I.J. Mathematical Sciences and Computing, **2017, 2, 22-36**
Published Online April 2017 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2017.02.03

Available online at ***http://www.mecs-press.net/ijmsc***

Handling Numerical Missing Values Via Rough Sets

Elsayed Sallam [a], T. Medhat [b,*], A.Ghanem [c], M. E. Ali [d]

[a] *Computer and Automatic Control Department, Faculty of Engineering, Tanta University, Tanta, Egypt.*
[b] *Electrical Engineering Department, Faculty of Engineering, Kafrelsheikh University, 33516, Kafrelsheikh, Egypt.*
[c] *Portal Manager of Kafrelsheikh University, Kafrelsheikh University, 33516, Kafrelsheikh, Egypt*
[d] *Physics and Engineering Mathematics Department, Faculty of Engineering, Kafrelsheikh University, 33516, Kafrelsheikh, Egypt,*

Abstract

Many existing industrial and research data sets contain missing values. Data sets contain missing values due to various reasons, such as manual data entry procedures, equipment errors, and incorrect measurements. It is usual to find missing data in most of the information sources used. Missing values usually appear as "NULL" values in the database or as empty cells in the spreadsheet table. Multiple ways have been used to deal with the problem of missing data. The proposed model presents rough set theory as a technique to deal with missing data. This model can handle the missing values for condition and decision attributes, the web application was developed to predict these values.

Index Terms: Rough sets, missing values, prediction, and most common value.

1. Introduction

According to Pawlak, Z. [10,11,12] a set is a collection of any objects, which according to some law can be considered as a whole. A lot of existing industrial and research data sets contain missing values. Missing values usually appear as "NULL" values in the database or as empty cells in spreadsheet table.

There are general ways that have been used to deal with the problem of missing data [7,8,13]. The proposed model can predict the missing values of any attributes for the numerical information system table. This model depends on distance functions of all values of attributes between the complete information system table and incomplete information system table. This model depends on rough set to deal with the repeated small distance by eliminating an attribute, which has the smallest effect on the complete information system table. If rough set

* Corresponding author.
E-mail address: sallam@f-eng.tanta.edu.eg, tmedhatm@eng.kfs.edu.eg, eng_ahmedabdou@kfs.edu.eg, manal.ali@eng.kfs.edu.eg

and distance function cannot get the result, the common value method will be used.

The article is organized as follows: In the next section, we review research on information analysis. Related work is presented in section 3. In Section 4, we discuss our proposed model. We present a case study by using the proposed model in section 5. Section 6 presents the URL of web application that used for simulating the model. Comparison between the model and KNN imputation [5,15] is shown in section 7. Conclusion is presented at the last section.

2. Information Analysis

The classical rough set theory developed by Professor Z.Pawlak [6,11,2,3,9,14] in 1982. This theory has made a great success in knowledge acquisition in recent years [1,11]. In Rough set theory, knowledge is represented in information systems. An information system is a data set represented in a table [4]. Information systems with missing data are called incomplete information systems. The main goal of the rough set analysis is the induction of approximations of concepts. Rough set identifies partial or total dependencies in data, eliminates redundant data. It can be used for missing data, data reduction, decision rule generation, and pattern extraction.

2.1. Information System

In rough sets theory, a data set is represented as a table and each row represents a state, an event or simply an object. Each column represents a measurable property for an object (a variable, an observation, etc.). This table is called an information system [1].

An information system can be defined as [13]:

$$IS = (U, A, \rho, V_b) \tag{1}$$

- *U is the universe (a finite set of objects)* $U=\{x_1, x_2, ..., x_m\}$
- *B is the set of attributes (features, variables)*
- V_b *is the set of values a, called the domain of attribute b.*
- $\rho : U \times B \to V_b$

2.2. Indiscernibility Relation

Let $IS = (U, B, \rho, V_b)$ be an information system, then with any, $P \subseteq B$ there is an associated equivalence relation [3]:

$$\mathrm{IND}(P)=\{(x, y) \in U \times U \mid \forall b \in P,\ f_b(x)= f_b(y)\} \tag{2}$$

Where $IND(P)$ is called the P -indiscernibility relation. The partition of U , generated by $IND(P)$ is denoted U/P . If $(x,y) \in IND(P)$, then x and y are indiscernible by attributes from P .

2.3. Approximation

The rough set theory depends on two basics concepts, namely the lower and the upper approximation.

Suppose the given an information system $IS = (U, A, \rho, V_b)$. U called the *universe*. Let X be a subset of U $(X \subseteq U)$. Let P is a subset of B. The basic concepts of the rough set theory [3] will be given below as shown in Fig.1:

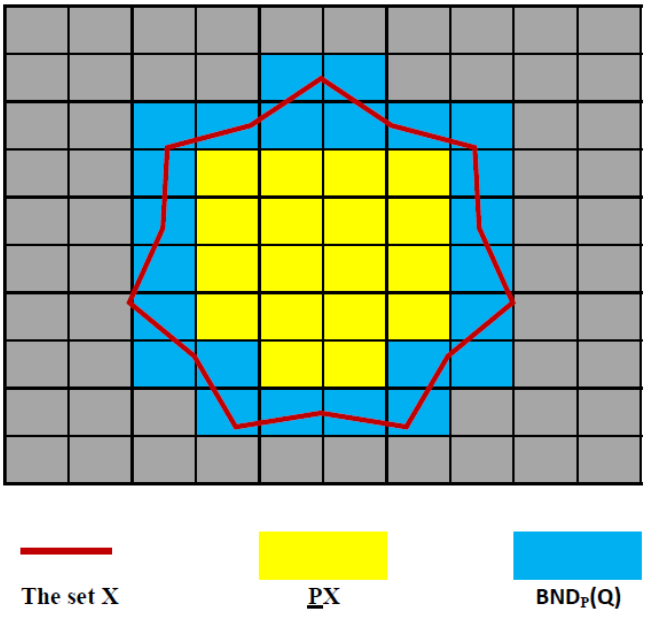

Fig.1. A Rough Set in Rough Approximation Space.

- For a set of attribute P, the lower approximation of a set X is the set of all objects, which can be for certain classified as X, using P:

$$\underline{P}X = \{x \in U \mid [x]_P \subseteq X\} \tag{3}$$

- For a set of attribute P, the upper approximation of a set X is the set of all objects which can be possibly classified as X, using P:

$$\overline{P}X = \{x \in U \mid [x]_P \cap X \neq \phi\} \tag{4}$$

- Let $P, Q \subseteq A$ be equivalence relations over U, then the positive and boundary regions can be defined as follows:

The boundary region is the difference between the upper approximation and lower approximation.

$$BND_P(Q) = \bigcup_{X \in U/Q} \overline{P}X - \bigcup_{X \in U/Q} \underline{P}X \tag{5}$$

The positive region of X, using B is:

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}X \tag{6}$$

- Degree of dependency:

Let $C, D \subseteq A$. D depends on C in a degree k $(0 \le k \le 1)$ denoted $C \Rightarrow_k D$ The positive region of the partition U/D with respect to C , $POS_c(D)$, is the set of all objects of U that can be certainly classified to blocks of the partition U/D by means of C [13].

$$k = \gamma_c(D) = \frac{|POS_c(D)|}{|U|} \tag{7}$$

If $k = 1$, D depends totally on C , if $0 < k < 1$, D depends partially on C , and if $k = 0$ then D does not depend on C . When C is a set of condition attributes and D is the decision, $\gamma_C(D)$ is the quality of classification [5].

2.4. The Most Common Value of an Attribute

In this method, one of the simplest methods to handle missing attribute values, such values are replaced by the most common value of the attribute. In different words, a missing attribute value is replaced by the most probable known attribute value, where such probabilities are represented by relative frequencies of the corresponding attribute.

2.5. Distance Function:

The distance between the complete decision table and incomplete decision table can be calculated by the following function [13]:

$$dis(X_{\text{incomp}}, X_{\text{comp}}) = \sqrt{\sum_{i=1}^{N} [b_i(X_{\text{incomp}}) - b_i(X_{\text{comp}})]^2} \tag{8}$$

$\forall X_{\text{incomp}}, X_{\text{comp}} \in U$

X_{incomp} is an incomplete case

X_{comp} is a complete case

$b_i \in B$; attributes

$N = \|B\|$; number of attributes

Where $b_i(X_{\text{incomp}})$ is the value of attribute b with respect to the case X_{incomp} .

In this method, one of the simplest methods to handle missing attribute values, such values are replaced by the most common value of the attribute. In different words, a missing attribute value is replaced by the most probable known attribute value, where such probabilities are represented by relative frequencies of the corresponding attribute.

3. Related Work

We find in paper [13] that the author can predict the missing values for the decision attribute values only, and can't predict some missing values, but in our proposed model, we can predict missing values for any attributes (condition or decision attributes) and can predict all missing values.

4. The Proposed Model

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.

4.1. Proposed Algorithm:

The steps that algorithm follows to predict the missing value are shown in Figure 2:

1. Separation the decision table to two tables(complete information system table and incomplete information system table)
2. Getting the most common value of each attribute.
3. calculation of Degree of dependency:
    - The model calculates Indiscernibility relation for the complete attribute.
    - The model calculates Indiscernibility relations for the complete attribute except for each attribute individual.
    - The model calculates the POS`s of the complete attribute except for each attribute individual.
    - The model calculates the degree of dependency by dividing each $\|POS\|$ by the $\|U\|$.The degree of dependency is between 0 and 1
    - The model eliminates the attribute b if $k_{B-\{b\}}$ is the biggest.
4. Calculations of the distance function between every case in incomplete information system table and complete information system table.
5. Getting the smallest distance for every case in the incomplete information system table.
6. If the smallest distance unique, then the missing attribute value equal the value of the same attribute which its case has the smallest distance.
7. If the smallest distance is repeated and its records of complete information system table have the same value of the missing attribute, then the missing attribute value equal this repeated attribute value.
8. If the smallest distance is repeated in more records in the complete information system table and the records have different values of the missing attribute, then:

a) The attribute which has small effects on the information system table will be done by using the degree of dependency.
b) Calculation of the distance is repeated again and the above sequence will be repeated but with only the cases of complete information system table which have the smallest distance.

9. If there is no matching case, the algorithm supposes that the missing attribute value is the most common attribute value of this attribute.

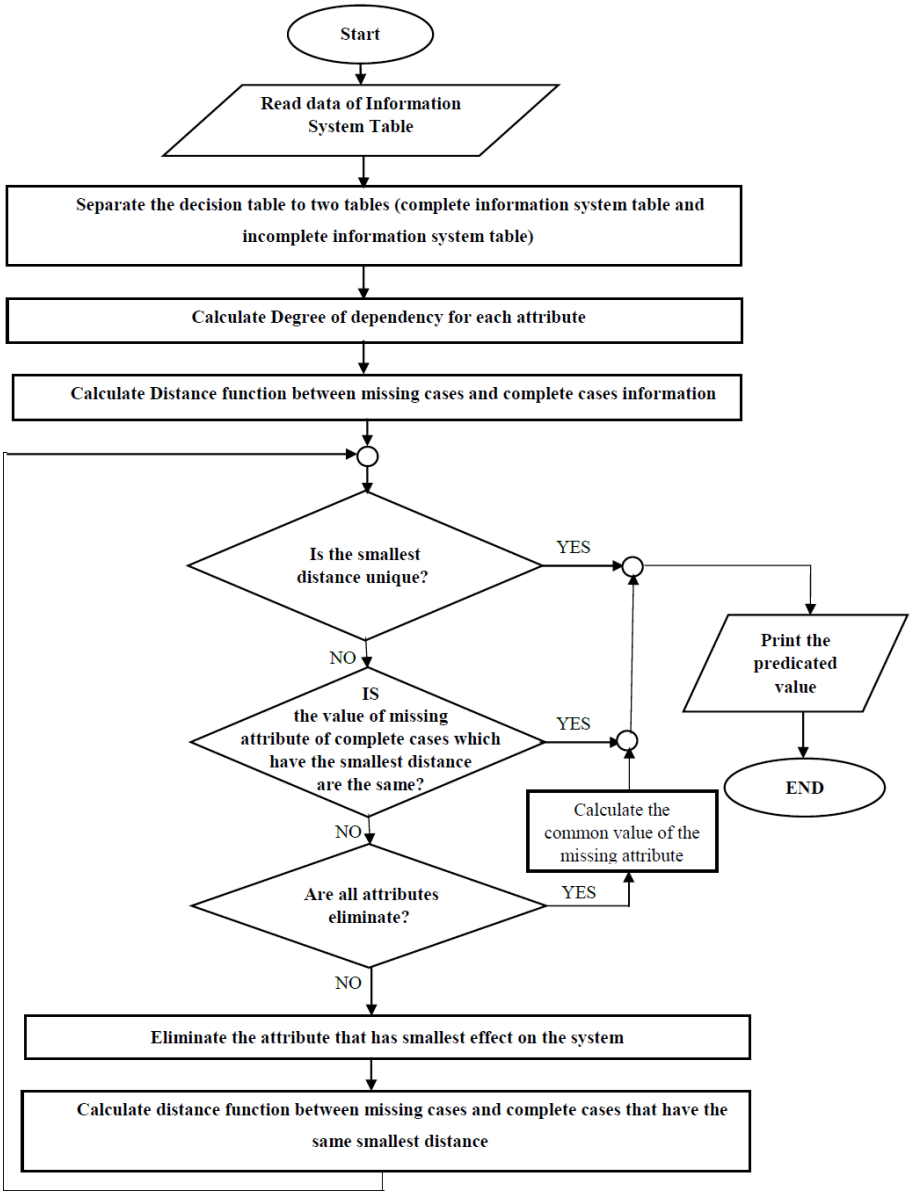

Fig.2. Flowchart of the Proposed Algorithm.

5. Case Study

This case study presents a method in which natural radionuclide concentrations of beach sand minerals are traced along a stretch of coast by rough sets analysis. This analysis yields a classification of groups of mineral deposit with different origins. From the following Table 1 which indicates the activity concentrations in Bq/Kg for mineral fractions in samples collected along the coast. We find that: attributes are B= {U, Ra, K, Ph, Tom}, and the locations are:

U = { c_1,c_2,c_3,c_4 , c_5 , c_6 , c_7 , c_8 , c_9 , c_{10} , c_{11} , c_{12} , c_{13}, c_{14} , c_{15} , c_{16} , c_{17} , c_{18} , c_{19} , c_{20} , c_{21} , c_{22} , c_{23} , c_{24} , c_{25} , c_{26} , c_{27} , c_{29} , c_{30} , c_{31} , c_{32} , c_{33} , c_{34} ,c_{35} , c_{36} , c_{37} , c_{38} , c_{39} , c_{40} , c_{41} , c_{42}}

The values of information table are shown in Table 1.
The supposed table (After deleting repeated cases) will be:

U = { c_4 , c_5 , c_6 , c_7 , c_8 , c_{14} , c_{19} , c_{20} , c_{22} , c_{23} , c_{24} , c_{25} , c_{26} , c_{27} , c_{29} , c_{35} , c_{36} , c_{38} , c_{39} , c_{41} , c_{42}}
as shown in Table 2.

5.1. Dividing that table into two tables

In this step, the algorithm divides the information system table in two tables. The first table is complete Information system table as shown in Table 3. The second table is incomplete Information system table which has the cases that have missing values as shown in Table 4.

5.2. Calculating the degree of dependency:

The following steps will calculate the degree of dependency:

5.2.1. Indiscernibility relation:

Calculate Indiscernibility relation for the complete attribute and the complete attribute except each attribute individual.

U/IND(B) = { {c_4}, {c_5}, {c_7}, {c_8}, {c_{19}}, {c_{20}}, {c_{23}}, {c_{24}}, {c_{25}}, {c_{26}}, {c_{27}}, {c_{29}}, {c_{35}}, {c_{38}}, {c_{39}}, {c_{41}} }
U/IND(B-{u}) = { {c_4}, {c_5}, {c_7}, {c_8}, {c_{19}}, {c_{20}}, {c_{23}}, {c_{24}}, {c_{25}}, {c_{26}}, {c_{27}}, {c_{29}}, {c_{35}}, {c_{38}}, {c_{39}}, {c_{41}} }
U/IND(B-{Ra}) = { {c_4}, {c_5}, {c_7}, {c_8}, {c_{19}}, {c_{20}}, {c_{23}}, {c_{24}}, {c_{25}}, {c_{26}}, {c_{27}}, {c_{29}},{c_{35}}, {c_{38}}, {c_{39}}, {c_{41}} }
U/IND(B-{K}) = {{c_{20},c_{26}}, {c_{23},c_{41}}, {c_{25},c_{38}}, {c_{27}}, {c_{24},c_{35},c_{39}}, {c_{29}}, {c_{19}}, {c_8}, {c_7}, {c_5}, {c_4}}
U/IND(B-{Ph}) = { {c_{20},c_{24}}, {c_{23}}, {c_{41}}, {c_{26},c_{29},c_{35}}, {c_{25}}, {c_{27}}, {c_{39}}, {c_{38}}, {c_{19}}, {c_8}, {c_7}, {c_4,c_5} }
U/IND(B-{Tom}) = { {c_4}, {c_5}, {c_7}, {c_8}, {c_{19}}, {c_{20}}, {c_{23}}, {c_{24}}, {c_{25},c_{27}}, {c_{26}}, {c_{29}}, {c_{35}}, {c_{38}}, {c_{39}}, {c_{41}} }

Table 1. Information System Table

U/B	U	Ra	K	Ph	Tom
C_1	1	1	1	2	1
C_2	1	1	1	1	1
C_3	2	1	1	2	1
C_4	4	4	1	2	1
C_5	4	4	1	1	1
C_6	2	2	1	2	1
C_7	3	3	1	2	1
C_8	2	2	1	3	1
C_9	1	1	1	2	1
C_{10}	1	1	1	2	1
C_{11}	1	1	1	2	1
C_{12}	1	1	1	2	1
C_{13}	1	1	1	2	1
C_{14}	1	1	2	2	1
C_{15}	1	1	1	1	1
C_{16}	1	1	1	2	1
C_{17}	1	1	1	2	1
C_{18}	1	1	1	2	1
C_{19}	2	1	1	2	1
C_{20}	1	1	1	1	1
C_{21}	1	1	1	2	1
C_{22}	1	1	1	2	1
C_{23}	1	1	1	2	2
C_{24}	1	1	1	3	1
C_{25}	1	1	3	2	3
C_{26}	1	1	3	1	1
C_{27}	1	1	3	2	4
C_{28}	1	1	3	2	1
C_{29}	1	1	3	4	1
C_{30}	1	1	3	3	1
C_{31}	1	1	4	3	1
C_{32}	1	1	4	2	1
C_{33}	1	1	4	3	1
C_{34}	1	1	3	2	1
C_{35}	1	1	3	3	1
C_{36}	1	1	3	2	1
C_{37}	1	1	4	2	1
C_{38}	1	1	4	2	3
C_{39}	1	1	4	3	1
C_{40}	1	1	4	2	1
C_{41}	1	1	2	2	2
C_{42}	1	1	4	2	1

Table 2. Information System Table after Deleting Repeated Cases and Supposing Missing Values As"?".

U/B	U	Ra	K	Ph	Tom
C_4	4	4	1	2	1
C_5	4	4	1	1	1
C_6	2	2	1	2	?
C_7	3	3	1	2	1
C_8	2	2	1	3	1
C_{14}	1	1	2	2	?
C_{19}	2	1	1	2	1
C_{20}	1	1	1	1	1
C_{22}	1	1	?	2	1
C_{23}	1	1	1	2	2
C_{24}	1	1	1	3	1
C_{25}	1	1	3	2	3
C_{26}	1	1	3	1	1
C_{27}	1	1	3	2	4
C_{29}	1	1	3	4	1
C_{35}	1	1	3	3	1
C_{36}	?	1	3	2	1
C_{38}	1	1	4	2	3
C_{39}	1	1	4	3	1
C_{41}	1	1	2	2	2
C_{42}	1	?	4	2	1

Table 3. Complete Information System Table

U/B	U	Ra	K	Ph	Tom
C_4	4	4	1	2	1
C_5	4	4	1	1	1
C_7	3	3	1	2	1
C_8	2	2	1	3	1
C_{19}	2	1	1	2	1
C_{20}	1	1	1	1	1
C_{23}	1	1	1	2	2
C_{24}	1	1	1	3	1
C_{25}	1	1	3	2	3
C_{26}	1	1	3	1	1
C_{27}	1	1	3	2	4
C_{29}	1	1	3	4	1
C_{35}	1	1	3	3	1
C_{38}	1	1	4	2	3
C_{39}	1	1	4	3	1
C_{41}	1	1	2	2	2

Table 4. Incomplete Information System Table

U/K	U	Ra	K	Ph	Tom
c_6	2	2	1	2	?
c_{14}	1	?	2	2	1
c_{22}	1	1	?	2	1
c_{36}	1	1	3	2	?
c_{42}	1	?	4	2	1

5.2.2. Calculating the POS`s:

$POS_{B-\{U\}}(B) = |\{c_4\}, \{c_5\}, \{c_7\}, \{c_8\}, \{c_{19}\}, \{c_{20}\}, \{c_{23}\}, \{c_{24}\}, \{c_{25}\}, \{c_{26}\}, \{c_{27}\}, \{c_{29}\}, \{c_{35}\}, \{c_{38}\}, \{c_{39}\}, \{c_{41}\}| = 16$

$POS_{B-\{Ra\}}(B) = |\{c_4\}, \{c_5\}, \{c_7\}, \{c_8\}, \{c_{19}\}, \{c_{20}\}, \{c_{23}\}, \{c_{24}\}, \{c_{25}\}, \{c_{26}\}, \{c_{27}\}, \{c_{29}\}, \{c_{35}\}, \{c_{38}\}, \{c_{39}\}, \{c_{41}\}| = 16$

$POS_{B-\{K\}}(B) = |\{c_{27}\}, \{c_{29}\}, \{c_{19}\}, \{c_8\}, \{c_7\}, \{c_5\}, \{c_4\}| = 7$

$POS_{B-\{Ph\}}(B) = |\{c_{23}\}, \{c_{41}\}, \{c_{25}\}, \{c_{27}\}, \{c_{39}\}, \{c_{38}\}, \{c_{19}\}, \{c_8\}, \{c_7\}| = 9$

$POS_{B-\{Tom\}}(B) = |\{c_4\}, \{c_5\}, \{c_7\}, \{c_8\}, \{c_{19}\}, \{c_{20}\}, \{c_{23}\}, \{c_{24}\}, \{c_{26}\}, \{c_{29}\}, \{c_{35}\}, \{c_{38}\}, \{c_{39}\}, \{c_{41}\}| = 14$

5.2.3. Degree of dependency calculation:

$K_{B-\{U\}}(B) = \Upsilon_{B-\{U\}}(B) = POS_{B-\{U\}}(B)/U = 16/16 = 1$

$K_{B-\{Ra\}}(B) = \Upsilon_{B-\{Ra\}}(B) = POS_{B-\{Ra\}}(B)/U = 16/16 = 1$

$K_{B-\{K\}}(B) = \Upsilon_{B-\{K\}}(B) = POS_{B-\{B\}}(C)/U = 7/16 = 0.4375$

$K_{B-\{Ph\}}(B) = \Upsilon_{B-\{Ph\}}(B) = POS_{B-\{Ph\}}(B)/U = 9/16 = 0.5625$

$K_{B-\{Tom\}} = \Upsilon_{B-\{Tom\}}(B) = POS_{B-\{Tom\}}(B)/U = 14/16 = 0.875$

5.3. Getting the most common attribute values:

The most common values of each attribute will be displayed in Table 5:

Table 5. The Most Common Attribute Value

Attributes	No. of Repeated attribute value				Most common attribute value
	(1)	(2)	(3)	(4)	
U	15	3	1	2	1
Ra	16	2	1	2	1
K	10	2	6	3	1
Ph	3	13	4	1	2
Tom	16	2	2	1	1

5.4. Prediction of Missing Values:

In this part the model will predict the values of the missing values.

5.4.1. Prediction of Missing Value: case 22

The distance function between case 22 and complete information system table will be calculated in Table 6.

Table 6. Distance Function between case 22 and Other Complete Cases

U/B	U	Ra	K	Ph	Tom	Distance function
C_{22}	1	1	?	2	1	
C_{4}	4	4	1	2	1	4.242641
C_{5}	4	4	1	1	1	4.358899
C_{7}	3	3	1	2	1	2.828427
C_{8}	2	2	1	3	1	1.732051
C_{19}	2	1	1	2	1	1
C_{20}	1	1	1	1	1	1
C_{23}	1	1	1	2	2	1
C_{24}	1	1	1	3	1	1
C_{25}	1	1	3	2	3	2
C_{26}	1	1	3	1	1	1
C_{27}	1	1	3	2	4	3
C_{29}	1	1	3	4	1	2
C_{35}	1	1	3	3	1	1
C_{38}	1	1	4	2	3	2
C_{39}	1	1	4	3	1	1
C_{41}	1	1	2	2	2	1

The smallest distance function is "1" and there are eight cases (case 19, case 20, case 23, case 24, case 26, case 29, case 35, case 39, case 41) that have this distance function but they have different values for the attribute K so we need to eliminate attribute which has small effects on the information system table by using the degree of dependency. Reference to the degree of dependency k value, the attribute U will be deleted and then only the cases that have the similar small distance (case 19, case 20, case 23, case 24, case 26, case 29, case 35, case 39, case 41) will be compared with the missing case (see Table 7):

Table 7. Distance Function between case 22 and Other Complete Cases after Deleting Attribute U

U/B	Ra	K	Ph	Tom	Distance function
C_{22}	1	?	2	1	
C_{19}	1	1	2	1	0
C_{20}	1	1	1	1	1
C_{23}	1	1	2	2	1
C_{24}	1	1	3	1	1
C_{26}	1	3	1	1	1
C_{35}	1	3	3	1	1
C_{39}	1	4	3	1	1
C_{41}	1	2	2	2	1

After eliminating the attribute Ra the smallest distance function is "0" and only the (case 19) has this distance function, so the value of the attribute (K) for the (case 22) equal the value of the attribute (K) for the (case 19) equal " 1".

5.4.2. Prediction of Missing Values of (case 36):

The distance function between (case 36) and complete information system table will be calculated in Table 8.

Table 8. Complete Cases Compared with case 36

U/B	U	Ra	K	Ph	Tom	Distance function
C_{36}	1	1	3	2	?	
C_4	4	4	1	2	1	4.690416
C_5	4	4	1	1	1	4.795832
C_7	3	3	1	2	1	3.464102
C_8	2	2	1	3	1	2.645751
C_{19}	2	1	1	2	1	2.236068
C_{20}	1	1	1	1	1	2.236068
C_{23}	1	1	1	2	2	2
C_{24}	1	1	1	3	1	2.236068
C_{25}	1	1	3	2	3	0
C_{26}	1	1	3	1	1	1
C_{27}	1	1	3	2	4	0
C_{29}	1	1	3	4	1	2
C_{35}	1	1	3	3	1	1
C_{38}	1	1	4	2	3	1
C_{39}	1	1	4	3	1	1.414214
C_{41}	1	1	2	2	2	1

The smallest distance function is "0" and there are two cases (case 25 and case 27) that have this distance function. The two cases have the different values for the attribute Tom, so we need to eliminate the attribute that has the small effect on the system (U) as shown in Table 9.

Table 9. Cases with the Similar Small Distance Compared with case 36 after Eliminating Attribute U

U/B	Ra	K	Ph	Tom	Distance function
C_{36}	1	3	2	?	
C_{25}	1	3	2	3	0
C_{27}	1	3	2	4	0

The smallest distance function is "0" and there are two cases (case 25 and case 27) that have this distance function.The two cases have the different values for the attribute Tom, so we need to eliminate the next attribute that has the small effect on the system (Ra) as shown in Table 10.

Table 10. Cases with the Similar Small Distance Compared with case 36 after Eliminating Attribute Ra

U/B	U	K	Ph	Tom	Distance function
C_{36}	1	3	2	?	
C_{25}	1	3	2	3	0
C_{27}	1	3	2	4	0

The smallest distance function is "0" and there are two cases (case 25 and case 27) that have this distance function.The two cases have the different values for the attribute Tom, so we need to eliminate the next

attribute that has the small effect on the system (Ph) as shown in Table 11.

Table 11. Cases with the Similar Small Distance Compared with case 36 after Eliminating Attribute Ph

U/B	U	Ra	K	Tom	Distance function
C_{36}	1	1	3	?	
C_{25}	1	1	3	3	0
C_{27}	1	1	3	4	0

The smallest distance function is "0" and there are two cases (case 25, case 27) that have this distance function.The two cases have the different values for the attribute Tom, so we need to eliminate the next attribute that has the small effect on the system (K) as shown in Table 12.

Table 12. Cases with the Similar Small Distance Compared with case 36 after Eliminating Attribute K

U/B	U	Ra	Ph	Tom	Distance function
C_{36}	1	1	2	?	
C_{25}	1	1	2	3	0
C_{27}	1	1	2	4	0

The most common value will be used to get the missing value. So the missing value of this case will equal the most common attribute value of (Tom) so it will be equal "1".

6. MVP

A web application MVP (Missing Value Prediction) has been developed to do the proposed algorithm using c#, asp.net, and SQL server as shown in Fig.3 which be published on the internet with this URL: www.kfs.edu.eg/mvp

Fig.3. Missing Value Prediction (MVP)

7. Comparison

To test the algorithm 100 tables are made with random 5 missing cases for each table from Table 1. The random tables are uploaded on the next URL:

http://www.kfs.edu.eg/pdf/tables.pdf

The proposed method predicts 60.4% of the missing attributes values with accuracy 100%. The proposed method has the following advantages by comparing with The KNN Imputation [5] as shown in Table 13.

Table 13. Comparing the Proposed Method with KNN Imputation

Methods	No. of true predicted values	Used Time
KNN Imputation	38%	0.36 SEC
The proposed method	60.4%	7.80 SEC

8. Conclusion

In this model, missing values can be predicted by calculating the distance function. If repeated distance function presents with different attribute values, the proposed model eliminates an attribute which has the smallest effect on the complete information system and repeats the calculation of the distance function again. If there is no matching case, the proposed model will suppose the missing value equal the most common attribute value of the missing attribute. The web application was developed to predict the missing values.

References

[1] Ahmed Tariq Sadiq, Mehdi Gzar Duaimi, Samir Adil Shaker" Data Missing Solution Using Rough Set Theory and Swarm Intelligence"International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2(3), pp.1-16, 2012.

[2] Boby P. Mathew, Sunil Jacob John,"ON ROUGH TOPOLOGICAL SPACES", International Journal of Mathematical Archive, Vol.3(9), pp.3413-3421,2012.

[3] H. Nasiri, M. Mashinchi," Rough Set and Data Analysis in Decision Tables", Journal of Uncertain Systems, Vol.3(3), pp.232-240, 2008.

[4] Hala S. Own, Aboul Ella Hassanien, "Rough Wavelet Hybrid Image Classification Scheme", JCIT, Vol. 3(4), pp.65-75, 2008.

[5] Hulse Jason Van, Khoshgoftaar Taghi M, "Incomplete-Case Nearest Neighbor Imputation In Software Measurement", Vol. 259, pp. 596–610, 2014.

[6] Mert Bal," Rough Sets Theory as Symbolic Data Mining Method: An Application on Complete Decision Table ", Information Science Letters, Vol. 2(1), pp.35-47, 2012.

[7] Michinori Nakata, Hiroshi Sakai, "Applying Rough Sets to Data Tables Containing Missing Values", Lecture Notes in Computer Science, Vol. 4585, pp. 181-191, 2007

[8] M.K. Sabu, G. Raju "Rough Set Approaches for Mining Incomplete Information Systems", Vol. 5227, pp.914-921,2008

[9] N. Senthilkumaran, R. Rajesh"A Study on Rough Set Theory for Medical Image Segmentation", International Journal of Recent Trends in Engineering, Vol.2(2),pp.236-238,2009.

[10] Pawlak Z., "Rough Sets", International Journal of Information and Computer Sciences, Vol.11(5), pp.341-356, 1982

[11] Pawlak Z, "Rough set approach to multi-attribute decision analysis", European Journal of Operational Research, Vol. 72(3), pp. 443-459, 1994.

[12] Pawlak Z., "Rough Sets and Intelligent Data Analysis", Information Sciences, Vol.147, pp. 1–12, 2002.

[13] T. Medhat, "Prediction of missing values for decision attribute", International Journal of Information Technology and Computer Science, Vol. 4(11), pp.58-66, 2013.

[14] Roman W.WINIARSKI,"ROUGH SETS METHODS IN FEATURE REDUCTION AND CLASSIFICATION", Vol.11(3), pp.565-582, 2001.

[15] G.Vamsi Krishna,"Prediction of Rainfall Using Unsupervised Model based Approach Using K-Means Algorithm", International Journal of Mathematical Sciences and Computing (IJMSC), Vol.1(1), pp.11-20, 2015.

Authors' Profiles

Elsayed Sallam, born in Egypt, received his M.Sc. and Ph.D degree in Faculty of Engineering, Bremen University, W. Germany in 1987. He is currently an Emeritus professor in the Faculty of Engineering, Tanta University, Egypt. His current research interests include Distributed Systems, fuzzy, networks, robotics and clustering. He is an IEEE member.

Tamer Medhat, born in Kafrelsheikh, Egypt, in July 13th, 1974, and received his Ph.D degree in Faculty of Engineering, Tanta University, Tanta, Egypt in 2007. He is currently a lecturer in the Faculty of Engineering, Kafrelsheikh University, Egypt. His current research interests include Information Systems, Augmented Reality, Decision Making, Computer Science, and Rough Set Theory Applications.

Ahmed Mohamed Abdou, born in Kafrelsheikh, Egypt, in january 18th, 1985, and graduarted from Faculty of Engineering, Tanta University, Tanta, Egypt in 2006, and received his M.Sc. in Faculty of Engineering, Tanta University in 2016. He is currently a Kafrelsheikh university portal manger in Kafrelsheikh University, Egypt. His current research interests include: information systems, computer science, and rough set theory applications.

Manal El-Said Ali, born in Kafrelsheikh, Egypt, and received her Ph.D degree in Faculty of Engineering, Tanta University, Tanta, Egypt in 2011. She is currently a lecturer in the Faculty of Engineering, Kafrelsheikh University, Egypt. Her current research interests include engineering mathematics, information systems, decision-making, granular computing, fuzzy sets, and rough set theory applications.

How to cite this paper: Elsayed Sallam, T. Medhat, A.Ghanem, M. E. Ali,"Handling Numerical Missing Values Via Rough Sets", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.3, No.2, pp. 22-36, 2017.DOI: 10.5815/ijmsc.2017.02.03