Received: June 17, 2017

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Core Reduct Based Preprocessing Approach to Incomplete Data

Pallab Kumar Dey[1]* **Sripati Mukhopadhyay[2]**

[1]*Department of Computer Science, Kalna College, Kalna-713409, India*
[2]*Department of Computer Science, The University of Burdwan, Bardhaman-713104, India*
* Corresponding author's Email: pallabkumardey@gmail.com

Abstract: Most of the data mining algorithm's application hampers due to missing attribute values. Inadequate treatment of missing values seriously affects the data mining and classification accuracy. A useful technique has been proposed to deal with missing attribute values. Rough set approach to incomplete information system has been shown. Application of discernible matrix for incomplete information to compute core and reduct has been shown. Imputation based preprocessing approach depends on relation between present attribute value and incomplete attribute value, so have to found most similar object to impute missing value. To find similar objet importance or priority should be given in core attributes, after that reduct attributes if matching occur in corresponding attributes value and other attributes may be neglected. In this paper this concept of core and reduct attributes of rough set has been utilized to fill missing values using the proposed Core Reduct Based (CRB) algorithm. Efficiency of the CRB algorithm in the completeness analysis to incomplete data has been shown by comparing its performance with other existing algorithms using some real life data sets.

Keywords: Data mining, Missing values, Rough set, Core, Reduct.

1. Introduction

With advancement of E-technology collection of raw data has been increased rapidly but, finding useful information from such a large data collection is a challenging issue. Since most of the data mining technology is based on ideal data. But real data sets are inconsistent and incomplete. In this paper incomplete data sets are considered. Data may be incomplete for many reasons. At the time of experiment or data collection it may not be available or due to time constraints and cost efficiency it is not collected. It is not possible to use existing data mining methods to those data sets directly. Inadequate treatment of missing values seriously affects the data analysis and classification accuracy. Missing data hamper application of data mining technology. Some data mining algorithms may handle missing value directly, but a pre-processing step to handle missing attribute value is more appropriate to use already available data mining algorithms effectively to improve performance. In this paper Rough set approach to incomplete information systems has been analyzed and it has been used as pre-processing tool to handle missing attribute values. The main advantage of using rough sets is that it does not need any additional or prior information about data. Imputation based preprocessing approach depends on relation between present attribute value and incomplete attribute value, so have to found most similar object to impute missing value. To find most similar objet importance or priority should be given in core attributes, after that reduct attributes if matching occur in corresponding attributes value and other attributes may be neglected considering as redundant attributes. Using this concept, Core Reduct Based (CRB) algorithm has been proposed to impute missing data. In this imputation based approach missing data is replaced if most similar object present with complete value. So after application of CRB algorithm there is no chance to generate misleading information. Proposed CRB algorithm has been used in different real life data set. Its performance has been compared with other

existing methods. Experimental result shows its efficiency over other methods. Also it is clear from experiment that its classification accuracy is better than other methods. Main advantage of the proposed method is, to impute missing data most similar object (considering main attributes) has been chosen.

Organization of the paper is as follows: Section 2 is devoted to literature overview. Section 3 deals with Rough set theory to compute reduct and core for incomplete information. Section 3 deals with computation modelling, and the experimental results based on the proposed algorithm are shown in section 4.

2. Literature Overview

Depends on nature of missing value [1-2] it can be divided into three categories. When probability of being missing is same for every value then MCAR (missing completely at random), so a value to be missing does not depends on either of the observe data or missing data. When probability of missing depends on other attributes value then MAR (Missing at random), so to be missing depend on the observe data. And not missing at random (NMAR) when probability of missing value depends on missing value itself, it is non ignorable and have to solve by going back to the source only. To handle incomplete information, natures of missing value have to consider. Object consist of missing values may be deleted [3] list-wise or pair-wise but we lost resources. List-wise deletion may be useful when data set is too is too large, missing values are completely random and missing rate is low. Due to computational complexity of covariance matrix pair-wise deletion is not so popular, though in pair-wise deletion all available information has been considered. But data analysis and classification accuracy may be biased by deletion. We may consider missing value as a special value and proceed in the same way as other values [4]. It also effect classification accuracy and data analysis. So data sets may be pre-processed to change in complete data or have to use data mining algorithm which can handle missing data. Rule may be generated or knowledge can be extracted directly from incomplete data sets [5-8] though available data mining algorithm cannot be used here.

In C4.5 method [5] decision tree has been used to classify new records. Extension of KNN classifier [8] or instance based learning algorithms can be used to classify incomplete data set. Modified LEM2 algorithm [6-7] has been used by computing block of the attributes with the objects of known values and then induced certain rules using original LEM2 method.

Main concept of data pre-processing is, to change incomplete data based on imputation of existing data. So for pre-processing there should be a relation between missing data and complete data. So missing data may be replaced or fetched from existing data if suitable matching found. In pre-processing step missing value is replaced with mean, of all complete values of the attribute for numeric type, or by mode of that attribute considering complete data for linguistic attributes [9, 2]. Missing values may be replaced randomly by retaining standard deviation same [10] but complex to implement. In pre-processing of data mining missing values may be managed by different strategy like, maximum occurring attribute value or maximum occurring attribute value considering same class value [11-12], all feasible domain values of the attribute or all feasible domain values considering same class value [4, 13] or by various statistical methods [9-10, 14-15]. These methods are useful to predict missing values but due to limitation of application, cannot be used in all data sets. Most similar instance is used in k nearest neighbour (KNN) imputation method [16] to impute the missing values of an instance considering a given number of instances. But computation cost of neural networks is very high.

Hot deck imputation [17] is also very popular. In this approach each missing value is replaced from the similar case. 'Hot deck' term derived from punch card. Imputed value comes from other cards in the current deck being processed (so hot). Similarly in 'cold deck' imputations data comes from previously collected databases i.e. from decks of card currently not processed (so cold). Different forms of hot deck methods are available. These methods not always fill the same value which have a significant impact on variance estimations and no extra value introduce which are not present and successfully used in larger data set but for smaller data set sample variance increased and produce poor result. Expectation maximization (EM) algorithm [18] is an iterative procedure for computing maximum likelihood estimation for incomplete data. Available informations within the data set are used by EM algorithm. The EM algorithm consists of two steps, E-step (expectation) and M-step (maximization), which are iterative and alternates between this two steps until convergence. Conditional expectations of the complete data likelihood estimated at E-step based on observe data. Computed expected likelihood is maximized at M-step with respect to the parameters. Implementation of EM algorithm is

very complicated so it has limited uses. In Multiple imputations (MI) method [19-20] missing value is replaced by more than one value derived from non missing values, in contrast to single imputation where missing value is replaced by only one value. MI produce unbiased estimates of missing values since it consider all possibility. For low sample size or high missing rates MI can produce better result than single imputation. Multiple imputations may face difficulties for competition with larger number of missing values. It also time consuming and not cost efficient. Iterative model based algorithm IRMI [21] is a popular robust imputation method for automatic imputation. It has been shown by experiment that algorithm usually converges in a few iterations and proposes better result.

Rough set theory is also emerging tool to deal with missing value. Indiscernibility relation and discernibility matrix of rough set has been used to fill missing values [22]. Here filling ratio is not considerable. Rough set based tolerance relation has been proposed [23]. Here dispensability of attributes, indispensability of attributes, core, and functional dependency between attributes of rough set has been redefined for incomplete information. It has been shown to fetch decision rule directly from such an incomplete decision table. So rough set approach has been approved to reasoning with incomplete information system also. To calculate similarity degree of tolerance relation, value tolerance relation [24] has been proposed considering uniform probability. Extended valued tolerance relation [25] have been used to anticipate missing values considering filling capacity with similarity. Characteristic relations are introduced to describe decision tables with missing attribute values [26]. Computations of characteristic relations, using an idea of block of attribute-value pairs have been shown and definitions of lower and upper approximations are defined in three different ways for incompletely specified decision tables. In a uniform way three approaches to missing attribute values are presented [27]. It has been shown that attribute value blocks are main concept of these definitions. It is also shown that for computing characteristic sets, characteristic relations, lower and upper approximations and for rule induction attribute value blocks may be used. Local approximations and global approximations for incomplete data have been introduced [28] such that corresponding upper approximations are minimal, as other existing definitions of upper approximations are not minimal definable sets. It is also shown that for decision tables containing only lost missing attribute values, local and global approximations are equal to one another and they are unique. For incomplete decision tables attribute-value pair block used to determine characteristic sets, characteristic relations, lower and upper approximations, and rule induction [29]. Six different rough set approaches to missing attribute values are tested and concluded that lost values provide better results in terms of smaller error rate. Fuzzy-Rough nearest neighbour based [30] tool is also interesting to impute missing value.

But these methods do not considered attributes impact or significance of attributes on data sets at the time of anticipating missing values. In this paper core, reduct and other attributes significance at the time of anticipating missing values have been considered differently according to their impact on data set.

3. Rough set theory to compute reduct and core for incomplete information

3.1 Basic concepts of Rough set

To analyze inexact, uncertain and vague knowledge pawlak's rough set theory [31] is the most prominent tool. The rough sets theory provides a technique to deal with vague and imprecise data. Objects with the same information are indiscernible considering available information. In this paper rough set approach to incomplete information system has been considered. Data sets are presented by decision tables where columns label denote attribute and rows by object. Attributes are categories into conditional attribute and decision attribute. Independent attributes are called conditional attribute and dependent attributes are called conditional attribute.

Information system may be represented as four tuple *(U,A,V,f)* where U is a non empty finite set of objects, A is a non empty finite set of attributes, $\forall\, a \in A$ V_a is the domain of attribute *'a'*. $V = \cup V_a$ is the domain of *A*, *f* is a mapping $f: U \times A \rightarrow V$, $f(x,a) \in V_a$ is the value that x holds on a. Indiscernibility relation is the fundamental idea of rough set theory. Any subset B of *A* determines a binary relation *I(B)* called indiscernibility relation defined as:

$(x_i, x_j) \in I(B)$ if and only if $a(x_i) = a(x_j)$ for every $a \in B$

where $a(x_i)$ is the value of attribute a for element x_i. (1)

Indiscernibility relation is an equivalence relation. Equivalence classes of *I(B)* i.e., block of partition determined by B are called elementary sets and are denoted by *U/B* or *B(x)*. In rough set approach elementary sets are basic building block of our knowledge.

Reduction of attributes is a important application of Rough Set. If removing of some attributes does not effect basic properties of a table then these attributes are redundant. If attribute $a \in B$ and *I(B)=I(B-{a})* then attribute *'a'* is dispensable, otherwise attribute *'a'* is indispensable.

Minimal subsets of attributes that maintain same partition as whole set of attributes is called reduct. Set of all indispensable attributes of the universe is called core, also it may be defined as intersection of all reduct. Subset of attributes that cannot be removed without effecting classification power of attributes are called core.

To compute reduct and core discernible matrix may be used. Discernible matrix *M(B)* of *B* is a $n \times n$ matrix defined as :

$$C_{ij} = \{a \in B: a(x_i) \neq a(x_j)\} \quad for\ i,j = 1,2,..n. \qquad (2)$$

C_{ij} is the set of all attributes that discern object x_i and x_j.

Core is the set of all single element entries of the discernible matrix *M(B)*.

$$Core(B) = \{a \in B, C_{ij} = \{a\}, for\ some\ i,j\} \qquad (3)$$

Discernibility function [28] f is a boolean function of m boolean variables $a_1^*, \ldots, a_m^*$ (corresponding to the attributes $a_1, \ldots, a_m$) computed from discernibility matrix *M(B)* , may defined as:

$$f(a_1^*, \ldots, a_m^*) = \wedge \{\vee\ c_{ij}^* | 1 \leq j \leq i \leq |U|,\ C_{ij} \neq \emptyset\ \} \qquad (4)$$

where $c_{ij}^* = \{a^* | a \in C_{ij}\}$.

All minimal reduct of a system may be found by simplification of discernibility function, using boolean laws of boolean algebra. Core may also be found by intersections of all reduct.

3.2 Rough set approach to incomplete information

Decision table with incomplete information i.e., containing missing values will be called incomplete information system. If some values of *f(x,a)* are missing (denoted by '?'), then it can be described by incomplete information system(IIS), defined as *(U, A,V', f)* where *U* is a non empty finite set of objects, *A* is a non empty finite set of attributes, $V'=V \cup \{?\}$, $\forall\ a \in A\ V_a$ is the domain of attribute *a*, $V= \cup V_a$ is the domain of *A*, *f* is a mapping $f:U \times A \rightarrow V'$, $\exists$ *some* $x \in U$ *and* $a \in A$ *such that* $f(x,a)=$*'?'*. Example of Incomplete decision table has been shown in table 1.

For incomplete decision table similarity relations are used in the same way as indiscernibility relations are used to define complete decision table. Similarity relation indicates possibly indiscernible i.e., objects that cannot be certainly said that they are different. Any subset *B of A* determines a binary relation *S(B)* called similarity relation defined as:

$(x_i, x_j) \in S(B)$ if and only if $a(x_i) = a(x_j)$ *or* $a(x_i) =?$ *or* $a(x_j) =?$ for every $a \in B$

where $a(x_i)$ is the value of attribute a for element x_i. (5)

Similarity relation may not be equivalence relation. *U/S(B)* do not constitute a partition of *U* , they may be subset or superset of each other or may overlap.

Table 1. Incomplete decision table

Cases	Temperature	Headache	Cough	Flu
X_1	'High'	'?'	'no'	'yes'
X_2	'veryhigh'	'yes'	'yes'	'yes'
X_3	'?'	'no'	'no'	'no'
X_4	'High'	'yes'	'yes'	'yes'
X_5	'High'	'yes'	'?'	'no'
X_6	'Normal'	'?'	'?'	'no'
X_7	'Normal'	'?'	'yes'	'no'
X_8	'?'	'yes'	'yes'	'yes'
X_9	'veryhigh'	'?'	'?'	'yes'
X_{10}	'Normal'	'no'	'no'	'no'

3.3 Discernibility function for computing reduct and core

Main properties of discernibility functions are, they are monotonic and their prime implicants determine reduct uniquely [23]. It is also shown by example to determine all reduct for decision table by prime implicants of discernibility functions. It is also shown that rough set approach is suitable for reasoning in incomplete information system. So discernibility function may be used to compute reduct and core. In incomplete information system, attributes set A may also be divided into conditional attribute set and decision attribute set *(D)* with no intersection. In this paper for incomplete information, it has been considered missing attribute value exist only on conditional attribute, decision attributes are complete. According to that all definitions have been redefined. $(x_i, x_j) \in I(D)$, denotes objects x_i and x_j belongs to same decision class. $n\times n$ discernible matrix for incomplete decision table is denoted by $IM_D(B)$ where $B \subseteq A$ and defined as :

$$IC_{ij} = \{a \in B: a(x_i) \neq ?,\ a(x_j) \neq ?,\ a(x_i) \neq a(x_j)\ and\ (x_i, x_j) \notin I(D)\} \quad for\ i,j = 1,2,..n. \qquad (6)$$

i.e., IC_{ij} is all attributes set which possibly discern object x_i and x_j.

Set of all single element entries of *IM(B)* is the possible core for incomplete decision system and defined by

$$Core(B) = \{a \in B, IC_{ij} = \{a\}, for\ some\ i,j\} \qquad (7)$$

Discernibility function (If_D) for incomplete decision system is a boolean function of m boolean variables $a_1^*,\ldots,a_m^*$ (corresponding to the attributes $a_1,\ldots,a_m$) computed from discernibility matrix may defined as:

$$f_D(a_1^*,\ldots,a_m^*) = \wedge\{\vee\ IC_{ij}^* | 1 \leq j \leq i \leq |U|,\ IC_{ij} \neq \emptyset\} \qquad (8)$$

where $IC_{ij}^* = \{a^* | a \in IC_{ij}\}$.

All possible minimal reduct of a system may be found by simplification of discernibility function using boolean laws of boolean algebra. In the same way possible Core may also be found by intersections of all reduct.

For application of the above definitions, example of incomplete decision table of table 1 has been chosen first. After that these concept of rough set approach to incomplete information, have been applied to real life data sets to impute missing data with proposed algorithm. Using these definitions of discernibility matrix for incomplete data set of Table 1 has been shown in table Table 2.

The computation of discernibility function for incomplete information, from table 2 may be denoted as:

$$T(T+C)(H+C)(T+H+C).$$

Using boolean algebra, simplification form of discernibility function is

$$TH+TC$$

So there are two reduct *TH* and *TC*. *{T}* is the core.

Table 2. Discernibility matrix

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1						T	TC			T
X_2			HC		T	T	T			THC
X_3		HC		HC				HC		
X_4			HC			T	T			THC
X_5		T							T	
X_6	T	T		T					T	
X_7	TC	T		T					T	
X_8			HC							HC
X_9					T	T	T			T
X_{10}	T	THC		THC				HC	T	

4. Computational modelling

Now a model will be proposed to impute missing data based on rough set approach to incomplete information. Core attributes are main or essential feature of a data set and reduct attributes are required feature of a data set. Other attributes may be ignored. Considering this idea, core attributes have to give importance than others to impute missing data. Other reduct attributes have to consider also.

4.1 Significance relation

In incomplete information system *(U, A, V', f)*, if $x_i \in U$ then the missing set of attributes w.r.t. object 'x_i ' may be defined as:

$$MA_i = \{m; a_m(x_i) = ?, j = 1,2..m\} \qquad (9)$$

and missing object set by:

$$MO = \{i;\ MA_i \neq \emptyset, i = 1,2..n\} \qquad (10)$$

Significance relation $S_m(i,j)$ can predict similarity of x_i and x_j with respect to attribute a_k. Core attributes significance (by value 3) have been given much more importance than reduct attribute (by value 1) to consider similarity. Again any missing attribute values insignificance has been consider (by value -1). Other unimportant or extranious attributes value have been neglected (by value 0) in significance relation. Values 3, 2, 1,-1 have been used just to calculate most suitable object to a missing object values. These values denote priority of attributes to impute missing data for prediction of most similar object.

Now using $S_m(i,j)$, priority significance relation $P(i,j)$ may be defined to predict most suitable object to fill missing attribute :

$$P(i,j) = \begin{cases} 0 & if\ \ MA_i \subseteq MA_j \vee (x_i, x_j) \notin I(D) \\ \sum_{a_{m\in A}} S_m(i,j) & otherwise \end{cases}$$

Where

$$S_m(i,j) = \begin{cases} 3 & if\ a_m(x_i) = a_m(x_j) \neq ?\ \wedge a_m\ core \\ 1 & if\ a_m(x_i) = a_m(x_j) \neq ?\ \wedge a_m\ reduct \\ -1 & if\ a_m(x_i) = a_m(x_j) = ? \\ 0 & otherwise \end{cases} \qquad (11)$$

Here computation of all reduct is not mandatory; reducts which are easily computable may be used. Previous knowledge of core and reduct attributes can be used. According to situation, knowledge regarding mandatory attributes (like temperature for flu) may be used as core and other essential attributes (like cough for flu) may be used as reduct.

If $j \in max\{P(i,j)\}$ then it can be said x_i is most similar with 'x_j' and attributes value of object 'x_j' may be used to fill missing values of 'x_i'.

4.2 Proposed CRB algorithm

So most similar object has been calculated based on core and reduct attributes values similarity. According to the proposed mathematical model, Core Reduct Based (CRB) algorithm has been proposed to impute missing data using most similar object. CRB algorithm is shown in Algorithm 1.

Algorithm 1: CRB algorithm

Input :{ incomplete information System *S*,
IIS= (U, A=C ∪ D, V', f)}

Output: {Information System IS= (U, A=C ∪ D, V', f) where $v_{ij} \in$ V' become completed if suitable object present}

Step 1. $compute\ MA_i\ and\ MO$;
Step 2. $compute\ P(i,j)$;
Step 3. for ($each\ object\ i \in MO$)
Step 4. $if(max_k(P(i,k)) > 0)$
Step5. $find\ j \leftarrow k\ for\ max_k(P(i,k))\ exist$;
Step 6. $for(each\ attribute\ m \in C$){
Step 7. $v'(i,m) = \begin{cases} \mathrm{v(j,m)} & if\ \mathrm{v(i,m)} = ? \\ \mathrm{v(i,m)} & if\ \mathrm{v(i,m)} \neq ? \end{cases}$
Step 8. end for step 6
Step 9. end if step 4
Step 10. end for step 3
Step 11. $stop$

5. Analysis of algorithm complexity

Let m is the number of attribute and *n* is the number of Object. To assign a value to a missing value worse case time complexity for step 1 and 3 to 7 is $O(m)*O(n)=O(m*n)$. Worse case time complexity for step 2 is $O(n^2)$. So worse case time complexity for the above algorithm is $O(n^2+ m*n)$. But depends on number of objects with missing value and computation of core and reduct, average case and best case time complexity is less than that. Clearly Space complexity is also $O(1)$.

6. Experimental Result

After application of the above algorithm to table 1, result has been shown in table 3. From table 3, it is clear that CRB algorithm can fill missing value if matching object values present otherwise left it as missing without misguiding it.

Three data sets IRIS, Hayes-Roth, and Blogger have been selected from UCI machine learning database [33] to test the performance of proposed CRB algorithm. Also CRB algorithms performance has been compared with popular and recent IRMI algorithm [21], mostly used mean-mode (MN) algorithm [9] and popular well known KNN based method [8] and C4.5 method [5]. There are no missing values in these data sets, so missing values have been generated with certain ratio on

Table 3. Information system after CRB algo

Cases	Temperature	Headache	Cough	Flu
X_1	'High'	'yes'	'no'	'yes'
X_2	'veryhigh'	'yes'	'yes'	'yes'
X_3	'Normal'	'no'	'no'	'no'
X_4	'High'	'yes'	'yes'	'yes'
X_5	'High'	'yes'	'?'	'no'
X_6	'Normal'	'no'	'no'	'no'
X_7	'Normal'	'no'	'yes'	'no'
X_8	'veryhigh'	'yes'	'yes'	'yes'
X_9	'veryhigh'	'yes'	'yes'	'yes'
X_{10}	'Normal'	'no'	'no'	'no'

conditional attributes. Data sets with same missing value and same percentage of missing have been used in all algorithms for fair comparison. Discretization method with equal-width binning has been used on IRIS data set before application of the algorithm. Ten-fold cross validation with k-nearest neighbours classifier has been used to measure performance of methods by computing accuracy (Ac), kappa statistic (Ka), mean absolute error (Ma) and root mean squared error (Rt). Brute force search algorithm has been utilized for nearest neighbour search. Euclidean distance and *k*'s value one (1) have been used, as only one instances affinity with classes taken into account.

Result of experiments for CRB and MN methods are presented in Table 4. Result of C4.5 and KNN methods are presented in Table 5 and IRMI algorithm's result described in Table 6. Experiment shows proposed CRB algorithm utilizes equivalent data set considering attribute significance more precisely to fill missing data. These tables' data shows that for every data set CRB algorithms accuracy is better than other methods. Kappa statistics is an important measure on classifier performance for prediction. Kappa statistics value of CRB algorithm itself describes its substantial or almost perfect classification ability. Comparison of kappa statistics value precisely describes its reliability over other methods.

Mean absolute error and root mean squared error, measure magnitude of error in a set of predictions with negative oriented score, describe that CRB algorithms prediction has lower error rate than others method. The comparison result of accuracy among CRB, MN, KNN, C4.5 and IRMI methods are shown in Fig.1. Fig.1 show that proposed CRB algorithms accuracy is better than other algorithms. So by considering all evaluation parameter, it can be concluded that CRB algorithms predictions is almost perfect and better than others. So it may be used as an imputation method.

Table 4. Experimental result of CRB and MN algorithm

	5% mis		10% mis		15% mis		20% mis		25% mis		30% mis		35% mis	
Iris Data Set														
	CRB	**MN**	**CRB**	**MN**	**CRB**	**MN**	**CRB**	**MN**	**CRB**	**MN**	**CRB**	**MN**	**CRB**	**MN**
Ac	92%	90%	92%	87%	87%	86%	95%	85%	93%	79%	96%	86%	97%	77%
Ka	0.88	0.86	0.88	0.81	0.79	0.80	0.93	0.77	0.89	0.69	0.94	0.79	0.95	0.66
Ma	0.06	0.09	0.06	0.10	0.12	0.08	0.04	0.12	0.06	0.15	0.04	0.12	0.03	0.19
Rt	0.20	0.23	0.20	0.26	0.26	0.28	0.16	0.28	0.18	0.31	0.14	0.27	0.12	0.35
Hayes-Roth Data Set														
Ac	66%	64%	75%	58%	72%	59%	78%	47%	84%	61%	75%	51%	86%	44%
Ka	0.45	0.42	0.61	0.33	0.56	0.35	0.65	0.16	0.75	0.39	0.61	0.21	0.79	0.10
Ma	0.20	0.25	0.16	0.28	0.16	0.30	0.17	0.37	0.14	0.31	0.18	0.34	0.13	0.34
Rt	0.35	0.40	0.32	0.43	0.32	0.45	0.65	0.50	0.30	0.45	0.34	0.48	0.30	0.51
Blogger Data Set														
Ac	85%	82%	86%	77%	87%	83%	84%	73%	89%	73%	92%	75%	83%	67%
Ka	0.61	0.54	0.64	0.39	0.67	0.57	0.59	0.26	0.74	0.25	0.80	0.32	0.55	0.08
Ma	0.21	0.25	0.19	0.27	0.20	0.22	0.19	0.36	0.15	0.34	0.12	0.31	0.20	0.40
Rt	0.37	0.41	0.34	0.41	0.34	0.37	0.35	0.47	0.32	0.46	0.26	0.45	0.34	0.52

Table 5. Experimental result of C4.5 and KNN algorithm

	5% mis		10% mis		15% mis		20% mis		25% mis		30% mis		35% mis	
Iris Data Set														
	C4.5	**KNN**	**C4.5**	**KNN**	**C4.5**	**KNN**	**C4.5**	**KNN**	**C4.5**	**KNN**	**C4.5**	**KNN**	**C4.5**	**KNN**
Ac	90%	91%	91%	89%	87%	89%	91%	89%	85%	84%	85%	90%	79%	87%
Ka	0.85	0. 87	0.86	0.84	0.81	0.84	0.87	0.84	0.78	0.76	0.78	0.85	0.69	0.80
Ma	0.10	0.07	0.10	0.08	0.12	0.09	0.11	0.09	0.19	0.12	0.17	0.09	0.21	0.12
Rt	0.21	0.21	0.21	0.22	0.22	0.24	0.21	0.24	0.27	0.28	0.26	0.23	0.29	0.27
Hayes-Roth Data Set														
Ac	64%	61%	67%	63%	61%	58%	61%	59%	66%	61%	55%	50%	46%	52%
Ka	0.43	0.38	0.48	0.41	0.38	0.32	0.38	0.35	0.47	0.38	0.27	0.20	0.14	0.24
Ma	0.26	0.23	0.26	0.25	0.31	0.29	0.33	0.32	0.31	0.30	0.38	0.33	0.39	0.33
Rt	0.39	0.38	0.36	0.40	0.39	0.43	0.41	0.46	0.37	0.45	0.43	0.46	0.44	0.45
Blogger Data Set														
Ac	77%	84%	73%	78%	70%	83%	67%	73%	67%	73%	70%	76%	68%	68%
Ka	0.44	0.60	0.28	0.44	0.17	0.57	0.00	0.28	0.02	0.30	0.16	0.35	0.00	0.08
Ma	0.32	0.21	0.38	0.26	0.37	0.23	0.44	0.32	0.44	0.30	0.41	0.31	0.44	0.38
Rt	0.42	0.38	0.45	0.40	0.45	0.37	0.47	0.47	0.47	0.48	0.46	0.45	0.47	0.50

Table 6. Experimental result of IRMI Algorithm

IRMI Algo, Misssing value % →							
	5%	10%	15%	20%	25%	30%	35%
Iris Data Set							
Ac	91%	87%	91%	87%	85%	87%	79%
Ka	0.87	0.80	0.86	0.81	0.78	0.80	0.68
Ma	0.07	0.09	0.09	0.10	0.11	0.10	0.14
Rt	0.20	0.26	0.24	0.26	0.27	0.29	0.33
Hayes-Roth Data Set							
Ac	62%	55%	49%	45%	60%	51%	55%
Ka	0.39	0.29	0.19	0.14	0.37	0.21	0.30
Ma	0.25	0.29	0.31	0.34	0.29	0.36	0.34
Rt	0.39	0.43	0.43	0.47	0.43	0.47	0.48
Blogger Data Set							
Ac	85%	80%	77%	65%	68%	72%	73%
Ka	0.61	0.46	0.42	0.03	0.10	0.24	0.26
Ma	0.21	0.27	0.28	0.41	0.38	0.33	0.37
Rt	0.37	0.41	0.41	0.51	0.48	0.46	0.48

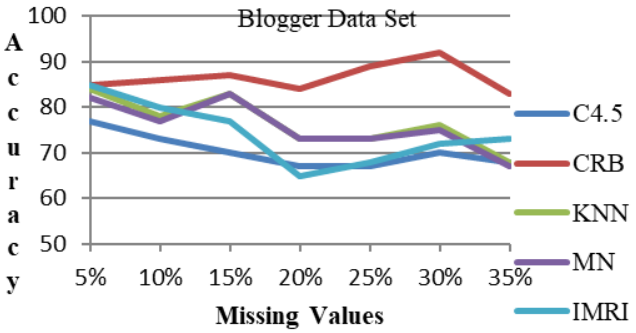

Figure.1 Comparison graph for accuracy vs missing data

7. Conclusions

Core and reduct attributes of Rough set, main part of a decision table used in this paper for incomplete data set. Rough set approach has been used for incomplete data set. Computation of core and reduct using discernibility matrix has been shown for incomplete data set and it is used, to impute missing data. For imputation based pre-processing approach, always it is better to fill only those missing data which are supported by available object information. This concept has been used and missing data without similar object kept as missing. So after application of CRB algorithm there is no chance to generate misleading information. Utilization of Core and reduct attributes in CRB algorithm enhance the efficiency of filling missing data by considering most suitable object. Proposed algorithm may be used as preprocessing tool for missing data. Discernible matrix has been used to compute core and reduct, so above algorithm better suited for small and medium size data set. But it may be used for large data set if we compute core and reduct using other method. Previous knowledge of core and reduct attributes may be used in CRB algorithm for better and quick result. This work may be enhanced for joint applications of imputation-feature reduction method for achieve more suitable data for data mining in less time.

References

[1] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley and Sons, 1987.

[2] P. K. Dey and S. Mukhopadhyay, "Modified Deviation Approach to Deal with Missing Attribute values in Data Mining with Different percentage of Missing Values", *International Journal of Computer Applications,* Vol.73, No. 5, pp. 1-6, 2013

[3] A. C. Acock, "Working with Missing Values", *Journal of* Marriage and Family, Vol.67, No.4, pp. 1012-1028, 2005.

[4] J. W. Grzymala-Busse and H. Ming, "A Comparison of Several Approaches to Missing Attribute Values in Data Mining", *Rough Sets and Current Trends in Computing, Springer, Lecture Notes in Computer Science,* pp.378-385, 2001

[5] J. R. Quinlan, *C4.5: Programs for Machine Learning,* Morgan Kaufmann Publishers, 1993.

[6] J. W. Grzymala-Busse and A. Y. Wang, "Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values", In: *Proc. of Fifth international Workshop on Rough Sets and Soft Computing (RSSC'97), Third Joint Conference on Information Sciences (JCIS'97),* pp.69-72, 1997.

[7] J. W. Grzymala-Busse, "Data with missing attribute values: Generalization of idiscernibility relation and rule induction", *Transactions on Rough Sets, Lecture Notes in Computer Science Journal Subline, Springer-Verlag*, vol. 1 , pp. 78–95, 2004.

[8] D. Aha, D. Kibler and M. Albert, "Instance-based learning algorithms", *Machine Learning*, Vol.6, pp. 37-66, 1991.

[9] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.

[10] D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, 1999.

[11] P. Clark and T. Niblett, "The CN2 induction algorithm", *Machine Learning*, Vol.3, 1989.

[12] I. Knonenko, I. Bratko and E. Roskar, *Experiments in automatic learning of medical diagnostic rules*, Technical Report, Jozef Stefan Institute, Lljubljana, Yugoslavia, 1984.

[13] J. W. Grzymala-Busse, "On the unknown attribute values in learning from examples", *Lecture Notes in Artificial Intelligence, Springer-verlag*, vol.542, pp.368-377, 1991.

[14] W. Young, G. Weckman and W. Holland, "A survey of methodologies for the treatment of missing values within datasets: limitations and benefits", *Theoretical Issues in Ergonomics Science*, Vol.12, pp.15-43, 2011.

[15] P. K. Dey and S. Mukhopadhyay, "Modified Deviation Approach to Deal with Missing Attribute values in Data Mining with Different percentage of Missing Values", *International Journal of Computer Applications,* Vol.73, No. 5, pp. 1-6, 2013.

[16] J. V. Hulse and T. M. Khoshgoftaar, "Incomplete-case nearest neighbor imputation in software measurement data", *Information Sciences*, Vol. 259, No. 2, pp. 596-610, 2014.

[17] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response", *International Statistical Review*, Vol.78, No.1, pp. 40-64, 2010.

[18] A. Dempster, N. Laired and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, Vol.39, No.1, pp. 1-38, 1977.

[19] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.

[20] J. L. Schafer, "Multiple imputation: a primer", *Statistical Methods in Medical Research,* Vol. 8, pp. 3–15, 1999.

[21] M. Templ, A. Kowarik and P. Filzmoser, "Iterative stepwise regression imputation using standard and robust methods", *Computational Statistics & Data Analysis, Elsevier*, Vol. 55, No. 10, pp. 2793-2806, 2011.

[22] W. Zhou, W. Zhang and Y. Fu , "An incomplete data analysis approach using rough set theory", *Intelligent Mechatronics and Automation, IEEE*, pp.332-338, 2004.

[23] M. Kryszkiewicz, "Rough set approach to incomplete information systems", *Information Sciences, Elsevier,* Vol. 112, pp.39-49, 1998.

[24] J. Stefanowski and A. Tsoukias, "On the Extension of Rough sets Under Incomplete Information",In: *Proc. of the 7th Int'l workshop on New Directions in Rough Sets, Data Mining and Granular Soft Computing, Berlin:Spinger-Verlag,* pp.73-81,1999.

[25] Z. Zhang, R. Li and Z. Li, "An Incomplete Data Analysis Approach Based on the Rough set Theory and Divide and Conquer Idea", In: *Proc. of Fourth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE,* 2007.

[26] J. W. Grzymala-Busse, "A Rough Set Strategies to Data with Missing Attribute Values", In: *Proc. Of the Workshop on Foundations and New Directions in Data* Mining, *associated with the third, IEEE International Conference on Data*

Mining, November 19–22, Melbourne, FL, USA, pp. 56–63, 2003.

[27] J. W. Grzymala-Busse, "Three Approaches to Missing Attribute Values— A Rough Set Perspective", In: *Proc. of the Workshop on Foundations of Data Mining, associated with the fourth IEEE International Conference on Data Mining, Brighton, UK*, November 1–4, 2004.

[28] J. W. Grzymala-Busse and W. Rzasa, "Local and Global Approximations for Incomplete Data", In: *Proc. of the Transactions on Rough Sets VIII, LNCS 5084, Springer-Verlag Berlin Heidelberg*, pp. 21–34, 2008.

[29] J. W. Grzymala-Busse and S. Siddhaye, "Rough Set Approaches to Rule Induction from Incomplete Data", In: *Proceedings of the IPMU'2004, the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy*, July 4–9, vol. 2, 923–930, 2004.

[30] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods", *Neurocomputing, Elsevier*, Vol.205, pp. 152-164, 2016.

[31] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht, Vol.9, 1991.

[32] R. Jensen and Q. Shen, "New Approach to Fuzzy-Rough Feature Selection", *IEEE Transactions on Fuzzy Systems*, Vol.17, No.4, pp. 824-838, 2009.

[33] http://archive.ics.uci.edu/ml/datasets.html.