Neurocomputing 133 (2014) 122–140

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Selection of genes mediating certain cancers, using a neuro-fuzzy approach

Anupam Ghosh [a], Bibhas Chandra Dhara [b], Rajat K. De [c,*]

[a] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152, India
[b] Department of Information Technology, Jadavpur University, Kolkata, India
[c] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 9 May 2013
Received in revised form
24 September 2013
Accepted 4 November 2013
Communicated by L. Kurgan
Available online 23 January 2014

Keywords:
Artificial neural networks
Fuzzy membership functions
GO attributes

ABSTRACT

In this article, we propose a methodology for selecting genes that may have a role in mediating a disease in general and certain cancers in particular. The methodology, first of all, groups an entire set of genes. Then the important group is determined using two neuro-fuzzy models. Finally, individual genes from the most important group are evaluated in terms of their importance in mediating a cancer, and important genes are selected. A method for multiplying existing data is also proposed to create a data rich environment in which neuro-fuzzy models are effective. The effectiveness of the proposed methodology is demonstrated using five microarray gene expression data sets dealing with human lung, colon, sarcoma, breast and leukemia. Moreover, we have made an extensive comparative analysis with 22 existing methods using biochemical pathways, *p*-value, *t*-test, *F*-test, sensitivity, expression profile plots, *pi*-GSEA, Fisher-score, KOGS, SPEC, *W*-test and *BWS*, for identifying biologically and statistically relevant gene sets. It has been found that the proposed methodology has been able to select genes that are more biologically significant in mediating certain cancers than those obtained by the others.

1. Introduction

In the present context, the selection of disease mediating genes is based on their expression patterns in the normal as well as in the diseased sample. A genome wide expression pattern for a particular tissue is generated through microarray experiments. These experiments involve uncertainty in the preparation of microarray chip as well as sample collection and hybridization experiments. Moreover, a microarray experiment provides an indication of average expression values of genes but not their expression values. Thus, uncertainty is involved in this kind of data.

Gene selection refers to the task of selecting some important genes that best explain experimental variations [1]. It is much cheaper to focus on a small number of important genes that have different expression patterns in diseased samples. Therefore, using effective gene selection methods, a small list of highly important genes can be isolated from the whole genome, which have a direct/indirect role in causing diseases [2,3]. We call these important genes disease mediating genes. From a gene expression point of view, disease mediating genes refer to those that have changed their behavior from normal conditions in which symptoms of the disease under consideration are not present. Then, these genes can be explored to identify the cause of the disease and thereby may aid rational drug design.

Fuzzy set theory enables one to deal with uncertainties arising from deficiencies like inexactness, vagueness, and uncertainty in information in an efficient manner. Thus, fuzzy set theory forms a tool for handling these uncertainties. Artificial neural networks (ANN), having the capability of fault tolerance, adaptivity, generalization, and scope for massive parallelism, are widely used in dealing with learning and optimization tasks. Moreover, artificial neural networks (ANNs) are used to solve problems which are intractable in nature. The neuro-fuzzy models, a hybridization of the concepts of fuzzy sets and artificial neural networks, can tackle such intractability and uncertainty, in an efficient way [4]. Thus, we have used neuro-fuzzy models here.

The present article is an attempt in this regard and provides a new methodology involving neuro-fuzzy models for identifying genes mediating a disease in general and certain cancers in particular. The methodology involves grouping of genes using correlation coefficient, followed by selecting the most important group using the neuro-fuzzy models. We call these models neuro-fuzzy Model-1 (NFM-1) and neuro-fuzzy Model-2 (NFM-2). Then the most important genes from the selected most important group are identified using neuro-fuzzy models again. It is to be

* Corresponding author.
E-mail addresses: anupam.ghosh@rediffmail.com (A. Ghosh), bibhas@it.jusl.ac.in (B. Chandra Dhara), rajat@isical.ac.in (R.K. De).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.023

mentioned here that these neuro-fuzzy models have been developed in [5–7] for the purpose of feature selection. The methodology is applicable in a data rich environment, i.e., if the number of samples is quite large compared to the dimension of each sample. However, in the present problem, the number of microarray measurements (samples) is quite low compared to the number of genes (dimension). In order to overcome this problem, we have proposed a way of generating more data from the given microarray gene expression measurements.

The effectiveness of the proposed methodology, along with its superior performance over several other methods, has been demonstrated using five microarray gene expression data sets dealing with cancers related to human lung, colon, breast, sarcoma and leukemia. An initial set of results using lung expression data has been published in [8]. The existing methods, with which the results have been compared, are Bayesian regularization (BR) model [9,10], significance analysis of microarray (SAM) [11], signal-to-noise ratio (SNR) [12], neighborhood analysis (NA) [13], support vector machine (SVM) [14,1], Gaussian mixture model (GMM) [15], hidden Markov model (HMM) [16], constructive approach for feature selection (CAFS) [17], entropy based penalized logistic regression (Entropy-PLR) [18], minimum sum of square of the correlation (MSC) and maximum value of square of the correlation (MMC) with Naive Bayes classifier (NBC) and nearest mean scaled classifier (NMSC) (i.e., NBC-MSC, NBC-MMC, NMSC-MSC and NMSC-MMC) [19], leave-one-out calculation sequential forward selection (LOOCSFS) [20], gradient based leave-one-out gene selection (GLGS) [20], least square (LS) bound measure with sequential forward selection (SFS) and sequential floating forward selection (SFFS) [21], a method in the R package (VarSelRF) [22], and partial least squares (PLS) SlimPLS [23]. The performance comparison has been made using *t*-test, *F*-test and *p*-value (in terms of the number of enriched attributes or GO (gene ontology) attributes). In addition, we have used biological and statistical measurements like *pi*-GSEA [24], Fisher-score [25], KOGS [26], SPEC [27], *W*-test [28,30], *BWS* [29] for identifying the biologically and statistically relevant gene set.

2. Some existing methods

In the present study, we have proposed neuro-fuzzy models for identification of cancer mediating genes. We have made a survey on existing gene selection methods for comparative analysis. Among them, we have found some gene selection methods that are using SAM, SNR, BR, NA, SVM, GMM, HMM, CAFS, Entropy-PLR, NBC-MSC, NBC-MMC, NMSC-MSC, NMSC-MMC, LOOCSFS, GLGS, SFS-LSBOUND and SFFS-LSBOUND, VarSelRF, and SlimPLS. Significance analysis of microarray (SAM) [11] identifies genes with statistically significant changes in expression values by using a set of gene-specific *t* tests. Each gene is assigned a score on the basis of its change in the gene expression value. Genes with scores greater than a threshold value are deemed potentially significant. The percentage of such genes identified by chance is the false discovery rate (FDR). In order to estimate FDR, nonsense genes are identified by analyzing permutations of the measurements. The threshold value can be adjusted to identify smaller or larger sets of genes, and FDRs are calculated for each set. The disadvantage of SAM lies in the permutation stage where all the genes are put into one group for evaluation. This requires an expensive computation. Moreover, it probably confuses the analysis because of the noise in gene expression data.

The method based on the signal-to-noise ratio (SNR) [12] is applied to rank the correlated genes according to their discriminative power. The method starts with the evaluation of a single gene, and iteratively searches for other genes based on some statistical criteria. The genes with high SNR scores are chosen as the important ones. A limitation of this method is that many genes with very low correlation coefficient values are removed by the ranking criterion, because the correlation coefficient of genes is only measured by one gene to others. However, it is very likely that some of these abandoned genes may be useful when they are combined for measuring the correlation values. SNR measurement is affected by the size of the variables. When there are more variables, the mean and variance of the remaining variables of other classes are dependent on the data dispersion and the number of variables, which affects SNR ranking of the significant variables due to the general increase in noise in the data. If the number of variables can be reduced significantly, the method is more capable of detecting and ranking a smaller number of significant variables.

Neighborhood analysis (NA) [13] is a method for clustering multivariate data into distinct classes based on a given distance metric over the data. Functionally, it serves the same purposes as the K-nearest neighbor algorithm. Golub [13] applied NA to identify predictor classes defined on the different responses to therapy. The main disadvantage of the method is that it cannot detect any significant correlation. Although this failure may be due to the limited number of genes included in the study [13], it is also possible that the "response" phenotype is too complex to be associated with a cluster of genes, and a more elaborate relationship may exist between response to therapy and gene expression.

Shevade and Keerthi [9] have developed a gene selection algorithm based on sparse logistic regression (SLogReg) and provide a simple but efficient training procedure. The degree of sparsity is determined by the value of a regularization parameter, which must be carefully tuned in order to get an optimal performance. This normally involves a model selection stage, based on a computationally intensive search for the minimization of the cross-validation error. In [10] a simple Bayesian approach has been incorporated to eliminate this regularization parameter.

SVM [14] is a machine learning methodology which separates two classes by maximizing the margin between them. A novel type of regularization in support vector machines (SVMs) is used to identify important genes for cancer classification [14]. The standard SVM and Lasso (L1) SVM are often considered using quadratic programming and linear programming methods. An iterative algorithm is used to solve the smoothly clipped absolute deviation (SCAD) SVM efficiently. It is reported that the SCAD-SVM selects a smaller and a more stable (with smaller standard errors) number of genes than the L1-SVM in almost all the cases [14]. Recursive feature elimination (RFE) SVM is another algorithm of gene selection using the weight magnitude as the ranking criterion [1]. The SVM-RFE method ranks all the genes according to some scoring function, and eliminates one or more genes with the lowest scores. This process is repeated until the highest classification accuracy is achieved.

A Gaussian mixture model (GMM) is based on a parametric probability density function that is represented as a weighted sum of Gaussian component densities [15,16]. Since GMM has been used for parameter selection, we have considered it for our comparison. In our study, we have implemented it on microarray gene expression data for gene selection. Like GMM, we have implemented HMM on microarray gene expression data for identification of genes. Generalized HMMs provide an intuitive framework for representing genes with their various functional features, and efficient algorithms can be built to use such models to recognize genes [32].

A constructive approach for feature selection (CAFS) [17] is based on the concept of the wrapper approach and sequential search strategy. As a learning model, CAFS employs a typical three layered feed-forward neural network for selecting genes. In another