Dimensionality Reduction with Adaptive Kernels

Shuicheng Yan and Xiaoou Tang
Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

Abstract[1]

A kernel determines the inductive bias of a learning algorithm on a specific data set, and it is beneficial to design specific kernel for a given data set. In this work, we propose a kind of new kernel, called Locality-Adaptive-Kernel (LAKE), which adaptively measures the data similarity by considering the geometrical structure of the data set. In theory, we prove that the LAKE is a special marginalized kernel; and intuitively, when the local kernel in LAKE is constrained to be linear, it has the explicit semantic of merging multiple local linear analyzers into a single global nonlinear one. We show in a toy problem that the kernel principal component analysis with LAKE well captures the intrinsic nonlinear principal curve of the data set. Moreover, a large set of experiments are presented to verify that the classification performance is sensitive to the kernel variation; and the extensive face recognition experiments on different databases demonstrate that KPCA and KDA based on LAKE are both superior to those based on traditional fixed kernels.

1. Backgrounds

In the past decades, many techniques have been proposed for dimensionality reduction task and well applied in face recognition task. In the following, we first introduce a common formulation for the dimensionality reduction problem and then discuss the kernel trick.

1.1 Kernel-based Dimensionality Reduction

Assume that the training sample set is denoted as $X = [x_1, x_2, \cdots, x_N] \in \mathrm{R}^{m \times N}$. In supervised learning problems, the class label of x_i is denoted as $l_i \in \{1, 2, \cdots, N_c\}$ and denote n_c as the sample number for the c-th class. Let $G=\{X, W\}$ be an undirected weighted graph with the sample matrix X as the vertex set. The similarity matrix W measures the similarities of different vertex pairs. The general goal of dimensionality reduction is to pursue a lower dimensional feature space that best characterizes the adjacency relationship of the graph measured by the similarity matrix W.

Denote the lower dimensional representations of the training samples as matrix Y, where the i-th row of Y is the low dimensional representation for sample x_i. For simplicity, we take the one-dimensional as example in this paper and replace Y by y. A starting point to preserve the adjacency relationship of a graph is to minimize the objective function as follows [9]

$$y^* = arg \min_{y'B\,y=c} \sum_{i,j} \| y_i - y_j \|^2 W_{ij} = arg \min_{y'B\,y=c} y'Ly \ . \tag{1}$$

Here B is a constraint matrix, c is a const number, D ($D_{ii} = \sum_j W_{ij}$) is a diagonal matrix with the diagonal element being the row sum of the matrix W and $L = D - W$ is called the Laplacian matrix of a graph.

Assume that the low dimensional representations are obtained from a linear projection $y = X'w$; then

$$w^* = arg \min_{\substack{w'XBX'w=c \\ or \ \ w'w=c}} w'XLX'w \, . \tag{2}$$

Define the kernel function as $k(x, y) = \phi(x) \cdot \phi(y)$. Let $w = \sum_i \alpha_i \phi(x_i)$ and K is the kernel Gram matrix with $K_{ij} = k(x_i, x_j)$, then $y = K\alpha$ and we have

$$a^* = arg \min_{\substack{\alpha'KBK\alpha=c \\ or \ \ \alpha'\alpha=c}} \alpha'KLK\alpha \, . \tag{3}$$

The problems defined in (1-3) can all be solved using the generalized eigenvalue decomposition method [9]. Most previous dimensionality reduction algorithms, like Principal Component Analysis (PCA) [4], Kernel-PCA (KPCA), and Linear / Kernel Discriminant Analysis (LDA/KDA) [9] all follow the above formulations.

1.2 Kernel Selection Issue

A kernel determines the inductive bias of a learning algorithm on a specific data set [5]. Therefore, selecting proper kernel for specific data is critical for optimal performance; yet how to determine the optimal kernel or the parameter of a kernel is still an unsolved problem. Most kernel-based algorithms assumed that the kernel was manually defined and the kernel parameter was determined experientially. Although there were some attempts [3] to select the optimal parameter of a kernel, they are restricted for specific algorithm.

As described above, many popular dimensionality reduction algorithms can be understood and explained in a unified formulation, that is, they provide the low dimensional feature space preserving the adjacency relationship of some characteristic graph. Moreover, a graph characterizes the geometry structure of the data set to some extent; thus, a kernel taking into account the geometric structure of the data has the potential to im-

[1] The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region.

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

prove the performance of the kernel-based dimensionality reduction algorithm. In the following section, we will introduce a kernel design procedure to automatically construct a Locality-Adaptive-Kernel aiming at combining a set of local kernels into a global one, and then discuss the characteristics of this new kernel.

2. Locality-Adaptive-Kernel

A kernel implies a mapping function which presents a new distribution for the input data set [7]. This new distribution may change the geometrical properties of the data, *e.g.* the principal component directions, and consequently may degrade the algorithm performance. To overcome this issue, we propose a new procedure to automatically design adaptive kernel for specific data set by taking into the geometry structure of the data set and the kernel is defined in a locality manner. The intuition of the new kernel is to be locally adaptive.

2.1. Locality-Adaptive-Kernel Construction

The basic idea of LAKE is to reside multiple kernels on the data set, and then adaptive it locally, finally combine them for similarity measure. The procedure to construct the kernel consists of three steps: 1) reside multiple kernels on the data space; 2) adapt the local kernels; and 3) merge these kernels into a global one.

Reside multiple kernels. In this work, we apply the Cluster-Balanced K-Means clustering algorithm [1] owing to its computational efficiency and the ability to provide balanced clustering result.

Assume that the final clustering results of a given data set are $C=\left[c_1,c_2,...,c_N\right]$, $c_i\in\{1,2,...,K\}$ and K is the number of the clusters. And the conditional probability of the cluster c with given data x, $p(c\mid x)$, can be obtained using a simple formulation:

$$p(c\mid x)=f_x^c/\sum\nolimits_{j=1}^{K}f_x^j\ , \tag{4}$$

where $f_x^c=\exp\{-\alpha^c(x)\}$, $\alpha^c(x)$ is the *activity signal* of the data for cluster c and conventionally $\alpha^c(x)$ is set as the multiply of the normalized distance from the data to cluster center. For ease of representation, in the following, we set $p_x^c=p(c\mid x)$.

In each cluster, PCA is conducted for dimensionality reduction. In all our experiments, we determine the PCA dimensions by retaining 99% information in the sense of reconstruction error. Denote $W_{pca}^c\in\mathbb{R}^{m\times m_c}$ as the principal component matrix within cluster c and $\overline{x}^c$ is the mean of the samples belonging to cluster c. Then, for cluster c, the data x is transformed into

$$x^c=W_{pca}^c{}'(x-\overline{x}^c)\qquad c=1,\ldots,K\ . \tag{5}$$

Adapt local kernels. A kernel presents a new similarity measure for the data space, yet a global kernel maybe can not well characterize the geometry-based similarity due to the possibly complex global geometry structure. Therefore, we propose to design specific kernel in different local clusters/patches; and within the c-th cluster, the local kernel is defined based on the low dimensional representation from the local PCA, that is,

$$K^c(x,y)=<\phi^c(x^c),\phi^c(y^c)>+1\qquad c=1,\ldots,K\ . \tag{6}$$

Here the number *one* is for the convenience of seamless merging of difference local kernels as analyzed later for special linear local kernel in section 2.2.

Therefore, the local kernel well characterizes the local distribution property by utilizing the dimension reduced representations from the local PCA for specific clusters. Moreover, the kernel parameter for the local kernel can be further locally adapted. For example, for the Gaussian kernel, it is generally acceptable that the data standard deviation is a stable value for the kernel parameter in most cases, thus we can adapt the kernel parameter within a local cluster.

Merge different local kernels. The final similarity of a data pair can be computed as the sum of the conditional cluster probability weighted local kernel outputs

$$k(x,y)=\sum\nolimits_{c=1}^{K}p_x^c p_y^c(<\phi^c(x^c),\phi^c(y^c)>+1)\ . \tag{7}$$

The final kernel is called Locality-Adaptive-Kernel since for each cluster, different local map function that sufficiently considers the local data distribution property will be adaptively selected.

2.2. Justification and Analysis

Justification. *Marginalized Kernel* [8] was proposed to define a kernel between two visible variables x,x' with a hidden variable $h\in H$, where H is a finite set. Let $z=(x,h)$, the marginalized kernel is derived by taking expectation with respect to the hidden variable

$$K(x,x')=\sum\nolimits_{h\in H}\sum\nolimits_{h'\in H}p(h\mid x)p(h'\mid x')K_z(z,z')\ , \tag{8}$$

in which the joint kernel $K_z(z,z')$ is designed between two combined variables z and z'.

From the definition of the Locality-Adaptive-Kernel in (7), we find it is a specialized marginalized kernel with the hidden variable being the cluster label, that is $h=c$, $p(h\mid x)=p_x^c$, $H=\{1,2,...,K\}$ and the corresponding joint kernels are defined as

$$K_z(z,z')=\delta_{cc'}(<\phi^c(W_{pca}^c{}'x),\phi^c(W_{pca}^c{}'x)>+1)\ , \tag{9}$$

where $\delta_{cc'}=1$ if $c'=c$; 0, else.

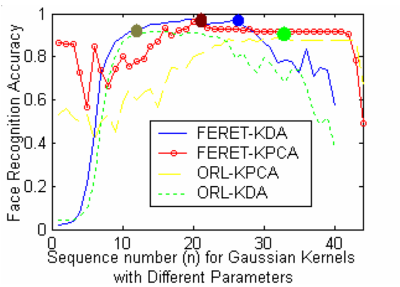


Figure 1. Face recognition accuracy distribution with respect to the Gaussian kernel parameter. Note that n-th kernel is defined as $k_n(x,y)=\exp\{-\|x-y\|^2/2\delta_n^2\}$ with $\delta_n^2=2^{n/2.5-8}\delta^2$ where δ is the standard derivation of the training samples. The solid circles are the kernels with maximum recognition accuracies.

Semantic for LAKE with linear local kernels. When the local kernels of LAKE are linear (LAKE/L)

$$K^c(x,y)=x'W_{PCA}^c{}'W_{PCA}^c y+1 \quad c=1,\ldots,K\text{ ,} \tag{10}$$

the kernel function has explicit representation as

$$\phi(x)=\left[\begin{pmatrix} p_x^c W_{PCA}^c{}'x \\ p_x^c \end{pmatrix}\downarrow\right]. \tag{11}$$

That is, the original feature space is transformed into a finite dimensional space, thus we can directly compute $w=\sum_i \alpha_i \phi(x_i)$ in (2) instead of to compute α as in (3). Let $Z=[\phi(x_1),\phi(x_2),...,\phi(x_N)]$, $\alpha' KLK\alpha = w'ZLZ'w$, and as in (11), we can rearrange w as

$$w = Z\alpha = \left[\begin{pmatrix} w^c \\ t^c \end{pmatrix}\downarrow\right]. \tag{12}$$

Here $w^c \in \mathrm{R}^{m_c}$ represents the projection direction in the c-th cluster; and t^c is a scale number presents a translation within the c-th cluster. If there is only one cluster, the derived solution is the same as the direct linear projection method defined in (2). For multiple clusters, there will be specific projection direction for each cluster; meanwhile there is a translation parameter within each cluster to make different clusters consistent in the marginal areas.

3. Experiments

In this section, we present three sets of experiments to discuss the kernel selection problem and evaluate the effectiveness of LAKE for both unsupervised and supervised problems. All the images were aligned by fixing the locations of two eyes. Histogram equilibrium was applied as the preprocessing step. The nearest neighbor method is applied for final classification.

3.1. Kernel Selection Issue

We conducted a series of experiments to evaluate the sensitivity of the kernel parameter selection for the classification problems. To this end, the face database FERET [2] and ORL [2] were applied for evaluation; and Gaussian-Kernel was used owing to its popularity. On FERET, seventy persons are used, and four images were randomly selected for training and the other two images for testing. On the ORL database, there are forty persons and ten images for each person; four of them are used for training and the other for testing.

Figure 1 plots the face recognition accuracies for kernel-based algorithms with different kernel parameters. From these comparative results, we can easily observe that when the algorithm type or data set is changed, the optimal kernel parameters will also change; and for the parameter where one algorithm is optimal, the results of other algorithms may be extremely bad, such as for the optimal parameter of case ORL-KDA, the accuracy of ORL-KPCA is very low. These experiments show that the kernel selection is necessary and very important.

3.2. A Toy Problem: Principal Curve Learning

Previous work [6] has shown that KPCA can capture nonlinear structure in the data. In this experiment, we compare the linear PCA with the Kernel-PCA using four different kernels: Gaussian, polynomial, inverse multiquadric kernel and our proposed LAKE/L. The data set are sampled from a structure integrating the symbol “V” and a horizontal line, *i.e.* with both nonlinear and linear structures. Figure 2 demonstrates the distributions of the data projections to the first principal component of KPCA and the lines are the contours of the projection values. It is evident that KPCA algorithm with LAKE/L best captures the intrinsic nonlinear structure of the data set. PCA cannot capture the nonlinear structure of the data as the linear property.

3.3. Face Recognition with LAKE

The effectiveness of the LAKE and its linearly simplified LAKE/L with linear local kernel is evaluated on face recognition problem based on the benchmark database CMU PIE [2] and our own database. The used local kernels in LAKE are Gaussian kernels and the parameters were adaptively set as the standard deviation of the samples belonging to the corresponding clusters. For fair comparison, 10 different parameters are tested and the best result is reported. For the Polynomial kernel, we test the parameters from 3 to 10.

CMU PIE. All the images with 21 different illumination conditions at pose-37 and 9 of all 68 persons were used. Ten images each person are randomly selected for model training and the other images are used for testing. The results in Figure 3 demonstrate that the LAKE based KPCA and KDA outperform all the other corre-

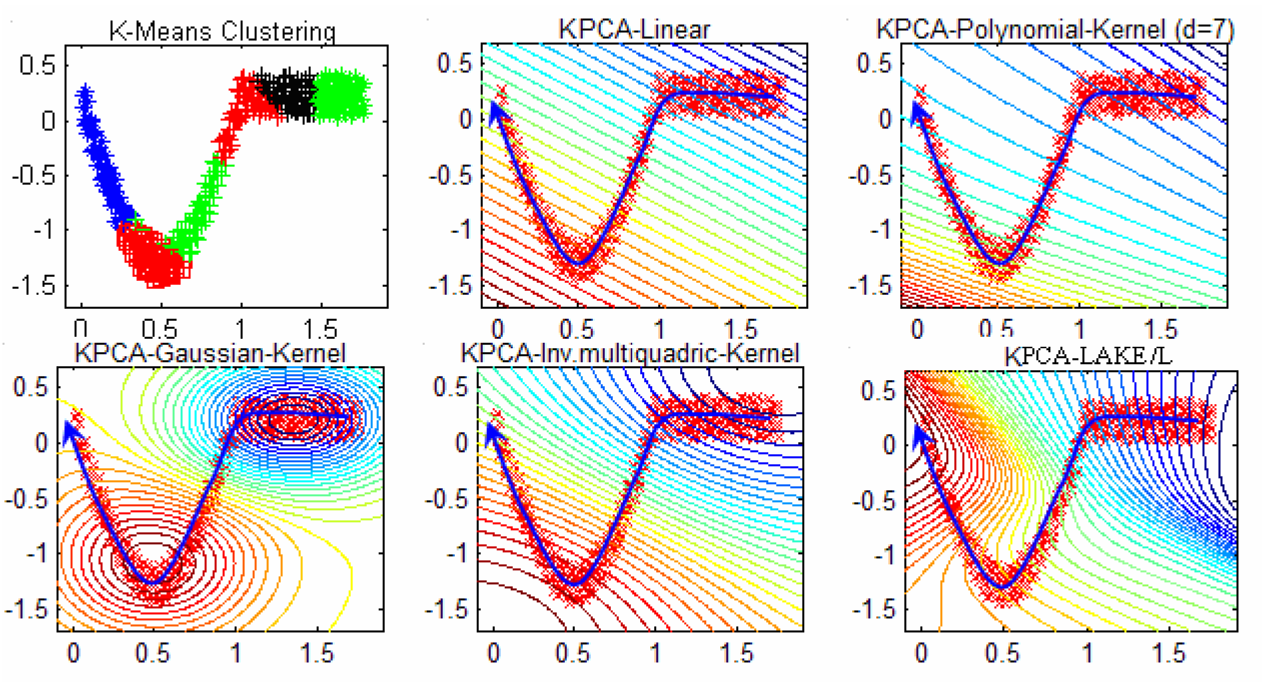
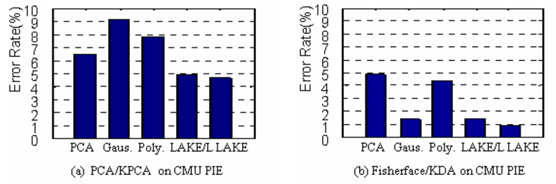


Figure 2. A toy problem to illustrate the capability of KPCA+LAKE/L to capture the principal curve of the nonlinearly distributed data. From left to right and top to bottom are the contour line image of the projection to the first principal component direction of PCA and KPCA with different types of kernels. Note that contour value at one point is computed by projecting its coordinates to the first principal direction. The blue curve with arrow means the desired principal curve direction; and for a good result, the contour line should be orthogonal to the blue line and the projection value should change monotonously along the direction of the blue line. Therefore, KPCA with LAKE/L is the best.

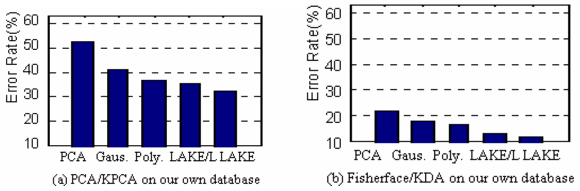
sponding linear and kernel based algorithms. Meanwhile the kernel sensitiveness issue is also very clear in these experiments, and the kernel based KPCA algorithm is even worse in accuracy than PCA.



(a) PCA/KPCA on CMU PIE

(b) Fisherface/KDA on CMU PIE

Figure 3. Face recognition error rates at CMU PIE database compared between (a) PCA/KPCA and b) Fisherface/KDA. The kernels include Gaussian kernel, Polynomial kernel, LAKE with linear local kernels (LAKE/L) and LAKE.



(a) PCA/KPCA on our own database

(b) Fisherface/KDA on our own database

Figure 4. Face recognition error rates on our own database compared between (a) PCA/KPCA and b) Fisherface/KDA. The kernels include Gaussian kernel, Polynomial kernel, LAKE with linear local kernels (LAKE/L) and LAKE.

Our Own Database. It contains 12 individuals and was captured in two different sessions with different backgrounds and illuminations and all the faces are frontal. In these experiments, 32 images for each person in session one were used for training; and the other session with 64 images each person were used for testing. The comparative experimental results plotted in Fig. 4 reveal that the algorithm with tailored LAKE is consistently superior to others in both KPCA and KDA.

References

[1] D. Cheung, S. Lee and Y. Xiao, "Effect of Data Skewness and Workload Balance in Parallel Data Mining", IEEE Transaction on Knowledge and Data Engineering, v14, pp. 498-513, 2002.

[2] http://www.face-rec.org/databases/

[3] J. Huang, C. Pong, W. Chen and J. Lai, "Kernel subspace LDA with optimized kernel parameters on face recognition", Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004.

[4] I. Joliffe, "Principal Component Analysis", Springer-Verlag, New York, 1986.

[5] H. Lodhi, C. Saunders, J. Taylor, N. Cristianini, C. Watkins, "Text classification using string Kernels", The Journal of Machine Learning Research, V2, pp. 419-444, March 2002.

[6] K. Mtiller, S. Mika, G. Riitsch,, K. Tsuda, B. Sch61kopf, "An Introduction to kernel-based learning algorithms", IEEE Transactions on Neural Networks, v12, pp.181 201, 2001.

[7] C. Sauders, J. Talor, A. Vinokourov, "String Kernels, Fisher Kernels and Finite State Automata", *Advances in Neural Information Processing Systems 16, 2003.*

[8] K. Tsuda, T. Kin and K. Asai, "Marginal Kernels for Biological Sequences", BIOINFORMATICS, Vol. 1 no. 1 2002, pp 1–8.

[9] S. Yan, D. Xu, B. Zhang, and H. Zhang. "Graph embedding: A general framework for dimensionality reduction", in CVPR05.