Classifier Adaptation with Non-representative Training Data

Sriharsha Veeramachaneni and George Nagy

Rensselaer Polytechnic Institute, Troy, NY 12180, USA,
veeras@rpi.edu

Abstract. We propose an adaptive methodology to tune the decision boundaries of a classifier trained on non-representative data to the statistics of the test data to improve accuracy. Specifically, for machine printed and handprinted digit recognition we demonstrate that adapting the class means alone can provide considerable gains in recognition. On machine-printed digits we adapt to the typeface, on hand-print to the writer. We recognize the digits with a Gaussian quadratic classifier when the style of the test set is represented by a subset of the training set, and also when it is not represented in the training set. We compare unsupervised adaptation and style-constrained classification on isogenous test sets of five machine-printed and two hand-printed NIST data sets. Both estimating mean and imposing style constraints reduce the error-rate in almost every case, and neither ever results in significant loss. They are comparable under the first scenario (specialization), but adaptation is better under the second (new style). Adaptation is beneficial when the test is large enough (even if only ten samples of each class by one writer in a 100-dimensional feature space), but style conscious classification is the only option with fields of only two or three digits.

1 Introduction

The design objective of document recognition systems is to yield high accuracy on a large variety of fonts, typefaces and handwriting styles. The most obvious approach, which is also the most popular amongst design engineers, is to collect patterns from all possible styles to train the classifier. Another approach is to design features that obscure differences between various styles. Due to the proliferation of fonts and typefaces with the advent of digital font design and the decreasing emphasis on neat handwriting, new styles encountered in the field often render these methods ineffective. Typically, OCR systems are overhauled periodically and retrained with patterns from newly encountered styles.

Statistical training of OCR engines is based on the assumption that the training set is representative of the statistics of the *test set*, i.e, the text that is encountered in the field. The only departure from this assumption has been the study of small-sample estimation problems, i.e., the *variance* of the classifier due to different draws from the population.

Although there are an immense variety of glyphs that correspond to each class, within a given document we expect to see a certain consistency owing to

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 123–133, 2002.

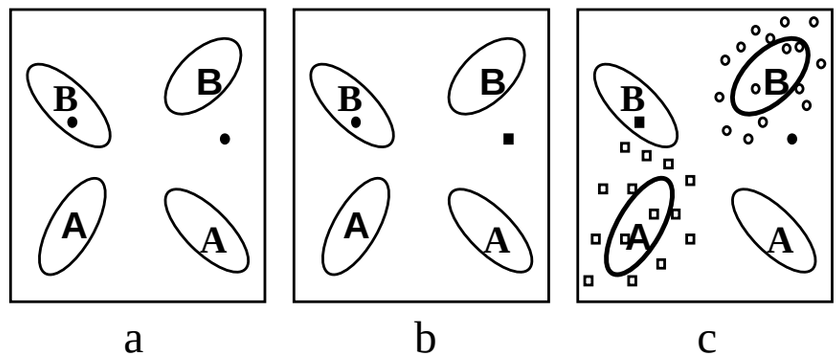


Fig. 1. Illustrative example. The equi-probability contours of the class-conditional feature distributions, for two typefaces and two classes {A, B}, estimated from the training data, are shown. The squares represent patterns classified as A's and the circles represent patterns classified as B's by each of the methods. The same two input patterns are shown filled in all three subfigures. (a) The conventional singlet classifier assigns the label (B, B) to these two patterns independently, oblivious to any style-consistency. (b) The style-conscious classifier assumes that both patterns are from the same typeface and assigns the label (B, A) to the left and right patterns respectively. (c) The existence of a large test set drawn from a single typeface causes an adaptive classifier to assign the label (A, B).

the common source. We call this consistency *style* or *spatial context*. The commonality denoted by style may arise from the processes of printing, scanning or copying as well as consistency of writer or typeface. Even documents composed with multiple scripts and fonts contain only a negligible fraction of all existing glyphs. In such scenarios, even if the style in which the current document is rendered is represented in the training data, the classification accuracy suffers from the relatively small weight given to the particular style by the classifier (which was trained on patterns drawn from a large number of styles). It is therefore appealing to consider the possibility of adapting the classifier parameters to the test set. Figure 1 illustrates the concepts of style-conscious classification and adaptive parameter estimation using a simple example.

Although little use of adaptive methods has been reported for OCR, there has been considerable work done in the field of communications, adaptive control and more recently in speech recognition [1]. Castelli and Cover explore the relative value of labeled and unlabeled samples for pattern classification [2]. Nagy and Shelton proposed a heuristic self-corrective character recognition algorithm that adapts to the typeface of the document to improve accuracy [3], which was later extended by Baird and Nagy to a hundred-font classifier [4]. Sarkar exploits style consistency in short documents (fields) to improve accuracy, under the assumption that all styles are represented in the training set, by estimating style-and-class-conditional feature distributions using the EM algorithm [5] [6]. We

have proposed a style-conscious quadratic discriminant classifier that improves accuracy on short fields under essentially the same assumptions [7].

Mathis and Breuel propose a hierarchical Bayesian approach very similar to our method to utilize the test data to improve accuracy [8]. They recursively apply EM estimation by combining the training and test data. We believe that this method introduces an avoidable classifier bias, especially if the size of the training data is commensurate with the size of the document being classified.

In the following sections we define the problem formally and describe a partial solution. We then present an experimental comparision of style-constrained classification and adaptation under different scenarios, and discuss the implications of the results.

2 Classifier Parameter Adaptation

We consider the problem of classifying the patterns in a large test set $T = \{x_1, \ldots, x_t\}$ where each x_i is a d-dimensional feature vector, into one of N classes $\{\omega_1, \ldots, \omega_N\}$. The test set T is drawn according to the class-conditional feature distributions $p(x|\,\omega_i) = f_i(x) \sim \mathcal{N}(\mu_i, \Sigma_i),\ i = 1, \ldots, N$ and *a priori* class probabilities $p(\omega_i) = p_i,\ i = 1, \ldots, N$.

We postulate the existence of a training set for estimating the class-conditional feature distributions given by $f_i^{(0)}(x) \sim \mathcal{N}(\mu_i^{(0)}, \Sigma_i^{(0)}),\ i = 1, \ldots, N$ and
$p^{(0)}(\omega_i) = p_i^{(0)},\ i = 1, \ldots, N$.

T is classified using a quadratic discriminant function classifier constructed with the estimated parameters. Clearly, the expected error-rate of the classifier on T is higher than the Bayes error-rate due to the discrepancies between the estimated parameters and the true parameters. We wish to adapt the classifier to the true parameters of the test set T.

We will assume that only the class-conditional feature means are misrepresented in the training set. That is, $\mu_i^{(0)} \neq \mu_i$ for some i, but $p_i^{(0)} = p_i$ and $\Sigma_i^{(0)} = \Sigma_i$ for all $i = 1, \ldots, N$. We adapt the estimate of the mean according to the following EM update formula

$$\mu_i^{(k+1)} = \frac{\sum_{x \in T} x p^{(k)}(\omega_i|\,x)}{\sum_{x \in T} p^{(k)}(\omega_i|\,x)},\ i = 1, \ldots, N$$

$$\text{where } p^{(k)}(\omega_i|\,x) = \frac{p_i f_i^{(k)}(x)}{\sum_{i=1}^{N} p_i f_i^{(k)}(x)},\ f_i^{(k)}(x) \sim \mathcal{N}(\mu_i^{(k)}, \Sigma_i)$$

Although convergence to the means of the test distribution is not guaranteed with arbitrary initialization, it appears that the true mean is a fixed point of the algorithm with good covariance estimates.

3 Experimental Results on Machine-Printed Data

A database of multi-font machine-printed numerals was generated as follows [6]. Five pages, containing the ten digits 0-9 spaced evenly and replicated 50 times,

were prepared using Microsoft Word 6.0. Each page was rendered in a different 6 pt typeface, namely Avant Garde (A), Bookman Old Style (B), Helvetica (H), Times New Roman (T), and Verdana (V), and printed on a 600 dpi Apple LaserWriterSelect. Each page was scanned 10 times at 200 dpi into 10 bilevel bitmaps using an HP flatbed scanner. This yielded a total of 25,000 samples (5000 samples per typeface). A few of the samples are shown in Figure 2. The resulting scanned images were segmented and for each digit sample 64 blurred directional (chaincode) features were extracted and stored [9]. We used only the top 8 principal component features for experimentation so that the gains in accuracy are significant. For each typeface, 2500 samples were included in the training set, while the remaining 2500 samples were used for testing. That is, the number of errors in each cell of the tables below are based on 2500 test patterns.

Fig. 2. Samples of the machine-printed digits, reproduced at approximately actual size.

Figure 3 shows the scatter plot of the top two principal component features of the test samples from typefaces Helvetica and Verdana. For clarity only some of the classes are shown.

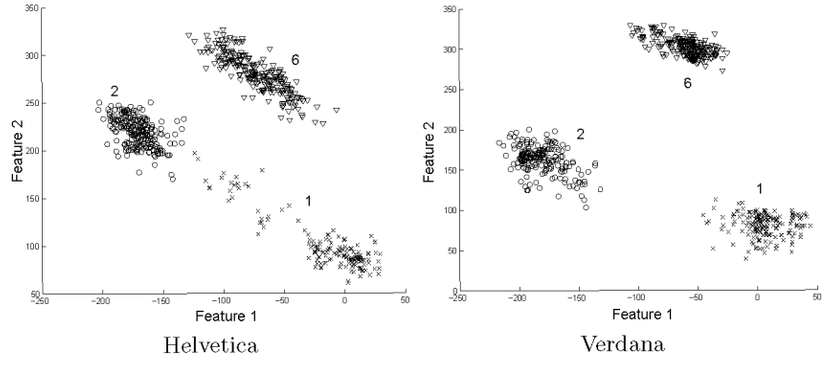


Fig. 3. Scatter plot of the top two principal component features for test samples from classes '1', '2' and '6'.

3.1 Multiple-Style Training, Test Style Represented

Here we consider the case when the classifier is trained on patterns from multiple typefaces including the typeface of the test set. We use the training data from all typefaces to train the classifier and adapt to each typeface separately. The recognition results for iterated adaptation of the mean are presented in Table 1.

Table 1. Error counts on different typefaces for successive EM iterations, 2,500 samples per typeface for testing (All-typefaces training)

	Test typeface				
Iterations	A	B	H	T	V
0	17	3	34	1	33
1	7	2	37	1	3
5	6	2	39	1	3
10	6	2	39	1	3

We now compare the results in Table 1 with our style-conscious quadratic classification [7]. The error-rates for various field lengths are presented in Table 2.

Table 2. Error counts using style-conscious field classification, 2,500 samples per typeface for testing (All-typefaces training)

	Test typeface				
Field length	A	B	H	T	V
1	17	3	34	1	33
2	6	4	35	0	7
3	1	4	33	0	2

We observe from Tables 1 and 2 that even when the test style is represented in the training set, utilizing the information that the entire test set is drawn from the same style can lead to improved accuracy. The style-conscious quadratic classifier outperforms the adaptive scheme. We attribute this anomaly to the violation of our assumption that the estimates of the covariance matrices from the training data are representative. Actually, the estimated feature variances are 'larger' than the typical single-typeface variance due to the variation in means across typefaces. We have tried using the average typeface-specific covariance matrix estimated from the training set. The error rates after adaptation were higher due to the higher initial error rate (the covariance matrix including the variance of the means is more representative of all the typefaces than the average typeface-specific covariance matrix). Also because of the high degree of consistency in machine-printed numerals, a few test patterns (i.e., short fields) are sufficient to specialize to the test style. The best that we can hope to achieve

with either the style-conscious classifier or the adaptive classifier is accuracy equaling typeface-specific singlet classification (diagonal entries in Table 5).

3.2 Multiple-Style Training, Test Style Not Represented

Here the classifier is trained on patterns from multiple typefaces, but excluding the typeface of the test data. We trained the classifier five times, each time excluding the training data from the typeface of the test data. The recognition results for the mean adaptation are presented in Table 3.

Table 3. Error counts on different typefaces (Leave-one-typeface-out training)

	Test typeface				
Iterations	A	B	H	T	V
0	102	113	146	33	141
1	7	14	76	5	6
5	5	2	44	4	4
10	5	2	43	4	4

Table 4 shows the error counts of the style-conscious quadratic field classifier when the test style is not represented in the training set.

Table 4. Error counts with style-conscious field classification (Leave-one-typeface-out training)

	Test typeface				
Field length	A	B	H	T	V
1	102	113	146	33	141
2	109	115	160	34	115
3	97	119	166	29	98

The potential of the adaptive scheme is more evident when the test style is *not* represented in the training data. Table 3 indicates that the classifier converges after only a few iterations and yields startling improvement in accuracy. As expected, the style-conscious classifier is impotent here, performing poorly even when classifying triples (Table 4).

3.3 Single-Style Training

This is a more challenging task for the adaptive algorithm. We train on only one typeface and classify the test sets of each typeface. The recognition results are presented in Table 5 for 0, 1 and 5 iterations of the mean adaptation algorithm.

The experimental results presented in Table 5 explore the most extreme case of non-representative training data. We observe that the adaptive classifier

Table 5. Error counts with cross-training (Each row is for the same training set, each column is for the same test set)

		Test typeface				
	Iter	A	B	H	T	V
A	0	**0**	520	440	276	486
	1	**0**	326	305	133	230
	5	**0**	40	179	2	0
B	0	145	**1**	111	54	92
	1	15	**1**	36	1	3
	5	8	**2**	51	0	1
H	0	17	154	**13**	164	555
	1	5	8	**13**	2	73
	5	5	8	**13**	1	0
T	0	163	324	433	**0**	3
	1	73	290	404	**0**	0
	5	0	162	402	**0**	0
V	0	251	481	517	34	**0**
	1	98	289	401	4	**0**
	5	4	321	408	3	**0**

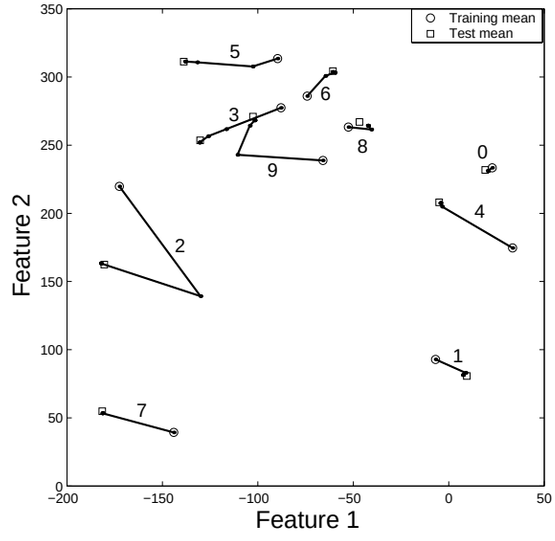
improves for every pairing of training and test sets, although not uniformly. For finite sized test set the maximum likelihood estimates do not necessarily yield the minimum error. In Table 3 there are cases where the accuracy after 5 iterations is lower that after one iteration. The gains obtained from adaptation are not symmetric because the convergence properties of the EM algorithm depend upon the initial estimates of the parameters. Figure 4 shows the loci of the class-conditional feature means of the top two principal component features for five iterations. The adaptative estimation of the means is much less effective when Verdana is used for training and Helvetica for testing than vice-versa because of convergence to a local minimum.

In Table 5 the error counts along the diagonal (in boldface) represent the lower bounds attainable with same-typeface training. They are, as expected, stable under EM iteration.

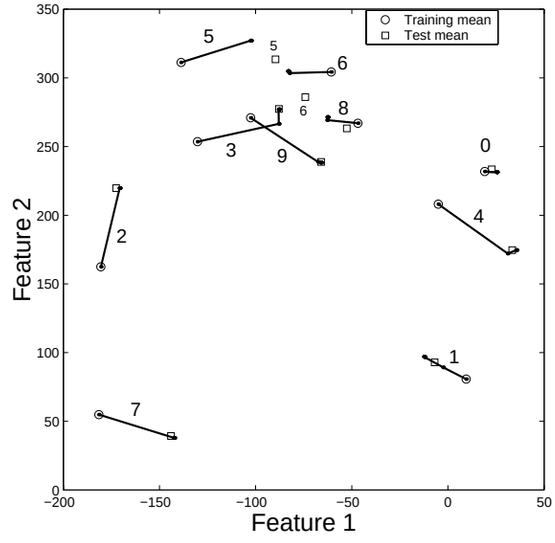
4 Experimental Results on Handwritten Data

We used the databases SD3 and SD7, which are contained in the NIST Special Database SD19 [10]. The database contains handwritten numeral samples labeled by writer and class (but not of course by style). 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 6. Each writer has approximately 10 samples per class.

We extracted 100 blurred directional (chaincode) features from each sample [9]. We then computed the principal components of the SD3-Train+SD7-



Training on Helvetica, Testing on Verdana



Training on Verdana, Testing on Helvetica

Fig. 4. Loci of the class-conditional feature means, during adaptation, of the top 2 principal component features. The circles represent the class means of the training data and the squares represent the class means of the test data. The dots represent the means after each iteration.

Table 6. Handwritten numeral datasets

	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

Train data onto which the features of all samples were projected to obtain 100 principal-component features for each sample.

Since the writers were arbitrarily chosen to form the training and test sets we did not expect any significant improvement in accuracy with mean adaptation when the *entire* test data was assumed to be from the same style. Our belief was confirmed by the recognition rates obtained.

Table 7. Error-rates in % with mean adaptation on handwritten data (Each row is for the same training set, each column is for the same test set)

		Test set	
	Iter	SD3-Test	SD7-Test
SD3-Train	0	2.2	8.0
	1	2.0	7.0
	5	1.9	6.5
SD7-Train	0	3.7	3.6
	1	2.8	3.2
	5	2.6	3.0
SD3-Train +SD7-Train	0	1.7	4.7
	1	1.5	4.0
	5	1.5	3.7

When the test data is known to be from a single writer, we do expect a good adaptive scheme to specialize the decision regions to the said writer. Table 7 shows the recognition rates for various iterations of the mean adaptation algorithm when samples of the test set are adaptively classified, operating on one writer at a time. For each writer in the test set the means are initialized before adaptation to those of the entire training set.

The recognition results on handwritten data (Table 7) indicate that even when the test data is small (approximately 10 samples per class) adapting the mean improves accuracy. The adaptive classifier that averages over the approximately 10 samples per digit available from each writer is better than the style-conscious classifier operating on fields of only two digits owing to the large variation in handwriting styles. The style-conscious classifier cannot fully exploit style consistency with such short fields.

Table 8. Error-rates on handwritten data before and after adaptation (5 iterations), showing the percentage of test writers that improved or worsened with adaptation.

		Error-rate (%)		% writers	
Training set	Test set	Before Adaptation	After Mean Adaptation	Accuracy increased	Accuracy decreased
SD3-Train	SD3-Test	2.2	1.9	19.5	1.3
	SD7-Test	8.0	6.5	54.0	2.0
SD7-Train	SD3-Test	3.7	2.6	43.9	1.3
	SD7-Test	3.6	3.0	39.0	4.0
SD3-Train	SD3-Test	1.7	1.5	15.8	2.0
+SD7-Train	SD7-Test	4.7	3.7	54.0	2.0

Table 8 shows the percentage of writers in the test set on which the accuracy increased and decreased with adaptive classification (after 5 iterations). The maximum improvement for any particular writer was approximately 10% while the maximum decrease in accuracy was about 2%.

5 Discussion and Future Work

The above results confirm the value of adaptive classification in the presence of a large volume of style-consistent test data. It is possible to design OCR systems that improve with use. For machine-printed data, when the style of the test data is represented in the training set, but the size of the test data is small, it is advantageous to use a style-conscious classifier over an adaptive methodology. Either method can, of course, be combined with language context.

We intend to extend the adaptive methodology to recursively estimating the covariance matrices as well. We believe that when only a moderate sized test data is available, the EM algorithm is unstable if used to estimate the covariances, and therefore intend to explore methods that exploit the configuration of the class-conditional densities in the feature space. We also intend to explore the possibility of using adaptation as a substitute for covariance matrix regularization in small training sample scenarios. Another important problem is to identify, at run-time, situations when the adaptive classifier degrades accuracy. This problem is related to the convergence properties of the EM algorithm and depends on the initialization strategy. Although we currently intialize the EM algorithm to the parameters estimated from the training set, we intend to explore other intialization methods. We also plan to study the conditions under which the adaptation can be guarateed to improve accuracy.

Acknowledgements. We thank Dr Hiromichi Fujisawa, Dr. Cheng-Lin Liu and Dr. Prateek Sarkar for the informative discussions we had over the years.

References

1. C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, April 1995.
2. V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42:2102–2117, November 1996.
3. G. Nagy and G. L. Shelton Jr. Self-corrective character recognition system. *IEEE Transactions on Information Theory*, IT-12(2):215–222, April 1966.
4. 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.
5. P. Sarkar. *Style consistency in pattern fields.* PhD thesis, Rensselaer Polytechnic Institute, Troy, NY, 2000.
6. 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.
7. S. Veeramachaneni, H. Fujisawa, C.-L. Liu, and G. Nagy. Style-conscious quadratic field classifier. In *Proceedings of the Sixteenth International Conference on Pattern Recognition*, 2002. (Accepted).
8. C. Mathis and T. Breuel. Classification using a Hierarchical Bayesian Approach. Submitted for publication.
9. C.L. Liu, H. Sako, and H. Fujisawa. Performance evaluation of pattern classifiers for handwritten character recognition. *International Journal on Document Analysis and Recognition*, 4(3):191–204, 2002.
10. P. Grother. Handprinted forms and character database, NIST special database 19, March 1995. Technical Report and CDROM.