A New Approximation Method of the Quadratic Discriminant Function

Shin'ichiro Omachi[1], Fang Sun[2], and Hirotomo Aso[1]

[1] Graduate School of Engineering, Tohoku University
Aoba 05, Aramaki, Aoba-ku, Sendai-shi, 980-8579 Japan
{machi, aso}@ecei.tohoku.ac.jp
[2] Faculty of Science and Technology, Tohoku Bunka Gakuen University
6-45-16, Kunimi, Aoba-ku, Sendai-shi, 981-8551 Japan
fan@ait.tbgu.ac.jp

Abstract. For many statistical pattern recognition methods, distributions of sample vectors are assumed to be normal, and the quadratic discriminant function derived from the probability density function of multivariate normal distribution is used for classification. However, the computational cost is $O(n^2)$ for n-dimensional vectors. Moreover, if there are not enough training sample patterns, covariance matrix can not be estimated accurately. In the case that the dimensionality is large, these disadvantages markedly reduce classification performance. In order to avoid these problems, in this paper, a new approximation method of the quadratic discriminant function is proposed. This approximation is done by replacing the values of small eigenvalues by a constant which is estimated by the maximum likelihood estimation. This approximation not only reduces the computational cost but also improves the classification accuracy.

1 Introduction

In conventional statistical pattern recognition methods, features are extracted from objects. The features are expressed in a form of feature vectors, and the probability density function of distribution of feature vectors is estimated for each category. An unknown input pattern is assigned to the category with the maximum probability [1, 2]. The estimation methods of the probability density function are classified into two types: parametric estimation and nonparametric estimation.

In parametric density estimation, the forms for the density function is assumed to be known, and parameters of the function are estimated using the training sample vectors. The multivariate normal distribution is usually used as the density function. It is because the multivariate normal distribution is easy to handle and in many cases the distribution of sample vectors can be regarded as normal if there are enough samples. Mean vector and covariance matrix are calculated from the vectors. However, if there are not enough training sample vectors, covariance matrix cannot be estimated accurately. The estimation errors will increase in eigenvalue expansion, especially for the higher dimensions

[3]. Moreover, the computational cost will reach $O(n^2)$ for n-dimensional vectors. In the case that the dimensionality is large, these disadvantages markedly reduce classification performance.

On the other hand, nonparametric density estimation is used without assuming that the forms for the density function is known. Many researchers have tried to estimate the distribution by nonparametric methods. In many cases, k nearest neighbor (k-NN) [4, 5] or Parzen kernel-type [6, 7] is used. Fukunaga et al. estimated the probability density function by using either k-NN or Parzen procedures, and discussed the estimation method of Bayes error [8]. Furthermore, for dimensional reduction, Buturović used the k-NN estimate of the Bayes error in transformed low-dimensional space as an optimization criterion for constructing the linear transformation [9]. Since these methods estimate arbitrary probability density functions which are not normal distributions, the computation time is significant and it is difficult to find optimal parameters.

In this paper, we focus on the parametric density estimation using probability density function of multivariate normal distribution. In order to avoid the disadvantages mentioned above, a new approximation method of the quadratic discriminant function is proposed. This approximation is done by replacing the values of small eigenvalues by a constant which is estimated by the maximum likelihood estimation. By applying this approximation, a new discriminant function, called *simplified quadratic discriminant function*, is defined. This function not only reduces the computational cost but also improves the classification accuracy.

2 Approximation of the Quadratic Discriminant Function

First we give a brief review about the quadratic discriminant function, and then propose a new approximation method of the function.

2.1 Quadratic Discriminant Function

Let n be the dimension of feature vector. The well-known probability density function of n-dimensional normal distribution is,

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \tag{1}$$

where $\boldsymbol{x}$ is an n-component vector, $\boldsymbol{\mu}$ is the mean vector, and Σ is the $n \times n$ covariance matrix. The quadratic discriminant function (QDF) is derived from Eq.(1) as follows.

$$g(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \log|\Sigma| \tag{2}$$

$$= \sum_{i=1}^{n} \frac{((\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\phi}_i)^2}{\lambda_i} + \sum_{i=1}^{n} \log \lambda_i, \tag{3}$$

where λ_i is the ith eigenvalue of Σ sorted by descending order, and $\boldsymbol{\phi}_i$ is the eigenvector that corresponds to λ_i. This will be the minimum-error-rate classifier if the distributions are normal, prior probabilities of all categories are equal, and the parameters $\boldsymbol{\mu}$ and Σ are known. However, in general, since the parameters are unknown, the sample mean vector $\hat{\boldsymbol{\mu}}$ and sample covariance matrix $\hat{\Sigma}$ are used.

$$\hat{g}(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \hat{\Sigma}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}) + \log |\hat{\Sigma}| \tag{4}$$

$$= \sum_{i=1}^{n} \frac{((\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \hat{\boldsymbol{\phi}}_i)^2}{\hat{\lambda}_i} + \sum_{i=1}^{n} \log \hat{\lambda}_i. \tag{5}$$

Here, $\hat{\lambda}_i$ is the ith eigenvalue of $\hat{\Sigma}$ and $\hat{\boldsymbol{\phi}}_i$ is the eigenvector. It is known that small eigenvalues in Eq.(5) usually contain many errors that cause the reduction of recognition accuracy [3]. Moreover, the computational cost of Eq.(5) is $O(n^2)$ for n-dimensional vectors. In the case that n is large, it requires enormous computational cost.

2.2 Simplified Quadratic Discriminant Function

To avoid the bad influence caused by small eigenvalues and to reduce the computational cost, one considerable solution is replacing small eigenvalues by a constant. Eq.(5) is approximated by the following function.

$$g_s(\boldsymbol{x}) = \sum_{i=1}^{k} \frac{((\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \hat{\boldsymbol{\phi}}_i)^2}{\hat{\lambda}_i} + \sum_{i=k+1}^{n} \frac{((\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \hat{\boldsymbol{\phi}}_i)^2}{\lambda}$$

$$+ \sum_{i=1}^{k} \log \hat{\lambda}_i + \sum_{i=k+1}^{n} \log \lambda. \tag{6}$$

Here, λ is a constant and $k \leq n$. Eq.(6) is called *simplified quadratic discriminant function*, or *SQDF*. In the case of $k = n$, SQDF is the same as QDF.

The value of λ is determined by the maximum likelihood estimation. For simplicity, the first and third terms of Eq.(6) are fixed, and the second and fourth terms are considered. In other words, the maximum likelihood estimation is performed in the $(n-k)$-dimensional subspace determined by $\{\hat{\boldsymbol{\phi}}_{k+1}, ..., \hat{\boldsymbol{\phi}}_n\}$. Replacing small eigenvalues with λ means that the variance on each axis in this subspace is assumed to be λ. We define

$$\boldsymbol{y} = (y_{k+1}, ..., y_n), \tag{7}$$

where

$$y_i = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \hat{\boldsymbol{\phi}}_i. \tag{8}$$

Since the variance of y_i is assumed to be λ, the probability density function of $\boldsymbol{y}$ is,

$$p(\boldsymbol{y}) = \frac{1}{(2\pi\lambda)^{(n-k)/2}} \exp\left\{ -\frac{1}{2\lambda} \sum_{i=k+1}^{n} y_i^2 \right\}. \tag{9}$$

Note that y_i and $\boldsymbol{y}$ are random variables. Let m be the number of samples and y_{ij} $(1 \leq j \leq m)$ be the jth observation value of y_i. Likelihood of λ is

$$L = \frac{1}{(2\pi\lambda)^{(n-k)m/2}} \exp\left\{ -\frac{1}{2\lambda} \sum_{i=k+1}^{n} \sum_{j=1}^{m} y_{ij}^2 \right\}. \tag{10}$$

Solving the equation

$$\frac{\partial}{\partial \lambda} \log L = 0,$$

we get

$$\lambda = \frac{1}{n-k} \sum_{i=k+1}^{n} \frac{1}{m} \sum_{j=1}^{m} y_{ij}^2 \tag{11}$$

$$= \frac{1}{n-k} \sum_{i=k+1}^{n} \hat{\lambda}_i. \tag{12}$$

In other words, λ is the mean value of $\hat{\lambda}_i$ $(i = k+1, ..., n)$. Since $\mathrm{tr}\hat{\Sigma} = \sum_{i=1}^{n} \hat{\lambda}_i$ and $\|\boldsymbol{x} - \hat{\boldsymbol{\mu}}\| = \sum_{i=1}^{n} ((\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \hat{\boldsymbol{\phi}}_i)^2$, Eq.(6) can be rewritten as,

$$g_s(\boldsymbol{x}) = \sum_{i=1}^{k} \frac{(\lambda - \hat{\lambda}_i)((\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \hat{\boldsymbol{\phi}}_i)^2}{\lambda \hat{\lambda}_i} + \frac{\|\boldsymbol{x} - \hat{\boldsymbol{\mu}}\|^2}{\lambda} + \sum_{i=1}^{k} \log \hat{\lambda}_i + (n-k) \log \lambda, \tag{13}$$

where

$$\lambda = \frac{\mathrm{tr}\hat{\Sigma} - \sum_{i=1}^{k} \hat{\lambda}_i}{n-k}, \tag{14}$$

which can be calculated with k eigenvectors and k eigenvalues. Comparing with Eq.(5), the computational cost of Eq.(13) is reduced from $O(n^2)$ to $O(nk)$.

Next, we investigate the form for the density function of Eq.(3) and Eq.(6). The first term of Eq.(3), and the first and second terms of Eq.(6) are only considered, because the other terms are just the normalizing terms. Let e_1, e_2, e_3 be the expected values of the first term of Eq.(3), the first term of Eq.(6), and the second term of Eq.(6), respectively. For simplicity, the case that there are enough samples is considered. Since $(\boldsymbol{x} - \boldsymbol{\mu})^t \boldsymbol{\phi}_i / \sqrt{\lambda_i}$ follows normal distribution $N(0, 1)$, the first term of Eq.(3) will follow χ^2 distribution with n degrees of freedom. Then,

$$e_1 = n.$$

In the case that there are enough samples, the first term of Eq.(6) will follow χ^2 distribution with k degrees of freedom,

$$e_2 = k.$$

Since $\hat{\lambda}_i$ represents the variance of the component projection onto the vector $\hat{\boldsymbol{\phi}}_i$, and the expected value of $((\boldsymbol{x} - \hat{\boldsymbol{\mu}})^t \hat{\boldsymbol{\phi}}_i)^2$ is $\hat{\lambda}_i$. Then the expected value of the second term of Eq.(6) is

$$e_3 = \sum_{i=k+1}^{n} \frac{\hat{\lambda}_i}{\lambda}.$$

Substituting Eq.(12),

$$e_3 = n - k,$$

is obtained. Therefore we get

$$e_1 = e_2 + e_3,$$

namely, the expectation values of the two expressions are equal. This means Eq.(6) gives a good approximation of Eq.(3).

As related approaches, quasi-Mahalanobis distance (QMD) [10] and modified Mahalanobis distance (MMD) [11] have been proposed. The QMD neglects the third and fourth terms of Eq.(6) and replaces λ by $\hat{\lambda}_{k+1}$. The MMD only uses the first term of Eq.(6), and instead of $\hat{\lambda}_i$, $\hat{\lambda}_i + b$ is employed, where b is a bias determined experimentally. The modified quadratic discriminant function (MQDF) [12] is derived from the Bayesian estimation of the covariance matrix, and $\hat{\lambda}_i$ of the first and third terms of Eq.(6) is replaced by $\hat{\lambda}_i + \lambda$. The value of λ is determined experimentally. All of these methods have been proposed to improve recognition accuracy but not to approximate the quadratic discriminant function. SQDF is an approximation of the quadratic discriminant function, and can describe the form of the distribution. Moreover, since SQDF is derived from the maximum likelihood estimation, it is not only appropriate as a classifier, but it also can be used for model complexity identification with information criterion such as Akaike's Information Criterion (AIC) [13] or Minimum Description Length (MDL) [14].

2.3 Model Identification

In SQDF, the only parameter that is not determined is k, that is, the number of reliable eigenvalues. The other parameters are calculated automatically with samples. Of course, k can be chosen arbitrarily or experimentally. In recognition systems which handle large number of categories with high dimensional vectors, small value of k should be chosen in order to limit the computational cost.

However, if we attach greater importance to the form of distribution, the value of k can be determined by information criterion. This is a kind of model identification. Let the numbers of parameters of mean vector, eigenvalues and eigenvectors be N_1, N_2 and N_3, respectively. Since mean vector is n-dimensional, $N_1 = n$. Since the number of eigenvalues is $k + 1$ $(k < n)$ or k $(k = n)$, $N_2 = \min(k + 1, n)$. The number of parameters of eigenvectors is,

$$N_3 = (n-1) + (n-2) + ... + (n-k) = \frac{2kn - k^2 - k}{2}. \tag{15}$$

Total number of parameters of SQDF is

$$N_1 + N_2 + N_3 = \frac{(2n-k)(k+1) + 2\min(k+1, n)}{2}. \tag{16}$$

Let $\boldsymbol{x}_j$ be a sample ($j = 1, 2, ..., m$). The AIC is written as,

$$AIC = 2\sum_{j=1}^{m} g_s(\boldsymbol{x}_j) + (2n-k)(k+1) + 2\min(k+1, n), \tag{17}$$

while the MDL is written as,

$$MDL = \sum_{j=1}^{m} g_s(\boldsymbol{x}_j) + \frac{(2n-k)(k+1) + 2\min(k+1, n)}{4} \log m. \tag{18}$$

The value of k is determined to minimize the criterion AIC or MDL.

3 Experiments

In order to confirm the effectiveness of SQDF, three types of experiments are carried out.

3.1 Effectiveness as a Classifier

The first experiment is done to confirm the effectiveness of SQDF as a classifier. Character recognition is performed using character images included in the NIST Special Database 19 [15]. The database includes over 800000 handprinted digit and alphabetic character images. Digit character images of '0' and '1' are used in the experiment. The numbers of samples of '0' and '1' are both 40000. As the feature vector, the improved directional element feature [16] is used. This feature is 196-dimensional vector.

For each category, m images out of the first 10000 images are used as training sample data, and the next 10000 images are used for evaluation. From the training sample data, feature vectors are extracted, and mean vectors and covariance matrices are calculated. Then SQDF and QDF are used as discriminant functions. The results are shown in Fig.1. Fig.1(a) shows error rates of various dimensionality k of SQDF. The number of training samples is fixed to $m = 10000$. Here, the case of $k = 196$ of SQDF equals to QDF. From the figure, the error rate of SQDF in the case of $k = 30$ is much smaller than the case of $k = 196$, that is the result of QDF. Fig.1(b) shows error rates of various number of training samples. The dimensionality is fixed to $k = 30$. These results show that SQDF is much more effective than QDF if the number of samples is small. In the case of $m < 2000$, the error rates of QDF becomes extremely large, however, the error rate of SQDF changes little.

All of these results clarifies that SQDF not only reduces the computational time but also improves classification accuracy. It is especially effective in the case of small number of training samples.

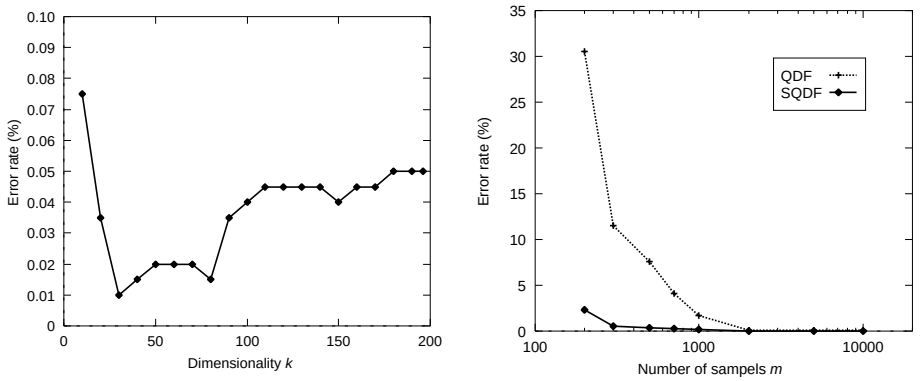

Fig. 1. Results of character recognition. (a) Error rates of various dimensionality. $m = 10000$. (b) Error rates of various number of samples. $k = 30$.

3.2 Validity of Approximation

Next, in order to confirm the validity of approximation, experiment using artificial data is carried out. Since Eq.(6) is supposed to approximate Eq.(3), it is required that the difference between Eq.(3) and Eq.(6) should be small.

Suppose $\boldsymbol{\mu}_0$ is an n-dimensional vector, and Σ_0 is an appropriate $n \times n$ covariance matrix. Here, we use the mean vector and the covariance matrix which are calculated with 10000 character images of '0' in Section 3.1. Since the improved directional element feature is adopted, $n = 196$. By producing random numbers, m training vectors that follow n-dimensional normal distribution $N(\boldsymbol{\mu}_0, \Sigma_0)$ are produced. The sample covariance matrix $\hat{\Sigma}$ and sample mean vector $\hat{\boldsymbol{\mu}}$ are calculated with m training vectors. Other 10000 vectors that follow $N(\boldsymbol{\mu}_0, \Sigma_0)$ are randomly obtained to be evaluation vectors. The value of Eq.(6) (SQDF) of each evaluation vector is computed with $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$. Suppose the value g_{true} obtained by Eq.(3) with Σ_0 and $\boldsymbol{\mu}_0$ is the true value of QDF. The error e is given as $e = |(g_s - g_{true})/g_{true}|$. The average of e of evaluation vectors is calculated. The error of QDF is calculated in the same manner.

Fig.2 shows the errors of SQDF and QDF computed with various m. Note that QDF can be calculated only if $m \geq n$, however, SQDF can be calculated even in the case of $m < n$ if $k \leq m$. In all cases, the larger m, the more similar the estimated value to the true value. In the case that the number of training samples is small, the error of SQDF is smaller than that of QDF. In the case of $m = 1000$, that means the number of samples is about five times larger than the dimensionality, the error of SQDF becomes slightly larger than that of QDF. However, the big difference of computational time between SQDF and QDF still offers the attraction of SQDF.

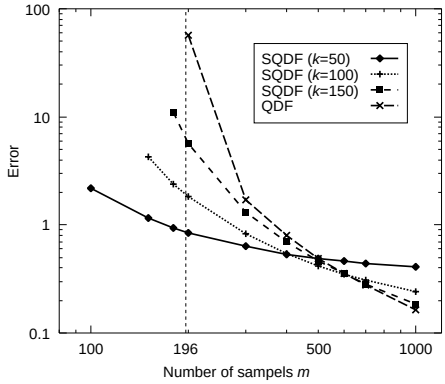

Fig. 2. Errors of each discriminant function.

3.3 Validity of Model Identification Method

The third experiment is carried out to confirm the validity of model identification method described in Section 2.3. Suppose Σ_1 is an appropriate $n \times n$ covariance matrix, and $\boldsymbol{\mu}_1$ is an n-dimensional mean vector. Here, we consider the following diagonal matrix and vector.

$$\Sigma_1 = \mathrm{diag}(\underbrace{1, 1, 1, ..., 1}_{n/2}, \underbrace{2, 3, 4, ..., n/2+1}_{n/2}), \tag{19}$$

$$\boldsymbol{\mu}_1 = (0, 0, ..., 0)^t. \tag{20}$$

Σ_1 consists of $n/2$ components those values are 1 and $n/2$ components those values are larger than 1. Because Σ_1 is a diagonal matrix, each diagonal component corresponds to eigenvalue. In this section, the dimensionality is $n = 16$.

m training vectors that follow n-dimensional normal distribution $N(\boldsymbol{\mu}_1, \Sigma_1)$ are produced in the same manner as described in Section 3.2. The sample covariance matrix $\hat{\Sigma}$ and sample mean vector $\hat{\boldsymbol{\mu}}$ are calculated with the training vectors. Then the value of k is determined by AIC or MDL as described in Section 2.3. In this case, the number of small eigenvalues that are regarded as constant is $n/2$. Since SQDF regards k small eigenvalues as constant, the value of k is expected to be determined that $k = n/2$.

Figs.3(a) and (b) show the values of *AIC* (Eq.(17)) and *MDL* (Eq.(18)) in the case of $m = 10000$, respectively. Both kinds of criterion become small if $k \geq 8(= n/2)$. MDL becomes smallest when the value of k is $8(= n/2)$. while AIC becomes smallest when the value of k is 10. Fig.3(c) shows the relationship between the number of samples m and the selected value of k. These results show an appropriate value can be selected by MDL.

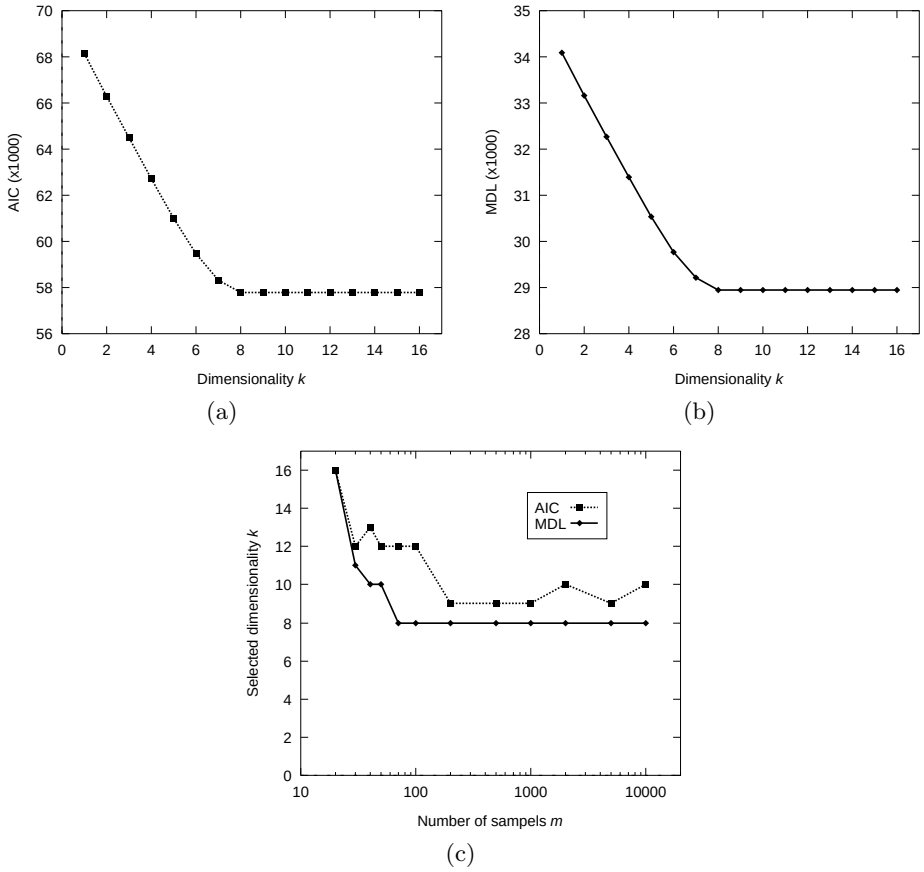

Fig. 3. Model identification by AIC and MDL. (a) The values of AIC in the case of $m = 10000$. (b) The values of MDL in the case of $m = 10000$. (c) Selected dimensionality by AIC and MDL.

4 Conclusions

In this paper, we have focused on the parametric density estimation using probability density function of multivariate normal distribution. In order to avoid the disadvantages of the quadratic discriminant function, we have proposed a new approximation method of the quadratic discriminant function. This approximation is done by replacing the values of small eigenvalues by a constant which is estimated by the maximum likelihood estimation. By applying this approximation, a new discriminant function, simplified quadratic discriminant function, or SQDF, has been defined. This function not only reduces the computational cost but also improves the classification accuracy.

In order to clarify the effectiveness of SQDF, three types of experiments have been carried out. Experimental results of classification using character images

have clarified that SQDF not only reduces the computational time but also improves classification accuracy. The second experiment has displayed that SQDF gives a good approximation of QDF. The third experimental results have shown that the parameter of SQDF can be determined by information criterion.

Applying SQDF to various pattern recognition problems is a future work.

References

1. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley & Sons, New York London Sydney Toronto (1973)
2. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Second Edition, Academic Press, San Diego (1990)
3. Takeshita, T., Kimura, F., Miyake, Y.: On the Estimation Error of Mahalanobis Distance. Trans. IEICE **J70-D** (1987) 567–573
4. Fix, E., Hodges, J.L.: Discriminatory analysis, nonparametric discrimination, consistency properties. Report No.4, School of Aviation Medicine, Randolph Field, Texas (1951)
5. Cover, T.M., Hart, P.E.: Nearest Neighbor Pattern Classification. IEEE Trans. Inform. Theory **IT-13** (1967) 21–27
6. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Ann. Math. Stat. **27** (1956) 832–837
7. Parzen, E.: On the estimation of a probability density function and the mode. Ann. Math. Stat. **33** (1962) 1065–1076
8. Fukunaga, K., Hummels, D.M.: Bayes Error Estimation Using Parzen and k-NN Procedures. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-9** (1987) 634–643
9. Buturović, L.J.: Toward Bayes-Optimal Linear Dimension Reduction. IEEE Trans. Pattern Anal. Mach. Intell. **16** (1994) 420–424
10. Kurita, M., Tsuruoka, S., Yokoi, S., Miyake, Y.: Handprinted "Kanji" and "Hiragana" character recognition using weighting direction index histograms and quasi-Mahalanobis distance. IEICE Technical Report **PRL82-79** (1983) 105–112
11. Kato, N., Abe, M., Nemoto, Y.: A Handwritten Character Recognition System Using Modified Mahalanobis Distance. Trans. IEICE **J79-D-II** (1996) 45–52
12. Kimura, F., Shridhar, M.: Handwritten numerical recognition based on multiple algorithms. Pattern Recognition **24** (1991) 969–983
13. Akaike, H.: A New Look at the Statistical Model Identification. IEEE Trans. Automatic Control **19** (1974) 716—723
14. Rissanen, J.: Stochastic Complexity and Modeling. The Annals of Statistics **14** (1986) 1080–1100
15. Grother, P.J.: NIST Special Database 19 Handprinted Forms and Characters Database. National Institute of Standards and Technology. (1995)
16. Kato, N., Suzuki, M., Omachi, S., Aso, H., Nemoto, Y.: A Handwritten Character Recognition System Using Directional Element Feature and Asymmetric Mahalanobis Distance. IEEE Trans. Pattern Anal. Mach. Intell. **21** (1999) 258–262