COMMENTARY

Lung Cancer Risk Prediction Method Based on Feature Selection and Artificial Neural Network

Nan-Nan Xie*, Liang Hu, Tai-Hui Li

Abstract

A method to predict the risk of lung cancer is proposed, based on two feature selection algorithms: Fisher and ReliefF, and BP Neural Networks. An appropriate quantity of risk factors was chosen for lung cancer risk prediction. The process featured two steps, firstly choosing the risk factors by combining two feature selection algorithms, then providing the predictive value by neural network. Based on the method framework, an algorithm LCRP (lung cancer risk prediction) is presented, to reduce the amount of risk factors collected in practical applications. The proposed method is suitable for health monitoring and self-testing. Experiments showed it can actually provide satisfactory accuracy under low dimensions of risk factors.

Keywords: Lung cancer prediction - feature selection - artificial neural network

Asian Pac J Cancer Prev, **15 (23)**, 10539-10542

Introduction

Lung cancer is one of the most serious cancers, and in recent years, the mortality rate has a significant growth. There are many researches on the reasons of lung caner, but have no generally accepted results. Such like smoking is recognized as one of the risk factors of lung cancer, studies have shown that passive smoking, air pollution, household smoke from solid fuels, alcohol drinking, and cancer genetic history are also related to lung cancer.

For risk factors study in lung cancer, current researches mostly based on the clinical diagnostic statistical analysis. For example, the related factors of radiation pneumonitis in locally advanced NSCLC patients treated with radiotherapy was analyzed (Feng et al., 2014). Research by Su Qiang (Su et al., 2014) investigated the risk factors of pulmonary infection with advanced NSCLC patients, and the patients' information were collected form 2011 to 2012 in oncology department of the hospital. In 2011, the research(Huang et al., 2011) used case-control study, 781 cases and 781controls matched for age and gender were recruited, constructed Decision Trees and unconditional Logistic Regression models, in order to explore the influencing factors of lung cancer and their interaction effects.

There are few researches about the cancer prediction. The research (Thunyarat et al., 2014) developed and validated a breast cancer risk prediction model for Thai women, and concluded that "age, menopausal status, body mass index, use of oral contraceptive" were significantly associated with breast cancer. About Lung cancer, the paper (Kawsar et al., 2013) proposed an easy, cost effective and time saving lung cancer risk prediction system, they used K-means clustering algorithm for identifying relevant and non-relevant data.

In current studies, the exact relationship between risk factors and lung cancer is not clear, thus the prediction of cancer risk based on the risk factors is complex. In this paper, we use the computer algorithms for predicting the risk of lung cancer. A method of lung cancer risk prediction is proposed and the correspond algorithm LCRP (Lung Cancer Risk Prediction) is implemented. The differences between the above works and ours are in two aspects, the first is the reduction of dimensions of risk factors is based on feature selection algorithm, and to the second, the proposed method in this paper has scalability, which can be used in other cancers. Experiments on the actual dataset show the satisfied accuracy. Although the proposed method cannot directly used in diagnosis, it can be used in health monitoring and self-testing.

Feature Selection and BP Neural Network

Feature selection

Feature Selection (Wang et al., 2005) is one of the hot topics in pattern recognition. From the perspective of classification, pattern recognition is to classify objects. For the data samples, firstly design classifier with training data, and then use the classifier to identify the classification data to help decision making. In the training process, the classifier training is based on the similarity of samples, and features is the key to evaluate similarity. In some practical applications, features are not easy to collect, or difficult to measure, which make it difficult to construct precise classifiers. Generally, features can be divided into key features and secondary ones, for example, smoking is often

*Computer Science and Technology College, Jilin University, Changchun, China *For correspondence: xienn1113@163.com*

considered as one of the key risk factors in lung cancer.

The purpose of feature selection is to select optimized feature subsets, and consists of two steps: the first have to decide the search strategy, and the second is determine the evaluation criteria, which is used to evaluate the efficiency of selected features. From the different implementation of the two steps, feature selection methods can be classified.

According to the search strategy of feature selection algorithms can be divided into global search strategy, random search strategy, and heuristic search strategy (Mao et al., 2007). The typical global search algorithm is "Branch and Bound", the representatives of random search algorithms are combining feature selection with "Simulated Annealing", "Tabu Search", or "Genetic Algorithm". As to the typical heuristic search algorithms, including "Sequential Forward Selection", "Sequential Backward Selection", "Plus-L Minus-R Selection", and so on.

According to the feature sets evaluation strategy, feature selection methods can be divided into the Filter method and Wrapper method. Filter uses appropriate filter criteria to quickly evaluate the features, and wrapper combines the process of feature selection with classifier results, the selected feature evaluation based on the classifier's performance.

Fisher feature selection (Wang et al., 2007) is based on the Fisher criteria, and based on this principle: the features have great ability to identify have the small intra-class distance and the large inter-class distance, otherwise the feature is not appropriate for classification. Using single feature's fisher ratio as criterion, sorting the features, ans select the better features for identification, in order to reduce the feature dimension and get better identification performance, is the goal of fisher feature selection.

ReliefF (Huang et al., 2012) select features by statistical methods, based on distance measurements which developed from Relief algortithm. "The traditional Relief selects an sample object R by random sampling, and calculate two nearest neighbors: similar nearest neighbor "Nearest Hit (H)" and nearest different types neighbor "Nearest Miss (M)". For the feature i, diff (i, R, H) and diff (i, R, M) indicts the distance of sample and H and M. If iff (i, R, H)<diff (i, R, M), the feature i is considered as advantageous to the classification, and increase its weight, otherwise, decrease the weight. Finally, repeat the process above, and calculate the average of the iteration results. In order to deal with multi-class problem, ReliefF selects neighbors in every class, and provide a new formula to calculate weights.

Recent years, feature selection have rapid development, which widely used in text processing, gene analysis, drug diagnosis and so on. In lung cancer data, the risk factors are the features referred in feature selection. We use the two typical feature selection algorithms, Fisher and ReliefF, to reduce the risk factors dimensions of lung cancer data, in order to reduce the complexity of prediction.

BP neural network

In 1986, Rumelhart et al. proposed BP Neural Network, a multi-layer feed forward neural network algorithm based on error back propagation, which is one of the most widely used neural network models (Jin et al., 1999). The basic principle is to transform the input vector through hidden layer, resulting in an output vector, build mapping between input and output. The input data propagate forward, and error propagate back, modifying the weight by the feedback error. The topology of BP neural network is shown in Figure 1.

The key of BP neural network is hidden layer, mainly on the processing of extracting features. The forward propagation of BP is from input layer, through hidden layer, and reach output layer. When passing the hidden layer, neurons in one layer only affect the neighboring next layer. If the difference between the desired and actual output is large, then go to the error back-propagation, which need to define error function, to adjust the neurons' weight. There are some applications about BP neural network in medical fields. For example, "Application on predicting of cancer mortality (Zeng et al., 1996) ", but it is different with the method proposed in this paper.

Lung Cancer Risk Prediction Method

Method framework

We use the two feature selection algorithms and BP neural network into lung cancer risk prediction, propose the method framework, and realize the algorithm LCRP (Lung Cancer Risk Prediction Algorithm). The principle of the method is mainly divided into two steps: Firstly using both Fisher and ReliefF to sort risk factors. In order to avoid sorting deviation from separate algorithm, we combine these two algorithms to get the final result. There BP neural network is first used to decide which features finally selected. The second step is classification, and BP neural network is used again to give the value of the predict lang cancer risk. In the aspect of implementation, the method framework as in Figure 2.

i) Data Preprocessing. Including data cleaning, feature processing, and numeralization. Data cleaning is dealing with the incomplete, fault and repeat data. Feature processing refers to separate features, process features without enough effective values, and combine risk factors. Numeralization is to disperse non-numerical data, and process zero values.

ii) Feature Selection. Firstly using Fisher and ReliefF independently to sorting the risk factors, then combine

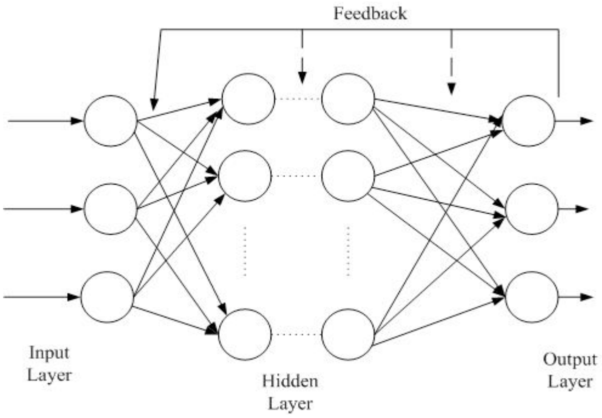

Figure 1. BP Neural Network Topology

the two results to gain a final sorting, and last use BP neural network to determine the risk factors used in the next prediction step.

iii) Feature Selection. Firstly using Fisher and ReliefF independently to sorting the risk factors, then combine the two results to gain a final sorting, and last use BP neural

Lung cancer risk predication algorithm (LCRP)

We describe the LCRP realization process with four steps, including Fisher sorting, ReliefF sorting, define the final features, and the risk prediction. We use the pseudo-code and text description to describe the algorithm.

Fisher sorting:

Step1: Input data set, calculate the number of line and column, initialize the zero output matrix out.W.

Step2: Calculate the average value of feature i (i from 1 to feature numbers n).

Step3: For the class j (j from 1 to class number m), temp1 calculate the variance between the average value of all classes under this feature and the average of all samples, iteration and add all as the inter-class variance. Temp2 calculated the sum of variances of all classes as intra-class variance.

Step3: If temp1=0, it indicates the feature has no distinction degree, out.W(i)=1; if temp2=0, it indicates high intra-class similarity and low inter-class similarity, and the feature has good distinction degree, then set out.W(i)=100; otherwise, out.W(i)=temp1/temp2;if i not equals n, back to step2.

Step4: Descending the features by important degree.

Step5: End.

ReliefF sorting:

Step1: Input data set, set iteration number m,initialize zero output matrix out.W;

Step2: for 1=1:m select samples Ri randomly; find k nearest Hj in the same class; find k nearest Mj(C) in different classes;

Step3: for j=1:n(feature numbers) update the weight W[j];

Step4: Descending the features by the weight;

Step5: end.

Define the final features:

Since different feature selection algorithms have different adaptation to the data set, we can use multiple algorithms and weighting the results. The calculate formula as follows.

$$y = \alpha_1 f_1(x) + \alpha_2 f_2(x) + ... + \alpha_n f_n(x) = \sum \alpha_i f_i(x)$$

In which n indicates the number of feature selection algorithms, f(x) refers to as the important score of each feature, and α indicts the weights of different algorithms. In this paper, we choose two algorithms, and set n=2, and $\alpha_1=\alpha_2=1$, f(x) equals to the sequence number of features, then we calculate the comprehensive ranking for all the features.

The final number of features is selected by BP neural network, using K-fold cross-validation. According to the acc`uracy of classification, we set the acceptance threshold and select the proper features. This process is similar to the risk prediction in the next step.

(4) Risk prediction

Step1: Set neural network parameters, define the relevant variables, and initialize connection weights, input training samples.

Step2: Calculate the output model of the network and compare with the respected model, if have big difference, go to Step3, otherwise, go to Step4.

Step3: Calculate the error in the same layer, update the weight and threshold. If the error less than threshold, go to Step4, otherwise, back to Step2.

Step4: The training is over, and save the training result file.

Step5: Input testing data, simulate the training classifier to classify the data.

Step6: Count the recognition accuracy, the prediction process is over.

Experiments

Data set

The data set in these experiments is from the First Affiliated Hospital of Jilin University in China in 2011 and 2012, which consist of the lung cancer data and the control health data. Each data have three kinds of features: Basic information like name and address; class labels like the case group and the control group; risk factors or features, such as smoking, alcohol drinking, and coffee intake. After preprocessing of raw data, the dataset contains 705

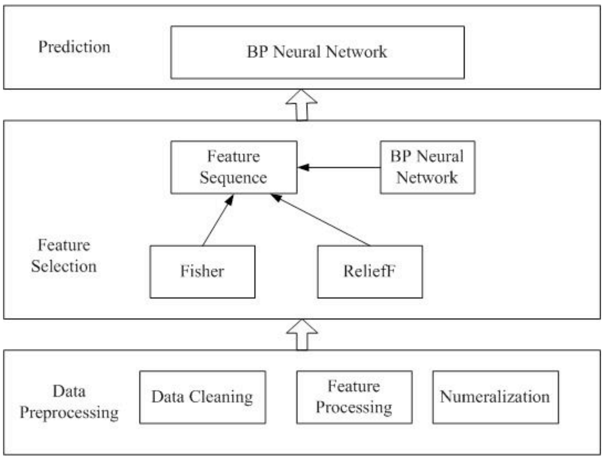

Figure 2. Lung Cancer Prediction Method Framework

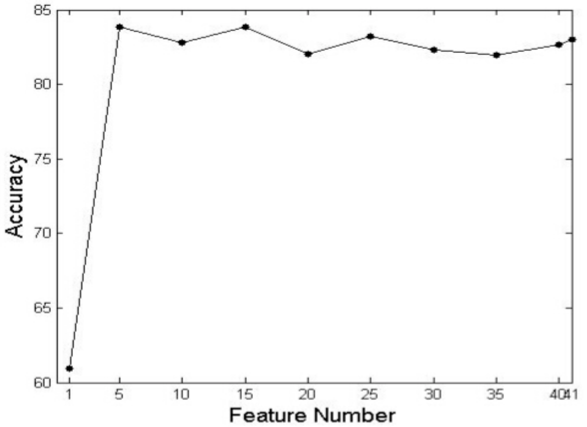

Figure 3. Accuracy of BP Neural Network

cases and 998 controls. Except the class label, the data set include 41 risk factors, such as age, education level, marital status, income and so on.

Results and analysis

Experiments show the proposed LCRP algorithm to process the data set. The 41 risk factors are ranking by feature selection algorithms. Using BP neural network and 5-fold cross validation, we get the accuracy of ranking risk factors, as in Figure 3. When add to 15 features, the accuracy is 83.816%, and we finally selected 15 risk factors. In the prediction process, we use BP neural network to predict the new data which include 50 cancer data and 50 healthy data, the finally prediction accuracy is 83.816%

Discussion

In this paper, we present a lung cancer risk prediction methods, including a method framework and an algorithm LCRP. We use feature selection to reduce risk factors dimension and BP neural network for prediction. A complete framework for prediction is constructed and the corresponding algorithm is realized. This study attempts to explore the effectiveness of feature selection and neural network used in cancer risk prediction. The focus of the method is how to determine the finally feature numbers, and it is related to the accuracy threshold which can be set based on the real applications. In the future work, we consider two points to improve this work. The one is about data sets, introducing experts' advices when decide the risk factor numbers, in order to improve the classification effectiveness of the features. The other is under the framework presented, try other feature selection and classification algorithms, comparing to this result, in order to improve the accuracy and practicality.

Acknowledgements

This research is supported by the National High Technology R&D Program 863 of China under Grant No. 2011AA010101, and by Key Science Technology Program of Jilin Province of China under Grant No. 2011ZDGG007.

References

Feng Zhijun, Wu Chaoyang (2014). Analysis of Related Facrtors of Radiation Pneumonitis in Locally Advanced NSCLC Patients Threated with Radiotherapy. *The Practical J Cancer*, **4**, 463-6.

Huang Meng, Chen Xing, Qiu Yue-feng et al (2011). Study on influencing factors and their interactions for lung cancer. *Chin J Dis Control Prev*, **2**, 91-4.

Huang Lili, Tang Jin, Sun Dengdi et al (2012). Feature selection algorithm based on multi-label ReliefF. *J Computer Applications*, **10**, 2888-91.

Jin Zhong, Hu Zhongshan, Yang Jingyu (1999). A Face Recognition Method Based on the BP Neural Network. *J Computer Res Development*, **3**, 274-7.

Kawsar Ahmed, Adbullah-Al-Emran, Tasnuba Jesmin et al (2013). Early Detection of Lung Cancer Risk Using Data Mining, *Asian Pac J Cancer Prev*, **1**, 595-9.

Mao Yong, Zhou Xiaobo, Xia Zheng et al (2007). A Survey for Study of Feature Selection Algorithms. *PR&AI*, **2**, 211-8.

Su Qiang, Ma Ni'na, Yu Jing et al (2014). Risk Factors of Advanced Non-small Lung Cancer with Pulmonary Infection. *Cancer Res Prev Treat*, **1**, 31-4.

Thunyarat Anothaisintawee, Yot Teerawattananon, Cholatip Wiratkapun et al (2014). Development and Validation of a Breast Cancer Risk Prediction Model for Thai Women: A Cross-Sectional Study. *Asian Pac J Cancer Prev*, **16**, 6811-7.

Wang Juan, Ci Linlin, Yao Kangze (2005). A Survey of feature Selection. *Computer Engineering&Science*, **12**, 68-71.

Wang Sa, Zheng Lian (2007). Feature selection method based on Fisher Criterion and feature clustering. *Computer Applications*, **11**, 2812-4.

Zeng Bixin, Ye Xiangyun (1996). Application of BP neural network approach on predicting of cancer mortality. Systems Engineering Theory Methodology Applications, **3**, 78-80.