Volume 5, Issue 08, August 2020

ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

More Accurate Approach for Recovering Missing Value Using Rough Set

Arpita Lasod
M Tech(C.S.E.) 4th Semester
Department of Computer Science and Engineering
Lord Krishna College of Technology Indore
Indore M.P. India

Mr Rahul Pawar
Assistant Professor
Department of Computer Science and Engineering
Lord Krishna College of Technology Indore
Indore M.P. India

Abstract: ***Rough set theory (RST) is a major mathematical method developed by Pawlak in 1982 (Pawlak, 1982). The RST has been applied in several fields including image processing, data mining, pattern recognition, medical informatics, knowledge discovery and expert systems. Rough set have been combined with methods such as neural networks, fuzzy logic etc resulting some good results. The use of rough set theory to solve a specific complex problem and has attracted world-wide attention of further research and development, extending the original theory and increasingly widening fields of application. Rough set as a computationally efficient technique it presents a basic significance to many theoretical developments and practical applications of computing and automation, especially in the areas of machine learning and data mining, decision analysis and intelligent control. In this paper we proposed a new approach based on rough set theory to recover missing values. We compare proposed approach with other approach to compare the performance. By the experimental analysis we found that proposed approach recover missing values more accurately as compare to other approach.***

Keywords: - Rough set, lower approximation ,upper approximation ,boundary, discernibility

I. Introduction

Rough set theory (RST) is a major mathematical method developed by Pawlak in 1982 (Pawlak, 1982). This method has been developed to manage uncertainties from information that presents some incompleteness and noises. When the available information is insufficient to determine the exact value of a given set, lower and upper approximations can be used by rough set for the representation of the concerned set. The approximation synthesis of concepts from the acquired data is the main objective of the rough set analysis. For example, if it is difficult to define a concept in a given knowledge base, rough sets can approximate with respect to that knowledge. In decision making, it has confirmed that rough set methods have a powerful essence in dealing with uncertainties.

The RST has been applied in several fields including image processing, data mining, pattern recognition, medical informatics, knowledge discovery and expert systems. Rough set have been combined with methods such as neural networks, fuzzy logic etc resulting some good results. The use of rough set theory to solve a specific complex problem and has attracted world-wide attention of further research and development, extending the original theory and increasingly widening fields of application. Rough set as a computationally efficient technique it presents a basic significance to many theoretical developments and practical applications of computing and automation, especially in the areas of machine learning and data mining, decision analysis and intelligent control.

Among other computational problems, rough set addresses problems such as data significance evaluation, hidden pattern discovery from data, decision rule generation, data reduction and data-driven inference interpretation.

Basic Concepts of Rough Sets

Rough set theory proposes a new mathematical approach to imperfect knowledge, i.e. to vagueness (or imprecision). In this approach, vagueness is expressed by a boundary region of a set. Rough set concept can be defined by means of topological operations, interior and closure, called approximations.

Let a finite set of objects U and a binary relation $R \subseteq U \times U$ be given. The sets U, R are called the universe and an indiscernibility relation, respectively.

The discernibility relation represents our lack of knowledge about elements of U. For simplicity, assume that R is an equivalence relation. A pair (U,R) is called an approximation space, where U is the universe and R is an equivalence relation on U.

Let X be a subset of U, i.e. $X \subseteq U$. Our goal is to characterize the set X with respect to R. In order to do it, we need additional notation and basic concepts of rough set theory which are presented below.

By $R(x)$ we denote the equivalence class of R determined by element x. The indiscernibility relation R describes lack of knowledge about the universe U. Equivalence classes of the relation R, called *granules*, represent an elementary portion of knowledge we are able to perceive due to R. Using only the indiscernibility relation, in general, we are not able to observe individual objects from U but only the accessible granules of knowledge described by this relation.

- The set of all objects which can be with *certainty* classified as members of X with respect to R is called the *R-lower approximation* of a set X with respect to R, and denoted by

$$\underline{R}(X) = \{x : R(x) \subseteq X\}$$

- The set of all objects which can be only classified as *possible* members of X with respect to R is called the *R-upper approximation* of a set X with respect to R, and denoted by

$$\overline{R}(X) = \{x : R(x) \cap X \neq \emptyset\}$$

- The set of all objects which can be decisively classified neither as members of X nor as members of - *is X with respect to R* called the *boundary region* of a set X with respect to R, and denoted by

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

Now we are ready to formulate the definition of the rough set notion.

A set X is called *crisp* (*exact*) with respect to R if and only if the boundary region of X *is* empty**.**

A set X is called rough (inexact) with respect to R if and only if the boundary region of X is nonempty.

The definitions of set approximations presented above can be expressed in terms of granules of knowledge in the following way. The lower approximation of a set is union of all granules which are entirely included in the set; the upper approximation − is union of all granules which have non-empty intersection with the set; the boundary region of a set is the difference between the upper and the lower approximation of the set.

Rough Sets In Data Analysis

Information Systems

A data set is represented as a table, where each row represents a case, an event, a patient, or simply an object. Every column represents an attribute (a variable, an observation, a property, etc.) that can be measured for each object; the attribute may be also supplied by a human expert or the user. Such table is called an information system. 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 non-empty finite set of attributes such that $a : U \to V_a$ for every $a \in A$. The set V_a is called the value set of a.

Let us consider a very simple information system shown in Table 1. The set of objects *U* consists of seven objects: x_1, x_2, x_3, x_4, x_5, x_6, x_7 and the set of attributes includes two attributes: *Age* and *LEMS*.

Table 1.1 Simple information systems

Objects	Age	LEMS
x_1	16-30	50
x_2	16-30	0
x_3	31-45	1-25
x_4	31-45	1-25
x_5	46-60	26-49
x_6	16-30	26-49
x_7	46-60	26-49

Consider a decision system presented in Table 2. The table includes the same seven objects as in Example 2.1 and one decision attribute (*Walk*) with two values: Yes, No.

Table 1.2 Simple information systems with decision

Objects	Age	LEMS	Walk
x_1	16-30	50	Yes
x_2	16-30	0	No
x_3	31-45	1-25	No
x_4	31-45	1-25	Yes
x_5	46-60	26-49	No
x_6	16-30	26-49	Yes
x_7	46-60	26-49	No

Indiscernibility Relation

A decision system expresses all the knowledge about the model. This table may be unnecessarily large in part because it is redundant in at least two ways. The same or indiscernible objects may be represented several times, or some of the attributes may be superfluous. We shall look is called the *B*-indiscernibility relation. It is easy to see that *IND* (*B*) *S* is equivalence relation.

If (*x x*) *IND* (*B*) *S* , ' ∈ , then objects *x* and *x*' are indiscernible from each other by attributes from *B.* The equivalence classes of the B-indiscernibility relation are denoted [ ]*B x* . The subscript *S* in the indiscernibility relation is usually omitted if it is clear which information system is meant. Some extensions of standard rough sets do not require from a relation to be transitive (see, for instance, [8]). Such a relation is called tolerance relation or similarity into these issues now. Let $S = (U, A)$ be an information system, and $B \subseteq A$. A binary relation *IND* (*B*) *S* defined in the following way.
$IND (\{Age \}) =\{\{ x_1 , x_2 , x_6 \} \{ x_3 , x_4 \} \{ x_5 , x_7 \}\}$
$IND (\{LEMS \})= \{\{ x_1 \} \{ x_2 \} \{ x_3 , x_4 \} \{ x_5 , x_6 , x_7 \}\}$
$IND (\{Age, LEMS \}) \{\{ x_1\} \{ x_2 \} \{ x_3 , x_4 \} \{ x_5 , x_7 \} \{ x_6 \}\}$

Set Approximation

In this subsection, we define formally the approximations of a set using the discernibility relation.
Let $S = (U, A)$ be an information system and let $B \subseteq A$, and $X \subseteq U$.
Now, we can approximate a set *X* using only the information contained in the set of attributes *B* by constructing the B-lower and B-upper approximations *of X*, denoted and respectively, where $=\{x| [x]_B \subseteq X\}$ and $= \{x |[x] \cap X \neq \varphi \}$.
Let $X = \{x: Walk (x) = Yes\}$, as given by Table 2. In fact, the set *X* consists of three objects*:* x_1 , x_4 , x_6 . Now, we want to describe this set in terms of the set of conditional attributes $A =\{Age, LEMS\}$. Using the above definitions, we obtain the following approximations:
The *A*-lower approximation $AX =\{x_1 , x_6 \}$
The *A*-upper approximation $AX =\{ x_1 , x_3 , x_4 , x_6 \}$

II. LITERATURE SURVEY

In 2011 Yiyu Yao_ and Xiaofei Deng proposed **"Sequential Three-way Decisions with Probabilistic Rough Sets"**. When approximating a concept, probabilistic rough set models use probabilistic positive, boundary and negative regions. Rules obtained from the three regions are recently interpreted as making three-way decisions, consisting of acceptance, deferment, and rejection. A particular decision is made by minimizing the cost of correct and incorrect classifications. This framework is further extended into sequential three-way decision making, in which the cost of obtaining required evidence or information is also considered. They generalize three-way decisions with probabilistic rough sets into a sequential three-way framework[1].

In 2012 M. E. Abd El-Monsef et al **"A Comprehensive Study of Rough Sets and Rough Fuzzy Sets on Two Universes".** They presents a framework for the study of rough sets and rough fuzzy sets on two universes of discourse. By means of a binary relation between two universes of discourse, a class of revised rough sets and revised rough fuzzy sets based on two universes has been proposed. Some properties of the new model are revealed. The proposed model will be more natural in the sense that rough sets approximated by sets on the same universe. results, examples and counter examples are provided. They presented a new definition of the lower approximation and upper approximation on two universes through the use of the intersection of right neighborhoods[2].

In 2013 Faziye Yüksel et al **"Soft Covering Based Rough Sets and Their Application"** Soft rough sets which are a hybrid model combining rough sets with soft sets are defined by using soft rough approximation operators. Soft rough sets can be seen as a generalized rough set model based on soft sets. They combined the covering soft set with rough set, which gives rise to the new kind of soft rough sets. They showed that a new type of the soft covering upper approximation operator is smaller than soft upper approximation operator. They also presented an example in medicine which aims to find the patients with high prostate cancer risk[3].

In 2014 Y.H.Qian et al Proposed **"An efficient accelerator for attribute reduction from incomplete data in rough set framework" .**They used attribute reduction from large-scale incomplete data is a challenging problem in areas such as pattern recognition, machine learning and data mining. In rough set theory, feature selection from incomplete data aims to retain the discriminatory power of original features. To address this issue, many feature selection algorithms have been proposed, however, these algorithms are often computationally time-consuming. To overcome this shortcoming, they introduced a theoretic framework based on rough set theory, which is called positive approximation and can be used to accelerate a heuristic process for feature selection from incomplete data[4].

In 2015 M. Pushpalatha et al proposed **"A Survey: Rough Set Theory in Incomplete Information Systems"**. Rough Set theory has been conceived as a tool to conceptualize, organize and analyze various types of data, in particular, to deal with inexact, uncertain or vague knowledge in applications related to Artificial Intelligence. They proposed a surveyed over the several real word datasets, the information collected to represent various decisions along with variables contains ambiguity. This survey r presents an overview of the rough set theory, terms used in the rough sets. Rough sets can be applied to the important process of feature selection and learning [5].

In 2016 Qinghua Zhang et al proposed **"A survey on rough set theory and its applications".** After probability theory, fuzzy set theory and evidence theory, rough set theory is a new mathematical tool for dealing with vague, imprecise, inconsistent and uncertain knowledge. They discussed the basic concepts, operations and characteristics on the rough set theory are introduced firstly, and then the extensions of rough set model, the situation of their applications, some application software and the key problems in applied research for the rough set theory are presented. The rough set theory has been researched for more than thirty years[6].

In 2017 B.S.Panda et al proposed **"Retrieving the Missing Information from Information Systems Using Rough Set, Covering Based Rough Set and Soft Set"**. They proposed a study the various aspects of rough set theory, covering-based rough set and soft set theory to handle the missing information systems. They summarized the basic concepts of rough set, covering based set and soft set the manner in which rough set are related to covering based set and soft set. They presented a detailed theoretical study of soft set, which led to the definition of missing data handling. If the mapping set of an attribute includes incomplete data, we filled the data according to the value in the corresponding attributes. This work focused on imputes the missing values through the above techniques[7].

In 2018 Andrzej Skowron et al proposed **"Rough sets: past, present, and future"**. They outline some selected past and present research directions of rough sets. They emphasize the importance of searching strategies for relevant approximation spaces as the basic tools in achieving computational building blocks required for approximation of complex vague concepts. They also discuss new challenges related to problem solving by intelligent systems (IS) or complex adaptive systems (CAS). They have discussed some issues related to the development of rough sets over 35 years, together with some challenges for the rough set approach, especially in the environment where computations are progressing due to interactions between physical and abstract granules, and where they can be controlled by performing actions activated on the basis of satisfiability (to a degree) of complex vague concepts, modeled by approximations [8].

In 2019 Yonca Yazirli et al proposed **" Comparison of Algorithms Based on Rough Set Theory for A 3-Class Classification".** There are various data mining techniques to handle with huge amount of data sets. Rough set based classification provides an opportunity in the efficiency of algorithms when dealing with larger datasets. The selection of eligible attributes by using an efficient rule set offers decision makers save time and cost. They presents the comparison of the performance of the rough set based algorithms: Johnson' s, Genetic Algorithm and Dynamic reducts. In the process of determining the best reduction algorithm based on rough set theory, the classification performance of the test data was taken into consideration and the genetic algorithm was chosen as the most successful reduction algorithm [9].

III. Proposed Approach

The Proposed Model depends on the distance function to detect any missing attributes values. This will be done by calculating the distance function between complete information system table and incomplete information system table. When the small distance is repeated with more than one case and the attribute - which the missing value on it - has a different value, then the method eliminates one of the attributes which has a small effect on the information system by using the degree of dependency. If the Model eliminates the last attribute that has a bigger effect on the system and there is no single value of the smallest distance, the most common attribute value will be supposed to be the missing value.
The distance between the complete decision table and incomplete decision table can be calculated by the following function

$$dis(X_{Iincomp}\,,\ X_{comp}) = \sqrt{\sum_{i=1}^{N}\left[b_i(X_{incomp}) - b_i(X_{comp})\right]^2}$$

where
$\forall X_{incomp}, X_{COMP} \in U$
X_{incomp} is an incomplete case
X_{comp} is an complete case
$b_i \in B$ attribute
$N = ||B||$ number of attributes

Conclusion

Rough set theory is a new method that deals with vagueness and uncertainty emphasized in decision making. Data mining is a discipline that has an important contribution to data analysis, discovery of new meaningful knowledge, and autonomous decision making. The rough set theory offers a viable approach for decision rule extraction from data. Rough sets have been proposed for a very wide variety of applications. In particular, the rough set approach seems to be important for Artificial Intelligence and cognitive sciences, especially in machine learning, knowledge discovery, data mining, expert systems, approximate reasoning. In this paper we proposed the basic concepts of rough et theory vaios term used in rough set theory and application of rough set theory

REFERENCES

1. Yiyu Yao_ and Xiaofei Deng Department of Computer Science, University of Regina Regina, Saskatchewan, Canada S4S 0A2 2011.
2. M. E. Abd El-Monsef, A. M. Kozae, A. S. Salama, R. M. Aqeel Journal Of Computing, Volume 4, Issue 3, March 2012, ISSN 2151-9617.
3. Faziye Yüksel,1 Zehra Güzel Ergül,2 and Naime Tozlu3 "Soft Covering Based Rough Sets and Their Application" Hindawi Publishing Corporation · e Scientific World Journal Volume 2014, Article ID 970893, 9 pages http://dx.doi.org/10.1155/2014/970893.
4. Y. H. Qian, J. Y. Liang, Pattern Recognit., vol. 44, no. 8, pp. 1658–1670, Aug. 2014.
5. M. Pushpalatha, Dr. V. Anuratha International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 8, August 2015.
6. Qinghua Zhang , Qin Xie www.sciencedirect.com Science Direct CAAI Transactions on Intelligence Technology (2016) .
7. B.S.Panda, S.S.Gantayat, Covering Based Rough Set and Soft Set B.S.Panda et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1403-1407.
8. Andrzej Skowron Rough sets: past, present, and future Natural Computing (2018) 17:855–876 Published online: 25 July 2018.
9. Yonca Yazirli , Betül Kan-Kilinç 3-Class Classification [Kilinç et. al., Vol.7 (Iss.8): August 2019] ISSN- 2350-0530(O), ISSN- 2394-3629(P) DOI: 10.5281/zenodo.3401362 International Journal of Research – Granthaalayah