Kernel Maximum a Posteriori Classification with Error Bound Analysis

Zenglin Xu, Kaizhu Huang, Jianke Zhu, Irwin King, and Michael R. Lyu

Dept. of Computer Science and Engineering,
The Chinese Univ. of Hong Kong,
Shatin, N.T., Hong Kong
{zlxu,kzhuang,jkzhu,king,lyu}@cse.cuhk.edu.hk

Abstract. Kernel methods have been widely used in data classification. Many kernel-based classifiers like Kernel Support Vector Machines (KSVM) assume that data can be separated by a hyperplane in the feature space. These methods do not consider the data distribution. This paper proposes a novel Kernel Maximum A Posteriori (KMAP) classification method, which implements a Gaussian density distribution assumption in the feature space and can be regarded as a more generalized classification method than other kernel-based classifier such as Kernel Fisher Discriminant Analysis (KFDA). We also adopt robust methods for parameter estimation. In addition, the error bound analysis for KMAP indicates the effectiveness of the Gaussian density assumption in the feature space. Furthermore, KMAP achieves very promising results on eight UCI benchmark data sets against the competitive methods.

1 Introduction

Recently, kernel methods have been regarded as the state-of-the-art classification approaches [1]. The basic idea of kernel methods in supervised learning is to map data from an input space to a high-dimensional feature space in order to make data more separable. Classical kernel-based classifiers include Kernel Support Vector Machine (KSVM) [2], Kernel Fisher Discriminant Analysis (KFDA) [3], and Kernel Minimax probability Machine [4,5]. The reasonability behind them is that the linear discriminant functions in the feature space can represent complex separating surfaces when mapped back to the original input space. However, one drawback of KSVM is that it does not consider the data distribution and cannot directly output the probabilities or confidences for classification. Therefore, it is hard to be applied in systems that reason under uncertainty.

On the other hand, in statistical pattern recognition, the probability densities can be estimated from data. Future examples are then assigned to the class with the Maximum A Posteriori (MAP) [6]. One typical probability density function is the Gaussian density function. The Gaussian density functions are easy to handle. However, the Gaussian distribution cannot be easily satisfied in the input space and it is hard to deal with non-linearly separable problems.

M. Ishikawa et al. (Eds.): ICONIP 2007, Part I, LNCS 4984, pp. 841–850, 2008.

To solve these problems, we propose a Kernel Maximum a Posteriori (KMAP) Classification method under Gaussianity assumption in the feature space. Different from KSVM, we make the Gaussian density assumption, which implies that data can be separated by more complex surfaces in the feature space. Generally, distributions other than the Gaussian distribution can also be assumed in the feature space. However, under a distribution with a complex form, it is hard to get a close form solution and easy to trap in over-fitting. Moreover, with the Gaussian assumption, a kernelized version can be derived without knowing the explicit form of the mapping functions for our model. In addition, to indicate the effectiveness of our assumption, we calculate a separability measure and the error bound for bi-category data sets. The error bound analysis shows that Gaussian density distribution can be more easily satisfied in the feature space.

This paper is organized as follows. Section 2 derives the MAP decision rules in the feature space, and analyzes its separability measures and upper error bounds. Section 3 presents the experiments against other classifiers. Section 4 reviews the related work. Section 5 draws conclusions and lists possible future research directions.

2 Main Results

In this section, our MAP classification model is derived. Then, we adopt a special regularization to estimate the parameters. The kernel trick is used to calculate our model. Last, the separability measure and the error bound are calculated in the kernel-induced feature space.

2.1 Model Formulation

Under the Gaussian distribution assumption, the conditional density function for each class $C_i (1 \leq i \leq m)$ is written as below:

$$p(\Phi(\mathbf{x})|C_i) = \frac{1}{(2\pi)^{N/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\Phi(\mathbf{x}) - \mu_i)^T \Sigma_i^{-1}(\Phi(\mathbf{x}) - \mu_i)\right\}, \quad (1)$$

where $\Phi(\mathbf{x})$ is the image of $\mathbf{x}$ in the feature space, N is the dimension of the feature space (N could be infinity), μ_i and Σ_i are the mean and the covariance matrix of C_i, respectively, and $|\Sigma_i|$ is the determinant of the covariance matrix. According to the Bayesian Theorem, the posterior probability of class C_i is calculated by

$$P(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{\sum_{j=1}^{m} p(\mathbf{x}|C_j)P(C_j)}. \quad (2)$$

Based on Eq. (2), the decision rule can be formulated as below:

$$\mathbf{x} \in C_w \text{ if } P(C_w|\mathbf{x}) = \max_{1 \leq j \leq m} P(C_j|\mathbf{x}). \quad (3)$$

This means that a test data point will be assigned to the class with the maximum of $P(C_w|\mathbf{x})$, i.e., the MAP. Since the MAP is calculated in the kernel-induced feature space, the output model is named as the KMAP classification. KMAP can provide not only a class label but also the probability of a data point belonging to that class. This probability can be viewed as a confidence of classifying new data points and can be used in statistical systems that reason under uncertainty. If the confidence is lower than some specified threshold, the system can refuse to make an inference. However, many kernel learning methods including KSVM cannot output these probabilities.

It can be further formulated as follows:

$$g_i(\Phi(\mathbf{x})) = (\Phi(\mathbf{x}) - \mu_i)^T \Sigma_i^{-1} (\Phi(\mathbf{x}) - \mu_i) + \log|\Sigma_i|. \tag{4}$$

The intuitive meaning of the function is that a class is more likely assigned to an unlabeled data point, when the Mahalanobis distance from the data point to the class center is smaller.

2.2 Parameter Estimation

In order to compute the Mahalanobis distance function, the mean vector and the covariance matrix for each class are required to be estimated. Typically, the mean vector (μ_i) and the within covariance matrix (Σ_i) are calculated by the maximum likelihood estimation. In the feature space, they are formulated as follows:

$$\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \Phi(\mathbf{x}_j), \tag{5}$$

$$\Sigma_i = S_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (\Phi(\mathbf{x}_j) - \mu_i)(\Phi(\mathbf{x}_j) - \mu_i)^T, \tag{6}$$

where n_i is the cardinality of the set composed of data points belonging to C_i.

Directly employing S_i as the covariance matrix, will generate quadratic discriminant functions in the feature space. In this case, KMAP is noted as KMAP-M. However, the covariance estimation problem is clearly ill-posed, because the number of data points in each class is usually much smaller than the number of dimensions in the kernel-induced feature space. The treatment of this ill-posed problem is to introduce the regularization. There are several kinds of regularization methods. One of them is to replace the individual within-covariance matrix by their average, i.e., $\Sigma_i = S = \frac{\sum_{i=1}^{m} S_i}{m} + rI$, where I is the identity matrix and r is a regularization coefficient. This method can substantially reduce the number of free parameters to be estimated. Moreover, it also reduces the discriminant function between two classes to a linear one. Therefore, a linear discriminant analysis method can be obtained.

Alternatively, we can estimate the covariance matrix by combining the above linear discriminant function with the quadratic one. Instead of estimating the

covariance matrix in the input space [7], we try to apply this method in the feature space. The formulation in the feature space is as follows:

$$\Sigma_i = (1-\eta)\tilde{\Sigma}_i + \eta\frac{trace(\tilde{\Sigma}_i)}{n}I, \tag{7}$$

where $\tilde{\Sigma}_i = (1-\theta)S_i + \theta S$.

In the equations, θ $(0 \leq \theta \leq 1)$ is a coefficient linked with the linear discriminant term and the quadratic discriminant one. Moreover, η $(0 \leq \eta \leq 1)$ determines the shrinkage to a multiple of the identity matrix. This approach is more flexible to adjust the effect of the regularization. The corresponding KMAP is noted as KMAP-R.

2.3 Kernel Calculation

We derive methods to calculate the Mahalanobis distance (Eq. (4)) using the kernel trick, i.e., we only need to formulate the function in an inner-product form regardless of the explicit mapping function. To do this, the spectral representation of the covariance matrix, $\Sigma_i = \sum_{j=1}^{N} \Lambda_{ij}\Omega_{ij}\Omega_{ij}^T$ where $\Lambda_{ij} \in \mathcal{R}$ is the j-th eigenvalue of Σ_i and $\Omega_{ij} \in \mathcal{R}^N$ is the eigenvector relevant to Λ_{ij}, is utilized. However, the small eigenvalues will degrade the performance of the function overwhelmingly because they are underestimated due to the small number of examples. In this paper, we only estimate the k largest eigenvalues and replace each left eigenvalue with a nonnegative number h_i. Thus Eq. (4) can be reformulated as follows:

$$\begin{aligned} g_i(\Phi(\mathbf{x})) &= \frac{1}{h_i}[g_{1i}(\Phi(\mathbf{x})) - g_{2i}(\Phi(\mathbf{x}))] + g_{3i}(\Phi(\mathbf{x})) \\ &= \frac{1}{h_i}\left(\sum_{j=1}^{N}[\Omega_{ij}^T(\Phi(\mathbf{x}) - \mu_i)]^2 - \sum_{j=1}^{k}\left(1 - \frac{h_i}{\Lambda_{ij}}\right)[\Omega_{ij}^T(\Phi(\mathbf{x}) - \mu_i)]^2\right) \\ &\quad + \log\left(h_i^{N-k}\prod_{j=1}^{k}\Lambda_{ij}\right). \end{aligned} \tag{8}$$

In the following, we show that $g_{1i}(\Phi(\mathbf{x}))$, $g_{2i}(\Phi(\mathbf{x}))$, and $g_{3i}(\Phi(\mathbf{x}))$ can all be written in a kernel form. To formulate these equations, we need to calculate the eigenvalues Λ_i and eigenvectors Ω_i. The eigenvectors lie in the space spanned by all the training samples, i.e., each eigenvector Ω_{ij} can be written as a linear combination of all the training samples:

$$\Omega_{ij} = \sum_{l=1}^{n}\gamma_{ij}^{(l)}\Phi(\mathbf{x}_l) = U\gamma_{ij} \tag{9}$$

where $\gamma_{ij} = (\gamma_{ij}^{(1)}, \gamma_{ij}^{(2)}, \ldots, \gamma_{ij}^{(n)})^T$ is an n dimensional column vector and $U = (\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n))$.

It is easy to prove that γ_{ij} and Λ_{ij} are actually the eigenvector and eigenvalue of the covariance matrix $\Sigma_{G^{(i)}}$, where $G^{(i)} \in \mathcal{R}^{n_i \times N}$ is the i-th block of the kernel matrix G relevant to C_i . We omit the proof due to the limit of space.

Accordingly, we can express $g_{1i}(\Phi(\mathbf{x}))$ as the kernel form:

$$\begin{aligned} g_{1i}(\Phi(\mathbf{x})) &= \sum_{j=1}^{n} \gamma_{ij}^T U^T (\Phi(\mathbf{x}) - \mu_i)^T (\Phi(\mathbf{x}) - \mu_i) U \gamma_{ij} \\ &= \sum_{j=1}^{n} \left[\gamma_{ij}^T \left(K_{\mathbf{x}} - \frac{1}{n_i} \sum_{l=1}^{n_i} K_{\mathbf{x}_l} \right) \right]^2 \\ &= \left\| K_{\mathbf{x}} - \frac{1}{n_i} \sum_{l=1}^{n_i} K_{\mathbf{x}_l} \right\|_2^2, \end{aligned} \tag{10}$$

where $K_{\mathbf{x}} = \{K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_n, \mathbf{x})\}^T$.

In the same way, $g_{2i}(\Phi(\mathbf{x}))$ can be formulated as the following:

$$g_{2i}(\Phi(\mathbf{x})) = \sum_{j=1}^{k} \left(1 - \frac{h_i}{\Lambda_{ij}} \right) \Omega_{ij}^T (\Phi(\mathbf{x}) - \mu_i)(\Phi(\mathbf{x}) - \mu_i)^T \Omega_{ij}. \tag{11}$$

Substituting (9) into the above $g_{2i}(\Phi(\mathbf{x}))$, we have:

$$g_{2i}(\Phi(\mathbf{x})) = \sum_{j=1}^{k} \left(1 - \frac{h_i}{\Lambda_{ij}} \right) \gamma_{ij}^T \left(K_{\mathbf{x}} - \frac{1}{n_i} \sum_{l=1}^{n_i} K_{\mathbf{x}_l} \right) \left(K_{\mathbf{x}} - \frac{1}{n_i} \sum_{l=1}^{n_i} K_{\mathbf{x}_l} \right)^T \gamma_{ij}. \tag{12}$$

Now, the Mahalanobis distance function in the feature space $g_i(\Phi(\mathbf{x}))$ can be finally written in a kernel form, where N in $g_{3i}(\Phi(\mathbf{x}))$ is substituted by the cardinality of data n. The time complexity of KMAP is mainly due to the eigenvalue decomposition which scales as $\mathcal{O}(n^3)$. Thus KMAP has the same complexity as KFDA.

2.4 Connection to Other Kernel Methods

In the following, we show the connection between KMAP and other kernel-based methods.

In the regularization method based on Eq. (7), by varying the settings of θ and η, other kernel-based classification methods can be derived. When $(\theta = 0, \eta = 0)$, the KMAP model represents a quadratic discriminant method in the kernel-induced feature space; when $(\theta = 1, \eta = 0)$, it represents a kernel discriminant method; and when $(\theta = 0, \eta = 1)$ or $(\theta = 1, \eta = 1)$, it represents the nearest mean classifier. Therefore, by varying θ and η, different models can be generated from different combinations of quadratic discriminant, linear discriminant and the nearest mean methods.

We consider a special case of the regularization method when $\theta = 1$ and $\eta = 0$. If both classes are assumed to have the same covariance structure for a

binary class problem, i.e., $\Sigma_i = \frac{\Sigma_1 + \Sigma_2}{2}$, it leads to a linear discriminant function. Assuming all classes have the same class prior probabilities, $g_i(\Phi(\mathbf{x}))$ can be derived as: $g_i(\Phi(\mathbf{x})) = (\Phi(\mathbf{x}) - \mu_i)^T(\frac{\Sigma_1 + \Sigma_2}{2})^{-1}(\Phi(x) - \mu_i)$, where $i = 1, 2$. We reformulate the above equation in the following form: $g_i(\Phi(\mathbf{x})) = \mathbf{w}_i\Phi(\mathbf{x}) + b_i$, where $\mathbf{w}_i = -4(\Sigma_1 + \Sigma_2)^{-1}\mu_i$, and $b_i = 2\mu_i^T(\Sigma_1 + \Sigma_2)^{-1}\mu_i$. The decision hyperplane is $f(\Phi(\mathbf{x})) = g_1(\Phi(\mathbf{x})) - g_2(\Phi(\mathbf{x}))$, i.e.,

$$f(\Phi(\mathbf{x})) = (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)\Phi(\mathbf{x}) - \frac{1}{2}(\mu_1 - \mu_2)^T(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 + \mu_2). \quad (13)$$

Eq. (13) is just the solution of KFDA [3]. Therefore, KFDA can be viewed as a special case of KMAP when all classes have the same covariance structure.

Remark. KMAP provides a rich class of kernel-based classification algorithms using different regularization methods. This makes KMAP as a flexible framework for classification adaptive to data distribution.

2.5 Separability Measures and Error Bounds

To measure the separability of different classes of data in the feature space, the Kullback-Leibler divergence (a.k.a. K-L distance) between two Gaussians is adopted. The K-L divergence is defined as

$$d_{KL}[p_i(\Phi(\mathbf{x})), p_j(\Phi(\mathbf{x}))] = \int P_i(\Phi(\mathbf{x})) \ln \frac{p_i(\Phi(\mathbf{x}))}{p_j(\Phi(\mathbf{x}))}. \quad (14)$$

Since the K-L divergence is not symmetric, a two-way divergence is used to measure the distance between two distributions

$$d_{ij} = d_{KL}[p_i(\Phi(\mathbf{x})), p_j(\Phi(\mathbf{x}))] + d_{KL}[p_j(\Phi(\mathbf{x})), p_i(\Phi(\mathbf{x}))] \quad (15)$$

Following [6], it can be proved that:

$$d_{ij} = \frac{1}{2}(\mu_i - \mu_j)^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) + \frac{1}{2}trace(\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I), \quad (16)$$

which can be solved by using the trick in Section 2.3.

The Bayesian decision rule guarantees the lowest average error rate as presented in the following:

$$P(correct) = \sum_{i=1}^{m} \int_{R_i} p(\mathbf{\Phi}(\mathbf{x})|C_i)P(C_i)d\mathbf{\Phi}(\mathbf{x}), \quad (17)$$

where R_i is the decision region of class C_i.

We implement the Bhattacharyya bound in the feature space for the Gaussian density distribution function. Following [6], we have

$$P(error) \leq \sqrt{P(C_1)P(C_2)}\exp^{-q(0.5)}, \quad (18)$$

where

$$q(0.5) = \frac{1}{8}(\mu_2 - \mu_1)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{|\frac{\Sigma_1 + \Sigma_2}{2}|}{\sqrt{|\Sigma_1||\Sigma_2|}}. \quad (19)$$

Using the results in Section 2.3, the Bhattacharyya error bound can be easily calculated in the kernel-induced feature space.

3 Experiments

In this section, we report the experiments to evaluate the separability measure, the error bound and the prediction performance of the proposed KMAP.

3.1 Synthetic Data

We compare the separability measure and error bounds on three synthetic data sets. The description of these data sets can be found in [8]. The data sets are named according to their characteristics and they are plotted in Fig. 1.

We map the data using RBF kernel to a special feature space where Gaussian distributions are approximately satisfied. We then calculate separability measures on all data sets according to Eq. (16). The separability values for the **Overlap**, **Bumpy**, and **Relevance** in the original input space, are 14.94, 5.16, and 22.18, respectively. Those corresponding values in the feature space are 30.88, 5.87, and 3631, respectively. The results indicate that data become more separable after mapped into the feature space, especially for the **Relevance** data set.

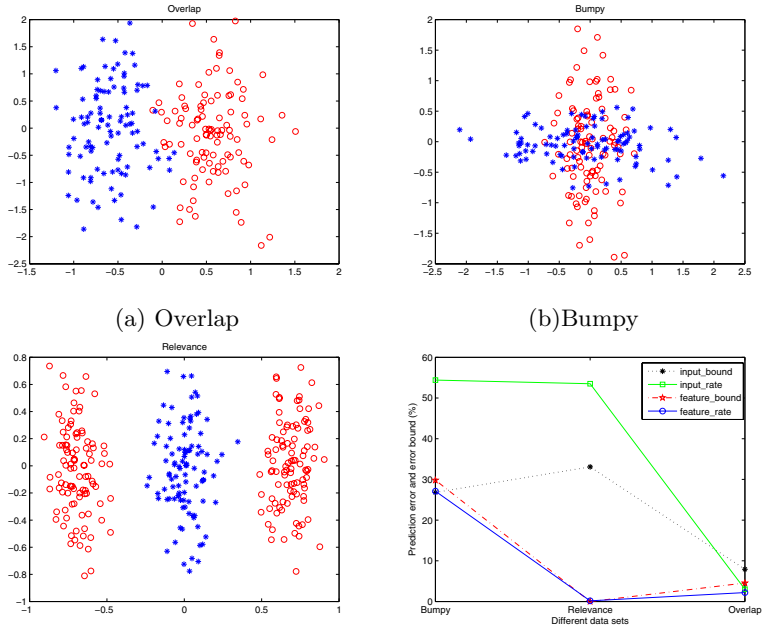
For data in the kernel-induced feature space, the error bounds are calculated according to Eq. (18). Figure 1 also plots the prediction rates and the upper error bounds for data in the input space and in the feature space, respectively. It can be observed that the error bounds are more valid in the feature space than those in the input space.

3.2 Benchmark Data

Experimental Setup. In this experiment, KSVM, KFDA, Modified Quadratic Discriminant Analysis (MQDA) [9] and Kernel Fisher's Quadratic Discriminant Analysis (KFQDA) [10] are employed as the competitive algorithms. We implement two variants of KMAP, i.e., KMAP-M and KMAP-R.

The properties of eight UCI benchmark data sets are described in Table 1.

In all kernel methods, a Gaussian-RBF kernel is used. The parameter $\mathbf{C}$ of KSVM and the parameter γ in RBF kernel are all tuned by 10-cross validation. In KMAP, we select k pairs of eigenvalues and eigenvectors according to their contribution to the covariance matrix, i.e., the index $j \in \{\ell : \frac{\ell}{\sum_{q=1}^{n}} \geq \alpha\}$; while in MQDF, the range of k is relatively small and we select k by cross validation. PCA is used as the regularization method in KFQDA and the commutative decay ratio is set to 99%; the regularization parameter r is set to 0.001 in KFDA.



(a) Overlap (b)Bumpy

(c) Relevance (d) Separability Comparison

Fig. 1. The data plot of Overlap, Bumpy and Relevance and the comparison of data separability in the input space and the feature space

Table 1. Data set information

Data Set	# Samples	# Features	# Classes	Data Set	# Samples	# Features	# Classes
Iono	351	34	2	Breast	683	9	2
Twonorm	1000	21	2	Sonar	208	60	2
Pima	768	8	2	Iris	150	4	3
Wine	178	13	3	Segment	210	19	7

In both KMAP and MQDF, h takes the value of Λ_{k+1}. In KMAP-R, extra parameters (θ, η) are tuned by cross-validation. All experimental results are obtained in 10 runs and each run is executed with 10-cross validation for each data set.

Experimental Results. Table 2 reports the average prediction accuracy with the standard errors on each data set for all algorithms. It can be observed that both variants of KMAP outperform MQDF, which is an MAP method in the input space. This also empirically validates that the separability among different classes of data becomes larger and that the upper error bounds get tighter and more accurate, after data are mapped to the high dimensional feature space.

Moreover, the performance of KMAP is competitive to that of other kernel methods. Especially, the performance of KMAP-R gets better prediction accuracy than all other methods for most of the data sets. The reason is that the regularization methods in KMAP favorably capture the prior distributions of

Table 2. The prediction results of KMAP and other methods

Data set	KSVM	MQDF	KFDA	KFQDA	KMAP-M	KMAP-R
Iono(%)	94.1±0.7	89.6±0.5	94.2±0.1	93.6±0.4	95.2±0.2	**95.7**±0.3
Breast(%)	96.5±0.4	96.5±0.1	96.4±0.1	96.5±0.1	96.5±0.1	**97.5**±0.1
Twonorm(%)	96.1±0.4	97.4±0.4	96.7±0.5	97.3±0.5	**97.6**±0.7	97.5±0.4
Sonar(%)	86.6±1.0	83.7±0.7	88.3±0.3	85.1±1.9	87.2±1.6	**88.8**±1.2
Pima(%)	**77.9**±0.7	73.1±0.4	71.0±0.5	74.1±0.5	75.4±0.7	75.5±0.4
Iris(%)	96.2±0.4	96.0±0.1	95.7±0.1	96.8±0.2	96.9±0.1	**98.0**±0.0
Wine(%)	98.8±0.1	99.2±1.3	99.1±0.1	96.9±0.7	**99.3**±0.1	99.3±0.6
Segment(%)	**92.8**±0.7	86.9±1.2	91.6±0.3	85.8±0.8	90.2±0.2	92.1±0.8
Average(%)	92.38	90.30	91.62	90.76	92.29	**93.05**

data, since the Gaussian assumption in the feature space can fit a very complex distribution in the input space.

4 Related Work

In statistical pattern recognition, the probability density function can first be estimated from data, then future examples could be assigned to the class with the MAP. One typical example is the Quadratic Discriminant Function (QDF) [11], which is derived from the multivariate normal distribution and achieves the minimum mean error rate under Gaussian distribution. In [9], a Modified Quadratic Discriminant Function (MQDF) less sensitive to estimation error is proposed. [7] improves the performance of QDF by covariance matrix interpolation. Unlike QDF, another type of classifiers does not assume the probability density functions in advance, but are designed directly on data samples. An example is the Fisher discriminant analysis (FDA), which maximizes the between-class covariance while minimizing the within-class variance. It can be derived as a Bayesian classifier under Gaussian assumption on the data. [3] develops a Kernel Fisher Discriminant Analysis (KFDA) by extending FDA to a non-linear space by the kernel trick.

To supplement the statistical justification of KFDA, [10] extends the maximum likelihood method and Bayes classification to their kernel generalization under Gaussian Hilbert space assumption. The authors do not directly kernelize the quadratic forms in terms of kernel values. Instead, they use an explicit mapping function to map the data to a high dimensional space. Thus the kernel matrix is usually used as the input data of FDA. The derived model is named as Kernel Fisher's Quadratic Discriminant Analysis (KFQDA).

5 Conclusion and Future Work

In this paper, we present a novel kernel classifier named Kernel-based Maximum a Posteriori, which implements Gaussian distribution in the kernel-induced feature space. Comparing to state-of-the-art classifiers, the advantages of KMAP include that the prior information of distribution is incorporated and that it can output probability or confidence in making a decision. Moreover, KMAP can

be regarded as a more generalized classification method than other kernel-based methods such as KFDA. In addition, the error bound analysis illustrates that Gaussian distribution is more easily satisfied in the feature space than that in the input space. More importantly, KMAP with proper regularization achieves very promising performance.

We plan to incorporate the probability information into both the kernel function and the classifier in the future work.

Acknowledgments

The work described in this paper is fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4205/04E and Project No. CUHK4235/04E).

References

1. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge (2002)
2. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons, Chichester (1998)
3. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Muller, K.: Fisher discriminant analysis with kernels. In: Proceedings of IEEE Neural Network for Signal Processing Workshop, pp. 41–48 (1999)
4. Lanckriet, G.R.G., Ghaoui, L.E., Bhattacharyya, C., Jordan, M.I.: A robust minimax approach to classification. Journal of Machine Learning Research 3, 555–582 (2002)
5. Huang, K., Yang, H., King, I., Lyu, M.R., Chan, L.: Minimum error minimax probability machine. Journal of Machine Learning Research 5, 1253–1286 (2004)
6. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-Interscience Publication (2000)
7. Friedman, J.H.: Regularized discriminant analysis. Journal of American Statistics Association 84(405), 165–175 (1989)
8. Centeno, T.P., Lawrence, N.D.: Optimising kernel parameters and regularisation coefficients for non-linear discriminant analysis. Journal of Machine Learning Research 7(2), 455–491 (2006)
9. Kimura, F., Takashina, K., S., T., Y., M.: Modified quadratic discriminant functions and the application to Chinese character recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 9, 149–153 (1987)
10. Huang, S.Y., Hwang, C.R., Lin, M.H.: Kernel Fisher's discriminant analysis in Gaussian Reproducing Kernel Hilbert Space. Technical report, Academia Sinica, Taiwan, R.O.C. (2005)
11. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego (1990)