Relevant Gene Selection Using Normalized Cut Clustering with Maximal Compression Similarity Measure

Rajni Bala[1], R.K. Agrawal[2], and Manju Sardana[2]

[1] Deen Dayal Upadhyaya College, University of Delhi, Delhi, India
[2] School of Computer and System Science, Jawaharlal Nehru University, New Delhi, India

Abstract. Microarray cancer classification has drawn attention of research community for better clinical diagnosis in last few years. Microarray datasets are characterized by high dimension and small sample size. To avoid curse of dimensionality good feature selection methods are needed. Here, we propose a two stage algorithm for finding a small subset of relevant genes responsible for classification in high dimensional microarray datasets. In first stage of algorithm, the entire feature space is divided into k clusters using normalized cut. Similarity measure used for clustering is maximal information compression index. The informative gene is selected from each cluster using t-statistics and a pool of non redundant genes is created. In second stage a wrapper based forward feature selection method is used to obtain a set of optimal genes for a given classifier. The proposed algorithm is tested on three well known datasets from Kent Ridge Biomedical Data Repository. Comparison with other state of art methods shows that our proposed algorithm is able to achieve better classification accuracy with less number of features.

Keywords: Cancer Classification, Microarray, Normalized Cut, Representative Entropy, Gene Selection.

1 Introduction

DNA microarrays have provided the opportunity to measure the expression levels of thousands of genes simultaneously. One of the most common application of microarray is to classify the samples such as healthy versus diseased by comparing the gene expression levels. Microarray data which is characterized by high dimension and small sample size suffers from curse of dimensionality[1]. For better classification there is a need to reduce the dimension. In general, among thousands of genes(features) which are monitored simultaneously only a fraction of them are biologically relevant. Therefore, efficient feature selection methods are needed to identify a set of discriminatory genes that can be used for effective class prediction and better clinical diagnose. In literature, various feature selection methods have been proposed. These methods broadly fall into two

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 81–88, 2010.

categories[2]: filter and wrapper methods. Most filter methods independently measure the importance of features without involving any classifier. So, they may not select the most relevant set of features for the learning algorithm. Also, the features set selected by filter methods may contain correlated(redundant) features which may degrade the performance of classifier. On the other hand, wrapper methods directly use the classification accuracy of some classifier as the evaluation criteria. They tend to find features better suited to the predetermined learning algorithm resulting in better performance. But, they are computationally more expensive . The conventional wrapper methods are hard to apply directly to high dimensional datasets as they require large computation time. Reducing the search space for wrapper methods will decrease the computation time. This can be achieved by first selecting a reduced set of non-redundant features from the original set of features without losing any informative feature.

In this paper, a novel two-stage approach is proposed to determine a subset of relevant and non-redundant genes for better cancer classification. Our approach first groups correlated genes and then select one informative gene from each one of these groups to reduce redundancy. This requires partitioning of the original gene set into some distinct clusters so that the genes within a cluster are highly similar(correlated) while those in different clusters are dissimilar. At the second stage a Sequential Forward Feature Selection(SFFS) method is applied to select a smaller set of discriminatory genes which can provide maximum classification accuracy.

This paper is organized as follows. Section 2 describes related work. In section 3 we present our proposed algorithm for selecting a set of informative and non-redundant genes. Experimental results on some well-known datasets are presented in Section 4. Section 5 contains conclusions.

2 Related Work

In order to achieve better classification of high dimensional microarray data, we need to determine a smaller set of discriminatory genes from a given set of genes without loosing any information. In literature, many gene selection methods have been proposed which are based on a gene ranking that assigns a score for each gene which approximates the relative strength of the gene. These methods return a set of top ranked genes and classifier is built on these genes. Among them, Golub et. al.[3] selected top genes using measure of correlation which emphasizes that a discriminatory gene must have close expression levels in samples within a class, but significantly different expression levels in samples across different classes. Other approaches that adopt the same principle with modifications and enhancements include[4] and [5]. Using ranking method, one cannot select a smallest set of discriminatory genes as the selected subset may contain many correlated genes. Few wrapper based approaches are also suggested in literature which works better for small and middle dimensional data. However they cannot be applied directly on high dimensional microarray dataset as it is computationally expensive. We can overcome this by determining a smaller set

of genes for wrapper approach. This is possible if we can group correlated or similar genes into clusters and then select a gene from each cluster which can provide us a reduced set of independent and informative genes.

In literature clustering has been employed for grouping correlated or similar genes. Many diverse clustering techniques have been suggested in literature. The most widely used techniques include hierarchical[6], k-means clustering[7] and Self-organized- maps(SOM)[8]. Each one of them is associated with advantages and disadvantages. Shi and Malik[9] have proposed an efficient normalized cut(NCUT) method based on graph theoretic approach for image segmentation. The normalized cut criterion measures both the total dissimilarity between the different groups as well as total similarity with in the groups. This can also be used for clustering of correlated genes in microarray data. In NCUT a given graph G=(V, E), where $v_i \epsilon$ V represents a gene and $e(v_i, v_j) \epsilon$ E represents similarity between two genes v_i and v_j, is divided into two disjoint sets A and B. For partitioning of the genes into A and B, the capacity of the normalized cut, Ncut is defined as

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{1}$$

where $cut(A, B) = \Sigma_{u \epsilon A, v \epsilon B} w(u, v)$ and $assoc(A, V) = \Sigma_{u \epsilon A, t \epsilon V} w(u, t)$

To determine a better partition of a cluster the value of Ncut should be minimized which is a NP-hard problem. Shi and Malik[9] have shown that this problem can be reformulated as eigenvalue problem which is given by

$$D^{-1/2}(D - W)D^{-1/2}x = \lambda x \tag{2}$$

It has been shown by Shi and Malik[9] that second smallest eigenvector of the above generalized eigenvalue system is the real valued solution to our minimum normalized cut problem. Hence, the second smallest eigenvector can be used to partition the original cluster into two clusters.

In general euclidean distance and Pearsons correlation are used as the distance or similarity measure for clustering. However, euclidean distance is not suitable to capture functional similarity such as positive and negative correlation, and interdependency[10]. It is also pointed out that it is suitable only for a data which follows a particular distribution[11]. On other hand, Pearson coefficient is not robust to outliers and it may assign a high similarity score to a pair of dissimilar genes[12]. Also both these measures are sensitive to scaling and rotation. A similarity measures called maximal information compression index[13] is suggested in literature for measuring redundancy between two features. Given two random variables x_1 and x_2, the maximal information compression index $\lambda_2(x_1, x_2)$ is defined as

$$\lambda_2(x_1, x_2) = \frac{\sigma_1 + \sigma_2 + \sqrt{((\sigma_1 + \sigma_2)^2 - 4\sigma_1\sigma_2(1 - \rho(x_1, x_2)^2)}}{2} \tag{3}$$

where σ_1, σ_2 are the variance of x_1, and x_2 respectively and $\rho(x_1, x_2)$ is the correlation between x_1 and x_2.

The value of λ_2 is zero when the features are linearly dependent and increases as the amount of dependency decreases. The measure λ_2 possesses several desirable properties such as symmetry, sensitivity to scaling and invariance to rotation which are not present in the commonly used euclidean distance and correlation coefficient.

Further splitting of a cluster, from a set of available clusters, can be decided on the basis of representative entropy measure. Representative entropy measures the amount of redundancy among genes in a given cluster. For a cluster containing p genes with covariance matrix Σ, representative entropy, H_R of a cluster is given by

$$H_R = -\Sigma_{l=1}^{p} \overline{\lambda_l} log(\overline{\lambda_l}) \tag{4}$$

where $\overline{\lambda_l} = \frac{\lambda_l}{\Sigma_{l=1}^{p} \lambda_l}$ and $\lambda_l, l = 1, 2, \ldots, p$ are the eigen values of the matrix Σ.

H_R attains a minimum value(zero) when all the eigenvalues except one are zero, or in other words when all the information is present along a single direction. If all the eigenvalues are equal, i.e. information is equally distributed among all the genes, H_R is maximum. High value of H_R represents low redundancy in the cluster. Since we are interested in partitioning the original subspace into homogeneous clusters, each cluster should have low H_R. So we split a cluster which has maximum H_R among a given set of clusters as it contains more non-redundant genes.

3 Proposed Method for Gene Selection

Here we propose a two stage algorithm to select a set of discriminatory genes to achieve better classification. Our proposed algorithm consists of two phases. The first phase involves partitioning of the original gene set into some distinct clusters so that the genes within a cluster are highly correlated to each other while those in different clusters are less correlated. The similarity measure used in NCUT clustering algorithm is maximal information compression index. We have used a hierarchical clustering in which we start with a single cluster. We split the original cluster into two clusters such that the normalized cut value is minimized. To determine which candidate cluster to further partition from the existing set of clusters, we have used representative entropy. The cluster with the maximum H_R(low redundancy) is partitioned. This process is repeated till we get the required number of clusters. Representative gene from each cluster is chosen using t-statistics. In the second phase a Sequential Forward Feature selection(SFFS) method is applied to select a smaller set of discriminatory genes which provides maximum accuracy. The criterion used in the SFFS is the accuracy of the classifier. The outline of the proposed algorithm is the following:

```
Proposed Algorithm
Input : Initial Set of genes, Class Labels C, Classifier M,
        Cluster_Size
PHASE  1 // to determine a subset of relevant and independent
             genes S
```

1. Intialization : Set G=initial set of genes ;
2. S = empty set; No_of_clusters=2; /*Set of Selected Attributes*/
3. Calculate the Similarity Matrix W using Maximal information compression index.
4. Define D where $D(i) = \sigma_j w(i,j)$
5. Solve eigenvalue problem $D^{-1/2}(D-W)D^{-1/2}x = \lambda$
6. Use the eigenvector with second smallest eigenvalues to divide the original cluster C into two clusters.
7. While (no_of_clusters≤Cluster_Size)
8. Begin
9. For each cluster calculate the representative entropy H_R
10. Choose the Cluster C_i having the maximum entropy
11. Repeat step (3)-(6) for Cluster C_i
12. No_of_clusters=No_of_clusters+1
13. End
14. For each cluster
15. Find the informative gene g_i from cluster C_i using t-statistics
16. S=S U g_i

`PHASE 2 // to determine subset of genes which provides max accuracy`

1.Initialization R=empty set
2.For each $x_j \in S$ calculate classification accuracy for classifier M.
3.$[x_k, max\_acc] = max_j\ Classification\_accuracy(x_j)$;
4.$R = R \cup x_k; S = S - x_k; R\_min = R$
5. For each x_j calculate classification_accuracy of $S \cup x_j$ for classifier M
6. $[x_k, max\_acc] = max_j\ Classification\_accuracy(x_j)$;
7. $R = R \cup x_k; S = S - x_k$
8. If new_max_acc ≥ max_acc then R_min=R;max_acc=new_max_acc;
9. Repeat 5-9 until max_acc=100 or S = empty set
10. Retum R_min, max_acc

4 Experimental Setup and Results

To test the effectiveness of our proposed algorithm, we have carried out experiments on three well known datasets from Kent Ridge Biomedical Data Repository[14]. The details of these datasets are given in Table 1. Datasets are normalized using Z-score before carrying out experiments.

Table 1. Datasets Used

Dataset	Samples	Genes	Classes
Colon	62	2000	2
SRBCT	83	2308	4
Prostate	102	5967	2

Table 2. Maximum classification accuracy along with number of genes for different classifiers using different cluster size methods

No.of Clusters	LDC	QDC	KNN	SVM
30	93.54(18)	91.93(24)	95.16(13)	95.16(14)
40	91.93(4)	93.54(8)	95.16(6)	93.54(5)
50	91.93(4)	95.16(6)	96.77(11)	95.16(10)
60	98.38(32)	95.16(7)	95.16(8)	96.77(19)

a. Colon dataset

No.of Clusters	LDC	QDC	KNN	SVM
30	97.59 (20)	96.38 (10)	100 (7)	100 (4)
40	100 (31)	97.59 (11)	100 (6)	100 (4)
50	100 (33)	97.59 (11)	100 (6)	100 (5)
60	98.79 (9)	97.59 (12)	100 (6)	100 (6)

b. SRBCT dataset

No.of Clusters	LDC	QDC	KNN	SVM
30	93.13 (3)	96.07 (3)	98.03 (7)	97.06 (14)
40	96.07 (8)	96.07 (3)	96.07 (3)	99.01 (15)
50	96.07 (8)	96.07 (3)	96.07 (3)	98.03 (17)
60	97.05 (5)	97.05 (19)	99.01 (7)	96.07 (3)

c. Prostate dataset

Table 3. Comparison of Maximum Classification accuracy and number of genes selected with other state of art methods

SRBCT	PROSTATE	COLON
Proposed Method 100(4)	Proposed Method 99.01(7)	Proposed method 98.38(32)
GS2+SVM[4] 100(96)	GAKNN[17] 96.3(79)	PSO+ANN[4] 88.7
GS1+SVM[4] 98.8(34)	BIRS[18] 91.2(3)	Yuechui and Yao[20] 90.3
Chos+SVM[4] 98.8(80)		BIRSW[18] 85.48(3.50)
Ftest + SVM[4] 100(78)		BIRSF[18] 85.48(7.40)
Fu and Liu[15] 100(19)		
Tibsrani[19] 100(43)		
Khan[16] 100(96)		

Genes are clustered using NCUT based on maximal information compression index as similarity measure. From each cluster the most informative gene is selected using t-statistics. After collecting a pool of genes, a Forward Feature Selection method is applied to get a sub-optimal set of genes which provides maximum classification accuracy. Classification accuracy is calculated using leave-one-out cross validation. The different classifiers used in our experiments are linear discriminant classifier(LDC), quadratic discriminant classifier(QDC), k-nearest neighbor(KNN) and support vector machine(SVM). For KNN the optimal value of k is chosen. Linear kernel is used in SVM. The experiment was conducted for different cluster sizes. The cluster sizes considered in our experiments are 30, 40, 50 and 60. Table 2 depicts the maximum classification accuracy along with the number of genes obtained by our proposed algorithm for different cluster sizes. We can observe the following from Table 2:

1. For Colon dataset a maximum accuracy of 98.38% is achieved with 32 genes for LDC classifier. The maximum accuracy of 96.77% is achieved for KNN and SVM with 11 and 19 genes respectively. For QDC a maximum accuracy of 95.16% is achieved with 6 genes.
2. For SRBCT dataset maximum classification accuracy of 100% is achieved for LDC, KNN and SVM with 31 , 6 and 4 genes respectively. For QDC a maximum accuracy of 97.59% is achieved with 11 genes.
3. For prostate dataset maximum classification accuracy of 99.01% is achieved for KNN and SVM with 7 and 15 genes respectively. For QDC and LDC a maximum accuarcy of 97.05% is achieved with 19 and 5 genes respectively.
4. The performance of KNN is better in terms of number of genes in comparison to other classifiers LDC, QDC and SVM for all three data sets.

It is observed that our proposed algorithm is able to achieve a high classification accuracy with small number of genes. In Table 3, we have also compared performance of our proposed method in terms of classification and number of genes with some already existing gene selection methods in literature[15],[16],[17],[18], [19],[4]and [20]. From Table 3, it can be observed that the performance of our proposed algorithm is significantly better in terms of both classification accuracy and number of genes selected.

5 Conclusion

In this paper, we have proposed a two stage algorithm for finding a small subset of discriminatory genes responsible for classification in high dimensional microarray datasets. The first stage involves partitioning of the original gene set into some distinct subsets or clusters so that the genes within a cluster are highly correlated to each other while those in different clusters are less correlated. We have used NCUT clustering algorithm which is based on graph theoretic approach and requires computation of similarity measures between genes. We have used a novel similarity measure maximal information compression index which is not used for microarray datasets earlier. Most informative gene from each cluster is then selected to create a pool of non-redundant genes. The size of this set is significantly small which allows us to use a wrapper approach at the second stage. The use of wrapper method at the second stage gives a smaller subset of genes which provides better classification accuracy. Experimental results show that our proposed method is able to achieve a better accuracy with a small number of genes. Comparisons with other state of art methods show that our proposed algorithm is able to achieve better or comparable accuracy with less number of genes with all the three datasets.

References

1. Bellman, R.: Adaptive Control Processes. A Guided Tour. Princeton University Press, Princeton (1961)
2. Guyon, I., Elisseeff, A.: An Introduction to Variable and feature Selection. Journal of Machine Learning Research (3), 1157–1182 (2003)

3. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
4. Yang, K., Cai, Z., Li, J., Lin, G.H.: A stable gene selection in microarray data analysis. BMC Bioinformatics 7, 228 (2006)
5. Cho, J., Lee, D., Park, J.H., Lee, I.B.: New gene selection for classification of cancer subtype considering within-class variation. FEBS Letters 551, 3–7 (2003)
6. Eisen, M.B., Spellman, T.P.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95(25), 14863–14868 (1998)
7. Tavazoie, S., Huges, D., Campbell, M.J., Cho, R.J., Church, G.M.: Systematic determination of genetic network architecture. Nature Genet., 281–285 (1999)
8. Kohonen, T.: Self-organizing maps. Springer, Berlin (1995)
9. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern analysis and machine Intelligence 22(8), 888–903 (2000)
10. Jiang, D., tang, C., Zhang, A.: Cluster Analysis for gene expression data: A survey. IEEE Trans. Knowledge and Data Eng. 16, 1370–1386 (2004)
11. Yu, J., Amores, J., Sebe, N., Tian, Q.: Toward Robust Distance Metric analysis for Similarity Estimation. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (2006)
12. Heyer, L.J., Kruglyak, S., Yooseph, S.: Exploring Expression Data: identification and analysis of coexpressed genes. Genome Research 9, 1106–1115 (1999)
13. Mitra, P., Murthy, C.A., Pal, S.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Analysis and Machine Intelligence 24(3), 301–312 (2002)
14. Kent Ridge Biomedical Data Repository, `http://datam.i2r.a-star.edu.sg/datasets/krbd/`
15. Fu, L.M., Liu, C.S.F.: Evaluation of gene importance in microarray data based upon probability of selection. BMC Bioinformatics 6(67) (2005)
16. Khan, J., Wei, S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F.: Classification and diagnosis prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7, 673–679 (2001)
17. Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G.: Gene Selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17(12), 1131–1142 (2001)
18. Ruiz, R., Riqueline, J.C., Aguilar-Ruiz, J.S.: Incremental wrapper based gene selection from microarray data for cancer classification. Pattern Recognition 39(12), 2383–2392 (2006)
19. Tibsrani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centriods of gene expression. Proc. Natl Acad. Sci., USA (99), 6567–6572 (2002)
20. Yuechui, C., Yaou, Z.: A novel ensemble of classifiers for microarray data classification. Applied Soft computing (8), 1664–1669 (2008)