Research Article

Clustering-Based Multiple Imputation via Gray Relational Analysis for Missing Data and Its Application to Aerospace Field

Jing Tian, Bing Yu, Dan Yu, and Shilong Ma

State Key Laboratory of Software Development Environment, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China

Correspondence should be addressed to Jing Tian; tianjing@nlsde.buaa.edu.cn

Received 7 March 2013; Accepted 9 April 2013

Academic Editors: Y.-P. Huang, P. Melin, M. F. G. Penedo, and D. Rodriguez

Copyright © 2013 Jing Tian et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A large number of scientific researches and industrial applications commonly suffer from missing data. Some inappropriate techniques of missing value treatment compromise data quality, which detrimentally influences the knowledge discovery. In this paper, we propose a missing data completion method named CBGMI. Firstly, it separates the nonmissing data instances into several clusters by excluding the missing-valued entries. Then, it utilizes the entropy of the proximal category for each incomplete instance in terms of the similarity metric based on gray relational analysis. Experiments on UCI datasets and aerospace datasets demonstrate that the superiority of our algorithm to other approaches on validity.

1. Introduction

In a variety of application domains, machine learning and data mining algorithms proved to be of great value [1–3]. However, people using real-world databases or datasets repeatedly encounter the data imperfection issue in the form of incompleteness [4, 5]. Therefore, a plenty of resolutions have been devised to cope with the unfavorable phenomenon. Despite the fact that missing data might not cause any major issue particularly when the missing rate is not significantly high, it could not be considered as an ideal case to ensure the data quality. Additionally, some people argue that the fragmentary data should be excluded from further consideration. Nevertheless, the opinion remains as an obvious shortcoming which is articulated that the other observed factual values of the same instance may simultaneously be absent [6]. In some high-missing-rate environment, this strategy is presumed unreasonable and infeasible. Consequently, the handling for substitution or replacement draws increasing attentions, termed as imputation. In broad outline, the methods available can be separated into two categories: single imputation and multiple imputation methods. On one hand, single imputation, that is, filling in precisely one value for each missing one, intuitively has many appealing features. For example, standard complete-data methods can be applied directly, and

the substantial effort required to create imputations needs to be carried out only once [7, 8]. On the other hand, multiple imputation generates a quantity of simulated values for each missing item, in order to reflect properly the uncertainty attached to missing data [9, 10]. This has been advocated as a statistically sound approach, but so far its use has been limited mainly to the social and medical sciences.

Recently, multiple imputation has emerged as an interesting and quite visible alternative in missing data analyses. The versions of the sophisticated approach are advantageous to these conventional techniques because they require less stringent assumptions and mitigate the pitfalls of traditional ones [11, 12]. Nevertheless, for the most part of existing solutions, the following facets retain defective. (a) The clustering strategy combining complete instances with incomplete instances violates the formation of good clusters. In other words, the entire instances involved clustering generates unbiased values due to the imperfection [13]. (b) The traditional *Minkowski's* L_p ($p = 1, 2, \infty$) metric is imprecise to scale the similarity among different instances. (c) The current methodologies are hardly applicable to handling the missing aerospace data due to the their underperformance on validity.

In this paper, data imputation is formulated as a problem of estimation of missing values by multiple operations based

on clustering. Furthermore, the prime contribution of this paper could be described as follows. (a) Dividing nonmissing items into a finite number of well-partitioned clusters contributes to make the completion in the optimal tailored area. (b) The gray relational analysis, which signifies the situational variation of the curve, could characterize the relative discrepancy more precisely. (c) CBGMI is adopted to the practical aerospace with accurate performance.

The rest of this paper is structured as follows. Section 2 first briefly introduces both the missingness mechanisms and the emblematic patterns of missing data treatment and then reviews the diverse-related literatures about imputation. In Section 3, the detailed process of the CBGMI algorithm is illustrated in three primary procedural subitems. Section 4 demonstrates a series of experimental results on both UCI datasets and empirical aerospace datasets to compare the performance with other methods. Finally, conclusions are given in Section 5.

2. Related Work

2.1. Missingness Mechanisms. Three different mechanisms, which lead to the introduction of missing values, can be categorized as follows [13, 14].

2.1.1. Missing Completely at Random (MCAR). When the distribution of an example having a missing value for an attribute does not depend on either the observed data or the missing data. When MCAR happens, evidently the set of subjects with no missing data is also a random sample from the source population. Hence, most simple techniques for handling missing data, including complete and available case analyses, yield unbiased results.

2.1.2. Missing at Random (MAR). This mechanism establishes once the distribution of an example having a missing value for an attribute depends on the observed data but does not depend on the missing data. Under this scenario, a complete or available case analysis is no longer based on a random sample from the source population, and selection bias likely occurs. Generally, when missing data are MAR, all simple techniques for handling missing data give biased results.

2.1.3. Not Missing at Random (NMAR). It implies that the pattern of data missingness is nonrandom, and it is not predictable from other variables in the database. If missing data are NMAR, valuable information is lost from the data and there are no universal methods of handling the missing data properly. For instance, participants who are unsatisfied with their company are more likely to not answer the questions about company satisfaction.

2.2. Methods for Missing Data Analysis. Current managements of processing missing data can be approximately divided into three categories: tolerable procedures, procedures based on deletion of cases, and imputation-based procedures.

2.2.1. Tolerance. The straightforward method aims to maintain the source entries in the incomplete fashion. Consequently, the ulterior analysis is directly designed based on the raw data [15]. It is poor when the percentage of missing values per attribute varies considerably.

2.2.2. Ignoring. Missing data ignorance often refers to “*Case Deletion*.” It is the most frequently applied procedure, but it is prone to diminish the data quality. Its strength lies in the ease of application; it simply proposes to delete elements with missing data. The procedure can be applied in two manners [12, 16].

- (a) Listwise/Casewise Deletion: it performs indiscriminately deleting from the database any elements with missing data for any of the attributes being examined.
- (b) Pairwise Deletion: incomplete cases are removed on an analysis-by-analysis basis, such that any given case may contribute to some analyses but not to others.

2.2.3. Imputation

Mean/Mode Substitution (MMS). This is a simple way to impute the missing data. It replaces the missing values by the mean or mode of all the observations or a subgroup at the same variable. It consists of replacing the unknown value for a given attribute by the mean (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute. Replacing all missing records with a single value distorts the input data distribution [15].

Hot-Deck/Cold-Deck Imputation [13]. Given an incomplete pattern, *hot-deck imputation* (HDI) replaces the missing data with the values from the input vector that is the closest in terms of the attributes that are known in both patterns. This method attempts to preserve the distribution by substituting different observed values for each missing item. Another possibility is the *cold-deck imputation* (CDI) method, which is similar to hot deck, but the data source must be other than the current dataset. For example, in a survey context, the external source can be a previous realization of the same survey.

Regression Imputation. This method uses multiple linear regression to obtain estimates of the missing values. It is applied by estimating a regression equation for each variable, using the others as predictors. This solves the problems concerning variance and covariance raised by the previous method but leads to polarization of all the variables if they are not linked in a linear fashion. Possible errors are due to the insertion of highly correlated predictors to estimate the variables. The advantage of this method is that existing relationships between the variables can be used to calculate missing data, but it is rarely used as it amplifies the correlation between variables [13, 16].

Expectation Maximization Estimation (EME). The technique is on the basis of *expectation maximization* (EM) algorithm proposed by Dempster, Laird, and Rubin. The algorithm can handle parameter estimation in the presence of missing

data. These methods are generally superior to case deletion methods, because they utilize all the observed data. However, they suffer from a strict assumption of a model distribution for the variables, such as a multivariate normal model, which has a high sensitivity to outliers [2, 14].

Machine Learning-Based Imputation. It acquires the features of interested unknown data by behavior evolution after sample data processed. The essence is to automatically learn sample for complicated pattern cognition and intelligently predict the missing values. The methods mainly includes decision tree-based imputation, association rules-based imputation, clustering-based imputation, and so forth [6, 8, 17].

Multiple Imputation. It replaces each missing value with two or more plausible values that represent the uncertainty about the right value to impute. Each of the two or more resulting complete datasets is then analyzed using standard complete-data methods. All the analyses become combined to reflect both the interimputation variability and intrainputation variability [16, 18, 19].

2.2.4. State-of-the-Art for Missing Data Imputation. Statistical analysis with missing data has been noted in the literature for more than 70 years. Allison [13] pointed that Walks initiated a study on the maximum likelihood estimation for multivariate normal models with fragmentary data. Thereafter, extensive discussions on this topic continue. A useful reference for general parametric statistical inferences with missing data can be found in Little and Rubin [14].

Zhu et al. [19] made use of Magnani's reviewing on the main missing data techniques, including conventional methods, global imputation, local imputation, parameter estimation, and direct management of missing data. They tried to highlight the advantages and disadvantages for all kinds of missing data mechanisms. However, the main problem of these techniques is the need for strong model assumptions. In recent years, many researchers focused on the topic of imputing missing values. S. M. Chen and H. H. Chen [20] developed an estimating null value method, where a fuzzy similarity matrix is used to represent fuzzy relations, and the method is used to deal with one missing value in an attribute. Embedded methods consist of casewise deletion, lazy decision tree, dynamic path generation, and some popular methods such as C4.5 and CART. Nonetheless, these methods are not a completely satisfactory way to handle missing value problems [21]. Firstly, they are merely designed to deal with the discrete values, and the continuous ones are discretized before imputing the missing value, which may lose the true characteristic during the converting process from the continuous value to discretized one. Secondly, they usually studied the problem of missing covariance.

Huang and Lee [22] employed a grey-based nearest neighbor method to handle the missing data problem. In their opinion, the gray association analysis is employed to determine the nearest neighbors of an instance with missing values. And those unknown values are inferred by the known attribute values derived from these nearest neighbors. Hruschka Jr. et al. [23] used Bayesian networks to

FIGURE 1: The flowchart of the CBGMI algorithm.

fulfill missing values in a hybrid model, which applies the clustering genetic algorithm in objects without missing values and generates Bayesian networks to substitute the missing values. Chen and Huang [24] used the weighted fuzzy rules to estimate null values in relational database. Li et al. [25] borrowed the idea from fuzzy K -means clustering and applied it to the problem of missing data imputation with superior performance to the basic K -means especially when the percentage of missing values are high. Meesad and Hengpraprom [26] combined K -nearest neighbor-based feature selection and K -nearest neighbor-based imputation, including feature selection and estimation of new values. The results showed that the proposed method had powerful estimation ability on microarray datasets. Di Nuovo [2] made the comparisons among four solutions of the fuzzy c -means (FCM) in the psychological research environment. The result revealed that the FCM based on optimal completion strategy leads to effective data imputation instead of deleting elements with missing values. Zhang et al. [18] utilized the information within the incomplete instances since the second imputation iteration. The *non-parametric iterative imputation* (NIIA) is an improvement of the classic multiple imputation, which is based on kernel function. The experimental results on UCI datasets unfolded that the NIIA could easily capture the distribution of a dataset even when there is no prior knowledge of the datasets.

3. The CBGMI Algorithm

In this section, the global procedure of the CBGMI algorithm is schematized in Figure 1. And each of the key components is detailed correspondingly. Firstly, the clustering technique is explained, and then the computation of gray relational analysis with missing values is presented with formulations. After

the instructions of the entropy-based multiple imputation comprising of initial and successive estimations, the entire algorithmic information descriptively listed.

3.1. The Clustering Strategy. The specific clustering schema utilizes the standard FCM [1, 27], which attempts to minimize the following objective function with respect to fuzzy memberships $U^{(r)} = [u_{ij}^{(r)}]$ and cluster centroids $C^{(r)} = c_j^{(r)}$: $J = \sum_{j=1}^G \sum_{i=1}^M (u_{ij}^{(r)})^s d(x_i, c_j^{(r)})$. In the function, r is the ordinal number of the iterations with x_k and $c_j^{(r)}$, respectively, denoting the k th complete data instance and the j th cluster, while $d(\cdot, \cdot)$ is the distance metric between two instances, and $u_{ij}^{(r)}$ is the degree of membership in which the i th instance is subordinate to the j th cluster under the “fuzzier” s , as G defines the total number of clusters, and M represents the number of data instances. The algorithm would immediately end with formed clusters under the circumstance that $\|U^{(r)} - U^{(r-1)}\| < \varepsilon$ or r accumulatively reaches the predefined number.

3.2. The Classification of Incomplete Instances. Each of the incomplete instances is individually incorporated into the closest cluster according to the maximal value of *gray relational grade* (GRG) [22] in (1) and (2)

$$\begin{aligned} \text{GRC}(x_k^{\text{mis}}, c_i) &= \frac{\min_M \min_N |x_{kp}^{\text{mis}} - c_{ip}| + \zeta \max_M \max_N |x_{kp}^{\text{mis}} - c_{ip}|}{|x_{kp}^{\text{mis}} - c_{ip}| + \zeta \max_M \max_N |x_{kp}^{\text{mis}} - c_{ip}|} \\ i &= 1, 2, \dots, M; \quad p = 1, 2, \dots, N; \quad 0 \leq \zeta \leq 1, \end{aligned} \quad (1)$$

where x_k^{mis} is the k th incomplete instance and p is the p th attribute with nonmissing values, while N is the number of attributes, and coefficient ζ is used to decrease the effect of $\max_M \max_N$, which is the maximal value in the matrix.

Consider the following:

$$\text{GRG}(x_k^{\text{mis}}, c_i) = \frac{1}{N} \sum_{p=1}^N \text{GRC}(x_k^{\text{mis}}, c_i), \quad i = 1, 2, \dots, M. \quad (2)$$

3.3. The Entropy-Based Multiple Imputation. When each time one instance has been assigned to the most proximate group, an internal multiple imputation strategic approach starts as follows.

3.3.1. First Imputation. Algorithms like C4.5 and kNN could be used in the initial round of imputation. Although the MMS was doubted for its potential bias in terms of distributions in the situation, Zhang et al. [18] emphasized that the value of such imputation would be reasonable unless it runs the extraiterative imputation. For this reason, MMS is employed to initialize missing values in the first imputation.

3.3.2. Successive Imputation. $R = (r_{ij})_{m \times n}$ associates with the data matrix of the cluster, into which $x_i^{\text{mis}} \in X_{ic}$ is attached. That is, it includes $m-1$ complete elements and one initialized element.

Step 1. Calculate the entropy value of the f th data instance [3]

$$\begin{aligned} I_f &= -k * h_f * \ln p_f, \\ k &= \frac{1}{\ln m}, \quad h_f = \frac{\|r_{fl} - r_{il}\|}{\sum_{i=1}^m \|r_{fl} - r_{il}\|}, \quad (l \neq j). \end{aligned} \quad (3)$$

Step 2. Compute the coefficient of difference for the f th instance

$$t_f = 1 - I_f, \quad f = 1, 2, \dots, n. \quad (4)$$

Step 3. Elicit the coefficient of weight for the f th copy

$$w_f = \frac{t_f}{\sum_{f=1}^n t_f}. \quad (5)$$

Step 4. Estimate the j th attributive missing value of x_i^{mis}

$$x_{ij}^{\text{mis}} = \sum_{q=1, q \neq j}^n w_q x_{iq}^{\text{mis}}. \quad (6)$$

If the estimated values of the individual instance vary beyond a tolerable interval compared with the calculated value of the last iteration or the number of iteration times does not reach the maximal value, the operations from (3) to (6) continue iteratively. On the contrast, the iterative process mentioned above terminates as the assessed value is considered as the imputed one. Consequently, the imputed instance is aggregated into the corresponding cluster afterwards with updated centroids.

3.4. The Framework of the CBGMI Algorithm

Procedure. CBGMI

Input. X_{raw} , the $n \times m$ dimensional dataset with missing values G , the number of clusters

Output. X_{full} , the $n \times m$ dimensional complete dataset with imputed values

$X_{\text{raw}} \rightarrow X_{\text{obs}}, X_{\text{mis}}$, where $X_{\text{raw}} = X_{\text{obs}} \cup X_{\text{mis}}$ and $\emptyset = X_{\text{obs}} \cap X_{\text{mis}}$

FCM(X_{obs}, G) $\rightarrow C = \{C_1, C_2, \dots, C_G\}$ according to Section 3.1.

For each element x_k in X_{mis}

Allocate x_k to the closest cluster c_q according to Section 3.2.

Complete the missing values of x_k according to Section 3.3.

Integrate the x_k into corresponding cluster, and update c_q according to Section 3.3.

$$X_{\text{full}} \leftarrow \bigcup_{i=1}^G c_i. \quad (7)$$

FIGURE 2: The RMSE in dataset Wine under (a) MCAR, (b) MAR, and (c) NMAR with different missing rates.

4. Experimental Evaluation

In this section, the assessment criteria are primarily explained in terms of the types of the attributes in Section 4.1. Then, the general effectiveness of our algorithmic approach is presented by a comparative experiment on two UCI datasets [28], remaining superior to MMS, HDI, CDI, C4.5, and EME in Section 4.2.1. Section 4.2.2 shows the technique which also outperforms these aforementioned approaches by applying CBGMI to a real case analysis in two aerospace datasets.

4.1. The Evaluation Criterion

4.1.1. Missing Data on Numeric Attributes. The root mean square error (RMSE) is used to evaluate the predictive ability of the various data imputation algorithms within which the attributes are quantitative

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (e_i - \tilde{e}_i)^2}, \quad (8)$$

FIGURE 3: The CA in dataset Thyroid Disease under (a) MCAR, (b) MAR, and (c) NMAR with different missing rates.

where e_i is the original value, $\tilde{e}_i$ is the predicted plausible value, and m is the total number of estimations. The larger value of RMSE suggests the less accuracy that the algorithm holds [18, 29].

4.1.2. Missing Data on Nominal Attributes. The performances of the algorithms for categorical attributes are appraised by the *classification accuracy* (CA)

$$CA = \frac{1}{n} \sum_{i=1}^n l(EC_i, TC_i), \quad (9)$$

where EC_i and TC_i are the estimated and true class label for the i th missing value, respectively, with n indicating the total

number of the missing values. The function $l(x, y) = 1$ if $x = y$, otherwise $l(x, y) = 0$. For this reason, the larger value of function l indicates the most correct imputed value [18, 29].

4.2. Empirical Result Analysis

4.2.1. UCI Datasets. Two datasets from UCI, that is, *Wine* and *Thyroid Disease*, are selected to test the validity of the algorithms. *Wine* contains 178 instances and 13 attributes. The variable values are either real or integer. *Thyroid Disease* includes 7200 instances and 21 attributes. The multivariate factual data are either categorical or real.

To intrinsically examine the effectiveness and validity and ensure the systematic nature of the research, we artificially

FIGURE 4: The performances of distance metrics on (a) Wine and (b) Thyroid Disease.

generated a lack of data at four distinct missing ratios, that is, 5%, 10%, 15%, and 20% under three different modalities, namely, MCAR, MAR, and NMAR in the complete datasets via the means that Twala did [30].

As we specify $G = 3$, fuzzier $s = 1.3$, and $\zeta = 0.5$, which are mentioned in 3.1 and 3.2, the RMSE produced by six methodologies on the two datasets is illustrated in Figures 2 and 3, respectively.

Figure 2 shows that CBGMI performs better than the other five approaches under all the three missingness mechanisms on Wine, since it is prone to identify the corresponding line from the other five tangled ones. Although the RMSE arises when the missing rate increases from a single subfigure, it indicates that the absence of observed values directly undermines the effect of imputation, as more information could promote the final predictions. Moreover, for each method, the minimum value of RMSE at the same missing ratio always appears when data are NMAR distributed, while MCAR yields the maximum value of RMSE. Therefore, CBGMI is effective and superior to the above algorithms in accuracy for numeric attributes.

The result in Figure 3 demonstrates that CBGMI outperforms the other five approaches under all the three missing modalities on *Thyroid Disease*. The phenomenon that increasing proportion of missing instances deteriorates the CA also states that incomplete nominal values negatively impact on the completion. Similarly, each imputation method functions best under NMAR but operates worst under MCAR. Thus, CBGMI is also applicable to categorical attributes.

To clarify the different distance metrics that influence the accuracy of the results, we suppose that the test happens under MAR. Then, the gray relational analysis metric is practically compared with the *Minkowski distance*, which

refers to *Manhattan distance* (MD), *Euclidean distance* (ED), and *Chebyshev distance* (CD).

From Figure 4, we could assume that *gray relational analysis*- (GRA-) based distance metric generates the least bias at different missing rates compared with the other three variants of *Minkowski distance*. Furthermore, the discrimination is even more significant when gray relational analysis is contrasted with CD according to RMSE or CA.

4.2.2. Aerospace Datasets. Since Section 4.2.1 testifies the effectiveness of the proposed CBGMI on distinct mechanisms, missing rates, and distance metrics, in this subsection, the missingness is artificially simulated under MAR at missing rate 15% based on gray relational analysis.

To the authors' best knowledge, there are a few of hybrid models integrating multiple imputation into clustering. Accordingly, we selected the method proposed by Zhang et al. [5] (denoted as CGKMI) and replaced our Section 3.3 by NIIA [18] (denoted as CNIIA) as the competitors in this part of experiment.

The *remote controlling for spacecraft flying* (RCSF) dataset covers the data generated by some particular unmanned spaceship on real-time condition when flying in the outer space with the remote controlling by the experts. Due to the huge amount of the raw data, we just extract the data which was produced within one minute. Subsequently, the experiment is designed on the 953 records of 20 continuous attributes.

When CBGMI is applied to RCSF dataset, the maximum times of the iteration in all the clusters are 18 loops, which is faster than CNIIA's 19 times and CGKMI's 22 times iterations, respectively, in Figure 5(a). What is more, the RMSE achieves slightly lower than the other counterparts.

FIGURE 5: The RMSE influenced by (a) imputation times and (b) number of clusters on RCSF dataset.

FIGURE 6: The CA influenced by (a) imputation times and (b) number of clusters on SOMD dataset.

As versions of clustering principles, interrelationship between RMSE and the number of clusters in these techniques should be discussed. In Figure 5(b), it appears that when the whole data is agglomerated into 6 groups, the RMSE declines to the minimum. Differently, CGKMI performs best with 8 clusters, while CNIIA requires 4 partitions.

The *spacecraft overall mechanical design* (SOMD) dataset comprises the data related to the assembling and fabrication of one specific model of the manned spaceships. Both the numeric values and categorical values are mixed in the dataset. The total number of instances is beyond 300,000. 1,221 elements with the 30 variables belonging

to a certain step of the entire manufacturing process are chosen.

It is easy to perceive that the three algorithms advance CA as the number of iteration aggrandizes until the convergence emerges in Figure 6(a). Simultaneously, CBGMI attains the best CA in the minimum time of the repetitions compared with the other counterparts on SOMD.

When the amount of clusters rises, the CA fluctuates irregularly in the interval (0.83, 0.88). And CBGMI reaches the maximum of CA when 11 clusters exist in Figure 6(b). Generally, CGKMI and CNIIA undulate in an inferior range of CA to CBGMI, which demands the different optimal number of clustered groups, respectively.

5. Conclusion

By investigating missing data analysis techniques, this study advocates the clustering-based imputation via partitioning original data into two nonoverlapped subsets, that is, the missing-valued subsets and the complete-valued subsets. Then, the iterative imputation strategy is combined within the categorized groups after each entry including missing values has been merged into the closest cluster through gray relational analysis-based distance metric. The experimental results demonstrate that CBGMI exceeds the existing methods, for example, MMS, HDI, CDI, C4.5, and EME, in terms of the RMSE (for continuous missing attributes) and the CA (for discrete missing attributes) at different missing ratios in two canonical UCI datasets, namely, *Wine* and *Thyroid Disease*. In particular, CBGMI algorithm has been applied into the aerospace datasets. The RMSE and CA affected by the iteration times indicate that CBGMI converges more rapidly than the other iterative imputation techniques with better accuracy in the real application environment. In future research, we will focus on how to more effectively estimate and impute missing values under massive data circumstance.

Acknowledgments

This work is supported by Project of the State Key Laboratory of Software Development Environment, Beihang University (SKLSDE-2011ZX-09) and National Natural Science Foundation of China (61003016).

References

- [1] J. C. Bezdek, J. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, The Handbooks of Fuzzy Sets, Springer, 1999.
- [2] A. G. Di Nuovo, "Missing data analysis with fuzzy C-Means: a study of its application in a psychological scenario," *Expert Systems with Applications*, vol. 38, pp. 6793–6797, 2011.
- [3] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.
- [4] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *Journal of School Psychology*, vol. 48, no. 1, pp. 5–37, 2010.
- [5] S. Zhang, Z. Jin, X. Zhu, and J. Zhang, "Missing data analysis: a kernel-based multi-imputation approach," *Transactions on Computational Science*, vol. 5300, pp. 122–142, 2009.
- [6] A. Farhangfar, L. Kurgan, and W. Pedrycz, "Experimental analysis of methods for imputation of missing values in databases," in *Proceedings of the Intelligent Computing: Theory and Applications II*, vol. 5421 of *Proceedings of SPIE*, pp. 172–182, April 2004.
- [7] P. J. García-Laencina, J. L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," *Neurocomputing*, vol. 72, no. 7–9, pp. 1483–1493, 2009.
- [8] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [9] M. Di Zio, U. Guarnera, and O. Luzzi, "Imputation through finite Gaussian mixture models," *Computational Statistics and Data Analysis*, vol. 51, no. 11, pp. 5305–5316, 2007.
- [10] Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "POP algorithm: kernel-based imputation to treat missing values in knowledge discovery from databases," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2794–2804, 2009.
- [11] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*, vol. 38, no. 18, pp. 2895–2907, 2004.
- [12] J. L. Schafer, *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC Press, London, UK, 1997.
- [13] P. D. Allison, *Missing Data*, Sage University Papers, Quantitative Applications in the Social Sciences, Thousand Oaks, Calif, USA, 2001.
- [14] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2002.
- [15] C. Enders, S. Dietz, M. Montague, and J. Dixon, "Modern alternatives for dealing with missing data in special education research," in *Advances in Learning and Behavioral Disorders*, T. E. Scruggs and M. A. Mastropieri, Eds., vol. 19, pp. 101–130, Elsevier, New York, NY, USA, 2006.
- [16] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: a gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.
- [17] J. F. Muñoz and M. Rueda, "New imputation methods for missing data using quantiles," *Journal of Computational and Applied Mathematics*, vol. 232, no. 2, pp. 305–317, 2009.
- [18] S. Zhang, Z. Jin, and X. Zhu, "Missing data imputation by utilizing information within incomplete instances," *Journal of Systems and Software*, vol. 84, no. 3, pp. 452–459, 2011.
- [19] 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.
- [20] S. M. Chen and H. H. Chen, "Estimating null values in the distributed relational databases environment," *Cybernetics and Systems*, vol. 31, no. 8, pp. 851–871, 2000.
- [21] G. J. McLachlan, K. A. Do, and C. Ambrose, *Analyzing Microarray Gene Expression Data*, John Wiley & Sons, 2004.
- [22] C. C. Huang and H. M. Lee, "A grey-based nearest neighbor approach for missing attribute value prediction," *Applied Intelligence*, vol. 20, no. 3, pp. 239–252, 2004.
- [23] E. R. Hruschka Jr., E. R. Hruschka, and N. F. F. Ebecken, "A Bayesian imputation method for a clustering genetic algorithm," *Journal of Computational Methods in Sciences and Engineering*, vol. 11, pp. 173–183, 2011.
- [24] S. M. Chen and C. M. Huang, "Generating weighted fuzzy rules from relational database systems for estimating null values

- using genetic algorithms,” *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp. 495–506, 2003.
- [25] D. Li, J. Deogun, W. Spaulding, and B. Shuart, “Towards missing data imputation: a study of fuzzy K-means clustering method,” in *Proceedings of the 4th International Conference (RSCTC '04)*, pp. 573–579, June 2004.
 - [26] P. Meesad and K. Hengpraprom, “Combination of KNN-based feature selection and KNN-based missing-value imputation of microarray data,” in *Proceedings of the 3rd International Conference on Innovative Computing Information and Control (ICICIC '08)*, pp. 341–344, June 2008.
 - [27] R. J. Hathaway and J. C. Bezdek, “Fuzzy c-means clustering of incomplete data,” *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 31, no. 5, pp. 735–744, 2001.
 - [28] <http://archive.ics.uci.edu/ml/>.
 - [29] C. Zhang, Y. Qin, X. Zhu, J. Zhang, and S. Zhang, “Clustering-based missing value imputation for data preprocessing,” in *Proceedings of the IEEE International Conference on Industrial Informatics (INDIN '06)*, pp. 1081–1086, August 2006.
 - [30] B. Twala, “An empirical comparison of techniques for handling incomplete data using decision trees,” *Applied Artificial Intelligence*, vol. 23, no. 5, pp. 373–405, 2009.