CLASSIFICATION BY PROBABILISTIC CLUSTERING

Thomas M. Breuel

Xerox PARC
3333 Coyote Hill Road
Palo Alto, CA, USA
tbreuel@parc.xerox.com

ABSTRACT

This paper describes an approach to classification based on a probabilistic clustering method. Most current classifiers perform classification by modeling class conditional densities directly or by modeling class-dependent discriminant functions. The approach described in this paper uses class-independent multi-layer perceptrons (MLPs) to estimate the probability that two given feature vectors are in the same class. These probability estimates are used to partition the input into separate classes in a probabilistic clustering. Classification by probabilistic clustering potentially offers greater robustness to different compositions of training and test sets than existing classification methods. Experimental results demonstrating the effectiveness of the method are given for an optical character recognition (OCR) problem. The relationship of the current approach to mixture density estimation, mixture discriminant analysis, and other OCR and handwriting recognition techniques is discussed.

1. INTRODUCTION

Current classifiers for the recognition of handwriting, printed characters, phonemes, and similar signals can achieve very high performance (often exceeding that of humans) when given sufficiently large and representative training sets. Techniques have also been used to synthesize additional training examples from a given training set to further increase the effective training set (and ability to generalize) for the classifier. A key limitation of such approaches is still that they can be sensitive to novel data whose distribution is significantly outside the training set. Yet, in many cases, human performance on such novel examples can be quite good. A particularly common example is that of generalization of OCR systems to previously unseen fonts. Novelty fonts that pose little problem to human readers will often be difficult to recognize for OCR systems.

To address this problem, an approach to OCR based on cryptanalysis has been proposed in the literature [1, 2]. In this approach, characters of similar appearance are grouped together. Each such group corresponds to an unknown character. The text is thereby transformed into a sequence of tokens, each of which corresponds to an unknown character. By determining the correspondence between tokens and characters, the textual transcription of the document can be obtained. Determining this correspondence can be carried out by techniques from cryptanalysis and statistical language processing.

In general, however, grouping together “characters of similar appearance” is a hard problem. The simple techniques used for determining similarity of appearance in the past are not robust enough to reliably identify related characters in degraded documents. We can address this problem by a more careful statistical modeling of the distributions that degradations of documents induce in the appearance of characters. One example of such a model, based on Gaussian mixtures, and its use for recognition by clustering has been described in [3]. That work was also motivated by the application of clustering OCR methods to OCR applied in the compressed domain for document images compressed using token-based methods.

This paper describes a new approach to recognition by clustering that is less dependent on the parametric form of the noise model. The approach is based on modeling, using a multilayer perceptron (MLP), the probability that two given images represent the same character. These probabilities are then integrated into an overall interpretation of a document using the maximum likelihood assignment of character identities to the individual images in the maximum entropy distribution compatible with the pairwise probability estimates derived from the MLP.

This work is related to a number of other approaches in pattern recognition. Character recognition by training networks to distinguish between classes of characters has been very successful. But that approach is different from the approach described here because such discriminatory training is character and font dependent, while the method described in this paper can learn class-independent discriminatory models. Work has been performed in treating font or style consistency constraints as hidden parameters, effectively representing the global class conditional distributions as mixtures of font- and style-specific class conditional distributions[4, 5]. There has been considerable work on trying to identify noise and degradation parameters (skew, scale, etc.) that can be fit to the data present within a particular document[6]. Both of these approaches allow classifiers to take advantage of regularities in font, style, or degradation for characters found on a single page or single document, but unlike the method described in this paper, they do not provide any special advantages for generalization outside the distribution of training samples. The probabilistic clustering method used in this work is also closely related to probabilistic segmentation and grouping methods in computer vision[7].

This paper will stay mostly within the framework of optical character recognition (OCR), although the techniques are applicable to many other classification problems.

2. PREVIOUS APPROACHES

Consider the recognition of a page of text. A page of text consists of a collection of letters and digits in a number of different fonts. Each letter or digit has a character class c_i and a font f_i. Corresponding to each such character on the page is a subimage I_i that represents the character. OCR systems generally normalize the size and appearance of these subimages and perform feature extraction. Thereby, each character subimage I_i is transformed into a feature vector v_i. The task of the character recognition component of an OCR system is then to determine $P(c_i, f_i|v_i)$ (or the marginal $P(c_i|v_i)$, if font information is not relevant) and for each feature vector v_i identify the c_i and f_i that maximizes this posterior probability.

A traditional OCR system will take a direct approach to this problem. Given a large set of training examples consisting of pairs of feature vectors v_i and corresponding labels c_i, f_i, estimate a functional model $\hat{P}(c, f|v)$ and use it to predict $P(c, f|v)$ for characters appearing in new documents. To approach this modeling task, we might assume, for example, that each character class and font determine a prototype vector $v_{c,f}$ and that the actually observed feature vector is obtained by adding to this prototype a noise vector $\mathcal{N}$: $\tilde{v} = v_{c,f} + \mathcal{N}$ Within this framework, a mixture model naturally suggests itself for performing the recognition task.

As has been pointed out in the literature, such an approach ignores an important additional source of information for the OCR problem (similar sources of information exist in other domains): while there is a large set of possible classifications c_i, f_i (due to the large set of fonts), within a single document or context, only a small set of fonts occur.

For simplicity of discussion, let us assume that only a single font occurs on each page. Then, to recognize each character on a page, instead of maximizing $P(c_i, f_i|v_i)$ for each v_i on the page, we maximize $P(c_i, f|v_i)$ for a global f, a considerably more constrained problem [4, 5]. A serious limitation of such approaches is that we still need to know beforehand the set of classes and fonts that can occur and have training examples at least for a representative set of fonts.

3. RECOGNITION BY PROBABILISTIC CLUSTERING

In recognition by probabilistic clustering, rather than estimating $P(c, f|v)$, we estimate the pairwise probabilities $P(c = c', f = f'|v, v')$, i.e., the probabilities that two feature vectors represent the same character. The motivation for this approach is that we can imagine that determining whether two character images are similar or different may be considerably easier to perform in a font-independent manner than determining whether a given character image actually represents a particular character. For example, empirically, a simple but already fairly good statistic for determining the identity of two bilevel characters is to look at the minimum of the total area of their symmetric difference under arbitrary translations, normalized by the area of the larger of the two characters. This statistic can be computed completely independently of the font and distinguishes characters in a wide variety of fonts well.

If characters were perfectly distinguishable from their feature vectors, so that this probability only assumes values of 0 or 1, this would allow us to divide the set of feature vectors corresponding to characters on a page into equivalence classes. Each such equivalence class would then correspond to a single character class. Of course, we would still have to determine the identity of this equivalence class using some other means.

If $P(c = c', f = f'|v, v')$ can assume values other than zero or one, then the interpretation is more complex. An optimal interpretation of the whole document would be based on the joint conditional probability $\prod_i P(c_i, f_i|v_i)$ for all characters in the document. The conditional probability $P(c_i = c_j, f_i = f_j|v_i, v_j)$ is a marginal probability of this distribution. It is given by

$$P(c_i = c_j, f_i = f_j|v_i, v_j) =$$

$$\sum_{c,f,c_i=c_j,f_i=f_j} P(c_i, f_i|v_i)P(c_j, f_j|v_j)$$

If we have estimates for the pairwise probabilities $P(c_i = c_j, f_i = f_j|v_i, v_j)$ (e.g., from a MLP), we could try to determine the $P(c_i, f_i|v_i)$ by solving a system of equations. If N is the number of characters on the page and k the number of distinct classes c, f, then there are $\frac{N(N-1)}{2}$ estimates for $P(c_i = c_j, f_i = f_j|v_i, v_j)$ and kN unknown probabilities $P(c_i, f_i|v_i)$. For a given page, by assumption, $k \ll N$, so that we have more equations than unknowns. This would appear to let us determine the $P(c_i, f_i|v_i)$ up to a permutation of the class labels c, f.

In practice, however, since we are only using estimates of the pairwise probabilities, there is almost always no probability distribution that is consistent with the estimates for the pairwise probabilities. In order to address this problem, we need to formulate the problem of assigning classes to the different feature vectors as an optimization problem. A simple method that suggests itself is to find an assignment of class labels to feature vectors that maximizes the product of all the pairwise probability estimates (we will not attempt a formal justification in this paper). We can solve this optimization problem simply by simulated annealing, which appears to converge quickly in our experiments.

A practical issue is determining the value of k, the number of distinct classes actually present on the page. For this paper, let us assume that k is given. In general, estimating k is similar to the cluster validity problem; optimization of a cluster validity measure can be incorporated into the simulated annealing step.

4. THE METHOD

Recognition by probabilistic clustering therefore can be summarized as follows:

- estimate $P(c = c'\hat{f} = f'|v, v')$ based on a set of training examples $\{(c = c'\hat{f} = f', v, v'), ...\}$ (for many different fonts)
- when faced with the problem of recognizing a new collection of feature vectors v_i, compute $\hat{P}(c_i = c_j\hat{f}_i = f_j|v_i, v_j)$ for each pair of feature vectors v_i, v_j
- assign cluster labels χ_i to the feature vectors v_i such as to maximize $\prod \hat{P}(\chi_i = \chi_j|v_i, v_j)$, for example using simulated annealing
- determine the correspondence between the cluster labels χ_i and the actual classes (and fonts, if desired)

5. EXPERIMENTS

The dataset used in these experiments consisted of 71700 images of digits derived from 717 TrueType fonts from a commercial collection of type fonts. This dataset was split into 64600 training images representing 10 degraded samples of each digit from each of 646 fonts, and 7100 test images representing 10 degraded samples of each digit from each of 71 fonts. The character images were rendered using the Freetype[8] engine, which performed antialiased rendering of greyscale images of characters under affine transformation. Character images were degraded using the Baird character degradation model[6] with its standard settings, a widely used and studied model for modeling degradation of printed text under a variety of common document imaging conditions. Characters were rescaled to fit into a 16×16 square, giving rise to a 256 dimensional feature vector.

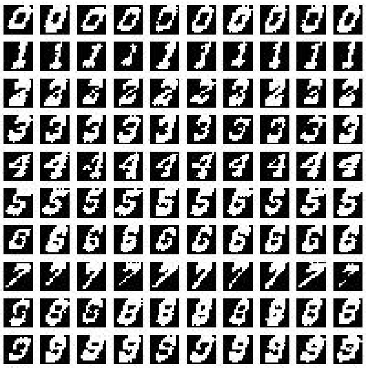

Fig. 1. Examples of degraded characters.

In a first step, to characterize the dataset, this feature vector was used as input to a multilayer perceptron The MLP had 256 input units, 15 hidden units, and 10 output units. The test set error of the MLP was 9.46%. This may appear like a high error rate for OCR, but it is important to keep in mind that this test is different from most traditional OCR tests. In particular, the test set and the training set consist of different fonts. Furthermore, the characters themselves are degraded significantly (see Figure 1). Of course, more sophisticated feature extraction methods would probably be used to improve the performance of this simple MLP-based classifier somewhat (preprocessing by PCA did not seem to make a significant difference in similar experiments).

For classification based on probabilistic clustering, the probability $P(c = c'|v, v')$ was estimated as follows. For each font in the dataset, the 4950 pairs of non-identical character images representing the same digits, as well as a random set of 4950 pairs of non-identical character images representing different digits were selected. For each pair, a translation of one image relative to the other was found that minimized the sum of the absolute differences between the digit images. At this translation, two 16×16 images were computed: the absolute difference between the two images, and the sum of the two images. These two images were used as a 512 dimensional feature vector and input into a MLP. The MLP had 512 input units, 15 hidden units, and 1 output unit. Training proceeded by training the output unit to “1” for pairs of character images representing the same digit and to “0” for pairs of character images representing different digits. It is well known that this training procedure will asymptotically converge to an estimate of the conditional probability that $c = c'$ given the input feature vector.

For testing, the input to the system consisted of 100 digit images from each font. For each pair of digit images, the pairwise probabilities $P(c = c'|v, v')$ was computed. For the simulated annealing step, a classification derived from the "traditional" classifier $\hat{P}(c|v)$ (modeled by the MLP described above) was used as the starting configuration. This biases the simulated annealing process to converge towards cluster labels that correspond directly to class labels when the pairwise probabilities are good estimates and avoids the need to adopt some other procedure for finding the permutation that brings the cluster labels into correspondence with the actual class labels. To verify that the annealing process did not merely "freeze" this initial assignment, the annealing process was also carried out with random pairwise probabilities, and an informal inspection suggests that the order of the initial assignment is quickly destroyed under the annealing schedule used.

When this procedure was carried out for the test set, the performance of the system improved from 9.46% for the traditional MLP-based classifier to 7.66% for the clustering classifier. A large fraction of the errors in both cases are due to a few "hard" fonts, for which the traditional MLP classifier misclassifies more than 25% of the samples. Furthermore, since the basis for assigning labels in the clustering classifier is the use of a starting configuration derived from the traditional MLP classifier, we cannot expect the clustering classifier to outperform the traditional MLP classifier in these cases; the experimental design limits the performance of the clustering classifier. It is therefore instructive to look at the subset of fonts that are classified more reliably by the traditional MLP classifier and evaluate the performance improvement we obtain from the clustering classifier. If we look only at fonts on which the traditional MLP classifier achieves an error rate of 25% or better (?? out of 71 fonts), then the overall error rate of the traditional MLP classifier on that subset is 5.06%, while the performance of the clustering classifier on the same set of fonts is 3.50%, a substantial improvement. Furthermore, looking at the performance on specific fonts, we find that the clustering classifier results in very substantial performance improvements (e.g., 19% error rate to 3% error rate for one font) on several fonts, while only making performance significantly worse in one case. Thus, classification by probabilistic clustering does indeed show promise in allowing a classifier system to generalize to novel or unusual fonts (of course, further experiments need to be designed to support this assertion more formally).

6. DISCUSSION

This paper describes an approach to classification based on the estimation of a class-independent probabilistic model of the similarity of two feature vectors, followed by a probabilistic clustering method. Future work will include better cluster assignment methods, a more formal analysis and better parametric models of character similarity, and automatic ways of assessing cluster validity. Perhaps most importantly, the assignment of labels to clusters by initializing the simulated annealing process is suboptimal because its performance is limited intrinsically by the quality of the traditional classifier (significantly incorrect initial assignments will result in permuted label assignments in the output). Several better methods offer themselves: use of the traditional classifier as a prior, greedy assignment of cluster labels based on predominant classifications of the members of each cluster, and the use of statistical language models. It will also be desirable to design experiments more specifically to explore and demonstrate the ability of the approach to handle variations in font, degradation, and robustness to samples outside the training set. Nevertheless, while it will be desirable to apply and evaluate the method on a much larger variety of problems, classification by clustering holds the promise of being a general approach to addressing problems that are very hard for traditional classifiers: coping with stylistic variations and generalization to samples outside the training set.

References

[1] G. Nagy, S. Sharad, K. Einspahr, and T. Meyer, "Efficient algorithms to decode substitution ciphers with applications to OCR," in *8th International Conference on Pattern Recognition*. 1986, vol. 1, pp. 352–355, IAPR.

[2] R.G. Casey, "Text OCR by Solving a Cryptogram," in *8th International Conference on Pattern Recognition*. 1986, vol. 1, pp. 349–351, IAPR.

[3] T. M. Breuel, "Modeling the Sample Distribution for Clustering OCR," in *SPIE Conference on Document Recognition and Retrieval VIII*, Jan. 2001.

[4] P. Sarkar, *Style Consistency in Pattern Fields*, Ph.D. thesis, Rensselaer Polytechnic Institute, May 2000.

[5] W. T. Freeman and J. B. Tenenbaum, "Learning bilinear models for two-factor problems in vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[6] T Kanungo, H Baird, and R Haralick, "Estimation and validation of document degradation models.," in *Proc. 4th Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, April 1995.

[7] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," 1984.

[8] "Freetype Renderer," available at http://www.freetype.org.