Journal of Biomedical Informatics 53 (2015) 381–389

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A fast gene selection method for multi-cancer classification using multiple support vector data description

Jin Cao [a], Li Zhang [a,b,*], Bangjun Wang [a], Fanzhang Li [a], Jiwen Yang [a,b]

[a] School of Computer Science and Technology & Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, Jiangsu, China
[b] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210000, Jiangsu, China

ARTICLE INFO

Article history:
Received 4 July 2014
Accepted 18 December 2014
Available online 27 December 2014

Keywords:
Support vector data description
Gene selection
Multi-class classification
Gene expression data
Support vector machine

ABSTRACT

For cancer classification problems based on gene expression, the data usually has only a few dozen sizes but has thousands to tens of thousands of genes which could contain a large number of irrelevant genes. A robust feature selection algorithm is required to remove irrelevant genes and choose the informative ones. Support vector data description (SVDD) has been applied to gene selection for many years. However, SVDD cannot address the problems with multiple classes since it only considers the target class. In addition, it is time-consuming when applying SVDD to gene selection. This paper proposes a novel fast feature selection method based on multiple SVDD and applies it to multi-class microarray data. A recursive feature elimination (RFE) scheme is introduced to iteratively remove irrelevant features, so the proposed method is called multiple SVDD-RFE (MSVDD-RFE). To make full use of all classes for a given task, MSVDD-RFE independently selects a relevant gene subset for each class. The final selected gene subset is the union of these relevant gene subsets. The effectiveness and accuracy of MSVDD-RFE are validated by experiments on five publicly available microarray datasets. Our proposed method is faster and more effective than other methods.

1. Introduction

Cancer classification is one of the conventional problems in microarray gene expression data [1–7]and includes tumor detection and prediction of some rare diseases [8–11]. Many methods have been applied to cancer classification, such as naive Bayes classifier (NBC) [12], partial least squares discriminant analysis (PLSDA) [13], support vector Machines (SVMs) [14,15], and *k* nearest neighbor (*k*NN) [16].

The accuracy of cancer classification largely depends on the biological relevance of genes [17]. Thus, gene selection can be viewed as a key stage for cancer classification based on microarray data and feature selection algorithms have been rapidly developed in the past few decades. For the gene expression data, the expression level of some genes is highly correlated, which plays an important role in biological evolution. When these genes are located on the same biological path, this correlation is more pronounced [18]. In this case, traditional feature selection methods ignore the relationships between genes, and choose only a few from these highly related genes. The irrelevant genes not only result in lower classification performance, but also add extra difficulties in finding informative genes [19,20].

As a general learner, SVM could be applied to the problems of classification [21], regression [22], and feature extraction [23]. Here, we focus on feature selection. Considering whether the evaluation criterion involves classification models, we can divide SVM-based feature selection methods into three groups: wrapper feature selection based on SVM, embedded feature selection based on SVM, and hybrid feature selection of filter and wrapper based on SVM.

Weston et al. proposed the wrapper feature selection algorithm based on SVM, which finds useful features by minimizing bounds on the leave-one-out error using gradient descent [24]. Guyon et al. proposed a SVM-RFE (recursive feature elimination) feature selection algorithm, which is the most representative algorithm [23]. Duan et al. presented a gene selection method similar to SVM-RFE, called MSVM-RFE. At each step, MSVM-RFE trains multiple linear SVMs on subsamples of training data and computes the feature ranking scores from statistical analysis of the weight vectors [25]. However, MSVM-RFE is computationally more expensive than SVM-RFE. Embedded feature selection algorithms based on SVM are similar to other embedded methods. Li et al. proposed an embedded feature selection algorithm [26] that can adaptively identify important features through introducing data driven

* Corresponding author at: School of Computer Science and Technology & Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, Jiangsu, China.

http://dx.doi.org/10.1016/j.jbi.2014.12.009
1532-0464/

weights, which simultaneously implements classification and gene selection. However, this method requires adjusting more parameters, and its performance largely depends on those parameters. How to design the objective function based on the standard SVM is the key and difficult issue of this type of algorithm. Lee and Leu proposed a hybrid feature selection algorithm based on SVM for microarray data analysis [27]. Specifically, this method uses the genetic algorithm to generate a number of subsets of genes, the chi square test to select a proper number of the top-ranked genes for data analysis, and SVM to verify the efficiency of the selected genes.

Among the methods mentioned above, SVM-RFE has attracted considerable attention for its simplicity and intuition. However, since SVM-RFE can only be applied to binary classification problems, Jeong et al. proposed feature selection algorithms based on support vector data description (SVDD) for one-class classification problems [28]. SVDD can describe a target data distribution, also called one-class SVM [29–31]. Two feature selection algorithms based on SVDD were presented in [28], or SVDD-radius-RFE and SVDD-dual-objective-RFE. SVDD-radius-RFE minimizes the boundary of target samples measured by its radius squared [28]. SVDD-dual-objective-RFE maximizes a dual-objective function of SVDD to provide a compact description boundary. However, it is very time-consuming to apply both SVDD-based methods to gene selection.

SVDD-based feature selection methods were proposed for one-class classification problems, and they cannot be applied to multi-class problems. Although SVM-RFE can be extended to multi-class classification problems by using the strategies of one-against-all [32], one-against-one [33], decision tree [34], and so on, its training speed is not optimistic.

To solve multi-class classification problems and reduce the time complexity of both SVM-RFE and the SVDD-based feature selection methods, we propose a multiple SVDD-RFE (MSVDD-RFE) method. Specifically, we first learn multiple feature selection models via multiple SVDD models, where each class has a corresponding model. For each model, we remove features according to the direction energy of its center vector. If the energy in some direction is small, the feature corresponding to this direction is eliminated. In doing so, multiple feature subsets are obtained by multiple models. Then, we combine these subsets into the selected feature subset. The concept of MSVDD-RFE can be generalized to SVDD-radius-RFE and SVDD-dual-objective-RFE when handling multi-class classification problems. We validate that MSVDD-RFE provides more precise classification performance and less time consumption, by experiments on five public microarray datasets.

Section 2 introduces SVDD and two SVDD-based feature selection algorithms, and proposes multiple SVDD for feature selection. Simulation experiments are presented in Section 3 and our conclusions are presented in Section 4.

2. Methods

We introduce SVDD and the two feature selection methods based on SVDD, and propose the multiple SVDD-based feature selection method.

2.1. Support vector data description

SVDD is a one-class classifier [29–31]. Compared with SVM, SVDD only allows learning from one-class data. SVDD can be implemented using hyperplane or hypersphere methods. The former takes the origin as an abnormal point and makes an optimal hyperplane away from it as far as possible [35], and the latter constructs a hypersphere to contain as many targets as possible [29,31]. For SVDD, we only need one-class data or target samples to construct the learning model expressed by a hypersphere. If a point falls within the hypersphere, it belongs to the target sample set; otherwise, it would be an abnormal point, or outlier.

Given a set of target samples $\{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the target sample, D is the dimensionality of the target sample, and n is the number of target sample, we try to find a hypersphere with minimum volume containing all (or most of) the data. To achieve this, we need to know two parameters, the hypersphere center, $\mathbf{a}$, and radius, R. The initial form of optimization problem is

$$\begin{aligned} \min_{R,\mathbf{a},\xi_i} \quad & R^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \|\mathbf{x}_i - \mathbf{a}\|^2 \leqslant R^2 + \xi_i, \xi_i \geqslant 0, i = 1,\ldots,n, \end{aligned} \tag{1}$$

where ξ_i is a slack variable, and $C > 0$ is the penalty factor which allows trade-off between the volume of hypersphere and the number of target objects rejected.

Using the Lagrange multiplier technique, we obtain the dual programming form of (1),

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{n}\alpha_i\mathbf{x}_i^T\mathbf{x}_i \\ \text{s.t.} \quad & \sum_{i=1}^{n}\alpha_i = 1\ , 0 \leqslant \alpha_i \leqslant C, i = 1,\ldots,n, \end{aligned} \tag{2}$$

where α_i is the Lagrange multiplier. The hypersphere center, $\mathbf{a}$, and radius, R, can be expressed by Lagrange multipliers,

$$\mathbf{a} = \sum_{i=1}^{n}\alpha_i\mathbf{x}_i, \tag{3}$$

and

$$R^2(\mathbf{x}_{sv}) = \|\mathbf{x}_{sv} - \mathbf{a}\|^2 = \mathbf{x}_{sv}^T\mathbf{x}_{sv} - 2\sum_{i=1}^{n}\alpha_i\mathbf{x}_{sv}^T\mathbf{x}_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j, \tag{4}$$

where $\mathbf{x}_{sv}$ is a support vector with its corresponding Lagrange multiplier $0 < \alpha_{sv} < C$.

2.2. SVDD-based feature selection methods

Two SVDD-based feature selection methods were proposed in [28] and are described briefly here. There are two cases, considering either a few outliers and targets, or only considering targets. We will only discuss the latter.

2.2.1. SVDD-radius-RFE

The performance of SVDD strongly depends upon how compactly the constructed hypersphere describes the target samples, while discriminating outliers [28]. The size of hypersphere can be characterized by its radius. The criterion for feature selection in SVDD-radius-RFE is related to the hypersphere radius. The term $R^2(\mathbf{x}_{sv})$ in (4) is the radius square based on $\mathbf{x}_{sv}$. Let SV be the support vector set. Then the average radius square, J_r, is defined as

$$J_r = \sum_{\mathbf{x}_{sv}\in SV}\frac{R^2(\mathbf{x}_{sv})}{|SV|}, \tag{5}$$

which may be rewritten as

$$J_r = \frac{1}{|SV|}\sum_{\mathbf{x}_{sv}\in SV}\left(\mathbf{x}_{sv}^T\mathbf{x}_{sv} - 2\sum_{i=1}^{n}\alpha_i\mathbf{x}_{sv}^T\mathbf{x}_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j\right). \tag{6}$$

Let $J_r(-p)$ be the size of hypersphere excluding feature p. The worst feature is the one with the largest $J_r(-p)$. That is, feature p has little effect on the size of the hypersphere. The worst feature [28] is the one with the smallest value of $(J_r - J_r(-p))$. Thus, the criterion function of SVDD-radius-RFE is