PERGAMON

Pattern Recognition 36 (2003) 1703 – 1709

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Feature extraction based on the Bhattacharyya distance

Euisun Choi, Chulhee Lee*

Department of Electrical and Electronic Engineering, Yonsei University, 134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-749, South Korea

Received 9 April 2002; accepted 5 November 2002

Abstract

In this paper, we present a feature extraction method by utilizing an error estimation equation based on the Bhattacharyya distance. We propose to use classification errors in the transformed feature space, which are estimated using the error estimation equation, as a criterion for feature extraction. The construction of linear transformation for feature extraction is conducted using an iterative gradient descent algorithm, so that the estimated classification error is minimized. Due to the ability to predict error, it is possible to determine the minimum number of features required for classification. Experimental results show that the proposed feature extraction method compares favorably with conventional methods.

Keywords: Feature extraction; Bhattacharyya distance; Bayes error estimation; Optimization; Classification error

1. Introduction

In general, feature extraction can be considered as the process of transforming high dimensional data into a low dimensional feature space based on an optimization criterion. In other words, reducing dimensionality without a serious loss of class separability is the key to feature extraction. Canonical analysis (CA) is one of the most well-known examples of feature extraction [1]. CA introduces a within-class scatter matrix, Σ_w, and a between-class scatter matrix, Σ_b, to maximize a separation criterion known as the Fisher criterion:

$$J = tr(\Sigma_w^{-1}\Sigma_b) \tag{1}$$

Unfortunately, the criterion of Eq. (1) is not directly related to the classification error which has been the ultimate measure of feature effectiveness for classification in the transformed space, although there has been an attempt to connect the criterion to the Bayes error using a pairwise weighting function for multiclass problems [2]. For pattern classification, it may be optimal to use the minimum achievable error, the Bayes error, as a criterion [3]. However, the Bayes error cannot be easily expressed in an analytical form except in special cases. Instead, an estimate of the Bayes error is often used as a criterion. For instance, Juang and Katagiri proposed a recursive algorithm based on the minimum classification error (MCE) learning for speech recognition, where the classification error is measured by misclassification over the training samples [4]. On the other hand, Buturović proposed to use the k-NN estimate of Bayes error in the transformed space as an optimization criterion [5]. Here, the Bayes error is approximated by upper and lower error bounds for the appropriate range of k, and then minimized using the simplex algorithm in the space spanned by transformation matrix coefficients. Apparently, the performance of this method depends on k. However, determining an optimal value of k is not an easy problem.

Recently, it has been reported that an accurate estimation of classification error is possible using the Bhattacharyya distance [6]. Thus, we propose to approximate the classification error using the error estimation equation based on the Bhattacharyya distance as an optimization criterion assuming normal distributions. A benefit of this approach is that the number of features necessary for classification without serious information loss can be predicted. We present this

* Corresponding author. Tel.: +82-2-2123-2779; fax: +82-2-312-4584.

E-mail address: chulhee@yonsei.ac.kr (C. Lee).

0031-3203/03/$30.00 © 2003 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0031-3203(03)00035-9

estimation method in Section 2. For the feature extraction, we construct a linear transformation using an iterative gradient descent algorithm proposed in Ref. [7] based on the minimization of the estimated error. This technique is discussed in Section 3. Experimental results are provided in Section 4. Finally, conclusions are presented in Section 5.

2. Error prediction using Bhattacharyya distance

The Bhattacharyya distance has been used as a class separability measure for feature selection and is known to provide the upper and lower bounds of the Bayes error [3]. However, the bounds are not tight enough for practical applications as can be seen in Fig. 1.

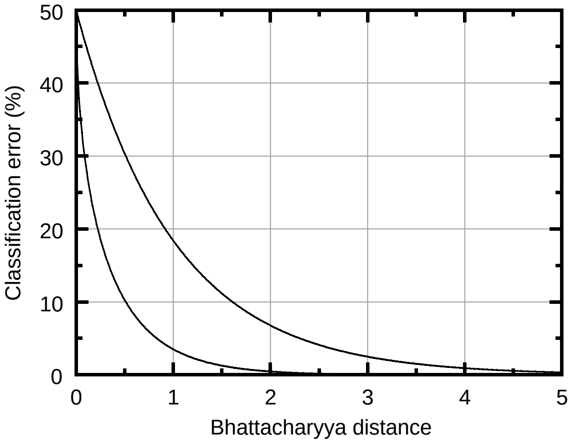

Fig. 1. Theoretical upper and lower bounds of Bayes error using the Bhattacharyya distance.

For two normally distributed classes, the Bhattacharyya distance is defined as follows:

$$b = \frac{1}{8}(\mu_2-\mu_1)^{\mathrm{T}}\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_2-\mu_1) + \frac{1}{2}\ln\frac{|(\Sigma_1+\Sigma_2)/2|}{|\Sigma_1|^{1/2}|\Sigma_1|^{1/2}}, \tag{2}$$

where μ_i and Σ_i are the mean vector and covariance matrix of class i, respectively. Since the Bhattacharyya distance is invariant under any non-singular linear transformation and translation, it can be rewritten as Eq. (3) in the N-dimensional diagonalized coordinate system [3].

$$b = \frac{1}{2}\sum_{l=1}^{N}\left\{\frac{m_l^2}{2(1+\lambda_l)} + \ln\left(\frac{1+\lambda_l}{2}\right) - \frac{1}{2}\ln\lambda_l\right\}. \tag{3}$$

It can be seen that the Bhattacharyya distance is a function of m_l and λ_l. In Ref. [6], Lee systematically sampled the space spanned by these variables, the *mean-lambda space*, under certain restrictions and generated pairs of classes with normal distributions, $N(0,I)$ and $N(M,\Lambda)$. Here, I is an identity matrix, $M=(m_1,m_2,\dots,m_N)^{\mathrm{T}}$, and Λ is a diagonal matrix with diagonal elements, $\lambda_1,\lambda_2,\dots,\lambda_N$. Then, the relationship between the Bhattacharyya distance and the error of the Gaussian ML classifier was investigated empirically using over 160 million pairs of classes, and the following error estimation equation was proposed:

$$\hat{\varepsilon} = 40.219 - 70.019b + 63.578b^2 - 32.766b^3 + 8.7172b^4 - 0.91875b^5, \tag{4}$$

where $\hat{\varepsilon}$ is an estimated classification error and b the Bhattacharyya distance. Using Eq. (4), it was shown that it is possible to predict the classification error from the Bhattacharyya distance within 1–2% margin. Fig. 2 shows the relationship between the classification error and the Bhattacharyya distance. The upper and lower bounds are also shown [6].

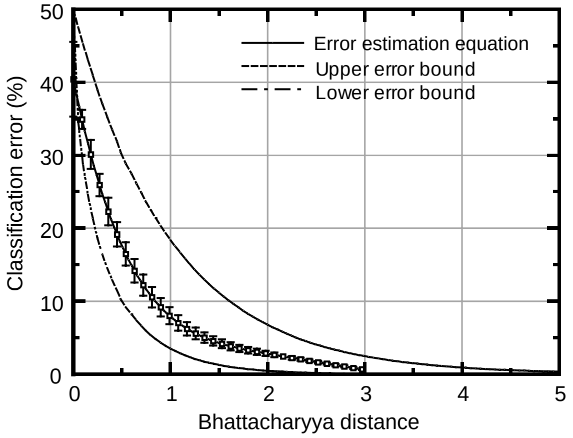

Fig. 2. Relationship between the Bhattacharyya distance and the classification error.

3. Feature extraction

3.1. Two-class problems

We begin the development of the feature extraction algorithm by considering two-class problems and then extend it to multiclass problems. Let Φ be an $N\times M$ matrix which transforms an observation vector X in the N-dimensional space into a new M-dimensional feature vector Y. In other words

$$Y = \Phi^{\mathrm{T}}X = [\phi_1,\phi_2,\dots,\phi_M]^{\mathrm{T}}X. \tag{5}$$

Here, the column vectors of Φ (ϕ_i, $i=1,\dots,M\leqslant N$) form an orthonormal basis for a subspace. Then, the mean vector and the covariance matrix of class i in the subspace are given by $\Phi^{\mathrm{T}}\mu_i$ and $\Phi^{\mathrm{T}}\Sigma_i\Phi$, respectively. Consequently, the

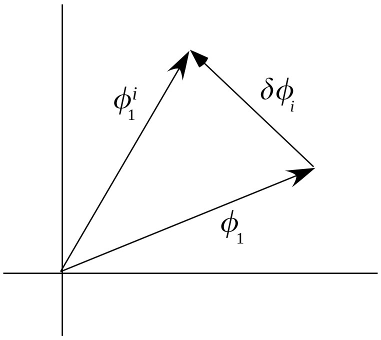

Fig. 3. Movement of feature vector in the sequential search.

estimate of the classification error in the subspace depends on the matrix Φ. In other words, the criterion is given by

$$J = \hat{\varepsilon}(\Phi), \tag{6}$$

where $\hat{\varepsilon}(\Phi)$ is an estimate of the classification error based on the Bhattacharyya distance in the subspace spanned by Φ. However, a closed form solution for Φ that minimizes J is not available, though it is possible to obtain the derivative of b with respect to Φ [3]. Therefore, we use a simple gradient descent scheme for minimizing J in order to find the subspace where the classification error is minimized. The algorithm discussed in this paper is called the *sequential search*, which is an optimization method for feature extraction proposed in Ref. [7]. Initially, we start with an $N \times N$ identity matrix Ψ:

$$\Psi = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{N\times N}.$$

In fact, Ψ can be any orthonormal matrix or can be obtained using another feature extraction algorithm. Then, the first M columns are used as the basis for the subspace:

$$\Psi_{N\times M} \equiv \Phi.$$

In the sequential search, we first move vector ϕ_1 slightly in the direction of the remaining basis vectors in order to find the direction in which the classification error decreases most rapidly.

As illustrated in Fig. 3, a new vector ϕ_1^i is given by

$$\phi_1^i = \phi_1 + \delta\phi_i, \tag{7}$$

where δ is the step size. Then, we compute the Bhattacharyya distance between the classes, which are projected onto this vector and predict the classification error using Eq. (4). By repeating this for ϕ_i $(i = 2, \ldots, N)$, we obtain the following vector whose elements are the rate of error change in each direction:

$$\mathbf{r} = [r_1, r_2, \ldots, r_N]^{\mathrm{T}} \tag{8}$$

where $r_i = (\hat{\varepsilon}(\phi_1^i) - \hat{\varepsilon}(\phi_1))/\delta$, $(i = 2, \ldots, N)$ and $r_1 = 0$ in the case of updating ϕ_1 (i.e. $M = 1$). We note that the finding of $\mathbf{r}$ can be done more efficiently than in Ref. [7], since we use the error estimation equation for the computation of the rate of error change. In Ref. [7], however, this step was computationally expensive, since all training samples were classified in order to obtain the classification error during the learning. Finally, the update rule for ϕ_1 is given by

$$\phi_{1,new} = \phi_{1,old} - \alpha(\Psi\mathbf{r}), \tag{9}$$

where α is a positive constant. It is noted that the matrix Ψ with the $\phi_{1,new}$ is no longer orthonormal. Therefore, we apply the Gram–Schmidt procedure so that each iteration can be started off with an orthonormal Ψ. We repeat this procedure until the change of error becomes smaller than a threshold.

However, the algorithm can be stuck in local minima in the worst case. To address this problem, one can start with a different initial matrix, Ψ, and select the best Φ that provides the lowest J. Additional feature vectors can be obtained similarly. If $M = 2$, for example, we move the vector ϕ_2 in the direction of ϕ_i $(i = 3, \ldots, N)$ and update it by using Eq. (9) recursively. In this case, the first 2 columns of Ψ form the transformation matrix $\Phi_{N\times 2}$, and the ϕ_1, which is already determined, remains unchanged. It is noted that the degree of freedom for the ϕ_2 decreases by one, which means $\mathbf{r} = [0, 0, r_3, \ldots, r_N]^{\mathrm{T}}$ in Eq. (9).

With the proposed feature extraction method based on the Bhattacharyya distance, it is possible to predict the minimum number of features required for classification without serious information loss. This can be achieved by using the error estimation equation as follows:

$$M_{\min} = \min\left\{ M \,\middle|\, \frac{1 - \hat{\varepsilon}(\Phi_{N\times M})}{1 - \hat{\varepsilon}(\Psi_{N\times N})} \geqslant t,\ 0 < t \leqslant 1 \right\}, \tag{10}$$

where t is a constant and has the value of 0.99 in this paper for this purpose. Here, $\hat{\varepsilon}(\Psi_{N\times N})$ denotes the estimated classification error in the N-dimensional space, which is estimated from the statistics of the original data using the error estimation equation before feature extraction. Since the classification error generally decreases when more features are added, we try to find the minimum number of features by increasing M until the condition of Eq. (10) is satisfied.

3.2. Multiclass problems

For multiclass problems, we present a simple extension by introducing the *Bhattacharyya distance feature matrix* [8]. Since the Bhattacharyya distance is defined between two classes, we consider each pair of classes for multiclass problems. Assuming that there are L classes which are assumed to have normal distributions in the N-dimensional space, let Σ_{BDFM}^{ij} be the Bhattacharyya distance feature matrix obtained from the pair of class i and j $(i \neq j)$. Then,

Σ^{ij}_{BDFM} is given by

$$\Sigma^{ij}_{BDFM} = \Phi^{ij} \mathbf{W}^{ij} \Phi^{ij\mathrm{T}}, \tag{11}$$

where Φ^{ij} is the transformation matrix that was obtained from the sequential search for each pair of classes, and $\mathbf{W}^{ij}$ is a diagonal matrix with diagonal elements, $w_1, w_2, \ldots, w_{M(t)}$. Here, $\mathbf{W}^{ij}$ is used to take into account the contribution of the individual feature vector and is computed as follows:

$$w_k = A_k - A_{k-1}, \tag{12}$$

where A_k equals the classification accuracy in the subspace spanned by the first k columns of Φ ($A_0 = 0$). It is noted that Σ^{ij}_{BDFM} is an $N \times N$ matrix with rank $M(t)$ ($M(t) \leqslant N$). If the constraint of Eq. (10) is used for multiclass problems, we have control over the contribution of M for each pair of classes by changing the value of t in Eq. (10). Obviously, if t is fixed (i.e. $t = 0.99$), the rank of Σ^{ij}_{BDFM} will be different for each pair of classes. Thus, the Bhattacharyya distance feature matrix for multiclass problems is achieved by summing Σ^{ij}_{BDFM} of every pair of classes

$$\Sigma_{BDFM} = \sum_{i=1}^{L-1} \sum_{j>i}^{L} \Sigma^{ij}_{BDFM} \tag{13}$$

and the eigenvectors corresponding to the desired M largest eigenvalues are taken as the feature vectors for multiclass problems.

4. Experiments and results

The proposed feature extraction method was tested on both generated data and real remotely sensed data. In the experiments, the Gaussian ML classifier was used to classify the transformed data. The parameters used in the sequential search were chosen $\delta = 0.1$ and $\alpha = 1$. First, we generated two pairs of classes with the statistics shown in Table 1. In Case 1, there is almost no difference between the mean vectors (equal means). On the other hand, the covariance matrices are identical in Case 2 (equal covariances).

Each class has a Gaussian distribution and there are 1000 samples per class. We used 300 samples as training data and the rest as testing data. The Foley–Sammon method, which applies the generalized Fisher criterion between two classes, was used for performance comparison [9]. We applied the proposed algorithm 100 times with different randomly initialized Ψ's.

Table 2 shows the average classification accuracies with standard deviations for the two pairs of classes in Table 1, where the proposed method is denoted by BHAFE. As can be seen in Table 2, the Foley–Sammon method provided a poor performance when the mean difference is small (Case 1). In general, the Fisher criterion mainly utilizes the mean differences. If there is no difference between the mean vectors, the Foley–Sammon method will fail to find the correct feature vectors. On the other hand, both the proposed method and the Foley–Sammon method achieved almost maximum classification accuracies with one feature in the equal covariance case (Case 2). In this case, the decision boundary is a straight line and the mean difference is dominant. This means that just one feature is needed to achieve the same classification accuracy as in the original space. However, the Fukunaga–Koontz method, which utilizes the covariance differences, will provide a poor performance in this example (equal covariance case) [10]. Based on these results, it can be said that the proposed method does not suffer from the issue of equal mean or equal covariance for two-class problems.

Table 1
Statistics of the generated data for two-class problems

	Case 1				Case 2			
	Class 1		Class 2		Class 1		Class 2	
Mean		0.01		−0.01		−1		1
		0		0		1		−1
Covariance	4	0	4	0	4	3.6	4	3.6
	0	4	0	1	3.6	4	3.6	4

In the next experiment, we generated data with Gaussian distributions using the statistics of classes drawn from real remotely sensed data (FSS data). Table 3 shows the major parameters of the field spectrometer system (FSS) [12]. Then, we reduced the dimension of data to $N = 10$ by combining adjacent bands. Each class has 1000 samples, out of which 500 samples were used as training data. We compared the estimated classification accuracy with the observed classification accuracy in the training stage. Fig. 4 shows the average estimated and observed classification accuracies of 10 pairs of classes. As can be seen in Fig. 4, the error estimation equation using the Bhattacharyya distance provides an accurate estimation of classification accuracies. The maximum difference is 1.37% when the number of features is $M = 1$.

In this experiment, we tried to find the minimum number of features, $M_{\min}$, for each pair of classes by applying Eq. (10) with $t = 0.99$. Table 4 shows the minimum number of features for each pair of classes assuming that 99% of the original classification accuracy (i.e. $N = 10$) needs to be achieved with these $M_{\min}$ features. It is noted that we can terminate the search algorithm at $M = M_{\min}$ without introducing a noticeable error between the pair of classes.

Fig. 5 shows the average classification accuracies of the proposed method and the conventional feature extraction methods. In this problem, we used the Foley–Sammon method and the principal component analysis (PCA) [1] for performance comparison. In Fig. 5, the proposed feature extraction method provided superior performance except $M = 1$.

Next, we tested the proposed method on 8-dimensional real data shown in Table 5. In this case, we used all samples

Table 2
Averages and standard deviations of classification accuracies (%) for the two pairs of classes generated with the statistics in Table 1

	BHAFE					Foley and Sammon	
	Training			Testing		Training	Testing
	M	Avg.	Std.	Avg.	Std.		
Case 1	1	63.33	1.76	64.04	0.83	57.7	57.3
	2	65.17	—	65.43	—	65.2	65.4
Case 2	1	98.74	0.1	98.41	0.32	98.8	98.6
	2	98.82	—	98.64	—	98.8	98.6

Table 3
Parameters of FSS

Number of bands	60 bands
Spectral coverage	0.4–2.4 μm
Altitude	60 m
IFOV (ground)	2.5 m

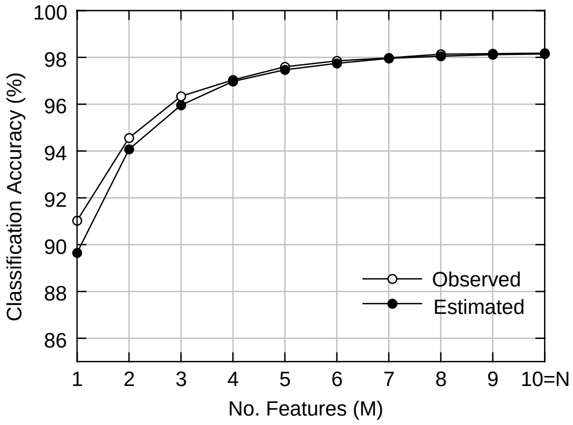

Fig. 4. The estimated accuracy and the observed accuracy.

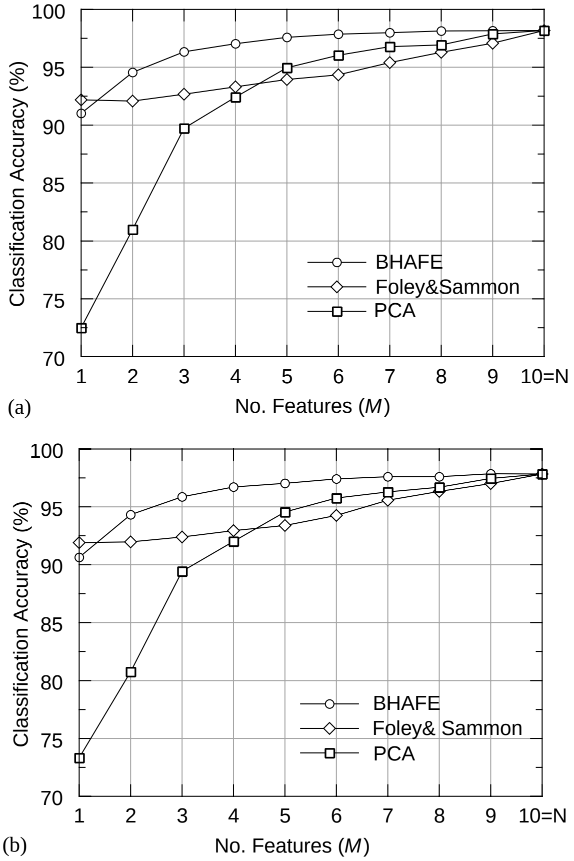

Fig. 5. Performance comparison for the two-class problems in Table 4: (a) training accuracy; and (b) testing accuracy.

for both training and testing, since the number of samples is not sufficiently large. We also used the Foley–Sammon method, PCA and the decision boundary feature extraction (DBFE) [11] for performance comparison. Fig. 6 shows the classification accuracies of the four feature extraction methods. In this experiment, the proposed method outperformed the DBFE, Foley–Sammon method and the PCA.

In the final experiment, we selected 40 classes that have sufficiently large number of samples from the FSS data. First, we randomly chose 8 classes from the 40 classes and then generated samples using the statistics of the 8 classes assuming Gaussian distributions. We generated 1000 samples for each class, out of which 500 samples were used as training data and the rest as testing data. This test was repeated 150 times using different sets of 8 classes. For performance comparison, we also tested the PCA, DBFE and the CA. Average classification accuracies for training data and testing data are shown in Fig. 7. As expected, PCA showed poor performances compared with other methods. The proposed method, DBFE, and CA provided similar performances as shown in Fig. 7. We observed that there is no

Table 4
Description of 10 pairs of classes drawn from remotely sensed data and the minimum number of features for each pair of classes ($t = 0.99$)

No.	Species	Date	No. samples	$M_{\min}$ $(N = 10)$
1	Native grass pas Pasture	78.08.16 78.10.26	212 217	5
2	Summer fallow Winter wheat	78.07.26 78.09.21	200 292	6
3	Oats Oats	78.07.26 78.09.21	173 182	6
4	Winter wheat Summer fallow	77.05.03 77.06.26	657 643	4
5	Spring wheat Spring wheat	78.07.09 78.07.26	454 515	3
6	Winter wheat Winter wheat	77.05.03 77.06.26	657 677	2
7	Spring wheat Spring wheat	78.07.26 78.08.16	515 464	5
8	Winter wheat Spring wheat	77.09.20 77.10.18	292 313	5
9	Barley Oats	78.07.26 78.07.26	102 173	4
10	Unknown crops Unknown crops	78.05.15 78.06.02	253 270	7

Table 5
Description of 8-dimensional real data for two-class problem

Dim.	Species	Date	No. samples
8	Oats Winter wheat	78.06.02 78.09.21	259 292

noticeable performance improvement for multiclass problems. This may be due to the pairwise approach to constructing the Bhattacharyya distance feature matrix. Thus, if the weighting scheme to construct the Bhattacharyya distance feature matrix is devised well, the performance of the proposed method can be improved for multiclass problems. However, it is noted that a main advantage of the proposed method over the previous feature extraction method is that it can predict the minimum number of features for a given problem.

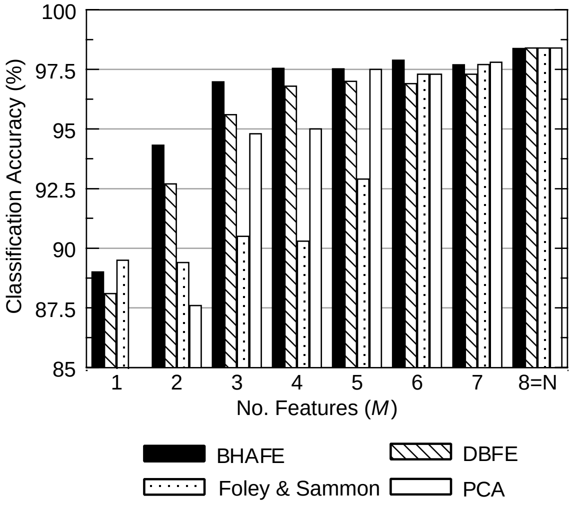

Fig. 6. Performance comparison for two-class problem (real data).

5. Conclusion

In this paper, we propose a feature extraction method based on the Bhattacharyya distance. As an optimal criterion, we approximate the classification error using the error estimation based on the Bhattacharyya distance and find a subspace of reduced dimensionality where the classification error is the minimum. We also extended this algorithm to multiclass problems by introducing the Bhattacharyya

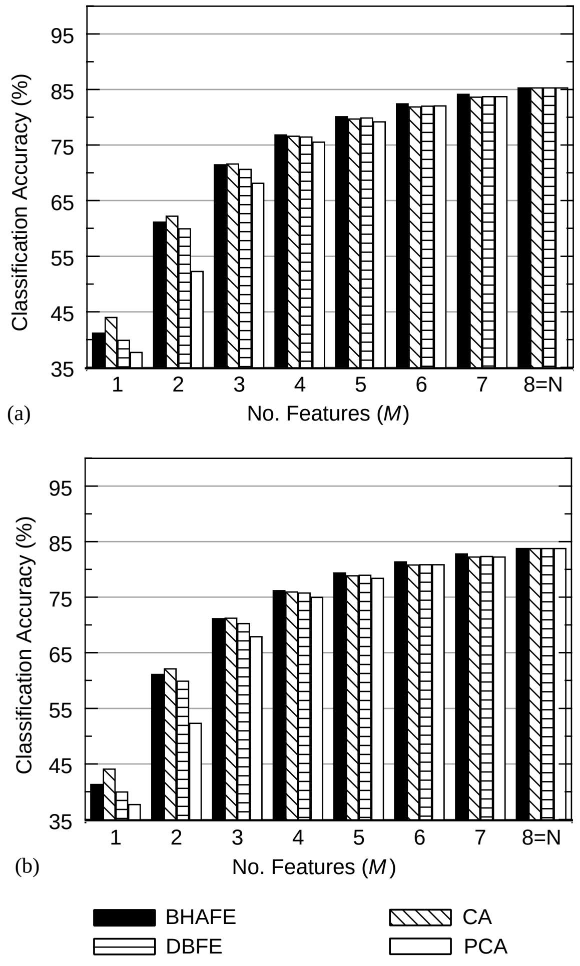

Fig. 7. Performance comparison for 8-class problems: (a) training accuracy; and (b) testing accuracy.

distance feature matrix. Experiments with generated data and real remotely sensed data showed that the proposed method was effective for two-class problems and compared favorably with conventional feature extraction methods. An important property of the proposed method for two-class problems is that it is possible to predict the minimum number of features needed for classification without significant information loss.

References

[1] J.A. Richards, Remote Sensing Digital Image Analysis, Springer, Berlin, 1993.

[2] M. Loog, R. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise Fisher criteria, IEEE Trans. Pattern Anal. Mach. Intell. 23 (7) (2001) 762–766.

[3] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd Edition, Academic Press, New York, 1990.

[4] A. Biem, S. Katagiri, B. Juang, Pattern recognition using discriminative feature extraction, IEEE Trans. Signal Process. 45 (2) (1997) 500–504.

[5] L. Buturović, Towards Bayes-optimal linear dimension reduction, IEEE Trans. Pattern Anal. Mach. Intell. 16 (4) (1994) 420–424.

[6] C. Lee, E. Choi, Bayes error evaluation of the Gaussian ML classifier, IEEE Trans. Geosci. Remote Sens. 38 (3) (2000) 1471–1475.

[7] E. Choi, C. Lee, Optimizing feature extraction for multiclass problems, IEEE Trans. Geosci. Remote Sens. 39 (3) (2001) 521–528.

[8] E. Choi, C. Lee, Feature extraction based on the Bhattacharyya distance, Proceedings of the International Conference on Geoscience and Remote Sensing, Hawaii, USA, 2000, pp. 2146–2148.

[9] D.H. Foley, J.W. Sammon, An optimal set of discriminant vectors, IEEE Trans. Comput. C-24 (3) (1975) 281–289.

[10] K. Fukunaga, W.L.G. Koontz, Application of the Karhunen–Loeve expansion to feature selection and ordering, IEEE Trans. Comput. C-19 (4) (1970) 311–318.

[11] C. Lee, D.A. Landgrebe, Feature extraction based on decision boundaries, IEEE Trans. Pattern Anal. Mach. Intell. 15 (4) (1993) 388–400.

[12] L.L. Biehl, et al., A crops and soils data base for scene radiation research, Proceedings of the Machine Processing of Remotely Sensed Data Symposium, West Lafayette, IN, USA, 1982.

About the Author—EUISUN CHOI received the B.S. and M.S. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1998 and 2000, respectively. He is currently working toward the Ph.D. degree in electronic engineering. His current research interests include pattern recognition and image/signal processing.

About the Author—CHULHEE LEE received the B.S. and M.S. degrees in electronics engineering from Seoul National University in 1984 and 1986, respectively, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1992.

From 1986 to 1987, he was a researcher in the Acoustic Laboratory at Technical University of Denmark (DTH). From 1993 to 1996, he worked with National Institutes of Health, Bethesda, MD. In 1996, he joined the faculty of the Department of Electrical and Computer Engineering, Yonsei University, Seoul, South Korea. His research interests include image/signal processing, pattern recognition, and neural networks.

Dr. Lee is a member of Tau Beta Pi, Eta Kappa Nu, and KSEA.