Prediction of Antimicrobial Resistance Based on Random Forest Algorithms

Xiujuan Xie[1,a,*], Xiangju Li[1,b], Yu Sheng[1,c], Bing Gu[2,d]

[1]Department of Computer Engineering,Southeast University Chengxian College, Nanjing 221006, China

[2]Department of Laboratory Medicine, Affiliated Hospital of Xuzhou Medical University , Xuzhou 221006, China

[a]21788194@qq.com, [b]215142863@qq.com, [c]727829995@qq.com, [d]gb20031129@163.com

*Corresponding author

Keywords: Drug Resistance Prediction, Antibiotics, Random Forest Algorithm, Data Mining

Abstract: Most hospitals use traditional bacterial culture methods to detect bacterial resistance, which has a long cycle and delays doctors' understanding of patients' antimicrobial resistance, and brings challenges and difficulties to clinical drug use. Therefore, this paper proposes to apply random forest algorithm of data mining classification method in antimicrobial drug resistance detection. Using a large number of historical data of bacterial susceptibility testing in a third-class A hospital as the original data set, the classification model of antimicrobial resistance is obtained through pre-treatment, model training and model evaluation, and then the new strains can be predicted. In addition, some representative data sets are selected and compared with the traditional decision tree C4.5 algorithm. The experimental results show that the random forest algorithm has better prediction effect and performance, and has certain practical application value.

1. Introduction

After years of development, antibiotics have been widely used in clinic, but at the same time, the problem of irrational use of antibiotics has become more and more serious, abuse occurs from time to time, which has a negative impact on human life and health and social and economic development, and the abuse of antibiotics has brought about various adverse reactions. According to statistics, adverse drug reactions or events caused by antibiotics account for the largest proportion of all kinds of drugs in China, nearly half of them [1]. In addition, doctors have the following and experience in the choice and use of antibiotics, ignoring the influence of patients' different physiological indications on the resistance of antibiotics, so irrational use of drugs not only affects the therapeutic effect, but also leads to the increase of adverse drug reactions, treatment costs and treatment time. At present, the traditional bacterial culture method is widely used in the detection of bacterial resistance in major hospitals. Bacterial culture usually takes more than one day to identify bacteria, and it takes more than two days to complete the drug sensitivity test. It takes five days or even a week to meet difficult bacteria. This traditional detection method is delayed. The opportunity for doctors to know the antimicrobial resistance of patients brings challenges and difficulties to clinical drug use. Therefore, how to use antibiotics quickly and accurately is of great significance, which is a difficult problem facing all medical institutions[2].

Data Mining is a new applied research field of artificial intelligence and machine learning technology. It has a very broad development prospect. It can extract potentially useful information and knowledge from a large number of incomplete, noisy, ambiguous and random application data, which people do not know beforehand. Today, with the rapid development of Internet technology, the data of various industries is growing at an explosive speed. Traditional data analysis can only complete simple data query and statistical operation. The introduction of data mining technology can help us obtain valuable information quickly and accurately from vast amounts of data. At present, data mining technology has been widely used in financial industry, insurance industry, retail industry, education, communications, scientific research and other fields. In recent years, more and more researchers began to explore the application of data mining technology in the

  DOI: 10.25236/cmph.2019.026

medical and health industry. Among them, clinical medical diagnosis and medical imaging applications are more typical [3], such as: Jin Xian et al. [4] using FP-growth association rule algorithm to mine the relationship between coronary heart disease and the etiology of the disease, Li Ge et al. [5] Established a prediction model of diabetic complications based on artificial neural network algorithm, Qian Yan et al. [6] Construct a classification model between diabetes mellitus and its influencing factors by using decision tree algorithm, Wang Lijun et al. [7] carried out the correlation analysis of the features of abdominal CT image pixels based on the algorithm of classification tree and and get the image rules for identifying normal and diseased tissues and organs, Li Xian et al. [8] proposed a method of MR image segmentation of nasopharyngeal tumors based on random forest feature selection algorithm to optimize the selection of original manual features and construct a segmentation model, so as to better segment nasopharyngeal tumors MR images.

From the existing statistical data, data mining of antimicrobial agents is relatively rare, but it has very important practical significance. Therefore, this paper proposes a classification algorithm based on data mining, which trains a large number of historical data of bacterial susceptibility testing to obtain a classification model of antimicrobial resistance, and uses this model to judge the degree of the tested bacterial resistance to antimicrobial drugs, so as to provide doctors with the most suitable decision-making for individual patients in the first time, as well as supporting the rational use of antibiotics.

2. Relevant background technology

2.1 Introduction of related technologies

Classification algorithm is an important supervised analysis method in data mining. By analyzing the training set of known categories in sample data, a suitable classification model (i.e. classifier) is established, from which classification rules are found to predict the classification of new data. Classifiers can be divided into single classifier and multi-classifier. The common single classifiers are Bayers and Decision Tree. The single classifier often has some problems such as local optimum and over-fitting. Therefore, according to the idea of ensemble learning, researchers put forward multi-classifier, also known as ensemble classifier, aiming at improving the performance of the single classifier. The aim is to combine several weak individual classifiers into a strong classifier according to some integration strategy, and synthesize the predicted results of multiple individual classifiers, so as to make decisions.

According to the different ways of generating individual classifiers (learners), ensemble learning methods are divided into two categories: Boosting method and Bagging method. In the former, there are dependencies between individual classifiers, so it needs to be generated serially, in the latter, there is no dependency between classifiers, so it can be generated in parallel. The main idea of Boosting method is to dynamically adjust the weight of each sample in the classifier according to the results of the previous round of classification, reconstruct the classifier for the adjusted training set, and finally judge the final result according to the prediction results of multiple weak classifiers. Bagging method, also known as Bootstrap, uses the sampling method with playback to extract different training sets from the initial sample data, constructs classifiers for each training set, and then synthesizes all classifier results to make final judgment.

2.2 Random Forest Algorithms

Random Forest (RF) is an integrated learning method based on Bagging, which was first proposed by Breiman [9] in 2001. The Bootstrap method is used to extract multiple sample sets from the initial data, and the decision tree model is built for each sample set. The prediction of multiple decision trees is combined. Finally, the voting method is used to get one of the categories with the highest number of votes as the final output.

Random forest is essentially an integrated classifier composed of multiple decision trees. The construction process of each decision tree is as follows:

Step 1: From the initial data with n samples, RF uses Bootstrap method to extract n samples randomly and playback as a new training set.

Step 2 uses the strategy of random subspace partition to randomly select m attributes from M original attributes as a new subset of attributes.

Step 3 builds classification tree according to CART algorithm of decision tree. For each node of the tree, Gini coefficient (Gini) is used as the basis for partitioning, and the attribute with the smallest Gini coefficient value is selected to divide the data on the node into its left and right sub-nodes. Gini coefficient is defined as formula (1).

$$Gini(D,A)= \frac{|D_1|}{|D|}Gini(D_1)+ \frac{|D_2|}{|D|}Gini(D_2) \quad (1)$$

$$Gini(D)=1- \sum_{k=1}^{2} \frac{|c_k|}{|D|} \quad (2)$$

Among them, D is the initial data set, A is the attribute set, D1 and D2 represent the left and right children of the current parent node, $\frac{|D_i|}{|D|}$ represents the proportion of left/right children in the parent data, $\frac{|c_k|}{|D|}$ represents the proportion of K samples.

3. Application of Random Forest Algorithms in Antimicrobial Resistance Prediction

In this paper, random forest algorithm is applied to predict the resistance of antibiotics. Before the traditional artificial results of bacterial resistance test come out, the tolerance value of current bacteria to various antibiotics is predicted at the first time, which provides assistant decision-making for doctors in clinical medication. Specific steps include data preprocessing, model training, model evaluation and new strain prediction. The flow chart is shown in Figure 1.

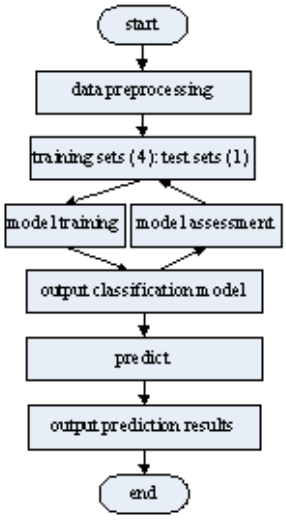

Figure 1 antimicrobial resistance prediction process

3.1 Data Preprocessing

The data in this paper are based on the historical data of bacterial susceptibility in the intensive care unit (ICU) of a tertiary A hospital in Jiangsu Province from 2014 to 2015, totaling 12107 items. Among them, there are 156 kinds of bacteria and 53 kinds of antibiotics. The actual production data is relatively complex, and data mining can only be carried out after preprocessing. The data Preprocessing work in this paper includes data filtering, outlier detection and standardization.

Firstly, data filtering: Considering the different sensitivity of the same strain to different antimicrobial agents, and the different sensitivity standard values of the same antimicrobial agent to different strains, such as: the minimum inhibitory concentration (MIC) of bacterial ABA to AMP and SXT is different. The sensitivity or resistance criteria of AMP to bacterial ABA and bacterial

ACA are different, so it is necessary to establish a one-to-one classification model for each strain and each antibiotic. With the help of professionals, the original data were screened, irrelevant indicators were removed, and four attributes that might affect the tolerance of antimicrobial agents under specific strains were selected, they are: patient's gender, patient's age, source department and specimen type, and a classification label,that is the minimum inhibitory concentration of antimicrobial agents(MIC).

Second, anomaly detection: using the experience of experts'historical detection, we can check whether there are unreasonable detection results in the original data. Because the data comes from rigorous and regular hospital detection institutions, there are fewer anomaly records, and the data with more missing values are also included in the anomaly data, and the records with anomaly values are deleted directly.

Thirdly, standardization: In order to facilitate the later data processing, four features are quantitatively represented by numerical values in order to achieve the unification of quantization methods. The results are shown in Table 1.

Table 1 Quantitative results and descriptions

variable name	value range	quantitative value	variable description
sex	male	1	sex of patients
	female	2	
age	<14	1	age of patients
	15-20	2	
	21-30	3	
	31-40	4	
	41-50	5	
	>50	6	
department	new born ICU	1	source departments
	pediatrics ICU	2	
	emergency treatment ICU	3	
	critical care medicine ICU	4	
	neurological intensive care unit	5	
specimenType	ab	1	specimen type
	at	2	
	ba	3	
	bi	4	
	bl	5	
	…	…	
	sp	16	
	ur	17	

3.2 Model Performance Evaluation Indicators

In this paper, F measure [10] is used as the evaluation criterion of classification results. This method takes both precision (P) and recall (R) into account. P, R and F are calculated by the following formulas respectively.

$$P=TP/(TP+FP) \quad (3)$$

$$R=TP/(TP+FN) \quad (4)$$

$$F=(2*P)/(P+R) \quad (5)$$

Among them, TP is the sample number of positive classes, FP is the sample number of negative

classes, FN is the number of samples with positive classes divided into negative classes, TP + FP is the sample number of actual classes, TP + FN is the sample number of due classes.

3.3 Model training and analysis

In the original sample data, an average of about 20 kinds of antimicrobial agents are tested annually by one bacterium. The types of antimicrobial agents tested are different each year, but there is a large intersection. Moreover, from the test results, the distribution of MIC categories of some antimicrobial agents by some bacteria shows great imbalance. For the sake of objectivity,on the basis of the pre-processed data, three bacterial strains-antibiotics data sets with large sample size and balanced sample distribution were further screened as the experimental data sets for the training and analysis of this model, as shown in Table 2.

Table 2 Representational experimental data sets

data sets	sample size	characteristic number	number of label categories
aba-SXT	2339	4	10
kpn-ATM	1779	4	10
sau-TCY	1495	4	12

Based on the above data sets, random forest algorithm (RF) is compared with traditional decision tree algorithm (C4.5). The two superparametric values of RF algorithm are determined by combining empirical values and OOB (out-of-pocket data) mean square error. The size of candidate feature set M = log2 (M), i.e. 2, the number of decision trees L = 200.For the sake of objectivity, the experimental process is validated by five fold cross validation. The results of comparison between RF algorithm and C4.5 algorithm (without pruning) on three data sets are shown in Table 3.

Table 3 Contrast results

data sets	training time /s		Forecasting time /ms		F measure/%	
	RF	*C4.5*	*RF*	*C4.5*	*RF*	*C4.5*
aba-SXT	4.59	21.73	53.07	5.07	85.76	76.32
kpn-ATM	3.49	16.33	40.37	3.85	80.92	72.26
kpn-ATM	2.93	13.72	33.93	3.24	78.31	70.34

From Table 3, it is easy to see that the F value and training time of RF algorithm are better than that of C4.5 algorithm. In the process of five fold cross validation, the F value of RF algorithm is relatively stable and higher than that of C4.5 algorithm, so it is not easy to have over-fitting problem. In addition, the training time of the RF algorithm is much shorter than that of the C4.5 algorithm, which is mainly due to: firstly, the decision tree forest of the RF algorithm is constructed in parallel with high execution efficiency; secondly, the training data of each decision tree in the RF algorithm is a subset of the original data set, which reduces the time-consuming of a single decision tree. In terms of prediction time, RF algorithm is inferior to C4.5 algorithm. The main reason is that each decision tree in RF needs to be predicted separately, then combined, and finally voted to get classification results. However, as shown in Table 3, the total prediction time of RF is millisecond level, which has little impact. Therefore, the application of stochastic forest algorithm to the prediction of antimicrobial resistance has better overall classification effect and performance.

3.4 Drug sensitivity prediction

In this paper, a visual drug sensitivity prediction platform is constructed. After login to the system, input the patient's relevant characteristic information and click the result button, the MIC prediction values of current bacteria for different antimicrobial agents will be displayed in the list below the page, as shown in Figure2.

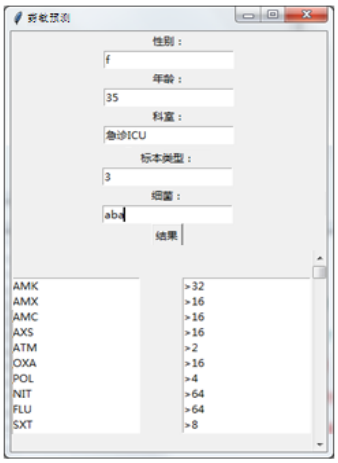

Figure 2 Prediction effect chart of new strains

4. Conclusion

In this paper, the random forest algorithm of classification method is applied to the prediction of antimicrobial resistance. A large number of historical data of antimicrobial susceptibility testing in a third-level hospital are used as data sets to train and get the classification model of antimicrobial resistance. Using this model, the tolerance of the tested bacteria to antimicrobial drugs can be predicted at the first time before the results of traditional artificial bacterial resistance test, so as to provide the most suitable decision-making support for individual patients for doctors' clinical medication. Some representative data sets are selected and compared with traditional decision tree algorithm C4.5. The experimental results show that the Random Forest algorithm has better classification effect and performance in the prediction of antimicrobial resistance.

Acknowledgements

Fund project: It is supported by the National Natural Science Foundation of China (81871734) and the Education Informatization Research Project of Jiangsu Province (20180054).

References

[1] An Minmei.Adverse drug reactions of common antibacterial drugs and analysis of influence factor[D]. Shandong University,2017. (in Chinese).

[2] Wang Yimin, Liang Zhigang. Data mining of antimicrobial drugs based on immune genetic algorithm [J]. Computer system & application, 2017,26(3)：156-161. (in Chinese).

[3] Liu Yanpei, Liu Enshun. Summary of Medical Data Mining [J].Guangming Traditional Chinese Medicine, 2018, 301 (12): 40-42. (in Chinese).

[4] Jin Yi. Application of correlation analysis in coronary heart disease diagnosis and treatment data [D]. Central South University, 2008. (in Chinese).

[5] Li Ge, Jin Lizhong. Establishment of a predictive model for diabetic complications based on learning vector quantization network [J].Chinese Journal of Virology, 2006, 8(4): 254-258. (in Chinese).

[6] Tan Yan. Data Mining Research on Diabetes Related Factors in Electronic Health Archives [D]. University of Electronic Science and Technology, 2013. (in Chinese).

[7] Wang Lijun. Medical image classification based on association rules [D]. Jiangsu University, 2006. (in Chinese).

[8] Li Xian, Wang Yan, Luo Yong. Nasopharyngeal neoplasm segmentation based on random forest

feature selection algorithm [J]. Computer applications, 2019, 39 (05): 245-249. (in Chinese).

[9] Beriman L.Random Forests[J].Machine Learning,2001,45(1):5-23.

[10] Zhong Jiang, Liu Ronghui. Improved KNN text categorization [J]. Computer Engineering and Applications, 2012, 48 (2): 142-144. (in Chinese).