Journal of Biomedical Informatics 62 (2016) 12–20

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A kernel-based clustering method for gene selection with gene expression data

Huihui Chen [a], Yusen Zhang [a,*], Ivan Gutman [b]

[a] School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China
[b] Faculty of Science, University of Kragujevac, P.O. Box 60, 34000 Kragujevac, Serbia

ARTICLE INFO

Article history:
Received 9 August 2015
Revised 8 May 2016
Accepted 19 May 2016
Available online 20 May 2016

Keywords:
Gene expression data
Kernel-based clustering
Adaptive distance
Gene selection
Cancer classification

ABSTRACT

Gene selection is important for cancer classification based on gene expression data, because of high dimensionality and small sample size. In this paper, we present a new gene selection method based on clustering, in which dissimilarity measures are obtained through kernel functions. It searches for best weights of genes iteratively at the same time to optimize the clustering objective function. Adaptive distance is used in the process, which is suitable to learn the weights of genes during the clustering process, improving the performance of the algorithm. The proposed algorithm is simple and does not require any modification or parameter optimization for each dataset. We tested it on eight publicly available datasets, using two classifiers (support vector machine, k-nearest neighbor), compared with other six competitive feature selectors. The results show that the proposed algorithm is capable of achieving better accuracies and may be an efficient tool for finding possible biomarkers from gene expression data.

1. Introduction

DNA microarray technology has proven to be a great breakthrough in molecular biology, monitoring thousands of gene expression in a single experiment [1–3]. One of its major applications is cancer diagnosis. With the gene expression data, identifying different cancer classes or subclasses with similar morphological appearances is of great importance for better treatment and prognosis. Since cancer microarray data consists of a large number of genes with small samples, and cancers are usually marked by a change in the expression levels of certain genes [4], gene selection technique is crucial in most studies of cancer classification.

Gene selection is a key pre-processing step in cancer classification in order to eliminate irrelevant and redundant genes, which can significantly improve the comprehensibility of classification models with the highest degree of accuracy. Furthermore, it is often the case that finding a small subset of biomarker genes is very important.

A wide variety of approaches have been developed for gene selection to extract relevant genes. There are three categories of feature selection techniques for microarray data: filter, wrapper and embedded [5]. Filter methods [5,6] select a gene subset from the original dataset using specific evaluation criteria mostly based on statistical methods that are independent of the classifier. Wrapper methods [7,8] select feature subsets employing the performance of the classifier to evaluate the importance of feature subsets. Embedded methods [4,9] combine the advantage of filter and wrapper techniques, using a pre-determined classification model algorithm to perform feature selection. Among the three kinds of methods, filter methods are widely used on large-scale data, such as gene expression data, for they are computationally faster and more general than wrapper methods. Though wrapper methods should provide more accurate classification results than filter methods in theory [10], wrapper methods have a higher risk of overfitting.

In Ref. [11], the authors proposed simultaneous clustering and attribute discrimination (SCAD) algorithm, based on fuzzy c-means (FCM) algorithm. SCAD algorithm performs clustering and feature weighting simultaneously, which learns different sets of feature weights for each cluster during the process of clustering. SCAD algorithm has two advantages. Firstly, it contributes to partition the dataset into more meaningful clusters. Secondly, it can be used as part of a more complex learning system to enhance its learning behavior. SCAD algorithm is effective when it used as part of supervised or unsupervised learning systems. SCAD algorithm is extended to simultaneous text document clustering and dynamic category-dependent keyword set weighting and it performs better than traditional text clustering methods [12].

* Corresponding author.
E-mail address: zhangys@sdu.edu.cn (Y. Zhang).

http://dx.doi.org/10.1016/j.jbi.2016.05.007
1532-0464/

The idea of SCAD algorithm can be used to select informative genes with gene expression data. However, it turns out that the performance of SCAD is not good. The reason is that SCAD algorithm considers that weights of one feature are different for each cluster (local adaptive distance). In some situations local adaptive distance may not be appropriate because they lead to falling into local minima, providing suboptimal solutions [13]. Considering this question, we developed our method assuming that the weight of one feature is same for all clusters (global adaptive distance).

In this paper, we present a new filter method for gene selection, Kernel-Based Clustering method for Gene Selection (KBCGS). In our method, we assign different weights to different genes, taking each class as a known cluster. Then the optimal weights of genes are obtained by minimizing the clustering objective function. In the process of clustering, dissimilarity measures are obtained as sums of Euclidean distances between samples and cluster centers, which computed individually for each gene by means of kernel functions [13,14]. The bigger weights of genes, the more likely that they are informative for sample classification. So top-ranked genes are selected to perform cancer classification.

Although the proposed method is similar to SCAD, compared to SCAD, the proposed method has some differences: (1) SCAD is unsupervised learning algorithm, while the proposed method is supervised learning algorithm. (2) SCAD uses local adaptive distance, which will fall into local minima, while the proposed method uses global adaptive distance that can avoid this problem. (3) The dissimilarity measure of SCAD is simple Euclidean distance, which only performs well when the dataset is linearly separable, while the dissimilarity measure of the proposed method is obtained by means of kernels, which can produce non-linear separating hyper-surfaces among clusters. The introduced modifications can improve the performance in this context.

The proposed method allows us to use adaptive distances which change at each step of the iteration. This kind of dissimilarity measure is suitable to learn the weights of the features during the clustering process, improving the performance of the algorithm. Furthermore, the proposed algorithm is simple to implement and does not require any modification or parameter optimization for each dataset. The validation of the KBCGS method is carried out on eight public cancer datasets, with two well-known classifiers (KNN, SVM), compared with other popular gene selection methods namely, Relief-F [15], minimal-redundancy-maximal relevance (MRMR) [16], Kruskal Wallis-Test [17], Gini Index [18], Information Gain [19], χ^2-Statistic [20]. The results show that our method is efficient comparing with other feature selection methods. And in order to illustrate the significance of KBCGS method, randomization studies are performed and the most important genes selected by the method are listed.

2. Materials and methods

Consider gene expression dataset $X \in \mathbb{R}^{n\times p}$, it is composed of n samples from different patients. Each sample is represented by the expression level of p genes, denoted as a row vector $x_j \in \mathbb{R}^p$, and labeled by an ordinal scale $y_j \in Y$, $Y = \{1, 2, \ldots, C\}$. Let $w = [w_1, w_2, \ldots, w_p]$ represents the weights of p genes, the relative importance of genes for classification. We treat the C classes as C established clusters, so cluster centers $v_i = [v_{i1}, v_{i2}, \ldots, v_{ip}]$ can be calculated as follows:

$$v_{ik} = \frac{\sum_{x_j \in C_i} x_{jk}}{|C_i|}, \tag{1}$$

in which $i = 1, 2 \ldots, C$, $j = 1, 2, \ldots, n$, $k = 1, 2, \ldots, p$, $|C_i|$ is the number of samples in class C_i.

Clustering takes intrinsic characteristics of data into account, organizing a set of samples into clusters such that patterns within a given cluster have a high degree of similarity, whereas patterns belonging to different clusters have a high degree of dissimilarity [13,21]. It is essential to determine the dissimilarity measures in clustering. Different dissimilarity measures may lead to different clustering results, so the appropriate measure is very important. Euclidean distance is the most commonly used, which performs well when the natural clusters are nearly hyper-spherical and linearly separable. But the clusters are not always in that shapes. In order to solve this problem, several methods have been proposed. In Ref. [14], the authors presented a survey of kernel and spectral clustering methods, two approaches able to produce nonlinear separating hypersurfaces between clusters. Kernel clustering methods are the kernel version of many classical clustering algorithms, e.g., SOM and neural gas. Spectral clustering arises from concepts in spectral graph theory and the clustering problem is configured as a graph cut problem where an appropriate objective function has to be optimized. Gustafson–Kessel clustering method was presented to detect clusters of different geometrical shapes [46]. In this paper we use kernel-based clustering methods, because kernel method can be used to reveal the intrinsic relationships that are hidden in the raw data.

Let $\Phi : X \to \mathcal{F}$ be a non-linear mapping from the input space X to a high-dimensional new feature space $\mathcal{F}$. Then the dot product $x_k^T x_l$ in the original input space is mapped to $\Phi(x_k)^T \Phi(x_l)$ in the new feature space. Each Mercer kernel can be expressed as [22]:

$$K(x_k, x_l) = \Phi(x_k)^T \Phi(x_l) \tag{2}$$

So, we can compute Euclidean distances in $\mathcal{F}$ as follows:

$$\begin{aligned} \|\Phi(x_k) - \Phi(x_l)\|^2 &= (\Phi(x_k) - \Phi(x_l))^T (\Phi(x_k) - \Phi(x_l)) \\ &= K(x_k, x_k) - 2K(x_k, x_l) + K(x_l, x_l) \end{aligned} \tag{3}$$

Then the dissimilarity function between samples and centers is obtained by means of kernels [50]:

$$\begin{aligned} \varphi^2(x_j, v_i) &= \sum_{k=1}^{p} \left\| \Phi(x_{jk}) - \Phi(v_{ik}) \right\|^2 \\ &= \sum_{k=1}^{p} \{K(x_{jk}, x_{jk}) - 2K(x_{jk}, v_{ik}) + K(v_{ik}, v_{ik})\} \end{aligned} \tag{4}$$

By using weighted dissimilarity measures for objects, the clustering methods gain competitive results, especially for analyzing large datasets [13,23]. In Ref. [13], the authors presented an algorithm labeled VKFCM-K-GS, which considers the global adaptive distance. Experiments with synthetic and benchmark datasets show that it is effective. We employ the global adaptive distance, by assigning different weights to different genes. Similar to SCAD [11] algorithm, our clustering objective function is defined:

$$\begin{aligned} J &= \sum_{i=1}^{C} \sum_{x_j \in C_i} \varphi^2(x_j, v_i) + \delta \sum_{k=1}^{p} w_k^2 \\ &= \sum_{i=1}^{C} \sum_{x_j \in C_i} \sum_{k=1}^{p} w_k \left\| \Phi(x_{jk}) - \Phi(v_{ik}) \right\|^2 + \delta \sum_{k=1}^{p} w_k^2 \end{aligned} \tag{5}$$

In which $w = (w_1, \ldots, w_p)$, subject to

$$\begin{cases} w_k \in [0, 1], \quad k = 1, \ldots, p \\ \sum_{k=1}^{p} w_k = 1 \end{cases} \tag{6}$$

As we can see in Eq. (5), the objective function contains two components. The first part is the sum of weighted dissimilarity distance between samples and their cluster centers by means of kernel functions, which allows us to obtain compact clusters. It is minimized when only one gene is completely relevant, and all