Style-conscious quadratic field classifier

Sriharsha Veeramachaneni[a], Hiromichi Fujisawa[b], Cheng-Lin Liu[b] and George Nagy[a]

[a]Rensselaer Polytechnic Institute, Troy, NY, USA
[b]Hitachi Central Research Laboratory, Tokyo, Japan
E-mail: veeras@rpi.edu

Abstract

[1] *When patterns occur in the form of groups generated by the same source, distinctions between sources can be exploited to improve accuracy. We present a method for exploiting such 'style' consistency using a quadratic discriminant. We show that under reasonable assumptions on the feature and class distributions, the estimation of style parameters is simple and accurate.*

1. Introduction

It is well known that an optical character recognition system which is tailored to a certain font (machine print) or personalized to a particular writer (handwriting) has higher accuracy on input drawn from the particular font or writer than one which is trained on multiple fonts or multiple writers. Confusions between classes across fonts or writers cause this increase in error rate. If, during classification, the identity of the font or the writer is known, a mixed font/writer classifier can perform font/writer-specific classification leading to higher accuracy. However, this information is often unknown and difficult to determine during operation time.

Nevertheless, if the input patterns appear in the form of isogenous fields, that is, a group of multiple patterns (henceforth called singlets) belonging to the same font or generated by the same writer, classification accuracy can be improved. This dependence or correlation between features of co-occurring patterns is called *style context*. Style context can exist independently of linguistic context (dependence between class labels of co-occuring patterns) and hence can be exploited concurrently. Methods to utilize linguistic context to improve word recognition accuracy have been well studied [12][9]. Many multi-source recognition methods attempt to improve accuracy by extracting features that are invariant to source-specific peculiarities such as size, slant, skew, etc. [13].

Bazzi *et al.* observe that due to the assumption that successive observations are statistically independent, HMM methods have higher error rate for infrequent styles [1]. They propose modifying the mixture of styles in the training data to overcome this effect.

Cha and Srihari propose a dichotomizer that classifies a pair of documents into the categories 'same author' or 'different author' [2]. This is irrespective of the classes that appear on the documents, which is converse to our problem where it is known that co-occurring patterns are from the same writer and the problem is to identify the class irrespective of the writer.

Nagy suggests that even in the absence of any linguistic context, spatial context, i.e., the stationary nature of typeset, typeface and shape deformations over a long sequence of symbols can be used to improve classification accuracy [9]. Sarkar models styles by multi-modal distributions with two layers of hidden mixtures corresponding to styles and variants [11]. The mixing parameters are estimated using the EM algorithm. Although under the assumptions made, this approach is optimal, the complex EM estimation stage is prone to small-sample errors. Under essentially the same assumptions as ours, Kawatani attempts to utilize style consistency by identifying 'unnaturalness' in input patterns based on other patterns by the same writer [5]. This is done by finding the distance to other patterns classified into the same class or correlations with patterns classified into different classes. Our method combines the above heuristic criteria into one classification decision using all the information available. Koshinaka *et al.* model style context as the dependence between subcategory labels of adjacent symbols [7].

In many OCR applications quadratic discriminant methods provide a robust yet versatile means for classification. In order to reduce the effect of small-sample estimation of covariance matrices on accuracy, various methods to smooth the estimates, such as the Modified Quadratic Discriminant Function [6] and Rectified Discriminant Function [10], have been proposed.

[1]Submitted to ICPR '02

2. Mathematical formulation

Consider the problem of classifying an isogenous field $\mathbf{y} = (x_1^T, \ldots, x_L^T)^T$ (each x_i represents d feature measurements for one of L patterns in the field) generated by one of S sources $s_1, s_2, \ldots, s_S$ (writers, fonts, etc.). Let $\Omega = \{\omega_1, \ldots, \omega_N\}$ be the set of singlet-class labels. Let c_i represent the identity of the i^{th} pattern of the field.

We make the following assumptions on the class and feature distributions.

1. $p(s_k|\, c_1, c_2, \ldots, c_L) = p(s_k)\ \forall k = 1, \ldots, S$. That is, any linguistic context is source independent.

2. $p(\mathbf{y}|\, c_1, c_2, \ldots, c_L, s_k) = p(\mathbf{x}_1|\, c_1, s_k) p(\mathbf{x}_2|\, c_2, s_k) \ldots \ldots\ p(\mathbf{x}_L|\, c_L, s_k)\ \forall k = 1, \ldots, S$. The features of each pattern in the field are class-conditionally independent of the features of every other in the same field.

Under the above assumptions, we will show that for any function $f(\cdot)$, we have for $L = 2$, $E\{f(\mathbf{x}_1)|\, c_1, c_2\} = E\{f(\mathbf{x}_1)|\, c_1\}$. This result will be useful while deriving the formulae for the field-class-conditional means and covariance matrices.

$$\begin{aligned} E\{f(\mathbf{x}_1)|\, c_1, c_2\} &= \sum_{k=1}^{S} E\{f(\mathbf{x}_1)|\, c_1, c_2, s_k\} p(s_k|\, c_1, c_2) \\ &= \sum_{k=1}^{S} E\{f(\mathbf{x}_1)|\, c_1, s_k\} p(s_k) \\ &= E\{f(\mathbf{x}_1)|\, c_1\} \end{aligned} \quad (1)$$

2.1. Expressions for means and covariances

Under the assumption of field-class-conditionally normally distributed features we can use a quadratic discriminant function for field classification. We will now derive the formulae for field-class-conditional means and covariance matrices. For now we set $L = 2$.

Let $\mathbf{c} = (c_1, c_2)^T = (\omega_i, \omega_j)^T$ be a field-class label.

The mean vector for the field class $\mathbf{c}$ is given by

$$\mu_{ij} = E\{\mathbf{y}|\, c_1, c_2\} = \begin{pmatrix} E\{\mathbf{x}_1|\, c_1\} \\ E\{\mathbf{x}_2|\, c_2\} \end{pmatrix} = \begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix} \quad (2)$$

Thus the field-class-conditional mean vector can be constructed by concatenating the component singlet-class-conditional mean vectors.

Let us compute the field-class-conditional covariance matrix for the class $\mathbf{c}$ which we will denote by K_{ij}.

$$\begin{aligned} K_{ij} &= E\{(\mathbf{y} - E\{\mathbf{y}|\, c_1, c_2\})(\mathbf{y} - E\{\mathbf{y}|\, c_1, c_2\})^T\} \\ &= E\left\{ \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} (\mathbf{x}_1^T \mathbf{x}_2^T) | c_1, c_2 \right\} \\ &\quad - E\left\{ \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} |\, c_1, c_2 \right\} E\left\{ (\mathbf{x}_1^T \mathbf{x}_2^T)|\, c_1, c_2 \right\} \\ &= \begin{pmatrix} C_i & C_{ij} \\ C_{ji} & C_j \end{pmatrix} \end{aligned} \quad (3)$$

$$\begin{aligned} \text{where } C_i &= E\{\mathbf{x}_1 \mathbf{x}_1^T|\, c_1 = \omega_i\} \\ &\quad - E\{\mathbf{x}_1|\, c_1 = \omega_i\} E\{\mathbf{x}_1^T|\, c_1 = \omega_i\} \\ C_{ij} &= E\{\mathbf{x}_1 \mathbf{x}_2^T|\, c_1 = \omega_i, c_2 = \omega_j\} \\ &\quad - E\{\mathbf{x}_1|\, c_1 = \omega_i\} E\{\mathbf{x}_2^T|\, c_2 = \omega_j\} \\ C_{ii} &= E\{\mathbf{x}_1 \mathbf{x}_2^T|\, c_1 = \omega_i, c_2 = \omega_i\} \\ &\quad - E\{\mathbf{x}_1|\, c_1 = \omega_i\} E\{\mathbf{x}_2^T|\, c_2 = \omega_i\} \end{aligned}$$

Thus, K_{ij} can be written as a block matrix where the diagonal blocks are just the class-conditional singlet covariance matrices. Also, note that the above derivations can be generalized to longer fields to yield field-class-conditional means and covariance matrices that can be constructed from the singlet-class-conditional means and from the N blocks $C_1, C_2, \ldots, C_N$ and the $N(N+1)/2$ blocks (since $C_{ij} = C_{ji}^T$) $C_{11}, C_{12}, \ldots, C_{NN}$. For example, the field-feature mean and covariance matrix for class $(\omega_2, \omega_1, \omega_3)$ are given by

$$\mu_{213} = \begin{pmatrix} \mu_2 \\ \mu_1 \\ \mu_3 \end{pmatrix}; K_{213} = \begin{pmatrix} C_2 & C_{21} & C_{23} \\ C_{12} & C_1 & C_{13} \\ C_{32} & C_{31} & C_3 \end{pmatrix}$$

Hence the computation of field-class-conditional covariance matrices for long fields is no more complicated than for $L = 2$.

2.2. Estimation of the covariance matrices

As mentioned previously the C_i matrices are the class-conditional singlet covariance matrices. We have, from Equation 1

$$E\{\mathbf{x}_1 \mathbf{x}_1^T|\, c_1, c_2\} = \sum_{k=1}^{S} E\{\mathbf{x}_1 \mathbf{x}_1^T|\, c_1, s_k\} p(s_k) \quad (4)$$

Hence, the singlet covariance matrices can be computed from the weighted sum of the source-conditional 'power' matrices, and the singlet class means.

In general, the accurate estimation of the off-diagonal C_{ij} matrices requires a large number of field-samples for each field-class.

We will show that under assumptions made at the beginning of Section 2 the computation of the cross covariance matrices is simplified. We have

$$\begin{aligned} & E\{\mathbf{x}_1 \mathbf{x}_2^T|\, c_1 = \omega_i, c_2 = \omega_j\} \\ &\quad = \sum_{k=1}^{S} E\{\mathbf{x}_1 \mathbf{x}_2^T|\, c_1 = \omega_i, c_2 = \omega_j, s_k\} p(s_k) \\ &\quad = \sum_{k=1}^{S} E\{\mathbf{x}_1|\, \omega_i, s_k\} E\{\mathbf{x}_2|\, \omega_j, s_k\}^T p(s_k) \end{aligned} \quad (5)$$

Thus, the cross covariance matrices can be computed from the source-specific singlet-class means ($E\{\mathbf{x}|\,\omega_i, s_k\}$) as follows.

$$C_{ij} = \sum_{k=1}^{S} E\{\mathbf{x}_1|\,\omega_i, s_k\}E\{\mathbf{x}_2|\,\omega_j, s_k\}^T p(s_k) - \mu_i\mu_j^T \quad (6)$$

Hence the estimates of the cross covariance matrices are believed to be accurate.

2.3. Classification

Having obtained the estimates for the field-class-specific means and covariance matrices, we can classify the input fields using a straightforward quadratic classifier.

If all the field classes are *a priori* equally likely, the quadratic discriminant function (QDF) for field-class $(\omega_i, \omega_j)^T$ is given by

$$g_{ij}(y) = (y - \mu_{ij})^T K_{ij}^{-1}(y - \mu_{ij}) + log|K_{ij}| \quad (7)$$

The field feature vector y is assigned the class label which yields the minimum discriminant value.

We will now show that when the off-diagonal blocks in the field-class covariance matrices are zero matrices, i.e., $C_{ij} = 0,\ \forall i, j = 1, \ldots, N$, the quadratic field classifier is equivalent to the quadratic singlet classifier. Such a scenario results from the presence of only one source, or when the source-specific class means are same for all classes.

$$\begin{aligned} g_{ij}(y) &= (y - \mu_{ij})^T K_{ij}^{-1}(y - \mu_{ij}) + log|K_{ij}| \\ &= ((x_1 - \mu_i)^T (x_2 - \mu_j)^T) \begin{pmatrix} C_i^{-1} & 0 \\ 0 & C_j^{-1} \end{pmatrix} \\ &\quad \begin{pmatrix} x_1 - \mu_i \\ x_2 - \mu_j \end{pmatrix} + log|C_i| + log|C_j| \\ &= (x_1 - \mu_i)^T C_i^{-1}(x_1 - \mu_i) + log|C_i| \\ &\quad + (x_2 - \mu_j)^T C_j^{-1}(x_2 - \mu_j) + log|C_j| \\ &= g_i(x_1) + g_j(x_2) \end{aligned} \quad (8)$$

This implies that the field-class that minimizes $g_{ij}(y)$ is the one formed by the singlet-classes that minimize $g_i(x_1)$ and $g_j(x_2)$ respectively. Hence, any style information used to improve singlet classification is derived from the variation in the class means between sources. This is intuitively appealing since within-source variation does not contribute any style information.

We note that the quadratic field classifier is computationally expensive due to the exponential increase in number of classes with L and also because the discriminant computation involves larger matrices and vectors.

2.4. Covariance matrix smoothing

In the absence of sufficient training samples the estimation error of the covariance matrices degrades the accuracy of the QDF classifier. In order to alleviate the adverse effect of small-sample estimation on accuracy Friedman proposed the Regularized Discriminant Function (RDF) [3]. In RDF, the covariance matrix of a class is an interpolation of the estimated covariance matrix and the identity matrix,

$$\hat{C}_i = (1 - \gamma)C_i + \gamma\sigma_i^2 I \quad (9)$$

where $\sigma_i^2 = trace(C_i)/d$, and $0 < \gamma < 1$.

For the field-class covariance matrices, RDF was generalized as follows. The singlet-class covariance matrices $C_i,\ \forall i = 1, \ldots, N$ (the diagonal blocks in the field-covariance matrices) were modified according to Equation 9, and the off-diagonal blocks were modified according to $\hat{C}_{ij} = (1 - \gamma)C_{ij}\ \forall i, j = 1, \ldots, N$.

3. Experiments

We used the databases SD3 and SD7, which are contained in the NIST Special Database SD19 [4]. The database contains handwritten numeral samples labeled by writer and class. SD3 was the training data released for the First Census OCR Systems Conference and SD7 was used as the test data. We constructed four datasets, two from each of SD3 and SD7, as shown in Table 1. Since we compute the field class-conditional covariance matrices from source-specific class-conditional matrices we require that each writer have at least two samples for each singlet class. We therefore deleted all writers not satisfying this criterion from the training sets. Some of test fields are shown in Figure 1 along with their true class labels and the classification results.

	Writers	Number of samples
SD3-Train	0-399 (395)	42698
SD7-Train	2100-2199 (99)	11495
SD3-Test	400-799 (399)	42821
SD7-Test	2200-2299 (100)	11660

Table 1. Handwritten numeral datasets

We extracted 100 blurred directional (chaincode) features from each sample [8]. We then computed the principal components of the SD3-Train+SD7-Train data onto which the features of all samples were projected to obtain 100 principal-component features for each sample. The samples of each writer in the test sets were randomly permuted to simulate fields.

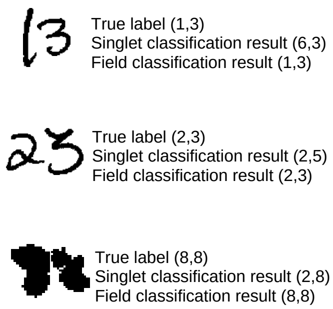

Figure 1. Some test fields with recognition results

The accuracy of the QDF and RDF classifiers was tested on various data sets for all 100 and the top 50 principal component (pca) features. The parameter γ for RDF was set to 0.2. The results are presented in Table 2.

Test set	Features	QDF		RDF	
		$L = 1$	$L = 2$	$L = 1$	$L = 2$
SD3-Test (42821)	Top 50	543	528	440	427
SD7-Test (11660)	Top 50	456	448	398	388
SD3-Test +SD7-Test	Top 50	999	976	838	815
SD3-Test (42821)	All 100	746	712	396	372
SD7-Test (11660)	All 100	551	534	369	352
SD3-Test +SD7-Test	All 100	1297	1246	765	724

Table 2. Number of character errors for various experiments (Training set = SD3-Train + SD7-Train)

The field classifier ($L = 2$) consistently outperforms the singlet classifier ($L = 1$). The results indicate that the regularization of the covariance matrices improves accuracy.

4. Discussion and future work

We have presented a model of style context in co-occurring patterns using second order correlations. We introduced a methodology to exploit such correlations to improve classification accuracy. We have demonstrated the efficacy of the scheme on handwritten data. We have also shown that our classifier is easy to train and is a natural extension of the widely used quadratic singlet classifier. As mentioned earlier, although the extension of the method to longer fields is theoretically possible, computational and storage constraints restrict its application. We are currently studying schemes to extend the pair classifier to longer fields without the need to store or manipulate large field-covariance matrices.

References

[1] I. Bazzi, R. Schwartz, and J. Makhoul. An omnifont open-vocabulary OCR system for English and Arabic. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-21(6):495–504, June 1999.

[2] S. Cha and S. Srihari. Writer identification: Statistical analysis and dichotomizer. In *Proceedings of the International Workshop on Structural and Syntactical Pattern Recognition*, pages 123–132, 2000.

[3] H. Friedman. Regularized discriminant analysis. *Journal of American Statistical Association*, 84(405):166–175, 1989.

[4] P. Grother. Handprinted forms and character database, NIST special database 19, March 1995. Technical Report and CDROM.

[5] T. Kawatani. Character recognition performance improvement using personal handwriting characteristics. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, volume 1, pages 98–103, 1995.

[6] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake. Modified quadratic discriminant functions and the application to Chinese character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(1):149–153, January 1987.

[7] T. Koshinaka, D. Nishikawa, and K. Yamada. A stochastic model for handwritten word recognition using context dependency between character patterns. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 154–158, 2001.

[8] C. Liu, H. Sako, and H. Fujisawa. Performance evaluation of pattern classifiers for handwritten character recognition. *International Journal on Document Analysis and Recognition*, in press 2002.

[9] G. Nagy. Teaching a computer to read. In *Proceedings of the Eleventh International Conference on Pattern Recognition*, volume 2, pages 225–229, The Hague, 1992.

[10] M. Sakai, M. Yoneda, and H. Hase. A new robust quadratic discriminant function. In *Proceedings of the Fourteenth International Conference on Pattern Recognition*, volume 1, pages 99–102, 1998.

[11] P. Sarkar and G. Nagy. Style consistency in isogenous patterns. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 1169–1174, 2001.

[12] R. Shingal and G. Toussaint. Experiments in text recognition with the modified Viterbi algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):164–172, April 1979.

[13] M. Shridhar and F. Kimura. Character recognition performance improvement using personal handwriting characteristics. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2341–2346, 1995.