Mansoura Journal of Computers and Information Sciences

Missing Value Management: Weighted Heuristic Similarity Estimation for Numeric Values

O. M. Elzeki
Faculty of Computers and Information, Computer Science Dept.
Mansoura University, Egypt
omar_m_elzeki@mans.edu.eg

M. F. Alrahmawy
Faculty of Computers and Information, Computer Science Dept.
Mansoura University, Egypt
mrahmawy@mans.edu.eg

S. Elmogy
Faculty of Computers and Information, Computer Science Dept.
Mansoura University, Egypt
mougy@mans.edu.eg

ABSTRACT

For businesses and technologies such as the Internet of Things (IoT) and digital banking that handles massive volumes of data, it is crucial to have all processed data values accurately recorded; for data values that are not recorded, they must be replaced using a reliable imputation method. The need for missing value imputation is of extreme importance in big data applications as data volumes tend to grow exponentially and their data structures change rapidly. This study proposes a reasonable distance function that is more effective in determining the best replacement values for missing data before applying a classifier on the objective dataset. In essence, the Weighted Heuristic Similarity Estimation mechanism (WHSE) consumes substantial effort in practical application fields. The WHSE method was benchmarked using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics. The evaluation process was conducted using three distinct classifiers: Nearest-Neighbor (NN), Linear-Regression (LR), and Multi-Layer Perceptron (MLP). WHSE method is applied on two different datasets: Iris and Forest Fires to estimate its impact in replacing missing value. Consequently, WHSE formula can direct the applied classifier to score at least similar performance -- if not ideal-- regardless of the characteristics of the imputed data. WHSE method is expected to be scalable, stable and applicable in big data analytics.

Keywords

Rough Sets, Information Gain, Missing Values, Missing Value Imputation, Machine Learning.

1. INTRODUCTION

Developing infrastructures for large-scale systems such as smart grids and cloud computing systems is dependent on the underlying networks. These large-scale infrastructures have been significantly proposed for producing big datasets as substantial streamed amounts of information are generated from an enormous range of connected objects over the network. These big datasets present a raw material for managers and specialists to produce software applications that are user-friendly for them [1], [2]. Generally, marketing, planning, manufacturing, and businesses involve huge volumes of high-variety data that are frequently updated [3], irrespective of what causes their generating application, whether it is a social networking service (SNS) or Internet of Things (IoT), etc [4]. Data in such applications are affected by their loosely coupled environments, which are rapidly changing, and are vulnerable to sensor faults such as a broken sensor, inaccurate sensor readings and so on. Serious conflicts with such datasets may lead to unpredictable information composition due to the existence of null values for some attributes[5]. Such shortcoming is known as the missing values problem and is common in many scientific research domains including biology [6], medicine [7] and climate science [8]. Causes of this problem include improper handling of samples, loss of responses from sensors, measurement error, low signal-to-noise ratio or deletion of abnormal values. Rubin [9] based his definition of missing data on three mechanisms [10]:

1. Missing Completely at Random (MCAR): the missed value of a variable doesn’t rely on missing data or known values (e.g. accidents or administrative errors),
2. Missing at Random (MAR): the variable with a missed value, this value may rely on the known values, but not on its missing data value (e.g. place, missing values of time, etc.), and
3. Missing not at Random (MNAR): the missed value of a variable may rely on a value of another variable (e.g. missing some related characteristics not available in the analysis).

Typically, databases industry could exhibit a substantial amount of missing data because of the system or human errors. The missing values can lead to a complication in the process of data mining whose algorithms couldn't directly be applied to incomplete data [11]. The need for reliable data is a major issue for knowledge discovery, data mining and machine learning[12]. High-quality level of data can be produced by a good data preprocessing; data cleaning is a data preprocessing and it is one of the biggest issues because it leads to miss some values in the database. Therefore, a significant attention by researchers to the process of imputing missed values seeking intelligent imputation algorithms that can be used by intelligent software systems to generate new values to replace the missing values. In such intelligent software systems, scalable classification methods are required for managing the big volumes and/or high velocities of big datasets [13]. Also, these systems need to be able to handle both single-missing value and multi-missing value problems with a reasonable level of accuracy [14].

Missing data, imputation methods, and datasets characteristics may be having an impact on overall performance of algorithms; used for classification [15]. So, it is difficult to get the best possible of an imputation process and classifier combination. As a matter of fact, there is no single existing a classifier and an imputation method combination capable of constantly providing successful classification, since the influence of the imputation process with the classifier differs in accordance with the dataset arranged [16]. The problems become more severe when a system is integrated with robust classifiers while using big data with missing values, as it would result in poor performance due to the constant patterns of data loss.

This paper proposes a novel intelligent imputation method that is based on artificial and statistical inference of missing values, named Weighted Heuristic Similarity Estimation (WHSE), and is capable of optimizing the performance of the classification process. It starts by introducing different mechanisms and causes of missing values occurrence as well as disadvantages of having missing values and incomplete datasets during learning and classification process. Related methods and strategies used as imputation methods are described in Section 2. WHSE method is outlined in Section 3 containing the experiments details and theoretical calculation. In Section 4, the proposed method performance, WHSE, is examined and empirical results are analyzed. Finally, conclusions are drawn and recommendations for future work are made based on the findings of this paper.

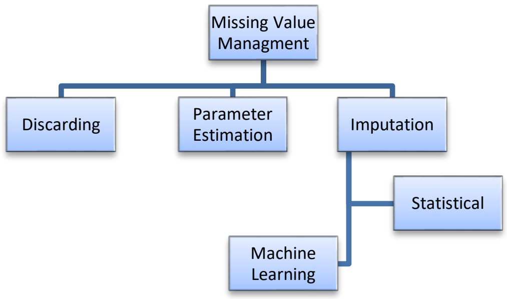

Figure 1. Missing Value Imputation Methods

2. Preliminaries

Many approaches have been developed for the classification of incomplete patterns [17], and they can be generally grouped into three different categories discarding data, parameter estimation, and imputation, as shown in Figure 1. In discarding data, the missing values are removed by complete case analysis, instance ignorance or attribute removal. The process best fits when the missing values are a small subset of instances. On the other hand, parameter estimation uses a model-based technique; e.g. probability density function (PDF), or a model-based method such as neural network ensemble method [18], decision trees [19], fuzzy approaches [20] and support vector machine classifier [21] to immediately overcome missing values without the need for imputing lost instances. The method of imputation is the third category for finding missing values, which is adopted in this paper. In this category, all methods consist of an arrangement of techniques to represent a class which has some estimated values, these values used as an alternative for missing values. A lot of work has been dedicated to imputing missing values using either statistical methods; e.g. regress imputation [10] and mean imputation [22], or machine learning methods; e.g. Fuzzy c-means imputation (FCMI) [23], [24], k-nearest neighbors imputation (KNNI) [17] and self-organizing map imputation (SOMI) [25].

In this paper, it is assumed that for any given information table T, T= (U, F), where U is the universe and F is the set of attributes/features defined by

$$F = A \cup \{c\}$$

where A is the attribute set and $\{c\}$ is the decision. Every single attribute $a \in A$ has a dependency on the class $\{c\}$ with a weight in the range [0, 1], where any independent attribute has weight=0, any core attribute has weight=1, and any other attribute is roughly dependent; i.e. its weight is greater than 0 and less than 1. There are many methodologies used for calculating the dependency weight; however, in this paper, both Rough Sets (RS) and Information Gain (IG) methods are used as weight evaluators. The basics of these methods are presented in the next section.

2.1 Rough Set Theory

RS is considerably a conceptual knowledge processor and is used to detect attribute dependencies for any given knowledge represented by the information table T= (U, F), which has two approximation sets for a given predefined relation Y: Lower approximation ↓, shown in Equation (1.i), and Upper approximation ↑, shown in Equation (1.ii) [26].

$$R \downarrow Y = \{y \in Y \mid [y]_R \subseteq Y\} \qquad \text{Equation (1.i)}$$

$$R \uparrow Y = \{y \in Y \mid [y]_R \cap Y \neq \emptyset\} \qquad \text{Equation (1.ii)}$$

where $[y]_R$ represents the equivalent class.

In contrast, the classification quality of any approximation is defined as in Equation (2).

$$\gamma_R(C) = \frac{|R \downarrow Y|}{|U|} \qquad \text{Equation (2)}$$

For every attribute subset, $A_i \subseteq A$ $\sigma_{RC}(A_i)$ is called importance factor as it refers to the importance of A_i for the decision C, and is calculated using the formula in Equation (3).

$$\sigma_{RC}(A_i) = \gamma_R(C) - \gamma_{A-A_i}(C) \qquad \text{Equation (3)}$$

2.2 Information Gain

The investigation of more succinct decision based on selecting a serious of attributes refers to IG. In this process, the statistical or heuristic measure basis is used to weight these attributes. IG relies on the entropy factor H(C) estimation. This factor used to evaluate the uncertainty degree, where lower and higher entropy refer to lower and higher uncertainty respectively. For any given information table K = (U, A ∪ {c}), Equation (4) used to define the relation of entropy for a given class; [27], C = {c}, and discrete instances $u_i \in U$ where i = 1... ||U||

$$H(C) = -\sum_{i=1}^{\|U\|} p(u_i) * \log_2 p(u_i) \qquad \text{Equation (4)}$$

since $p(u_i)$ is the propability of discrete values of C.

Each decision can be either a consideration of random variable with an arrangement of values has been predefined or a column of information gained through the entropy relation derived from the certainty relation. Equation (4) can be rewritten as an expected information measurement to be used for classification, as in Equation (5).

$$\text{Info}(C) = -\sum_{i=1}^{\|U\|} p(u_i) * \log_2 p(u_i) \qquad \text{Equation (5)}$$

Every attribute a in the table, K, has certain information which can affect the order of the overall attributes relied on the certainty level of information relative to C. Equation (6) defines the expected information value of an attribute a, which splits C into k partitions.

$$\text{Info}_a(C) = \sum_{j=1}^{k} \frac{|C_j|}{|C|} \times \text{Info}(C_j) \qquad \text{Equation(6)}$$

Finally, the total information gain of an attribute a using the class C is defined in Equation (7).

$$\text{Gain}(a) = \text{Info}(C) - \text{Info}_a(C) \qquad \text{Equation (7).}$$

3. The Proposed Methodology

Each dataset has an information table that contains complete data collected from various sources. The knowledge that constitutes the domain of interest can be represented using information table. Classification is an artificial intelligent field for mining in datasets. Our proposed method is mainly connected with various classification algorithms that are specialized in pre-specified fields. The following subsections describe the details of the proposed method.

3.1 Problem Formulation

In general, an information table (I), a knowledge base (K), and a time series (T) or a decision table (S) can be constructed using a nonempty set of finite objects called universe, U, and $F = A \cup \{c\}$, is a nonempty finite set of attributes in the information table. In which, Va is the set of all possible values for attribute a such that a: U→ Va for every a ∈ A. In information table, the missing value problem is defined by two sets: a set Va of complete values and another set V'a of missing values. Each missing value is denoted by a special mark: "?", empty, null, or any other mark. The missing values which may occur at category or class label are also considered.

We proposed a mechanism that adopts a hybrid mechanism where the missed value is imputed using both artificial intelligent and numerical formula in sequence. First, an artificially intelligence algorithm is applied to compute the overall weight of the current attribute using the column data. Next, in the dataset for every attribute, the same algorithm is repeated. A numerical formula is then carried out using the column weight and current row value to derive the new missed value. In the dataset for every attribute, the same process is repeated.

Table 1. A sample of Information Table

U=tᵢ ∈ W	Sr			C
	S1	S2	S3	
t_1	16	16.9	15.1	17.1
t_2	16.2	16.7	14.9	16.8
t_3	16.2	16.8	15.1	16.9
t_4	16	16.9	15.2	17.2
t_5	16.3	16.7	15.5	17.5
t_6	16.2	17	15.1	17.1
t_7	16	17	15.3	?

Data shown in Table 1 is extracted, using a window W = [1, 7], from within a discrete time series. T is the base time series and Sr is a set of three different sensors' readings. In the 7th timestamp, the value Va; i.e. T(7), is missing and it is denoted by (?). Hence, this time series table represents the information table and is defined as K = (T, Sr).

In this example, we intuitively assume that each sequence, or each line in the table, having values closer to the corresponding values in another sequence, should act in the same way. In other words, these values should collectively rise or fall drastically or roughly through the same level. Because of the data frequency heuristic, using a proximity spatial measure between record and collect data sensors under certain adjacent constraints, the measured values can become similar, and we can consider that in each information table, there is a list of n instances. This heuristic desirably completes the task in many instances though there are situations where the assumption purely will never hold.

3.2 Proposed methodology: Weighted Heuristic Similarity Estimation

Relied on the assumption of heuristic mentioned above for imputation methods, for any given information table K = (U, F ⋃ {c}), U is the universe and $u_i \in U$, where i = 1... ||U||; {c} represents the class whose decision is either multi-type or single and is assumed to be an ordinal attribute whereas F denotes the features used for describing instances in K. The total difference $\delta(u_\alpha - u_\beta)$ between two instances u_α having one or more missing values and all other instances u_β in K, where $\alpha, \beta = 1 \ldots \|U\|$ and $\alpha \neq \beta$ can be calculated using the following formula:

$$\delta(u_\alpha - u_\beta) = \sum_{a \in F \quad a \neq b} \sqrt[p]{\frac{\left(W_a * \left|Va(u_\alpha) - Va(u_\beta)\right|\right)^p}{s}} \qquad \text{Equation(8)}$$

where p is the power of the root operator, s is a variable scalable, W_a is a pre-calculated of dependency between the class {c} in table K and the attribute a, and $V_a(u_\alpha)$ denotes the assigned value to attribute a in the instance u_α.

The missing value in a certain instance u_α can be obtained using Equation 8 as follows: First, if there exists an instance of u_β in table K whose attribute values are closest to the existing values of the attributes of u_α, then the instance is recorded. Thus, the algorithm imputes the missing values in the instances u_α with their corresponding values found in u_β.

Figure 2 shows the overall process of imputing the missing values using the proposed methodology, WHSE. First, the process starts using information table which contains missing values; remarked by "?". The given table is splatted into complete instances and incomplete instances. Next, the complete instances set is used as ground truth samples to determine the dependency relations between the attributes and the associated class label. The calculated weight vector represents a weighting factor for every value in the feature vector. Using the weight vector, imputation stages is started for each imputing the incomplete instances. In which, for every incomplete instance, WHSE is applied to obtain the total difference distance between the current incomplete instance, u_α, and every complete instance, u_β where $\beta = 1 \ldots \|\text{complete instances}\|$. Then, the distance list is sorted in ascending order using quick sort to find the closest instance of complete instances to the incomplete, u_α. Finally, the missed value is assigned to be the same value already existed for the chosen instance.

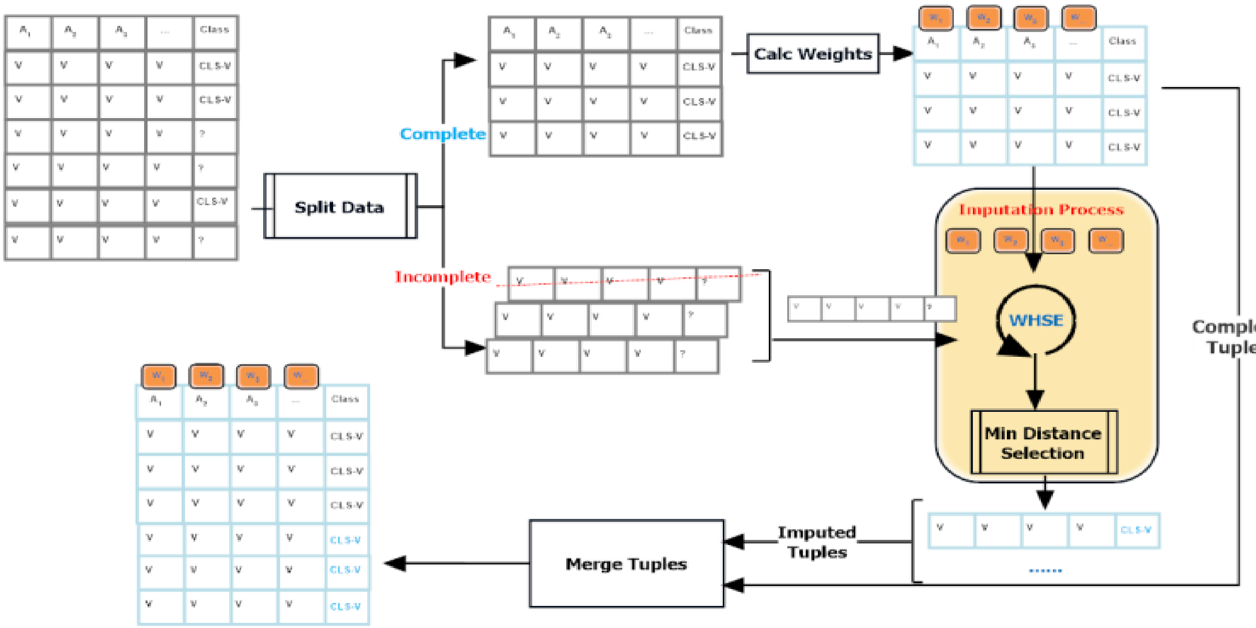

Figure 2. The proposed framework using WHSE method for imputing the missing class

3.3 Proof of Correctness

With the aim of showing the effectiveness and correctness of WHSE method for imputing the missing values, it is assumed that for any two values of the same attribute of two different instances u_α and u_β, the distance $\delta_a(u_\alpha - u_\beta)$ between them is calculated as the absolute difference between two distinct values $Va(u_\alpha)$ and $Va(u_\beta)$, as shown in Equation (9).

$$\delta_a(u_\alpha - u_\beta) = |Va(u_\alpha) - Va(u_\beta)| \qquad \text{Equation(9)}$$

This formula can be written in general form for any power value p as follows:

$$\delta_a(u_i - u_j) = \sqrt[p]{|Va(u_\alpha) - Va(u_\beta)|^p} \qquad \text{Equation(10)}$$

So, we can easily show that p=1, we can write

$$\sqrt[p]{|Va(u_\alpha) - Va(u_\beta)|^p} = |Va(u_\alpha) - Va(u_\beta)| \qquad \text{Equation(11)}$$

Now, for the weighting factor W_a, where $0 \le W_a \le 1$, we can easily conclude that $0 \le \sqrt[p]{W_a} \le 1$. Hence, embedding the value of $\sqrt[p]{W_a}$ in the left side of Equation (11) results in the inequality presented in Equation (12).

$$\sqrt[p]{(W_a * |Va(u_\alpha) - Va(u_\beta)|)^p} \le$$

$$|Va(u_\alpha) - Va(u_\beta)| \qquad \text{Equation(12)}$$

Also, for any scalar value s, if $s \ge 1$, it is trivial to show that $\sqrt[p]{\frac{1}{s}} \le 1$. So, by embedding the value $\sqrt[p]{\frac{1}{s}}$ in the left-hand side of Equation (12), the inequality shown in Equation (13) is obtained.

$$\sqrt[p]{\frac{(W_a * |Va(u_\alpha) - Va(u_\beta)|)^p}{s}} \le |Va(u_\alpha) - Va(u_\beta)| \qquad \text{Equation(13)}$$

Applying the sum operand on both sides of Equation (13) leads to the inequality shown in Equation (14).

$$\sum_{a \in F, \alpha \neq \beta} \sqrt[p]{\frac{(W_a * |Va(u_\alpha) - Va(u_\beta)|)^p}{s}} \le \sum_{a \in F, \alpha \neq \beta} |Va(u_\alpha) - Va(u_\beta)| \qquad \text{Equation(14)}$$

since $\sum_{a \in F, \alpha \neq \beta} |Va(u_\alpha) - Va(u_\beta)|$ represents the base form of Euclidian distance and $\sum_{a \in F, \alpha \neq \beta} \sqrt[p]{\frac{(W_a * |Va(u_\alpha) - Va(u_\beta)|)^p}{s}}$ is WHSE distance measure.

From Equation (14), it is apparent that the proposed, WHSE, method can result in more accurate values by using Equation (8).

3.4 Evaluation of the Effectiveness

We have used two effectiveness metrics to evaluate WHSE imputation method, these metrics called root mean squared error (RMSE) and mean absolute error (MAE). MAE is the measure between the true and the predicted value; it is a comparing between the imputed values that been estimated with their final outcomes or the provided true values using a classifier. On the other side, RMSE is probably the most usually recognized metrics for assessing a classifier's reasoning precision for both numeric or nominal classes. MAE and RMSE can be computed according to Equation (15) and Equation (16), respectively:

$$\mathrm{MAE} = \frac{\sum |x_i - y_i|}{N} \qquad \mathrm{Equation}(15)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum |x_i - y_i|^2}{N}} \qquad \mathrm{Equation}(16)$$

where x_i and y_i refers respectively to each individual actual and predicted values of the missed values in the given dataset(s). Next, how these two metrics are used to evaluate WHSE imputation method will be explained.

4. Implementation & Evaluation

4.1 Datasets

In order to evaluate WHSE imputation method, we used two datasets: Iris [28] and Forest Fires [29]; these two datasets are retrieved from the UCI online repository [30] and their details and important characteristics are shown in Table 2. These two datasets are chosen as both are used in multiclass classification problems and are used extensively to evaluate innovative knowledge mining algorithms along with options. Each one of these two datasets are composed of a row known as instance and a set of columns called attributes, where the Iris dataset is composed of 150 instances with 4 numeric attributes, whereas the second dataset, Forest Fires, is composed of 12 attributes (10 numeric and one nominal category of two fields) with 517 instances. To measure the effectiveness of the WHSE imputation method, missing values are arbitrarily generated in some cells and randomly selected in some instances. Randomly missing values were marked missing by setting their values to be "?".

Table 2. Summary details of the used datasets

Description	Forest Fires (FF)	Iris (I)
Features	Multivariate	Multivariate
Data Types	Real	Real
No. of samples	150 instances	517 instances
No. of features	4 conditions	13 conditions
Has missing values	Not Available	Not Available
Domain	Physical	Life

4.2 Implementation Details

Our implementation environment was built on a computer with an i7 processor and 8 GB DDRAM and running Ubuntu 14.04, where the implementation was conducted on Python 2.7 using SciPy, NumPy libraries. The imputation accuracy is evaluated by comparing the efficiency of WHSE against that of the Euclidian distance estimation method. The output of each of these two methods is fed to three different WEKA 3.7.2 classifiers [31]: Simple Neural Network (NN) classifier, Linear Regression (LR) classifier, and Multilayer Perceptron (MLP) classifier in order to compare the precision of the two methods. A cross-validation testing with 10 folds was used.

The same two evaluation metrics, namely MAE and RMSE, were used to evaluate both the precision of the imputed values using WHSE method and the impact of the imputed values on the precision of the classifiers. First, the metrics were used to measure how close the imputed values generated by the WHSE and Euclidean distance methods are to the original labels in the data set. Then, we use the same metrics to evaluate the impact of the use of the imputed values on the overall precision of the used classifiers. In order to perform a fair comparison and transparent empirical analysis of the experimental results, normalization procedures are applied on MAE and RMSE. In turn, subsequent tables represent NMAE and NRMSE instead of the standards.

In each experiment, the distance is measured using three different weights: default (1), IG-based, RS-based, and using three different values (1, 2, and 3) of the power p. The experiments were made on three classification algorithms resulting in a total of 27 experiments used for the evaluations.

4.3 Precision Evaluation of the Imputed missing values

The results of this set of experiments are shown in Table 3. There are nine cases: three cases when W = 1, the default Euclidian distance and the other six cases from using the IG-based and RS-based weights to calculate WHSE is shown in Equation 8 and three different values for the parameter p (1, 2 and 3). The experiments show that the parameter p has no effect on the precision of the imputed values, as the results are identical at p=1, 2 and 3. Hence, it is clear that any changes in p have no effect on the precision and, for simplicity, using p=1 would be the best choice when applying this method. On the contrary, as seen in Table 3, for the parameter W and both datasets, the use of IG-based weights scores the best distance values among all the methods: default (1), IG-based, RS-based, whereas the RS-based weights always result in a distance less than IG-based weights, but its scored distance was always either less than or equal to the distance scored by using the default distance (W=1). In summary, we recommend WHSE method using the IG-based weights and p=1 irrespective of the used dataset in order to get the best possible precision.

Table 3. Evaluation of the suggested imputed values

Settings		NMAE		NRMSE	
p	Wa	(FF)	(I)	(FF)	(I)
1	1	0.003958	0.020168	0.048527	0.053364
	IG	0.003768	0.017566	0.047705	0.047596
	RS	0.003958	0.019664	0.048527	0.052721
2	1	0.003958	0.020168	0.048527	0.053364
	IG	0.003768	0.017566	0.047705	0.047596
	RS	0.003958	0.019664	0.048527	0.052721
3	1	0.003958	0.020168	0.048527	0.053364
	IG	0.003768	0.017566	0.047705	0.047596
	RS	0.003958	0.019664	0.048527	0.052721

4.4 Precision Evaluation of the imputed datasets

The impact of WHSE method on the precision of the classification is evaluated using three conventional data classifiers: NN, LR, and MLP. Table 4 shows the results of the evaluation using MAE and RMSE distance metrics. There are 9 cases representing the base standard Euclidian distance; i.e. W=1, regardless of the value of p. The other 18 cases represent the use of WHSE method using IG-based and RS-based weights when p = 1, 2 and 3.

Table 4. Evaluation of the imputed dataset using different data mining techniques

Settings		Classifier	NMAE		NRMSE	
p	Wa		(FF)	(I)	(FF)	(I)
1	1	NN	0.065568	0.192654	0.229856	0.218419
		LR	0.066909	0.184509	0.230398	0.210301
		MLP	0.087508	0.193319	0.252876	0.232047
	IG	NN	0.066885	0.19638	0.230392	0.223929
		LR	0.066885	0.185414	0.230392	0.211259
		MLP	0.084803	0.197125	0.247146	0.242614
	RS	NN	0.065568	0.194677	0.229856	0.220415
		LR	0.066909	0.184642	0.230398	0.210008
		MLP	0.087508	0.195741	0.252876	0.234416
2	1	NN	0.065568	0.192654	0.229856	0.218419
		LR	0.066909	0.184509	0.230398	0.210301
		MLP	0.087508	0.193319	0.252876	0.232047
	IG	NN	0.065424	0.19638	0.229825	0.223929
		LR	0.066885	0.185414	0.230392	0.211259
		MLP	0.084803	0.197125	0.247146	0.242614
	RS	NN	0.065568	0.194677	0.229856	0.220415
		LR	0.066909	0.184642	0.230398	0.210008
		MLP	0.087508	0.195741	0.252876	0.234416
3	1	NN	0.065568	0.192654	0.229856	0.218419
		LR	0.066909	0.193319	0.230398	0.232047
		MLP	0.087508	0.193319	0.252876	0.232047
	IG	NN	0.065424	0.19638	0.229825	0.223929
		LR	0.066885	0.185414	0.230392	0.211259
		MLP	0.066885	0.197125	0.230392	0.242614
	RS	NN	0.065568	0.194677	0.229856	0.220415
		LR	0.066909	0.184642	0.230398	0.210008
		MLP	0.087508	0.195741	0.252876	0.234416

As shown in Table 4, the use of WHSE method with either of IG-based or RS-based weights, in most cases, results in a distance less than that scored by the Euclidian distance. The high accuracy is obtained because the suggested imputed values using WHSE strategy are more precise than the suggested values using the standard Euclidian method.

In summary, WHSE method is an efficient imputation method which guides the applied classifiers to achieve better, or at least equal, classification precision, regardless of dataset characteristics.

4.5 Discussion and future work

The proposed method suggests a novelty of the missing values in addition to suggest a meaningful values imputation for covering patterns, WHSE, plays a major benefit over traditional methods. In turn, a computable, reasonable, scalable, and stable method for obtaining ideal values is needed before classifiers are applied. From the results of the experiment of the conducted in this study, it is recommended that WHSE method is integrated into good-sized records analytics.

Sim at. al. [32] reported that the characteristics of data and the performance of the classifier have a relationship. They suggest only an adaptive matching classifier and imputation (AMCI) rely on classifier and imputation method without proposing a strategy to handle these changes. In AMCI, the ideal mixture of classification and imputation procedures is versatile selected to attest sufficient performance when identifying the data sets characteristics. Though literature reviews recently have led researchers to study the effect of the imputation process with distinct effectiveness, only two classifiers were proposed: genetic algorithms [33] or the nearest neighbor rule [34], [35]. Other researchers have assumed in their studies that there are no missing values and hence, the end result of their work may not be reliable in all circumstances [36]. Furthermore, quite a few earlier types of research were unable to identify a collection involving a classifier and its applications [15], [37].

Basically, the literature shows that there are several missing ideal designs, like horizontal dispersing, top to bottom scattering together with a higher standard distribute, substance metric and imbalance rate together with missing ratio represented [2]. In contrast, this study is also thought-out to be the first work to handle the dependency between attributes and fields as a characteristic of the missing pattern of values. According to the knowledge that we have gained, there no study has ever proposed a soft artificial, reasonable method to impute missing values prior to the classifier have been applied.

Moreover, WHSE method is also scalable, as the used datasets have various sizes, which happen to have excessive volume. It does not depend on the size of the dataset but on the pre-calculation of the columns' weights of the imputation procedure. The weights are calculated using IG or RS at fixed runtimes, $\theta(f(\mathrm{IG}))$, $\theta(f(\mathrm{RS}))$ respectively, and for estimating the missing value, it requires the execution of only multiplication, summation and subtraction operation per column, which are basic operations and have constant time; $\theta(\mathrm{C})$. In turn, the total runtime required for imputing a single instance $\mathrm{u_i}$ in a given information table $\mathrm{K} = (\mathrm{U.A})$ can be formalized as $\theta(\text{imputing\_instance}) = \theta\big(\mathrm{f}(\mathrm{IG})\big) + |\mathrm{A}| * \theta(\mathrm{C})$ when using IG as weight evaluator or $\theta(\text{imputing\_instance}) = \theta\big(\mathrm{f}(\mathrm{RS})\big) + |\mathrm{A}| * \theta(\mathrm{C})$ when using RS as weight evaluator. For imputing all missing instances n, costs $= \mathrm{n} * \theta(\text{imputing\_instance})$. This is a clear indication that WHSE method can handle the missing values in the dataset with no increase in runtime. The overall performance of imputation is changed relevant to the number of instances that have missing valuesn, which is also constant. Hence, WHSE method is incredibly realistic and is workable with big data analytic options.

5. Conclusion

It is concluded that our recommended method can be viewed as an artificial intelligent imputation approach for applicable data analytics of real-time fields as a valuable application for its ideal reasonability, accuracy, stability, scalability and minimized costs. We already identified that at this time there are in existence quite a few factors that will need to be considered in future studies such as processing time, parallel computing and nominal fields imputation. Additionally, though this study suggests further inquiries for missing value imputation, the weight evaluator still needs to be closely examined if WHSE method were to be implemented more passively to substantial internet real-time big data software and applications. Thirdly, it is recommended that different classifiers be evaluated using different datasets to improve the effectiveness of WHSE method in application domains. Finally, WHSE imputation method did not consider nominal values since it is beyond the scope of this study and is planned to be dealt with in future research.

6. REFERENCES

[1] M. A. Nia, R. E. Atani, and A. K. Haghi, “Ubiquitous IoT structure via homogeneous data type modelling,” in Telecommunications (IST), 2014 7th International Symposium on, 2014, pp. 283–288.

[2]K. Wellenzohn, H. Mitterer, J. Gamper, M. H. Böhlen, and M. Khayati, “Missing Value Imputation in Time Series Using Top-k Case Matching.,” in Grundlagen von Datenbanken, 2014, pp. 77–82.

[3]A. Bifet, “Mining big data in real time,” Informatica, vol. 37, no. 1, 2013.

[4] M. Chen, S. Mao, Y. Zhang, and V. C. Leung, “Big data applications,” in Big Data, Springer, 2014, pp. 59–79.

[5] S. Edge, D. Byrd, C. Compton, A. Fritz, F. Greene, and A. Trotti, “Cancer staging manual,” American Joint Committee on Cancer (AJCC). 7th ed. New York: Springer, 2010.

[6] O. Troyanskaya et al., “Missing value estimation methods for DNA microarrays,” Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[7] R. J. Little et al., “The prevention and treatment of missing data in clinical trials,” New England Journal of Medicine, vol. 367, no. 14, pp. 1355–1360, 2012.

[8] T. Schneider, “Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values,” Journal of Climate, vol. 14, no. 5, pp. 853–871, 2001.

[9] D. B. Rubin, “Inference and missing data,” Biometrika, vol. 63, no. 3, pp. 581–592, 1976.

[10] D. B. Rubin and R. J. Little, “Statistical analysis with missing data,” Hoboken, NJ: J Wiley & Sons, 2002.

[11] B. Zhu, C. He, and P. Liatsis, “A robust missing value imputation method for noisy data,” Applied Intelligence, vol. 36, no. 1, pp. 61–74, 2012.

[12] D. V. Patil and R. Bichkar, “Multiple imputation of missing data with genetic algorithm based techniques,” IJCA Special Issue on" Evolutionary Computation for Optimization Techniques, pp. 74–78, 2010.

[13] Y.-J. Jang and J. Kwak, “Social network service real time data analysis process research,” in Frontier and Innovation in Future Computing and Communications, Springer, 2014, pp. 643–652.

[14] K.-H. Kim and W. Tsai, “Social comparison among competing firms,” Strategic Management Journal, vol. 33, no. 2, pp. 115–136, 2012.

[15] J. Sim, J. S. Lee, and O. Kwon, “Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications,” Mathematical Problems in Engineering, vol. 2015, 2015.

[16] A. Farhangfar, L. Kurgan, and J. Dy, “Impact of imputation of missing values on classification error for discrete data,” Pattern Recognition, vol. 41, no. 12, pp. 3692–3705, 2008.

[17] G. E. Batista, M. C. Monard, and others, “A Study of K-Nearest Neighbour as an Imputation Method.,” HIS, vol. 87, no. 251–260, p. 48, 2002.

[18] P. K. Sharpe and R. Solly, “Dealing with missing values in neural network-based diagnostic systems,” Neural Computing & Applications, vol. 3, no. 2, pp. 73–77, 1995.

[19] J. R. Quinlan, “Induction of decision trees,” Machine learning, vol. 1, no. 1, pp. 81–106, 1986.

[20] R. J. Hathaway and J. C. Bezdek, “Fuzzy c-means clustering of incomplete data,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 31, no. 5, pp. 735–744, 2001.

[21] K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor, “Handling missing values in support vector machine classifiers,” Neural Networks, vol. 18, no. 5, pp. 684–692, 2005.

[22] D. J. Mundfrom and A. Whitcomb, “Imputing Missing Values: The Effect on the Accuracy of Classification.,” 1998.

[23] J. Luengo, J. A. Sáez, and F. Herrera, “Missing data imputation for fuzzy rule-based classification systems,” Soft Computing, vol. 16, no. 5, pp. 863–881, 2012.

[24] D. Li, J. Deogun, W. Spaulding, and B. Shuart, “Towards missing data imputation: a study of fuzzy k-means clustering method,” in International Conference on Rough Sets and Current Trends in Computing, 2004, pp. 573–579.

[25] F. Fessant and S. Midenet, “Self-organising map for data imputation and correction in surveys,” Neural Computing & Applications, vol. 10, no. 4, pp. 300–310, 2002.

[26] S. Kumar, N. Jain, and S. L. Fernandes, “Rough set based effective technique of image watermarking,” Journal of Computational Science, vol. 19, pp. 121–137, 2017.

[27] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, “Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model,” Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[28] “UCI Machine Learning Repository: Iris Data Set.” [Online]. Available:https://archive.ics.uci.edu/ml/datasets/iris. [Accessed: 08-October-2018].

[29] “UCI Machine Learning Repository: Forest Fires Data Set.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/forest+fires. [Accessed: 08-October-2018].

[30] “UCI Machine Learning Repository: Data Sets.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html. [Accessed: 08-October-2018].

[31] “Weka 3 - Data Mining with Open Source Machine Learning Software in Java.” [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 08-October-2018].

[32] J. Sim, O. Kwon, and K. C. Lee, “Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets,” Expert Systems with Applications, vol. 46, pp. 485–493, 2016.

[33] C. T. Tran, P. Andreae, and M. Zhang, “Impact of imputation of missing values on genetic programming based multiple feature construction for classification,” in 2015 IEEE Congress on Evolutionary Computation (CEC), 2015, pp. 2398–2405.

[34] C. Jiang and Z. Yang, “CKNNI: An Improved KNN-Based Missing Value Handling Technique,” in Advanced Intelligent Computing Theories and Applications: 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, 2015. Proceedings, Part III, D.-S. Huang and K. Han, Eds. Cham: Springer International Publishing, 2015, pp. 441–452.

[35] T. Orczyk and P. Porwik, “Investigation of the Impact of Missing Value Imputation Methods on the k-NN Classification Accuracy,” in Computational Collective Intelligence: 7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part II, M. Núñez, N. T. Nguyen, D. Camacho, and B. Trawiński, Eds. Cham: Springer International Publishing, 2015, pp. 557–565.

[36] S. Jones, D. Johnstone, and R. Wilson, “An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes,” Journal of Banking & Finance, vol. 56, pp. 72–85, 2015.

[37] P. Kang, “Locally linear reconstruction based missing value imputation for supervised learning,” Neurocomputing, vol. 118, pp. 65–78, 2013.