Evaluation Methods for Breast Cancer Prediction in Machine Learning Field

Zirui Zhang[1, *], Zixuan Li[2]

[1]Wenzhou-Kean University, Wenzhou, Zhejiang Province, 325000, China
[2]University of Nottingham Ningbo China, Ningbo Province, 315000, China

ABSTRACT: Breast cancer is the most common malignant tumor found in women, and there is no cure for advanced breast cancer. Early detection and treatment can effectively improve patient survival. This paper uses five machine learning classification models, namely Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors Algorithm (KNN). The training data for the five models are provided by the Wisconsin Breast Cancer Dataset (WBCD). By evaluating and comparing the performance of the five models in accuracy, F1Score, ROC curve, and PR curve, the study finds that LR has the best performance.

1. INTRODUCTION

Breast cancer is the most common malignancy for women, and advanced breast cancer is incurable [1]. Like most cancers, breast cancer is divided into two types: the benign type and the malignant type. Machine learning can provide patients with efficient self-examination in breast cancer prediction. Early prevention and treatment can improve the relative survival rate by 5 years [2]. In other words, if patients can be diagnosed and treated early, their chances of survival can be significantly improved. Therefore, the accuracy of machine learning models in breast cancer prediction is essential.

This study uses five classification models to distinguish breast cancer. The models are Support Vector Machine (SVM) which is based on grid search, Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors Algorithm (KNN). The training and accuracy results of the five models come from the use of the python programming language and the Sklearn package. The dataset used here is from Wisconsin Breast Cancer Database, which is created by Dr. William H. Wolberg [3].

The literature review on the common parts of the five machine learning models and Grid Search (GS) will be presented in Section 2. In Section 3, the introduction of the dataset and the preprocessing of the data are described. Experimental methods and evaluation methods (ROC curve, confusing matrix, F1 Score, and P-R curve) are introduced in Section 4. In the last section, the author summarizes this article and offers an outlook. The purpose of this study is to find a more accurate machine learning model in breast cancer prediction, since the more accurate the model is, the more effective the classification will be in breast cancer prediction. The research results of this paper could improve the efficiency of breast cancer detection, and this has reference value in the practical application of breast cancer prediction.

2. LITERATURE REVIEW

Data mining can reveal potentially valuable information and hidden features from large amounts of data. In the medical field, data mining can predict medical costs, diagnose diseases, and discover potential pharmacological techniques from large amounts of data [4]. The five machine learning models (SVM, LR, DT, RF, KNN) used in this paper are supervised learning. Supervised learning labels the data, classifies the test data (unclassified) and continuously adjusts the model to classify as much data as possible into the same class as the label [5].

SVM is a machine learning model used in this paper. The original data maps to upper dimensional spaces and forms a kernel function that significantly affects the classification accuracy [6]. In this paper, the author uses GS to adjust the hyper-parameter and chooses the best presentation as the score of SVM.

The Cross-Validation method is used to test the models when they are being trained. Most of the given modeling samples are taken out to build the model, and a small part of the sample is left to test with the newly established model. The method (Cross-Validation) yields better average performance than using a single classification strategy while reducing the risk of a model training failure [7].

*Corresponding author. Email: zhangzir@kean.edu

3. DATASET INTRODUCTION AND DATA PREPROCESSING

3.1. The source and description of the Dataset

All data used for model training comes from Wisconsin Breast Cancer Database. The database recorded 699 instances from January 1989 to November 1991. Each instance has ten attributes except its unique sample ID. The first nine attributes are independent: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. They are represented by values from 1 to 10. The last one, "Class," is a dependent attribute; it is represented by 2 and 4 (2 for benign, 4 for malignant).

3.2. Data preprocessing

Some instances in Wisconsin Breast Cancer Database have missing data for the "Bare Nuclei" attribute, which is set to "?". The instances with missing values were deleted during the processing to prevent the experimental results from being affected. After processing, 683 instances remain. In binary classification, generally, 0 and 1 are used to represent two types of cases, so 2 (benign) and 4 (malignant) are converted into 0 and 1.

4. EXPERIMENTAL METHOD AND EVALUATION METHOD

To control the variables, the five models applied in this paper all use 70% of the data as the training set and the remaining 30% of the data as the test set.

4.1. Support Vector Machine (SVM based on GS)

The SVM model is used to find the optimal hyperplane that separates classes. Its idea is to use the data to map out a hyperplane that makes the separated two classes have the greatest distance from it [8]. SVM has four kernel functions: RBF, linear, sigmoid, and polynomial. There are two parameters for the RBF kernel function, C and gamma. C is the penalty coefficient, which is the tolerance for error. The higher the C is, the less tolerance for errors, and the easier it is to overfit. The smaller the C is, the easier it is to underfit. If C is too large or too small, the generalization ability becomes poor. The relationship between sigma and gamma would be like as follows:

$$k(x,z) = \exp\left(-\frac{d(x,z)^2}{2*\sigma^2}\right) = \exp\left(-\text{gamma} * d(x,z)^2\right.$$

$$\text{gamma} = \frac{1}{2*\sigma^2}$$

Gamma is the width of the RBF, which affects the Gaussian range of each support-vector. If the gamma is too large, the test set accuracy will be lower; if the gamma is too small, the train set accuracy will be lower. To find the best parameter set, grid search is used. In the grid search test results of SVM, the combination of RBF kernel function, penalty coefficient C=100, and gamma=0.0001 have the highest accuracy. The evaluation of SVM in the following sections is based on the training results of this combination.

4.2. Logistic Regression (LR)

The prediction of breast cancer is a binary classification problem, and regularities are searched from a large amount of data using linear fitting. Then the probability of the two categories is obtained through the sigmoid function, and the data is merged into a class with a higher chance.

4.3. Decision Tree (DT)

The decision tree (DT) has a tree-like structure (a binary tree in this paper). All connections between nodes represent the output of a feature attribute in a given range, and each leaf node stores a classification result [9]. The decision tree starts from the root node, tests the corresponding feature attributes of the items to be classified, and selects the output branch according to its value until it reaches the leaf node; the leaf shows the classification result.

4.4. Random Forest (RF)

Random forest is built based on the decision tree by changing the sampling of original training samples and the selection of feature nodes to obtain trees with different classes.

4.5. K-Nearest Neighbors (KNN)

KNN is an evolution based on the Nearest Neighbor rule (NN) [10]. KNN does not just find the closest point to the query and marks its classification to the query. In KNN, the query will be classified by a majority vote of its k-nearest neighbors.

Experiments have been done for two different values of neighbor. When K=8, the model gets the best accuracy. The next part's evaluation will be based on the trained result of K=8.

Table 1. The accuracy of KNN with different number of neighbors.

Neighbor	6	8
Accuracy	.94761	.95238

4.6. Accuracy & ROC curve & Confusion matrix with F1-Score

In supervised learning, generally, the predicted results are compared with the actual results and get the $\text{accuracy} = \frac{\text{number of successful prediction}}{\text{number of total test}}$.

However, for datasets with unclear data distribution, in order to prevent skewed data from affecting the accuracy, the confusion matrix and ROC Curve are used to represent the accuracy of the prediction results.

For confusion matrix, the accuracy is calculated by:

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{FN} + \text{FP} + \text{TN} + \text{TP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

$$\text{F1Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Harmonic mean can reduce the impact of large outliers in the data. In the ROC curve, the farther the curve is from the pure chance line (diagonal), the stronger the identification ability of the model will be. From the ROC curves of the five machine models used in this paper, it is difficult to identify which model is more efficient, so the Area Under Curve (AUC) is used to compare the accuracy of the models.

4.7. P-R Curve

In the P-R curve, Recall is the x-axis and Precision is the y-axis. If the P-R curve of one machine learning is completely wrapped by the curve of another machine learning, then the performance of the latter is better than that of the former. When the P-R curve of the model is too close to distinguish, the model can be evaluated by the break-even point (precision = recall). The larger the value of the balance point of the curve, the better the performance of the model.

5. RESULT

In this section, the author compares and evaluates the outputs of five machine learning models (LR, SVM, DT, RF, KNN). First, the accuracy of the model and the harmonic mean of recall and precision are compared. Due to the small proportion of malignant or confirmed cases in disease datasets, the accuracy of model training results is greatly affected by samples. Accuracy is not the first criterion for judging model quality. The harmonic mean of recall and precision better reflects the training results. Based on the experimental results and theory shown in Figure 1 and Table 2, the LR and KNN models have better performance.

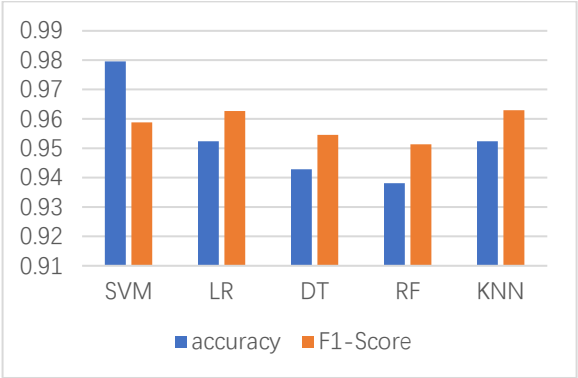

Figure 1 Accuracy and F1-score of five model.

Table 2. Accuracy and F1-score of five model.

	SVM	LR	DT	RF	K-NN
accuracy	0.97955	0.952381	0.942857	0.938095	0.952381
F1-Score	0.958801	0.962687	0.954545	0.951311	0.962963

Table 3. AUC of five model.

	SVM	LR	DT	RF	K-NN
AUC	0.99042	0.987654	0.965481	0.970815	0.976296

When the data is biased, the ROC curve and its area under the curve (AUC) can more effectively reflect the training results. Since the ROC curves generated by the five models are too close to be distinguished, the paper uses AUC to compare and evaluate (seen in Figure 2 and Table 3).

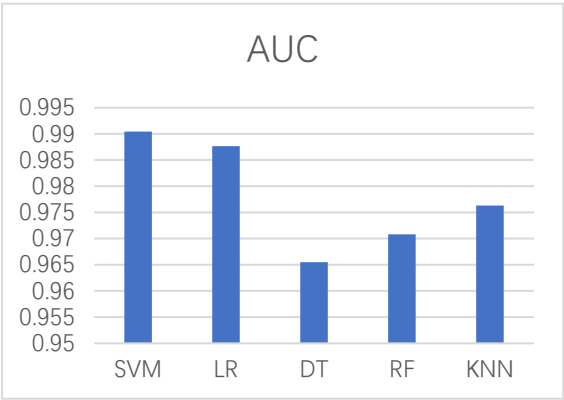

Figure 2 AUC of five model.

The closer the PR curve is to the upper right corner, the better the model training results are. As shown in Figure 3, KNN, LR, and SVM have better performance, while LR has a higher value in the intersection of the curves of these three models and the Break-Even point.

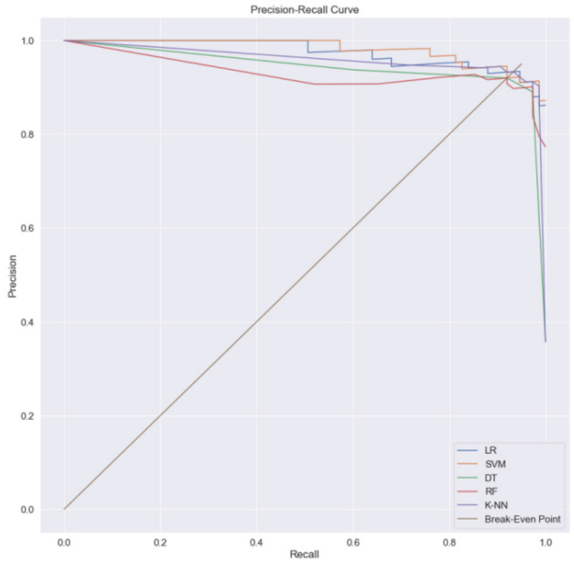

Figure 3 P-R Curve of the five models.

Since LR achieved excellent performance in AUC, F1Score, and P-R curves, a conclusion can be drawn that, based on the Wisconsin breast cancer dataset, when the training and test sets are divided by 70% and 30%, the logistic regression model performs the best among the five machine learning models.

6. CONCLUSION

This paper uses five machine learning models, LR, SVM, DT, RF, and KNN to achieve the breast cancer prediction. The data for training the model comes from the Wisconsin Breast Cancer Dataset, which contains 683 valid samples, each with 9 attributes and a benign-malignant classification label. By comparing and evaluating the performance of the five models in terms of accuracy, ROC curve (AUC), F1Score and P-R curve, it is found that LR has the best performance in classification based on the experimental data and training set segmentation. However, there are limitations in terms of the sample and data used in this paper. Each sample has only 9 attributes, each attribute is divided into only 10 levels, and the data set of this experiment records breast cancer cases in the 20th century. In the future, the medical field can detect and record more physical characteristics of breast cancer patients, and formulate a more standardized classification of each attribute. In addition to updating and refining the data, it is possible to combine LR with other techniques or optimize the LR algorithm and produce a more efficient predictive model.

REFERENCES

1. N. Harbeck, F. Penault-Llorca, J. Cortes et al., Breast cancer, Nature reviews Disease primers, 5(1), 2019, pp. 1-31.
2. Y.S. Sun, Z. Zhao, Z.N. Yang et al., Risk factors and preventions of breast cancer, International journal of biological sciences, 13(11), 2017, 1387.
3. Breast Cancer Wisconsin (Diagnostic) Data Set, (1995), Retrieved April 27, 2022, from: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29.
4. I. Yoo, P. Alafaireet, M. Marinov et al., Data mining in healthcare and biomedicine: a survey of the literature, Journal of medical systems, 36(4), 2012, pp. 2431-2448.
5. P. Cunningham, M. Cord, S.J. Delany, Supervised learning, In Machine learning techniques for multimedia, Springer, 2008, pp. 21-49.
6. C. Schaffer, Selecting a classification method by cross-validation, Machine Learning, 13(1), 1993, pp. 135-143.
7. V. Vapnik, Statistical Learning Theory, New York, NY, USA: Wiley, 1998.
8. A. Mert, N. Kilic, A. Akan, Breast cancer classification by using support vector machines with reduced dimension, ELMAR, IEEE, 2011, pp. 37-40.
9. H. Sharma, S. Kumar, A survey on decision tree algorithms of classification in data mining, International Journal of Science and Research (IJSR), 5(4), 2016, pp. 2094-2097.
10. T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 13 (1), 1967, pp. 21-27.