Ziqiang WANG[1,2], Xia SUN[1], Lijun SUN[1], Xu QIAN[2]

Henan University of Technology (1), China University of Mining & Technology(Beijing) (2)

Tissue Classification Using Efficient Local Fisher Discriminant Analysis

Abstract. *A novel scatter-difference-based local Fisher discriminant analysis(SDLFDA) algorithm for tissue classification is proposed in this paper. SDLFDA explicitly considers the local manifold structure and interclass discrimination in gene expression data space. By using SDLFDA, each gene expression data can be projected into a lower-dimensional discriminative feature space. In addition, SDFLDA reduces the computational cost through QR decomposition. Experimental results demonstrate the effectiveness and efficiency of the proposed SDLFDA algorithm.*

Streszczenie. *W artykule przedstawiono algorytm analizy lokalnym wyróżnikiem Fisher'a opartym na różnicach rozproszenia (ang. SDLFDA), służący do klasyfikacji tkanek. Proponowana metoda pozwala na zmniejszenie wymiarowości przestrzeni wyróżnika, określającego dane GXD, a także redukcję kosztów obliczeniowych dzięki dekompozycji QR. Wyniki badań eksperymentalnych potwierdzają skuteczność i sprawność algorytmu. (**Efektywna analiza lokalnego wyróżnika Fisher'a do klasyfikacji tkanek**).*

Keywords: tissue classification, gene expression data, dimensionality reduction, local Fisher discriminant analysis.
Słowa kluczowe: klasyfikacja tkanek, GXD, redukcja wymiarowości, analiza lokalnego wyróżnika Fisher'a.

Introduction

Tissue classification based on gene expression has received extensive attention in recent years because of its potential applications in many fields[1,2]. Usually, tissue classification is very difficult due to the curse of dimensionality, a common way to attempt to resolve this problem is to use dimensionality reduction techniques. Two of the most popular techniques are principal component analysis(PCA) and linear discriminant analysis(LDA)[3]. However, both PCA and LDA algorithms see only global Euclidean structure and cannot discover the underlying manifold structure hidden in the high-dimensional data.

Recently, a number of manifold learning algorithms have been proposed to discover the geometric property of high-dimensional data that lie on or near a submanifold of the observation space[4,5], and they have been successfully applied to face recognition, document analysis, and microarray data analysis. Unfortunately, all of these algorithms suffer from the out of sample problem. One common response to cope with this problem is to apply a linearization procedure to construct explicit maps over new samples. The most representative such algorithm is locality preserving projection(LPP)[6].

In this paper, we propose an efficient scatter-difference-based LFDA(SDLFDA) algorithm to overcome the above two problems. The main contributions of this paper include: 1) the scatter-difference-based local discriminant analysis is proposed to avoid the matrix singularity problem; and 2) By using QR-decomposition, SDLFDA casts local discriminant analysis into a much smaller matrix computation, which greatly facilitates efficient computation.

Brief review of LFDA

LFDA is a recently proposed manifold algorithm, it aims to search for the directions on which the between-class separability is maximized and at the same time the within-class local structure is preserved [7]. It is based on locality preserving projection and explicitly considers the interclass discrimination. Suppose we have a set of gene expression data $X=\{x_1,x_2,\cdots,x_n\}\subset\mathbb{R}^N$ belonging to c classes, the number of gene expression data in the ith class is n_i satisfying $\sum_{i=1}^{c} n_i = n$. The objective function of LFDA is

$$V_{opt}=\arg\max_{V}\frac{V^T S^{(b)} V}{V^T S^{(w)} V} \tag{1}$$

$$S^{(b)}=\frac{1}{2}\sum_{i,j=1}^{n} W_{ij}^{(b)}\left(x_i-x_j\right)\left(x_i-x_j\right)^T \tag{2}$$

$$S^{(w)}=\frac{1}{2}\sum_{i,j=1}^{n} W_{ij}^{(w)}\left(x_i-x_j\right)\left(x_i-x_j\right)^T \tag{3}$$

$$W_{ij}^{(b)}=\begin{cases}A_{ij}\left(\frac{1}{n}-\frac{1}{n_l}\right), & \text{if } c_i=c_j=l\\ \frac{1}{n}, & \text{if } c_i\neq c_j\end{cases} \tag{4}$$

$$W_{ij}^{(w)}=\begin{cases}\frac{A_{ij}}{n_l}, & \text{if } c_i=c_j=l\\ 0, & \text{if } c_i\neq c_j\end{cases} \tag{5}$$

where $S^{(b)}$ and $S^{(w)}$ denote the local between-class scatter matrix and local within-class scatter matrix, respectively, $W_{ij}^{(b)}$ and $W_{ij}^{(w)}$ denote the weight matrices of the local between-class adjacency graph and local within-class adjacency graph, respectively, c_i is the class label of the data point x_i, and $l\in\{1,2,\cdots,c\}$ is the class label. A_{ij} is the heat kernel weight. As can be seen from (1), by preserving the local geometric structure, LFDA aims to look for a transformation vector V such that the data pairs in the same class are made close and the data pairs in different classes are separated from each other. Finally, the optimal V's are the eigenvectors corresponding to the maximum eigenvalue of the generalized eigenvalue problem:

$$S^{(b)}V=\lambda S^{(w)}V \tag{6}$$

The solution can be readily computed by applying an eigen-decomposition on $\left(S^{(w)}\right)^{-1}S^{(b)}$, provided that the local within-class scatter $S^{(w)}$ is nonsingular.

Efficient scatter-difference-based LFDA algorithm

In this section, we propose a scatter-difference-based local Fisher discriminant analysis(SDLFDA) technique to overcome the matrix singularity problem for tissue classification. A scatter-difference-based local discriminant analysis scheme is introduced to produce discriminating features:

$$J(V)=V^T\left(\alpha S^{(b)}-(1-\alpha)S^{(w)}\right)V \tag{7}$$

where $\alpha \in [0,1]$ is nonnegative parameter to control the trade off between local between-class scatter $S^{(b)}$ and local within-class scatter $S^{(w)}$. By imposing $V^T V = 1$ on (7), the maximization problem of (7) is equivalent to solving the following maximization problem:

$$V_{opt} = \arg\max_{V^T V=1} V^T \left(\alpha S^{(b)} - (1-\alpha) S^{(w)}\right) V \tag{8}$$

In addition, as shown in [7], $S^{(b)}$ and $S^{(w)}$ can be easily rewritten as the following forms:

$$S^{(b)} = XL^{(b)}X^T, \; S^{(w)} = XL^{(w)}X^T. \tag{9}$$

where $L^{(b)} = D^{(b)} - W^{(b)}$, $L^{(w)} = D^{(w)} - W^{(w)}$, $D^{(b)}$ and $D^{(w)}$ are both diagonal matrices, and their entries are column sums of $W^{(b)}$ and $W^{(w)}$, respectively, $D_{ii}^{(b)} = \sum_{j=1}^{n} W_{ij}^{(b)}$ and $D_{ii}^{(w)} = \sum_{j=1}^{n} W_{ij}^{(w)}$.

By using (9), the maximization problem of (8) can be rewritten as follows:

$$\arg\max_{V^T V=1} V^T X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T V \tag{10}$$

We can use the Lagrange multipliers to transform the above objective function to include the constraint

$$G(V,\lambda) = V^T X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T V - \lambda \left(V^T V - 1\right) \tag{11}$$

The optimization is performed by setting the partial derivation of $G(V,\lambda)$ with respect to V to zero

$$\frac{\partial G(V,\lambda)}{\partial V} = 0 \Rightarrow X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T V = \lambda V \tag{12}$$

Thus, the transformation vector V can be regarded as the eigenvectors of the matrix $X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T$ associated with the largest eigenvalues. Since the matrix $X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T$ is symmetric, the obtained transformation vector V has the orthogonal columns. In addition, as can be seen from the above computation process, unlike original LFDA, our proposed SDLFDA successfully avoids the singular problem since no matrix inverse needs to be computed. For SDLFDA-based tissue classification, computing the transformation vector V in SDLFDA needs to solve the eigenvectors of the $N \times N$ matrix $X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T$, which is very expensive when the dimension N of gene expression data is very high. To reduce the computation in calculating the transformation vector V, we propose an efficient algorithm for performing SDLFDA through QR decomposition. Then we have

$$X = QR \tag{13}$$

where matrix $Q \in \mathbb{R}^{N \times t}$ has orthonormal columns, matrix $R \in \mathbb{R}^{t \times p}$ is an upper triangular matrix, t is the rank of the matrix X, and p is the reduced dimension of the matrix X.

Since the optimal transformation vector V has the orthogonal columns, it can be denoted as $V = QK$ for a certain matrix $K \in \mathbb{R}^{t \times d}$ satisfying $K^T K = I_d$. Then the optimal problem of computing V in (10) can be transformed into calculating the optimal K such that

$$\arg\max_{K^T K = I_d} \mathrm{Tr}\left(K^T \left(Q^T X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T Q\right) K\right) \tag{14}$$

where $\mathrm{Tr}(\,)$ denotes the matrix trace, I_d is the *d*-dimensional identity matrix.

In addition, since $Q^T Q = I_t$ and $X = QR$, we can obtain

$$\begin{aligned} & Q^T X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T Q \\ & = Q^T Q R \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) R^T Q^T Q \\ & = R \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) R^T \end{aligned} \tag{15}$$

Then, by using (15), the maximization problem of (14) can be rewritten as follows:

$$\arg\max_{K^T K = I_d} \mathrm{Tr}\left(K^T R \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) R^T K\right) \tag{16}$$

Thus, the optimization problem in (16) can be solved by using the above similar method of Lagrange multipliers. That is, computing the optimal K is translated into finding the eigenvectors of the matrix $R \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) R^T$ associated with the largest eigenvalues. Finally, the optimal transformation vector V can be obtained by

$$V = QK \tag{17}$$

Note that the matrix $R \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) R^T$ is of size $t \times t$, which is much smaller than the size $N \times N$ of the matrix $X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T$ since $t \ll N$. In addition, solving the eigenvectors of the matrix $R \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) R^T$ has time complexity $O(t^3)$, which is greatly lower than the time complexity $O(N^3)$ of directly solving the eigenvectors of the matrix $X \left(\alpha L^{(b)} - (1-\alpha) L^{(w)}\right) X^T$. Thus the proposed QR decomposition algorithm for SDLFDA can facilitate efficient computation on high-dimensional data.

According to the above statement, the algorithmic procedure of SDLFDA is summarized as follows:

Step1: Constructing the nearest-neighbor graph. Let G denotes a graph with n nodes. The *i*th node corresponds to the gene expression data x_i. We put an edge between node i and j if x_i is among the *k*-nearest neighbor of x_j or x_j is among the *k*-nearest neighbor of x_i, and set the weight matrix A of graph G as

$$A_{ij} = \begin{cases} e^{-\left(\|x_i - x_j\|^2 / t\right)}, & \text{if } x_i \text{ is among the } k \text{ nearest neighbor of } x_j \\ & \text{or } x_j \text{ is among the } k \text{ nerest neighbor of } x_i. \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Step2: Computing the local between-class weight matrix $W^{(b)}$ and the local within-class weight matrix $W^{(w)}$ according to (4) and (5).

Step3: Constructing the optimal objective function of SDLFDA in terms of (10), and transforming it into the eigenvector problem as in (12) via Lagrangian multiplier.

Step4: QR decomposition the data matrix X as in (17) with incomplete Cholesky decomposition.

Step5: Computing the eigenvectors associated with the largest eigenvalues of the eigen-problem $R\left(\alpha L^{(b)}-(1-\alpha)L^{(w)}\right)R^T K=\lambda K$, and obtaining vector K .

Step6: Computing the projection matrix V of SDLFDA according to (17).

Step7: Obtaining the lower-dimensional representations Y of high-dimensional gene expression data X based on

$$X \to Y = V^T X = (QK)^T QR = K^T R \tag{19}$$

Step8:Tissue classification in the lower-dimensional feature space. Now, we get lower-dimensional representations of the original high-dimensional gene expression data via (19). In the reduced feature space, those gene expression data belonging to the same class are close to one another while those gene expression data belonging to different classes are far away each other.

In summary, our proposed SDLFDA algorithm not only avoids the singularity problem by not computing matrix inverse, but also significantly reduces the computational cost on high-dimensional data via QR decomposition.

Experimental results

In this section, we investigate the performance of our proposed SDLFDA algorithm for tissue classification. The algorithm performance is compared with the PCA, LDA, LPP algorithms and the original LFDA algorithm, four of the most popular dimensionality reduction algorithms in tissue classification. We performed comparative study by repeated random splitting into training set and testing set. Each data set was partitioned randomly into a training set consisting of two-thirds of the whole data set and a testing set consisting of one-third of the whole data set.

Table 1. Classification accuracy(%) comparisons

Algorithm	ALL	GCM	SRBCT	LYMPHOMA
PCA	87.5	66.2	83.6	85.4
LDA	91.1	70.8	90.2	91.8
LPP	91.3	70.9	89.4	91.6
LFDA	95.8	75.6	97.5	97.8
SDLFDA	97.6	76.3	99.9	99.7

Table 2. Computational time(s) comparisons

Algorithm	ALL	GCM	SRBCT	LYMPHOMA
PCA	4.82	6.56	4.73	5.27
LDA	9.58	15.28	10.57	11.48
LPP	9.21	13.46	8.89	10.36
LFDA	8.76	11.95	7.48	9.52
SDLFDA	3.29	4.52	2.61	2.83

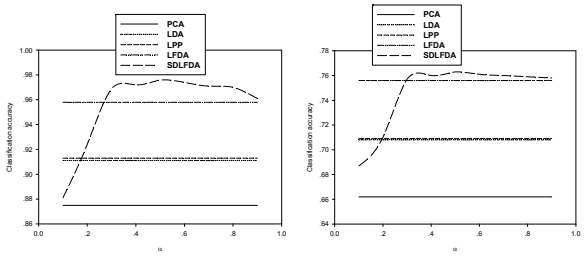

Fig.1. (1)Classification accuracy of SDLFDA with respect to α on the ALL data set; (2) Classification accuracy of SDLFDA with respect to α on the GCM data set

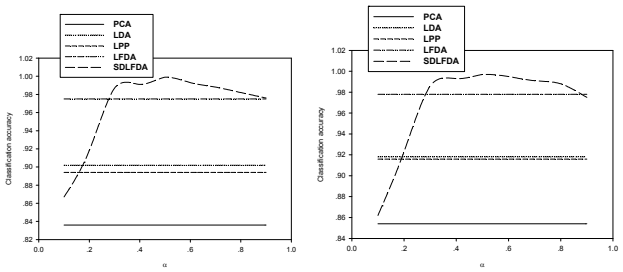

Fig.2. (1) Classification accuracy of SDLFDA with respect to α on the SRBCT data set; (2) Classification accuracy of SDLFDA with respect to α on the LYMPHOMA data set

Fig.1 to Fig.4 show the performance of SDLFDA as a function of the parameter α on the four data sets. It is easy to see that SDLFDA can achieve better performance than PCA, LDA, LPP and LFDA over a large range of α .Thus, the parameter selection is not a very crucial problem is our SDLFDA algorithm.

Conclusions

We have proposed a novel scatter-difference-based LFDA(SDLFDA) algorithm for tissue classification based on gene expression data. Experimental results demonstrate its effectiveness and efficiency.

Acknowledgements

This work is supported by NSFC (Grant No. 70701013), the National Science Foundation for Post-doctoral Scientists of China (Grant No. 2011M500035), and the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No.20110023110002).

REFERENCES

[1] Ben-Dor A., Bruhn L., Friedman N., Nachman I., Schummer M.,Yakhini Z., Tissue Classification with Gene Expression Profiles, Journal of Computational Biology, 7(2000), No.3-4,559-584

[2] Ye J., Li T., Xiong T., Janardan R., Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data, IEEE/ ACM Trans. Computational Biology and Bioinformatics, 1(2004),No.4,181-190

[3] Duda R.O.,Hart P.E.,Stork D.G., Pattern Classification(second edition),Wiley-Interscience, Hoboken, N.J., 2000

[4] Wang H., Chen S., Hu Z., Zheng W., Locality-Preserved Maximum Information Projection, IEEE Trans. Neural Networks, 19(2008),No.4, 571-585

[5] Belkin M., Niyogi P., Laplacian Eigenmaps for Dimensionality Reduction and Data Representations, Neural Comput., 15(2003), No.6, 1373-1396

[6] He X., Niyogi P., Locality Preserving Projections, Proceedings of NIPS'03,2003:585-592

***Authors**: Dr. Ziqiang Wang, Lecturer Xia Sun, Dr.Lijun Sun, School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, china; Dr.Xu Qian, College of Mechanical Electronic and Information Engineering, China University of Mining and Technology (Beijing), Beijing 100083, China. E-mail: wzqagent@126.com; wzqbox@gmail.com.*