Artificial Intelligence Technologies and Applications
C. Chen (Ed.)

doi:10.3233/FAIA231362

Breast Cancer Classification Based on KNNI Imputation of Missing Data

Yong WANG[a,1], Rong ZHU[b2]
[a]*Center for Experimental Teaching and Equipment Management*
Qufu Normal University
Rizhao, China
[b]*Member, IEEE, School of Computer Science Qufu Normal University*
Rizhao, China

Abstract. Some disease datasets have different degrees of missing, which will lead to the problem of low classification accuracy. To improve the effectiveness of breast cancer disease detection and diagnosis, a classification prediction method combining KNNI and XGBoost was proposed and applied to the classification and analysis of breast cancer data. First, the KNNI method is used to impute the missing data in the breast cancer patient dataset; Then, the original dataset is equalized by the SMOTE oversampling method; Finally, XGBoost is used to extract features that are strongly related to breast cancer malignancies as the input of the model, optimize the XGBoost model by grid search algorithm, find the optimal model parameters, and classify and diagnose the breast cancer dataset. The experimental results show that KNNI can effectively recover the lost data, improve the data quality, and improve the subsequent classification accuracy. Applying imputation methods to flexibly apply missing data to machine learning methods holds great promise.

Keywords. Breast cancer, KNNI; Data preprocessing; XGBoost

1. Introduction

Despite advances in medicine, breast cancer remains the second leading cause of death worldwide [1]. Early detection, early diagnosis, and early treatment can improve the survival rate of patients, which is of great significance to patients. However, given the fact that the diagnosis results may be interfered with by the subjective factors of doctors, there will be deviations such as misdiagnosis, which will affect the patients to make reasonable plans for further treatment, and medical work will face important challenges.

At present, some machine learning algorithms have been used for early diagnosis or prediction of cancer [2-4]. For example, Hosseinpour et al. [5] achieved superior performance by predicting overall breast cancer risk through the improved random forest algorithm. In the medical field, there is still a lack of medical data, resulting in insufficient useful research information. So when using machine learning for prediction and discrimination, it may cause certain error interference [6]. Therefore, it is necessary to perform preprocessing such as interpolation on missing datasets first. Data duplication

[1] Corresponding Author: Yong WANG, Center for Experimental Teaching and Equipment Management, Qufu Normal University, Rizhao, China; email: wyongsd@126.com
[2] Rong ZHU, School of Computer Science Qufu Normal University, Rizhao, China; email: zhurongsd@qfnu.edu.cn

and data inconsistency can be screened by algorithms, and data missing may be due to data inaccessibility or omission during operation. Without missing value processing, some machine learning algorithms cannot even be used directly. Therefore, missing value imputation is a practical and challenging problem in the field of medical bioinformatics computing.

Currently, there are usually two methods for preprocessing datasets with missing values, namely deletion and interpolation [7]. The deletion method is very simple. It deletes cases or variables with missing values. It is most commonly used when the missing value rate is low [8]. However, if there are many missing values in the dataset, the deletion method is not suitable, as it will result in a large amount of information loss and significant bias and overfitting in the models used for training [9]. In this case, the preferred method for handling missing values is the imputation method.

The currently widely used imputation methods mainly include K nearest neighbor imputation (KNNI) [10], MissForest [11], and Multivariate imputation by chained equations (MICE) [12].

KNNI is an interpolation method based on the K nearest neighbors algorithm. The basic idea of KNNI is to first divide the data set into two sets, one set contains all complete samples (that is, samples without missing values), and the other set contains all incomplete samples (that is, samples with missing values). For each incomplete sample, its K nearest neighbors can be found in the complete sample set, and if the missing value is a categorical attribute, it can be filled in the mode of the attribute value of the K nearest neighbor sample; For missing values that are numerical attributes, the average attribute values of the K nearest neighbor samples should be filled in. Since the missing values of incomplete samples are obtained from "adjacent" samples, the KNNI algorithm will not add too much new sample information.

The MissForest [11] method is one of the most widely used methods for imputation of datasets with missing values. The MissForest algorithm utilizes a random forest algorithm to interpolate missing data.

The MICE method is to use the imputation method to fill the data set m times (m>1), that is, to generate multiple intermediate imputation values, and then generate a complete data set. This imputation method is more comprehensive.

Most classification methods assume that the number of class observations is balanced. However, many datasets, in reality, are not balanced [13]. Therefore, the classification accuracy of the classifier will be affected by the uneven distribution of the data and cannot be correctly classified.

Aiming at the problem of how to effectively learn, train, and predict classification from class imbalanced data, more and more scholars have paid attention to [14], resulting in many methods for imbalanced data processing. One of the most widely used methods for handling imbalanced datasets is the synthesis of a few oversampling techniques (SMOTE) [15].

In this paper, KNNI and Extreme Gradient Boosting (XGBoost) are combined, the SMOTE method is used to overcome the imbalance problem of the dataset, and XGBoost is used for feature selection to realize the benign and malignant classification of breast cancer. Through experiments, the performance of the algorithm was compared and verified from the aspects of Accuracy, Sensitivity, Precision, and Matthews correlation coefficient (MCC), which provided a reference for the treatment and research of breast cancer.

2. Design and Usage of KNNI-XGBoost

2.1 The Overall Structure of KNNI-XGBoost

The flow of the KNNI-XGBoost classification method constructed in this paper is shown in Figure 1.

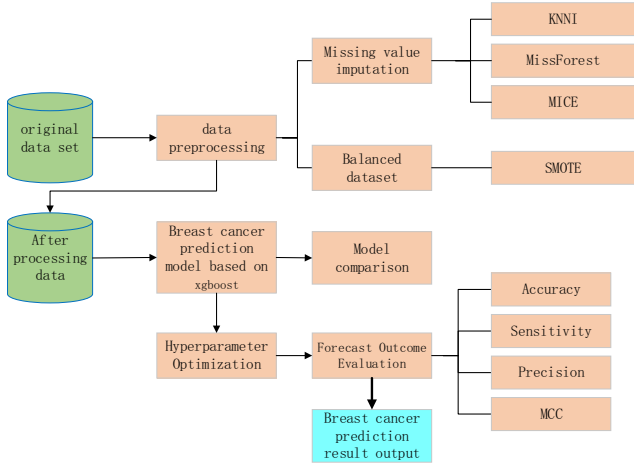

Fig. 1. The overall structure of KNNI-XGBoost.

2.2 Data Preprocessing

Since the integrity and availability of the data are uneven, the processing efficiency of such data is not high and the model operation results obtained are not good, so it is particularly important to preprocess the data. In this paper, data preprocessing is divided into two aspects: missing value processing and balancing the data set.

2.2.1 Missing Value Handling

The KNNI algorithm first selects the K samples closest to the missing samples based on distance measurement and then weights the K samples to estimate the missing data of the missing samples.

Suppose the distance between the two samples x_1 and x_2 is $d(x_1,x_2)$, the distance metric calculation uses the Heterogeneous Euclidean Overlap metric (HEOM), $d(x_1,x_2)$ can be defined as:

$$d(x_1,x_2)\sqrt{\sum_{k=1}^{p} d_j(x_{1k},x_{2k})^2} \tag{1}$$

where, $d_k(x_{1k},x_{2k})$ represents the distance between the kth variable of the sample x_1 and x_2.

If there is a missing value for the kth variable of the sample x_1 or x_2, the maximum distance value of 1 is returned. For KNNI, if there is a missing value in the kth variable

of the sample x_1, it is necessary to select the K samples that are closest to the sample x_1, and the kth variable of these K samples has no missing value. The set consisting of arranging the distances of the selected K nearest samples from near to far is expressed as follows:

$$\mathrm{S}_{x_1} = \{s_a\}_{a=1}^{K} \tag{2}$$

where s_1 is the sample closest to the sample x_1.

2.2.2 Balanced Dataset

When dealing with unbalanced datasets, there are generally two methods: under-sampling and oversampling. The under-sampling technique discards some large-class samples, resulting in a waste of data; Simple oversampling methods can lead to overfitting problems. The SMOTE method is an oversampling technique to solve the imbalance between classes. Aiming at the imbalance phenomenon in the sample, this paper uses the SMOTE method to oversample the experimental dataset.

2.3 XGBoost Model

XGBoost [16] is a boosting algorithm based on the CART regression tree to classify and predict data sets. The XGBoost algorithm expands the loss function by the second-order Taylor formula. After the original data is cleaned, such as missing value filling and dimensionality reduction, the samples are input into the XGBoost model for calculation.

For the jth sample x_j in the dataset, the predicted output is $\hat{y}_j$, which can be expressed as:

$$\hat{y}_j = \sum_{a=1}^{K} f_a(x_j) \tag{3}$$

where, x_j is the jth sample, $\hat{y}_j$ is the predicted value of the jth sample, K is the total number of regression trees and f_a is the a th regression tree. The objective function of the XGBoost model is represented as follows:

$$Obj(\theta) = \sum_{j=1}^{n} l(y_j, \hat{y}_j^{(t-1)} + f_t(x_j)) + \sum_{a} \Omega(f_a) \tag{4}$$

where, y_j is the real value, $\hat{y}_j$ is the predicted output, $\sum_{j=1}^{n} l(y_j, \hat{y}_j^{(t-1)} + f_t(x_j))$ represents the loss function, $\sum_{a} \Omega(f_a)$ represents the regularization term.

Using L_2 regularization to expand the regularization term, the objective function is represented as follows:

$$Obj(\theta)=\sum_{j=1}^{n} l(y_j, \hat{y}_j^{(t-1)} + f_t(x_j))+\gamma \mathrm{T} + \frac{1}{2}\lambda \left\| \theta \right\|^2 \tag{5}$$

where, T is the number of leaf nodes, θ is a vector composed of all leaf node values of the decision tree, γ controls the number of leaves, and λ is L_2 regular parameter.

The XGBoost algorithm adopts the gradient boosting strategy, and the model continuously adds new regression trees, $\hat{y}_j^{(t)}$ represents the predicted value of the j th sample in t round iteration, and adds a new tree function is represented by $f_t(x_j)$. The derivation process is as follows:

$$\begin{aligned}
&\hat{y}_j^{(0)} = 0 \\
&\hat{y}_j^{(1)} = f_1(x_j) = \hat{y}_j^{(0)} + f_1(x_j) \\
&\hat{y}_j^{(2)} = f_1(x_j) + f_2(x_2) = \hat{y}_j^{(1)} + f_2(x_j) \\
&\cdots \\
&\hat{y}_j^{(t)} = \sum_{a=1}^{t} f_a(x_j) = \hat{y}_j^{(t-1)} + f_t(x_j)
\end{aligned} \tag{6}$$

As an improved machine learning model, XGBoost has better prediction performance for small and medium data sets and has the advantages of strong algorithm scalability, strong tolerance for outliers, fast parallel speed, and adding regular terms to prevent overfitting [17].

3. Experimental Results and Analysis

3.1 Dataset Overview

The breast cancer dataset used in this paper is publicly available in the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)). The dataset contains 569 samples, and each sample contains 30 feature components. Disease status is divided into two categories: Benign (1) and malignant (0). The number of benign samples in the dataset is 357, and the number of malignant samples is 212.

3.2 Evaluation Indicators

According to the true class and the predicted class, it is divided into true class (TP), true negative class (TN), false positive class (FP), and false negative class (FN). Accuracy, Sensitivity, Precision, and MCC were used to evaluate the classification accuracy.
The calculation formulas of the four indicators are defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{7}$$

$$\Pr e = \frac{TP}{TP + FP} \tag{8}$$

$$Sen = \frac{TP}{TP + FN} \tag{9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{10}$$

3.3 Comparative Analysis of Mainstream Machine Learning Models

Firstly, all sample points in the cancer dataset are divided into training and testing sets in an 8:2 ratio. Then, the data is imported into 10 mainstream machine-learning models for classification training. The results of this experiment were verified through programming in Python language. The 10-fold cross-validation method was used in the experiment to evaluate the classification performance of different machine learning models. The classification performance of 10 mainstream machine learning models was compared in the experiment. Four indicators (Accuracy, Sensitivity, Precision, and MCC) are used for model evaluation, and each model uses the default parameters. The classification evaluation results of 10 mainstream machine learning algorithms are shown in Figure 2.

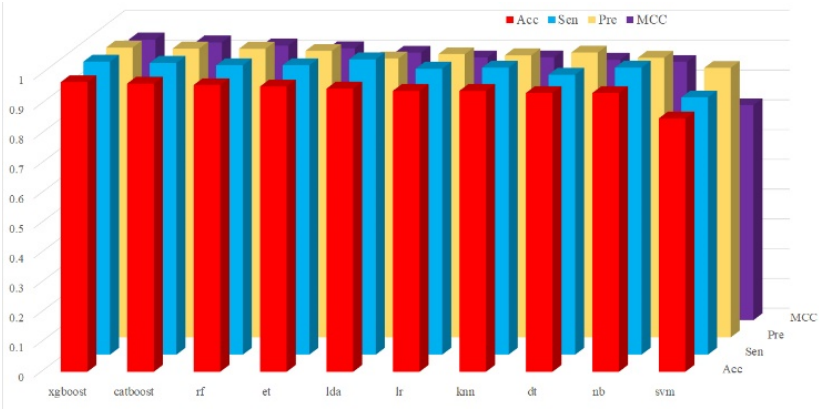

Fig. 2. Comparison of classification performance of mainstream machine learning models

In Figure 2, among the 10 mainstream machine learning models compared to the breast cancer dataset, the XGBoost model has the best classification results. Therefore, this paper chooses the XGBoost model as the model for breast cancer classification prediction.

3.4 Comparison of Experimental Results of Different Interpolation Methods

In this section, we use KNNI, MissForest, and MICE methods to interpolate the missing data of the breast cancer dataset, and then compare and analyze the XGBoost classification performance of the data interpolated by different methods. The loss rates of cancer data in the experiment were set at 10% and 20%, respectively. The comparison results of classification performance under different data loss ratios are shown in Figures 3-4.

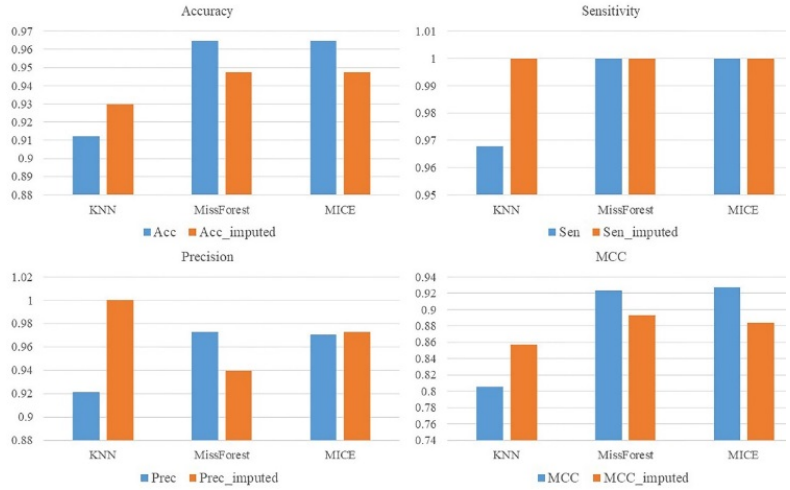

Fig. 3. XGBoost classification results at a 10% loss level

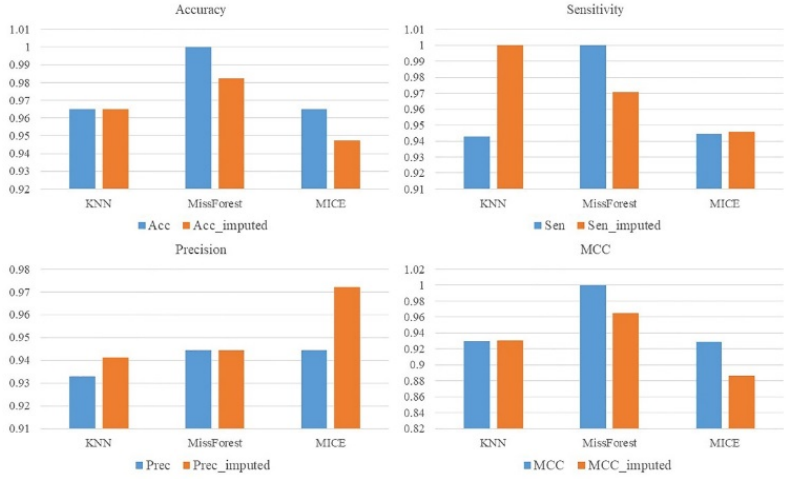

Fig. 4. XGBoost classification results at 20% loss level

It can be seen from Figures 3-4 that when the missing rates are 10% and 20%, the data classification results after the KNNI method imputation are the best. Therefore, the KNNI method was finally chosen to impute the missing data, and then put into the XGBoost classifier for classification and diagnosis of breast cancer data.

3.5 Feature Selection

In general, the more features, the higher the classification effect. However, too many features will seriously reduce the learning efficiency of the model and increase the amount of computation and computation time. Therefore, this paper performs feature selection on related data, and the XGBoost model can show the importance of different features according to the size of the value. Figure 5 shows the top ten most important features after sorting.

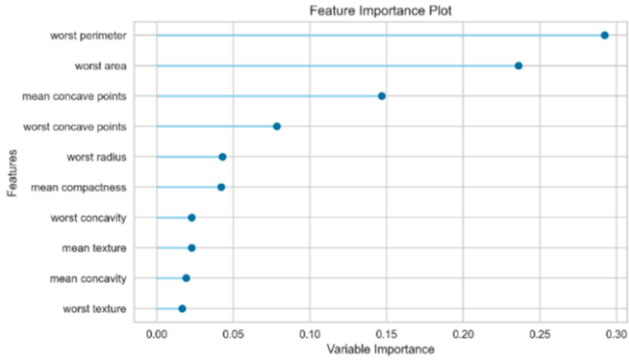

Fig. 5. The top 10 important features of the XGBoost model arrangement

As shown in Figure 5, features such as worst perimeter, worst area, mean concave points, worst concave points, and worst radius have significant effects on the model. Therefore, this paper selects the dataset composed of 10 features in Figure 5 for further classification prediction.

3.6 XGBoost Model Hyperparameter Tuning

Due to excessive parameter optimization, the optimization time can be too long. The number of parameters in the XGBoost model is relatively large. So, in this article, only a few main parameters were selected for optimization.

Grid Search is a commonly used parameter adjustment method, which can automatically arrange and combine to filter out the best parameter combination. Grid Search first selects the parameter that currently has the greatest impact on the tuning model, searches in order by giving the value interval until it is optimal, then tunes the next parameter with greater impact, and so on, until all the parameters tuning is over and the best parameter combination is selected. The parameter optimization range and results that make use of the Grid Search method can be seen in Table 1.

Table 1 The optimal parameter combination of the XGBoost model

XGBoost main hyperparameters	Value list of Grid Search	Final parameter value
n_estimators	[10 20 30 40 50 60 70 80 90]	50
gamma	[0.0, 0.1, 0.2, 0.3, 0.4]	0.1
colsample_bytree	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]	0.3
max_depth	[1,2,3,4,5,6,7,8,9]	4
reg_alpha	[0.0, 0.1, 0.2, 0.3, 0.4]	0.2
subsample	[0.6, 0.7, 0.8, 0.9]	0.7

4. Conclusion

In this paper, a breast cancer data classification analysis model is developed by combining KNNI and XGBoost. The performance of 10 mainstream machine learning classification analysis models is compared under the same training and test data, and it is found that XGBoost has the best performance. Therefore, XGBoost is applied to breast cancer data classification prediction. Compared with the three current mainstream imputation methods for dealing with missing data, the most suitable KNNI imputation method for

breast cancer data was selected. First, the dataset with missing values was preprocessed, and then the XGBoost method was used to classify and analyze the processed dataset. In the experiment, the ten most important features were selected for classification prediction, thereby reducing structural complexity and improving accuracy. The classification and analysis of machine learning methods on breast cancer datasets are of great significance to assist doctors in the screening and treatment of breast cancer-related diseases, and to detect patients' conditions earlier.

Acknowledgments

This work is supported in part by the Shandong Province Graduate Education Quality Improvement Plan Project (SDYAL21092).

References

[1] J. Ferlay et al., "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," International Journal of Cancer, Vol. 136, No. 5, pp. E359-E386, 2015.
[2] H. Asri, H. Mousannif, and H. Al Moatassim, "A Hybrid Data Mining Classifier for Breast Cancer Prediction," Cham, 2020, pp. 9-16: Springer International Publishing.
[3] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors," in IEEE Region 10 Humanitarian Technology Conference, 2018.
[4] A. Sharma, S. Kulshrestha, and S. Daniel, "Machine learning approaches for breast cancer diagnosis and prognosis," in International Conference on Soft Computing and its Engineering Applications.
[5] M. Hosseinpour, S. Ghaemi, S. Khanmohammadi, and S. Daneshvar, "A hybrid high-order type-2 FCM improved random forest classification method for breast cancer risk assessment," Applied Mathematics and Computation, Vol. 424, Jul 1 2022, Art. no. 127038.
[6] Myers and R. W., "Handling Missing Data in Clinical Trials: An Overview," Drug Information Journal, Vol. 34, No. 2, pp. 525-533, 2000.
[7] A. M. Sefidian and N. Daneshpour, "Missing value imputation using a novel grey-based fuzzy c-means, mutual information based feature selection, and regression model," Expert Systems with Applications, Vol. 115, No. JAN., pp. 68-94, 2019.
[8] Q. Lan, X. Xu, H. Ma, and G. Li, "Multivariable data imputation for the analysis of incomplete credit data," Expert Systems with Applications, Vol. 141, No. Mar., pp. 112926.1-112926.12, 2020.
[9] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," Expert Systems with Applications, Vol. 42, No. 13, pp. 5621-5631, 2015.
[10] T. Olga et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, Vol. 17, No. 6, pp. 520-525, 2001.
[11] D. J. Stekhoven and P. Buehlmann, "MissForest-non-parametric missing value imputation for mixed-type data," Bioinformatics, Vol. 28, No. 1, pp. 112-118, Jan 1, 2012.
[12] S. V. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate Imputation by Chained Equations in R," Journal of Statistical Software, Vol. 45, No. 3, pp. 1-67, 2011.
[13] X. Tao et al., "Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering," Information Sciences, Vol. 519, pp. 43-73, 2020.
[14] H. He and E. A. Garcia, "Learning from Imbalanced Data," Ieee Transactions on Knowledge and Data Engineering, Vol. 21, No. 9, pp. 1263-1284, 2009.
[15] M. Nakamura, Y. Kajiwara, A. Otsuka, and H. Kimura, "LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data," Biodata Mining, Vol. 6, Oct 2, 2013, Art. No. 16.
[16] A. Ogunleye and Q.-G. Wang, "XGBoost Model for Chronic Kidney Disease Diagnosis," Ieee-Acm Transactions on Computational Biology and Bioinformatics, Vol. 17, No. 6, pp. 2131-2140, Nov 1 2020.
[17] S. Ji, X. Wang, W. Zhao, and D. Guo, "An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise," Mathematical Problems in Engineering, Vol. 2019, Sep 16, 2019, Art. No. 8503252.