Comparison of Several Machine Learning Algorithms for Prediction

Yi Song*
Shanghai Southwest Weiyu Middle School, Shanghai, China
*Corresponding author e-mail: course@mfeducation.cn

Abstract. Currently, the analysis of bacterial drug resistance is a research hotspot in the biomedical field. However, due to the long culture experiment cycle for traditional multi-antibiotics, the tolerance degree of a specific bacterium to a specific antibiotic cannot be quickly analyzed and accurately determined. Therefore, how to analyze and predict drug resistance is still a problem to be solved. With the development of whole genome sequencing (WGS) technology, this paper proposed a variety of machine learning methods to analyze and predict the drug resistance of Staphylococcus aureus and Neisseria gonorrhoeae. First of all, the original data was preprocessed; secondly, the key features, such as information entropy and Gini index, were extracted; finally, the drug resistance of each drug was analyzed and predicted. Experimental results showed that the training of this model could achieve rapid convergence, and the recognition accuracy was as high as 97%. By comparing the experimental results of various machine learning methods, it can effectively predict bacterial drug resistance.

1. Introduction
Bacterial drug resistance refers to a fatal phenomenon in which bacteria develop drug resistance to specific drugs after large-scale use of antibiotics and lead to the decline or even disappearance of the effect. In the past 70 years, the use of antibiotics has spread rapidly and greatly reduced the risk of death caused by infectious diseases. In addition to humans, farm animals also use antibiotics in large quantities, which will cause a large number of bacteria to be resistant to drugs. For example, in the United States, at least 2 million people are infected with drug-resistant bacteria every year, and 23,000 of them die from infection.

Globally, the total number of deaths due to antimicrobial resistance (AMR) may exceed 700,000 every year. If the current trend continues, the annual death toll will rise to a devastating 10 million by 2050. Traditional antibiotics based on culture have a long experimental period, which may lead to the deterioration of patients' condition while waiting for the results, but doctors are helpless. Therefore, doctors and hospitals need a method that can quickly analyze and accurately determine the tolerance of a specific bacterium to a specific antibiotic.

For a long time, genome data has been used as a tool for medical research. With the development of whole genome sequencing (WGS) technology, it is easier to obtain genome data. If antibiotic resistance can be quickly correlated with genome-wide WGS data, it can not only save doctors precious time and prescribe correct antibiotics to help save patients' lives, but also prevent patients from infecting other bacteria.
With the rapid development of machine learning, the success of image analysis and chemical design has attracted the attention of biological researchers. Santerre et al. successfully used machine learning to predict drug resistance databases. In 2018, Andrew Zhang et al. proposed to use DCNN (Deep Convolution Neural Network) to predict the drug resistance of Rabbsiella pneumonia to tetracycline data and Acinetobacter baumannii to carbapenem data, and obtained good prediction results and accuracy.

Faced with a large amount of genetic data, machine learning shows a slight disadvantage in speed and accuracy. At the same time, deep learning is used to predict drug resistance picture data sets, but because the conversion of pictures is too complicated, a lot of time is needed for preliminary preparation.

For the problem of bacterial drug resistance analysis, this paper firstly analyzes the gene sequence data sets of Staphylococcus aureus and Neisseria gonorrhoeae, and then uses random forest, lightgbm, XGBoost and deep learning techniques in machine learning respectively for drug resistance analysis. Through experimental comparative analysis, the gene sequence resistance of Staphylococcus aureus and Neisseria gonorrhoeae was identified. Compared with other methods, the machine learning model proposed in this paper is more convenient and obviously improved.

2. Basic Theory

According to the analysis of bacterial drug resistance, this paper mainly adopts machine learning methods, namely random forest, XGBoost and CNN and DNN in deep learning. The specific theoretical basis is as follows.

2.1. Random forest
Random forest (RF), proposed by Breman W, combines the characteristics of Bagging algorithm and random subspace method, and is an algorithm model that can be used for classification and regression \[1\]. At present, this algorithm has been widely used in many fields such as biology, medicine, electricity, agriculture, finance and so on. And good results.

Random forest consists of multiple decision trees. If a new instance needs to be classified, the characteristics of the instance are presented to each decision tree in the forest, and each decision tree returns the classification value \[2\] and the vote for the class. Finally, after weighting the classification values given by each tree, the result that exceeds the classification values of decision trees in other forests is selected \[3\].

2.2. XGBoost
XGBoost belongs to boosting algorithm. Its main idea is to build multiple weak classifiers and gather them together to form a strong classifier \[7\].

The idea of the algorithm is to continuously add new trees, one tree at a time to fit the residual of the last prediction, that is, to retrain a new function, and then continuously generate new branches according to feature splitting. After the training is completed, K subtrees can be obtained. The purpose of this operation is to obtain a classification result, that is, according to the characteristics of the input data, each tree is classified several times to fall to a corresponding leaf node and a corresponding score \[8\]. Finally, only the score corresponding to each tree needs to be added up to be the predicted value \[9\] of the sample.

2.3. Deep Neural Network
Each neuron of Deep Neural Network (DNN) takes the output of all neurons in the previous layer as input, and its output will give each neuron in the next layer as input. Each neuron in the adjacent layer has "connection right". What the neural network has learned is contained in the connection weight and threshold (bias) \[13\]. Its expression is as follows:

\[
y = f(\sum_{i=1}^{n} \omega_i x_i + b)
\]  \hspace{1cm} (1)
2.4. Convolution neural network

Convolutional Neural Networks (CNN) was first proposed by LeCun in literature\textsuperscript{[15]}. The algorithms include AlexNet, LeNet and VGGNet. CNN is a neural network directly used to process images\textsuperscript{[15]}. By establishing a back propagation program, the learning process is also strengthened. The basic unit of CNN is neuron, which has input/output connection with other neurons. Each neuron can process the information collected from all input nodes through a linear function and an activation function. In order to distinguish the result from the expected value, the network also includes a loss function.

3. Data set production

A total of 3478 gene sequences of Neisseria gonorrhoeae were included in the data set of drug resistance analysis of Neisseria gonorrhoeae, of which 90% were artificially divided into training sets with 3130 sequences, and 10% were test sets with 348 sequences. Because the length of each gene sequence is inconsistent, it is sorted into a gene sequence with a length of 515, and the insufficient one is supplemented with 0, in which the test set does not participate in the training of the model. The predicted tag shows the probability that this gene fragment may develop drug resistance, where \([0, 1]\) indicates that it is completely impossible to have drug resistance, whereas \([1, 0]\) indicates that it will completely develop drug resistance. The following is a schematic diagram of Neisseria gonorrhoeae gene sequence data set as shown in Fig.1.

![Figure 1. Schematic diagram of Neisseria gonorrhoeae gene sequence data set](image)

Staphylococcus aureus is one of the food-borne pathogens and can spread widely in the natural environment. Staphylococcus aureus can produce enterotoxin at the same time, which is easy to cause food poisoning. Recently, there have been numerous incidents of food poisoning caused by Staphylococcus aureus. Vomiting, nausea, loss of appetite, abdominal pain, diarrhea, dysentery, collapse and slight fever caused by Staphylococcus aureus can endanger life in serious cases. The data set consists of 216 gene sequences, of which 194 are training sets and 22 are test sets. The original gene sequence data set is shown in Fig.1.

![Figure 2. Schematic diagram of original data set of Staphylococcus aureus](image)

4. Analysis of experimental results

The CPU of the experimental platform used in this paper is Intel Core i7-10750H, the graphics card model is NVIDIA GeForce GTX 2070s, the video memory capacity is 8GB, and the memory is 16GB.
The Tensorflow deep learning framework is used for development and testing, and CUDA is used to accelerate the GPU training process in parallel.

Using machine learning, analysis of drug resistance of Staphylococcus aureus can be started, and the analysis result is represented by [1, 0] or [0, 1], where [1, 0] indicates complete resistance to drugs, whereas [0, 1] indicates complete impossibility of drug resistance.

In this experiment, the drug resistance of Staphylococcus aureus was analyzed by using random forest of machine learning, XGBoost and DNN and CNN of deep learning respectively, and the experimental results were obtained to compare the advantages and disadvantages of the algorithms. Then, the drug resistance of Neisseria gonorrhoeae was analyzed and the performance of the algorithm was compared.

4.1. Analysis of drug resistance of Staphylococcus aureus
First of all, compared with the decision tree algorithm, the random forest algorithm has obviously better accuracy than the decision tree algorithm, reaching about 98%.

After parameter adjustment, the parameter configuration is as follows. We use information entropy as the judgment standard. The higher the information entropy, the more uncertain the classification result is, thus guiding the program to classify again. N_estimators represents the number of decision trees in the random forest. We set it to 3. Min_samples_split represents the minimum number of samples before continuing to bifurcate. It is set to 10. Fig. 3 is the structure diagram of random forest training.

![Figure 3. Structure diagram of random forest training](image)

Among them, entropy is the information entropy, samples is the number of samples owned by the classification, and the information entropy of each classification decreases layer by layer. This random forest has three decision trees, each decision tree makes judgment at the same time, and finally weighted and integrated into the final classification result.

Secondly, XGBoost algorithm is used to re-train and predict the above data set.

During the whole training process, the accuracy of the XGBoost recognition algorithm is close to 99%, and the parameters of the XGBoost algorithm are as follows: the number of threads when XGBoost runs is expressed as nthread, which is set to 4, and the loss function objective is set to softmax. The maximum depth of the tree is max_depth, set to 6. Subsample is used to divide the data set and is set to 0.7. Colsample_bytree is the ratio of XGBoost to feature sampling set to 0.7.
Min_child_weight is the sum of the minimum sample weights that continue to divide leaf nodes, and the value is 3.

The following figure shows the distribution of important features identified by XGBoost in the entire data set.

![Feature importance chart]

**Figure 4.** Distribution of XGBoost Important Features

To sum up, for the data set of Staphylococcus aureus, the random forest in machine learning, XGBoost algorithm and DNN and CNN algorithms in depth learning have similar performance, so the above algorithms can be used for drug resistance analysis.

### 4.2. Analysis of drug resistance of Neisseria gonorrhoeae

Firstly, the Neisseria gonorrhoeae data set is analyzed by using random forest. The data set is randomly divided into five categories, and a total of five operations are carried out. Four categories are selected for training and one category is used for verification in each operation. Finally, five results are obtained to be the best. The final calculation result is shown in the following figure 5.

![Confusion matrix]

**Figure 5.** Random forest confusion matrix

As shown in the figure, the initial manually given parameters are respectively the maximum depth of the tree is 3 or 5, the feature number is 52 or 258 or 412, and the number of decision trees is 25 or 50. After automatic iteration of the model to select parameters, the maximum depth of the tree is determined to be 3, the feature number is determined to be 258, and the number of decision trees is determined to be 50. The confusion matrix given in the above figure, that is, the diagonal line is the predicted value, which is consistent with the real label, and the accuracy rate is 97.12% after calculation.

Secondly, XGBoost is used to analyze the above data sets. The results obtained after model calculation are shown in the following figure 6.
As shown in the figure, the model after parameter adjustment is shown in the figure. Among them, alpha is the L1 regularization term of weight, taking 1e-05, colsample_bytree is the proportion of the number of randomly sampled columns per tree, taking 0.6, gamma is the penalty term combined with leaf nodes, taking a value of 0.05, and the accuracy rate is 97.12%, which is almost the same as that of random forest.

Finally, the DNN and CNN network models were used to analyze the drug resistance of the above data sets. The accuracy and loss rate of DNN model are shown in the following figure 7.

As shown in the above figure, the accuracy rate is directly proportional to the increase of epoch, while the loss rate is inversely proportional. The final accuracy rate is stable at about 97%.

Similarly, the accuracy and loss rate of CNN are shown in the following figure 8.

The accuracy and DNN are also directly proportional to epoch, the loss rate is inversely proportional to epoch, and the final accuracy is also stable at 97%. However, compared with DNN, the accuracy and loss rate of CNN have obvious floating trend, so compared with CNN, DNN model should be used in drug resistance analysis.
5. Conclusion
Aiming at the task of antimicrobial resistance analysis, this paper proposes a variety of classification and regression models based on machine learning. It includes random forest, XGBoost, deep neural network and convolution neural network. Firstly, the data sets of Staphylococcus aureus and Neisseria gonorrhoeae are processed, then the corresponding key features, such as information entropy, are extracted, and the parameters of each model are optimized and selected. Finally, the drug resistance of known drugs is predicted. Experimental results show that the detection accuracy of the machine learning model in this paper is about 97%, and it has a fast convergence speed, which can effectively predict drug resistance according to relevant characteristics.

References
[1] Breiman L . Random forest [J]. Machine Learning, 2001, 45:5-32.
[2] Liaw A , Wiener M . Classification and Regression by randomForest [J]. R News, 2002, 23 (23).
[3] Svetnik V . Random forest: a classification and regression tool for compound classification and QSAR modeling. [J]. Journal of Chemical Information & Computer ences, 2003, 43.
[4] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System [J]. 2016.
[5] Tomislav H , Jorge M D J , Heuvelink G B M , et al. SoilGrids250m: Global gridded soil information based on machine learning[J]. Plos One, 2017, 12(2):e0169748.
[6] General Chair-Krishnapuram B , General Chair-Shah M , Program Chair-Smola A , et al. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM, 2016.
[7] Kovtun D . Forecasting with Many Predictors [J]. 2016.
[8] Lecun Y , Bottou L . Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.
[9] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS. Curran Associates Inc. 2012.
[10] Karpathy A , Toderici G , Shetty S , et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// Computer Vision & Pattern Recognition. IEEE, 2014.
[11] Lawrence, Steve, Giles, et al. Face recognition: A convolutional neural-network approach.[J]. IEEE Transactions on Neural Networks, 1997.