INTERPOLATORY MERCER KERNEL CONSTRUCTION FOR KERNEL DIRECT LDA ON FACE RECOGNITION

Wen-Sheng Chen

College of Mathematics & Computational Science
Shenzhen University, China, 518060
chenws@szu.edu.cn

Pong C Yuen

Department of Computer Science
Hong Kong Baptist University
pcyuen@comp.hkbu.edu.hk

ABSTRACT

This paper proposes a novel methodology on Mercer kernel construction using interpolatory strategy. Based on a given symmetric and positive semi-definite matrix (Gram matrix) and Cholesky decomposition, it first constructs a nonlinear mapping Φ, which is well-defined on the training data. This mapping is then extended to the whole input feature space by utilizing Lagrange interpolatory basis functions. The kernel function constructed by inner product is proven to be a Mercer kernel function. The self-constructed interpolatory Mercer (IM) kernel keeps the Gram matrix unchanged on the training samples. To evaluate the performance of the proposed IM kernel, a popular kernel direct linear discriminant analysis (KDDA) method for face recognition is selected. Comparing with RBF kernel based KDDA method on two face databases, namely FERET and CMU PIE databases, the IM kernel based KDDA approach could increase the performance by around 20% on CMU PIE database.

Index Terms— Mercer kernel, KDDA, Face recognition

1. INTRODUCTION

Over the past decade, positive semi-definite (Mercer) kernel functions have been popularly applied to the areas of machine learning [1]-[7]. The basic idea of kernel method is to apply a nonlinear mapping $\Phi : x \in R^d \rightarrow \Phi(x) \in F$ to the input data vector x and then to perform linear classifiers on the mapped feature space F. However, its dimension could be arbitrarily large and possibly infinite. Direct applying linear method to feature space is impossible. Kernel trick can overcome this obstacle and avoid using nonlinear mapping directly. The inner products $\langle \Phi(x_i), \Phi(x_j) \rangle_F$ can be replaced with a kernel function $K(x_i, x_j)$, i.e. $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_F$, where $x_i, x_j \in R^d$ are input pattern vectors. So the nonlinear mapping Φ can be performed implicitly in input space R^d. In kernel based approaches, kernel function can measure the similarity between two pattern samples. The advantage of using Mercer kernel as a similarity measure is that it allows us to construct algorithms in inner product spaces. Gram matrix, also called kernel matrix, is generated by the inner product of mapped training samples and thus can be calculated by a kernel function. Gram matrix is a symmetric and positive semi-definite matrix and plays very important role in kernel based machine learning. The question is what kind of Gram matrix is good for kernel based classifier? It is natural to hope that the similarities are higher among within-class samples and lower among between-class samples. However, the Gram matrices, which are computed by the commonly used kernels such as RBF/polynomial kernels on the training data, are full matrices. It means that the between-class data possibly have higher similarity and this leads to degrading the performance of kernel based learning methods. So, it is reasonable to think that such a kernel is a better kernel, if its Gram matrix generated from the the training data is a block diagonal matrix. To overcome the drawback of commonly used RBF kernel, this paper first exploits a RBF kernel to generate a symmetric and positive definite block diagonal matrix $\mathbf{K}$ on the training samples, and then utilizes Cholesky decomposition technique to construct a feature mapping, which is just well-defined on training data. The feature mapping is subsequently expanded to the whole input space using Lagrange interpolatory strategy. It is shown that our self-constructed interpolatory kernel is indeed a Mercer kernel. The Gram matrix determined by our IM kernel on the training data is exactly the previous constructed block diagonal matrix $\mathbf{K}$. To evaluate the performance of our IM kernel, it is applied to KDDA for face recognition. Comparing with KDDA with RBF kernel, KDDA with IM kernel gives superior performance.

The rest of this paper is organized as follows. Section 2 describes the details on IM kernel construction and theoretically shows our interpolatory kernel is a Mercer kernel. Section 3 designs a IM kernel based KDDA algorithm. Section 4 reports kernel performance comparisons on FERET and CUM PIE databases by KDDA with IM kernel and RBF kernel. Finally, the conclusions are drawn in section 5.

2. PROPOSED METHODOLOGY

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

2.1. Some notations

Let d and C be the dimension of input feature space and the number of sample classes respectively, the total training sample set $X = \{X_1, X_2, \cdots, X_C\} \subset R^d$, the ith class X_i contains N_i samples, namely $X_i = \{x_1^i, x_2^i, \cdots, x_{N_i}^i\}$, $i = 1, 2, \cdots, C$, $N(= \sum_{i=1}^{C} N_i)$ be the total number of original training samples. If $\Phi(x)$: $x \in R^d \rightarrow \Phi(x) \in F$ is the kernel nonlinear mapping, where F is the mapped feature space, denote $df = \dim F$, the total mapped sample set is $\Phi(X) = \{\Phi(X_1), \Phi(X_2), \cdots, \Phi(X_C)\}$, and the ith mapped class is $\Phi(X_i) = \{\Phi(x_1^i), \Phi(x_2^i), \cdots, \Phi(x_{N_i}^i)\}$. If $K(x, y)$ is a Mercer defined on $R^d \times R^d$, then there exists a nonlinear mapping Φ, such that $K(x, y) = \langle \Phi(x), \Phi(y) \rangle_F$. We denote RBF kernel $K_{RBF}(x, y)$ by $K_{RBF}(x, y) = \exp(-\frac{\|x-y\|^2}{t})$ with $t > 0$. Define matrices $\mathbf{K}_i = (k_{jk}^i)_{N_i \times N_i} \in R^{N_i \times N_i}$, where $k_{jk}^i = K_{RBF}(x_j^i, x_k^i)$, $i = 1, 2, \ldots, C$. So, $\mathbf{K}_i$ ($i = 1, 2, \ldots, C$) all are symmetric and positive semi-definite matrices. If let

$$\mathbf{K} = \text{diag}\{\mathbf{K}_1, \ldots, \mathbf{K}_C\} \in R^{N \times N}, \tag{1}$$

then $\mathbf{K}$ is a symmetric and positive semi-definite matrix as well.

2.2. Cholesky decomposition

Let matrix $\mathbf{K}$ be the matrix define by (1). Since submatrices $\mathbf{K_i}$ ($i = 1, 2, \ldots, C$) are generated by RBF kernel and thus are symmetric and positive semi-definite matrix.

By performing Cholesky decomposition on matrix $\mathbf{K}_i$, we have that $\mathbf{K}_i = U_i^T U_i \in R^{N_i \times N_i}$, where U_i is a unique $N_i \times N_i$ upper triangular matrix. Denote that $U = \text{diag}\{U_1, U_2, \cdots, U_C\} \in R^{N \times N}$, then U is also a upper triangular matrix. The Cholesky decomposition of matrix $\mathbf{K}$ can be written as $\mathbf{K} = U^T U \in R^{N \times N}$. We rewrite matrix U as $U = [u_1^1, \cdots, u_{N_1}^1 | u_1^2, \cdots, u_{N_2}^2 | \cdots | u_1^C, \cdots, u_{N_C}^C]$, where $u_j^i \in R^N$ is the $(j + \sum_{k=1}^{C} N_k)$ column vector. Define nonlinear feature mapping Φ on the training data X set as:

$$\Phi(x_j^i) = u_j^i, \text{where } j = 1, 2, \ldots, N_i \text{ and } i = 1, 2, \ldots, C. \tag{2}$$

2.3. Interpolatory strategy

By using interpolatory technique, this subsection will expend the nonlinear mapping $\Phi(x)$ (see (2)), which is just well-defined on training sample set, to the whole input space. To this end, we define N Lagrange interpolatory basis functions $L_j^i(x)$ as

$$L_j^i(x) = \frac{\prod_{(p,q) \neq (i,j)} \|x - x_q^p\|_2}{\prod_{(p,q) \neq (i,j)} \|x_j^i - x_q^p\|_2}, \quad x \in R^d.$$

Apparently, above interpolatory basis functions satisfy the following property

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

Therefore, the nonlinear mapping $\Phi(x)$ can be expanded to the whole input feature space R^d as follows:

$$\Phi(x) = \sum_{i=1}^{C} \sum_{j=1}^{N_i} L_j^i(x) u_j^i. \tag{3}$$

2.4. Interpolatory Mercer kernel construction

Based on the nonlinear feature mapping defined in (3), we can construct the kernel function on $R^d \times R^d$ below:

$$\begin{aligned} K(x, y) &= \langle \Phi(x), \Phi(y) \rangle \\ &= \{\sum_{i=1}^{C} \sum_{j=1}^{N_i} L_j^i(x) u_j^i\}^T \cdot \{\sum_{p=1}^{C} \sum_{q=1}^{N_p} L_q^p(y) u_q^p\}. \end{aligned} \tag{4}$$

Obviously, function $K(x, y)$ is a symmetric function. The following theorem 1 demonstrates that above $K(x, y)$ is indeed a Mercer kernel function.

Lemma. *[8] If $K(x, y)$ is a symmetric function defined on $R^d \times R^d$, and for any finite data set $\{y_1, \cdots, y_m\} \subset R^d$, it always yields a symmetric and positive semi-definite matrix $\mathbf{K} = (k_{ij})_{m \times m}$, 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. *Function $K(x, y)$ defined by (4) is a Mercer kernel function.*

Proof. We just need to show that $K(x, y)$ is a positive semi-definite function. To this end, we first denote a column vector $\mathbf{L}(x) \in R^N$ as following:

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

then the function $K(x, y)$ can be written as

$$\begin{aligned} K(x, y) &= (U\mathbf{L}(x))^T \cdot (U\mathbf{L}(y)) \\ &= \mathbf{L}(x)^T \cdot (U^T U) \cdot \mathbf{L}(y) \\ &= \mathbf{L}(x)^T \mathbf{K} \mathbf{L}(y). \end{aligned}$$

For any finite training data set $\{x_l | l = 1, 2, \cdots, n\} \subset R^d$, the Gram matrix $\mathbf{G}$ generated by the kernel function $K(x, y)$ on this n training data set is $\mathbf{G} = [K(x_l, x_s)]_{n \times n}$, where $K(x_l, x_s) = \mathbf{L}(x_l)^T \mathbf{K} \mathbf{L}(x_s)$, $l, s = 1, 2, ...n$. Let $\mathbf{L}_n = [\mathbf{L}(x_l), \mathbf{L}(x_2), \cdots, \mathbf{L}(x_n)]_{N \times n}$, the Gram matrix $\mathbf{G}$ can be written as $\mathbf{G} = \mathbf{L}_n^T \mathbf{K} \mathbf{L}_n$. Thereby, $\mathbf{G}$ is a symmetric matrix. As $\mathbf{K}$ is a positive semi-definite matrix, for all $\theta \in R^n$, we have

$$\theta^T \mathbf{G} \theta = \theta^T \mathbf{L}_n^T \mathbf{K} \mathbf{L}_n \theta = (\mathbf{L}_n \theta)^T \mathbf{K} (\mathbf{L}_n \theta) \geq 0.$$

It means that Gram matrix $\mathbf{G}$ *is a positive semi-definite matrix. Hence by lemma 1, we know that* $K(x, y)$ *is a Mercer kernel.*

It is not difficult to verify that the Gram matrix $\mathbf{G}_X$, which is generated by our IM kernel (4) on the training data set X, is exactly the block diagonal positive semi-definite matrix $\mathbf{K}$. This indicates that the similarities among between-class data are zeros, while the similarities among within-class data are greater than zeros. Therefore, our IM kernel is good for measuring the similarity between two samples and will enhance the the classification power of Kernel based machine learning approaches.

3. ALGORITHM DESIGN

Based on analysis in above sections, our IM-KDDA algorithm is designed as follows.

Step 1: Construct symmetric and positive semi-definite matrix $\mathbf{K} = \text{diag}\{\mathbf{K}_1, \ldots, \mathbf{K}_C\} \in R^{N\times N}$, where $\mathbf{K}_i = [K_{RBF}(x_j^i, x_k^i)]_{N_i\times N_i}, x_j^i, x_k^i \in X_i$, and $K_{RBF}(x_j^i, x_k^i) = \exp\left(\frac{-\|x_j^i - x_k^i\|^2}{t}\right)$.

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

$$L_j^i(x) = \frac{\prod_{(p,q)\neq(i,j)} \|x - x_q^p\|_2}{\prod_{(p,q)\neq(i,j)} \|x_j^i - x_q^p\|_2},$$

$x_q^p \in X_p, x_j^i \in X_i$.

Step 3: The interpolatory Mercer kernel is constructed as $K(x, y) = \mathbf{L}^T(x)\mathbf{K}\mathbf{L}^T(x)$.

Step 4: KDDA [3] with IM kernel is performed for face recognition.

Remark. In the above algorithm, if the value of some interpolatory basis function exceeds a given large threshold, then its value is set to zero.

4. EXPERIMENTAL RESULTS

In this section, two databases, namely FERET and CMU PIE databases, are selected to evaluate the performance of our self-constructed IM kernel for kernel direct linear discriminant analysis algorithm.

4.1. Face image datasets

For FERET database, we select 120 people, 6 images for each individual. Face image variations in FERET database include pose, illumination, facial expression and aging. Images from one individual are shown in Figure 1.

Fig. 1. Six images of one person on FERET dataset

CMU PIE face database, includes totally 68 people. There are 13 pose variations ranged from full right profile image to full left profile image and 43 different lighting conditions, 21 flashes with ambient light on or off. In our experiment, for each people, we select 56 images including 13 poses with neutral expression and 43 different lighting conditions in frontal view. Several images of one people are shown in Figure 2.

Fig. 2. Parts images of one person on CMU PIE

In above two face databases, all images are aligned with the centers of eyes and mouth. The orientation of face is adjusted (on-the-plane rotation) such that the line joining the centers of eyes is parallel with x-axis. Also, the original images with resolution 112x92 are reduced to wavelet feature faces with resolution 30x25 after two-level D4 wavelet decomposition.

4.2. Results on FERET dataset

This section reports the results of proposed IM-KDDA method on FERET database. We randomly select n (n=2 to 5) images from each people for training , while the rest $(6-n)$ images of each individual are selected for testing. The experiments are repeated 10 times and the average accuracies are recorded in Table 1. It can be seen that the recognition rate of KDDA with IM kernel increases from 73.06% with training number 2 to 92.00% with training number 5, while the recognition accuracy of KDDA with RBF kernel increases from 69.13% with training number 2 to 91.50% with training number 5 respectively.

Comparing with KDDA with RBF kernel, KDDA with our IM kernel gives around 2.81% entire mean accuracy improvement.

Table 1. Average accuracy of rank 1 versus Training Number (TN) on FERET database

TN	2	3	4	5
RBF Kernel	69.13%	80.89%	89.17%	91.50%
Our Kernel	73.06%	84.08%	90.33%	92.00%

Table 2. Average accuracy (%) of rank 1 versus Training Number on CMU PIE database.

TN	5	6	7	8	9	10
RBF	67.51	68.11	70.79	72.34	72.74	72.91
Our	86.03	89.15	90.78	92.16	88.16	94.26

4.3. Results on CMU PIE dataset

The experimental setting on the CMU PIE database is similar with that of FERET database. As the number of images for each individual is 56, the number of training images is ranged from 5 to 10. The experiments are repeated 10 times and the average accuracy of KDDA with IM kernel is then calculated. The average accuracy are recorded and tabulated in the last row of Table 2. It can be seen from Table 2 that the recognition accuracy of proposed method increases from 86.03% with 5 training images to 94.26% with 10 training images. The results are encouraging.

The same experiments are implemented by using KDDA with RBF kernel function. The results are also recorded and tabulated in Table 2. It can be seen that when 5 images are used for training, the accuracy for KDDA with RBF kernel is 67.51%. When the number of training images is equal to 10, the accuracy for RBF kernel based KDDA increases to 72.91%. Comparing with RBF based KDDA method, KDDA with IM kernel gives around 19.36% entire average accuracy improvement.

In the 10 repeated experiments with training number 9, we found that the abnormal situations occurred in 2 times running, namely the value of some interpolatory basis function exceeds a given large threshold and probably attains infinite. So, we set its value to zero in practice. The 10 times mean accuracy with training number 9 is 88.16%. If excluding 2 abnormal cases, the mean accuracy of the rest 8 times running improves to 93.84%. Comparing with RBF based KDDA method, KDDA with IM kernel gives around 20.30% entire mean accuracy improvement. It can be seen that our IM kernel based KDDA approach gives the best performance for all cases.

5. CONCLUSIONS

This paper proposed a novel framework on Mercer Kernel construction using interpolatory strategy. Our IM kernel is constructed using Cholesky decomposition technique and then applied to KDDA for face recognition tasks. The results are encouraging on FERET and CMU PIE face databases. Comparing with RBK kernel based KDDA, experimental results show that the proposed self-constructed IM kernel based KDDA algorithm gives the best performance.

6. ACKNOWLEDGEMENT

This project is supported by the Hong Kong RGC General Research Fund HKBU2113/06E and NSF of China (60873168). The authors would like to thank for the US Army Research Laboratory for contribution of the FERET database and CMU for the CMU PIE database

7. REFERENCES

[1] S. Mika, G. Rätsch, J Weston, B Schölkopf, and K. R. Müller, "Fisher discriminant analysis with kernels," *Neural Networks for Signal Processing IX*, pp. 41–48, August, 1999.

[2] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach", *Neural Computation*, Vol.12, No.10, pp.2385-2404, 2000.

[3] J. Lu, K. N. Plataniotis, and A. N. Ventsanopoulos, "Face recognition using kernel direct discriminant analysis," *IEEE Trans. on Neural Network*, vol. 14, pp. 117–126, January 2003.

[4] C. J. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.26, Issue 5, pp.572–581, 2004.

[5] W. S. Chen, P. C. Yuen, J. Huang and D. Q. Dai, "Kernel Machine-based One-parameter Regularized Fisher Discriminant Method for Face Recognition, " *IEEE Transactions on System, Man and Cybernetics, Part B*, Vol. 35, pp 659–669, August 2005.

[6] T. Evgeniou, C. A. Micchelli, M. Pontil "Learning Multiple Tasks with Kernel Methods, " *Journal of Machine Learning Research*, Vol.6, pp.615–637, 2005.

[7] F. De la Torre, ; O. Vinyals, "Learning Kernel Expansions for Image Classification, " *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–7, 2007.

[8] B. Scholkopf and A. J. Smola, " Learning with kernels-Support vector machine, Regularization, Optimization, and Beyond," *The MIT Press* 2002.