Classification of style-constrained pattern-fields

Prateek Sarkar George Nagy
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy NY 12180, U.S.A.
E-mail: `sarkap@rpi.edu, nagy@ecse.rpi.edu`

Abstract

In some classification tasks, all patterns in a field, such as digits in a ZIP-code image, originate from the same, but unknown, source (writer/print style). The class-conditional feature distributions depend on the source of the patterns. Several sources may share the same distribution, or style. The style-conditional distributions are estimated from the training set. The optimal field-classifier computes the class-conditional field-feature-probabilities as the sum of class-and-style-conditional field-feature-probabilities, weighted by the prior probabilities of the styles. We compare the decision regions and error rates of style-weighted classification with both conventional singlet and top-style classification in a minimal family of examples, and discuss some related practical considerations.

1. Introduction

In many pattern recognition tasks, patterns appear in groups (*fields*) that have common traits owing to a common source or origin. For example, one can safely assume that in a single directory assistance call (field), every speech segment (pattern) is from the same speaker (source). Speech features depend on the gender of the speaker, and are more consistent within a gender than across genders. This induces underlying *styles* in fields of patterns. Feature measurements on different patterns of a field may not be statistically independent, but related through the underlying style. Style conscious classification exploits this phenomenon when the identity of the source is unknown [6].

The source, though unknown, provides some context in a field. We call this style context. This is different from linguistic context because we model the dependence of the feature measurements on patterns in a field, rather than the interdependence of their class-labels. Our method is also different from adaptive classification [1] because the style parameters are estimated in advance.

Styles arise in print (fonts) [7], script (writers), speech (speakers), vegetation (soil types or other locally uniform growing conditions), micrographs (dye concentrations and microscope characteristics).

It has been known for long in the pattern recognition community that using source or style specific classifiers can improve classification performance [4]. However, style or source recognition has traditionally been a separate step isolated from sample classification, often using a different set of features as in Optical Font Recognition [8]. We seek to find styles in the distribution of the pattern features, *i.e.*, the ones actually used for classifying patterns.

We illustrate some aspects of optimal style-conscious classification with a simple example with two classes, and two styles. We will use univariate, unit-variance Gaussians feature distributions conditioned on each class-style pair. The means of the conditional feature probabilities are either known, or estimated from training samples. Two scenarios will be considered for estimating the four means: training samples with class and style labels, and training samples with only class labels.

This simple framework gives rise to a surprisingly rich structure. We shall adhere to this framework throughout this paper, though it readily extends to more classes, styles, and more complex class-and-style conditional distributions, such as mixtures.

2. Formal problem statement

Our goal is to classify fields of two patterns. Each field is generated according to one of two styles, and thus has a (true) style-label, $s \in \{1, 2\}$. Each pattern in the field also has a (true) class-label $w_l \in \{1, 2\}$, where $l = 1, 2$ is an index to the position of the pattern in the field. The ordered pair of two class-labels is referred to as the field-label $(w_1, w_2) \in \{1, 2\} \times \{1, 2\}$. A single feature is measured on each pattern in a field, yielding a field feature (x_1, x_2). Neither the field-label nor the style-label is known for a test-field.

For each style and class, the pattern feature measurement is assumed to have a Gaussian distribution.

$$p(x_l|w_l, s) = N(\mu_{w_l,s}, \sigma) \text{ where } l{=}1, 2$$

The four means $\mu_{1,1}$, $\mu_{1,2}$, $\mu_{2,1}$, $\mu_{2,1}$, are either given, or can be estimated from the training samples. Same holds for prior class probabilities $P[w_l]$, $l{=}1,2$, $w_l{=}1,2$, and style-probabilities $P[s]$, $s{=}1,2$. For the sake of simplicity we assume that the prior class-probabilities do not depend on the position in the field ($P[w_1{=}1] = P[w_2{=}1]$), and also that there is no linguistic context ($P[(w_1, w_2)] = P[w_1]P[w_2]$). Neither assumption is a requirement for style-conscious classification.

A field classifier examines the field-pattern, and assigns to it a field-label (w_1^*, w_2^*). The classification is correct if $w_1^* = w_1$ *and* $w_2^* = w_2$, and erroneous otherwise. The conventional "singlet" classifier assigns field labels by considering each pattern-feature in the field independently of the other. For Bayesian classification, the class with the highest *a posteriori* probability is assigned to a pattern.

$$w_l^* = \operatorname*{argmax}_{w\in\{1,2\}} P[w] \sum_{s=1,2} p(x_l|w,s)P[s] \tag{1}$$

The summation indicates that class-conditional distributions of the features are mixtures induced by the styles. Since x_1 and x_2 are classified independently, the following is an equivalent, albeit computationally inefficient, method of field-label assignment.

$$\begin{aligned}(w_1^*, w_2^*) = \operatorname*{argmax}_{(w_1,w_2)} & \left[P[w_1]\sum_{s=1,2} p(x_1|w_1,s)P[s]\right] \times \\ & \left[P[w_2]\sum_{s=1,2} p(x_2|w_2,s)P[s]\right]\end{aligned} \tag{2}$$

The argument which is maximized above expands to

$$\begin{aligned}&P[w_1]P[w_2]\times \\ &\left[\begin{array}{rl} & p(x_1|w_1,1)P[s=1]\cdot p(x_2|w_2,1)P[s=1] \\ + & p(x_1|w_1,1)P[s=1]\cdot p(x_2|w_2,2)P[s=2] \\ + & p(x_1|w_1,2)P[s=2]\cdot p(x_2|w_2,1)P[s=1] \\ + & p(x_1|w_1,2)P[s=2]\cdot p(x_2|w_2,2)P[s=2] \end{array}\right]\end{aligned}$$

where $p(x_1|w_1,1)$ is a short-cut for $p(x_1|w_1,s=1)$.

The optimal style-conscious classifier differs in that it does not allow intermixing of styles within a field.

$$\begin{aligned}(w_1^*, w_2^*) = \operatorname*{argmax}_{(w_1,w_2)} & P[w_1]P[w_2]\times \\ & \sum_{s=1,2} p(x_1|w_1,s)p(x_2|w_2,s)P[s]\end{aligned} \tag{3}$$

The maximization argument in this case expands to

$$\begin{aligned}&P[w_1]P[w_2]\times \\ &\left[\begin{array}{rl} & p(x_1|w_1,1)p(x_2|w_2,1)\cdot P[s=1] \\ + & p(x_1|w_1,2)p(x_2|w_2,2)\cdot P[s=2] \end{array}\right]\end{aligned}$$

The argmax in (3) is taken over all possible field-labels (w_1, w_2), the number of which grows polynomially with the number of classes, and exponentially with field length. The classification can be sped up by selecting the most probable style instead of weighting the styles. Once the style is specified, each pattern can be classified individually. The resulting formula is sub-optimal but works for fields that carry dependable style traits.

$$\begin{aligned}(w_1^*, w_2^*) = \operatorname*{argmax}_{(w_1,w_2)} & P[w_1]P[w_2]\times \\ & \max_{s=1,2} p(x_1|w_1,s)p(x_2|w_2,s)P[s]\end{aligned} \tag{4}$$

The maximization over s replaces the summation in (3). The computation can now be made more efficient by changing the order in which the maximizations are performed.

$$\begin{aligned} s^* &= \operatorname*{argmax}_{s=1,2} P[s] \times \left[\max_{w_1} P[w_1]p(x_1|w_1,s)\right] \times \\ & \qquad \left[\max_{w_2} P[w_2]p(x_2|w_2,s)\right] \\ w_l^* &= \operatorname*{argmax}_{w_l} P[w_l]p(x_l|w_l,s^*) \text{ for } l = 1,2 \end{aligned} \tag{5}$$

Bazzi et al. [2] present results on using a sub-optimal classifier in the presence of styles. Their classifier is in essence a singlet classifier, but the weights in the style-induced mixture distributions are manipulated to counter the bias against less probable styles in long fields.

3. Example

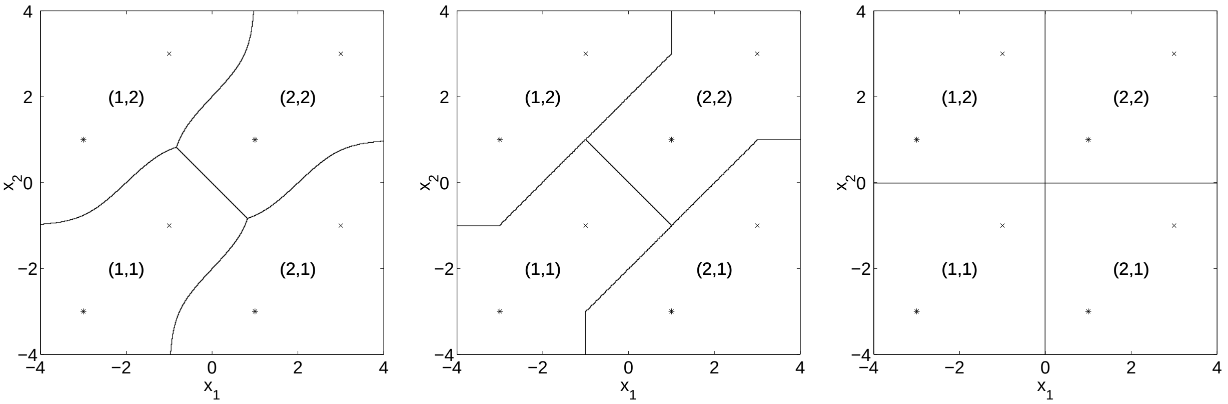
Let us consider the following example.

$$\begin{array}{ll} \mu_{1,1} = -3 & \mu_{1,2} = -1 \\ \mu_{2,1} = \mu_{1,1} + d = 1 & \mu_{2,2} = \mu_{1,2} + d = 3 \end{array}$$

Within each class, the style-specific distributions are separated by a distance of 2. The inter-class distance, d, is 4. All distributions are unit-variance Gaussians ($\sigma = 1$), and the class and style probabilities are equal.

$$\begin{array}{ll} P[w_l{=}1] = P[w_l{=}2] = 0.5 & \text{for } l{=}1,2 \\ P[s=1] = P[s=2] = 0.5 & \end{array}$$

Figure 1 shows the four decision regions for each of the three classifiers in the x_1-x_2 plane. In each region, the assigned field-label is shown in parentheses, and the means of the bi-variate field-label-conditional field-feature distributions are plotted as an asterisk and a cross for the styles 1 and 2 respectively. The field error rate for the conventional singlet classifier is 14.8%, that of the style conscious classifier is 10.5% (Table 1, $d = 4$). The error rate of the top-style classifier is 10.8%.



(a) (b) (c)

Figure 1. Decision regions for (a) style-conscious, (b) top-style, and (c) singlet classifiers.

Table 1. Performance of singlet and style-conscious classifier with supervised and unsupervised parameter estimates.

d	Percentage field error					
	Supervised		Unsupervised			
			40 training fields		400 training fields	
	Style	Sngl	Style	Sngl	Style	Sngl
0	74.7	74.7	75.0	75.3	74.5	74.8
1	57.2	60.3	61.6	62.6	57.3	60.7
2	38.5	45.1	39.8	47.9	38.8	44.8
3	22.3	28.7	23.6	32.8	22.4	28.7
4	10.5	14.8	11.0	18.2	10.6	14.7
5	3.9	6.0	4.2	6.6	3.9	6.0
6	1.1	1.9	1.4	4.2	1.1	2.0

The bivariate mixture-Gaussians make it difficult to compute the error rate for almost any decision boundary. We estimate error rates by generating 4000 random field-feature samples according to the known distribution, and then classifying them with the appropriate classifier.

We vary the inter-class distance, *d*, while the intra-class style-separation is fixed at 2. The reduction in field error rate achieved by the optimal and top-style classifiers over the singlet classifier (absolute gain), is plotted in Figure 2 as a function of *d*. “Relative gain” refers to the reduction, computed as a percentage of the singlet field-error. The absolute gain is highest when different classes from different styles have similar distributions (*e.g.* $d = 2$ and consequently $\mu_{1,2} = \mu_{2,1}$), because the conventional classifier cannot distinguish between these, while the style-conscious classifier can profit from information derived from the other pattern in the field. Of course, confusions within the same style cannot be resolved.

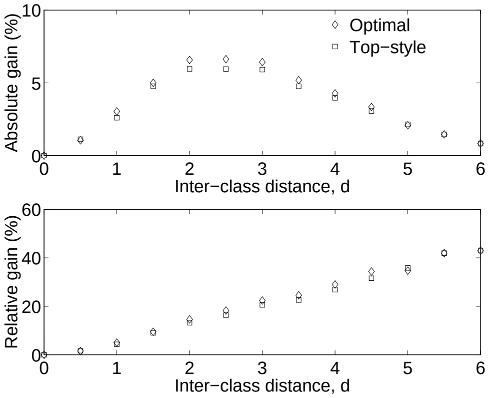


Figure 2. Improvement in classification performance provided by modeling styles as a function of shift *d*.

4. Training the style-conscious classifier

If the training sample is labeled with respect to style as well as class, then we can partition the training patterns by label and style and estimate the four means independently.

However, training data is seldom labeled by style. When only the class labels are given, then we must estimate the means of multi-modal mixtures. Unsupervised estimation is desirable further because

- Despite the presence of different styles, our ultimate goal is to classify patterns by class-labels, not styles.
- Obvious sources of style may not have a strong effect on patterns. Thus even though fonts may constitute different styles in printed characters, recognition may be affected more by the styles induced by quality of printing and scanning.
- It may not be obvious how human-perceived styles are

reflected in the feature distributions.

- While it is easy enough to understand the meaning of style in two documents with different type-faces, or hand-written by different writers, it is more difficult to assign five styles to one hundred type-faces or to a thousand writers. The number of styles that we can specify is limited by the number of training samples available, and by the computational complexity of the classification.

Thus hidden styles gives us the flexibility of defining styles in the most useful way, provided that we can find such styles automatically from training data. We have applied the EM algorithm [3, 5] to this end. Table 1 compares the classification performances with style-supervised and style-unsupervised estimates of parameters in our example with Gaussian mixtures.

5. Generalizations

We have hitherto confined our discussion to a simple case of our generalized framework for style conscious classification. The methods presented readily extend to:

Longer field lengths. Style-conscious classification benefits in situations when a class in some style is prone to be confused with other classes in other styles, because other patterns in the field furnish information on the underlying style. Fields formed entirely of such error-prone classes of a style are hard to classify. With increased field-length, probability of such fields reduces and the gain over singlet classification increases exponentially.

More styles. The modeling of additional styles helps only if they really exist.

Multidimensional features. Our EM algorithm estimates the means and variances from field samples, but so far we have not attempted to estimate covariance terms. Estimating full class-and-style-conditional covariance matrices for multidimensional features would require a large training set where all styles are populous.

Linguistic context. The use of linguistic context such as bigram or trigram class-probabilities is compatible with the use of styles.

Laboratory experiments on a six-font, ten-class, machine-printed digit recognition problem have shown style-conscious classifiers to be more accurate than singlet classifiers. A singlet classifier model, with six Gaussians per digit-class, yielded digit error rates of 19.8%. When the font-labels of test samples are known, and font-specific classifiers are used for classification, the error rate drops to 14.2%. When a six-style, one-Gaussian-per-class-per-style classifier is used for style-conscious classification of fields of length 4 and unlabeled style, the error rate is 14.9% [6].

6. Conclusions

Modeling styles can reduce the error rate on fields of patterns from the same source, provided that (1) styles are present in the data, and (2) some class of one style can be distinguished from the same class in another style.

Additional research is needed to find effective methods of determining the presence of styles, and the number of distinct styles, in a given collection of data. Other open problems include the application of styles to unsegmented patterns (perhaps in combination with HMMs), and the estimation of style parameters from training patterns without class labels.

Acknowledgment. We thank Dr. H. Fujisawa and his group, at the Hitachi Central Research Laboratories, Japan, for suggesting and supporting this project. This work has been conducted in the New York State Center for Automation Technologies (CAT), which is partially supported by a block grant the New York State Science and Technology Foundation.

References

[1] H. S. Baird and G. Nagy. A self-correcting 100-font classifier. In L. Vincent and T. Pavlidis, editors, *Document Recognition, Proceedings of the SPIE*, volume 2181, pages 106–115. 1994.

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

[3] R. A. Dempster, M. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[4] J. Rabinow. The present state of the art of reading machines. In L. N. Kanal, editor, *Pattern Recognition*, pages 3–30. Thompson Book Company, 1968.

[5] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[6] P. Sarkar. *Style consistency in pattern fields*. PhD thesis, Rensselaer Polytechnic Institute, USA, 2000.

[7] K. H. Warkentin. Classifying typefaces according to DIN. In P. Karow, editor, *Font technology methods and tools*, chapter 16. Springer-Verlag, 1994.

[8] A. Zramdini and R. Ingold. Optical Font Recognition using typograpgical features. *IEEE Transactions on PAMI*, 20(8), August 1998.