International Core Journal of Engineering Volume 7 Issue 2, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

Using Different Machine Learning Classification Models on the Prediction of Heart Disease

Feiran Wang[1], Tianda Fu[2], Shuzhou Li[3], Yiyao Yu[4]

[1]School of Automation, Shanghai University, Shanghai, 200444, China;

[2]School of Mathematics and Computer, Wuhan Textile University, Wuhan, Hubei 430200, China;

[3]Sandy Spring Friends School, Sandy Spring, Maryland 20860, USA;

[4]School of International Information and Software, Dalian University of Technology, Dalian, Liaoning 116024, China.

Abstract

In August 2020, our group use 4 different machine learning models to predict the heart disease. We would like to find the best algorithm for predicting the heart disease. Through this project, we compare the 4 different machine learning models. We found that the svm model is the best model.

Keywords

Heart Diseases; Machine Learning; Accuracy Comparison; Efficiency Analysis.

1. Introduction

Under the shadow of this ever-changing time, heart disease has become the biggest threat to human health. In China, thousands of people die from heart disease every year. According to CDC's research, one person dies every 37 seconds in the United States from heart disease. At this high rate, predicting heart disease and its cause seems particularly important. Therefore, if we can use machine learning to analyze the correlation between heart disease and the relevant indicators of the human body. It will play a vital role in predicting and preventing heart disease.

In the work, we use Random Forest, Naïve Bayes, SVM (support vector machine), and Logistic Regression to predict heart disease. Each group member decides to take on one of the models.

2. Proposed Algorithm/Model

2.1 Dataset

The dataset was found from Kaggle. It has 14 features and 1025 data points. In the work,80 percent of the data was used for the training set and 20 percent of the data was used for the testing set.

2.2 Random Forest

2.2.1Decision tree

2.2.1.1Information

Information is the foundation of entropy and information gain. To quote Shannon, information is used to eliminate random uncertainty. Although this sentence is classic, it is still hard to understand what this kind of thing is. From the perspective of mathematics, it may be clearer. Mathematics is originally an abstract theory. It may be more appropriate to use abstractions to explain abstractions. It is also a definition used in an algorithm of machine learning called decision trees.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

If the set of samples with classification labels can be divided into several categories, the information of category (xi) is defined as:

2.2.1.2 Entropy

Now that the information has been spoken, entropy is not as abstract as it may be. It is more likely to be defined in probability theory. Entropy is the nomenclature suggested by John von Neumann and it is the expected value of information and can be written as:

$$H（x） = \sum_{i=1}^{n} p(x_i)I(x_i) = -\sum_{i=1}^{n} p(x_i)log_b p(x_i)$$

Conditional entropy: $H(Y|X) = \sum_x p（x）H（Y|X = x）$

2.2.1.3 Information Gain

Information gain is an index used to select features in the decision tree algorithm. The greater the information gain, the better the selectivity of this feature[1]. It is defined in probability as:

$$IG(Y|X) = H(Y) - H（Y|X）$$

2.2.2 Random forest

2.2.2.1 Concept

Random forest is a supervised learning algorithm. According to its name, it can create a forest with decision trees and give it randomness. Most of those trees are trained by the "bagging" method. In short: Random Forest builds lots of different decision trees and merges them together to obtain higher accuracy and much more stable predictions. One of the great advantages of random forest algorithm is that it can handle both classification problems and regression problems. These two types of problems just constitute what machine learning need to face mostly.

2.2.2.2 Features

1) It can handle very high-dimensional (a lot of features) data without feature selection;

2) It can give the importance of feature after training;

3) It has fast training speed and is easy to be parallelized;

4) It can still maintain accuracy if a large part of the features are missing;

2.2.2.3 Building step

1) If the size of the training data is N, randomly select N samples from the total training data as the training set of each tree.

2) If the number of feature of each sample is M, choose a constant k<<M, randomly select k subsets of feature from all features, then calculate the optimal feature from the k features and take it as the node each time the tree is split.

3) Each tree grows to the maximum depth without pruning process.

Computational Experiment:

Build a model and tuning hyperparameters by using GridSearchCV

2.3 Naïve Bayes

In the scikit-learn, there are three models of Naïve Bayes for people to use.

The applicable classification conditions of the models are different and it can be chosen by the given data set.

2.3.1 GaussianNB

If the distribution of the sample are mostly continuously values, GaussianNB should be used.

$$P(X_j|Y = C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\exp(-\frac{(X_j - \mu_k)^2}{2\sigma_k^2})$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

2.3.2 MultinomialNB

If most of the data in the data set are multivariate discrete values, MultinomialNB is a better choice.

$$P(X_j = x_{jl}|Y = C_k) = \frac{x_{jl} + \lambda}{m_k + n\lambda}$$

2.3.3 BernoulliNB

If the sample characteristics are sparse multivariate discrete values or binary discrete values, BernoilliNB should be used.

$$P(X_j = x_{jl}|Y = C_k) = P(j|Y = C_k)x_{jl} + (1 - P(j|Y = C_k))(1 - x_{jl})$$

2.4 SVM (support vector machine)

2.4.1 Algorithm principle

Support vector machines (SVM) can be used when the data has to be classified into two classes. The model classifies data by finding the optimal hyperplane which separates all data points of another class. The optimal hyperplane for SVM is the one that maximizes the margin between the two classes.

Margin is the maximal width of the slab without interior points that is parallel to the hyperplane. The support vectors are the data points closest to the separating hyperplane and the points are on the boundary of the slab.

The definitions are illustrated in the Figure 1

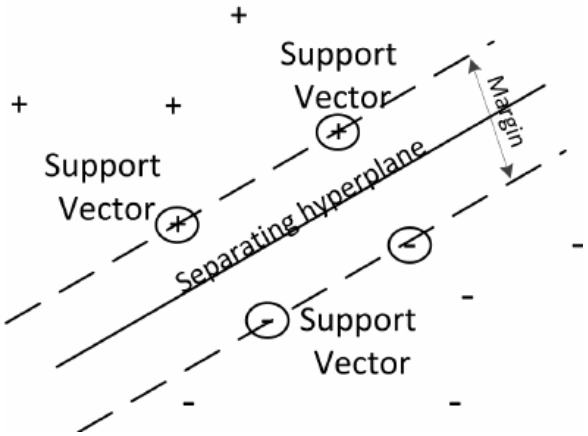

Figure 1 Represents data points of type –1, and + represents data points of type 1,[2]

When SVM is used to handle the complex and nonlinear optimization problems, a kernel function K is selected to map data to high-dimension space. When solving the linearly inseparable problem, SVM completes the calculation in low dimensional space. The SVM maps the input space from low dimensional space to high dimensional space by using a kernel function. Finally, constructing an optimal separating hyperplane in high dimensional space[3] to separate the data.

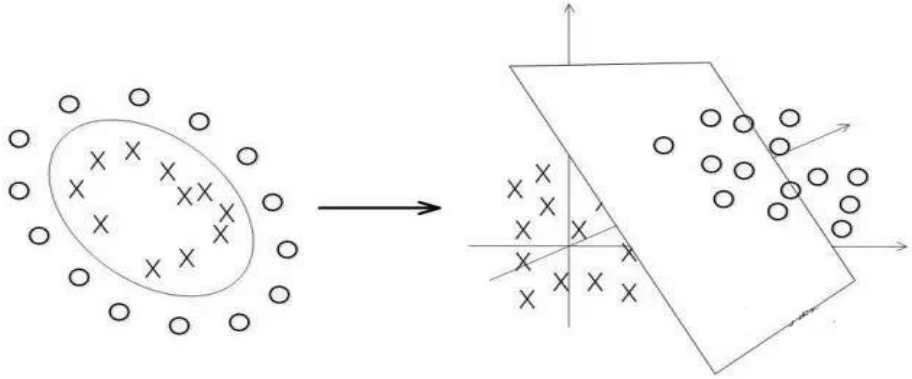

Figure 2 Represents the process of the SVM

ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

There is an example. Considering the x1 and x2 as following.

$$x1 = (\mu1 + \mu2)T, x2 = (\eta1 + \eta2)T$$

The dot product is the map in 5-dimension, therefore it is the inner product after mapping. And at the same time, the equation can be found in the low dimension.

$$(X_1 * X_2 + 1)^2 = 2\mu_1\eta_1 + 2\mu_2\eta_2 + \mu_1^2\eta_1^2 + \mu_2^2\eta_2^2 + 2\mu_1\eta_1\mu_2\eta_2 + 1$$

High dimensional equation

$$\emptyset(x_1) * \emptyset(x_2) = \mu_1\eta_1 + \mu_2\eta_2 + \mu_1^2\eta_1^2 + \mu_2^2\eta_2^2 + \mu_1\eta_1\mu_2\eta_2$$

low dimensional equation

It is easy to find that if we times a correlation coefficient, the results will be equal.

2.4.2 Adjust parameters

2.4.2.1 hyperparameters

There are different kinds of kernel functions. In this SVM model, two kinds of kernel functions are selected to solve the problem.

Linear kernel function: K(x,y)=x·y

Gaussian kernel function:

Generally is a kind of nonlinear function which can be shown as:

$$K(x_1, x_2) = < \emptyset(x_1), \emptyset(x_2) > \quad (1)$$

$$\left|\left|\emptyset(x_1) - \emptyset(x_j)\right|\right|^2 = < \emptyset(x_1) - \emptyset(x_2), \emptyset(x_1) - \emptyset(x_2) >$$

$$= < \emptyset(x_1), \emptyset(x_1) > -2 < \emptyset(x_1), \emptyset(x_2) > + < \emptyset(x_2) - \emptyset(x_2) >$$

$$= K(x_1, x_1) - 2K(x_1, x_2) + K(x_2, x_2) \quad (2)$$

$$K(x, z) = \exp\left(-\frac{\left|\left|x - z\right|\right|^2}{2\sigma^2}\right) \quad (3)$$

With equation (2) and (3), two cases can be separated:

$$\sigma \to 0, -\frac{\left|\left|x - z\right|\right|^2}{2\sigma^2} \to -\infty, K(x, z) \to 0$$

$$||\emptyset(x) - \emptyset(z)||^2 = k(x, x) - 2k(x, z) + k(z, z) = 2 - 2k(x, z) = 2$$

$$\sigma \to \infty, -\frac{\left|\left|x - z\right|\right|^2}{2\sigma^2} \to 0, K(x, z) \to 1$$

$$||\emptyset(x) - \emptyset(z)||^2 = k(x, x) - 2k(x, z) + k(z, z) = 2 - 2k(x, z) = 0$$

The smaller the parameter σ is, the finer the classification is, which means the more likely it is to lead to overfitting and vice versa. After adding the relaxation factor, the target function and constraint condition will be:

$$min\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi(i)$$

$$s, t, y^{(t)}\left(\omega^{(t)}\emptyset\left(z^{(t)}\right) + b\right) \geq 1 - \xi(i), i = 1,2 \ldots . n$$

Our goal is to minimize the objective function. When the C is very large, ξ(i) tends to 0. That is to say, on the border between the sample whose tolerance is low and the error points less. Therefore the function fits better, but the accuracy of prediction is not always good. When the value of C is small, the number of samples between the two boundaries becomes larger, and the possibility of misclassification increases. The fitting of samples decreases, but it may be more reasonable, because there may be noise between samples.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

2.5 Logistic regression

2.5.1 Usage

When the independent variable or the dependent variable is categorical, Logistic regression can be used. In our case, having heart disease which is "1" and not having heart disease which is "0" are the categorical variables. Logistic regression replicates the probability of the occurrence of an event depending on the values of the independent variable. It also predicts the probability of the occurrence of an event in a random observation, versus the probability that the event does not occur. Last but not least, it can classify observation by predicting the odds that an observation is in a selected category.

2.5.2 Theoretical background

In logistic regression, the dependent variable abide by the Bernoulli distribution and has an unknown probability, ***p.*** The model estimates an unknown ***p*** value for any given independent variables'' linear combination. So the model needs to connect together the independent variables to essentially form the Bernoulli distribution. The link is called the ***logit***.

Since ***p*** is unknown in logistic regression, the model is estimating ***p*** (the estimation of p is ***p-hat***) for a linear combination of the independent variables. Tying the links together, the function of this distribution is the natural log of the odds ratio.

In the logit function, 0 to 1 is along the x-axis. In order to have the values of the probabilities be represented on the y-axis, the model takes the inverse of the logit function.

3. Computational Experiments

3.1 Random Forest

```
# 4. todo train model
rf = RandomForestClassifier()
n_estimators = np.arange(10, 100, 10)
max_depth = np.arange(3, 30, 2)
criterion = ('gini', 'entropy')
grid = {"n_estimators": n_estimators, "max_depth": max_depth, "criterion": criterion}
gsc = GridSearchCV(rf, param_grid=grid, cv=5)
gsc.fit(X_train, y_train)
```

Figure 3 Building the model and tuning hyperparameters

3.2 Naïve Bayes

A log transformation are made for the feature age, trestbps, chol, and thalach. The transformation will let the result better. Because of the difference between the independent variable and the predicted target. The variable should be taken as 0 or 1 to make it naïve and linearly independent and then calculate their means to build a new model. If the result is very close this feature will be removed in my model. I removed the following features in my model.

```
'target','cp_0','restecg_0','slope_0','thal_0',"trestbps","chol","thalach"
```

Figure 4 The removed features

There is no parameter to adjust for the GaussianNB. For the MultinomialNB, there are three parameters to adjust.

Alpha: If you find that the fitting is not good, the parameter should be 1

ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

Bool Parameter fit_prior: It represents the prior probability that should be existed or not. If it is false, then consider that all categories have the same prior probability.

Class_prior: It is the prior probability.

The default parameter: alpha=1.0, fit_prior=True, class_prior=None

In my code, I use the following data to adjust.

for a in [0.1,1,10,20]:

for f in [True, False]:

For c in[None,[0.1,0.9],[0.2,0.8],[0.28,0.72],[0.29,0.71],[0.3,0.7],[0.31,0.69],[0.32,0.68],[0.33,0.67],[0.4,0.6]]:

For the BernoulliNB, there is another parameter to adjust which can help handle the Binomial distribution. The default binarize is 0. In my code, I use following data to adjust.

for a in [0.1,1,10,20]:

for f in [True,False]:

for c in [None,[0.1,0.9],[0.2,0.8],[0.28,0.72],[0.29,0.71],[0.3,0.7],[0.31,0.69],[0.32,0.68],[0.33,0.67],[0.4,0.6]]:

for b in [0,1,2,3,4,5,6,7,-1,-2]:

3.3 SVM (support vector machine)

3.3.1 Adjustment result

Table 1. The result of the adjustment

Delta C	4	5	6
0.1	98.54%	97.59%	92.68%
1	96.49%	100%	97.07%
10	98.32%	95.61%	96.59%

3.4 Logistic Regression

3.4.1Sigmoid function

First of all, I wrote out the main sigmoid function.

3.4.2 Error rate

For the logistic regression model, the error rate is extremely significant, because the weight and bias is the only unknown parameter.

3.4.3 LR Gradient Descent

The basic process of this method is as follows:

First of all, the model randomly selects to open an initial point. Secondly, the model chooses the direction of the gradient drop. Thirdly, the model selects the step length. Then, the model updates the point. Lastly, the model repeats (2), (3), (4) steps until the termination conditions are met.

3.4.4 Testing

First of all, I load the testing and training set. Then after loading the datasets and the weights, the model will predict the results.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

4. Results and Discussion

4.1 Results

4.1.1 Random Forest

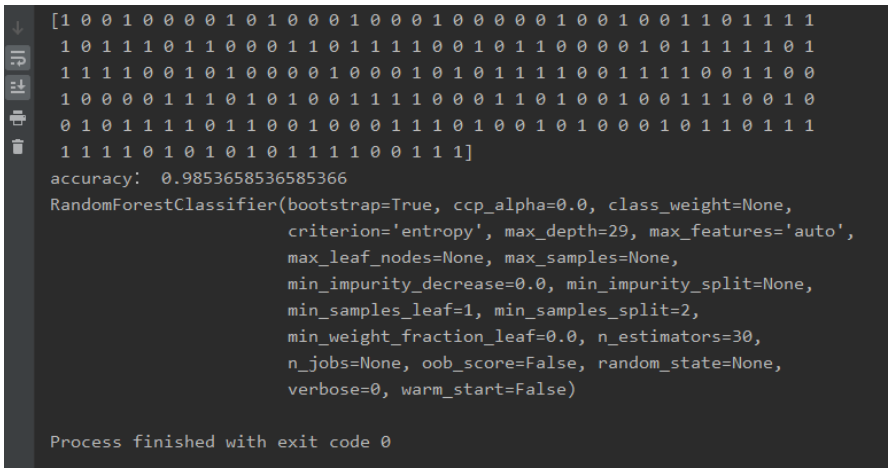

Figure 5. The result of the Random Forest

4.1.2 Naïve Bayes

GaussianNB

Accuracy: 0.8634146341463415

MultinomialNB

Accuracy: 0.8585365853658536

BernoulliNB

Accuracy: 0.8390243902439024

4.1.3 SVM (support vector machine)

4.1.3.1 Gaussian kernel

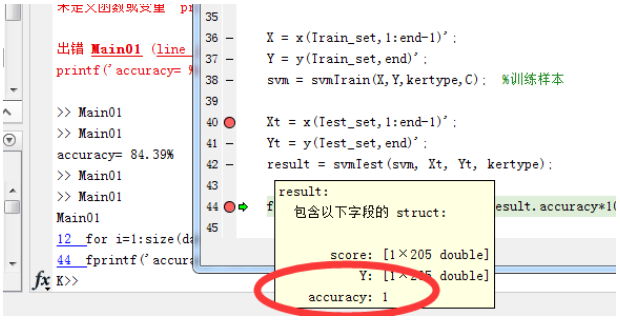

Figure 6. The result of Gaussian kernel

4.1.3.2 Linear kernel

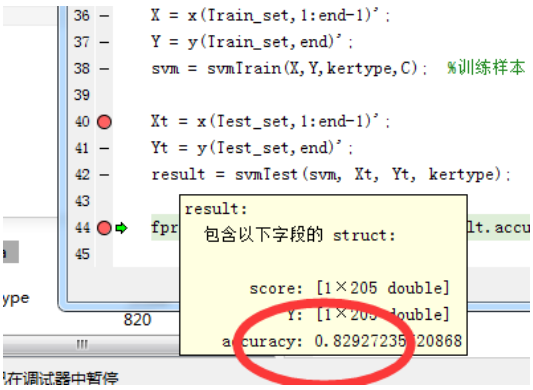

Figure 7. The result of Linear kernel

ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

4.1.3.3 The contribution rate of each feature

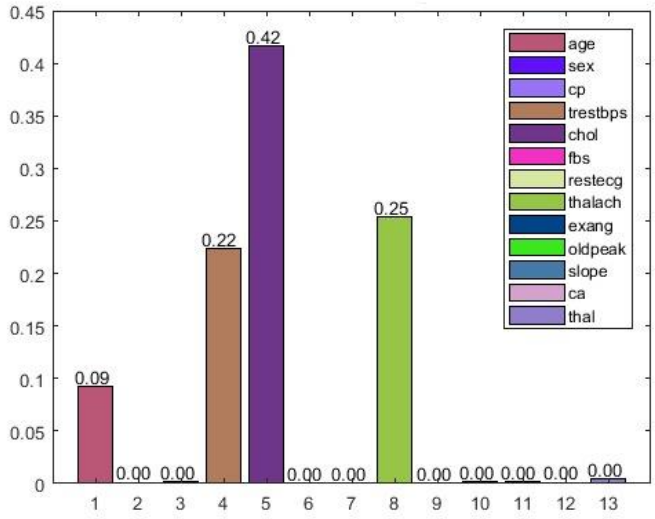

Figure 8. The contribution rate of each feature

4.1.4 Logistic regression

Table 2. The result of Logistic regression

	Precision	Recall	F1-score	Support
0	0.98	0.76	0.82	148
1	0.88	0.93	0.86	160
Accuracy			0.84	398
Macro avg	0.85	0.84	0.84	398
Weighted avg	0.85	0.84	0.84	398

4.2 Discussion

4.2.1 Naïve Bayes

The model can be used to make predictions and help the doctor prevent and treat heart disease. Most time the MultinomaialNb is better than BernoulliNb. Theoretically, comparing with other classification methods, the Naïve Bayes model has the minimum error rate and performs better. However, when there are many correlations between attributes, the classification result is worth than expected.

4.2.2 Support Vector Machine (SVM)

4.2.2.1 advantage of SVM

(1) The theoretical basis of SVM is nonlinear mapping. Instead of nonlinear mapping to high dimensional space, SVM uses inner product kernel function which is easier to understand and program.

(2) The idea of classification is simple, which is to maximize the interval between the samples and the decision surface.

(3) SVM can solve nonlinear classification problems with the kernel function.

(4) The effect of classification is good.

4.2.2.2 disadvantage of SVM

(1) It is difficult to implement for large-scale training samples with SVM.

(2) SVM is difficult to solve multiple classification problems.[4]

ISSN: 2414-1895 DOI: 10.6919/ICJE.202102_7(2).0062

(3) It is too sensitive to the selection of kernel functions, parameters and missing data

4.2.3 Logistic Regression

The Logistic Regression model is transparent and easy to use. Furthermore, it can give a probabilistic output which is useful in this project. However, the Logit can rarely hold when I did the experiment which is causing the accuracy stopping at 84%.

5. Conclusion

For this data set, the best model is SVM. The Gaussian kernel accuracy is 1. The worse is the Logistic Regression. We hope we can find more data set to test our models and use the models in a practical way. We believed that these models can be used to make predictions and help the doctor prevent and treat heart disease.

References

[1] Ying Tan, Yuhui Shi, Ben Niu. (2019) Advances in Swarm Intelligence: 10th International Conference, ICSI 2019, Chiang Mai, Thailand, July 26–30, 2019, Proceedings, Part II. Springer.

[2] M.A. Hadj-Youcef, M. Adnane, A. Bousbia-Salah. (2013) Detection of epileptics during seizure free periods. In: 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA). Algiers. pp. 212

[3] Satnam Singh, Wayne Blanding, Vishal Ravindra, Krishna Pattipati. (2006) Communication Channel Equalization-Pattern Recognition or Neural Networks? In: International Conference on Communication Technology.

[4] Xiaoyu Zhang, Rui Wang, Tao Zhang, Yajie Liu, Yabing Zha. (2018) Short-Term Load Forecasting Using a Novel Deep Learning Framework. Energies.