Improved Feature Selection Algorithm for Prognosis Prediction of Primary Liver Cancer

Yunxiang Liu, Qi Pan, Ziyi Zhou

Improved Feature Selection Algorithm for Prognosis Prediction of Primary Liver Cancer

Yunxiang Liu[1] and Qi Pan[2] and Ziyi Zhou[3]

School of Computer Science and Information Engineering,
Shanghai Institute of Technology, Shanghai 201418, China
yxliu@sit.edu.cn, 463728073@qq.com, 2634578954@qq.com

Abstract. Primary liver cancer, one of the most common malignant tumors in China, can only be roughly diagnosed through doctors'expertise and experience at present, making it impossible to resolve the health problem that people care about. A new method that applies machine learning to the medical filed is therefore presented in this paper. The decision tree algorithm and the random forest algorithm are used to classify the data, and decision tree algorithm and improved feature selection algorithm to select important features. Comparison shows that the performance of the random forest algorithm is better than that of the decision tree algorithm, and the improved feature selection algorithm can filter out more important features on the premise of retaining accuracy.

Keywords: Primary Liver Cancer,Machine Learning, Decision Tree,Random Forest.

1 Introduction

Primary liver cancer is one of the most common malignant tumors in China,with its mortality in patients being the third in malignant tumors[1].Typically the prognosis of this disease can only be roughly judged through doctors' professional knowledge and experience.The low accuracy, therefore,has a negative effect on both doctors and patients.At present,systematic researches conducted by machine learning are few both at home and abroad, and there is no corresponding model or software to verify the classification of liver cancer data[2].Even if the iterative updating of medical treatment equipment cannot predict the occurrence of liver cancer.For solving the health problem that people care about more effectively,we try to use the machine learning into analyzing the data about primary liver cancer, hoping it can be used in clinical prognosis assessment and treatment option.This paper studies two machine learning algorithms--the decision trees and random forests,which are the most typical representatives of symbol learning and ensemble learning.

[1] Corresponding Author: Yunxiang Liu (yxliu@sit.edu.cn)
[2] Corresponding Author: Qi Pan (463728073@qq.com)
[3] Corresponding Author: Ziyi Zhou (2634578954@qq.com)

In addition,it also improves and verifies the character choosing method based on random forest so as to reduce the overhead of model training and the difficulty in data acquisition.By using Python language,this paper implements the above algorithms and organizes and tests the system interface.Besides,the classification accuracy and feature selection of different algorithms are also analyzed in this paper,providing reference for the selection of models.To sum up, this topic has considerable research significance both in the field of computer science and medical domain.

2 Principle

2.1 Decision Tree Algorithm

During late 1970s and the early 1980s,J.Ross Quinlan developed the decision tree algorithm which was originally called ID3. In the application process,Quinlan found and improved the shortcomings of ID3,and put forward the C4.5.In 1984,many statisticians published the book named "Classification and Regression Trees" (CART).A decision tree is a tree structure similar to a flow chart with each internal node representing a test on an attribute[3],each branch the corresponding output of the test,and each leaf node the category.A typical decision tree model is shown in Figure 1.The inner node is represented by an ellipse,and the leaf nodes by rectangles.Most decision trees are built by top-down recursion,selecting instances of a class from the known training set by using changes in entropy.

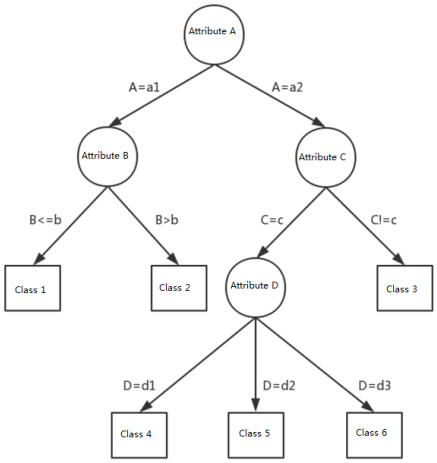

Fig. 1. Decision tree model

2.1.1 Information gain

ID3's attribute selection and measure are realized by information gain.Entropy[4], which is inversely proportional to the purity of the data set,is usually used to represent the uncertainty of random variables.Supposing that there are D samples and K

classes,with p_i being the probabilities of class I in D.The entropy of D is given in the following form:

$$H(D) = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{1}$$

When D is divided according to the attribute A,the feature A has n different values,and the new conditional entropy is defined as the subset of the D:

$$H(D \mid A) = \sum_{i=1}^{n} p_i' \cdot H(D \mid A = a_i) = \sum_{i=1}^{n} p_i' \cdot H(D_i) \tag{2}$$

By comparison,*H(D)*it can reflect the uncertainty of the original data set, *H(D|A)*indicate the uncertainty after division,and regard the difference between them as information gain.

$$G(D, A) = H(D) - H(D \mid A) \tag{3}$$

2.1.2 Gain rate

C4.5 uses an information gain called to expand the split[5]. In order to standardize the information gain,we use the "split information" value.The gain rate is defined as follows:

$$G_R(D, A) = \frac{G(D, A)}{H_A(D)} \tag{4}$$

among

$$H_A(D) = -\sum_{i=1}^{n} p_i' \log_2 p_i' \tag{5}$$

2.1.3 Gini index

Gini index is used to reflect the uncertainty of the number set,and Gini index is inversely proportional to the purity of the data set.The Gene index is defined as follows:

$$Gini(D) = \sum_{i=1}^{k} p_i(1 - p_i) = 1 - \sum_{i=1}^{k} p_i^2 \tag{6}$$

Under the condition of characteristic A,the Gini index of the sample set D is defined as:

$$Gini(D, A) = \sum_{i=1}^{n} p_i' \cdot Gini(D_i) \tag{7}$$

When the characteristics of the sample set have the smallest Gini index, the current feature is the best feature. Figure 2 below shows the relationship between the Gini index, the entropy half and the classification error rate in the two categories.

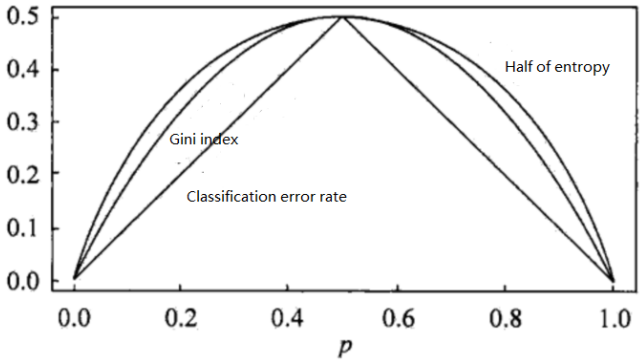

Fig. 2. Comparison of two indicators

2.2 Random Forest Algorithm

Random forest is built on the basis of decision tree.It is an integrated classifier model formed by multiple decision trees[6].To put it simple,multiple decision trees construct the random forest.Random forests have adopted the Bagging thought[7]and characteristic subspace thought,being of more anti-noise ability than a single decision tree.It will not over-fit and can significantly improve the generalization ability.The basic flow of a random forest is shown in Figure 3:

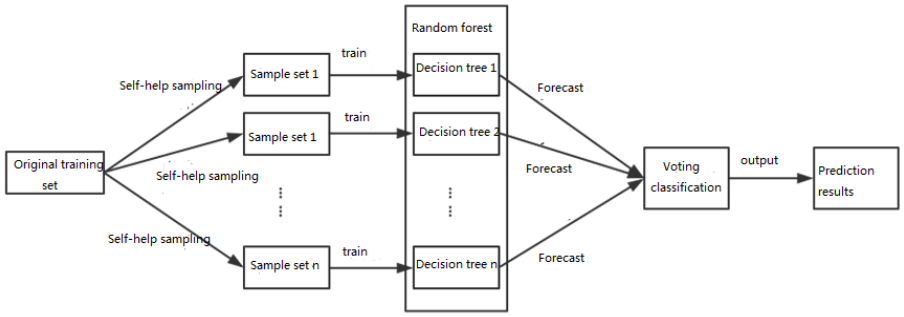

Fig. 3. The basic flow of a random forest

2.3 Improved Feature Selection Algorithm

The basic idea of selecting the characteristics by using the random forest is to sort the characteristics according to the importance first,then remove some features through generalized-sequence backward selection algorithm.Then we train random forest on the new feature set and calculate its accuracy rate;repeat the process,and finally use the

feature set with the highest accuracy as the output.In order to ensure the stability of each test result,cross validation is used to evaluate the newly established random forest after each round of screening,and the average accuracy rate is taken as the accuracy of that round.Compared with the wrapped method like LVW random selection feature subset,the algorithm's feature selection is heuristic and has higher efficiency.However,since the iteration will continue until the number of remaining features is reduced to the threshold,it still has a large time and space overhead;and because we make the final selection based on the highest test precision, the feature set is not necessarily the smallest one.

On this basis,a faster feature selection algorithm is designed to optimize the process.According to the error increment caused by each wheel screening,we judge whether to continue screening.Once it exceeds the specified threshold,the iteration is exited,and feature set selected in the last round is used as the result.For models trained on shrinking feature sets,the generalization performance tends to decrease, and the degree of reduction can be used as an evaluation criterion for feature sets. The essence of this strategy is to select the smallest feature subset in a given error range rather than the highest test precision,so that it can stop screening as soon as possible and save a lot of time.The reason why the error increment threshold is not simply set to 0 is that some weak correlation features are expected to be removed in addition to the unrelated features,and this can also allow small deviations of each test. The test results show that the selected feature set does not actually produce an error increment as large as the threshold value does,and the test accuracy on it can be even higher before screening.

3 Experimental Analysis.

3.1 Collection and Data Collation

The case data came from Eastern Hepatobiliary Surgery Hospital,Second Military Medical University, including malignancies, benign lesions and normal types.There were 588 groups of data, among which malignant tumor accounting for 246;benign lesions 149 and normal 193.The data itself has too many missing values, making it difficult to sort and classify the data.We delete many useless indexes with the help of professionals,leaving 39 anonymous indicators for each sort finally. As a result of privacy protection,the sample set has 693 missing values;Samples that have more than 5 missing values are automatically discarded by the program.The sample contains 6 discrete indexes and 519 effective sample groups.Different methods are used to treat the missing values in the decision tree model. Some endow the sample with the common value of the feature,while the method used in C4.5 gives a weight for each value of the feature,which is divided into the sub nodes with different probability.

3.2 Comparison between Decision Tree and Random Forest

After reading the data,70% of the data in the effective sample are selected as the training set and the rest as the test set.5 times of random training and testing are repeated,the

decision tree and the random forest model are created.We record the average value to realize a simple cross validation.The results of the recording were shown in Table 1.

Table 1. Comparison between decision tree and random forest

Model	Pruning algorithm	Training time	Prediction time	Test accuracy
decision tree	nothing	7.16s	1.69ms	86.13%
decision tree	PEP	7.16s +3.78ms	0.74ms	89.21%
Random forest	nothing	1.53s	17.72ms	92.16%

Obviously,in addition to forecasting time,the overall performance of random forests is better than that of decision trees, especially in training time.For training sets of the same size,the construction speed of the random forest is nearly 5 times faster than that of the decision tree.

Similarly,5 training and testing are conducted under random division,and the feature set selected after each training decision tree (PEP pruning) is recorded. F_i in the following table represents the characteristic subscript set of i test, and r_i the test accuracy that is used for reference:

Table 2. Features selected by the decision tree

i	F_i	r_i
1	{1,3,4,16,21,25,28,34}	90.12%
2	{1,2,3,4,16,21,33}	88.46%
3	{2,3,4,11,16,21,25,28,34,35}	87.92%
4	{2,3,4,11,15,16,21,28,33,34,37}	89.03%
5	{1,2,3,4,13,16,21,34,35}	89.35%

The intersection and union of $F_i(1 \le i \le 5)$ are resolved respectively as follows:

$$F_{\min} = \bigcap_{1 \le i \le 5} F_i = \{3,4,16,21\}$$

$$F_{\max} = \bigcup_{1 \le i \le 5} F_i = \{1,2,3,4,11,13,15,16,21,25,28,33,34,35,37\}$$

In the above formula, $F_{\min}$ represents the most important features that can be collected first in the collection of samples to avoid missing values as much as possible;while $F_{\max}$ can be used as a selected feature set.Then we use the improved feature selection algorithm for feature selection.

It better demonstrates a feature selection process with a default parameter.The filtered results of each group and the test accuracy robtained from the training are listed as follows,and those removed features will be discarded in the next round.

Table 3. Feature screening process

Rotation	Selected feature subscript (order of importance)	Characteristic number	*r*
0	4, 28, 14, 21, 15, 6, 26, 11, 16, 10, 3, 7, 25, 34, 30, 27, 36, 35, 17, 13, 38, 20, 24, 32, 22, 29, 37, 8, 33, 0, 12, 2, 18, ~~31, 5, 1, 19, 9, 23~~	39	91.82%
1	16, 4, 3, 28, 27, 6, 25, 14, 24, 21, 29, 13, 11, 8, 7, 26, 2, 22, 34, 15, 37, 20, 32, 17, 36, 33, 38, 10, ~~30, 12, 0, 18, 35~~	33	90.60%
2	4, 28, 25, 11, 16, 24, 21, 6, 3, 29, 15, 22, 26, 13, 8, 20, 17, 2, 34, 14, 7, 38, 32, ~~36, 37, 33, 27, 10~~	28	90.81%
3	6, 3, 4, 16, 21, 25, 28, 8, 15, 11, 34, 24, 7, 13, 26, 29, 20, 2, 14, ~~17, 38, 22, 32~~	23	92.73%
4	28, 3, 25, 16, 4, 21, 26, 15, 29, 13, 8, 6, 2, 34, 14, 7, ~~20, 24, 11~~	19	92.52%
5	4, 28, 2, 6, 3, 14, 21, 16, 25, 15, 13, 8, 26, ~~34, 29, 7~~	16	91.45%
6	4, 3, 21, 15, 28, 25, 16, 26, 13, 6, 14, ~~8, 2~~	13	91.87%
7	4, 3, 25, 21, 16, 15, 28, 26, 14, 6, 13	11	92.31%

As can be seen from the above table, the algorithm will eliminate some features with the lowest importance every time, and the number of discarded features will gradually decrease as a result of proportionate screening. The results obatined in seventh round are final results. The reason why we stop screening features is that the error increment in the eighth round exceeds fixed value 2%.It can be concluded that the accuracy of test before the screening stops does not change significantly with the decrease of feature numbers, but only fluctuates near the initial accuracy. This shows that the algorithm can effectively identify redundant and weak correlation features, and thus it can preserve the classifier performance while removing these features. It is also found during the test that the number of features having the highest test precision (23 in the upper case) is not stable and has a large randomness.

Table 4. Feature screening results

maxAccurDesc	Selected feature subscript (order of importance)	***r***

2.5	3, 28, 16, 4, 6, 27, 21	90.28%
	4, 15, 28, 11, 6, 21, 14	91.02%
	3, 4, 28, 16, 21, 25, 6, 2, 24, 34, 27	91.23%
	4, 3, 15, 28, 25, 21, 11, 14, 26	90.34%
2（默认）	4, 28, 21, 27, 15, 24, 3, 26, 16, 6, 14, 2, 5, 8, 11, 13	92.10%
	4, 25, 21, 15, 16, 6, 26, 28, 13, 24, 3, 2, 30, 22, 27, 19	92.52%
	4, 3, 25, 21, 16, 15, 28, 26, 14, 6, 13	92.31%
1.5	4, 28, 3, 21, 24, 2, 15, 16, 37, 6, 25, 14, 30	91.86%
	28, 4, 16, 3, 21, 26, 6, 27, 5, 13, 24, 15, 2, 35, 29, 14, 25, 20, 34	92.02%
	4, 28, 21, 27, 11, 3, 15, 16, 6, 24, 25, 30, 10, 13, 26, 2, 31, 20, 8	92.64%
	4, 21, 3, 15, 28, 25, 2, 14, 24, 16, 6, 22, 27	91.98%
	4, 28, 16, 21, 11, 25, 3, 6, 24, 15, 30, 13, 14, 38, 29, 33	93.01%

It can be seen from the table that the subscript, the size of the selected feature sets is generally between 7 and 19,and the average size is 13,more than half of that of the original feature capacity 39;However,their corresponding test accuracy does not show a sharp decline compares with 92.16% before screening,both within 2% and equaling that before selecting after adjusting the parameters. First of all,the establishment of random forests is a stochastic process in itself,which makes the order of importance always change and affects the screening results.The improved feature selection algorithm uses the generalized-sequence backward selection to eliminate the feature,which is actually based on a greedy strategy and usually leads to local optimum as a result of neglecting the correlation between features.

4 Conclusion

By applying the large data technology to the prognosis prediction of primary liver cancer, we can find that the accuracy of the random forest algorithm is obviously superior to that of the decision tree algorithm. The decision tree algorithm cannot avoid a low accuracy caused by over-fitting even if it uses the pruning technique. The improved feature selection algorithm, on the contrast, can significantly reduce the

feature set on the premise of guaranteeing the prediction accuracy, which lays the foundation for considering the correlation between features later. It further shows that it is an essential trend to apply the increasingly perfect big data technology to the medical field, and it is worth further exploring and studying.

References

1. LIU Qian,WANG Wenqi.:Liver cancer.Beijing:People's Medical Publishing House,(2000).
2. Han Yu,Shi Hailong,Qu Bo.:Application of random forest method in medicine.Chinese Journal of preventive medicine.15(1),79-81(2014).
3. Chen Xiao,Wang Shubao,Li Jianjing.:Application of weighted constraint based decision tree method in identifying poor students. computer application and software Parts.32(12).136-139(2014).
4. Wang Xiaowei,Jiang Yuming.:Analysis and improvement of decision tree ID3 algorithm.computer engineering and design.32(9).3070-3072(2011).
5. MIAO Yufei,ZHANG Xiaohong.:Improvement and application of C4.5 decision tree algorithm.Computer Engineering and Applications.32(9).3070-3072(2011).
6. Zhou Zhihua.:Machine learning.Beijing: Tsinghua University press(2016).
7. Huai Ting Ting.:Improvement and application of random forest algorithm.Hang zhou.Metrology University of China.