International Journal of Biology and Life Sciences
ISSN: 2957-9511 | Vol. 3, No. 2, 2023

Breast Cancer Prediction based on Support Vector Machine

Tianci Liu [1], Xiangjun Li [2, *]

[1] Computer Science College of Yangtze University, Jingzhou, Hubei, 434023, China
[2] Jingzhou Vocational College of Technology Institute of Software Engineering, Jingzhou, Hubei, 434023, China
*** Corresponding author:** Xiangjun Li (Email: 1007472895@qq.com)

Abstract: Breast cancer is the most common malignant tumor among women in the world, and its mortality rate ranks second. In this paper, the breath cancer Wisconsin (diagnostic) data set is used for prediction. First, exploratory data analysis was carried out, and 10 indicators related to breast cancer were selected as independent variables. Then, support vector machine (SVM) and logistic regression model were used as classifiers, and 75% of data sets were divided into training sets to build models. Finally, 25% of test sets were used as input models for prediction, and accuracy and recall were used as evaluation indicators to compare and analyze the advantages and disadvantages of the two algorithms. The results show that the accuracy of SVM is 16.2% higher than that of logistic regression, and the recall rate is 9.2% lower. The prediction application of SVM algorithm shows that SVM algorithm has low computational complexity and strong generalization ability. Therefore, support vector machine can be used to improve the diagnostic accuracy in the diagnosis of breast tumors.

Keywords: Breast Cancer; Data Analysis; SVM; Logistic Regression.

1. Introduction

Breast cancer is the most common malignant tumor among women in the world, and its mortality rate ranks second. When cells in breast tissue divide and grow, cell death and cell division are not normally controlled. Although scientists don't know the exact cause of most breast cancers, they do know some risk factors that increase the possibility of breast cancer in women. These factors include age, genetic risk and family history. The treatment of breast cancer can be divided into local treatment and systemic treatment. Surgery and radiotherapy are examples of local treatment, while chemotherapy and hormone therapy are examples of systematic treatment. Usually, for the best effect, these two treatments are used together. It is the serious influence of breast cancer that this paper studies the prediction of breast cancer based on support vector machine.

In the field of breast cancer prediction, with the emergence of various public data sets, the research on breast cancer prediction by domestic and foreign scholars has gradually increased. Hu Xuemei and others [1] used LASSO penalty logistic regression method to improve the diagnostic accuracy of breast tumors. Wangyi[2] and others used convolutional neural network to study molybdenum target examination images, and then made benign and malignant identification of breast cancer. Wang Jiaqing[3] and others based on principal component analysis and random forest algorithm fitting to classify and predict breast cancer model.

Based on related research, this paper carries out breast cancer prediction research based on support vector machine. The main contributions of this paper are as follows: (1) Exploratory data analysis of breast cancer data, and visualization method is used to show the correlation between feature variables. (2) Using logistic regression and support vector machine to classify and predict breast cancer data respectively, and calculate its accuracy and recall rate.

2. Related Technologies

2.1. SVM Algorithm Principle

support vector machines (SVM) [4]is a method to improve the generalization ability of learning machines according to the principle of structural risk minimization proposed by VAPNIK. The essence is to find the global optimal solution of convex quadratic optimization problem, which was originally proposed for linear binary classification problem. The basic idea is to find the most classified surface of two kinds of samples, and the advantage is that not only can the two kinds of samples be separated correctly, but also the classification interval can be kept maximum. Breast cancer prediction is a linear dichotomy problem.

2.2. SVR

SVR support vector regression [5] is a "tolerant regression model", and the model function of SVR is a linear function.

SVR supports vector regression. The algorithm is realized by creating an "interval band" on both sides of the linear function. Whether the sample is included in the loss function is related to the relationship between the sample and the interval band. If the sample falls in the middle of the interval band, it will not be included in the loss function. If the sample falls outside the interval band, it will be included in the loss function. The optimization model problem is solved by minimizing the width and total loss of the spacer. Only when it falls outside the edge of the separation zone or on the edge of the isolation zone will it be included in the final loss.

3. Exploratory Data Analysis

3.1. Experimental Data

The breast cancer Wisconsin (diagnostic) data set is used in this paper. There are 569 samples and 32 columns in the breast cancer data set. According to the diagnosis results, among 569 patients with breast cancer, 357 cases were benign and 212

cases were malignant, accounting for 62.7% of benign cases and 37.3% of malignant cases.

Table 1. Number of data sets

Type	Number
Benign	357
Malignant	212

3.1.1. Data Preprocessing

Diagnosing whether breast tumor is benign or malignant is essentially a binary problem. In this paper, the Wisconsin breast cancer (diagnosis) data set is used, in which the ID feature has no practical meaning and is deleted in the data preprocessing process. The values of feature diagnosis are m and b, where m is malignant, b is benign, 0 is benign, and 1 is malignant. The middle 3~32 columns are three kinds of values of 10 features: the first 10 columns in the field _mean represents the average value of 10 features, the middle 10 columns _se corresponds to the standard deviation of each feature in 10 features, and the last 10 columns _worst corresponds to the maximum value of each feature in 10 features.

3.1.2. Characteristic Correlation Analysis

Visualize the data set first, and then visualize it according to the value of feature diagnosis to observe the data distribution of benign and malignant data sets as a whole. as shown in figure.1.

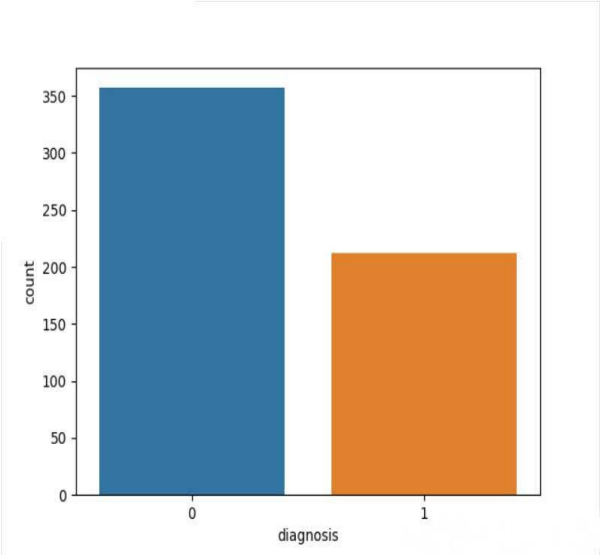

Figure 1. benign and malignant tumor

There are 10 features used in the experiment. In data processing, the features are divided into three groups according to _mean for average, _se for standard deviation and _worst for maximum. In the experiment, the correlation between the 10 features and feature diagnosis is shown by thermal diagram.

The correlation coefficient of a single variable on the diagonal in the heat map is 1, and the lighter the color, the greater the correlation. In the above figure, the characteristics of radius_mean, perimeter_mean and area_mean are highly correlated, and so are compactness_mean, concave_ points_ mean. In the heat map, it can be concluded that the three groups of features _mean, _se and _worst are different measures of the same group of content, and _mean is reserved in the feature selection.

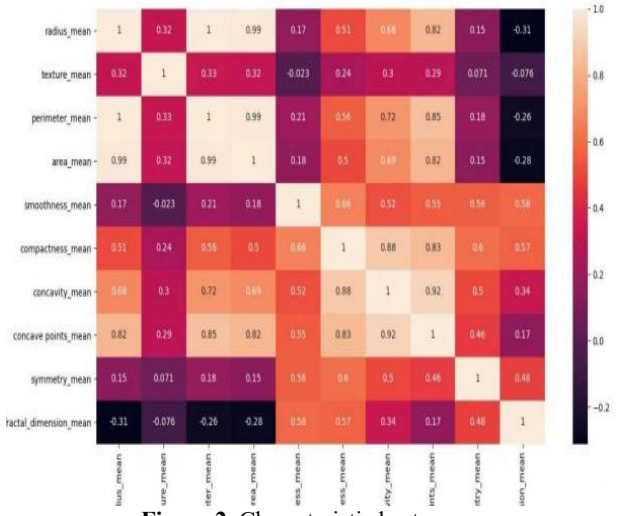

Figure 2. Characteristic heat map

4. Result Analysis

Two groups of comparative experimental models are designed, with accuracy as the main evaluation index and recall rate as the auxiliary evaluation indexes.

Table 2. Model Performance Test

Model category	Accuracy (%)	Recall rate (%)
LR	78.2	61.3
SVM	94.4	52.1

5. Summary

In this paper, the breast cancer Wisconsin (diagnostic) data set, is used to learn LR and SVM classification models, and the average value of indicators is used as the prediction variable to establish two classifiers: logistic regression and support vector machine, which show different prediction performance and prediction accuracy, among which support vector machine has better realization, higher prediction accuracy, poor performance of logistic regression and lower prediction accuracy.

References

[1] Hu Xuemei, Xie Ying, Jiang Huifeng. Breast cancer prediction based on penalty logistic regression [J]. Data collection and processing, 2021,36(06):1237-1249.

[2] Wangyi, Fang Wenjing, Zhang Zhongzheng. Study on breast cancer risk prediction based on convolutional neural network [J]. Science and Technology Innovation and Application, 2022,12(07):61-63.

[3] Wang Jiaqing, Wei Zhe, Zhang Taipeng, et al. Study on breast cancer prediction model based on random forest algorithm [J]. China Medical Equipment, 2022,19(01):119-123.

[4] Huang Jiang, Li Yuhan, Wu Shengbin, et al. Study on driving style recognition based on multivariate characteristic parameters and improved SVM algorithm [J]. Journal of Chongqing University of Technology (Natural Science), 2022, 36 (11): 8-19.

[5] Xin Guo, Hao Qian, Yang Feng. Apple multi-target detection method based on improved HOG and SVM [J]. Foreign electronic measurement technology, 2022,41(11):154-159.