Improving Penalized Logistic Regression Model with Missing Values in High-Dimensional Data

https://doi.org/10.3991/ijoe.v18i02.25047

Aiedh Mrisi Alharthi[1,2], Muhammad Hisyam Lee[1(✉)], Zakariya Yahya Algamal[3]
[1] Department of Mathematical Sciences, Universiti Teknologi Malaysia, Skudai, Malaysia
[2] Department of Mathematics, Taif University, Taif, Saudi Arabia
[3] Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
mhl@utm.my

Abstract—Analysis without adequate handling of missing values may lead to inconsistent and biased estimates. Despite multiple imputations becoming a widely used approach in handling missing data, manuscript researchers generally encounter missing data in their respective studies. In high-dimensional data, penalized regression is a popular technique for performing feature selection and coefficient estimation simultaneously. However, one of the most vital issues with high-dimensional data is that it often contains large quantities of missing data that common multiple imputation approaches may not work correctly. Therefore, this study uses imputations penalized regression models as an extension of the penalized methods to improve the performance and impute missing values in high-dimensional data. The method was applied to real-life high-dimensional datasets for the different number of features, sample sizes, and missing dataset rates to evaluate its efficiency. The method was also compared with other existing imputation penalized methods for high-dimensional data. The comparative experimental results indicate that the proposed method outperforms its competitors by achieving higher sensitivity, specificity, and classification accuracy values.

Keywords—high-dimensional data, feature selection, missing data, multiple imputations, penalized regression

1 Introduction

Missing data exist in almost all areas of biomedical, epidemiological, and social research. This may be due to various reasons including unavailability of measurements, survey nonresponse, and data loss [1]. Many statistical techniques often require complete cases without any missing data. This as inaccurate estimates and conclusions may result from an analysis that does not properly handle missing values [2]. Though, the problem of missing data may be addressed using a number of statistical approaches. Jiang et al. [1] argued that ignoring the observation with missing values is a straightforward solution. This is due to the fact that when there are a few observations with missing values, there is usually no significant problem. However, delete a high

number of observations with missing values, on the other hand, results in a considerable loss of data [3], [4]. It also has a negative impact on the data's statistical power and efficiency [5]. By filling in the missing values with some reasonable values, imputation produces the complete data without eliminating the missing cases for analysis. Some ad-hoc methods, including mean substitution, maximum likelihood approaches, single imputation, and multiple imputation (MI), can be used to impute missing data [6]. Therefore, to overcome the missing values in high-dimensional data, reliable imputation approaches are required.

High-dimensional data is another issue that is often faced in a wide range of scientific study domains, including genetics, health sciences, economics, chemometrics, sociological surveys, environmental sciences, finance, and machine learning, amongst others [7]. In high-dimensional data analysis, variable selection is crucial. Recently, the use of gene selection techniques in biological datasets has risen significantly, where the number of genes is usually more than the number of samples [8], which can lead to overfitting and have a detrimental influence on learning. Furthermore, only a few genes have relevant meanings and are directly related to the associated disease from both a biological and knowledge discovery standpoint [9]. As a result, identifying informative genes is an efficient technique to handle these challenges, which may be considered a machine learning feature selection problem [10], [11]. In the previous decade, there has been major progress in variable selection methods. Among these methods, penalized methods were identified. The penalized method is used to select features and classify them. Penalized logistic techniques are those that include a type of penalty term into the logistic regression in order to perform both selection and classification simultaneously. The logistic regression method has attracted a great deal of attention. A variety of logistic regression models with varying penalties may be utilized. "Least Absolute Shrinkage and Selection Operator" is the name of one of these penalties ("also known as Lasso"), which is based on the L_1-norm [12]. Another penalty that is based on the L_2-norm is ridge regression [13]. Other penalties are the so-called "Smoothly Clipped Absolute Deviation" (SCAD) [14], the elastic net [15], the adaptive Lasso method [16], and the adaptive elastic net methods [17], [18].

Consequently, in high-dimensional data, penalized regression is a popular technique for performing variable selection and coefficient estimation simultaneously. However, one of the most vital issues with high-dimensional data is that it often contains large quantities of missing data. According to previous researches, most microarray datasets are incomplete to varying degrees, ranging from fifty percent to ninety-five percent [19]. Multiple imputation (MI) [20], [21] has become a widely used approach in handling missing data, with significant improvement in the methods and software [22], [23]. However, MI approaches may not work correctly in high-dimensional data, where the number of variables (p) in the imputation model exceeds the sample size (n), i. e., $(p > n \text{ or } p \gg n)$ [2]. As it is now, the problem gets more critical, and conventional likelihood estimates become unavailable. It has also become challenging to apply sequential regression imputation in this situation. [6], [24].

In the existence of high-dimensional data, it is possible that the current MI methods and software packages may perform inadequately. To address this issue, this study uses imputations penalized regression models as an extension of the penalized meth-

ods to improve the performance and impute missing values in high-dimensional data. This is done by employing the "one-dimensional weighted Mahalanobis distance" (1-DWM) as an initial weight inside L_1-norm with imputing missing values for each predictor variable (feature). The proposed method referred by imputations adaptive penalized logistic regression (IAPLR) is compared with other existing imputation methods for high-dimensional data. The remainder of this article is arranged in the following. Detailed descriptions of the materials and methods are included in Section 2. Section 3 presents and debates the findings of the experimental investigation designed to assess the effectiveness of IAPLR compared to other penalized approaches. This paper is then concluded in Section 4.

2 Materials and methods

2.1 Missing data imputation

The missing data is one of the most prevalent problems in several fields of research. Traditional statistical methodologies demand entire cases without missing data in order to analyze the data. The removal of missing data is a loss of important data, and which could lead to an inaccurate statistic inference. Though, by imputing plausible values to the missing values, the imputation provides the whole data without removing the missing analytical data. Little and Rubin [20], [21] divided missing mechanisms into three main categories: First, missing completely at random (MCAR). This involves missing data independently of both observed and unobserved data. The second is Missing at Random (MAR). It is in the probability of a missing value, which is determined by the observed values but not by the data values that are missing., i.e., $P(missing/complete\ data) = P(missing/observed\ data)$. The third is missing not at random (MNAR), in which the probability of a missing data value is determined by the missing data, i.e., $P(missing/complete\ data) \neq P(missing/observed\ data)$.

For a missing real value in a dataset, single imputation approaches create a specified value. This method has a lower computational cost. The researchers have proposed a variety of single imputation strategies. The primary strategy is to analyze other replies and choose the most significant possible response. The value may be calculated using the mean, median, and mode of the variable's available values [3]. Imputed values are treated as actual values in single imputation. The uncertainty of the imputed values is ignored in single imputation-based approaches. Standard errors may exist for these values. As a result, the results are biased [25]. Also, for single imputation, other methodologies, such as machine learning-based methods, may be utilized [26].

However, using several simulation models, MI methods yielded various values for the imputation of a single missing data. The variability of imputed data is introduced in these approaches to find a range of reasonable responses. MI approaches are more complicated than single imputation, but they do not suffer from bias values. MI can be summarized into three steps. The first step is imputation, in which M independent

imputed values matching to missing data are obtained. The analysis is the second step, which involves analyzing each of the M imputed datasets using standard statistical techniques for complete data. The third step combines results of the analysis, in which M sets of desired estimates are combined into one set of parameter estimates using Rubin's rules [27]. Several previous studies have proposed packages in R to implement MI methods more efficiently. One of these packages is called "Multivariate Imputation via Chained Equations" (also known as packages "mic") [22]. Other packages are "mi" [23] and "Amelia" [28].

2.2 The penalized logistic regression model

The logistic regression is a statistical approach for predicting the value of a categorical response variable with just two potential values represented by 0 and 1. When dealing with low-dimensional data, logistic regression works well. Nevertheless, when dealing with high-dimensional data sets, such as those including gene expression data, it may become inefficient in terms of prediction accuracy and computational efficiency. Another issue that affects the use of logistic regression is overfitting, which occurs when the number of features exceeds the number of observed values [29]. The logistic regression with a penalty is used in various classification fields to perform gene selection and classification simultaneously [30]. This model is penalized, and its coefficients are shrunk as part of the regularization procedure [31]. Over the last decade, penalized regression approaches have gained popularity due to their superior prediction accuracy and computational efficiency.

For illustration purposes, suppose a set of data is designed as a matrix $X \in R^{n\times p}$ $(n \ll p)$, $X = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where each column indicates a feature, each row denotes a sample, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the i^{th} input sample, the entry $x_{i,j}$ denotes the value of the j^{th} feature of the i^{th}sample and $y = (y_1, \ldots, y_n)^T$ is the $n-$dimensional vector of binary responses coded as $\{0, 1\}$. The class posterior probability is defined in the logistic regression function as follows:

$$p(y_i = 1|x_{ij}) = \pi(x_i) = \frac{exp(x_j^T \beta_j)}{1+exp(x_j^T \beta_j)}, j = 1,2, \ldots, p \quad (1)$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p-$dimensional vector of the unknown parameters. Then, the estimator $\hat{\beta}$ is obtained as the minimizer of the log-likelihood function as follows:

$$\ell(\beta, y_i) = -\sum_{i=1}^{n} \{y_i \, log \, \pi \, (x_i) + (1 - y_i) \, log(\, 1 - \pi(x_i)\} \quad (2)$$

The classification method of logistic regression is a powerful discriminative tool (variable selection). However, logistic regression is not useful as a classification technique when the dataset is high dimensional since the design matrix is singular. As a result, it is unable to produce accurate regression coefficient estimations. Furthermore, overfitting occurs when datasets are high dimensional, such as when there are a large number of genes (or features in general). Furthermore, multicollinearity might affect its estimators [32], [33].

From a statistical point of view, other (unrelated) features may cause noise and reduce classification effectiveness. To increase classification accuracy, statisticians commonly use feature selection methods that can eliminate irrelevant and redundant features. Besides the logistic regression, there are other classification methods available, such as penalized logistic regression (PLR), which is used to reduce high dimensionality and enhance classification accuracy [34]. Although regularization methods are often applied to high-dimensional data, [35] claimed that they might also be applied to low-dimensional data.

The log-likelihood function is modified by the addition of a positive penalty term in penalized logistic regression, imposing certain coefficients to become zero to produce a sparse solution. The PLR penalizes a logistic model with too many features by adding a penalty term to the equation. As a result, when the coefficients are constrained, the coefficients of less essential features become either extremely near to zero or precisely zero. Regularization is another name for this technique. The following is the setting for the technique.

The penalized log-likelihood is represented as:

$$\text{PLR} = -\sum_{i=1}^{n}\{y_i \, log\, \pi\,(x_i) + (1-y_i)\, log(\,1-\pi(x_i)\} + \lambda g(\beta), \tag{3}$$

where, $g(\beta)$ indicates a regularization term that can be expressed in different forms and $\lambda > 0$ denotes a control parameter. Then the PLR of Eq. (3) is minimized with regard to λ to obtain estimates of the coefficients. The use of a penalty ensures that each parameter has a unique estimate and results in better predictions than the conventional "Maximum Likelihood Estimation" (MLE), with a reasonable balance between bias and variance [36]. Without loss of generality, y and the columns of X are considered to be standardized, $\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 0$, and $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1,2,\ldots,p$. Consequently, the intercept term(β_0) is not penalized. β is estimated employing Lasso technique by:

$$\hat{\beta}_{LASSO} = argmin_{\beta}\Big[-\sum_{i=1}^{n}\{y_i \, log\, \pi\,(x_i) + (1-y_i)\, log(\,1-\pi(x_i))\} + \lambda \sum_{j=1}^{p} |\beta_j|\Big], \tag{4}$$

where, λ is a control parameter. When $\lambda = 0$, Eq. (4) reduces the likelihood estimator to its lowest possible value. As $\lambda \to \infty$ penalization forces all features to be zero.

The adaptive Lasso (ALasso) method is an extension of Lasso. It was originally proposed by [16] to overcome the shortcomings of Lasso by combining the L_1 penalty with the weighted penalty [37]. Zou [16] modified the L_1-penalty by providing various weights to various coefficients in order to make it more efficient. Shrinkage techniques such as Ridge, Lasso, and other similar methods might be used to assign weights. The ALasso associated with the logistic regression is given by:

$$\hat{\beta}_{ALASSO} = argmin_{\beta}\Big[-\sum_{i=1}^{n}\{y_i \, log\, \pi\,(x_i) + (1-y_i)\, log(\,1-\pi(x_i))\} + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\left(|\hat{\beta}_j^{initial}|\right)^{\gamma}}\Big], \tag{5}$$

where, $\lambda, \gamma \geq 0$ and $\hat{\beta}_j^{initial}$ is an initial estimate for each β_j estimated using the Lasso technique or other shrinkage techniques. Here we set $\gamma = 1$, for simplicity.

2.3 The proposed method

Missing data is a problem that affects performance in data analytics. An inaccurate prediction might result from incorrect imputation of missing values. Recently, when a vast amount of data is created every second, data usage become a key concern for stakeholders. Thus, managing missing data efficiently becomes increasingly crucial. This research is also motivated by the fact that in PLR, the L_1-norm penalty may be used to apply the PLR approach to high-dimensional data sets. However, because the L_1-norm is inconsistent with feature selection, this technique may result in the selection of irrelevant and redundant features [38]. To put it another way, PLR estimates based on the L_1-norm may be biased for large coefficients since they receive more enormous penalties.

Peng et al. [39] employed the "one-dimensional weighted Mahalanobis distance" (1-DWM) as a criterion of gene efficiency to extend the effect of individual genes to the joint impact of multigene, that is defined as:

$$J(x_{.j}) = \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{\sigma_{wj}^2}, \tag{6}$$

where x_j is a column vector, denotes feature j across samples, and $\sigma_{wj}^2 = w_{1j}.\sigma_{1j}^2 + w_{2j}.\sigma_{2j}^2$, denotes the weighted variance of feature j, σ_{kj}^2 denotes variance of feature j in class k, w_k is the prior probability or weight of class k, where $k = 2$ in this study and $w_1 = w_2 = 0.5$.

Therefore, this study uses imputations adaptive penalized logistic regression (IAPLR) as an extension of the penalized methods to improve the performance and impute missing values in high-dimensional data. This is done by employing the (1-DWM) as an initial weight inside L_1-norm with imputing missing values for each feature (gene). The proposed method addresses missing values and improves feature selection in high-dimensional.

The j^{th} component of the p-dimensional vector of features is denoted as:

$$w_j = \frac{1}{|J(x_{.j})|},\ j = 1,2,\ldots,p, \tag{7}$$

where $J(x_{.j})$ is the weight for every feature j that is indicated as Eq. (6).

The proposed method imputes missing values by the "naniar" package in R. Moreover, to alleviate inconsistency in feature selection, the proposed weight in this work gives a relatively large amount of weight to the feature with a low ratio value while giving a small weight to the feature with a high ratio value. The IAPLR becomes capable of reliably picking related features after correctly assigning weights to features. Figure 1 shows the implementation algorithm for the IAPLR technique. The fact that the IAPLR equation is convex ensures the presence of a global maximum

point for the solution. The coordinate descent method can be used to find the IAPLR solution by:

$$\hat{\beta}_{IAPLR} = argmin_{\beta}[-\sum_{i=1}^{n}\{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\}+\lambda\sum_{j=1}^{p} w_j|\beta_j|]. \quad (8)$$

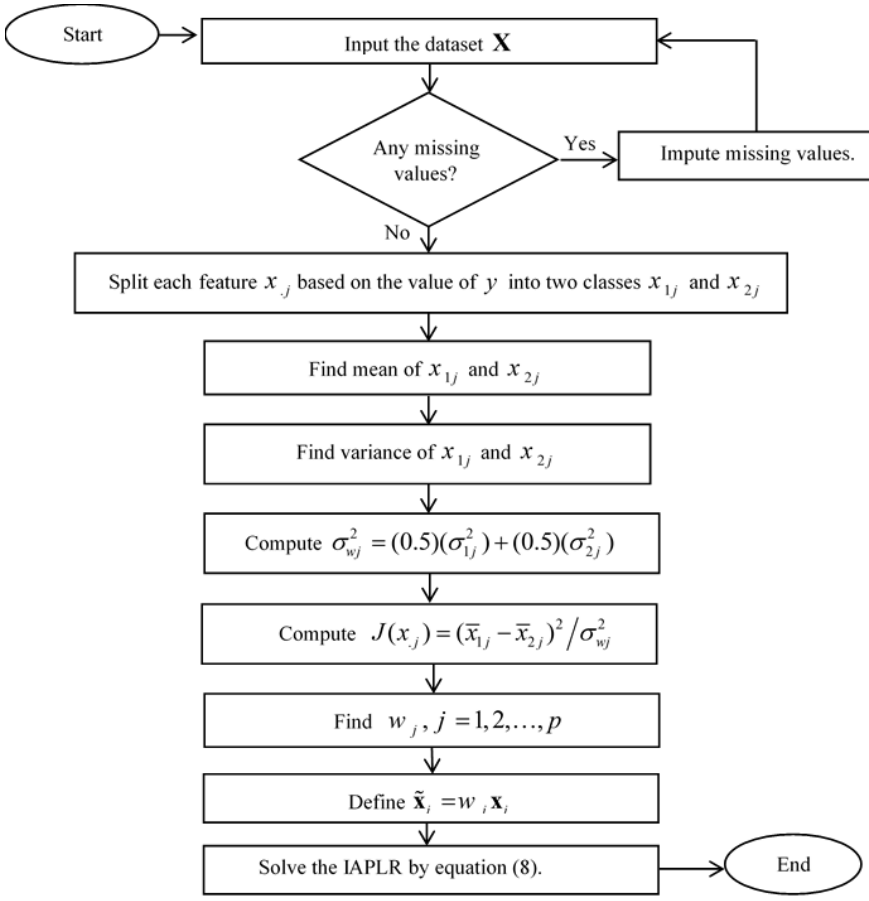

Fig. 1. Flowchart of IAPLR

2.4 Evaluation metrics

In this subsection, three evaluation metrics are used to evaluate the performance of the method. These criteria are widely used in the healthcare setting [40]. These criteria involve classification accuracy (CA), sensitivity (SEN), and specificity (SPE) that are given as:

$$CA = \frac{TN+TP}{FP+TP+TN+FN} \times 100\% \quad (9)$$

$$SEN = \frac{TP}{FN+TP} \times 100\% \tag{10}$$

$$SPE = \frac{TN}{TN+FP} \times 100\% \tag{11}$$

where *TP*, *FP*, *TN*, and *FN* are denoted in Figure 2. The greater the values of the applied assessment criteria, the better the classification performance is expected to be.

	Prediction (+)	Prediction (−)
Actual (+)	True Positive (TP)	False Negative (FN)
Actual (−)	False Positive (FP)	True Negative (TN)

Fig. 2. Confusion matrix of classification

2.5 Dataset description

In order to assess its effectiveness, the proposed method (IAPLR) is used for two datasets with varying numbers of genes and observations. These datasets are freely accessible and have been used by a large number of researchers in the previous. First, the colon cancer data set, in which the number of observations is 62 people (40 malignant tumors and 22 noncancerous cells) and 6500 genes. Affymetrix oligonucleotide array technology was used to get it. Only 2000 gene expressions were utilized in this data set, and they were selected based on the samples' lowest minimum intensity [41]. The second data set is the Bipolar disorder (Bip) dataset, which had a sample size of 61 observations, including 31 control observations and 30 bipolar disorder observations. Again, Affymetrix technology was used to capture the expression of 22,283 human genes. [42], [43].

3 Results and discussion

In this section, the datasets described above were considered to show various methods regarding feature selection with missing values. The proposed method (IAPLR) was demonstrated to be efficient throughout comparative experiments with Lasso and ALasso. We first applied these methods to complete data without missing data. We randomly partitioned each dataset into a training dataset with 70% of the samples and a test dataset with 30% of the samples to allow for a fair comparison. The 10-fold cross-validation (CV) was used with the training dataset 100 times in order to obtain the optimal value of λ, using the "glmnet" package in R. On the other hand, to evaluate the methods with missing values, the process as follows. First, we seed missing values in our datasets with the different rates (10%, 20%, 30%) using the "missForest" package in R. This study assumes no missing data in the response variable. Secondly, we used the "naniar" package of the programing language R to

impute the missing values. Thirdly, we applied penalized methods on imputing data as complete data. The average number of selected genes, the averaged CA, SEN, and SPE in both the training and testing datasets are presented in Tables 1 and 2.

It can be seen from the data in Tables 1 and 2. The proposed method selected fewer genes than the Lasso and ALasso in both colon and Bip datasets with different rates of missing values. For example, in the colon with 20% missing data, IAPLR selected 13 genes compared to 16 genes for Lasso and 15 genes for ALasso. On the contrary, we observed that Lasso usually produces the highest number of picked genes in both datasets.

Furthermore, we observe in Tables 1 and 2 that in both datasets used in this research, the average CA, SEN, and SPE in both the training and testing sets of IAPLR are much better than that of Lasso and ALasso. For instance, in colon data with 10% missing values, the CA of IAPLR in the training set is (96%), which is better than (93.91%) for Lasso and (93.82%) ALasso. Additionally, in Bip data with 30% missing values, the SEN of IAPLR is 87.93%, which is better than that of Lasso and ALasso, 81.39%, and 83.86%, respectively. The same conclusion can be made from the testing sets in the colon and Bip datasets with different rates of missing values.

To further highlight the performance of the IAPLR, it is required to conduct statistical tests in order to investigate whether the differences in classification accuracy obtained in Tables 1 and 2 are statistically significant or not. In this study, the paired t-test was utilized to analyze the data. Tables 3 and 4 present the findings. The relative improvement in the mean of average accuracy that the proposed method provides in comparison to the other methods is represented by the column "improvement". In addition, Tables 3 and 4 demonstrate that there is a statistically significant difference between our proposed method, IAPLR, and each competing approach at the 5% level of significance.

Table 1. The averaged criteria over 100 times for the training and testing colon dataset

Missing %	**Methods**	**Genes**	**Training set**			**Testing set**		
			% CA	***% SEN***	***% SPE***	***% CA***	***% SEN***	***% SPE***
No missing	Lasso	14	94.14	92.20	94.64	79.53	83.43	86.92
	ALasso	14	94.83	93.22	95.22	83.40	85.41	86.91
	Proposed	12	96.12	95.85	96.83	87.91	89.42	88.33
10%	Lasso	14	93.82	90.63	94.00	81.14	82.34	86.74
	ALasso	15	93.91	90.72	95.00	84.11	84.41	86.91
	Proposed	13	96.00	95.00	96.00	88.53	90.82	88.43
20%	Lasso	16	89.80	87.50	89.91	81.34	80.12	75.44
	ALasso	15	91.90	87.71	90.03	80.53	78.02	75.42
	Proposed	13	93.61	88.83	92.42	85.44	82.74	81.42
30%	Lasso	17	83.71	80.00	83.52	78.50	77.73	79.32
	ALasso	15	86.64	84.40	89.24	80.62	79.24	84.82
	Proposed	15	90.00	89.00	91.00	88.34	83.23	86.43

Table 2. The averaged criteria over 100 times for the training and testing Bip dataset

Missing %	Methods	Genes	Training set			Testing set		
			% CA	*% SEN*	*% SPE*	*% CA*	*% SEN*	*% SPE*
No missing	Lasso	18	90.69	91.70	92.97	79.53	83.53	86.12
	ALasso	16	92.62	94.31	94.51	83.40	85.65	86.11
	Proposed	15	94.70	95.55	95.98	88.91	89.30	88.15
10%	Lasso	18	90.19	90.14	91.36	77.97	81.77	84.58
	ALasso	15	91.42	93.31	94.00	82.46	84.72	85.41
	Proposed	15	93.60	94.55	94.88	88.45	88.69	90.80
20%	Lasso	19	85.39	86.38	87.82	77.07	79.64	80.55
	ALasso	16	87.28	88.48	90.94	81.14	83.46	86.64
	Proposed	14	90.64	90.87	91.28	85.78	86.57	89.64
30%	Lasso	20	80.63	81.39	81.84	76.58	78.36	78.63
	ALasso	18	84.39	83.86	82.18	80.78	79.23	80.40
	Proposed	16	88.93	87.80	87.55	83.72	83.97	82.87

Table 3. Significant test results of paired *t*-test for the training and testing colon dataset

Missing %	Methods	Training set: Average accuracy		Testing set: Average accuracy	
		Improvement	*p-value*	*Improvement*	*p-value*
No missing	Lasso	2.10%	0.0021 (*)	10.54%	0.0000 (*)
	ALasso	1.36%	0.0053 (*)	5.41%	0.0001 (*)
10%	Lasso	2.32%	0.0014 (*)	9.11%	0.0000 (*)
	ALasso	2.23%	0.0022 (*)	5.26%	0.0001 (*)
20%	Lasso	4.24%	0.0005 (*)	5.04%	0.0001 (*)
	ALasso	1.86%	0.0020 (*)	6.10%	0.0001 (*)
30%	Lasso	7.51%	0.0002 (*)	12.54%	0.0000 (*)
	ALasso	3.89%	0.0006 (*)	9.58%	0.0001 (*)

(*) significant at $\alpha = 0.05$

Overall, these results indicate that the IAPLR has been effectively applied to improve gene selection, classification, and dealing with missing values in high-dimensional data. It achieved higher CA, SEN, and SPE in both the training and testing datasets. Hence, IAPLR is nominated as a potential gene selection approach since it can simultaneously satisfy all three of these criteria. Furthermore, when compared to competitor approaches, the proposed penalized technique is the most effective classification technique. This illustrates that the weights of the genes are taken into account by the IAPLR method.

Table 4. Significant test results of paired *t*-test for the training and testing Bip dataset

Missing %	Methods	Training set: Average accuracy		Testing set: Average accuracy	
		Improvement	***p-value***	***Improvement***	***p-value***
No missing	Lasso	4.42%	0.0023(*)	11.79%	0.0000 (*)
	ALasso	2.25%	0.0061(*)	6.61%	0.0001 (*)
10%	Lasso	3.78%	0.0023(*)	13.44%	0.0000 (*)
	ALasso	2.38%	0.0057(*)	7.26%	0.0001 (*)
20%	Lasso	6.14%	0.0001(*)	11.30%	0.0000 (*)
	ALasso	3.85%	0.0021(*)	5.72%	0.0001 (*)
30%	Lasso	10.29%	0.0000(*)	9.32%	0.0000 (*)
	ALasso	5.38%	0.0001(*)	3.64%	0.0006 (*)

(*) significant at $\alpha = 0.05$

4 Conclusion

In data analytics, the imputation of missing data is extremely important. Unfortunately, it is difficult to find a missing data imputation method that works for all types of datasets. Although there has been significant progress in the methods and tools for variable selection, missing data often occurs in extensive, complicated research and which can make data analysis challenging. In this study, it is mainly focused on improving the performance of penalized logistic regression models and handling missing values in high-dimensional data through the IAPLR method. The IAPLR, Lasso, and ALasso were applied to two datasets (colon and dip) in the presence of the different rates of missing values. The findings of comparative experiment demonstrated that the efficiency of IAPLR in the presence of missing data is better than the efficiency of the other two techniques in terms of CA, SEN, and SPE. The findings also showed that the IAPLR method for classification and gene selection is a statistically significant one.

5 Acknowledgment

We would like to express our gratitude to Taif University for their financial sponsorship. Also, thanks to Universiti Teknologi Malaysia for making the facilities available.

6 References

[1] W. Jiang, J. Josse, M. Lavielle, and T. Group, "Logistic Regression with Missing Covariates -- Parameter Estimation, Model Selection and Prediction within a Joint-Modeling Framework," *Comput. Stat. Data Anal.*, vol. 145, p. 106907, May 2018. https://doi.org/10.1016/j.csda.2019.106907

[2] Y. Deng, C. Chang, M. S. Ido, and Q. Long, "Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data," *Sci. Rep.*, vol. 6, no. 1, p. 21689, Feb. 2016. https://doi.org/10.1038/srep21689

[3] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *J. Big Data*, vol. 7, no. 1, p. 37, Dec. 2020. https://doi.org/10.1186/s40537-020-00313-w

[4] Z. Zhang, "Missing values in big data research: some basic skills," *Ann. Transl. Med.*, vol. 3, no. 21, p. 323., 2015. https://doi.org/10.3978/j.issn.2305-5839.2015.12.11

[5] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean J. Anesthesiol.*, vol. 70, no. 4, p. 407, 2017. https://doi.org/10.4097/kjae.2017.70.4.407

[6] F. M. Zahid and C. Heumann, "Multiple imputation with sequential penalized regression," *Stat. Methods Med. Res.*, vol. 28, no. 5, pp. 1311–1327, May 2019. https://doi.org/10.1177/0962280218755574

[7] F. M. Zahid, S. Faisal, and C. Heumann, "Variable selection techniques after multiple imputation in high-dimensional data," *Stat. Methods Appt.*, vol. 29, no. 3, pp. 553–580, Sep. 2020. https://doi.org/10.1007/s10260-019-00493-7

[8] M. Hamim, I. El Moudden, M. D Pant, H. Moutachaouik, and M. Hain, "A Hybrid Gene Selection Strategy Based on Fisher and Ant Colony Optimization Algorithm for Breast Cancer Classification," *Int. J. Online Biomed. Eng.*, vol. 17, no. 02, p. 148, Feb. 2021. https://doi.org/10.3991/ijoe.v17i02.19889

[9] C. Liu and H. S. Wong, "Structured Penalized Logistic Regression for Gene Selection in Gene Expression Data Analysis," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 16, no. 1, pp. 312–321, Jan. 2019. https://doi.org/10.1109/TCBB.2017.2767589

[10] M. S. Al-Batah, B. M. Zaqaibeh, S. A. Alomari, and M. S. Alzboon, "Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers," *Int. J. Online Biomed. Eng.*, vol. 15, no. 08, p. 62, May 2019. https://doi.org/10.3991/ijoe.v15i08.10617

[11] M. S. Al-batah, "Ranked Features Selection with MSBRG Algorithm and Rules Classifiers for Cervical Cancer," *Int. J. Online Biomed. Eng.*, vol. 15, no. 12, p. 4, Aug. 2019. https://doi.org/10.3991/ijoe.v15i12.10803

[12] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, Jan. 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[13] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970. https://doi.org/10.1080/00401706.1970.10488634

[14] J. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *J. Am. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, Dec. 2001. https://doi.org/10.1198/016214501753382273

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[16] H. Zou, "The Adaptive Lasso and Its Oracle Properties," *J. Am. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006. https://doi.org/10.1198/016214506000000735

[17] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *Ann. Stat.*, vol. 37, no. 4, pp. 1733–1751, Aug. 2009. https://dx.doi.org/10.1214/08-AOS625

[18] S. Ghosh, "On the grouped selection and model complexity of the adaptive elastic net," *Stat. Comput.*, vol. 21, no. 3, pp. 451–462, Jul. 2011. https://doi.org/10.1007/s11222-010-9181-4

[19] Y. Chen *et al.*, "A global learning with local preservation method for microarray data imputation," *Comput. Biol. Med.*, vol. 77, pp. 76–89, Oct. 2016. https://doi.org/10.1016/j.compbiomed.2016.08.005

[20] D. B. Rubin, "Multiple imputation after 18+ years," *J. Am. Stat. Assoc.*, vol. 91, no. 434, pp. 473–489, 1996.

[21] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019.

[22] S. Van Buuren and K. Groothuis-Oudshoorn, "Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011. https://doi.org/10.18637/jss.v045.i03

[23] Y.-S. Su, A. E. Gelman, J. Hill, and M. Yajima, "Multiple imputation with diagnostics (mi) in R: Opening windows into the black box," *J. Stat. Softw.*, vol. 45, no. 2, pp. 1–31, 2011. https://doi.org/10.7916/D8VQ3CD3

[24] Y. Zhao and Q. Long, "Multiple imputation in the presence of high-dimensional data," *Stat. Methods Med. Res.*, vol. 25, no. 5, pp. 2021–2035, Oct. 2016. https://doi.org/10.1177/0962280213511027

[25] R. Holman and C. A. W. Glas, "Modelling non- ignorable missing- data mechanisms with item response theory models," *Br. J. Math. Stat. Psychol.*, vol. 58, no. 1, pp. 1–17, 2005. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

[26] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor, "Handling missing values in support vector machine classifiers," *Neural Networks*, vol. 18, no. 5–6, pp. 684–692, Jul. 2005. https://doi.org/10.1016/j.neunet.2005.06.025

[27] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons, 2004.

[28] J. Honaker, G. King, and M. Blackwell, "Amelia II: A program for missing data," *J. Stat. Softw.*, vol. 45, no. 7, pp. 1–47, 2011.

[29] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[30] A. M. Alharthi, M. H. Lee, Z. Y. Algamal, and A. M. Al-Fakih, "Quantitative structure-activity relationship model for classifying the diverse series of antifungal agents using ratio weighted penalized logistic regression," *SAR QSAR Environ. Res.*, vol. 31, no. 8, pp. 571–583, Aug. 2020. https://doi.org/10.1080/1062936X.2020.1782467

[31] X. Li, Y. Wang, and R. Ruiz, "A Survey on Sparse Learning Models for Feature Selection," *IEEE Trans. Cybern.*, pp. 1–19, 2020. https://doi.org/10.1109/TCYB.2020.2982445

[32] Z. Y. Algamal, M. H. Lee, A. M. Al-Fakih, and M. Aziz, "High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty," *J. Chemom.*, vol. 31, no. 6, p. e2889, Jun. 2017. https://doi.org/10.1002/cem.2889

[33] I. I. M. Manhrawy, M. Qaraad, and P. El- Kafrawy, "Hybrid feature selection model based on relief- based algorithms and regulizer algorithms for cancer classification," *Concurr. Comput. Pract. Exp.*, no. January, pp. 1–17, Jan. 2021. https://doi.org/10.1002/cpe.6200

[34] M. El Guide, K. Jbilou, C. Koukouvinos, and A. Lappa, "Comparative study of L 1 regularized logistic regression methods for variable selection," *Commun. Stat. - Simul. Comput.*, pp. 1–16, Apr. 2020. https://doi.org/10.1080/03610918.2020.1752379

[35] S. Doerken, M. Avalos, E. Lagarde, and M. Schumacher, "Penalized logistic regression with low prevalence exposures beyond high dimensional settings," *PLoS One*, vol. 14, no. 5, p. e0217057, May 2019. https://doi.org/10.1371/journal.pone.0217057

[36] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010. https://doi.org/10.18637/jss.v033.i01

[37] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[38] Z. Y. Algamal and M. H. Lee, "A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification," *Adv. Data Anal. Classif.*, vol. 13, no. 3, pp. 753–771, Sep. 2019. https://doi.org/10.1007/s11634-018-0334-1

[39] H. Peng, Y. Fu, J. Liu, X. Fang, and C. Jiang, "Optimal gene subset selection using the modified SFFS algorithm for tumor classification," *Neural Comput. Appl.*, vol. 23, no. 6, pp. 1531–1538, Nov. 2013. https://doi.org/10.1007/s00521-012-1148-2

[40] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, Jan. 2021. https://doi.org/10.1016/j.aci.2018.08.003

[41] U. Alon *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 12, pp. 6745–6750, 1999. https://doi.org/10.1073/pnas.96.12.6745

[42] M. M. Ryan, H. E. Lockstone, S. J. Huffaker, M. T. Wayland, M. J. Webster, and S. Bahn, "Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes," *Mol. Psychiatry*, vol. 11, no. 10, pp. 965–978, Oct. 2006. https://doi.org/10.1038/sj.mp.4001875

[43] Q. Shen, Z. Mei, and B.-X. Ye, "Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification," *Comput. Biol. Med.*, vol. 39, no. 7, pp. 646–649, Jul. 2009. https://doi.org/10.1016/j.compbiomed.2009.04.008

7 Authors

Aiedh Mrisi Alharthi is a Ph.D. candidate in statistics at the Department of Mathematical Sciences, Universiti Teknologi Malaysia (UTM). He received his M.Sc. degree from the Department of Mathematics and Statistics, Taif University, Saudi Arabia. His research interests are high-dimensional data, penalized (regularized) methods, gene selection in cancer classification, and missing values (Email: aiedh.harthi@gmail.com).

Muhammad Hisyam Lee is a Professor of Statistics in the Department of Mathematical Sciences, Faculty of Sciences, Universiti Teknologi Malaysia (UTM). He served as Vice President of the Malaysia Institute of Statistics between 2010 and 2014. Currently, he is serving as the Manager of Information Technology in the office of the Deputy Vice-Chancellor, Academic and International, UTM. His research interests include forecasting, time series analysis, and statistical quality control.

Zakariya Yahya Algamal received the B.S. degree in Statistics from University of Mosul, Mosul, Iraq, in 2001, the M.S. degree in Statistics from University of Mosul, Mosul, Iraq, in 2004, the Ph.D. degree in Mathematical Science/Statistics from Universiti Teknologi Malaysia (UTM), Malaysia, in 2016, and Post. Doctorate in Mathematical Science/Statistics from Universiti Teknologi Malaysia (UTM), Malay-

sia, in 2017. He is a professor in Statistics at University of Mosul. His research interest includes the development of high dimensional data, generalized linear models, bioinformatics, chemoinformatics, and optimization algorithms (Email: zakariya.algamal@uomosul.edu.iq).

Article submitted 2021-06-21. Resubmitted 2021-08-29. Final acceptance 2021-11-14. Final version published as submitted by the authors.