Chemometrics and Intelligent Laboratory Systems 139 (2014) 6–14

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique

Sadegh Karimi [a,*], Maryam Farrokhnia [b]

[a] Department of Chemistry, College of Sciences, Persian Gulf University, Bushehr, Iran
[b] The Persian Gulf Marine Biotechnology Research Center, Bushehr University of Medical Sciences,Bushehr,Iran

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form 25 August 2014
Accepted 5 September 2014
Available online 16 September 2014

Keywords:
Microarray gene expression
Linear discriminant analysis
Clustering of variable
Kohonen self-organizing map
Genetic algorithm

ABSTRACT

Using gene expression data in cancer classification plays an important role for solving the fundamental problems relating to cancer diagnosis. Because of high throughput of gene expression data for healthy and patient samples, a variable selection method can be applied to reduce complexity of the model and improve the classification performance. Since variable selection procedures pose a risk of over-fitting, when a large number of variables with respect to sample are used, we have proposed a method for coupling data dimension reduction and variable selection in the present study. This approach uses the concept of variable clustering for the original data set. Significant components of local principal component analysis models have just been retained from all clusters. Then, the variable selection algorithm is performed on these locally derived principal component variables. The proposed algorithm has been evaluated on two gene expression data sets; namely, acute Leukemia and small round blue-cell tumor (SRBCT). Our results confirmed that the classification models achieved on the reduced data were better than those obtained on the entire microarray gene expression profile.

1. Introduction

Cancer research is one of the most important research areas in the medical sciences. A correct prediction of different tumor types has noticeable value in providing better treatment and toxicity minimization on the patients. The early diagnosis of cancer can significantly reduce mortality rates among the patients [1]. On the other hand cancer classification and detection have always been morphological and clinical based while using conventional methods have own several restrictions in their diagnostic ability [1]. In order to increase a better insight into the problem of cancer classification, systematic approaches based on global gene expression analysis have been proposed [2]. One of the good criteria in cancer detection is the expression level of genes. This phenomenon contains the keys to address fundamental problems relating to the inhibition and treatment of diseases, biological evolution mechanisms and drug discovery.

The recent advent of microarray knowledge has assisted the scientists to quickly measure the levels of thousands of genes expressed in a biological tissue sample just in a single experiment [3–7]. This kind of data has some properties. One of their main characteristics is that microarray studies often generate massive amounts of data (usually contains tens to thousands of genes), which are difficult to be exhaustively examined by hand. Therefore bioinformatics analysis and interpretation are essential to extract genetic patterns from these data for gaining biological insights from experiments [8]. The second characteristic is related to the publicly available data size which is very small; for example some data have sizes below 100 samples. This subject led to small sample size problem (the ratio of variable to sample is high). In this condition, the classification methods such as LDA have a tendency to show over-fitting result [9]. Finally, most of genes are irrelevant to cancer distinction and should be discarded or removed.

On the other hand, the rise of chemometrics as an important subdiscipline of analytical measurement science results in providing powerful multivariate data analysis. In addition, rapid growth of analytical instrumentation produces huge data set. Consequently, applying chemometric methods in the analysis of these huge data leads to extraction of more information.

Interestingly, a similar revolution has also occurred in biological sciences resulting from new measurement technologies in the last two decades and subsequently the need for the effective data analysis tools such as chemometric methods has been increased. For instances, different classification methods from statistical and machine learning area have been applied to cancer classification, but there are some issues that make it difficult to perform. For example, it is evident that these conventional classification methods have not been designed to handle this kind of data efficiently and effectively. Some researchers proposed to do the gene selection prior to cancer classification [10–14] to reduce

[*] Corresponding author.
E-mail addresses: karimi.sadegh@gmail.com, sakarimi@pgu.ac.ir (S. Karimi).

http://dx.doi.org/10.1016/j.chemolab.2014.09.003
0169-7439/© 2014 Elsevier B.V. All rights reserved.

the variable space size. Therefore this approach can improve the running time. Significantly, gene selection removes a large number of irrelevant genes and results in better classification accuracy [15].

In the present study, due to the important role of removing unnecessary genes or factors which are irrelevant for cancer classification, we have suggested the new strategy based on local data dimension reduction with variable selection method for analyzing the gene expression data set. In this approach, clusters of the original variable concept are used to cluster the gene value and only significant components are retained, which is similar to segmented principal component analysis and regression (SPCAR) [16,17]. Finally, a variable selection algorithm (GA) combined with LDA is performed on those locally derived principal component variables instead of whole original data.

2. Theory

2.1. Linear discriminant analysis (LDA)

The LDA is one of the most used traditional classification techniques [18]. The method is a probabilistic parametric classification technique which performs dimensionality reduction by maximizing the variance between categories and minimizing the variance within categories. The classification index (discriminant function) is based on the Bayes minimum error rule, i.e. samples are classified into the class with the maximum a-posteriori probability and LDA makes the assumption that the classes have identical covariance matrices and fits a multivariate normal density to each group with a pooled estimate of the covariance. Since a pooled covariance matrix is calculated, the number of objects must be greater than the number of variables. In the other words when the number of variables is exceeded the number of samples, the LDA classifiers does not work [19], i.e. on the percentage of correctly assigned samples, evaluated both on cross-validation groups and external test samples.

2.2. Kohonen self-organizing map

Self-organizing map (SOM) [20] is from the category of artificial neural network (ANN) algorithms that uses unsupervised learning to create a two dimensional representation of training samples. This two dimension representation which is called map consists of components called nodes or neurons. Accompanying with each neuron, there is a weight vector with the same dimension as the input vectors in the map space. The interesting feature of SOM algorithm that distinguishes it from other artificial neural networks is the use of a neighborhood function. During the training step, the Kohonen network adjusts itself in such a way that similar input (here, gene value) is associated with the topological close neuron in the network. The arrangement of neurons is in two dimensions in a hexagonal or rectangular space, with size ($p \times p$) where p is a defined network sizes.

The mapping procedure is used to find the neuron in the created network with the closest weight vector to the input vector. The most similar neuron (small distance metric with input vector) has been selected as a winner. Then, the neighbor neurons (in the first and second neighborhood) also adjust their weights with respect to the winner neuron. Changes in neighboring neuron depend on the neighborhood function. As a result, when the process is completed, similar input vectors (gene values) are clustered in the space, based on their similarities. It should be noted that we have applied SOM to cluster the gene values not the objects. Hence, the variables in the original data matrix, including similar information, are mapped into one node or neighboring nodes. The variables in each node can be collected to form a sub-matrix D_i.

2.3. GA-LDA based on SOM

In some cases, we deal with data sets included many variables. For analysis (multivariate calibration and classification) of such data sets, we should careful about over-fitting problem. Although variable selection methods have been proposed for the aforementioned problem, but these methods (genetic algorithm, GA, and forward selection) are not appropriate solution when the number of variables is too large. Related to this issue, Ballabio and co-workers [21] have explained that, GA algorithm results in severe over-fitting or non-optimal solutions when the huge number of variables exists. Thus, they have suggested that a reduction of data dimension can be useful when dealing with high-dimensional data (MALDI–MS or GC–MS) in which the chemical rank is well below the dimension of the data.

In the present study we faced with the same problem of high dimensional data namely gene expression data set. With this aim, an efficient algorithm for the extraction of significant features from high-dimensional data has been proposed. First, the variables (gene values) have been clustered using Kohonen self-organizing map as clustering algorithm. Subsequently, PCA [22] has been applied on each cluster of the original gene expression profile, similar to the characteristics of the SPCAR algorithm [16,17]. Then, the original data set has been transformed into a new data set of which their columns are significant principal components retained from these local clusters. Finally; the specific PC selection algorithm combined with classification methods can be used for analysis this reduced data set.

Suppose we have a data matrix (D) with I rows (the samples) and J columns (the gene value). The data dimension reduction can be illustrated using the subsequent steps:

1- In the first part, the whole gene expression value has been partitioned in q cluster using Kohonen self-organizing map (SOM). Thus, the gene values have been clustered in a different sub-matrix (D_i) according to their similarities in information.

$$D = \left[[D_1][D_2]\ldots\left[D_q\right] \right] \tag{1}$$

2- In the next step, in order to calculate the principal components and loading of each cluster, PCA can be applied in each sub-matrix (D_i) separately. It should be noted that different preprocessing algorithms (depending on the type of data) can be used for each cluster.

$$D_i = T_i P_i^T \; i = 1 : q \tag{2}$$

The matrices T_i and P_i are the principal components and loadings of the each cluster (D_i) respectively. The superscript “T” indicates the matrix transpose notation.

3- In this step, the most important of PCs and corresponding loadings form each cluster should be selected. Different strategies can be considered for this screening. In the present study, explained variance (EV) and root mean square error of cross-validation (RMSCV) have been used as criteria to select the most significant PCs in each cluster. For explained variance criterion, the PCs of which their explained variances are higher than the specific value (95%) have been kept. By substitution of Eq. (2) into Eq. (1) we obtain:

$$D_r = \left[\left[T_1 P_1^T\right] \left[T_2 P_2^T\right] \ldots \left[T_q P_q^T\right] \right]. \tag{3}$$

The T_1 to T_q and $P_1{}^T$ to P_q^T are the PCs and loadings obtained from different sub-matrices (D_i).

The new data set, $D_{r,}$ which we called reduced data matrix, consists of all the obtained PCs from different clusters. Obviously the dimensions of D_r is ($I \times r$), where I is the number of samples and r is the total number of principal components obtained from previous step. Eq. (3) indicates that one can be able to separate the PCs and loadings of different clusters. If we have just rearranged Eq. (3), a new possibility can be obtained (Eq. (4)).

$$D_r = \left[T_1 T_2 \ldots T_q\right] V \tag{4}$$