Learning Quadratic Discriminant Function for Handwritten Character Classification

Cheng-Lin Liu, Hiroshi Sako, and Hiromichi Fujisawa
Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
E-mail: {liucl, sakou, fujisawa}@crl.hitachi.co.jp

Abstract

For handwriting recognition integrating segmentation and classification, the underlying classifier is desired to give both high accuracy and resistance to outliers. In a previous evaluation study, the modified quadratic discriminant function (MQDF) proposed by Kimura et al. was shown to be superior in outlier rejection but inferior in classification accuracy as compared to neural classifiers. This paper proposes a learning quadratic discriminant function (LQDF) to combine the advantages of MQDF and neural classifiers. The LQDF achieves high accuracy and outlier resistance via discriminative learning and adherence to Gaussian density assumption. The efficacy of LQDF was justified in experiments of handwritten digit recognition.

1. Introduction

Statistical techniques and neural networks are widely used for classification. As a statistical classifier, the modified quadratic discriminant function (MQDF) proposed by Kimura et al. improves the computation efficiency and classification performance of QDF via smoothing the eigenvalues of minor subspace [1]. In a performance evaluation study, the MQDF was shown to be superior to neural classifiers in outlier rejection but inferior in classification accuracy when trained on large sample size [2].

The outlier resistance of classifiers is important in pattern field recognition, such as handwriting recognition integrating segmentation and classification [3]. Since the characters cannot be reliably segmented prior to classification, multiple hypotheses of segmentation are generated and a classifier is used to verify the hypotheses and identify the optimal one. Frequently, the segmentation hypotheses contain non-character patterns, which should be rejected by the underlying classifier rather than classified to a character class. An outlier resistant classifier is expected to give low confidences (outputs) to all classes on an outlier input.

The outlier resistance of MQDF originates from the assumption of Gaussian density for each class; while neural classifiers yield high accuracies because the parameters are trained in discriminative learning with aim to separate the patterns of different classes. In this paper, we propose a learning quadratic discriminant function (LQDF) to combine the outlier resistance of MQDF and the high accuracy of neural classifiers. The parameters of LQDF are subject to Gaussian density assumption and are adjusted in discriminative learning under the minimum classification error (MCE) criterion [4].

In regard to discriminative learning of quadratic classifiers, Watanabe and Katagiri have applied the MCE criterion to eigenvector learning of the subspace method [5], which can be viewed as a generalization of the learning subspace method [6]. Kurosawa generally discussed the discriminative learning of quadratic classifiers [7]. This paper, however, describes a successful implementation of discriminative learning for the MQDF, which is unique in that strong constraints are imposed onto the parameters. Experimental results in handwritten digit recognition justified the efficacy of LQDF, as compared to the MQDF and neural classifiers.

2. MQDF

On an input pattern $x = (x_1, \ldots, x_d)^T$, the MQDF for class $\omega_i, i = 1, \ldots, M$, is computed by

$$g_i(x, \omega_i) = g_i([x - \mu_i]^T \phi_{ij}]^2 + \sum_{j=k+1}^{d} \frac{1}{\lambda_j} [(x - \mu_i)^T \phi_{ij}]^2 + \sum_{j=1}^{k} \log \lambda_j + (d - k) \log \delta_i$$

where $k(<d)$ is the number of principal components, $\mu_i$ denotes the mean vector of class $\omega_i$, $\lambda_j, j = 1, 2, \ldots, d$, denote the eigenvalues of the covariance matrix $\Sigma_i$. With the "outlier", we mean a pattern that is out of the hypothesized classes that the classifier aims to detect and classify.
sorted in decreasing order, and $\phi_{ij}, j = 1, 2, \ldots, d$, are the corresponding eigenvectors.

The MQDF2 is the smoothed version of QDF, which performs Bayesian classification under the assumptions of multivariate Gaussian density and equal a priori probabilities. Compared with the original QDF, the minor eigenvalues $\lambda_{ij}, j = k + 1, 2, \ldots, d$, have been replaced with a larger constant $\delta_i$. In this way, the memory space and computation cost of the QDF are largely reduced.

The parameter $\delta_i$ can be heuristically set as a class-independent constant. In ML sense, the class-dependent $\delta_i$ equals the average of the minor eigenvalues. We have also combined the MQDF2 with the principle of regularized discriminant analysis (RDA) \[8\] to obtain what we call MQDFS. Before replacing the minor eigenvalues with the average, the covariance matrix of each class is interpolated with an identity matrix by

$$\Sigma_i = (1 - \gamma)\Sigma_i + \gamma \sigma^2_i I, \quad (2)$$

where $\sigma^2_i = \frac{1}{d} \text{tr}(\Sigma_i)$, and $0 < \gamma < 1$.

### 3. LQDF

The initial parameters of LQDF are inherited from the MQDF2 of the same structure. Then in discriminative learning, the parameters are updated to minimize the MCE criterion. For convenience of illustration, the MQDF2 is re-written as:

$$d_Q(x, \omega_i) = g_0(x, \omega_i) = \sum_{j=1}^{k} \frac{1}{\lambda_{ij}} p_{ij}^2 +$$

$$\frac{1}{k} \sum_{j=1}^{k} p_{ij}^2 \frac{1}{\lambda_{ij}} d_E(x, \omega_i) - \sum_{j=1}^{k} \log \lambda_{ij} + (d - k) \log \delta_i, \quad (3)$$

where $p_{ij} = (x - \mu_i)^T \phi_{ij}$ denotes the projection onto an eigenvector, and $d_E(x, \omega_i) = \|x - \mu_i\|^2$ is the squared Euclidean distance.

The MCE criterion of Katagiri and Juang \[4\] is illustrated as follows. On a training pattern, a loss function is computed to approximate the classification error and on a training dataset, the empirical loss is minimized to optimize the classifier parameters. Let the discriminant function of class $i$ equals the negative of distance:

$$g_i(x) = -d_Q(x, \omega_i),$$

the misclassification measure of a pattern from class $c$ is given by

$$\mu_c(x) = -g_c(x) + \left[ \frac{1}{M - 1} \sum_{\omega \notin c} g_0(x) \right]^{\frac{1}{2}},$$

where $\eta$ is a positive number. When $\eta$ approaches infinity, it becomes

$$\mu_c(x) = -g_c(x) + g_r(x),$$

where $g_r(x)$ is the discriminant function of the closest rival class: $g_r(x) = \max_{\omega \notin c} g_0(x)$. And finally,

$$\mu_c(x) = d_Q(x, \omega_c) - d_Q(x, \omega_r). \quad (4)$$

The misclassification measure is transformed to give the loss:

$$l_c(x) = l_c(\mu_c) = \frac{1}{1 + e^{-\mu_c}}. \quad (5)$$

On a training dataset $\{(x^n, c^n) | n = 1, 2, \ldots, N\}$ (where $c^n$ is the class label of pattern $x^n$), the empirical loss is computed by

$$L_0 = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{M} l_i(x^n) I(x^n \in \omega_i) = \frac{1}{N} \sum_{n=1}^{N} l_c(x^n), \quad (6)$$

where $I(.)$ is an indicator function.

To maintain the ortho-normality of eigenvectors, we add a regularization term to the loss function:

$$L_1 = \frac{1}{N} \left[ \sum_{n=1}^{N} l_c(x^n) + \beta q(\Phi) \right], \quad (7)$$

where $q(\Phi)$ denotes the deviation of ortho-normality:

$$q(\Phi) = \sum_{i=1}^{M} \sum_{j=1}^{k} \left( \frac{1}{2} \| \phi_{ij} \|^2 - 1 \right)^2 + \sum_{i=1}^{M} \sum_{j=1}^{k} \frac{1}{2} (\phi_{ij} \cdot \phi_{ij})^2.$$ 

By stochastic gradient descent \[9\], the parameters of LQDF are updated on each training pattern:

$$\begin{align*}
\lambda_{ij}(t+1) &= \lambda_{ij}(t) - \alpha_1(t) \frac{\partial l_c(x^n)}{\partial \lambda_{ij}} \\
&= \lambda_{ij}(t) - \alpha_1(t) \frac{\partial g_0(x^n)}{\partial \lambda_{ij}} \\
\delta_i(t+1) &= \delta_i(t) - \alpha_1(t) \frac{\partial g_0(x^n)}{\partial \delta_i} \\
&= \delta_i(t) - \alpha_1(t) \frac{\partial g_0(x^n)}{\partial \delta_i} \\
\mu_i(t+1) &= \mu_i(t) - \alpha_2(t) \frac{\partial l_c(x^n)}{\partial \mu_i} \\
&= \mu_i(t) - \alpha_2(t) \frac{\partial g_0(x^n)}{\partial \mu_i} \\
\phi_{ij}(t+1) &= \phi_{ij}(t) - \alpha_3(t) \frac{\partial l_c(x^n)}{\partial \phi_{ij}} \\
&= \phi_{ij}(t) - \alpha_3(t) \frac{\partial g_0(x^n)}{\partial \phi_{ij}} \\
\end{align*} \quad (8)$$

where

$$\frac{\partial g_{ij}(\Phi)}{\partial \phi_{ij}} = \sum_{i=1}^{M} \left( \phi_{ij}^2 - 1 \right) \phi_{ij} + \sum_{i=1}^{M} \sum_{j=1}^{k} (\phi_{ij} \cdot \phi_{ij}) \phi_{ij}.$$ 

The partial derivatives of $l_c(x)$ with respect to the parameters are computed by

$$\begin{align*}
\frac{\partial l_c(x^n)}{\partial \mu_i} &= \xi_c (1 - l_c) \frac{\partial g_0(x^n)}{\partial \mu_i} \\
&= -\xi_c (1 - l_c) \frac{\partial g_0(x^n)}{\partial \mu_i} \\
\end{align*} \quad (9)$$

where $\theta_c$ and $\theta_r$ stand for the parameters of class $c$ and class $r$, respectively. Further, the derivatives of the distance $d_Q(x, \omega_i)$ with respect to the parameters are computed by

$$\begin{align*}
\frac{\partial d_Q(x^n, \omega_i)}{\partial \lambda_{ij}} &= \frac{1}{\lambda_{ij}} \left( \lambda_{ij} - \frac{1}{\lambda_{ij}} \right) p_{ij}^2 \\
\frac{\partial d_Q(x^n, \omega_i)}{\partial \delta_i} &= \frac{1}{\lambda_{ij}} \left( \lambda_{ij} - \frac{1}{\lambda_{ij}} \right) \delta_i^2 \\
\frac{\partial d_Q(x^n, \omega_i)}{\partial \phi_{ij}} &= \frac{1}{\lambda_{ij}} \left( \lambda_{ij} - \frac{1}{\lambda_{ij}} \right) \delta_i \phi_{ij} - \frac{1}{\lambda_{ij}} (x - \mu_i) \\
\frac{\partial d_Q(x^n, \omega_i)}{\partial \mu_i} &= 2 \left( \frac{1}{\lambda_{ij}} - \frac{1}{\lambda_{ij}} \right) \delta_i \phi_{ij} (x - \mu_i). \quad (10)$$

1051-4651/02 $17.00 (c) 2002 IEEE
In implementation of MCE learning, the learning rates \( \alpha_i, i = 1, 2, 3 \), decrease progressively and the hardness parameter \( \xi \) increases progressively, all in a pre-specified schedule. Considering the difference of dynamic ranges of partial derivatives, the learning rates were set such that \( \alpha_1(t) \geq \alpha_2(t) \geq \alpha_3(t) \).

4. Experimental Results

The performance of LQDF was tested in handwritten digit recognition under the same condition with our previous evaluation study [2]. The experiment database was extracted from the NIST Special Database 19 (SD19). The training dataset and the test dataset contain 66,274 and 45,998 digit patterns, by 600 and 400 different writers, respectively.

To test the outlier rejection capability of classifiers, we synthesized 10,000 outlier patterns by splitting and merging digit images. In addition, 16,000 outlier patterns were generated for training neural networks to improve the outlier resistance. The synthesized outlier data was called type 1 outliers. We further used some English letter images from NIST SD19 as type 2 outliers for testing. The type 2 outlier data contains 7,600 patterns with 200 from each of 38 classes (all English letters except “DGIOSSHgllopz”)².

Each pattern (either digit or outlier) was represented in a feature vector of 100 measurements, extracted in sequential operations of size normalization, contour direction assignment and measurement blurring [2; lo].

First, we tested different versions of LQDFs that update different groups of parameters without regularization (\( \beta = 0 \)): LQDFI updates the eigenvalues \( \gamma_i \); LQDF2 updates the eigenvalues, \( \gamma_i \), and mean vectors; and LQDF3 updates all the parameters. The LQDFs inherit initial parameters from the MQDF2 with \( \delta \) being the average of minor eigenvalues.

With variable number of eigenvalues, the classification accuracies on the test data are shown in Fig. 1. The results of MQDF2 (Const refers to that with class independent constant \( \gamma_i \), and Aver refers to that with \( \gamma_i \) being the average of minor eigenvalues) and MQDF3 are also given for comparison. We can see that the discriminative learning of parameters is efficient to improve the classification accuracy.

To test the effect of regularization, we set the regularization coefficient to variable values \( \beta = 0; 1; 10; 100 \) (the case \( \beta = 0 \) corresponds to the LQDF3 in the above). We have also tested the effect of initializing the parameters of LQDF from MQDF3 with interpolation coefficient \( \gamma = 0.1 \). The error rates on the test data are listed in Table 1. The rightmost column of the table shows the average ortho-normalization error of eigenvectors

\[
e(\Phi) = \frac{q(\Phi)}{M \cdot k(k+1)/2}
\]

at \( k = 40 \). We can see that when the regularization coefficient is within an appropriate range, the ortho-normality is slightly improved by regularization. However, even when the regularization is not applied, the ortho-normalization error is small and the recognition accuracy is very high. Compared to the error rate of MQDF2 (1.630/c), the LQDF reduces the error rate by nearly 50%.

The accuracy of LQDF is also compared favorably to that of neural classifiers. We trained three neural classifiers, multilayer perceptron (MLP), radial basis function (RBF) classifier [11], and polynomial classifier (PC) [12] on the same training data as LQDF and tested on the same dataset. The MLP has one hidden layer of 300 units, the RBF classifier has 300 spherical Gaussian kernels, and the PC uses the linear and bi-nomial terms of 70 principal components. On the test dataset, the error rates of MLP, RBF classifier, and PC are 1.07%, 0.97%, and 0.89%, respectively.

The outlier rejection performance is measured in terms of the tradeoff between the false reject of digit patterns and the false acceptance of outliers at variable thresholds to the top rank class output. We tested the outlier rejection performance of LQDF (with \( \beta = 0; \beta = 10; \) and \( \gamma = 0.1 \), for all \( k = 40 \)) with comparison to that of MQDF2 (\( k = 30 \)) and enhanced neural classifiers (EMLP, ERBF, and EPC). The enhanced neural classifiers were trained with outlier data as well as digit data, while the MQDF2 and LQDF were trained without outlier data.

The accuracy of LQDF is hardly influenced. The ortho-normality is slightly improved by regularization. However, even when the regularization is not applied, the ortho-normalization error is small and the recognition accuracy is very high. Compared to the error rate of MQDF2 (1.630/c), the LQDF reduces the error rate by nearly 50%.

![Figure 1. Test accuracies of LQDFs](image)

Table 1. Error rates of LQDF at regularization

<table>
<thead>
<tr>
<th>k, b</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>e(\Phi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.31</td>
<td>0.96</td>
<td>0.89</td>
<td><strong>0.88</strong></td>
<td>0.91</td>
<td>0.91</td>
<td>1.32E-1</td>
</tr>
<tr>
<td>1</td>
<td>1.32</td>
<td>0.96</td>
<td>0.89</td>
<td><strong>0.88</strong></td>
<td>0.91</td>
<td>0.92</td>
<td>1.30E-4</td>
</tr>
<tr>
<td>10</td>
<td>1.31</td>
<td>0.97</td>
<td>0.89</td>
<td><strong>0.86</strong></td>
<td>0.92</td>
<td>0.91</td>
<td>1.12E-1</td>
</tr>
<tr>
<td>100</td>
<td>1.41</td>
<td>1.03</td>
<td>0.98</td>
<td><strong>0.89</strong></td>
<td>0.96</td>
<td>0.90</td>
<td>0.30E-4</td>
</tr>
<tr>
<td>200</td>
<td>1.36</td>
<td>0.98</td>
<td>0.89</td>
<td><strong>0.88</strong></td>
<td>0.86</td>
<td><strong>0.84</strong></td>
<td>2.12E-4</td>
</tr>
</tbody>
</table>

²Compared with the type 2 outlier set of [2], we have removed six more letters.
The tradeoffs are plotted in Fig. 2 (outlier type 1) and Fig. 3 (outlier type 2). We can see that the outlier rejection performance of LQDFs is very close to that of MQDF2, and both LQDF and MQDF2 outperform the enhanced neural classifiers in this respect.

5. Conclusion

This paper proposed a learning quadratic discriminant function (LQDF) for handwritten character recognition. The LQDF has the same structure and inherit initial parameters from MQDF2 yet the parameters are optimized in discriminative learning. The experimental results show that the LQDF provides both high accuracy and outlier resistance, which are desirable for handwriting recognition integrating segmentation and classification.

References


