Dimensionality Reduction Using Kernel Pooled Local Discriminant Information

Peng Zhang & Jing Peng
EECS Department
Tulane University
New Orleans, LA 70118
{zhangp,jp}@eecs.tulane.edu

Carlotta Domeniconi
ISE Department
George Mason University
Fairfax, VA 22030
carlotta@ise.gmu.edu

Abstract

We study the use of kernel subspace methods for learning low-dimensional representations for classification. We propose a kernel pooled local discriminant subspace method and compare it against several competing techniques: generalized Fisher discriminant analysis (GDA) and kernel principal components analysis (KPCA) in classification problems. We evaluate the classification performance of the nearest-neighbor rule with each subspace representation. The experimental results demonstrate the efficacy of the kernel pooled local subspace method and the potential for substantial improvements over competing methods such as KPCA in some classification problems.

1 Introduction

Subspace analysis methods such as GDA and KPCA play an important role in classification and data mining. In data visualization and classification the principal modes are extracted and utilized for description, detection, and classification. Using these "principal modes" to represent data can be found in data preprocessing [5, 10] and linear discriminant analysis [7].

Subspace analysis often significantly simplifies tasks such as regression, classification, and density estimation by computing low-dimensional subspaces having statistically uncorrelated or independent variables. PCA [9] is a prime example that employs eigenvector-based techniques to reduce dimensionality and extract features. KPCA [12] and GDA [1] extend these linear techniques in a nonlinear fashion. In this paper we propose a kernel pooled local Fisher discriminant subspace method for learning low-dimensional representations for classification. We perform a nonlinear global dimensionality reduction by pooling local discriminant dimension information in feature space and applying the kernel trick [4] to capture nonlinearity. The resulting subspaces are nonlinear, discriminant and compact, whereby better classification performance and greater computational efficiency can be expected.

2 Subspace Methods

The objective of subspace analysis is to represent high-dimensional data in a low-dimensional subspace according to some optimality criteria. Classification then takes place on the chosen subspace. Here we briefly describe several methods for computing both linear and nonlinear subspaces and highlight their corresponding characteristics. We assume that the data can be captured by a compact and connected subspace, which is often the case, for example, in face recognition.

Kernel PCA is a nonlinear version of PCA [12]. KPCA applies a nonlinear mapping to the input $\phi(\mathbf{x}) : \Re^q \rightarrow \Re^N$ and then solve for a linear PCA in the induced feature space $\Re^N$, where $N \gg q$ and possibly infinite. In KPCA, the mapping ϕ is made implicit by the use of kernel functions satisfying Mercer's theorem [3] $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y})$. Since computing covariance involves only dot-products, performing a PCA in the feature space can be formulated with kernels in the input space without the explicit (and possibly prohibitive) direct computation of ϕ. A major advan-

tage of KPCA over principal curves is that KPCA does not require nonlinear optimization. On the other hand, selecting the optimal kernel (and its associated parameters) remains an engineering problem.

Similar to KPCA, GDA [1] is a kernelized version of Fisher discriminant analysis (FDA). The basic idea is to maximize the ratio of the between sum-of-squares matrix to the within sum-of-squares matrix in the feature space. The major problem associated with GDA (or FDA) is that the within sum-of-squares matrix is usually degenerated in practice. Often this problem is solved by using techniques such as pseudo inverse or PCA to remove the null space of the within sum-of-squares matrix. However, it can be shown that the null space potentially contains significant discriminant information [2].

3 Kernel Pooled local Discriminant Subspace Method

3.1 Pooled Local Subspace Method

Hastie and Tibshrani [8] propose a global dimension reduction technique by pooling local dimension information. For each training point, local pooling calculates the local centroid deviations $\tilde{\mathbf{x}}_j(i) = \mathbf{x}_j(i) - \hat{\mathbf{x}}(i)$, where $\mathbf{x}_j(i)$ denotes the mean of class j points in a neighborhood of the ith training point, and $\hat{\mathbf{x}}(i)$ the overall mean. Then it seeks a subspace that is close in average weighted squared distance to all these deviations. If U denotes an orthonormal basis for the subspace, this subspace can be computed by minimizing the total weighted residual sum of squares $RSS(U) = \sum_{i=1}^{l}\sum_{j=1}^{J}\pi_j(i)\tilde{\mathbf{x}}_j^t(i)(I-UU^t)\tilde{\mathbf{x}}_j(i)$, where $\pi_j(i)$ represents the local class membership proportions, l the number of training samples, and J the number of classes.

It turns out, as shown in [8], that this subspace is spanned by the largest eigenvectors of the average between-sum of squares matrix: $B = \frac{1}{l}\sum_{i=1}^{l} B_i$, where B_i denotes the local between-sum of squares matrices at the ith training point: $B_i = \sum_{j=1}^{J}\pi_j(i)(\bar{\mathbf{x}}_j^{(i)} - \bar{\mathbf{x}}^{(i)})(\bar{\mathbf{x}}_j^{(i)} - \bar{\mathbf{x}}^{(i)})^t$. The experimental results presented in [8] show that the pooled local subspace method is very promising.

It is important to note that local pooling does not sphere the data locally before calculating the centroid deviations. An argument given in [8] is that any local spherical window containing two classes will likely have a linear decision boundary orthogonal to the line joining the two means. As a result, local pooling will not suffer the small size sample problem (degenerate within class matrices) facing FDA or GDA [1, 2]. It is interesting to note that locally linear embedding also uses pooled locally linear constraints to compute a global subspace [11].

3.2 Kernel Pooled Local Subspace Analysis

We now show how to compute a nonlinear pooled local discriminant subspace by using the kernel trick [4]. Let $\phi : \mathbf{x} \to \phi(\mathbf{x})$ be the nonlinear mapping from $\Re^q$ to $\Re^N$. Also, let $\bar{\phi}_j(\mathbf{x}^{(i)}) = \frac{1}{l_{ij}}\sum_{y_k=j}\phi(\mathbf{x}_k^{(i)})$ be the mean of class j samples in a neighborhood of the ith training point in the feature space, and $\bar{\phi}(\mathbf{x}^{(i)}) = \frac{1}{l_i}\sum_{k=1}^{l_i}\phi(\mathbf{x}_k^{(i)})$ be the overall mean in the same neighborhood, where l_j represents the number of class j training samples in the neighborhood of the ith training point, and l_i the total number of training samples in the neighborhood. Then the local between-sum of squares matrix at the ith training point in the feature space is $B_i^{\phi} = \frac{1}{J}\sum_{j=1}^{J}(\bar{\phi}_j(\mathbf{x}^{(i)}) - \bar{\phi}(\mathbf{x}^{(i)}))(\bar{\phi}_j(\mathbf{x}^{(i)}) - \bar{\phi}(\mathbf{x}^{(i)}))^t$. The pooled local subspace method seeks a discriminant subspace that is close to all of B_is. The average between-sum of squares matrix B in the feature space is

$$B^{\phi} = \frac{1}{l}\sum_{i=1}^{l} B_i^{\phi} = \frac{1}{lJ}\sum_{i=1}^{l}\sum_{j=1}^{J}\tilde{\phi}_j(\mathbf{x}^{(i)})\tilde{\phi}_j(\mathbf{x}^{(i)})^t \qquad (1)$$

where $\tilde{\phi}_j(\mathbf{x}^{(i)}) = \bar{\phi}_j(\mathbf{x}^{(i)}) - \bar{\phi}(\mathbf{x}^{(i)})$.

Similar to KPCA [12], we have the eigenvector equation $\lambda\mathbf{v} = B^{\phi}\mathbf{v}$. Clearly all solutions must lie in the span of $\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_l)$. Therefore, there exist coefficients α_i $(i = 1, \cdots, l)$ such that

$$\mathbf{v} = \sum_{i=1}^{l}\alpha_i\phi(\mathbf{x}_i). \qquad (2)$$

It is also true that for all $k = 1, \cdots, l$ we have

$$\lambda(\phi(\mathbf{x}_k)\cdot\mathbf{v}) = (\phi(\mathbf{x}_k)\cdot B^{\phi}\mathbf{v}) \qquad (3)$$

Substituting (2) and (1) into (3), we obtain, the left hand side of (3) $\lambda(\phi(\mathbf{x}_k)\cdot\mathbf{v}) = \lambda\sum_{i=1}^{l}\alpha_i\phi(\mathbf{x}_k)\cdot\phi(\mathbf{x}_i)$. For the right hand side of (3), we have $\sum_{j=1}^{J}[\phi(\mathbf{x}_k)\cdot$

$\tilde{\phi}_j(\mathbf{x}^{(1)}) \cdots \phi(\mathbf{x}_k) \cdot \tilde{\phi}_j(\mathbf{x}^{(l)})] K_j \alpha$, for all $k = 1, \cdots, l$. Here $\alpha = (\alpha_1, \cdots, \alpha_l)^t$ and

$$K_j = \begin{pmatrix} \tilde{\phi}_j(\mathbf{x}^{(1)}) \cdot \phi(\mathbf{x}_1) & \cdots & \tilde{\phi}_j(\mathbf{x}^{(1)}) \cdot \phi(\mathbf{x}_l) \\ \cdots & \cdots & \cdots \\ \tilde{\phi}_j(\mathbf{x}^{(l)}) \cdot \phi(\mathbf{x}_1) & \cdots & \tilde{\phi}_j(\mathbf{x}^{(l)}) \cdot \phi(\mathbf{x}_l) \end{pmatrix}. \quad (4)$$

Define

$$K = [k_{ij}] = [\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)] \quad (5)$$

and $\tilde{K} = \sum_{j=1}^{J} K_j^t K_j$. We obtain

$$(lJ)\lambda K \alpha = \tilde{K} \alpha \quad (6)$$

which is a generalized eigenvector problem [6]. By solving (6), the ith principal component u_i of $\mathbf{x}$ can be calculated according to: $u_i = \mathbf{v}_i \cdot \phi(\mathbf{x}) = \sum_{k=1}^{l} \alpha_k^i k(\mathbf{x}, \mathbf{x}_k)$, where $\mathbf{v}_i$ denotes the ith eigenvector of the feature space.

4 Kernel Pooled Local Subspace Algorithm

Calculating K_j^c is the key step in the implementation. Here we show it is done. Notice that

$$(\bar{\phi}_j(\mathbf{x}^{(i)}) - \bar{\phi}(\mathbf{x}^{(i)})) \cdot \phi(\mathbf{x}_k) = \mathbf{1}_{l_{ij}}^t K_{j,k}^{(i)} - \mathbf{1}_{l_i}^t K_k^{(i)} \quad (7)$$

where $\mathbf{1}_{l_{ij}} = (\frac{1}{l_{ij}}, \cdots, \frac{1}{l_{ij}})^t$, $\mathbf{1}_{l_i} = (\frac{1}{l_i}, \cdots, \frac{1}{l_i})^t$, $K_{j,k}^{(i)} = (k(\mathbf{x}_{j,1}^{(i)}, \mathbf{x}_k), \cdots, k(\mathbf{x}_{j,l_{ij}}^{(i)}, \mathbf{x}_k))^t$, and $K_k^{(i)} = (k(\mathbf{x}_{i,1}^{(i)}, \mathbf{x}_k), \cdots, k(\mathbf{x}_{i,l_i}^{(i)}, \mathbf{x}_k))^t$, $k = 1, \cdots, l$. Thus, each row of K_j^c can be calculated according to: $\mathbf{1}_{l_{ij}} K_j^{(i)} - \mathbf{1}_{l_i} K^{(i)}$, where $K_j^{(i)} = (K_{j,1}^{(i)}, \cdots, K_{j,l}^{(i)})$ is a $l_{ij} \times l$ matrix, and $K^{(i)} = (K_1^{(i)}, \cdots, K_l^{(i)})$ is a $l_i \times l$ matrix.

Figure 1 Here we use a 2D toy example to illustrate subspace computation by KPooLS using Gaussian kernels. The left panel shows the 2D example, where the two class data are uniformly distributed in two dimensions, separated by a sinusoidal decision boundary. The middle panel shows the projection of the data onto the first two eigenvectors of the feature space computed by KPooLS, while the right panel shows the intensity values of the first principal component when the 2D space is projected onto the first eigenvector of the feature space.

5 Experimental Results

In the following we use two data sets to examine the classification performance of the nearest neighbor rule (3NN)

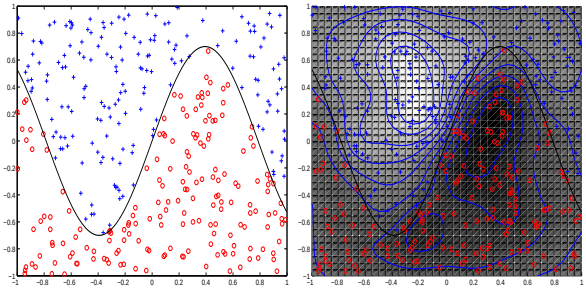

Figure 1. Left panel: 2D toy example. Right panel: First principal component values of KPooLS.

with each subspace representation, i.e., PCA, KPCA, and the proposed KPooLS algorithm. Since our focus here is on subspace methods, a simple classifier is prefered.

For each data set, we randomly select 60% of the data as training and remaining 40% as testing. This process is repeated 20 times and the average error distributions are reported. The dimensions of subspaces computed by each method are determined so that only eigenvectors remain whose eigenvalues are great than or equal to $0.01\lambda_{max}$, i.e., 90% of variations are retained. Also, we use Gaussian kernels for all nonlinear subspace methods and kernel parameters are determined through cross-validation.

Glass Data: This data set is taken from UCI Machine Learning Repository. It consists of $q = 9$ chemical attributes measured for each of $N = 214$ data of $J = 6$ classes. The left panel in Figure 2 shows the error distributions of each subspace methods.

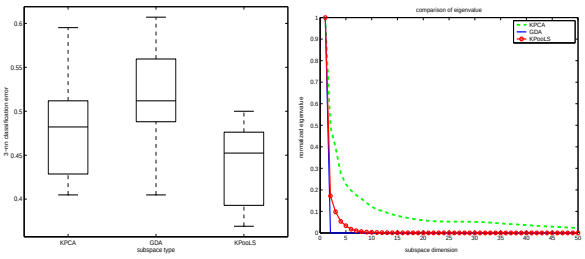

Figure 2. Left panel: Error distributions by each subspace method on the Glass data. Right panel: Normalized eigenvalues

The error distributions clearly favor KPooLS. It is interesting to note that while the normalized eigenvalues of the

principal subspaces computed by KPooLS and GDA follow a similar trend (right panel), KPooLS performed better than GDA. On average, both KPooLS and GDA used three principal modes to represent the subspace. In contrast, KPCA used 12. GDA performed slightly worse than KPCA. However, KPCA used substantially larger subspaces (12 dimensions on average).

Cat and Dog Image Data: In this experiment, the data set is composed of one hundred images of cat faces and dog faces. Each image is a black-and-white 64×64 pixel image, resulting in 4096 dimensional measurement space. These images have been registered by aligning the eyes.

The left panel in Figure 3 shows the error distributions obtained by each method on the cat and dog image data. KPooLS and GDA registered similar performance. However, KPCA performed significantly worse, as expected.

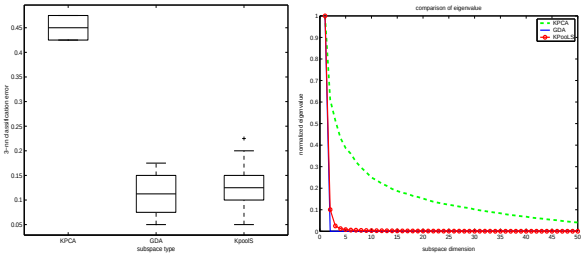

Figure 3. Left panel: Error distributions by each subspace method on the cat and dog data. Right panel: Normalized eigenvalues.

The right panel in Figure 3 shows normalized eigenvalues calculated by each method. The eigenvalues of the principal spaces computed by both KPooLS and GDA again decrease rapidly. On average, the subspaces computed by KPooLS and GDA were represented by two principal components. In contrast, KPCA used 30 principal components to represent the subspaces. It is rather surprising to see that KPCA failed to achieve significant dimensionality reduction.

6 Summary

This paper presents a kernel pooled local discriminant subspace method for learning low-dimensional representations for classification. This method performs a nonlinear global dimensionality reduction by pooling local dimension information and applying the kernel trick to capture nonlinearity. The resulting Subspaces are nonlinear, discriminant and compact, whereby better classification performance and greater computational efficiency can be achieved. The experimental results show that the KPooLS algorithm can learn discriminant subspaces that are much more compact than that computed by KPCA and can outperform competing methods such as GDA in some classification problems.

References

[1] G. Baudat and F. Anouar. Generalized discriminant analysis using akernel approach. *Neural Computation*, 12:2385–2404, 2000.

[2] L. Chen and et al. A new lda-based face recognition system which can solve the small samplesize problem. *Pattern Recognition*, 33:1713–1726, 2000.

[3] R. Courant and D. Hilbert, editors. *Methods of Mathematical Physiacs, vol. 1*. Interscience, New York, 1953.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.

[5] M. Dash and H. Liu. Feature selection methods for classifi - cation. *Intelligent Data Analysis: An International Journal*, 1, 1997.

[6] R. O. Duda and P. E. Hart. *Pattern Classifi cation and Scene Analysis*. John Wiley & Sons, Inc., 1973.

[7] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human faces. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 2148–2151, 1996.

[8] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classifi cation and regression. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 409–415. The MIT Press, 1996.

[9] I. Jolliffe. *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[10] J. Neter, M. Kutner, C. Nachtsheim, and L. Wasserman. *Applied Linear Statistical Models, 4th Edition*. Irwin, Chicago, 1996.

[11] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction bylocally linear embedding. *Science*, 290:2323–2326, 2000.

[12] B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.