Predicting Five-year Overall Survival in Patients with Non-small Cell Lung Cancer by ReliefF Algorithm and Random Forests

Xueyan Mei
School of Professional Studies
Columbia University
New York City, USA
Meixueyan836@gmail.com

***Abstract*— Non-small Cell Lung Cancer (NSCLC) is a leading death disease in many countries. Many studies are focusing on exact surgical approaches to treat the disease. The five-year overall survival rate for NSCLC patients is typically predicted by traditional regression models with small samples and data size. In this paper, we introduce machine learning tools with feature selection algorithms and random forests classifier to predict the five-year overall survival rate based on a large database. The results of this experiment show that our proposed framework is better than other machine learning approaches to predict the five-year overall survival rate.**

***Keywords*— ReliefF Algorithm; Feature Selection; Five-year Overall Survival; Non-small Cell Lung Cancer (NSCLC); Random Forests**

I. Introduction

Lung cancer is one of the most fatal diseases all over world [1]. Surgical resection is significant to offer the only curable option for most patients with non-metastatic non-small cell lung cancer (NSCLC) [2]. Study of the clinical data in predicting the five-year overall survival rates is essential for doctors and patients. However, previous studies on NSCLC surgery were small sample size as well as biased criteria. Due to the unbiasedness of the NSCLC database we used in our experiment [3], many features such as ID, sex, age, pathology, operation type, TNM stages, etc. are contained. However, redundant and unhelpful information might be provided by all presented features. Hence, it is necessary to choose useful features from branches of features in order to better predict the five-year overall survival rate.

The five-year overall survival is a two-class classification problem, for the target only has two labels. In our case, the two classes of the target are “yes” and “no”. The objective of our model is to predict whether the patient with NSCLC can live over five-years or not, based on feature selection algorithms and classification algorithms.

Feature selection is a key task for a classification problem. Since the raw data contains multiple features, it is necessary to select out the useful ones. Therefore, to find out an optimal combinatory of features should be addressed. Throughout many trials, we found out that the combination of the ReliefF [4] and random [5] forests will generate the best prediction of the five-year overall survival rate.

In this paper, we built a framework combining ReliefF algorithm and random forests. Fig 1. describes the detailed information of our proposed framework. We first input our raw dataset with all features. In step 1, we employ ReliefF algorithm to find an optimized combinatory of all features. The trained classifier obtained by step 2 will be used to predict the condition of the pulmonary nodule. The empirical results show that the combination of ReliefF algorithm and random forests can provide good performance in predicting the five-year overall survival for NSCLC patients.

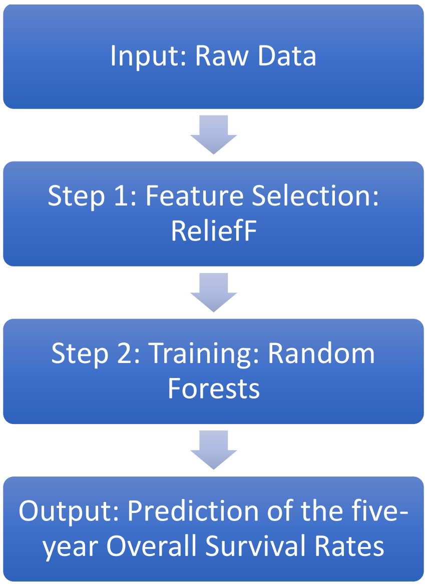
Input: Raw Data
Step 1: Feature Selection: ReliefF
Step 2: Training: Random Forests
Output: Prediction of the five-year Overall Survival Rates

Fig. 1. Proposed framework ReliefF Feature Selection and Random Forests

The rest of the paper are organized as follows. In section II, we reviewed previous work. In Section III, we present ReliefF feature selection algorithm. Section IV talks about random forests. Section V describes Experiments. A conclusion is drawn in section VI.

978-1-4673-8979-2/17/$31.00 ©2017 IEEE

II. Previous Work

A. Feature selection

Feature selection is a typical approach to solve problems that have many features. The purpose of feature selection is to obtain a small set of features that are relevant and provide necessary information from the original set. Kenji and Larry [6] proposed the Relief feature selection algorithm, which does not depend on heuristics and is accurate even if features interact, and is noise-tolerant. Kenji and Larry [4] proposed ReliefF algorithm, which can handle incomplete and noisy data and overcomes the limitations of two-class problems of the Relief algorithm. Kononenko et al. [7] proposed updated ReliefF algorithm by using the L1 norm and the absolute difference instead of the square of these differences. He et al. [8] demonstrated Laplacian feature selection that can solve both supervised and unsupervised problems, where a feature is evaluated by its power of locality preserving. Hall [9] introduced correlation-based feature selection that an apposite correlation measure and heuristic search strategy are paired by the feature evaluation formula.

B. Decision trees and random forests

Quinlan [10] proposed C4.5 algorithm to generate decision trees that can be used for classification. Breiman [11] suggested bagging predicators to generate many forms of a predictor in order to get an aggregated predictor, where the many forms are produced by making bootstrap replicates of the learning set and using these as new learning sets. Efron et al. [12] demonstrated a computer-based bootstrap method, which is a statistical inference approach that can solve many real-statistical problems without formulas. Breiman [5] proposed random forests, a blend of decision trees such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.

C. Non-small cell lung cancer

Alberti et al. [13] studied the overall survival in patients with non-small cell lung cancer from 52 randomized clinical trials, finding out that the cytotoxic chemotherapy may have a role in treating the disease. Kris et al. [14] proved that blocking EGFR tyrosine kinase with gefitinib led to symptom improvement and radiographic regression in patients with NSCLC. Mitsudomi et al. [15] interpreted that the selection by EGFR mutation will produce longer pregression-free survival for patients with lung cancer compared to those who receive gefitinib than cisplatin plus docetaxel.

III. ReliefF Feature Selection

Kira and Rendell [6] proposed Relief feature selection algorithm to be used for binary classification problems. Relief algorithm costs less computational time that can deal with continuous or two-class data as well as tolerating noises. The original purpose of Relief algorithm is to assess the quality of features based on values of features that are different from their neighbors. Therefore, the algorithm is aimed to seek for the two nearest neighbors from both the same and different class, which are called nearest hit and nearest miss respectively. The Relief algorithm evaluates the weights of features, but the weights are not stable since they could be fluctuated with observations due to the uncertainty and randomness of the observations. Additionally, the Relief feature selection algorithm is weak to handle incomplete data and multi-class problems.

To overcome the defects of the Relief algorithm, the ReliefF algorithm, an extension of the Relief algorithm, is designed to cope with incomplete and noisy data as well as solving multi-class problems. Like the Relief algorithm, the ReliefF algorithm chooses an instance r_i randomly and seeks for k of its nearest hits h_j and nearest misses $m_j(C)$ respectively. The quality estimation W(X) for all features X based on their values for r_i, h_j, and $m_j(C)$ is updated afterwards. The value for each class of the misses is weighted with the previous probability of that class P(*C*). To ensure that the values of hits and misses in each step is symmetric and between 0 and 1, the sum of probability of misses weights is equal to 1. Every probability weight is divided with parameter $1 - P(class(r_i))$ as a result of missing class of hits. The whole progress is repeated *m* times. The noise will be reduced due to the selection of *k* nearest hits and misses. The following describes the ReliefF algorithm:

Input: for each training observation a vector of attribute values and the class value

Output: the vector w of estimations of the qualities of attributes

```
Set all weights $w[X]$ = 0.0
For $i$= 1 to $m$
Randomly select an observation $r_i$
Find $k$-nearest hits $h_j$
For each class $C \neq\ class(r_i)$
Form class C find $k$-nearest misses $m_j(C)$
For $X$=1 to $a$
$W[X] = W[[X] - \sum_{j=1}^{k} \frac{diff(X,r_i,h_j)}{(m.k)} + \sum_{C \neq class\ r_i} \frac{\left[\frac{p(c)}{1-p(class(r_i))} \sum_{j=1}^{k} diff(a,r_i,h_j)\right]}{m.k}$
end
```

IV. Random Forests

A. Decision Tree

Decision tree classifier is a common supervised classification method in machine learning to build a predictive model by constructing a decision tree that maps instances of a case to its target values. Typically, observations are represented at the branches, where the classes are represented at the leaves. The goal of constructing a decision tree is to complete classification problems. Thus, a decision tree is a tree where an input feature is associated with each node, while each leaf is labeled with a class or a probability over the classes.

B. Random Forests

Ensemble learning methods use and associate with various learning algorithms to obtain improved predictive result than could be achieved by any single learning algorithm [16]. Two popular approaches for classification trees are boosting [12] and bagging [11]. Successive trees give additional weight to points that are wrongly predicted by previous predictors in boosting, while successive trees are independent on earlier trees by a bootstrap sample of the dataset in bagging.

Breiman [5] proposed random forests, where an extra level of randomness is added to bagging. A random forest is a training classifier comprised of a series of tree-structure classifiers $\{h(\boldsymbol{x}, \boldsymbol{\theta_k}), k = 1, \ldots\}$ where the $\boldsymbol{\theta_k}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\boldsymbol{x}$. Random forests are one of the ensemble learning methods to solve classification and regression problems , which construct multiple nodes decision trees when training for classification problems of the individual trees. Random forests could overcome the overfitting issues to the training set generated by decision trees.

Random features play an important role in terms of randomness in constructing random forests. Typically, the insertion of the randomness should diminish the correlation as well as taking advantages for the sake of accuracy. The forests are comprised by using randomly selected input or combinations of inputs at each node to grow each tree.

The random forests algorithm for both classification and regression can be described by the following steps [16]:

1. Draw n_{tree} bootstrap samples from the original data.

2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables.

3.Predict new data by aggregating the prediction of the n_{tree} trees.

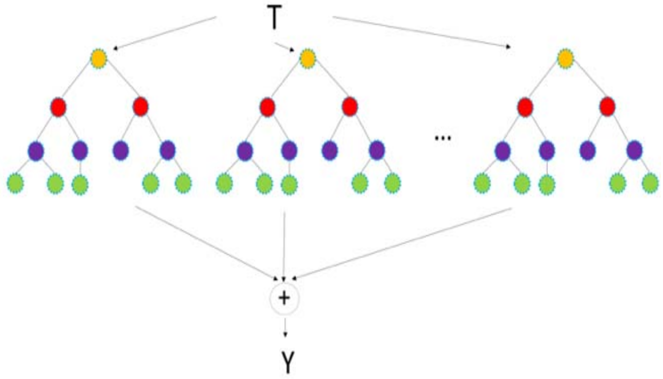

Fig 2. Description of random forests

V. Experiments

A. Description of the dataset

Surgical treatment of non-small cell lung cancer (NSCLC) creates concerns in recent years. The study of whether the patient can survive over five years is significant. The dataset we used was collected from multiple hospitals and areas in China containing various types of criteria [3]. Based on our proposed framework, we first used ReliefF feature selection algorithm to select out the optimal combinatory of features from the primary dataset. Detailed description of the selected features is expressed in Table I. The average of duration of operation is 150.95 ± 16.38 minutes, the average of drain age days is 5.48 ± 2.97 days, and the average of the length of hospitalization is 7.50 ± 2.97days.

TABLE I. Dedcription of Selected Features

Feature	**Description**
Duration of operation	100-180 mins
Operation type	1. lung lobe 2. total pneumonectomy 3. wedge-shape resection 4. sleeve resection
Pathology	1. squamous cell carcinoma 2. adenocarcinoma 3. adenosquamous carcinoma 4. large cell carcinoma 5. bronchioalveolar carcinoma 6. sarcoma pleomorphical carcinoma 7. others
Drain age days	0-50 days
Length of hospitalization	2-52 days
Status overall survival	1. Closed instance 2. Incomplete instance
Empyema	1. Yes 0. No
Chylothorax	1. Yes 0. No
Bronchopleural fistula	1. Yes 0. No
Wound infection	1. Yes 0. No

B. Comparison

We conduct experiments based on the framework we built on a realistic NSCLC dataset [3] to classify the five-year overall survival for patients. The data includes 5123 observations. Every observation is labeled with two classese, “yes” or “no”. The primary dataset contains 39 features, such as gender, length of hospitalization, pathology, operation type, TNM stages, and so on.

The original data is randomly separated into two subspaces, namely, training data (80% of observations) and testing data (20% of observations). The training part is intended for feature selection by ReliefF algorithm as well as random forest classifier training, while the testing data is used to evaluate the performance of the trained classifiers. The number of trees we used in the random forest is 60. We compare random forest with the decision tree method as a baseline method. As to feature selectors, we will also compare ReliefF algorithm with other commonly used feature selection algorithms, such as Laplacian algorithm [8], feature selection based on pairwise correlations [9], and local learning-based clustering [17]. The comparison of different combinations of classifiers and feature selectors is presented in Table II. We report the classification accuracies in the testing data and the numbers of selected feature subsets. As to our methods, the final selected features by ReliefF algorithm are (1) duration of operation, (2) optype2 (3) path (4) drain age days (5) length of hospitalization (6) status overall survival (7) empyema (8) chylothorax (9) bronchopleural fistula (10) wound infection. The compared results show that our framework generates higher predictive accuracies than other feature selection manners associated with decision tree and random forests. Our method selects out few useful and meaningful features that are highly relevant, which reduces the computational cost, meanwhile improving the predictive accuracies.

TABLE II. COMPARISON OF CLASSIFICATION PERFORMANCE IN TESTING DATA

classifier	Feature selector	# of features	Accuracy
Decision Tree	raw	36	0.68295
	laplacian	20	0.68295
	cfs	20	0.72183
	llcfs	20	0.69093
	ReliefF	10	0.78166
Random Forest	raw	36	0.79262
	laplacian	20	0.79362
	cfs	20	0.78863
	llcfs	20	0.78564
	ReliefF	**10**	**0.80259**

CONCLUSIONS

This paper proposes a framework to classify the five-year overall survival rate for NSCLC patients by random forests and ReliefF algorithm. We use the random forests in the five-year overall survival for NSCLC patients to classify whether a patient can live more than five years by learning random forests on a training data set. The projected method uses a ReliefF algorithm to find a combinatorial optimization of features before training classifiers. Feature selection is a necessary step to find out useful and meaningful features and get rid of redundant information of the primary dataset. The tests demonstrate our approach can produce higher classification accuracies compared to other feature selection algorithms associated with decision trees and random forests. The model we built achieves high classification accuracy as well as key representations of features.

REFERENCES

[1] A. Jemal A, et al., Cancer statistics, CA Cancer J Clin, vol. 60, pp: 277-300, 2010.

[2] National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in OncologyTM. Non-Small Cell Lung Cancer, vol. 2, 2013.

[3] W. Liang, et al., "Chinese multi-institutional registry (CMIR) for resected non-small cell lung cancer: survival analysis of 5,853 cases." Journal of thoracic disease, vol. 5, pp: 726-729, 2013.

[4] K. Kira, and L. Rendell, "A practical approach to feature selection," In Proceedings of the ninth international workshop on Machine learning, pp: 249-256, 1992.

[5] L. Breiman, “Random forests,” Machine learning, vol. 45, pp: 5-32, 2001.

[6] K. Kira, and L. Rendell, "The feature selection problem: Traditional methods and a new algorithm," AAAI, vol. 2. 1992.

[7] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, “Overcoming the myopia of inductive learning algorithms with RELIEFF,” Applied Intelligence, vol. 7, pp.39-55, 1997.

[8] X. He, D.Cai, and P. Niyogi, "Laplacian score for feature selection," In Advances in neural information processing systems, pp: 507-514, 2005.

[9] M. Hall, "Correlation-based feature selection for machine learning." PhD diss., The University of Waikato, 1999.

[10] J. Quinlan, “C4.5: Programs for Machine Learning,” Morgan Kaufmann Publishers. 1993.

[11] L. Breiman, "Bagging predictors." Machine learning, vol. 24, pp: 123-140, 1996.

[12] B. Efron, and R. Tibshirani, An introduction to the bootstrap, CRC press, 1994.

[13] W. Alberti, G. Anderson, A. Bartolucci, and D. Bell, "Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials," British Medical Journal,vol. 311, pp: 899, 1995.

[14] M. Kris, et al., "Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non–small cell lung cancer: a randomized trial," Jama, vol. 290, pp: 2149-2158, 2003.

[15] T. Mitsudomi, et al., "Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial," The lancet oncology, vol. 11, pp: 121-128, 2010.

[16] A. Liaw, and M. Wiener, "Classification and regression by randomForest," R news, vol. 2, pp: 18-22, 2002.

[17] Y. Sun, S. Todorovic, and S. Goodison, “Local-learning-based feature selection for high-dimensional data analysis,” IEEE transactions on pattern analysis and machine intelligence, vol. 32, pp: 1610-1626, 2010.