Computational Biology and Chemistry 67 (2017) 92–101

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

A L1-regularized feature selection method for local dimension reduction on microarray data

Shun Guo[a,b], Donghui Guo[a,*], Lifei Chen[c], Qingshan Jiang[b]

[a] Department of Electronic Engineering, Xiamen University, Fujian 361005, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China
[c] School of Mathematics and Computer Science, Fujian Normal University, Fujian, 350117, China

ARTICLE INFO

Article history:
Received 18 October 2015
Received in revised form 4 November 2016
Accepted 29 December 2016
Available online 31 December 2016

Keywords:
Local dimension reduction
Classification
L1-regularized logistic regression
Microarray data
Partial least squares (PLS)

ABSTRACT

Dimension reduction is a crucial technique in machine learning and data mining, which is widely used in areas of medicine, bioinformatics and genetics. In this paper, we propose a two-stage local dimension reduction approach for classification on microarray data. In first stage, a new L1-regularized feature selection method is defined to remove irrelevant and redundant features and to select the important features (biomarkers). In the next stage, PLS-based feature extraction is implemented on the selected features to extract synthesis features that best reflect discriminating characteristics for classification. The suitability of the proposal is demonstrated in an empirical study done with ten widely used microarray datasets, and the results show its effectiveness and competitiveness compared with four state-of-the-art methods. The experimental results on St Jude dataset shows that our method can be effectively applied to microarray data analysis for subtype prediction and the discovery of gene coexpression.

1. Introduction

Dimension reduction is a key technique in data mining and machine learning aimed at using a small number of features or synthesis features (component variables) to replace the feature subsets with strong correlations in the original data (Yan et al., 2006). This technique has been studied in various areas of applications, including image recognition, text mining and microarray data analysis. As high-dimensional microarray data typically have many more variables than observations and contain significant noise or outliers, it remains a challenging task for microarray data analysis. In this paper, we consider dimension reduction for the purpose of microarray data classification.

According to the relationship between the selected feature subset and the categories, dimension reduction methods can be divided into global, local two classes. The global dimension reduction methods select a few features or synthesis features (component variables) for all categories, such as PCA (Belhumeur et al., 1997), LDA (Duda et al., 2001) and LPP (Niyogi, 2004). The challenge of these methods arises from the fact that there is usually complex category structure in microarray data. For example, the coexpress genes for one category may not occur in another category. In local dimension reduction methods, features or synthesis features are selected for each category of the original data.

Recently, some local dimension reduction methods have been suggested (You et al., 2014a, 2014b; Chen and Wang, 2012; Guo and Guo, 2015). Generally, a two-stage strategy is used in such a method: examples include TotalPLS (You et al., 2014a, 2014b) and the centroid-based method (Guo and Guo, 2015). In first stage, a feature selection method is used to eliminate the irrelevant and redundant features. Subsequently, the selected features are transformed into a small number of synthesis features (component variables) by using feature extraction techniques (e.g., PLS (Barker and Rayens, 2003)) in the next stage. The first stage is especially crucial due to the noise or outliers contained in microarray data.

Many traditional feature selection methods lack an effective mechanism to deal with noise and outliers, such as Trace Ratio (Nie et al., 2008), F-test (Le Cao et al., 2009) and Similarity Preserving Feature Selection (SPFS) (Zhao et al., 2013). To handle this problem, some methods (e.g., PLSRFE (You et al., 2014a, 2014b) and MSVM-RFE (Zhou and Tuck, 2007)) use the Recursive Feature Elimination (RFE) strategy (Guyon et al., 2002), which heuristically removes the features with small feature weights in each iteration. However, there is no guarantee that a useful feature has to have a weight larger than other features and some useful features may be eliminated during the process of RFE (Sun et al., 2010).

L1 regularization, which adds a L1 regularized penalty to the objective function, is usually deemed as an effective way to deal

* Corresponding author.
E-mail address: dhguo@xmu.edu.cn (D. Guo).

http://dx.doi.org/10.1016/j.compbiolchem.2016.12.010
1476-9271/

with noise and outliers (Li et al., 2015). Popular L1-regularized methods include L1-SVM (Zhu et al., 2004) and L1 logistic regression (Ng, 2004). However, it is difficult to solve the L1-regularized objective function efficiently due to the non-differentiability of L1 regularization (Liu et al., 2013). Recently, several L1-regularized LDA methods (e.g., Li et al., 2015; Wang et al., 2014) have been developed in classification problems for high-dimensional data and they can solve the L1-regularized problem efficiently by using some iterative techniques. One main drawback of these methods is that they can only obtain local optimum instead of global optimum. Some L1-regularized logistic regression methods (e.g., Shi et al., 2010; Yuan et al., 2012; Tan et al., 2013) are able to obtain the optimal solution under some strict conditions. Nevertheless, it is still challenging for them to identify the most discriminative subset from millions of features (Tan et al., 2013). LLFS (Sun et al., 2010), which combines L1-regularized logistic regression with the large margin framework, shows interesting performance on high-dimensional data. However, LLFS has high computational cost when applied to multiclass problems. It is necessary to consider the balance between the computational cost and classification accuracy for microarray data classification.

In this paper, we propose an efficient L1-regularized feature selection method that combines LDA with L1-regularized logistic regression for local dimension reduction on microarray data. The objective function is solved by using Fletcher–Reeves conjugate gradient descent. And the theoretical analyses suggest that the global optimal solution can be obtained with a non-zero initial point. As PLS (Partial Least Squares) has been extensively used in high-dimensional microarray data analysis, such as SVA-PLS (Chakraborty and Datta, 2012) and PCA-PLS (Li et al., 2015), we also use it in the second stage of our method. However, most dimension reduction methods only focus on improving the recognition accuracy for microarray datasets. We try to predict the subtype of selected genes (features) by using PLS-based method and the experimental results on St Jude dataset verify the rationality of our method to some extent. Moreover, the performance of our dimension reduction method is evaluated on ten publicly available microarray datasets, and the results show its effectiveness. Compared with our previous work (Guo and Guo, 2015), the method presented in this paper is much motivated, analyzed and experimentally evaluated. In particular, we proved that the L1-regularized feature selection method can obtain an optimal solution with a non-zero initial point in this paper.

The rest of the paper is organized as follows: Section 2 gives detailed description of the proposed local dimension reduction method. Experimental evaluation is presented and discussed in Section 3. Subtype prediction on St Jude dataset is provided in Section 4. In Section 5, our conclusions and the future work are presented.

2. The proposed local dimension reduction method

The proposed local dimension reduction method contains two stages. First, a L1-regularized feature selection method (see Section 2.1) is used for feature elimination. After that, PLS-based feature extraction (see Section 2.2) is implemented on the selected features, which projects the features into low-dimensional space.

2.1. The feature selection method

This section presents a feature selection method using L1-regularized logistic regression. The method uses the class separability as a criterion for feature selection and it can obtain the optimal solution efficiently with a non-zero initial point.

2.1.1. Class separability measure

Class separability means large between-class distance but small within-class distance. Let $S = [x_1, x_2, \cdots, x_n]^T \in R^{n\times d}$ denote a training data set, where x_i is the i-th sample of the data containing d features, $Y = [y_1, y_2, \cdots, y_n]^T \in R^{n\times 1}$ is the corresponding class labels, and $d \gg n$. Let $x_i^{(j)}$ denotes the i-th sample in class j, $W = [w_1, w_2, \cdots, w_d] \in R^{1\times d}$ be a weight vector, the between-class distance and the within-class distance are defined as:

$$S_b = d_w(u_j, u_k) \tag{1}$$

and

$$S_w = \sum_{j=1}^{g}\sum_{i=1}^{n_j} d_w\left(x_i^{(j)}, u_j\right), \tag{2}$$

where g is the number of classes, n_j is the number of samples in class j, u_j is the centroid of the class j, i.e., $u_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i^{(j)}$ and $d_w(,)$ is a distance function about W. Varieties of distances $d_w(,)$ have been proposed in some references (Sun et al., 2010; Liu et al., 2013). For the purpose of this paper, $d_w(,)$ is defined as the following:

$$d_w(x, y) = \|x - y_w\|, \tag{3}$$

with $\|Z_W\| = \sum_i w_i\sqrt{z_i{}^2}$.

2.1.2. Objective function

Here, we first consider binary problems, while in Section 3.1.4 the method is extended to multiclass settings. In the same spirit as LDA (Fisher, 1936), which maximizes the between-class distance while minimizing the within-class distance, our objective function is given by:

$$\min_w J(W) = S_w - rS_b, \tag{4}$$

where r is the balance coefficient to make the second term of the function can get the same range of the first term. So, r can be set to equal to n (n is the number of all samples).

The objective function (4) can be simplified as:

$$\min_w J(W) = -\sum_{i=1}^{n} WZ_i^T, \tag{5}$$

where $Z_i = \left(\sqrt{(x_{i1}^{(j)} - u_{j1})^2} - \sqrt{(u_{11} - u_{21})^2}, \ldots, \sqrt{(x_{i1}^{(j)} - u_{jd})^2} - \sqrt{(u_{1d} - u_{2d})^2}\right)$, with $x_{ik}^{(j)}, u_{ik}$ the k-th element of $x_i^{(j)}$ and u_i, respectively.

Motivated by LLFS, we introduce L1 regularized logistic regression to formulate the new objective function as:

$$\min_w K(W) = \sum_{i=1}^{n}\log\left(1 + exp\left(-WZ_i^T\right)\right) + \lambda\|W\|_1, \tag{6}$$

where λ is the regularization parameter that controls the sparseness of the solution.

Let us consider the following optimization problem:

$$\begin{cases} \min_w K(W) = \sum_{i=1}^{n}\log\left(1 + exp\left(-WZ_i^T\right)\right) + \lambda\|W\|_1 \\ s.t. W \geq 0. \end{cases} \tag{7}$$

The introduction of the constraint $W \geq 0$ enables us to obtain the optimal solution, where it has been previously used in many methods (e.g., (Sun et al., 2010; Cai et al., 2010)).