Neuro-fuzzy Methodology for Selecting Genes Mediating Lung Cancer

Rajat K. De[1] and Anupam Ghosh[2]

[1] Department of Machine Intelligence Unit,
Indian Statistical Institute, Kolkata, India
rajat@isical.ac.in
[2] Department of Computer Science and Engineering,
Netaji Subhash Engineering College, Kolkata, India
anupam.ghosh@rediffmail.com

Abstract. In this article, we describe neuro-fuzzy models under supervised and unsupervised learning for selecting a few possible genes mediating a disease. The methodology involves grouping of genes based on correlation coefficient using microarray gene expression patterns. The most important group is selected using existing neuro-fuzzy systems [1,2,3,4,5]. Finally, a few possible genes are selected from the most important group using the aforesaid neuro-fuzzy systems. The effectiveness of the methodology has been demonstrated on lung cancer gene expression data sets. The superiority of the methodology has been established with four existing gene selection methods like SAM, SNR, NA and BR. The enrichment of each gene ontology category of the resulting genes was calculated by its P-value. The genes output the low P-value, and indicate that they are biologically significant. According to the methodology, we have found more true positive genes than the other existing algorithms.

1 Introduction

Gene selection refers to the task of selecting some informative genes. The goal of gene selection algorithms is to filter out a small set of informative genes that best explains experimental variations. It is much cheaper to focus on a small number of informative genes, from the whole genome, that can differentially express in various diseases. Therefore, using effective gene selection methods, a small list of highly informative genes can be discovered from whole gene set [6], which have direct/indirect role in causing diseases. Thus, these genes can be utilized to construct the classifier for discriminating disease patterns. From data mining point of view, the task of gene selection can be viewed as that of feature selection that is widely used in data preprocessing stage [7,8]. However, gene selection, unlike feature selection in the area of machine learning literature, is characterized by the great difference between a huge number of genes and very small number of samples.

Several attempts have been made during the past several years for developing methodologies or using feature selection algorithms that select informative

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 388–393, 2011.

genes from microarray gene expression data. These genes improve the efficiency of the system in terms of disease prediction accuracy. The attempts include Noise sampling method [9], Bayesian regularization model (BR) [10,11], Significance Analysis of Microarray (SAM) [12], Signal-to-Noise Ratio (SNR) [13], Neighborhood analysis (NA) [9]. Most of the above methods are claimed to be capable of extracting a set of highly informative genes [6].

The present article is an attempt in this regard and provides neuro-fuzzy methodology for gene selection. The methodology involves grouping of genes using correlation coefficient, followed by selecting the most important group using the neuro-fuzzy models. Then the most informative genes are selected using neuro-fuzzy methods again. It is to be mentioned here that neuro-fuzzy methods have been developed in [1,2,3,4,5] for the purposed of feature selection. Neuro-fuzzy methodologies are applicable in data rich environment, i.e., if the number of samples is quite large compared to the number of features. However, in the present problem, the number of microarray measurements (samples) is quite low compared to the number of genes (features). In order to tackle this situation, we have proposed a way of generating more data so that neuro-fuzzy systems can be effectively used.

Incorporation of fuzzy set theory enables one to deal with uncertainties in different tasks of a pattern recognition system, arising from deficiency (e.g., vagueness,incompleteness, etc.) in information, in an efficient manner. Artificial Neural Networks, having the capability of fault tolerance, adaptivity, and generalization, and scope for massive parallelism, are widely used in dealing with learning and optimization tasks. In the area of pattern recognition, neuro-fuzzy approaches have been attempted mostly for designing classification/clustering/feature selection or extraction methodologies; the problem of gene selection has not been addressed.

The effectiveness of the proposed methodology, along with its superior performance over several of other methods, is demonstrated using one microarray gene expression data set dealing with human lung. The performance comparison is made using t-test and P-value (in terms of the number of enriched attributes).

2 Methodology

Here we describe the proposed methodology for gene selection. The task of gene selection has been considered as the task of feature selection in pattern recognition literature. Since the number of genes is very large compared to the number of measurements (samples), we have grouped the genes based on the correlation coefficient. Then the groups are evaluated using neuro-fuzzy systems under supervised (NFS) [2,3,4,5] and unsupervised (NFU) [1,3,4,5] learning, and the most important group is selected. Finally, important genes are selected from the most important group using NFS and NFU. For details of NFS and NFU, one may refer to [1,2,3,4,5].

Let us consider a set $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ of n genes for each of which p expression values in normal samples and q expression values in diseased samples

are given. We now compute the correlation coefficient among these genes based on their expression values in normal samples. Thus the correlation coefficient r_{ij} between ith and jth genes is given by

$$r_{ij} = \frac{\sum_{k=1}^{p}(x_{i_k} - m_i) \times (x_{j_k} - m_j)}{(\sum_{i=1}^{p}(x_{i_k} - m_i)^2)^{1/2} \times (\sum_{i=1}^{p}(x_{j_k} - m_j)^2)^{1/2}} \tag{1}$$

Here m_i and m_j are the mean of expression values of ith and jth genes, respectively, over normal samples. The term x_{i_k} denotes kth expression value of ith gene. The correlation coefficient assumes values in the interval $[-1, 1]$. When $r_{ij} = -1$ $(+1)$, there is a strong negative (positive) correlation between ith and jth genes. Genes with high positive correlation are placed into the same group. The main idea of grouping is as follows. If a gene is strongly correlated with another gene, then the expression value of one of them is linearly dependent on that of the other. In that case, we may consider one of them as a representative gene and ignore the other.

We now find out the groups of genes in such a way that the genes in the same groups are strongly correlated. In order to do this, we have computed r_{ij} (Equation (1)) for each pair of genes. If $r_{ij} \geq 0.75$, then we place these genes in the same group. In this way, the first group of genes is created. Then we continue in the same way on the remaining genes, and the second group is created. We proceed in this way till all the genes are placed in one of the groups. Note that some singleton groups may also be formed by this process. Thus we have a few groups containing the genes. This process reduces the number of genes and hence reduces the curse of dimensionality. It is to be mentioned here that one may choose other high value (< 1) instead of 0.75 as the threshold.

We now use NFS or NFU [1,2,3,4,5], in the next step, for selecting the most important group. Since the number of measurements (samples) is quite low, we need to generate more data. This will be helpful to create a data rich environment where artificial neural networks are more effective. We proceed as follows.

After grouping, let us assume that we have K groups, viz., $G_1, G_2, \ldots, G_k, \ldots, G_K$ such that $|G_k| = n_k, \forall k$. Let us also assume that a member of G_k is represented by $\mathbf{g}_k = [g_{k1}, g_{k2}, \ldots, g_{kl}, \ldots, g_{kp}]^T$ such that and $\mathbf{g}_k = \mathbf{x}_j$, for some value of j. Then we choose one gene for each group and form a vector $\mathbf{v} = [v_1, v_2, \ldots, v_k, \ldots, v_K]^T$, where $v_k = g_{kl}$, lth sample value. That is, the components of vector $\mathbf{v}$ is the lth normal sample value of K genes that are drawn from each group G_k. Similarly, other $\mathbf{v}$s are formed by the other normal sample values and we have a total of p such vectors for each draw of K genes, one from each group. We thus create a set S of all such vectors from normal samples $\mathbf{v}$ so that the numbers of such vectors in S is

$$s = |S| = p \times \prod_{k=1}^{K} n_k \tag{2}$$

Similarly, another set S' of vectors $\mathbf{v}'$ is created from the diseased samples such that

$$s' = |S'| = q \times \prod_{k=1}^{K} n_k \tag{3}$$

Now we have two sets, S and S' of vectors $\mathbf{v}$ and $\mathbf{v}'$ respectively. For NFS, we consider that normal and diseased samples form two classes, viz., *normal* and *diseased*. We take the number of input nodes as K, and the other nodes along which the architecture of the system is decided automatically [2,3,4,5]. In the case of NFU, the number of input nodes is $2K$, and the other nodes along which its architecture is decided automatically [1,3,4,5]. The first K nodes receive the vectors $\mathbf{v}$ as their inputs and second K nodes receive $\mathbf{v}'$. Thus the number of such presentations is $s \times s'$. After learning in both the systems, we get weight values representing importance of each group. Thus the most important group is selected for which the weight value is the highest.

Once the most important group is selected, only the genes in this group are considered. If the number of genes in the most important group is N ($N << n$), the numbers of input nodes in NFS and NFU are N and $2N$, respectively. The remaining parts of the architecture of both the systems are determined automatically. As in the case of selection of groups, the number of classes for NFS is 2. For NFU, the first N input nodes receive expression values of genes (in the most important group) of normal samples and the next N nodes receive that of diseased samples. Thus the number of presentations in NFU is $p \times q$. After learning, we get weight values corresponding to each gene representing its importance. Then we select a few important genes based on the connection weights of NFS or NFU.

3 Results

In this section, the effectiveness of the proposed methodology is demonstrated on human lung expression data [14]. A comparative analysis with SAM, SNR, BR, NA is also included.

We have found 6 groups, containing 1659, 1247, 1290, 741, 666, and 1526 genes respectively. The group containing 1659 genes has been selected as the most important group by both NFS and NFU. Applying NFS and NFU, we have found 30 and 32 genes respectively. Among these genes, we have found 22 genes that are present in both the results. Finally, we have selected 20 most important genes based on the connection weights of NFS and NFU. These selected genes are then evaluated for their role in causing lung cancer through computing the number of functional enrichments. We have performed t-test for the genes identified by other gene selection algorithms like SAM, SNR, NA and BR. But highly significant (99.9% significance level) genes like PFKP, TYMS, IARS, and HLA-B are not present in the first twenty selected genes by these methods. This result suggests that NFS and NFU are able to find more significant genes than the existing methods.

Table 1. Comparative results on number of attributes of various sets of genes

No. of Attributes							
Dataset	**Gene set**	**NFS**	**NFU**	**SAM**	**SNR**	**NA**	**BR**
Lung	First 5	62	80	14	21	26	27
expression	First 10	75	76	13	9	15	21
Data	First 15	80	82	30	14	16	16
	First 20	82	75	28	13	15	16

In order to validate the results statistically, we have applied t-test on the genes identified by NFS, NFU, SAM, SNR, NA and BR. Here we have identified some important genes like CALCA (4.02), PFKP (5.78), TYMS (3.98), IGFBP3 (6.98), IARS (5.98), HBB (7.08), HLA-B (5.42), SFTPA2 (6.89), and TNF (4.23). The number in the bracket indicates t-value corresponding to the gene. The t-value of these set of genes exceeds the value for $p = 0.001$. It indicates that these set of genes are highly significant (99.9% level of significance). Similarly, genes like IGHG3 (2.67), PRKACA (2.89), SORT1 (2.76), MEN1 (3.15), SFTPA1 (2.92) and IGHM (3.25) exceeds the t-value for $p = 0.01$. This means that these genes are significant at the level of 99%. Likewise, RPLP0 (2.12), SMCIL1 (2.07), MGP (2.31), RNASE1 (2.43), SFTPC (2.37), and HLA-DRA (2.27) genes are important at the level of 95% significance.

In our study, the enrichment of each GO category [15] for each of the genes has been calculated by its P-value. A low P-value indicates that the genes belonging to the enriched functional categories are biologically significant. Here only functional categories with $P - value < 5.0 \times 10^{-5}$ are considered. We have made comparative study, with other methods, viz., SAM, SNR, NA, BR in terms of their ability to identify functionally enriched genes. Table 1 shows the number of functionally enriched attributes corresponding to these methods for different sets of genes. It is found that NFS and NFU performed the best. These results show that the proposed methodology has been able to select more important genes responsible for mediating a disease than the other methods considered here.

4 Conclusions

In this article, we have provided a methodology based on neuro-fuzzy models for the selection of genes whose over/under expression may cause diseases. The methodology, first of all, finds various groups of genes based on correlation values. This is followed by determining the most important group. The genes in this groups are evaluated using NFS and NFU. This results in important genes is mediating development of a particular disease. The effectiveness of the methodology is demonstrated on various gene expression data sets where each gene is treated as a feature. The most important genes obtained by the methodology are also verified using their P-values [15]. The superior performance of the methodology compared to some existing ones have been shown. The results are verified using t-test, some existing results.

References

1. Pal, S.K., De, R.K., Basak, J.: Unsupervised feature evaluation: A neuro-fuzzy approach. IEEE Trans. Neural Networks 11, 366–376 (2000)
2. De, R.K., Basak, J., Pal, S.K.: Neuro-fuzzy feature evaluation with theoretical analysis. Neural Networks 12, 1429–1455 (1999)
3. Basak, J., De, R.K., Pal, S.K.: Fuzzy feature evaluation index and connectionist realization-ii: Theoretical analysis. Information Sciences 111, 1–17 (1998)
4. Basak, J., De, R.K., Pal, S.K.: Unsupervised feature selection using neuro-fuzzy approach. Pattern Recognition Letters 19, 997–1006 (1998)
5. Pal, S.K., Basak, J., De, R.K.: Fuzzy feature evaluation index and connectionist realization. Information Sciences 105, 173–188 (1998)
6. Bezdek, J., Pal, S.K.: Fuzzy models for pattern recognition. IEEE Press, New York (1992)
7. Kraaijveld, M.A., Mao, J., Jain, A.K.: A nonlinear projection method based on kohonen topology preserving maps. IEEE Trans. Neural Networks 6, 548–559 (1995)
8. Rubner, J., Tavan, P.: A self-organizing network for principal component analysis. Europhys. Lett 10, 693–698 (1989)
9. Golub, T.R., Slonim, T.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Downing, J.R., Caliguri, M.A., Bloomeld, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
10. Shevade, S.K., Keerthi, S.S.: A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics 19, 2246–2253 (2003)
11. Cawley, G.C., Talbot, N.L.C.: Gene selection in cancer classification using sparse logistic regression with bayesian regularization. Bioinformatics 22, 2348–2355 (2006)
12. Goh, L., Song, Q., Kasabov, N.: A novel feature selection method to improve classification of gene expression data. In: APBC (2004)
13. Tusher, V.G., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA 98 (2001)
14. Beer, G.D., et al.: Gene-expression profilespredict survival of patients with lung adenocarcinoma. Nature Medicine 8, 816–823 (2002)
15. Kim, D.W., Lee, K.H., Lee, D.: Detecting clusters of different geometrical shapes in microarray gene expression data. Bioinformatics 21, 1927–1934 (2005)