Discriminant Kernels derived from the Optimum Nonlinear Discriminant Analysis

Takio Kurita
Hiroshima University

Abstract— **Linear discriminant analysis (LDA) is one of the well known methods to extract the best features for multi-class discrimination. Recently Kernel discriminant analysis (KDA) has been successfully applied in many applications. KDA is one of the nonlinear extensions of LDA and construct nonlinear discriminant mapping by using kernel functions. But the kernel function is usually defined a priori and it is not known what the optimum kernel function for nonlinear discriminant analysis is. Also the class information is not usually introduced to define the kernel functions. In this paper the optimum kernel function in terms of the discriminant criterion is derived by investigating the optimum discriminant mapping constructed by the optimum nonlinear discriminant analysis (ONDA).**

Otsu derived the optimum nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities similar with the Bayesian decision theory. He showed that the optimum non-linear discriminant mapping was obtained by using Variational Calculus. The optimum nonlinear discriminant mapping can be defined as a linear combination of the Bayesian *a posterior* probabilities and the coefficients of the linear combination are obtained by solving the eigenvalue problem of the matrices defined by using the Bayesian *a posterior* probabilities. This means that the ONDA is closely related to Bayesian decision theory. Also Otsu showed that LDA could be interpreted as a linear approximation of the ONDA through the linear approximation of the Bayesian *a posterior* probabilities.

In this paper, the optimum kernel function is derived by investigating the optimum discriminant mapping constructed by ONDA. The derived kernel function is also given by using the Bayesian *a posterior* probabilities. This means that the class information is naturally introduced in the kernel function. For real application, we can define a family of discriminate kernel functions can be defined by changing the estimation method of the Bayesian *a posterior* probabilities.

I. Introduction

LINEAR discriminant analysis (LDA) [1] is one of the well known methods to extract the best discriminating features for multi-class classification. LDA is formulated as a problem to find an optimum linear mapping by which the within-class scatter in the mapped discriminant feature space is made as small as possible relative to the between-class scatter. LDA is useful for linear separable cases, but for more complicated cases, it is necessary to extend it to nonlinear.

Recently the kernel discriminant analysis (KDA) has been successfully applied in many applications [2], [3], [4]. KDA is one of the nonlinear extensions of LDA and constructs a nonlinear discriminant mapping by using kernel functions. Usually the kernel function is defined a priori. For example, the polynomial functions or the radial bases functions are often used as a kernel function for KDA. However, it is not known what the best kernel function for nonlinear discriminant analysis is. Also the class information is usually not introduced in such kernel functions.

On the other hand, Otsu derived the optimum nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities [5], [6], [7] similar with the Bayesian decision theory[8]. He showed that the optimum non-linear discriminant mapping was obtained by using Variational Calculus and was closely related to Bayesian decision theory (The *posterior* probabilities). The optimum nonlinear discriminant mapping can be defined as a linear combination of the Bayesian *a posterior* probabilities and the coefficients of the linear combination are obtained by solving the eigenvalue problem of the matrices defined by using the Bayesian *a posterior* probabilities. This result is fundamental to understand the nature of the discriminant analysis.

Also Otsu pointed out that LDA could be interpreted as a linear approximation of this ultimate ONDA through the linear approximations of the Bayesian *posterior* probabilities. However the linear model used in the LDA is not suitable to estimate the *posterior* probabilities because the outputs of the linear model can not satisfy the constraints on the probabilities. To overcome this drawback, Kurita et al. [9] proposed Logistic discriminant analysis (LgDA) in which the *posteriori* probabilities are estimated by using the Multi-nominal logistic regression (MLR) instead of the linear model. MLR is known as one of the generalized linear models (GLM) which are a flexible generalization of the ordinary least squares regression. MLR can naturally estimate the *posteriori* probabilities by modifying the outputs of the linear predictor by the link function. Thus LgDA can be regarded as a natural extension of LDA by generalizing the linear model to the generalized linear model. It was shown that the discriminant spaces constructed by LgDA were drastically improved by this generalization for the several standard repository datasets [9].

This theory of ONDA suggests that many novel nonlinear discriminant mappings can be constructed if we change the estimation methods of the *posterior* probabilities. For example, Kurita et al. [11] proposed the neural network based non-linear discriminant analysis in which the outputs of the trained MLP were used as the estimates of the *posteriori* probabilities because the outputs of the trained multi-layered Perceptron (MLP) for pattern classification problems can be regarded as the approximations of the *posteriori* probabilities [10].

This paper investigates the kernel function used in the ONDA. The best kernel function is derived from the optimum

discriminant mapping constructed by ONDA by investigating the dual problem of the eigenvalue problem of ONDA. The derived kernel function is also given by using the *posteriori* probabilities. This means that the class information is naturally introduced in this kernel function. Since ONDA is optimum in terms of the discriminant criterion, the derived kernel function is also optimum in terms of the discriminant criterion. We call this kernel function the discriminant kernel function (DKF). For real application we can define a family of discriminate kernel functions by changing the estimation method of the Bayesian *a posterior* probabilities. This situation is similar with applications of the Bayesian decision theory or the optimum nonlinear discriminant analysis.

The rest of this paper is organized as follows: Section II reviews LDA and KDA. The theory of ONDA is introduced in Section III. Then the dual problem of ONDA is investigated to derive the discriminant kernel function (DKF) in Section IV. Section IV also shows some simple examples of KDFs in which the Bayesian *a posterior* probabilities are estimated from the training samples. To investigate the validity of the proposed DKF, the kernel matrix estimated by using the proposed DKF is compared with the one computed by using the Radial Basis functions in Section V. The discriminant spaces constructed by using the proposed DKF is also compared with the one constructed by using the Radial Basis functions in Section V. Finally, Section VI concludes the paper.

II. Linear and Kernel Discriminant Analysis

A. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) was proposed by Fisher in the original article "The use of multiple measures in taxonomic problems (1936)" [1]. LDA is defined as a method to find the linear combination of features which best separates two classes of objects. It is extended to find a subspace in which the within-class scatter in the mapped discriminant space is made as small as possible relative to the between-class scatter for multi-classes. LDA is regarded as one of the well known methods to extract the best discriminating features for multi-class classification.

Let an m dimensional feature vector be $\boldsymbol{x} = (x_1, \ldots, x_m)^T$. Consider K classes denoted by $\{C_1, \ldots, C_K\}$. Assume that we have N feature vectors $\{\boldsymbol{x}_i | i = 1, \ldots, N\}$ as training samples and they are labeled as one of the K classes. Then LDA constructs a dimension reducing linear mapping from the input feature vector $\boldsymbol{x}$ to a new feature vector $\boldsymbol{y}$ as

$$\boldsymbol{y} = A^T \boldsymbol{x} \tag{1}$$

where $A = [a_{ij}]$ is the coefficient matrix.

The discriminant criterion

$$J = \mathrm{tr}\left(\hat{\Sigma}_T^{-1} \hat{\Sigma}_B\right) \tag{2}$$

is used to evaluate the performance of the discrimination of the new feature vectors $\boldsymbol{y}$, where $\hat{\Sigma}_T$ and $\hat{\Sigma}_B$ are respectively the total covariance matrix and the between-class covariance matrix of the new feature vectors $\boldsymbol{y}$. The objective of LDA is to maximize the discriminant criterion J.

The optimal coefficient matrix A is then obtained by solving the following generalized eigenvalue problem

$$\Sigma_B A = \Sigma_T A \Lambda \qquad (A^T \Sigma_T A = I) \tag{3}$$

where $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_L)$ is a diagonal matrix of eigen values and I denotes the unit matrix. The matrices Σ_T and Σ_B are respectively the total covariance matrix and the between-class covariance matrix of the input feature vectors $\boldsymbol{x}$, and they are computed as

$$\Sigma_T = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)^T \tag{4}$$

$$\Sigma_B = \sum_{k=1}^{K} P(C_k)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T, \tag{5}$$

where $P(C_k)$, $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{x}}_T$ denote *a priori* probability of the class C_k, the mean vector of the class C_k and the total mean vector, respectively. Usually we compute the probability of the class C_k as $P(C_k) = \frac{N_k}{N}$, where N_k is the number of input vectors of the class C_k and N is the number of whole input vectors.

The j-th column of A is the eigenvector corresponding to the j-th largest eigenvalue. Therefore, the importance of each element of the new feature vector $\boldsymbol{y}$ is evaluated by the corresponding eigenvalues. The dimension of the new feature vector $\boldsymbol{y}$ is bounded by $\min(K-1, N)$ because the rank of the matrix Σ_B is bounded by $\min(K-1, N)$.

B. Kernel Discriminant Analysis

Recently the kernel discriminant analysis (KDA) has been successfully applied in many applications [2], [3], [4]. The KDA is one of the nonlinear extensions of LDA and constructs a nonlinear discriminant mapping as a linear combination of kernel functions. This is similar with the kernel Support Vector Machine (SVM) which constructs a nonlinear decision functions using kernel functions.

Consider a nonlinear mapping Φ from a input feature vector $\boldsymbol{x}$ to the new feature vector $\Phi(\boldsymbol{x})$. In KDA the discriminant features $\boldsymbol{y}$ are constructed as a linear combinations of the new feature $\Phi(\boldsymbol{x})$.

For the case of 1-dimensional feature extraction, the discriminant mapping can be given as

$$y = \boldsymbol{a}^T \Phi(\boldsymbol{x}). \tag{6}$$

Since the coefficient vector $\boldsymbol{a}$ can be expressed as a linear combinations of the training samples as

$$\boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \Phi(\boldsymbol{x}_i), \tag{7}$$

the discriminant mapping can be rewritten as

$$\begin{aligned} y &= \sum_{i=1}^{N} \alpha_i \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}) \\ &= \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\alpha}^T \boldsymbol{k}(\boldsymbol{x}), \end{aligned} \tag{8}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}) = \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x})$ and $\boldsymbol{k}(\boldsymbol{x}) = (K(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{x}_N, \boldsymbol{x}))$ are the kernel function defined by the nonlinear mapping $\Phi(\boldsymbol{x})$ and the vector of the kernel functions, respectively.

Then the discriminant criterion is given as

$$J = \frac{\sigma_B^2}{\sigma_T^2} = \frac{\boldsymbol{\alpha}^T \Sigma_B^{(K)} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \Sigma_T^{(K)} \boldsymbol{\alpha}}, \tag{9}$$

where

$$\sigma_T^2 = \frac{1}{N} \sum_{i=1}^{N} ||y_i - \bar{y}||^2 \tag{10}$$

$$\sigma_B^2 = \sum_{k=1}^{K} P(C_k) ||\bar{y}_k - \bar{y}||^2 \tag{11}$$

$$\Sigma_T^{(K)} = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{k}(\boldsymbol{x}_i) - \bar{\boldsymbol{k}}_T)(\boldsymbol{k}(\boldsymbol{x}_i) - \bar{\boldsymbol{k}}_T)^T \tag{12}$$

$$\Sigma_B^{(K)} = \sum_{k=1}^{K} P(C_k)(\bar{\boldsymbol{k}}_k - \bar{\boldsymbol{k}}_T)(\bar{\boldsymbol{k}}_k - \bar{\boldsymbol{k}}_T)^T. \tag{13}$$

In these definitions, $\bar{y}$, $\bar{y}_k$, $\bar{\boldsymbol{k}}_T$, and $\bar{\boldsymbol{k}}_k$ denote the total mean value of the discriminant feature y, the mean value of the class C_k of the discriminant feature y, the total mean vector of the kernel feature vector $\boldsymbol{k}(\boldsymbol{x})$, and the mean vector of the class C_k of the kernel feature vector $\boldsymbol{k}(\boldsymbol{x})$, respectively.

From these relations, it is noticed that the problem to find the optimum coefficients vector $\boldsymbol{\alpha}$ which maximizes the discriminant criterion J is equivalent to apply LDA to the kernel feature vector $\boldsymbol{k}(\boldsymbol{x})$. Thus the optimum coefficient vector $\boldsymbol{\alpha}$ can be obtained by solving the generalized eigenvalue problem

$$\Sigma_B^{(K)} \boldsymbol{\alpha} = \Sigma_W^{(K)} \boldsymbol{\alpha} \lambda. \tag{14}$$

For the multi-dimension case, the kernel discriminant mapping is given by

$$\boldsymbol{y} = A^T \boldsymbol{k}(\boldsymbol{x}), \tag{15}$$

where the coefficinet matrix A is defined by $A^T = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N)$. The optimum coefficient matrix A is obtained by solving the eigenvalue problem

$$\Sigma_B^{(K)} A = \Sigma_W^{(K)} A \Lambda. \tag{16}$$

Usually the kernel function is defined a priori in KDA. The polynomial functions

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^q \tag{17}$$

or the Radial Basis functions

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{y}||^2}{2\sigma^2}\right) \tag{18}$$

are often used as the kernel function for KDA. However it is not noticed what the best kernel function for nonlinear discriminant analysis is. Also the class information is usually not introduced in these kernel functions.

III. Optimum Nonlinear Discriminant Analysis

A. Optimal Nonlinear Discriminant Analysis

Otsu derived the optimal nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities [5], [6], [7]. This assumption is similar with the Bayesian decision theory. Similar with LDA, ONDA constructs the dimension reducing optimum nonlinear mapping which maximizes the discriminant criterion J. Namely ONDA finds the optimum nonlinear mapping in terms of the discriminant criterion J.

By using Variational Calculus, Otsu showed that the optimal non-linear discriminant mapping is obtained as

$$\boldsymbol{y} = \sum_{k=1}^{K} P(C_k|\boldsymbol{x}) \boldsymbol{u}_k \tag{19}$$

where $P(C_k|\boldsymbol{x})$ is the Bayesian *posterior* probability of the class C_k given the input $\boldsymbol{x}$. The vectors $\boldsymbol{u}_k (k = 1, \ldots, K)$ are class representative vectors which are determined by the following generalized eigenvalue problem

$$\Gamma U = P U \Lambda \tag{20}$$

where $\Gamma = [\gamma_{ij}]$ is a $K \times K$ matrix whose elements are defined by

$$\gamma_{ij} = \int (P(C_i|\boldsymbol{x}) - P(C_i))(P(C_j|\boldsymbol{x}) - P(C_j)) p(\boldsymbol{x}) d\boldsymbol{x} \tag{21}$$

and the other matrices are defined as

$$U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]^T \tag{22}$$

$$P = \mathrm{diag}\,(P(C_1), \ldots, P(C_K)) \tag{23}$$

$$\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_L)\,. \tag{24}$$

It is important to notice that the optimal non-linear mapping is closely related to Bayesian decision theory, namely the *posterior* probabilities $P(C_k|\boldsymbol{x})$. Along this line, Fukunaga et al. [12] discussed the various properties of the criterion from the viewpoint of such non-linear mappings.

By using the eigen vectors obtained by solving the generalized eigenvalue problem (20), we can construct the optimum nonlinear discriminant mapping from a given input feature $\boldsymbol{x}$ to the new discriminant feature $\boldsymbol{y}$ as shown in the equation (19) if we can know or estimate all the *posterior* probabilities. This means that we have to estimate the *posterior* probabilities for real applications. This situation is similar with the Bayesian decision theory. In other words, a family of nonlinear discriminant mapping can be defined by changing the estimation method of the *posterior* probabilities.

B. Linear discriminant analysis of posterior probabilities

The optimum nonlinear discriminant mapping (19) obtained by ONDA can be rewritten as

$$\boldsymbol{y} = \sum_{k=1}^{K} P(C_k|\boldsymbol{x})\boldsymbol{u}_k = U^T\boldsymbol{B}(\boldsymbol{x}), \tag{25}$$

where $\boldsymbol{B}(\boldsymbol{x}) = (P(C_1|\boldsymbol{x}), \ldots, P(C_k|\boldsymbol{x}))^T$ is the vector of *posterior* probabilities. This means that the optimum nonlinear discriminant mapping can be interpreted as a linear combination of *posterior* probabilities.

Here we consider the linear discriminant analysis of the vector of *posterior* probabilities $\boldsymbol{B}(\boldsymbol{x})$. Namely the vector of *posterior* probabilities $\boldsymbol{B}(\boldsymbol{x})$ is considered as the input feature vector and LDA is applied to this vector $\boldsymbol{B}(\boldsymbol{x})$. The linear discriminant mapping from this vector $\boldsymbol{B}(\boldsymbol{x})$ to the new discriminant feature vector $\boldsymbol{y}$ is defined by

$$\boldsymbol{y} = U^T\boldsymbol{B}(\boldsymbol{x}). \tag{26}$$

Then the optimum coefficient matrix U which maximizes the discriminant criterion J is obtained by solving the following generalized eigenvalue problem

$$\Phi_B U = \Phi_T U \Lambda. \tag{27}$$

The matrices Φ_T and Φ_B are respectively the total covariance matrix and the between-class covariance matrix of the vectors of *posterior* probabilities $\boldsymbol{B}(\boldsymbol{x})$, and they are given as

$$\Phi_T = \int (\boldsymbol{B}(\boldsymbol{x}) - \bar{\boldsymbol{B}}_T)(\boldsymbol{B}(\boldsymbol{x}) - \bar{\boldsymbol{B}}_T)^T p(\boldsymbol{x}) d\boldsymbol{x} \tag{28}$$

$$\Phi_B = \sum_{k=1}^{K} P(C_k)(\bar{\boldsymbol{B}}_k - \bar{\boldsymbol{B}}_T)(\bar{\boldsymbol{B}}_k - \bar{\boldsymbol{B}}_T)^T, \tag{29}$$

where $\bar{\boldsymbol{B}}_k$ and $\bar{\boldsymbol{B}}_T$ denote the mean vector of the class C_k and the total mean vector of the vector of *posterior* probabilities respectively. By using the relations between Γ, Φ_T, and Φ_B shown as

$$\Phi_T = \Gamma \tag{30}$$

$$\Phi_B = \Gamma P^{-1} \Gamma, \tag{31}$$

the eigenvalue problem (27) can be rewritten as

$$\Gamma U = P U \Lambda. \tag{32}$$

This is the same as the eigenvalue problem (20) of ONDA. In other words, the optimum nonlinear discriminant mapping obtained by ONDA is the same as the mapping obtained by the linear discriminant analysis of the vector of *posterior* probabilities $\boldsymbol{B}(\boldsymbol{x})$. By using the vector of the *posterior* probabilities as the new feature vector and applying LDA to such vectors, we can construct the optimum nonlinear discriminant mapping from the input vector $\boldsymbol{x}$ to the new discriminant feature vectors $\boldsymbol{y}$. Since the mapping is the same as the optimum nonlinear discriminant mapping, the constructed mapping is optimum in terms of the discriminant criterion J.

These results show that the estimation of the *posterior* probabilities is also very important in the context of the discriminant analysis like in the Bayesian decision theory.

C. Linear approximation of ONDA

In the previous subsections, ONDA is derived as the ultimate nonlinear extension of LDA and the optimum nonlinear discriminant mapping which maximizes the discriminant criterion J can be obtained by applying LDA to the vector of *posterior* probabilities. Then we may have the following question: in what sense does LDA approximate this optimum nonlinear discriminant mapping constructed by ONDA ? The answer to this question was also given by Otsu [7]. The discriminant mapping of LDA can be interpreted as a linear approximation of the one of ONDA through the linear approximations of the *posterior* probabilities $P(C_k|\boldsymbol{x})$.

Consider a linear approximation of the Bayesian *posterior* probabilities $P(C_k|\boldsymbol{x})$ which is expressed as

$$L(C_k|\boldsymbol{x}) = \boldsymbol{b}_k^T\boldsymbol{x} + b_k^{(0)}. \tag{33}$$

To determine the coefficients $\boldsymbol{b}_k^T$ and $b_k^{(0)}$, we minimize the mean squared errors between the Bayesian *posterior* probabilities $P(C_k|\boldsymbol{x})$ and their linear approximations $L(C_k|\boldsymbol{x})$

$$\varepsilon^2 = \int (P(C_k|\boldsymbol{x}) - L(C_k|\boldsymbol{x}))^2 p(\boldsymbol{x}) d\boldsymbol{x}. \tag{34}$$

The optimum linear approximation of the Bayesian *posterior* probabilities which minimizes the mean squared errors is given by

$$L(C_k|\boldsymbol{x}) = P(C_k)[(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T \Sigma_T^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_T) + 1] \tag{35}$$

where Σ_T denotes the total covariance matrix of the input feature vectors $\boldsymbol{x}$.

It is interesting to note that this function has unit-sum property as

$$\sum_{k=1}^{K} L(C_k|\boldsymbol{x}) = 1. \tag{36}$$

This is similar with the property of the probabilities but it's value happens to be greater than 1 or less than 0. Namely this function $L(C_k|\boldsymbol{x})$ is an approximation of the Bayesian *posterior* probabilities $P(C_k|\boldsymbol{X})$ but it does not satisfy some properties of the probability.

Consider the approximation of the optimum nonlinear discriminant mapping obtained by ONDA by substituting these linear approximations $L(C_k|\boldsymbol{x})$ for the Bayesian *posterior* probabilities $P(C_k|\boldsymbol{x})$ in (19) and (20). By this substitution, the equation (19) becomes

$$\begin{aligned} \boldsymbol{y} &= \sum_{k=1}^{K} L(C_k|\boldsymbol{x})\boldsymbol{u}_k \\ &= U^T P M^T \Sigma_T^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_T) + U^T\boldsymbol{p} \end{aligned} \tag{37}$$

where

$$M = [(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_T), \ldots, (\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_T)]^T \tag{38}$$

$$\boldsymbol{p} = (P(C_1), \ldots, P(C_K))^T. \tag{39}$$

Also by substituting these linear approximations $L(C_k|\boldsymbol{x})$ for the Bayesian *posterior* probabilities $P(C_k|\boldsymbol{x})$, the matrix Γ in Eq. (20) of ONDA becomes

$$\Gamma = P M^T \Sigma_T^{-1} M P. \tag{40}$$

By multiplying M from the left and substituting A for $\Sigma_T^{-1}MPU$, we have

$$\Sigma_B A = \Sigma_T A \Lambda. \tag{41}$$

This is the same as the eigenvalue problem (3) of LDA. This means that the linear discriminant mapping of LDA is the linear approximation of the one of ONDA through the linear approximation $L(C_k|\boldsymbol{x})$ of the *posterior* probabilities $P(C_k|\boldsymbol{x})$. This also shows the importance of the estimation of the *posterior* probabilities to construct the discriminant mapping.

Linear approximation $L(C_k|\boldsymbol{x})$ of the *posterior* probabilities are used in LDA. However a linear model is not suitable to estimate the *posterior* probabilities because they are not satisfy the property on the probability. To overcome this drawback, Kurita et al. [9] proposed Logistic discriminant analysis (LgDA) in which the *posteriori* probabilities are estimated by Multi-nominal logistic regression (MLR) instead of the linear model in LDA. MLR is known as one of the members of the generalized linear model (GLM) which is flexible generalization of the ordinary least squares regression. MLR can naturally estimate the *posteriori* probabilities by modifying the outputs of the linear predictor by the link function [13]. Thus LgDA can be regarded as a natural extension of LDA by generalizing the linear model to the generalized linear model. It was shown that the discriminant spaces constructed by the LgDA were drastically improved for the several standard repository datasets.

Kurita et al. [11] also proposed the neural network based non-linear discriminant analysis in which the outputs of the trained MLP are used as the estimates of the *posteriori* probabilities because the outputs of the trained multi-layered Perceptron (MLP) for pattern classification problems can be regarded as the approximations of the *posteriori* probabilities [10].

IV. Discriminant Kernels

A. Dual Problem of ONDA

In the KDA, usually the kernel function is defined a priori. The polynomial functions or the Radial Basis functions are often used as the kernel functions but such kernel functions are general and are not related to the discrimination. Thus the class information is usually not introduced in these kernel functions. Also it is not known what the optimum kernel function for nonlinear discriminant analysis is.

This paper derives the optimum kernel function in terms of the discriminant criterion by investigating the optimum nonlinear discriminant mapping constructed by ONDA. The optimum kernel function can be derived by investigating the dual problem of the eigenvalue problem of ONDA. Since ONDA is optimum in terms of the discriminant criterion, the derived discriminant kernel function is also optimum in terms of the discriminant criterion.

The eigenvalue problem of ONDA given by the equation (20) is the generalized eigenvalue problem. By multiplying $P^{-1/2}$ from the left, this eigen equations can be rewritten as the usual eigenvalue problem as

$$P^{-1/2}\Gamma P^{-1/2}P^{1/2}U = P^{1/2}U\Lambda. \tag{42}$$

By denoting $\tilde{U} = P^{1/2}U$, we have the following usual eigenvalue problem as

$$(P^{-1/2}\Gamma P^{-1/2})\tilde{U} = \tilde{U}\Lambda. \tag{43}$$

Then the optimum nonlinear discriminant mapping of ODNA is rewritten as

$$\boldsymbol{y} = U^T\tilde{\boldsymbol{B}}(\boldsymbol{x}) = \tilde{U}^T P^{-1/2}\tilde{\boldsymbol{B}}(\boldsymbol{x}) = \tilde{U}^T\boldsymbol{\phi}(\boldsymbol{x}) \tag{44}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = P^{-1/2}\tilde{\boldsymbol{B}}(\boldsymbol{x})$ and $\tilde{\boldsymbol{B}}(\boldsymbol{x}) = (P(C_1|\boldsymbol{x}) - P(C_1), \ldots, P(C_K|\boldsymbol{x}) - P(C_K))^T$.

For the case of N training samples, the eigenvalue problem to determine the class representative vectors (43) is given by

$$(\Phi^T\Phi)\tilde{U} = \tilde{U}\Lambda, \tag{45}$$

where $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_N))^T$.

The dual eigenvalue problem of (45) is then given by

$$(\Phi\Phi^T)V = V\Lambda. \tag{46}$$

From the relation on the singular value decomposition of the matrix Φ, these two eigenvalue problems (45) and (46) have the same eigenvalues and there is the following relation between the eigenvectors $\tilde{U}$ and V as

$$\tilde{U} = \Phi^T V\Lambda^{-1/2}. \tag{47}$$

By inserting this relation into the nonlinear discriminant mapping (44), we have

$$\begin{aligned}\boldsymbol{y} &= \Lambda^{-1/2}V^T\Phi\boldsymbol{\phi}(\boldsymbol{x}) \\ &= \sum_{i=1}^{N}\Lambda^{-1/2}\boldsymbol{v}_i\phi((\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) \\ &= \sum_{i=1}^{N}\boldsymbol{\alpha}_i K(\boldsymbol{x}_i,\boldsymbol{x}) - \boldsymbol{\alpha}_0\end{aligned} \tag{48}$$

where

$$\begin{aligned}K(\boldsymbol{x}_i,\boldsymbol{x}) &= \phi((\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) + 1 \\ &= \sum_{k=1}^{K}\frac{(P(C_k|\boldsymbol{x}_i) - P(C_k)(P(C_k|\boldsymbol{x}) - P(C_k)}{P(C_k)} + 1 \\ &= \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x}_i)P(C_k|\boldsymbol{x})}{P(C_k)}.\end{aligned} \tag{49}$$

This shows that the kernel function of the optimum nonlinear discriminant mapping is given by

$$K(\boldsymbol{x},\boldsymbol{y}) = \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x})P(C_k|\boldsymbol{y})}{P(C_k)}. \tag{50}$$

We call this function the discriminant kernel function (DKF).

The derived DKF is defined by using the Bayesian *a posterior* probabilities $P(C_k|\boldsymbol{x})$. This means that the class information is explicitly introduced in this kernel function.

B. Discriminant Kernel Functions

For real application, we have to estimate the Bayesian *a posterior* probabilities similar with the Bayesian decision theory, but this means that we can defined a family of discriminate kernel functions by changing the estimation method of the Bayesian *a posterior* probabilities.

There are many ways to estimate the Bayesian *a posterior* probabilities. Depending on the estimation method, we can define the corresponding DKF. Here two simple examples of the DKFs are shown. One is the DKF in which the Bayesian *a posterior* probabilities are estimated by assuming the probability densities of each class as multivariate normal distribution. The other is the DKF in which they are estimated by applying the K-nearest-neighbor density estimation technique to each class separately.

1) The Gaussian Disctibution: The one of the most simple methods to estimate the Bayesian *a posterior* probabilities is to assume the probability densities of each class as multivariate normal distribution. If the Bayesian *a posterior* probabilities are estimated from the training samples, we can easily define the discriminant kernel function by using the equation (50).

If the probability densities $p(\boldsymbol{x}|C_k)$ of each class C_k can be defined as multivariate normal $N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)$, that is

$$N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_k)^T\Sigma_k(\boldsymbol{x}-\bar{\boldsymbol{x}}_k)\right] \tag{51}$$

and the parameters $\bar{\boldsymbol{x}}_k$ and Σ_k are estimated from the training samples, the Bayesian *a posterior* probabilities are given by

$$P(C_k|\boldsymbol{x}) = \frac{P(C_k)N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)}{p(\boldsymbol{x})}, \tag{52}$$

where the probability density of $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} P(C_k)N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k). \tag{53}$$

This is the most simple way to estimate the Bayesian *a posterior* probabilities and is known as parametric method.

Then the corresponding DKF is given as

$$K_{Gauss}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} P(C_k)\frac{N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)N(\boldsymbol{y}|\bar{\boldsymbol{x}}_k, \Sigma_k)}{p(\boldsymbol{x})p(\boldsymbol{y})}. \tag{54}$$

For more complicated distributions, we can estimate the probability densities $p(\boldsymbol{x}|C_k)$ by using the mixtures of Gaussians as

$$p(\boldsymbol{x}|C_k) = \sum_{j=1}^{J_k} \pi_j^{(k)} N(\boldsymbol{x}|\bar{\boldsymbol{x}}_j^{(k)}, \Sigma_j^{(k)}) \tag{55}$$

with the components $N(\boldsymbol{x}|\bar{\boldsymbol{x}}_j^{(k)}, \Sigma_j^{(k)})$, where $\bar{\boldsymbol{x}}_j^{(k)}$ and $\Sigma_j^{(k)}$ are their own mean and covariance.

2) Nearest-neighbor: We can also estimate the Bayesian *a posterior* probabilities by applying the K-nearest-neighbor density estimation technique to each class separately.

Assume we have a training data set comprising N_k samples in the class C_k with N samples in total. Consider a sphere centered on $\boldsymbol{x}$ containing precisely K samples in the sphere. Suppose this sphere has volume $V(\boldsymbol{x})$ and contains $K_k(\boldsymbol{x})$ samples from the class C_k. Then the estimate of the probability density associated with the class C_k is estimated by

$$p(\boldsymbol{x}|C_k) = \frac{K_k(\boldsymbol{x})}{N_k V(\boldsymbol{x})}. \tag{56}$$

Similarly the unconditional density is given by

$$p(\boldsymbol{x}) = \frac{K}{NV(\boldsymbol{x})}. \tag{57}$$

The probabilities of each class are also given by

$$P(C_k) = \frac{N_k}{N}. \tag{58}$$

Using Bayes's theorem, the Bayesian *a posterior* probabilities are obtained as

$$P(C_k|\boldsymbol{x}) = \frac{K_k(\boldsymbol{x})}{K}. \tag{59}$$

Thus the DKF for this estimation method is given by

$$K_{k-NN}(\boldsymbol{x}, \boldsymbol{y}) = \frac{N}{K^2}\sum_{k=1}^{K}\frac{K_k(\boldsymbol{x})K_k(\boldsymbol{y})}{N_k}. \tag{60}$$

This is one of the examples of the non-parametric estimation of the DKF. Similarly we can estimate the Bayesian *a posterior* probabilities by using Kernel density estimators and use them to define the DKF.

V. Experiments

To investigate the validity of the proposed Discriminant Kernel Functions, the kernel matrix was computed for Fisher's Iris data set. The dataset consists of 50 samples from each of three species of Iris flowers. Four features are measured from each sample, they are the length and the width of sepal and petal, in centimeters. Here the data was rearranged such that the first 25 samples are from the class C_1, the next 25 samples are from the class C_2, the next 25 samples are from the class C_3, the next 25 samples are from the class C_1, and so on.

The Bayesian *a posterior* probabilities are estimated by assuming the probability densities of each class as multivariate normal distribution. The estimated kernel matrix is shown in Figure 1 (a). It is noticed that kernel function values of samples with the same class label are high and they are low for the samples with different class labels.

The kernel matrix computed by using the Radial Basis functions is also shown in Figure 1 (b). The kernel parameter σ was set to 1.0. By comparing these two figures of the kernel matrices, the kernel matrix estimated by the proposed method gives the clear discrimination between the different classes and the values are more deterministic, namely differences

between the high values and the low values are clearer than the kernel matrix of the Radial Basis functions.

The kernel matrix computed by using the linear model is also shown in Figure 1 (c). Since the LDA is the linear approximation of the ONDA in which the Bayesian *posterior* probabilities are estimated by using linear model, this kernel matrix corresponds to the one used in LDA. The discrimination between the classes of this kernel matrix is also not clear.

Comparisons of these kernel matrices shows that the proposed discriminant kernel function gives better discrimination than the Radial Basis Kernel or the kernel used in LDA.

The discriminant space constructed by using the kernel functions shown in Figure 1 are shown in Figure 2. Figure 2 (a) shows the discriminant space constructed by using the discriminant kernel function estimated by the proposed method assuming the Normal Distribution. It is noticed that the samples of the class C_1 gather on a single point and the samples from the class C_2 and C_3 are aligned on a straight line.

The discriminant space constructed by KDA is shown in Figure 2 (b), where the Radial Basis Kernel ($\sigma = 1$) is used as the kernel function. The configuration of the samples is roughly similar with the discriminant space constructed by using the proposed discriminant kernel function. But the samples from the class C_1 are scattered and the samples from the class C_2 and C_3 are not aligned on a straight line.

The discriminant space constructed by LDA is also shown in Figure 2 (c). In this case, the samples are scattered more than the other two discriminant spaces.

From these comparisons, we can say that the discriminant kernel function proposed in this paper is effective for discriminant analysis.

VI. Conclusions

Based on the theory on the ONDA, the optimum kernel function was derived by investigating the optimum discriminant mapping constructed by the ONDA. Since the ONDA is optimum in terms of the discriminant criterion, the proposed discriminant kernel function is optimum in terms of the discriminant criterion. The derived kernel function is defined by using the Bayesian *a posterior* probabilities. This means that the class information is naturally introduced in the proposed kernel function.

For real application, we have to estimate the Bayesian *a posterior* probabilities similar with the Bayesian decision theory, but this means that we can defined a family of discriminate kernel functions by changing the estimation method of the Bayesian *a posterior* probabilities. In this paper, two simple examples of the DKFs are shown. One is the DKF in which the Bayesian *a posterior* probabilities are estimated by assuming the probability densities of each class as multivariate normal distribution. The other is the DKF in which they are estimated by applying the K-nearest-neighbor density estimation technique to each class separately. Since the theory of the discriminant kernel functions is general, any other methods to estimate the Bayesian *a posterior* probabilities can be utilized to define the discriminant kernel functions.

The effectiveness for discrimination of the proposed DKF was confirmed by comparing the kernel matrix computed by the proposed DKF with the one computed by the Radial Basis Kernel. Also the discriminant space constructed by the proposed DKF was better than the one constructed by using the Radial Basis Kernel.

As the future works, we would like to investigate the relations between the effectiveness of the discrimination and the estimation methods of the Bayesian *a posterior* probabilities. Also we would like to show the merit of the proposed DKF as the kernel function of the support vector machines (SVM).

References

[1] R.A.Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, Vol.7, pp.179–188, 1936.

[2] S.Mika, G.Ratsch, J.Weston, B.Scholkopf, A.Smola, and K.Muller, "Fisher discriminant analysis with kernels," Proc. IEEE Neural Networks for Signal Processing Workshop, pp.41-48, 1999.

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

[4] S.Akaho, "Kernel Multivariate Data Analysis," Iwanami Shoten, 2008 (in Japanease).

[5] N.Otsu, "Nonlinear discriminant analysis as a natural extension of the linear case," Behavior Metrika, Vol.2, pp.45-59, 1975.

[6] N.Otsu, "Mathemetical Studies on Feature Extraction In Pattern Recognition," Researches on the Electrotechnical Laboratory, Vol.818, 1981 (in Japanease).

[7] N.Otsu, "Optimal linear and nonlinear solutions for least-square discriminant feature extraction," Proceedings of the 6th International Conference on Pattern Recognition, pp557-560, 1982.

[8] C.K.Chow, "An optimum character recognition system using decision functions," IRE Trans., Vol.EC-6, pp.247–254, 1957.

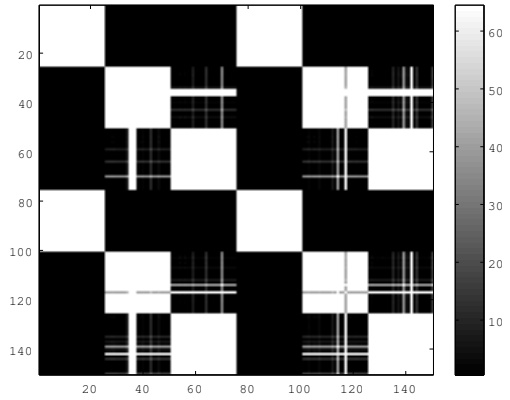
[9] T.Kurita, K.Watanabe, and N.Otsu, " Logistic Discriminant Analysis, !GG Proc. of 2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, Texas, USA., October 11-14, pp.2236-2241, 2009.

[10] D.W.Ruck, S.K.Rogers, M.Kabrisky, M.E.Oxley and B.W.Suter, "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," IEEE Transactions on Neural Networks, Vol.1, pp.296-298, 1990.

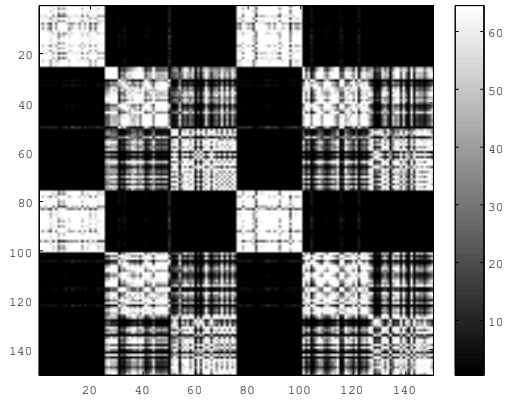
[11] T.Kurita, H.Asoh and N.Otsu, "Nonlinear discriminant features constructed by using outputs of multilayer perceptron," Proceeding of the International Symposium on Speech, Image Processing and Neural Networks (ISSIPNN' 94), vol.2, pp.417-420, 1994.

[12] K.Fukunaga and S.Ando, "The optimum nonlinear features for a scatter criterion in discriminant analysis," IEEE Transactions on Information Theory, Vol. 23, pp.453- 459, 1977.

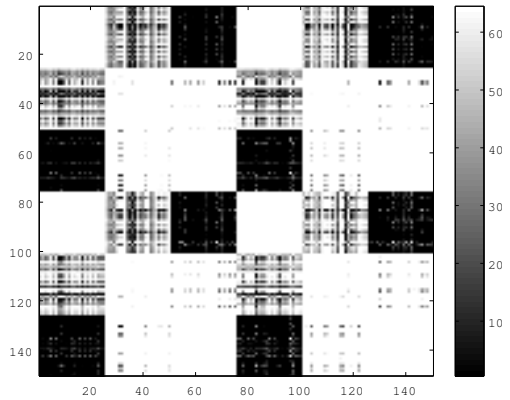
[13] P.McCullagh, J.Nelder, "Generalized Linear Models," Chapman and Hall, 1989.



(a) Kernel matrix estimated by the proposed method assuming the Normal Distribution

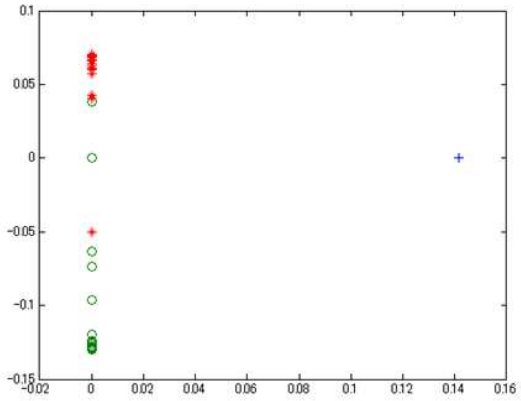


(b) Kernel matrix computed by the Radial Basis Kernel ($\sigma = 1$)

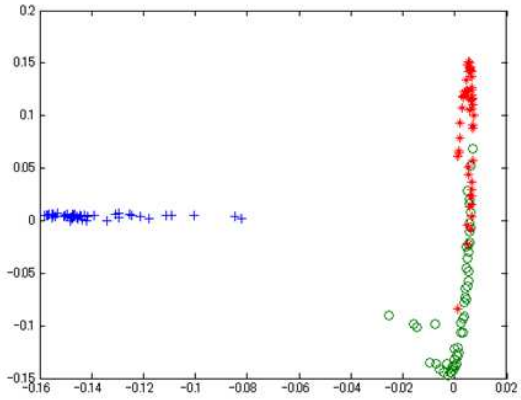


(c) Kernel matrix estimated by using the linear model (Kernel matrix used in LDA)

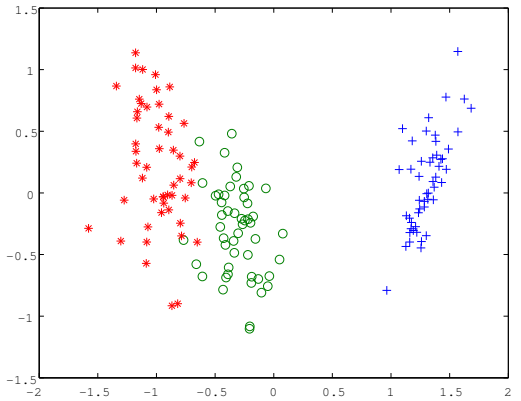
Fig. 1. Comparison of kernel matrices



(a) Discrimination space constructed by using the discriminant kernel function estimated by the proposed method assuming the Normal Distribution



(b) The discriminant space constructed by KDA (the Radial Basis Kernel with $\sigma = 1$)



(c) The discriminant space constructed by LDA

Fig. 2. Comparison of the discriminant spaces.