LEARNING BASED ON KERNEL DISCRIMINANT-EM ALGORITHM FOR IMAGE CLASSIFICATION

Qi Tian, Jie Yu, Ying Wu[1]*, Thomas S. Huang*[2]

Department of Computer Science, University of Texas at San Antonio, TX 78249
[1]Department of Electrical and Computer Engineering, Northwestern University, IL 60208
[2]Beckman Institute, University of Illinois, Urbana, IL 61801

ABSTRACT

In image classification and other learning-based object recognition tasks, it is often tedious and expensive to label large training data sets. Discriminant-EM (DEM) proposed a semi-supervised learning framework which takes both labeled and unlabeled data to learn classifiers. This paper extends the linear D-EM to nonlinear kernel algorithm, KDEM and evaluates KDEM on both benchmark image databases and synthetic data. Various comparisons with other state-of-the-art learning techniques are investigated.

1. INTRODUCTION

Content-based image retrieval (CBIR), a technique that uses visual content to search images from large-scale image database according to user's interest, has been an active and fast advancing research area since the 1990s.

One of the difficulties of CBIR is the gap between high-level semantics in human mind and low-level image features, due to the rich content but subjective concepts of an image [1]. The mapping between them would be nonlinear such that it is impractical to represent it explicitly. A promising approach to this problem is machine learning, by which the mapping could be learned through a set of examples. In our proposed approach, image retrieval is cast as a statistical learning problem.

The task of image retrieval is to find as many as possible "similar" images to the query images in a given database. The retrieval system acts as a classifier to divide the images in the database into two classes, relevant or irrelevant. In image retrieval, there are a limited number of labeled training samples given by the query and relevance feedback. Pure supervised learning from such a small training dataset will have poor generalization performance. If the learning classifier is over-trained on the small training dataset, *over-fitting* will probably occur.

This problem can be alleviated by *semi-supervised* or *self-supervised* learning techniques which take hybrid training datasets. For image retrieval, there are a large number of unlabeled images in the given database. Unlabeled data contain information about the joint distribution over features which can be used to help supervised learning. Discriminant-EM (DEM) [2] is a self-supervised learning algorithm for such purposes by taking a small set of labeled data with a large set of unlabeled data. The basic idea is to learn discriminating features and the classifier simultaneously by inserting a multi-class linear discriminant step in the standard expectation-maximization (EM) [3] iteration loop. DEM makes assumption that the probabilistic structure of data distribution in the lower dimensional discrimination space is simplified and could be captured by lower order Gaussian mixture.

Contrary to the traditional two-class, i.e., fisher discriminant analysis (FDA) and multi-class discriminant analysis (MDA) which treats every class equally when finding the optimal projection subspaces, Zhou and Huang in [4] proposed a biased discriminant analysis (BDA) which treats all positive, i.e., relevant, examples as one class and negative, i.e., irrelevant, examples as different classes. The intuition behind BDA is that "all positive examples are alike, each negative example is negative in its own way". Compared with the state-of-the-art methods such as Support Vector Machines (SVM) [5], BDA and its kernel version (KBDA) [4] outperform SVM when the size of negative examples is small (<20).

However, one drawback of BDA is its ignorance of unlabeled data in semi-supervised learning. Unlabeled data could improve the classification under the assumption that nearby data are to be generated by the same class. In the past years there has been a growing interest in the use of unlabeled data for enhancing classification accuracy in supervised learning such as text classification [6], face expression recognition [7], and image retrieval [2, 8]. Recent work [7, 9] shows that unlabeled data can improve or degrade the classification performance depending on whether the model assumption is correct, and also on whether the labeled and unlabeled data has the same distribution.

0-7803-8484-9/04/$20.00 ©2004 IEEE

2. NONLINEAR DISCRIMINANT ANALYSIS

Preliminary results of applying DEM for CBIR have been shown in [2]. Because the discrimination step in DEM is linear, it has difficulty handling data sets which are not linearly separable. Image distribution is likely, e.g., mixture of Gaussians, which is highly non-linear-separable. In this paper, we generalize the DEM from linear setting to a nonlinear one. We first map the data **x** via a nonlinear mapping ϕ into some high, or even infinite dimensional feature space *F* and then apply linear DEM in feature space *F*. To avoid working with the mapped data explicitly (being impossible if *F* is of an infinite dimension), we will adopt the well-known *kernel trick* [10]. The kernel functions $k(\mathrm{x},\mathrm{z})$ compute a dot product in a feature space *F*: $k(\mathrm{x},\mathrm{z}) = (\phi(\mathrm{x}) \cdot \phi(\mathrm{z}))$. Formulating the algorithms in *F* using ϕ only in dot products, we can replace any occurrence of a dot product by the kernel function *k*, which amounts to performing the same *linear* algorithm as before, but *implicitly* in a kernel feature space *F*. Kernel principle is quickly gaining attention in CBIR in recent years [8, 4].

2.1. Linear Multiple Discriminant Analysis

Multiple discriminant analysis (MDA) is a natural generalization of Fisher's linear discriminant analysis (FDA) for multiple classes [3]. The goal is to find a lower dimensional space in which the ratio of between-class scatter over within-class scatter is maximized.

$$W_{opt} = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} \qquad (1)$$

where

$$S_B = \sum_{j=1}^{C} N_j \cdot (m_j - m)(m_j - m)^T \qquad (2)$$

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (x_i^{(j)} - m_j)(x_i^{(j)} - m_j)^T \qquad (3)$$

we use $\{x_i^{(j)}, i = 1,\ldots,N_j\}, j = 1,\ldots,C$ ($C = 2$ for FDA) to denote the feature vectors of training samples. *C* is the number of classes, N_j the number of the samples of the *j*th class, $x_i^{(j)}$ is the *i*th sample from the *j*th class, and m_j is mean vector of the *j*th class, and *m* grand mean of all examples. $W_{opt} = [w_1, w_{2,} \cdots, w_{C-1}]$ will contain in its columns *C-1* eigenvectors corresponding to *C-1* eigenvalues, i.e., $S_b w_i = \lambda_i S_w w_i$. [3]

2.2. Kernel Discriminant Analysis

We will have the similar formulae Eqs. (1–3) for kernel-based approaches except now MDA is performed in the *feature space F* and *x* is replaced by $\phi(x)$.

In general, there is no other way to express the solution $W_{opt} \in F$, either because *F* is too high or infinite dimension, or because we do not even know the actual *feature space* connected to a certain kernel. Referring to [10], we know that any column of the solution W_{opt}, must lie in the span of all training samples in *F*, i.e., $w_i \in F$. Thus for some expansion coefficients $\vec{\alpha} = [\alpha_1, \cdots, \alpha_N]^T$,

$$w_i = \sum_{k=1}^{N} \alpha_k \phi(x_k) \;\; i = 1,\ldots,N \quad (4)$$

The use of the above expansion makes things tractable and with some reformulation [10], it is now a quotient in terms of expansion coefficients $\vec{\alpha}$, and not in terms of $w_i \in F$. Therefore, Kernel MDA becomes

$$A_{opt} = \arg\max_{A} \frac{|A^T K_B A|}{|A^T K_W A|} \qquad (5)$$

where $A = [\vec{\alpha}_1, \cdots, \vec{\alpha}_{C-1}]$, *C* is the total number of classes , *N* the size of training samples, and K_B and K_W are $N \times N$ matrices which require only kernel computations on the training samples [10].

2.3. Biased Discriminant Analysis

BDA [4] differs from regular MDA defined in (1)-(3) in a modification on the computation of scatter matrices S_B and S_W. They are replaced by $S_{N \to P}$ and S_P, respectively, where $S_{N \to P}$ is the scatter matrix between the negative examples towards the centroid of the positive examples, and S_P is the scatter matrix within the positive examples. $N \to P$ indicates the asymmetric property of this approach, i.e., the user's biased opinion towards the positive class, thus the name of biased discriminant analysis (BDA) [4].

3. KERNEL D-EM ALGORITHM

Kernel DEM (KDEM) is a generalization of DEM [2] in which instead of a simple linear transformation to project the image into discriminant subspaces, the image data is first projected nonlinearly into a high dimensional feature space *F* where the data is better separately linearly. Then the linear D-EM is applied in the feature space.

Empirical observations suggest that the transformed image data often approximates a Gaussian in discriminant subspace, and so in our implementation, we use low-order Gaussian mixture to model the transformed data. KDEM loops between three steps until some appropriate convergence criterion:

- E-step: set $\hat{Z}^{(k+1)} = E[Z \mid D; \hat{\Theta}^{(k)}]$
- D-step: set $A_{opt}^{k+1} = \arg\max_{A} \frac{|A^T K_B A|}{|A^T K_W A|}$, and project a data point **x** to a linear subspace of feature space F.
- M-Step: set $\hat{\Theta}^{(k+1)} = \arg\max_{\Theta} p(\Theta \mid D; \hat{Z}^{(k+1)})$

The same notation is used as in [2]. The E-step gives probabilistic labels to unlabeled data which are then used by the D-step to separate the data.

4. EXPERIMENTS

In this section, we compare KMDA and KDEM with other supervised learning techniques on both benchmark image datasets and synthetic data for image classification.

4.1. Benchmark Test

Kernel functions that have been proven useful are e.g., Gaussian RBF, $k(\mathrm{x},\mathrm{z}) = \exp(-\|\mathrm{x}-\mathrm{z}\|^2 / c)$, or polynomial kernels, $k(\mathrm{x},\mathrm{z}) = (\mathrm{x}\cdot\mathrm{z})^{\mathrm{d}}$, for some positive constants $c \in R$ and $d \in N$, respectively [10].

Several benchmark data sets[1] were used in the experiments. For comparison, KMDA is compared to a single RBF classifier (RBF), AdaBoost, a support vector machine (SVM), and the kernel Fisher discriminant (KFD) and the linear MDA on the benchmark dataset [11]. RBF kernels were used in all kernel-based algorithms. #KV is the number of kernel vectors, i.e., the size of the training samples. Two sampling schemes of selecting the training samples are PCA-based and by an iterative procedure.

Table 1: The average test error (%) and standard deviation

Benchmark	Banana	Breast-Cancer	Heart
RBF	10.8±0.06	27.6±0.47	17.6±0.33
AdaBoost	12.3±0.07	30.4±0.47	20.3±0.34
SVM	11.5±0.07	26.0±0.47	16.0±0.33
KFD	10.8±0.05	25.8±0.48	16.1±0.34
MDA	38.43±2.5	28.57±1.37	20.1±1.43
KMDA-*pca*	10.7±0.25	27.5±0.47	16.5±0.32
KMDA-*iter.*	10.8±0.56	26.3±0.48	16.1±0.33
# KV	120	40	20

Table 1 shows that the proposed KMDA achieves comparable performance as other state-of-the-art techniques over different training datasets. Table 1 also shows that KMDA performs better than linear MDA.

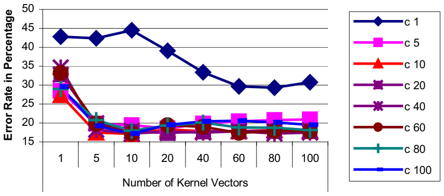

Figure 1: The average error rate for KMDA on Heart data

It should be noted that a proper selection of kernel function is critical and till now there is no good method on how to choose a kernel function and its parameter. Figure 1 shows the error rate of KMDA with RBF kernel under different degrees c and number of kernel vectors used on Heart data. By experimental observation, we found that 10 for c and 20 for #Kernel Vectors gives almost the best performance. Similar results are also obtained for Breast-Cancer data and Banana data. Thus this setting will be used in the rest of our experiments.

4.2. KDEM vs. KBDA for Image Classification

As mentioned in Section 1, biased discriminant analysis (BDA) and kernel BDA [4] have achieved great success in CBIR when the number of training samples is small (<20). BDA differs from traditional MDA in that it tends to cluster all the positive samples and scatter all the negative samples from the centroid of positive examples. This works very well with relatively small training samples. However, BDA ignores unlabeled images and is biased tuned toward the centroid of the positive examples. It is effective only if these positive examples are the *most-informative* images, i.e., images close to the classification boundary, instead of *most-positive* images, i.e., images far away from the classification boundary. Optimal transformation found based on the *most-positive* images won't improve classification for images on the boundary.

In this second experiment, Kernel DEM (KDEM) is compared with Kernel BDA (KBDA) on both real image databases and synthetic data. The real image database consists of the face images from MIT facial image database[2] (2358 images) and non-face images from Corel database[3] (2598 images).

For training sets, the face images are randomly selected from MIT database with fixed size 100, and non-

[1] The data sets are obtained from http://mlg.anu.edu.au/~raetsch/

[2] MIT facial database is obtained from http://www.ai.mit.edu/projects/cbcl/software-datasets/FaceData2.html

[3] Corel database is benchmark dataset used in CBIR

face images are randomly selected from Corel database with varying size from 5 to 200. The testing set consists of 200 randomly images (100 faces and 100 non-faces) from two databases. The image features used are extracted from size-reduced images and feature dimension is 256.

Figure 2 shows the average classification error rate in percentage for KDEM and KBDA under the same RBF for face and non-face classification.

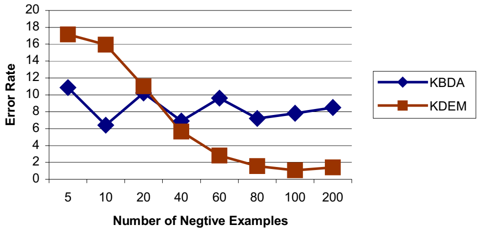

Figure 2: Comparison of KDEM and KBDA for face and non-face classification

Figure 2 shows that KDEM performs better than KBDA when more negative examples are provided while KBDA performs better when the size of negative examples is small (<20). This agrees with our expectation.

To further examine the effect of increasing size of negative classes (in Fig. 2, all negative examples are possible from the same class) on the performance of KDEM and KBDA, we investigated on synthetic data for which we have more controls over data distribution.

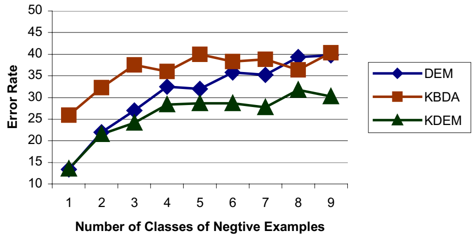

Figure 3: Comparison of KDEM, DEM and KBDA on synthetic data

Figure 3 shows that with the increasing size of negative classes from 1 to 9, KDEM always performs better than KBDA. Even linear DEM works better than KBDA. The reason is that learning in both DEM and KDEM is on hybrid data, while only a small set of labeled data is used in KBDA. This shows proper incorporation of unlabeled data does improve classification.

5. CONCLUSIONS

We presented a semi-supervised discriminant analysis technique, Kernel DEM, which employs both labeled and unlabeled data in training. Kernel DEM not only out-performs linear DEM on the benchmark data set but also out-performs Kernel BDA [4] on both real image database and synthetic data.

Our future work include (1) the selection of a representative subset of training samples to reduce the computation complexity of KDEM, (2) further connection between KDEM and support vector machines, and (3) incorporation of unlabeled data in BDA.

6. ACKNOWLEDGEMENT

This work was supported in part by National Science Foundation Grant, EIA 99-75019.

7. REFERENCES

[1] A. Smeulder, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Trans. on PAMI*, pp. 1349-1380, Dec. 2000.

[2] Y. Wu, Q. Tian, and T. S. Huang, “Discriminant EM Algorithm with Application to Image Retrieval,” *IEEE Int’l Conf. CVPR 2000*, South Carolina, June 13-15, 2000.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2001.

[4] X. Zhou, and T.S. Huang, “Small Sample Learning during Multimedia Retrieval Using BiasMap,” *IEEE Int’l Conf. CVPR 2001*, Hawaii, December 2001.

[5] V. Vapnik, *The nature of statistical learning theory*, second edition, Springer-Verlag, 2000.

[6] K. Nigram, A. K. McCallum, S. Thrun, and T. M. Mitchell, “Text Classification from Labeled and Unlabeled Documents Using EM,” *Machine Learning*, 39(2/3):103-134, 2000.

[7] Cohen, N. Sebe, F. G. Cozman, M. C. Cirelo, and T. S. Huang, “Learning Bayesian Network Classifiers for Facial Expression Recognition with Both Labeled and Unlabeled Data,” *IEEE Int’l Conf. CVPR 2003*, Madison, WI, June 2003.

[8] L. Wang, K. L. Chan, and Z. Zhang, “Bootstrapping SVM Active Learning by Incorporating Unlabelled Images for Image Retrieval,” *IEEE Int’l CVPR 2003*, Madison, WI, June 2003.

[9] Q. Tian, J. Yu, Q. Xue, and N. Sebe, “A New Analysis of the Value of Unlabeled Data in Semi-Supervised Learning for Image Retrieval,” submitted to *IEEE Int’l Conf. Multimedia and Expo (ICME’2004)*, Taipei, Taiwan, June 27-30, 2004.

[10] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Mass: MIT Press, 2002.

[11] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K. Müller, “Fisher Discriminant Analysis with Kernels,” *IEEE Workshop on Neural Networks for Signal Processing*, 1999.