Supervised Kernel Construction for Unsupervised PCA on Face Recognition

Yang Zhao, Wen-Sheng Chen*, Binbin Pan, and Bo Chen

College of Mathematics and Computational Science, Shenzhen University
Shenzhen Key Laboratory of Media Security
Shenzhan, 518060, China
chenws@szu.edu.cn

Abstract. This paper aims to establish a novel framework for high-performance Mercer kernel construction. Based on a given kernel matrix incorporated the class label information, a nonlinear mapping is firstly generated and well-defined on the training samples. The partial data-defined mapping can be extended and well-defined on the entire pattern space by means of interpolatory technology. The analytic expression of the nonlinear mapping is then obtained. It theoretically shows that the function $K(x, y)$, created by the inner product of the nonlinear mapping, is a supervised Mercer kernel function. Our supervised kernel is successfully applied to unsupervised principal component analysis (PCA) method for face recognition. Two face databases, namely ORL and FERET databases, are selected for evaluations. Compared with KPCA with RBF kernel (RBF-PCA) method, experimental results demonstrate that KPCA with our supervised kernel (SK-PCA) has superior performance.

Keywords: Face Recognition, Supervised Mercer Kernel, Kernel PCA.

1 Introduction

Over the past decades, face recognition has become one of the most challenging technologies in the area of pattern recognition and computer vision because of variations of facial images, such as pose and illumination variations. These variations incur that the distribution of the facial data in original feature space is very complicated and usually nonlinear. So, the linear feature extraction methods, say PCA [1] and LDA [2], cannot achieve satisfactory performance. Kernel method is an effective means to tackle the nonlinear problem of face recognition [3]-[10]. The basic idea of kernel method is to find a nonlinear mapping Φ which maps the input samples into a high dimensional feature space F, and then performs a linear classifier in the kernel feature space F. However, it is very difficult to learn a nonlinear mapping and the dimensionality of F is also large and perhaps infinite. Thereby, direct computation in F is infeasible. Fortunately, the linear feature extraction methods conducted in F just need to calculate the inner

* Corresponding author.

S. Li et al. (Eds.): CCPR 2014, Part II, CCIS 484, pp. 351–359, 2014.

product $\langle \Phi(x), \Phi(y) \rangle_F$. Based on Mercer kernel theory [12], the inner product $\langle \Phi(x), \Phi(y) \rangle_F$ can be replaced with a kernel function $K(x, y)$, where x and y are two samples from the input feature space. This kernel trick allows us to execute the kernel method in kernel space F without knowing the nonlinear mapping. The kernel matrix, which is a symmetric and positive semi-definite matrix, plays an important role in kernel based machine learning. This paper will discuss how to construct a kernel matrix using the training data and further design a high-performance kernel function. It is known that the kernel matrices determined by the popular kernel functions, such as linear, polynomial and RBF kernels, do not make use of the class label information. The performances of these unsupervised kernel based learning methods will be degraded. To overcome this limitation, we previously proposed a kernel construction method using Lagrange interpolation strategy [8]. But this method is numerically unstable because of the Runge phenomenon caused by Lagrange interpolation [11].

To remedy the drawbacks of existing kernel learning algorithms, this paper proposes a new framework to design the supervised Mercer kernel with high-performance. As we know, kernel matrix associated with a certain kernel function records the similarity scores among the training samples. It is desired for high-performance kernel construction that the similarities among the intra-data have large scores, while the inter-data possess small similarities. To this end, the class label information is utilized in this paper to model a supervised kernel matrix, which is a block diagonal matrix generated using a radial basis function. Unlike Lagrange interpolation, we propose a methodology to construct some new interpolatory basis functions such that their values range in the interval $[0, 1]$. So, our interpolatory strategy can automatically eliminate the Runge phenomenon since the proposed method avoids the values of interpolatory basis functions from growing unboundedly. Based on the supervised kernel matrix and the new interpolatory strategy, this paper obtains the analytic expression of the nonlinear mapping and theoretically proves that the function $K(x, y)$, defined by the inner product of the nonlinear mapping, is a supervised Mercer kernel function. The supervised kernel (SK) is tested using kernel PCA approach for face recognition. Two publicly available face databases, namely ORL and FERET, are chosen for performance evaluations. Experimental results show that KPCA with our supervised kernel (SK-PCA) surpasses KPCA with RBF kernel (RBF-PCA).

The rest of this paper is organized as follows. Section 2 briefly introduces the related work. Section 3 gives the theoretic discussions on interpolatory basis functions and the supervised Mercer kernel constructions. Section 4 develops our SK-PCA algorithm. Experimental results are reported in section 5. Finally, section 6 draws the conclusions.

2 Related Work

This section will briefly introduce some related work such as PCA and Kernel PCA.

2.1 PCA

PCA is an unsupervised linear feature extraction and dimensionality reduction method, which finds the orthogonal linear transformation such that the mapped data along the principal component directions have the largest variances. In face recognition, PCA discovers the principal facial features which are called eigenfaces [1]. The facial images can be linearly expressed by the eigenfaces. However, PCA is a linear method which cannot uncover the underlying nonlinear structure of the facial images. Moreover, as an unsupervised method, PCA is used in a label-independent manner and thus not suitable for classification tasks.

2.2 Kernel PCA

To solve nonlinear problems, the classic PCA has been generalized to its kernel version, namely Kernel PCA (KPCA) [3,4]. In this method, the data from the input space are mapped into a high dimensional kernel space using a nonlinear mapping, where different classes of objects are supposed to be linearly separable. And then the classic PCA can be performed in the high dimensional feature space using kernel trick. However, the kernel matrices, which are computed by the commonly used kernels on the training data, are full matrices and cannot reflect the class label information. So, KPCA with the commonly used kernels, such as linear, polynomial and RBF kernels, is still an unsupervised learning approach. The accuracy of unsupervised KPCA will be affected in face recognition.

In the following sections, this paper discusses how to design a supervised Mercer kernel with high-performance.

3 The Proposed Method

This section proposes a theoretical framework on supervised Mercer kernel construction. Details are below.

3.1 Some Notations

Let d be the dimension of original feature space and C be the number of sample classes. The total original sample set $X = \bigcup_{i=1}^{C} X_i$, where the ith class $X_i = \{x_j^i\}_{j=1}^{N_i}$ which contains N_i training samples. $N(= \sum_{i=1}^{C} N_i)$ is the number of total training samples. Assume $\Phi(x) : x \in R^d \to \Phi(x) \in F$ is the kernel nonlinear mapping, where F is the mapped feature space with dimension $df(= \dim F)$. The total mapped sample set is $\Phi(X) = \bigcup_{i=1}^{C} \Phi(X_i)$, and the ith mapped class is $\Phi(X_i) = \{\Phi(x_j^i)\}_{j=1}^{N_i}$. If $K(x, y)$ is a Mercer kernel defined on $R^d \times R^d$, then there exists a nonlinear mapping Φ, such that $K(x, y) = \langle \Phi(x), \Phi(x) \rangle_F$. We denote RBF kernel $K_{RBF}(x, y)$ by $K_{RBF}(x, y) = \exp(-\frac{\|x-y\|^2}{t})$ with $t > 0$.

3.2 Basis Function Construction

We construct N interpolatory basis functions $L_j^i(x)$ as follows

$$L_j^i(x) = \frac{\omega_j^i(x)}{\sum_{J=1}^{N_I} \sum_{I=1}^{C} \omega_J^I(x)}, \tag{1}$$

where $\omega_j^i(x) = \prod_{(p,q)\neq(i,j)} \|x - x_q^p\|, q = 1, 2, \cdots, N_p, p = 1, 2, \cdots, C, x \in R^d$. It can be easily verified that $\omega_j^i(x)$ has the following property:

$$\omega_j^i(x) = \begin{cases} \prod_{(p,q)\neq(i,j)} \|x_j^i - x_q^p\|, & x = x_j^i \\ 0, & x \in X \backslash \{x_j^i\}. \end{cases}$$

Therefore, the interpolatory basis function $L_j^i(x)$ satisfies that:

$$L_j^i(x_q^p) = \begin{cases} 1, (p,q) = (i,j) \\ 0, (p,q) \neq (i,j) \end{cases}, \text{ for all } x_q^p \in X.$$

From (1), we can see that basis function $L_j^i(x)$ is a bounded function and ranges in the internal $[0, 1]$. For convenience, all basis functions are formed as a interpolatory basis vector function $L(x)$ denoted by

$$L(x) = [L_1^1(x), \cdots, L_{N_1}^1(x) | \cdots | L_1^C(x), \cdots, L_{N_C}^C(x)]^T. \tag{2}$$

3.3 Supervised Mercer Kernel Construction

Assume matrices $K_i = (k_{jk}^{(i)})_{N_i \times N_i}$, where $k_{jk}^{(i)} = K_{RBF}(x_j^i, x_k^i), i = 1, 2 \cdots C$. Let block diagonal matrix K be

$$K = \text{diag}(K_1, K_2, \cdots, K_C) \in R^{N \times N}, \tag{3}$$

then K is a symmetric and positive semi-definite matrix, which is able to serve as a kernel matrix. It can be seen from matrix K that the similarities of the intra-data are large and that of the inter-data are small. Thus, K encodes the class label information of the training data.

By performing eigenvalue decomposition on K_i, we have $K_i = \tilde{U}_i^T \Lambda_i \tilde{U}_i$, where $\tilde{U}_i$ is a $N_i \times N_i$ orthogonal matrix and Λ_i is a diagonal matrix with non-negative diagonal entries. Let $U_i = \Lambda_i^{\frac{1}{2}} \tilde{U}_i$ and $U = \text{diag}(U_1, U_2, \cdots, U_C) \in R^{N \times N}$, matrix K has the decomposition $K = U^T U \in R^{N \times N}$, where

$$U = [u_1^1, \cdots, u_{N_1}^1 | u_1^2, \cdots, u_{N_2}^2 | \cdots | u_1^C, \cdots, u_{N_C}^C] \tag{4}$$

and u_j^i is the $j + \sum_{k=1}^{i-1} N_k$ column of matrix U. A nonlinear mapping Φ is defined on the training data set X by:

$$\Phi(x_j^i) = u_j^i, \quad j = 1, 2, \cdots, N_i, i = 1, 2, \cdots, C.$$

Above Φ can be extended and well-defined on the whole input feature space by using interpolatory technique. In details, we extend the nonlinear mapping Φ to the entire input feature space R^d as follows:

$$\Phi(x) = U \cdot L(x), \ x \in R^d \tag{5}$$

where $L(x)$ is the interpolatory basis vector function defined by (2) and matrix U is determined by (4). Then, we denote function $K(x, y)$ by:

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_F,$$

where $\langle \cdot, \cdot \rangle_F$ is a inner product in feature space F. It can be directly derived that

$$K(x, y) = L^T(x) K L(y). \tag{6}$$

In order to show that the function $K(x, y)$ defined by (6) is a Mercer kernel, we need the following lemma.

Lemma 1. *[12] If $K(x, y)$ is a symmetric function defined on $R^d \times R^d$, and for any finite data set, it always yields a symmetric and positive semi-definite matrix $K = (k_{ij})_{m \times m}$, where $k_{ij} = k(y_i, y_j), i, j = 1, 2, \cdots, m$, then function $K(x, y)$ is a Mercer kernel function.*

Theorem 1. *Function $K(x, y) = L^T(x) K L(y)$, defined on $R^d \times R^d$, is a Mercer kernel function, where $L(\cdot)$ and K are determined by (2) and (3) respectively.*

Proof. Since function $K(x, y)$ is apparently a symmetric function, it is merely to show that the Gram matrix G generated by $K(x, y)$ on any finite training data is a positive semi-definite matrix. For any finite training data set $\{y_l | l = 1, 2, \cdots, n\} \subset R^d$, the Gram matrix G can be calculated as $G = [K(y_l, y_s)]_{n \times n}$. If we denote matrix L_n by $L_n = [L(y_1), L(y_2), \cdots, L(y_n)]_{N \times n}$, then Gram matrix G can be rewritten as $G = L_n^T K L_n$. For any column vector $\alpha \in R^n$, we have $\alpha^T G \alpha = (L_n \alpha)^T K (L_n \alpha) \geq 0$ because K is a positive semi-definite matrix. It indicates that G is a symmetric and positive semi-definite matrix. The theorem is concluded from Lemma 1 immediately.

The constructed kernel $K(x, y)$ has adopted the class label information and thus becomes a supervised kernel (SK) function, which will be evaluated using kernel PCA method.

4 Algorithm

This section develops a KPCA algorithm using SK function. Details are as follows.

Step 1: Construct symmetric and positive semi-definite matrix $K = \text{diag}(K_1, \cdots, K_C) \in R^{N \times N}$, where $K_i = [K_{RBF}(x_j^i, x_k^i)]_{N_i \times N_i}$ for $x_j^i, x_k^i \in X_i$.

Step 2: Let $L(x) = [L_j^i(x)] \in R^{N \times 1}$, where $L_j^i(x)$ are the intepolatory basis functions defined by

$$L_j^i(x) = \frac{\omega_j^i(x)}{\sum_{J=1}^{N_I} \sum_{I=1}^{C} \omega_J^I(x)},$$

where $\omega_j^i(x) = \prod_{(p,q) \neq (i,j)} \|x - x_q^p\|, q = 1, 2, \cdots, N_p, p = 1, 2, \cdots, C, x_q^p \in X_p$.

Step 3: The supervised kernel is constructed as

$$K(x, y) = L^T(x) K L(y).$$

Step 4: KPCA [4] with SK is performed for face recognition.

5 Experimental Results

This section will evaluate the performance of the proposed kernel based KPCA method for face recognition. In our experiments, linear PCA (PCA) [1] (as a benchmark here), PCA with RBF kernel (RBF-PCA) [4] and our method (SK-PCA) are chosen for comparisons on two datasets, namely ORL dataset and FERET dataset. The parameter t in RBF kernel is fixed to 1e3.

5.1 Facial Image Datasets

The ORL database contains 400 images of 40 persons and each person consists of 10 images with different facial expressions (open or closed eyes, smiling or not smiling), small variations in scales and orientations. The resolution of each image is 112×92, and with 256 gray levels per pixel. Facial image variations of one person from ORL database are shown in Figure 1.

For FERET database, we select 120 people, 6 images from each person. The resolution of each facial is also 112×92. FERET dataset is more challenging than ORL dataset since the variations in FERET database include pose, illumination, facial expression and aging. Images from two individuals are shown in Figure 2.

Fig. 1. Images of one person from ORL database

Fig. 2. Images of two persons from FERET database

5.2 Comparisons on ORL Dataset

The experimental setting is as follows. The $n(n = 2, 3, \cdots, 9)$ images are randomly selected from each people for training, and the remaining $(10 - n)$ images of each individual are for testing. The experiments are repeated 10 times and the mean accuracies are recorded in Table 1 and plotted in Figure 3 (left) respectively. It can be seen that the recognition rate of SK-PCA increases from 83.34% with training number 2 to 98.75% with training number 9, while the recognition rates of PCA and RBF-PCA increase from 83.06% and 83.25% with training number 2 to 95.50% and 97.00% with training number 9 respectively.

Experimental results show that the proposed SK-PCA method gives the best performance on ORL dataset.

Table 1. Mean accuracy(%) versus Training Number (TN) on ORL database

TN	2	3	4	5	6	7	8	9
PCA	83.06	89.39	89.50	92.40	91.56	93.25	93.88	95.50
RBF-PCA	83.25	89.50	89.88	93.20	92.63	94.33	94.37	97.00
SK-PCA	**83.34**	**91.64**	**93.88**	**95.25**	**96.69**	**97.42**	**97.63**	**98.75**

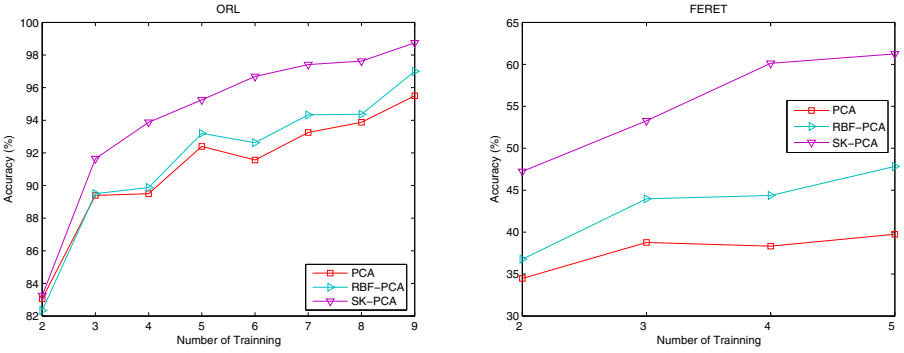

Fig. 3. Recognition rate on ORL face database (left) and FERET face database (right)

5.3 Comparisons on FERET Dataset

We randomly choose $n(n = 2, 3, \cdots, 5)$ images from each people for training, while the rest $(6 - n)$ images of each individual are selected for testing. The experiments are also run 10 times and the average accuracies are tabulated in Table 2 and plotted in Figure 3 (right) respectively. It can be seen that the recognition rate of SK-PCA increases from 47.23% (TN=2) to 61.25% (TN=5). In contrast, the accuracy of PCA ascends from 34.48% (TN=2) to 39.75% (TN=5) , while the accuracy of RBF-PCA increases from 36.77% (TN=2) to 47.83% (TN=5).

Compared with PCA and RBF-PCA, our SK-PCA gives around 17.64% and 12.23% entire mean accuracy improvements on FERET dataset, respectively.

Table 2. Mean accuracy (%) versus Training Number on FERET database

TN	2	3	4	5
PCA	34.48	38.78	38.33	39.75
RBF-PCA	36.77	43.97	44.37	47.83
SK-PCA	**47.23**	**53.28**	**60.13**	**61.25**

6 Conclusions

In this paper, we propose a novel methodology to construct supervised kernel function with high-performance. The class label information is incorporated into the kernel matrix and the interpolatory basis functions, range in [0 1], are established to obtain the analytic expression of nonlinear mapping. The kernel function, generated using the inner product of nonlinear mapping, is theoretically proven to be a supervised Mercer kernel. The constructed supervised kernel is tested using kernel PCA for face recognition. Experimental results on ORL and FERET databases show that our SK-PCA method surpasses linear PCA and RBF-PCA methods. Especially, our supervised kernel can be applied to all the kernel based machine learning tasks.

Acknowledgements. This paper is partially supported by NSF of China Grant (61272252) and Science & Technology Planning Project of Shenzhen City (JCYJ20130326111024546). We would like to thank Olivetti Research Laboratory and Amy Research Laboratory for providing the face image databases.

References

1. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neurosicence 3, 71–86 (1991)
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 711–720 (1997)

3. Schökopf, B., Smola, A., Müller, K.R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation 10, 1299–1319 (1998)
4. Kim, K., Jung, K., Kim, H.J.: Face Recognition Using Kernel Principal Component Analysis. IEEE Signal Processing Letters 9, 40–42 (2002)
5. Yang, J., Jin, Z., Yang, J.Y., Zhang, D., Frangi, A.F.: Essence of Kernel Fisher Discriminant: KPCA plus LDA. Pattern Recognition 37, 2097–2100 (2004)
6. Eftekhari, A., Forouzanfar, M., Abrishami Moghaddam, H., Alirezaie, J.: Block-Wise 2D Kernel PCA/LDA for Face Recognition. Information Processing Letters 110, 761–766 (2010)
7. Ebied, R.M.: Feature Extraction Using PCA and Kernel-PCA for Face Recognition. In: 8th International Conference on Informatics and Systems, pp. 72–77. IEEE Press, New York (2012)
8. Chen, W.S., Yuen, P.C.: Interpolatory Mercer Kernel Construction for Kernel Direct LDA on Face Recognition. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 857–860. IEEE Press, New York (2009)
9. Chu, W.S., Chen, J.C., James, L.J.J.: Kernel Discriminant Transformation for Image Set-Based Face Recognition. Pattern Recognition 44, 1567–1580 (2011)
10. Chan, C.H., Tahir, M.A., Kittler, J., Pietikäinen, M.: Multiscale Local Phase Quantization for Robust Component-Based Face Recognition Using Kernel Fusion of Multiple Descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 1164–1177 (2013)
11. Süli, E., Mayers, D.F.: An Introduction to Numerical Analysis. Cambridge University Press, Cambridge (2003)
12. Schölkopf, B., Smola, A.J.: Learning with Kernels-Support Vector Machine, Regularization, Optimization, and Beyond. The MIT Press, Cambridge (2002)