Data & Knowledge Engineering 66 (2008) 338–347

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Cancer classification by gradient LDA technique using microarray gene expression data

Alok Sharma [a,b,*], Kuldip K. Paliwal [a]

[a] Signal Processing Lab, Griffith University, Brisbane, Australia
[b] University of the South Pacific, School of Engineering and Physics, Suva, Fiji

ARTICLE INFO

Article history:
Received 18 December 2007
Received in revised form 29 March 2008
Accepted 8 April 2008
Available online 22 April 2008

Keywords:
DNA microarray
Linear discriminant analysis (LDA)
Gradient LDA (GLDA)
Dimensionality reduction
Cancer classification
Feature selection
Feature extraction

ABSTRACT

Cancer classification is one of the major applications of the microarray technology. When standard machine learning techniques are applied for cancer classification, they face the small sample size (SSS) problem of gene expression data. The SSS problem is inherited from large dimensionality of the feature space (due to large number of genes) compared to the small number of samples available. In order to overcome the SSS problem, the dimensionality of the feature space is reduced either through feature selection or through feature extraction. Linear discriminant analysis (LDA) is a well-known technique for feature extraction-based dimensionality reduction. However, this technique cannot be applied for cancer classification because of the singularity of the within-class scatter matrix due to the SSS problem. In this paper, we use Gradient LDA technique which avoids the singularity problem associated with the within-class scatter matrix and shown its usefulness for cancer classification. The technique is applied on three gene expression datasets; namely, acute leukemia, small round blue-cell tumour (SRBCT) and lung adenocarcinoma. This technique achieves lower misclassification error as compared to several other previous techniques.

1. Introduction

The advent of microarray technology has enabled the researchers to rapidly measure the levels of thousands of genes expressed in a biological tissue sample in a single experiment [19]. One important application of this microarray technology is to classify the tissue samples using their gene expression profiles as one of the several types (or subtypes) of cancer. Compared with the standard histopathological tests, the gene expression profiles measured through microarray technology provide accurate, reliable and objective cancer classification.

The DNA microarray data for cancer classification consists of large number of genes (dimensions) compared to the number of samples or feature vectors. The high dimensionality of the feature space degrades the generalisation performance of the classifier and increases its computational complexity. This problem is popularly in known as the small sample size (SSS) problem in the literature [10]. It restricts direct application of conventional statistical and machine learning techniques for classification purposes. This situation, however, can be overcome by first reducing the dimensionality of feature space, followed by classification in the lower-dimensional feature space. Different methods used for dimensionality reduction can be grouped into two categories: feature selection methods and feature extraction methods. Feature selection methods retain

* Corresponding author. Address: University of the South Pacific, School of Engineering and Physics, Suva, Fiji. Tel.: +679 3232870; fax: +679 3231538.
E-mail addresses: sharma_al@usp.ac.fj (A. Sharma), K.Paliwal@griffith.edu.au (K.K. Paliwal).

0169-023X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2008.04.004

only a few useful features and discard others. Feature extraction methods construct a few features from the large number of original features through their linear (or nonlinear) combination. For the classification of the lower-dimensional feature vector, the Bayes decision rule provides the most optimal solution. But, since the amount of training data available for designing the classifier is limited and small, other classifiers (e.g., nearest centroid classifier, k-nearest neighbour classifier, etc.) are used for cancer classification. A number of papers have been reported in the past for the cancer classification task using the microarray data. We provide a brief description of a small sample of these papers to highlight different techniques used for dimensionality reduction and classification for this task.

Golub et al. [13] adopted gene selection criteria based on correlation of genes prior to the classification. The selected genes were utilized in weighted voting (WV) approach for cancer classification. Furey et al. [11] applied similar technique as of Golub et al. [13] for gene selection and demonstrated the use of support vector machine (SVM) for cancer classification. Dudoit et al. [8] compared the performance of different discrimination methods for classification of tumours. These methods included nearest neighbour (NN) classifier, linear discriminant analysis (LDA), diagonal discriminant analysis, quadratic classifiers and classification trees. They considered bagging [5] and boosting [9] approaches to select relevant genes, which were used in the classification. Nguyen and Rocke [18] proposed partial least square (PLS) method for human tumor classification. They used PLS and principal component analysis (PCA) for dimension reduction as well as quadratic discriminant analysis (QDA) and logistic discrimination (LD) for classification task. Guyon et al. [14] proposed a gene selection criterion utilizing SVM methods based on recursive feature elimination (RFE). Lee et al. [16] developed a hierarchical Bayesian (HB) model for variable gene selection. Instead of fixing the number of selected genes (dimensions), a prior distribution over it was assigned. Bee and Mallick [2] pointed out that this approach is sensitive toward the choice of some hyper-parameters. Consequently, they considered a multivariate Bayesian regression model and assigned *priors* that favour sparseness in terms of number of genes used. They introduced the use of different *priors* to promote different degree of sparseness using a two-level hierarchical Bayesian (2L-HB) model. Zhou et al. [28] proposed a Bayesian approach to gene selection and classification using logistic regression model [1]. They used Gibbs sampling and Markov chain Monte Carlo (MCMC) methods to discover important genes. Geman et al. [12] introduced top scoring pair (TSP), which is based on pairwise comparison between two gene expression levels. This TSP algorithm was extended by Tan et al. [24] to *k*-TSP, which uses *k* pairs of genes for classifying gene expression data. They investigated the performance of TSP and *k*-TSP for three different schemes namely one-vs-others scheme, one-vs-one scheme and hierarchical classification (HC) scheme. Yeung et al. [26] used Bayesian model averaging (BMA) to address multi-class cancer classification problem. A typical gene selection and classification procedure ignores model uncertainty and uses a single set of relevant genes to predict the class. On the other hand, BMA accounts for the uncertainty by averaging over multiple sets of potentially overlapping relevant genes. Tan et al. [25] addressed the small sample size problem with microarray data by proposing total principal component regression (TPCR). It can classify human tumors by extracting the lateral variable structure underlying microarray data from the augmented subspace of both independent variables and dependent variables. Zhang et al. [27] developed a type of regularization in SVM to identify important genes for cancer classification. Leng and Müller [17] used functional logistic regression tool based on functional principal components for classifying temporal gene expression data.

From this short survey, it is clear that most of the techniques described above employ feature (or gene) selection for dimensionality reduction. Since feature extraction method always give better performance than feature selection method [7] we will investigate techniques based on feature extraction methods in this paper. In the feature (or gene) selection methods, several genes are discarded. Only a few genes are retained based on some criterion function, which tries to rank genes for classification purpose. The genes that are discarded may contain crucial information for classification of cancer types. In addition, the choice of the number of genes to be selected in these papers is usually arbitrary. Theoretically it is difficult to identify the number of selected genes that provides optimal performance for the classification of cancerous tissues, though empirical arguments can be made to justify the number of genes to be selected.

In the present paper, we concentrate on the feature extraction methods for dimensionality reduction. Feature extraction methods can provide better classification performance over feature selection methods since in feature extraction the subspace is a linear combination of original feature space [7]. The two popular feature extraction methods for dimensionality reduction are principal component analysis (PCA) and linear discriminant analysis (LDA). The former method concentrates on the representation of data and is not very powerful in discriminating the cancer classes. The LDA method, on the other hand, is discriminative in nature and could help in classifying a tissue sample more accurately. In LDA, we attempt to maximize between-class scatter with respect to within-class scatter. This process requires solving generalized eigenvalue decomposition problem, which involves the computation of the inverse of within-class scatter matrix. Due to the high dimensionality of microarray data compared to the number of samples available, the scatter matrix becomes singular and its inverse computation is not feasible. Looking at the limitations of LDA, we adapted gradient LDA (GLDA) technique [21] which resolves this type of limitation. The GLDA technique utilizes gradient descent algorithm to do dimensionality reduction. Once the dimension is reduced through the GLDA algorithm, the *k*-nearest neighbour classifier with Euclidean distance measure is used to classify a tissue sample. Experiments on several microarray gene expression datasets using the GLDA technique show very encouraging results for cancer classification.

The paper is organized as follows. Section 2 describes the three datasets used in the experimentation, Section 3 illustrates the GLDA technique, Section 4 describes the experiments and results, and Section 5 presents our conclusions.