IEICE TRANS. INF. & SYST., VOL.E99–D, NO.11 NOVEMBER 2016

PAPER

Optimum Nonlinear Discriminant Analysis and Discriminant Kernel Support Vector Machine

Akinori HIDAKA[†a)], ***Member and*** **Takio KURITA**[††], ***Fellow***

SUMMARY Kernel discriminant analysis (KDA) is the mainstream approach of nonlinear discriminant analysis (NDA). Since it uses the kernel trick, KDA does not consider its nonlinear discriminant mapping explicitly. In this paper, another NDA approach where the nonlinear discriminant mapping is analytically given is developed. This study is based on the theory of optimal nonlinear discriminant analysis (ONDA) of which the nonlinear mapping is exactly expressed by using the Bayesian posterior probability. This theory indicates that various NDA can be derived by estimating the Bayesian posterior probability in ONDA with various estimation methods. Also, ONDA brings an insight about novel kernel functions, called discriminant kernel (DK), which is defined by also using the posterior probabilities. In this paper, several NDA and DK derived from ONDA with several posterior probability estimators are developed and evaluated. Given fine estimation methods of the Bayesian posterior probability, they give good discriminant spaces for visualization or classification.

key words: *discriminant analysis, nonlinear discriminant analysis, kernel method, support vector machine, discriminant kernel*

1. Introduction

Fisher's linear discriminant analysis (FLDA) [6] is one of the well known methods of extracting the best discriminating features for multi-class classification. FLDA is formulated as a problem of finding an optimum linear mapping which maximizes a discriminant criterion defined as a ratio of within-class scatter and between-class scatter in the mapped discriminant feature space.

FLDA is useful for linear separable cases, but for more complicated cases, it should be improved to express nonlinear boundaries. There are several approaches to obtain nonlinear discriminant analysis (NDA) by extending FLDA. NDA based on neural network (NN) model were studied in the 1990's [7], [14], [22], and kernel-based discriminant analysis (KDA) has been the mainstream since 2000 [1], [16], [19]. By using the kernel trick, KDA efficiently gives nonlinear, high-dimensional and discriminative feature space without explicitly knowing a nonlinear discriminant mapping $\Phi(\boldsymbol{x})$ for an input feature $\boldsymbol{x}$. Instead of the map Φ, the kernel functions which can be implicitly represented as $K(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x})^T \Phi(\boldsymbol{x}')$ are considered.

There are several approaches to determine the kernel function K. The most simple way is to use *a priori* functions such as the polynomial kernel or the radial bases functions (RBF) kernel. Since these functions are primarily not designed for the tasks of classification or discrimination, this approach usually has some unavoidable problems: How to select the kernel function? Is the selected function really suitable to the problem?

On the other hand, there are many studies on the discriminative kernel approach in which the user's notion or knowledge about a target problem is incorporated into kernel functions [9], [12], [20], [21]. For example, Fisher kernel [12] and Tsuda's marginalized kernel (MK) [20], [21] are strong frameworks to integrate a generative approach and a discriminative approach. However, in those methods, usually the kernel function K is directly defined based on the user's notion of similarity, without considering the nonlinear mapping Φ explicitly.

As one of the drawbacks of the kernel approach, hyperparameters in kernel functions should be tuned with expensive computational costs. To handle these problems, many papers studied optimizing kernel functions, for instances [13], [24], [26]. However, those studies also did not directly turn their attention to the nature of the nonlinear discriminant mapping Φ.

As far as we know, there have been few reports on nonlinear discriminant analysis which are not based on the common (NN or KDA) approaches. As a few exceptions, Otsu proposed optimum nonlinear discriminant analysis (ONDA) [17], [18] of which the exact expression of the optimum nonlinear discriminant mapping Φ was derived based on variational calculus. ONDA is closely related to Bayesian decision theory [3]. Interestingly, the optimal mapping Φ can be defined as a linear combination of the Bayesian *posterior* probabilities, and their coefficients are obtained by solving the eigenvalue problem of the matrices defined by using the Bayesian posterior probabilities. Also Otsu pointed out that FLDA could be interpreted as a linear approximation of ONDA through the linear approximations of the Bayesian posterior probabilities. These results are fundamental to understand the nature of the discriminant analysis.

It is important to note that ONDA is based on the ideal probabilistic assumption. Hence, it is required to know all of the Bayesian posterior probability $P(C|\boldsymbol{x})$ of class C, which are usually regarded as *an output* of classification tasks, to obtain the discriminant mapping Φ. However, when viewed from a different angle, this theory of ONDA also suggests

Manuscript received February 22, 2016.
Manuscript revised June 21, 2016.
Manuscript publicized August 4, 2016.

[†]The author is with Tokyo Denki University, Saitama-ken, 350–0394 Japan.
[††]The author is with Hiroshima University, Higashihiroshima-shi, 739–8521 Japan.
a) E-mail: hidaka.akinori@mail.dendai.ac.jp

DOI: 10.1587/transinf.2016EDP7081

that various NDA can be constructed depending on specific approximation or estimation methods of the posterior probabilities [11], [14].

Based on this suggestion, in this paper, we develop several NDA which have the exact expression of the discriminant map Φ, and evaluate their performance. We use two estimation methods of the Bayesian posterior probability; the Gaussian model as a simple estimator and support vector machine (SVM) as a more complex estimator. We show some properties and discrimination performance of our Gaussian NDA and SVM NDA by using standard benchmark data sets [8]. Our NDA have good discrimination and visualization performance compared with RBF KDA when sufficiently good estimation of the Bayesian posterior probability could be obtained.

The theory of ONDA also brings us another insight about novel kernel functions. By investigating the dual problem of the eigenvalue equation of ONDA, kernel functions which can be interpreted as being incorporated in ONDA is derived from the optimum mapping Φ. The derived kernel function, called discriminant kernel (DK) [15], is also defined by using the posterior probabilities. This means that the class information is naturally introduced in this kernel. Since ONDA is optimum in terms of the discriminant criterion, DK can be considered as the kernel function which is designed to optimize the discriminant criterion.

According to this theory, we develop a novel DK, called linear discriminant kernel (LDK), which is derived from linear approximation of ONDA. LDK can be regarded as the kernel function which is implicitly incorporated in FLDA. Also, we develop another DK derived from Gaussian NDA, called Gaussian DK [10]. In this paper, we also show a theoretical relationship between DK and Tsuda's MK [21], and compare the performance of Gaussian DK and Gaussian MK.

In experiments, we use DK and MK as the kernel function of SVM. We evaluate and compare the performance of DK SVM, MK SVM and usual (linear and RBF kernel) SVM by using standard benchmark data set [8].

The rest of this paper is organized as follows: Section 2 reviews FLDA, KDA, ONDA and MK. Then our ONDA-based NDA are described in Sect. 2. The discriminant kernel functions and our DK SVM are introduced in Sect. 4. The experiments are described in Sect. 5. Finally, Sect. 6 concludes the paper.

2. Discriminant Analysis

2.1 Fisher's Linear Discriminant Analysis

Fisher's linear discriminant analysis (FLDA) [6] is one of the well known methods to extract the best discriminating features for multi-class classification. FLDA 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.

Consider K classes denoted by $C = \{C_1, \ldots, C_K\}$. Assume that we have n training samples $\{\boldsymbol{x}_i, t_i\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbf{R}^{m\times 1}$ is an m dimensional feature vector and $t_i \in C$ is a class label. Then FLDA 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 the within-class and the between-class covariance matrix of the new feature vectors $\boldsymbol{y}$, respectively.

The discriminant criterion J is the objective function of FLDA to obtain the optimal coefficient A in Eq. (1). This criterion can be maximized by solving the following generalized eigenvalue problem

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

where Λ is a diagonal matrix of eigenvalues, I denotes the unit matrix, and Σ_T and Σ_B are the total and the between-class covariance matrix of the input feature vectors $\boldsymbol{x}$, respectively.

2.2 Kernel Discriminant Analysis

There are several approaches to obtain nonlinear discriminant analysis (NDA) by extending FLDA. Especially, the kernel discriminant analysis (KDA) seems to be one of the most powerful and popular approaches [1], [16], [19].

Let us consider a nonlinear mapping Φ from an input feature vector $\boldsymbol{x}$ to the new feature vector $\Phi(\boldsymbol{x})$.

By considering $\boldsymbol{\alpha}_i = [\alpha_1, \cdots, \alpha_n]^T$ which is a coefficient vector for the sample $\boldsymbol{x}_i$, the kernel discriminant mapping is given by

$$\boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i K(\boldsymbol{x}_i, \boldsymbol{x}) = A^T \boldsymbol{k}(\boldsymbol{x}) \tag{4}$$

where

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}_i), \tag{5}$$

$$A = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n)^T, \tag{6}$$

$$\boldsymbol{k}(\boldsymbol{x}) = [K(\boldsymbol{x}_1, \boldsymbol{x}), \cdots, K(\boldsymbol{x}_n, \boldsymbol{x})]^T. \tag{7}$$

The optimum coefficient matrix A is obtained by solving the eigenvalue problem

$$\Sigma_B^{(K)} A = \Sigma_W^{(K)} A \Lambda \tag{8}$$

where $\Sigma_B^{(K)}$ and $\Sigma_T^{(K)}$ are the total and the between-class covariance matrix of the kernel feature vector $\boldsymbol{k}(\boldsymbol{x})$, respectively.

Typically, the kernel function $K(\boldsymbol{x}, \boldsymbol{x}')$ is defined a priori. The polynomial kernel $K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}' + h)^p$ and the RBF kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{\sigma}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right)$ are often used. Generally kernel functions have several hyper-parameters such as p and h in the polynomial kernel or σ in the RBF kernel.

In real applications, those kernel functions must be manually selected without theoretical validity, and their hyper-parameters must be experimentally determined with expensive computational costs. This is seen as one of the drawbacks of the kernel approach. Also, the kernel functions are primarily not designed for the tasks of classification or discrimination, namely they do not contain in themselves any information about target objects or classes.

In order to deal with those drawbacks, there are many studies to incorporate user's notion or knowledge about a target problem into kernel functions [9], [12], [20], [21]. Fisher kernel [12] embeds appropriate distance metric into a space of probability distributions. Marginalized kernel [20], [21] is defined by using the *posterior* distribution $p(\boldsymbol{h}|\boldsymbol{x})$ where $\boldsymbol{x}$ and $\boldsymbol{h}$ are visible and hidden variable, respectively. It can incorporate information or knowledge of target problems in the kernel, by estimating $p(\boldsymbol{h}|\boldsymbol{x})$ from given training samples.

However, KDA approach usually consider not the nonlinear mapping Φ but the kernel function K. In this article, we show a novel viewpoint about NDA and kernel approach via paying attention to the non-linear mapping Φ.

2.3 Optimal Nonlinear Discriminant Analysis

KDA efficiently gives their nonlinear feature space by considering the kernel functions $K(\boldsymbol{x}, \boldsymbol{x}')$ instead of the nonlinear discriminant mapping Φ. The articles directly considering the nature of the nonlinear discriminant mapping Φ is limited [17], [18], [25].

As the few exceptions, Otsu proposed optimal nonlinear discriminant analysis (ONDA) [17], [18]. By assuming the ideal probabilistic condition similar to the Bayesian decision theory, the exact expression of the optimal nonlinear mapping $\Phi(\boldsymbol{x})$ which maximizes the discriminant criterion can be derived.

Let us consider the dimension reducing nonlinear mapping

$$\boldsymbol{y} \approx \Phi(\boldsymbol{x}). \tag{9}$$

Similar to FLDA, ONDA constructs the dimension reducing optimum nonlinear mapping which maximizes the discriminant criterion

$$J = \mathrm{tr}(\hat{\Sigma}_T^{-1}\hat{\Sigma}_B) \tag{10}$$

where $\hat{\Sigma}_T$ and $\hat{\Sigma}_B$ are the total covariance and the between covariance of $\boldsymbol{y}$, respectively. They are computed as

$$\Sigma_T = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)^T, \tag{11}$$

$$\Sigma_B = \sum_{k=1}^{K} P(C_k)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T, \tag{12}$$

where $P(C_k)$, $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{x}}_T$ denote the prior probability of the class C_k, the mean vector of the class C_k and the total mean vector, respectively. Typically we compute the probability of the class C_k as $P(C_k) = \frac{n_k}{n}$, where n_k is the number of the training samples of the class C_k.

By using variational calculus, the optimal nonlinear discriminant mapping can be obtained as

$$\boldsymbol{y} = \sum_{k=1}^{K} P(C_k|\boldsymbol{x})\boldsymbol{u}_k \tag{13}$$

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

$$\Gamma U = \Psi U \Lambda \tag{14}$$

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

$$\begin{aligned}\gamma_{ij} &= \int (P(C_i|\boldsymbol{x}) - P(C_i))(P(C_j|\boldsymbol{x}) - P(C_j))p(\boldsymbol{x})d\boldsymbol{x} \\ &= \int P(C_i|\boldsymbol{x})P(C_j|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} - P(C_i)P(C_j) \\ &= \gamma(C_i, C_j) - P(C_i)P(C_j).\end{aligned} \tag{15}$$

$\gamma(C_i, C_j) = \int p(\boldsymbol{x})P(C_i|\boldsymbol{x})P(C_j|\boldsymbol{x})d\boldsymbol{x}$ can be regarded as the similarity between the class C_i and C_j. The other matrices in Eq. (14) are defined as

$$U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]^T, \tag{16}$$

$$\Psi = \mathrm{diag}\,(P(C_1), \ldots, P(C_K)), \tag{17}$$

$$\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_L). \tag{18}$$

Then the optimal nonlinear discriminant mapping (9), (13) can be rewritten as

$$\boldsymbol{y} = \Phi(\boldsymbol{x}) = U^T\boldsymbol{B}(\boldsymbol{x}) \tag{19}$$

where

$$\boldsymbol{B}(\boldsymbol{x}) = [P(C_1|\boldsymbol{x}), \cdots, P(C_K|\boldsymbol{x})]^T \tag{20}$$

is a vector of the Bayesian posterior probability. This means that the optimum mapping can be interpreted as a linear combination of the posterior probabilities.

Interestingly, the optimum nonlinear discriminant mapping Φ from a given input feature $\boldsymbol{x}$ to the new discriminant feature $\boldsymbol{y}$ obtained by ONDA is equivalent to the linear mapping A obtained by applying FLDA to the posterior probability vector $\boldsymbol{B}(\boldsymbol{x})$ instead of $\boldsymbol{x}$ (see [15] for details).

As shown in the Eqs. (13) and (19), we can construct the optimum mapping Φ if we know $P(C_k|\boldsymbol{x})$ for all classes $C_1, \cdots, C_K$. This means that we must approximate or estimate the posterior probabilities for real applications. It also

implies various nonlinear discriminant mapping can be defined by changing the estimation methods of the posterior probabilities.

It is important to note that the estimation of the posterior probabilities is also very important in the context of the discriminant analysis like in the Bayesian decision theory.

2.4 Linear Approximation of ONDA

According to the study of Otsu [18], FLDA can be regarded as the linear approximation of ONDA; the linear discriminant mapping of FLDA can be interpreted as a linear approximation of the nonlinear discriminant mapping of ONDA through the linear approximations of $P(C_k|\boldsymbol{x})$.

Consider a linear approximation of the Bayesian posterior probabilities $P(C_k|\boldsymbol{x})$ as follows:

$$P(C_k|\boldsymbol{x}) \approx L(C_k|\boldsymbol{x}) = \boldsymbol{b}_k^T \boldsymbol{x} + b_k^{(0)}. \tag{21}$$

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

$$\epsilon^2 = \int (P(C_k|\boldsymbol{x}) - L(C_k|\boldsymbol{x}))^2 p(\boldsymbol{x}) d\boldsymbol{x}. \tag{22}$$

The optimum linear approximation of $P(C_k|\boldsymbol{x})$ which minimizes the mean squared errors is given by

$$L(C_k|\boldsymbol{x}) = P(C_k)\left[(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T \Sigma_T^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_T) + 1\right] \tag{23}$$

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

$$\sum_{k=1}^{K} L(C_k|\boldsymbol{x}) = 1. \tag{24}$$

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 (13) and (14). By this substitution, the Eq. (13) becomes

$$\boldsymbol{y} = \sum_{k=1}^{K} L(C_k|\boldsymbol{x})\boldsymbol{u}_k = U^T \Psi M^T \Sigma_T^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_T) + U^T \boldsymbol{\psi} \tag{25}$$

where

$$M = [(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_T), \cdots, (\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_T)]^T, \tag{26}$$

$$\boldsymbol{\psi} = [P(C_1), \cdots, P(C_K)]^T. \tag{27}$$

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. (14) of ONDA becomes

$$\Gamma = \Psi M^T \Sigma_T^{-1} M \Psi. \tag{28}$$

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

$$\Sigma_B A = \Sigma_T A \Lambda. \tag{29}$$

This is the same as the eigenvalue problem (3) of FLDA. This means that the linear mapping $A^T\boldsymbol{x}$ of FLDA can be considered as the linear approximation of the nonlinear mapping $U^T B(\boldsymbol{x})$ of ONDA through the linear approximation $P(C_k|\boldsymbol{x}) \approx L(C_k|\boldsymbol{x})$.

3. Nonlinear Discriminant Analysis Based on ONDA

As far as we know, there has been few reports about NDA which are not based on the kernel approach. In this paper, we develop another approach for NDA, based on the approximation or the estimation of the Bayesian posterior probability in ONDA.

As described in the previous section, linear approximation of $P(C_k|\boldsymbol{x})$ in ONDA becomes equivalent to FLDA. Since FLDA is suitable to only linear separable problems, $P(C_k|\boldsymbol{x})$ should be approximated or estimated based on nonlinear ways when we treat more complex problems. In this section, we introduce nonlinear estimation methods of the posterior probability to construct more reliable NDA than FLDA for more complicated problems.

3.1 Gaussian NDA

One of the most simple methods to estimate the Bayesian posterior probabilities $P(C_k|\boldsymbol{x})$ is to assume the conditional probability densities of each class as multivariate Gaussian distribution. Let us consider multivariate Gaussian distribution

$$\mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)^T \Sigma_k^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)\right]}{\sqrt{(2\pi)^d |\Sigma_k|}} \tag{30}$$

where the parameters $\bar{\boldsymbol{x}}_k$ and

$$\Sigma_k = \frac{1}{n_k} \sum_{t_i = C_k} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)^T \tag{31}$$

have to be estimated from the given training samples.

If the conditional probability densities $P(\boldsymbol{x}|C_k)$ of each class C_k can be expressed as a Gaussian distribution (30), the Bayesian posterior probabilities are given by

$$P(C_k|\boldsymbol{x}) = \frac{P(C_k)p(\boldsymbol{x}|C_k)}{p(\boldsymbol{x})} \tag{32}$$

$$= \frac{P(C_k)\mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)}{\sum_{k=1}^{K} P(C_k)\mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)}. \tag{33}$$

By substituting this into Eq. (15) and solving the eigenvalue equation (14), the optimal coefficient matrix U in Eq. (19) is obtained. We call it Gaussian NDA [11].

3.2 SVM NDA

The Bayesian posterior probabilities $P(C_k|\boldsymbol{x})$ can also be estimated by using more advanced classification models. Wu et al. proposed probability estimation algorithm for multi-class classifier by pairwise coupling [23]. Their algorithm can estimate the posterior probabilities $P(C_k|\boldsymbol{x})$ by using a set of classifiers H_{ij} which classifies sample x into class C_i or class C_j. This estimation algorithm, in which support vector machine (SVM) is used as the classifier H_{ij}, is implemented in libsvm [4].

Let $P_{SVM}(C_k|\boldsymbol{x})$ be the estimated posterior probabilities by Wu's algorithm for a given input feature vector $\boldsymbol{x}$ which belongs to the class C_k. Then the posterior probability can be simply written as

$$P(C_k|\boldsymbol{x}) = P_{SVM}(C_k|\boldsymbol{x}). \tag{34}$$

We call NDA which is brought by this estimation by the name of SVM NDA.

Wu's probability estimation algorithm [23] is available in libsvm [4] by using the training option '-b 1'.

4. Discriminant Kernel

In the previous sections, we described our ONDA-based NDA. The theory of ONDA also brings us the idea of the novel kernel function [15]. By investigating the dual problem of the eigenvalue equation of ONDA, an optimum kernel function is derived from the optimum discriminant mapping (19).

4.1 Dual Problem of ONDA

By multiplying $\Psi^{-1/2}$ from the left of (14) which is the generalized eigenvalue problem of ONDA, it can be rewritten as the usual eigenvalue problem as

$$\Psi^{-1/2}\Gamma\Psi^{-1/2}\Psi^{1/2}U = \Psi^{1/2}U\Lambda. \tag{35}$$

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

$$(\Psi^{-1/2}\Gamma\Psi^{-1/2})\tilde{U} = \tilde{U}\Lambda. \tag{36}$$

Then the optimum nonlinear discriminant mapping of ODNA is rewritten as

$$\boldsymbol{y} = U^T\tilde{\boldsymbol{B}}(\boldsymbol{x}) = \tilde{U}^T\Psi^{-1/2}\tilde{\boldsymbol{B}}(\boldsymbol{x}) = \tilde{U}^T\boldsymbol{\phi}(\boldsymbol{x}) \tag{37}$$

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

For the case of n training samples, the eigenvalue problem to determine the class representative vectors (36) is given by

$$(\Phi^T\Phi)\tilde{U} = \tilde{U}\Lambda, \tag{38}$$

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

The dual eigenvalue problem of (38) is then given by

$$(\Phi\Phi^T)V = V\Lambda. \tag{39}$$

From the relation on the singular value decomposition of the matrix Φ, these two eigenvalue problems (38) and (39) have the same eigenvalues and there is the following relation between the eigenvectors $\tilde{U}$ and V as $\tilde{U} = \Phi^T V\Lambda^{-1/2}$.

By inserting this relation into the nonlinear discriminant mapping (37), we have

$$\begin{aligned}\boldsymbol{y} &= \Lambda^{-1/2}V^T\Phi\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n}\Lambda^{-1/2}\boldsymbol{v}_i\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) \\ &= \sum_{i=1}^{n}\boldsymbol{\alpha}_i K(\boldsymbol{x}_i, \boldsymbol{x}) - \boldsymbol{\alpha}_0\end{aligned} \tag{40}$$

where

$$\begin{aligned}&K(\boldsymbol{x}_i, \boldsymbol{x}) = \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) + 1 \\ &= \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x}_i) - P(C_k)(P(C_k|\boldsymbol{x}) - P(C_k))}{P(C_k)} + 1 \\ &= \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x}_i)P(C_k|\boldsymbol{x})}{P(C_k)}.\end{aligned} \tag{41}$$

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

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x})P(C_k|\boldsymbol{x}')}{P(C_k)}. \tag{42}$$

This is called the discriminant kernel (DK) function. This theory shows that ONDA can be interpreted as one of KDA using DK as the kernel function. DK is defined by using the Bayesian posterior probabilities. This means that the class information is directly introduced in our kernel.

Usually, kernel functions are designed to incorporate prior knowledges of target problems into classifiers. In this sense, it can be said that DK is naturally designed to maximize the discriminant criterion J (10). For real applications, we must approximate or estimate the posterior probabilities from given samples. Instead, since DK has no kernel parameters, we do not need to tune them.

4.2 Mercer's Condition

By using the Bayes' theorem (32), Eq. (42) can be rewritten as

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{K}P(C_k)\frac{P(C_k|\boldsymbol{x})}{p(\boldsymbol{x})}\frac{P(C_k|\boldsymbol{x}')}{p(\boldsymbol{x}')}. \tag{43}$$

Then, it can be expressed as the matrix form,

$$K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{D}(\boldsymbol{x})^T\Psi\boldsymbol{D}(\boldsymbol{x}') \tag{44}$$

where

$$\boldsymbol{D}(\boldsymbol{x}) = \left[\frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x})}, \cdots, \frac{p(\boldsymbol{x}|C_K)}{p(\boldsymbol{x})} \right]^T \tag{45}$$

is the vector of the likelihood ratio. Thus, the discriminant kernel can be interpreted as the inner product of the likelihood ratio weighted by Ψ which has the class prior $P(C_k)$ as the k-th diagonal component.

Since the diagonal matrix Ψ (17) has only non-negative components $P(C_k) \geq 0$, every eigenvalue of Ψ, $\lambda_k = P(C_k)$, is obviously non-negative. Therefore, the matrix Ψ is positive semidefinite. Also, Ψ is symmetric because it is a diagonal matrix. Thus, the kernel function (44) is a valid kernel (see [2]). In other words, the discriminant kernel is a Mercer kernel.

4.3 Relationship between DK and MK

Tsuda proposed the marginalized kernel (MK) [21],

$$K_M(\boldsymbol{x}, \boldsymbol{x}') = \sum_{h \in \mathcal{H}} \sum_{h' \in \mathcal{H}} p(h|\boldsymbol{x}) p(h'|\boldsymbol{x}') K_z(z, z'), \tag{46}$$

where h is a hidden variable in a finite set $\mathcal{H}$, $z = (\boldsymbol{x}, h)$ is a combined variable and $K_z(z, z')$ is a joint kernel function. The posterior probability $p(h|\boldsymbol{x})$ is unknown in general, and has to be estimated from given samples. By using the joint kernel

$$K_z(z, z') = I(h = h')(\boldsymbol{x}^T \Sigma_h^{-1} \boldsymbol{x}) \tag{47}$$

where Σ_h is a covariance matrix of the variable h and I is the indicator function

$$I(H) = \begin{cases} 1 & \text{if } H \text{ is true} \\ 0 & \text{otherwise,} \end{cases} \tag{48}$$

Equation (46) can be rewritten as

$$K_M(\boldsymbol{x}, \boldsymbol{x}') = \sum_{h \in \mathcal{H}} p(h|\boldsymbol{x}) p(h|\boldsymbol{x}') \boldsymbol{x}^T \Sigma_h^{-1} \boldsymbol{x}'. \tag{49}$$

By regarding the hidden variable h in Eq. (49) as the class label C_k, MK (49) can be rewritten as

$$K_M(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{K} P(C_k|\boldsymbol{x}) P(C_k|\boldsymbol{x}') \boldsymbol{x}^T \Sigma_k^{-1} \boldsymbol{x}' \tag{50}$$

where Σ_k is a covariance matrix calculated from samples in the class C_k. This function can be transformed into a matrix form using $\boldsymbol{B}$ in Eq. (20),

$$K_M(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{B}(\boldsymbol{x})^T W \boldsymbol{B}(\boldsymbol{x}') \tag{51}$$

where $W = \mathrm{diag}(\boldsymbol{x}^{\mathrm{T}} \Sigma_1^{-1} \boldsymbol{x}', \cdots, \boldsymbol{x}^{\mathrm{T}} \Sigma_{\mathrm{K}}^{-1} \boldsymbol{x}')$.

On the other hand, by using Eqs. (17) and (20), Eq. (42) can also be expressed as another matrix form,

$$K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{B}(\boldsymbol{x})^T \Psi^{-1} \boldsymbol{B}(\boldsymbol{x}'). \tag{52}$$

Equations (51) and (52) show the relationship between MK and DK; they are expressed in the same form that is the inner product of the vector of the Bayesian posterior probability with a certain weighting matrix (W or Ψ^{-1}). DK uses the weighting matrix Ψ^{-1}, which includes the prior information of each class C_k. On the other hand, MK uses the matrix W, which includes the information of normalized original features.

These weight matrices will indicate the characteristics of DK and MK; DK purely relies on the probabilistic information, $p(C_k|\boldsymbol{x})$ and $P(C_k)$. In other words, it no longer relies on the original feature $\boldsymbol{x}$. On the other hand, MK forms the intermediate expression between the posterior probability $p(C_k|\boldsymbol{x})$ and the original feature $\boldsymbol{x}$.

4.4 A Family of Discriminant Kernel Functions

Similar to the derivation of ONDA-based NDA, we can define various DK by changing the estimation method of the Bayesian posterior probability $P(C_k|\boldsymbol{x})$. In this paper, we develop two types of DK and evaluate their performance. The one is a new discriminant kernel which is derived from the linear approximation of the posterior probability of ONDA. The other one is derived from Gaussian NDA [11].

4.4.1 Linear Discriminant Kernel

By regarding $L(C_k|\boldsymbol{x})$ in Eq. (23) as the approximation of $P(C_k|\boldsymbol{x})$ in Eq. (42), we obtain

$$\begin{aligned} K_L(\boldsymbol{x}, \boldsymbol{x}') &= \sum_{k=1}^{K} \frac{L(C_k|\boldsymbol{x}) L(C_k|\boldsymbol{x}')}{P(C_k)} \\ &= \sum_{k=1}^{K} P(C_k) \left[(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T \Sigma_T^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_T) + 1 \right] \\ &\qquad \times \left[(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T \Sigma_T^{-1} (\boldsymbol{x}' - \bar{\boldsymbol{x}}_T) + 1 \right] \\ &= (\boldsymbol{x} - \bar{\boldsymbol{x}}_T)^T \Sigma_T^{-1} \Sigma_B \Sigma_T^{-1} (\boldsymbol{x}' - \bar{\boldsymbol{x}}_T) + 1. \end{aligned} \tag{53}$$

We call it linear discriminant kernel (LDK).

When we recall that FLDA can be given by linear approximation of ONDA through the linear approximation $P(C_k|\boldsymbol{x}) \approx L(C_k|\boldsymbol{x})$, LDK is considered as the kernel function which is implicitly used in FLDA.

As can be seen in Eq. (53), LDK could be regarded as one of the normalizing operation for input $\boldsymbol{x}$ and $\boldsymbol{x}'$ by the linear transformation using Σ_T and Σ_B. It seems to be consistent with the essence of FLDA to maximize the ratio of the between covariance and the total covariance.

4.4.2 Gaussian Discriminant Kernel

Similar to Gaussian NDA, by using multivariate Gaussian distribution $\mathcal{N}$ for the class C_k, the vector of the likelihood ratio (45) can be estimated as

$$\boldsymbol{D}_G(\boldsymbol{x}) = \left[\frac{\mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)}{p(\boldsymbol{x})}, \cdots, \frac{\mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)}{p(\boldsymbol{x})} \right]^T \tag{54}$$

where $p(\boldsymbol{x}) = \sum_{k=1}^{K} P(C_k) \mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)$. Thus, Gaussian DK

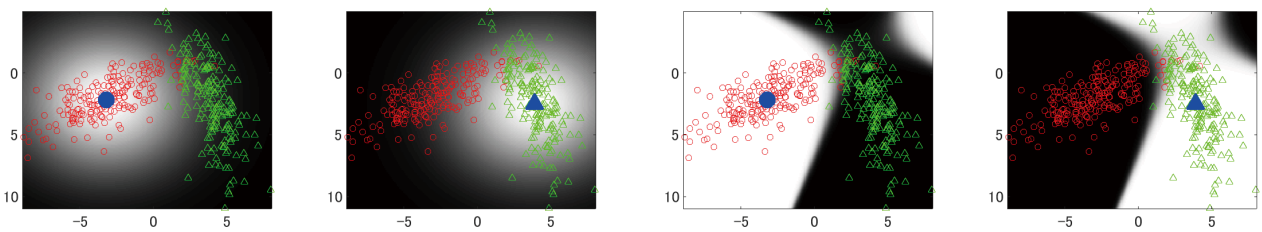

Fig. 1 The maps of kernel values from fixed points. The left two and the right two figures are the maps of RBF kernel and GDK, respectively. Circle and triangle markers indicate the training samples of class 1 and 2, respectively. The grayscale gradations show the kernel values, i.e. the similarity measure, between the class center (illustrated as the filled big markers) and each location. Brightness indicates high similarity.

Table 1 Statistics of the data sets

	Tic-Tac-Toe	German	Iris	Wine	Balance	Vehicle	Vowel
# of classes	2	2	3	3	3	4	11
# of samples	958	1000	150	178	625	846	990
# of features	9	24	4	13	4	18	10

Table 2 Comparisons of classification performance. Averaged classification rate (%), standard deviation and P values of paired t-test are shown. The classification rate of the winner is written in bold if the P value < 0.01 (=1.00E-02).

	Tic-Tac-Toe	German	Iris	Wine	Balance	Vehicle	Vowel
FLDA	57.6 (3.93)	72.1 (2.09)	83.2 (4.61)	98.6 (1.35)	57.5 (24.56)	76.2 (2.35)	49.1 (2.72)
Gaussian NDA	**74.2 (2.41)**	70.8 (2.12)	**97.3 (2.05)**	98.3 (1.91)	**67.6 (36.26)**	**80.8 (4.28)**	**84.8 (2.00)**
P value	1.25E-12	1.62E-02	1.01E-11	5.69E-01	1.70E-03	3.46E-04	2.22E-23
FLDA	57.6 (3.93)	72.1 (2.09)	83.2 (4.61)	**98.6 (1.35)**	57.5 (24.56)	**76.2 (2.35)**	49.1 (2.72)
LSVM NDA	54.9 (3.36)	**73.8 (1.64)**	**96.3 (1.79)**	96.8 (1.70)	**67.6 (36.30)**	71.2 (2.16)	**79.7 (2.80)**
P value	2.45E-02	1.68E-06	3.57E-11	6.13E-04	1.75E-03	9.69E-06	2.11E-19
RBF KDA	98.6 (0.65)	70.0 (1.50)	95.7 (2.82)	96.4 (2.09)	68.6 (29.40)	**81.6 (2.02)**	97.4 (1.06)
RBFSVM NDA	98.7 (0.82)	**73.9 (2.03)**	96.1 (1.80)	97.1 (1.70)	70.7 (39.68)	74.5 (2.74)	**98.1 (0.98)**
P value	4.53E-01	2.85E-08	5.51E-01	1.30E-01	1.40E-01	9.70E-10	4.37E-03

(GDK) can be obtained as

$$K_G(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{D}_G(\boldsymbol{x})^T \Psi \boldsymbol{D}_G(\boldsymbol{x}'). \tag{55}$$

To illustrate the property of GDK, we demonstrate a preliminary experiment using 2-dimensional artificial data where samples of class 1 and 2 are obtained from different Gaussian distributions. Figure 1 shows the maps of kernel value from fixed points. The left two figures indicate the similarity measure given by RBF kernel with $\sigma = 32$ which is the value manually tuned. Since RBF kernel assigns same σ to all samples regardless of their class labels, the similarity measure must become monotonous circle. On the other hand, the right two figures indicate the similarity measure given by our GDK. Since GDK can incorporate class information via the estimation of the Bayesian posterior probability using given training samples, the similarity measure given by GDK naturally express the decision boundary.

5. Experiments

We evaluated the performance of our NDA and DK SVM by using several data sets in UCI machine learning repository [8]: Tic-Tac-Toe, German, Iris, Wine, Vehicle and Vowel. We divided each data into a training set (2/3 of all samples) and a test set (remaining samples) at random. We made twenty different divisions of the training and test sets. All values shown in Tables 2-4 were calculated as the average of the results of twenty trials.

For all experiments, we used the class prior $P(C_k) = N_k/N$ where N_k is the number of samples in C_k. For RBF KDA and RBF SVM, the kernel parameter σ was tuned by grid search with 10-fold cross validation. The grid was set to $\sigma = 2^{-15}, 2^{-14}, \cdots, 2^{+15}$. Also, for linear and RBF SVM, the soft-margin parameter c was tuned by the same way. The grid was also set to $c = 2^{-15}, 2^{-14}, \cdots, 2^{+15}$. Thus, Linear SVM and RBF KDA were tuned over 31 grid points and RBF SVM was tuned over 961 points. All experiments were performed on a standard PC with the Intel Core i7 X980 CPU (3.33GHz) and 24GB RAM.

For several classification experiments, we used paired t-test to evaluate statistical significance of comparison results. In this article, if the P value of the t-test is smaller than 0.01 (1.0E-2), we consider that the corresponding mean values are statistically different. Also, we consider that there is the possibility that the means are statistically different if $0.01 \leq \mathrm{P} \leq 0.05$.

Table 3 Comparisons of classification performance. Averaged classification rate (%), standard deviation and P values of paired t-test are shown. The classification rate of the winner is written in bold if the P value < 0.01 (=1.00E-02).

	Tic-Tac-Toe	German	Iris	Wine	Balance	Vehicle	Vowel
Linear SVM	65.3 (0.87)	75.7 (1.89)	96.6 (2.19)	97.7 (1.57)	91.4 (1.54)	**79.4 (2.39)**	**80.4 (2.39)**
LDK SVM	**67.0 (1.65)**	76.1 (1.96)	96.4 (2.14)	98.6 (1.73)	90.5 (1.67)	77.6 (1.88)	77.4 (2.84)
P values	6.16E-04	2.73E-01	6.94E-01	6.09E-02	8.80E-02	1.32E-04	4.85E-08
GMK SVM	**78.0 (1.90)**	73.8 (2.33)	97.3 (2.05)	98.1 (1.82)	91.0 (1.71)	83.3 (2.24)	**87.5 (1.90)**
GDK SVM	75.3 (2.49)	73.2 (2.01)	97.2 (2.06)	98.2 (1.87)	91.0 (1.43)	83.4 (2.31)	85.1 (1.94)
P values	2.52E-06	2.43E-02	3.30E-01	5.77E-01	6.73E-01	5.36E-01	9.36E-06
RBF SVM	98.7 (0.77)	75.2 (1.67)	96.6 (2.19)	97.5 (1.27)	98.8 (1.03)	83.4 (2.63)	98.3 (0.92)

Table 4 Averaged training time (sec) of usual SVM and DK SVM, and its standard deviation

	Tic-Tac-Toe	German	Iris	Wine	Balance	Vehicle	Vowel
Linear SVM	1126 (84.3)	9212 (589.8)	0.46 (0.26)	0.69 (0.04)	363.2 (53.5)	1094 (74.4)	471.6 (49.8)
LDK SVM	335.2 (115.1)	452.9 (84.4)	0.59 (0.51)	0.38 (0.04)	11.9 (4.85)	254.0 (24.2)	464.5 (70.3)
GMK SVM	5950 (190.5)	14912 (438.2)	2.08 (0.27)	1.17 (0.05)	507.0 (202.4)	4193 (464.6)	1049 (49.4)
GDK SVM	93.1 (46.4)	596.7 (175.4)	0.39 (0.26)	0.32 (0.03)	52.4 (19.5)	97.0 (49.5)	125.1 (48.8)
RBF SVM	2075 (160.9)	2067 (100.6)	25.62 (0.16)	46.05 (0.17)	489.1 (11.4)	1593 (75.0)	1897 (72.5)

5.1 Comparison of Visualized Discriminant Space

To compare the property of FLDA, KDA and our NDA, the first 2 dimensions of the discriminant spaces are illustrated in Figs. 2 and 3. In Fig. 2, as one of a property of our NDA, it is noticed that each NDA space compose triangular forms. Since our NDA based on the posterior probability, i.e., the value from 0 to 1, their discriminant spaces form $K-1$ dimensional hyper-tetrahedron (simplex) for K classes problems. This property gives us a probabilistic interpretation such as how a sample is close or far to each class.

It is interesting to note that the 4-th and the 5-th columns in Fig. 3 probably show the typical difference between discriminant analysis and SVM. In the case of RBF KDA for the training set, almost all samples precisely locate at their class center. On the other hand, for the test set, RBF KDA shows the scattered distribution. In contrast, although RBFSVM NDA shows slightly perturbed distribution for the training set, the concentration to the class centers is still kept well in the test set. These differences were probably caused by the difference in the generalization performance of discriminant analysis and SVM; the training result of RBF KDA seems to be over-fitting.

5.2 Performance Evaluation of ONDA

Table 2 shows the average of the 20 trials of the classification rates for the test sets. The classification rate is the result of nearest mean classification in each discriminant spaces. The table also shows the statistical significance in which it is assumed that "*the average classification rates of corresponding two methods have no difference*," for several pairs of methods.

The statistical test showed that Gaussian NDA is better than FLDA for 5 data excepting German and Wine. For German (P = 0.0162), FLDA has a possibility that it is better than Gaussian NDA. For Wine (P = 0.569), their performance have no significant difference. Since the computation procedure to make Gaussian model (i.e., computation of the mean vector and the covariance matrix for each class) is usually a part of the computation to construct the linear discriminant mapping, Gaussian NDA might become a useful substitute for FLDA, especially for problems which depend on Gaussian distributions.

The test also showed that LSVM NDA is better for 4 data (German, Iris, Balance and Vowel) and worse for 2 data (Wine and Vehicle) than FLDA. For Tic-Tac-Toe (P = 0.0245), FLDA has a possibility that it is better than LSVM NDA. Also, RBFSVM NDA is better than RBF KDA for 3 data (German, Vehicle and Vowel), and for other 4 data (Tic-Tac-Toe, Iris, Wine and Balance), their performance have no significant difference. Though our SVM NDA seem to work well, we of course require the expensive training of LSVM or RBF SVM with grid search and cross validation. Especially in the case of RBFSVM NDA, the improvements seem to be not sufficient to pay those additional costs.

5.3 Comparison of Visualized Kernel Matrices

In this section, we show the visualized kernel matrices of DK, MK and usual RBF kernel. For a comparison of DK and MK, we use Gaussian DK (GDK) and Gaussian MK (GMK) in which the Bayesian posterior probability in Eqs. (52) and (51) are estimated by a simple Gaussian model. The hyper-parameter of RBF kernel, σ, was tunned by the grid search and the cross validation for RBF SVM together with the soft margin c.

The left columns in Figs. 4 and 5 show the kernel matrices of *an ideal discriminant kernel*. These are computed from the ideal Bayesian posterior probability, namely, K dimensional vector whose k-th component is 1 and other components are 0 if the sample belongs to the class C_k. In this experiment, samples are sorted by their class number.

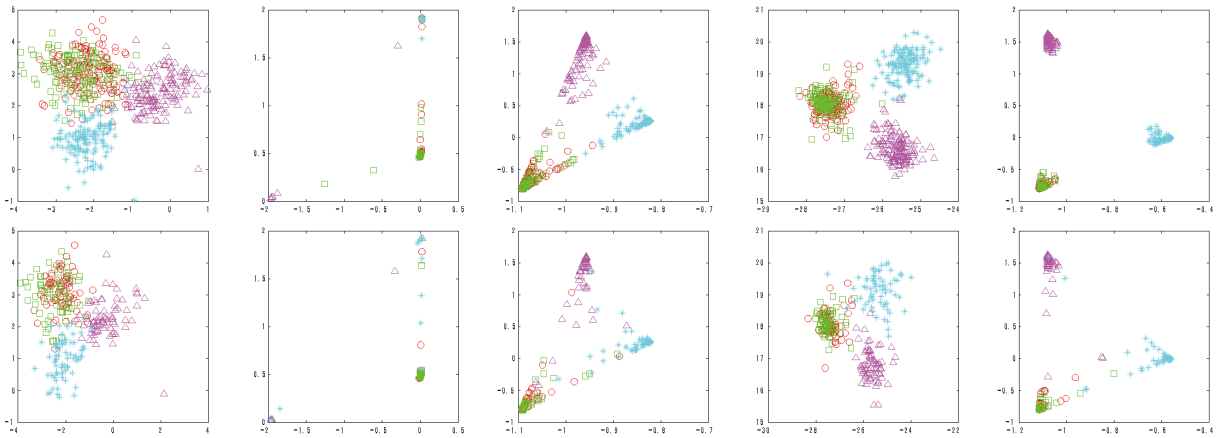

Fig. 2 Discriminant spaces (from left, FLDA, Gaussian NDA, LSVM NDA, RBF KDA and RBFSVM NDA) of Vehicle (Top: training, Bottom: test).

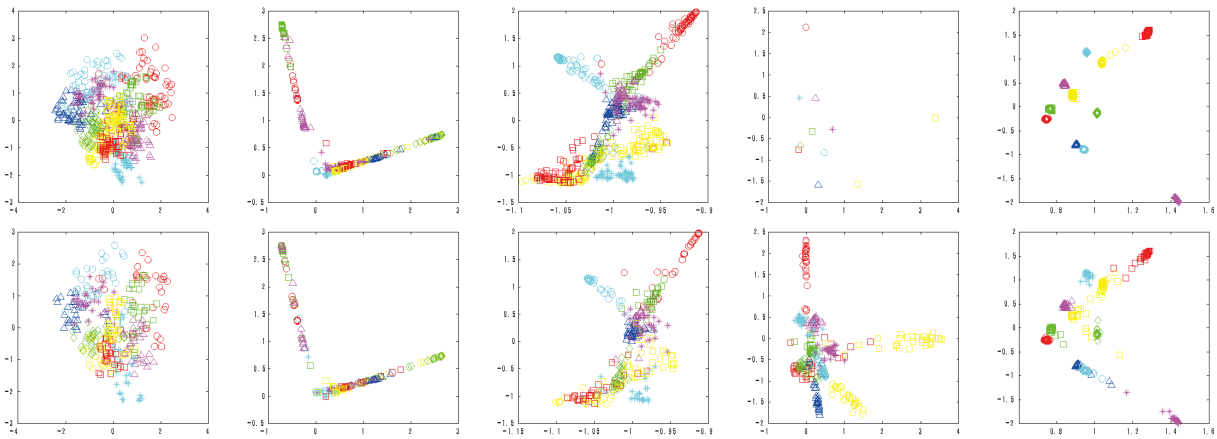

Fig. 3 Discriminant spaces (from left, FLDA, Gaussian NDA, LSVM NDA, RBF KDA and RBFSVM NDA) of Vowel (Top: training, Bottom: test).

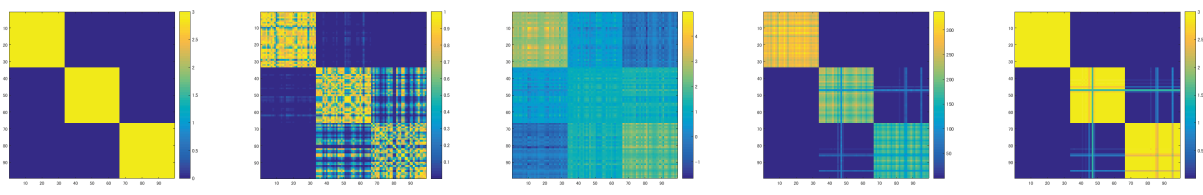

Fig. 4 Kernel matrices (from left, ideal DK, RBF kernel, LDK, GMK and GDK) of Iris (training).

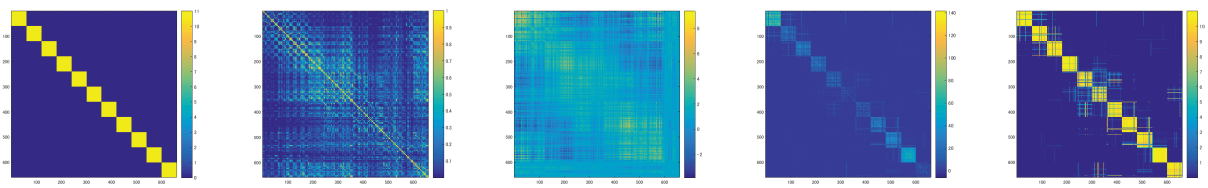

Fig. 5 Kernel matrices (from left, ideal DK, RBF kernel, LDK, GMK and GDK) of Vowel (training).

Therefore, the diagonal block in the ideal DK shows the group of each class.

We can see that the kernel matrices of RBF and LDK show complicated responses, and the diagonal blocks which imply the groups of classes are unclear. On the other hand, GMK and GDK show clear responses similar to the ideal DK. The class structures clearly illustrated in GMK and GDK will be helpful to analyze the relationship between classes visually.

In the figure, also we can see the difference of the properties between MK and DK; Since GDK consists of only the information of the Bayesian posterior probability, it is more similar to the ideal DK than GMK. On the other hand, since the intermediate form between the posterior probability and the original feature, GMK seems to include more practical information for classification tasks.

5.4 Performance Evaluation of DK SVM

In this section, we describe the performance evaluation for usual (linear and RBF) SVM, MK SVM and our DK SVM. In this experiment, we mainly focus on the comparison of two pairs of similar classifiers; (1) Linear SVM and LDK SVM, and (2) GMK SVM and GDK SVM.

Table 3 shows the averaged classification rates and the statistical tests for them. The experimental results showed that (1) LDK SVM is better for Tic-Tac-Two and worse for Vehicle and Vowel than Linear SVM. For other 4 data (German, Iris, Wine and Balance), the performance of LDK SVM and linear SVM have no significant difference. In this comparison, their performance seem to be in the almost same level.

The tables also showed that (2) GMK SVM is better than GDK SVM for Tic-Tac-Toe and Vowel. For German (P = 0.0243), GMK SVM has a possibility that it is better than Gaussian GDK SVM. And for other data (Iris, Wine, Balance and Vehicle), their performance have no significant difference. Though the difference of the averages were relatively small (less than 3%), the performance of GMK SVM seems to be consistently equal to or greater than GDK SVM.

Though it is not directly related to the discussion here, RBF SVM showed the best performance in many cases; For Tic-Tac-Toe, Balance and Vowel, RBF SVM is clearly better than other SVM. Meanwhile, RBF kernel of course needs expensive training costs to tune the hyper-parameter σ.

Table 4 shows the average of the training time of the grid search with 10-fold CV for each SVM. Note that these times will depend on the implementation and usage of SVM library. We used LibSVM [4] for Matlab with probability option '-b 1'. For Linear SVM and RBF SVM, we use kernel option '-t 0' and '-t 2', respectively. Also for GDK and GMK SVM, we use pre-computed kernel option '-t 4'. Note that we did not perform any scaling or normalizing process for original feature $\boldsymbol{x}$.

In this situation, about the grid search training, LDK was faster than linear SVM in many cases. Perhaps, one of the cause of this result is the scaling or normalizing effect of LDK which is described in Sect. 4.4.1. Also, grid search for GDK SVM is faster than GMK SVM in all cases. This is probably because our GDK SVM has strongly normalized kernel matrices in which each component takes the value in $[0, K]$.

Due to their low training cost and moderate performance, LDK SVM and GDK SVM may become good substitutes for linear SVM.

6. Conclusions

In this article, we showed a novel viewpoint of NDA and kernel methods by paying attention to the non-linear mapping Φ. We developed ONDA-based NDA where the nonlinear discriminant mapping $\Phi(\boldsymbol{x})$ can be analytically obtained. Also, we developed discriminant kernel (DK) which is the novel kernel function derived from ONDA. Our NDA and DK are defined by using the Bayesian *posterior* probability. It implies that our NDA and DK explicitly include class information of target problems.

Given fine estimation methods of the Bayesian posterior probability, our NDA and DK can be used as the good visualization tools to illustrate the probabilistic relationship between classes. Gaussian NDA, LDK SVM and GDK SVM showed modest classification accuracies in spite of their low training costs. Thus, for simple objectives (e.g. problems which have simple distribution, preliminary experiments of new data, and so on), they may be used as good substitutes for FLDA or linear SVM.

Also, we theoretically and experimentally showed the relationship between our DK and the special case of Tsuda's MK; They are both formulated as the weighted inner product of the vector $\boldsymbol{B}$ which consists of the Bayesian posterior probability, but their weight matrices are different from each other. For the generalization ability, Gaussian MK may be better than Gaussian DK because MK relies on both of the probabilistic information and original feature information while DK relies only on the probabilistic information.

These findings may bring a novel direction to develop kernel methods; We can consider and may develop a family of kernel functions $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{B}(\boldsymbol{x})^T \mathcal{W} \boldsymbol{B}(\boldsymbol{x}')$ where $\mathcal{W}$ is an arbitrary weight matrix calculated from original feature $\boldsymbol{x}$.

While maximizing Fisher's discriminant criterion, our NDA and DK do not consider any other criterions related with the generalization ability. Thus, there will be room to incorporate some kind of mechanisms to improve such ability. In our methods, we can introduce regularization terms or criterions by two different ways; To introduce it into the matrix form of our NDA or DK directly, or into the estimation algorithm (such as Gaussian model, SVM, logistic regression, etc....) for the Bayesian posterior probability. As Mika et al. [16] pointed out, the regularization term introduced into the covariance matrix of KDA improves the performance of KDA. Also, Clemmensen et al. [5] proposed sparse LDA which is based on L1-regularization. Such regularizations will also be helpful for our methods. Also we

might be able to consider and develop a way to improve such ability by appropriately designing the weight matrix $\mathcal{W}$.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23500211.

References

[1] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," Neural Computation, vol.12, no.10, pp.2385–2404, 2000.
[2] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[3] C.K. Chow, "An optimum character recognition system using decision functions," IRE Trans., vol.EC-6, no.4, pp.247–254, 1957.
[4] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm
[5] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," Technometrics, vol.53, no.4, pp.406–413, 2011.
[6] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, vol.7, no.2, pp.179–188, 1936.
[7] V. Fontaine, C. Ris, and J.-M. Boite, "Nonlinear discriminant analysis for improved speech recognition," Proc. Eurospeech-97, vol.4, pp.2071–2074, 1997.
[8] A. Frank and A. Asuncion, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]," University of California, School of Information and Computer Science.
[9] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," Tenth IEEE International Conference on Computer Vision (ICCV 2005), vol.2, pp.1458–1465, 2005.
[10] A. Hidaka and T. Kurita, "Discriminant Kernels based Support Vector Machine," The First Asian Conference on Pattern Recognition (ACPR 2011), pp.159–163, Beijing, China, Nov. 28-30, 2011.
[11] A. Hidaka and T. Kurita, "Nonlinear Discriminant Analysis based on Probability Estimation by Gaussian Mixture Model," Proc. of IAPR Joint International Workshops on Statistical Techniques in Pattern Recognition (SPR2014) and Structural and Syntactic Pattern Recognition (SSPR2014), vol.8621, pp.133–142, 2014.
[12] T.S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," NIPS, vol.11, pp.487–493, 1999.
[13] S.-J. Kim, A. Magnani, and S. Boyd, "Optimal kernel selection in kernel fisher discriminant analysis." In Int. Conf. Machine Learning, pp.465–472, 2006.
[14] 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.
[15] T. Kurita, "Discriminant Kernels derived from the Optimum Nonlinear Discriminant Analysis," Proc. of 2011 International Joint Conference on Neural Networks, pp.299–306, San Jose, California, USA, July 31 - Aug. 5, 2011.
[16] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, and K.R. Muller, "Fisher discriminant analysis with kernels," Proc. IEEE Neural Networks for Signal Processing Workshop, pp.41–48, 1999.
[17] N. Otsu, "Nonlinear discriminant analysis as a natural extension of the linear case," Behavior Metrika, vol.2, pp.45–59, 1975.
[18] N. Otsu, "Optimal linear and nonlinear solutions for least-square discriminant feature extraction," Proc. 6th International Conference on Pattern Recognition, pp557–560, 1982.
[19] V. Roth and V. Steinhage, "Nonlinear discriminant analysis using kernel functions," in Advances in Neural Information Processing Systems, ed. S.A. Solla, T.K. Leen, and K.-R. Muller, vol.12, pp.568–574, MIT Press, Cambridge, MA, 1999.
[20] K. Tsuda, "The fisher kernel and beyond," IEICE Technical Report, PRMU2001-108, Oct. 2001.
[21] K. Tsuda, T. Kin, and K. Asai, "Marginalized kernels for biological sequences," Bioinformatics, 18(90001), pp.S268–S275, 2002.
[22] A.R. Webb and D. Lowe, "The Optimised Internal Representation of Multilayer Classifier Networks Performs Nonlinear Discriminant Analysis," Neural Networks, vol.3, no.4, pp.367–375, 1990.
[23] T.-F. Wu, C.-J. Lin, and R. Weng, "Probability estimates for multiclass classification by pairwise coupling," J. Machine Learning Research 5, pp.975–1005, Aug. 2004.
[24] H. Xiong, M.N.S. Swamy, and M.O. Ahmad, "Optimizing the kernel in the empirical feature space." IEEE Trans. on Neural Networks, vol.16, no.2, pp.460–474, 2005.
[25] S. Yan, Y. Hu, D. Xu, H.-J. Zhang, B. Zhang, and Q. Cheng, "Nonlinear Discriminant Analysis on Embedded Manifold," IEEE Trans. Circuits Syst. Video Technol., vol.17, no.4, pp.468–477, April 2007.
[26] D. You and A.M. Martinez, "Bayes optimal kernel discriminant analysis." In Proceedings of the IEEE Computer Vision and Pattern Recognition, pp.3533–3538, 2010.

Akinori Hidaka received the B.Sci degree from Ibaraki University in 2004, and the M.Eng. and D.Eng. degree from the University of Tsukuba, in 2006 and 2009, respectively. He is currently an assistant professor of Tokyo Denki University, in Japan. His current research interests include computer vision based on statistical pattern recognition, especially object detection, tracking, generic object recognition and robot vision. He is a member of IEEE, IEICE and ISCIE.

Takio Kurita received the B.Eng. degree from Nagoya Institute of Technology and the Dr. Eng. degree from the University of Tsukuba, in 1981 and in 1993, respectively. He joined the Electrotechnical Laboratory, AIST, MITI in 1981. From 1990 to 1991 he was a visiting research scientist at Institute for Information Technology, National Research Council Canada. From 2001 to 2009, he was a deputy director of Nueroscience Research Institute, National Institute of Advanced Industrial Science and Technology (AIST). Also he was a Professor at Graduate School of Systems and Information Engineering, University of Tsukuba from 2002 to 2009. He is currently a Professor at Hiroshima University. His current research interests include statistical pattern recognition and its applications to image recognition. He is a member of the IEEE, the IPSJ, the IEICE of Japan, Japanese Neural Network Society, The Japanese Society of Artificial Intelligence.