Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7): 2792-2800

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Research on classification and detection of colon cancer's gene expression profiles

Jun Yao[1], Yuchun Zhang[1] and Pingbo Hao[2]

College of science, Shenyang Ligong University, Shenyang, China
Qier Machine Tool Group Co. Ltd, Qiqihar, Heilongjiang, China

---

ABSTRACT

With the large-scale development of the technology—Gene Expression Profiles, the diagnostic method based on gene expression profiles is now becoming a quick and effective method in clinical medicine. But because of gene expression data's high dimension, small sample size and large noise, extracting the information about cancer correctly becomes the key point. In this paper, the gene expression data of colon tumor as an example put forward the mixed information gene extraction method combining Fisher Weight Function, discrete Fourier transform and principal component analysis and take multiple Logistic regression analysis together with Bayesian decision as classifier to do tumor classification and detection. The experiment results show that, the accuracy of 96.80% is achieved on CV recognition for colon cancer's data set using this method.

Keywords: Colon cancer, gene expression profiles, information gene, tumor classification.

INTRODUCTION

Cancer is one of the main diseases which influence human's health. Nowadays the world has more than 10 million cancer patients, and the death rate is high. Although early tumor's cure rate has been increasing along with the improvement of the treatment method and technology, tumor diagnosed by today's method tends to develop into the middle or late and naturally the treatment effect is not good. With the rapid development of molecular biology, people's understanding of tumor has developed to the genetic level and people have discovered many tumor-related genes. The occurrence and development of the tumor performance in the differences in gene expression and changes in gene expression of tumor cells. DNA Micro-array, also called gene chip, is a new technology which could detect DNA sequence and gene expression level rapidly and efficiently. This technology developed in recent years. With DNA Micro-array gene expression data becoming rich, a lot of classification forecasting and clustering technology began to be used to analyze gene expression data. There are many researchers using DNA Micro-array data to classify the cancer. Because every tumor has its genes' character expression profile, finding a group of gene "label" which could decide the category of the sample, namely informative genes among hundreds and thousands genes measured from DNA chip is the key to correctly recognize the type of tumor, diagnose reliably and simplify the experiment analysis. Meanwhile, it provides a shortcut for the development of anti-cancer drugs.

Tumor detection is actually a classification problem. In addition to study the classification rule of the difference between colon cancer tissue and normal tissue, Alon and others [1] used hierarchical clustering and some other methods to analyze and study the colon cancer sample's data. They choose 22 normal samples with 2000 different genes and 40 tumor samples. The article analyze the colon cancer genes' expression dataset of Alon and other people. (http://www.molbio.princeton.edu/colondata). This Dataset has 62 samples, in which 40 samples are colon cancer tissue's samples and the other 20 samples are normal samples of corresponding tissue. In the dataset, each sample recorded corresponding genes or expressed sequence tag (EST) of 2000 gene probe in DNA chip, namely

gene expression profile. This study based on gene expression profile, start from the classification of colon cancer and normal tissue samples. We used Fisher weight function and main constituent analysis as well as discrete cosine transform to extract information genes. We also used multiple Logistic regression analysis and Bayesian decision as classifier. On this basis, we reveal abnormal expression genes in colon cancer tissue and make classification detection.

TUMOR DETECTION METHOD

Algorithmic model

In this paper there are 5 steps to complete the tumor detection algorithmic frame model. The following 5 sections describe the steps particularly. Because extracting classification character combines Fisher weight function, discrete Fourier transform and main constituent analysis method, we called it hybrid information gene extraction method.

Step 1: Fisher weight function pretreats data. Use Fisher weight function to calculate gene weight function F_i of Number i. Arrange F_i according to ascending order and choose the higher F_i gene to be the candidate set M_s of information gene.

Step 2: Principal Component analysis(PCA) extracts information gene. Using the method PCA to reduce dimension of the sample data we choose, then we got the sample's principal component matrix M_w.

Step 3: Classify samples by Logistic regression analysis. Use Logistic classifier to classify M_w whose dimension has been reduced and foregone sample category information.

Step 4: Noise analysis. Use DCT(Discrete Cosine Transform) to remove noise.

Step 5: Bayesian decision with minimum error rate. Combine prior probability and mutant gene information based on gene expression profile to make classified detection of cancer.

The progress to extracting information gene from normal samples and colon cancer samples of colon cancer gene is as Figure 1.

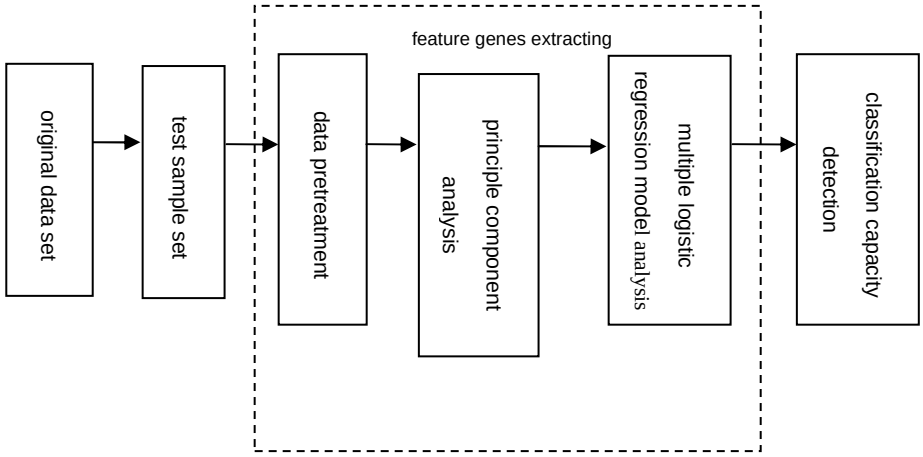

Figure1: DNA chip gene expression data's information gene extracting progress

Fisher weight function pretreat data

Because the number of genes are large, in the progress of judging tumor gene label, we need to remove a lot of "unrelated genes" to narrow the scope of oncogene that we need to research greatly. Actually, in gene expression profile, some genes' expression level are very close among all the samples. But some genes present difference obviously in normal samples and colon cancer samples. Fisher weight function is more efficient to two classification problem's character extraction. It started with known pattern of classification information. Calculate Fisher weight function F_i of Number i to judge the contribution character i make for classification. The definition of Fisher weight function [2]:

$$F_i = \frac{(\overline{x}_{i1} - \overline{x}_{i2})^2}{\sigma_{i1}^2 + \sigma_{i2}^2} \quad (1)$$

$\overline{x}_{i1}^2$ and $\overline{x}_{i2}^2$ separately express the average expression level of No. i gene in normal samples and colon cancer samples. σ_{i1}^2 and σ_{i2}^2 separately express the level variance of No. i gene in normal samples and colon cancer samples. Use formular (1) to calculate the score of colon cancer genes' normal samples and colon cancer samples and census the score to get the candidate set M_s of information gene.

Principle component analysis extracts information gene

Principle component analysis [3−5] is a common and efficient method to dispose, compress and extract information based on variable covariance matrix. Analyze the principle component of M_s and extract it. In order to minimum the square error produced in the progress of reducing sample set M_s's dimensions, we do two aspects of work. One is to transform the coordinate, that is to say, solve the orthogonal transformation matrix by Jacobi method. The other is to choose w component product of principle component. The calculation progress of PCA has three steps,

Step 1: Standardize the data in matrix M_s, namely, transform the samples' concentration elemen x_{ik}

$x_{ik} = (x_{ik} - \mu_k) / \sigma_k$, $i = 1,2,...,m, k = 1,2,...,p$,

among $\sigma_k^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_{ik} - \mu_k)^2$, $\mu_k = \frac{1}{m}\sum_{i=1}^{m} x_{ik}$, the purpose is to eliminate dimension's influence to evaluation result and get the standardized matrix M_s and its related coefficient matrix R.

Step 2: To the related coefficient matrix R, use Jacobi method to solve p non-negative eigenvalue $\lambda_1 > \lambda_2 > ... > \lambda_p \geq 0$ of the equation $\det(R - \lambda I) = 0$ and the eigenvector v_i of $\lambda_i, i = 1, 2, ..., p$, also $v_1, v_2, ..., v_p$ is standard orthogonal vector.

Step 3: Choose w component product of principle component and make w variance of principle component and the proportion of the total variance $\eta = \sum_{i=1}^{w}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_j$ approach 1. And make the w principle component we choose keep the original p gene information as much as possible in order to achieve dimensionality reduction and get the principle component matrix M_w.

Logistic regression analysis classify colon cancer samples

Use $x = (X_1, X_2, ..., X_{P-1})^T$ to express the factor that influence the occurrence probability of colon cancer, $\pi(x)$ to express the probability of colon cancer. Establish the function relationship between $\pi(x)$ and $x = (X_1, X_2, ..., X_{P-1})^T$:

$$\pi(x) = f(X_1, X_2, ..., X_{P-1}) \quad (2)$$

$\pi(x)$ expressed the probability of colon cancer's occurrence and establish the relationship between $\pi(x)$ and $f(X_1, X_2, ..., X_{P-1})$. Transform $\pi(x)$ as followed:

$$\theta[\pi(x)] = ln(\frac{\pi(x)}{1-\pi(x)}) \quad (3)$$

We can get the multiple Logistic regression model [6] between the incidence probability of colon cancer and mutant genes [6]:

$$\pi(x)=\frac{\exp(\beta_0+\sum_{k=1}^{p-1}\beta_k X_k)}{1+\exp(\beta_0+\sum_{k=1}^{p-1}\beta_k X_k)} \quad (4)$$

Apply Logistic regression to classify colon cancer samples, the progress is as followed:
Step 1: Confirm evaluation index. A discriminant function's function of sample classification largely depends on the selection of indicators. Too little or too much are not necessarily appropriate. On one hand we need to filter index according to specialized knowledge and experience. On the other hand we use statistical analysis method to detect the property of index.

Step 2: Get study samples. To the 62 known colon cancer samples data, set the 22 normal samples' probability value as 0, the other 40 tumor samples' probability value as 1.

Step 3: Estimation of regression coefficient. Use maximum likelihood method to estimate regression coefficient. To multiple Logistic regression analysis, coefficient estimation got from maximum likelihood method is often not unbiased estimation. In order to get unbiased estimation, we use multiple recursion method to revised the maximum likelihood estimation value constantly.

Step 4: Some independent variable x_k 's Wald detection to entirety classification influence. Keep the smaller variables of Sig. Repeat Step 1, Step 2 and Step 3 until find the minimum variables that could produce the right classification.

Noise analysis
Noise elimination in tumor genes expression profile is a hard work. The noise that gene expression profile data may influence classification are the error during the progress of data collection, the error in the matrix calculation progress of gene expression level, the obvious difference between specific sample data and similar sample, mutual interference between gene groups. What we have mentioned, such as, principle extraction, data standardization and other work could all be used to reduce the influence of noise on classification to some degree. But noise still has some influence in the property of classification. Using DCT could further eliminate the noise in gene expression profile.

2.5.1. Discrete Cosine Transform
DCT is a kind of transform who is similar to Discrete Fourier Transform and related to Fourier. It has a strong energy compression feature and almost get to the optimum value in compression efficiency. The math expression of single-dimensional DCT transformation could be expressed as formula (5) [7]

$$X(k)=\alpha(k)\sum_{n=0}^{N-1}x(n)\cos[\frac{\pi(2n+1)k}{2N}] \quad 0 \le k \le N\text{-}1 \quad (5)$$

Inverse DCT transformation could be expressed by formular (6) [7]

$$x(n)=\sum_{n=0}^{N-1}\alpha(k)X(n)\cos[\frac{\pi(2n+1)k}{2N}] \quad 0 \le k \le N\text{-}1 \quad (6)$$

In the formular above, $\alpha(0)=\sqrt{1/N}$, $\alpha(k)=\sqrt{2/N}$, $1 \le n \le N-1$ 。

2.5.2. Discrete cosine transform removes noise
To the gene expression profile data disposed by Fisher weight function, standardize the data in matrix M_s to get the standardized matrix M_b. Use formular (5) to do discrete cosine transformation against to the standardized matrix M_b. We get the new data matrix L_b. Using L_b to do principle component analysis again according to the steps in 3.3.

Bayesian Decision Theory based on minimum error rate

Bayes decision theory is a kind of pattern classification method when prior probability and kind of condition probability are known. The classification result of the to be classified samples depends on the entire samples in all fields. Bayesian theory's obvious character as decision is that it uses information that could be collected. Bayesian method's application progress [8] is as Figure 2.

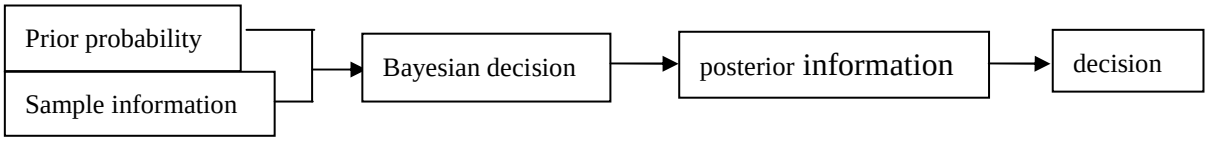

Figure 2: the steps of Bayesian decision

To the pretreated sample that need to be recognized, extract p character. And become a p -dimensional space's vector x. The purpose is to classify x to normal (ω_1) tissue sample or colon cancer(ω_2) tissue sample. Suppose two kind of prior probability are $P(\omega_1)$ and $P(\omega_2)$, kind of condition probability are $p(x|\omega_1)$ and $p(x|\omega_2)$. Use Bayesian formula to get the following posterior probability

$$P(\omega_i|x)=\frac{p(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{2}p(x|\omega_j)P(\omega_j)},i=1,2 \quad (7)$$

Then, the Bayesian decision rule based on minimum error rate is [9],

If $P(\omega_i|x)=\max_{j=1,2}P(\omega_j|x)$

Denominator in formula (7) has nothing to do with category. So when we compare the maximum, it could be ignored. Just calculating prior probability $P(\omega_i),i=1,2$ and kind of condition probability density $p(x|\omega_i),i=1,2$ could complete the classification.

RESULTS AND DISCUSSION

Calculate the score of colon cancer genes' normal samples and colon cancer samples separately using Fisher weight function and make a statistics. As is shown in Figure 3, abscissa expresses Fisher weight function's score of gene expression level and ordinate expresses the corresponding level's genes' number. We can see, most Fisher weight function's scores of the genes' expression level concentrate on the area that less than 0.2. These genes contribution less to classification. As is shown in Figure 4, rank ordering Fisher weight function's scores to colon cancer genes' samples and choose the Fisher weight function scores of 293 genes which are more than 0.2 as the candidate set of feature gene.

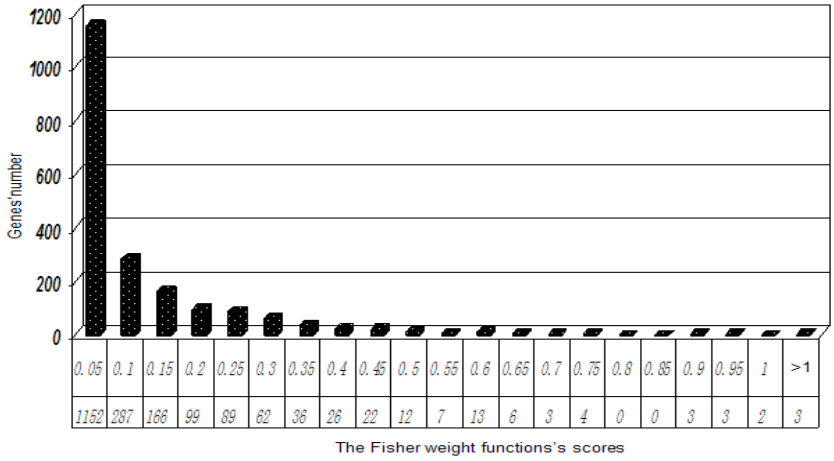

Figure 3 :The Fisher weight function statistics of gene's expression level

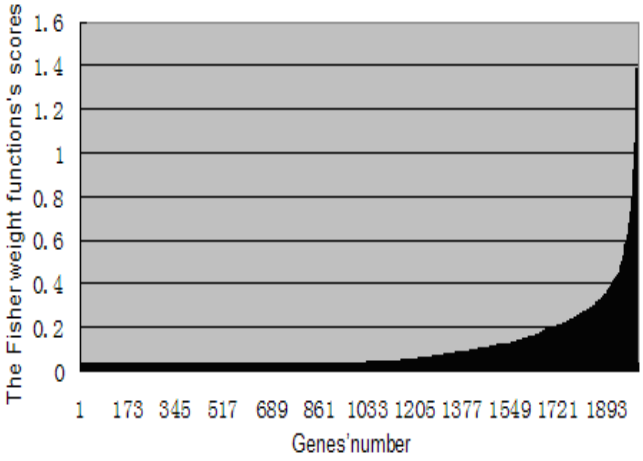

Figure 4: Ranking of Fisher weight functions's scores

Figure5 shows the principle component quantity that has not been discrete cosine transformed while Chart 6 shows the principle component quantity that has been discrete cosine transformed.

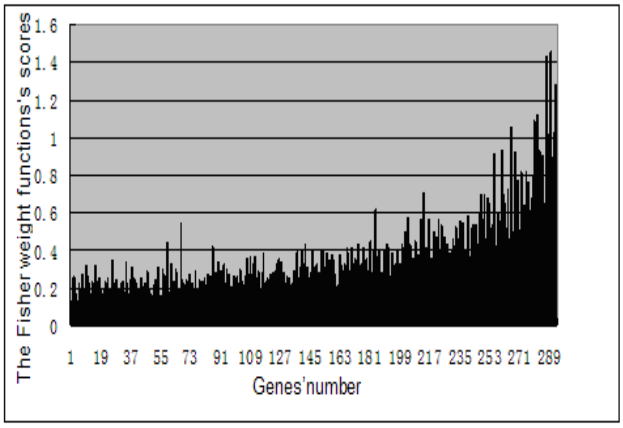

Figure 5: Principle component quantity without DCT

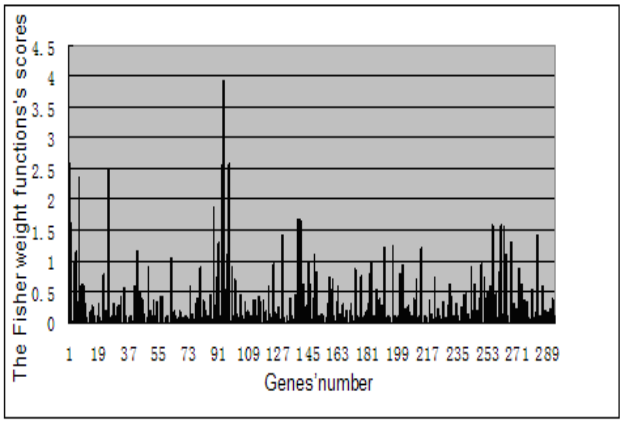

Figure 6: Principle component quantity with DCT

Compared Figure5 with Figure 6, it is not hard to see that, principle component quantity with DCT is more obvious and outstanding. Some gene expression profile's information which expresses not obviously before DCT begins to emerge. And other gene expression profile's information which expresses obviously begins to descend after DCT. It means the principle component quantity's differences are more obvious through DCT.

Analyze the principle component of 293 genes' expression level scores after discrete cosine transformation, we get 61 principle component quantities. As is shown in Figure 7, the new and old component quantity's expression value discriminate more obviously compared to before.

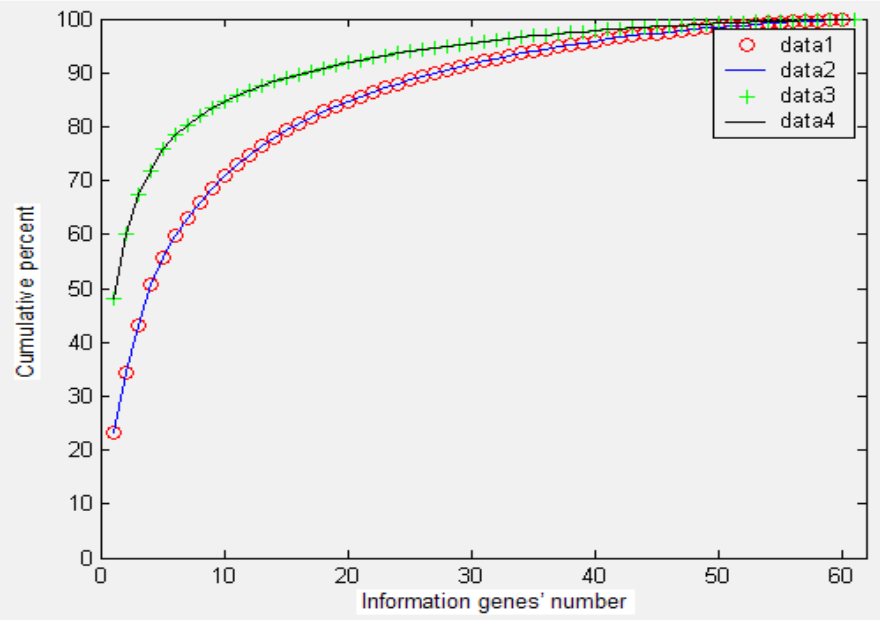

Figure7: Principle component quantity's accumulation expression level before and after DCT

Abscissa expresses the number of principle component quantity's accumulation while ordinate expresses accumulation expression level. The upper curve shows the principle component quantity's expression level without DCT while the nether shows the principle component quantity's expression level with DCT. The principle component quantity's expression level with DCT improved to some degree which means after discrete cosine transformation, noise's interference decreased.

Their expression rate to samples is as Figure1. We can see that, the more principle component quantities we choose, the more complete the sample's expression is. So in order to make the genes group we choose express the information of colon cancer samples as complete as possible, we choose 61 principe component quantities as information gene set.

Table 1: The relationship between the number of principle component quantity and Samples' expression rate.

The number of principle component quantity	Samples' expression rate(%)	The number of principle component quantity	Samples' expression rate(%)
61	100.000	15	89.118
40	97.782	8	82.084
28	94.924	4	71.881

According to Bayesian decision's rule based on the minimum error rate, make the 61 principle component quantities we choose which express gene feature as feature and integrate the general information, sample information and prior information about unknown paraments.(about 90% colon cancer is Chromosome 5 long arm APC gene inactivation in early stage, and about 40%~50% 's RAS related genes mutate, we take 45%). We can get the posterior information, then according to the posterior information we correct APC genes an RAS related genes. Do Logistic regression analysis with SPSS16.0 software. Finally we got 11 feature genes which could classify samples very well. In Table 2, we got the classification result wich adopt 11 feature genes. In Table 3, there is the Logistic regression analysis's result and 11 genes' parament feature which is used to classify.

Table 2: the result adopting 11 feature genes

Classification Table[a]

Observed		Predicted		
		VAR00074		
		0	1	Percentage Correct
Step 1 VAR00074	0	22	0	100.0
	1	2	38	95.0
Overall Percentage				96.8

a. The cut value is .500

Table 3: 11 genes' parament feature which is used to classify

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	VAR00054	-9.765	5.191	3.538	1	.060	.000	.000	1.506
	VAR00057	5.148	4.251	1.466	1	.226	172.163	.041	7.158E5
	VAR00060	13.355	7.474	3.193	1	.074	6.312E5	.274	1.452E12
	VAR00062	-15.296	9.112	2.818	1	.093	.000	.000	12.967
	VAR00063	4.907	3.076	2.544	1	.111	135.172	.325	5.615E4
	VAR00064	-6.891	4.705	2.145	1	.143	.001	.000	10.289
	VAR00065	-4.644	3.691	1.583	1	.208	.010	.000	13.338
	VAR00067	-1.032	1.318	.613	1	.434	.356	.027	4.719
	VAR00068	4.911	4.091	1.441	1	.230	135.733	.045	4.120E5
	VAR00072	8.161	5.827	1.961	1	.161	3.500E3	.038	3.193E8
	VAR00073	1.982	3.856	.264	1	.607	7.258	.004	1.389E4
	Constant	8.878	5.353	2.750	1	.097	7.170E3		

a. Variable(s) entered on step 1: VAR00054, VAR00057, VAR00060, VAR00062, VAR00063, VAR00064, VAR00065, VAR00067, VAR00068, VAR00072, VAR00073.

From Table 4 we can see that, when we choose Fisher weight function, DCT, PCA, Logistic discrimination and Bayesian decision as classification detecting method, the accuracy rate of CV recognition is higher, reaching 96.8%.

Table 4：Compared CV recognition accuracy rate using different classification detecting method to classify colon cancer data

Classification detecting method	Number of feature genes	CV recognition accuracy rate（%）
Fisher weight function+ PCA+Logistic discrimination	3	83.90
	4	93.50
	5	91.90
	8	91.90
	12	93.50
Fisher weight function+DCT + PCA+Logistic discrimination	2	88.70
	4	93.50
	5	93.50
	6	95.20
	9	95.20
Fisher weight function+DCT+ PCA +Logistic discrimination+Bayesian decision	11	96.80

To tumor classification method, except for CV recognition accuracy rate as the evaluation rule, there is no uniform and normative rule to evaluate it as so far. So in this article, we mainly use CV recognition accuracy rate as the main evaluation rule to compare different classification method. Table 5 gave us 9 kinds of conditions that using different feature selection methods and different classifiers to classify uniformed colon cancer gene expression data set. From this chart, we can clearly see that, the hybrid method we use in this article has a higher CV recognition accuracy rate.

Table 5: Accuracy rates' comparison of different feature gene selection method and different classifiers to classify colon cancer gene expression data

Feature gene selection method	Classifier adopted	CV recognition accuracy rate（%）	Reference
TNOM Score	SVM	74.20	[10]
PCA	Logistic discrimination	87.10	[11]
SNR	SVM	90.30	[12]
PLS	Quadratic equation discriminant analysis	91.90	[11]
ICA	Calculate the specific value between tumor and normal samples' independent component and build the classification model	91.10	[13]
PLS	Logistic discriminant	93.50	[11]
GA	KNN	94.10	[14]
RFSC and PCA	Gaussian radio basis function's SVM classifier	95.16	[15]
Fisher weight function +DCT+PCA	Logistic discriminant and Bayesian decision	96.80	This article

CONCLUSION

This paper studied about the selection of gene expression data classification's information genes. We did data pretreatment, principle component analysis, multiple Logistic regression analysis, noise analysis, Bayesian decision and so on. The information genes we extracted has a good classification capacity to samples set. The experiment shows, the algorithm to colon cancer data set, the CV recognition accuracy rate could reach to 96.8%; Compared to other related classification method to colon cancer tissue samples' classification condition, our method has obvious advantages in classification performance.

Aknowledgement
the National Natural Science Foundation of China under Grant No.11004138

REFERENCES

[1] Alon U, Barkai N, Notterman D A, et al, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[C]//Proc of the NationaI Academy of Sciences of the United States of America, 96(12):6745-675,**1999**.
[2] Shuxia Liu, Rongchuan Chen, Yannli Liu, Hui Chang,The Ultrasonic Indicator's Discriminant analysis of Ovarian Masses's Nature.*Chinese Medical Imaging Technology*,26(4)：737-740, **2010**
[3] Wang S L, Wang J, Chen H W, et aI. SVM-based tumor classification with gene expression data[C]//Internaional Conference on Advanced Data Mining and Applications. *Berlin Heidelberg*:*Springer-Verlag*,864-870, **2006**
[4]Nishimura K, Abe K, Ishikawa S, et aI. *Genome Informatics*, 14：346-347, **2003**
[5]Liu Z Q, Chen D C, Bensmail H, *Journal of Biomedicine and Biotechnology*, 2:155-159, **2005,**
[6]Jichuan Wang，Zhigang Guo, Logistic Regression Models—Method and Application. *Higher Education Press*, **2001**
[7]Lei Sha，Xia Ye,The method of image compression using discrete cosine transform. *Chengdu University of Technology* , 7:109-114, **1997**
[8]Gangzheng Guo, The application of bayes method in decision analysi.*statistics and applicaion*, 2013.16, **2013**,
[9]Zhaoqi Bian, Xuegong Zhang ,*Pattern Recognition*[M]. *Tsinghua University Press*, **2006**,
[10]Ban-Dor A，Bruhn L，Friedman N，et aI. *Journal of Computational Biology*, 7(3-4)：559-584, **2000**
[11]Nguyen D V,Rocke D M, *Bioinformatics*,18(1): 39-45, **2002**
[12]Furey T S，Cristianini N, Duffy N,et al. *Bioinformatics*, 16(10)：906-914, **2000**
[13]Zhang Xue Wu，Yap Yee Leng,Wei Dong, et al. *European Journal of Human Genetics*, 5(9)：1018-4813, **2005**
[14]Li L，Weinberg C R，Darden T A,et al. *Bioinformatics*, 17(12)：1131-1142, **2001**,
[15]Shulin Wang, Ji Wang, Huowang Chen, Boyun Zhang, *Computer Engineering and Science* 29(9)：84-90, **2007**