Combining Dichotomizers for MAP Field Classification

Srinivas Andra and George Nagy
DocLab, ECSE, Rensselaer Polytechnic Institute, Troy, NY, USA 12180
{andras@, nagy@ecse}.rpi.edu

Abstract

A new method for combining dichotomizers like SVMs is proposed for classifying multi-class pattern fields. The novelty lies in the estimation of the style-constrained posterior field class probabilities from the frequencies of the training patterns in the regions of the feature space engendered by the pairwise decision boundaries of the dichotomizers. We show that on simulated data, this non-parametric field classifier is nearly optimal. On scanned printed digits, its accuracy is comparable to that of state-of-the-art style classifiers.

1. Introduction

Style context was defined and formalized as the statistical dependence between patterns generated by the same source [1]. Style context, unlike language context, is independent of the order of the patterns in the field. Previous style-classifiers were based on Gaussian mixture densities [1, 2]. In contrast, the proposed method, like Parzen Windows and *k*-Nearest Neighbors classifiers, is a non-parametric MAP classifier [3, 4]. Like all style-constrained classifiers, the style-code field classifier described herein exploits the constraint that each pattern in a test field, regardless of its class, has the same style. Our underlying premise is that the dichotomizer (binary classifier) for each class pair also provides some style information about the other classes.

Each pattern is represented in some (arbitrary) feature space. The training patterns are labeled by class and style, and the test patterns are unlabeled. A set of dichotomizers assigns each training pattern to one of a set of regions indexed by the outputs of the dichotomizers. Therefore the output for each pattern is a binary vector of length equal to the number of dichotomizers. We use either class-pair or class-and-style-pair Support Vector Machines (SVMs) as dichotomizers [5], but the method is applicable to any linear or non-linear set of dichotomizers, to any number of classes and styles, and to any field length.

The patterns of each class and style in each region are counted. These pattern frequencies are estimates of the joint class-and-style posterior probabilities of each region. Each pattern can then be classified by frequency coding [6]. This is closely related to stacking pairwise classifiers [7], but formulated probabilistically to enable style-constrained field classification. As the number of training samples grows to infinity, the estimates will converge uniformly to the actual (unknown) posterior probabilities [6]. The field classifier is constructed by computing the field-class-and-style posterior probabilities from the corresponding singlet probabilities under certain class-and-style conditional independence assumptions. Combining diverse and moderately accurate classifiers − so called ensemble methods − to yield a highly accurate classifier is an active area of research [8, 9]. Ensemble methods can be applied to character recognition tasks as well [10, 11]. However, all these methods consider either singlet classification [10] or order-dependent field classification [11] and none incorporate style context.

We first present a formal description of the style-code classifier. Following this, we present a simulation with three classes *A*, *B* and *C* and two styles S_1 and S_2 that illustrates our method, and some experimental results on printed digits.

2. The style-code classifier

Each training pattern is labeled with one of N_c class labels $\{C_1, C_2, \ldots, C_{Nc}\}$, and one of N_s style labels $\{S_1, S_2, \ldots, S_{Ns}\}$. The output of dichotomizer Y_k on training pattern $\mathbf{x}_i$ is $Y_k(\mathbf{x}_i) = 1$ or $Y_k(\mathbf{x}_i) = 0$:
$\mathbf{Y} = \mathbf{Y}(\mathbf{x}_i) = (Y_1(\mathbf{x}_i), Y_2(\mathbf{x}_i), \ldots, Y_K(\mathbf{x}_i))$.

The dichotomizers together assign pattern $\mathbf{x}_i$ to a region in the feature space. A region can be described either by its *K*-element *binary region vector* **Y** or by its scalar *region index m*, $m = 1, \ldots, M$, $M = 2^K$. With linear dichotomizers (i.e., hyperplanes) in a *d*-

dimensional space, the maximum number of (convex) regions is only [12]:

$$N_x = \binom{K}{0} + \binom{K}{1} + \binom{K}{2} + \dots + \binom{K}{d}.$$

The index m can be the value of the binary integer formed by concatenating the elements Y_k of $\mathbf{Y}$ in any fixed order. The result of assigning all the training patterns $\mathbf{x}_i$ is a set of *region assignment matrices* $\{B^m, m = 1,2, \dots, M\}$ corresponding to the M possible regions. Each matrix B^m has elements $b^m{}_{i,j}$, where $b^m{}_{i,j}$ is the number of training patterns of class C_i and style S_j assigned to region m by the dichotomizers.

A *field* is a sequence of L patterns $\mathbf{x} = (\mathbf{x}^1, \dots, \mathbf{x}^\ell, \dots, \mathbf{x}^L)$ of the same style. Its field label is: $C(\mathbf{x}) = (C^1, \dots, C^\ell, \dots, C^L)$, where $C^\ell \in \{C_1, C_2, \dots, C_{Nc}\}$. C has $(N_c)^L$ possible values. With any set of dichotomizers, we can build a *style-code field classifier*. The output of the field classifier is a field assignment. For a test field $\mathbf{x}$, the *candidate field assignment of class labels* $E(\mathbf{x}) = (E^1, \dots, E^\ell, \dots, E^L)$, where $E^\ell \in \{C_1, C_2, \dots, C_{Nc}\}$, is a sequence of class labels selected by the field classifier according to the output of the dichotomizers on the test field $\mathbf{x}$ and on all the training patterns $\mathbf{x}_i$ (represented by the B^m matrices).[1]

Following [1] and [2], we assume that there is *no linguistic context* (i) $P[E] = P[E^1]P[E^2]\cdots P[E^L]$ and *field class is independent of style* (ii) $P[E \mid S_j] = P[E]$. Also that patterns are *class-and-style-conditionally independent:*

(iii) $P[\mathbf{x} \mid E, S_j] = P[\mathbf{x}^1 \mid E^1, S_j]P[\mathbf{x}^2 \mid E^2, S_j]\cdots P[\mathbf{x}^L \mid E^L, S_j]$.

Then the style-constrained posterior probability of a *candidate field assignment E* is

$$\begin{aligned} P[E \mid \mathbf{x}] &= \sum_{j=1}^{N_s} P[E, S_j \mid \mathbf{x}] \\ &= \sum_{j=1}^{N_s} P[E^1, E^2, \dots, E^L, \mathbf{x} \mid S_j]P[S_j] \big/ P[\mathbf{x}] \\ &= \frac{1}{P[\mathbf{x}]} \sum_{j=1}^{N_s} P[S_j] \left(\prod_{\ell=1}^{L} P[E^\ell, S_j \mid \mathbf{x}^\ell] \frac{P[\mathbf{x}^\ell]}{P[S_j]} \right) \end{aligned}$$

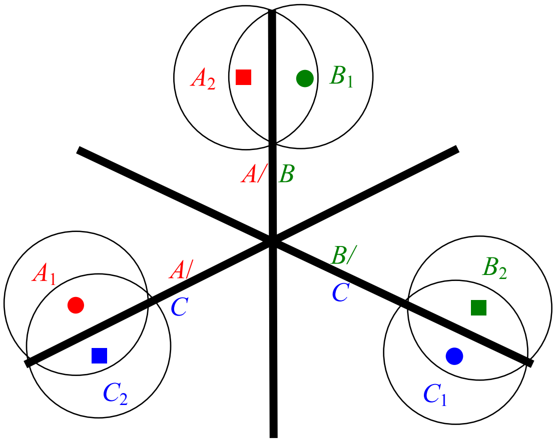
$$\begin{aligned} &= \frac{\prod_{\ell=1}^{L} P[\mathbf{x}^\ell]}{P[\mathbf{x}]} \sum_{j=1}^{N_s} \frac{1}{P^{L-1}[S_j]} \left(\prod_{\ell=1}^{L} P[E^\ell, S_j \mid \mathbf{x}^\ell] \right) \\ &= \beta \sum_{j=1}^{N_s} \frac{1}{\left(\sum_{m=1}^{M} \sum_{i=1}^{N_c} b^m_{i,j} \right)^{L-1}} \prod_{\ell=1}^{L} b^{m(\ell)}_{i^\ell, j} \\ &= \beta Z(i^1, \dots, i^\ell, \dots, i^L),\ i^\ell = 1, 2, \dots, N_c \qquad (1) \end{aligned}$$

where $m(\ell)$ is the region to which the ℓth pattern $\mathbf{x}^\ell$ of a test field $\mathbf{x}$ is assigned, and β is a constant of proportionality. The final field assignment consists of field class $E^*(\mathbf{x}) = (C_{\underset{\sim}{i}^1}, \dots, C_{\underset{\sim}{i}^\ell}, \dots, C_{\underset{\sim}{i}^L})$, where $Z(\underset{\sim}{i}^1, \dots, \underset{\sim}{i}^\ell, \dots, \underset{\sim}{i}^L) = \underset{i^1, \dots, i^\ell, \dots i^L}{\arg\max}\ Z(i^1, \dots, i^\ell, \dots, i^L)$.

A field error occurs when $E^*(\mathbf{x}) \neq C(\mathbf{x})$.

3. Simulation

The centroids of the six class-and-style distributions are shown in Fig. 1. The two styles of each class (S_1 and S_2) are widely separated. The centroids of S_1 are shown by disks, of S_2 by squares. We generated 2-D data according to this configuration. The inter-style distance (distance between a square and a disk of the same color) is d_s. The class-and-style distributions are identical Gaussians with covariance matrix $\sum = \sigma^2 I$. Under this assumption, the *linear* class-pair dichotomizers, shown as thick lines in Fig. 1, yield the optimal singlet classification rule.



[1] If a test pattern falls into an empty region m, i.e., if no training patterns were assigned to it, then its $b^m{}_{i,j}$ are set to $b^{m'}{}_{i,j}$, corresponding to the frequencies of the nearest non-empty region m' as measured by the Hamming distance between the region vectors $\mathbf{Y}$. Ties between equally-near regions are broken first by dominant class among all these regions, second by testing the next-nearest regions.

Figure 1. Centroids of class-and-style distributions. A centroid of class B and style S_1 is denoted by B_1.

Table 1. Region occupancy by class and style ($d_s = 0.6$)

Regions	$i = 1$ (A)		$i = 2$ (B)		$i = 3$ (C)	
m	$j = 1$	$j = 2$	$j = 1$	$j = 2$	$j = 1$	$j = 2$
110	346	13	1	0	1	122
111	11	338	123	0	0	0
011	*1*	*144*	*364*	*12*	2	0
001	0	2	11	363	119	0
000	1	1	*1*	*125*	*370*	*11*
100	*141*	*2*	0	0	*8*	*367*

We generated two training sets of 3000 samples each (500 samples drawn from each class-and-style distribution), with $\sigma = 0.15$, $d_s = 0.6$ and $\sigma = 0.15$, $d_s = 0.8$. An SVM with a linear kernel and capacity constant $C = 1$ was used for class-pair dichotomization. The distribution of training patterns in the regions ($b^m_{i,j}$) is shown in Table 1 for $d_s = 0.6$.

We consider an example with the locations in feature space of three patterns $\mathbf{x}^1$, $\mathbf{x}^2$, and $\mathbf{x}^3$ of a field
$\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3)$ as shown in Fig. 2. Its field label must assume one of the $3^3 = 27$ field labels *AAA*, *AAB*, *AAC*,
…., *CCB*, *CCC*. The relative values of the posterior probabilities of the field classes can be calculated from Equation (1). One relative posterior probability is

$$P[ACB \mid \mathbf{x}] = \beta\,(1/1500)^2\,(141 \text{ x } 370 \text{ x } 364 + 2 \text{ x } 11 \text{ x } 12)$$

$$= 8.44\beta, \text{ and another is}$$

$$P[CBA \mid \mathbf{x}] = \beta\,(1/1500)^2\,(8 \text{ x } 1 \text{ x } 1 + 367 \text{ x } 125 \text{ x } 144)$$

$$= 2.93\beta$$

In fact, if all three patterns have the same style, then these are the two highest among the 27 calculated values. Therefore the style-constrained field classifier will assign the label *ACB* to this field, whereas a singlet classifier would call it *CCB*, because *C* is the most probable label for $\mathbf{x}^1$ considered in isolation. The style constraint on the posterior probability computations accounts for the superiority of the field classifier over a style-unaware singlet classifier operating on the same dichotomizer outputs. Longer fields, more classes, and more dichotomizers favor the style-constrained classifier, because they yield more regions near the optimal decision boundaries, and therefore finer quantization of the underlying distributions.

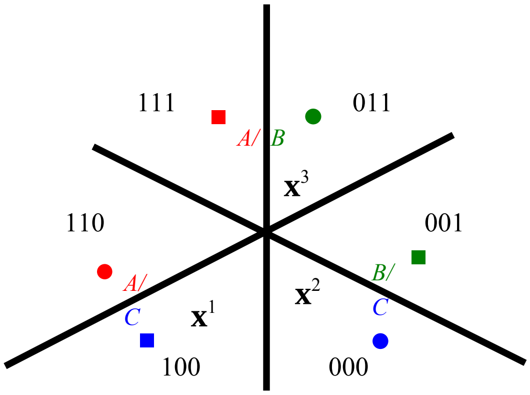


Figure 2. Example of a field $\mathbf{x}^1\mathbf{x}^2\mathbf{x}^3$ (i.e., a triplet of patterns), showing in which region of the feature space each of the patterns $\mathbf{x}^1$, $\mathbf{x}^2$, and $\mathbf{x}^3$ fell.

We constructed two different style-code classifiers. (i) *The simple style-code classifier* uses only 3 class-pair dichotomizers (solid lines in Fig. 3), and (ii) *the extended style-code classifier* uses both class-pair and class-and-style-pair dichotomizers (both solid and dashed lines, altogether 3 + 2 x 3 = 9 dichotomizers, in Fig. 3). The simple style-code classifier and extended style-code classifier are identical at field length $L = 1$ to the class-pair region-frequency classifier and class-and-style-pair region frequency classifier described in [6].

The distance between the most confused class centroids (A_1 & C_2, B_1 & A_2, and C_1 & B_2) is $(1 - d_s)/2$. This is within 0.67σ for $d_s= 0.8$ and 1.33σ for $d_s = 0.6$ respectively, therefore, high error rates are expected. The test sets, like the training sets, consisted of 3000 samples, with 500 samples drawn from each class-and-style distribution. An SVM with a linear kernel and capacity constant $C = 1$ was used for class-and-style-pair dichotomization as well.

The singlet error rates of the two style-code classifiers are compared with discrete style classifier [1], and the SQDF classifier [2] in Table 2. When parametric forms of the underlying class-and-style-conditional distributions are known, as in this example, the optimal discrete style field classifier gives the lowest field error rates and, usually, low singlet error rates. The benefits of style-code classifiers are greater for $d_s = 0.8$, because there are more inter-style

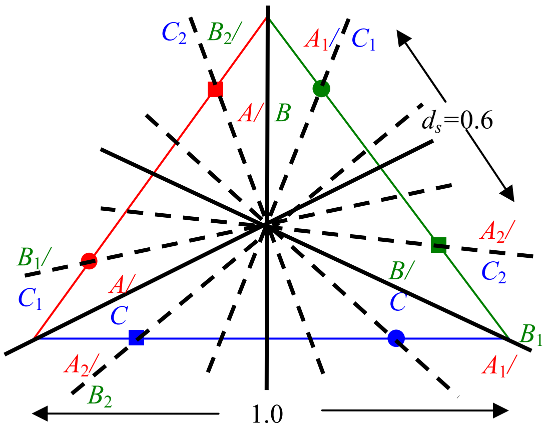


Figure 3. Class-pair decision boundaries (solid lines) and class-and-style-pair boundaries (dashed lines) for simulated Gaussian mixtures.

Table 2: Singlet error rates (%) of four field classifiers on simulated data

Field Length Classifier	L=1	L=2	L=3	L=4
d_s = 0.6				
Discrete style	25.4	18.9	15.0	12.0
SQDF	26.3	20.3	15.9	12.7
Simple style-code	25.3	26.4	18.8	18.9
Extended style-code	25.3	21.4	17.0	14.3
d_s = 0.8				
Discrete style	36.7	32.0	29.0	25.8
SQDF	40.2	36.7	33.4	30.4
Simple style-code	36.6	37.3	30.8	31.4
Extended style-code	37.0	34.3	30.4	28.5

confusions. At d_s = 0.8, the extended style-code classifier yields lower error rates than the SQDF classifier. This is due to the SQDF classifier's poor unimodal approximation of the bimodal class distributions.

A surprising result is that there is no reduction in the error rate of the simple style-code classifier as the field length is increased from L=1 to L=2, and, from L=3 to L=4 (the error rate, in fact, increases). This result is due to the symmetrical arrangement of class-and-style distributions of the simulated data. Fields of length L=2 are classified either accurately or both singlets are misclassified, whereas the singlet classifier misclassifies one singlet in each field. Therefore, there is no reduction in the singlet error rate of a simple style-code classifier. Such symmetry does not arise with odd field lengths as is evident from the reduction in the error rates. The extended style-code classifier does not suffer from the cancellation effect because of finer resolution in the MAP assignments to regions.

It is worthwhile to note that the training patterns did not leave any empty regions in the feature space, obviating the need for nearest-neighbor calculations during the test phase. Such an idealized scenario is unique to the simulated data. For a detailed description and results with other parameter settings and nonlinear dichotomizers, see [6].

4. Experiments on printed digits

The data consisted of 24,000 6-pt printed digits scanned at 200 dpi. Only 5 directional edge features were used. The printed digits were evenly distributed among five different fonts divided into two styles (serif and sans-serif). Therefore, the simple style-code

Table 3: Singlet error rates (%) of four field classifiers on printed data

Field Length Classifier	*L*=1	*L*=2	*L*=3
Discrete style	2.4	1.9	1.5
SQDF	2.9	2.2	2.0
Simple style-code	3.0	2.3	2.0
Extended style-code	2.2	1.6	1.4

classifier required 45 dichotomizers and the extended style-code classifier required 135 (45+2×45) dichotomizers. The training and test sets consisted of 12,000 digits each. The dichotomizers were again SVMs with a linear kernel and a capacity constant of $C = 1$.

The error rates of the two style-code classifiers are compared with the discrete style and SQDF classifiers in Table 3. The extended style-code classifier yields the lowest error rate, with the discrete style classifier a close second. The simple style-code classifier performs comparably to the SQDF classifier. The classifiers assuming the presence of a discrete number of styles in the data perform better than the SQDF classifier, even though the data may have more than two styles. The higher error rates of the simple style-code classifier can be attributed to coarser approximation of the posterior probabilities.

The simple style-code classifier does not use style-specific dichotomizers, but yields considerable reduction in the error rate with increase in field length on both simulated (only for odd field lengths) and printed data. It is easier to understand this phenomenon for simulated data. For each one of the three classes, only two dichotomizers are needed to uniquely identify the class of a test pattern. E.g., to identify a pattern of class *A*, only the decisions of *A/B* and *A/C* need be known for either majority-voting or frequency-coding based classification. However, the additional dichotomizer provides the necessary style discrimination for style-constrained classification.

The splitting of classes into style components by "third-party" dichotomizers is a corollary of the tetrahedral class-and-style arrangement postulated in [13]. The reductions in the error rates on printed digits render credence to this hypothesis. The datasets considered in this paper exhibit style, which is critical to the success of any style-constrained classifier. The amount of style in a dataset can be quantified to ascertain the suitability of style-constrained classification in a given application [14, 15].

Acknowledgement

We are grateful for helpful suggestions to Drs. C-L. Liu, P. Sarkar, and S. Veeramachaneni.

References

[1] P. Sarkar and G. Nagy, "Style consistent classification of isogenous patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 1, pp. 88-98, 2005.

[2] S. Veeramachaneni and G. Nagy, "Style context with second-order statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 1, pp. 14-22, 2005.

[3] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics,* vol. 33, no. 3, pp. 1065-1076, 1962.

[4] T.M. Cover, P.E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory,* vol. 13, no. 1, pp. 21-27, 1967.

[5] V. Vapnik, *Statistical Learning Theory,* Wiley, New York, 1998.

[6] S. Andra, Combining dichotomizers for MAP field classification, Walter J. Karplus Summer Research Grant Report, 2005.

[7] P. Savicky and J. Fürnkranz, "Combining pairwise classifiers with stacking," In *Proceedings of the 5th International Symposium on Intelligent Data Analysis (IDA)*, Berlin, Germany, 2003.

[8] T. G. Dietterich, "Ensemble methods in machine learning," In J. Kittler and F. Roli (Ed.) *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science,* Springer Verlag, New York, 2000, pp. 1-15.

[9] L. Breiman, "Random forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[10] L. Lam, Y.-S. Huang, and C. Suen, "Combination of multiple classifier decisions for optical character recognition," In H. Bunke and P. Wang (Ed.), *Handbook of Character Recognition and Document Image Analysis,* World Scientific, 1997, pp. 79-101.

[11] S. Günter and H. Bunke, "Evaluation of classical and novel ensemble methods for handwritten word recognition," In A. Fred et al. (Ed.): *Structural, Syntactic, and Statistical Pattern Recognition, Proc. Joint IAPR Int. Workshops SSPR and SPR,* Springer LNCS 3138, 2004, pp. 583 - 573.

[12] T. Zaslavsky, “Facing up to arrangements: face-count formulas for partitions of space by hyperplanes,” *AMS Memoirs*, vol. 1, no. 154, 1975.

[13] S. Veeramachaneni and G. Nagy, “Towards a Ptolemaic model for OCR,” In *Proc. the Sixth International Conference on Document Analysis and Recognition (ICDAR),* Edinburgh, Scotland, 2003, pp. 1060-1064,

[14] X. Zhang and S. Andra, “Towards quantifying the amount of style in a dataset,” In *Proc. SPIE Electronic Imaging, Document Recognition & Retrieval XIII,* vol. 6067, San Jose, CA, 2006.

[15] X. Zhang and G. Nagy, “Style quantification of scanned multi-source digits,” accepted, *International Conference on Pattern Recognition (ICPR),* 2006.