Z., Classification of objects by means of attributes, Institute Comp. Sci. Polish Acad. Sci. Rep. No.429, Warsaw,(1981).
10. Pawlak, Z., Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991.
11. Pawlak,Z. and Skowron,A., Rudiments of rough sets, Information Sciences-An International Jounral.Elsevier,177(1)(2007),3-27.
12. Pawlak,Z. and Skowron,A., Rough sets: Some extensions, Information Sciences-An International Jounral.Elsevier,177(1)(2007),28-40.
13. Pawlak,Z. and Skowron,A., Rough sets and Boolean reasoning, Information Sciences-An International Jounral.Elsevier,177(1)(2007),41-73.