Nonlinear Discriminant Analysis Based on Probability Estimation by Gaussian Mixture Model

Akinori Hidaka and Takio Kurita

[1] School of Science and Engineering, Tokyo Denki University, Saitama, Japan
[2] Department of Information Engineering, Hiroshima University, Hiroshima, Japan

Abstract. The Bayesian a posterior probability is a very important element in pattern recognition. In classification problems, the posterior probabilities reflect the uncertainty of assessing an example to particular class. Such residual information will be useful for more deep understanding or analysis of examples. In this paper, we propose a nonlinear discriminant analysis based on the probabilistic estimation of the Gaussian mixture model (GMM). We use GMM to estimate the Bayesian a posterior probabilities of any classification problems. Then we use posterior probabilities estimated by GMM to construct discriminative kernel function. The performance of the proposed kernel function is confirmed by several experiments using UCI machine learning repository.

Keywords: Fisher's Linear Discriminant Analysis, Gaussian Mixture Model, Bayesian a posterior probabilities, Discriminant Kernel.

1 Introduction

The Bayesian *a posterior* probability is a very important element in pattern recognition. The task that classifies unknown example $\boldsymbol{x}$ can be interpreted as the maximization procedure to the posterior probability $P(C_k|\boldsymbol{x})$ which implies the probability that $\boldsymbol{x}$ belongs to the $k-$th class C_k. Furthermore, in classification problems, the posterior probabilities reflect the uncertainty of assessing an example $\boldsymbol{x}$ to the class C_k. Such residual information will be useful for more deep understanding or analysis of examples.

There are many ways to estimate the Bayesian *a posterior* probabilities. Naive Bayes [16] is one of the most simple probabilistic classifier. Logistic regression is a generalized linear model and it has saturated outputs which is suitable to represent probabilities [2]. Several classifiers can also perform the estimation of the posterior probability simultaneously with the classification task. Wu et al. proposed how to presume the posterior probability from the output of SVM [15]. The one of the most efficient methods to estimate the Bayesian *a posterior* probabilities $P(C_k|\boldsymbol{x})$ is to assume the probability densities of each class as multivariate Gaussian distribution. To treat multi-modal distributions, Gaussian mixture model is widely used many real application [3,17].

P. Fränti et al. (Eds.): S+SSPR 2014, LNCS 8621, pp. 133–142, 2014.
© Springer-Verlag Berlin Heidelberg 2014

Fisher's Linear discriminant analysis (FLDA) [4] is one of the well known methods to extract the best discriminating features for multi-class classification. FLDA is useful for linear separable cases, but for more complicated cases, it is necessary to extend it to nonlinear.

As one of the nonlinear extensions of FLDA, kernel discriminant analysis (KFDA) has been successfully applied in many applications [9,1]. The polynomial kernel, sigmoidal kernel or radial basis function (RBF) are popular and widely used. However these functions are defined a priori and selected without the clear reason. Also these functions are general and not related to probabilistic inference.

In recent years, discriminant kernel function (DKF) which is based on the Bayesian *a posterior* probability estimation is proposed [8]. This kernel is derived from the theory of optimum nonlinear discriminant analysis (ONDA) [11,12]. Since ONDA gives the optimum nonlinear mapping that maximizes the Fisher's discriminant criterion [4], the DKF derived from ONDA is also optimum in terms of the discriminant criterion. The DKF is defined by explicitly using the Bayesian *a posterior* probability $P(C_k|\boldsymbol{x})$. Similar with the Bayesian decision theory, we have to presume $P(C_k|\boldsymbol{x})$ by a certain estimation method to use DKF for real application.

In this paper, we propose a nonlinear discriminant analysis based on the probabilistic estimation of the Gaussian mixture model. We use GMM to estimate the Bayesian *a posterior* probabilities $P(C_k|\boldsymbol{x})$ of any classification problems. Then we use $P(C_k|\boldsymbol{x})$ estimated by GMM to construct discriminative kernel function which is optimal in terms of the Fisher's discriminant criterion. We call this Gaussian mixture (GM) kernel.

We investigate the performance of the proposed GM kernel by several experiments using UCI machine learning repository [5]. We compare the discriminative power of the discriminant spaces which are constructed from the proposed kernel and usual kernels. The visualization experiments for the discriminant spaces or kernel matrices show some good properties of our discriminant kernels.

The rest of this paper is organized as follows: Section 2 reviews FLDA, KFDA and discriminant kernels. Section 3 reviews Gaussian mixture model. Section 3.2 describes our proposed Gaussian mixture kernel. The experiments are described in Section 4. Finally, Section 5 concludes the paper.

2 Discriminant Analysis

2.1 Fisher's Linear Discriminant

Fisher's linear discriminant analysis (FLDA) [4] 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.

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

FLDA constructs a dimension reducing linear mapping from the input feature vector $\boldsymbol{x}$ to a new feature vector $\boldsymbol{y}$ as

$$\boldsymbol{y} = A^T(\boldsymbol{x} - \bar{\boldsymbol{x}}_T) \tag{1}$$

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

The discriminant criterion

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

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

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

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

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_W and Σ_B are respectively the within-class covariance matrix and the between-class covariance matrix of the input feature vectors $\boldsymbol{x}$, and they are computed as

$$\Sigma_W = \sum_{k=1}^{K} P(C_k)\Sigma_k \tag{4}$$

$$\Sigma_k = \frac{1}{n_k}\sum_{l_i=C_k}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)^k \tag{5}$$

$$\Sigma_B = \sum_{k=1}^{K} P(C_k)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^T, \tag{6}$$

where n_k, $P(C_k)$, $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{x}}_T$ denote the number of training samples of the class C_k, *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 $P(C_k) = \frac{n_k}{n}$.

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 $\min(K-1, n)$.

2.2 Kernel Discriminant Analysis

FLDA is useful for linear separable cases, but for more complicated cases, it is necessary to extend it to nonlinear. Kernel discriminant analysis (KFDA) [1,9] is

one of the nonlinear extensions of FLDA and constructs a nonlinear discriminant mapping as a linear combination of kernel functions.

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

The discriminant mapping can be given as

$$\boldsymbol{y}(\boldsymbol{x}) = U^T \Phi(\boldsymbol{x}). \tag{7}$$

Similar with the case of the kernel PCA, the coefficient matrix U can be expressed as a linear combinations of the training samples as

$$U = \sum_{j=1}^{n} \Phi(\boldsymbol{x}_j) \boldsymbol{\alpha}_j^T, \tag{8}$$

the discriminant mapping can be rewritten as

$$\boldsymbol{y}(\boldsymbol{x}) = \sum_{j=1}^{n} \boldsymbol{\alpha}_j \Phi(\boldsymbol{x}_j)^T \Phi(\boldsymbol{x}) = \sum_{j=1}^{n} \boldsymbol{\alpha}_j K(\boldsymbol{x}_j, \boldsymbol{x}) = A^T \boldsymbol{k}(\boldsymbol{x}), \tag{9}$$

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 empirical kernel vector, respectively.

Then the discriminant criterion is given as

$$J = \mathrm{tr}\left(\hat{\Sigma}_W^{-1} \hat{\Sigma}_B \right), \tag{10}$$

where $\hat{\Sigma}_W$ and $\hat{\Sigma}_B$ are the within-class covariance matrix and the between-class covariance matrix of the new feature vectors $\boldsymbol{y}(\boldsymbol{x})$, respectively.

The polynomial functions

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^q \tag{11}$$

or the Radial Basis functions

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{y}||^2}{2\sigma^2} \right) \tag{12}$$

are often used as the kernel function for KFDA.

2.3 Discriminant Kernel Functions

In the KFDA, usually the kernel functions are defined a priori and selected without the clear reason. Also such kernel functions are general and not related to the probabilistic inference.

Recently, Kurita proposed the discriminant kernel function (DKF) which is based on the Bayesian *a posterior* probability estimation [8]. This kernel function is defined as

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

where $P(C_k|\boldsymbol{x})$ is the Bayesian *a posterior* probability which is presumed by a certain estimation method, and $P(C_k)$ is the prior of the k-th class C_k.

The Eq. (13) is derived from the theory of optimum nonlinear discriminant analysis (ONDA) [11,12]. Since ONDA gives the optimum nonlinear mapping that maximizes the discriminant criterion, DKF derived from ONDA is also optimum in terms of the discriminant criterion.

As shown in Eq. (13), DKF is defined by using the Bayesian *a posterior* probability $P(C_k|\boldsymbol{x})$. Similar with the Bayesian decision theory, we have to estimate $P(C_k|\boldsymbol{x})$ by a certain estimation method to use DKF for real application. Conversely, DKF can be used as one of the optimal way to construct kernel functions maximizing the discriminant criterion from the Bayesian *a posterior* probability estimation.

There are many ways to estimate the Bayesian *a posterior* probabilities. Depending on the estimation method, we can define the corresponding discriminant kernel function. In this paper we propose discriminant kernel function based on Gaussian mixture model (GMM).

3 Gaussian Mixture Model

Multivariate Gaussian distribution is defined as

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\} \tag{14}$$

where m is a number of variables, $\boldsymbol{\mu}$ is a mean vector and Σ is a covariance matrix.

Gaussian mixture model (GMM) is a linear combination of multiple Gaussian distributions. In GMM, each elemental Gaussian distribution is called component. GMM is formulated as

$$p(\boldsymbol{x}) = \sum_{j=1}^{J} \pi_j \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_j, \Sigma_j) \tag{15}$$

where J is a number of components, $\boldsymbol{\mu}_j$ and Σ_j is a mean vectors and a covariance matrix of the j-th component respectively, and π_j is coefficient of the linear combination.

The parameters $\boldsymbol{\mu}_j, \Sigma_j$ and π_j are usually estimated by Expectation Maximization (EM) algorithm [2].

3.1 The Bayesian *A Posterior* Probability Estimation by GMM

To estimate the Bayesian *a posterior* probability $P(C_k|\boldsymbol{x})$ by GMM, we define the probability densitiy $p(\boldsymbol{x}|C_k)$ of each class C_k as

$$p(\boldsymbol{x}|C_k) = \sum_{j=1}^{J_k} \pi_{k,j} \mathcal{N}_{k,j}(\boldsymbol{x}) \tag{16}$$

where $\mathcal{N}_{k,j}(\boldsymbol{x})$ represetns $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{k,j}, \Sigma_{k,j})$, the j-th Gaussian component for the class C_k. J_k is a number of components for the class C_k. The coefficient $\pi_{k,j}$, the mean vector $\boldsymbol{\mu}_{k,j}$ and the covariance matrix $\Sigma_{k,j}$ are estimated by using given samples $\boldsymbol{x}$ belongs to the class C_k.

Then the posterior probability can be written as

$$P(C_k|\boldsymbol{x}) = \frac{P(C_k)p(\boldsymbol{x}|C_k)}{p(\boldsymbol{x})} = \frac{P(C_k)\sum_{j=1}^{J_k} \pi_{k,j}\mathcal{N}_{k,j}(\boldsymbol{x})}{p(\boldsymbol{x})} \tag{17}$$

where

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} P(C_k)p(\boldsymbol{x}|C_k) = \sum_{k=1}^{K} P(C_k) \sum_{j=1}^{J_k} \pi_{k,j}\mathcal{N}_{k,j}(\boldsymbol{x}). \tag{18}$$

3.2 Gaussian Mixture Kernel

As described in Sec. 2.3, the estimation of the Bayesian *a posterior* probability $P(C_k|\boldsymbol{x})$ can be used to construct the kernel function which is optimum in terms of the discriminant criterion. We obtain kernel function based on Gaussian mixture model by substituting Eq. (17) for Eq. (13):

$$\begin{aligned} K_{GM}(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{k=1}^{K} \frac{P(C_k|\boldsymbol{x})P(C_k|\boldsymbol{y})}{P(C_k)} = \frac{\sum_{k=1}^{K} P(C_k)p(\boldsymbol{x}|C_k)p(\boldsymbol{y}|C_k)}{p(\boldsymbol{x})p(\boldsymbol{y})} \\ &= \frac{\sum_{k=1}^{K} P(C_k) \sum_{i=1}^{J_k} \sum_{j=1}^{J_k} \pi_{k,i}\pi_{k,j}\mathcal{N}_{k,i}(\boldsymbol{x})\mathcal{N}_{k,j}(\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \end{aligned} \tag{19}$$

We call this the Gaussian mixture (GM) kernel.

The matrix $\hat{K}$ having the component $k_{mn} = K_{GM}(\boldsymbol{x}_m, \boldsymbol{x}_n)$ is regarded as the kernel matrix of GM kernel. We can perform a novel nonlinear discriminant analysis by applying FLDA to the matrix $\hat{K}$. We call it GMM based kernel discriminant analysis (GM KDA).

After several deformations, Eq. (19) can be rewritten as

$$K_{GM}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{k=1}^{K} P(C_k) \sum_{i=1}^{J_k} \sum_{j=1}^{J_k} \alpha_{k,i,j} \exp\left\{-\frac{M_{k,i}(\boldsymbol{x}) + M_{k,j}(\boldsymbol{y})}{2}\right\}}{p(\boldsymbol{x})p(\boldsymbol{y})}, \tag{20}$$

$$\alpha_{k,i,j} = \frac{\pi_{k,i}\pi_{k,j}}{(2\pi)^D \sqrt{|\Sigma_{k,i}||\Sigma_{k,j}|}}, \tag{21}$$

$$M_{k,i}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_{k,i})^T \Sigma_{k,i}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{k,i}). \tag{22}$$

Table 1. Specifications of data sets

data set	# of classes	# of samples	# of features
heart	2	270	13
breast cancer	2	683	10
australian	2	690	14
wine	3	178	13
vehicle	4	846	18
vowel	11	990	10

$M_{k,i}(\boldsymbol{x})$ represents the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_{k,i}$, the mean vector of the i-th Gaussian component for the class C_k. Then the proposed kernel function can be interpreted as the sum of exponential of negative averaged Mahalanobis distances.

4 Experiments

The performance of our GMM based nonlinear discriminant analysis is evaluated by using six standard data sets (**heart**, **breast cancer**, **australian**, **wine**, **vehicle** and **vowel**) from UCI Machine Learning Repository [5]. Table 1 shows the statistics of these data sets.

For classification experiments, each data set is divided into a training set (2/3 of all samples) and a test set (remaining samples) at random. A training and testing task is repeated 10 times with different random seeds, and the averaged classification rate for the test sets are shown in the following sections. 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.

4.1 Evaluation of the Number of Components

Gaussian mixture model has the hyper-parameter J which implies the number of Gaussian components. We confirm the relationship between the number of components and classification accuracy. In this section we express the Gaussian mixture model having J components as J-GMM.

For the dataset **heart** and **vowel**, five Gaussian mixture models (1-GMM to 5-GMM) are trained. Each model is used to make GM kernel, and these kernels are used to do the GMM based discriminant analysis.

Tab. 2 shows the training and testing accuracy. Although the performances to the training samples are improving with the number of components, the performances to the test samples are not always increasing.

To avoid the over-fitting problem, we have to reduce the unnecessary components. In this paper, we manually determine the appropriate number of components based on the preliminary experiments. For all classes C_k, we use $J_k = 1$ for **heart**, **breast cancer**, **australian**, **wine**, **vehicle** and use $J_k = 3$ for **vowel**.

Table 2. Relationship between the number of components and classification accuracy

	1-GMM	2-GMM	3-GMM	4-GMM	5-GMM
heart (train)	88.28%	90.11%	92.06%	93.50%	94.06%
heart (test)	81.56%	79.11%	77.22%	75.44%	76.00%
vowel (train)	94.11%	98.56%	99.26%	99.33%	99.38%
vowel (test)	85.67%	91.88%	94.18%	93.15%	94.03%

Table 3. Classification rates (and standard deviations) of 9-NN in discriminant spaces

	Fisher's LDA	RBF KDA	GM KDA (proposed)
heart	81.11% (1.57)	77.56% (8.84)	81.56% (3.02)
breastcancer	97.06% (1.48)	96.40% (1.79)	96.67% (1.51)
australian	85.48% (1.76)	84.87% (1.75)	85.74% (1.28)
wine	98.50% (1.46)	98.17% (2.00)	98.33% (1.76)
vehicle	76.95% (2.06)	84.49% (1.19)	82.45% (1.39)
vowel	75.52% (1.64)	97.03% (1.94)	94.18% (1.97)
Average	85.77% (1.66)	89.75% (2.92)	89.82% (1.82)

4.2 Visualization of Kernels

To compare the property of the proposed and the existing kernel functions, the feature spaces or kernel matrices of the **wine** are illustrated in Fig. 1, 2.

Fig. 1 shows the PCA space of the original features or the discriminant spaces of RBF or GM kernel. It shows a goodness of the proposed kernel. It is noticed that samples of the GM kernel are distributing only on the triangle regions. Generally, for K classes problems, the discriminant spaces of the proposed discriminant kernel forms the $K-1$ dimensional hyper-tetrahedron (simplex) which is expected to be ideal. Since the GM kernel is defined by the Bayesian *a posterior* probabilities, it easily gives a probabilistic interpretation such as how a sample is close to each class. On the other hand, samples of the original features and the RBF kernel are freely and widely distributing in the two dimensional plane.

Fig. 2 shows the visualization result for three types of kernel matrices. The first one is linear kernel; it is constructed from just a inner product of the pair of the original features. Others are the RBF or GM kernel. The color of the i-th row and the j-th column shows the similarity between sample i and j. Since the samples are sorted in order of a class label beforehand, ideally, these matrices should have a block diagonal structure. Such diagonal class structure more clearly appears in the GM kernel than the Linear or the RBF kernel.

4.3 Comparison of Classification Accuracy

We compare the performances of the proposed GMM based discriminant analysis with usual Fisher's Linear Discriminant Analysis (FLDA) and RBF Kernel Discriminant Analysis (RBF KDA). For the classification method in their discriminant spaces, k-nearest neighbor method is adopted. We use $k = 9$ for all dataset and all discriminant spaces.

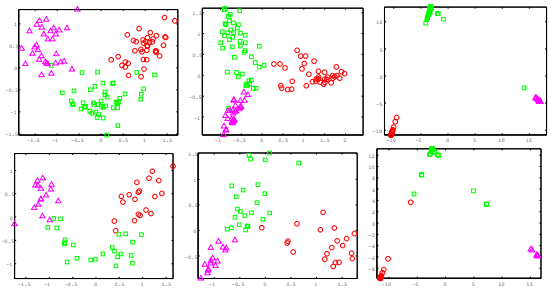

Fig. 1. Sample distributions of **wine** data. The top row and the bottom row show the training and test sets, respectively. (Left) PCA spaces of original features. (Center) Discriminant spaces obtained from RBF kernel matrices. (Right) Discriminant spaces obtained from GM kernel matrices.

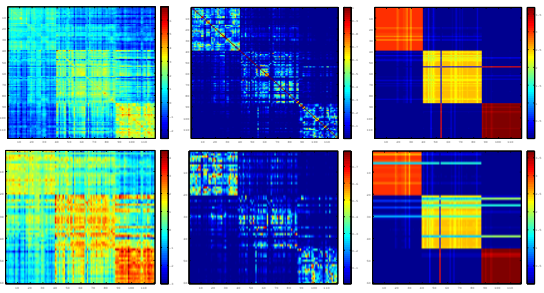

Fig. 2. Visualized kernel matrices of **wine** data. The top row and the bottom row show the results of training and test sets, respectively. (Left) Linear kernel (inner product) of original features. (Center) RBF kernel matrices. (Right) GM kernel matrices.

The parameters of RBF KDA, i.e. the coefficient σ in Eq. (12), are determined by grid search and 10-fold cross validation. We search the best σ from 31 candidates $\sigma = 2^{-15}, 2^{-14}, 2^{-13}, \cdots, 2^{+14}, 2^{+15}$.

Table 3 shows the classification rates for test samples of the proposed and existing methods. The proposed GMM based discriminant analysis has a better performance about averaged accuracy for six datasets. RBF KDA and GM KDA have almost comparable accuracy, but GM KDA shows good (smaller) averaged variance.

5 Conclusion

In this paper we propose GMM based nonlinear discriminant analysis which is formulated by the Bayesian *a posterior* probabilities estimated by Gaussian mixture model. The GM kernel has comparable classification performance with RBF kernel while GM kernel has more good stability (smaller variance).

In the experiment, we manually determined the hyper-parameter J which is the number of individual Gaussian distributions. We shoud automatically

determine J by using cross validation or several statistical validation methods as the future work.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number 23500211.

References

1. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Computation 12(10), 2385–2404 (2000)
2. Bishop, C.M.: Pattern Recognition and Machine Learning, Springer (2006)
3. Bilmes, J.: A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden markov models. Technical Report ICSI-TR-97-02, University of Berkeley (1997)
4. Fisher, R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7, 179–188 (1936)
5. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science
6. Hidaka, A., Kurita, T.: Discriminant Kernels based Support Vector Machine. In: The First Asian Conference on Pattern Recognition (ACPR 2011), Beijing, China, November 28-30, pp. 159–163 (2011)
7. Kurita, T., Watanabe, K., Otsu, N.: Logistic Discriminant Analysis. In: Proc. of 2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, Texas, USA, October 11-14, pp. 2236–2241 (2009)
8. Kurita, T.: Discriminant Kernels derived from the Optimum Nonlinear Discriminant Analysis. In: Proc. of 2011 International Joint Conference on Neural Networks, San Jose, California, USA, July 31-August 5 (2011)
9. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Smola, A., Muller, K.: Fisher discriminant analysis with kernels. In: Proc. IEEE Neural Networks for Signal Processing Workshop, pp. 41–48 (1999)
10. Nishida, K., Kurita, T.: RANSAC-SVM for Large-Scale Datesets. In: Proc. of International Conference on Pattern Recognition, December 8-11. Tampa Convention Center, Tampa (2008)
11. Otsu, N.: Nonlinear discriminant analysis as a natural extension of the linear case. Behavior Metrika 2, 45–59 (1975)
12. Otsu, N.: Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In: Proceedings of the 6th International Conference on Pattern Recognition, pp. 557–560 (1982)
13. Scholkopf, B., Burges, C.J.C., Smola, A.J.: Advances in Kernel Methods - Support Vector Learning. MIT Press (1999)
14. Viola, P., Jones, M.: Robust real time object detection. In: IEEE ICCV Workshop on Statistical and Computational Theories of Vision (July 2001)
15. Wu, T.F., et al.: Probability Estimates for Multi-class Classification by Pairwise Coupling. Journal of MLR 5, 975–1005 (2004)
16. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–137 (1997)
17. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3, 72–83 (1995)