ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 12, No. 3, 2017, *pp*.210-215

Gene expression data classification using exponential locality sensitive discriminant analysis

Chunming Xu [1]
[1] *School of Mathematics and Statistical, Yancheng Teachers University,*
Yancheng 224002,PR China
E-mail: ycxcm@126.com.

(Received April 05, 2017, accepted June 20, 2017)

Abstract. Locality sensitive discriminant analysis is a typical and very effective graph-based dimensionality reduction method which has been successfully applied in pattern recognition problems. LSDA aims to find a projection which maximizes the margin between data points from different classes at each local area. As a result, it can discover the local geometrical structure of the data samples. However, just as linear discriminant analysis, it has the small sample size (SSS) problem. To overcome this limitation, we propose a novel exponential locality sensitive discriminant analysis algorithm in this paper. The proposed algorithm can make nearby objects with the same labels in the input space also nearby in the new representation; while nearby objects with different labels in the input space should be far apart. In addition, it can also deal with the SSS problem. The experiments on gene expression data sets verify the effectiveness of the proposed algorithm.

Keywords: gene expression data classification; dimensionality reduction; locality sensitive discriminant analysis; exponential locality sensitive discriminant analysis

1. Introduction

In recent years, with the rapid development of microarray gene-expression technology, it is now possible to simultaneously monitor the expression of all genes in the genome with a single experiment. One important application of gene expression data is the classification of cancer or other diseases, which draws a great number of researchers' attention [1-3]. Typically, the gene expression data sets are characterized by thousands of variables on only a few observations. It has been observed that although there are a lot of genes for each observation, the number of tissue samples ranges from tens to hundreds. In other words, there is much redundant information resided in the high-dimensional gene-expression data. To remove redundant information, dimensionality reduction technique is an effective way.

During the past decades, many dimensionality reduction algorithms have been developed. Classical dimensionality reduction methods can be categorized into two class: unsupervised methods and supervised methods. Representatives of unsupervised methods are principal component analysis (PCA) [4], independent component analysis (ICA) [5] and locality preserving projections (LPP) [6]. Unfortunately, unsupervised dimensionality reduction methods don't utilize any class information so they are not suitable for classification problems. For supervised methods, linear discriminant analysis (LDA) [7] is one of the most popular dimensionality reduction techniques. LDA seeks the optimal transformation that maximizing the between-class scatter while at the same time minimizing the within-class scatter. LDA has been widely used in many practical applications such as image retrieval and face recognition due to the fact that it can extract the most discriminatory features. Many extended LDA algorithms have been developed, for example, regularized discriminant analysis (RDA) [8], kernel linear discriminant analysis (KLDA) [9], two dimensional linear discriminant analysis (2DLDA) [10], locality sensitive discriminant analysis (LSDA) [11] and so on.

However, in many case, the number of samples is smaller than the dimensionality of the samples which will leads to the SSS problem in linear discriminant analysis based methods. In this paper, the SSS problem of LSDA algorithm is considered and a novel exponential LSDA (ELSDA) algorithm is proposed to overcome the shortcoming. The proposed ELSDA method not only inherits the advantages of the LSDA algorithm, but also avoids the SSS problem.

The rest of this paper is organized as follows. Section 2 gives a brief review of LSDA. Our ELSDA approach is proposed in section 3. Experimental results and some conclusions are provided in sections 4 and 5, respectively.

2. Locality sensitive discriminant analysis

LDA is a supervised learning algorithm, which can take advantage of the classification information of samples, and has received wide attention in the field of pattern recognition. However, LDA does not consider the distribution of samples, so it is not very good to deal with the data with nonlinear geometric distribution. To resolve this drawback, a local sensitive discriminant analysis (LSDA) method is proposed in literature [11]. LSDA can make full use of both the label information and the local manifold structure information of labeled samples to guide the dimensionality reduction process.

Given l data points $\boldsymbol{x_1},\dots,\boldsymbol{x}_l$ that are distribduted on a underlying submanifold. Let $l(\boldsymbol{x_i})$ be the class label of $\boldsymbol{x_i}$ and its k nearest neighbors be $N(\boldsymbol{x_i})=\{\boldsymbol{x_i^1},\boldsymbol{x_i^2},\dots,\boldsymbol{x_i^k}\}$ .By the label information, the set $N(\boldsymbol{x_i})$ can be further divided into two non overlapping subsets, $N_b(\boldsymbol{x_i})$ and $N_w(\boldsymbol{x_i})$. $N_w(\boldsymbol{x_i})$ contains the neighbors having the same label with $\boldsymbol{x_i}$, while $N_b(\boldsymbol{x_i})$ contains the neighbors sharing different labels. Specifically,

$$N_w(\boldsymbol{x_i})=\{\boldsymbol{x_i^j}\,|\,l(\boldsymbol{x_i^j})=l(\boldsymbol{x_i}),1\le j\le k\} \tag{1}$$

$$\mathrm{N_b}(\boldsymbol{x_i})=\{\boldsymbol{x_i^j}\,|\,\mathrm{l}(\boldsymbol{x_i^j})\ne \mathrm{l}(\boldsymbol{x_i}),1\le \mathrm{j}\le \mathrm{k}\} \tag{2}$$

Define the weight matrices W_b and W_w respectively as follows:

$$W_{b,ij}=\begin{cases}1,\boldsymbol{x_i}\in N_b(\boldsymbol{x_j}) \quad or \quad \boldsymbol{x_j}\in N_b(\boldsymbol{x_i})\\ 0, otherwise\end{cases} \tag{3}$$

$$W_{w,ij}=\begin{cases}1,\boldsymbol{x_i}\in N_w(\boldsymbol{x_j}) \quad or \quad \boldsymbol{x_j}\in N_w(\boldsymbol{x_i})\\ 0, otherwise\end{cases} \tag{4}$$

The goal of LSDA is to maximize $\sum_{ij}(\boldsymbol{y_i}-\boldsymbol{y_j})^2W_{b,ij}$ while at the same time minimize

$$\sum_{ij}(\boldsymbol{y_i}-\boldsymbol{y_j})^2W_{w,ij}\ .$$

It is easy to derive that

$$\sum_{ij}(\boldsymbol{y_i}-\boldsymbol{y_j})^2W_{b,ij}=\frac{1}{2}(\boldsymbol{p}^T\boldsymbol{x_i}-\boldsymbol{p}^T\boldsymbol{x_j})^2W_{b,ij}=\frac{1}{2}\boldsymbol{p}^T\boldsymbol{X}(\boldsymbol{D_b}-\boldsymbol{W_b})\boldsymbol{X}^T\boldsymbol{p},$$

where $y_i=\boldsymbol{p}^T\boldsymbol{x_i}$; $\boldsymbol{D_b}$ is a diagonal matrix which is satisfied that $D_{b,ii}=\sum_j W_{b_{ij}}$. On the other hand,

$$\sum_{ij}(\boldsymbol{y_i}-\boldsymbol{y_j})^2W_{w,ij}=\frac{1}{2}(\boldsymbol{p}^T\boldsymbol{x_i}-\boldsymbol{p}^T\boldsymbol{x_j})^2W_{w,ij}=\frac{1}{2}\boldsymbol{p}^T\boldsymbol{X}(\boldsymbol{D_w}-\boldsymbol{W_w})\boldsymbol{X}^T\boldsymbol{p},$$

where $\boldsymbol{D_w}$ is a diagonal matrix which is satisfied that $D_{w,ii}=\sum_j W_{w,ij}$.

Thus the objective function of LSDA can be written as:

$$J(p)=\arg\max_p \alpha\boldsymbol{p}^T\boldsymbol{X}(\boldsymbol{D_b}-\boldsymbol{W_b})\boldsymbol{X}^T\boldsymbol{p}-(1-\alpha)\boldsymbol{p}^T\boldsymbol{X}(\boldsymbol{D_w}-\boldsymbol{W_w})\boldsymbol{X}^T\boldsymbol{p} \tag{5}$$

where α is a positive parameter and $0\le\alpha\le 1$.

Denote that $\boldsymbol{L_b}=\boldsymbol{D_b}-\boldsymbol{W_b}$, $\boldsymbol{L_w}=\boldsymbol{D_w}-\boldsymbol{W_w}$, then $J(p)$ can be further written as

$$J(p)=\arg\max_p \boldsymbol{p}^T\boldsymbol{X}(\alpha\boldsymbol{L_b}-(1-\alpha)\boldsymbol{L_w})\boldsymbol{X}^T\boldsymbol{p} \tag{6}$$

Constrain that $\boldsymbol{p}^T\boldsymbol{X}\boldsymbol{D_w}\boldsymbol{X}^T\boldsymbol{p}=1$, then the objective function (4) can be recast as the following optimization problem: