A Novel Feature Selection Method Based on CFS in Cancer Recognition

Xinguo Lu[1,2], Xianghua Peng[1,3], Ping Liu[4], Yong Deng[1], Bingtao Feng[1], Bo Liao[1]
1. School of Information Science and Engineering, Hunan University, Changsha, 410082, China
2. College of Mechatronics and Automation, National University of Defense Technology, Changsha, 410073, China
3. Hunan Industry Polytechnic, Changsha, 410208, China
4. Hunan Want Want Hospital, Changsha, 410016, China

***Abstract*—In recent years, the gene expression profiles are used for cancer recognition. But the researchers are disturbed by their large variables and small observes. In this paper, a novel feature selection method based on correlation-based feature selection(CFS) was proposed. Firstly, the measures of variable to variable and variable to observe were calculated respectively. Then we utilized heuristic search method to search the space of variable for selecting informative gene subset and the subset weight was computed using these measures. Through regression we obtained a subset of distinguished genes. Finally, the stratified sampling strategy was presented to obtain the most informative genes. And classification performance was tested to evaluate the proposed method. Ten-fold cross-validation experiment was performed in three datasets including leukemia, colon cancer and prostate tumor. The experimental results show that the proposed method can obtain the distinguished gene subset and different classifier can acquire better classification performance with this subset.**

I. Introduction

Microarray techonology which produces gene expression data is a powerful tool for gene function studying and can analyze thousands of genes at the same time. Analysis of gene expression data helps us quickly explore gene expression differences between patient samples and healthy controls. These studies can used not only for cancer diagnosis, but also for rapid treatment and drug research[1]. However, it is very difficult to identify the distict genes in the cancer recognition due to the character of high dimensions, small samples and great redundant in the microarray gene expression data([2], [3], [4]). Therefore, feature selection becomes a research focus in analysis of gene expression data.

Feature selection methods can be divided into two categories: filters, evaluating the features according to the heuristic function based on general characteristics of the data; and wrappers, evaluating the features using the characteristics of the data joint with the learning algorithm. Filters has the following characteristics: relatively low computational complexity, suitable for large-scale database, and the selected features with moderate classification capability. But for wrappers, more informative genes are selected whereas it is suitable for small-scale database due to the high computational complexity([5], [6]). Using gene features transformation, Lu et al. proposed a novel method of gene features extraction in cancer recognition. In this method, the cancerogenic factors are extracted to different cancers and a relative space is built to the cancerthen the gene features are extracted for cancers with them[7]. Yu et al. presented a novel feature gene selection approach combining improved discrete particle swarm optimization with support vector machine[4]. Cho et al. attempted to explore the relationship between the feature selection methods and machine learning classifiers. In three datasets, the performances of many feature selections and classifiers were evaluated systematically[8].

Due to the different searching mechanism and evaluation strategy, the selected significant genes with different approaches are extremely different. None of feature selection method is proved to be the most optimal one[7]. So this presented method which combined wrappers with filters would reach a good performance.

Correlation-based feature selection(CFS) is an effective feature selection method, and the set of features mostly related to some class can be selected from the gene expression data. It reduces the data in dimensionality by more than sixty percent in most cases without negatively affecting accuracy[5]. However, in CFS the features are selected only by calculating the correlation between features and classes, features and features. It does not take into account the characters of various classifiers. When the selected features are applied for cancer diagnosis it cannot achieve a satisfying performance. So a Stratified Sampling feature selection method based on CFS (CFS-SS) was proposed in this paper. CFS-SS was applied in three gene expression datasets including leukemia, colon cancer and prostate tumor. The results of ten-fold cross-validation experiments showed that the selected genes gained a better classification performance compared to other approaches in cancer recognition.

II. Backgrounds

A. Microarray expression profile

Gene expression profile produced by DNA microarrays is usually illustrated in the matrix form. Set X is a mn (usually $m>>n$) gene expression matrix. X_{ij} is the observed gene expression value of ith gene in jth sample. In the gene expression dataset, there are some characters of large dimension, small samples and great noise.

2012 IEEE 6th International Conference on Systems Biology (ISB)
978-1-4673-4398-5/12/$31.00 ©2012 IEEE

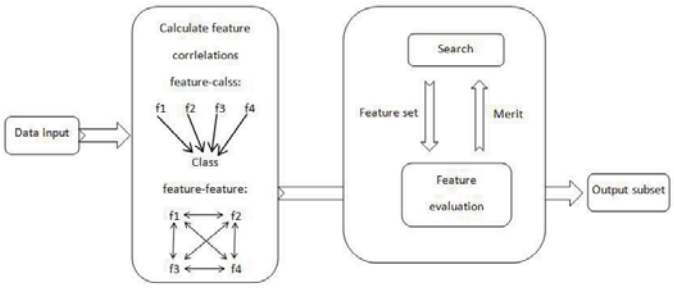

Fig. 1. The process of CFS feature selection

B. Correlation-based feature selection(CFS)

CFS is a fast, correlation-based filter algorithm that can be applied to continuous and discrete problems[5]. The CFS algorithm is a heuristic for evaluating the worth or merit of a subset of features. This heuristic algorithm takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The hypothesis on which the heuristic based is:

Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other.

In test theory, the same principle is used to design a composite test (the sum or average of individual tests) for predicting an external variable of interest. In this situation, the features are individual tests which measure traits related to the variable of interest (class). Equation 1 formalises the heuristic:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (1)$$

where $Merit_s$ is the heuristic merit of a feature subset S containing k features, $\overline{r_{cf}}$ is the average feature-class correlation, and $\overline{r_{ff}}$ is the average feature-feature intercorrelation. Equation 1 is, in fact, Pearson's correlation, where all variables have been standardised. The numerator can be thought of as giving an indication of how predictive a group of features are; the denominator of how much redundancy there is among them. The heuristic handles irrelevant features as they will be poor predictors of the class. Redundant attributes are discriminated against as they will be highly correlated with one or more of the other features.

The purpose of feature selection is to decide which of the initial (possibly large) number of features to include in the final subset and which to ignore. If there are n possible features initially, then there are 2^n possible subsets. The only way to find the best subset would be to try them all this is clearly prohibitive for all but a small number of initial features. Various heuristic search strategies such as hill climbing and best first are often applied to search the feature subset space in reasonable time. CFS uses a stopping criterion of five consecutive fully expanded non-improving subsets. In this paper, the search strategy is best first. The process of CFS feature selection as shown in Fig.1.

C. Stratified Sampling(SS)

Stratified sampling is a method that separates the subsets with N features into L groups according to some strategies. There are $N_1, N_2, ..., N_L$ subsets in L groups respectively, and the total subsets of the L groups is 2^N [9].In this paper, the subsets with same features size will put into a group.

III. A Stratified Sampling Feature Selection Method Based on CFS (CFS-SS)

CFS-SS is a feature selection method based on CFS. It is a wrapper whereas CFS is a filter. The steps of CFS-SS select features are listed as follows:

Step 1: preproccess the inputting data: Firstly, preproccess the data include adding missing value, normalizing the data, ranking the gene by variance in descend. Then select the top S genes to next step.

Step 2: pre-select features by CFS: From the top S genes, select the best gene subset (S_{cfs}) containing features highly correlated with the class, yet uncorrelated with other genes in the subset by CFS. After this step, the number of genes in this subset will be reduced to less than 10% of the inputting data.

Step 3: stratified sampling from the S_{cfs}. Stratified sampling the all subsets of the S_{cfs} and put the subsets with the same gene size into a group. Select k groups(S_{ss}) and $\sum_{i=N-k}^{N-1} C_N^i$ subsets are included.

Step 4: acquire the best feature subset S_{cfs-ss}: Select a classifier to test each element of the S_{ss} using ten-fold cross-validation test. Return the element with the best performance.

The detail of the CFS-SS algorithm is as follows:

algorithmCFS-SS (S, k, s, classifier)
inputting S, k, s, classifier
outputtingsubset S_{cfs-ss}
steps
(1) Data pre- preprocessing
(2) S_{temp}=filter(S , *s*)
(3) S_{cfs}=CFS (S_{temp})
(4) S_{ss}=SS (S_{cfs} , *k*)
(5) for *i*=0 to length(S_{ss})-1 do
(6) Acc[*i*]= evaluation (S_{ss}[*i*],classifier).pctCorrect()
(7) Max_pctCorrect=Acc[*i*]>Max_pctCorrect? Acc[*i*]: Max_pctCorrect
(8) endfor
(9) for *i*=0 to length(S_{ss})-1 do
(10) if Acc[*i*] ==Max_pctCorrect then
(11) S_{cfs-ss}.add(S_{ss}[*i*])
(12) endif
(13) endfor
(14) return S_{cfs-ss}

IV. Experimental results and analysis

A. Datasets

Three datasets including leukemia dataset, colon cancer dataset and prostate tumor dataset were applied to evaluate the proposed method.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

1) Leukemia dataset: consists of 72 samples in which 25 samples are acute myeloid leukemia (AML) and 47 samples are acute lymphoblastic leukemia (ALL). The source of the gene expression measurements was taken from 63 bone marrow samples and 9 peripheral blood samples. Gene expression levels in these 72 samples were measured using high density oligonucleotide microarrays. Each sample contains 7129 gene expression levels.

2) Colon cancer dataset: consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains 2000 gene expression levels. Although original data consists of 6000 gene expression levels, 4000 out of 6000 were removed based on the confidence in the measured expression levels. 40 of 62 samples are colon cancer samples(CCS) and the remaining are normal colon samples(NCS). Each sample was taken from tumors and normal healthy parts of the colons of the same patients and measured using high density oligonucleotide arrays.

3) Prostate tumor dataset: Consists of 136 samples in which 77 samples are prostate tumor(PTS) and 59 samples are normal prostate(NPS). Each sample contains 12600 gene expression levels.

B. Experimental environment and parameters

In this paper, the classification results on the gene set which selected by CFS-SS were compared with the results on the feature set selected by information gain(IG), principal component analysis(PCA) and CFS. K-nearest neighbor (KNN), support vector machine (SVM), multilayer perceptron (MLP), native bayes(NB), decision tree(DT) were used to recognize the samples in the experimental. In CFS, the best first was used as searching strategy. The number of top genes(S) was pre-defined to 500. After preprocess, the numbers of genes selected by CFS were 34, 13, 15 respectively. In the method of IG, the top of 50 genes were selected. In the stratified sampling, the *k* was set to 3, which means that the subsets in N-3,N-2,N-1 layer space were selected. Ten-fold cross-validation experiment was performed in this three datasets. Each experiment was run ten times, and the mean of these ten times were calculated.

C. Classifiers

KNN, SVM, MLP, NB, DT were used to recognize the samples.

1) KNN: K-nearest neighbor (KNN) is one of the most common methods among memory based induction. Given an input vector, KNN extracts *k* closest vectors in the reference set based on similarity measures, and makes decision for the label of input vector using the labels of the *k* nearest neighbors. Pearsons coefficient correlation and Euclidean distance have been used as the similarity measure. When we have an input X and a reference set D = d_1, d_2, , d_N, the probability that X may belong to class c_j, P(X, c_j) is defined as follows:

$$P(X, c_j) = \sum_{di \in kNN} Sim(X, d_i) P(d_i, c_j) - b_j \qquad (2)$$

where *Sim*(*X*, d_i) is the similarity between *X* and d_i and b_j is a bias term.

2) SVM: Support vector machine (SVM) estimates the function classifying the data into two classes. SVM builds up a hyperplane as the decision surface in such a way to maximize the margin of separation between positive and negative examples. SVM achieves this by the structural risk minimization principle that the error rate of a learning machine on the test data is bounded by the sum of the training-error rate and a term that depends on the Vapnik-Chervonenkis (VC) dimension. Given a labeled set of M training samples (X_i, Y_i), where $X_i \in R^N$ and Y_i is the associated label, $Y_i \in$ -1, 1, the discriminant hyperplane is defined by:

$$f(X) = \sum_{i=1}^{M} Y_i \alpha_i k(X_i.X_i) + b \qquad (3)$$

where $k(X_i.X_i)$ is a kernel function and the sign of *f(X)* determines the membership of *X*. Constructing an optimal hyperplane is equivalent to finding all the nonzero (support vectors) and a bias *b*.

3) MLP: Error back propagation neural network is a feed-forward multilayer perceptron (MLP) that is applied in many fields due to its powerful and stable learning algorithm. The neural network learns the training examples by adjusting the synaptic weight of neurons according to the error occurred on the output layer. The power of the backpropagation algorithm lies in two main aspects: local for updating the synaptic weights and biases, and efficient for computing all the partial derivatives of the cost function with respect to these free parameters. The weight-update rule in backpropagation algorithm is defined as follows:

$$\Delta w_{ji}(n) = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}(n-1) \qquad (4)$$

where $\Delta w_{ji}(n)$ is the weight update performed during the *n*th iteration through the main loop of the algorithm, η is a positive constant called the learning rate, δ_j is the error term associated with *j*, x_{ji} is the input from node *i* to unit *j*, and $0a<1$ is a constant called the *momentum*.

4) NB: NB is optimal when the features are conditionally independent. i.e., when the probability density function for class, denoted, can be decomposed as. In this case, the densities can be estimated separately for each feature which simplifies the training and makes NB feasible for very large feature sets. NB has been deemed surprisingly accurate. Even when the independence assumption is clearly false. NB may produce linear boundaries between the classes. This will happen if the individual densities are assumed to be Gaussian with the same variance (called Gaussian NB with shared variance). Only the means for the c classes need be estimated for each feature. Alternatively, variances for the classes can be estimated together with the means (Gaussian NB with distinct variance).

5) DT: J48 is a kind of decision tree (DT), each attribute in the tree is completely independent. A DT model was developed using a variant of the classification and regression tree

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

(CART) method, which consists of two stepsłtree construction and tree pruning. In the process of the tree construction, the algorithm identifies the best predictor variables that divide the sample in the parent node into two child nodes. The split maximizes the homogeneity of the sample population in each child node (e.g., one node is dominated by the cancer samples, and the other is populated with the noncancer samples). Then, the child nodes become parent nodes for further splits, and splitting continues until samples in each node are either in one classification category or cannot be split further to improve the quality of the DT model. To avoid overfitting the training data, the tree is then cut down to a desired size using tree cost-complexity pruning. In the end of the process, each terminal node contains a certain percentage of cancer samples. This percentage specifies the probability of a sample to be the cancer sample.

D. Results and analysis

In this paper, the classification results on the feature set which selected by CFS-SS were compared with the results on the feature set which selected by IG, PCA and CFS.

1) Classification performance evaluation: The samples used in the experiment are divided into two categories: the positive samples and the negative samples. The positives samples are ALL, CCS and PTS and the negative samples are AML, NCS and NTS. Accuracy(Acc), precision(Prec), sensitivity(sn) and specificity(sp) were used to evaluate the performance of the feature subset in the classification. Acc, Prec, sn, sp are defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Prec = \frac{TP}{TP + FP} \tag{6}$$

$$Sp = \frac{TN}{TN + FP} \tag{7}$$

$$Sn = \frac{TP}{TP + FN} \tag{8}$$

Where TP is the number of true positive samples, FP is the number of false positive samples. TN is the number of true negative samples, and FN is the number of false negative samples. The confusion matrix defined as Table I.

TABLE I
CONFUSION MATRIX OF THE CLASSIFICATION PERFORMANCE EVALUATION

$Testsample$	$Predictsample$	
	$Positivesample$	$Negativesample$
$Positivesample$	TP	FN
$Negativesample$	FP	TN

2) Results and analysis: The results of recognition rate on the test data are shown in Tables II,III and IV. In the first column are the feature selection methods, and in the first row are the classifiers. The classification accuracy on feature subsets S_{cfs-ss} chosen by CFS-SS are better than that on feature sets selected by IG, PCA and CFS. KNN and SVM on feature sets S_{cfs-ss} reaches the best classification accuracy 100% on leukemia dataset. The accuracy of the SVM method based on the feature set select by IG reaches 91.91% while the accuracies on other three feature sets are under 80% in prostate dataset. It shows that not all classifiers and feature selection methods are suitable for all datasets.

TABLE II
THE ACCURACY OF TEN-FOLD CROSS VALIDATION IN LEUKEMIA DATASET(%)

	KNN	SVM	MLP	NB	$J48$
IG	94.44	97.22	95.83	95.83	86.11
PCA	68.56	84.72	81.94	84.72	93.06
CFS	97.22	98.61	100	100	84.72
$CFS-SS$	100	100	100	100	94.44

TABLE III
THE ACCURACY OF TEN-FOLD CROSS VALIDATION IN COLON DATASET(%)

	KNN	SVM	MLP	NB	$J48$
IG	80.65	87.10	80.65	83.87	85.48
PCA	69.35	77.42	70.97	66.13	64.52
CFS	85.48	85.48	82.26	83.87	85.48
$CFS-SS$	93.55	90.32	90.32	88.71	91.94

TABLE IV
THE ACCURACY OF TEN-FOLD CROSS VALIDATION IN PROSTATE DATASET(%)

	KNN	SVM	MLP	NB	$J48$
IG	87.5	91.91	91.91	58.09	90.44
PCA	77.94	67.65	89.71	57.35	80.88
CFS	93.38	72.06	94.12	55.88	91.18
$CFS-SS$	95.59	78.68	96.32	58.82	94.85

Fig.2, Fig.3 and Fig.4 show the precision of ten-fold cross validation with different classifiers respcecting to the gene selection methods including IG, PCA, CFS and CFS-SS in three datasets respectively. From Fig.2, Fig.3 and Fig.4, it can be seen that S_{cfs-ss} get better precision in these three datasets excluding prostate tumor dataset. Strinking, in leukemia dataset it reaches 100% in precision excluding J48. In prostate tumor dataset, KNN, MLP, NB and J48 with CFS-SS achieve better precisions than with other gene selection methods.

Fig.5, Fig.6 and Fig.7 illustrate the sensitivity(sn) and specificity(sp) of ten-fold cross validation with different classifiers respcecting to the gene selection methods including IG, PCA, CFS and CFS-SS in three datasets respectively. From Fig.5, Fig.6 and Fig.7, we can see that in leukemia dataset the sn and sp on S_{cfs-ss} is better than on other selected gene sets

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

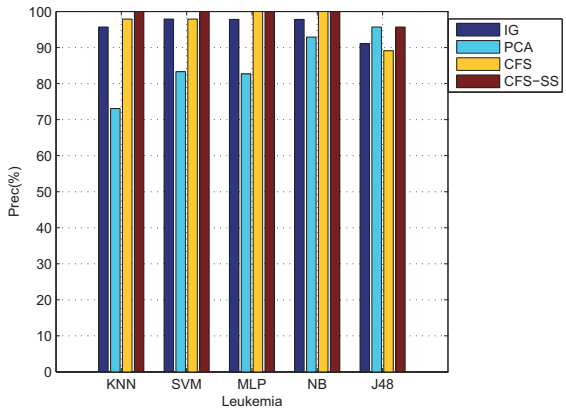

Fig. 2. The *Prec* of ten-fold cross validation in leukemia dataset

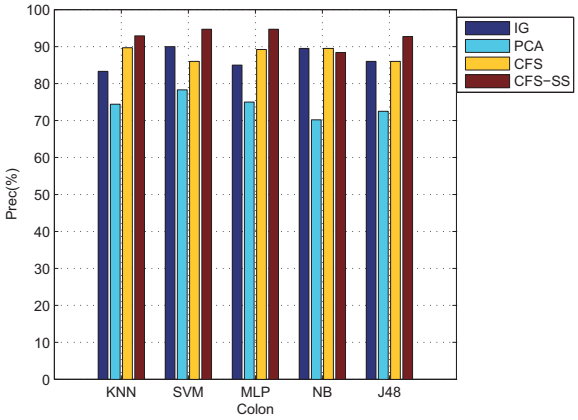

Fig. 3. The *Prec* of ten-fold cross validation in Colon dataset

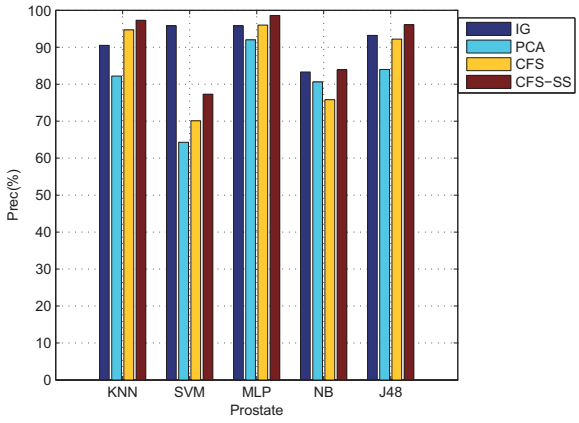

Fig. 4. The *Prec* of ten-fold cross validation in Prostate dataset

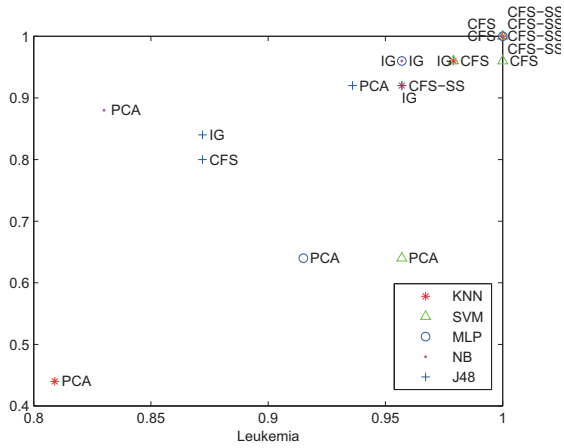

Fig. 5. The *Sn* and *Sp* of ten-fold cross validation in leukemia dataset

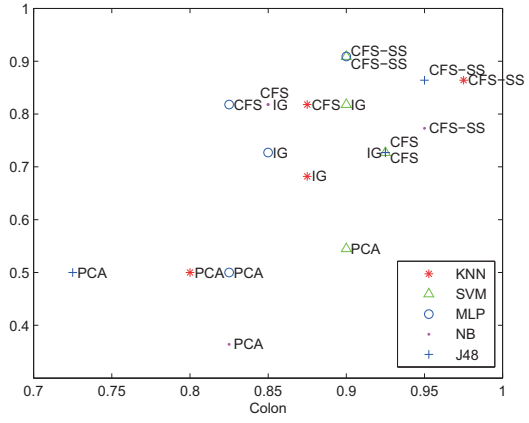

Fig. 6. The *Sn* and *Sp* of ten-fold cross validation in Colon dataset

via IG, PCA and CFS. In prostate dataset the results using NB are not good in comparison with using other classifiers, but NB combined with CFS-SS get the better sn and sp.

To evaluate validity of the proposed method, S was set to 50, 100, 150, 200, 250 and 500 respectively. The experiment results are shown in Table V, VI and VII.

From Table V, we can see that in leukemia dataset using CFS-SS the classification accuracies to different classifiers are better than CFS. Striking, when S is set to 200 and 250, the classification accuracies of KNN, MLP and NB on feature subset selected by CFS-SS reach 100%. And when S is set to 500, the classification accuracies of SVMKNNMLP and NB on feature subset selected by CFS-SS reach 100%. The performance of J48 is not good as the other classifiers, it also achieves the best accuracy to 95.83% at S=150, 200, 250.

Table VI shows the classification accuracies of different classifier on S_{cfs-ss} compared to S_{cfs} in the colon dataset. When S is set to 250, J48 based on S_{cfs-ss} reaches the best accuracy to 93.55%. The KNN gains the same performance at S=500. The best accuracies of SVM, MLP and NB on S_{cfs-ss} are achieved when s=200, which is better than on S_{cfs}

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

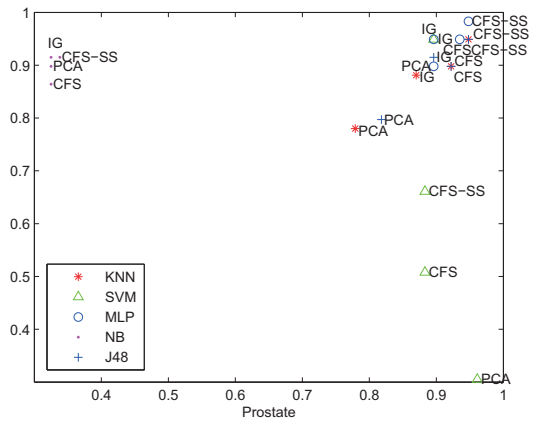

Fig. 7. The *Sn* and *Sp* of ten-fold cross validation in Prostate dataset

TABLE V
THE ACCURACY OF TEN-FOLD CROSS VALIDATION IN S_{cfs} AND S_{cfs-ss} OF LEUKEMIA DATASET

Classifier	*Feature set*	50	100	150	200	250	500
KNN	*CFS*	93.56	93.06	94.44	97.22	98.61	97.22
	CFS − SS	98.61	97.22	98.61	100	100	100
SVM	*CFS*	94.44	93.06	97.22	98.61	98.61	98.61
	CFS − SS	97.22	97.22	98.61	98.61	98.61	100
MLP	*CFS*	94.44	90.28	94.44	97.22	94.44	100
	CFS − SS	98.61	97.22	98.61	100	100	100
NB	*CFS*	95.83	95.33	97.22	100	98.61	100
	CFS − SS	98.61	98.61	98.61	100	100	100
J48	*CFS*	94.44	93.06	88.89	87.5	90.28	84.72
	CFS − SS	94.44	94.44	95.83	95.83	95.83	94.44

TABLE VI
THE ACCURACY OF TEN-FOLD CROSS VALIDATION IN S_{cfs} AND S_{cfs-ss} OF COLON DATASET

Classifier	*Feature set*	50	100	150	200	250	500
KNN	*CFS*	82.26	80.65	75.81	79.03	82.26	85.48
	CFS − SS	80.65	88.71	83.87	90.32	87.10	93.55
SVM	*CFS*	82.26	87.10	87.10	85.48	87.10	85.48
	CFS − SS	80.64	88.71	90.32	91.94	88.71	90.32
MLP	*CFS*	85.48	90.32	85.48	85.48	80.65	82.26
	CFS − SS	85.48	91.94	90.32	91.94	90.32	90.32
NB	*CFS*	75.81	80.65	82.26	83.87	80.65	83.87
	CFS − SS	85.48	87.10	87.10	90.32	87.10	88.71
J48	*CFS*	79.03	85.48	83.87	80.65	88.71	85.48
	CFS − SS	85.48	87.10	88.71	87.1	93.55	91.94

In Table VII, when S is set to 500, the MLP achieves the best classification accuracy 96.32% on S_{cfs-ss}. The other classifiers also gain the best performance when S=500. All classifiers on S_{cfs-ss} achieve better classification performance than on S_{cfs}. The only exception is MLP at S=150. Compared to SVM and NB, KNN, MLP and J48 obtain better classification accuracies.

TABLE VII
THE ACCURACY OF TEN-FOLD CROSS VALIDATION IN S_{cfs} AND S_{cfs-ss} OF PROSTATE DATASET

Classifier	*Feature set*	50	100	150	200	250	500
KNN	*CFS*	76.47	84.56	89.71	89.7	91.18	93.38
	CFS − SS	83.82	86.03	91.18	90.44	93.38	95.59
SVM	*CFS*	52.20	52.94	58.82	61.02	66.18	72.06
	CFS − SS	57.35	58.82	63.97	67.64	66.91	78.68
MLP	*CFS*	77.94	79.41	85.29	93.83	91.18	94.12
	CFS − SS	78.68	82.35	76.47	94.85	91.91	96.32
NB	*CFS*	55.15	55.15	55.14	55.88	55.88	55.88
	CFS − SS	55.88	55.15	55.14	57.35	57.35	58.82
J48	*CFS*	79.41	81.62	85.29	85.29	86.02	91.18
	CFS − SS	86.03	87.5	87.5	88.24	92.65	94.85

V. CONCLUSION

Due to the high dimension, small samples and great noise, the researchers can not make a good decision in cancer recognition via analyzing gene expression data. In the present work, a novel feature selection method based on correlation-based feature selection(CFS) was proposed. In this method, filters and wrappers were combined to eliminate the noise and redundance in gene expression data. And the experimental results show that the method can gain the better performance in comparison with corresponding approaches.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (61202288), the Ph.D. Programs Foundation of Ministry of Education of China (20100161120023), the National Science Foundation for Post-doctoral Scientists of China (20100471790), the Fundamental Research Funds for the Central Universities and the Young Teachers Program of Hunan University.

REFERENCES

[1] Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. Nature Genet, 1999, 21. 48-50.
[2] Li JZ, Yang K, Gao H, et al. Model-Free Gene Selection Method by Considering Unbalanced Samples. Journal of Software,Vol 17, No 7, July 2006. 1485-1493.
[3] Lu XG, Lin YP, Wang HJ, Zhou SW, Li XL. A novel relative space based gene feature extraction and cancer recognition. In: Zhou ZH, Li H, Yang Q, eds. Proc. of the 11th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2007). Berlin: Springer-Verlag, 2007. 712-719.
[4] Yu HL, Gu GC, Liu HB, et al. Feature gene selection by combining an improved discrete PSO and SVM. Journal of Harbin Engineering University, Dec.2009, vol.30, No.12: 1399-1403.
[5] Hall MA. Correlation-Based Feature selection for discrete and numeric class machine learning. In: Langley P, et al., eds. Proc. of the 17th Intl Conf. Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2000. 359-366.
[6] Ji ZW, Hu M, Yin JX. A survey of feature selection algorithm. Electronic Design Engineering, Vol.19, No.9, May. 2011.
[7] Lu XG, Peng XH, Li D, et al. Novel method of gene features extraction in cancer recognition. Computer Engineering and Application, 2010, 46(30): 237-240.
[8] Cho SB, Won HH. Machine Learning in DNA microarray analysis for cancer classification Proc of Bioinformatics 2003 First Asia-Pacific Bioinfomatics Conference(APBC), 2003: 189-198.
[9] Jin YJ, Jiang Y, Li XY. sampling technique. [Chinese People's University Press], 2002. 61-215.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE