MAXIMUM CORRELATION MINIMUM REDUNDANCY IN WEIGHTED GENE SELECTION

Morva Ebrahimpour[1], Hamid Mahmoodian[2], Rahim Ghayour[3]

[1]Department of Electrical Engineering, Science and Research branch, Islamic Azad University, Fars, Iran, cms.sru@gmail.com
[2]Electrical Faculty, Najafabad Branch, Islamic Azad University, Daneshgah Blvd., Najafabad, Iran, h_mahmoodian@pel.iaun.ac.ir
[3]Department of Electrical Engineering, Science and Research branch, Islamic Azad University, Fars, Iran, ghayour_r@hotmail.com

ABSTRACT

Microarray technology has been recently used to analyze the behavior of thousands of genes simultaneously, and have an important role in diagnosis, detection and treatment methods. Reducing the size of the attributes (genes) with high potential for classification of microarray data analysis is thus an important goal. In this paper, we propose a new feature selection method based on maximum correlation and minimum redundancy (MCMR). In addition, a new method for weighting the genes has been introduced to select a final set of genes within all participated genes in cross validation procedure. The performance of proposed have been analyzed on two microarray data sets: colon cancer and breast cancer dataset. The results show that MCMR can increase the classification accuracy as well as reducing the number of selected genes significantly, compare to some other gene selection methods such as SNR (signal to noise ratio), PCC (Pearson Correlation Coefficient) and Fisher score.

Index Terms— Gene selection, correlation, redundancy, weighting

1. INTRODUCTION

Feature selection is usually used in pattern recognition problems such as face detection and similarity detection. Gene selection is also viewed as a data mining technique to reveal the relation between genes and diseases such as cancer. The goal of these techniques is to reduce the number of features to decrease the entire feature space. The elimination of the features or genes may improve the performance of data mining algorithms to be used, in terms of speed, accuracy, and simplicity [1]. Feature selection techniques can be categorized in two types: wrapper and filter methods [2]. In wrapper methods, the best subset of features or genes have been selected based on the prediction performance of the classification procedure and are embedded in the learning algorithm. In Filter methods, the features are assessed by their ability in discrimination of the samples individually. In gene selection, it is clear that the combinations of individually effective genes do not necessarily lead to high accuracy classification since the genes or input attributes to a classifier can be categorized into two types: redundant genes or relevant. The relevancy of a gene is measured with respect to the output class labels and relates to the importance of the gene for the classification task [3]. In addition, highly correlated genes tend to deteriorate the performance and become redundant for classification [4].

In many previous works mutual information theory has been used to measure the information of the genes [5], [6]. Usually, optimal classification accuracy is achieved by a set of maximally relevant and minimally redundant genes.

Max-Relevance and Min-Redundancy (MRMR) is a filter method which has been proposed by Ding and Peng to reduce the number of selected genes in final dataset [7]. The proposed method in [8] has a high mathematical computation cost which is considered in our work to reduce this expenditure. In this paper, a new simple method based on maximum correlation and minimum redundancy has been developed and the results show that the acceptable accuracy in the classification rates of tumor samples have been achieved despite the selection of a smaller subset of 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. As 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, in this paper we have introduced a simple method to weight the union of genes which are participated in all cross validation iterations. In this method after weighting the genes participated in the subsets of genes, final subset of genes have been selected based on the top highest ranked genes. In addition, these weighting values have been used in classification.

2. GENE SELECTION

Given a matrix of gene expression profile $G^{m\times n}$ where element $g_{ji}\ (i = 1,2,\dots n)$ and $j = 1,2,\dots m$ is the expression amount of gene j in i^{th} samples and the elements of vector $C^{n\times 1}(\mathrm{C} = \{c_i | c_i \in \{1,0\}\}, i = 1,2,\dots n)$ show the class labels of the samples, and $I_j = \{1,2,\dots,\mathrm{m}\}$ be the indexed set representing the genes.

2.1. Pearson Correlation Coefficient

Pearson Correlation Coefficient (PCC) is a measure of the linear dependence between two vectors X and Y with N dimension. This parameter is defined as:

$$PCC(X,Y) = \frac{\left(\sum XY - \left(\sum X \sum Y / N\right)\right)}{\sqrt{\left(\sum X^2 - (\sum X)^2/N\right)\left(\sum Y^2 - (\sum Y)^2/N\right)}} \quad (1)$$

The correlation is 1 in the case of an increasing linear relationship, -1 in the case of a decreasing linear relationship, and some value in between for 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. This method was proposed by van't Veer to sort genes in breast cancer dataset [9]. If the gene expression values of gene j^{th} in all samples are considered as elements of vector X, and the vector of class labels C is considered such as vector Y, the absolute value of PCC can be used for gene ranking [10].

2.2. Signal to Noise Ratio

The Signal to noise ratio (SNR) ranks the genes according to the following equation:

$$SNR(g_j) = \frac{|\mu_{j1} - \mu_{j2}|}{s_{j1} + s_{j2}} \quad (2)$$

where μ_{j1} and μ_{j2} are the mean of expression values of j^{th} gene in two different classes and S_{j1} and S_{j2} are their standard deviations. The high positive or high negative value for genes give the difference between the expression values in two classes, so, the genes with highest absolute value of SNR have more ability in classification [11].

2.3. Fisher Score

Fisher score is a statistical criterion for gene ranking that defines according to the following equation:

$$F(j) = \frac{(\mu_{j1} - \mu_{j2})^2}{{s_{j1}}^2 + {s_{j2}}^2} \quad (3)$$

where μ_{j1} and μ_{j2} are the mean of expression values of j^{th} gene in two different classes and $(S_{j1})^2$, $(S_{j2})^2$ are their variances.

3. THE PROPOSED METHOD

Our proposed algorithm has been based on selecting the maximum correlation with classes and minimum correlation with the other selected genes. According to this idea, we try to maximize the following equation:

$$g_{si} = \max_j \frac{|PCC(g_j,c)|}{\frac{1}{L}\sum_{K=1}^{L}|PCC(g_j,g_k)|} \quad (4)$$

where L is the number of selected genes in the subset, g_j are the genes which should be analyzed and g_k are the genes in the selected subset. The following figure shows the proposed algorithm:

TABLE I
THE ALGORITHM OF PROPOSED METHOD

Step 1:
O_G = original data set

Step 2: set $G_s = \{.\}$

Step 3: set $i = 1$

Step 4: Select the gene which has $\max_j(|PCC(g_j,c)|)$ and call it g_{si} where $g_j \in O_G - G_S \forall j$

Step 5:
$G_s = G_s \cup g_{si}$

Step 6:
$A(G_s)$ = the rate of accuracy in classification

Step 7:
$i = i + 1$

Step 8:
Select the gene $(\max_j \frac{|PCC(g_j,c)|}{\frac{1}{L}\sum_{K=1}^{L}|PCC(g_j,g_k)|})$ and call it g_{si} where $g_j \in (O_G - G_S) \forall j$ and $g_k \in G_s \forall k$ and L is the number of genes in G_S

Step 9:
$G_s = G_s \cup g_{si}$

Step 10: If G_s is equal to O_G stop procedure, otherwise go to step 6

Step 11: introduce the subset of genes with respect to the maximum accuracy

4. EXPERIMENTAL RESULT

4.1. Dataset

We experimented with two well known gene expression microarray datasets.

4.1.1. Colon cancer dataset
It consists of the gene expression profiles of 2,000 genes for 62 colon tumor samples. Among them, 40 tumor biopsies

are from tumors (labelled as "negative") and 22 normal biopsies are from healthy parts of the colons of the same patients (labelled as "positive") [12].

4.1.2. Breast cancer dataset

It consists of gene expression profiles of 5,166 genes for 44 low risk tumor samples and 33 high risk tumor samples [13].

4.2. Result

We applied the proposed method on two public datasets to test and compare with some other common methods. In all gene selection methods, SVM (support vector machines with linear kernel function), Knn (k-nearest neighborhood) and a classification method based on minimum distances to the mean of classes (termed Mean_C) have been used to compare the results. To increase the validity of the experiment, external tenfold cross validation has been used to calculate the rate of accuracy in 2 iterations. It is noted that the results from external cross validation are more valid with respect to internal cross validation used in some previous works such as [14]. The accuracy rates of different methods of gene selection with 95% confidence interval have been shown in tables II and III. In these tables GS1, GS2, GS3 and GS4 are representing four types of gene selection methods respectively (MCMR, Fisher Score, SNR and PCC). NG and Acc have been used to show the number of selected genes and accuracy of the models.

TABLE II
ACCURACY OF 4 TYPES OF GENE SELECTION FOR COLON DATASET

GS	NG	Acc. (%)	NG	Acc. (%)	NG	Acc. (%)
	SVM		Knn		Mean-C	
GS1	6	88.72±0.8	7	89.85±1.3	**7**	**90.63±0.7**
GS2	93	85.38±0.8	27	89.03±0.7	60	88.13±0.4
GS3	80	85.43±0.9	13	86.36±0.8	55	88.76±0.4
GS4	30	82.42±1.1	16	88.65±0.6	62	89.03±0.5

TABLE III
ACCURACY OF 4 TYPES OF GENE SELECTION FOR BREAST DATASET

GS	NG	Acc. (%)	NG	Acc. (%)	NG	Acc. (%)
	SVM		Knn		Mean-C	
GS1	39	83.60±1.9	37	85.79±1.9	**33**	**87.56±1.5**
GS2	62	75.89±1.3	25	75.63±1.1	15	76.41±1.3
GS3	131	76.80±1.1	103	77.18±1.5	8	78.31±1.0
GS4	137	76.36±1.3	18	75.62±0.8	8	77.32±1.2

Results present that the best accuracy (Bold values) have been achieved by proposed method with Mean-C classifier.

Also in table IV, V comparisons between proposed method and some previous studies are demonstrated. Results in these tables illustrate that in spite of equality in accuracy rates in most of them, the numbers of selected genes in final dataset have been reduced to 7 and 33 in colon and breast datasets respectively by our proposed method.

TABLE IV
COMPARING THE RESULT OF PROPOSED METHOD WITH THREE PREVIOUS STUDIES

colon dataset	NG	Acc. (%)
Result of proposed method	**7**	**90.63±0.73**
Result in [15]	90	91.00±5.85
Result in [16]	16	88.7
Result in [10]	14	88.25

TABLE V
COMPARING THE RESULT OF PROPOSED METHOD WITH THREE PREVIOUS STUDIES

breast dataset	NG	Acc. (%)
Result of proposed method	**33**	**87.56±1.52**
Result in [16]	70	89.47
Result in [17]	57	89.47
Result in [10]	41	89.47

4.3. 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 weighting method has been proposed to sort the union of all selected genes in cross validation procedure. Finally the top high ranked of genes has been selected based on the best number of genes introduced in table II, III. Two important features in the weighting of a gene which is in the union of all genes gathered by cross validation are: 1) times of repeat of the gene in the folds and 2) the rank of the gene in the folds of cross validation. Suppose that gene A is repeated n times $(n \leq 10)$ in10 fold CV and its position in each fold, which the gene A is exist in it, is $k_n$where k is pointing to the position of gene A in the fold n^{th}. So the weight of the gene A before normalization is calculated by:

$$WG_A = \sum_{j=1}^{n} \frac{1}{k_j} \quad (5)$$

For instance, suppose that gene A in a 3 fold cross validation is repeated 3 times such as follow:

$$\begin{bmatrix} \dots & A & \dots \\ A & \dots & \dots \\ \dots & \dots & A \end{bmatrix}$$

so the weight of gene A before normalization is:

$$WG_A = \frac{1}{2} + \frac{1}{1} + \frac{1}{3} = \frac{11}{6}$$

Normalization has been done after calculation of the weights of all genes such that the weighting factors are in range [0, 1].Weighting factors have been used not only for final subset selection of the genes, but also for Euclidean distance calculation which is used in classification such as equation (6) where w_i, g_i and $\bar{g}_i$ are weighting factors, gene expression value of test samples and mean of gene expression values for i^{th}gene respectively. To analyze the effect of weighting in classification, accuracy rates with and without weighting have been illustrated in table VI and

show the superiority of accuracy rates in weighted Euclidean distance.

$$\|.\|_w = \sqrt{\sum_{i=1}^{n} w_i (g_i - \bar{g}_i)^2} \quad (6)$$

TABLE VI
EFFECT OF GENE WEIGHTING IN CLASSIFICATION

Dataset	Acc (%)	NG	Acc (%)
Colon	90.63±0.7	7	**95.93±0.3**
Breast	87.56±1.5	33	**91.61±0.32**

Finally, 7 significant selected genes of colon cancer dataset which have been introduced by this study are: *R87126, M35531, T47377, R54097, R60217, R62549* and M26383, and 33 significant selected genes of breast cancer dataset are: *NM_002820, Contig44278_RC, NM_003561, Contig31771_RC, D26362, NM_000991, NM_013262, Contig37063_RC, Contig34634_RC, Contig47102_RC, NM_012429, NM_006366, Contig18502_RC, NM_002690, NM_006314, NM_016361, Contig38438_RC, Contig48328_RC, AF201951, NM_003882, Contig67229_RC, NM_020672, Contig36106_RC, NM_004165, NM_004887, NM_003820, NM_012168, Contig7558_RC, NM_003315, NM_015362, Contig52134_RC, NM_001747 and Contig45591_RC*. Actually, more studies are needed to recognize the effect of these genes in colon and breast cancer diseases.

4.4. Conclusions

In this study, we proposed a new feature selection method: MCMR. To assess our theory, we carried computational modelling on two well known datasets: colon cancer and breast cancer dataset. The results demonstrated an equal or higher level of accuracy with a reduced number of genes. In addition, we proposed a weighting method to overcome the challenge of gene selection in cross validation procedure and also to increase more the role of higher ranked genes in classification.

5. REFERENCES

[1] C. W. D. Justin and R. J. Victor, "Feature subset selection with a simulated annealing data mining algorithm," *J. Intell. Inform. Syst*., vol. 9, pp. 57-81, 1997.

[2] I. Inza, P. Larranaga, R. Blanco, and A. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artif. Intell. Med*., vol. 31, pp. 91-103, 2004.

[3] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artif Intell*, vol. 97, pp. 273-324, 1997.

[4] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res,* vol. 5, pp. 1205–1224, 2004.

[5] O. Kursun , C.O. Sakar, O. Favorov, N. Aydin, F. Gurgen, "Using covariates for improving the minimum Redundancy Maximum Relevance feature selection method," *Turk J Elec Eng & Comp Sci*, vol. 18, No. 6, pp. 975-989, 2010.

[6] M.J. Abdi, S.M. Hosseini, and M. Rezghi, "A Novel Weighted Support Vector Machine Based on Particle Swarm Optimization for Gene Selection and Tumor Classification," *Computational and Mathematical Methods in Medicine,* vol. 2012, 2012.

[7] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and minredundancy," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[8] P.A. Mundraand, J.C. Rajapakse, "SVM-RFE With MRMR Filter for Gene Selection," *NanoBioscience, IEEE Transactions*, pp. 31-37, 2010.

[9] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He , A.A. M. Hart , M. Mao ,H.L. Peterse , K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards and S.H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature,* vol. 415, pp. 530-536, 2002.

[10] H. Mahmoodian, M.H. Marhaban ,R. Abdulrahim, R. Rosli, I. Saripan, "Using fuzzy association rule mining in cancer classification," *Australas Phys Eng Sci Med*, vol. 34, pp. 41-54, 2011.

[11] Duda, R.O., P.E. Hart and D.G. Stork, *Pattern Classification,* Wiley-Interscience Publication, 2001.

[12] U. 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. Nat. Acad. Sci. USA*, vol. 96, pp. 6745-675,1999.

[13] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He , A.A. M. Hart , M. Mao ,H.L. Peterse , K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards and S.H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature,* vol. 415, pp. 530-536, 2002.

[14] E. Alba, , J. Garcia-Nieto, L. Jourdan, E. Talbi, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms," *Evolutionary Computation, CEC 2007. IEEE Congress*, pp. 284-290, 2007.

[15] W. Xiong, Z. Cai, J. Ma, "DSRPCL-SVM approach to informative gene analysis," *Genomics Proteomics Bioinform.*, vol. 6, No.2, pp.83-90, 2008.

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

[17] L. van't Veer, H. Dai and M.J. van de Vijver ,"Gene expression profiling predicts clinical outcome of breast cancer," *Nature* ,vol. 415, pp. 530-536, 2002.