International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A New Method for Missing Value Imputation in Incomplete Information Systems Using Hybrid Rough Set Theory

G.V. Suresh[1], Dr. E. Sreenivasa Reddy[2]
[1]Research Scholar, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India
[2]Professor, University College of Engineering, Acharya Nagarjuna University, Andhra Pradesh, India

Article Info
Volume 6, Issue 4
Page Number: 580-592
Publication Issue :
July-August-2020

Article History
Accepted : 25 Aug 2020
Published : 31 Aug 2020

ABSTRACT

Decision making has become a main reason for data analysis in the current scenario. Before analysis, the data must be freed from noise by applying data pre-processing techniques to the raw data. Missing value imputation is one of the data cleaning methods in data preprocessing. This article presents a new data imputation technique with the concepts of approximate set theory. A Hybrid Rough Set Theory for Missing Value Imputation (HRST-MVI) imputation algorithm is developed. The performance of the proposed algorithm is carried out by comparing the classification accuracy obtained, after the imputation of the missing value. The proposed method and comparative methods were compared using different classifiers in terms of accuracy, precision, recall, and F1 score. The performance of the classifiers shows that the HRST-MVI can impute missing values from multiple patterns more efficiently than other comparative methods.

Keywords : Decision Making, Incomplete Information Systems, Rough Set Theory (RST), Missing Value Imputation.

I. INTRODUCTION

The study of huge datasets using novel ideas, methods, and tools is called data mining, a developing branch of computer intelligence. It aids contemporary businesses who are grappling with the difficult task of deciding what to do with vast amounts of data in order to better understand their markets, clients, vendors, operations, processes, medical diagnoses, operational effectiveness, etc[1]. A large body of evidence is now available thanks to the development of technology and the refinement of data gathering techniques. In addition to getting a high probability of experiencing questionable data, the number of real-time datasets is growing daily in both domains.

1.1. Characteristics of the Data

Since they come from several, multiple perspectives, real-world data references are very sparse, chaotic, and unreliable. Retrieval of insight from inaccurate data yields useless judgments and little importance[2].

- **Incomplete data** results out of missing attributes, absence of some desirable qualities, or availability of simply aggregate data. It could

happen because the data value was "not appropriate" while being acquired, because of the delay among both data gathering and its analysis, or because of issues with people, systems, or algorithms[3].

- **Noisy data** results out of errors or outliers causing imbalanced datasets. Errors in data transmission, user or software malfunction during entering data, or flawed data collecting are the main causes.
- **Data** inconsistency results out of discrepancies in codes or names that typically happens because the data comes from many sources.
- **Vagueness** refers to the inability to distinguish clearly or precisely in the real world which is also known as ambivalence[4].
- **Uncertainty** is caused by low knowledge or inadequate evidence.
- **Ambiguity** refers to existence of more than one specific choice or viewpoint, leaving the decision indefinite.

Whenever data pretreatment procedures are used prior to mining, pattern reliability is ensured while subsequent decision-making may both be significantly improved.

- **Information based uncertainty:** Absence of data leads to doubt that really is information-based, wherein conflict and uncertainty are other categories. Conflict results from deciding between various choices for the attribute, and probability theory can help with this. RST, which provides analytical methods to find hidden patterns and relationships, may deal with ambiguity. It finds cause-and-effect relationships (i.e., partial, or complete dependency) in databases, gets rid of unnecessary data, and provides solutions for null values, missing data, dynamic data, etc[5].

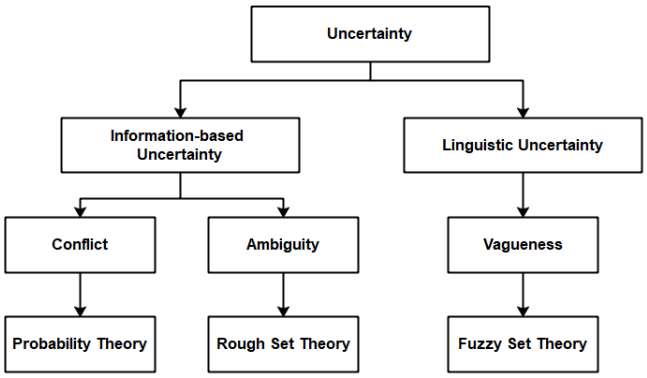

Figure. 1. Classification of Uncertainty

- **Linguistic uncertainty:** It is based on ambiguous human speech, and with time, the exact meaning of the term might alter. A linguistic variable refers to values related to words or phrases in a language, whether it be natural or manufactured, rather than numbers. Vagueness serves as a representation of it [6]. The border region method, or the presence of items that cannot be uniquely categorized in relation to a set or its complement, is typically linked to vagueness. Fuzzy set theory can deal with ambiguity. The membership function, which is stated as, describes the extent to which an item is a member of that group, shown as;

$$\mu_X^I(x) = \frac{|X \cap I(x)|}{|I(x)|}$$

wherein, I represents binary relation on U universe, $I(x)$ stands for entire set of $y \in U$ for $\mu_X^I(x) \in [0,1]$. As certain words contain several meanings that makes unclear interpretation as opposed to the intended one, ambiguity results. Ambiguity is defined as subset $X \subseteq U$ while binary relation R is based on U. An object $X \in U$ could appear unclear due to its relationship with X in both cases[7].

- When $x \notin X$, yet $y \notin X$ shows that x is identical to y, wherein the information given by R indicates converting x into X.

- b) If $x \notin X$, yet $y \in X$ while x is identical to y, wherein R provides information to include x into X.

Uncertainty is a major hurdle for decision making in medical diagnosis. Application of computational intelligence techniques for handling uncertainty in medical data may reveal new insights in medical diagnosis and drug discovery[8].

1.2. Information and Decision Systems

Every row in a table that describes a data set is a case, an event, a pattern, or just a plain item. Each column contains a measurement-able feature (a variable, an observation, a characteristic, or a trait) which can be given to objects by users or experts, while a system is commonly referred to as tables[9]. In other words, $I = (U, A)$ is a pair in which U is referred to as a non-empty finite set of objects named 'Universe' and A contains non-empty finite attribute sets as $a : U \to V_a$ for every $a \in A$. The value set of a. Set is nothing but, V_a.

Table. 1.Information Table Specimen (N stands for No; Y for Yes)

Attributes			
Id	**Headache**	**Nausea**	**Temperature**
1	N	Y	High
2	Y	N	Very High
3	Y	Y	High
4	N	Y	High
5	Y	N	Normal
6	N	Y	Normal

The classification of the attributes related to numerous patterns (or entities) is commonly described upfront. Training data is the collection of patterns. Supervised learning refers to technique of automatic classification of an unknown patterns or test data out of past experience of training data. Decision systems represent the above. In case, A as an attribute set is sliced into sets of condition attributes C and set of decision attributes D such that $A \to C \cup D$ and $C \cap D$ is empty, the information system is known as a decision table. Table.1 contains an illustration of a decision system. This table includes a decision feature (D) also class, eight objects, and four conditional features (Migraine, Anxiety, and Temperature) (or patterns). A decision system seems coherent when related decision features remain similar in any collection of objects with almost similar attribute value[12].

Table. 2. Decision Table specimen (N stands for No; Y for Yes)

$x \in U$	Headache	Nausea	Temperature	Decision (D)
0	N	Y	High	Yes
1	Y	N	Very High	Yes
2	Y	Y	High	Yes
3	N	Y	High	No
4	Y	N	Normal	No
5	N	Y	Normal	Yes
6	N	Y	High	Yes

1.3. Incomplete Information Systems

A quadruple $I = (U, A, V, f)$ that has U as a non-empty finite set of objects $a \in A : U \to V_a$, A as a non-empty finite object set V_a is referred to as an insufficient information system a. Every attribute domain V_a may include symbols other than the special "*" to denote attributes with uncertain values. V is considered the value set of all characteristics in I and V ought to satisfy $V = \cup_{a \in A} V_a$. Define f as an information function in I and there will be $f(x, a) \in V_a$ for any $a \in A$ and $x \in U$.

- if $"f(x, a) = *"$, then we assume that the unknown value x holds on a is "do not care".

- if "$\mathbf{f(x,a)=?}$" , then we assume that the unknown value $\mathbf{x}$ holds on $\mathbf{a}$ is lost.

Table. 3. Example of Incomplete Information Table (N stands for No; Y for Yes)

Attributes			
Id	**Headache**	**Nausea**	**Temperature**
1	N	Y	High
2	*	N	Very High
3	Y	Y	*
4	N	*	High
5	*	N	Normal
6	N	Y	*

A partial decision system refers to such information systems like $\mathbf{I=(U,A\bigcup D,V,f)}$ where $\mathbf{A\bigcap D=\phi}$.

Table. 4. Incomplete Decision Table specimen(N stands for No; Y for Yes)

$\mathbf{x\in U}$	Headache	Nausea	Temperature	Decision (D)
0	N	Y	High	Y
1	*	N	Very High	*
2	Y	Y	*	Y
3	N	*	High	N
4	*	N	Normal	*
5	N	Y	*	Y
6	N	Y	High	Y

1.4. Problem with Incomplete Information

Missing values may appear in databases that hold survey or health records the data mining part of data storing operation. Instances of how data might disappear include issues with clinical data and failure to respond to a questionnaire. The issue of missing data makes it hard to analyze underlying condition and make decisions based upon such data, necessitating precise and effective estimating techniques. Solutions for authentic computing that rely heavily on data frequently struggle with the issue of missing input variables. Missing data could exhibit a predictable pattern. Investigating such patterns is crucial to enable identification of situations, factors influencing missing data[10]. A suitable estimating technique could be chosen when factors predicting required patterns have been determined. Attributing a "null" value to every missing attribute value has been shown to be problematic in decision and data analysis since missing or incomplete values are simply "missed," even if they occur and affect choices. The content of the present article is presented in various sections. The approaches for missing data have been described in Section 2. Section 3 presents a novel approach to impute missing data. Section 4 presents experimental results that are computationally obtained by using available datasets. Lastly, Section 5 presents discussions, while section 6 presents the conclusion.

II. Related Work

2.1. Mechanisms of Missing Data

Missing data mechanism is classified as ignorable or non- ignored. Ignorable case where missing data probability is dependent upon a scenario wherein observed data is free from missing data. In non-ignorable, the chance of missing data is dependent upon missing data not being found in observed data. Further in line to ignorable and non-ignorable missing data mechanisms classified into three types and is shown in Fig.2. The process in which missing data connection is maintained between missingness and variable values in the data array is described [15].

2.1.1. Missing Completely at Random (MCAR)

If each value has the same missing probability, it is referred as MCAR or Missing Completely at Random. So, here, probability of a value as missing is not dependent on such value as well as all others values

as well. In that case such missing data forms random model of all cases[16].

2.1.2. Missing at Random (MAR)

While missing value probability depends on other attributes value, then it may be known as MAR or Missing at Random. So, missing value probability is not dependent upon that value, though the dependence may be on rest of the values[18].

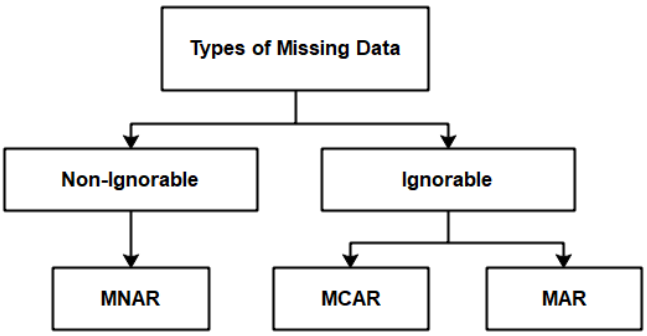

Figure.2. Classification of Missing Data

2.1.3. Missing Not at Random (MNAR)

When probability of missing value probability depends upon missing value itself then it is referred as MNAR or Missing Not at Random. For Example, If a questionnaire does not have age of a women that remains unanswered[20].

There is no mechanism to test missingness characteristics MCAR, MNAR or MAR, so we have to depend on assumption to handle missing data. Missingness characteristics and different conditions of missing value handling methods have to consider selecting best methods.

2.2. Missing Value Imputation?

Typically, any missing attribute value often gives more than one distinct meaning. One scenario may reflect just loss of missing attribute value. The absence of data is what caused such value loss. "Do not care" circumstances offer alternative as the second scenario. At data collection point, any missing attribute value is regarded as unimportant. A specialist may determine, for instance, that such attribute values was not essential to a proper evaluation. Missing Value Imputation (MVI) techniques substitute parameter estimates for missing values predicated upon the existing and possible info contained in the dataset[21].

2.2.1. Case deletion using Missing Attribute Value

Such approach relies on disregarding instances when an attribute value is absent. In statistical data, this is often known as ranking removal, specific instance omission, or detailed overview. This data collection contains no cases having missing attributes.

2.2.2. Missing Value Imputation with Mean

Regarding discrete data, our approach substitutes the mean of attributes ascribed to all missing values. Such a method has drawbacks, including weakening of data's covariance and correlation estimations and reduction in variability since its non-focus on variable relations.

2.2.3. Missing Value Imputation with Most Common Attribute Value

For symbolically presented attribute, such a technique assigns the value that is frequently present, while in case of numerical attribute, it uses mean to fill in the gaps left by missing value.

2.2.4. Missing Value Imputation with Concept Most Common Attribute Value

In this case, values chosen to represent unknown missing attributes appear most often. Such a method restricts how the majority of attribute values are filled up by idea or choice.

2.2.5. Missing Value Imputation by deleting cases or ignoring Missing

Such a procedure involves eliminating any instance (case) that have at minimum single attribute having missing data. It analyses the leftover instances (scenarios), while leaving out ones having missing data. Once qualities are randomly absent in their whole, such technique is constrained. Whenever it's realistic to forgo making a forecast on certain situations, it could be useful in practice to dismiss cases with missing attribute values at inference time.

2.3.Rough Set Theory

Topological procedures, such as interior and closure, could be used to construct rough set idea in a very generic way. Let's say the cosmos is a collection of objects $\mathbf{U}$, and the unnormalized connection $\mathbf{R} \subseteq \mathbf{U} \times \mathbf{U}$ represents our ignorance of its constituent parts $\mathbf{U}$. We presume that is an equivalence relation of $\mathbf{U}$ just for convenience. Assume that $\mathbf{X}$ is a subset of $\mathbf{U}$ [30]. As regard to, we wish to describe the set $\mathbf{X}$ in conjunction with $\mathbf{R}$. The following list contains rough set theory fundamentals[5].

- The Lower Approximation of a set $\mathbf{X}$ with respect to $\mathbf{R}$ refers to a set consisting objects that may be surely classified as $\mathbf{X}$ in relation to $\mathbf{R}$ (certainly $\mathbf{X}$ in relation to $\mathbf{R}$).
- The Upper Approximation of a set $\mathbf{X}$ with respect to refers to a set consisting of objects that may be classified with possibility as $\mathbf{X}$ related to $\mathbf{R}$).
- The Boundary Region of a set $\mathbf{X}$ related to $\mathbf{R}$ refers to a set consisting of objects not classified as $\mathbf{X}$ or not- as related to $\mathbf{R}$.
- Set $\mathbf{X}$ turns Crisp (exact to $\mathbf{R}$) when having empty boundary region of $\mathbf{X}$.
- Set $\mathbf{X}$ turns Rough (inexact to $\mathbf{R}$)when having nonempty boundary region of $\mathbf{X}$.

Let $\mathbf{R}$ be considered as indiscernibility or similarity relation among unnoticeable values. $[\mathbf{X}]_{\mathbf{R}}$ denotes the equivalence class of $\mathbf{R}$ containing $\mathbf{x}$. The elementary sets are the equivalence classes of $\mathbf{R}$. The lower and higher approximations of a set over the universe's constituent parts are provided.

R-lower approximation of $\mathbf{X}$

$$\underline{\mathbf{RX}}=\{\mathbf{x} \in \bigcup \mid [\mathbf{x}]_{\mathbf{R}} \subseteq \mathbf{X}\}$$

R-upper approximation of $\mathbf{X}$

$$\overline{\mathbf{RX}}=\{\mathbf{x} \in \bigcup \mid [\mathbf{x}]_{\mathbf{R}} \cap \mathbf{X} \neq \varnothing\}$$

R-boundary region of $\mathbf{X}$

$$\mathbf{R_{NR}(X)}=\underline{\mathbf{RX}}-\overline{\mathbf{RX}}$$

Accuracy of Approximations given by

$$\alpha_{\mathbf{R}}(\mathbf{X})=\frac{|\bar{\mathbf{R}}(\mathbf{X})|}{|\underline{\mathbf{R}}(\mathbf{X})|}[0 \leq \alpha_{\mathbf{R}} \leq 1]$$

Where $|\mathbf{X}|$ denotes the cardinality of $\mathbf{X}$. If $\alpha_{\mathbf{R}}(\mathbf{X})=1$, $\mathbf{X}$ is crisp related to $\mathbf{R}$ ($\mathbf{X}$ is accurate in $\mathbf{R}$), and otherwise, if $\alpha_{\mathbf{R}}(\mathbf{X})<1$, $\mathbf{X}$ is rough in $\mathbf{R}$ ($\mathbf{X}$ is vague} in $\mathbf{R}$).

2.4. Dependency Attributes

Let $\mathbf{A}$ consist of $\mathbf{C}$ and $\mathbf{D}$ as subsets. It may be said that $\mathbf{D}$ is dependent upon $\mathbf{C}$ in a degree $\mathbf{k}(\mathbf{0} \leq \mathbf{k} \leq \mathbf{1})$ as specified by $\mathbf{C} \rightarrow_{\mathbf{k}} \mathbf{D}$ such that

$$\mathbf{k}=\gamma(\mathbf{C},\mathbf{D})=\frac{|\mathbf{POS_c(D)}|}{|\mathbf{U}|}$$

Wherein $\mathbf{POS_c(D)}$ is designated as $\mathbf{D}$'s C-positive region, while degree of the dependency becomes coefficient . In case of $\mathbf{k=1}$, $\mathbf{D}$ is known to be having total dependence on $\mathbf{C}$, whereas, in case of $\mathbf{k<1}$, $\mathbf{D}$ has partial dependence (in a degree $\mathbf{k}$) on $\mathbf{C}$. In other words, the ratio of all items in the universe that may be correctly categorized as, while partitioned blocks $\mathbf{U/D}$ are expressed by the coefficient, that makes use of features $\mathbf{C}$.[7]

2.5. Indiscernibility Relation

Let $\mathbf{I}=(\mathbf{U},\mathbf{A})$ be an information system, then with any $\mathbf{B} \subseteq \mathbf{A}$, there is associated an equivalence relation $\mathbf{IND_I(B)}$.

$$\mathbf{IND_I(B)}=\left\{(\mathbf{x},\mathbf{x}') \in \mathbf{U}^2 \mid \forall \mathbf{a} \in \mathbf{B}\ \mathbf{a(x)}=\mathbf{a(x)}\right\}$$

$\mathbf{IND_I(B)}$ refers to B-indiscernibility relation. When $(\mathbf{x},\mathbf{x}') \in \mathbf{IND_I(B)}$, the object $\mathbf{x}$ and $\mathbf{x}'$ become inseparable with use of attributes from $\mathbf{B}$. The similarity classes of the B-indiscernibility relation are denoted $[\mathbf{x}]_{\mathbf{B}}$.[6]

III. Proposed Work

Here, our paper suggested an imputation method that is premised on Hybrid Rough Set for estimating as well as substituting missing values for a given datasets.

3.1. Missing Value Imputation using Hybrid Rough Set Theory (HRST-MVI)

The pre-requisitions for the proposed method are:

- **Indiscernibility Relation (IND)** recognizes a relationship among two or more groups as each object's value match to a group of conditional attribute values $\mathbf{A}$.
- **Lower approximation** is used to determine the links among conditional attribute $\mathbf{A}$ and decision features $\mathbf{D}$ that describe which items unquestionably pertain to the notion $\mathbf{X} \subseteq \mathbf{U}$.
- **Upper approximation** is needed to find the links among conditional attribute and decision attribute, that indicate whether objects could belong to the notion $\mathbf{X} \subseteq \mathbf{U}$ or not.
- **Positive Region** refers to any object $\mathbf{U}$ that may be grouped into classes that use the data in conditional attribute.
- Dependency only happens only in positive region if an object's equivalence class is in the positive area. An estimation of the relevance of an attribute may be determined by calculating the change in dependence when that characteristic is excluded from the list of potential attributes. This characteristic is more relevant the greater the shift in reliance.

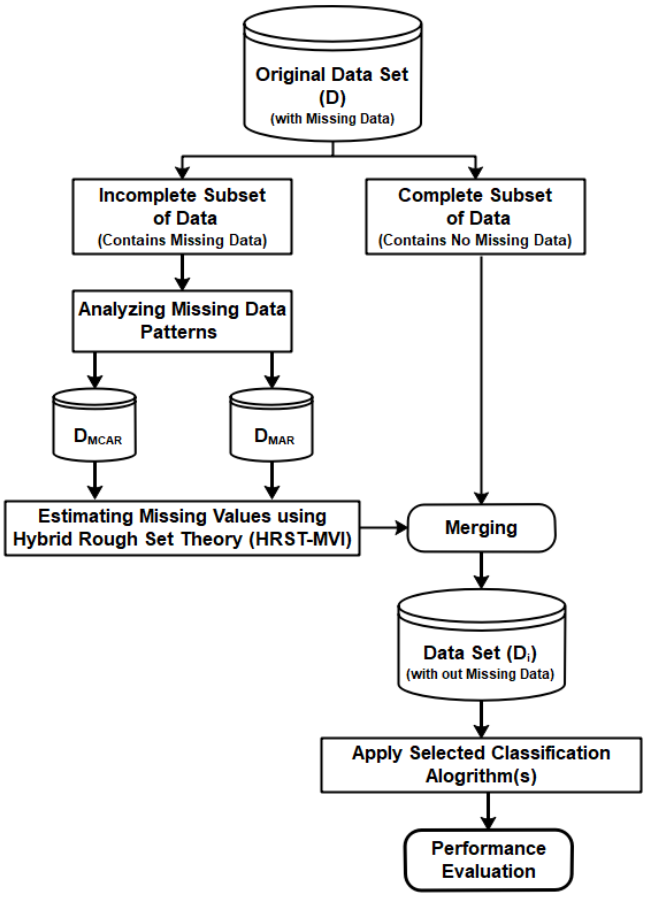

Figure.3. Experimental procedure for proposed Model

Algorithm: Missing Value Imputation using Hybrid Rough Set Theory (HRST-MVI)

Input: An incomplete decision information system, suppose that $\mathbf{I} = (\mathbf{U}, \mathbf{A}, \mathbf{V}, \mathbf{f})$ and $\mathbf{A} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_m)$ is a family of $\mathbf{m}$ attribute subsets of $\mathbf{A}$, $\mathbf{U}/\mathbf{D} = \{\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_k\}$.

Output: Vector containing possible Missing Values.

- Given dataset is shown as an information system $\mathbf{I} = (\mathbf{U}, \mathbf{A} \cup \mathbf{D})$ wherein $\mathbf{U}$ is a finite, non-empty object sets known as universe of discourse, $\mathbf{A}$ is finite, non-empty attribute sets like $\mathbf{a} : \mathbf{U} \rightarrow \mathbf{V}_a$ for each $\mathbf{a} \in \mathbf{A}$, wherein $\mathbf{V}_a$ refers to values for attribute '$\mathbf{a}$' to assume, whereas $\mathbf{D} \notin \mathbf{A}$ refers to decision attribute; $\mathbf{B} \subseteq \mathbf{A}$.
- Indiscernibility relation **IND** An operation is carried out to divide the universe of objects $\mathbf{U}$ into equivalence classes $\mathbf{D}$.
- The given steps to be repeated in case of every

attribute in $\mathbf{A}$.

- Compute equivalence class family of every attribute
- Calculate degree of association for extracting exact relation among object sets $\mathbf{U}$. When $\mathbf{x_i}$ contains identical conditional attributes with $\mathbf{x_j}$ other than missing value, missing value, $\mathbf{x_{miss}}$ is replaced, with the value $\mathbf{v_j}$, from $\mathbf{x_j}$, wherein $\mathbf{j}$ represents an index to another case.
- Else, let $\underline{\mathbf{AX}}$ be computed with every conditional attribute A from the data of previous case having missing value.
- Replace value in a delayed time when one exceeding $\mathbf{v_j}$ value is found suitable to estimate the replacement.
- Let $\overline{\mathbf{AX}}$ be compute upper approximations of every subset partition before performing missing data computation and imputation with the $\underline{\mathbf{AX}}$.
- Apply Selected Classification Algorithms.
- Performance Evaluation

IV. Experimental Results

The Cleveland heart disease dataset that was obtained from the UCI machine learning repository was used in this study. The data mining community uses a variety of datasets and data generators from this library to conduct empirical research.

4.1. Classifiers used for Analysis

The performance of the proposed method was evaluated using three classifiers multi-layer perceptron (MLP), Support Vector Machine (SVM), and K-nearest neighbors (KNN). In addition, k-fold cross validation with k = 5 was considered to alleviate the bias caused by the random selection of the dataset.

4.2. Performance Measures

In each experiment set, the accuracy, precision, and recall of the findings of HRST-MVI and comparison approaches are evaluated when a separate classifier is applied.

$$\mathbf{Accuracy} = \sum_{i=1}^{1} \frac{\mathbf{TN_i} + \mathbf{TP_i}}{\mathbf{TN_i} + \mathbf{FN_i} + \mathbf{TP_i} + \mathbf{FP_i}}$$

$$\mathbf{Precision} = \frac{\sum_{i=1}^{1} \mathbf{TP_i}}{\sum_{i=1}^{1} \mathbf{TP_i} + \mathbf{FP_i}}$$

$$\mathbf{Recall} = \frac{\sum_{i=1}^{1} \mathbf{TP_i}}{\sum_{i=1}^{1} \mathbf{TP_i} + \mathbf{FN_i}}$$

In this experiment, the Expectation-Maximization Imputation (EMI),Fuzzy Rough Set Imputation (FRSI), and proposed HRST-MVI, , and imputation methods are considered to impute missing values with MCAR and MAR patterns in the Cleveland heart disease dataset. Based on this determined the $\mathbf{D_{MAR}}$ and Topper $\mathbf{D_{MCAR}}$ datasets by comparing the performance of different classifiers SVM, KNN, and MLP. The KNN, SVM, and MLP classifiers estimate the classification accuracy rate of all the imputed datasets. The most accurate dataset has been chosen. The chosen dataset is the most accurately imputed dataset and comprises data having least amounts of unreliability as compared to unquantified data. Tables 5 and 6 present the findings from such an experiment. In comparison rest of the approaches in the dataset, HRST method performed much better.

Table 5. Imputation Comparison on the MCAR

Metrics	Classifier	10%			20%			30%		
		HRST-MVI	FRSI	EMI	HRST-MVI	FRSI	EMI	HRST-MVI	FRSI	EMI
Accuracy	SVM	87.55	86.92	83.42	87.08	86.46	83.03	85.79	85.18	81.78
	KNN	88.72	88.18	82.66	87.23	87.74	82.29	86.92	86.45	81.05
	MLP	89.76	88.29	86.27	88.25	87.81	85.83	87.93	86.51	84.55
Precision	SVM	87.47	86.42	86.03	87.47	85.97	85.59	85.71	84.69	84.32
	KNN	90.98	86.93	83.09	90.44	86.47	82.71	89.12	85.19	81.46
	MLP	92.35	89.09	83.67	91.78	88.59	83.28	90.45	87.28	82.03
Recall	SVM	87.57	86.39	81.89	87.17	85.94	81.53	85.81	84.66	80.31
	KNN	85.32	80.86	78.49	84.89	80.52	78.24	83.63	79.38	77.72
	MLP	90.55	86.92	82.42	90.02	86.46	82.05	88.71	85.18	80.81

Evaluation of the HRST-MVI imputation method showed in Table 5 and Table 6, the HRST-MVI with MLP methods, which yielded the best results compared to other imputed data sets. In the experiment, hybrid imputation method HRST-MVI is put in a comparative mode with FRST and EM imputation, on the heart dataset.

Table.6. Imputation Comparison on the MAR

Metrics	Classifier	10%			20%			30%		
		HRST-MVI	FRSI	EMI	HRST-MVI	FRSI	EMI	HRST-MVI	FRSI	EMI
Accuracy	SVM	88.36	87.73	84.23	87.72	87.10	83.67	86.21	85.60	82.20
	KNN	89.53	88.99	83.47	88.87	88.34	82.93	87.34	86.82	81.47
	MLP	92.57	89.10	87.08	89.89	88.45	86.47	88.35	86.93	84.97
Precision	SVM	88.28	87.23	86.84	87.64	86.61	86.23	86.13	85.11	84.74
	KNN	91.79	87.74	83.90	91.08	87.11	83.35	89.54	85.61	81.88
	MLP	93.16	89.90	84.48	92.42	89.23	83.92	90.87	87.70	82.45
Recall	SVM	88.38	87.20	82.70	87.74	86.58	82.17	86.23	85.08	80.72
	KNN	86.13	81.67	79.30	85.53	81.16	78.84	84.05	79.72	77.42
	MLP	91.36	87.73	83.23	90.66	87.10	82.69	89.12	85.60	81.23

Figure.4 and Figure.5 visibly demonstrate the proposed method's superior performance over some powerful imputation methods in terms of Accuracy.

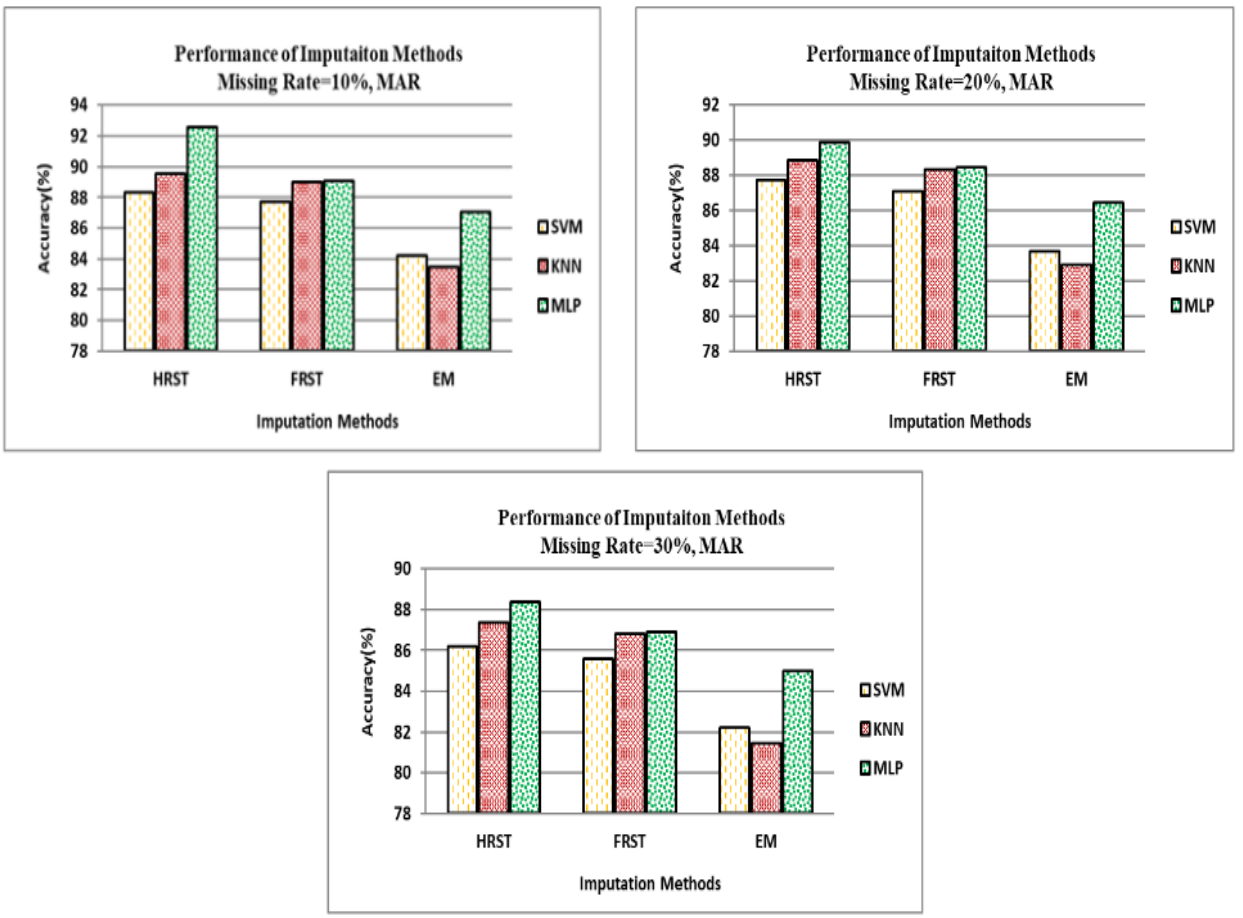

Figure.3. Performance of Imputation Methods with 10%,20%,&30% missingness on the MAR

The classification accuracy of 92.57% makes the proposed approach better than other methods shown in Table 6 in case of the MAR. Similarly, in case of the MCAR, classification accuracy of 89.76% makes the proposed approach better than other methods shown in Table 5. In case of missing proportion 20%, classification accuracy of 89.98% makes the proposed approach better than other methods shown in Table 6 in case of the MAR with 10% missing rate. Similarly, in case of the MCAR, classification accuracy of 88.25% makes the proposed approach better than other methods shown in Table 5

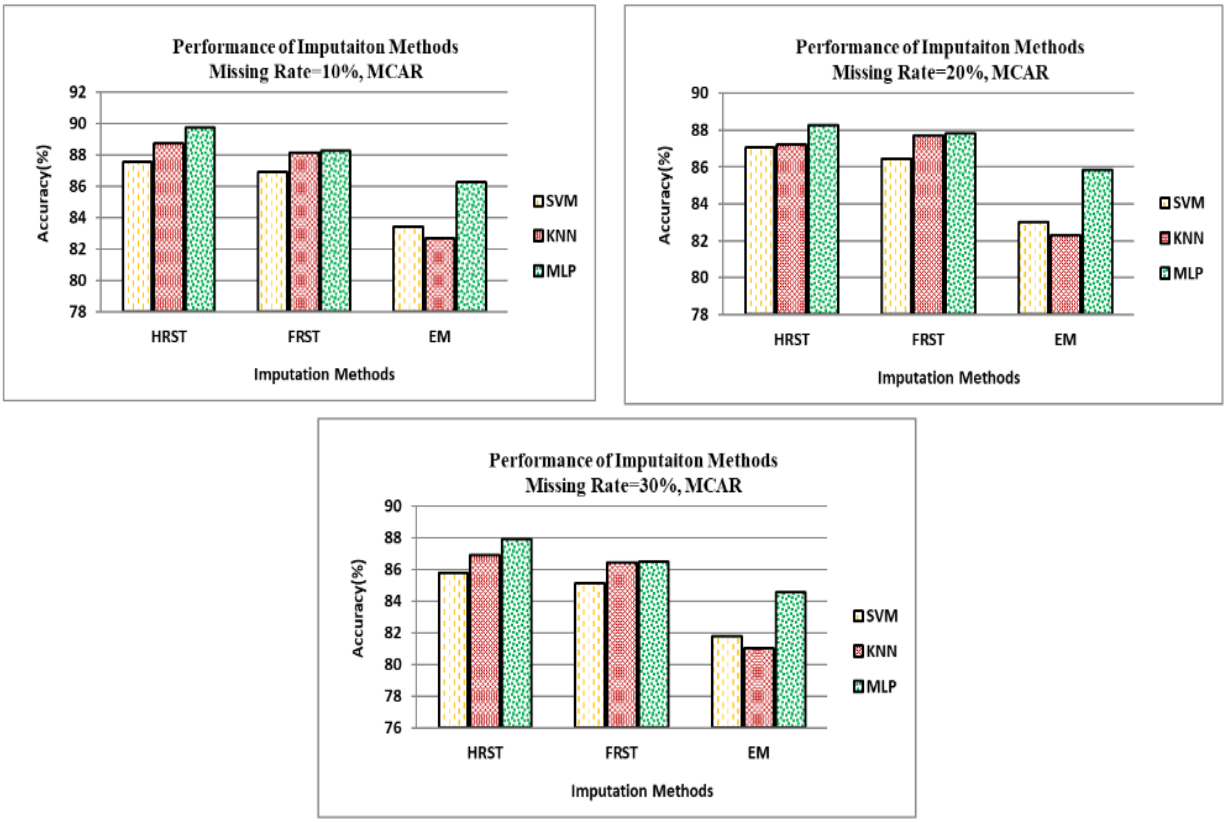

Figure.4. Performance of Imputation Methods with 10%,20%,&30% missingness on the MCAR

In case of missing proportion 30%, classification accuracy of 88.35% makes the proposed approach better than other methods shown in Table 6 in case of the MAR. Similarly, in case of the MCAR, classification accuracy of 87.93% makes the proposed approach better than other methods shown in Table 5. Similar to accuracy, recall and precision results also shows that proposed method dominates standard techniques. Also, it is proved that proposed method is good for predicting early diseases for heart issues.

V. CONCLUSION

Owing to huge size and possible emergence from many sources, actual data are especially sensitive to noise and missing data. Preprocessing data may improve the accuracy in making decisions. Data preparation benefits significantly from data cleansing. Usually, many data sets, regardless of kind, have much missing value. It might be difficult to impute missing data with acceptable value. The judgments made may be impacted by incorrect imputations. Our study illustrates the use of RST ideas to fill in datasets with missing vales. Medical information is typically unclear, imprecise, and vague in character. The suggested imputation approach might produce an effective analysis when applied to health data. Employing various classifiers, the efficiency, accuracy, recall, and F1-score of the suggested technique and comparable methods was contrasted. The results of the classifiers demonstrate that the HRST-MVI can much more successfully do imputation of multi-pattern missing data over rest of the comparison techniques. Additionally, it has been found as the proportion of missing data grew, the performance of the classifier increased when missing values were imputed utilizing HRST-MVI.

VI. REFERENCES

[1]. El-Hasnony IM, El-Bakry HM, Saleh AA (2016) Classification of breast cancer using soft computing techniques. Int J Electron Inf Eng 4(1):45–54.

[2]. Wang H, Wang S (2010) Mining incomplete survey data through classification. Knowledge. Inf Syst 24(2):221–233.

[3]. J.W. Grzymala-Busse, L.K. Goodwin, W.J. Grzymala-Busse, X. Zheng, Handling missing attribute values in preterm birth data sets, Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, (Springer, 2005), pp.342-351.

[4]. T. Schneider, Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, Journal of Climate, 14 (2001) 853-871.

[5]. R. Jensen, C. Cornelis, Q. Shen, Hybrid fuzzy-rough rule induction and feature selection, (IEEE2009), pp.1151-1156.

[6]. Pawlak, Z., 1991. Rough Set-Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht.

[7]. 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 University Press.

[8]. Kim, J., Curry, J., 1997. The treatment of missing data in multivariate analysis. Socio-logical Methods and Research 6, 215–241.

[9]. Kim, J., Curry, J., 1997. The treatment of missing data in multivariate analysis. Socio-logical Methods and Research 6, 215–241.

[10]. Ahmed;T.,et.al.,"Data Missing Solution Using Rough Set Theory and Swarm Intelligence", JIJACSIT, Vol.2, No. 3, Page: 1-16, ISSN :2296-1739,2013.

[11]. Yu;J.,et.al.,"Andrzej Skowron Rough Set and Knowledge Technology", 5th International Conference, pp.135 ,2010 .

[12]. Qian,Y., et al.: Multi-granulation decision-theoretic rough sets. Int. J.Approx.Reason.55(1), 225–237(2014)

[13]. Zhang,R., et al.: Rough set attribute reduction algorithm based on tabu discrete particle swarm optimization. Chin.Comput.Syst.38(008),1840–1844(2017).

[14]. He, L., et al.: Hybrid multi-granulation rough sets of variable precision based on tolerance .J. Intell.FuzzySyst.31(2), 717–725(2015)

[15]. Pawlak,Z.: Roughset. Int. J. Comput.Inf. Sci. 11(5), 341–356(1982)

[16]. Pawlak,Z., Skowron,A.: Rudiments of rough sets. Inf. Sci. 177(1), 3–27(2007)

[17]. Jaddi,N., Abdullah,S.: Hybrid of genetic algorithm and great deluge algorithm for rough set attribute reduction. Turk. J. Electr. Eng. Comput.Sci. 21(6), 1737–1750(2013)

[18]. Sang,Y., Qian,Y.: Granular structure reduction approach to multi-granulation decision-theoretic rough sets. Comput.Sci.44(005),199–205(2017).

[19]. Chebrolu,S., Sanjeevi,S.: Attribute reduction in decision-theoretic rough set model using particle swarm optimization with the threshold parameters determined using LMS training rule. Procedia Comput. Sci. 57,527–536(2015)

[20]. Chen, Y, Miao, D, Wang, R and Wu, K. "A Rough Set Approach to Feature Selection Based On Power Set Tree". Knowledge-Based System, Vol. 24, PP. 275–281, 2011.

[21]. Guo, Q.L and Zhang, M. "Implement web learning environment based on data mining", Knowledge-Based Systems, Vol. 22, No. 6, PP. 439-442, 2009.

[22]. Wroblewski, J. "Finding minimal reducts using genetic algorithms". In Proceedings of the second annual join conference on information science, PP. 186-189, 1995.

Cite this article as :

G. V. Suresh, Dr. E. Sreenivasa Reddy, "A New Method for Missing Value Imputation in Incomplete Information Systems Using Hybrid Rough Set Theory", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 580-592, July-August 2020.
Journal URL : https://ijsrcseit.com/CSEIT2064119