Rough Sets Computations to Impute Missing Data

Fulufhelo Vincent Nelwamondo and Tshilidzi Marwala

School of Electrical and Information Engineering,

University of the Witwatersrand,

Private Bag 3, Wits, 2050, South Africa

February 1, 2008

Abstract

Many techniques for handling missing data have been proposed in the literature. Most of these techniques are overly complex. This paper explores an imputation technique based on rough set computations. In this paper, characteristic relations are introduced to describe incompletely specified decision tables.It is shown that the basic rough set idea of lower and upper approximations for incompletely specified decision tables may be defined in a variety of different ways. Empirical results obtained using real data are given and they provide a valuable and promising insight to the problem of missing data. Missing data were predicted with an accuracy of up to 99%.

Key words: Indiscernibility, membership, missing data, rough sets, set approximation

1 Introduction

There are three general ways that have been used to deal with the problem of missing data (Little and Rubin, 1987). The simplest method is known as 'listwise deletion' which, simply deletes instances with missing values. The major disadvantage of this method is the dramatic loss of information in data sets. Kim and Curry (1997) found that when 2% of the features are missing and the complete observation is deleted, up to 18 percent of the total data may be lost. The second common technique imputes the data by finding estimate of the values and missing entries are replaced with these estimates. Various estimates have been used and these estimates include zeros, means and other statistical calculations. These estimations are then used as if they were the observed values. Another common technique assumes some models for the prediction of the missing values and uses the maximum likelihood approach to estimate the missing values.

A graet deal of research has been conducted to find new ways of approximating the missing values. Among others, Abdella and Marwala (2006) and Mohamed and Marwala (2005) have used neural networks together with Genetic Algorithms (GA) to approximate missing data. Qiao et al. (2005) have used neural networks and Particle Swarm Optimization (PSO) to keep track of the dynamics of a power plant in the presence of missing data. Nauck and Kruse (1999) and Gabrys (2002) have used Neuro fuzzy for learning in the presence of missing data. A different approach was taken by Wang (2005) who replaced incomplete patterns with fuzzy patterns. The patterns without missing values are, along with fuzzy patterns, used to train the neural network. In his model, the neural network learns to classify without actually predicting the missing data. Special attention in the literature has been given to imputation techniques such as the Expectation maximisation as well as the use of neural networks, coupled with an optimisation technique such as genetic algorithms. The use of neural networks comes with a greater cost in terms of computation and in that data has to be made available before the missing condition occurs. This paper proposes a new algorithm based on rough set theory for missing data estimation. Although other simmillar methods have been mentioned in the literature (Nakata and Sakai, 2006; Grzymala-Busse, 2004), this paper also applies a rough set technique for missing data imputation to a large and real database for the first time. It is envisaged in this work that in large databases, it is more likely that the missing values could be correlated to some other variables observed somewhere in the same data. Instead of approximating missing data, it might therefore be cheaper to spot similarities between the observed data instances and those that contain missing attributes.

2 Applications of Rough Sets

There are many applications of rough sets reported in literature. Most of the applications assume that complete data is available (Grzymala-Busse, 2004). This is, however, not often the case in real life situations. There is also a great deal of information regarding various applications of rough sets in medical data sets. Rough sets have been used mostly in prediction cases and Rowland et al. (1998) compared neural networks and rough sets for the prediction of ambulation following a spinal cord injury. Although rough sets performed slightly lower than neural networks, they proved that they can still be used in prediction problems. Rough sets have also been used in learning Malicious Code Detection (Zhang et al., 2006) and in Fault diagnosis (Tay and Shen, 2003). Grzymala-Busse and Hu (2001) have presented nine approaches of imputing up missing values. Among others, the presented methods include selecting the most common attribute, *concept most common* attribute, assigning all possible values related to the current concept, deleting cases with missing values, treating missing values as special values and imputing for missing values using other techniques such as neural networks, and maximum likelihoods approaches. Some of the techniques proposed come with expense either in terms of computation time or loss of information.

3 Rough Set Theory

The rough sets theory provides a technique of reasoning from vague and imprecise data (Goh and Law, 2003). The technique is based on the assumption that some information is associated somehow with *some information* of the universe of the discourse (Komorowski et al., 1999; Yang and John, 2006). Objects with the same information are *indiscernible* in the view of the available information. An elementary set consisting of indiscernible objects forms a basic granule of knowledge. A union of elementary set is referred to as a crisp set, otherwise the set is considered to be rough. The next few subsections briefly introduce concepts that are common to rough set theory.

3.1 Information System

An information system (Λ), is defined as a pair $(\mathbf{U}, A)$ where $\mathbf{U}$ is a finite set of objects called the universe and A is a non-empty finite set of attributes as shown in Eq 1 below (Yang and John, 2006).

$$\Lambda = (\mathbf{U}, A) \tag{1}$$

Every attribute $a \in A$ has a value which must be a member of a value set V_a of the attribute a.

$$a : \mathbf{U} \rightarrow V_a \tag{2}$$

A rough set is defined with a set of attributes and the indiscernibility relation between them. Indiscernibility is discussed next.

3.2 Indiscernibility Relation

Indiscernibility relation is one of the fundamental ideas of rough set theory (Grzymala-Busse and Siddhaye, 2004). Indiscernibility simply implies similarity (Goh and Law, 2003). Given an information system Λ and subset $B \subseteq A$, B determines a binary relation $I(B)$ on $\mathbf{U}$:

$$(x, y) \in I(B) \quad iff \quad a(x) = a(y) \tag{3}$$

for all $a \in B$ where $a(x)$ denotes the value of attribute a for element x. Eq (3) implies that any two elements that belong to $I(B)$ should be identical from the point of view of a. Suppose $\mathbf{U}$ has a finite set of N objects $\{x_1, x_2, \ldots, x_N\}$. Let Q be a finite set of n attributes $\{q_1, q_2, \ldots, q_n\}$ in the same information system Λ, then,

$$\Lambda = \langle \mathbf{U}, Q, V, f \rangle \tag{4}$$

where f is the *total decision function* called the information function. From the definition of Indiscernibility Relation given in this section, any two objects have a similarity relation to attribute a if they have the same attribute values everywhere except for the missing values.

3.3 Information Table and Data Representation

An Information Table (IT) is used in rough sets theory as a way of representing the data. The data in the IT are arranged based on their condition attributes and decision attribute ($\mathcal{D}$). Condition attributes and decision attribute are analogous to the independent variables and dependent variable (Goh and Law, 2003). These attributes are divided into $C \cup \mathcal{D} = Q$ and $C \cap \mathcal{D} = \emptyset$. An IT can be classified into complete and incomplete classes. All objects in a complete class have known attribute values whereas an IT is considered incomplete if at least one attribute variable has a missing value. An example of an incomplete IT is given in Table 1.

Data is represented by a table where each row represents an instance, sometimes referred to as an object. Every column represents an attribute which can be a measured variable. This kind of a table is also referred to as Information System (Komorowski et al., 1999).

Table 1: An example of an Information Table with missing values

	x_1	x_2	x_3	$\mathcal{D}$
1	1	1	0.2	B
2	1	2	0.3	A
3	0	1	0.3	B
4	?	?	0.3	A
5	0	3	0.4	A
6	0	2	0.2	B
7	1	4	?	A

3.4 Decision Rules Induction

Rough sets also involve generating decision rules for a given IT. The rules are normally determined based on condition attributes values (Goh and Law, 2003). The rules are presented in an *if* CONDITION(S)-*then* DECISION format. This paper will not directly focus on rule induction since the major interest of this work is to estimate the missing data as opposed to taking the decision.

3.5 Set Approximation

There are various properties of rough sets that have been presented in (Pawlak, 1991) and (Pawlak, 2002). Some of the properties are discussed below.

3.5.1 Lower and Upper Approximation of Sets

The lower and upper approximations are defined on the basis of indiscernibility relation discussed above. The lower approximation is defined as the collection of cases whose equivalent classes are contained in the cases that need to be approximated whereas the upper approximation is defined as the collection of classes that are partially contained in the set that needs to be approximated (Rowland et al., 1998).

Let **concept** X be defined as a set of all cases defined by a specific value of the decision. Any finite union of elementary set, associated with B is called a $B-definable$ set (Grzymala-Busse and Siddhaye, 2004). The set X is approximated by two $B-definable$

sets, referred to as the B-lower approximation denoted by $\underline{B}X$ and B-upper approximation, $\overline{B}X$. The B-lower approximation is defined as (Grzymala-Busse and Siddhaye, 2004)

$$\{x \in \mathbf{U} | [x]_B \subseteq X\} \tag{5}$$

and the B-upper approximation is defined as

$$\{x \in \mathbf{U} | [x]_B \cap X \neq \emptyset\} \tag{6}$$

There are other methods that have been reported in the literature for defining the lower and upper approximations for a completely specified decision tables. Some of the common ones include approximating the lower and upper approximation of X using Equations 7 and 8 respectively as follows (Grzymala-Busse, 2004):

$$\cup \{[x]_B | x \in \mathbf{U}, [x]_B \subseteq X\} \tag{7}$$

$$\cup \{[x]_B | x \in \mathbf{U}, [x]_B \cap X \neq \emptyset\} \tag{8}$$

The definition of definability is modified in cases of incompletely specified tables. In this case, any finite union of characteristics sets of B is called a $B - definable$ set. Three different definitions of approximations have been discussed Grzymala-Busse and Siddhaye (2004). Again letting B be a subset of A of all attributes and $R(B)$ be the characteristic relation of the incomplete decision table with characteristic sets $K(x)$, where $x \in U$, the following are defined:

$$\underline{B}X = \{x \in \mathbf{U} | K_B(x) \subseteq X\} \tag{9}$$

and

$$\overline{B}X = \{x \in \mathbf{U} | K_B(x) \cap X \neq \emptyset\} \tag{10}$$

Equations 9 and 10 are referred to as *singletons*. The *subset* lower and upper approximations of incompletely specified data sets are then defined as:

$$\cup \{K_B(x) | x \in \mathbf{U}, K_B(x) \subseteq X\} \tag{11}$$

and

$$\cup \{K_B(x) | x \in \mathbf{U}, k_B(x) \cap X \neq \emptyset\} \tag{12}$$

More information on these methods can be found in (Grzymala-Busse, 2004; Grzymala-Busse and Hu, 2001; Grzymala-Busse, 1992; Grzymala-Busse and Siddhaye, 2004).

It follows from the properties that a crisp set is only defined if $\underline{B}(X) = \overline{B}(X)$. Roughness therefore is defined as the difference between the upper and the lower approximation.

3.5.2 Rough Membership Functions

Rough membership function is a function $\mu_A^x : \mathbf{U} \rightarrow [0, 1]$ that when applied to object x, quantifies the degree of overlap between set X and the indiscinibility set to which x belongs. The rough membership function is used to calculate the plausibility, defined as

$$\mu_A^X(X) = \frac{|[x]_B \cap X|}{|[x]_B|} \tag{13}$$

4 Missing Data Imputation Based on Rough Sets

The algorithm implemented here imputes the missing values by presenting a list of all possible values, based on the observed data. As mentioned earlier, the hypothesis here is that in most finite databases, a case similar to the missing data case could have been observed before. It therefore should be cheaper to use such values, instead of computing missing values with complex methods such as neural networks. The algorithm implemented is shown in Algorithm 1, followed by a *work-through example* demonstrating how the missing values are imputed. There are two approaches to reconstructing the missing values. The missing values can either be probabilistically interpreted or be possibilistically interpreted (Nakata and Sakai, 2006).

```
Algorithm 1: Rough sets based missing data imputation algorithm
  input       : Incompete data set $\Lambda$ with $a$ attributes and $i$ instances.
                All these instances should belong to a desision $\mathcal{D}$
  output      : A vector containing possible missing values
  Assumption: $\mathcal{D}$ and some attributes will always be known
  forall $i$ do
      → Partition the input space according to $\mathcal{D}$ → Arrange all attributes according
      to order of availability, with $\mathcal{D}$ being first.
  end
  foreach attribute do
      → Without directly extracting the rules, use the available information to
      extract relationships to other instances $i$ in the $\Lambda$.
      → The family of equivalent classes $\varepsilon(a)$ containing each object $o_i$ for all input
      attributes is computed.
      → The degree of belongingness $\kappa(o[A]1/|dom(a_{i_{missing}})|$ where $o \neq o'$ and
      $dom(x_{1_4})$ denotes the domain of attribute $x_{1_4}$, which is the forth instance of $x_1$,
      and $|dom(x_{1_4})|$ is the cardinality of $dom(x_{1_4})$  while extracting relationships do
          If $i$ has the same attribute values with $a_j$ everywhere except for the missing
          value, replace the missing value, $a_{missing}$, with the value $v_j$, from $a_j$, where
          $j$ is an index to onother instance.
          Otherwise proceed to the next step
      end
      → Complete the lower approximation of each attribute,given the available data
      of the same instance with the missing value.
      while doing this do
          IF more than one $v_j$ values are suitable for the estimation, postpone the
          replacement for later when it will be clear which value is appropriate
      end
      → Compute the incomplete upper approximations of each subset partition.
      → Do the computation and imputation of missing data as was done with the
      lower approximation.
      → Either crips sets will be found, otherwise, rough sets can be used and
      missing data can be heuristically be selected from the obtained rough set.
  end
```

In our example, the degree of belongingness $\kappa(o[x_{1_4}] = o[x_{1_4}] = 1/|dom(x_{1_4})|$ where $o \neq o'$ and $dom(x_{1_4})$ denotes the domain of attribute x_{1_4},which is the forth instance of x_1, and $|dom(x_{1_4})|$ is the cardinality of $dom(x_{1_4})$. If the missing values were to be possibilistically interpreted, all attributes have the same possibilistic degree of being the actual one.

The algorithm in this study is fully dependent on the available data and makes no additional assumptions about the data or the distribution thereof. As presented in the algorithm, a list of possible values is given in a case where a crisp set could not be found. It is from this list that possible values may be heuristically chosen. A justification to this is that it is not always the case that we need to know the *exact* value. As a result, it may be cheaper to have a *rough* value. The possible imputable values are obtained by collecting all the entries that lead to a particular decision $\mathcal{D}$. The algorithms used in this application is a simplified version of the algorithm of Hong et al. (2002).

The algorithm will now be illustrated using an example. Missing values will be denoted by the question mark (?) symbol. Attribute values of attribute a are denoted as V_a. Using the notation defined in Gediga and Duntsch (2003), we let $rel_Q(x)$ represent a set of all *Q-relevant attributes* of x. Assuming an IT as presented in Table 1, where x_1 is in binary form, $x_2 \in [1:5]$ and being integers and x_3 can either be 0.2, 0.3 or 0.4.

The algorithms firstly seeks relationship between variables. Since this is a small database, it is assumed that the only variable that will always be known is the decision. The first step will be to partition the data according to the decision and this could be done as follows:

$$\varepsilon(D) = \{o_1, o_3, o_6\}, \{o_2, o_4, o_5, o_7\}$$

Two partitions are obtained due the binary nature of the decision in the chosen example. The next step is to extract indiscernible relationships within each attribute.

For x_1, the following is obtained:

$$\begin{aligned} IND(x_1) = \{&(o_1, o_1), (o_1, o_2), (o_1, o_4), (o_1, o_7), (o_2, o_2), (o_2, o_4), (o_2, o_7), \\ &(o_3, o_3), (o_3, o_4), (o_3, o_5), (o_3, o_6), (o_4, o_4), (o_4, o_5), (o_4, o_6)(o_4, o_7), \\ &(o_5, o_5), (o_5, o_6), (o_6, o_6), (o_7, o_7)\} \end{aligned}$$

The family of equivalent classes $\varepsilon(x_1)$ containing each object o_i for all input variables is computed as follows:

$$\varepsilon(x_1) = \{o_1, o_2, o_4, o_7\}, \{o_3, o_4 o_5, o_6\}$$

Similarly,

$$\varepsilon(x_2) = \{o_1, o_3, o_4\}, \{o_2, o_4, o_6\}, \{o_4, o_5\}, \{o_, o_7\}, \{o_4\}\{0_7\}$$

and

$$\varepsilon(x_3) = \{o_1, o_6, o_7\}, \{o_2, o_3, o_4, o_7\}, \{o_5, o_7\}$$

In our example, the degree of belongingness $\kappa(o[x_{1_4}] = o[x_{1_4}] = 1/|dom(x_{1_4})|$ where $o \neq o'$ and $dom(x_{1_4})$ denotes the domain of attribute x_{1_4}, which is the fourth instance of x_1, and $|dom(x_{1_4})|$ is the cardinality of $dom(x_{1_4})$. If the missing values were to be possibilistically interpreted, each attribute has the same possibilistic degree of being the actual one. The lower approximations is defined as:

$$\underline{A}(X_{miss}, \{X_{avail}, \mathcal{D}\}) = \{E(X_{miss}) | \exists (X_{avail}, \mathcal{D}), E(X) \subseteq (X_{avail}, \mathcal{D})\} \tag{14}$$

whereas the upper approximation is defined as

$$\overline{A}(X_{miss}, \{X_{avail}, \mathcal{D}\}) = \{E(X_{miss}) | \exists (X_{avail}, \mathcal{D}), E(X) \cap X_{avail} \cap \mathcal{D}\} \tag{15}$$

Using $IND(x_1)$, the families of all possible classes containing o_4 are given by

$$\begin{aligned} Poss\varepsilon(x_1)_{o_i} &= \{o_1, o_2, o_7\}, \{o_1, o_2, o_4, o_7\}, i = 1, 2, 7 \\ Poss\varepsilon(x_1)_{o_i} &= \{o_3, o_5, o_6\}, \{o_3, o_4, o_5, o_6\}, i = 3, 5, 6 \\ Poss\varepsilon(x_1)_{o_4} &= \{o_4, o_1, o_2, o_7\}, \{o_3, o_4, o_5, o_6\} \end{aligned}$$

The probabilistic degree to which we can be sure that the chosen value is the right one is given by (Nakata and Sakai, 2006)

$$\begin{aligned} \kappa((\{o_i\}) \in \varepsilon(x_1)) &= 1/2, i = 1, 2, 7 \\ \kappa((\{o_i\}) \in \varepsilon(x_1)) &= 1/2, i = 3, 5, 6 \\ \kappa((\{o_i\}) \in \varepsilon(x_1)) &= 1/2, i = 4 \\ & else \\ \kappa(\{o_i\}) \in \varepsilon(x_1)) &= 0 \end{aligned}$$

The else part applies to all other conditions such as $\kappa(\{o_1, o_2, o_3\}) \in \varepsilon(x_1)) = 0$. A family of weighted equivalent classes is now computed as follows:

$$\varepsilon(x_1) = \{\{o_1, o_2, o_4, o_7\}\{1/2\}\}, \{\{o_3, o_4 o_5, o_6\}\{1/2\}\}$$

The values $\varepsilon(x_2)$ and $\varepsilon(x_3)$ are computed in a similar way. We then use these families of weighted equivalent classes to obtain the lower and upper approximations as presented above. The degree to which the object o has the same value as object o' on the attributes is referred to as the degree of belongingness and is defined in terms of the binary relation for indiscernibility as (Nakata and Sakai, 2006):

$$
\begin{aligned}
IND(X) = \{((o, o'), \kappa(o[X] = o'[X]))|(\kappa(o[X] = o'[X]) \\
\neq 0) \wedge (o \neq o')\} \cup \{((o, o), 1)\}
\end{aligned}
$$

where $\kappa(o[X] = o'[X])$ is the indiscernibility degree of the objects o and o' and this is equal to the degree of belongingness,

$$\kappa(o[X] = o'[X]) =_{A_i \in X}^{\otimes} \kappa(o[A_i] = o'[A_i])$$

where the operator $\otimes$ depends on whether the missing values are possibilistically or probabilistically interpreted. For probabilistic interpretation, the parameter is a product denoted by ×, otherwise the operator *min* is used.

5 Experimentatal Evaluation

5.1 Database

The data used in this test was obtained from the South African antenatal sero-prevalence survey of 2001. The data for this survey is obtained from questionnaires answered by pregnant women visiting selected public clinics in South Africa. Only women participating for the first time in the survey were eligible to answer the questionnaire.

Data attributes used in this study are the *HIV status, education level, gravidity, parity, age, age of the father, race* and *region* . The HIV status is the decision and is represented in a binary form, where 0 and 1 represent negative and positive respectively. Race is measured on the scale 1 to 4 where 1, 2, 3, and 4 represent African, Coloured, White and Asian, respectively. The data used was obtained in three regions and are referred to as region A, B and C in this investigation. The education level was measured using integers representing the highest grade successfully completed, with 13 representing tertiary education. Gravidity is the number of pregnancies, complete or incomplete, experienced by a female, and this variable is represented by an integer between 0 and 11. Parity is the number of times the individual has given birth and multiple births are counted as one. Both parity and gravidity are important, as they show the reproductive activity as well as the reproductive health state of the women. Age gap is a measure of the age difference between the pregnant woman and the prospective father of the child. A sample of this data set is shown in Table 2.

Table 2: Extract of the HIV database used, with missing values

Race	Region	Educ	Gravid	Parity	Age	Father's age	HIV
1	C	?	1	2	35	41	0
2	B	13	1	0	20	22	0
3	?	10	2	0	?	27	1
2	C	12	1	?	20	33	1
3	B	9	?	2	25	28	0
?	C	9	2	1	26	27	0
2	A	7	1	0	15	?	0
1	C	?	4	?	25	28	0
4	A	7	1	0	15	29	1
1	B	11	1	0	20	22	1

5.2 Data Preprocessing

As mentioned in a previous section, the HIV/AIDS data that is used in this work is obtained from a survey performed on pregnant women. Like all data in raw form, there are several steps that need to be taken in order to ensure the data is in usable form. There are several types of outliers that have been identified in the data. Firstly, some of the data records were not complete. This is probably due to the fact that the people being surveyed omitted certain information and also errors made by the person who manually recorded the surveys onto a spreadsheet. The outliers were from incorrectly entered variables. For instance *Gravidity* is defined as the number of times a woman has been pregnant and *parity* is described as the number of times a woman has given birth. Any instance whereby the value of parity is greater than that of parity, the whole observation was considered an outlier and was removed. The justification to this is that it is not possible for a woman to give birth more than she has been pregnant.

5.3 Variable Discretisation

The discretisation defines the granularity with which we would like to analyse the universe of discourse. If one chooses to discretise the variables into a large number of categories the rules extracted are more complex to analyse. Therefore, if one would like to use rough sets for rule analysis and interpretation rather than for classification it is advisable that the number of categories be as small as possible. For the purposes of this work the input variables have been discretised into four categories. A description of the categories and their definition is shown in Table 3. Table 4 shows the simplified version of the information system shown in Table 2.

Table 3: A table showing the discretised variables.

Race	**Age**	**Education**	**Gravidity**	**Parity**	**Father's Age**	**HIV**
1	≤ 19	Zero (0)	Low (≤ 3)	Low (≤ 3)	≤ 19	0
2	$[20 - 29])$	P $(1 - 7)$	High (> 3)	High (> 3)	$([20 - 29])$	1
3	$[30 - 39])$	S $(8 - 12)$	-	-	$([30 - 39])$	-
4	≥ 40	T (13)		-	- ≥ 40	-

5.4 Results and Discussion

The experimentation was performed using both the original and the simplified data sets. Results obtained in both cases are summarised in Table 5.

It can be seen that the prediction accuracy is much higher for the generalised data set. This is because the states have been reduced. Furthermore, instead of being exact, the likelihood of being correct is even higher if one has to give a rough estimate. For

Table 4: Extract of the HIV database used, with missing values after discretisation

Race	Region	Educ	Gravid	Parity	Age	Father's age	HIV
1	C	?	≤ 3	≤ 3	[31:40]	[41:50]	0
2	B	T	≤ 3	≤ 3	≤ 20	[21:30]	0
3	?	S	≤ 3	≤ 3	?	[21:30]	1
2	C	S	≤ 3	?	≤ 20	[31:40]	1
3	B	S	?	≤ 3	[21:30]	[21:30]	0
?	C	S	≤ 3	≤ 3	[21:30]	[21:30]	0
2	A	P	≤ 3	≤ 3	≤ 20	?	0
1	C	?	> 3	?	[21:30]	[21:30]	0
4	A	P	≤ 3	≤ 3	≤ 20	[21:30]	1
1	B	S	≤ 3	≤ 3	≤ 20	[21:30]	1

Table 5: Missing data estimation results for both the original data and the generalised data

	Education	Gravidity	Parity	Father's age
Original	83.1	86.5	87.8	74.7
Generalised	99.3	99.2	99	98.5

instance, instead of saying that someone has a highest level of education of 10, it is much safer to say, *They have secondary education.* Although this approach leaves details, it is often the case that the left-out details are not required. In a decision system such as the one considered in this chapter, knowing that the prospective father is 19 years old may carry the same weight as saying that the father is a *teenager*.

6 Conclusion

Rough sets have been used for missing data imputation and characteristic relations are introduced to describe incompletely specified decision tables. It has been shown that the basic rough set idea of lower and upper approximations for incompletely specified decision tables may be defined in a variety of different ways. The technique was tested with a real database and the results with the HIV database are acceptable with accuracies ranging from 74.7% to 100%. One drawback of this method is that it makes no extrapolation or interpolation and as a result, can only be used if the missing case is similar or related to another case with full or more observation.

7 Acknowledgement

The financial assistance of the National Research Foundation (NRF) of South Africa and the Council for Scientific and Industrial Research (CSIR)towards this research is hereby acknowledged.

References

Abdella, M., Marwala, T., 2006. The use of genetic algorithms and neural networks to approximate missing data in database. Computing and Informatics 24, 1001–1013.

Gabrys, B., 2002. Neuro-fuzzy approach to processing inputs with missing values in pattern recognition problems. International Journal of Approximate Reasoning 30, 149–179.

Gediga, G., Duntsch, I., 2003. Maximum consistency of incomplete data via non-invasive imputation. Artificial Intelligence Review 19, 93–107.

Goh, C., Law, R., 2003. Incorporating the rough sets theory into travel demand analysis. Tourism Management 24, 511–517.

Grzymala-Busse, J. W., 1992. LERSA system for learning from examples based on rough sets. Handbook of Applications and Advances of the Rough Sets Theory: Kluwer Academic Publishers.

Grzymala-Busse, J. W., 2004. Three approaches to missing attribute values - a rough set perspective. In: IEEE Fourth International Conference on Data Mining. Brighton, United Kingdom, pp. 57–64.

Grzymala-Busse, J. W., Hu, M., 2001. A comparison of several approaches to missing attribute values in data mining 205, 378–385.

Grzymala-Busse, J. W., Siddhaye, S., 2004. Rough set approaches to rule induction from incomplete data. In: Proceedings the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Vol. 2. Perugia, Italy, pp. 923–930.

Hong, T., Tseng, L., Wang, S., 2002. Learning rules from incomplete training examples. Expert Systems With Application 22, 285–293.

Kim, J., Curry, J., 1997. The treatment of missing data in multivariate analysis. Sociological Methods and Research 6, 215–241.

Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A., 1999. A Rough Set Perspective on Data and Knowledge. The Handbook of Data Mining and Knowledge Discovery, Oxford Univesrity Press.

Little, R. J. A., Rubin, D. B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

Mohamed, S., Marwala, T., 2005. Neural network based techniques for estimating missing data in databases. In: The 16th Annual Symposium of the Pattern Recognition Association of South Africa. Langebaan, South Africa, pp. 27–32.

Nakata, M., Sakai, H., 2006. Rough sets approximations to possibilistic information. In: IEEE International Conference on Fuzzy Systems. Vancouver, BC, Canada, pp. 10804–10811.

Nauck, D., Kruse, R., 1999. Learning in neuro-fuzzy systems with symbolic attributes and missing values. In: Proceedings of the IEEE International Conference on Neural Information Processing. Perth, pp. 142–147.

Pawlak, Z., 1991. Rough sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht.

Pawlak, Z., 2002. Rough sets and intelligent data analysis. Information Science 147, 1–12.

Qiao, W., Gao, Z., Harley, R. G., 2005. Continuous online identification of nonlinear plants in power systems with missing sensor measurements. In: IEEE International Joint Conference on Neural Networks. Montreal, pp. 1729–1734.

Rowland, T., Ohno-Machado, L., Ohrn, A., 1998. Comparison of multiple prediction models for ambulation following spinal chord injury. In Chute 31, 528–532.

Tay, F. E. H., Shen, L., 2003. Fault diagnosis based on rough set theory. Engineering Applications of Artificial Intelligence 16, 39–43.

Wang, S., 2005. Classification with incomplete survey data: a Hopfield neural network approach. Computers & Operations Research 24, 53–62.

Yang, Y., John, R., 2006. Roughness bound in set-oriented rough set operations. In: IEEE International Conference on Fuzzy Systems. Vancouver, Canada, pp. 1461–1468.

Zhang, B., Yin, J., Tang, W., Hao, J., Zhang, D., 2006. Unknown malicious codes detection based on rough set theory and support vector machine. In: IEEE International Joint Conference on Neural Networks. Vancouver, Canada, pp. 4890–4894.