ISSN 2959-6157

Generalization Investigation in Heart Disease Prediction: Comparative Analysis of MLP, Random Forest, and SVM with Cross-Domain Datasets

Shansong Zhou

[1]Computer Science and Technology (Data Science), Lanzhou University, Lanzhou, China
zhoushs21@lzu.edu.cn

Abstract:

The traditional way of predicting heart disease is usually through the subjective judgment of doctors, which is high subjectivity. Current machine learning methods do not pay enough attention to its universality. The two datasets used in this study are from Kaggle and preprocessed in advance, including standardization by using Z-Score and dimensionality reduction by KPCA. The two datasets were designated as the source domain and the target domain. The source domain dataset was further divided into two subsets with an 80:20 split, where 80% was used for training the model and the remaining 20% served as the test set. The trained model was then applied to the target domain to compare the differences in prediction results. Through the exploration of Multilayer Perceptron (MLP), Random Forest and Support Vector Machine (SVM), for the data set used in this study, when MLP uses a simple structure, the difference in prediction accuracy dropped from 13.8% to 5.07. For Random Forest, by increasing the number of decision trees and decreasing the minimum number of samples required for splitting, the difference is reduced from 15.6% to 9.32%. By modifying the penalty parameter C value of SVM, the difference on different datasets is reduced from 13.39% to 4.46%. This study is one of the few to explore the generalizations of heart disease prediction. The results demonstrate that the generalization performance of the model for heart disease prediction can be significantly enhanced through appropriate modifications to its structure and hyperparameters.

Keywords: Machine learning; heart disease; generalization.

1. Introduction

Heart disease is one of the foremost causes of mortality on a global scale and imposes a substantial burden on healthcare systems. According to the World Heart Federation, approximately 20.5 million people die from cardiovascular disease each year globally [1]. Thus, the prediction and prevention of cardiovascular disease will make a significant contribution to both the medical field and public health. Anticipating the risk early can help reduce the risk of heart disease or detect it earlier. However, the traditional disease prediction only relies on the subjective judgment of doctors. Such methods have the disadvantages of high subjectivity and relatively low accuracy, which often fails to detect the existence of diseases in time.

With the advancement of information technology, an increasing number of machine learning algorithms have provided a viable approach for the healthcare industry, demonstrating significant potential in enhancing medical services and disease prevention [2]. An individual's likelihood of developing a disease can be predicted by the machine learning model with some specific features. In recent years, there has been a lot of research in this field or other diseases. For example, using data from 974 elderly patients with Coronary Heart Disease (CHD) to develop and validate a robust predictive model by their demographics, tests, comorbidities, and drugs for one-year mortality in elderly coronary CHD patients with anemia using machine learning methods [3]. Madrid et al. identified distinct clusters of CAD individuals based on Electrocardiographic (ECG) morphological phenotypes using unsupervised learning with the data from 15-second ECGs (lead I) from participants in the UK Biobank imaging study with prevalent CAD (N = 1,198) and test the association of each cluster with the risk of prevalent AF, HF, or VA, and evaluate their performance in predicting incident events during the follow-up period [4]. Based on CACS and clinical factors, machine learning models including Random Forest (RF), Radial Basis Function Neural Network (RBFNN), Support Vector Machine (SVM),

k-Nearest Neighbors (KNN), and Kernel Ridge Regression (KRR) were utilized to assess the risk of CHD [5]. Neural network also demonstrated strong performance in disease prediction. For example, a system was developed using multi-layer neural networks trained with convolutional and simulated convolutional neural networks, which achieved an accuracy of 89% in studies on the UCI StatLog Heart Disease dataset [6]. These methods do play an important role in disease prediction, but there also has a problem is that these researches often involve model development and prediction based on only one data set with the same distribution. This can lead to models that work well on one data set that may perform poorly on another.

This study aims to use two data sets, one as the training set and another as the test set. The model trained on the training set is applied to the test set to investigate the difference in results across different datasets. And research focuses on finding a method of improving the accuracy of prediction on different data sets by modifying the model structure, hyperparameters and other factors to get a superior performance. This research enhances the practicality of the model in different scenarios to help people have a more accurate prediction.

2. Method

2.1 Dataset Preparation

In this study, two distinct datasets from Kaggle are utilized to investigate methods for enhancing the robustness and applicability of the model to different scenarios. The dataset employed for training and as the source domain is the Heart Disease Dataset, which is a comprehensive dataset combined from 5 popular heart disease datasets: Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog (Heart) Data Set [7]. It has 1, 190 rows and 12 columns of data. Another dataset used for comparison testing and as the target domain is the heart disease dataset consisting of medical information of patients with heart diseases [8]. It has 1, 049 rows and 13 columns of data. In this research, 10 features common to the two datasets were selected, such as age, sex and chest pain type.

Data preprocessing used in this study consists of three parts. The first step is the missing value processing and check for duplicated values. The absence of certain eigenvalues will cause the model cannot fully utilize the data, thereby diminishing its accuracy. Duplication of data values can lead to overfitting; the model will play well in the training set and have a poor performance in the test set. After removing missing and duplicate values, 272 pieces of data were removed from the training set, while the test set remained unchanged. After that, the data was standardized. This research use Z-score normalization to transform data from different features into a distribution with a mean of 0 and a standard deviation of 1, preventing imbalances due to different feature scales and mitigating the impact of outliers on data distribution.

In addition, some data visualization work was performed to better understand the data, such as observations of numerical variables data distribution (Fig. 1).

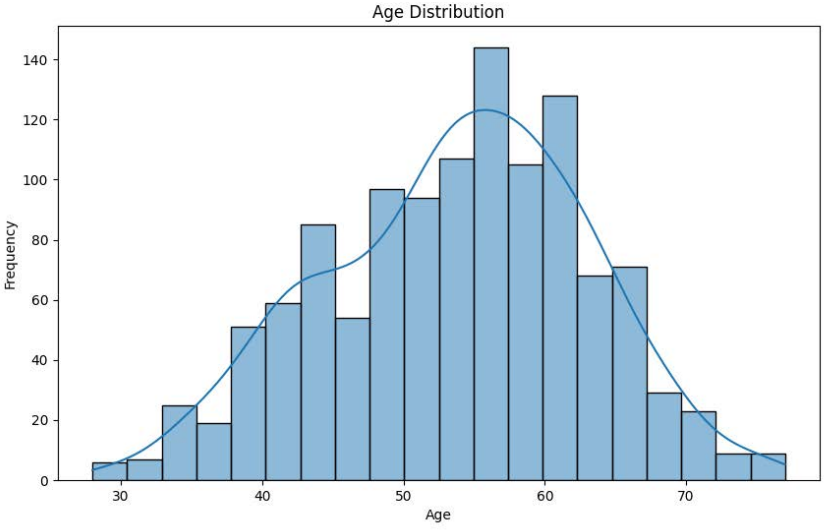

Fig. 1 Age Distribution of Source Domain (Photo/Picture credit: Original).

2.2 Machine Learning Model-based Prediction

This study used Multilayer Perceptron (MLP) model,

Random Forest model and Support Vector Machine (SVM) model to explore how to improve the robustness and its applicability to different data sets of the model. Kernel Principal Component Analysis (KPCA) is employed in this research to reduce the dimensionality of the data, mitigate the complexity of the feature space, and enhance computational efficiency. KPCA is a nonlinear extension of Principal Component Analysis (PCA) and a commonly used data dimensionality reduction technique. It captures the nonlinear features of the data by mapping it into a high-dimensional feature space using the kernel trick. Through KPCA with Radial Basis Function (RBF) as the core function, the final data dimension is reduced to 9 after tuning.

For the data set used for training, it is first divided into 8:2, 80% of the data is used to train the model, and 20% of the data is used to test. Then, for the data set to test, the model trained by the training data set is used to test. The same model is used to test 20% of the data divided into the training data set and the test data set to observe the difference in results. In this study, accuracy is primarily used as the evaluation index to assess the precision of the model's predictions.

2.2.1 MLP

MLP is a basic feedforward artificial neural network, which consists of an input layer, an output layer, and one or more intermediate hidden layers [9]. The simplest MLP contains only one hidden layer. The layers of the MLP are fully connected to each other, and the neurons in each layer are connected to all the neurons in the last layer. It introduces nonlinearity through the activation function. In the process of forward propagation, each layer of neurons performs weighted summation of inputs and activates them to generate the final prediction result. During the training process, MLP uses the loss function to assess the accuracy of the model and updates the weight and bias by backpropagation.

In this study, the model structure initially chosen was a neural network structure that connects an input layer, 9 fully connected layers after the input layer, an output layer and 9 residual blocks added after each fully connected layer. The adjusted model is a simple model with one input layer, three hidden layers, and one output layer. And the activation function used in the hidden layers of both models is the Rectified Linear Unit (ReLU).

2.2.2 Random Forest

Random forest algorithm is an ensemble learning algorithm based on decision tree, which combines multiple decision trees together and synthesizes the results of multiple trees to improve its performance [10]. When constructing each decision tree, some features not all the features are randomly selected to find the best split, which increases the difference between trees and reduces the risk of overfitting [11]. It can be used for both classification tasks and regression tasks.

In this research, the initial model is a model with 50 decision trees, the depth of the tree is unlimited, the minimum number of samples required for splitting is 14, and the minimum number of samples for leaf nodes is 1. The adjusted model increases the number of decision trees to 100 and reduces the minimum number of samples required for splitting to 2.

2.2.3 SVM

SVM is a classical supervised learning algorithm that finds the maximum boundary between different classes by finding an optimal hyperplane [12]. SVM makes the data linearly separable by mapping the samples into a high-dimensional feature space. For cases where a linear hyperplane could not be found to separate the data in the original input space, SVM introduces kernel functions to shift the computational complexity from the high-dimensional feature space to the original input space. In general, SVM is a powerful machine learning algorithm that can handle data in high-dimensional spaces and perform classification and regression analysis by finding the optimal hyperplane.

In this study, penalty parameter C of the initial model is chosen to be 1 and the kernel function is RBF. Gamma is a scale which means the value of gamma is automatically calculated based on the number and variance of the data features. The adjusted model increases the number of C to 10 and 100.

3. Results and Discussion

3.1 The Performance of Models

For the results of MLP shown in Table 1, with the complex structure of the initial model, the test of the source domain shows a good result, the accuracy reaches 0.8804, but when the target domain is used as the test set, it shows a great difference, only 0.7424. And for the adjusted model, it has an accuracy of 0.8370 on the test data of the source domain but achieves 0.7863 on the target domain.

Table 1. The Accuracy of MLP

Model	Accuracy of Source Domain	Accuracy of Target Domain
initial model	0.8804	0.7424
adjusted model	0.8370	0.7863

For the results of Random Forest shown in Table 2, as the initial model with 50 decision trees and the minimum number of samples required for splitting is 14, the accuracy of the test on the source domain is 0.8967, but when the target domain is used as the test set, it just reaches 0.7461. And for the adjusted model, it achieves an accuracy of 0.8804 on the source domain and 0.7872 on the target domain.

Table 2. The Accuracy of Random Forest

Model	Accuracy of Source Domain	Accuracy of Target Domain
initial model	0.8967	0.7461
adjusted model	0.8804	0.7872

For the SVM model shown in Table 3, the initial model achieves an accuracy of 0.8695 on the source domain but only 0.7356 on the target domain. With the adjusted model(C=10), the accuracy of source domain is 0.8423 while the accuracy of the target domain is 0.7929. And when the penalty parameter C is changed to 100, the accuracy of target domain remained unchanged, while the accuracy of the target domain increased to 0.7977.

Table 3. The Accuracy of SVM

Model	Accuracy of Source Domain	Accuracy of Target Domain
initial model	0.8695	0.7356
adjusted model(C=10)	0.8423	0.7929
adjusted model(C=100)	0.8423	0.7977

3.2 Analyze and Discuss the Results

Through observation of the results, it can be found that for the MLP although the simpler structure does not perform as well on the training set as the more complex structure, it performs much better on the test set than the complex structure. For the initial model, there is a 13.8% difference in accuracy between the two datasets, while for the modified model, the difference is only 5.07%. The reason for this phenomenon is that for the initial model, its structure is complex and leads to overfitting. Although the accuracy is high on the training data, the performance is significantly worse on the datasets with different distributions. For the adjusted model, due to its simpler structure, it has improved generalization ability and is more conducive to the data sets with different distributions. It has better robustness and applicability. For Random Forest, the accuracy on different datasets is reduced from 15.06% to 9.32%. The phenomenon can be attributed to two factors. On one hand, the increase in the number of decision trees enhances the model's stability. On the other hand, the reduction in the minimum number of samples required for splitting allows the trees to fit the data more precisely. These changes improve the generalization ability of the Random Forest model and make it perform better when facing different datasets. Finally, for the SVM model, the increase of the penalty parameter C value enhances its generalization ability, and the difference on different datasets is reduced from 13.39% to 4.46%. This is because the Increasing value of C causes the classifier to become more inclined to correctly classify all training samples, which enhances the model's performance across different datasets.

In summary, improving the generalization capability of a model requires consideration from multiple aspects, such as the inherent characteristics of the model self and the potential for overfitting. It is essential to adjust the model's parameter structure based on specific circumstances to achieve better generalization, thereby enhancing its robustness and applicability. At present, there are still some shortcomings in this study, such as not exploring many models like KNN and not trying too large data sets. For these shortcomings, further research will aim to explore

more widely applicable methods to improve the generalization ability of the model in disease prediction.

4. Conclusion

In this research, the enhancement of generalization in machine learning models for heart disease was explored by the modification and comparison of the structure and parameters of the model. The prediction results of two different heart disease datasets were compared by modifying the structures of MLP, Random Forest, and SVM. Through exploration, it is found that different models have different characteristics. For example, when MLP is applied to a small data set, selecting a simpler model can achieve better generalization. In contrast, employing a complex model will lead to overfitting, resulting in large differences in the results on different datasets. The results on the target domain will be significantly worse. Currently, the study has some limitations, including the absence of exploration into other models and the lack of experimentation with larger datasets. Future research will focus on investigating a broader range of methods to enhance the model's generalization capability in disease prediction.

References

[1] World Heart Federation. World Heart Federation, 2024, https://world-heart-federation.org/about-whf/.

[2] Liao Y, Tang Z, Gao K, et al. Optimization of resources in intelligent electronic health systems based on internet of things to predict heart diseases via artificial neural network. Heliyon, 2024, 10(11): e32090.

[3] Cheng L, Nie Y, Wen H, et al. An ensemble machine learning model for predicting one-year mortality in elderly coronary heart disease patients with anemia. Journal of Big Data, 2024, 11(1): 99.

[4] Madrid J, Duijvenboden V S, Munroe B P, et al. PO-02-049 ECG-based unsupervised clustering in coronary artery disease detects and predicts heart failure. Heart Rhythm, 2024, 21(5S): S278-S278.

[5] Yue H, YingBo R, Hai Y, et al. Using a machine learning-based risk prediction model to analyze the coronary artery calcification score and predict coronary heart disease and risk assessment. Computers in Biology and Medicine, 2022, 151(PB): 106297.

[6] Varshney K, Paliwal M. Heart Disease Diagnosis by Neural Networks. Journal of Pharmaceutical Research International, 2021: 202-208.

[7] Mexwell. Heart Disease Dataset, 2023. https://www.kaggle.com/datasets/mexwell/heart-disease-dataset/data.

[8] Winson. Heart Disease Dataset, 2023. https://www.kaggle.com/datasets/winson13/heart-disease-dataset/data.

[9] Cybenko G. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems, 1989, 2(4): 303-314.

[10] Ho T K. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995: 278–282.

[11] Ho T K. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[12] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273-297.