Neurocomputing 73 (2010) 2375–2383

Contents lists available at ScienceDirect

Neurocomputing

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

A hybrid LDA and genetic algorithm for gene selection and classification of microarray data

Edmundo Bonilla Huerta [a], Béatrice Duval [b], Jin-Kao Hao [b,*]

[a] Instituto Tecnologico de Apizaco, av Instituto Tecnologico S/N, Apizaco, Tlaxcala 90300, Mexico
[b] LERIA, Université d'Angers, 2 Boulevard Lavoisier, 49045 Angers, France

ARTICLE INFO

Available online 31 May 2010

Keywords:
Gene selection
Classification
Dedicated genetic algorithm
Linear discriminant analysis

ABSTRACT

In supervised classification of Microarray data, gene selection aims at identifying a (small) subset of informative genes from the initial data in order to obtain high predictive accuracy. This paper introduces a new embedded approach to this difficult task where a genetic algorithm (GA) is combined with Fisher's linear discriminant analysis (LDA). This LDA-based GA algorithm has the major characteristic that the GA uses not only a LDA classifier in its fitness function, but also LDA's discriminant coefficients in its dedicated crossover and mutation operators. Computational experiments on seven public datasets show that under an unbiased experimental protocol, the proposed algorithm is able to reach high prediction accuracies with a small number of selected genes.

1. Introduction

The DNA microarray technology permits to monitor and to measure gene expression levels for 10s of 1000s of genes simultaneously in a cell mixture. This technology enables to consider cancer diagnosis based on gene expressions [3,5,1,16]. Given the very high number of genes, it is useful to select a limited number of relevant genes for classifying tissue samples.

Three main approaches have been proposed for gene selection. *Filter methods* achieve gene selection independently of the classification model. They rely on a criterion that depends only on the data to assess the importance or relevance of each gene for class discrimination. A relevance scoring provides a ranking of the genes from which the top-ranking ones are generally selected as the most relevant genes. As many filter methods are univariate, they ignore the correlations between genes and provide gene subsets that may contain redundant information.

In *wrapper approaches*, gene subset selection is performed in interaction with a classifier. The goal is to find a gene subset that achieves the best prediction performance for a particular learning model. To explore the space of gene subsets, a search algorithm (e.g. a genetic algorithm—GA) is "wrapped" around the classification model which is used as a black box to assess the predictive quality of the candidate gene subsets. Wrapper methods are computationally intensive since a classifier is built for each candidate subset.

Embedded methods are similar to wrapper methods in the sense that the search of an optimal subset is performed for a specific learning algorithm, but they are characterized by a deeper interaction between gene selection and classifier construction.

Generally, wrapper and embedded methods use a search algorithm for the purpose of exploring a given space of gene subsets. Among different options, GAs are probably the most popular choice. Indeed, one finds from the literature a large number of studies using GAs (often in combination with other techniques) for gene selection. For some representative examples, see the following references [23,34,31,20,24,6,29,50,19,26]. All these methods share a set of common characteristics: they represent a pre-selected gene subsets by a binary vector, employ standard (blind) or specific crossover and mutation operators to evolve a population and use a specific classifier (kNN, SVM…) for fitness evaluation.

In this paper, we propose a new embedded approach to gene subset selection for Microarray data classification. Starting with a set of pre-selected genes using a filter (typically more than 100 genes), we use a dedicated genetic algorithm combined with Fisher's linear discriminant analysis (LDA) to explore the space of gene subsets. More specifically, our dedicated GA uses the LDA classifier to assess the fitness of a given candidate gene subset and LDA's discriminant coefficients to inform the genetic search operators. This LDA-based GA is to be contrasted with many GAs that rely only on random search operators without taking into account problem knowledges.

To evaluate the usefulness of the proposed LDA-based GA, we carry out extensive experiments on seven public datasets and compare our results with 15 best performing algorithms from the literature. We observe that our approach is able to achieve a high prediction accuracy (from 96% to 100%) with a small number of

* Corresponding author.
E-mail address: hao@info.univ-angers.fr (J.-K. Hao).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.03.024

informative genes (often less than 20). To understand the behavior of the proposed algorithm, we also study the impact of the LDA-based genetic operators on the evolutionary process as well as other important parameters of the proposed algorithm.

The remainder of this paper is organized as follows. Section 2 recalls the main characteristics of Fisher's LDA and discusses the calculus that must be done in the case of samples of small size. Section 3 presents our LDA-based GA for gene selection and classification. Section 4 shows the experimental results and comparisons. Section 5 is dedicated to additional studies on the LDA-based crossover operator and other important parameters of the proposed algorithm. Finally conclusions are provided in Section 6.

2. LDA and small sample size problem

2.1. Linear discriminant analysis

LDA is a well-known dimension reduction and classification method, where the data are projected into a low dimension space such that the classes are well separated. LDA has recently been used for Microarray data analysis [12,47,48].

As we use this method for binary classification problems, we shall restrict the explanations to this case. We consider a set of n samples belonging to two classes C_1 and C_2, with n_1 samples in C_1 and n_2 samples in C_2. Each sample is described by q variables. So the data form a matrix $X=(x_{ij})$,i$=1,\ldots,n$; $j=1,\ \ldots,q$. We denote by μ_k the mean of class C_k and by μ the mean of all the samples:

$$\mu_k = \frac{1}{n_k}\sum_{x_i \in C_k} x_i \quad \text{and} \quad \mu = \frac{1}{n}\sum_{x_i} x_i = \frac{1}{n}\sum_k n_k \mu_k$$

The data are described by two matrices S_B and S_W, where S_B is the between-class scatter matrix and S_W the within-class scatter matrix defined as follows:

$$S_B = \sum_k n_k(\mu_k-\mu)(\mu_k-\mu)^t \tag{1}$$

$$S_W = \sum_k \sum_{x_i \in C_k} (x_i-\mu_k)(x_i-\mu_k)^t \tag{2}$$

If we denote by S_V the covariance matrix for all the data, we have $S_V=S_B+S_W$.

LDA seeks a linear combination of the initial variables on which the means of the two classes are well separated, measured relatively to the sum of the variances of the data assigned to each class. For this purpose, LDA determines a vector w such that w^tS_B w is maximized while w^tS_W w is minimized. This double objective is realized by the vector w_{opt} that maximizes the criterion:

$$J(w) = \frac{w^tS_Bw}{w^tS_Ww} \tag{3}$$

One can prove that the solution w_{opt} is the eigen vector associated to the sole eigen value of $S_W^{-1}S_B$, when S_W^{-1} exists. Once this axis w_{opt} is determined, LDA provides a classification procedure (classifier), but in our case we are particularly interested in the *discriminant coefficients* of this vector: the absolute value of these coefficients indicates the importance of the q initial variables for the class discrimination.

2.2. Generalized LDA for small sample size problems

When the sample size n is smaller than the dimensionality of samples q, S_W is singular, and it is not possible to compute S_W^{-1}. To overcome the singularity problem, recent works have proposed different methods like the null space method [48], orthogonal LDA [46], uncorrelated LDA [47,46] (see also [33] for a comparison of these methods). The two last techniques use the pseudo inverse method to solve the small sample size problem and this is the approach we apply in this work. When S_w is singular, the eigen problem is solved for $S_w^+S_b$, where S_w^+ is the pseudo inverse of S_w. The pseudo-inverse of a matrix can be computed by singular value decomposition. For a matrix A of size $m\times p$, the singular values of A, noted σ_i, are the non-negative square roots of A^tA. The singular value decomposition of A is $A=U\Sigma V^t$, where U of size $m\times m$ and V of size $n\times n$ are orthogonal and

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

with D a diagonal matrix containing the elements σ_i. If we note Σ^+ the matrix

$$\Sigma^+ = \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

(where the 0 blocks have the appropriate sizes), the pseudo inverse of A is then defined as $A^+ = V\Sigma^+U^t$.

2.3. Application to gene selection

Microarray data generally contain less than 100 samples described by at least several 1000s of genes. We limit this high dimensionality by a first pre-selection step, where a filter criterion (t-statistic here) is applied to determine a subset of relevant genes (see Section 4.2 for more details). In this work, we typically retain 100 genes from which an intensive exploration is realized using a genetic algorithm to select smaller subsets. In this process, LDA is used as a classification method to evaluate the classification accuracy that can be achieved on a candidate gene subset. Moreover the coefficients of the eigen vector calculated by LDA are used to evaluate the importance of each gene for class discrimination.

For a selected gene subset of size p, if $p\le n$, we rely on the classical LDA (Section 2.1) to obtain the projection vector w_{opt}, otherwise we apply the generalized LDA (Section 2.2) to obtain this vector. We explain in Section 3 how the LDA-based GA explores the search space of gene subsets.

3. LDA-based genetic algorithm

Given a set of p top ranking genes pre-selected with a filter (typically $p\ge 100$, in this work, $p=100$ or 150), our LDA-based GA is used to conduct a combinatorial search within the space of size 2^p. The purpose of this search is to determine among the possible gene combinations small sized gene subsets allowing a high predictive accuracy. In what follows, we present the general procedure and then describe the components of the LDA-based genetic algorithm. In particular, we explain how LDA is combined with the Genetic Algorithm.

3.1. General GA procedure

Our LDA-based genetic algorithm (LDA-GA) follows the conventional schema of a generational GA and uses also an elitism strategy.

- *Initial population*: The initial population is generated randomly in such a way that each chromosome contains a number of genes ranging from $p\times 0.6$ to $p\times 0.75$. The population size is fixed at 100 in this work.
- *Evolution*: The chromosomes of the current population P are sorted according to the fitness function (see Section 3.3). The "best" 10% chromosomes of P are directly copied to the next