Contents lists available at ScienceDirect

Neurocomputing

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

# A novel forward gene selection algorithm for microarray data

Dajun Du [a],*, Kang Li [b], Xue Li [a], Minrui Fei [a]

[a] Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronical Engineering and Automation, Shanghai University, Shanghai 200072, China
[b] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT9 5 AH, UK

ARTICLE INFO

*Article history:*
Received 2 February 2013
Received in revised form
19 October 2013
Accepted 8 December 2013
Communicated by M. Chetty
Available online 10 January 2014

*Keywords:*
Computational complexity analysis
Data augmentation
Fast regression algorithm
Gene selection
Small samples
Variant correlation

ABSTRACT

This paper investigates the gene selection problem for microarray data with small samples and variant correlation. Most existing algorithms usually require expensive computational effort, especially under thousands of gene conditions. The main objective of this paper is to effectively select the most informative genes from microarray data, while making the computational expenses affordable. This is achieved by proposing a novel forward gene selection algorithm (FGSA). To overcome the small samples' problem, the augmented data technique is firstly employed to produce an augmented data set. Taking inspiration from other gene selection methods, the $L_2$-norm penalty is then introduced into the recently proposed fast regression algorithm to achieve the group selection ability. Finally, by defining a proper regression context, the proposed method can be fast implemented in the software, which significantly reduces computational burden. Both computational complexity analysis and simulation results confirm the effectiveness of the proposed algorithm in comparison with other approaches.

## 1. Introduction

Due to their ability to measure the expression levels of thousands of genes simultaneously in a single experiment, DNA microarray technology has been widely employed to obtain microarray data which provides useful information for the molecular investigation of various diseases [1–5]. Microarray data usually have the following properties:

(i) Small samples. In conventional data-driven models from engineering, various sensors can usually be attached to the system fairly easily to collect as many samples as needed. Unfortunately, the measurement of biological systems invariably involves destroying the actual system or cells, which means that the sample size in biological data sets is small (typically the number of samples $N < 100$). The problem is exacerbated for the case of gene expression modelling, where measurements from high-throughput DNA microarrays can simultaneously measure the expression of thousands of genes (or variables $M$). Therefore, one of the main challenges for gene selection is to overcome what is known as the '$M \ll N$' problem.

(ii) Variant correlation. Usually when attempting to model a system or process in engineering, highly correlated variables are considered to be redundant and in order to obtain a compact interpretable model, only one of the correlated variables is assigned a parameter, while the redundant variables are discarded. However, in systems biology, many important variables can be highly correlated. A common biological example is the problem of gene selection from microarray data. Analysis of this type of gene expression data shows that genes share certain biological 'pathways' where they are co-regulated. Genes in the same pathway can therefore be naturally grouped together as they exhibit high correlations. When performing regression in these cases it is important that the entire group of genes is added to the model, rather than each one in isolation. Thus another challenge is how to model with variant correlation.

The above properties indicate that microarray data characteristically have a high dimension, while some of the thousands of genes show strong correlation with a certain phenotype. How to find the relevant genes to a clinicopathological feature or phenotype from microarray data is a very crucial issue. Most researchers mainly show interest in two types of descriptive analyses [6–8]. Firstly, the identification of relevant genes that should include genes even if they perform similar functions and are highly correlated. Secondly, the selection of small sets of genes that may be responsible for a certain phenotype; this involves obtaining the smallest possible set of genes that can achieve good predictive performance.

* Corresponding author.
*E-mail address:* ddj@shu.edu.cn (D. Du).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.012

Discriminant analysis of microarray data can generally be posed in feature selection. The existing feature classification methods can be classified into three categories: filters, wrappers, and embedded methods. The filter approach [9] is widely used based on gene ranking such as *t*-test, Relief-F, information gain, and Kruskal–Wallis rank. However, the drawback is that such a selection procedure is independent of the specific required prediction/classification task. The wrapper method, such as sequential forward selection [10] and particle swarm optimization [11], usually consists of the search procedure and the evaluation criterion. However, an exhaustive search of all subsets is too expensive to implement from a high-dimensional feature space.

Unlike the filter and wrapper methods that separate the variable selection and training processes, the embedded methods incorporated feature selection into the construction procedure of the classifier or regression model. The support vector machine recursive feature elimination method [12] and the boosting method [13] are typical embedded methods. Independent component analysis (ICA) [6,14,15] minimizes both second-order and higher-order dependencies to find the basis along which the data are statistically independent. Suppose that the gene expression data is a linear combination of some independent components, ICA has been applied to microarray gene expression data analysis, i.e., the gene expression data matrix is decomposed into a latent variable matrix and a gene signature matrix, and then the genes are selected only if genes with loading exceed the set thresholds. The ICA method can solve small samples′ problem, but it cannot show clearly the solution paths of the group genes. Specially, if there exists some priori structure knowledge, the structure learning methods such as a scale-free structure prior for graphical models [16] can be developed. The role of the structure prior is to direct inference toward models consistent with prior knowledge, which may come in the form of a priori topological considerations or from a posteriori sources apart from the data set. These learning methods have been applied to functional genomics [16], denoising of musical audio [17], etc.

Recently, some new embedded learning approaches have been proposed to achieve the group effect by inducing the penalty term into the cost function. The most popular algorithm is the Elastic Net [18]. By transforming the system into a Lasso-type problem using the data augmentation technique, the Elastic Net can not only be solved using the least angle regression (LARS) algorithm [19,20], but also overcome the $M \ll N$ problem. Further effort has been made to improve the prediction performance [21,22]. Unfortunately, these methods still require extensive computational effort.

However, for gene expression modelling, the system variables (genes) can generally be naturally grouped together according to regulatory pathways where the within-group correlations are very high. In this scenario it is more important that the entire group be selected for inclusion in the model rather than just a single representative which would best reduce the prediction error. Therefore, a gene selection method should have the following three properties:

(i) Averaging of parameters: when the pairwise correlations between regressors in a group $G$ are very high, the group should undergo an averaging of parameter estimates towards almost equal values.
(ii) Unrestricted model size: the method should be able to select all $M$ variables, and not be restricted by the sample size $N$, even in cases where $M > N$.
(iii) Automatic selection: the modelling procedure should automatically group regressors during the selection process without requiring to pre-specify the grouping structure.

In this paper, a novel forward gene selection algorithm for microarray data is proposed, which can not only solve the small samples and variant correlation problem but also integrate the key properties of the group ability into the modelling procedure. The main contributions of the paper include the following: (1) the small samples′ problem is solved by the augmented data technique; (2) the group selection ability is achieved effectively by introducing the $L_2$-norm penalty; and (3) the proposed approach is faster while maintaining good generalization performance in comparison with the popular Elastic Net method.

The paper is organized as follows. Section 2 gives some preliminaries. Section 3 presents the main results of the paper, including overcoming small samples, proposed forward gene selection algorithm, complete algorithm procedure, computational complexity analysis and remarks on the relationship between the proposed method and the existing work. The simulation results are presented in Section 4, followed by concluding remarks in Section 5.

## 2. Preliminaries

### 2.1. Formulation of the problem

Suppose that an $N \times M$ matrix $X$ denotes the microarray gene expression data with $M$ genes under $N$ samples. $x_{ij}$ in $X$ represents the expression level of the $j$th gene in the $i$th sample, $x_j^c = [x_{1j}, \ldots, x_{Nj}]^T$ represents the expression level of the $j$th gene. The outputs, or corresponding responses, are denoted by $y_i \in y$ and can be either a discrete class label (e.g., $y_i \in [+1, -1]$ for binary classification) in classification problems like discriminating between disease subtypes or a continuous real variable in regression problems such as measurement of some biological parameter and assessment of the gravity of the illness [21,22]. After a location and scale transformation, it is obtained that $y$ is centred and $x_j$, $j = 1, \ldots, M$, are standardized [12], i.e.,

$$\sum_{i=1}^{N} y_i = 0, \quad \sum_{i=1}^{N} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{N} x_{ij}^2 = 1. \tag{1}$$

For the regression problems, the linear-in-the-parameters (LITP) model for gene expression data can be employed directly. However, for the binary classification problem, a LITP classifier for gene expression data can be given by

$$\begin{cases} \hat{y}_k = I_d(f_k) \\ f_k = \sum_{i=1}^{M} \theta_i \varphi_i(x_k^r), \end{cases} \tag{2}$$

where $\theta_i$ is the estimated coefficients, $\varphi_i(\cdot)$ denotes the classifier′s kernels with a known nonlinear basis function like the radial basis function, $\hat{y}_k$ is the model predicted class label, $I_d(\cdot)$ denotes the indicator function and $I_d(f_k \geq 0.5) = 1$ or $I_d(f_k < 0.5) = -1$. The error between the true class label and the classifier′s output signal is expressed as $\varepsilon_k = y_k - f_k$, which can be re-written in the matrix form as

$$y = X\Theta + \Xi, \tag{3}$$

where $X = [x_1^c, \ldots, x_M^c]$, $\Theta = [\theta_1, \ldots, \theta_M]^T$, $\Xi = [\varepsilon(1), \ldots, \varepsilon(N)]^T$, and $y = [y_1, \ldots, y_N]^T$.

The above designed classifiers (2) and (3) focus on the two classes. For multiclass classification problems, one-against-all (OAA) and one-against-one (OAO) methods (e.g., support vector machine) are mainly employed [23,24]. The OAA method consists of $m$ classifiers, where the $i$th classifier is trained with all of the samples in the $i$th class with positive labels and all the other examples from the remaining $m$ classes with negative labels. The OAO method