Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Logistic Discriminant Analysis

Takio Kurita
Neuroscience Resarch Institute
AIST
Tsukuba, Japan
takio-kurita@aist.go.jp

Kenji Watanabe
AIST
Tsukuba, Japan
kenji-watanabe@aist.go.jp

Nobuyuki Otsu
AIST Fellow
AIST
Tsukuba, Japan
otsu.n@aist.go.jp

***Abstract*—Linear discriminant analysis (LDA) is one of the well known methods to extract the best features for the multi-class discrimination. Otsu derived the optimal nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities and showed that the ONDA was closely related to Bayesian decision theory (*the posterior* probabilities). Also Otsu pointed out that LDA could be regarded as a linear approximation of the ONDA through the linear approximations of the Bayesian *posterior* probabilities. Based on this theory, we propose a novel nonlinear discriminant analysis named logistic discriminant analysis (LgDA) in which *the posterior* probabilities are estimated by multi-nominal logistic regression (MLR). The experimental results are shown by comparing the discriminant spaces constructed by LgDA and LDA for the standard repository datasets.**

***Keywords*— linear discriminant analysis, nonlinear discriminant analysis, multi-nominal logistic regression, logistic discriminant analysis, Bayesian decision theory**

I. INTRODUCTION

Feature extraction is one of the most important problems in pattern recognition. Linear discriminant analysis (LDA) is one of the well known methods to extract the best features for multi-class discrimination. LDA is formulated as a problem to find an optimal linear mapping by which the within-class scatter in the mapped feature space is made as small as possible relative to the between-class scatter. LDA is useful for linear separable cases, but for more complicated cases, it is necessary to extend it to non-linear.

Otsu derived the optimal nonlinear discriminant analysis (ONDA) by assuming the underlying probabilities [1, 2, 3]. He showed that the optimal non-linear discriminant mapping was closely related to Bayesian decision theory (The *posterior* probabilities). Also Otsu pointed out that LDA can be regarded as the linear approximation of the ultimate ONDA through the linear approximations of the Bayesian *posterior* probabilities. This theory suggests that we can construct a novel nonlinear discriminant mapping if we utilize nonlinear estimates of the *posterior* probabilities. Since the outputs of the trained multi-layered Perceptron (MLP) for pattern classification problems can be regarded as the approximations of the *posteriori* probabilities [4], Kurita et al. [5] proposed the neural network based non-linear discriminant analysis by using the outputs of the trained MLP. Recently non-linear discriminant space can be constructed by the kernel discriminant analysis [6, 7]. This is also interpreted as an approximation of the ultimate ONDA.

LDA is the linear approximation of the ONDA through the linear approximations of the Bayesian *posterior* probabilities. However, a linear model is not suitable to estimate the *posterior* probabilities. Logistic regression (LR) is one of the simplest models for binary classification and can directly estimate the *posterior* probabilities. Multi-nominal logistic regression (MLR) is a natural extension of LR to multi-class classification problems. They are known as the members of the generalized linear model (GLM) which is a flexible generalization of ordinary least squares regression. By modifying the outputs of the linear predictor by the link function, MLR can more naturally estimate the *posterior* probabilities.

In this paper, we propose a novel nonlinear discriminant analysis in which the Bayesian *posterior* probabilities are estimated by MLR. The proposed method is named as logistic discriminant analysis, in short LgDA. It is expected that the discriminant space constructed by LgDA is better than the one constructed by LDA, because MLR is more natural as the probability estimator than the linear approximation of the *posterior* probabilities used in LDA. The experimental results are shown by comparing the discriminant spaces constructed by LgDA and LDA for the standard repository datasets.

II. LINEAR AND NON-LINEAR DISCRIMINANT ANALYSIS

A. Linear Disctiminant Analysis

Let an m dimensional feature vector be $\boldsymbol{x} = (x_1, \cdots, x_m)^{\mathrm{T}}$. Consider K classes $\{C_k\}_{k=1}^{K}$. As training samples, we have N feature vectors and they are labeled as one of the K classes. Then LDA constructs a dimension reducing linear mapping from the input feature vector $\boldsymbol{x}$ to a new feature vector $\boldsymbol{y}$

$$\boldsymbol{y} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}, \tag{1}$$

where $\boldsymbol{A}$ = $[a_{ij}]$ is the coefficient matrix. The discriminant criterion

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

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

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

The optimal coefficient matrix $\boldsymbol{A}$ is then obtained by solving the following eigen equation

$$\Sigma_B \boldsymbol{A} = \Sigma_T \boldsymbol{A}\Lambda \quad \left(\boldsymbol{A}^{\mathrm{T}} \Sigma_T \boldsymbol{A} = \boldsymbol{I}\right), \tag{3}$$

where Λ is a diagonal matrix of eigen values and $\boldsymbol{I}$ denotes the unit matrix. The matrices Σ_T and Σ_B are respectively the total covariance matrix and the between-class covariance matrix of the input vectors $\boldsymbol{x}$, and they are computed as follows:

$$\begin{aligned} \Sigma_T &= \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_T)^{\mathrm{T}}, \\ \Sigma_B &= \sum_{k=1}^{K} P(C_k)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^{\mathrm{T}}. \end{aligned} \tag{4}$$

Where $P(C_k)$, $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{x}}_T$ denote *a priori* probability of the class C_k ($P(C_k) = N_k / N$, N_k is the number of input vectors of the class C_k and N is the number of input vectors), the mean vector of the class C_k and the total mean vector, respectively.

The j-th column of $\boldsymbol{A}$ 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 $min(K-1, N)$.

B. Optimal Nonlinear Discriminant Analysis

Otsu derived the optimal nonlinear discriminant analysis (NDA) by assuming the underlying probabilities [1, 2]. Similarly to the LDA, the ONDA constructs the dimension reducing nonlinear mapping which maximizes the discriminant criterion J. The optimal non-linear discriminant mapping is given by

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

where $P(C_k|\mathbf{x})$ is the Bayesian *posterior* probability of the class C_k given the input $\boldsymbol{x}$. The vectors $\boldsymbol{u}_k$ (k = 1,…, K) are class representative vectors which are determined by following eigen-equation:

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

where Γ is a $K \times K$ matrix whose elements are γ_{ij}

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

and the other matrices are represented as follows:

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

It is important to notice that the optimal non-linear mapping is closely related to Bayesian decision theory, namely the *posterior* probabilities $P(C_k|\mathbf{x})$. Along this line, Fukunaga *et al* discussed the various properties of the criterion from the viewpoint of non-linear mappings [8].

Thus, we can construct optimal nonlinear discriminant features by ONDA from a given input features if we can know or estimate all the Bayesian *posteriori* probabilities correctly. However, it is usually difficult to estimate them from the input features.

C. Linear approximation of NDA

In the previous subsection, we explained ONDA as a ultimate nonlinear extension of LDA. Then we may have the following question: in what sense does LDA approximate NDA?

Let

$$L(C_k|\mathbf{x}) = \mathbf{b}^{(k)}\mathbf{x} + b_0^{(k)} \tag{9}$$

be a linear approximation of the Bayesian *posteriori* probabilities which minimizes the mean square error as follows:

$$\varepsilon^2 = \int \left\{P(C_k|\mathbf{x}) - L(C_k|\mathbf{x})\right\}^2 p(\mathbf{x}) d\mathbf{x} . \tag{10}$$

Otsu [2, 11] already pointed out that the optimal linear function is given by

$$L(C_k|\boldsymbol{x}) = P(C_k)\left\{(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}_T)^{\mathrm{T}} \Sigma_T^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_T) + 1\right\}, \tag{11}$$

where Σ_T denotes the total covariance matrix. It is interesting to note that this function has unit-sum property from $\boldsymbol{x}$ as follows:

$$\sum_{k=1}^{K} L(C_k|\boldsymbol{x}) = 1 . \tag{12}$$

Let us substitute these linear approximations $L(C_k|\boldsymbol{x})$ for the Bayesian *posterior* probabilities $P(C_k|\mathbf{x})$ in (5) and (6) of ONDA. By this substitution, (5) becomes

$$\begin{aligned}\mathbf{y} &= \sum_{k=1}^{K} L(C_k|\mathbf{x})\mathbf{u}_k \\ &= \mathbf{U}^{\mathrm{T}}\mathbf{P}\mathbf{M}^{\mathrm{T}}\Sigma_T^{-1}(\mathbf{x}-\bar{\mathbf{x}}_T)+\mathbf{U}^{\mathrm{T}}\mathbf{p}\end{aligned}, \quad (13)$$

where $\boldsymbol{M} = [(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_T),\ldots,(\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_T)]^{\mathrm{T}}$ and $\mathbf{p} = (P(C_1),\ldots,P(C_K))^T$. Γ in (6) becomes as follows:

$$\Gamma = \mathbf{P}\mathbf{M}^{\mathrm{T}}\Sigma_T^{-1}\mathbf{M}\mathbf{P}. \quad (14)$$

By multiplying $\boldsymbol{M}$ from the left and substituting $\Sigma_T^{-1}\mathbf{M}\mathbf{P}U$ for A, we have the same eigen-equation with (3). This means that LDA is the linear approximation of ONDA through the linear approximation $L(C_k|\boldsymbol{x})$ of the *posteriori* probabilities $P(C_k|\mathbf{x})$.

III. LOGISTIC DISCRIMINANT ANALYSIS

A. Multi-nominal Logistic Regression

Logistic regression (LR) is one of the simplest models for binary classification and can directly estimate the *posterior* probabilities. Multi-nominal logistic regression (MLR) is a natural extension of LR to multi-class classification problems [9]. It is known as one of the generalized linear model (GLM) which is a flexible generalization of ordinary least squares regression. By modifying the outputs of the linear predictor by the link function, MLR can naturally estimate the *posterior* probabilities.

For K-class classification problem, let $D = \{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^{N}$ be the given training data, where $\mathbf{x}_i = (x_{i1},\cdots,x_{im})^T$ is the i-th input vector, and $\boldsymbol{t}_i \in \boldsymbol{T} = \{\boldsymbol{t} \mid \boldsymbol{t} \in \{0,\ 1\}^k,\ \|\boldsymbol{t}\|_{L1} = 1\}$ is the class representative vector for the i-th input vector. The outputs of MLR estimate the *posterior* probabilities $P(t_i^k \mid \mathbf{x}_i)$. They are defined as follows:

$$\hat{P}(t_i^k \mid \mathbf{x}_i) = y^k(\mathbf{x}_i) = \frac{\exp(\eta_i^k)}{1+\sum_{j=1}^{K-1}\exp(\eta_i^j)}, \quad (15)$$

$$\eta_i^k = \mathbf{x}_i^T\hat{\mathbf{w}}^k + b_k = \hat{\mathbf{x}}_i^T\mathbf{w}^k. \quad (16)$$

Where $\hat{\mathbf{w}}^k = (w_{1k},\cdots,w_{mk})^T$ and b_k are the weight vector and the bias term of k-th class, respectively. To simplify the notation, we include the bias term in the vectors as $\mathbf{w}^k = (w_{1k},\cdots,w_{mk},b_k)^T$ and $\hat{\mathbf{x}}_i = (x_{i1},\cdots,x_{im},1)^T$. In matrix notation, we use $\mathbf{W} = (\mathbf{w}^1,\cdots,\mathbf{w}^{K-1})$ and $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_1,\cdots,\hat{\mathbf{x}}_N)$. The optimal parameters of MLR are obtained by minimizing the negative log-likelihood

$$\boldsymbol{W} = \arg\min_{\boldsymbol{w}} E_D, \quad (17)$$

$$E_D = \sum_{i=1}^{N}\sum_{j=1}^{K-1}\left\{t_i^j\ log\left(1+\sum_{l=1}^{K-1}exp(\eta_i^l)\right) - t_i^j\eta_i^j\right\}. \quad (18)$$

Equation (17) represents a convex optimization problem and it has only a single, global minimum. Again the optimal parameter $\boldsymbol{W}$ can be efficiently found using Newton-Raphson method or an iterative re-weighted least squares (IRLS) procedure. In each iteration step, $\boldsymbol{W}$ is updated by

$$\boldsymbol{W}^{t+1} = \boldsymbol{H}^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{Z}, \quad (19)$$

where $\mathbf{H} = \mathbf{G}^T\mathbf{R}\mathbf{G}$ is the block Hessian matrix, and $\boldsymbol{H}^{-1}$ is the inverse matrix of $\boldsymbol{H}$. $\mathbf{G} = diag(\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{K-1})$ is the block diagonal matrix of $\hat{\boldsymbol{X}}$, and $\hat{\boldsymbol{X}}^k = \hat{\boldsymbol{X}}$. The matrix $\mathbf{R}$ is the block matrix defined as follows:

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{11} & \cdots & \boldsymbol{R}_{1(K-1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{R}_{(K-1)1} & \cdots & \boldsymbol{R}_{(K-1)(K-1)} \end{pmatrix},$$

$$\boldsymbol{R}_{jk} = diag(r_1^{jk},\ \cdots,\ r_N^{jk}), \quad (20)$$

$$r_n^{jk} = y_n^j(\delta_{jk} - y_n^k),\quad \delta_{jk} = \begin{cases} 1 & if \quad j=k \\ 0 & otherwise \end{cases}.$$

The vector $\mathbf{Z}$ is the block vector with elements

$$\boldsymbol{z}_k = \sum_{j=1}^{K-1}\boldsymbol{R}_{kj}\boldsymbol{\eta}^j - (\boldsymbol{y}^k - \boldsymbol{t}^k). \quad (21)$$

Equation (19) is repeated until it converges

B. Regularization of MLR

In general, the regularization term is introduced to control the over-fitting. The regularization methods of MLR were proposed such as the shrinkage method (regularized MLR) and locality preserving multi-nominal logistic regression (LPMLR) [10]. In shrinkage method, unnecessary growth of the parameters is penalized by introducing the regularization term E_W defined as follows:

$$E_W = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}\boldsymbol{w}_j{}^{\mathrm{T}}\boldsymbol{w}_k. \quad (22)$$

Then the optimal parameters of the regularized MLR is determined by minimizing the negative log-likelihood as

$$\boldsymbol{W} = \arg\min_{\boldsymbol{w}}(E_D + \lambda_{\boldsymbol{w}}E_W). \quad (23)$$

Equation (23) represents a convex optimization problem, and λ_w is the pre-specified regularization parameter of E_W.

The multiplicative update rule for the regularized MLR is the same as (19). However, the elements of the block Hessian matrix $\mathbf{H}$ are different from MLR [10]. the block Hessian matrix $\mathbf{H}$ of the regularized MLR is defined as follows:

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \cdots & \mathbf{H}_{1(K-1)} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{(K-1)1} & \cdots & \mathbf{H}_{(K-1)(K-1)} \end{pmatrix}, \tag{24}$$

$$\mathbf{H}_{jk} = \begin{cases} \hat{\mathbf{X}}^{\mathrm{T}} \mathbf{R}_{jk} \hat{\mathbf{X}} + 2\lambda_{\mathrm{w}} \mathbf{I} & if \quad j = k \\ \hat{\mathbf{X}}^{\mathrm{T}} \mathbf{R}_{jk} \hat{\mathbf{X}} + \lambda_{\mathrm{w}} \mathbf{I} & otherwise \end{cases}.$$

Where $\mathbf{I}$ is the identity matrix. $\mathbf{R}$ is the block matrix similar to (20). $\mathbf{Z}$ is the block vector with elements similar to (21).

C. Logistic Discriminant Analysis

After the training of the parameters using the sufficient number of samples, the outputs of the ordinal MLR or the regularized MLR can be interpreted as estimates of the Bayesian *posterior* probabilities $\left(P(C_1|\mathbf{x}),\ldots,P(C_K|\mathbf{x})\right)^{\mathrm{T}}$. By substituting the Bayesian *posterior* probabilities in the ONDA with the outputs of the ordinal MLR or the regularized MLR, we can directly construct an approximation of ONDA. We call this method logistic discriminant analysis (LgDA). It is expected that the discriminant space constructed by LgDA is better than the one constructed by LDA, because MLR is more natural as the estimates of the *posterior* probabilities than the linear approximation of them used in LDA.

Let the outputs of the ordinal MLR or the regularized MLR for an input vector $\boldsymbol{x}$ be $\mathbf{y}(\mathbf{x}) = \left(y^1(\mathbf{x}),\ldots,y^K(\mathbf{x})\right)^{\mathrm{T}}$. Then *a priori* probability $P(C_k)$ is approximated as follows:

$$\widetilde{P}(C_k) = \frac{1}{N}\sum_{i=1}^{N} y^k(\mathbf{x}_i) = \bar{y}_k \quad (k = 1,\ldots,K). \tag{25}$$

The approximation of the matrix Γ is also given by

$$\widetilde{\Gamma} = \frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{y}(\mathbf{x}_i) - \bar{\mathbf{y}}\right)\left(\mathbf{y}(\mathbf{x}_i) - \bar{\mathbf{y}}\right)^{\mathrm{T}}. \tag{26}$$

Thus the non-linear discriminant mapping is obtained as

$$\tilde{\mathbf{y}} = \sum_{k=1}^{K} y^k(\mathbf{x})\tilde{\mathbf{u}}_k. \tag{27}$$

The representative vectors of each class $\tilde{\mathbf{u}}_k$ are determined by the following eigen equation

$$\widetilde{\Gamma}\widetilde{U} = \widetilde{P}\widetilde{U}\widetilde{\Lambda}, \tag{28}$$

where the matrices such as $\widetilde{P}$, $\widetilde{U}$ and $\widetilde{\Lambda}$ are defined as follows:

$$\begin{aligned} \widetilde{\mathbf{U}} &= [\tilde{\mathbf{u}}_1,\ldots,\tilde{\mathbf{u}}_K]^{\mathrm{T}}, \\ \widetilde{\mathbf{P}} &= diag\left(\widetilde{P}(C_1),\ldots,\widetilde{P}(C_K)\right), \\ \widetilde{\Lambda} &= diag\left(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_K\right). \end{aligned} \tag{29}$$

If the outputs of the ordinal MLR or the regularized MLR can give sufficiently good approximation to the Bayesian *posterior* probabilities, it is expected that the nonlinear discriminant mapping defined by (27) constructs the good approximation of the ultimate nonlinear discriminant mapping ONDA in terms of the discriminant criterion.

IV. Experiments

To show the effectiveness of the proposed LgDA, the discriminant space was compared with LDA using the standard repository datasets for multiclass classification [11]. Fig.1 shows the 2-dimensional discriminant spaces constructed by LDA and LgDA for Satimage dataset which has 36 dimensional features from 6 classes and consists of 4435 training samples and 2000 test samples. The regularization parameter λ_w of LgDA was determined by grid search. Fig.2 shows the 2-dimensional discriminant spaces for Balance dataset which has 4 dimensional features from 3 classes and consists of 625 samples. This dataset was randomly divided into 90 training samples and 535 test samples. The test samples are plotted in the constructed discriminant spaces. It is noticed that samples of each class are gathered around the class representative vectors in the discriminant space constructed by LgDA but samples are more spread in the discriminant space by LDA. Especially in the case of Balance dataset, the discriminant space constructed by LgDA is more class-dependent while that by LDA inherits the topology in the input feature space.

TABLE I. shows the values of the discriminant criteria of the constructed discriminant space. It is noticed that the proposed LgDA achieves higher values than LDA. This means that the discriminant space constructed by LgDA is better than that constructed by LDA in terms of the discriminant criterion which is the objective function of the discriminant analysis. Especially, the improvement of the discriminant criterion is large in the case of Balance dataset.

TABLE II. shows the recognition rates of Satimage and Balance datasets obtained by using LDA, MLR and LgDA. They are calculated by using k nearest neighbor (k-NN) classifier in the discriminant space for the test samples. In TABLE II., LgDA (λ_W = 0) denotes LgDA by MLR without regularization. It is noticed that the recognition rates of Satimage by LgDA are higher than that by LDA. Especially, LgDA by MLR without regularization gave higher recognition

rate than MLR and it gave the best recognition rate for Satimage. The recognition rate by LgDA with regularization was slightly lower than that by the regularized MLR. The reason of this is probably because the regularization parameter of LgDA is not tuned and is set to the same value with the regularized MLR. The recognition rate of LgDA with regularization is probably improved by tuning the regularization parameter for LgDA classifier. Also the recognition rate by LgDA with regularization was slaighlty lower than that by LgDA with regularization. The reason is also similar. The recognition rate of LgDA with regularization is probably improved by tuning the reguralization parameter for LgDA classifier. In the case of Balance dataset, the recognition rate by LDA was the lowest and that by other methods was equal to 92.15%. These results suggest that LgDA can achieve higher recognition rates than LDA even if the sample shows the structured distribution in input feature space.

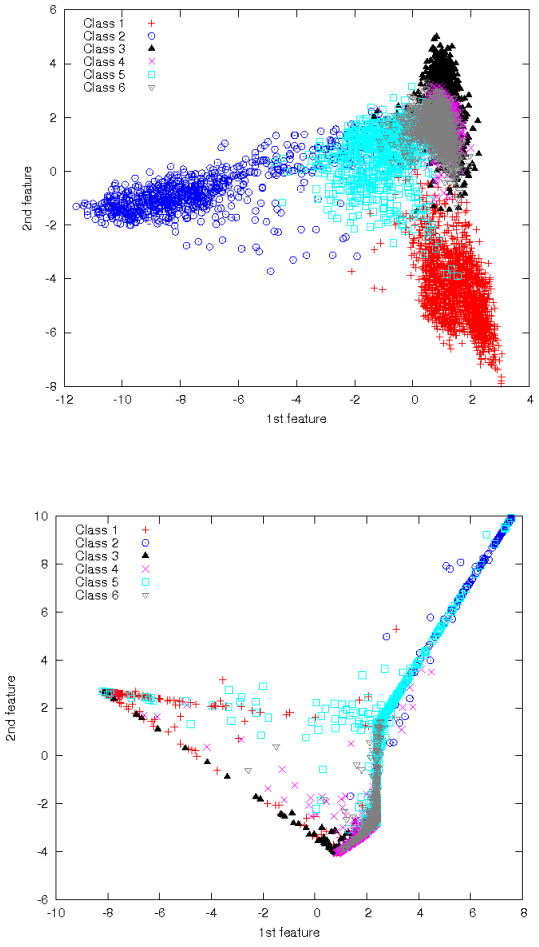
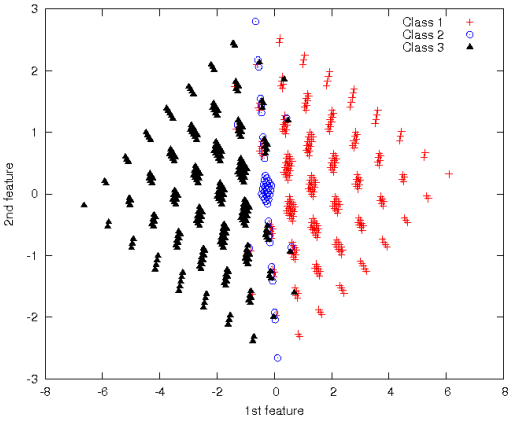


Figure 1. Discriminant spaces by LDA (above) and b LgDA (below) for Satimage



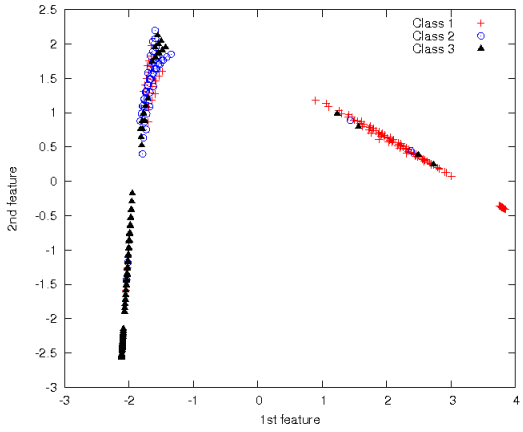


Figure 2. Discriminant spaces by LDA (above) and by LgDA (below) for Balance

TABLE I. DISCRIMINANT CRITERIONS

	Satimage	Balance
LDA	0.4900	0.3333
LgDA ($\lambda_W = 0$)	**0.7541**	**0.6783**
LgDA	0.7462	0.6773

TABLE II. RECOGNITION RATES FOR THE TEST DATASETS

	Satimage	Balance
LDA	0.8395	0.8150
MLR	0.8375	**0.9215**
Regularized MLR	0.8435	**0.9215**
LgDA ($\lambda_W = 0$)	**0.8455**	**0.9215**
LgDA	0.8400	**0.9215**

V. DISCUSSION

This paper proposes a novel nonlinear discriminant analysis named logistic discriminant analysis (LgDA). The Bayesian *posterior* probabilities are estimated by the multi-nominal logistic regression (MLR) and the non-linear discriminant mapping is constructed based on Otsu's theory of the optimal non-linear discriminant analysis (ONDA). MLR is known as one of the generalized linear model which is a flexible generalization of ordinary least squares regression. By modifying the outputs of the linear predictor by the link function, MLR can naturally estimate the Bayesian *posterior* probabilities. Since linear discriminant analysis (LDA) can be regarded as the linear approximation of the ONDA through the linear approximations of the Bayesian *posterior* probabilities, the proposed LgDA can be regarded as the natural extension of LDA substituting the generalized linear model for the linear model in LDA.

The experimental results show that the discriminant space constructed by LgDA is better than the one obtained by LDA. Especially in the case of Balance dataset, the discriminant space constructed by LDA inherits the topology in the input feature space but LgDA constructs more class dependent discriminant space. Also the recognition rates obtained by k nearest neighbor classifier in the discriminant space constructed by LgDA are better than LDA as shown in TABLE I. These results suggest that LgDA can construct the better discriminant feature space than LDA. Also these results show the importance of the link function in MLR to estimate the *posterior* probabilities.

Since the MLR can be extend to non-linear by using kernel trick, it is natural to extend the LgDA to the kernel logistic discriminant analysis by using kernel MLR as the estimator of the *posterior* probabilities. For future works, we would like to investigate the discriminant spaces constructed by the kernel LgDA.

REFERENCES

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

[2] N. Otsu, "Mathmatical Studies on Feature Extruction in Pattern Recognition," Researches of the Electrotechnical Laboratory, No. 818, 1981 (in Japanese)

[3] N. Otsu, "Optimal linear and nonlinear solutions for least-square discriminant feature extraction," Proceedings of the 6th International Conference on Pattern Recognitoin, pp557-560, 1982

[4] D.W. Ruck, S.K. Rogers, M. Kabrisky, M.E. Oxley and B.W. Suter "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," IEEE Transactions on Neural Networks, Vol. 1, pp.296-298, 1990.

[5] T. Kurita, H. Asoh and N. Otsu, "Nonlinear discriminant features constructed by using outputs of multilayer perceptron," Proceeding of the International Symposium on Speech, Image Processing and Neural Networks (ISSIPNN' 94), vol. 2, pp.417-420, 1994

[6] S.Mika, G.Ratsch, J.Weston, B.Scholkopf, A.Smola, and K.Muller, "Fisher discriminant analysis with kernels," Proc. IEEE Neural Networks for Signal Processing Workshop, pp.41-48, 1999.

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

[8] K. Fukunaga and S. Ando, "The optimum nonlinear features for a scatter criterion in discriminant analysis," IEEE Transactions on Information Theory, Vol. 23, pp.453- 459, 1977

[9] McCullagh, Peter; Nelder, John, Generalized Linear Models. London: Chapman and Hall, 1989.

[10] K. Watanabe and T. Kurita, "Locality preserving multi-nominal logistic regression," Proceeding of the 19th International Conference on Pattern Recognition, pp.1-4, December 2008

[11] A. Asuncion and D.J. Newman. "UCI Machine Learning Repository," http://www.ics.uci.edu/~mlearn/MLRepository.html., 2007