ISSN 2959-6157

Diagnose Breast Cancer Using Two Machine Learning Methods and Comparing and Analyzing Them

Chengjun Hou

Department of Computer Science and Technology, TongJi University, Shanghai 201800, China
Corresponding author: 2152828@tongji.edu.cn

Abstract:

The breast cancer is currently the number one cancer in the world and has a high early treatment rate. Therefore, early screening and diagnosis of breast cancer is very important. Machine learning is very strong in analyzing and processing massive data. In this paper, we train and optimize decision tree and SVM models to identify malignant breast cancer and compare and analyze the performance of the two models. The accuracy of the earliest optimized models reaches about 98%, and it is found that the two models focus on different classification effects for the two samples, which can be further optimized by stacking and other ways. The dataset used to train the models in this paper is from the Wisconsin breast cancer database (WBCD), a benchmark dataset commonly used to compare different algorithms, which is of some significance for exploring the use of machine learning in the field of medical diagnosis.

Keywords: Breast Cancer; Machine Learning; Decision Tree; SVM.

1. Introduction

The development of human technology has overcome many diseases, but there are still a few diseases that cannot be completely cured at present, including cancer. About 4,824,700 new cancer cases and 2,574,200 new cancer deaths occurred in China in 2022[1], which could be related to the aging population, lifestyle changes and environmental factors such as air pollution. Among them, breast cancer is the most common malignant tumor worldwide with the highest mortality rate among female patients [2]. With the development of modern medicine, the five-year survival rate of early breast cancer can reach 95.1% [3]. Therefore, early diagnosis and screening of breast cancer is essential for the development of treatment plans and improving the survival rate of patients.

Currently, breast cancer detection methods include imaging tests (ultrasound, MRI, etc.), biomarker tests (e.g., gene mutation, protein variation, etc.) and body fluid tests (e.g., blood samples to detect tumor markers or tumor DNA, etc.). However, these methods still have some limitations, such as MRI: In the systematic review [4], MRI sensitivity is 75%, which means that only 25% of true breast cancers are missed, and the specificity of MRI is about 96.1%, which means that 3.9% of patients with non-breast cancers will be misdiagnosed as breast cancers. As another example, ultrasound testing is operated and interpreted by a sonographer, so the reliability of the test relies heavily on the operator. With the development of modern computers, machine learning techniques are maturing and showing advantages in various fields. Examples include processing valid information and patterns from large-scale data, and predicting future events based on historical data and patterns. In the medical field, machine learning can be used to construct models that can be predicted to support decision-making through the aggregation, integration, and analysis of massive amounts of data, which can help doctors diagnose breast cancer earlier and more accurately. Alireza Osareh and Bita Shadgar used algorithms such as the K-nearest neighbors and SVM in breast cancer diagnosis, in which SVM was used in two widely used breast cancer benchmark datasets achieving 98.80% and 96.33% accuracy, respectively [5,6].

In this paper, two machine learning algorithms, decision tree, and svm, are used to construct a breast cancer diagnosis model, to compare and explore the value of different models in breast cancer diagnosis as well as their respective advantages and disadvantages, and to try to optimize the model in terms of algorithms and parameter tuning, to improve the robustness and accuracy of the model.

2. Data and Methods

2.1 Dataset

The Breast Cancer Wisconsin (Diagnostic) Data Set(Breast Cancer Wisconsin (Diagnostic) Data Set (kag-

gle.com)) from the kaggle platform was used for the dataset. The dataset is from the University of Wisconsin. The dataset has a total of 569 data items, each of which contains the id, diagnosis (benign/malignant), and the features of the real data. These features were calculated from the images of the breast glands detected by FNA (fine needle aspirate). relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The dataset was examined and found to be very clean, with no empty rows and no NaN values to be processed much. Checking the column names, it is found that there is an unneeded data column 'Unnamed: 32', removing this column gives the dataset which is ready for training. The data in the 'diagnosis' column is taken as y and the data after removal of the 'diagnosis' column is taken as x.

2.2 Model Selection

2.1.1 Decision Tree

A decision tree is a commonly used machine learning algorithm for classification and regression tasks.

A decision tree is a tree structure in which each internal node represents an attribute, each branch represents a classification result of node attribute, and each leaf node represents a predicted value. Data can be categorized or regressed by testing along the path of the tree from the root node to the leaf nodes. In this experiment, it is used to accomplish achieve classification purposes.

Constructing a decision tree is a recursive process, which starts from the root node, selects the best splitting attribute to divide the dataset into subsets. Then it repeats the process on each subset until the stopping conditions are met, such as reaching the maximum depth or the node contains fewer than a threshold number of samples. A splitting criterion is generally used in each selection of splitting attributes, such as information gain (ID3 algorithm): after each new attribute node is determined, the change in information entropy of the data as a whole is calculated, and the more the information entropy decreases, the better the attribute can categorize the categories to a greater extent.

Finally, to prevent overfitting, the tree may be pruned to remove some unnecessary nodes to improve the model's generalization ability.

Each node of the decision tree has traceable attributes, so it is highly interpretable and intuitive to understand. And the decision tree is insensitive to missing values and can automatically handle missing values. However, the decision tree is easy to overfit when the depth is large or the number of training samples is small, and at the same time, small data changes may lead to generating different decision trees, which is not stable enough.

2.1.2 Support Vector Machine(SVM)

The core idea of SVM is to find an optimal hyperplane that can separate the samples of different categories in the data set and maximize the interval between two categories. In the binary classification case, SVM tries to find an optimal decision boundary such that the samples of two categories are furthest away from this boundary. If the dataset is not linearly separable, the SVM can make the data linearly separable in a high-dimensional space by mapping the data into a high-dimensional space via the kernel trick.

SVM can solve nonlinear problems by mapping the data to higher dimensions through kernel functions. Meanwhile, because the principle of SVM is to find an optimal hyperplane that can maximize the interval between two classes, it is robust to noisy data. However, the performance of SVM is greatly affected by parameters and kernel functions, so good parameter tuning is required. And the computational complexity of SVM is high, so it may lead to longer training time when dealing with large-scale datasets.

2.3 Evaluation Metrics

After building the model, its performance needs to be evaluated to determine whether it is accurate and reliable enough. At the same time, the strengths and weaknesses of different models can be analyzed and compared to determine which type of model performs better on a given task. In this paper, the following metrics will be used to evaluate the models:

(1) precision: precision = number of correctly predicted samples/number of correctly predicted samples. The higher the precision, the more samples the model predicts as positive examples are actually positive examples.

(2) recall: recall rate = number of correctly predicted samples/number of predicted samples. The higher the recall rate, the more true positive cases the model has identified.

(3) f1_score: 2 * precision * recall / (precision + recall). f1-score, is the reconciled average of precision and recall. The higher the f1-score, the better balance between precision and recall the model has achieved.

(4) macro avg: unweighted average of the evaluation metrics for each category

(5) weighted avg: The weighted average for each category.

(6) confusion matrix: The confusion matrix is a scenario analysis table that summarizes the predictions of the classification model, aggregating the records in the dataset in matrix form according to two criteria: the true categories and the category judgments predicted by the classification model. Thus, from the confusion matrix, it is possible to visualize which category of samples the model performs

well and which category of samples it does not perform well.

3. Results

Table 1. Evaluation of Decision Tree

	Precision	Recall	f1_score	Support
B	0.96	0.90	0.93	121
M	0.84	0.93	0.88	67
Accuracy			0.91	188
Macro avg	0.90	0.91	0.90	188
Weighted avg	0.91	0.91	0.91	188

The results obtained using the decision tree are shown in Table 1. The metrics in the first row are precision, recall, f1_score, and support, with support indicating the amount of data. In the first column, the first two letters B and M represent that the sample is actually BENIGN and MALIGNANT, respectively, and the last three indicators are thus ACCURACY, MACRO AVG and WEIGHTED AVG. The table is divided into two parts. The first two rows under the heading of the first part, record the indicators precision, recall, f1_score. the last three rows of the second part, record accuracy, macro avg and weighted avg.

In Decision Tree, there is 0.96 precision in the benign class sample and only 0.84 precision in the malignant class sample, which indicates that the model is not good at judging the malignant class correctly. The recalls for the two classes were 0.90 and 0.93, respectively, identifying most of the correct examples, indicating that the model is pretty good at judging breast cancer overall. Finally, the accuracy of the Decision Tree model is 0.91 and the exact accuracy is 0.9096.

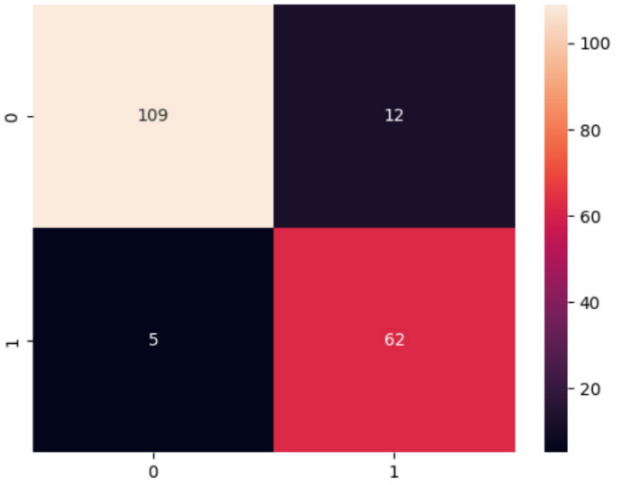

Fig. 1 confusion matrix of Decision Tree

Fig.1 shows the confusion matrix of Decision Tree. From the confusion matrix, the accuracy of Decision Tree is about 90%, and there are 12 samples with a true value of 1 but a prediction of 0; there are 5 samples with a true value of 0 but a prediction of 1, which indicates that the model is easy to miss the judgment, and it is not strong enough for identifying malignant tumors.

When the SVM model is trained, the first training time is very long, and the training is not finished after about 30 minutes, which may be a problem with the input data. The calculation process of SVM calculates the distance from the sample points to the decision boundary and finally minimizes the decision vector w. And the computational complexity of the decision boundary is greatly affected by the input data. The experimental dataset is real data computed from images, and the data range of each FEATURE is uncertain and varies greatly, leading to a long training time. So, the data needs to be normalized before the SVM is trained.

Table 2. Evaluation of SVM

	Precision	Recall	f1_score	Support
B	0.98	0.97	0.97	121
M	0.94	0.97	0.96	67
Accuracy			0.97	188
Macro avg	0.96	0.97	0.97	188
Weighted avg	0.97	0.97	0.97	188

The results obtained from the SVM model are shown in Table 2. In the SVM model, the accuracy has 0.98 in the BENIGN class sample and only 0.94 in the MALIGNANT class sample, the model performs well in both samples and has a slight defect in the MALIGNANT class. The recall for both classes reached 0.97, identifying most of the correct examples, indicating that the model judged breast cancer well overall. Finally, the accuracy of the SVM model is 0.97, and the exact accuracy is 0.9681.
From the confusion matrix (Fig. 2), the accuracy of SVM is about 97%, which is quite higher than the decision tree. There are only 4 samples with a true value of 1 but a prediction of 0. There are 2 samples with a true value of 0 but a prediction of 1, which indicates that the model still has minimal omissions and misjudgments, but the overall accuracy is already higher.

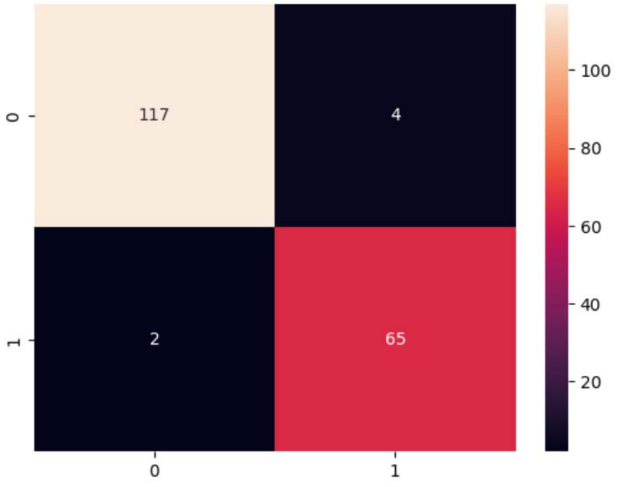

Fig. 2 confusion matrix of SVM

4. Discussion

The decision tree results are not ideal, which can be expected. For complex classification situations, having just one decision tree for up to 30 features is difficult in itself. So, we can optimize this by using the AdaBoost classifier algorithm. This algorithm uses multiple weak classifiers for co-training and finally synthesizes the results of all classifiers to arrive at the final classification result.
In this experiment, each weak classifier is chosen a decision tree with a maximum depth of 2. The statistics of the classification results obtained are shown in Table 3:

Table 3. Evaluation of AdaBoost

	Precision	Recall	f1_score	Support
B	0.98	0.98	0.98	121
M	0.97	0.97	0.97	67
Accuracy			0.98	188
Macro avg	0.98	0.98	0.98	188
Weighted avg	0.98	0.98	0.98	188

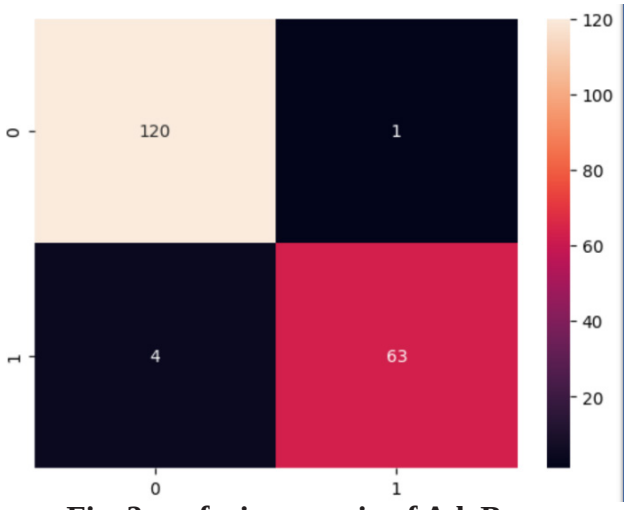

Fig. 3 confusion matrix of AdaBoost

After optimization, AdaBoost showed a significant improvement compared to the decision tree, reaching around 0.98 in both sample precision and recall, while accuracy came to 0.98, with an exact accuracy of 0.9734.

Compared with the decision tree, the accuracy is improved by about 7%. As can be seen in Fig.3, the number of missed judgments is significantly reduced by a lot, with only 1 case of missed judgment.

At the same time, we can also take certain optimization for SVM because the initial svm does not choose too many parameters, and the kernel function is also the simplest linear kernel function used. The grid search cv can be used for the SVM to tune the parameters, calculate the accuracy results for each combination of parameters, and select the set of parameters with the highest accuracy. In this experiment, the most commonly used RBF is used as the kernel function, and the parameters chosen for tuning are C, and gamma, C is the penalty coefficient, that is, the tolerance of error, the larger C, the more intolerable error. If C is too large, the model is easy to overfitting. If C is too small, the model is easy to underfitting. gamma is a self-contained parameter after choosing the RBF as the kernel, the larger the value of the gamma, the higher the dimensionality of the map, the higher the accuracy of the training. However, it's more likely to cause overfitting, i.e., low generalization ability. On the contrary, the smaller the gamma value is, the lower the dimension of the mapping is, and the underfitting may occur. The final range of parameters chosen is as follows: C ranges from [0.1, 1, 10, 100, 1000]; gamma ranges from [1, 0.1, 0.01, 0.001, 0.0001]; kernel as ' rbf ' .

The parameter combination with the highest accuracy obtained was C=100, gamma=0.001 The classification result statistics are shown in Table 4:

Table 4. Evaluation of SVM(fine tuning)

	Precision	Recall	f1_score	Support
B	0.98	0.98	0.98	121
M	0.97	0.97	0.97	67
Accuracy			0.98	188
Macro avg	0.98	0.98	0.98	188
Weighted avg	0.98	0.98	0.98	188

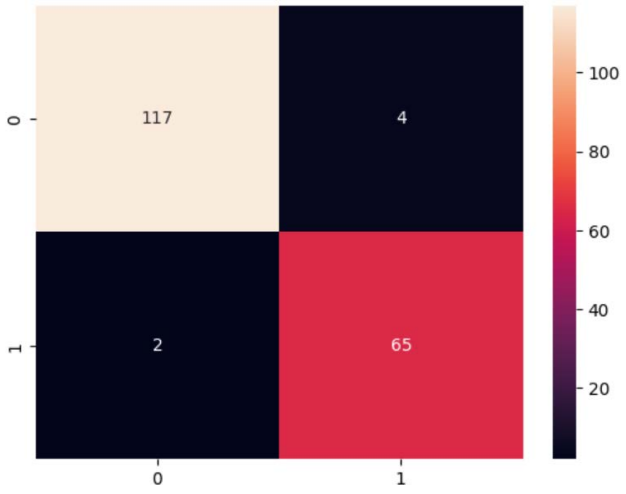

Fig. 4 confusion matrix of SVM(fine tuning)

In Table 4, the SVM has increased in all indicators after parameterization, and both precision and recall have reached about 0.98, and the final accuracy has reached 0.98, with a precise accuracy of 0.9787. It is a little higher than the AdaBoost classifier and is the most accurate model so far.

From Fig.4, we can see four cases of missed judgment and two cases of wrong judgment. For both AdaBoost and optimized SVM, they are more accurate, but they have more errors in opposite places. AdaBoost is more likely to misjudge, while SVM is more likely to miss. If you can combine the parts that they are each good at, you may get better results. Voting can be used to vote on the predictions of the two models, and the category with the majority of votes is chosen as the final prediction. Soft voting can be chosen, which is weighted voting with weights adjusted according to the model performance. Alternatively, stacking can be used, where the predictions of the two models are used as new features to be input into a meta-model for training. This metamodel can be a simple linear or more complex model such as a SVM or neural network.

5. Conclusion

In this paper, decision tree and SVM models are trained to analyze breast data to identify malignant breast cancer, compare the effectiveness of the two models, and optimize the two models using AdaBoost and parameter tuning, respectively, and both of them improve the accuracy to a certain extent. It is found that the accuracy of decision tree model before optimization 90% is significantly lower than SVM 97%, but after optimization both AdaBoost and SVM achieve about 98% accuracy. But simultaneously, the two models have different effects for different cases. adaBoost misses fewer cases and misjudges more, while SVM misses more cases and misjudges fewer. Integrated learning methods such as Stacking and Bagging can be further used to combine multiple models to improve the overall performance.

By training a large amount of medical imaging data and clinical data to train the machine learning model, and then assisting doctors in the diagnosis of breast cancer, the accuracy and efficiency of diagnosis can be improved, the misdiagnosis rate caused by human factors is reduced, and it helps to detect the lesion at an early stage and take the corresponding therapeutic measures. Meanwhile, according to the prediction model, breast cancer's pathogenesis and lesion characteristics can be inferred, providing certain reference and guidance for clinical practice.

References

[1] Bingfeng Han, Rongshou Zheng, Hongmei Zeng, Shaoming Wang, Kexin Sun, Ru Chen, Li Li, Wenqiang Wei, et al. Cancer incidence and mortality in China, 2022. Journal of the National Cancer Center, Vol 4, Issue 1, 2024; 47-53.

[2] Gabriel N. Hortobagyi, et al. The Global Breast Cancer Burden: Variations in Epidemiology and Survival. Clinical Breast Cancer, Vol. 6, No. 5, 2005; 391-401.

[3] Taylor C, McGale P, Probert J, et al. Breast cancer mortality in 500 000 women with early invasive breast cancer diagnosed in England, 1993-2015: population based observational cohort study. BMJ. 2023 Jun 13; 381:e074684.

[4] Warner E, et al. Systematic review: using magnetic resonance imaging to screen women at high risk for breast cancer. Ann Intern Med. 2008 May 6; 148(9):671-9.

[5] A. Osareh, B. Shadgar, Machine learning techniques to diagnose breast cancer, 2010 5th International Symposium on Health Informatics and Bioinformatics, Ankara, Turkey, 2010, pp. 114-120.

[6] Benson J, Jatoi I, Keisch M, Esteva F, Makris A, Jordan VC, Early breast cancer. Lancet. 2009; 373: 1463-79