Thyroid Cancer Gene Detection Algorithm Based on Feature Selection ELM

Wang Jining*, Mi Haoyang, Wu Yubo, Li Xingtong, Lin Chaohui

Sino-Dutch Biomedical and Information Engineering School,
Northeastern University, Shenyang, 110819, China
*email: 498628528@qq.com

Key words: Extreme Learning Machine; Gene Expression Data; Correlation Analysis

Abstract. At present, the detection and diagnosis of thyroid cancer has always been depend on the puncture cytology of thyroid gland. However, this method (the puncture cytology of thyroid gland) demands high accuracy of the instrument and costs may stay at a relative high level. Under this circumstance, aiming at solving this problem, our team comes up with a method, which is the thyroid cancer gene detection algorithm based on feature, to assist medical detection. The experiment of 100 genetic samples from DDBJ human gene bank shows that, the thyroid cancer gene detection algorithm based on feature selection ELM method can effectively improve the result of the thyroid cancer detection.

Thyroid cancer is one of the most common types of cancer, if the patients are not cured at an early-stage, once developed to end-stage, the cancer will lead to lifelong medication or life-threatening. In addition, some of the complications of thyroid cancer such as the damage to liver and kidney functions, which are seriously affected the daily life of patients. But now medical diagnostic tool for thyroid cancer diagnosis can not be 100%, so the supplementary medical means to help diagnose thyroid cancer are needed.

The current means of computer-aided diagnosis of thyroid cancer associated with gene expression data is the use of the R language for DNA microarray data to identify and to give the results of computer-aided diagnosis. The literature describes the feasibility of using DNA microarray data for gene selection, classification and machine learning. Finally achieve the feasibility of auxiliary diagnosis. But the disadvantage of this approach is that the success rate is not high, after the completion of the diagnosis, a lot of conventional physical and chemical examination are still required to diagnose patients whether suffering from thyroid cancer or not.

To solve this problem, this article proposes some methods to increase the success rate:

A. Using the correlation analysis to filter the7 data present in the DNA microarray with genes involved in thyroid cancer screening;

B. Proposed thyroid cancer diagnosis method based on gene expression data of ELM;

C. The use of the data downloaded from the NCBI demonstrates the feasibility of the method described above and obtained a better diagnosis.

Research Background

Computer Aided Diagnosis is a medical approach through the using the computer imaging, image processing technology and gene expression data to improve the diagnosis rate. Classification of genetic samples plays an extremely important part in computer-aided diagnosis of gene expression data detection field, effective classification can simplify the genetic data operation, and post-classification using machine learning methods to speed up the diagnostic process, shortening diagnosis period. Currently, the main machine learning methods are support vector machine (SVM), BP Neural Networks and Extreme Learning Machine (ELM). Generally speaking, SVM is usually used for pattern recognition , but only for two types of problems, the result of handling the multi types problem is relatively bad, which is very likely to produce sub-optimal solution; BP neural network is one kind of multilayer feedforward network which is trained based on error back propagation algorithm, however, this method costs long calculation time, and low learning rate; On the other hand, ELM, is widely used for classification of medical imaging and genetic data depends on its fast learning speed and classification accuracy rate in the biomedical field.

The Pretreatment of The Gene Data

For the sake of simplify the algorithm from time, space and calculation three aspects, the characteristic extraction of the obtained gene expression data is needed. The method is based on maintaining the gene expression feature, meanwhile, reducing the data size. After comparing the results of several characteristic extraction methods, the correlation analysis method is determined to be adapted. Correlation analysis means analyze two or more correlative variable elements, and measure how related these two variable elements are.

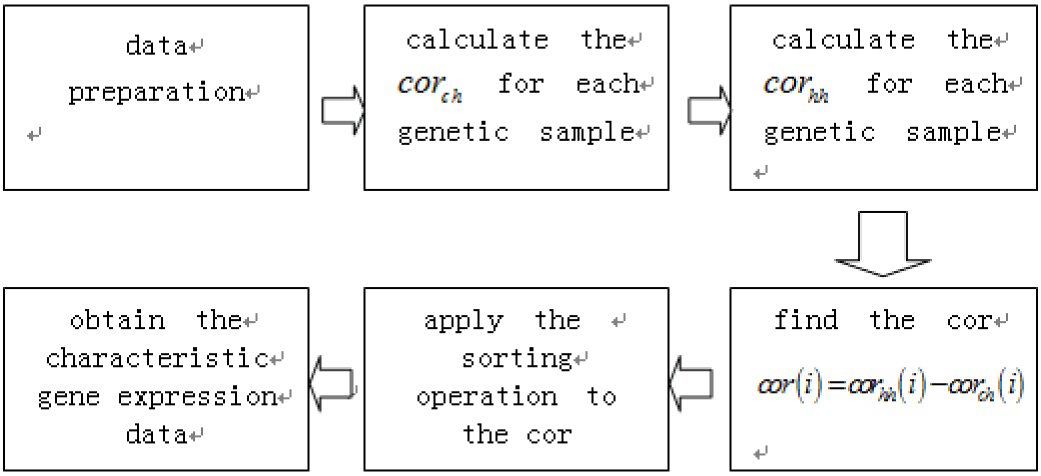
The Preparation For Gene Expression

Since the obtained gene expression data are all in the form of microarray, it's inconvenient to operate the data directly. To solve this problem, the microarray data should be transformed into matrix form, because matrix-form data make it easier for classification, transportation and summation.

Simultaneously, the obtained microarray gene expression data will be divided in to two classes, according to whether suffer from thyroid cancer or not. The two classes are named as cancer group and healthy group. Furthermore, the gene samples in the two groups will be separated into several size-equal panels.

The Correlation Between Thyroid Cancer and Gene Expression Data

All the panels in the two groups will be divided into sample pairs and obtain the correlation of the corresponding column vectors of the gene samples from each pairs. In this way, the correlation of the whole sample pairs can build a vector, simultaneously, all the vectors can be made as a matrix. Find the average of each column's element and obtain several vectors represent the correlation of the sample pair (which called cor_{ch}). Next, the panels in the healthy group will be divided in pairs, and obtain vectors which represent each panels' correlation coefficient according to the method mentioned above. (Which called cor_{hh}). Subtract cor_{ch} and cor_{hh}, obtaining the vector which represent the correlation between each gene expression data and thyroid cancer (which called cor. And then obtain the vector which represent the correlation between gene expression data and thyroid cancer using the method above according to the raw vector. Apply the sorting operation to the vector cor, remain the gene expression data that the correlation is greater than the threshold value, get rid of the other data, and complete the feature extraction of data.


data preparation
calculate the cor_{ch} for each genetic sample
calculate the cor_{hh} for each genetic sample
find the cor
$cor(i)=cor_{hh}(i)-cor_{ch}(i)$
apply the sorting operation to the cor
obtain the characteristic gene expression data

Step one：data preparation.

Step two：assume that the i-th column vector in the sample matrix of the cancer group is $m_c(i)$, the i-th column vector in the sample matrix of the healthy group is $m_h(i)$.Then the equation

$$cor_{ch}(i)=\frac{1}{n}\sum cor(m_c(i),m_h(i)) \qquad （1）$$

can be obtained.。

Step three：Assume that the i-th column vectors from the two gene samples from the healthy group are $m_{h1}(i)$ and $m_{h2}(i)$,separately. Then the equation

$$cor_{hh}(i)=\frac{1}{n}\sum cor(m_{h1}(i),m_{h2}(i)) \qquad （2）$$

can be obtained.

Step four：According to the formula

$$cor(i)=cor_{hh}(i)-cor_{ch}(i) \qquad （3）$$

the vector which represent the correlation between the i-th correlation of the gene expression data and the thyroid cancer can be obtained.

Step five：Apply the sorting operation to the correlation value which obtained last step，and select the value according to the set threshold value, finally the feature value in the gene expression data can be found

Step six：Maintain the obtained feature data, get rid of the other data and find the characteristic gene expression data.

The Auxiliary Diagnosis of Thyroid Cancer Based on the ELM and Gene Expression Data

Training

First, apply the pre-experiment operation to the gene expression data, and then select the original matrix(OM) to obtain the characteristic matrix ）（L1-7）.

Second，divide the CM samples from cancer and healthy groups into several sets randomly, entering the ELM and calculate, this process called training.

The classified data set can be obtained.T(L8-13)

```
1For 1 to Hmax I1=total(corr(Hx(i),UHy(i);
2For 1 to UHmax I2=total(corr(UHx(i),UHy(i);
3 I=I2/Hmax-I1/UHmax;
4For 1 to Hmax J1=total(corr(Hx(j),UHy(j);
5For 1 to UHmax J2=total(corr(UHx(j),UHy(j);
6 J=J2/Hmax-J1/UHmax;
7CM=SolveS(OM,I,J);
8 ELM(CM);
9 for    to L do;
10 Randomly generate hidden node parameters    ;
11 Calculate the hidden layer output matrix H;
12 Calculate the output weight vector    ;
13 Classification the test dataset T=H
```

The characters’ meaning:

Hx: the x-th diseased sample matrix
Hmax: the total amount of the H-UH pairs.
UHy:the y-th healthy sample matrix
UHmax:the total amount of the UH pairs.
I:raw correlation data
J:column correlation data

Auxiliary Diagnosis

First, characteristic extract the data, which is to be diagnosis according to the correlation extraction algorithm, and obtain the CM.

And then classify the trained categorical data, judging the gene expression data whether affected by thyroid cancer or not, if so, do the auxiliary diagnosis

Experiment

Data

The gene expression data samples of the experiment are collected from NCBI gene bank. The form of the samples are microarray data. Finally, 20 diseased gene expression data and 2000 healthy gene expression data are collected for the experiment.

To simplify the calculation, the collected sample data will be transformed 1164*1164 digital matrix.

Result

The result of the experiment is tested by the five-fold Cross Validation. Simultaneously, to ensure the universality, the experiment has been carried out fifty time under the conditions that the parameters are same. And obtain the average of the fifty results as the final result of the experiment.

Under the conditions that parameters are varied, the accuracy of the auxiliary diagnosis is obtained as followed:

Threshold	ELM hidden nodes	Accuracy
0.1	200	0.8555
0.1	400	0.9155
0.1	600	0.9252
0.1	800	0.9162
0.1	1000	0.9245
0.2	200	0.8872
0.2	400	0.9307
0.2	600	0.9310
0.2	800	0.9379
0.2	1000	0.9476
0.3	200	0.9093
0.3	400	0.9341
0.3	600	0.9428
0.3	800	0.9421
0.3	1000	0.9469
0.4	200	0.7645
0.4	400	0.8021
0.4	600	0.7938
0.4	800	0.7890
0.4	1000	0.8055
0.5	200	0.7497
0.5	400	0.7700
0.5	600	0.7507
0.5	800	0.7497
0.5	1000	0.7562

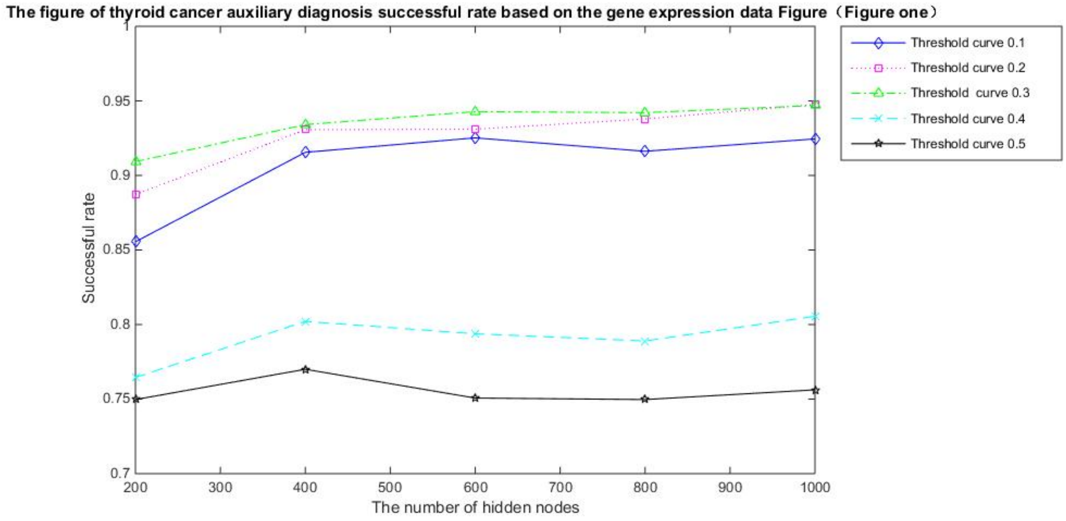
Outcomes

The Selection of the Threshold Value of the Gene Expression Data Pre-experiment

In order to select the threshold with the highest accuracy to carry out the experiment, under the different threshold selection, carry out several experiment with different hidden nodes. Comparing the auxiliary diagnosis accuracy, and obtained the figure one.

The Selection of the ELM Hidden Nodes

In order to select the hidden node with the highest accuracy to carry out the experiment, under the different hidden nodes selection, carry out several experiment with different threshold values. Comparing the auxiliary diagnosis accuracy, and obtained the figure one.



Conclusion

Aiming at the problem: time- consuming, low speed and accuracy of the traditional computer-aided diagnosis method the main body of the paper talks about a new approach: the auxiliary diagnosis of thyroid cancer based on the correlation analysis. Using correlation analysis to extract the characteristic point in the DNA microarray which is related to the thyroid cancer, to reduce the matrix size. And using ELM for machine learning. The experiment shows that, using the correlation analysis method to deal with the DNA microarray, and using ELM for machine learning can effectively shorten the diagnostic period, and increase the success rate of the auxiliary diagnosis.

Acknowledgement

This research was funded in SIVI project team. The team puts forward many valuable opinions for our research and paper writing. Here we express our sincere gratitude to them.

References

[1]Yu Hualong. The cancer classification research based on DNA microarray data.[D].Harbin Engineering University,2010.

[2]Huang Danfeng. The research based on the data microarray characteristic selection and classification methods.[D].Jiangsu University of Science And Technology,2012.

[3]Xu Chungui.The classification of tumor research based on the microarray data .[D].University of Science And Technology of China,2009.

[4]Wang Zhiqiong, Kang Yan, Yu Ge, Zhao Yingjie. The breast lump detection algorithm based on the characteristic selection ELM.[J]. Journal of Northeastern University (Natural Science),2013,06:792-796.

[5]Zhu Chan, Zou Xianxia, Xu Longfei. The classification research based on the gene expression data of DNA microarray.[J].Computer Science and Application ,2005,06:171-174.

[6]Peng Hongyi, Ye Yanrui, Zhang Junhui, Luo Zeju, Feng Guohe. The research based on the data microarray characteristic selection and classification methods [J]. Computer Science and

Application,2010,28:40-42.

[7]Jin Feiming. The tumor gene expression data classification algorithm research based on ELM.[D].Northeastern University,2013.