วารสารวิชาการพระจอมเกล้าพระนครเหนือ ปีที่ 31, ฉบับที่ 1 ม.ค.–มี.ค. 2564
The Journal of KMUTNB., Vol. 31, No. 1, Jan.–Mar. 2021

บทความวิจัย

ความถูกต้องในการแทนค่าข้อมูลสูญหายในการจำแนกประเภทกรณีข้อมูลสองกลุ่ม

จำลอง วงษ์ประเสริฐ*
สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุบลราชธานี

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 0 4535 2000 ต่อ 1430 อีเมล: jumlong.v@ubru.ac.th DOI: 10.14416/j.kmutnb.2020.07.002
รับเมื่อ 2 มกราคม 2563 แก้ไขเมื่อ 28 กุมภาพันธ์ 2563 ตอบรับเมื่อ 7 เมษายน 2563 เผยแพร่ออนไลน์ 2 กรกฎาคม 2563

บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบความถูกต้องของการจำแนกประเภทกรณีข้อมูลสองกลุ่ม ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชิน (Support Vector Machines; SVM) โครงข่ายประสาทเทียม (Artificial Neural Networks; ANN) แรนดอมฟอร์เรส (Random Forests; RF) การแทนค่าแบบพหุ (Multiple Imputation; MI) และการแทนค่าแบบแบ็กทรี (Bagged Tree Imputation; BTI) โดยใช้ชุดข้อมูล 3 ชุด ได้แก่ ข้อมูลชุดที่ 1 ประกอบด้วย ตัวแปรอิสระที่เป็นข้อมูลเชิงกลุ่ม 7 ตัวแปร และข้อมูลต่อเนื่องจำนวน 9 ตัวแปร ข้อมูลชุดที่ 2 ประกอบด้วย ตัวแปรอิสระที่เป็นข้อมูลเชิงกลุ่ม 9 ตัวแปร และข้อมูลชุดที่ 3 ประกอบด้วย ตัวแปรอิสระที่เป็นข้อมูลต่อเนื่องจำนวน 9 ตัวแปร การเปรียบเทียบดำเนินการภายใต้เงื่อนไข 1) ข้อมูลจำนวน 3 ชุด 2) ข้อมูลสูญหาย 3 ประเภท ได้แก่ การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random; MCAR) การสูญหายแบบสุ่ม (Missing at Random; MAR) และการสูญหายแบบไม่สุ่ม (Not Missing at Random; NMAR) 3) ร้อยละของข้อมูลสูญหาย ได้แก่ ร้อยละ 5, 10, 15, 20, 25 และ 30 ผลการวิเคราะห์ความถูกต้องของการจำแนกประเภทพบว่า ในภาพรวมภายใต้ทุกเงื่อนไขของการทดลอง แนะนำให้ใช้วิธีแรนดอมฟอร์เรส และซัพพอร์ตเวกเตอร์แมชชิน ภายใต้เงื่อนไขการสูญหายแบบสุ่มอย่างสมบูรณ์ และการสูญหายแบบสุ่ม แนะนำให้ใช้วิธีซัพพอร์ตเวกเตอร์แมชชิน ภายใต้เงื่อนไขการสูญหายแบบไม่สุ่ม แนะนำให้ใช้วิธีแรนดอมฟอร์เรส

คำสำคัญ: ข้อมูลสูญหาย การแทนค่า การจำแนกประเภทกรณีข้อมูลสองกลุ่ม

การอ้างอิงบทความ: จำลอง วงษ์ประเสริฐ, "ความถูกต้องในการแทนค่าข้อมูลสูญหายในการจำแนกประเภทกรณีข้อมูลสองกลุ่ม," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 31, ฉบับที่ 1, หน้า 90–98, ม.ค.–มี.ค. 2564.

Research Article

Missing Data Imputation Based on Accuracy of Binary Classification

Jumlong Vongprasert*
Applied Statistics Department, Faculty of Science, Ubon Ratchathani Rajabhat University, Ubon Ratchathani, Thailand

* Corresponding Author, Tel. 0 4535 2000 Ext. 1430, E–mail: jumlong.v@ubru.ac.th
DOI: 10.14416/j.kmutnb.2020.07.002
Received 2 January 2020; Revised 28 February 2020; Accepted 7 April 2020; Published online: 2 July 2020

Abstract

The purpose of this study was to compare accuracy of binary classification based on missing data imputations methods namely: Support Vector Machines (SVM); Neural Networks (NN); Random Forests (RF); Multiple Imputation (MI) and Bagged Tree Imputation (BTI). Three data sets comprise: 1) 7 categorical and 9 continuous independent variables, 2) 9 categorical independent variables and 3) 9 continuous independent variables. The comparisons were made with the following conditions: 1) Three data sets; 2) three types of missing data: Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR); 3) six levels of percentage of missing data (5, 10, 15, 20, 25 and 30). We analyze which imputation method influences most the classifiers' accuracy. The best imputations in overall were obtained using RF and SVM, the imputation under MAR and MCAR were obtained using SVM, the imputation under NMAR were obtained using RF.

Keywords: Missing Data, Imputation, Binary Classification

Please cite this article as: J. Vongprasert, "Missing data imputation based on accuracy of binary classification," *The Journal of KMUTNB*, vol. 31, no. 1, pp. 90–98, Jan.–Mar. 2021 (in Thai).

1. Introduction

Missing values are unavoidable in real world datasets, there is a variety of causes why data may be missing. Anyone who does statistical data analysis of any kind runs into the problems of missing data. In a characteristic dataset we always land up in some missing values for attributes. The most serious concern is that missing data can introduce bias into estimates derived from a statistical model [1]–[3]. If the responses are not ignorable, however, estimation of the propensity scores is complicated and often requires additional surrogate [4] or instrumental variables [5] to estimate the model parameters consistently. Missing data analysis is importance since an inference based ignoring the missingness may not only misleading conclusions, but also lose efficiency and lead to biased results. [6]

1.1 Missing Data

Let $\mathbf{y} = (y_1, y_2, ... y_n)'$ denote the complete set of the outcome variables, and $\boldsymbol{\delta} = (\delta_1, \delta_2, ... \delta_n)'$ be the vector of missing data indicators such that δ_1 = 1 when y_i is observed and δ_i = 0 when y_i is missing. We note that each y_i and the corresponding δ_i can also be vectors. Let $\mathbf{y}_{\text{obs}}$ denote the observed and $\mathbf{y}_{\text{mis}}$ missing components of $\mathbf{y}$. With the above notation, the missing data mechanisms are characterized by the conditional distribution of $\boldsymbol{\delta}$ given $\mathbf{y}$, say $f(\boldsymbol{\delta}|\mathbf{y},\phi)$, where ϕ denotes some unknown parameters.
Three major types of missing data are [6]:

Missing Completely at Random (MCAR) denotes the mechanism that missingness does not depend on the values of the data $\mathbf{y}$, missing or observed.

$$f(\boldsymbol{\delta}|\mathbf{y},\phi) = f(\boldsymbol{\delta}|\phi) \forall \mathbf{y}, \phi \quad (1)$$

Missing at Random (MAR) denotes the mechanism that missingness only depends on the components of y_{obs} that are observed, and not on the components that are missing.

$$f(\boldsymbol{\delta}|\mathbf{y},\phi) = f(\boldsymbol{\delta}|\mathbf{y}_{\text{obs}},\phi) \forall \mathbf{y}_{\text{mis}}, \phi \quad (2)$$

Not missing at random (NMAR) denotes the one that the distribution of $\mathbf{y}$ does depend on the missing values in the data.

$$f(y_i|x_i, \delta_i = 0, \phi) \neq f(y_i|x_i, \delta_i = 1, \phi) \quad (3)$$

2. Materials and Methods

2.1 Data Set

In this section, we introduce and describe the data set. We used three data set from University of California Irvine Machine Learning Repository [7], 1) Bank Marketing data set with 9 categorical (type of job, marital status, education, has credit in default?, has housing loan?, has personal loan?, contact communication type, last contact month of year, outcome of the previous marketing campaign) and 7 continuous independent variables (age, average yearly balance, last contact day of the month, last contact duration, number of contacts performed during this campaign and for this client, number of days that passed by after the client was last contacted from a previous campaign, number of contacts performed before this campaign and for this client) and 1,000 instances by simple random sampling from 45,211 instances, 2) Wisconsin Breast Cancer Database with 9 categorical independent variables (Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single

Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses) and 700 instances and 3) Breast Cancer Coimbra Data Set with 9 continuous independent variables (Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1) and 116 instances.

2.2 Research Methodology

2.2.1 Methods

In this section, we introduced and described the methods applied to impute the original incomplete data set. The five imputation techniques applied are: Support Vector Machines (SVM); Neural Networks (NN); Random Forests (RF); Multiple Imputation (MI) and Bagged Tree Imputation (BTI).
Support Vector Machines (SVM)

SVM are learning machines based on the statistical learning theory, which can use linear and nonlinear kernels for the classification. They minimize the structure risk in a higher dimensional feature space, searching for the hyperplane with the largest margin between the classes. [8] SVM are useful approach for solving data classification and recognition problems. In this work we used the SVM implementation from the Gaussian kernel from R package 'kernlab'.

2.2.2 Neural Networks (NN)

A number of approaches have been investigated and applied to solve the missing data of this research includes NN, because of their flexibility, fault tolerance and capability to handle incomplete data. NN models have previously been applied to solve different tasks of missing data comprised of neural networks as a key classifier [9]. In this work we used the NN implementation from the R package 'nnet'.

2.2.3 Random Forests (RF)

RF [10] is a machine learning technique that builds a multitude of weak decisional trees at training time and outputs the class that is the mode of the classes (classification) or average prediction (regression) of the individual trees. Each tree is individually trained on a sample of the training data, and at each node, the algorithm only searches across a random subset of the features to determine a split. The input vector to be classified is submitted to each of the decision trees in the forest and the prediction is then formed using a majority vote. [11]. In this work we used the RF with ntree 500 implementation from the R package 'randomForest.

2.2.4 Multiple Imputation (MI)

MI is a statistical technique for analyzing incomplete data sets, that is, data sets for which some entries are missing. Application of the technique requires three steps: imputation, analysis and pooling. The process of MI follows these steps,

Imputation: Impute the missing entries of the incomplete data sets, not once, but m times. Imputed values are drawn for a distribution. This step results are m complete data sets.

Analysis: Analyze each of the m completed data sets. This step results in m analyses.

Pooling: Integrate the m analysis results into a final result. Simple rules exist for combining the m analyses.

For imputing the missing data, we use MI algorithm [12], implemented in the amelia function from the Amelia package

2.2.5 Bagged Tree Imputation (BTI)

Bagging predictors approach generates manifold versions of a predictor to get an aggregated one.

The aggregation function usually is the average value over all predictor estimations for a numerical outcome, or employs a majority vote when the desired output is a categorical one. The multiple predictions are estimated by bootstrapping from the training set and subsequently using these as new learning sets. Tests on real and artificial data sets, using classification and regression trees and subset bootstrap with linear regression, show that bagging can be beneficial for the accuracy. Vital component of this technique is the instability of the prediction model, but if perturbing the learning set can cause significant changes in the constructed predictor, then bagging can improve the accuracy. [11], [13], [14]. In this work we used the BTI implementation from the R package 'caret'.

2.2.6 Model Evaluation

The accuracy of missing data imputation methods is evaluated by accuracy of classification.

$$\text{Accuracy of classification} \quad \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

2.2.7 Structural Flow of the Work

Referring to data set,

First, we divide each data set using MCAR (1) MAR (2) and NMAR (3) method with the percentages of the missing at 5, 10, 15, 20, 25 and 30, complete data set (training set) and incomplete data set (testing set) split.

For Bank Marketing data set with 1,000 instances, number of instances in training set and test set for missing data 5, 10, 15, 20, 25 and 30 are, training set 950, 900, 850, 800, 750 and 700 instances respectively, test set 50, 100, 150, 200, 250 and 300 instances respectively.

For Wisconsin Breast Cancer Database data set with 700 instances, number of instances in training set and test set for missing data 5, 10, 15, 20, 25 and 30 are, training set 665, 630, 595, 560, 525 and 490 instances respectively, test set 35, 70, 105, 140, 175 and 210 instances respectively.

For Breast Cancer Coimbra data set with 116 instances, number of instances in training set and test set for missing data 5, 10, 15, 20, 25 and 30 are, training set 110, 104, 99, 93, 87 and 81 instances respectively, test set 6, 12, 17, 23, 29 and 35 instances respectively.

Next, for complete data set (training set) with each missing data type and missing percentage, we fit data with SVM, NN, RF, MI and BTI algorithms to obtain a classification model.

Finally, we impute the missing values in incomplete data set (testing set) by running SVM, NN, RF, MI and BTI algorithms with each missing data type and missing percentage.

For executing the tests, we wrote the codes in R-programming and retrieved some equations relating to those techniques from CRAN projects, and used 1,000 replicated for each condition. Next, we tested the results from simulations with the estimators by accuracy of classification. The simulations and results are described in the next section.

3. Results

Missing data imputation methods: SVM, NN, RF, MI and BTI were applied to impute missing data. The goal was to analyze the improvements in accuracy of classification when different algorithms were applied to impute missing data values. Table 1–3 indicates the average of accuracy classified by

percentage of missing data and missing type for Bank Marketing data set, Wisconsin Breast Cancer database and Breast Cancer Coimbra data set respectively. Table 4–5 indicates the accuracy of classified percentage of missing data and type of missing respectively. Figure 1–5 shows the accuracy of SVM, NN, RF, MI and BTI respectively.

Table 1 Average of accuracy classified by percentage of missing data and missing type for Bank Marketing data set

Type	Missing	SVM	NN	RF	MI	BT
MAR	5	0.7646	0.7759	0.7944	0.7444	0.7415
	10	0.8576	0.8335	0.8565	0.7721	0.7690
	15	0.8291	0.8286	0.8398	0.7660	0.7789
	20	0.8544	0.8319	0.8789	0.7730	0.7604
	25	0.8388	0.8331	0.8544	0.7471	0.7103
	30	0.8424	0.8113	0.8594	0.7524	0.7757
	Average	0.8412	0.8262	0.8581	0.7596	0.7563
MCAR	5	0.8897	0.8764	0.8907	0.8207	0.7999
	10	0.8869	0.8821	0.8871	0.8060	0.7814
	15	0.8904	0.8835	0.8933	0.8108	0.7909
	20	0.8894	0.8846	0.8990	0.8013	0.8147
	25	0.8906	0.8748	0.8915	0.7981	0.7896
	30	0.8929	0.8767	0.8988	0.7935	0.7955
	Average	0.8900	0.8803	0.8939	0.8019	0.7944
NMAR	5	0.4600	0.4657	0.4800	0.5762	0.5105
	10	0.5400	0.5522	0.5332	0.5848	0.5597
	15	0.6060	0.5848	0.6080	0.6354	0.6214
	20	0.6368	0.6329	0.6561	0.6203	0.5923
	25	0.7120	0.6847	0.7295	0.6474	0.5804
	30	0.7384	0.7224	0.7619	0.6852	0.6631
	Average	0.6155	0.6071	0.6281	0.6249	0.5879
Average		0.7672	0.7566	0.7780	0.7198	0.7017

Table 2 Average of accuracy classified by percentage of missing data and missing type for Wisconsin Breast Cancer database

Type	Missing	SVM	NN	RF	MI	BT
MAR	5	1.0000	0.8517	0.9331	0.6168	0.9805
	10	1.0000	0.8316	0.9245	0.6044	0.9818
	15	1.0000	0.8409	0.9231	0.6017	0.9819
	20	1.0000	0.8392	0.9220	0.6003	0.9827
	25	1.0000	0.8461	0.9223	0.6047	0.9823
	30	1.0000	0.8553	0.9241	0.5987	0.9789
	Average	1.0000	0.8441	0.9249	0.6044	0.9813
MCAR	5	0.9614	0.9435	0.9695	0.8004	0.9548
	10	0.9628	0.9423	0.9690	0.7845	0.9537
	15	0.9619	0.9420	0.9684	0.7782	0.9520
	20	0.9625	0.9407	0.9686	0.7728	0.9527
	25	0.9627	0.9404	0.9686	0.7717	0.9508
	30	0.9624	0.9402	0.9683	0.7703	0.9503
	Average	0.9623	0.9415	0.9687	0.7797	0.9524
NMAR	5	0.9444	0.8291	0.9999	0.8765	0.9715
	10	0.9252	0.7883	0.9467	0.8521	0.9356
	15	0.9252	0.7813	0.9470	0.8519	0.9348
	20	0.9165	0.5766	0.9317	0.8267	0.9356
	25	0.8975	0.6403	0.9246	0.8150	0.9261
	30	0.1846	0.4595	0.9097	0.7866	0.9197
	Average	0.7989	0.6792	0.9433	0.8348	0.9372
Average		0.9204	0.8216	0.9456	0.7396	0.9570

Table 3 Average of accuracy classified by percentage of missing data and missing type for Breast Cancer Coimbra data set

Type	Missing	SVM	NN	RF	MI	BT
MAR	5	0.9065	0.6204	0.8948	0.9689	0.7985
	10	0.8946	0.6142	0.8907	0.9535	0.8020
	15	0.8774	0.5992	0.8892	0.9475	0.8000
	20	0.8654	0.5985	0.8898	0.9452	0.7953
	25	0.8597	0.5964	0.8825	0.9390	0.7926
	30	0.8492	0.5939	0.8793	0.9344	0.7906
	Average	0.8629	0.5970	0.8852	0.9415	0.7946
MCAR	5	0.7365	0.5722	0.7205	0.9667	0.7120
	10	0.7402	0.5785	0.7243	0.9402	0.7228
	15	0.7417	0.5687	0.7329	0.9375	0.7249
	20	0.7380	0.5750	0.7220	0.9226	0.7151
	25	0.7297	0.5679	0.7181	0.9215	0.7104
	30	0.7310	0.5674	0.7159	0.9165	0.7045
	Average	0.7361	0.5715	0.7226	0.9277	0.7155
NMAR	5	0.8333	0.7772	0.9947	0.9807	0.7405
	10	0.9167	0.6908	0.9238	0.9480	0.6310
	15	0.8889	0.5749	0.9282	0.9163	0.6639
	20	0.8737	0.5368	0.7083	0.9181	0.6167
	25	0.8614	0.4642	0.6272	0.9127	0.5254
	30	0.7431	0.4993	0.7310	0.9160	0.5968
	Average	0.8528	0.5905	0.8189	0.9320	0.6290
Average		0.8166	0.5859	0.8045	0.9331	0.7020

Table 4 Average of accuracy classified by percentage of missing data

Missing	SVM	NN	RF	MI	BT
5	*0.8329*	0.7458	0.8530	0.8168	0.8011
10	*0.8582*	0.7459	0.8506	0.8051	0.7930
15	*0.8578*	0.7338	0.8589	0.8050	0.8054
20	*0.8596*	0.7129	0.8418	0.7978	0.7962
25	*0.8614*	0.7164	0.8354	0.7953	0.7742
30	0.7716	0.7029	0.8498	0.7948	0.7972
Average	0.8400	0.7264	0.8493	0.8007	0.7943

Table 5 Average of accuracy classified by type of missing

Type	SVM	NN	RF	MI	BT
MAR	0.9014	0.7558	0.8894	0.7685	0.8441
MCAR	0.8628	0.7978	0.8617	0.8364	0.8208
NMAR	0.7557	0.6256	0.7968	0.7972	0.7180
Average	0.8400	0.7264	0.8493	0.8007	0.7943

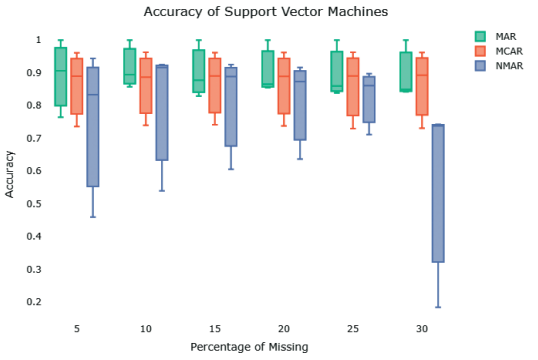

Figure 1 Accuracy of Support Vector Machines.

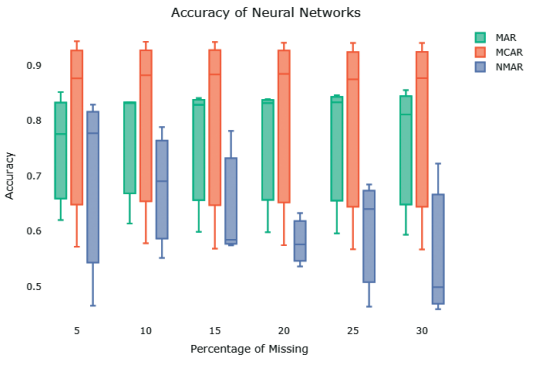

Figure 2 Accuracy of Neural Networks.

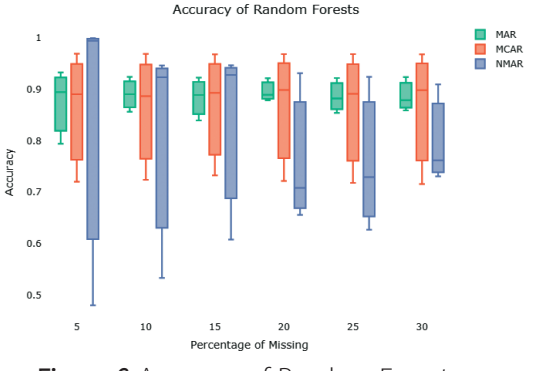

Figure 3 Accuracy of Random Forests.

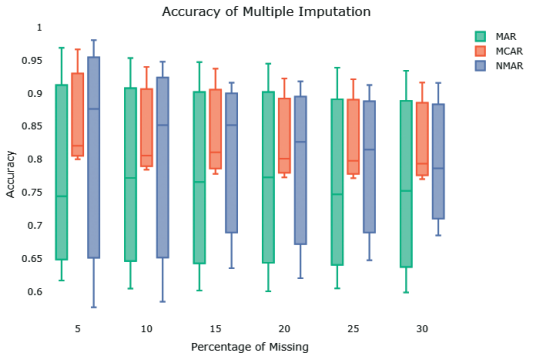

Figure 4 Accuracy of Multiple Imputation.

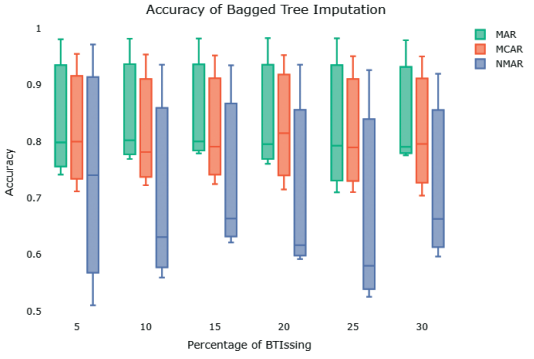

Figure 5 Accuracy of Bagged Tree Imputation.

4. Discussion and Conclusions

We applied five imputation methods to treat the problem of missing data. We reviewed and provided technical details of the different methods used included SVM, NN, RF, MI and BTI. As depicted in Table 1-5, all imputation methods led to an improvement in accuracy prediction, as measured by accuracy of classification. The best imputations in overall were obtained using RF and SVM, the imputation under MAR and MCAR were obtained using SVM, the imputation under NMAR were obtained using RF.

5. Acknowledgement

This work is fully achieved by the collaborations of the Doctor of Philosophy (Educational Research and Evaluation) Department, Faculty of Education, Ubon Ratchathani Rajabhat University, Thailand.

References

[1] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons Inc, 1987.

[2] W. E Becker and W. B. Walstad. "Data loss from pretest to posttest as a sample selection problem," *The Review of Economics and Statistics*, vol. 72, no. 1, pp. 184–188, 1990.

[3] W. Becker and J. Powers, "Student performance, attrition, and class size given missing student data," *Economics of Education Review*, vol. 20, no. 4, pp. 377–388, 2001.

[4] S. X. Chen, D. H. Leung, and J. Qin. "Improving semiparametric estimation by using surrogate data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 4, pp. 803–823, 2008.

[5] P. S. Kott and T. Chang, "Using calibration weighting to adjust for nonignorable unit nonresponse," *Journal of the American Statistical Association*, vol. 105, no. 491, pp. 1265–1275, 2010.

[6] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., New York: John Wiley & Sons Inc, 2020, pp. 408.

[7] D. Dua and C. Graff, "UCI machine learning repository," Irvine, CA: University of California, School of Information and Computer Science, 2019.

[8] Z. H. O. U. Xin, W. U. Ying, and Y. A. N. G. Bin, "Signal classification method based on support vector machine and high-order cumulants,"

Wireless Sensor Network, vol. 2, no. 1, pp. 48–52, 2010.

[9] N. K. Ibrahim, R. S. A. Raja Abdullah, and M. I. Saripan, "Artificial neural network approach in radar target classification," *Journal of Computer Science*, vol. 5, no. 1, pp. 23–32, 2009.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] I. Jordanov, N. Petrov, and A. Petrozziello. "Classifiers accuracy improvement based on missing data imputation," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 8, no. 1, pp. 31–48, 2018.

[12] S. Verboven, K. V. Branden, and P. Goos, "Sequential imputation for missing values," *Computational Biology and Chemistry*, vol. 31, no. 5–6, pp. 320–327, 2007.

[13] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of Machine Learning Research*, vol. 8, pp. 1623–1657, 2007.

[14] G. Rahman and Z. Islam, "A decision tree-based missing value imputation technique for data pre-processing," in *Proceedings of the Ninth Australasian Data Mining Conference*, 2011, pp. 41–50.