MAP Classifier with BDA features

Jiyong Oh and Chong-Ho Choi
School of Electrical Engineering and Computer Science, ASRI,
Seoul National University
Seoul, Korea
{*yyongman, chchoi*}*@csl.snu.ac.kr*

Nojun Kwak
Division of Electrical and Computer Engineering,
Ajou University
Suwon, Korea
nojunk@ajou.ac.kr

***Abstract*—In this paper, we derive a maximum a posteriori (MAP) classifier using the features extracted by biased discriminant analysis (BDA) in multi-class classification problems. Using the one-against-the-rest scheme we construct several feature spaces, where the MAP classifier is formulated. Although the maximum likelihood (ML) classifier is generally equivalent to the MAP classifier when the prior probability of each class is the same, an additional assumption is needed for the ML classifier to have the same results as the MAP classifier using the features extracted by BDA. We also show that the ML classifier is the same as the nearest to the mean classifier under some assumption. In order to estimate the distribution of negative samples in each reduced space, we can use the Parzen window density estimation or the Gaussian mixture model. Experimental results on several data sets indicate that the MAP classifier with BDA features provides better classification result than using the features extracted by linear discriminant analysis (LDA) or LDA using the Chrenoff criterion.**

***Keywords*-MAP; BDA; classification**

I. Introduction

In classification problems with large number of input variables, input features are extracted so that the input dimension is reduced and the resulting classifier is simple and has robust performance. Among the linear dimensionality reduction (LDR) techniques, the linear discriminant analysis (LDA) [1], [2] is the most popular and successful. In LDA, it is to find a transformation so that the ratio of the between class scatter and the within class scatter is maximized. It is very simple and powerful, especially when the differences among the class means are large compared to class variances.

In LDR methods, it is usually assumed that there are more than one class and each sample of the class is clustered around the class center. However, in many classification problems we are only interested in discriminating one class from the others. This is the one-against-the-rest problem and some of typical examples are face detection, eye detection, and content-based image retrieval. In these problems, positive samples can be assumed to be located close together, but no assumption can be made on the distribution of negative samples. For these problems the conventional LDA may not extract good features because it tries to cluster negative samples as well as positive samples and this may not contribute to the goal of the problem. In addition, for two-class problems, the conventional LDA can extract only one feature because the rank of the between-class scatter matrix is one. Several techniques have been introduced to extract more than one feature [3], [4]. In [3], the Chrenoff criterion is incorporated in the conventional LDA to make use of the covariance differences among different classes as well as the mean differences. However, it still assumes that the samples in each class are clustered around its center. Thus, such techniques does not fit to the problems where negative samples are not clustered.

For the one-against-the-rest problem, the biased discriminant analysis (BDA) was proposed in [5]. BDA finds a transformation that makes positive samples close together and negative samples far away from the mean of positive ones. It was shown in [5] that BiasMap, a kernel version of BDA, is generally comparable to the kernel support vectors machine (SVM) and is superior to kernel SVM for various kernel parameters when the number of negative sample is less than one hundred. In order to solve the small sample size (SSS) problem, which occurs when the number of sample is less than the input dimension, a regularization method was also proposed in [5]. In such a case the performance of BDA can be increased by adopting the methods in [6] and [7]. Even though there is a potential that BDA can provide better features than LDA in classification problems, BDA is not widely used compared to LDA.

In this paper, the maximum a posteriori (MAP) classifier is investigated in using the features extracted from BDA for muiti-class classification problem. According to the Baysian decision theory [8], the MAP classifier is equivalent to the maximum likelihood (ML) under the assumption that all classes have the same prior probability. However, it will be shown that an additional assumption is needed in order for the ML classifier to be equivalent to the MAP classifer in multi-class classification problems. It will be also shown that the nearest to the mean classifier is equivalent to the ML classifer under some assumption.

This paper is organized as follows. In section 2, we briefly overview the conventional BDA and the one-against-the rest scheme for applying BDA to multi-class classification problems. In section 3, we derive the MAP classifier and show that the MAP classifier is equivalent to the nearest

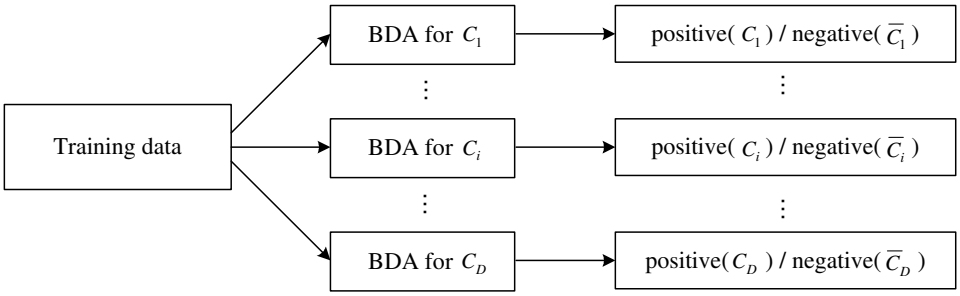

Figure 1. The one-against-the-rest classifier using BDA in multi-class classification problem

to the mean classifier under some condition. Experimental results are shown in section 4, and conclusion follows in section 5.

II. Preliminaries

A. Biased Discriminant Analysis

BDA was proposed as a variation of LDA for multimedia retrieval problems in [5]. It focuses on only one class (positive) and considers the rest to be negative. Therefore, BDA tries to cluster positive samples and to place negative samples far away from positive ones with the assumption that positive samples are close to each other and there is no information on the distribution of negative samples. The problem is to find m projection vectors $\mathbf{w}_i$, $i = 1, \ldots, m$ that maximize the following objective function

$$\begin{aligned} \mathbf{W}^* &= \arg\max_W \frac{|\mathbf{W}^T\mathbf{S}_y\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_x\mathbf{W}|}, \\ \mathbf{W} &= [\mathbf{w}_1, \ldots, \mathbf{w}_m], \\ \mathbf{S}_x &= \sum_{i=1}^{N_x}(\mathbf{x}_i - \mathbf{m}_x)(\mathbf{x}_i - \mathbf{m}_x)^T, \\ \mathbf{S}_y &= \sum_{i=1}^{N_y}(\mathbf{y}_i - \mathbf{m}_x)(\mathbf{y}_i - \mathbf{m}_x)^T. \end{aligned} \tag{1}$$

Here, $\{\mathbf{x}_i : i = 1, \ldots, N_x\}$ and $\{\mathbf{y}_i : i = 1, \ldots, N_y\}$ denote the positive and negative samples respectively, and $\mathbf{m}_x$ is the mean vector of positive samples. This optimization problem is equivalent to the following generalized eigenvalue problem,

$$\mathbf{S}_y\mathbf{w}_i = \lambda_i\mathbf{S}_x\mathbf{w}_i, \quad i = 1, \ldots, m. \tag{2}$$

This problem can be solved by diagonalizing $\mathbf{S}_x$ and $\mathbf{S}_y$ simultaneously [2], [9]. In LDA, the number m_{LDA} of features that can be extracted is limited by $min(D-1, l)$, where D is the number of classes and l is the input dimension. The number m_{BDA} of features that can be extracted by BDA is greater than m_{LDA} because the rank of $\mathbf{S}_y$ is $min(N_y, l)$ $(> min(D-1, l))$.

The Euclidean distance from the positive mean in the reduced space is used to determine whether a test sample belongs to a positive class in [5]. However, a classifier based on Euclidean distance from the positive mean may not show good performance when the features extracted by BDA are used. In the next section, we will show that the MAP classifier can be related to the nearest-to-the-mean classifier under some condition.

B. Multi-Class Classification using BDA

Let us consider a set C_i of samples each of which belongs to one of D classes, $\{C_i : i = 1, \ldots, D\}$. Class C_i has N_i samples and the total number of samples are N ($\sum_{i=1}^{D} N_i = N$). Since BDA has been proposed for image retrieval or detection problems, it focuses on one class (positive) and does not pay attention on the statistical properties of negative samples. It considers each negative sample as a class, which results in $(1+N_y)$ classes. However this is not good way to solve multi-class classification problem. In [10], D reduced spaces have been constructed by using the one-against-the-rest scheme in order to apply BDA to D-class classification problem. As shown in Fig. 1, the samples in the i-th class are considered positive and the other samples in the remaining $D-1$ classes are considered negative in the i-th reduced space. Let $\mathbf{W}^i$ denote the transformation matrix found in (1) for the i-th class. Then, a sample $\mathbf{z}$ is projected into the i-th reduced space as

$$\mathbf{z}^i = (\mathbf{W}^i)^T(\mathbf{z} - \mathbf{m}_i). \tag{3}$$

Here, $\mathbf{m}_i$ is the mean of the samples belonging to C_i in the original input space. Note that we use $\mathbf{z} - \mathbf{m}_i$ in (3) so that positive samples have zero-mean in its reduced space. Also, by using the simultaneous diagonalization technique as in [2] and [9], the covariance matrix $\mathbf{\Sigma}_i^j$ of the in C_j samples in the i-th reduce space can be made to be an identity matrix when $i = j$. Hearafter, we assume that $\mathbf{\Sigma}_i^i = \mathbf{I}$.

III. MAP and ML classifiers

In Baysian decision theory [8], a sample $\mathbf{z}$ is assigned to the class C_i that has the maximum a posteriori probability $P(C_i|\mathbf{z})$. In the one-against-the-rest scheme for BDA, we must use $P(C_i|\mathbf{z}^i)$ instead of $P(C_i|\mathbf{z})$ because the projection $\mathbf{z}^i$ varies depending on i. For the MAP classifier, a test

sample $\mathbf{z}$ is assigned to class $C_{MAP}(\mathbf{z})$ as the following

$$\begin{aligned} C_{MAP}(\mathbf{z}) &= \arg\max_i P(C_i|\mathbf{z}^i) \\ &= \arg\max_i \frac{P(C_i)p(\mathbf{z}^i|C_i)}{p(\mathbf{z}^i)}. \end{aligned} \tag{4}$$

Since $p(\mathbf{z}^i)$ depends on the argument i and can be represented as

$$p(\mathbf{z}^i) = P(C_i)p(\mathbf{z}^i|C_i) + P(\overline{C}_i)p(\mathbf{z}^i|\overline{C}_i),$$

where $\overline{C}$ represents the classes except C_i, $P(C_i) = N_i/N$, and $P(\overline{C}_i) = 1 - p(C_i)$. While $p(\mathbf{z}^i|C_i)$ can be assumed to be the Gaussian distribution with zero-mean and identity-covariance matrix, $p(\mathbf{z}^i|\overline{C}_i)$ is unknown since there is no information on the distribution of negative samples. However, we can estimate $p(\mathbf{z}^i|\overline{C}_i)$ by using the Parzen window density estimation [11]. Given a set of negative samples $\mathbf{y}_1^i, \cdots, \mathbf{y}_{N_y}^i$ in the i-th reduced space, $p(\mathbf{z}^i|\overline{C}_i)$ can be calculated as

$$p(\mathbf{z}^i|\overline{C}_i) = \frac{1}{N_y}\sum_{j=1}^{N_y}\phi(\mathbf{z}^i - \mathbf{y}_j^i, h), \tag{5}$$

where $\phi(\cdot)$ and h are the window function and the window width parameter, respectively. The Gaussian window function is given by

$$\phi(\mathbf{v}, h) = \frac{1}{(2\pi)^{l/2}h^n|\mathbf{\Sigma}|^{1/2}}exp(-\frac{\mathbf{v}^T\mathbf{\Sigma}^{-1}\mathbf{v}}{2h^2}), \tag{6}$$

where $\mathbf{\Sigma}$ is a covariance matrix of n-dimensional vectors. However, it requires a lot of computational effort to estimate the density by the Parzen window. If it is assumed that the samples in each class has Gaussian distribution in the reduced space, $p(\mathbf{z}^i|\overline{C}_i)$ can be represented as a mixture of $D-1$ Gaussian distributions.

$$p(\mathbf{z}^i|\overline{C}_i) = \sum_{\substack{j=1 \\ j\neq i}}^{D}\frac{N_j}{N - N_i}p^i(\mathbf{z}^i|C_j) \tag{7}$$

where

$$p^i(\mathbf{z}^i|C_j) = \frac{exp[-\frac{1}{2}(\mathbf{z}^i - \boldsymbol{\mu}_j^i)^T(\mathbf{\Sigma}_j^i)^{-1}(\mathbf{z}^i - \boldsymbol{\mu}_j^i)]}{\sqrt{(2\pi)^m}|\mathbf{\Sigma}_j^i|^{\frac{1}{2}}}. \tag{8}$$

Here, $\boldsymbol{\mu}_j^i$ is the mean of the samples belonging to the j-th class in the i-th reduced space and $|\cdot|$ is the determinant operator. As a result, we can assign a test sample $\mathbf{z}$ to the class $C_{MAP}(\mathbf{z})$ that maximizes a posteriori probability $P(C_i|\mathbf{z}^i)$ by using either (5) and (6) or (7) and (8).

According to the Bayesian decision theory [8], the MAP classifier is equivalent to the maximum likelihood (ML) classifier under the assumption that each class has the same prior probability. In order for the maximum likelihood classifier $C_{ML}(\mathbf{z})$ to be equivalent to the maximum a posteriori classifier $C_{MAP}(\mathbf{z})$, it needs to be assumed that $p(\mathbf{z}^1) = \cdots = p(\mathbf{z}^D)$ as well as $P(C_1) = \cdots = P(C_D)$.

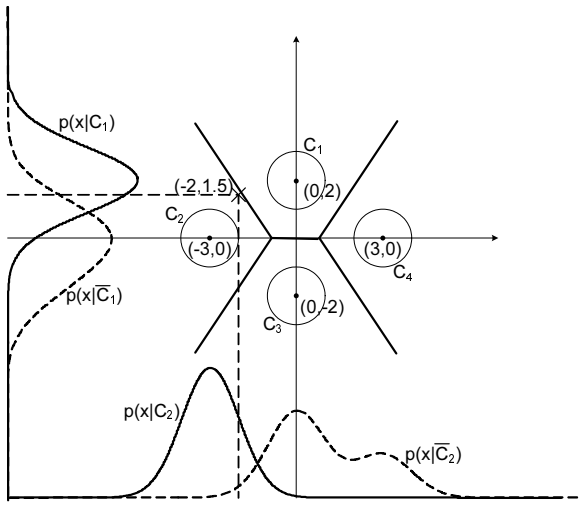

Figure 2. An example of four classes and the optimal decision boundary

$$\begin{aligned} C_{MAP}(\mathbf{z}) &= \arg\max_i P(C_i|\mathbf{z}^i) \\ &= \arg\max_i \frac{P(C_i)p(\mathbf{z}^i|C_i)}{p(\mathbf{z}^i)} \\ &= \arg\max_i p(\mathbf{z}^i|C_i) \\ &= C_{ML}(\mathbf{z}) \end{aligned}$$

Under these the assumptions,

$$\begin{aligned} C_{ML}(\mathbf{z}) &= \arg\max_i p(\mathbf{z}^i|C_i) \\ &= \arg\max_i \frac{exp[-\frac{1}{2}(\mathbf{z}^i - \boldsymbol{\mu}_i^i)^T(\mathbf{\Sigma}_i^i)^{-1}(\mathbf{z}^i - \boldsymbol{\mu}_i^i)]}{\sqrt{(2\pi)^m}|\mathbf{\Sigma}_i^i|^{\frac{1}{2}}} \\ &= \arg\min_i \frac{1}{2}(\mathbf{z}^i - \boldsymbol{\mu}_i^i)^T(\mathbf{\Sigma}_i^i)^{-1}(\mathbf{z}^i - \boldsymbol{\mu}_i^i) \\ &= \arg\min_i \|\mathbf{z}^i\|, \end{aligned} \tag{9}$$

because $\boldsymbol{\mu}_i^i = 0$ and $\mathbf{\Sigma}_i^i = \mathbf{I}$.

Note that $\|\mathbf{z}^i\|$ is the Euclidean distance between a test sample and the positive mean in the i-th reduced space. Therefore, the ML classifier gives the same result as the Euclidean distance-based classifier under the assumption. Even though $C_{ML}(\mathbf{z})$ needs only D distances to classify a test sample, the assumption $p(\mathbf{z}^1) = \cdots = p(\mathbf{z}^D)$ is not true in general. To obtain better classification performance, we should use the MAP classifier instead of the ML classifier. An example of this is illustrated in Fig. 2. There are four classes and the samples in each class has the Gaussian distribution with the mean $[0 \quad 2]^T$, $[-3 \quad 0]^T$, $[0 \quad -2]^T$, $[3 \quad 0]^T$ and the covariance matrices $\mathbf{I}$. Each circle represents the contour corresponding to one standard deviation from

Table I
UCI DATA SETS USED IN THE EXPERIMENTS

Data set	# of variables	# of class	# of instances
Balance	4	3	625
Breast cancer	9	2	683
Heart disease	13	2	297
Ionosphere	34	2	351
Pima	8	2	768
Sonar	60	2	208
Vehicle	18	4	846

the center of a class and the lines represent the optimal decision boundary. We assume that each class has the same prior probability. Though this problem is difficult to solved by using LDA, BDA can find the feature which can well separate the samples in one class from the others.

Let us consider a test sample $\mathbf{z} = [-2 \quad 1.5]^T$ which should be classified as C_2. As can be seen in the figure, the transformation vector which makes the samples in C_1 close together and the others far away from the mean of C_1 is $[0 \quad 1]^T$ and similarly it is $[1 \quad 0]^T$ for the samples in C_2, and these can be found from (1). The reduced spaces for C_1 and C_2 are represented on the left and the bottom in Fig. 2. Note that the distributions of the negative samples in the reduced spaces for C_1 and C_2 are different from each other. The ML classifier assigns C_1 to $\mathbf{z}$ since $p(\mathbf{z}^1|C_1) = 0.3521 > p(\mathbf{z}^2|C_2) = 0.2420$, which is wrong. However, $\mathbf{z}$ can be classified as C_2 by the MAP classifier because $P(C_1|\mathbf{z}^1) = 0.5754 < P(C_2|\mathbf{z}^2) = 0.6914$.

IV. EXPERIMENTS

In this section, we evaluate the performance of MAP classifier using the Parzen window (Par) and Gaussian mixture model (GMM), and the nearest to the mean (N2M) classifier by using seven data sets from the UCI machine learning repository [12] as shown in Table I. The features were extracted by BDA, LDA, and LDA using the Chrenoff distance (Chre-LDA) [3]. In finding the solution for (1), we used $\mathbf{S}'_x = (1-\mu)\mathbf{S}_x + \mu\frac{tr(\mathbf{S}_x)}{l}\mathbf{I}$ instead of $\mathbf{S}_x$ as in [5]. Here, μ is a regularization parameter and was set to 0.1. For each data set, we performed 10-fold cross validation 10 times and computed the average classification rate and its standard deviation. Each input variable in the training set was normalized to have zero mean and unit variance, and so were the variables in the test set. For Parzen window estimation, the window width parameter h was set to $0.3\sqrt{m}$, where m denotes the dimension of feature vectors.

The result for each data set is shown in Table II, along with the optimal number of features in the parentheses. The best result for each data set is indicated in boldface.

Table II shows that MAP (Par) classifier gives better performance than the nearest to the mean (N2M) classifier in most cases. There are some cases that N2M classifier gives better classification rate, but the differences are very negligible. Generally, the MAP (Par) classifier using the features extracted by BDA gives better than the MAP (Par) classifier using the features extracted by LDA or LDA with the Chrenoff criterion. Since the MAP (GMM) classifier gives nearly the same results as the MAP (Par) classifier, one can use the MAP (GMM) classifier to save computation cost without sacrificing performance.

V. CONCLUSION

In this paper, we derived the MAP classifier which can be used with the features extracted by BDA in multi-class classification problems. We constructed several subspaces from the one-against-the-rest scheme, and formulated MAP classifier based on the Baysian decision rule. We showed that under some assumption, the MAP classifier gives the same result as the ML classifier, which is also equivalent to the nearest to the mean classifer under another assumption. We could use the Parzen window density estimation or the Gaussian mixture model to estimate the distribution of negative samples, which is used in the MAP classifier. Through experimental results, we showed the MAP classifier using the features extracted by BDA, rather than LDA or LDA with the Chrenoff criterion, gave the best performance. We also demonstrated that GMM can be an effective alternative to the Parzen window for the MAP classifier to reduce the computational time.

ACKNOWLEDGMENT

This work was partially supported by the Ajou University under Research Grant 20083770.

REFERENCES

[1] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, vol. 8, pp. 376–386, 1938.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic Press, 1990.

[3] M. Loog and R. P. Duin, "Linear dimensionality reduction via a heteroscedastic extension of lda: The chernoff criterion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 732–739, June 2004.

[4] T. Okada and S. Tomita, "An optimal orthonormal system for discriminant analysis," *Pattern Recognition*, vol. 18, no. 2, pp. 139–144, 1985.

[5] X. S. Zhou and T. S. Huang, "Small sample learning during multimedia retrieval using biasmap," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Dec. 2001.

[6] D. Tao and X. Tang, "Kernel full-space biased discriminant analysis," in *Proc. IEEE Conf. Multimedia and Expo*, June 2004.

Table II
OVERVIEW OF THE EXPERIMENTAL RESULTS

Data set	BDA			LDA		Chre-LDA	
	N2M	MAP (Par)	MAP (GMM)	N2M	MAP (Par)	N2M	MAP (Par)
Balance	**93.73±0.35** (3)	93.60±0.42 (3)	91.68±0.00 (1)	70.66±0.61 (2)	87.22±0.46 (2)	70.66±0.61 (4)	89.87±0.19 (4)
Breast	96.50±0.08 (1)	96.57±0.17 (1)	**96.90±0.13** (1)	96.19±0.14 (1)	96.75±0.13 (1)	96.19±0.35 (9)	96.81±0.22 (1)
Heart	82.46±0.64 (1)	**84.18±0.48** (1)	84.04±0.55 (1)	83.43±0.41 (1)	83.43±0.69 (1)	83.67±0.64 (1)	84.04±0.71 (1)
Ionosphere	91.05±0.60 (5)	94.05±0.59 (5)	**94.59±0.64** (5)	87.04±0.71 (1)	86.92±0.81 (1)	87.82±0.80 (22)	88.21±0.24 (3)
Pima	75.10±0.60 (8)	77.20±0.54 (1)	77.16±0.40 (1)	76.04±0.29 (1)	77.15±0.34 (1)	76.29±0.39 (1)	**77.30±0.34** (1)
Sonar	**81.78±1.02** (8)	81.54±1.25 (7)	81.49±0.99 (17)	74.62±1.91 (1)	74.42±1.47 (1)	74.95±1.59 (4)	81.25±1.94 (35)
Vehicle	78.03±0.46 (10)	**85.56±0.37** (18)	85.33±0.42 (18)	78.14±0.27 (3)	78.69±0.42 (3)	78.27±0.45 (10)	83.38±0.37 (11)
Average	85.52	**87.53**	87.31	80.87	83.50	81.12	85.84

[7] D. Tao, X. Tang, X. Li, and Y. Rui, "Direct kernel biased discriminant analysis: A new content-based image retrieval relevance feedback algorithm," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 716–727, Aug. 2006.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[9] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition." *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 467–476, April 2002.

[10] N. Kwak and J. Oh, "Feature extraction for one-class classification problems: Enhancements to biased discriminant analysis," *Pattern Recognition*, vol. 42, no. 1, pp. 17–26, 2009.

[11] E. Parzen, "Analyzing appearance and contour based methods for object categorization," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[12] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html