Pattern Recognition 43 (2010) 3144–3150

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Reduction of processing time for optimal and quadratic discriminant analyses

Michitaka Suzuki *, Akiyoshi Itoh

College of Science and Technology, Nihon University, 7-24-1 Narashinodai, Funabashi 247-8501, Japan

ARTICLE INFO

Article history:
Received 1 April 2009
Received in revised form
3 February 2010
Accepted 20 March 2010

Keywords:
Optimal discriminant analysis
Quadratic discriminant analysis
Fast algorithm
Character recognition

ABSTRACT

A fast algorithm is presented for optimal discriminant analysis and quadratic discriminant analysis. In this algorithm, the discriminant function of an input feature vector for each category is calculated via a monotonically increasing sequence, and when the sequence value exceeds a certain value, then you can assert that the current category cannot be the classification result and omit the redundant calculation of the remaining terms for the category, thus making the calculation faster. Applying this algorithm to the recognition experiment on handwritten characters, we could reduce the processing time to 4% of the conventional simple method. Since both discriminant analyses assume the normal distribution of the features, disnormality contained in real-world data affects the accuracy of the two discriminant analyses. We also compared the accuracy performances of the two discriminant analyses using real-world data and artificial data.

1. Introduction

In discriminant analysis, an input feature vector $\boldsymbol{x}$ is assigned to a category, c, by the condition that its probability $P(c|\boldsymbol{x})$ is maximized, given in advance training data $\{\boldsymbol{x}_i, c_i\}$. The simplest approach is to calculate the probabilities for all the categories and to find their maximum. The time complexity of this approach is $\mathcal{O}(Cd^2)$, where C is the number of the categories and d is the dimensionality of the feature vector. When the number of the categories is large, the dimensionality must be large accordingly, possibly making the processing time enormous. For example, there are more than 3000 categories of characters used in some oriental countries, and the sufficient dimensionality is about 200 or more.[1] In order to evade the bulky calculation, approximations have been used at the expense of accuracy: the modified Bayes classifier, the subspace method, etc. [1]. Additionally, the preclassification for narrowing down the candidate categories is often adopted with the use of linear discriminant analysis [1], or some other less accurate but easily calculable metric [2].

When each element of the feature vector shows the normal distribution, the predictive distribution is the multivariate t-distribution [3]. The t-distribution is well approximated by the normal distribution in some cases. The discriminant analyses based on the t-distribution and the normal distribution are referred to as optimal discriminant analysis (ODA) and quadratic discriminant analysis (QDA), respectively. The present authors developed a fast classification algorithm for QDA without using further approximations, and reduced the processing time to as low as 4% of that by the simple conventional approach [4]. In this algorithm, the discriminant function for each category is calculated via a monotonically increasing sequence and when the sequence value exceeds the minimum of the discriminant function values so far calculated for the input feature vector, then the current category cannot possibly be the classification result and the redundant calculation of the remaining terms for the category can be omitted, thus making the calculation faster. The main objective of this paper is to extend the fast algorithm for QDA to apply to ODA.

Since QDA is an approximation to ODA, QDA is generally considered less accurate than ODA, especially when the numbers of training samples are small or nonuniform across the categories. It is not so simple, however, because both ODA and QDA assume normal distribution for the features, although real-world data often show distributions quite different from the normal distribution [5]. We study the performances of the two discriminant analyses in their accuracy and speed using real-world data and artificial data.

This paper is organized as follows. In Section 2, a unified description of ODA and QDA is given. In Section 3, a fast algorithm for ODA and QDA is described. In Section 4, experimental results on accuracy and speed are given and discussed. In Section 5, conclusions are given.

* Corresponding author.
E-mail address: msuzuki@mxh.mesh.ne.jp (M. Suzuki).

[1] Another example where you need to deal with many categories is to identify many persons by their faces or fingerprints. In simple classification issues with a few categories, however, the algorithm proposed here may not be imperative.

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.03.017

2. Discriminant analyses

We describe ODA and QDA as well as diagonal approximation in a unified manner so that the proposed algorithm can handle them equally.

2.1. Posterior probability

We assume that each element of the features in a category shows the normal distribution. Given n_c samples of training data of category c, $\{\boldsymbol{x}_{ci}|i=1,\ldots,n_c\}$, the resultant posterior probability that an input, $\boldsymbol{x}$, belongs to c can be written as the product of the prior probability, P_c, and the predictive distribution (t-distribution) [3,6]

$$P(c|\boldsymbol{x}) = P_c \frac{b_c}{|\Sigma_c|^{1/2}} \left[1+\frac{(\boldsymbol{x}-\boldsymbol{m}_c)^{\mathrm{t}}\Sigma_c^{-1}(\boldsymbol{x}-\boldsymbol{m}_c)}{n_c'}\right]^{-(n_c'+1)/2}, \tag{1a}$$

$$b_c = \frac{\Gamma\left(\frac{n_c'+1}{2}\right)}{(n_c'\pi)^{d/2}\Gamma\left(\frac{n_c'-d+1}{2}\right)}, \tag{1b}$$

where $\boldsymbol{m}_c$ is the mean vector: $\boldsymbol{m}_c = n_c^{-1}\sum_{i=1}^{n_c}\boldsymbol{x}_i$, $\Gamma(\cdot)$ the gamma function and Σ_c the estimate of the covariance matrix. According to the hierarchical Bayes method [7], we can put Σ_c in the following form:

$$\Sigma_c = (n_c S_c + v_1 \operatorname{diag} S + v_2 S)/n_c', \tag{2}$$

$$n_c' = n_c + v_1 + v_2, \tag{3}$$

where S_c and S are the sample covariance matrix and the pooled covariance matrix, respectively,

$$S_c = \frac{1}{n_c}\sum_{i=1}^{n_c}(\boldsymbol{x}_i-\boldsymbol{m}_c)(\boldsymbol{x}_i-\boldsymbol{m}_c)^{\mathrm{t}}, \tag{4}$$

$$S = \frac{1}{n}\sum_c n_c S_c, \tag{5}$$

where n is the total number of the training samples: $n=\sum_{c=1}^{C} n_c$. The hyperparameters v_1 and v_2 are positive and n_c' is considered as an effective number of training samples for Σ_c. Similar expressions to (2) had been proposed as regularized discriminant analysis (RDA) [8,9]. The prior probability is taken to be

$$P_c = n_c/n. \tag{6}$$

By diagonalizing Σ_c, we can rewrite (1a) with the eigenvalues e_{ci} and the eigenvectors $\boldsymbol{v}_{ci}$ of Σ_c:

$$P(c|\boldsymbol{x}) = P_c \frac{b_c}{|\Sigma_c|^{1/2}} \left(1+\frac{1}{n_c'}\sum_{i=1}^{d}\frac{[\boldsymbol{v}_{ci}^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{m}_c)]^2}{e_{ci}}\right)^{-(n_c'+1)/2} \tag{7}$$

for ODA.

Taking the limit of $n_c\to\infty$, we have the probability with the form of the normal distribution:

$$P(c|\boldsymbol{x}) = P_c \frac{1}{(2\pi|\Sigma_c|)^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{[\boldsymbol{v}_{ci}^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{m}_c)]^2}{e_{ci}}\right) \tag{8}$$

for QDA.

Furthermore, if we neglect the off-diagonal elements of Σ_c, we have

$$P(c|\boldsymbol{x}) = P_c \prod_{i=1}^{d}\frac{1}{(2\pi)^{1/2}\sigma_{ci}} \exp\left[-\frac{1}{2}\left(\frac{x_i-m_{ci}}{\sigma_{ci}}\right)^2\right] \tag{9}$$

for what we refer to as diagonal approximation, where $\sigma_{ci}^2=(\Sigma_c)_{ii}$ since $e_{ci}=\sigma_{ci}^2$ and $\boldsymbol{v}_{ci}$ becomes the unit vector directed along the ith axis of the original coordinate system.

2.2. Discriminant function

The recognition result, $\tilde{c}(\boldsymbol{x})$, is determined so that its probability should be maximized:

$$\tilde{c}(\boldsymbol{x}) = \arg\max_c P(c|\boldsymbol{x}). \tag{10}$$

It is usually convenient to define a discriminant function instead of dealing with the probability directly. We define the same form of the discriminant function for the above probabilities as

$$f(\boldsymbol{x},c) = -2\ln P(c|\boldsymbol{x}), \tag{11}$$

so that

$$\tilde{c}(\boldsymbol{x}) = \arg\min_c f(\boldsymbol{x},c). \tag{12}$$

Furthermore, we calculate the discriminant functions for all the discriminant analyses in the same form

$$f(\boldsymbol{x},c) = \phi(g(\boldsymbol{x},c)), \tag{13}$$

$$g(\boldsymbol{x},c) = a_c + \sum_{i=1}^{d} a_{ci}[\boldsymbol{v}_{ci}(\boldsymbol{x}_i-\boldsymbol{m}_c)]^2, \tag{14}$$

where a_{ci}'s are in ascending order, i.e., the eigenvalues e_{ci}'s are in descending order. We give specific expressions for $\phi(\cdot)$ and a_{ci}'s and a_c of each discriminant analysis in the following.

For ODA, we have

$$\phi_c(y) = (n_c'+1)\ln y, \quad a_{ci} = \frac{a_c}{n_c' e_{ci}}, \quad a_c = \left(\frac{P_c b_c}{|\Sigma_c|^{1/2}}\right)^{-2/(n_c'+1)}. \tag{15}$$

For QDA, we have

$$\phi_c(y) = y, \quad a_{ci} = 1/e_{ci}, \quad a_c = -2\ln P_c + \ln|\Sigma_c|. \tag{16}$$

For diagonal approximation, we have

$$\phi_c(y) = y, \quad a_{ci} = 1/\sigma_{ci}^2, \quad a_c = -2\ln P_c + \ln\prod_{i=1}^{d}\sigma_{ci}. \tag{17}$$

In this case the discriminant function becomes the weighted Euclidean distance corrected by the logarithm term.

3. Reduction of the processing time

It is a matter of course to avoid redundant calculations to attain high-speed performance. You can actually spot the redundant calculations while calculating the discriminant function via a monotonically increasing sequence which converges to it. In order to make clear the essence of our fast algorithm, we first present obvious inequality relations in the following propositions.

Proposition 1. *Let $g_k(c)$ be a monotonically increasing sequence with respect to k which converges to $g(c)$. If $g_k(c) > g(\tilde{c})$ for any k, then $g(c) > g(\tilde{c})$.*

Proof. From the monotonicity of the sequence, we have $g(c) \geq g_k(c)$. Add this inequality and the inequality of the hypothesis, and we obtain the conclusion. □

Proposition 2. *Let $\phi_c(y)$ be a monotonically increasing function with respect to y and $f(c)=\phi_c(g(c))$. If $g_k(c) > \phi_c^{-1}(f(\tilde{c}))$ for any k, then $f(c) > f(\tilde{c})$.*

Proof. Define a sequence $f_k(c) \equiv \phi_c(g_k(c))$, which is also a monotonically increasing sequence converging to $f(c)$. From the

hypothesis, we have $f_k(c) > f(\tilde{c})$. Apply Proposition 1 to this inequality, and we obtain the conclusion. □

To use these propositions to find redundant calculations, we need a monotonically increasing sequence.

3.1. Monotonically increasing sequence

In order to obtain a monotonically increasing sequence, we consider an approximation to $g(\boldsymbol{x},c)$. We approximate the $(d-k)$ largest values of $\{a_{ci}\}$ by α_c independent of i. The resulting value, $g_k(\boldsymbol{x},c)$, is expressed as

$$g_k(\boldsymbol{x},c) = a_c + \sum_{i=1}^{k} a_{ci} w_i(\boldsymbol{x},c) + \alpha_c \sum_{i=k+1}^{d} w_i(\boldsymbol{x},c), \tag{18}$$

$$w_i(\boldsymbol{x},c) = [\boldsymbol{v}_{ci}^{\mathrm{t}}(\boldsymbol{x}-\boldsymbol{m}_c)]^2. \tag{19}$$

When we choose $\alpha_c = 0$, the resulting sequence, $h_k(\boldsymbol{x},c)$, is given by

$$h_k(\boldsymbol{x},c) = a_c + \sum_{i=1}^{k} a_{ci} w_i(\boldsymbol{x},c). \tag{20}$$

The sequence $h_k(\boldsymbol{x},c)$ certainly increases monotonically and converges to $g(\boldsymbol{x},c)$ at $k=d$, but it converges slowly. To get a faster sequence, we choose $\alpha_c = a_{c,k+1}$. In other words, we approximate the $(d-k)$ smallest eigenvalues of the covariance matrix by their maximum. Thus, $g_k(\boldsymbol{x},c)$ is expressed as

$$g_k(\boldsymbol{x},c) = h_k(\boldsymbol{x},c) + a_{c,k+1} j_k(\boldsymbol{x},c), \tag{21}$$

$$j_k(\boldsymbol{x},c) = |\boldsymbol{x}-\boldsymbol{m}_c|^2 - \sum_{i=1}^{k} w_i(\boldsymbol{x},c), \tag{22}$$

where the sequence $j_k(\boldsymbol{x},c)$, which represents the contribution from the terms with larger values of $\{a_{ci}\}$, is expressed by the sum of the terms with smaller values of $\{a_{ci}\}$ using the orthonormality of the eigenvectors:

$$\sum_{i=1}^{d} w_i(\boldsymbol{x},c) = |\boldsymbol{x}-\boldsymbol{m}_c|^2.$$

Using (21) and (22), we can calculate $g_k(\boldsymbol{x},c)$ sequentially via $h_k(\boldsymbol{x},c)$ and $j_k(\boldsymbol{x},c)$, which are calculated from $h_{k-1}(\boldsymbol{x},c)$, $j_{k-1}(\boldsymbol{x},c)$, and $w_k(\boldsymbol{x},c)$. This sequence $g_k(\boldsymbol{x},c)$ increases monotonically, converges faster than $h_k(\boldsymbol{x},c)$, and reaches $g(\boldsymbol{x},c)$ at $k=d-1$. The proof of these properties is given in Appendix A.

Though our objective is to get a fast algorithm without approximation, this kind of sequence has been used for an approximation to the discriminant function. If you choose $\alpha_c = 0$ and approximate $g(\boldsymbol{x},c)$ by $h_k(\boldsymbol{x},c)$ with $k < d$, you get principal components analysis. Kimura et al. [1] chose $\alpha_c = a_{cd}$ to get their sequence $g'_k(\boldsymbol{x},c) = h_k(\boldsymbol{x},c) + a_{cd} j_k(\boldsymbol{x},c)$ and approximated $g(\boldsymbol{x},c)$ by $g'_k(\boldsymbol{x},c)$ with $k \simeq d/2$. This has been the state-of-art technique in the field of character recognition. Their sequence decreases monotonically, as can be proved in a similar way. They claim that their approximation is insensitive to the error rates, but their reduction of the processing time is much less than our result in Section 4.

In order to get clear pictures of the above sequences, we take ridge features [10] for character recognition as an example. Ridge features is a set of directional features averaged in the vicinity of each of the sampling centers forming grids. The directional features represent the orientation of the ridge as its components of the four fixed orientations: horizontal, vertical, and two slants. Ridge features is special in that the directions are calculated via the ridge lines extracted form a gray-scale image. In contrast, the boundary lines are often used as the source of directional information about a binary image. The ridge lines, shown in Fig. 1(b), are extracted from a character image in the ETL9G data set [11], shown in Fig. 1(a). The positions of the sampling centers are calculated for each image so that the numbers of strokes in the vicinities of the sampling centers should be as uniform as possible [12]. The sampling centers are shown by the "+" marks in Fig. 1(b). The dimensionality of the ridge features is 196 $(=7\times7\times4)$.

Fig. 2 shows how the sequences $g_k(\boldsymbol{x},c)$, $h_k(\boldsymbol{x},c)$, and $g'_k(\boldsymbol{x},c)$ converge with respect to k. We choose the first sample of category 鳥 in the ETL9G data set as the test sample: $\boldsymbol{x} = \boldsymbol{x}_{鳥1}$, and study the sequences for $c = $ 鳥,雀,烏. These characters have the meanings of bird, sparrow, raven, respectively. Note the last one 烏 is different from the first one 鳥 in that 烏 lacks one horizontal stroke in the upper part. To avoid any misperceptions, we will always write 烏! adding "!" in this paper. The expression g_k(鳥1,c) means $g_k(\boldsymbol{x}_{鳥1},c)$. The thick curves represent g_k(鳥1,c), and the thin curves h_k(鳥1,c), and the dotted curves g'_k(鳥1,c) for reference. In Fig. 2(a), the solid curves represent $c = $ 雀, and the dashed curves $c = $ 鳥. In Fig. 2(b), the solid curves represent $c = $ 烏! instead of $c = $ 雀. We see $g_k(\boldsymbol{x},c)$ and $h_k(\boldsymbol{x},c)$ both increases monotonically, with $g_k(\boldsymbol{x},c)$ faster than $h_k(\boldsymbol{x},c)$, while $g'_k(\boldsymbol{x},c)$ decreases monotonically. In Fig. 2(b), the discriminant function $g(\boldsymbol{x}_{烏!})$ is very close to $g(\boldsymbol{x}_{鳥})$ reflecting their shape resemblance.

3.2. Main procedure

Using the propositions and the increasing monotonicity of the sequence (21), we can frame a fast algorithm as shown in Fig. 3. In its line 2 after inputting a feature $\boldsymbol{x}$, an initial value is set at c as a trial. Then each of the other categories is examined. If a category with a lower discriminant value appears, c and $f_{\min}$ are updated in lines 17 and 18. The GOTO in line 15 prevents redundant calculations according to Proposition 2.

Let us see Fig. 2 again to understand the procedure visually. We consider QDA first for simplicity, where $g(\boldsymbol{x},c)$ is equal to the discriminant function. Suppose that the value of g(鳥1,鳥) is already obtained by either of the thick or thin dashed curve. In Fig. 2(a), while calculating g_k(鳥1,雀) sequentially, it exceeds g(鳥1,鳥) at $k=2$, and then we can assert that g(鳥1,雀) > g(鳥1,鳥), namely, 雀 is not the answer. To get the same assertion, h_k(鳥1,雀) must be calculated until $k=20$. Moreover, using g'_k(鳥1,雀), it must be calculated until $k=d-1=195$, i.e., completing the discriminant function.

In Fig. 2(b) the category 烏! is shown by the solid curves instead of 雀. The dashed curves show the same g_k(鳥1,鳥) as in Fig. 2(a), though it looks different because the scale of the vertical axis is three times larger. Since the discriminant function g(鳥1,烏!) is narrowly over g(鳥1,鳥), the sample is not misclassified as 烏! by a slim margin. Even in this delicate case, calculating g_k(鳥1,烏!) until $k=53$ leads to the assertion that g(鳥1,烏!) > g(鳥1,鳥), i.e., 烏! is not the answer, To get the same assertion, h_k(鳥1,烏!) must be calculated until $k=112$.

As for ODA, this kind of visual explanation would be complicated, but if $g_{\min}$ in line 6 and $f_{\min}$ in line 18 are calculated in accordance with the first equation of (15), the redundant calculations can be avoided in the same way according to Proposition 2.

We have seen so far in details for two categories 雀 and 烏! that we do not necessarily need to calculate the sequence until the end but until k^* which is much smaller than d. Fig. 4 shows the frequency distribution of k^* for all the 3035 categories by QDA.

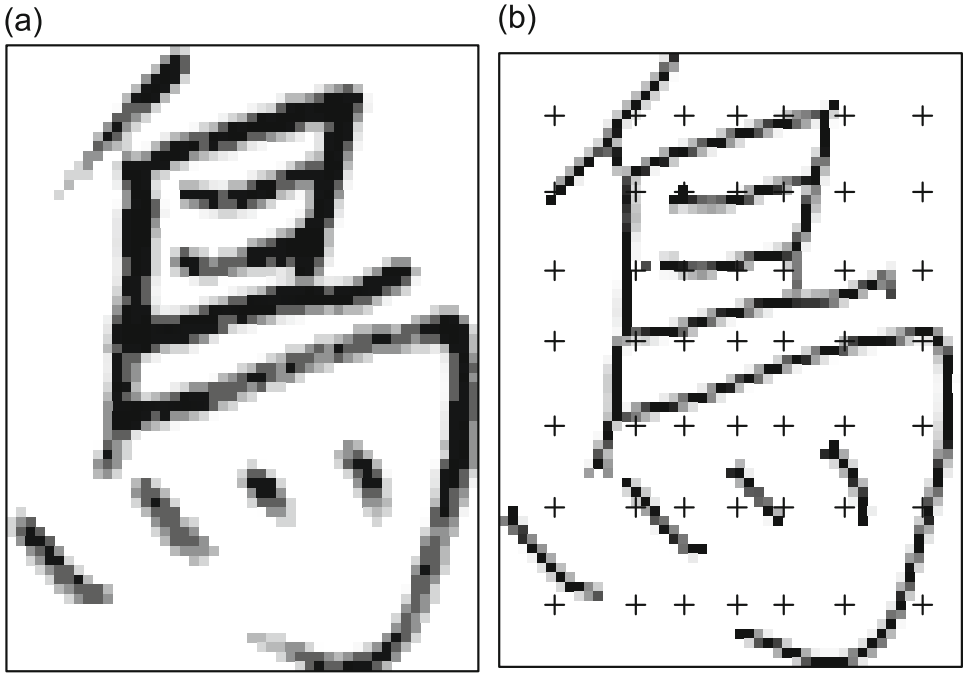

Fig. 1. Sample ETL9G-1954-1: (a) original image and (b) ridge features and sampling centers.

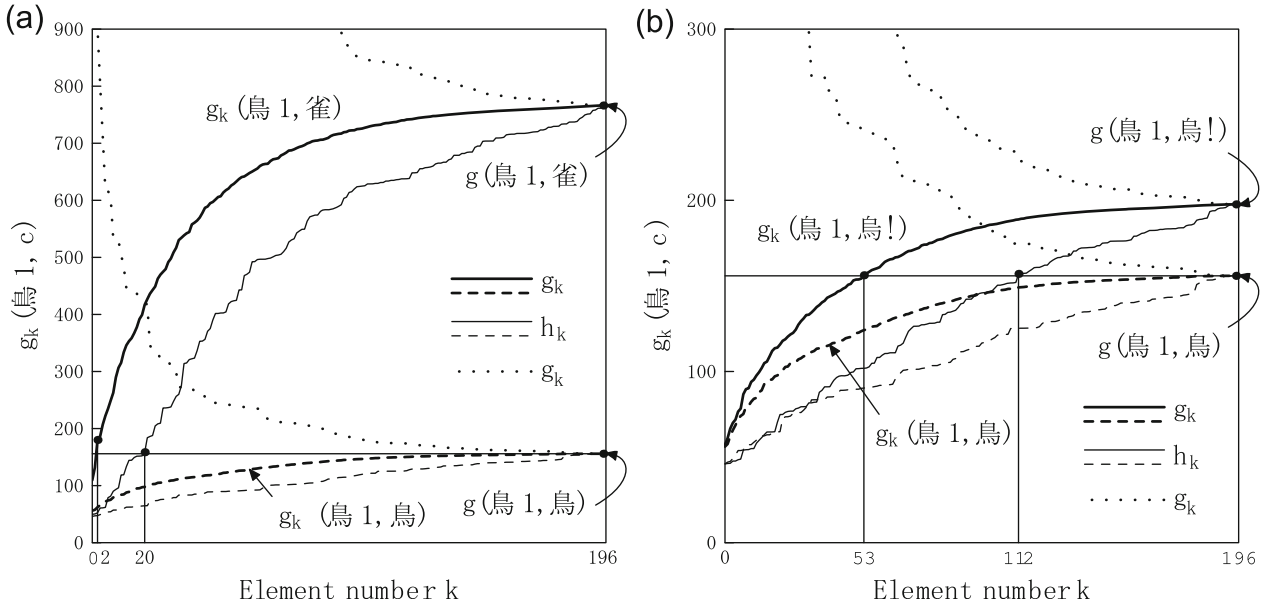

Fig. 2. Sequences converging to discriminant functions. After g(鳥1,鳥) is evaluated, (a) g(鳥1,雀) > g(鳥1,鳥) is asserted when g_2(鳥1,雀) is evaluated (b) and g(鳥1,鳥) > g(鳥1,鳥) is asserted when g_{53}(鳥1,鳥) is evaluated: (a) c=鳥 vs c=雀 and (b) c=鳥 vs c=烏!.

The large points connected with the solid lines represent the frequency of k^* by g_k(鳥1,c), while the smaller points connected with the dashed lines by h_k(鳥1,c). The distribution of k^* by g_k(鳥1,c) resembles the Poisson distribution with its mean $\overline{k}^* = 2.1$. The frequencies at larger k^* are mostly zero, with some occasional occurrences of similar characters annotated by the arrows. For the distribution of k^* by h_k(鳥1,c), $\overline{k}^* = 18.8$. The time complexity of this algorithm for measuring recognition rates is $\mathcal{O}(Cd\overline{k}^*)$, not $\mathcal{O}(Cd^2)$. As we see from Fig. 4, $\overline{k}^*$ is much smaller than d, but what determines $\overline{k}^*$ is not clear at present.

3.3. Setting initial category

In the explanation in the previous subsection, we assumed that g(鳥1,鳥) is evaluated in advance. If the order of examining categories is 雀→鳥, however, both of the full discriminant functions must be evaluated. Thus, the processing time depends on the order of examining categories. The worst case is when the categories are in order of decreasing discriminant value, and then the discriminant functions must be calculated for all the categories. Therefore, to avoid the worst case the initial category should be a category of a lower discriminant value. In the

c: category number
$\boldsymbol{m}_c$: mean vector
a_{ci}: coefficients
$\boldsymbol{v}_{ci}$: eigenvector of the covariance matrix
d: dimensionality of the feature vector

1. input $\boldsymbol{x}$
2. c = initial category
3. $f_{\min} = f(\boldsymbol{x}, c)$
4. c' = next category
5. IF c' = "end," GOTO 20.
6. $g_{\min} = \phi_{c'}^{-1}(f_{\min})$
7. $h = a_{c'}$
8. $j = |\boldsymbol{x} - \boldsymbol{m}_{c'}|^2$
9. $g = h + a_{c'k}j$
10. $i = 1$
11. $w = [\boldsymbol{v}_{c'i}^{t}(\boldsymbol{x} - \boldsymbol{m}_{c'})]^2$
12. $h = h + a_{c'i}w$
13. $j = j - w$
14. $g = h + a_{c',i+1}j$
15. IF $g > g_{\min}$, GOTO 4.
16. IF $i < d$, THEN $i = i + 1$ and GOTO 11.
17. $c = c'$
18. $f_{\min} = \phi_c(g)$
19. GOTO 4.
20. output c

Fig. 3. Fast classification algorithm.

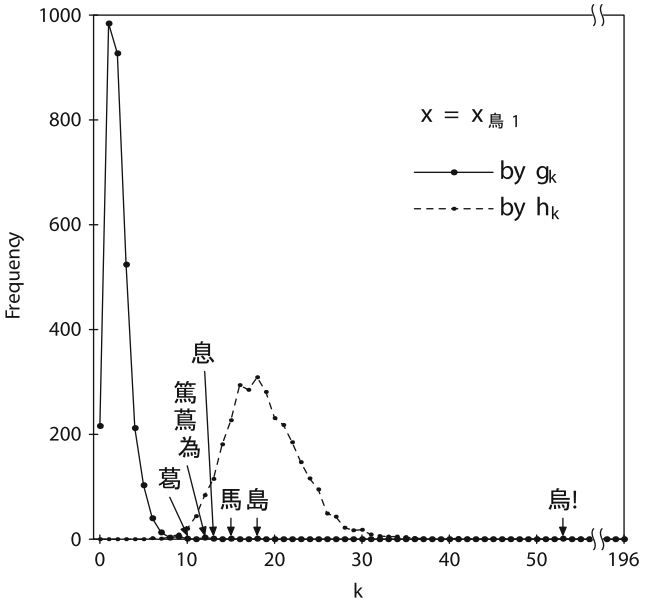

Fig. 4. Frequency distribution of k^*. The calculation up to $g_{10}(x, c)$ is sufficient for most candidate categories dissimilar to 鳥 in the classification of $x = x_{鳥1}$. Values of k^* for similar categories are annotated by the arrows.

experiments measuring recognition rates in a research environment, the experimentalist knows the correct answer. Then the answer should be set at the initial category. Moreover, if an update on line 17 happens, then the accuracy-measuring program can move on to the next input, because it is then clear that a misclassification happened.

In the actual application of a classifier, where the answer is unknown, the initial category should be set in a reasonable way to reduce the number of the calculations of the full discriminant functions. The easiest way may be by the prior probability (6), but it has effect only when one category dominantly excels in prior probability [4]. For more general cases diagonal approximation suits the purpose, because it is better in accuracy and not much slower than the Euclidean distance method. The same algorithm in Fig. 3 is applicable to this approximation, but in this case, the procedure using $h_k(\boldsymbol{x},c)$ is faster than $g_k(\boldsymbol{x},c)$. This is because by using $g_k(\boldsymbol{x},c)$ the calculation of $|\boldsymbol{x}-\boldsymbol{m}_c|$ is necessary at first and its processing time is not small by ratio to the total processing time for this simple procedure. The time complexity is $\mathcal{O}(Cd)$.

4. Experiments

We measured recognition rates and processing times on several data sets by several methods. We performed the experiments on a Java Virtual Machine on a personal computer with two CPUs running at 1.8 GHz and a memory of 2 Gbytes.

4.1. Data specification

We show the specifications of the data sets we use in the experiment in Table 1(a). Etl9g3036 and Etl9g300 are data sets of ridge features [10] extracted from the character images in data set ETL9G [11]. ETL9G consists of 3036 categories, and each category consists of 200 samples. Etl9g3036 is the whole set of ridge features of all categories excluding 930 inappropriate samples of misentry or erroneous writing. Etl300 is a partial set of Etl3036 consisting of categories from no. 1001 to 1300. We also used data from UCI machine learning repository [13]. We chose four data sets with large dimensions made of numerical data: Isolet, Letter, Landsat, and Musk2.

Table 1(a) shows several quantities about the data sets which we think important in affecting the results. Data sufficiency S/Cd is the number of samples per category per dimensionality. To perceive how the samples are distributed across the categories, we define uniformity as

$$u = \frac{C}{C-1}\left(1 - \frac{1}{n^2}\sum_c n_c^2\right). \tag{23}$$

Table 1
Experimental results.

	Data set name					
	Etl9g 3036	Etl9g 300	Isolet	Letter	Landsat	Musk 2
(a) Data specification						
Item						
1. Number of categories, C	3036	300	26	26	6	2
2. Dimensionality of the feature, d	196	196	617	16	36	166
3. Number of samples, S	606270	59971	7797	20000	6425	6598
4. Data sufficiency, S/Cd	1.02	1.02	**0.32**	48.1	29.7	28.9
5. Uniformity, u	1.00	1.00	1.00	1.00	0.97	**0.52**
6. Skewness, η	0.4 ± 1.2	0.5 ± 1.3	0.6 ± 2.7	$0.3 \pm$ **0.7**	**0.2** ± 0.9	0.5 ± 1.7
7. Kurtosis, κ	1.7 ± 9.3	2.1 ± 10.1	8.0 ± 33.4	**0.6** $\pm$ **1.7**	1.2 ± 2.0	3.1 ± 17.7
(b) Recognition rate (%)						
Method						
1. ODA	99.53	99.92	96.87	**88.57**	83.95	87.59
2. QDA	**99.55**	**99.93**	**96.90**	88.53	**84.09**	**90.91**
3. Diagonal approximation	97.93	99.50	89.60	64.14	78.75	79.59
(c) Processing time per sample for measuring accuracy (ms (%))						
Method						
1. ODA, All discriminant functions	473 (100)	47 (100)	41.6 (100)	0.039 (100)	0.039 (100)	0.24 (100)
2. ODA, by $g_k(\boldsymbol{x},c)$	**17** (4)	**2.4** (5)	**3.9** (9)	**0.019** (51)	**0.015** (39)	**0.19** (81)
3. QDA, by $g_k(\boldsymbol{x},c)$	**16** (3)	**2.0** (4)	**3.8** (9)	**0.017** (45)	**0.015** (39)	**0.19** (80)
4. QDA, by $h_k(\boldsymbol{x},c)$	65 (14)	6.2 (13)	8.8 (21)	0.023 (59)	0.017 (44)	0.17 (71)
(d) Processing time per sample for classification by ODA (ms (%))						
Initialization						
1. Random selection	39 (8)	7.1 (15)	10.3 (25)	**0.027** (70)	0.024 (61)	0.21 (90)
2. Diagonal approximation	**21** (5)	**2.8** (6)	**4.4** (11)	0.028 (71)	**0.020** (52)	**0.19** (82)
2.1 Diagonal approximation itself	4 (1)	0.5 (1)	0.1 (0)	0.007 (17)	0.004 (11)	0.00 (2)
2.2 Main classification	17 (4)	2.4 (5)	4.2 (10)	0.021 (54)	0.016 (41)	0.19 (81)

Note: All the measured processing times fluctuate according to the operating condition of the computer with relative errors of about 5%.

According to this definition, there are exactly the same number of samples in every category when $u=1$, and all the samples are in one category when $u=0$. When data sufficiency and uniformity are small, ODA is considered to dominate in accuracy. The smallest value in each quantity is written in bold.

In the rows of Skewness and Kurtosis are the means and standard deviations of the following quantities:

$$\eta_{ci} = \frac{1}{\sigma_{ci}^3} E[(X_{ci} - m_{ci})^3], \tag{24}$$

$$\kappa_{ci} = \frac{1}{\sigma_{ci}^4} E[(X_{ci} - m_{ci})^4] - 3, \tag{25}$$

where $E[X]$ denotes the average of X throughout the samples, X_{ci} is the random variable of the sample feature element with category c and element number i. Addition of -3 in (25) assures that the kurtosis κ_{ci} is zero for the normal distribution. When both skewness and kurtosis are close to zero, the two discriminant analyses, ODA and QDA, are accurate since they assume the normal distribution. The smallest mean and standard deviation in each row are written in bold.

4.2. Recognition rate

We measured the recognition rates via 10-fold cross validation. Table 1(b) shows the results. We used the same parameter values, v_1, v_2, throughout the experiment for each discriminant analysis, which were determined so that the recognition rate was maximized. The parameter values for ODA, in left-to-right fashion of the table, are $(v_1, v_2) = (110,80)$, (80, 60), (500, 200), (0.0, 0.7), (20, 14), (0, 0). The parameter values for QDA [4] are not much different from those for ODA. We used the same parameter values in the experiment for diagonal approximation as for ODA. QDA outperforms ODA in recognition rate for 5 out of the 6 data sets. The five data sets include Isolet and Musk2, which have the smallest data sufficiency and the smallest uniformity, respectively. In such cases ODA is generally considered to outperform QDA contrary to the results. On the other hand, Letter, the only data set for which ODA outperforms QDA, is characterized by the small means and deviations of skewness and kurtosis. This fact was confirmed also by experiments using artificial data with various values of skewness and kurtosis generated by pseudorandom-number generators. Therefore, we deduce that the normality of the distribution is crucial for ODA to outperform QDA.

The recognition rates by diagonal approximation are substantially lower in all the data sets.

4.3. Processing time for accuracy measurement

Table 1(c) shows the processing times for the measurements of recognition rates. The time required for making dictionary and IO are excluded from the processing time. Line 1 shows the processing times for calculating the discriminant functions of all the categories by ODA. These values are used as denominators when calculating the percentages in parentheses. Lines 2 and 3 show the processing times for ODA and QDA using $g_k(\boldsymbol{x},c)$. ODA requires more processing time than QDA, but the differences are small. Line 4 shows the processing times by QDA using the sequence $h_k(\boldsymbol{x},c)$. They are substantially larger than those when using $g_k(\boldsymbol{x},c)$ for the data sets with a large number of categories.

4.4. Processing time for classification

In classification, where the answer is unknown unlike accuracy measurement, setting an appropriate initial category is preferable. Table 1(d) shows the processing times for cases where the

initialization is done by two ways: random selection and diagonal approximation. Random selection actually means no meaningful initialization; it is just for smoothing out accidental irregularities. As seen from lines (d)1 and (d)2, the initialization by diagonal approximation has effect in reducing processing time, with one exception, Letter. The processing times for the main classification after the initialization by diagonal approximation, are shown in line (d)2.2. They are always smaller than the processing times by random selection given in line (d)1. However, the initialization itself requires processing time, given in line (d)2.1. Therefore, the total processing time is not much different from random selection for the data sets with smaller numbers of categories or with low recognition rates, like Landsat, Musk2, or Letter.

The initialization affects the processing time, but not the recognition rate, which remains to be the same value given in Table 1(c), because it does not screen out any possibility of a category unlike the preclassification. Since the processing time for the main classification is only four times of the diagonal approximation for Etl3036, there is little point in a preclassification which screens out certain possibilities of categories.

5. Conclusions

The proposed algorithm drastically reduces the processing time for ODA and QDA without lowering the accuracy. In experiments measuring recognition rates of handwritten oriental characters, the processing time for ODA and QDA by this algorithm was 4% of the simple method calculating the discriminant functions for all the categories. The algorithm is more effective when the number of the categories is larger. This algorithm evades the redundant calculations by checking in every step of calculating a monotonically increasing sequence converging to the discriminant function. The best sequence $g_k(\boldsymbol{x},c)$ is obtained by replacing the $(d-k)$ smallest eigenvalues of the covariance matrix by their maximum.

When the features contain much disnormality and the training samples are distributed uniformly across the categories, QDA can possibly outperform ODA in accuracy.

Acknowledgments

The authors would like to thank Chisato Hayashi and Hideto Watanabe for their valuable discussions on hierarchical Bayes method.

Appendix A. The properties of the sequence $g_k(\boldsymbol{x}, c)$

We will omit the function argument $(\boldsymbol{x},c)$ in this section. First for $0 \le k \le d-2$, we have

$$\begin{aligned} g_{k+1}-g_k &= h_{k+1}-h_k+a_{c,k+2}j_{k+1}-a_{c,k+1}j_k \\ &= a_{c,k+1}w_{k+1}+a_{c,k+2}j_{k+1}-a_{c,k+1}(j_{k+1}+w_{k+1}) \\ &= (a_{c,k+2}-a_{c,k+1})j_{k+1} \ge 0. \end{aligned}$$

For $0 \le k \le d-1$, we have

$$g_k-h_k = a_{c,k+1}j_k > 0.$$

Furthermore, we have

$$g_{d-1} = h_{d-1}+a_{cd}j_{d-1} = a_c+\sum_{i=1}^{d-1} a_{ci}w_i+a_{cd}w_d = g.$$

Therefore, the sequence $g_k(\boldsymbol{x},c)$ increases monotonically, converges faster than $h_k(\boldsymbol{x},c)$, and reaches to $g(\boldsymbol{x},c)$ at $k=d-1$.

References

[1] F. Kimura, T. Wakabayashi, S. Tsuruoka, Y. Miyake, Improvement of handwritten Japanese character recognition using weighted direction code histogram, Pattern Recognition 30 (1997) 1329–1337.
[2] N. Kato, M. Suzuki, S. Omachi, H. Aso, Y. Nemoto, A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance, IEEE Trans. PAMI 21 (1999) 258–262.
[3] D. Keehn, A note on learning for Gaussian properties, IEEE Trans. 11 (1965) 126–132.
[4] M. Suzuki, A. Itoh, Reduction of processing time for quadratic discriminant analysis, IPSJ J. 50 (2009) 1789–1797 (in Japanese).
[5] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. PAMI 13 (1991) 252–264.
[6] X. Han, T. Wakabayashi, F. Kimura, The optimum classifier and the performance evaluation by Bayesian approach, Lecture Notes 1876 (2000) 591–600.
[7] P.J. Brown, T. Fearn, M.S. Haque, Discrimination with many variables, J. Am. Stat. Assoc. 94 (1999) 1320–1329.
[8] J.H. Friedman, Regularized discriminant analysis, J. Am. Stat. Assoc. 84 (1989) 165–175.
[9] J.P. Hoffbeck, D.A. Landgrebe, Covariance matrix estimation and classification with limited training data, IEEE Trans. PAMI 18 (1996) 763–767.
[10] M. Suzuki, C. Hayashi, A. Itoh, Recognition of grey-scale handwritten characters by ridge features, Tech. Rep. IEICE PRMU 106 (2007) 85–90 (in Japanese).
[11] T. Saito, H. Yamada, K. Yamamoto, On the data base ETL9 of handprinted characters in JIS Chinese characters and its analysis, Trans. IEICE J68-D (1985) 757–764 (in Japanese).
[12] H. Yamada, K. Yamamoto, T. Saito, A nonlinear normalization method for handprinted Kanji character recognition—line density equalization, Pattern Recognition 23 (1990) 1023–1029.
[13] A. Asuncion, D.J. Newman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2007 〈http://www.ics.uci.edu/mlearn/MLRepository.html〉.

About the Author—MICHITAKA SUZUKI received his B.E. degree in Physics from Chiba University, Japan, in 1975, and his M.E. degree and Doctor of Science degree in Physics from Tohoku University, Japan, in 1977 and 1987, respectively. In 1985, he joined Iwasaki College of Information Science, Yokohama, Japan, where he engaged in research in Computer Graphics. In 1991, he joined EXA Corporation, Kawasaki, Japan, where he engaged in IT research and development. He is currently with Research Institute of Science & Technology, Nihon University, Japan. He is a member of the Information Processing Society of Japan (IPSJ) and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

About the Author—AKIYOSHI ITOH received his B.E. degree in Electrical Engineering from College of Science and Technology, Nihon University, Japan, in 1966, with the highest distinction. He received his M.E. degree and Doctor of Engineering degree in Electrical Engineering from College of Science and Technology, Nihon University, Japan, in 1968 and 1978, respectively. In 1971, he joined Department of E.E. of College of Science and Technology, Nihon University. He was a Visiting Associate Professor in Department of ECE of Carnegie Mellon University during 1987 to 1988. He is currently a Professor at Computer Science Department, Graduate School of Science and Technology, Nihon University, Japan. His major is Materials Science and System Engineering for ultra high capacity information devices, and also he is interested in the pattern recognition of hand written characters and image processing technologies for medical applications. He is a member of the Magnetics Society of IEEE, Magnetics Society of Japan and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.