Discriminant Kernels based Support Vector Machine

Akinori Hidaka
Tokyo Denki University

Takio Kurita
Hiroshima University

***Abstract*—Recently the kernel discriminant analysis (KDA) has been successfully applied in many applications. KDA is one of the nonlinear extensions of Linear Discriminant Analysis (LDA). But the kernel function is usually defined a priori and it is not known what the optimum kernel function for nonlinear discriminant analysis is.**

Otsu derived the optimum nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities similar with the Bayesian decision theory. Kurita derived discriminant kernels function (DKF) as the optimum kernel functions in terms of the discriminant criterion by investigating the optimum discriminant mapping constructed by the ONDA. The derived kernel function is given by using the Bayesian *posterior* probabilities. For real applications we can define a family of discriminant kernel functions by changing the estimation method of the Bayesian *posterior* probabilities.

In this paper, we propose and evaluate the support vector machine (SVM) in which the discriminant kernel functions are used. We call this SVM the discriminat-kernel-based support vector machine (DKSVM). In the experiments, we compare the proporsed DKSVM with the usual SVM.

I. Introduction

LINEAR discriminant analysis (LDA) [5] is one of the well known methods to extract the best discriminating features for multi-class classification. 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 [10], [2], [1]. 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, and it is not known what the best kernel function for nonlinear discriminant analysis (NDA) 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 [11], [12], [13] similar with the Bayesian decision theory[3]. He showed that the optimum nonlinear 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 *posterior* probabilities and the coefficients of the linear combination are obtained by solving the eigenvalue problem of the matrices defined by using the Bayesian *posterior* probabilities.

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. [8] proposed Logistic discriminant analysis (LgDA) in which the *posteriori* probabilities are estimated by using the multinominal logistic regression instead of the linear model.

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. [7] proposed the neural network based NDA in which the outputs of the trained multilayered Perceptron (MLP) were used as the estimates of the *posteriori* probabilities because the outputs of the trained MLP for classification problems can be regarded as the approximations of the *posteriori* probabilities [14].

Kurita showed that 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 [9]. The derived kernel function, called the discriminant kernel function (DKF), is also given by using the *posteriori* probabilities. This means the class information is naturally introduced in the kernel function. As like ONDA, the DKF is also optimum in terms of the discriminant criterion. Kurita also showed that a family of DKFs can be defined by changing the estimation method of the Bayesian *posterior* probabilities [9].

Since the discriminant kernel function is optimum in terms of the discriminant criterion, it is expected to be effective for other classification models such as support vector machine (SVM). In this paper, we propose to use the DKF as the kernel function of the SVM. We call this the discriminat-kernel-based support vector machine (DKSVM). In the experiments, we compare the porposed DKSVM with the usual SVM using several data sets in UCI machine learning repository.

The rest of this paper is organized as follows: Section II reviews LDA, KDA and ONDA. The discriminant kernel functions are introduced in Section III. Section IV shows our DKSVM. The experiments are described in Section V. Finally, Section VI concludes the paper.

II. Optimal Nonlinear Discriminant Analysis

A. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [5] is defined as a method to find the linear combination of features which best separates two classes of objects. LDA is regarded as one of the well known methods to extract the best discriminating features for multi-class classification.

Let an $m-$D 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 $\boldsymbol{y} = A^T \boldsymbol{x}$ where $A = [a_{ij}]$ is the coefficient matrix.

The objective of LDA is to maximize the discriminant criterion $J = \text{tr}(\hat{\Sigma}_T^{-1} \hat{\Sigma}_B)$ 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 optimal coefficient matrix A is then obtained by solving the following generalized eigenvalue problem

$$\Sigma_B A = \Sigma_T A \Lambda \quad (A^T \Sigma_T A = I) \tag{1}$$

where $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_L)$ is a diagonal matrix of eigen values and I shows 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 $\boldsymbol{x}$.

B. Kernel Discriminant Analysis

The kernel discriminant analysis (KDA) is one of the nonlinear extensions of LDA. Consider a nonlinear mapping Φ from a input feature vector $\boldsymbol{x}$ to the new feature vector $\Phi(\boldsymbol{x})$. For the case of $1-$D feature extraction, the discriminant mapping can be given as $y = \boldsymbol{a}^T \Phi(\boldsymbol{x})$. Since the coefficient vector $\boldsymbol{a}$ can be expressed as a linear combinations of the training samples as $\boldsymbol{a} = \sum_{i=1}^N \alpha_i \Phi(\boldsymbol{x}_i)$, the discriminant mapping can be rewritten as

$$y = \sum_{i=1}^N \alpha_i \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}) = \sum_{i=1}^N \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\alpha}^T \boldsymbol{k}(\boldsymbol{x}), \tag{2}$$

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

$$J = \frac{\sigma_B^2}{\sigma_T^2} = \frac{\boldsymbol{\alpha}^T \Sigma_B^{(K)} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \Sigma_T^{(K)} \boldsymbol{\alpha}}, \tag{3}$$

where σ_T^2 and σ_B^2 are respectively the total variance and the between-class variance of the discriminant feature y, and $\Sigma_T^{(K)}$ and $\Sigma_B^{(K)}$ are respectively the total covariance matrix and the between-class covariance matrix of the kernel feature vector $\boldsymbol{k}(\boldsymbol{x})$ (details are denoted in [9]).

The optimum coefficient vector $\boldsymbol{\alpha}$ can be obtained by solving the generalized eigenvalue problem $\Sigma_B^{(K)} \boldsymbol{\alpha} = \Sigma_W^{(K)} \boldsymbol{\alpha} \lambda$.

For the multi-dimension case, the kernel discriminant mapping is given by $\boldsymbol{y} = A^T \boldsymbol{k}(\boldsymbol{x})$, 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

$$\Sigma_B^{(K)} A = \Sigma_W^{(K)} A \Lambda. \tag{4}$$

Usually the kernel function is defined a priori in KDA. The polynomial functions $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^q$ or the Radial Basis functions $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-||\boldsymbol{x} - \boldsymbol{y}||^2 / 2\sigma^2)$ 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.

C. Optimal Nonlinear Discriminant Analysis

Otsu derived the optimal nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities [11], [12], [13]. 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 nonlinear discriminant mapping is obtained as

$$\boldsymbol{y} = \sum_{k=1}^K P(C_k | \boldsymbol{x}) \boldsymbol{u}_k \tag{5}$$

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

$$\Gamma U = P U \Lambda \tag{6}$$

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

$$\gamma_{ij} = \int (P(C_i | \boldsymbol{x}) - P(C_i))(P(C_j | \boldsymbol{x}) - P(C_j)) p(\boldsymbol{x}) d\boldsymbol{x} \tag{7}$$

and the other matrices are defined as

$$\begin{aligned} U &= [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]^T & (8) \\ P &= \text{diag}\,(P(C_1), \ldots, P(C_K)) & (9) \\ \Lambda &= \text{diag}\,(\lambda_1, \ldots, \lambda_L)\,. & (10) \end{aligned}$$

It is important to notice that the optimal nonlinear mapping is closely related to Bayesian decision theory, namely the *posterior* probabilities $P(C_k | \boldsymbol{x})$.

By using the eigen vectors obtained by solving the generalized eigenvalue problem (6), 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 (5) if we can know or estimate all the *posterior* probabilities. This means that we have to estimate the *posterior* probabilities for real applications. It also implies a family of nonlinear discriminant mapping can be defined by changing the estimation method of the *posterior* probabilities.

The important theoretical relationship between LDA, KDA and ONDA is described in [9] but omitted in this paper.

III. 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.

Kurita showed the optimum kernel function, called discriminant kernel function (DKF), can be derived by investigating the dual problem of the eigenvalue problem of ONDA [9]. The DKF is also optimum in terms of the discriminant criterion.

The eigenvalue problem of ONDA given by the equation (6) 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

$$P^{-1/2}\Gamma P^{-1/2}P^{1/2}U = P^{1/2}U\Lambda. \tag{11}$$

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

$$(P^{-1/2}\Gamma P^{-1/2})\tilde{U} = \tilde{U}\Lambda. \tag{12}$$

Then the optimum nonlinear discriminant mapping of ODNA is rewritten as

$$\boldsymbol{y} = U^T\tilde{\boldsymbol{B}}(\boldsymbol{x}) = \tilde{U}^T P^{-1/2}\tilde{\boldsymbol{B}}(\boldsymbol{x}) = \tilde{U}^T\boldsymbol{\phi}(\boldsymbol{x}) \tag{13}$$

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

For the case of N training samples, the eigenvalue problem to determine the class representative vectors (12) is given by

$$(\Phi^T\Phi)\tilde{U} = \tilde{U}\Lambda, \tag{14}$$

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

The dual eigenvalue problem of (14) is then given by

$$(\Phi\Phi^T)V = V\Lambda. \tag{15}$$

From the relation on the singular value decomposition of the matrix Φ, these two eigenvalue problems (14) and (15) 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 (13), we have

$$\begin{aligned}\boldsymbol{y} &= \Lambda^{-1/2}V^T\Phi\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{N}\Lambda^{-1/2}\boldsymbol{v}_i\phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) \\ &= \sum_{i=1}^{N}\boldsymbol{\alpha}_i K(\boldsymbol{x}_i, \boldsymbol{x}) - \boldsymbol{\alpha}_0 \end{aligned} \tag{16}$$

where

$$\begin{aligned} & K(\boldsymbol{x}_i, \boldsymbol{x}) = \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}) + 1 \\ = & \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x}_i) - P(C_k)(P(C_k|\boldsymbol{x}) - P(C_k))}{P(C_k)} + 1 \\ = & \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x}_i)P(C_k|\boldsymbol{x})}{P(C_k)}. \end{aligned} \tag{17}$$

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

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K}\frac{P(C_k|\boldsymbol{x})P(C_k|\boldsymbol{y})}{P(C_k)}. \tag{18}$$

This is called the discriminant kernel function (DKF).

The derived DKF is defined by using the Bayesian *posterior* probabilities $P(C_k|\boldsymbol{x})$. This means that the class information is explicitly introduced in this kernel function. Also there is no kernel parameters. This means that we do not need to estimate the kernel parameters.

B. Discriminant Kernel Functions

There are many ways to estimate the Bayesian *posterior* probabilities. Depending on the estimation method, we can define the corresponding DKF. Kurita proposed two examples of the DKFs which are based on the assumption of Gaussian distribution or the K-nearest-neighbor density estimation [9]. Likewise, the Bayesian *posterior* probabilities estimated by using the prediction result of support vector machine or multi-nominal logistic regression can be used to define a variant of the DKFs. In following subsections, DKFs based on two estimation methods are described.

1) Gaussian Discriminant Kernel Function: The one of the most simple methods to estimate the Bayesian *posterior* probabilities is to assume the probability densities of each class as multivariate Gaussian distribution. If the probability densities $p(\boldsymbol{x}|C_k)$ of each class C_k can be defined as multivariate Gaussian $N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)$, that is

$$N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}}\exp\left[-\frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)^T\Sigma_k^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)\right] \tag{19}$$

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

$$P(C_k|\boldsymbol{x}) = \frac{P(C_k)N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)}{p(\boldsymbol{x})}, \tag{20}$$

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

$$p(\boldsymbol{x}) = \sum_{k=1}^{K}P(C_k)N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k). \tag{21}$$

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

Then the corresponding DKF (Gaussian-DKF) is given as

$$K_{Gauss}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K}P(C_k)\frac{N(\boldsymbol{x}|\bar{\boldsymbol{x}}_k, \Sigma_k)N(\boldsymbol{y}|\bar{\boldsymbol{x}}_k, \Sigma_k)}{p(\boldsymbol{x})p(\boldsymbol{y})}. \tag{22}$$

2) SVM Discriminant Kernel Function: The Bayesian *posterior* probabilities can also be estimated by using the SVM classifier. Wu et al. proposed the Bayesian *posterior* probability estimation algorithm by SVM for multi-class problem [15]. For the input vector $\boldsymbol{x}$, their algorithm outputs probabilistic vector $(P_1(\boldsymbol{x}), \cdots, P_k(\boldsymbol{x}))$ as the estimation of the Bayesian *posterior* probabilities $(P(C_1|\boldsymbol{x}), \cdots, P(C_k|\boldsymbol{x}))$.

Then the corresponding DKF (SVM-DKF) is given as

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K}\frac{P_k(\boldsymbol{x})P_k(\boldsymbol{y})}{P(C_k)}. \tag{23}$$

TABLE I
COMPARISON OF AVERAGED CLASSIFICATION ACCURACY FOR 10 TEST SETS

	breast-cancer	german	splice	iris	wine	vehicle	vowel
# of classes	2	2	2	3	3	4	11
# of samples	683	1000	3175	150	178	846	990
# of features	10	24	60	4	13	18	10
LSVM	96.84 %	75.95 %	84.17 %	96.27 %	98.00 %	79.47 %	79.82 %
LSVM-DKF	96.80 %	75.74 %	84.31 %	96.47 %	97.83 %	79.50 %	80.52 %
Gaussian-DKF	94.74 %	73.69 %	88.50 %	98.04 %	98.67 %	84.61 %	86.10 %

TABLE II
COMPARISON OF AVERAGED CLASSIFICATION ACCURACY FOR 10 TEST SETS

	breast-cancer	german	splice	iris	wine	vehicle	vowel
RBFSVM	96.41 %	74.90 %	91.10 %	96.86 %	96.67 %	84.28 %	98.16 %
RBFSVM-DKF	96.53 %	75.25 %	91.03 %	96.86 %	97.17 %	84.07 %	98.18 %

Wu's probability estimation algorithm [15] is available in libsvm [4] by using the training option ’-b 1’.

IV. DISCRIMINANT KERNEL SVM

Once the discriminant kernel function is derived, the DKF can be used as a kernel function of any kernel based approach. In this paper, we propose discriminant kernel based support vector machines (DKSVM). We introduce two DKSVMs as follows:

1) DKSVM using Gaussian-DKF: In this method, Gaussian-DKF (eq. (22)) is used as the kernel function of SVM.

2) DKSVM using SVM-DKF: In this method, SVM-DKF (eq. (23)) is used as the kernel function of SVM. In the case, the training of SVM has two stage; (i) Linear or some kernel SVM which is used to calculate SVM-DKF is trained by cross validation and grid search to optimize the soft margin and kernel parameters; (ii) SVM using SVM-DKF is trained by cross validation and grid search to optimize the soft margin. In the experiment, linear SVM and RBF SVM are used in stage (i).

V. EXPERIMENTS

We confirmed the performance of DKSVMs by using several data sets in UCI machine learning repository [6]: Breast-cancer, german, splice, iris, wine, vehicle and vowel data. We divided each data into a training set (2/3 of all samples) and a test set (remaining samples) at random. For classification experiments, we made ten different divisions of the training and test sets. For all experiments, we used class prior $P(C_k) = N_k/N$ where N_k is the number of samples in C_k.

At first, we compare the classification accuracies of linear SVM (LSVM), LSVM-DKF SVM and Gaussian-DKF SVM. For each SVM, the soft margin c is optimized by grid search where $c = 2^{-20}, 2^{-19}, 2^{-18}, \cdots, 2^{19}, 2^{20}$. Each grid is evaluated by 9-fold cross validation. The averaged classification accuracies for 10 test sets are shown in Table I. LSVM and LSVM-DKF show almost same performance. In the results of splice, vehicle and vowel data, Gaussian-DKF shows clearly higher performance against other two methods.

Next, we compare the classification accuracy of RBF SVM and RBF-DKF SVM. For RBF SVM, the soft margin c and the kernel parameter g are optimized by grid search where $c = 2^{-30}, 2^{-27}, 2^{-24}, \cdots, 2^{27}, 2^{30}$ and $g = 2^{-20}, 2^{-18}, 2^{-16}, \cdots, 2^{18}, 2^{20}$. Each grid is evaluated by 4-fold cross validation. The accuracies are shown in Table II. RBF SVM and RBFSVM-DKF show almost same performance.

At last, we visualize the kernel matrices of LSVM, Gaussian-DKF, RBF SVM and RBF-DKF which are calculated from four data sets: Breastcancer, iris, wine, vowel. The comparisons of LSVM vs Gauss DKF and RBFSVM vs RBF-DKF are respectively shown in Fig. 1 and Fig. 2. The DKF matrices show the structures of classes more clearly than the linear or RBF kernel matrices.

VI. CONCLUSIONS

In this paper, we proposed the support vector machine (SVM) in which the discriminant kernel functions are used. The classification experiments show that the discriminant kernel gives the same levels of the classification performance with the tuned general kernels such as the linear or RBF kernels of the usual SVM. The visualization results of the kernel matrices show that DKSVMs have more clear kernel matrices against linear or RBF kernel.

REFERENCES

[1] S.Akaho, “Kernel Multivariate Data Analysis,” Iwanami Shoten, 2008 (in Japanease).

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

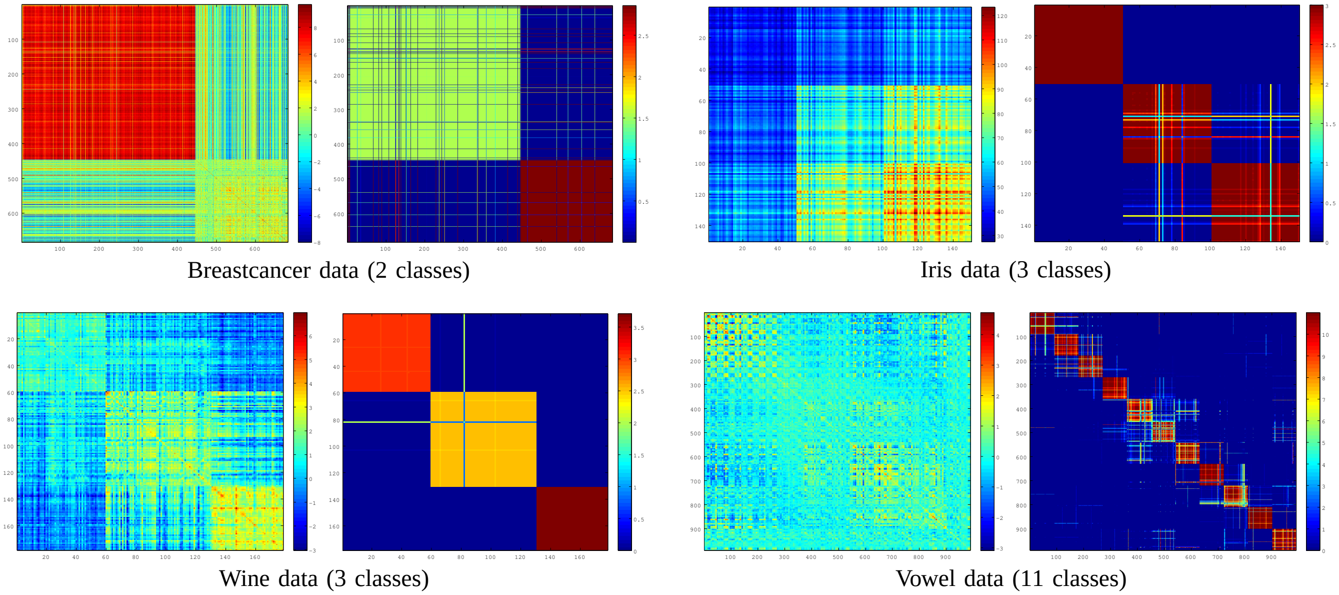
[3] C.K.Chow, “An optimum character recognition system using decision functions,” IRE Trans., Vol.EC-6, pp.247–254, 1957.

[4] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: a library for support vector machines”, 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

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

[6] A. Frank, A. Asuncion, “UCI Machine Learning Repository [http://archive.ics.uci.edu/ml],” University of California, School of Information and Computer Science.

[7] 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.



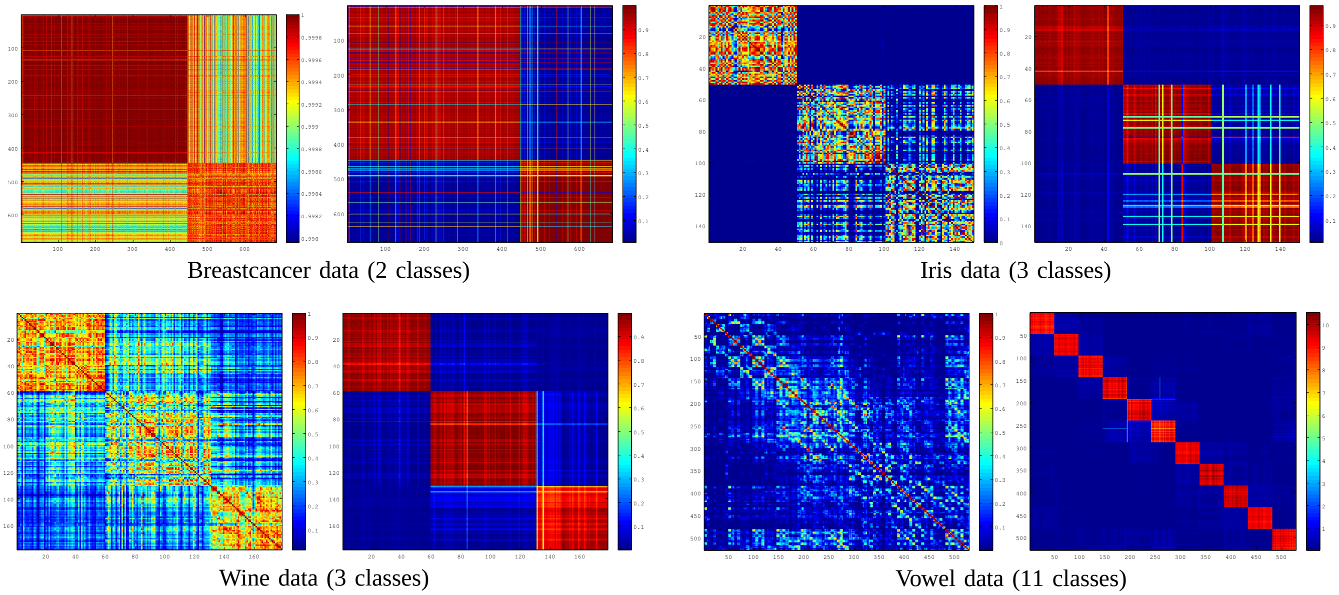
Breastcancer data (2 classes)

Iris data (3 classes)

Wine data (3 classes)

Vowel data (11 classes)

Fig. 1. Comparison of the kernel matrices of linear kernel (left) and Gaussian-DKF (right)



Breastcancer data (2 classes)

Iris data (3 classes)

Wine data (3 classes)

Vowel data (11 classes)

Fig. 2. Comparison of the kernel matrices of RBF kernel (left) and RBF-DKF (right)

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

[9] T.Kurita, "Discriminant Kernels derived from the Optimum Nonlinear Discriminant Analysis," Proc. of 2011 International Joint Conference on Neural Networks, San Jose, California, USA, July 31 - August 5, 2011 (accepted).

[10] 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.

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

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

[13] 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.

[14] 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.

[15] T.-F. Wu, C.-J. Lin and R. Weng, "Probability estimates for multi-class classification by pairwise coupling," Journal of Machine Learning Research 5 (August) (2004), pp. 975-1005.