A Comparative Study of Handling Missing Data in Student Data Analysis using Rough Set and Soft Set

B. S. Panda*
Research Scholar, CUTM, Odisha
bspanda.cse@gmail.com

S. S. Gantayat
GMRInstitute of Technology, Rajam, AP.
sasankosekhar.g@gmrit.org

Ashok Misra
CUTM, Odisha.
ashok.misra@cutm.ac.in

Abstract - **When analyze the student data some of the data may seem to be missing. By using a number of techniques such missing data can be brought to line. Among the methods which are used to handle this issue to recover the missing values some popular methods can be adopted, such as CDC (Complete Delete Case), LOCF (Last Observation Carried Forward), RST (Rough Set Technique) and (SS) Soft Set. When data is analyzed in reality mainly three types of missing data occurs depending on the reasons of missingness. They can be classified as MCAR, MAR and MNAR. In order to evaluate the performance of this three usual missing data cases number of similar experiments under a variety of situations where conducted to bring the responsible factors to the surface in this paper. It was noted that the CDC method cannot be used in the data analysis of educational trials. The methods of LOCF and RST are successfully performed under the missing mechanism of MCAR. The methods which can be reliably and comfortably grounded under the MAR missing mechanism are the RST and SS methods.**

Keywords: MCAR, MAR, Complete Delete Case, Mean Substitution, LOCF, Rough Set, Soft Set

I. INTRODUCTION

While conducting educational study, observations are repeatedly measure in the same subject and at each schedule visit. The educational trials may sometimes face the problems of missing observations and resultant information. The issue of missing data arises from bringing together of class test marks, percentage of the student, assignments, attendance, and grading of the students. It may further lead to serious problems like reducing the statistical power and the inability to rule out biases in the estimates. The problem is so serious that an immediate and complete solution can not be reached in statistical practice. The traditional statistical methods are marred by intentional faulty designs. Special attention should be directed to conduct the incomplete data details.

II.BACKGROUND

The details of the various methods to handle missing data are discussed in a conference paper by Panda et.al.[1]. Some of the summarized methods that are used in this paper are discussed below. Due to various reasons the missing data mechanism is classified in to three different types [2].

if the messiness is not cover under the ambit of any observed and unobserved factors complete data missing may happen at certain times and certain circumstances this factor is called MCAR (Missing Complete at Random).

1. If there are conditions missingness due to lack of efficacy this missing case is called MAR (Missing at Random).
2. If the unobserved quantity cause the missingness along with some observed factors it is called MNAR (Missing Not at Random)

The MNAR missing technique can be un-erroneously utilized for student results. Between the MCAR technique and the MAR technique the later is accepted by researchers as the more effective and popular one. However if one goes by definition MCAR is a suitable alternative to MAR. In actuality the every MCAR [3] missing format can be used as an alternative from MAR but not vice-versa. In actually the MCAR exemption can be systematically tested against its alternative hypothesis not MCAR [4]. Every time the MAR or MNAR format is used additional information has to be added without which the test will not be possible. The sensitivity of results cannot be usefully accessed by MNAR without MAR [5]. Therefore MAR [6] has highly to be incorporated in to the analysis.

In the paper a number of approaches have been applied in the statistical analysis of educational data with missing values. These approached should be applied basing on appropriate methods along with the data missing mechanism. The applications of different statistical approaches are valid only under selected situations (missing mechanism) in terms of specified missing rates to put it the other way. None of the methods is unique to be used for all situations. In the analysis of educational study it is not easy to test the missing mechanism as there is no clear rule to specify as to how much available missing data is in access to the required study. If the study is directed in its pragmatic and a more explanatory dimensions a particular method of handling missing data is adopted. Too much of missing data [7] can stand as a problem for the study.

Sprint and Dupin-Sprint [8] have prescribed an amount of missing data which can be tolerable excepted and it would not lead to an effect in the opposite direction from the performance of the worst case analysis of missing data it can be determent whether the right level of missing data has been reached.

as well as some observed factors. The MNAR missing mechanism is usually used to describe the students result. Researchers have pointed out that the MAR assumption may be more in practice than that of the MCAR. In fact, by definition MCAR is only a special case of MAR. In other words, a MCAR missing mechanism is also a MAR one, but not every MAR is a MCAR. Actually, it is possible to formally test the MCAR assumption against its alternative hypothesis not MCAR. However, it is not possible to test MAR or MNAR without using additional (external) information. MNAR is particularly useful in assessing the sensitivity of the results that are not MAR and it is highly recommended to be incorporated into the analysis.

In the literature, several alternative statistical approaches have been applied to the analysis of educational data with missing values. These appropriate methods for analysis should be selected based on the data missing mechanism, since different statistical methods are valid only under certain situations (missing mechanisms) with specified missing rates. In other words, there is no unique best method available for all situations. However, it is difficult to test the missing mechanism in a educational study and there is also no clear rules regarding how much is qualified as too much missing data[7]. In general, the choice of a particular method for handling missing data depends largely on whether one is considering a more pragmatic or a more explanatory perspective. There is often the question of whether there are too many missing data. Sprint and Dupin-Sprint[8] pointed out that the tolerable amount of missing data is that would not conceal an effect in the opposite direction. In order to determine whether this level of missing data has been reached, one can perform what was called the “worst case” analysis.

Despite these difficulties, several researchers have considered and constructed simulation studies for the proof of strong consistency of imputation methods to check the efficiency of the imputation methods. For example, Myers [9] compared the results of two imputation methods (that is, the complete case method and the multiple imputation method) based on simulated data sets with a dropout rate ranging from 20% to 60%, and they concluded that MI method provided results that are more closely mimicked the complete data set.

Hening and Koonce [10] investigated five imputation methods (i.e., mean substitution, median substitution, zero value, hot-deck, and MI) and a first-year-student retention data with more than 20% missing values is used. The results shown that multiple and hot-deck imputations perform poorly in an accuracy comparison test, but they can slightly increase the predication accuracy rate compared with other methods.

Ali, et al. [11] performed a survival analysis in which missing data were simulated under MCAR and MAR to compare four imputation methods—complete case analysis (CCA), means substitution (MS), and multiple imputation (MI)with the inclusion of the outcome (MI− and MI+). The simulation results suggested that in general MI+ is likely to be the best method. Patrician [12] pointed out that MI is the best approach and should be considered to handle missing data compared with CCA and MS by an empirical investigation of AIDS care longitudinal data outcomes.

Recently, Nakai, et al. [13] have shown that MI is the most effective imputation method in longitudinal data setting under MCAR via a simulation study. This indeed provides useful information about the performance of imputation methods under MCAR, but it is limited and restricted to clinical situations where MAR is more plausible. For example, Lavori, et al.[14] have pointed out that the MCAR assumption is often not plausible in most clinical trial settings. The purpose of this paper is through a simulation approach to analytically evaluate the performance of four imputation methods for different missing mechanisms (MCAR and MAR) with various missing rates.

For simplicity and also without loss of generality, a monotone pattern of missing data (meaning that once a patient has a missing response at an assessment visit, his or her data will be missing for all subsequent visits) is assumed. Under such assumptions, this paper primarily concentrates on the following four imputation methods:

Complete delete case (CDC), Last observation carried forward (LOCF), Rough Set Techniques (RST) and Soft Set (SS).

To compare the performance of these methods, RST and Soft Sets are used as evaluation criteria.

Table-1: Student information system with missing values

Student No.	Class Test 1	Class Test 2	Assignment 1	Assignment 2	Lab test	Result
S01	First	Second	?	First	First	First
S02	Second	First	Second	First	Third	Second
S03	Third	First	Third	Third	Third	Third
S04	First	?	First	First	Second	First
S05	Second	Second	Second	Second	Second	Second
S06	Second	Second	?	Third	?	Second
S07	Third	?	Third	Second	Third	Third
S08	Second	Second	Second	First	First	Second
S09	Second	Third	Second	Second	Second	Second
S10	First	Second	First	?	First	First
S11	Second	?	Second	Third	Third	Second
S12	Second	Second	Third	?	Second	Second

III. APPROACHES TO HANDLING MISSING DATA

There are so many techniques in handling missing data discussed in the literature. Especially, many methods have been proposed and developed to handle missing data in longitudinal clinical trials. However, there are few methods that are actually used in real trials with missing data. The purpose of this paper is to study four most frequently used methods for dealing with missing data and they will be described as follows.

A. *Rough Set Analysis*

Rough set theory introduced by Pawlak in 1982 is a mathematical tool to deal with vagueness and uncertainty of information [15]. It has been proved to be very effective in many practical applications. The rough set uses the basic relation operations known as Equivalence relation, which is reflexive, symmetric and transitive. Using this relation the data table is classified according to the attributes and this classification is analyzed using core and reduct of the relation.

S = (R, X, Y) is independent if all are indispensable in $x \in X$.

The set of attributes is called a reduct of X, if S' = (R, Z, Y) is independent and

$$POS_Z(Y) = POS_X(Y)$$

The set of all the condition attributes indispensable in S is denoted by CORE(X).

$$CORE(X) = \cap RED(X)$$

Where RED(X) is the set of all reducts of X

T = (U, C, D) is independent if all are indispensable in $c \in C$.

The set of attributes is called a reduct of C, if T' = (U, R, D) is independent and

$$POS_R(D) = POS_C(D)$$

The set of all the condition attributes indispensable in T is denoted by CORE(C).

$$CORE(C) = \cap RED(C)$$

Where RED(C) is the set of all reducts of C

However, in rough set theory, the deterministic mechanism for the description of error is very simple [16]. Therefore, the rules generated by rough sets are often unstable and have low classif cation accuracy. Rough set methods can greatly accelerate the network training time and improve its prediction accuracy. In [17] Rough set method was also applied to generating rules from trained neural networks. In these hybrid systems, rough sets were used only as a tool to speed up or simplify the process of mining knowledge from the databases and impute the missing values in the database.

For example, in [18], a rule set, a part of knowledge, is f rst generated from a database by rough sets. In the prediction phase, a new object is f rst predicted by the rule set, if it does not match any of the rules, the model can get high classif cation accuracy. In this paper, from a new perspective we develop a hybrid system of rough sets and neural networks to mine classif cation rules from large databases. Compared with previous research works our study has the following contributions:

- *Basic Concepts of Rough Sets*

An information system is a 10-tuple $S = \{U,A,V,F \ldots\}$ where U is a finite set of objects, called the universe, A is a finite set of attributes, $V= U_{a\in A} V_a$ is a domain of attribute a and $f : U \times A \rightarrow V$ is called an information function such that $f=(x,a)\in V_a$ *for* $\forall_a \in A, \ \forall_x \in U$.

In the classification problems, an information system is also seen as a decision table assuming that A = $C\cup D$ and $C\cap D=\phi$ where C is a set of condition attributes and D is a set of decision attributes.

Let $S= \{U,A,V,F \ldots\}$ be an information system, every P $\subseteq$ A generates a indiscernibility relation *IND(P)* on U, which is defined as follows:

$$IND(P) = \{(x,y) \in U\times U: f(x,a)=f(y,a) \ \forall a\in P\}$$

- *Attribute Rule Generation by Rough Sets*

Attribute rule generation (feature selection) is a process of finding an optimal subset of all attributes according to some criterion so that the attribute subset is good enough to represent the classification relation of data. A good choice of attribute subset provided to a classifier can increase its accuracy, save the computational time, and simplify its results [19].

In general, rough set theory provides useful techniques to reduce irrelevant and redundant attributes from a large database with a lot of attributes [20] and [21]. However, it is not so satisfactory for the reduction of noisy attributes because the classification region defined by rough set theory is relatively simple and rough set based attribute rule generation criteria lack effective validation method.

There are different approaches to generate rules, direct and indirect methods. Direct methods generate rule from training data like sequential covering algorithms. Indirect methods build the classification model from which rule are extracted, e.g. decision tree, Neural Network, Genetic Algorithms etc.

From the Table-1 the missing attribute classes can be generated using rough set concept as follows.

R1: Class Test-1= {{S01,S04,S10},{S02,S05,S06,S08,S09,S11,S12},{S03,S07}}

*R2: Class Test-2={{S02,S03},{S01,S05,S06,S08,S10,S12},{S09},{**S04,S07,S11**}}*

*R3: Assignment-1={{S04,S10},{S02,S05,S08,S09,S11},{S03,S07,S12},{**S01,S06**}}*

*R4: Assigment-2={{S01,S02,S04,S08},{S05,S07,S09 },{S03,S06, S11},{**S10,S12**}}*

*R5: Lab Test={{S01,S08,S10},{S4,S05,S09,S12},{S02,S03,S07,S11},{**S6**}}*

In the above classification the classes are First, Second, Third and the bold classes are the missing data, which will be filled in the common attribute rough set technique.

The bold classes are replacement of approximated information with the missing data in the Table-2. After replacing proper common values to the missing data, we get the different classes from the Table-1 as follows.

Table-2: Student information system with RST analysis

Student No.	Class Test 1	Class Test 2	Assignment 1	Assignment 2	Lab test	Result
S01	First	Second	**Second**	First	First	First
S02	Second	First	Second	First	Third	Second
S03	Third	First	Third	Third	Third	Third
S04	First	**Second**	First	First	Second	First
S05	Second	Second	Second	Second	Second	Second
S06	Second	Second	**Second**	Third	**Second**	Second
S07	Third	**Second**	Third	Second	Third	Third
S08	Second	Second	Second	First	First	Second
S09	Second	Third	Second	Second	Second	Second
S10	First	Second	First	**First**	First	First
S11	Second	**Second**	Second	Third	Third	Second
S12	Second	Second	Third	**First**	Second	Second

B. Soft Set (SS)

The models such as theory of probability, Interval mathematics, fuzzy set theoryIntuitionistic fuzzy set theory are inadequate to handle some uncertainty problems. D.A.Moldtsov [22] and Majji et.al [23], noticed that, the problem might be due to inadequacy of parameterization tools in those models. So, Molodtsov initiated the concept of Soft Set theory which is a fusion of the notions of topology and set theory as a new mathematical tool to deal with uncertainties and free from some previous difficulties. Soft Sets can be called as (Binary, Basic, and Elementary) neighborhood systems. Soft set gives a general mathematical tool to deal with uncertain, fuzzy, not clearly defined (vague) objects.

Let ***U*** be an initial universal set and let ***E*** be a set of parameters. A pair *(F, E)* is called a *soft set* (over U) iff *F* is a mapping of *E* into the set U. The pair *(U, E)* is often regarded as a soft universe. Members of the universe and the parameter set are generally denoted by x and e respectively. Let A be the subset of E. A soft set over the soft universe (*U,E)* is denoted by *(F, A)*, where $F{:}A \rightarrow P(U)$. In other words, the soft set is a parameterized family of subsets of the set U. Every set *F(e)*, $e \in E$, from this family may be considered as the set of *e*-approximate elements of the soft set. The Sets of *F(e)*may be arbitrary. Some of them may be empty, some may have non-empty intersection.

Analysis of Missing Data for Student Data

In the soft set study analysis, the missing data values are considered to impute the corresponding missing values. A count is added at the end of the table to count the number of missing values for each tuple. After that, the count is sorted in descending order. A threshold value for the count is considered to eliminate non-missing valued tuples.

From the below Table-3, count=0 is eliminated as these tuples are not containing any missing values. Minimum threshold>=1 for missing data.

Table-3: Student Information System with Soft Set Analysis

Student No.	**Class Test 1**	**Class Test 2**	**Assignment 1**	**Assignment 2**	**Lab test**	**Result**	**Count for Missing Attributes**
S06	Second	Second	?	Third	?	Second	2
S01	First	Second	?	First	First	First	1
S04	First	?	First	First	Second	First	1
S07	Third	?	Third	Second	Third	Third	1
S10	First	Second	First	?	First	First	1
S11	Second	?	Second	Third	Third	Second	1
S12	Second	Second	Third	?	Second	Second	1

Case 1: The no. of missing data =1Classify according to attributes,

Class Test1= {{1,4,10},{6,11,12},{7}} Result={{1,4,10},{6, 11,12},{7}}

Lab Test={{**6**},{1,10},{4,12}, {7,11}}

Max possibility for S06, is in Result and Class Test 1= {6, 11,12}. So the missing value is **Second**.

Case 2: The no. of missing data =2Classify according to attributes,

Class Test1= {{1,4,10},{6,11,12},{7}} Result={{1,4,10},{6, 11,12},{7}}

Assignment 1= {{**1,6**},{4,10},{7,12},{11}

Max possibility for S01, is in Result and Class Test 1= {1,4,10}. So the missing value is **First**

Max possibility for S06, is in Result and Class Test 1= {6,11,12}. So the missing value is **Second**

Assignment 2= {{**10,12**},{1,4},{6,11},{7}}

Max possibility for S10, is in Result and Class Test 1= {1,4,10}. So the missing value is **First**

Max possibility for S12, is in Result and Class Test 1= {6,11,12}. So the missing value is **Second**

Case 3: The no. of missing data =3Classify according to attributes,

Class Test1= {{1,4,10},{6,11,12},{7}} Result={{1,4,10},{6, 11,12},{7}}

Class Test 2={{**4, 7,11**},{1,6,10,12}}

Max possibility for S04, is in Result and Class Test 1= {1,4,10}. So the missing value is **First.**

Max possibility for S07, is in Result and Class Test 1= {7}. So the missing value is **Third.**

Max possibility for S11, is in Result and Class Test 1= {6,11,12}.So the missing value is **Second.**

Result: From the above analysis using Rough Set and Soft Set, it is found that there is no certainty to impute the correct missing value and we have use our own intuitionistic knowledge to select the value. So it is necessary to use some other techniques to replace the missing data with an appropriate value. Some of these techniques are discussed in the following sections.

Table-4: Analysis using Soft Set

Student No.	Class Test 1	Class Test 2	Assignment 1	Assignment 2	Lab test	Result
S01	First	Second	**First**	First	First	First
S02	Second	First	Second	First	Third	Second
S03	Third	First	Third	Third	Third	Third
S04	First	**First**	First	First	Second	First
S05	Second	Second	Second	Second	Second	Second
S06	Second	Second	**Second**	Third	**Second**	Second
S07	Third	**Third**	Third	Second	Third	Third
S08	Second	Second	Second	First	First	Second
S09	Second	Third	Second	Second	Second	Second
S10	First	Second	First	**First**	First	First
S11	Second	**Second**	Second	Third	Third	Second
S12	Second	Second	Third	**Second**	Second	Second

C. Complete Delete Case (CDC) Analysis

This method deletes all cases with missing data and then performs statistical analyses on the remaining complete data set (which has a smaller sample size). Since all cases containing missing data have been removed, there is no missing data problem to handle. Therefore, all statistical methods can be used to analyze the smaller data set. Obviously, one major advantage of this method is its ease of use. In fact, virtually all statistical programs incorporate this method as a default method because it accommodates any type of statistical analysis [24]. The method may be preferred under the situation in which the sample size is large, the proportion of missing data is small, and the missing data mechanism is MCAR [25]. For MCAR missing data, the method will yield unbiased parameter estimates and larger standard errors due to the smaller sample size. However, even when data are MCAR, loss of data will result in loss of precision (larger standard errors), particularly in multivariate data analyses.

In general, the major disadvantage of the method is that it could possibly lead to losing statistical power due to the reduction of the sample size [26].

Table-5: Student information system with CDC analysis

Student No.	Class Test 1	Class Test 2	Assignment 1	Assignment 2	Lab test	Result
S02	Second	First	Second	First	Third	Second
S03	Third	First	Third	Third	Third	Third
S05	Second	Second	Second	Second	Second	Second
S08	Second	Second	Second	First	First	Second
S09	Second	Third	Second	Second	Second	Second

D. Last Observation Carried Forward (LOCF)

The simplest imputation approach is the LOCF method that replaces every missing value with its corresponding last observed value. LOCF method is often used in longitudinal studies of continuous outcomes under MCAR. Conceptually, this method assumes that the outcome would not change after the last observed value. Therefore, there is no time effect since the last observed data. In fact, LOCF has been a popular method that is frequently used in handling missing data problems because it is easy to understand and can be implemented easily as well. Also, unlike the CDC method, the sample size does not change. For example, in an educational trial (see the data below), the bold words are newly imputed in missing attributes.

If there are more attributes missing then this method might give a biased conclusion about the effect of the student group. In our example, the measurement of student 1,4,6,7,10,11 and 12 are missing randomly. After impute the missing data the following table3 shows the bold letters.

Table-6: Student information system with LOCF analysis

Student No.	Class Test 1	Class Test 2	Assignment 1	Assignment 2	Lab test	Result
S01	First	Second	**Second**	First	First	First
S02	Second	First	Second	First	Third	Second
S03	Third	First	Third	Third	Third	Third
S04	First	**First**	First	First	Second	First
S05	Second	Second	Second	Second	Second	Second
S06	Second	Second	**Second**	Third	**Third**	Second
S07	Third	**Third**	Third	Second	Third	Third
S08	Second	Second	Second	First	First	Second
S09	Second	Third	Second	Second	Second	Second
S10	First	Second	First	**First**	First	First
S11	Second	**Second**	Second	Third	Third	Second
S12	Second	Second	Third	**Third**	Second	Second

Rigorously speaking, LOCF is not an analytic approach, but it is a method that is very easy to impute missing values. Analytic proofs [27] and studies in simulated data [28] and [29] have been clearly shown that LOCF can bias results and lead to either overestimation or underestimation of the parameter estimates.

IV. MISSING DATA GENERATION

After the original data sets were created, the measurements at different time points for different subjects were set to missing, according to the MCAR or MAR missing mechanism. However, the measurement at the first time point of each subject was assumed always observed. In the MCAR setting, missing data were generated randomly at visits 2 through 5 based on the missing probabilities listed in Table 1. Therefore, the missing probabilities do not depend on either observed or unobserved data. Furthermore, Little's MCAR test was performed to make sure the missing mechanism is indeed MCAR otherwise that data set was discarded and another data set was generated a new [30].

V. SIMULATION RESULTS

The simulation results are summarized in Table-6. In this table, Original data set, RST analysis, SS analysis and LOCF analysis are shown. In the following Comparison, the numbers 1,2 and 3 are represented as First, Second and Third respectively.

Table-7: Comparison of different Methods

Method	Student No.	Class Test 1	Class Test 2	Assignment 1	Assignment 2	Lab test	Result	Maching %
Original Data	S01	1	2	1	1	1	1	
	S04	1	1	1	1	2	1	
	S06	2	2	2	3	3	2	
	S07	3	2	3	2	3	3	
	S10	1	2	1	1	1	1	
	S11	2	3	2	3	3	2	
	S12	2	2	3	2	2	2	
Rough Set Analysis	S01	1	2	**2**	1	1	1	38%
	S04	1	**2**	1	1	2	1	
	S06	2	2	**2**	3	**2**	2	
	S07	3	**2**	3	2	3	3	
	S10	1	2	1	**1**	1	1	
	S11	2	**2**	2	3	3	2	
	S12	2	2	3	**1**	2	2	
Soft Set Analysis	S01	1	2	**1**	1	1	1	62%
	S04	1	**1**	1	1	2	1	
	S06	2	2	**2**	3	**2**	2	
	S07	3	**3**	3	2	3	3	
	S10	1	2	1	**1**	1	1	
	S11	2	**2**	2	3	3	2	
	S12	2	2	3	**2**	2	2	
LOCF Analysis	S01	1	2	**2**	1	1	1	50%
	S04	1	**1**	1	1	2	1	
	S06	2	2	**2**	3	**3**	2	
	S07	3	**3**	3	2	3	3	
	S10	1	2	1	**1**	1	1	
	S11	2	**2**	2	3	3	2	
	S12	2	2	3	**3**	2	2	
CDC Analysis	**No Results came as it is removing all the tuples with missing values retaining the non-missing tuples.**							0%

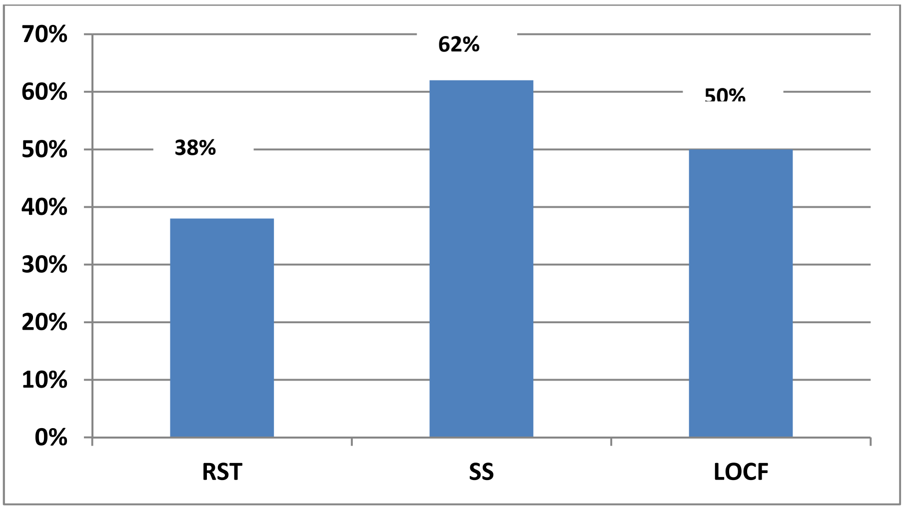

Fig-1. Comparison of different Methods

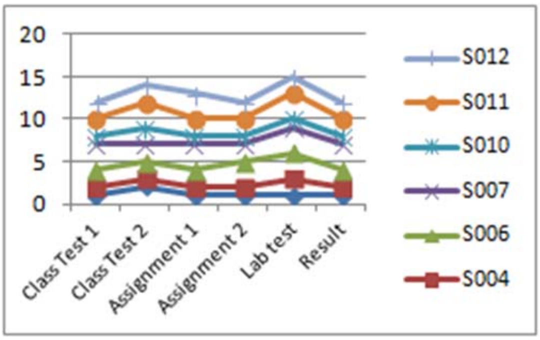

Fig-2. Original Data Set

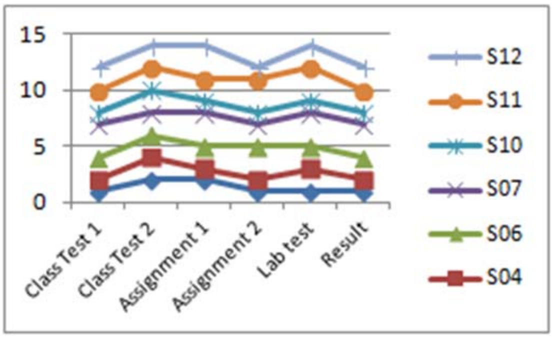

Fig-4. RST Analysis

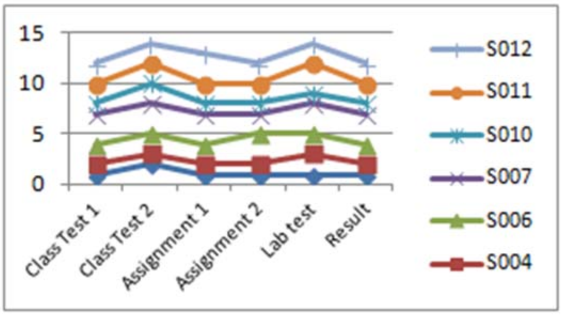

Fig-3. LOCF Analysis

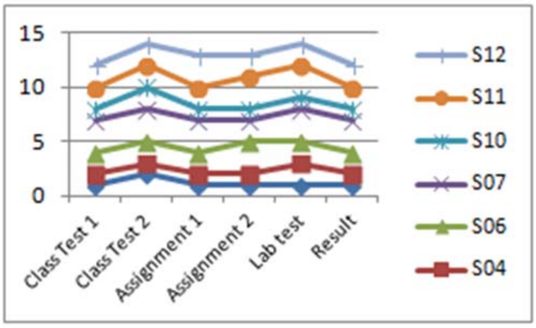

Fig-5. Soft Set Analysis

The simulation results suggested that the there is no single method available that is the best under all situations. In the above experiments the original student data set contains 12 rows and 7 attributes; in this the data set randomly missed 8 attributes. In CDC method the missing data are completely deleted and result data set will 5 rows. While Soft Set method was superior to LOCF and RST also the LOCF method better than RST because in Soft Set analysis the missing data rate matched 62.5% where as in LOCF analysis the missing data matching is 50% and RST analysis matching only 38%.

VI. DISCUSSION AND CONCLUSIONS

This paper discusses different methods like Rough Set Analysis, Soft Set Analysis [31], LOCF Analysis and CDC Analysis to impute the missing values. It is found that none of the methods are giving exact information of missing values to impute. Missing values are replaced by probability distributions over possible values for the missing feature, which allows the corresponding transaction to support all item sets that could possibly match the data. Transactions which do not exactly match the candidate item set may also contribute a partial amount of support this behavior is beneficial for databases with many missing values or containing numeric data. Handling missing values using the most probable information for all the samples belonging to the same class gives better result as compare to other techniques. Missing values filled with better accuracy leads to better results, this phenomenon is also observed.

REFERENCES

[1] Gantayat S. S., Misra Ashok, Panda B. S., (2014) “A Study of Incomplete Data - A Review”, International Conference on Frontiers in Intelligent Computing: Theory and Applications-2013, (FICTA-2013), Published in Advances in Intelligent Systems and Computing, Springer Verlag, Switzerland, 247, pp. 401-408.

[2] Little, R.J.A. and Rubin, D.B. (1987), Statistical Analysis with Missing Data, John Wiley & Sons, New York.

[3] Collins, L.M., Schafer, J.L. and Kam, C.M. (2001), A Comparison of Inclusive and Restrictive Missing-Data Strategies in Modern Missing-Data Procedures, Psychological Methods, 6, pp. 330-351.

[4] Little, R.J.A. (1988), A Test of Missing Completely at Random for Multivariate Data with Missing Values, Journal of the American Statistical Association, 83, pp. 1198-1202.

[5] Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002), Analysis of Longitudinal Data. 2nd Edition, Clarendon Press, Clarendon.

[6] Carpenter, J.R., Kenward, M.G. and Vansteelandt, S. (2006), A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data, Journal of the Royal Statistical Society, Series A (Statistics in Society), 169, pp. 571-584.

[7] Musil, C.M., Warner, C.B., Yobas, P.K. and Jones, S.L. (2002), A Comparison of Imputation Techniques for Handling Missing Data, Western Journal of Nursing Research, 24, pp. 815-829.

[8] Sprint, A. and Dupin-Sprint, T. (1993), Imperfect Data Analysis, Drug Information Journal, 27, pp. 995-994.

[9] Myers, W.R. (2000), Handling Missing Data in Clinical Trials: An Overview, Drug Information Journal, 34, pp. 525-533.

[10] Hening, D. and Koonce, D.A. (2014), Missing Data Imputation Method Comparison in Ohio University Student Retention Database, Proceeding of the 2014 International Conference on Industrial Engineering and Operations Management, Bali, Indonesia.

[11] Ali, A.M.G., Dawson, S.J., Blows, F.M., Provenzano, E., Ellis, I.O., Baglietto, L., Huntsman, D., Caldas, C. and Pha- roah, P.D. (2011), Comparison of Methods for Handling Missing Data on Immunohistochemical Markers in Survival Analysis of Breast Cancer, British Journal of Cancer, 104, pp. 693-699.

[12] Patrician, P.A. (2002), Focus on Research Methods Multiple Imputation for Missing Data, Research in Nursing & Health, 25, pp. 76-84.

[13] Nakai, M., Chen, D.G., Nishimura, K. and Miyamoto, Y. (2014), Comparative Study of Four Methods in Missing Value Imputations under Missing Completely at Random Mechanism, Open Journal of Statistics, 4, pp. 27-37.

[14] Lavori, P.W., Dawson, R. and Shera, D. (1995), A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data, Statistics in Medicine, 14, pp.1913-1925.

[15] Pawlak, Z. (1982), Rough Sets, J. Inf. & Comp. Sc. II, pp. 341- 356.

[16] Grzymala-Busse, J.W., Grzymala-Busse, W.J., Goodwin, L.K. (2001), Coping with Missing Attribute Values Based on Closest Fit in Preterm Birth Data: A Rough Set Approach, Computation Intelligence. 17, pp. 3425-434.

[17] Grzymala-Busse, J.W., (2003) "Rough set strategies to data with missing attribute values, Workshop Notes, Foundations and New Directions of Data Mining", the 3rd International Conference on Data Mining, Melbourne, FL, USA, November 19–22, pp. 56–63.

[18] Grzymala-Busse, J. W., (2004), Three Approaches to Missing Attribute Values- A Rough Set Approach, Workshop on Foundations of Data Mining, Associated with 4th IEEE International Conference on Data Mining, Brighton, UK.

[19] Panda B. S., GantayatS. S., Misra Ashok, (2015), Rough Set Rule Based Technique for the Retrieval of Missing Data in Malaria Diseases Diagnosis, In: Springer Briefs in "Applied Sciences and Technology - Forensic and Medical Bioinformatics", pp-59-71, 2015.

[20] Schafer, J.L. (1997), The Analysis of Incomplete Multivariate Data, Chapman & Hall, London.

[21] Schafer, J.L. (2000), Analysis of Incomplete Multivariate Data, Chapman & Hall/CRC, Boca Raton.

[22] Molodtsov, D. (1999), Soft Set Theory First Results, Computers and Mathematics with Applications, 37(4-6), pp. 19–31.

[23] Maji, P.K., Biswas R. and Roy A.R. (2003), Soft Set Theory, Computers and Mathematics with Applications, 45(4-5), pp. 555-562.

[24] Allison, P.D. (2001), Missing Data, Sage Publications, Thousand Oaks.

[25] Kim, J.O. and Curry, J. (1977), The Treatment of Missing Data in Multivariate Analysis, Sociological Methods Research, 6, pp. 215-240.

[26] Allison, P.D. (1998) Multiple Regressions: A Primer, Pine Forge Press, Thousand Oaks.

[27] Molenberghs, G., Thijs, H., Jansen, I., et al. (2004), Analyzing Incomplete Longitudinal Clinical Trial Data, Biostatistics, 5, pp. 445-464.

[28] Shao, J. and Zhong, B. (2003), Last Observation Carry-Forward and Last Observation Analysis, Statistics in Medicine, 22, pp.2429-2441.

[29] Mallinckrodt, C.H., Clark, W.S. and David, S.R. (2001), Accounting for Dropout Bias Using Mixed-Effects Models. Journal of Biopharmaceutical Statistics, 11, pp. 9-21.

[30] Graham, John W., (2012), Missing Data-Analysis and Design, Springer-Verlag, New York.

[31] B. S. Panda, S. S. Gantayat, Ashok Misra "Retrieving the Missing Information from Information Systems Using Rough Set, Covering Based Rough Set and Soft Set " In: International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 7 (3), 2016, 1403-1407-2016.