MACHINE LEARNING APPROACHES FOR PREDICTION OF LARYNGEAL CANCER BASED ON LABORATORY TEST RESULTS

By

SIRU LIU

A Thesis Submitted to The Honors College

In Partial Fulfillment of the Bachelors degree
With Honors in

Computer Science

THE UNIVERSITY OF ARIZONA

MAY 2017

Approved by:

Dr. Lingling An, Dr Richard T. Snodgrass

ABSTRACT

Laryngeal cancer is approximately the twentieth most common cancer in the world with more than 150,000 new cases diagnosed annually. Laryngeal cancer, a prognostic serious disease associated with high mortality, is one of the most debilitating forms of cancer. Despite advances in therapy and novel surgical and non-surgical approaches, early diagnosis remains the best predictor of survival. Although cancer classification using gene expression data analysis has recently emerged in the research field, little is known of the relationship between pathology report results and final clinical results. In reality, vocal cord polyps are a common benign lesion, having the same voice disorder symptom as early laryngeal cancer. In this project, we use several popular machine learning techniques (logistic regression, random forest, PCA, etc.) to develop relevant prediction models to classify vocal cord polyps and early laryngeal cancer. The data set contains 63 variables for 5,000 patients. The k-fold cross-validation methodology is used in model evaluation and comparison. We compare the results from each method and provide some helpful instructions to support physician diagnosis.

ACKNOWLEDGMENTS

Special thanks to my advisors Dr. Lingling An and Dr. Richard T. Snodgrass for giving me patient guidance and help. I greatly appreciate the valuable time they spent on my research.

I am grateful to researchers at the West China Hospital, who provided the high-quality data for my research. They made a great effort and without their help, this paper could not have been completed.

I am also grateful for my friends, who made me understand the meaning of friendship. At the same time, I thank my family for supporting me in every aspect of my life, so that I can complete this thesis.

TABLE OF CONTENTS

Chapter 1 INTRODUCTION

Laryngeal cancer is approximately the twentieth most common cancer in the world with more than 150,000 new cases diagnosed annually. Laryngeal cancer is a serious prognostic disease associated with high mortality, and is among the most debilitating forms of cancer. Despite advances in therapy and novel surgical and non-surgical approaches, early diagnosis remains the best predictor of survival. Although cancer classification by using gene expression data analysis has recently emerged in the research field, little is known about the relationship between results in pathology reports and the final clinical results. In reality, vocal cord polyps are common benign lesions, having the same voice disorder symptom with early laryngeal cancer. This thesis explores the possibility to predict first stage laryngeal cancer using laboratory tests results. The object of this research is to build a predictive framework that can be embedded into healthcare organizations' clinical decision support system.

Beginning with chapter 2, background information of laryngeal cancer and vocal cord polyps is briefly introduced. This section also contains an introduction to machine learning techniques we used in this thesis, including logistic regression, PCA, and random forest.

In chapter 3, data cleaning methods and the building process of predictive models based on machine learning methods are discussed. The dataset contains around 5000 records and 63 variables. Then, three models are evaluated and compared in terms of ROC and cross-validation. And then an explanation of these models is given based on clinical knowledge, finally it is demonstrated how these models could significantly help clinicians to distinguish laryngeal cancer from vocal cord polyps and diagnosis laryngeal cancer at the early stage.

Lastly, in chapter 4, the thesis and suggested the direction of future research are summarized.

Chapter 2 BACKGROUND

2.1 Background of Laryngeal Cancer Diagnosis and Predication

Laryngeal cancer is a worldwide common malignant tumor, which is the second highest incidence of head and neck cancers. Its incidence accounts for about 5.7% to 8.4% of human malignancies, and 40% of patients are already at Phase III or IV when diagnosed [1,2]. In developed countries the incidence of laryngeal cancer was 5.5 per 100,000 and the mortality rate was 2.4 per 100,000. The incidence of laryngeal cancer was 3.5 per 100,000 in the developing countries, with a mortality rate of 2.1 per 100,000[3]. With environmental pollution, and food safety problems, the incidence of laryngeal cancer has increased significantly and continues to increase year by year [4]. Additional reasons for the occurrence of laryngeal cancer are still unknown, as the causes could be a variety of simultaneous factors. Vocal cord polyp is a benign proliferative lesion that occurs in the superficial layer of the vocal cords and is also a special type of chronic laryngitis. Vocal cord polyps and laryngeal cancer, at early stages, both have the similar common clinical manifestations: hoarseness, pharyngeal discomfort. It is very difficult to do the early diagnosis. Therefore, vocal cord polyps have been selected as a control group in this topic. (Fig 1-1)

Vocal Cord Polyp

Laryngeal Cancer

Fig 1-1 Vocal Cord Polyp and Laryngeal Cancer

As the throat is the body's vital breathing and vocal organs, laryngeal cancer has become a serious disease affecting the quality of life. The early clinical symptoms of laryngeal cancer are complex and non-specific, easily misdiagnosed or missed by physicians. Once a physician misses the early diagnosis, the patient's treatment and quality of life would be adversely affected. At present, the diagnosis of laryngeal cancer mainly uses electronic fiber laryngeal mirror and throat tissue pathological examination. Because these methods are invasive examinations, they are not suitable as screening methods for early laryngeal cancer. Some articles reported the use of tumor biomarkers (such as carcinoembryonic antigen, tumor necrosis factor, etc.) for the early diagnosis and monitoring of laryngeal cancer, but because of its lack of specificity and sensitivity, the tumor biomarkers failed to be widely used clinically.

With the success of data mining on the success of influenza by Google, large data and data mining has become an important issue in the field of medical health research and content. Data mining technology in the early diagnosis and prediction of malignant tumors has been under rapid development. It has become a hot topic of international research and focus. Early diagnosis and treatment for reducing the risk of laryngeal cancer are critical. On the one hand, it can improve the survival rate of patients, and may also try to retain the throat pronunciation function, reduce postoperative complications, and reduce medical costs.

To make a large number of data into useful information, data mining and other statistical models are gradually becoming indispensable technology and tools. Logistic regression analysis is commonly used in the analysis of risk factors for malignant tumors [5]. Random forest is often used in clinical diagnosis of cancer and medical data analysis based on decision trees [6,7]. In this paper, logistic regression analysis, and PCA combined with random forest were used to establish the prediction model of vocal cord polyp and laryngeal carcinoma. The establishment of the model can assist in the analysis of laryngeal cancer and vocal cord polyp by using the differences in laboratory test results and providing reference and help for the early diagnosis and prediction of laryngeal cancer.

2.2 Machine Learning Introduction

2.2.1 Two-Class Logistic Regression Model

Logistic Regression is a special case of the generalized linear model, using a logistic function based on a linear regression method. It is widely used in two class and multi-class classification problems in biostatistics and epidemiology fields. In this thesis, the binary dependent variable is 0 or 1:

$$f(x) = \begin{cases} 0 & \text{Laryngeal Cancer} \\ 1 & \text{Vocal Cord Polyp} \end{cases}$$

The probability of y=1 is assumed π, therefore the probability of y=0 is 1-π.

The probability π can be regressed on covariate variables x through a logistic regression:

$$logit(\pi) = x'\beta$$

i.e., $$\frac{\pi}{1-\pi} = exp\{x'\beta\}$$

where β is a vector of regression coefficients.

In details, the logistic regression model can predict the odds of a result case according to the value of corresponding independent variables. Odds are the ratio of the probability of that result case happening and the probability of not happening.

$$odds = \frac{\pi}{1-\pi}$$

Odds ratio is the ratio of two odds, which is used in a comparison of two result case happening probabilities. Building binary Logistic Regression model:

$$ln\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Each regression coefficient β_i represents the change of ln odds ratio when the corresponding independent variable changes by one unit while the rest are kept the same.

2.2.2 Random Forest Model

As one of the most popular machine learning methods, the random forest model is robust when training size smaller than feature size and parallelizability. Random forest model is a set of decision trees, built by many subsets of data and taking the average values to improve overall model performance.

The main algorithm is as follows:

Sample *n* random batches with size K with replacement in the overall dataset:

$$(S_1, S_2, S_3, \ldots, S_n)$$

For each random batch, train decision trees are independent. On each node split, the Gini index or entropy is calculated to minimize error. For each new tree, prediction and regression results are made. If it is a classifying random forest model, we use a vote for each classifier. If it is a regression random forest model, we use average values of each tree. In this project, classifying random forest model was used.

Each decision tree is based on the tree model and greedy algorithm, which is the mapping of observation values to the target values. It can quickly classify different clusters, discover relationships between each cluster, and predict the future results. Decision models are included with regression trees (continuous target values) and classified trees (discrete target values). For each test or decision, it will cause different outcomes and multiple input branches. A decision tree uses flow-charts to represent the process of decision. The leaf nodes represent predicted classes labels; the internal nodes represent a test on the attributes and the predicted classes labels; the branches represent objects that satisfied both attributes.

Based on geometric aspects, each observation value in the training set is a point in the nth dimensional space. Building each decision tree is the process of partitioning n-dimensional spaces into some areas. Each zone represents a label class. When predicting by using the decision tree model, it will pass through different nodes in the tree according to the various values of an independent variable, finally arrive at the leaf node, which is class label [8].

The core part of the decision tree model is the growth problem and the pruning problem. Different algorithms have various methods to solve these problems, which both have advantages and shortages. The growth problem is how the decision tree chooses the optimal variables and the optimal thresholds in the top-down process while minimizing the depth of the tree to achieve fast classification. "Optimum" refers to the high purity of the data set through the decision tree, where "purity" has different definitions in different algorithms. A wide range of data purity representation methods is information gain and Gini impurity indicators.

The information gain is applied to the growth of ID3, C4.5, and C5.0 decision trees. The core is the information entropy in information theory. The information entropy is a method to measure the uncertainty. For a source that sends an indeterminate signal, it can be measured by the probability of occurrence of different signals. The greater the probability, the more the likelihood of occurrence and the less uncertainty. Therefore, Shannon defines the entropy of the random variable X according to Boltzmann's H-theorem and the probabilistic mass function P as follows:

$$H(X) = \sum_{i} P(x_i)I(x_i) = -\sum_{i} P(x_i)log_b P(x_i),$$

I (X) is self-information. In this paper, it is assumed that the sample data set to construct the decision tree model is D, the number of classes is c, the ratio of the ith sample to the total data set is P_i, and the information entropy before the definition is:

$$Info(D) = -\sum_{i}^{c} P(x_i)log_b P(x_i),$$

After selecting the variable A to be the branching variable of the decision tree, the information entropy of the sample data set D becomes $Info_A(D)$ where K represents the number of samples to be divided, as defined below:

$$Info_A(D) = -\sum_{j=1}^{k} \frac{|D_j|}{|D|} Info(D_j)$$

After the sample data set D passes through a variable A of the decision tree, the reduction of the information entropy is the information gain Gain (A) of the sample data set, defined as follows:

$$Gain(A) = Info(D) - Info_A(D)$$

Therefore, the optimal branching variable selected for the decision tree growth is the variable that makes the information gain Gain (A) obtain the maximum value, which means the dataset has the smallest uncertainty and the highest purity.

Gini impurity is the probability that an event becomes an opposing event, that is, the probability that a random sample is divided into an error subset, which is applied to CART (Classification and Regression Trees). The smaller the Gini purity, the higher the purity of the sample data set D. The number of classes is c, the ratio of the i-th class to the total data set is Pi. The definition of the original Gini impurity is as follows:

$$Gini(D) = 1 - \sum_{i}^{c} P_i$$

After selecting variable A to be the branching variable of the decision tree, the Gini impurity of the sample data set D becomes $Gini_A(D)$, and k represents the number of samples to be divided, defined as follows:

$$Gini_A(D) = \sum_{j=1}^{k} \frac{|D_j|}{|D|} Gini(D_j)$$

After the sample data set D passes through variable A of the decision tree, the reduction of Gini impurity is the sample data set's Gini impurity gain ΔGini(A), defined as follows:

$$\Delta Gini(A) = Gini(D) - Gini_A(D),$$

Therefore, the optimal branching variable selected for the decision tree growth is the variable that makes the Gini impurity gain ΔGini (A) obtain the maximum value. At that point, the data set has the smallest uncertainty and the highest purity.

In the process of building a decision tree, if the tree branches are too numerous or the depth is too deep, overfitting will occur, meaning while the training sample has a high precision, the accuracy of the test sample is very low. The overfitting situation can be repaired by pruning, including pre-pruning and post-pruning. Pre-pruning can be divided into depth threshold pre-pruning and entropy threshold pre-pruning. In practice, pre-pruning is often difficult to achieve; the depth threshold pre-pruning refers to the decision tree reaching the specified depth to stop the growth. The specified depth of the variable needs to have a clear understanding of the distribution of variables, as well as the need for repeated attempts under different parameters, selecting the optimal parameters. Information entropy increase rate is different from the growth of the decision tree growth; we meet the same difficulty to find the appropriate threshold. In this paper, we use the post-pruning method to construct a decision tree. We construct the decision tree at first, and then consider whether the entropy gain of the merged information is smaller than the entropy threshold for a group of nodes with the same parent node. If the merged information entropy gain is less than the threshold, then we do pruning. In this paper, the decision tree analysis model used the CART algorithm. The decision tree constructed by the CART algorithm is a binary tree, and the leaf node is a function rather than a specific category. Choosing the branching variable is based on the minimum Gini impurity gain. Suppose that there is a sample in a node t in the decision tree, the sample set is $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the average value of the node's dependent variable is

$$\bar{y} = \frac{1}{N(t)} \sum_{i=1, x_i \in t}^{N(t)} y_i,$$

Define its squared residuals in the t node:

$$SS(t) = \sum_{i=1, x_i \in t}^{N(t)} (y_i - \bar{y})^2,$$

According to the selection of the branching variable A which minimizes the Gini impurity gain, the left and right nodes are divided according to their attributes F:

$$\phi(S, F) = SS(t) - SS(t_L) - SS(t_R),$$

When Φ (S, F) is the largest, we have the optimal division F^*.

2.2.3 Principal Components Analysis

Principal components analysis is one of the most widely used statistical methods to reduce the dimension of the data by using projection. It is often used in Machine Learning's pipeline, working as a selecting feature. Assume our input dataset's dimensions is N, we prefer decreasing to D vectors while maximizing the variance of the projection data. Our dataset is $\{x_n\}, n = 1, \dots, N$

First, assume D =1 is fixed. In general, we can use the M dimensional vector v_1 to define the direction of that projection space. $v_1^T v_1 = 1$. Each data point x_n can be projected to a specific direction by using $v_1^T x_n$. The average value of the projection data is $v_1^T \bar{x}$, and $\bar{x}$ is the average value of the sample set.

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

The variance of the projection data is

$$\frac{1}{N} \sum_{n=1}^{N} \{v_1^T - v_1^T \bar{x}\}^2 = v_1^T S v_1$$

S is the covariance matrix,

$$S = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$$

Now, we can maximize the projection variance $v_1^T S v_1$. Obviously, the maximization process should meet certain constraints to prevent v_1 increasing to infinitely. We will use $v_1^T v_1 = 1$ to be our constraint. In order to satisfy this limit, we introduce Lagrangian multiplier λ_1, then we have

$$v_1^T S v_1 + \lambda_1 (1 - v_1^T v_1)$$

Setting its derivation to be zero, we have

$$S v_1 = \lambda_1 v_1$$

This indicates that v_1 is a characteristic vector of S. If we left multiply v_1^T, we have

$$v_1^T S v_1 = \lambda_1$$

So when we set v_1 to be the same as the eigenvector with the largest eigenvalue λ_1, the variance will reach the maximum. This eigenvector is called the first principal component.

2.2.4 Combined PCA and Random Forest Model

One of the most important applications of PCA is data pre-processing. In this case, our aim is not only to reduce the dimension. Instead, the data set is transformed to standardize certain properties. It is important for the subsequent successful application of pattern recognition algorithms to datasets. Often, when the original variable is measured using different units, or if the change is quite different, we will transform the data set as such.

In this paper, we will combine PCA model with the Random Forest Model to improve its performance.

Chapter 3 BUILDING PREDICATION MODEL AND ANALYSIS

3.1 Data Source and Data Cleaning

3.1.1 Data Source

In order to provide a better early detection and prediction of laryngeal cancer, this project uses laryngeal cancer patients and vocal cord polyp patients' EHR (Electric Health Record) data from West China Hospital from January 2013 to December 2016. West China Hospital is one of the biggest hospitals in the world and it is the medical center in Southwestern China. Laryngeal cancer patients and vocal cord polyp patients mostly come from the southwest part, giving the data some representativeness.

All the data are retrieved from Hospital Information System (HIS). Diagnosis code is based on the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), confirmed by experts in otolaryngology. We use this code as the label of our dataset. Using SQL to search information in EHRs, the cleaned, transformed data was loaded into the objective data warehouse to get a laryngeal cancer and vocal cord polyp dataset. That data warehouse has a total of 5,180 laryngeal cancer and vocal cord polyp patients' labs with 63 test results (blood, biochemistry, liver function, renal function).

3.1.2 Data Cleaning

In the original dataset, each lab test result has the collection date, although some of the lab results were missing values, affecting the building of the following models. Based on this consideration, a Python script was written to delete the collection date and convert the diagnosis results into corresponding two-classes numerical variables. 0 means vocal cord polyp and 1 means laryngeal cancer. Then the patients who have missing values in the 63 lab tests were filtered out. After the data cleaning, the new dataset had 3,769 patients' records, including 2,813 laryngeal cancer patients and 956 vocal cord polyp patients. The attributes are 63 numerical values in different lab tests.

3.2 Prediction Model Based on Logistic Regression

3.2.1 Building a predication model based on Logistic Regression

Two-classes logistic regression analysis is used to solve the problem that the dependent variable is a binary variable. The laryngeal cancer predication model is intended to classify patients with laryngeal cancer and vocal cord polyp through a large number of biochemical tests results. Therefore, this model results in binary dependent variable: “0” for vocal cord polyps, and “1” for laryngeal cancer based on the predicted probability. That is, the predicted status will be laryngeal cancer if the probability is greater than 0.5, otherwise, vocal cord polyps. The logistic analysis model was established using scikit-learn package in Python 3.5.

3.2.2 Result and Evaluation

Classification Table

Observed		Predicted		
		Diagnosis		Percentage Correct
		0	1	
Diagnosis	0	422	14	96.8
	1	47	83	63.8
Overall Percentage				89.22

Table 3-1. Classification Table (Logistic Regression Model)

The 10-fold cross validation accuracy is 86.44%. 422 of the 463 cases of vocal cord polyp were successfully predicted with a prediction accuracy rate of 96.8%, and 83 cases of 130 cases of laryngeal cancer were successfully predicted with a prediction accuracy rate of 63.8%.

Summary of the final results from logistic regression model can be found in the table below:

	B	S.E.	Wald	df	Sig.	Exp(B)
RDWCV	1.865	0.671	7.717	1	0.005	6.453
NEUT	-12.339	2.205	31.306	1	0.000	0.000
EO#	-2.191	0.786	7.773	1	0.005	0.112
LYMPH	-15.056	2.045	54.212	1	0.000	0.000
HBEAG	-7.052	6.028	1.369	1	0.242	0.001
HBCAB	-1.501	0.459	10.695	1	0.001	0.223
HBSAB	-0.625	0.205	9.286	1	0.002	0.535
CO2CP	3.955	0.696	32.264	1	0.000	52.212
LDLC	4.532	0.662	46.849	1	0.000	92.925
TTR	3.735	0.764	23.898	1	0.000	41.901
URIC	-1.695	0.520	10.633	1	0.001	0.184
TBIL	-13.306	4.458	8.909	1	0.003	0.000
TP	-2.183	0.813	7.215	1	0.007	0.113
CL	-1.731	0.883	3.847	1	0.050	0.177
APTT	-43.943	20.715	4.500	1	0.034	0.000
ALB	-6.345	0.835	57.735	1	0.000	0.002
HCT	9.449	15.829	0.356	1	0.551	12689.872
RBC	-3.117	0.636	23.986	1	0.000	0.044
FIB	6.041	0.627	92.885	1	0.000	420.498
HBDH	-5.508	1.958	7.916	1	0.005	0.004
CREA	5.202	1.982	6.887	1	0.009	181.624
GLU	4.991	1.061	22.137	1	0.000	147.123
PLT	2.616	0.712	13.517	1	0.000	13.685
CA	-3.237	1.507	4.616	1	0.032	0.039
AG	3.433	0.926	13.728	1	0.000	30.960
Constant	10.235	2.407	18.085	1	0.000	27853.238

Table 3-2. Variables in the Final Equation

It is clear that the model is only limited by analyzing the correctness of the forecast. For example, one probability of prediction is 0.6 and another probability of prediction is 0.9. These two cases are attributed to the same class when calculating the correctness of prediction of probability> 0.5, but they clearly have different meaning in statistics. To avoid the large amount of information lost during this process, the ROC curve method is used for representing the model's prediction probability. ROC curve (Receiver Operating Characteristic Curve) is an analysis tool based on coordinate diagram first applied to the signal detection theory to select the best signal detection model, or in the same model set the best threshold. When the prediction effect is best, the ROC curve will rise vertically from the lower left corner to the top and then horizontally to the right. On the contrary, if the ROC curve coincides with the main diagonal, then the correct and false classification rate is 50% with no predictive value. If the ROC is between those two lines, then the prediction model has a certain reliability. The ROC plot by logistic regression model results is shown in Figure. 3-1.

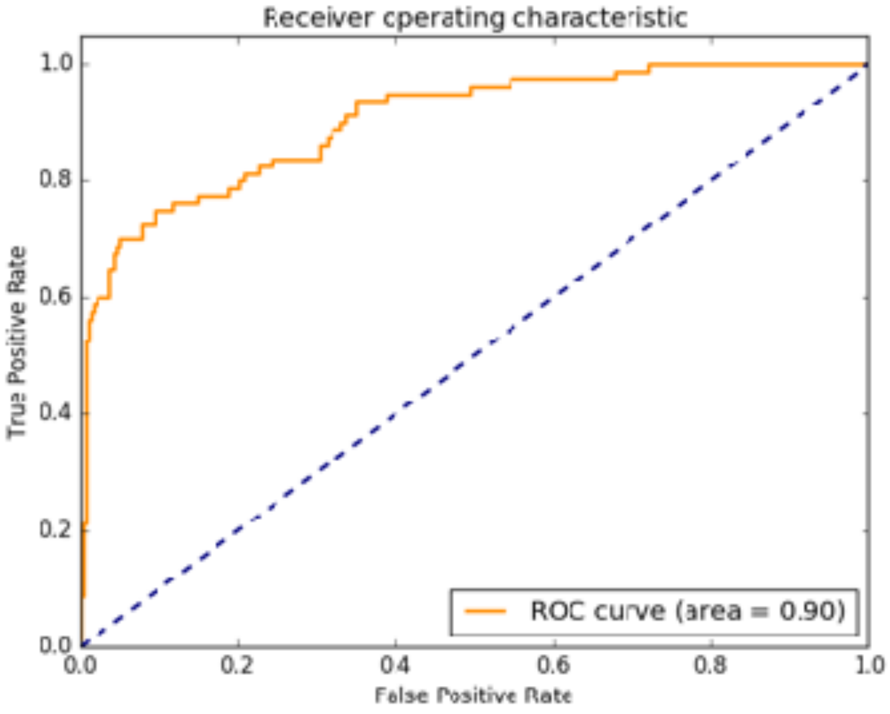


Figure. 3-1 ROC plot from logistic regression results

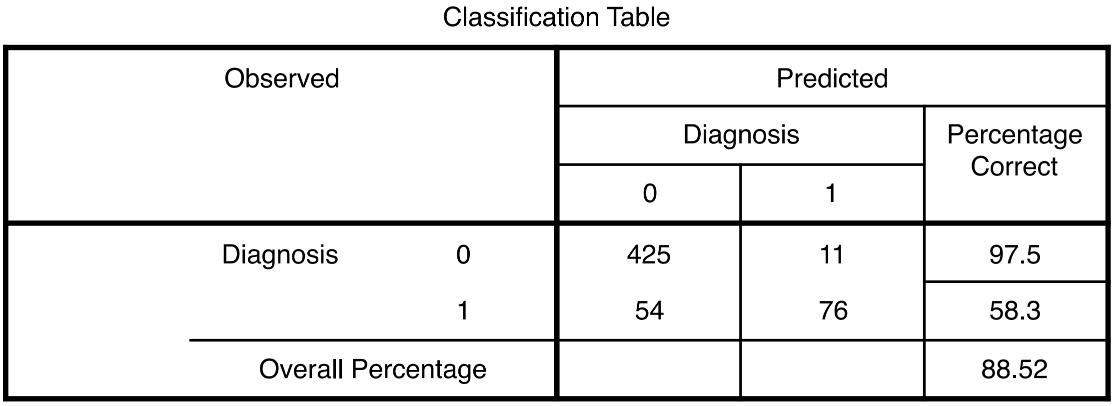
Consider the area under the ROC curve (AUC), because ROC must exist in a 1*1 grid, the value of AUC should be 0~1. If a threshold is determined, then the sample with the predicted probability below the threshold is defined as the negative sample, and the sample with the predicted probability above the threshold is defined as the positive sample. Both the negative sample and the positive sample are randomly selected from their corresponding groups. The value of AUC is equal to the probability that the positive sample's predicated value is higher than the negative sample's predicated value. Therefore, the larger the AUC value, the more accurate the classification model is. Logistic Regression Model's AUC value is 0.9, which shows this model can make good predications.

3.3 Prediction Model Based on Random Forest Model

3.3.1 Building a predication model based on Random Forest Model

First, we only use the Random Forest method to build the predication model. The laryngeal cancer predication model is intended to classify patients with laryngeal cancer and vocal cord polyp through a large number of biochemical tests results. Therefore, this model uses bivariate dependent variables, "0" for vocal cord polyps, and "1" for laryngeal cancer. The model has 1000 trees with the maximum depth at 100 and established using the RandomForestClassifer package in sklearn in Python3.5.

3.3.2 Result and Evaluation



Classification Table

Observed			Predicted		
			Diagnosis		Percentage Correct
			0	1	
	Diagnosis	0	425	11	97.5
		1	54	76	58.3
	Overall Percentage				88.52

Table 3-3. Classification Table (Random Forest Model)

Its 10-fold cross validation accuracy is 86.34%. 425 of the 463 cases of vocal cord polyp were successfully predicted, the prediction accuracy rate was 97.5%, and 83 cases of 130 cases of laryngeal cancer were successfully predicted, the prediction accuracy rate was 53.8%.

The ROC plot by random forest results is shown in Figure. 3-2, and its AUC value is 0.85.

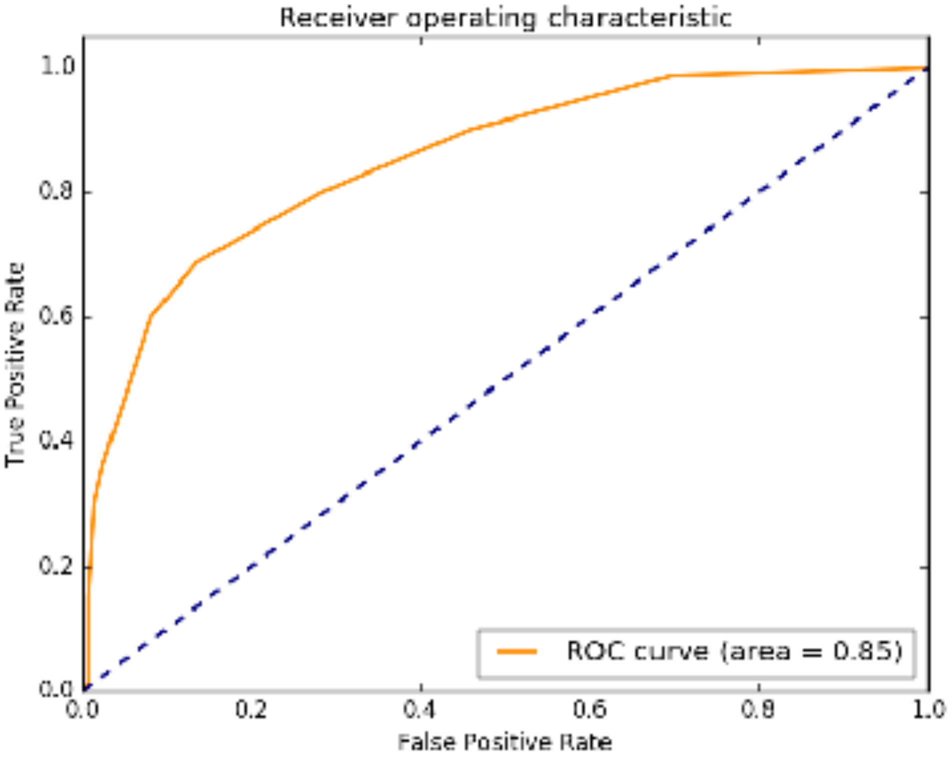


Figure. 3-2 ROC plot from random forest results

3.4 Predication model based on PCA and Random Forest Model

3.4.1 Building a predication model based on PCA and Random Forest Model

We consider combining PCA with Random Forest Model, using PCA to reduce the number of dimensions and transform the dataset to standardize some properties.

The following graph draws the trend of the number of components and its related explained variance. It demonstrates that the last 20 components explained very low variance (<0.5), which means they make a smaller contribution to the model. Thus, the first 40 components were used as new variables training into the random forest model.

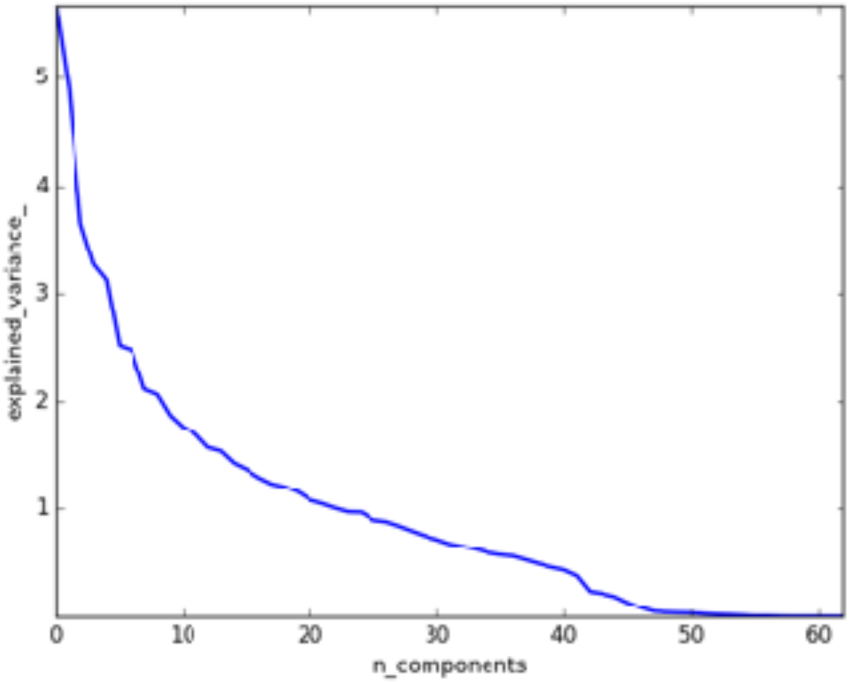


Figure. 3-3 PCA components

3.4.2 Result and Evaluation

Classification Table

Observed			Predicted		
			Diagnosis		Percentage Correct
			0	1	
	Diagnosis	0	416	20	95.40
		1	49	81	62.30
	Overall Percentage				87.80

Table 3-4. Classification Table (PCA and Random Forest Model)

The 10-fold cross validation accuracy is 85.43%. 416 of the 436 cases of vocal cord polyp were successfully predicted, the prediction accuracy rate was 95.40%, and 81 cases of 130 cases of laryngeal cancer were successfully predicted, the prediction accuracy rate was 62.30%.

3.5 Comparison of Prediction Models

In the binary logistic regression model, we finally have the formula to calculate the probability of being diagnosed with laryngeal cancer. It is included RDWCV, NEUT, etc. Those 25 laboratory test results are significantly related to laryngeal cancer. Based on our knowledge and after searching articles on Medline, we did not find any known relationship between these laboratory test and laryngeal cancer. Thus, the logistic regression model not only has the highest accuracy, but it also provides some new direction on studying these unknown relationships. This model also has a shortage: its correct percentage on classifying observed laryngeal cancer patients with predicted laryngeal cancer patients is not very high.

About the PCA followed by Random Forest model, its accuracy is 87.80%, a bit lower than the accuracy of only using Random Forest model. However, it has a better performance on classifying the laryngeal cancer patients.

Chapter 4 CONCLUSION AND FUTURE DIRECTION

With the development of electronic health records and medical information, machine learning techniques have become a hotspot in the field of medical and healthcare research. Using informatics technology in the medical, health area may greatly change and subvert the traditional medical and healthcare. It may provide new models and methods for the early diagnosis of disease, treatment, and prevention, ultimately to improve human health. After more than one year of research, we have studied the method of machine learning methods in the early diagnosis and prediction of laryngeal cancer and opened up the ideas and fields of clinical informatics research, In summary, the thesis mainly made the following research results:

1. Built a logistics regression model to predict the laryngeal cancer, which has 89.22% accuracy. Variables occurring in the regression formula can provide useful information to help physicians make better diagnosis.

2. Built a predication model that use PCA method followed by Random Forest method. It has 87.60% accuracy, but has a better performance on classifying laryngeal cancer (improving accuracy 58.3% to 62.30%).

Due to factors of ability, time and sources, there are some places in this study that require further research. In the future, we need to explore more on the following areas:

1. Increase the size of the dataset, optimizing the relevant random forest model, making it more suitable for the prediction of laryngeal cancer patients, and improving the accuracy and performance..

2. Add the text mining from laryngeal cancer patients' progress notes, combine with natural language processing[10].

APPENDIX A: VARIABLE INFORMATION

VARIABLE	NAME (ENGLISH)
RDWCV	RBC DISTRIBUTION WIDTH CV
RDWSD	RBC DISTRIBUTION WIDTH SD
NEUT	NEUTROPHILS PERCENTAGE
NEUT#	NEUTROPHILS ABSOLUTE VALUES
MONO#	MONONUCLEAR CELLS ABSOLUTE VALUE
MONO	MONONUCLEAR CELLS PERCENTAGE
BASO	BASOPHIL PERCENTAGE
BASO#	ABSOLUTE VALUE OF BASOPHILS
EO	EOSINOPHIL PERCENTAGE
EO#	EOSINOPHIL ABSOLUTE VALUE
MCH	AVERAGE ERYTHROCYTE HGB CONTENT
MCHC	MEAN ERYTHROCYTE HGB CONCENTRATION
LYMPH	LYMPHOCYTE PERCENTAGE
LYMPH#	LYMPHOCYTES ABSOLUTE VALUE
AA	ANTI-SSA ANTIBODY
ALT	ALANINE AMINOTRANSFERASE
HBEAB	HEPATITIS B E ANTIBODY SEMI-QUANTITATIVE
HBEAG	HEPATITIS B E ANTIGEN SEMI - QUANTITATIVE
HBCAB	HEPATITIS B CORE ANTIBODY SEMI - QUANTITATIVE
HBSAB	HEPATITIS B SURFACE ANTIBODY QUANTIFICATION
HBSAG	HEPATITIS B SURFACE ANTIGEN SEMI - QUANTITATIVE
LDH	LACTATE DEHYDROGENASE
CO2CP	CARBON DIOXIDE BINDING CAPACITY
LDLC	LOW-DENSITY LIPOPROTEIN
PT	PROTHROMBIN TIME

VARIABLE	NAME (ENGLISH)
TT	THROMBIN TIME
TTR	THROMBIN TIME RATIO
INR	INTERNATIONAL STANDARDIZED RATIO
BUN	UREA
URIC	URIC ACID
MCV	AVERAGE ERYTHROCYTE VOLUME
TBIL	TOTAL BILIRUBIN
TP	TOTAL PROTEIN
CL	CHLORINE
APTT	ACTIVATED PARTIAL THROMBOPLASTIN TIME
APTTR	ACTIVATION OF PARTIAL THROMBOPLASTIN TIME RATIO
GLB	GLOBULIN
AG_A	ANIONIC GAP
WBC	WHITE BLOOD CELL COUNT
ALB	24 HOURS URINARY ALBUMIN
DBIL	DIRECT BILIRUBIN
ALP	ALKALINE PHOSPHATASE
HCT	HEMATOCRIT
RBC	RED BLOOD CELLS
FIB	FIBRINOGEN
HBDH	HYDROXYBUTYRATE DEHYDROGENASE
CREA	CREATININE
CK	CREATINE KINASE
CHOL	CHOLESTEROL
GLU	BLOOD SUGAR
PLT	PLATELET COUNT
PO4	SERUM INORGANIC PHOSPHORUS

VARIABLE	NAME (ENGLISH)
CysC	DETERMINATION OF SERUM CYSTATIN C
HGB	HEMOGLOBIN
GGT	GLUTAMYL TRANSPEPTIDASE
CA	CALCIUM
Na	SODIUM
K	POTASSIUM
MG	MAGNESIUM
AST	AST / ALT
IBIL	INDIRECT BILIRUBIN
AG	ANIONIC GAP
HDLC	HIGH DENSITY LIPOPROTEIN
Diagnosis	

APPENDIX B: PYTHON CODE FOR LOGISTIC REGRESSION APPROACH

logReg.py

```
'''
Data Source: "../cancerD/cancer_Com_Var_Eng.csv" "../cancerD/
nonCancer_Com_Var_Eng.csv" Each file has 100 datapoints.
'''
import numpy as np

from sklearn.metrics import confusion_matrix
from sklearn.utils import shuffle
from sklearn import linear_model
from sklearn.linear_model import ElasticNet

from data_handle import load_dataset
from CV import CV
from draw_roc import draw_roc

file_cancer = "../cancerD/cancer_Com_Var_Eng_clean.csv"
file_nonCancer = "../cancerD/nonCancer_Com_Var_Eng_clean.csv"

X1, y1, featureName = load_dataset(file_nonCancer,0)
X2, y2, featureName =load_dataset(file_cancer,1)
X, y = shuffle(np.row_stack((X1,X2)), np.append(y1,y2),
random_state =7)

num_training = int(0.8 * len(X))
X_train, y_train = X[:num_training], y[:num_training]
X_test, y_test = X[num_training:], y[num_training:]

#  The C parameter controls the regularization strength.
#  A lower value indicates higher regularization strength
classifier =
linear_model.LogisticRegression(solver='liblinear', C=1)
y_score=classifier.fit(X_train,
y_train).decision_function(X_test)
```

```
y_pred_lr = classifier.predict(X_test)


accuracy = 100.0 * np.sum(y_test == y_pred_lr)/
X_test.shape[0]
print ("\n#### Logistic regressor performance ####")
print ("Accuracy of the classifier =", round(accuracy, 2),
"%")
CV(classifier,X,y)
print(confusion_matrix(y_test, y_pred_lr))
draw_roc(y_test, y_score)
```

draw_roc.py

```
import matplotlib
matplotlib.use('TkAgg')
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc
# Compute ROC curve and ROC area for each class

def draw_roc(y_test, y_score):
    fpr, tpr, thresholds = roc_curve(y_test, y_score)
    roc_auc = auc(fpr, tpr)


    plt.figure()
    lw = 2
    plt.plot(fpr, tpr, color='darkorange',lw=lw, label='ROC
curve (area = %0.2f)' % roc_auc)
    plt.plot([0, 1], [0, 1], color='navy', lw=lw,
linestyle='--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('Receiver operating characteristic')
    plt.legend(loc="lower right")
    plt.show()
```

APPENDIX C: PYTHON CODE FOR PCA AND RANDOM FOREST MODEL

PCA.py

```
import numpy as np
from sklearn.decomposition import PCA
from data_handle import load_dataset
from sklearn.utils import shuffle
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import
explained_variance_score,confusion_matrix
from CV import CV
from draw_roc import draw_roc


file_cancer = "../cancerD/cancer_Com_Var_Eng_clean.csv"
file_nonCancer = "../cancerD/nonCancer_Com_Var_Eng_clean.csv"
X = []
y = []

X1, y1, featureName = load_dataset(file_nonCancer,0)
X2, y2, featureName =load_dataset(file_cancer,1)
X, y = shuffle(np.row_stack((X1,X2)), np.append(y1,y2),
random_state =7)

scaler = StandardScaler()
scaler.fit(X)
X = scaler.transform(X)

num_training = int(0.90 * len(X))

X_train, y_train = X[:num_training], y[:num_training]
X_test, y_test = X[num_training:], y[num_training:]

pca = PCA(n_components=60)
pca.fit(X)
```

```
'''
plt.figure()
plt.plot(pca.explained_variance_, linewidth=2)
plt.axis('tight')
plt.xlabel('n_components')
plt.ylabel('explained_variance_')
plt.show()
'''


X = pca.transform(X)

X_train, y_train = X[:num_training], y[:num_training]
X_test, y_test = X[num_training:], y[num_training:]


classifier_rf = RandomForestClassifier(max_depth=1000,
n_estimators=10)
classifier_rf.fit(X_train, y_train)

y_pred_rf = classifier_rf.predict(X_test)
accuracy = 100.0 * np.sum(y_test == y_pred_rf)/
X_test.shape[0]
print ("\n#### Random Forest regressor performance ####")
print ("Accuracy of the classifier =", round(accuracy, 2),
"%")

print(confusion_matrix(y_test, y_pred_rf))
CV(classifier_rf,X,y)
print(classifier_rf.decision_path(X_test))
y_score= classifier_rf.predict_proba(X_test)[:, 1]

draw_roc(y_test, y_score)

'''
from sklearn import tree

i_tree = 0
for tree_in_forest in classifier_rf.estimators_:
    with open('tree_' + str(i_tree) + '.dot', 'w') as my_file:
        my_file = tree.export_graphviz(tree_in_forest,
out_file = my_file)
    i_tree = i_tree + 1
‘''
```

BIBLIOGRAPHY

[1] Pfister, D. G., Laurie, S. A., & Weinstein, G. S.American Society of Clinical Oncology clinical practice guideline for the use of larynx preservation strategies in the treatment of laryngeal cancer[J].*J Clin Oncol*, 2006, 24(22): 3693-3704.

[2] Hao, J., Zhao, P., &Chen, W.Q. *2011 China Cancer Registration Annual Report*[M]. Version 1. 2011:94-129.

[3] Ferlay, J., Shin, H.R., &Bray, F. Estimates of worldwide burden of cancer in 2008:GLOBOCAN 2008. *Int J Cancer*[J].2010,127(12):2893-917.

[4] Du, L.B., Mao, W.M., &Chen, W.Q. Analysis of incidence and mortality of laryngeal carcinoma in China from 2003 to2007. *Journal of Epidemiology* [J]. 2012, 33(4):395-398.

[5] Qiu, H., Zhang, Y., &Lei, H.K. Logistic Regression Analysis of Malignant Tumor Risk Factors. *Journal of Chongqing Medical University* [J]. 2012,37(3): 251-255.

[6] He, Q.D., Wei, X.L., &Zhang, H.Q. The value of decision tree model based on "optimal tumor marker group" for the diagnosis of lung cancer. *Journal of Zhengzhou University (Medical Science Edition)* [J]. 2014(1):37-40.

[7] Liu, K. Medical Data Analysis Based on Decision Tree. [J].*Hefei University of Technology*, 2001,28(2):41-43.

[8] Kurt, I., Ture, M., & Kurum, A. T, “Comparing performances of logistic regression, classification and regression tree, and neural networks for

predicting coronary artery disease,” *Expert. Syst. Appl.*, vol. 34, no. 1, pp. 366–374, 2008.

[9] Bishop, C. M. (2013). *Pattern recognition and machine learning*. New Delhi: Springer.

[10] Ma, L., & Zhang, Y. (2015). Using Word2Vec to process big text data. *2015 IEEE International Conference on Big Data (Big Data)*. doi:10.1109/bigdata.2015.7364114

[11] Rogers, S., & Girolami, M. (2017). *A first course in machine learning*. Boca Raton: CRC Press.

[12] Baiju R. Shah, Lorraine L. Lipscombe, Clinical Diabetes Research Using Data Mining: A Canadian Perspective, *Canadian Journal of Diabetes*, Volume 39, Issue 3, June 2015, Pages 235-238, ISSN 1499-267.

[13] Cameron D. Wright, John C. Kucharczuk, Sean M. O'Brien, Joshua D. Grab, Mark S. Allen, Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: A Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model, The Journal of Thoracic and Cardiovascular Surgery, Volume 137, Issue 3, March 2009, Pages 587-596, ISSN 0022-5223.