ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

# Ensemble Neural Classifier Design for Face Recognition

Terry Windeatt
Centre for Vision, Speech and Signal Proc (CVSSP),
School of Electronics and Physical Sciences
University of Surrey, Guildford, Surrey, United Kingdom GU2 7XH
t.windeatt@surrey.ac.uk

**Abstract.** A method for tuning MLP learning parameters in an ensemble classifier framework is presented. No validation set or cross-validation technique is required to optimize parameters for generalisability. In this paper, the technique is applied to face recognition using Error-Correcting Output Coding strategy to solve multi-class problems.

## 1 Introduction

In the past decade, the method of multiple classifiers systems (MCS) has been developed to improve classifier accuracy and efficiency. Informally, the idea in MCS is that for some complex classification problems it may be better to combine relatively simple classifiers (base classifiers) with diverse opinions rather than designing a single complex classifier. If classifiers are not too well correlated and a suitable combining rule can be found, it has been shown that simpler and more accurate systems may result. A Multi-layer perceptron (MLP) with random starting weights is a suitable base classifier since randomisation has shown to be beneficial in the MCS context. Random selection has been successfully applied to training sets (Bootstrapping), to feature sets (random subsets [1]) and to output labels [2]. Traditional MLP problems of local minima and computational slowness may be alleviated by the MCS approach of pooling together the decisions obtained from locally optimal classifiers, but there is still the problem of tuning base classifiers.

MLPs make powerful classifiers that may provide superior performance compared with other classifiers, but are often criticized for the number of free parameters. Most commonly, parameters are set with the help of either a validation set or cross-validation techniques [3]. However, there is no guarantee that a pseudo-test set is representative, and for many problems there is insufficient data to rely on this approach. Cross-validation can also be time-consuming and biased. In this paper, we present a base classifier tuning technique that has previously been extensively tested on benchmark problems and on face identification [4]. Facial images are a popular source of biometric information since they are relatively easy to acquire. However, automated face recognition systems often perform poorly and improving them is known to be a difficult task. Ensemble methods are among the best-performing solutions to achieving high face recognition rates [12]. The Error-Correcting Output Coding (ECOC) method is applied to face identification and verification in Section 4.

## 2 Two-class problems and diversity measures

For a two-class supervised learning problem, assume the label given to each pattern $X_m$ is denoted by $\omega_m = f(X_m)$ where $m = 1 \ldots \mu$ and $\omega_m \in \{0,1\}$. Here $f$ is the unknown function that maps $X_m$ to the target label $\omega_m$. It is assumed that there are $B$ parallel single hidden-layer MLP base classifiers and that $X_m$ is a $B$-dimension vector formed from the outputs of the $B$ classifiers ($\xi_{mi}$, $i = 1 \ldots B$ ) applied to the original patterns which in general are real-valued and of arbitrary dimension. Therefore, we may represent the *mth* pattern by

$$X_m = (\xi_{m1}, \xi_{m2}, \ldots, \xi_{mB}) \tag{1}$$

where $\xi \in \{x^s, x, x^d\}$, defined by

$x^s \in [0,1]$ is the soft decision in the interval

$x \in \{0,1\}$ is the hard (binary) decision formed by hardening $x^s$

$x^d \in \{0,1\}$ is the binary decision conventionally used for calculating diversity measures, where a correct classification is indicated by $x_{mi}{}^d = 1$ if and only if $x_{mi}{}^d = \omega_m$

Let the *jth* classifier output for the *pth* pattern using $x^d$ in (1) be a μ-dimensional binary vector given by $x_{pj}^d$ *where* $p = 1, \ldots \mu$.. The following counts are defined for *ith* and *jth* classifiers

$$N_{ij}^{ab} = \sum_{p=1}^{\mu} \psi_{pi}^{a} \wedge \psi_{pj}^{b} \qquad a,b\in\{0,1\}, \qquad \psi^1 = \overline{x}^d, \psi^0 = x^d \tag{2}$$

where $\wedge$ is logical AND and $\overline{x}^d$ is the logical complement of $x^d$

The Q statistic is a pair-wise diversity measure [5] that is defined by

$$Q = \frac{2}{B(B-1)} \sum_{i=1}^{B-1} \sum_{j=i+1}^{B} Q_{ij}\ , \qquad Q_{i\,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{3}$$

Note from (3), that target labels are not explicitly incorporated in defining Q. Now consider a pair-wise measure that incorporates diversity and accuracy and which is calculated over patterns between the two classes using counts as follows

$$\tilde{N}_m^{ab} = \sum_{n} \sum_{j=1}^{B} (\psi^a{}_{mj} \wedge \psi_{nj}^b), \omega_m \neq \omega_n \tag{4}$$

In (4) the *mth* pattern chosen from one class is paired with all patterns of the other class. Consider a measure $\sigma'$, interpreted in [6] as a measure of class separability (more details in [4]) and defined by

$$\sigma' = \frac{1}{\mu} \sum_{n=1}^{\mu} \sigma'_n\ , \ \sigma'_n > 0 \tag{5}$$

where $\sigma'_n = \frac{1}{\tilde{K}}\left(\frac{\tilde{N}_n^{11}}{\sum_{m=1}^{\mu}\tilde{N}_m^{11}} - \frac{\tilde{N}_n^{00}}{\sum_{m=1}^{\mu}\tilde{N}_m^{00}}\right)$, $\tilde{K} = \left(\frac{\tilde{N}_n^{11}}{\sum_{m=1}^{\mu}\tilde{N}_m^{11}} + \frac{\tilde{N}_n^{00}}{\sum_{m=1}^{\mu}\tilde{N}_m^{00}}\right)$

The relationship between ensembles and diversity measures is not well understood and the consensus is that such measures cannot predict ensemble performance [5]. However, in [4] the class separability measure defined in (5) was shown to correlate well enough with ensemble test error to predict optimal base classifier complexity.

## 3 Multi-class and Error-correcting Output Coding (ECOC)

Error-Correcting Output Coding (ECOC) is a well-established method [7] for solving multi-class problems by decomposition into complementary two-class problems. It is a two-stage process, coding followed by decoding. The coding step is defined by the binary *k x B* code word matrix Z that has one row (code word) for each of *k* classes, with each column defining one of B sub-problems that use a different labeling. Assuming each element of Z is a binary variable z, a training pattern with target class $\omega_l$ *(l = 1... k)* is re-labeled as class $\Omega_1$ if $Z_{ij} = z$ and as class $\Omega_2$ if $Z_{ij} = \bar{z}$. The two super-classes $\Omega_1$ and $\Omega_2$ represent, for each column, a different decomposition of the original problem. For example, if a column of Z is given by $[0\ 1\ 0\ 0\ 1]^T$, this would naturally be interpreted as patterns from class 2 and 5 being assigned to $\Omega_1$ with remaining patterns assigned to $\Omega_2$. This is in contrast to the conventional One-per-class (OPC) code, which can be defined by the diagonal *k x k* code matrix {$Z_{ij}$ = 1 if and only if i = j}.

In the test phase, the *jth* classifier produces an estimated probability $\hat{q}_j$ that a test pattern comes from the super-class defined by the *jth* decomposition. The *pth* test pattern is assigned to the class that is represented by the closest code word, where distance of the *pth* pattern to the *ith* code word is defined as

$$D_{pi} = \sum_{j=1}^{B} \alpha_{jl} \left| Z_{ij} - \hat{q}_{pj} \right| \qquad l = 1,...k \tag{6}$$

where $\alpha_{jl}$ allows for *lth* class and *jth* classifier to be assigned a different weight. Hamming decoding is denoted in (6) by $\{\alpha=1,\ \hat{q} \equiv x)$ and $L^1$ norm decoding by $\{\alpha=1,\ \hat{q} \equiv x^s)$ where *x* and $x^s$ are defined in (1). Many types of decoding are possible, but theoretical and experimental evidence indicates that, providing a problem-independent code is long enough and base classifier is powerful enough, performance is not much affected. In this paper, a random code is used, which is shown to perform almost as well as a pre-defined code, optimised for its error-correcting properties [8]. Issues surrounding design of optimal codes were discussed in [9].

## 4 EXPERIMENTS ON FACE DATABASES

In the first set of experiments on face identification, it is shown that the number of epochs for optimal generalization may be selected using class separability measure defined in (5). The second set of experiments applies ECOC to the problem of face verification

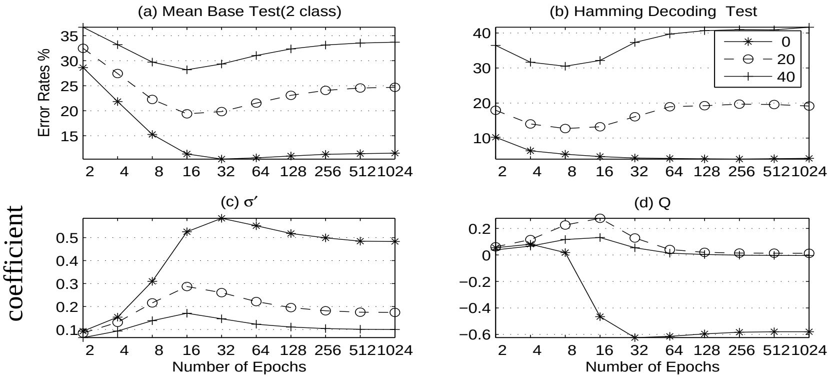

Figure 1: Test error, σ′, Q for ORL 50/50 database using 16 hidden-node base classifiers for [0,20,40] % classification noise.

**The ORL database** (Olivetti Research Laboratory http//www.cam-orl.co.uk), consists of four hundred images of forty individual faces with some variation in lighting, facial expression, facial hair, pose and spectacles. The background is controlled with subjects in an upright frontal position, although small variation in rotation and scale is allowed. The advantage of this database is that it can be used without need for face detection algorithm or any other pre-processing, so that there is a fairer comparison with the results obtained by other researchers. Although it is possible to use gray levels directly, normally features are first extracted. A popular approach is Linear Discriminant Analysis (LDA) which is used in our experiments. We compute the between-class scatter matrix, $S_B$ and the within-class scatter matrix, $S_W$. The objective of LDA is to find the transformation matrix, $W_{opt}$, that maximises the ratio of determinants $\left|w^T s_B w\right| / \left|w^T s_W w\right|$. $W_{opt}$ is known to be the solution of the following eigenvalue problem $S_B$ - $S_W \Lambda = 0$ where $\Lambda$ is a diagonal matrix whose elements are the eigenvalues of matrix $S_W^{-1} S_{B.}$ Since in practice $S_W$ is nearly always singular, dimensionality reduction is achieved by Principal Components Analysis (PCA) before solving the eigenvalue problem.

In our experiments, images have been projected to forty-dimensions using PCA and subsequently to a twenty-dimension feature space using LDA. It is treated as a forty-class face identification problem with the four hundred images randomly split into training/testing patterns.

Figure 1 shows test error rates, σ′, Q (defined in (3) and (5)) for 50/50 random train/test split with 16 hidden node base classifiers. The effect of classification noise [0 20 40] % (class label selected at random from other classes) demonstrates the ability to predict the number of epochs at which base classifier test error is minimum. The correlation of σ′ with base classifier test error is significant (ninety-five percent confidence that the correlation would not be as large as the observed value by random chance). Each combination of training epoch and noise is repeated twenty times, giving 10 x 3 x 20 runs. ECOC with random 40 x 200 code is used to solve 40-class learning problem, and 200 base classifiers are trained using Levenberg-Marquardt algorithm with default parameters.

**The extended M2VTS (XM2VTS)** database contains 295 subjects. The subjects were recorded in four separate sessions uniformly distributed over a period of 5 months, and within each session a number of shots were taken including both *frontal-view* and *rotation* sequences. Further details of this database can be found in [10]. The experimental protocol (Lausanne protocol) given in [11] provides a framework within which the performance of the XM2VTS database can be measured. The protocol assigns 200 clients and 95 impostors. Two shots of each session for each subject's frontal or near frontal images are selected to compose two configurations. We used the first configuration, in which each client has 3 training, 3 evaluation and 2 test images. The impostor set is partitioned into 25 evaluation and 70 test impostors. Within the protocol, the verification performance is measured using the false acceptance (FA) and the false rejection (FR) rates. Since no validation is required, we join training and validation sets.

The face images differ in both shape and intensity, so *shape alignment* (geometric normalisation) and *intensity correction* (photometric normalisation) can improve performance. Our approach to geometric normalisation is based on eye position, using *manually localised* eye coordinates to eliminate the dependency of the experiments on processes which may lack robustness. For photometric normalisation we have used histogram equalisation. For our experiments, images have been projected to a lower dimension feature space using PCA and LDA as described in [12], so that each image is represented by a vector with 199 elements. Each client *i* is represented by a set $X_i$ of *N* ECOC classifier output vectors, that is $X_i = \{x_i^{s(l)} \mid l = 1,2,...N\}$, where *N* is the number of *ith* client patterns available for training. In order to test the hypothesis that the client claim is authentic the average distance $d_i(\underline{x}^s)$ based on $L_1$ norm is adopted, that is

$$d_i(x^s) = \frac{1}{N}\sum_{l=1}^{N}\sum_{j=1}^{B}\left|x_j^{s(l)} - x_j^s\right| \tag{7}$$

where $x_j^s$ is the *jth* binary classifier output for the probe image and $x_j^{s(l)}$ is the *jth* classifier output for the *lth* member of class *i*. The distance is checked against a threshold, to determine if the client's claim is accepted or rejected. To find the required threshold for verification used with the distance measure defined in (7), |*FA*+*FR*| on the validation plus training set is minimised. The distances of the probe image to all images in the combined set are found and a label is assigned to the image that has minimum distance to the probe image.

A two-class MLP base classifier having one hidden layer containing 199 input nodes and 35 hidden nodes was used with ECOC. The Back-propagation algorithm with fixed learning rate, momentum and number of epochs was used for training. The error rates FA and FR were found to be 1.3% and 0.8% respectively [12], which is among the best results for XM2VTS using this protocol [13].

## 4 Conclusion

ECOC with MLPs as base classifiers has been successfully applied to problems in face identification and verification. MLPs are powerful but have the disadvantage that parameter tuning is difficult. The proposed approach in this paper enables the optimal number of training epochs of base classifier MLPs to be selected based only on performance of the training set, thereby obviating the need for validation.

[1] T. K. Ho, The Random Subspace Method for Constructing Decision Forests , IEEE Trans. PAMI, 1998, 832 - 844 [

[2] L. Breiman, Randomizing outputs to increase prediction accuracy, Machine Learning 40 (3), 2000, 229-242.

[3] L.K. Hansen, P. Salamon, Neural network ensembles, IEEE Trans. PAMI,12, 1990, 993-1001.

[4] Windeatt T., Accuracy/ Diversity and Ensemble Classifier Design, IEEE Trans Neural Networks 17(4), September, 2006

[5] L. I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles, Machine Learning 51, 2003, 181-207.

[6] T. Windeatt, Vote Counting Measures for Ensemble Classifiers, Pattern Recognition 36(12), 2003, 2743-2756.

[7] T. G. Dietterich ,G. Bakiri, Solving multiclass learning problems via error-correcting output codes, J. Artificial Intelligence Research 2, 1995, 263-286.

[8] T Windeatt and R Ghaderi., Multi-class learning and error-correcting code sensitivity, Electronics Letters 36(19), 2000, 1630-1632.

[9] T Windeatt and R Ghaderi, Coding and Decoding Strategies for multiclass learning problems, Information Fusion, 4(1), 2003, pp 11-21.

[10] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre., 1999, XM2VTS DB: The extended M2VTS database, *Proc. of AVBPA'99*, 72-77.

[11] J Luettin and G. Maitre. 1998, Evaluation Protocol For The Extended M2VTS Database (XM2VTS). Dalle Molle Institute, Switzerland,. IDIAP-Com 98-05.

[12] J. Kittler, R. Ghaderi, T. Windeatt and J. Matas Face verification via error correcting output codes, Image and Vision Computing, 21 (13-14), 2003, 1163-1169.

[13] T Windeatt, G Ardeshir, Boosted ECOC ensembles for face recognition, Proc. IEE Conf Visual Information Engineering , July 2003, Univ. of Surrey, 165-168.