Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 3, 2024

Prediction of Postoperative Survival Time of Hepatocellular Carcinoma Patients Based on Machine Learning

Zhiyao Wang[1, *]

[1]Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[*]Corresponding author's e-mail: wzy040429@outlook.com

Abstract: The survival rate of patients with primary liver cancer after operation is very low. Accurate prediction of postoperative survival of cancer patients plays a key role in the whole treatment period, both for patients and medical staff. For cancer patients, accurate prediction of survival time can make patients have a positive and optimistic attitude when fighting against the disease, and can also reasonably plan the rest of life to realize some wishes. For medical staff, doctors can only predict the survival time of patients through their experience before, and patients' expectations for curing diseases are generally high. Accurate survival prediction can avoid the contradiction between doctors and patients. According to the difference of prognosis of patients with clinical data, medical staff can make individualized postoperative treatment plan for patients to prolong the survival time of patients. To sum up, it is necessary to accurately predict the postoperative survival time of patients through clinical data. In this paper, survival time of 437 hepatocellular carcinoma patients was divided into 1, 3, 5 years and the unbalanced data set is oversampled and shuffled. Then 784 samples are divided into training set and test set by ten fold cross validation. Six machine learning algorithms, were used to predict the postoperative survival time of patients. The performance of the model is evaluated by calculating the accuracy. The optimal machine learning model is selected and the XGboost algorithm will be further optimized by Grid Search. Finally, the survival time of hepatocellular carcinoma patients will be well predicted.

Keywords: Primary hepatocellular carcinoma; Unbalanced datasets; XGboost algorithm; Prediction model of postoperative survival time.

1. Introduction

Liver cancer is one of the most common malignant tumors in the world, and its mortality rate accounts for 8.3% of all cancer deaths in the world. Primary liver cancer has been paid more and more attention by the society, and the survival rate of patients with primary liver cancer after operation is very low. Accurate prediction of postoperative survival of cancer patients plays a key role in the whole treatment period, both for patients and medical staff [1]. At present, most studies on the prognosis of liver cancer patients by machine learning focus on medical imaging data and genetic data, and few people study clinical data .Different from other malignant tumors, the liver function and physical condition of patients will affect the operation result. According to the clinical data, the medical staff can make individualized postoperative treatment plan for the patient to prolong the patient's survival time.

At present, Logistic regression and Cox proportional risk model are mainly used to predict the prognosis of liver tumors. But these are all based on linear assumptions, and the prediction effect is limited. In recent years, using various machine learning algorithms to predict disease risk has become a research hotspot in the field of medical big data.

In 2001, He Jia and others used Cox regression to screen 11 clinical features including tumor size, therapy from 34 clinical features, and then used BP neural network to predict the survival time of hepatocellular carcinoma patients after radiotherapy [2]. In 2004, Shen Yu and others divided the survival time of patients into six months and one year for prediction. Based on 192 clinical cases of primary liver cancer, 16 characteristics including IBIL2, GIB1*, GIB2*, WBC1* were screened out from 41 clinical characteristics by looking for attributes with large statistical distribution differences in different categories, then using Bayesian classification to predict the postoperative survival time of patients [3].

Shen et al. used Artificial Neural Network, Logistic regression to establish a survival prediction model for patients with early liver cancer after operation, and found that the AUC of Artificial Neural Network model was higher than that of other models [4]. Ho et al. used Logistic regression, Artificial Neural Network and Decision Tree to construct the 1,3,5 year disease-free survival model of hepatocellular carcinoma patients undergoing hepatectomy, and further found that the accuracy of the Artificial Neural Network model is higher [5]. Chiu et al. screened 21 features by Cox regression model and then trained Artificial Neural Network model and Logistic regression model to predict the 1,3,5 year survival rate of patients after hepatectomy. The results showed that Artificial Neural Network was superior to Logistic regression [6,7].

In 2019, Lin Li divided patients into two categories with a threshold of three years to predict the survival time of postoperative hepatocellular carcinoma patients. MSVM-RFE feature selection was performed in 343 males and 43 females , to screen out important clinical features. Three-year survival prediction model and three-year recurrence model were established based on the variables after feature selection using seven classification algorithms. The results show that SVM and Random Forest have better prediction effect [8]. In 2022, Ding Kexin combined gene expression characteristics and clinical characteristics to predict the tag of survival time of hepatocellular carcinoma patients by using XGBoost classification algorithm, and optimize the parameters of

XGBoost algorithm. Finally, BO-XGBoost algorithm with better prediction effect is selected for classification prediction [9].

2. Datasets

The dataset used in this paper contains 110 clinical features and survival time (OS.time) of 438 patients with hepatocellular carcinoma from UCSC Xena database.

2.1. Data preprocessing

2.1.1. Fill in missing values

The lack of clinical features is relatively serious. If the missing value of a clinical feature is higher than 20%, it is considered that the feature is of little significance and deleted it. For the remaining missing clinical features, the numerical data are filled with mean value; then the nonnumerical data, the features that are almost different from each sample are deleted first, and others are filled with mode value, and then the string data are converted into numeric data.

2.1.2. Data normalization

In this paper, Max-Min normalization method is used to make the fluctuating clinical features have the same dimensional values and all fall on [0, 1]. The calculation formula is as follows:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

2.1.3. Variance filtering

If the variance of a feature itself is very small, the fluctuation of the value on the feature will be very small. The influence of such features on the target can be ignored. In this study, features with variance less than 0.01 are deleted.

2.1.4. Sample label processing

The survival data were classified with a threshold of 1 year (365 days), 3 years (1095 days) and 5 years (1825 days), and then the samples were labeled. Patients with less than 1 year of survival labeled 0, 1 to 3 years patients labeled 1, 3 to 5 years patients labeled 2, more than 5 years patients labeled 3, and then remove samples with empty labels. In this paper, 437 samples are used.

Table 1. Label classification

Class	Label	Number of samples
less than 1 year	0	123
1 to 3 years	1	196
3 to 5 years	2	64
more than 5 years	3	54

2.1.5. Unbalanced dataset processing

The number of samples in each class is quite different, which affects the prediction effect of the classifier. In this paper, random oversampling is used to deal with the unbalanced data set, that is, random replication of minority samples to make the number of minority samples equal to the number of majority samples, so that a new balanced data set finally contains 784 samples.

2.1.6. Shuffle dataset

The order of the data set may contain some biases. By shuffling the data set, the model can get more different types of data in training, which is helpful for the model to better generalize the data that has not been seen before.

3. Feature Selection

The 17 features with the strongest correlation were selected from 35 clinical features by Spearman correlation coefficient.

Table 2. Feature selection

Feature
creatinine_lower_level
creatinine_upper_limit
platelet_result_lower_limit
platelet_result_upper_limit
sample_type_id
year_of_initial_pathologic_diagnosis
new_tumor_event_after_initial_treatment
pathologic_M
pathologic_N
pathologic_T
pathologic_stage
person_neoplasm_cancer_status
post_op_ablation_embolization_tx
residual_tumor
specimen_collection_method_name
tissue_retrospective_collection_indicator
vital_status

4. Machine Learning Algorithm

4.1. Classification algorithm

4.1.1. KNN

The K nearest training samples are found out by calculation and the classes of the K training samples are known. Then the unknown class samples are predicted according to the information of the K training samples. Euclidean distance is used to calculate the distance, and the voting method is used to quickly judge which class the unclassified samples belong to.

4.1.2. Naive Bayes

A classification algorithm using probability statistics knowledge. In essence, Bayesian algorithm actually infers a prior probability according to the existing knowledge, and then constantly adjusts this probability according to the new evidence. For the unclassified samples, the probability of each class under the condition that the sample appears is calculated, and the sample belongs to the class with the greatest probability.

4.1.3. Logistic regression

Using hypothesis testing to statistically infer and analyze data. The variables are screened by calculating the regression coefficient of each variable, and establish regression model with maximum likelihood method.

4.1.4. SVM

The SVM model represents instances as points in space such that mappings allow instances of individual categories to be separated by as wide a significant interval as possible. The new instances are then mapped to the same space. The category they belong to is predicted based on which side of the interval they fall on.

4.1.5. Random forest

A multi-classifier with Decision Tree algorithm as meta-classifier. To generate a new training sample set, RF uses Bootstrap resampling technique to randomly extract K

samples from the original training sample set. Random sampling is repeated and then N classification trees are generated based on the sample set sampled by Bootstrap. Finally a random forest is formed.

4.1.6. XGboost

In the construction of multiple weak classifiers, boosting adjusts the probability of each sample being selected according to the prediction results of the previous weak classifier for each training set sample. XGBoost is a machine learning algorithm based on Gradient Boosting framework [10].

4.2. Grid Search parameter optimization

Optimize several common parameters of XGBoost. In this study, the five parameters including learning_rate, subsample, colsample_bytree, max_depth, min_child_weight are used for Grid Search. The basic idea of grid search algorithm is exhaustive search, that is, to try every possibility through loop traversal among all candidate parameter choices. The best parameter is the final result.

5. Model Performance

5.1. Ten fold cross validation

Ten fold cross validation is to divide liver cancer patients into 10 subsets and take 9 sample subsets as training sets every time, and use classifier to label the survival time of liver cancer patients in the training set. The remaining subset is used to test. The output result is the average of Precision, recall and F1_score on 10 test subsets. Through comparison,obtain the optimal classification algorithm [11].

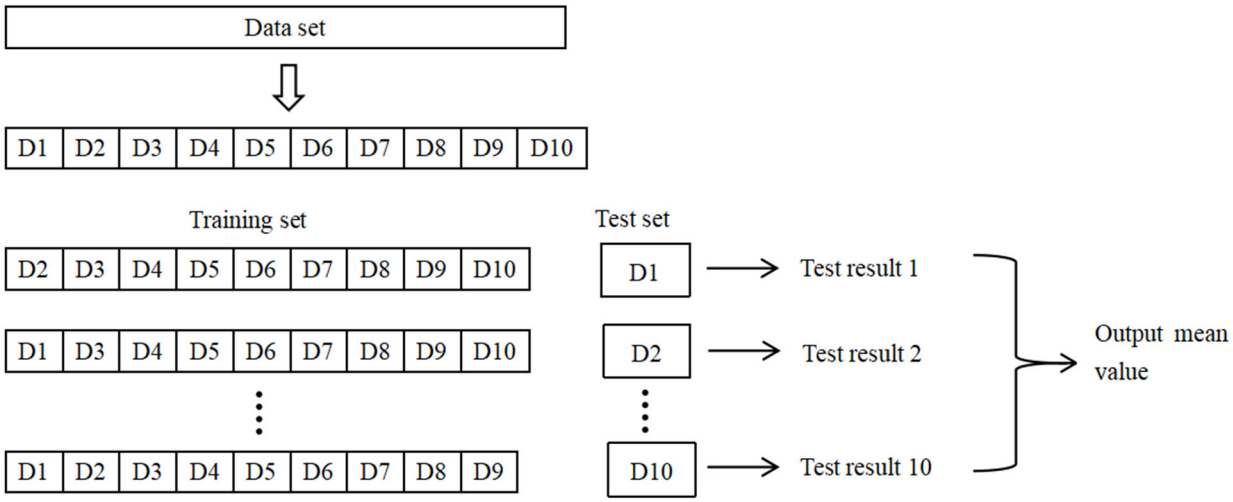

Figure 1. Ten fold cross validation flow chart

5.2. Common evaluation indexes of classification models

Accuracy: The proportion of the samples whose real survival labels are consistent with the predicted survival labels of hepatocellular carcinoma patients in all samples.

Take Classification 1 as an example (at this time, Classification 1 is regarded as a positive sample and the rest are negative samples):

Precision: The proportion of samples with both predicted and real survival labels of 1 to all samples with predicted labels of 1.

Recall: The proportion of samples with both predicted and real survival labels of 1 to all samples with real labels of 1.

Precision and recall often have contradictory situations, which need to be considered comprehensively. F1_score is the harmonic average of precision and recall.

$$\text{F1\_score} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \quad (2)$$

6. Results

6.1. Comparison of classification algorithms

KNN, Naive Bayes, Logistic regression, SVM, Random forest and XGboost are used as classifiers to predict the survival time of patients with liver cancer. By comparing the accuracy under ten fold cross validation, the optimal classification algorithm is obtained.

Table 3. Accuracy of classification algorithms

Algorithm	Accuracy
KNN	0.592
Naive Bayes	0.468
Logistic regression	0.632
SVM	0.603
Random forest	0.899
XGboost	0.901

The conclusion is that Random forest algorithm and XGboost algorithm have obvious performance in multi-classification prediction.

6.2. XGboost optimization

The parameters of XGboost algorithm are optimized by Grid Search.

Table 4. Accuracy of optimization algorithm

	Accuracy
XGboost	0.901
GS-XGboost	0.904

Table 5. Comparison of precision, recall and F1_score after optimization

	Class 0		Class 1		Class 2		Class 3	
	XGboost	GS	XGboost	GS	XGboost	GS	XGboost	GS
precision	0.875	0.864	0.864	0.879	0.913	0.917	0.942	0.951
recall	0.857	0.872	0.781	0.781	0.964	0.964	1	1
F1_score	0.866	0.868	0.820	0.827	0.938	0.940	0.970	0.975

After optimizing the parameters of XGBoost by Grid Search, the average value of F1_score of each classification is increased from 0.899 to 0.903, which is better than the default parameters. This shows that the model optimization is very necessary.

7. Summary

This is a study on the survival time of patients with hepatocellular carcinoma after operation based on machine learning. This article uses 1 year, 3 years and 5 years as the threshold to divide the samples, so that the prediction time is more accurate. The dimension of clinical feature data is low, so we use Spearman correlation coefficient to find out the features with strong correlation to predict. By comparing six classification algorithms, we finally choose XGBoost model after Grid Search parameter optimization to classify samples. Finally, the average value of the four classifications of F1_score can reach 0.903.

References

[1] Zhou, L.J., Cui, J., Zhao, J.J. (2009) Research progress of survival evaluation tools for dying cancer patients. Chinese modern nursing, 01:94-96.

[2] He, J., He, X.M., Liu, Q., et al. (2001) Prediction of survival time of liver cancer patients by BP Neural Network. Health statistics in China, 01:17-19.

[3] Shen, Y., Zhuang, T.G., Cheng, H.Y., et al. (2004) Prediction of prognosis of primary liver cancer by naive Bayes algorithm. Space Medicine & Medical Engineering, 05:350-354.

[4] Qiao, G., Li, J., Huang, A., et al. (2014) Artificial neural networking model for the prediction of post-hepatectomy survival of patients with early hepatocellular carcinoma. Journal of Gastroenterology and Hepatology, 29:2014-2020.

[5] Ho, W.H., Lee, K.T., Chen, H.Y., et al. (2012) Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network. PLoS One, 7: e29179.

[6] Chiu, H.C., Ho, T.W., Lee, K.T., et al. (2013) Mortality predicted accuracy for hepatocellular carcinoma patients with hepatic resection using artificial neural network. The Scientific World Journal, 2013:201976.

[7] Zeng, J.J. (2021) Prediction model of postoperative recurrence of liver cancer. Thesis of Fujian Medical University.

[8] Li, L. (2019) Prediction model of prognosis of liver cancer based on machine learning. Thesis of Xinjiang Medical University.

[9] Ding, K.X. (2022) Prediction of survival time of liver cancer based on machine learning. Thesis of Huazhong Agricultural University.

[10] Li, Z.S., Liu, Z.G. (2019) Feature selection algorithm based on XGBoost. Journal of Communications, 10:101-108.

[11] Yang, X.L. (2021) Overview of performance metrics of classification learning algorithms. Computer Science, 08:209-219.