Supervised feature extraction based on orthogonal discriminant projection

Bo Li², Chao Wang², De-Shuang Huang², *

² School of Computer Science of Technology, Wuhan University of Science of Technology, Wuhan, Hubei 430081, China
² Intelligent Computing Lab, Institute of Intelligent Machine, Chinese Academy of Science, P.O. Box 1130, Hefei, Anhui 230031, China

ABSTRACT

In this paper, a supervised feature extraction method, named orthogonal discriminant projection (ODP), is presented. As an extension of spectral mapping method, the proposed algorithm maximizes the weighted difference between the non-local scatter and the local scatter. Moreover, the weights between two nodes of a graph are adjusted according to their class information and local information. Experiments on FERET face data, Yale face data and MNIST handwriting digits data validate that ODP can offer better recognition rate than some other feature extraction methods, such as local preserving projection (LPP), unsupervised discriminant projection (UDP) and orthogonal LPP (OLPP).

1. Introduction

Recently, manifold learning based methods are becoming the most promising feature extraction approaches. Among them, one representative is Laplacian Eigenmap (LE) [1], which is also a spectral mapping method. LE aims to find a low dimensional representation that preserves the local properties of the data lying on a low dimensional manifold. Based on the training data points, a weighted graph can be constructed with edges connecting nearby points. LE can find the optimal feature subspace by solving an optimization problem which shows connections to the graph Laplacian, the Laplacian Beltrami operator on the manifold and the heat equation. However, LE is also a nonlinear dimensionality reduction technique, whose generalization ability is much weak. That is to say, the image of a test point in low dimensional space cannot be easily obtained with the projection results of the training set. This problem is also named out-of-sample problem. Linearization, kernelization, tensorization and some other tricks have been introduced to avoid the problem [2,3]. LPP [4,5] is a linear approximation to LE. Due to introducing linearization, the LPP algorithm shows its prosperities on lowering the computational cost and enhancing the robustness to noise and outliers [4]. In the original LPP, the linear transformation matrix is not under the orthogonal constraint. In order to solve the problem, D. Cai proposed an orthogonal LPP algorithm, which shows more preserving power than LPP [8]. Recently, J. Yang et al. presented a UDP algorithm [6], which can be viewed as simplified LPP on the assumption that the local density is uniform [7]. UDP characterizes the local scatter and the non-local scatter and looks for a linear projection that not only maximizes the non-local scatter but also minimizes the local scatter simultaneously. This property contributes to make UDP more intuitive and more powerful than most of up-to-date methods. However, it must be noted that UDP is a linear approximation to the manifold learning approaches without taking the class information into account. Moreover, the class information has been considered to have much to do with discriminant features. At the same time, both LPP and UDP can discover the optimized feature subspace by carrying out the generalized eigen-decomposition, which will bring another problem. It is that the local scatter matrix is not invertible when the number of features of the training data is larger than the number of the samples, which is also named small sample size (SSS) problem.

In this paper, a new supervised feature extraction method, named orthogonal discriminant projection (ODP), is proposed to overcome the problems mentioned above. Instead of the generalized eigen-decomposition, in the proposed algorithm, the optimized features can be obtained by solving an eigen-equation, where the SSS problem is naturally avoided. Unlike the original manifold learning methods, ODP follows the supervised techniques and the label information is taken to model the manifold. Furthermore, combined to labels, both local information and non-local information is introduced to define the weights of any two points, which can explore the intrinsic structure of original data and enhance the recognition ability.

The paper is organized as follows: Section 2 simply describes LPP and UDP. In Section 3, the principle of ODP is addressed. Experiments are offered in Section 4 and the paper is finished with some conclusions in Section 5.
2. LPP and UDP

Assume that there are \( n \) \( m \)-dimensional data set \([X_1, X_2, \ldots, X_n]\), it is desired to project these points into a linear subspace where the points with the same label will be clustered closer and the points belonging to different classes will be located farther. Let points \( Y_1, Y_2, \ldots, Y_n \) denote images of points \( X_1, X_2, \ldots, X_n \) in \( d \)-dimensional subspace, thus the projection can be expressed to \( Y_i = A^T X_i \), where \( A \) is a linear transformation matrix. Both LPP and UDP seek to find the linear transformation by the following steps. Firstly an adjacency graph \( G = (V, E) \) is constructed using \( k \) nearest neighbor criterion, where \( G \) denotes the graph, \( V \) is the node set and \( E \) is the edge set. Then an adjacency matrix \( H \) is defined, whose elements are given below:

\[
H_{ij} = \begin{cases} 
1 & \text{If both } X_i \text{ and } X_j \text{ are } k \text{ nearest neighbors each other} \\
0 & \text{otherwise} 
\end{cases}
\]

(1)

Sometimes, the elements of \( H \) can also be defined by a \( \varepsilon \)-ball criterion. In this study, the \( k \) nearest neighbor criterion is adopted.

Due to introducing the adjacency matrix \( H \), the local scatter can be expressed to:

\[
J_L(A) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} H_{ij} (Y_i - Y_j)^2 \\
= \frac{1}{2} \sum_{i=1}^{n} A^T \sum_{j=1}^{n} H_{ij} (X_i - X_j)(X_i - X_j)^T A = A^T S_L A
\]

(2)

where

\[
S_L = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} H_{ij} (X_i - X_j)(X_i - X_j)^T = \frac{1}{n} L X L^T,
\]

\( L \) is the Laplacian matrix with definition of \( L = D - H \) and \( D \) is a diagonal matrix, i.e. \( D_{ii} = \sum H_{ii} \).

After characterizing the local scatter, the non-linear scatter can be characterized by the following expression,

\[
J_N(A) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (1 - H_{ij}) (Y_i - Y_j)^2 \\
= \frac{1}{2} \sum_{i=1}^{n} A^T \sum_{j=1}^{n} (Y_i - Y_j)^2 \\
- \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} H_{ij} (Y_i - Y_j)^2 \\
= A^T S_L A - A^T S_N A
\]

(3)

where

\[
S_N = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (1 - H_{ij}) (X_i - X_j)(X_i - X_j)^T
\]

and

\[
S_T = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)(X_i - X_j)^T
\]

LPP is derived by the direct linear approximation of LE, which seeks an optimal linear subspace to minimize the following constrained objective function,

\[
A_{LPP} = \arg \min_{A} A^T X L X^T A
\]

(4)

Then the transformation matrix \( A \) that minimizes the objective function is obtained by solving the following generalized eigenvalue problem,

\[
X L X^T A = \lambda X D X^T A
\]

(5)

However, UDP aims to find a transformation matrix \( A \) which can maximize \( J_N(A) \) and minimize \( J_L(A) \) simultaneously, that is to say, the objective function of UDP should have the following form,

\[
A_{UDP} = \max \frac{J_N(A)}{J_L(A)} = \max \frac{A^T S_N A}{A^T S_L A}
\]

(6)

So it can be found that \( A \) is composed of eigenvectors associated with \( d \) top eigenvalues of the following generalized eigen-equation,

\[
S_N A = \lambda S_L A
\]

(7)

3. Orthogonal discriminant projection

As discussed in Section 1, both LPP and UDP ignore the class information. Moreover, LPP not only pays attention to local information, but also concentrates on non-local information. As a simplified version of LPP [7], UDP also takes advantage of both local information and non-local information. However, UDP, along with LPP, is often encountered the SSS problem. Some techniques are introduced to overcome the problem at the cost of discarding some useful information [9–11]. In order to avoid those limitations mentioned above, an orthogonal discriminant projection method is put forward. In the proposed algorithm, the weights between any two nodes are defined based on their local information and label information. Then corresponding local scatter and non-local scatter can be obtained. In order to avoid the SSS problem, a small trick is adopted in the proposed ODP. It can be found that the objective function of UDP is a ratio of the non-local scatter to the local scatter, which often results in the SSS problem when the sample number is less than the dimensions of the samples. Under such circumstance, the local scatter \( S_L \) will be not invertible [6]. So the optimal solutions cannot be obtained by solving the eigen-equation \( S_L^{-1} S_N A = \lambda A \) in UDP. Thus in the proposed algorithm, we change the form of a ratio into the form of a difference between the non-local scatter and the local scatter, i.e. \( S_N - S_L \), where the local scatter needs not positive-definite [12]. Moreover, both the objective function \( \max(S_L^{-1} S_N) \) and the objective function \( \max(S_N - S_L) \) have the same motivation that it is to maximize \( S_N \) and minimize \( S_L \) simultaneously. At the same time, in the proposed algorithm, we change the form to weighted difference, thus the contributions of \( S_N \) and \( S_L \) to the objective

Fig. 1. Typical plot of \( W_0 \) as a function of \( d^2(X_i, X_j)/\beta \).
function can be adjusted with a parameter. In the following Subsections, the definitions of the weight function and linear feature extraction method are offered, respectively.

3.1. The weight between two points

In the original LE or LPP, the weight between two nodes is defined to be a heat kernel or simply either 1 or 0, which cannot reflect the class information. In the proposed algorithm, the weight between two points is defined based on their local information and class information. The definition is stated below in details.

\[
W_{ij} = \begin{cases} 
\exp \left( -\frac{d^2(X_i, X_j)}{\beta} \right) & \text{if both } X_i \text{ and } X_j \text{ are } k \text{ nearest neighbors each other and have the same label;} \\
\exp \left( -\frac{d^2(X_i, X_j)}{\beta} \right) \left( 1 - \exp \left( -\frac{d^2(X_i, X_j)}{\beta} \right) \right) & \text{if both } X_i \text{ and } X_j \text{ are } k \text{ nearest neighbors each other and have different labels;} \\
0 & \text{otherwise}
\end{cases}
\]

where \(d(X_i, X_j)\) denotes the geodesic distance between points \(X_i\) and \(X_j\), \(\beta\) is a parameter which is used as a regulator.

Shown in Fig. 1 is the typical plot of \(W_{ij}\) as a function of \(d^2(X_i, X_j)/\beta\). In Fig. 1, S1 denotes the case that both \(X_i\) and \(X_j\) are \(k\) nearest neighbors each other and \(X_i, X_j\) have the same label; S2 denotes the case that both \(X_i\) and \(X_j\) are \(k\) nearest neighbors each other and \(X_i, X_j\) have different labels and S3 denotes the other cases. The weight \(W_{ij}\) displays the discriminant similarity between \(X_i\) and \(X_j\). The similarity integrates the local neighborhood structure and the class information. From Fig. 1, the properties of the weight function can be summarized as follows:

**Property 1.** When the distance is equal, the inter-class weight is larger than the intra-class weight, which reflects the class information of data points. That is to say, two points with larger weight will be more possibility to have the same label. On the contrary, two data with different labels will have small weight. This benefits to classification;

**Property 2.** Both the intra-class weight and the inter-class weight are located between value 0 and value 1. Compared to the original weighted function \(H, j_{ij}(A)\) will be smaller by adopting the weighted function \(W\), which results in the goal of locality preserving;

**Property 3.** The weighted function heavily depends on the regulator \(\beta\) instead of the geodesic distance \(d(X_i, X_j)\). For any data set, the geodesic distances between points can be computed with some formulas. However, the optimal parameter \(\beta\) for classification can only achieved by adjusting. For a fixed \(d(X_i, X_j)\), when the regulator \(\beta\) is small, \(d^2(X_i, X_j)/\beta\) will be large, thus the weighted function \(W\) will be small. With the increasing of \(\beta\), the weighted function \(W\) will increase accordingly. When \(\beta\) is very large even infinite, the weighted function \(W\) will attain the value of 1 or 0 under different circumstances, respectively.

**Table 1**

Performance comparison by using LPP, UDP, OLPP and ODP on FERET face.

<table>
<thead>
<tr>
<th>Methods</th>
<th>LPP</th>
<th>UDP</th>
<th>OLPP</th>
<th>ODP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate (%)</td>
<td>78.4</td>
<td>81.0</td>
<td>81.4</td>
<td>82.6</td>
</tr>
<tr>
<td>Dimensions</td>
<td>92</td>
<td>108</td>
<td>110</td>
<td>100</td>
</tr>
</tbody>
</table>

Due to those properties, the weight function \(W\) can be introduced to improve the discriminant ability and to preserve the local neighborhood structure of the original data.

Thus the local scatter \(J_L = A^T S_L A\) and non-local scatter \(J_N = A^T S_N A\) can be defined based on Eqs. (2) and (3) by only substituting \(W_{ij}\) for \(H_{ij}\).

3.2. Linear feature extraction

After the local scatter and the non-local scatter have been constructed, an optimization objective function can be devised to maximize the difference between the non-local scatter and the local scatter. That is:

\[
J'(A) = J_N(A) - J_L(A) = A^T (S_N - S_L) A = A^T (S_T - 2S_L) A
\]

We can also get the weight difference by adjusting the contributions of \(S_T\) and \(S_L\) with the constraint of \(A^T A = I\). So the objective function can be expressed as follows:

\[
\arg\max_{\beta \in \mathbb{R}} J'(A) = \arg\max_{A \in \mathbb{R}} A^T ((1 - \pi)S_T - 2S_L) A
\]

where \(\pi\) is an adjustable parameter.

Then we can easily find that \(A\) consists of the eigenvectors associated with \(d\) top eigenvalues of the following eigen-equation,

\[
((1 - \pi)S_T - 2S_L) A = \lambda A
\]

Thus the optimal linear features \(Y\) can be obtained by the following linear transformation,

\[
Y = A^T X.
\]

4. Experiments

In this Section, FERET face data set, Yale face data and MNIST handwriting digits data are applied to evaluate the performance of the proposed ODP algorithm, which is compared with those of LPP, UDP and OLPP.
4.1. Experiment on FERET face database

The subset of FERET face database contains 100 individuals and seven images for each person. It is composed of images whose names are marked with two-character strings: “bd”, “bj”, “bf”, “bc”, “ba”, “bk”. This subset involves two facial expression images, two left pose images, two right pose images and an illumination image. All the images in subset are to be the size of 80 × 80. The first four images are selected as training samples and the rest three images as test set. Shown in Fig. 2 are sample images of one person in the experiments. When constructing the $k$ nearest neighborhood graph, $k$ is set to be 4. After features have been extracted by performing LPP, UDP, OLPP and ODP, the nearest neighbor classifier is adopted to predict the labels of the test data.

Shown in Table 1 is the recognition rate comparison on FERET data set by performing LPP, UDP, OLPP and ODP respectively. The best recognition rates are achieved at 92, 108, 110 and 100 dimensions for LPP, UDP, OLPP and ODP, respectively. From the experimental results, it can be found that ODP outperforms the other three methods.

To find how the weight parameter $\alpha$ affects the recognition performance, we change $\alpha$ from 0.0 to 1.0 with step 0.1. Fig. 3 displays the recognition rates with varied parameter $\alpha$ by carrying out ODP. From Fig. 3, it can be found that UDP obtains the best recognition rate when $\alpha = 0.8$.

The above experiments are conducted with a fixed $k$. In the following experiment, we varied $k$ from 2 to 6 with $\alpha = 0.8$. The performance with different $k$ is displayed in Fig. 4. It can be found that when $k$ is set to 4, the best recognition rate can be achieved.

4.2. Experiment on Yale face database

The Yale face database was constructed at the Yale Center for Computation Vision and Control. There are 165 images about 15 individuals in YALE face data sets, where each person has 11 images. The images demonstrate variations in lighting condition (left-right, center-light, right-light), facial expression (normal, happy, sad, sleepy, surprised and wink), and with or without glasses. Each image is cropped to be the size of 32 × 32. Fig. 5 shows the cropped images of one person in Yale face database.

<table>
<thead>
<tr>
<th>Methods</th>
<th>LPP</th>
<th>UDP</th>
<th>OLPP</th>
<th>ODP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate (%)</td>
<td>88.5</td>
<td>90.4</td>
<td>91.6</td>
<td>93.5</td>
</tr>
<tr>
<td>Dimensions</td>
<td>28</td>
<td>22</td>
<td>26</td>
<td>20</td>
</tr>
</tbody>
</table>
We select the first six images as training set and the rest five images as test set for each class. The parameter $k$ can be set to 5 when constructing the neighborhood graph. LPP, UDP, OLPP and ODP are carried out to extract the features. At last we use the nearest neighbor method for classification. Shown in Table 2 are the optimal recognition rates at corresponding dimensions for different feature extraction methods. It is found that the proposed method outperforms the other techniques. Fig. 6 discloses the impacts of $a$ on recognition rate. When $a$ equals to 0.7, the optimal recognition rate can be obtained. Shown in Fig. 7 is performance curve with varied $b$. In Fig. 7, it can be found that after $b$ attains 200, $b$ shows little impact on the recognition rate.

4.3. Experiment on MNIST handwriting digits data

The MNIST database of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples. In this experiment, for each class, we select 120 and 50 images as training and test samples, respectively. Each image is transformed to a vector with 784 dimensions. Fig. 8 illustrates some samples of zero from MNIST handwriting digits data database. Then $k$ nearest neighbor criterion is adopted to construct the adjacency graph and $k$ is set to be 10. At last, the nearest neighbor classifier is also taken to classify the test data.

From Table 3, it can be found that ODP obtains the best recognition rate compared with LPP, UDP and DLPP. Moreover, parameter $x$ also shows much impact on recognition rate. When $x$ equals to 0.6, ODP achieves the best performance, which is displayed in Fig. 9. At last, we set $x$ to 0.6 and vary $b$ from 50 to 1000. Shown in Fig. 10 is the performance trend with the increasing of $b$. The parameter $b$ shows its impact on the recognition rate when $b$ is small. However, when $b$ equals to or is larger than 300, the recognition rate almost keeps unchanged.

We also test the impact of $k$ on the performance, which can be found in Fig. 11. We varied $k$ from 8 to 16 with step 1 after fixing $x = 0.6$ and $b = 500$. Then it can be found that when $k$ equals to 10, the recognition rate gains the best value.

<table>
<thead>
<tr>
<th>Methods</th>
<th>LPP</th>
<th>UDP</th>
<th>OLPP</th>
<th>ODP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate (%)</td>
<td>91.3</td>
<td>93.5</td>
<td>95.8</td>
<td>97.4</td>
</tr>
<tr>
<td>Dimensions</td>
<td>20</td>
<td>18</td>
<td>22</td>
<td>18</td>
</tr>
</tbody>
</table>

Fig. 9. Recognition rates with varied $x$ on MNIST handwriting digits.

Fig. 10. Recognition rates with varied $b$ on MNIST handwriting digits.

Fig. 8. Some sample images of zero from MNIST handwriting digits.
Fig. 11. Recognition rates with varied k on MNIST handwriting digits.

5. Conclusion

In this paper, an improved version of graph spectral mapping, namely ODP, is proposed for classification. The proposed algorithm uses the local information and the non-local information as well as the class information of the data to model the manifold learning. So the proposed algorithm becomes more suitable for the tasks of classification. This result is validated either from the theoretical analysis or from experiments on real-world data set.

Acknowledgment

This work was supported by the grants of the National Science Foundation of China, nos. 60772130, 60705007 and 60805021, the grant from the National Basic Research Program of China (973 Program), no. 2007CB31100, the grants from the National High Technology Research and Development Program of China (863 Program), nos. 2007AA01Z167 and 2006AA02Z309, the grant of the Guide Project of Innovative Base of Chinese Academy of Sciences (CAS), no. KSCX1-YW-R-30, and the grant of Oversea Outstanding Scholars Fund of CAS, no. 2005-1-18.

References


De-Shuang Huang (S’98) received the B.Sc., M.Sc. and Ph.D. degrees all in Electronic Engineering from Institute of Electronic Engineering, Hefei, China, National Defense University of Science and Technology, Changsha, China and Xidian University, Xi’an, China, in 1986, 1989 and 1993, respectively. During 1993–1997 period he was a postdoctoral student respectively in Beijing Institute of Technology and in National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In September, 2000, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences as the recipient of “Hundred Talents Program of CAS”. From September 2000 to March 2001, he worked as research associate in Hong Kong Polytechnic University. From April 2002 to June 2003, he worked as research fellow in City University of Hong Kong. From August to September 2003, he visited the George Washington University as visiting professor, Washington DC, USA. From October to December 2003, he worked as research fellow in Hong Kong Polytechnic University. From July to December 2004, he worked as the University fellow in Hong Kong Baptist University. Dr. Huang is currently a senior member of the IEEE.

Chao Wang was born in AnHui province, China, in 1981. He received Bachelor degree on Communication Engineering in AnHui University. He is now in pursuit for M.Sc. degree in Pattern Recognition and Intelligent System in University of Science and Technology of China. His research interests include Semi-Supervised Learning, manifold learning and image processing.

Bo Li was born in Hubei province, China, in 1975. He received M.Sc. degree in Mechanic and Electronic Engineering in 2003, from Wuhan University of Technology. He is now in pursuit for Ph.D. degree in Pattern Recognition and Intelligent System in University of Science and Technology of China. His research interests include pattern recognition, manifold learning and image processing.