Discriminant Analysis in Correlation Similarity Measure Space

Yong Ma ma@ari.ncl.omron.co.jp
Shihong Lao lao@ari.ncl.omron.co.jp
Erina Takikawa erinat@ari.ncl.omron.co.jp
Masato Kawade kawade@ari.ncl.omron.co.jp
Sensing & Control Lab., Omron Corporation, Kyoto, 619-0283, Japan

Abstract

Correlation is one of the most widely used similarity measures in machine learning like Euclidean and Mahalanobis distances. However, compared with proposed numerous discriminant learning algorithms in distance metric space, only a very little work has been conducted on this topic using correlation similarity measure. In this paper, we propose a novel discriminant learning algorithm in correlation measure space, Correlation Discriminant Analysis (CDA). In this framework, based on the definitions of within-class correlation and between-class correlation, the optimum transformation can be sought for to maximize the difference between them, which is in accordance with good classification performance empirically. Under different cases of the transformation, different implementations of the algorithm are given. Extensive empirical evaluations of CDA demonstrate its advantage over alternative methods.

1. Introduction

Correlation (also termed as normalized correlation, correlation coefficient, Pearson correlation, or cosine similarity, hereafter correlation for simplicity) is a widely used measure to describe the similarity s between two vectors, $\mathbf{u}$ and $\mathbf{v}$, in pattern classification and signal processing problems like Euclidean distance and Mahalanobis distance.

$$s = \frac{\mathbf{u}^T \mathbf{v}}{\sqrt{\mathbf{u}^T \mathbf{u} \mathbf{v}^T \mathbf{v}}}, \quad (1)$$

Appearing in *Proceedings of the 24dthInternational Conference on Machine Learning*, Corvallis, OR, 2007.

For example, in face recognition, the framework (Kittler, 2000) combining principal component analysis (PCA), linear discriminant analysis (LDA) (Fukunaga, 1990), and nearest neighbor (NN) based on correlation measure achieved obviously superior performance to other frameworks, such as the combination of PCA, LDA, and NN based on Euclidean metric or weighted Euclidean metric, on a large scale comparative evaluations such as BANCA (Kittler, 2000) etc. In gene expression analysis (Brown, 2000) and document categorization problems (Han, 2001; Peterson, 2005), correlation was found to be an effective measure. Besides these areas, in signal processing community, correlation and correlation-based filters (Kumar, 1986; Mahalanobis, 1987; Xie, 2005) were also widely used to detect the existence of some specific signals.

However, compared with the Euclidean or Mahalanobis distance, the correlation is not a metric due to that it cannot satisfy the non-negativity and the triangle inequality. And different from the situation that there are so many discriminant learning algorithms in distance metric space such as LDA, relevant component analysis (RCA) (Hillel, 2005), distance metric learning with side-information (DMLSI) (Xing, 2003) etc, there is only a very little discriminant learning work based on correlation similarity measure. To the best of our knowledge, this type of research includes canonical correlation analysis (CCA) (Hardoon, 2004), correlation filter design such as minimum average correlation energy (MACE) filter (Mahalanobis, 1987), etc. So it is very interesting to extend the discriminant learning research in correlation similarity space from a new way.

On the other hand, from the research (Belhumeur, 1997; Kittler, 2000; Martinez, 2005) on face recognition about the framework of PCA, LDA and correlation-based NN classifier, we knew that PCA was mainly used to reduce feature dimension and solve the singularity problem of LDA due to small sample size, LDA was used to find the optimum discriminant projections to make the Euclidean metric-based within-class scatter denser and between-class more scattered over training set, and the correlation-based NN is used to find the best match in the gallery set for a probe face according to their correlation similarity during evaluation stage. We could notice that although

correlation-based NN does not match the Euclidean metric-based LDA very much, it can achieve better performance than Euclidean metric-based NN. Naturally, we would wonder how about the performance if we directly substitute correlation measure for Euclidean metric during the discriminant learning stage.

In this paper, we will explore the above questions and propose a novel discriminant learning algorithm in correlation measure space, Correlation Discriminant Analysis (CDA). CDA combines the discriminant learning with correlation measure together, seeks for the optimum transformation to maximizing the difference between within-class correlation and between-class correlation. This difference can be empirically considered as a criterion for the classification performance. In extensive experimental studies on publicly available databases and real applications we show that the proposed learning algorithm can consistently improve the performance of original correlation-based classification further, and achieve the comparable or even better performance with alternatives.

The rest of the paper is organized as follows. In section 2, we will briefly introduce the related work and compare them with CDA. In section 3, the details of CDA algorithm and optimization methods under different situation are given. In section 4, thorough experiments are made among CDA and other supervised discriminant learning algorithms. Finally, section 5 gives the conclusions.

2. Related Work

In this section, we will first review the most popular discriminant learning techniques in distance metric space, such as LDA, relevant component analysis (RCA) (Hillel, 2005), distance metric learning with side-information (DMLSI) (Xing, 2003), and then introduce some related work in similarity measure domain.

As we know, LDA find the best projection directions on which the within-class scatter matrix is minimized while the between-class scatter matrix is maximized under the Fisher criterion (Fukunaga, 1990). To solve its singularity problem when the number of samples is much smaller than the feature dimension, PCA is often used prior to LDA; Not considering the maximization of between-class scatter matrix which stands for the negative equivalence constraints, RCA is only concerned with the minimization of sum of squared Mahalanobis distances of positive equivalence constraints; under this criterion, the optimum transformation is the inverse of within-class scatter matrix. Different from RCA, DMLSI tries to learn an optimum Mahalanobis metric under both positive and negative constraints, in which, the sum of squared Mahalanobis distances from positive constraints is minimized given the sum of Mahalanobis distances from negative constraints being over the constant value.

LDA, RCA, and DMLSI are only concerned with global Euclidean or Mahalanobis structure. Sometimes in the context of sparse data such as face recognition, these distance metrics were not optimal (Kittler, 2000). Several attempts have made to take the advantage of non-linear structure to provide more discriminatory information while keeps LDA structure untouched. In the Smith et al.'s (2006) work, an angular LDA (ALDA) was presented. In their method, the LDA transformation was first applied to the probe vector, and then a non-linear transformation was used to project the probe LDA vector into the probe ALDA vector whose *i*th component was the angle between the probe LDA vector and the *i*th LDA component axis. Standard classifier can be designed in this ALDA space. This work still separated the optimization procedure from the metric or measure learning.

Correlation similarity measure was also widely utilized together with *k*-nearest neighbor algorithms to replace traditional Euclidean/Mahalanobis distance. Han et al., (2001) proposed a weight adjusted *k*-nearest neighbor classification method to learn the weights for different features in correlation similarity; Peterson et al., (2005) presented a genetic optimization algorithm to solve the above problem to optimize both the feature weights and offsets to features according to a simple empirical criterion. The main drawback of these methods is that the transformation is only in diagonal form and the weight of every feature is assumed to be uncorrelated.

Canonical correlation analysis (CCA) (Hardoon, 2004) might be the most successful dicriminant learning methods dedicated to correlation measure till now. It can be viewed as the problem of finding basis vectors for two sets of variables such that the correlations between the projections of the variables are mutually maximized. So if one set of variables is taken as class labels, CCA can be explicitly utilized to realize a supervised linear feature extraction and subsequent classification (Loog, 2004). The similar method is also straightforwardly used in partial least squares (PLS) (Barker 2003) for classification in which it also involves the correlation between two sets of variables with different constraints.

There are two main problems for CCA and PLS to be applied to classification problems. The first one is that different encoding modes for class labels can result in different performances. Johansson (2001) proposed the *one-of-c* label encoding to associate a sample with its corresponding class label; Sun (2007) proposed a soft label encoding method based on fuzzy *k*-nearest neighbor method. The second problem is that they are not suitable for open-set verification applications (Phillips, 2000). In open-set verification, the classes in gallery or probe set might never appear in the training set (for example, the verification of a person and his claimed passport photo, which both never appear in the training face database). However, the transformations learned by CCA or PLS during training stage are class-specific and related to the

classes which appear in the training set. When verifies two new samples, the system needs to first find projections separately for two samples online to discriminate them from training faces and then decide whether they belong to the same class or not. That is unfavorable for real applications.

The correlation-based filter design methods have the similar problem when used in the open-set verification applications and need to design the class-specific correlation filter. This type of methods include the minimum variance synthetic discriminant function filter (Kumar, 1986), which minimizes the correlation output noise variance and typically suppresses high frequencies in order to achieve noise tolerance, the minimum average correlation energy (MACE) filter (Mahalanobis, 1987), which minimizes the average correlation output energy and emphasizes high spatial frequencies in order to produce sharp correlation peaks, and the optimal tradeoff filter (Refregier,1990) which is designed to balance these two criteria by minimizing a weighted criterion.

3. Correlation Discriminant Analysis

Correlation Discriminant Analysis (CDA) is a method that seeks a global linear transformation to maximize the correlation of samples from the same class and minimize the correlation of samples from different classes in the transformed space. Compared with LDA, RCA and DMLSI, CDA combines the similarity measure learning with the supervised discriminant learning in a very simple way to explore the nonlinear information instead of only linear information. And compared with CCA, PLS and correlation-based filter design methods, CDA only cares the correlation of samples between-class and within-class, instead of the correlation between samples and their corresponding labels; the trained projection matrix is class-irrelevant. So the verification between two new samples which both do not appear in the training set can be easily conducted without modifying the trained projection matrix. The details of CDA are given as follows.

For general multi-class classification problems, given a training set $\{\mathbf{x}_i, t_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$ is the sample feature vector, t_i is the corresponding class label, the total class number is c and the jth class has n_j samples. Under a transformation $\mathbf{w} \in \mathbb{R}^{d \times d}$, the feature vector $\mathbf{x}$ is projected into $\mathbf{y} = \mathbf{w}\mathbf{x}$.[1] In the transformed space, the within-class correlation S_w, between-class correlation S_b and total correlation S_t over the training set can be defined separately as follows:

$$S_w = \frac{1}{N_w} \sum_{(i,j)|_{t_i=t_j}}^{n} \frac{\mathbf{y}_i^T \mathbf{y}_j}{\|\mathbf{y}_i\| \|\mathbf{y}_j\|} = \frac{1}{N_w} \sum_{(i,j)|_{t_i=t_j}}^{n} \frac{\mathbf{x}_i^T \mathbf{w}^T \mathbf{w} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^T \mathbf{w}^T \mathbf{w} \mathbf{x}_i \mathbf{x}_j^T \mathbf{w}^T \mathbf{w} \mathbf{x}_j}} \tag{2}$$

$$S_b = \frac{1}{N_b} \sum_{(i,j)|_{t_i \neq t_j}}^{n} \frac{\mathbf{y}_i^T \mathbf{y}_j}{\|\mathbf{y}_i\| \|\mathbf{y}_j\|} = \frac{1}{N_b} \sum_{(i,j)|_{t_i \neq t_j}}^{n} \frac{\mathbf{x}_i^T \mathbf{w}^T \mathbf{w} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^T \mathbf{w}^T \mathbf{w} \mathbf{x}_i \mathbf{x}_j^T \mathbf{w}^T \mathbf{w} \mathbf{x}_j}} \tag{3}$$

$$S_t = \frac{1}{n^2} \sum_{(i,j)}^{n} \frac{\mathbf{y}_i^T \mathbf{y}_j}{\|\mathbf{y}_i\| \|\mathbf{y}_j\|} = \frac{1}{n^2} \sum_{(i,j)}^{n} \frac{\mathbf{x}_i^T \mathbf{w}^T \mathbf{w} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^T \mathbf{w}^T \mathbf{w} \mathbf{x}_i \mathbf{x}_j^T \mathbf{w}^T \mathbf{w} \mathbf{x}_j}} \tag{4}$$

Here N_w and N_b are the numbers of sample pair which is from the same class or different classes. So

$$N_w + N_b = n^2, \tag{5}$$

$$N_w S_w + N_b S_b = n^2 S_t, \tag{6}$$

$$S_w - S_b = \frac{n^2}{N_b}(S_w - S_t). \tag{7}$$

Define $\mathbf{A} = \mathbf{w}^T \mathbf{w}$, then $\mathbf{A}$ should be symmetric and positive semi-definite, $\mathbf{A} \geq 0$.

Similar to LDA algorithm which maximizes the within-class scatter matrix and minimizes the between-class scatter matrix, a simple way of defining a criterion for the desired transformation is to demand the larger difference between S_w and S_b, which generally stands for better discriminant power to separate the sample pairs from the same class from the pairs from different classes. This gives the optimization problem:

$$\max_{\mathbf{A}} \quad S_w - S_b \quad \text{or} \quad \max_{\mathbf{A}} \quad S_w - S_t \tag{8}$$

$$\text{s.t.} \quad \mathbf{A} \geq 0. \tag{9}$$

This problem has an objective that is nonlinear and non-convex in the parameters $\mathbf{A}$. So the optimization problem is different from the convex optimization

[1] Note that, by putting the original dataset through a non-linear kernel function ϕ, non-linear transformation can also be learned.

problems of LDA, CCA and DMLSI. We also note that, while one might consider various criteria to (8), $S_w/(S_b+1)$ would not be a good choice mainly due to its much more complex optimization process.

3.1 The Case of Diagonal *A*

If we restrict $\mathbf{A}$ to be diagonal, this corresponds to learning a transformation in which the different axes are given different "weights" to the correlation similarity in the transformed space. We define $\mathbf{A} = \mathrm{diag}(\rho_1^2, \cdots, \rho_d^2)$ and $\mathbf{w} = \mathrm{diag}(\rho_1, \cdots, \rho_d)$. So the optimization of problem (8), (9) becomes the following unconstrained maximization problem

$$\begin{aligned} f(\mathbf{A}) &= f(\rho_1, \cdots, \rho_d) = S_w - S_t \\ &= \sum_{i=1}^{c} p_i \mathbf{m}_i^T \mathbf{m}_i - \mathbf{m}^T \mathbf{m} \end{aligned} \tag{10}$$

Here p_i is the percentage of the number of sample pairs from *i*th class in N_w; $\mathbf{m}_i$ and $\mathbf{m}$ are the mean value of *i*th class and total samples after projection and normalization;

$$p_i = \frac{n_i^2}{N_w}$$

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{k|_{l_i}} \frac{\mathbf{w}\mathbf{x}_k}{\left\| \mathbf{w}\mathbf{x}_k \right\|}$$

$$\mathbf{m} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{w}\mathbf{x}_i}{\left\| \mathbf{w}\mathbf{x}_i \right\|}$$

Using Newton-Raphson gradient-based optimization method, the above maximization problem can be solved.

$$\frac{\partial f}{\partial \rho_k} = 2 \sum_{i=1}^{c} p_i \mathbf{m}_i^T \frac{\partial(\mathbf{m}_i)}{\partial \rho_k} - 2\mathbf{m}^T \frac{\partial(\mathbf{m})}{\partial \rho_k} \tag{11}$$

In the above equation,

$$\frac{\partial(\mathbf{m})}{\partial \rho_k} = \frac{1}{n} B - \frac{1}{n} \rho_k \left(\frac{\mathbf{w}\mathbf{x}_1}{\left\| \mathbf{w}\mathbf{x}_1 \right\|} \quad , \cdots, \quad \frac{\mathbf{w}\mathbf{x}_n}{\left\| \mathbf{w}\mathbf{x}_n \right\|} \right) C \tag{12}$$

$$B = \left(0, \quad \cdots, \quad \sum_{i=1}^{n} \frac{x_{ik}}{\left\| \mathbf{w}\mathbf{x}_i \right\|} \quad , \cdots \quad , 0 \right)^T$$

$$C = \left(\frac{x_{1k}^2}{\left\| \mathbf{w}\mathbf{x}_1 \right\|} \quad , \cdots, \quad \frac{x_{nk}^2}{\left\| \mathbf{w}\mathbf{x}_n \right\|} \right)^T$$

Here the x_{ik} is the *k*-th component of sample $\mathbf{x}_i$. $\dfrac{\partial(\mathbf{m}_i)}{\partial \rho_k}$ can be computed in the similar way.

Because only the relative values of $\rho_1, \cdots, \rho_d$ affect the correlation value, we can set $\rho_1 = 1$ and only need to estimate $\rho_2, \cdots, \rho_d$.

3.2 The Case of Full *A*

More generally, in the case of learning a full matrix A, the constraint that $\mathbf{A} \geq 0$ becomes slightly difficult to enforce, and Newton's method often becomes prohibitively expensive (require $O(n^6)$ time to invert the Hessian of n^2 parameters). Similar to the optimization problem in DMLSI (Xing, 2003), using gradient descent and iterative projections (Rockafellar, 1970), we derive a different algorithm for this setting.

Initialize **A**;
Repeat
  Repeat
    $\mathbf{A} := \mathrm{argmin}_{\mathbf{A}'} \{ \left\| \mathbf{A}' - \mathbf{A} \right\|_F : \mathbf{A}' \in C \}$
  Until **A** converges
  $\mathbf{A} := \mathbf{A} + \alpha \nabla_{\mathbf{A}} f(\mathbf{A})$
Until convergence

Figure 1. Gradient ascent +Iterative projection algorithm for full CDA. Here $\left\| \bullet \right\|_F$ is the Frobenius norm defined on matrices ($\left\| M \right\|_F = (\sum_i \sum_j M_{ij}^2)^{1/2}$).

We will use a gradient ascent step on *f*(**A**) to optimize (8), followed by the method of iterative projections just to ensure that the constraints (9) hold. Specifically, we will repeatedly take a gradient step $\mathbf{A} := \mathbf{A} + \alpha \nabla_{\mathbf{A}} f(\mathbf{A})$, and then repeatedly project A into the sets $C = \{ \mathbf{A} : \mathbf{A} \geq 0 \}$. This gives the algorithm shown in Figure 1.

To calculate derivatives of $f(\mathbf{A})$ with respect to A_{kl} , which denotes the *k*-th row and *l*-th column component of A , we just need to compute the derivatives of the correlation of *i*-th and *j*-th samples $\dfrac{\partial s_{ij}}{\partial A_{kl}}$ and sum up all the derivatives between-class and with-in class separately:

$$s_{ij} = \frac{\mathbf{x}_i^T \mathbf{A} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^T \mathbf{A} \mathbf{x}_i} \sqrt{\mathbf{x}_j^T \mathbf{A} \mathbf{x}_j}} \tag{13}$$

If *k*=*l*,

Table 1. The comparison of classification accuracies on various UCI data sets (%)

Method	balance	glass	lenses	sonar	thyroid	vehicle	wine
Ed-NN	78.59	68.45	76.64	83.14	95.04	68.59	94.79
Co-NN	81.30	68.94	78.72	84.19	89.44	69.52	94.38
LDA+Ed-NN	87.74	58.91	78.03	68.60	94.24	73.07	**98.11**
LDA+Co-NN	75.62	59.01	76.76	58.52	85.48	72.03	97.66
CCA+Ed-NN	87.50	57.24	76.09	71.52	92.83	73.26	97.35
d-CDA+Co-NN	84.34	71.23	81.81	**84.56**	90.65	71.91	95.90
f-CDA+Co-NN	87.93	72.55	82.72	81.38	91.29	74.35	97.04
Ke-SVM	91.76	**74.27**	**83.85**	83.98	**96.88**	**78.01**	95.80
Ke-RVM	**95.85**	73.15	76.94	80.90	93.45	76.58	96.65

$$\frac{\partial s_{ij}}{\partial A_{kl}} = \frac{x_{ik}x_{jk}}{\sqrt{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}} - \frac{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j}{2\sqrt{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}} \left(\frac{x_{ik}^2}{\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j} + \frac{x_{jk}^2}{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i}\right) \quad (14)$$

If $k \neq l$

$$\frac{\partial s_{ij}}{\partial A_{kl}} = \frac{x_{ik}x_{jl} + x_{il}x_{jk}}{\sqrt{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}} - \frac{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j}{\sqrt{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}} \left(\frac{x_{ik}x_{il}}{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i} + \frac{x_{jk}x_{jl}}{\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}\right) \quad (15)$$

In the inner iteration of this algorithm, the projection step onto *C*, the space of all positive-semi definite matrices, is done by first finding the diagonalization $\mathbf{A} = \mathbf{U}^T\mathbf{\Lambda}\mathbf{U}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of $\mathbf{A}$'s eigenvalues and the columns of $\mathbf{U} \in \mathbb{R}^{n\times n}$ contains $\mathbf{A}$'s corresponding eigenvectors, and taking $\mathbf{A}' = \mathbf{U}^T\mathbf{\Lambda}'\mathbf{U}$, where $\mathbf{\Lambda}' = \mathrm{diag}(\max(0,\lambda_1), \ldots, \max(0,\lambda_n))$ (Golub, 1996).

Also we can set $A_{11} = 1$ and only need to estimate other parameters.

4. Experimental Result

To evaluate the performance of proposed CDA algorithm, some classification experiments are performed on two types of data: (a) 7 UCI machine learning repository: balance, glass, lense, sonar, thyroid, vehicle and wine (Blake & Merz, 1998). (b) ORL human face dataset for face recognition (available at http://www.uk.research.att.com/facedatabase.html); these datasets are different in the feature dimension (from 4 to 3280), sample size (from 30 to 800), the class of number (from 3 classes to 40 classes) and the physical meaning. So the experimental result on this large scale of datasets should be reliable.

The methods used in the following comparisons include:

Ed-NN: Euclidean distance based 1-nearest neighbor classifier;

Co-NN: Correlation measure based 1-nearest neightbor classifier;

PCA: Principal component analysis;

LDA: Linear discriminant analysis;

CCA: Canonical correlation analysis for classification. Here we employ one-of-c label encoding (Johansson, 2001);

d-CDA: The case of diagonal correlation discriminant analysis.

f-CDA: The full matrix correlation discriminant analysis;

Ke-SVM: Gaussian Kernel Support Vector Machines (Vapnik, 1995) for classification. Here we use *one-vs-others* strategy to solve the multi-class classification problem by learning one binary classifier for each class; The Gaussian kernel parameters and *C* are decided by experiments;

Ke-RVM: Gaussian Kernel Relevance Vector Machines for classification (Tipping, 2001). *One-vs-other* strategy is also used here to design multi-class classifiers; The Gaussian kernel parameters are decided by experiments.

The details of experiments are reported separately.

4.1 Experiments on UCI Data

All contrastive experiments are based on the identical partition of the training/test set for each dataset of UCI. In each round of experiment, half of total samples are

randomly selected for training, and other samples are used for testing. The experiments are repeated 100 times.

The recognition results on all datasets are given in Table 1, from which we can observe that almost on all datasets except sonar dataset, the d-CDA and f-CDA outperform Co-NN; That demonstrates the inclusion of discriminant information in correlation measure can improve the performance. However, on sonar datasets, f-CDA is worse than d-CDA. That is mainly due to that the feature dimension of 60 and the number of training sample is only 104 and in f-CDA, there are about 1800 parameters to be estimated. So over-fitting problem happened on training set and cause the performance degrading on test set.

We also can observe d-CDA+Co-NN framework can achieve overall better performance than LDA+Co-NN, especially on sonar and glass datasets. That is mainly due to the nonlinear distribution of these two datasets. Obviously, d-CDA or f-CDA algorithm can explore and utilize more information under this situation.

For the purpose of comparison, in the above table, we also give the result based on the state-of-art kernel-methods, such as Kernel SVM and Kernel RVM. Although it is not very fair to directly compare d-CDA and f-CDA with these two powerful nonlinear machine learning methods, we can see that on some data sets, their performances are still comparable.

In computation efficiency, both the optimization of d-CDA and that of f-CDA are the gradient-based iteration processes. Specifically, for d-CDA, each iteration consists of computing the gradient $\nabla f(\boldsymbol{\rho})$, and the optimum step along the gradient direction. If the number of iteration is *m*, the total computation would be $O(mdn^2)$. For f-CDA, the computation is much more complex due to the computation of the gradient of $d(d+1)/2$ parameters, the sum of $O(n^2)$ correlations between-class and within-class and also the eigen-decomposition of the full matrix. So the computation will be $O(m(d^2n^3+d^4))$. In the experiment, the running time of d-CDA algorithm (in MATLAB, on a PIV 3.4GHz) is less than 9 seconds for glass, and about 26 seconds for the sonar problems in one round, and the corresponding times for f-CDA are about 70 seconds and 200 seconds. As mentioned earlier, the target function of CDA is not convex. The gradient-based methods can not guarantee to get the global optimum point. To partially solve this problem, we run the d-CDA and f-CDA several times with random initial values and select the best result. This also makes the training process a little longer.

4.2 Application of CDA to Face Recognition

The characteristics of face recognition different from previous classification problems are that there are more classes and fewer training samples. So in this experiment we want to evaluate the performance of CDA on the problems with large class number and sparse data distribution.

The experiments are randomly repeated 10 times for ORL. All contrastive experiments are based on the identical partition of the training/test set for each dataset. For ORL dataset, the training set size 200 is far less than the sample dimension using Gabor wavelets about 3000 dimensions. To overcome this problem, all the algorithms are preceded by PCA and then performed in the transformed 120-dimensional PCA subspace which accounts for 95% of total energy. The comparison is shown in table 2 (note in this experiment, due to the optimization complexity, we only evaluate the PCA+d-CDA+Co-NN framework and do not evaluate the performance of f-CDA)

Table 2 Face recognition accuracies on ORL databases (%).

METHOD	ORL
PCA+Ed-NN	93.75
PCA+Co-NN	94.10
PCA+LDA+Ed-NN	97.7
PCA+LDA+Co-NN	98.5
CCA+Ed-NN	97.1
PCA+d-CDA+Co-NN	99.5

From the table 2, we first can confirm that correlation measure-based methods are always better than Euclidean metric-based methods nomatter in PCA space or PCA+LDA space, which coincides with the conclusion of Kittler (2000) very well. Besides this, we can find the similar trend that by incorporating the discriminant information, d-CDA can achieve much better performance than PCA+Co-NN framework, and better than the popular PCA+LDA+Co-NN framework for face recognition.

5. Discussion and Some Further Research Topics

We propose a novel discriminant learning algorithm in correlation measure space, Correlation Discriminant Analysis (CDA). In this framework, based on the definitions of within-class correlation and between-class correlation, the optimum transformation can be sought for to maximize the difference between them, which is in accordance with good classification performance empirically. Under different cases of the transformation, diagonal and full metrics implementations of the algorithm are given. Extensive empirical evaluations of CDA demonstrate its advantage over alternatives methods.

Although in this paper both the d-CDA and f-CDA are limited to their linear transformation forms, in fact, they

can be generalized to the corresponding kernel versions via the kernel trick. However, compared with the simple extension of CCA to kernel CCA and LDA to kernel LDA, the extension of CDA to kernel CDA is not very easy. The similar method to (Yeung, to appear) to represent the projection matrix in feature space can be adopted to solve this problem.

And another interesting topic might be to combine the sparse Bayesian learning (Tipping, 2001) with CDA to solve the over-fitting problems and at the same time fulfill the feature selection by CDA.

Acknowledgments

We would like to express our grateful thanks to the anonymous reviewers for their valuable comments and suggestions to improve the quality of our papers.

References

Barker, M., & Rayens W.(2003). Partial least squares for discrimination. *Journal of Chemometrics*, Vol 17, pp.166-173.

Belhumeur, P. N., Hespanha, J., & Kriegman, D. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8, 19(7):711–720.

Blake, C. L. & Merz, C. J. (1998) *UCI repository of machine learning databases*, URL http://www.ics.uci. edu/ ~mlearn/MLRepository.html.

Brown, M. P, Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Science*, 97, pp.262-267.

Fukunaga, K. (1990). *Statistical pattern recognition.* Academic Press, San Diego, 2nd edition.

Golub, G. H. & Van Loan, C. F. (1996) *Matrix computations*. Johns Hopkins University Press.

Han, E.H., Karypis, G. & Kuma, V. (2001), Text categorization using weight adjusted k-Nearest Neighbor classification. *5th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD), pp.53-65.

Hardoon, D. R., Szedmak, S. & Taylor, J. S (2004). Canonical correlation analysis: an overview with application to learning method, *Neural Computation*, Vol.16, pp: 2639-2664.

Hillel, A. B., Hertz, T., Shental, N. & Weinshall, D. (2005). Learning a Mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research* Vol 6, pp.937-965.

Johansson, B. (2001). *On classification: simultaneously reducing dimensionality and finding automatic representation using canonical correlation*. Technical report LiTH-ISY-R-2375, ISSN 1400-3902, Linköping University

Kittler, J., Li, Y.P. & Matas, J. (2000) On matching scores for LDAbased face verification. *Proceedings of British Machine Vision Conference* (BMVC00), pp. 42-51.

Kumar, B.V. (1986). Minimum variance synthetic discriminant functions. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, Vol 3, Issue 10, pp.1579-1584.

Loog, M., Ginneken, B.V. & Duin, R.W. (2004) Dimensionality reduction by Canonical Contextual Correlation projections, *the 8th European Conference on Computer Vision*, pp.562 – 573.

Mahalanobis, A., Kumar, B.V, & Casasent, D. (1987). Minimum average correlation energy filters. *Applied Optics*, Vol. 26, Issue 17, pp. 3633-3645.

Martinez, A. M. & Zhu, M. L. (2005). Where are linear feature extraction methods applicable. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol 27, Issue 12 pp.1934-1944.

Peterson, M.R., Doom, T.E. & Raymer, M. L. (2005). Facilitated KNN classifier optimization with varying similarity measures. *IEEE Congress on Evolutionary Computation*, vol 3, pp. 2514-2521.

Phillips, P. J., Moon, H., Rauss, P. J. & Rizvi, S.,(2000) The FERET evaluation methodology for face recognition algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10, pp.1756-1760.

Refregier, P. (1990). Filter design for optical pattern recognition: multi-criteria optimization approach. *Optics Letters*, Vol 15, Issue 15, pp.854-856

Rockafellar, R. (1970) *Convex analysis*. Princeton University Press.

Smith, R.S., Kittler, J., Hamouz, M.& Illingworth, J. (2006). Face recognition using angular LDA and SVM ensembles. *Proceedings of the 18th International Conference on Pattern Recognition* (ICPR'06), Vol 3, pp.1008-1012.

Sun, T.K. & Chen, S.C. (2007), Class label- versus sample label-based CCA. *Applied Mathematics and Computation* (in press).

Tipping M.E. (2001) Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, pp.211-214.

Xie, C.Y., Savvides, M. & Kumar, B.V. (2005). Redundant class-dependence feature analysis based on correlation filters using FRGC2.0 data. *Proceedings of*

the Computer Vision and Pattern Recognition, Vol 3, pp.153-153.

Xing, E.P., Ng, A.Y., Jordan, M.I. & Russell, S. (2003). Distance metric learning, with application to clustering with side-information. *Advances in Neural Information Processing Systems* 15, pp. 521–528.

Yeung, D.Y. & Chang H. (To appear) A kernel approach for semi-supervised metric learning. *IEEE Transactions on Neural Networks*.

Vapnik V. (1995), *The nature of statistical learning theory*. New York: Springer Verlag.