Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FUZZY C-MEANS AND ENTROPY BASED GENE SELECTION BY PRINCIPAL COMPONENT ANALYSIS IN CANCER CLASSIFICATION

[1]SOMAYEH ABBASI, [2]HAMID MAHMOODIAN
[1]Department of Electrical Engineering, Najafabad branch, Islamic Azad University, Isfahan, Iran
[2]Faculty of Electrical Engineering, Najafabad Branch, Islamic Azad University, Isfahan, Iran
E-mail: [1]so.abbasi90@gmail.com , [2] h_mahmoodian@pel.iaun.ac.ir

ABSTRACT

Microarray analysis is used in human cancer diagnosis and tumor classification. However, microarray data often have high dimensionality and small sample size. Gene selection is a significant preprocessing of the discriminant analysis of microarray data to select the most informative genes from thousands of genes. In this paper, a gene selection method proposed for cancer classification in two stages. First, the initial reduction of data by Fuzzy C-Means clustering (FCM) and filter method (T-test) is performed for dimensionality reduction**.** The next stage is to perform gene selection based on an entropy measure of eigenvalues from Principal Component Analysis on a leave-one-out basis, which is called PCA-entropy. Colon cancer, leukemia and lung datasets have been classified based on proposed gene selection algorithm by SVM and KNN classifiers. In most cases, the results show a good performance compared to the other recent studies.

Keywords: *Classification, Clustering, Entropy, Gene selection, Principal Component Analysis.*

1. INTRODUCTION

Microarray technology is a tool for analyzing gene expressions data of the genome under different conditions or in different phenotypes [1]. A gene expression dataset is sparse so that a dataset has a small number of tissue samples with thousands of genes. This extreme sparseness lead to significantly worse performance of cancer classification [2]. In addition, the expression data of some genes may contain noises that deteriorate the prediction accuracy [3]. Recent studies have shown that a small set of genes can result high prediction accuracy in cancer classification [4, 5].

As a result, the extraction of informative genes between irrelevant or redundant genes is critical for accurate classification. Gene selection is a popular tool to reduce the size of data that reduces the computational complexity and improves the classification accuracy and interpretation of the learning results [6]. Lu and Han (2003) divided gene selection methods into two categories: individual gene ranking and gene subset ranking [7].

Individual-gene-ranking method contains a criterion function that measures the discriminative power of individual genes. Examples of this method are maximum likelihood ratio [8], BW ratio [9] and Information Gain (IG) [10]. This ranking is simple but does not consider the correlation between genes [11].

Gene subset assessment searches for a subset of genes that achieves the best criterion. If the criterion function is based on the classification, it called wrapper approach. While, those methods that their functions are independent of the classification named filter methods in feature selection studies [3, 11, 12]. Typically, in wrapper approach, gene subset selection is done in interaction with classifiers such as K-Nearest Neighbors (KNN) [13] and Support Vector Machine (SVM) [4]. The goal is to find a gene subset of genes that contribute to achieve the best prediction performance for a specific learning model. To achieve this goal, there is a search algorithm around the classification model for evaluating the prediction quality of the candidate gene subsets [14]. Examples of a wrapper method are SVM-RFE [2, 15, 16] and genetic algorithm (GA) [5, 14].

The wrapper method is not as efficient as the filter method because it has very computational expensive and it runs algorithm on the dataset with high dimensions. When the data set is small, overuse of accuracy as the goodness measure may occurs over-fitting [11, 17].

Filter methods are based on criterion that depends only on the data for evaluating the significance or

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

relevance of each gene to discriminate classes. Genes are ranked base on the relevance score and genes that the top-ranking ones selected as the most relevant genes. These methods do not have much exploration on gene subset evaluation [11, 14]. Common methods of gene filtering include T-test [13, 16, 18], PCC [19], ranking of genes according to variance [20], selection according to the highest rank of the first principal component [21], and other statistical criteria. All these methods estimate the importance of each gene independently of all others while our propose method ranks genes based on criteria that take into account the other genes. This criterion is information entropy that is calculated of the eigenvalues from Principal Component Analysis (PCA). Because the computational complexity of PCA is high. Therefore, we apply Fuzzy C-Means clustering (FCM) and filter gene selection to reduce the dimensions.

In this work, we use 10-fold external cross validation to analyze the validity of the selected genes. Besides, there is an important challenge in all previous gene selection, which the final subset of genes have been generated in a cross validation method. It is clear, that the subset of genes is usually altered in cross validation procedure and there is no systematic way to select the best set of genes. Therefore, we apply a simple method for selecting informative genes from the union of genes that are participated in all cross validation iterations and genes selected have been tested on the independent samples to achieve classification accuracy. Which, Results show that the acceptable accuracy in samples classification have been achieved despite selecting a small set of genes.

The rest of the paper is organized as follows: the mathematical framework explained in section 2, Gene selection methods are described in Section 3, proposed algorithm for gene selection is presented in section 4, experimental results and conclusion are stated in section 5 and 6 respectively.

2. MATHEMATICAL FRAMEWORK

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a popular tool for dimensioning reduction and feature extraction with applications in engineering and recently it has been used in gene expressing data analysis [22].

PCA is a linear projection method that determines a new dimensional space, which captures the maximum information present in the original matrix by minimizing the mean square error of approximation. It identifies the full set of eigenvectors from given matrix. The eigenvectors can be sorted by their eigenvalues, which assigns weights to the vectors [23, 24].

Let us consider a matrix of gene expression profile $G^{m\times n}$ where element g_{ji} (j= 1,..., m and i = 1,..., n) is the expression amount of gene j in i^{th} samples and the elements of vector $c^{n\times 1}(c=\{c_i \mid c_i \in \{0,1\}, i=1,...,n)$ show the class labels of the samples.

To calculate the PCA from original matrix, first the covariance matrix of G is computed according to

$$\hat{\Sigma} = \frac{1}{m}\sum_{j=1}^{m}(g_j - \hat{\mu})(g_j - \hat{\mu})^T \quad (1)$$

Where $\hat{\Sigma}$ is the covariance matrix and $\hat{\mu}$ is the mean of gene j. The equation for PCA of G is the following:

$$\hat{\Sigma} = \phi\Lambda\phi^T \quad (2)$$

In which the columns of matrix ϕ are the eigenvectors of the matrix G so that $\phi^T\phi = I$ and Λ is a matrix which has the corresponding eigenvalues λ_j on the diagonal and zeros elsewhere (λ_j is the corresponding j^{th} eigenvalue of the matrix G), More details can be found in [24]. That eigenvalues in gene expression data called eigengene, Therefore information contained in the eigengenes in order to gene selection.

2.2 Entropy

Entropy is an amount of information that may be gained by an observation of a system and it measures variation or changes in a series of events. Alter et al., (2000) have defined SVD-based entropy of the dataset [25]. While, in this paper we calculate the entropy of the eigengenes of microarray dataset. Let us define the normalized relative values p_j of eigengenes (λ_j), as follows:

$$p_j = \frac{\lambda_j}{\sum_{k=1}^{L}\lambda_k} \quad (3)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

That L is the rank of the matrix. We use Shannon's entropy to calculate the entropy of the eigengene in following equation:

$$H = -\sum_{j=1}^{L} p_j \log(p_j) \quad (4)$$

Where p_j is normalized relative values, and H is Shannon's entropy bounded between (0, 1). $H=0$ means that a dataset has only one eigenvector, and $H=1$ means that a dataset has a uniform distribution on the eigengenes. Therefore, when a dataset is identified by only a few large eigengenes while others are significantly smaller, expected to be very low entropy and reflects large redundancy in the data. On the other hand, non-redundant datasets lead to uniformity on the eigengenes distribution and they have high entropy [26].

In this study, to investigate the effect of each gene on the eigengenes distribution we can calculate entropy by a leave-one-out principle [27]. So that entropy is computed on the set of all eigengenes for the corresponding set of the matrix without gene j (g_j). Then the j^{th} gene is replaced and $(j+1)^{th}$ gene is removed. Thus, the entropy is calculated by removing each gene.

3. GENE SELECTION METHODS

T-test, PCC (Pearson Correlation Coefficient) and SVM-RFE (Support Vector Machine Recursive Feature Elimination) methods are used to select informative genes. T-test and PCC are known as filter methods While, SVM-RFE is considered as a wrapper gene selection method.

3.1 T-test

T-test is a well-known statistical test applied to data containing two or more groups. The test assesses whether the means of two groups are statistically different from each other. Given the replicas of particular treatment and control samples, it is possible to compute the T-test for any gene g_j for differential expression [18]. T-test ranks the genes according the following equation:

$$T(g_j) = \frac{(\mu_{j1} - \mu_{j2})}{\sqrt{\frac{s_{j1}^2}{n_1} + \frac{s_{j2}^2}{n_2}}} \quad (5)$$

Let n_c be the number of training instances with class c for $c = 0, 1$, and let μ_{jc} and s_{jc}^2 be the mean and the variance of gene j calculated from the training instances with class c respectively. A gene with a larger absolute T value is more relevant to the class value [13].

3.2 Pearson correlation coefficient

Pearson Correlation Coefficient (PCC) has been used to show linearity dependence of two vectors X and Y with N dimension. This method was proposed by van't Veer to sort genes in breast cancer dataset [19]. If the gene expression values of gene j^{th} (g_j) in all samples are considered as elements of vector X, and the vector of class labels $(c^{n\times 1})$ considered such as vector Y with n dimensions, this parameter defined as:

$$PCC(g_j, c) = \frac{(\sum g_j c - (\frac{\sum g_j - \sum c}{n}))}{\sqrt{(\sum g_j{}^2 - \frac{(\sum g_j)^2}{n})(\sum c^2 - \frac{(\sum c)^2}{n})}} \quad (6)$$

The correlation is about 1 in the case of an increasing linear relationship, -1 in the case of a decreasing linear relationship, and some value in between all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables. The absolute value of PCC can be used for gene ranking.

3.3 SVM-RFE

Guyon et al. [4] presented a gene selection method by employing SVM classifier based on Recursive Feature Elimination (RFE). The SVM-RFE method, starting with the gene set containing the full genes, then eliminates repeatedly the gene that is least significant to the classifier from the gene set [16].

Decision functions that are simple weighted sums of the training samples plus a bias are called linear discriminate functions according to the following equation:

$$D(x) = W \times x + b \quad (7)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Where W is weighting vector and b is the bias value, the genes can be ranked by amount of $\|W_j\|^2$ or $|W_j|$ where W_j is j^{th} element of vector W. in (7), $x = g_j^T$ (g_j^T is the transpose of j^{th} row of matrix G) and W can be measured based on the proposed method in support vector machine classification problem and it is equal to:

$$W = \sum_{j=1}^{m} c_j \alpha_j g_j \quad (8)$$

Where c_j and α_j can be calculated by the following optimization problem:

$$\min_{\alpha} \left(\frac{1}{2}\sum_{j=1}^{m}\sum_{f=1}^{m} c_j c_f \alpha_j \alpha_f g_j^T g_f - \sum_{j=1}^{m} \alpha_j\right) \quad (9)$$

$$\text{Subject to } \sum_{j=1}^{m} c_j \alpha_j = 0 \text{ and } 0 < \alpha_j < \varepsilon$$

that ε is a constant that shows the rate of penalize misclassification and Margin errors. More details about SVM-RFE can be found in [4].

4. PROPOSED METHOD

Our proposed method is based on an entropy measure applied to eigengenes from principal component analysis (PCA). Because PCA have complex computations, so to reduce the computation time before the gene selection process, the initial reduction of the gene expression data is done.

Procedure works in two stage. At the first stage, genes are grouped into different Homogeneous groups by the Fuzzy C-Means Clustering (FCM). Potential advantage with this fuzzy clustering is that genes regulating cancers in different ways can be extracted in balance. Please refer to [2] for more details about the FCM. Then, irrelevant genes to the class value in each cluster can be eliminated based on T-test method, the combination of FCM and T-test named FCT-test.

Second stage has been considered for selecting informative genes that reduce redundancy. To select these genes, entropy from the eigengenes has been computed and genes ranked based on their contribution to the entropy. To have a set of genes with uniform distribution, the lower ranked genes should be removed. This procedure has been repeated until a set of non-redundant genes have finally selected with high accuracy in classification.

Figure 1 shows hybrid Algorithm of FCT-test and PCA-entropy that used to extract highly informative gene subsets by decreasing information loss. As mentioned in Figure 1 the process can be represented as follows:

(1) Using 10-fold external cross validation in all steps of gene selection. All procedures have been repeated 10 times and the mean of misclassification error considered as the error rate. It means that the samples in original gene expression dataset (G_O) were randomly divided into 10 subsets (9 of them as training and the left out one as the test subset).

(2) Applying FCM to group genes into h clusters ($C_1, \ldots, C_h$) in the training tissue samples space.

(3) T-test method will be employed to rank genes in each cluster C and genes with high rank are extracted, that $\mu_{TC} + \alpha\sigma_{TC}$ is set to be the threshold where μ_{TC} and σ_{TC} are mean and standard deviation of T-test, also α is the constant value between [-1, 1].

(4) All high rank genes in these clusters are union (G_S) for next step.

(5) Removing a gene g_j ($j=1,\ldots,m$) from G_S: G_R

(6) PCA is performed on the data matrix G_R to extract eigengenes for measuring entropy in other steps, as shown in (4).

(7) Steps 4, 5 and 6 is repeated until is calculate entropy of all genes inG_S.

(8) Ranking genes based on their relative contribution to the entropy.

(9) Removing a gene that has lowest rank: G_U, this step is used to select genes that have a uniform behavior on the eigengenes.

(10) Classification with test samples for measuring misclassification error.

(11) Repeat steps 4-10 until the number of genes in the G_U will be 1.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(12) Mean of misclassification error is calculated from ten times and subset of genes that has the least error and the minimum number of genes as the final set to introduce.

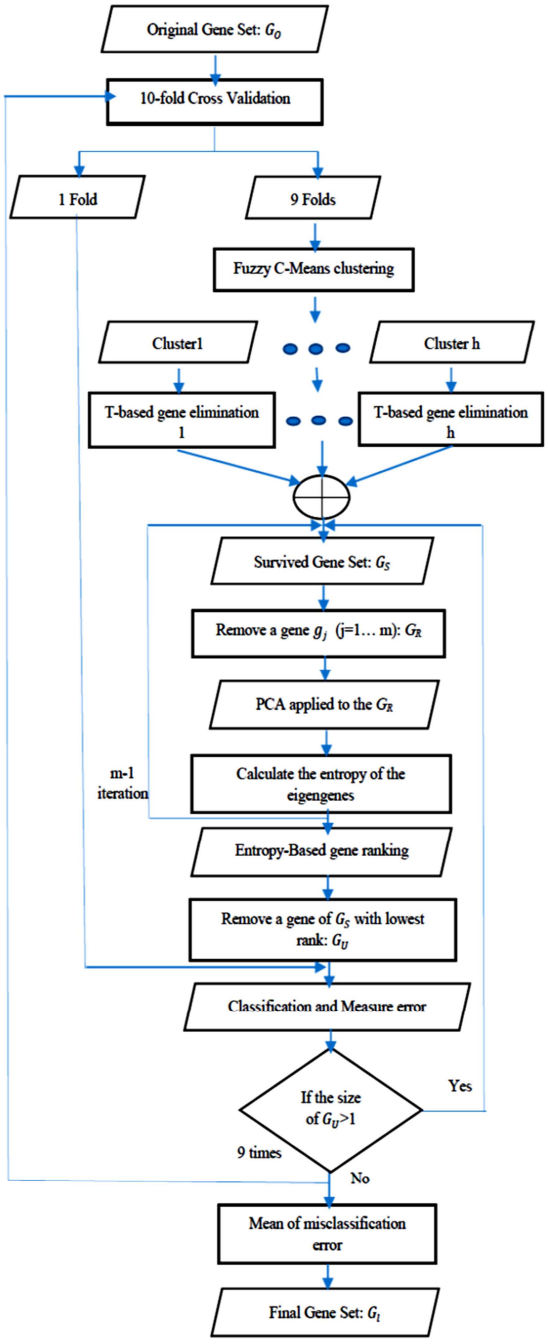

Figure 1: Algorithm For Gene Selection

5. EXPERIMENTAL RESULTS

The original dataset is simply normalized by decreasing the mean of the corresponding gene vector from each gene's expression data and then dividing by the corresponding standard deviation. To avoid over-fitting, the mean and standard deviation are calculated by using the training dataset. As a result, each gene vector has a mean of 0 and a standard deviation of 1.

5.1 Dataset

We experimented with three well-known gene expression datasets. All datasets are available at http://sdmc.lit.org.sg/GEDatasets/Datasets.html.

Colon cancer dataset

The colon cancer data set used by Alon et al. [28] contains 62 samples collected from colon-cancer of 12 patients. Among them, 40 tumor biopsies are from tumors, and 22 normal biopsies are from healthy parts of the colons of the same patients. It consist of 2000 genes. 4 of normal and 6 of tumor samples have been randomly separated as independent samples (the model was trained with 52 samples).

Leukemia cancer dataset

The Leukemia data set studied by Golub et al. [29]. It consists of two types of acute leukemia: 47 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples with over 7129 probes from 6817 human genes, which contains 38 training samples(27 ALL and 11 AML) and 34 independent samples (20 ALL and 14 AML).

Lung cancer dataset

The Lung data set studied by Gordon et al. [30]. It consists of 12533 genes (after filtering some missed data) and 181 samples, including 31 malignant pleural mesothelioma (MPM) and 150 adenocarcinoma (ADCA). Which contains 32 training samples (16 MPM and 16 ADCA) and 149 independent samples (15 MPM and 134 ADCA).

5.2 Results

We applied the proposed algorithm on the above datasets to test and compare with some other common methods. In all gene selection methods, SVM (Support Vector Machines) and KNN (K-Nearest Neighborhood) have been used as classifier to compare the results. SVM is a classifier with linear kernel function that value of *C* is 1 and for the

Journal of Theoretical and Applied Information Technology
10th October 2014. Vol. 68 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

KNN Classifier, the *K* is set to 5. To increase the validity of the experiment, 10-fold external cross validation has been used to calculate the rate of accuracy average in 10 iterations.

We consider *h* for colon dataset equal 6, leukemia dataset is 5 and lung dataset equal 4 (step 2 of section 4). These values have maximum accuracy through 2 to 10 clusters. Figures 2, 3 and 4 show accuracy rate of two classifiers (SVM, KNN) with specified number of clusters for colon, leukemia and lung datasets respectively.

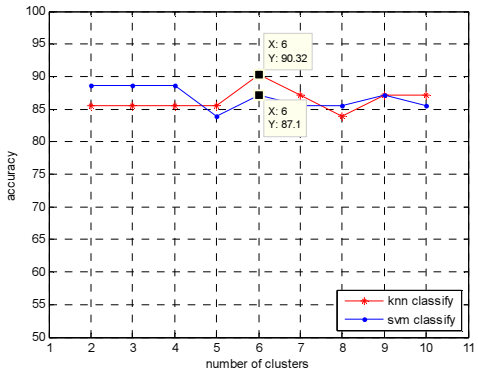

Figure 2: Accuracy from number of clusters in colon dataset.

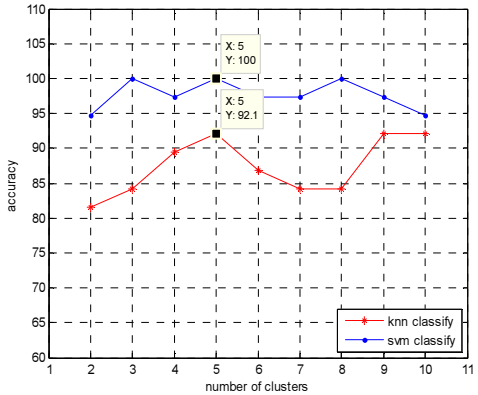

Figure 3: Accuracy from number of clusters in leukemia dataset.

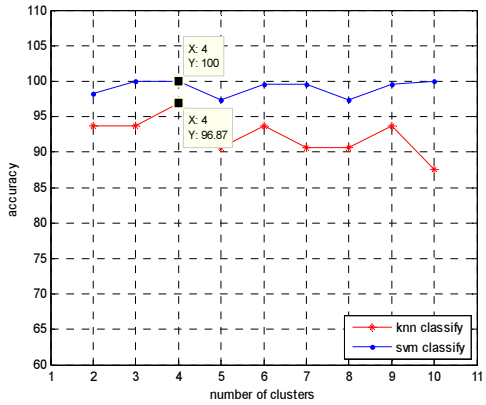

Figure 4: Accuracy from number of clusters in lung dataset.

In step 3, α is set to be 0.5. The misclassification error in the training data of this algorithm is calculated based on the mean of measured errors in 150 iteration, the error in each iteration calculated based on mean of errors in 10-fold cross validation. The percentage of the accuracy is measured by:

Percentage of accuracy (%) *=100-percentage of misclassification error (%)*

Tables 1, 2 and 3 show the percentage of accuracy in SVM and KNN classifiers of training samples with colon, leukemia and lung datasets respectively and in these tables comparing our proposed method (PCA-entropy) with 3 other gene selection methods such as PCC, T-test and SVM-RFE. The bold values in the rows illustrate the superiority accuracy of our method by subset of 4, 8 and 4 genes respectively (NG and Acc have been used to show the number of selected genes and accuracy of the methods).

Table 1: Prediction Accuracies From Different Gene Selection Methods In Colon Dataset When The Classifiers Are SVM And KNN.

Gene Selection	SVM		KNN	
	NG	Acc. (%)	NG	Acc. (%)
PCA-entropy	**14**	**90.32**	**4**	**90.32**
SVM-RFE	53	87.10	38	88.71
T-test	21	88.71	20	87.10
PCC	16	87.10	17	88.71

Table 2: Prediction Accuracies From Different Gene Selection Methods In Leukemia Dataset When The Classifiers Are SVM And KNN.

Gene Selection	SVM		KNN	
	NG	Acc. (%)	NG	Acc. (%)
PCA-entropy	**8**	**97.37**	**8**	**100**
SVM-RFE	17	94.74	49	92.11
T-test	49	94.74	26	89.47
PCC	8	94.74	54	94.74

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Prediction Accuracies From Different Gene Selection Methods In Lung *Dataset When The Classifiers Are SVM And KNN.*

Gene Selection	SVM		KNN	
	NG	Acc. (%)	NG	Acc. (%)
PCA-entropy	**4**	**100**	**4**	**100**
SVM-RFE	12	96.88	7	93.75
T-test	17	93.75	22	90.63
PCC	8	96.88	15	93.75

Table 4 compares the proposed method on colon dataset with some previous works. The accuracy rate in Table 4 shows 90.32% accuracy in classification that the proposed model in colon dataset has superiority in accuracy rate and number of selected genes comparing to the results in [15, 32]. Although there is no advantage in the accuracy rate comparing to the result in [31], the number of associated genes is the advantage of the proposed model compared with it.

Table 4: Comparing The Result Of Proposed Model With Three Recent Studies (Colon Dataset)

Colon Dataset	NG	Acc. (%)
Result of proposed method	**4**	**90.32**
Result in [15]	21	82.62
Result in [31]	35	90.63
Result in [32]	52	82.7

The accuracy rate in Table 5 on leukemia dataset shows that the proposed model has supremacy in accuracy rate and number of selected genes comparing to the results in [15, 32, 33].

Table 5: Comparing The Result Of Proposed Model With Three Recent Studies (Leukemia Dataset)

Leukemia Dataset	NG	Acc. (%)
Result of proposed method	**8**	**100**
Result in [15]	14	94.33
Result in [32]	9	90.3
Result in [33]	50	94.4

Table 6 compares the proposed model on lung dataset with some previous works and it shows that the proposed method has 100% accuracy in classification that it has superiority in accuracy rate and number of selected genes comparing to the results in [14, 32, 33].

Table 6: Comparing The Result Of Proposed Model With Three Recent Studies (Lung Dataset)

Lung Dataset	NG	Acc. (%)
Result of proposed method	**4**	**100**
Result in [14]	20	95.3
Result in [32]	16	98.1
Result in [33]	30	95.9

5.3 Final subset selection

Since the list of selected genes in cross validation procedure is different in each fold, it is unclear to select the final subset of significant genes. Therefore, a simple method has been proposed to sort the union of all selected genes in cross validation procedure. This means that genes based on the number of times that have been participated in the cross-validation folds are arranged in order. Finally, the top ranked of genes has selected.

Tables 7, 8 and 9 report the highly informative gene subset with 4, 8 and 4 genes respectively (Where NT is the number of times that the desired gene is involved in the cross validation folds).

Table 7: Gene Subset Selected By The Proposed Model (Colon Dataset)

index	NT	Probe ID
748	10	X84700
1062	8	Y00097
533	6	R42837
88	5	R50158

Table 8: Gene Subset Selected By The Proposed Model (Leukemia Dataset)

index	NT	Probe ID	index	NT	Probe ID
2642	9	U05259	5191	7	Z69881
532	7	D63874	6184	6	M26708
2354	7	M92287	1394	5	L20941
4936	7	Y00433	1909	4	M29696

Table 9: Gene Subset Selected By The Proposed Model (Lung Dataset).

index	NT	Probe ID
6821	6	36781
6980	6	36938
7069	4	37027
6571	3	36533

Table 10 shows the percentage of accuracy in the classification of independent samples with final gene subset selected. From table 10 we can find that the best accuracy for independent samples is achieved by a subset of 4, 8 and 4 genes on colon, leukemia and lung datasets respectively.

Table 10: Accuracy Independent Samples With Selected Genes.

Dataset	NG	Acc. (%)
Colon Dataset	4	90
Leukemia Dataset	8	97.06
Lung Dataset	4	95.30

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ROC analysis of 4 independent samples of colon cancer dataset has been shown in Table 11 which presents that the false negative error is less than false positive error in class prediction of independent dataset (TP, FP, FN and TN are the number of True Positive, False Positive, False Negative and True Negative errors respectively).

Table 11: ROC Analysis Of Independent Samples (Colon Dataset)

True Positive(TP)=3	**False Positive(FP) =1**
True Negative(TN)=6	**False Negative(FN) =0**
Overall Error (%)	10
Accuracy (%)	90
False positive error (%)	25
False negative error (%)	0
Sensitivity	0.75
Specificity	1

Table 12 has been shown ROC analysis of 34 independent samples of leukemia dataset.

Table 12: ROC Analysis Of Independent Samples (Leukemia Dataset)

True Positive(TP)=14	**False Positive(FP) =0**
True Negative(TN)=19	**False Negative(FN) =1**
Overall Error (%)	2.94
Accuracy (%)	97.06
False positive error (%)	0
False negative error (%)	5
Sensitivity	1
Specificity	0.95

ROC analysis of 149 independent samples of lung dataset has been shown in Table 13, which presents False Negative error is less than False Positive error in class prediction of independent dataset.

Table 13: ROC Analysis Of Independent Samples (Lung Dataset)

True Positive(TP)=127	**False Positive(FP) =7**
True Negative(TN)=15	**False Negative(FN) =0**
Overall Error (%)	4.70
Accuracy (%)	95.30
False positive error (%)	5.22
False negative error (%)	0
Sensitivity	0.95
Specificity	1

6. CONCLUSION

To extract informative gene subsets for reliable cancer classification, the combination of FCT-test and PCA-entropy has been used in this work. First, FCT-test (FCM cluster and T-test method) was applied to dimension reduction and preselect an effective gene subset from the candidate set. Second, the PCA-entropy is applied to select genes that cause a uniform distribution on the eigengenes and reduce redundancy.

Our gene selection method is tested on three microarray data sets. The experimental results show that in public dataset, a small subset has a high classification accuracy with top ranked genes on entropy ranking.

This study can be developed in two directions in the future. First, several clustering methods can be applied to divide genes into clusters. The second is to consider other methods of gene ranking in clusters like information gain to reduce the dimensions and choose significantly different gene subsets. The gene ranking methods for finding an initial gene subset may be critical for the PCA-based mechanism to discover a gene subset that can result high prediction accuracy.

REFRENCES:

[1] N. Banerjee and M. Zhang, "Identifying cooperative among transcription factors controlling cell cycle in yeast," *Nucleic Acids Research,* vol.31 , pp.7024-7031, 2003.

[2] Y. Tang, Y. Zhang, Z. Huang, X. Hu and Y. Zhao, "Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification," *IEEE Transaction on Information Technology in Biomedicine*, vol. 12, no. 6, November 2008.

[3] T. Wong, and K. Liu, "A Probabilistic mechanism based on clustering analysis and distance measure for subset gene selection," *Expert Systems with Applications*, vol. 37, Issue. 3, pp. 2144–2149, 2010.

[4] I. Guyon, J.Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn*, vol. 46, no. 1–3, pp. 389–422, 2002.

[5] S. Li, X. Wu, and M. Tan, "Gene selection using hybrid particle swarm optimization and genetic algorithm," *Soft Computing*, vol. 12, Issue. 11, pp. 1039-1048, 2008.

[6] M. Kantardzic. "Data mining: Concepts, models, methods, and algorithms," *New York: John Wiley and IEEE*, 2002.

[7] Y. Lu and J. Han, "Cancer classification using gene expression data," *Information Systems,* vol. 28, pp. 243–268, 2003.

[8] W. Li, and Y. Yang, "Zipf's law in importance of genes for cancer classification using

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

microarray data," *journal of Theoretical Biology*, vol. 219, no. 4, pp. 539-51, 2002.

[9] S. Dudoit, and J. Fridlyand, "Speed Comparison of discrimination methods for the classification of tumor using gene expression data," *Journal of the American Statistical Association*, vol. 97, pp. 77–87, 2002.

[10] Y. Su, , T. Murali, V. Pavlovic, M.Schaffer , and S. Kasif, "RankGene: identification of diagnostics genes based on expression data," *Bioinformatics,* vol. 19, no. 12, pp. 1578–1579, 2003.

[11] M. Ng and L. Chan, "Informative Gene Discovery for Cancer Classification from Microarray Expression Data," IEEE, pp.393-398, 2005.

[12] R. Ruiz, J. Riquelme, and J. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," *journal of Pattern Recognition Society*, vol.39, Issue. 12, pp. 2383-2392, 2006.

[13] T. Wong, and D. Chen, "A gene selection method for microarray data based on risk genes*," Expert Systems with Application*s, vol. 38, issue. 1, pp. 14065–14071, 2011.

[14] E. Huerta, B. Duval and J. Hao, "A hybrid LDA and genetic algorithm for gene selection and classification of microarray data," *Neurocomputing*, vol. 73, Issues. 13–15, pp. 2375–2383, 2010.

[15] E. Tapia and P. Bulacio, L. Angelone, "Sparse and stable gene selection with consensus SVM-RFE," Pattern Recognition Letters, vol. 33, Issue. 2, pp. 164–172, 2012.

[16] X. Li, S. Peng, J. Chen, B. Lü, H. Zhang and M. Lai, "SVM–T-RFE: A novel gene selection algorithm for identifying metastasis-related genes in colorectal cancer using gene expression profiles," Biochemical and Biophysical Research Communications, vol. 419, Issue. 2, pp. 148–153, 2012.

[17] H. Hong, L. Jiuyong, W. Hua, and D. Grant, "Combined Gene Selection Methods for Microarray Data Analysis," *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 4251, pp. 976-983, 2006.

[18] M. Alshalalfa, R. Alhajj, J. Rokne, "Combining Singular Value Decomposition and t-test into Hybrid Approach for Significant Gene Extraction from Microarray Data," *8th IEEE International Conference on BioInformatics and BioEngineering*, 2008.

[19] L. Van't Veer, H. Dai, M. Van de Vijver et al, "Gene expression profiling predicts clinical outcome of breast cancer," 2002

[20] J. Herrero, R. Diaz-Uriarte, and J. Dopazo, "Gene expression data preprocessing," *Bioinformatics*, vol. 19, pp. 655–656, 2003.

[21] C.H.Q. Ding, "Unsupervised Feature Selection Via Two-way Ordering in Gene Expression Analysis," Bioinformatics, vol. 19, pp. 1259–1266, 2003.

[22] H. Zou, T. Hastie and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[23] M. Wall, A. Rechtsteiner, and L. Rocha, "Singular Value Decomposition and Principal Component Analysis," *in A Practical Approach to Microarray Data Analysis*, pp. 91-109, 2003.

[24] I.T. Jolliffe, "Principal Component Analysis," *Statistical Theory and Methods*, 2nd ed., Springer, NY, 2002.

[25] O. Alter, P. Brown, and D. Botstein, , "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *in Proceedings of the National Academy of Sciences*, vol. 97, pp. 10101–10106, 2000.

[26] R. Varshavsky, A. Gottlieb, M. Linial and D. Horn, "Novel Unsupervised Feature Filtering of Biological Data," *bioinformatics*, vol. 22, no. 14, pp. 507–513, 2006.

[27] A. Gottlieb, R. Varshavsky, M. Linial and D. Horn, "UFFizir:tigeneric platform for ranking informative features," *BMC Bioinformatics*, 2010.

[28] A. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci*, vol. 96, no. 12, pp. 6745–6750, 1999.

[29] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[30] G. Gordon, R. jensen, L. Hsiao, S. Gullans, J. Blumenstock, S. Ramasvamy, W. Richards, D. Sugarbaker, and R. Bueno, " Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in lung Cancer and Mesothelioma," *Cancer research*, vol. 62. pp. 4963-67, 2002.

[31] Y. Cui, C. Zheng, J.Yang and W. Sha, "Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data," *Computers in Biology and Medicine*, vol. 43, pp. 933–941, 2013.

[32] H. mahmoodian, M.H Marhaban, R.Abdul 99ahim, R, Rosli and Saripan. I, "New Entropy-Based Method for Gene Selection," *IETE Journal of Research*, vol. 55, issue. 4, 2009.

[33] M. Seo, S. Oh, "Derivation of an artificial gene to improve classification accuracy upon gene selection," *Computational Biology and Chemistry*, vol. 36, pp. 1–12, 2012.