IJAMML 11:1-2 (2019) 1-12 September and December 2019
ISSN: 2394-2258
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/ijamml_7100122097

A SIMPLE AND EFFECTIVE DISCRIMINANT FEATURE CONSTRUCTION METHOD FOR IMAGE RECOGNITION

Huan Xu[a], Wenjing Yan[b], Gang Zhu[a], Yinghao Deng[a] and Ruonan Zhang[a]

[a]College of Computer Science and Engineering, Anhui University of Science & Technology, Huainan, Anhui, 232001, P. R. China

[b]Information Management Department, School of Computer and Information Engineering, Beijing, 100037, P. R. China

Abstract

Raw data from real-world applications are usually high-dimensional data with noise and redundancy information. How to obtain low-dimensional features of raw data is crucial to pattern recognition. In this paper, we propose a simple and effective discriminant feature construction (DFC) method, which exploits class labels and nonlinear similarity information to directly construct discriminant similarity features of training samples. The features keep the discriminating power of class labels and similarity information as much as possible. Extensive experiments on several real-world image datasets have demonstrated the superior performance of our proposed method.

Keywords: feature construction, discriminant similarity feature, image recognition.

*Corresponding author.

E-mail address: hxu@aust.edu.cn (Huan Xu).

2010 Mathematics Subject Classification: 68W25.
Submitted by Jianqiang Gao.
Received August 24, 2019.

1. Introduction

Feature learning is a prevalent research field in pattern recognition and machine learning. Typical feature learning algorithms include principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2, 3], locality preserving projections (LPP) [4], etc. The algorithms focus on finding a set of projection directions using a certain criterion to extract low-dimensional features from high-dimensional data.

PCA is an unsupervised linear method for seeking a subspace where the projections of data possess maximum variance. Generally, linear PCA may have difficulty catering complex nonlinear data in many real-world applications, and thus kernel PCA (KPCA) [5] was proposed to solve the problem. KPCA firstly maps original data into a higher (even infinite) dimensional kernel space, and then implements linear PCA in the kernel space. LDA is a supervised feature learning method, which can obtain a discriminant subspace where within-class scatter of data is minimized and at the same time between-class scatter is maximized. In [6], kernel discriminant analysis (KDA) was proposed to extract nonlinear low-dimensional face features. In addition, LPP preserves local information hidden in data. Likewise, kernel-based LPP algorithms [7] have also been presented for better capturing nonlinear relationships among data.

In this paper, we propose a simple and effective discriminant feature construction (DFC) method. Different from traditional feature learning methods, DFC directly employs class labels and nonlinear similarity information to construct discriminant similarity features of training samples. Since the discriminating power of supervised and similarity information is kept as much as possible, our discriminant similarity features are well discriminative. Due to the lack of out-of-sample class label information, discriminant similarity features of out-of-sample data are difficultly constructed. For the issue, we employ an existing method, i.e., kernel propagation strategy (KPS) [8], and the key idea of KPS is that discriminant similarity features of an out-of-sample data should be similar to ones of its nearest neighbour data in kernel spaces. By means of KPS, we can obtain discriminant similarity features of out-of-sample data. Moreover, extensive experiments have been implemented on several real-world image datasets, and the promising experimental results have demonstrated the effectiveness of our proposed method.

The rest of the paper is organized as follows. In Section 2, we briefly review linear discriminant analysis (LDA). We give a detailed description of DFC in Section 3. In Section 4, extensive experiments are designed to evaluate our algorithm. We conclude the paper in Section 5.

2. Linear Discriminant Analysis

LDA seeks to find a linear subspace such that the projections of within-class samples become more compress and the projections of between-class samples become far apart. Suppose $X = [x_1, x_2, \dots, x_n] \in R^{d \times n}$ is a training sample set with c different classes, where d denotes the dimensionality of samples and n is the number of samples. The within-class and between-class scatter matrices S_w and S_b are defined by

$$S_w = \frac{1}{n} \sum_{i=1}^{c} \sum_{x_j \in C_i} (x_j - m_i)(x_j - m_i)^T, \tag{1}$$

$$S_b = \frac{1}{n} \sum_{i=1}^{c} q_i (m_i - m_0)(m_i - m_0)^T, \tag{2}$$

where C_i is a set of samples belong to the i-th class, m_i denotes the sample mean of C_i, q_i denotes the number of samples in the i-th class, and m_0 represents the total mean of X.

Conventional LDA aims to find an optimal projection matrix $P = [p_1, p_2, \ldots, p_{c-1}] \in R^{d\times(c-1)}$, and the optimal projection matrix can be obtained by maximizing the following objective function:

$$J(P) = \frac{Tr(P^T S_b P)}{Tr(P^T S_w P)}, \tag{3}$$

where $Tr(\cdot)$ denotes the trace of a matrix.

If S_w is nonsingular, the optimal projection matrix P can be gained by computing the eigenvectors of $S_w^{-1} S_b$ corresponding to the $(c-1)$ largest eigenvalues. But in many applications, the algorithm has to be confronted with the difficult problem that S_w is singular, because the number of training samples is often much lower than the dimensionality of samples. One popular method of overcoming the problem is to utilize PCA as a preprocessing step to reduce the dimensionality of samples.

3. The Discriminant Feature Construction Method

For the training samples X, we can directly construct a discriminant similarity feature $y_i = [y_{i1}, y_{i2}, \ldots, y_{in}]^T \in R^{n\times1}$ for any x_i $(i = 1, 2, \ldots, n)$. In DFC, we utilize Gaussian scheme to compute similarities among samples, and thus y_{ij} can be obtained by

$$y_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}, & \text{if } \text{label}(x_i) = \text{label}(x_j), \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where y_{ij} denotes the j-th element of y_i, and $\text{label}(x_i)$(or $\text{label}(x_j)$) is the class label of x_i (or x_i).

We use the first three samples of X, i.e., x_1, x_2, and x_3 to intuitively exhibit how to construct their discriminant similarity features. Suppose x_1 and x_2 belong to the first class, and x_3 comes from the second class, and discriminant similarity features of the three samples are as following:

$$y_1 = \begin{bmatrix} 1 \\ e^{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}} \\ 0 \end{bmatrix}, \; y_2 = \begin{bmatrix} e^{-\frac{\|x_2 - x_1\|^2}{2\sigma^2}} \\ 1 \\ 0 \end{bmatrix}, \; y_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Our discriminant similarity features directly employ class labels and nonlinear similarity information, and the discriminating power of supervised and similarity information can be kept as much as possible. Therefore, then on linear discriminant similarity features have well discriminating power.

Due to the lack of out-of-sample class label information, discriminant similarity features of out-of-sample data are difficultly computed. For the issue, we employ an existing method, KPS based on sample distribution similar principle. More concretely, suppose the testing samples are $\widetilde{X} = \{\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_N\} \in R^{d \times N}$, and $\widetilde{Y} = [\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_N] \in R^{n \times N}$ denotes the corresponding discriminant similarity features. For the sample set $\widehat{X} = [X, \tilde{X}] \in R^{d \times (n+N)}$, we firstly construct a neighbour weight graph G by the following Gaussian weighting scheme:

$$w_{ts} = \begin{cases} e^{-\frac{\|\hat{x}_t - \hat{x}_s\|^2}{2\sigma^2}}, & \text{if } \hat{x}_t \in Nei_k(\hat{x}_s) \text{ or } \hat{x}_s \in Nei_k(\hat{x}_t), \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where w_{ts} is the (t, s) element in the weight matrix W of G, $\hat{x}_t$ and $\hat{x}_s$ are respectively, the t-th and s-th samples of $\widehat{X}$, and $Nei_k(\hat{x}_s)$(or $Nei_k(\hat{x}_t)$) is a sample set whose samples belong to the k nearest neighbours of $\hat{x}_s$ (or $\hat{x}_t$).

The main idea of KPS is that the discriminant similarity feature of $\hat{x}_t$ is similar to discriminant similarity features of the k nearest neighbours of $\hat{x}_t$, and the similarities are quantified by the neighbour weights of G. Note that the k nearest neighbours contain not only training samples but also testing samples. Based on the idea, we give an iterative propagation criterion of $\widetilde{Y}$ at time $r+1$:

$$\widetilde{Y}(r+1) = [Y, \widetilde{Y}(r)]\begin{bmatrix} L_{lu} \\ L_{uu} \end{bmatrix} = \widetilde{Y}(r)L_{uu} + YL_{lu}, \tag{6}$$

where $Y = [y_1, y_2, \ldots, y_n] \in R^{n\times n}$ is the discriminant similarity features of X, and $\widetilde{Y}(r)$ is the r-th iterative result of $\widetilde{Y}$. $L = D^{-1}W$ is a stochastic matrix, in which D is a diagonal matrix whose elements on diagonal are the row sum of W. Then, we rewrite L as a partitioned matrix $L = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix}$, where L_{ll} of size $n \times n$ corresponds to X, and L_{uu} of size $N \times N$ corresponds to $\widetilde{X}$. Via the iteration equation (6), we can obtain:

$$\widetilde{Y}(r) = \widetilde{Y}(0)(L_{uu})^r + \sum_{\nu=1}^{r-1} Y L_{lu}(L_{uu})^\nu, \tag{7}$$

where $\widetilde{Y}(0)$ is the initial matrix of $\widetilde{Y}$, $(L_{uu})^r$(or $(L_{uu})^\nu$) denotes the r-th (or υ-th) power of L_{uu}. Due to $L = D^{-1}W$, every element of the matrix L is a non-negative real number that is less than one. By the limit operation for $\widetilde{Y}(r)$ of Equation (7), we can easily observe that the iteration process converges to

$$\begin{aligned}\widetilde{Y} &= \lim_{r\to\infty} \widetilde{Y}(r)\\ &= \widetilde{Y}(0)\left(\lim_{r\to\infty}(L_{uu})^{r}\right) + YL_{lu}\left(\lim_{r\to\infty}\sum_{\nu=1}^{r-1}(L_{uu})^{\nu}\right)\\ &= YL_{lu}(I - L_{uu})^{-1}, \end{aligned} \tag{8}$$

where I is the $N \times N$ identity matrix. Although KPS is an iterative process, we obtain an analytical convergent solution by means of the limit operation, i.e., $\widetilde{Y}$ can be directly obtained by Equation (8). In addition, the discriminant similarity features $\widetilde{Y}$ obtained by KPS not only include the integral sample distribution information from the graph G, but also further inherit supervised information from Y.

In order to clearly exhibit DFC, its brief steps are given as follows:

Step 1. Construct the discriminant similarity features Y of training samples by Equation (4).

Step 2. Compute the discriminant similarity features $\widetilde{Y}$ of out-of-sample data by Equation (8).

Step 3. A classifier is used for final image recognition tasks.

4. Experiments

In this section, DFC is compared to two typical supervised feature extraction algorithms, i.e., LDA and KDA. In addition, all the algorithms are implemented in three real-world image datasets, i.e., COIL20 object dataset [9], Sheffield face dataset [10], and ORL face dataset [11]. To avoid the singular problem of LDA and reduce the time complexity of all the algorithms, we employ PCA to reduce the dimensionality of every image, where 99 percent energy is kept. Since the neighbour weight graph of DFC is constructed by the Gaussian weighting scheme [12], DFC unavoidably contains two parameters, i.e., the neighbour parameter k and the Gaussian parameter σ. In the experiments, the parameter k is empirically set to 10 to avoid exhaustive search, and the Gaussian parameter σ is selected from the range of {0.2*r*, 0.4*r*, 0.6*r*, 0.8*r*, *r*, 2*r*, 4*r*, 6*r*, 8*r*, 10*r*}, and *r* is set as the averaged Euclidean distance from each sample to its ten nearest neighbours. Moreover, KDA is based on an empirical kernel method with Gaussian kernel function, and the Gaussian parameter σ is assigned according to the above way. For all the algorithms, the nearest neighbour classifier with Euclidean distance metric is taken in final recognition tasks, and the best recognition rates of the two contrastive algorithms will be reported on all possible dimensions.

On the COIL20 and Sheffield datasets, we randomly choose $q(q = 3, 4, 5)$ samples from each class for training and the rest are treated as the testing samples. To guarantee the randomness of the experiments, every random experiment is repeatedly performed 10 times, and Table 1 and Table 2 exhibit the average recognition rates. On ORL face dataset, the first q (q = 2, 3, 4, 5, 6) samples from each class are treated as the training samples, and the rest are used for testing, and we tabulate the recognition rates of all the algorithms in Table 3.

Table 1. The average recognition rates (%) on the COIL-20 object dataset

	3Train	**4Train**	**5Train**
DFC	94.04 ± 1.72	96.10 ± 3.02	97.37 ± 1.73
LDA	72.95 ± 2.3	77.63 ± 2.86	78.69 ± 2.02
KDA	79.23 ± 1.99	83.59 ± 2.40	85.42 ± 1.61

A ± *B*: *A* denotes the average recognition rate and *B* denotes the corresponding standard deviation.

Table 2. The average recognition rates (%) on the Sheffield face dataset

	3Train	**4Train**	**5Train**
DFC	90.54 ± 3.36	94.73 ± 2.26	95.09 ± 1.37
LDA	59.44 ± 2.54	67.33 ± 3.68	69.92 ± 4.65
KDA	79.83 ± 2.75	87.54 ± 2.34	91.75 ± 1.67

A ± *B*: *A* denotes the average recognition rate and *B* denotes the corresponding standard deviation.

Table 3. The recognition rates (%) on the ORL face dataset

	2Train	**3Train**	**4Train**	**5Train**	**6Train**
DFC	89.38	91.43	94.17	96.50	97.50
LDA	83.13	85.36	89.58	91.00	91.88
KDA	89.06	91.07	95.00	96.00	96.88

From all the experimental results, it can be seen that DFC obviously outperforms the other algorithms on recognition rates. We employ discriminant and nonlinear similarity information to directly construct discriminant similarity features of training samples, and thus these features are more discriminative for recognition tasks, which are underlying reasons why DFC possesses the highest recognition rates. For LDA and KDA, the class label information is also utilized. However, unlike our method, the two methods do not directly take supervised information into account in the feature extraction process of training samples and out-of-sample data. Compared to LDA that only exploits label information, KDA further considers nonlinear relationships of samples and always has higher recognition rates, which reveals the importance of nonlinear information in image recognition to some extent. In addition, KDA even outperforms DFC when each class has four training samples on the ORL datasets.

Compared to the other algorithms, DFC is relatively robust for the sample randomness, which can be clearly seen from the standard deviation of Tables 1 and 2. For all the datasets, the superiority of DFC is most outstanding on the COIL20 object dataset. By analyzing all the experimental results, we also make two interesting observations. The smaller the ratios of training samples in all the samples are, the more obvious the advantages of our algorithm in recognition rates are. In addition, with the increase of training samples, the recognition rate differences between our algorithm and the other algorithms have a tendency to decrease, but our algorithm always has excellent recognition rates. In summary, the experimental results in the three real-world image datasets give a reasonable observation that DFC is an effective and relatively robust method for image recognition tasks.

5. Conclusion

Different from conventional feature learning methods, our DFC directly constructs discriminant similarity features of training samples by means of class labels and similarity information among samples, and the discriminating power of discriminant and similarity information is kept as much as possible. Although the discriminant similarity features process well discriminating power, we difficultly construct discriminant similarity features of out-of-sample data due to the lack of out-of sample class label information. With the help of KPS, this issue is solved and discriminant similarity features of out-of-sample data are obtained, which makes DFC can be utilized in image recognition tasks. To evaluate our DFC, we design extensive experiments on the three real-world image datasets, which reveal that our DFC is an effective and relatively robust method for image recognition.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 61806006), the Natural Science Research Project of Colleges and Universities in Anhui Province (Grant No. KJ2018A0083), and Youth Science Research Project of Anhui University of Science & Technology (Grant No. QN2017208).

References

[1] W. Sun and Q. Du, Graph-regularized fast and robust principal component analysis for hyperspectral band selection, IEEE Transactions on Geoscience and Remote Sensing 56(6) (2018), 3185-3195.

DOI: https://doi.org/10.1109/TGRS.2018.2794443

[2] P. Deng, H. Wang, T. Li, S.-J. Horng and X. Zhu, Linear discriminant analysis guided by unsupervised ensemble learning, Information Sciences 480 (2019), 211-221.

DOI: https://doi.org/10.1016/j.ins.2018.12.036

[3] J. Wen, X. Fang, J. Cui, L. Fei, K. Yan, Y. Chen and Y. Xu, Robust sparse linear discriminant analysis, IEEE Transactions on Circuits and Systems for Video Technology 29(2) (2019), 390-403.

DOI: https://doi.org/10.1109/TCSVT.2018.2799214

[4] G. F. Lu, Y. Wang, J. Zou and Z. Wang, Matrix exponential based discriminant locality preserving projections for feature extraction, Neural Networks 97 (2018), 127-136.

DOI: https://doi.org/10.1016/j.neunet.2017.09.014

[5] B. Chen, J. Yang, B. Jeon and X. Zhang, Kernel quaternion principal component analysis and its application in RGB-D object recognition, Neurocomputing 266 (2017), 293-303.

DOI: https://doi.org/10.1016/j.neucom.2017.05.047

[6] H. K. Min, Y. Hou, S. Park and I. Song, A computationally efficient scheme for feature extraction with kernel discriminant analysis, Pattern Recognition 50 (2016), 45-55.

DOI: https://doi.org/10.1016/j.patcog.2015.08.021

[7] G. Shikkenawis and S. K. Mitra, On some variants of locality preserving projection, Neurocomputing 173(Part 2) (2016), 196-211.

DOI: https://doi.org/10.1016/j.neucom.2015.01.100

[8] S. Su, H. Ge and Y. H. Yuan, Kernel propagation strategy: A novel out-of-sample propagation projection for subspace learning, Journal of Visual Communication and Image Representation 36 (2016), 69-79.

DOI: https://doi.org/10.1016/j.jvcir.2016.01.007

[9] A. A. S. Najafabadi and F. T. Azar, Removing redundancy data with preserving the structure and visuality in a database, Signal, Image and Video Processing 13(4) (2019), 745-752.

DOI: https://doi.org/10.1007/s11760-018-1404-8

[10] A. Khalili Mobarakeh, J. A. Cabrera Carrillo and J. J. Castillo Aguilar, Robust face recognition based on a new supervised kernel subspace learning method, Sensors 19(7) (2019); Article 1643.

DOI: https://doi.org/10.3390/s19071643

[11] X. Song, Z. H. Feng, G. Hu, J. Kittler and X.-J. Wu, Dictionary integration using 3D morphable face models for pose-invariant collaborative-representation-based classification, IEEE Transactions on Information Forensics and Security 13(11) (2018), 2734-2745.

DOI: https://doi.org/10.1109/TIFS.2018.2833052

[12] S. Su, H. Ge and Y. Tong, Multi-graph embedding discriminative correlation feature learning for image recognition, Signal Processing: Image Communication 60 (2018), 173-182.

DOI: https://doi.org/10.1016/j.image.2017.10.005

■