Computers in Biology and Medicine 64 (2015) 292–298

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

New feature selection for gene expression classification based on degree of class overlap in principal dimensions

Somsak Rakkeitwinai [a,*], Chidchanok Lursinsap [a], Chatchawit Aporntewan [a], Apiwat Mutirangura [b]

[a] Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Patumwan, Bangkok 10330, Thailand
[b] Department of Anatomy, Faculty of Medicine, Chulalongkorn University, Patumwan, Bangkok 10330, Thailand

ARTICLE INFO

Article history:
Received 19 May 2014
Accepted 30 January 2015

Keywords:
Micro-array data
Dimension reduction
Feature selection
Feature extraction
principal component
Analysis
Support Vector Machine

ABSTRACT

Micro-array data are typically characterized by high dimensional features with a small number of samples. Several problems in identifying genes causing diseases from micro-array data can be transformed into the problem of classifying the features extracted from gene expression in micro-array data. However, too many features can cause low prediction accuracy as well as high computational complexity. Dimensional reduction is a method to eliminate irrelevant features to improve the prediction accuracy. Typically, the eigenvalues or dimensional data variance from principal component analysis are used as criteria to select relevant features. This approach is simple but not efficient since it does not concern the degree of data overlap in each dimension in the feature space. A new method to select relevant features based on degree of dimensional data overlap with proper feature selection was introduced. Furthermore, our study concentrated on small sized data sets which usually occur in reality. The experimental results signified that this new approach can achieve substantially higher prediction accuracy when compared with other methods.

1. Introduction

Classification techniques have been used for analysis and interpretation of micro-array gene expression data. The information from the pattern of micro-array gene expression data has provided valuable insights for many biological problems. The main challenge in classifying gene expression data is the curse of dimensionality problem. There are a large number of genes (in biological context) or features (in classification context) compared to small sample sizes [1]. In the aspect of computation, the number of features can be treated as the number of dimensions in the feature space. Mathematically, if the number of dimensions is larger than the number of data, then no unique solution exists and achieving the high prediction accuracy is impossible. To alleviate this curse of dimensionality, some irrelevant features must be eliminated from the prediction or classification process.

Dimension reduction usually consists of two types of methods, i.e. feature selection and feature extraction [2]. Feature selection chooses a subset from original features according to classification performance. Feature selection can be organized into two categories: filter and wrapper methods. The filter method basically uses a criterion relating to rank and selects key genes for classification such as Pearson correlation coefficient method [3,4], *t*-statistics methods [5,6], signal-to-noise ratio method [7,8]. The wrapper method integrates the feature selection process with the evaluation of the selected features for classification such as forward feature selection [9,10], sequential backward selection [11,12], genetic algorithm [13,14]. Feature extraction projects the whole data into a low dimensional space and constructs the new dimensions (components) by analyzing the statistical relationship hidden in the data set. Many transformation methods have been applied to reduce the dimension of the data such as independent component analysis [15,16], principal component analysis [17,18], wavelet analysis [19,20].

Another popular technique is based on the evolutionary computation or algorithm. Evolutionary algorithm (EA) finds optimal solution in a complex high dimensional space. Dimensional reduction was applied to solve this problem. Genetic Algorithm (GA) [21] generates acceptable solutions by using techniques inspired by natural evolution such as selection, crossover and mutation. A population of chromosomes representing encoded candidate solutions to the optimization problem is evolved toward the better solutions as some examples reported in [22,23]. Rough set theory [24] is another concept used for dimensional reduction in large data. Rough set is a formal approximation of a crisp set similar to fuzzy set [25]. Rough set can be combined with other methods such

---

[*] Corresponding author. Tel.: +66 8 1208 8801.
E-mail addresses: somsak.rak@gmail.com, Phisetphong.S@student.chula.ac.th (S. Rakkeitwinai).

http://dx.doi.org/10.1016/j.compbiomed.2015.01.022
0010-4825/

as GAs and multi-objective genetic algorithms (MOGA) to search for optimal solutions [26,27]. Particle swarm optimization (PSO) [28] based on the behavior of swarm is another popular method for finding an acceptable solution. PSO was adopted to the problem of identifying important gene from gene expression [29,30].

The most widely used dimensional reduction method is principal component analysis (PCA) [31]. PCA [32] rotates the bases of the feature space to the positions so that the data variance of each dimension is minimum. PCA is a popular unsupervised statistical method to find useful eigenassays [33]. One goal of PCA is to find a "better" set of eigenassays, so that the PCA coefficients are uncorrelated on the new basis, which cannot be linearly predicted from each other [34]. However, PCA does not consider the degree of overlap among classes in each dimension. For any dimension, if there is no overlap among classes, then the feature in this dimension can obviously be used as one of the principal features for achieving maximum classification accuracy.

To improve efficiency of PCA for use in feature selection, we proposed a new method called *principal feature analysis of minimum distribution overlap* with forward feature selection based on Support Vector Machine. The degree of class overlap is measured and used to determine the candidate bases in the feature space. The propose method was tested with four gene expression data sets and compared with the recently proposed methods.

The rest of this paper is organized as follows. Section 2 summarizes the background and related work. Section 3 discusses the concept of our proposed method. Section 4 explains the data set and experimental results. Section 5 concludes the paper with some discussions.

2. Background and related work

Our proposed method selects each feature based on the degree of overlap between classes and also the result of preliminary classifying process. Recently, Luo [9] adopted the statistical measure similar to *t*-test with different distance measures, e.g. Euclidean, to improve the feature selection in the original feature space. Zhang [34] proposed a weight measure which can improve the discriminant ability and preserve the local neighborhood structure of the original data. Hence the methods of Luo's and Zhang's with the concept of support vector machine are briefly summarized in the following section.

2.1. Feature selection based on signal-to-noise ratio

Luo [9] proposed two forward feature selection methods named as FFS-ACSA1 and FFS-ACSA2 in dichotomy classification. The target value of each feature vector is in $\{+1,-1\}$. The concept of this approach is to deploy the regular technique of average ensemble classifiers with a new feature selection. The proposed feature selection of Luo's is based on the measure of signal-to-noise ratio of each feature and a cost function defined as follows. Suppose that the considered feature corresponds to dimension i. Let a_i be the value of the ith feature:

$$f(a_i)=\mathrm{sign}[s_i(a_i-b_i)] \tag{1}$$

$$s_i=\frac{\mu_i^{(A)}-\mu_i^{(B)}}{\sigma_i^{(A)}+\sigma_i^{(B)}} \tag{2}$$

$$b_i=(\mu_i^{(A)}+\mu_i^{(B)})/2 \tag{3}$$

where $\mu_i^{(A)}$, $\mu_i^{(B)}$, $\sigma_i^{(A)}$, and $\sigma_i^{(B)}$ are the mean and standard deviation of data distribution in dimension i of classes A and B, respectively. Function $f(a_i)$ is the cost function and s_i is the signal-to-noise ratio for the ith feature. The difference between the values of $f(a_i)$ and the values of actual targets is compared by using some distance measures such as Euclidean distance and loss function. Those features with the differences less than a predefined threshold are selected.

2.2. Dimension reduction based on manifold learning

Zhang [34] proposed an improved supervised orthogonal discriminant projection (SODP). The concept of this approach is based on the nearest neighbor classifier with a manifold learning. The proposed manifold learning of Zhang's maximizes the weighted difference between the non-local scattering and the local scattering. Zhang defined the weight $w_{i,j}$ between two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ based on local information and class information as follows:

$$w_{i,j}=\begin{cases}\exp\left(\dfrac{-\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\beta^2}\right)\left[1+\exp\left(\dfrac{-\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\beta^2}\right)\right] & \text{if } (\mathbf{x}_i\in K(\mathbf{x}_j)) \text{ or } (\mathbf{x}_j\in K(\mathbf{x}_i)) \\ & \text{and } (t_i=t_j) \\ \exp\left(\dfrac{-\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\beta^2}\right) & \text{if } (\mathbf{x}_i\in K(\mathbf{x}_j)) \text{ or } (\mathbf{x}_j\in K(\mathbf{x}_i)) \\ & \text{and } (t_i\neq t_j) \\ 0 & \text{otherwise}\end{cases} \tag{4}$$

where $\|\mathbf{x}_i-\mathbf{x}_j\|$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$; $K(\mathbf{x}_i)$ is the set of k nearest neighbors of $\mathbf{x}_i$; t_i is the class label of $\mathbf{x}_i$; and β is a tuning parameter. The local scattering and the non-local scattering are defined as follows. Let $\mathbf{A}$ be a linear transformation matrix. The local scattering $J_L(\mathbf{A})$ is defined as

$$J_L(\mathbf{A})=\mathbf{A}^T\mathbf{S}_L\mathbf{A} \tag{5}$$

where $\mathbf{S}_L=(1/2n^2)\sum_{i=1}^{n}\sum_{j=1}^{n}w_{i,j}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T$. The non-local scattering $J_N(\mathbf{A})$ is defined by

$$J_N(\mathbf{A})=\mathbf{A}^T\mathbf{S}_N\mathbf{A} \tag{6}$$

where $\mathbf{S}_N=(1/2n^2)\sum_{i=1}^{n}\sum_{j=1}^{n}(1-w_{i,j})(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T$. The optimal transformation matrix $\mathbf{A}$ can be obtained by

$$\arg\max J(\mathbf{A})=\arg\max \mathbf{A}^T((1-\gamma)\mathbf{S}_N-\gamma\mathbf{S}_L)\mathbf{A} \tag{7}$$

where the constant γ is an adjustable parameter. Then the optimal linear features of all data $\mathbf{x}_i\in\mathbf{X}$ can be obtained by

$$\mathbf{Y}=\mathbf{A}^T\mathbf{X} \tag{8}$$

For any $\mathbf{y}_i\in\mathbf{Y}$, a data point $\mathbf{y}^*\in\mathbf{Y}$ can be assigned to the same class as $\mathbf{y}_i$ if the following condition is satisfied:

$$\mathbf{y}^*=\underset{\forall j}{\arg}\min(\|\mathbf{y}_j-\mathbf{y}_i\|) \tag{9}$$

2.3. Support Vector Machine

Support Vector Machine (SVM) [35, 36] has been very popular in solving dichotomy classification problems. It mapped the data in feature space onto a higher dimensional feature space by a set of kernel functions so that the data are absolutely separable. The values of targets are in $\{+1,-1\}$. The data are separated by a hyperplane and the class of each datum is determined by this condition. Let $\mathbf{x}_j$ and b be a datum and a bias, respectively. If $(\mathbf{w}\mathbf{x}_j+b)\le -1$, then it is in class -1. But if $(\mathbf{w}\mathbf{x}_j+b)\ge +1$, then it is in class $+1$. Any datum whose value $(\mathbf{w}\mathbf{x}_j+b)$ in between -1 and $+1$ is indeterminate. The most commonly used kernel function is inner product of two feature vectors.

2.4. Forward feature selection

Forward feature selection method selects one feature at a time which gives the best prediction accuracy in combination with the