Comparison of Combined Probabilistic Connectionist Models in a Forensic Application

Edmondo Trentin[1], Luca Lusnig[1], and Fabio Cavalli[2]

[1] Dip. di Ingegneria dell'Informazione, Università di Siena, Italy
[2] Research Unit of Paleoradiology and All. Sci., AOUTS Trieste, Italy
trentin@dii.unisi.it

Abstract. A growing interest toward automatic, computer-based tools has been spreading among forensic scientists and anthropologists wishing to extend the armamentarium of traditional statistical analysis and classification techniques. The combination of multiple paradigms is often required in order to fit the difficult, real-world scenarios involved in the area. The paper presents a comparison of combination techniques that exploit neural networks having a probabilistic interpretation within a Bayesian framework, either as models of class-posterior probabilities or as class-conditional density functions. Experiments are reported on a severe sex determination task relying on 1400 scout-view CT-scan images of human crania. It is shown that connectionist probability estimates yield higher accuracies than traditional statistical algorithms. Furthermore, the performance benefits from proper mixtures of neural models, and it turns up affected by the specific combination technique adopted.

Keywords: Multiple classifier, neural net, density estimation, forensics.

1 Introduction

In recent times, a growing interest toward automatic, computer-based tools has been spreading among forensic scientists and anthropologists wishing to extend the armamentarium of traditional statistical analysis and classification techniques [8]. In particular, reliable methods for the determination of the sex from human skeletal remains is of fundamental importance, for identification in forensic cases and for paleodemographic studies on ancient populations [2]. The sexual dimorphism is better recognizable in the pelvis, but (because of its complex shape) the latter is often found in very poor condition. A fundamental alternative is thus represented by the skull, which is generally better preserved and more easily reconstructed if found fragmented [7]. The paper copes with sex classification from scout-view computerized tomography (CT)-scan images of male and female human skulls, relying on 1400 images collected on the field. In particular, the goal is twofold: (i) searching for a reliable solution to the problem, applying either statistical or neural network approaches within a Bayesian framework; (ii) investigating and comparing different techniques for combining connectionist estimates of the probabilistic quantities involved in the maximum-a-posteriori classification strategy.

F. Schwenker and E. Trentin (Eds.): PSL 2011, LNAI 7081, pp. 128–137, 2012.

As reviewed in Section 2, a probabilistic interpretation of the output of a neural network can be given in terms of a supervised, discriminative posterior-probability setup, or in terms of unsupervised class-conditional density estimation. While the former is the traditional practitioner's choice in pattern recognition applications of neural networks, the latter is far less investigated in the literature, mostly due to the intrinsic difficulties which arise in dealing with the unsupervised estimation task. Nonetheless, robust class-conditional density estimates can be used *per se* within Bayes theorem as viable classification tools. Moreover, they can capture and convey relevant information that can be combined with the class-posterior estimates in order to improve the performance of the overall multiple-classifier system. To this end, we rely on a neural network approach to the density estimation task that we proposed in [10] (reviewed in Section 2, as well). Note that in [10] the experimental evaluation of the model was carried out on illustrative, univariate synthetic datasets generated with probability density functions (pdf) having known form. Therefore, an additional aim of this paper is the evaluation of the approach in a multivariate, real-world task.

The combination techniques evaluated in the paper are presented in Section 3. They rely on two common, somewhat complementary notions. First, having models of probabilistic quantities may ease the definition of meaningful combination schemes that benefit from the homogeneous nature of the underlying classifiers (possibly, turning themselves out to undergo a plausible interpretation in terms of probabilities). Second, on the other way around, posterior probability models and individual class-conditional density functions are the carrier of non-completely overlapping information, providing the combination algorithm with the opportunity to perform better than the separate models actually do. Sex determination experiments on an original, real-scale dataset are reported in Section 4. Some conclusions, relevant to the machine learning as well as to the anthropology/forensic sciences communities are drawn in Section 5.

2 Probabilistic Interpretation of Neural Networks

Artificial neural networks (ANNs) [4,1] have been widely applied to pattern classification tasks [1]. In most cases, their application takes the form of a connectionist discriminant function, which is trained to yield a high "score" on the correct class, along with low scores on all the wrong classes. No probabilistic interpretation of such a discriminant function is usually given, neither it is even expected. As a matter of fact, minimum classification error is gained when a maximum class-posterior probability is chosen as a discriminant within a Bayesian framework [3]. This is accomplished relying on the popular Bayes theorem[3], i.e. $P(\omega_i \mid \mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)/p(\mathbf{x})$, where $\mathbf{x}$ is a pattern (real-valued feature vector) to be assigned to one out of c distinct and disjoint classes $\omega_1, \ldots, \omega_c$. The theorem transforms a prior knowledge on the probability of individual classes, i.e. the prior probability $P(\omega_i)$, into a posterior knowledge upon observation of a certain feature vector $\mathbf{x}$, namely the posterior probability

$P(\omega_i \mid \mathbf{x})$. Such a transformation relies on the evaluation of the so-called class-conditional pdf $p(\mathbf{x} \mid \omega_i)$. Theorems confirm that, under rather mild conditions, ANNs can be trained as optimal estimates of Bayes posterior probabilities [1]. These theorems give a mathematical foundation to the popular heuristic decision rules that we mentioned at the beginning of this section. Roughly speaking, it can be shown that a multi-layer perceptron (MLP) [4] having c output units and trained via regular backpropagation (BP) [4] over a labeled training set $\mathcal{T} = \{(\mathbf{x}_k, \mathbf{y}_k) \mid k = 1, \ldots, n\}$ where $\mathbf{y}_k = (y_{k1}, \ldots, y_{kc})$ and $y_{ki} = \begin{cases} 1 \text{ if } \mathbf{x}_k \in \omega_i \\ 0 \text{ otherwise} \end{cases}$ is an "optimal" non-parametric estimation of the left-hand-side of Bayes theorem. In practice, it is not necessary to know the class-posterior probabilities in advance in oder to create target outputs for the BP training, since a crisp 0/1 labeling (which reminds us of the good, old Widrow-Hoff labeling for linear discriminant [3]) drives the ANN weights to convergence towards the same result. Since a probabilistic interpretation of the MLP outputs is sought, some constraints are required. First, output values are limited to the $(0, 1)$ range. This is readily accomplished by relying on the usual sigmoid activation functions. Then, since $\sum_{i=1}^{c} P(\omega_i \mid \mathbf{x}) = 1$, a normalization of the MLP outputs is needed.

Whilst estimation of posterior probabilities via ANNs is feasible due to the simplicity of satisfying the probability constraints, connectionist estimation of pdfs–i.e., class-conditional pdfs to be used in the right-hand-side of Bayes Theorem–is much harder, since: (i) a pdf may possibly take any non-negative, unbounded value; (ii) its integral over the feature space shall equal 1; (iii) above all, pdf estimation is an intrinsically unsupervised learning problem, and standard training algorithms do not do. Yet, due to their flexibility and generalization capabilities, neural models of pdfs could improve over parametric and non-parametric statistical estimation techniques. In [10] we proposed a connectionist model for density estimation which overcomes the major limitation of statistical techniques. A concise review of the approach follows. Let us consider a pdf $p(\mathbf{x})$, defined over a real-valued, d-dimensional feature space. The model is introduced along the usual line followed in the traditional kernel-based nonparametric pdf estimates, such as the Parzen window (PW) [3]. These techniques are built on the observation that the probability that a pattern $\mathbf{x}' \in \mathcal{R}^d$, drawn from $p(\mathbf{x})$, falls in a certain region R of the feature space is $P = \int_R p(\mathbf{x}) d\mathbf{x}$. Let then $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be an unsupervised sample of n patterns, identically and independently distributed (i.i.d.) according to $p(\mathbf{x})$. If k_n patterns in $\mathcal{T}$ fall within R, an empirical estimate of P can be obtained as $P \simeq k_n/n$. If $p(\mathbf{x})$ is continuous and R is small enough to prevent $p(\mathbf{x})$ from varying its value over R in a significant manner, we are also allowed to write $\int_R p(\mathbf{x}) d\mathbf{x} \simeq p(\mathbf{x}')V$, where $\mathbf{x}' \in R$, and V is the volume of region R. An estimated value of the pdf $p(\mathbf{x})$ over pattern $\mathbf{x}'$ is thus given by:

$$p(\mathbf{x}') \simeq \frac{k_n/n}{V_n} \tag{1}$$

where V_n denotes the volume of region R_n, assuming that smaller regions around $\mathbf{x}'$ are considered as the sample size n increases. This is expected to allow

equation (1) to yield improved estimates of $p(\mathbf{x})$, i.e. to converge to the exact value of $p(\mathbf{x}')$ as n (hence, also k_n) tends to infinity (a discussion of the asymptotic behavior of nonparametric models of this kind can be found in [3]). The basic instance of the PW technique assumes that R_n is a hypercube having edge h_n, such that $V_n = h_n^d$. The edge h_n is usually defined as a function of n as $h_n = h_1/\sqrt{n}$, in order to ensure a correct asymptotic behavior. The value h_1 has to be chosen empirically, and it heavily affects the resulting model. The formalization of the idea requires to define a unit-hypercube window function in the form $\varphi(\mathbf{y}) = \begin{cases} 1 & \text{if } \mid y_j \mid \leq 1/2, j = 1, \ldots, d \\ 0 & \text{otherwise} \end{cases}$, such that $\varphi(\frac{\mathbf{x}'-\mathbf{x}}{h_n})$ has value 1 iff $\mathbf{x}'$ falls within the d-dimensional hyper-cubic region R_n centered in $\mathbf{x}$ and having edge h_n. This implies that $k_n = \sum_{i=1}^{n} \varphi(\frac{\mathbf{x}'-\mathbf{x}_i}{h_n})$. Using this expression, from equation (1) we can write

$$p(\mathbf{x}') \simeq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_n} \varphi(\frac{\mathbf{x}' - \mathbf{x}_i}{h_n}) \tag{2}$$

which is the PW estimate of $p(\mathbf{x}')$ from the sample $\mathcal{T}$. The model is usually refined by considering smoother window functions $\varphi(.)$, instead of hypercubes. The idea for training a MLP to estimate $p(\mathbf{x})$ from $\mathcal{T}$ is to use the PW model as a target output for the ANN, and to apply standard BP to the MLP. A unbiased variant of this idea is proposed, according to the following unsupervised algorithm (expressed in pseudo-code):

Input: $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, h_1.
Output: $\tilde{p}(.)$ /* the connectionist estimate of $p(.)$ */

1. Let $h_n = h_1/\sqrt{n}$
2. Let $V_n = h_n^d$
3. For i=1 to n do /* loop over $\mathcal{T}$ */
3.1 Let $\mathcal{T}_i = \mathcal{T} \setminus \{\mathbf{x}_i\}$
3.2 Let $y_i = \frac{1}{n-1} \sum_{\mathbf{x} \in \mathcal{T}_i} \frac{1}{V_{n-1}} \varphi(\frac{\mathbf{x}_i - \mathbf{x}}{h_{n-1}})$ /* target output */
4. Let $\mathcal{S} = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, n\}$ /* supervised training set */
5. Train the ANN via BP over $\mathcal{S}$
6. Let $\tilde{p}(.)$ be the function computed by the ANN
7. Return $\tilde{p}(.)$

Since the ANN output is assumed to be an estimate of a pdf, it must be non-negative, yet unbounded. For this reason, sigmoids with adaptive amplitude λ (i.e., in the form $y = \frac{\lambda}{1+e^{-x}}$), as described in [9], are used as output activation functions. As in several statistical nonparametric models, such as the k_n-nearest neighbor technique [3], the ANN is not necessarily a pdf (in general, the integral of $\tilde{p}(.)$ over the feature space is not 1), but a good (i.e., useful) approximation of the desired density is obtained, overcoming the limitations of traditional estimation methods [10]. We refer to this model as the Parzen-ANN (P-ANN). In this

paper, the P-ANN is applied to the estimation of the class-conditional density functions $p(\mathbf{x} \mid \omega_i)$ to be used within Bayes theorem. This means that individual, class-specific networks have to be trained over the data belonging to the corresponding class. Standard Gaussian kernels will be applied in the experiments (step 3.2 of the algorithm).

3 Combination Techniques

The probabilistic interpretation of different neural models provides us with a number of simple yet well-grounded combination techniques for a multiple classifier system. For notational convenience, for each class $i = 1, \ldots, c$ we write $\hat{P}(\omega_i \mid \mathbf{x})$ to denote the posterior estimate of $P(\omega_i \mid \mathbf{x})$ yielded by the i-th output of the supervised MLP, and $\tilde{P}(\omega_i \mid \mathbf{x})$ to refer to the quantity $\tilde{p}(\mathbf{x} \mid \omega_i)P(\omega_i)/\tilde{p}(\mathbf{x})$, where $\tilde{p}(\mathbf{x} \mid \omega_i)$ is the P-ANN for the class-conditional $p(\mathbf{x} \mid \omega_i)$ and $\tilde{p}(\mathbf{x})$, the estimate of the evidence $p(\mathbf{x})$, is obtained as $\sum_{j=1}^{c} P(\omega_j)\tilde{p}(\mathbf{x} \mid \omega_j)$, as usual. Plausible combination techniques may be defined as follows.

1. *Pseudo-joint probability*: let ξ_1 and ξ_2 be the random quantities yielded by two distinct functions (or, regression models) of a given random vector $\mathbf{x} \in \Re^d$. We refer to ξ_1 and ξ_2 as the "models", and the following discussion can be extended straightforwardly to an arbitrary number of models. For any generic state of nature ω_i, $i = 1, \ldots, c$, we can write:

$$P(\omega_i \mid \xi_1, \xi_2) = \frac{p(\xi_1, \xi_2 \mid \omega_i)P(\omega_i)}{p(\xi_1, \xi_2)} = \frac{p(\xi_1 \mid \omega_i)p(\xi_2 \mid \xi_1, \omega_i)P(\omega_i)}{p(\xi_1, \xi_2)}. \tag{3}$$

Under the assumption that the models are independent of each other, equation (3) can be rewritten as follows:

$$\begin{aligned} P(\omega_i \mid \xi_1, \xi_2) &= \frac{p(\xi_1 \mid \omega_i)p(\xi_2 \mid \omega_i)P(\omega_i)}{p(\xi_1)p(\xi_2)} \\ &= \frac{P(\omega_i \mid \xi_1)p(\xi_1)}{p(\xi_1)P(\omega_i)} \frac{P(\omega_i \mid \xi_2)p(\xi_2)}{p(\xi_2)P(\omega_i)} P(\omega_i) \\ &= \frac{P(\omega_i \mid \xi_1)P(\omega_i \mid \xi_2)}{P(\omega_i)} \end{aligned} \tag{4}$$

which has the form of a pseudo-joint probability (the product of quantities at the numerator) normalized by the class-prior. The use of the expression "pseudo" is enforced by the observation that in real-world scenarios the models are hardly independent, yet equation (4) can still be fruitfully applied in a naive-Bayes fashion. If the classes are equally alike a priori (as in the experiments reported in the paper), i.e. if $P(\omega_i) = P(\omega_j)$ for each $i, j \in$

$\{1, \ldots, c\}$, then a discriminant function $g_i(.)$ can be defined for each class ω_i by taking the usual maximum-a-posteriori probability given the models, i.e. $\max_i P(\omega_i \mid \xi_1, \xi_2)$, and dropping the denominator from Eq. (4). In so doing, discriminant functions are defined as pseudo-joint probabilities in the form $g_i(\mathbf{x}) = P(\omega_i \mid \xi_1(\mathbf{x}))P(\omega_i \mid \xi_2(\mathbf{x}))$, and the corresponding decision rule assigns a pattern $\mathbf{x}$ to class i if $g_i(\mathbf{x}) \geq g_j(\mathbf{x})$ for each $j \neq i$, as usual. In the experiments we assume that $\xi_1(.)$ is the supervised MLP and $\xi_2(.)$ is realized via P-ANN (and Bayes theorem), and we let $P(\omega_i \mid \xi_1(\mathbf{x})) \approx \hat{P}(\omega_i \mid \mathbf{x})$ and $P(\omega_i \mid \xi_2(\mathbf{x})) \approx \tilde{P}(\omega_i \mid \mathbf{x})$, according to the notation above.

2. *Maximum confidence*: when we assign a pattern $\mathbf{x}$ to class ω_i according to the maximum-a-posteriori criterion, i.e. $i = argmax_j P(\omega_j \mid \mathbf{x})$, we face a certain Bayesian risk, namely the probability of misclassification given the pattern. The latter can be written as $P(error \mid \mathbf{x}) = \sum_{j=1, j\neq i}^{c} P(\omega_j \mid \mathbf{x})$. It is seen that the higher the posterior probability of ω_i, the lower the probability of error. In the present setup, a minimum-risk combination strategy for the two connectionist models follows in a natural manner: if the neural networks agree on the decision of assigning pattern $\mathbf{x}$ to ω_i, just do it. Otherwise, if $\hat{P}(\omega_i \mid \mathbf{x}) \geq \hat{P}(\omega_j \mid \mathbf{x})$ for all $j \neq i$ and $\tilde{P}(\omega_k \mid \mathbf{x}) \geq \tilde{P}(\omega_j \mid \mathbf{x})$ for all $j \neq k$, then the decision $d(\mathbf{x})$ between ω_i and ω_k is taken as:

$$d(\mathbf{x}) = \begin{cases} \omega_i \text{ if } \hat{P}(error \mid \mathbf{x}) \geq \tilde{P}(error \mid \mathbf{x}) \\ \omega_k \text{ otherwise} \end{cases} \tag{5}$$

where $\hat{P}(error \mid \mathbf{x}) = \sum_{j=1, j\neq i}^{c} \hat{P}(\omega_j \mid \mathbf{x})$ and $\tilde{P}(error \mid \mathbf{x}) = \sum_{j=1, j\neq k}^{c} \tilde{P}(\omega_j \mid \mathbf{x})$. In other words, the classification relies eventually on the model which exhibits the highest confidence in its own decision.

3. *Minimum expectation*: albeit appealing, the combination based on maximum confidence has a major drawback. In fact, a rough model of the Bayesian posterior probability turns implicitly out to be a rough estimator of its own Bayesian risk, as well (e.g., by over-estimating the class-posterior over a certain pattern, resulting in an under-estimate of the corresponding probability of error). This may suggest taking a somewhat complementary approach, discarding the (overwhelmingly optimistic) maximum-confidence decision and opting for the (possibly more realistic) minimum expectation strategy. In this framework, the latter takes the following form: if the two models are in disagreement, say $d(\mathbf{x}) = \omega_i$ based on $\hat{P}(\omega_i \mid \mathbf{x})$ and $d(\mathbf{x}) = \omega_k$ based on $\tilde{P}(\omega_k \mid \mathbf{x})$, then assign $\mathbf{x}$ to ω_i if $\hat{P}(\omega_i \mid \mathbf{x}) \leq \tilde{P}(\omega_k \mid \mathbf{x})$, else assign $\mathbf{x}$ to ω_k. Albeit heuristic, this conservative strategy reveals to be backed up by empirical evidence.

4. *Average*: a natural, simple alternative is represented by the average between the two estimates, namely taking $P(\omega_i \mid \mathbf{x}) \approx \frac{1}{2}\hat{P}(\omega_i \mid \mathbf{x}) + \frac{1}{2}\tilde{P}(\omega_i \mid \mathbf{x})$ for each $i = 1, \ldots, c$. The straightforward extension of the technique relies on a *weighted average* over the models in the form $P(\omega_i \mid \mathbf{x}) \approx \alpha\hat{P}(\omega_i \mid \mathbf{x}) + (1 - \alpha)\tilde{P}(\omega_i \mid \mathbf{x})$ for each class, where the relative weight $\alpha \in (0, 1)$ can be determined empirically via model selection techniques, contributing to compensate for possible biases and/or numeric mismatches between the two models.

5. *Rejection on* ξ_ι: a variation on the theme of the maximum confidence, which outsprings from the same background reasoning and from a long-standing tradition in practical development of classifiers which include the reject option (i.e. reject current pattern $\mathbf{x}$, refusing to take any decision, whenever the estimated value of the discriminant functions $g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})$ are all below a given rejection threshold θ, with θ in the $(0, 1)$ interval) can be introduced as follows. Let $\xi_1(\mathbf{x}) = \hat{P}(\omega_i \mid \mathbf{x})$ and $\xi_2(\mathbf{x}) = \tilde{P}(\omega_k \mid \mathbf{x})$, where ω_i and ω_k are the decisions taken by models ξ_1 and ξ_2 over $\mathbf{x}$, respectively. We say that a rejection on ξ_1 decision strategy assigns $\mathbf{x}$ to ω_i if $\xi_1(\mathbf{x}) \geq \theta$, and to ω_k otherwise (regardless of the value of $\xi_2(\mathbf{x})$). On the other way around, the rejection on ξ_2 assigns by default to ω_k, unless $\xi_2(\mathbf{x}) < \theta$ (in the latter case $\mathbf{x}$ is assigned to ω_i). It is seen that these decision rules do not coincide with the maximum-confidence approach. Suitable values for θ are found empirically, within a proper model selection framework.
6. *Mixture of experts*: in principle, the most flexible combination technique simply avoids arbitrary choices on the explicit combination strategy, and lets the machine learn its own "optimal" recipe from examples. A straightforward, yet sound realization of this principle relies on a committee of neural experts [4]. In the present setup we consider a third MLP which, for each pattern $\mathbf{x}$, is fed with the estimates $\xi_1(\mathbf{x})$ and $\xi_2(\mathbf{x})$ and is expected to yield in output a more robust estimate of $P(\omega_i \mid \mathbf{x})$. We refer to this third connectionist module as the gating network. More precisely, in a c-class problem $\xi_1(\mathbf{x})$ has c output units, forming an input vector $(\hat{P}(\omega_1 \mid \mathbf{x}), \ldots, \hat{P}(\omega_c \mid \mathbf{x}))$ while $\xi_2(\mathbf{x})$ is better described as an ordered collection of c separate P-ANNs, say $(\tilde{p}(\mathbf{x} \mid \omega_1), \ldots, \tilde{p}(\mathbf{x} \mid \omega_c))$. The aggregate vector $(\hat{P}(\omega_1 \mid \mathbf{x}), \ldots, \hat{P}(\omega_c \mid \mathbf{x}), \tilde{p}(\mathbf{x} \mid \omega_1), \ldots, \tilde{p}(\mathbf{x} \mid \omega_c))$ defines the input space for the gating network, whose target output is the usual, Widrow-Hoff-like binary coding (0/1) of the correct class whom the current training pattern belongs to. In so doing, as remarked in Section 2, the gating network approximates the Bayesian class-posterior probability, learning the combination law of its inputs which best fits its training criterion. To practical ends, $\xi_1(\mathbf{x})$ and $\xi_2(\mathbf{x})$ are separately trained first, as usual. Later on, the gating network is trained (with regular BP) on the outputs yielded by $\xi_1(\mathbf{x})$ and $\xi_2(\mathbf{x})$ over the original training data.

4 Experiments

For this study, a total of 1400 scout-view CT scanogram (of healthy, adult, Caucasian subjects) were selected at random from our PACS database, including 700 males and 700 females within an age range of 25–92. The scanogram was chosen because it is routinely performed before a cranial CT examination, and since for our purposes (i.e., the determination of the external shape of calvarium in norma lateralis) it is basically as reliable as the cephalometric lateral radiograph. The patients were chosen on the basis of their residence in the province of Trieste (Italy), since the population of this geographic area is the result of

complex historical genetic crossover between Italic, Germanic and Slavic populations. Lateral cranial scanograms were automatically selected and anonymized by our PACS facilities (registering only the sex and the age) among the CT examinations performed between the years 2005 and 2010 in the radiological structures of the Department of Diagnostic Imaging of the Hospital Corporation at the University of Trieste with similar multislice computed tomography (MSCT) equipment. Lateral CT scanograms were taken on an Aquilion 16 Toshiba multislice CT scanner, using the standard preset (120 kVp, 150 mAs, matrix size 512x512). The images were automatically transformed from DICOM to JPG format, maintaining the original matrix size.

Visual feature extraction from the images underwent the following procedure. A smoothing Gaussian filter (with discrete Weierstrass transform relying on a 5×5 convolution matrix) is applied first [6], in order to reduce additive noise. It is followed by a sharpening filter. Starting from the filtered image, edge detection and edge connection are accomplished by a technique relying on Canny algorithm, followed by thresholding. Upon removal of the maxilla and mandible area, the contour of the *cranium* is extracted automatically (including the *glabella*, *calvarium*, and *opisthion* areas). The centroid-distance signature function is then extracted [12], ensuring translation-invariance. In order to reduce the dimensionality significantly, and to resort to a fixed-dimensionality representation, sub-sampling of the overall set of signatures is accomplished via the equal points sampling technique. Features are finally extracted from the sub-sampled signatures by application of the usual fast Fourier transform (FFT), retaining the first 64 parameters. This results in a 64-dim feature space which ensures rotation invariance and scale invariance (by proper normalization of the magnitude of the first half of the FFT coefficients), as described in [12].

The data were split first into a training and a validation set, for model selection purposes. Once the selection process was completed and upon replacement of the original data, the patterns were then randomly partitioned again into a training set (1000 patterns), and a test set (400 patterns), having an equal balance between the relative frequencies of male and female samples. Results are reported in Table 1 (the same notation used in the previous section is used to refer to the specific models). Linear discriminant analysis was applied first (applying the pseudo-inverse method based on singular value decomposition), in order to fix a baseline. The results confirm the high-nonlinearity of the classification task. A more significant baseline was yielded by a regular k-nearest neighbor (k-NN) classifier with $k = 5$. The performance turnt out to be improved by the PW approach. Standard Gaussian kernels were used, with initial width $h_1 = 9.77 \times 10^{-2}$. Connectionist approaches follow, starting from the individual classifiers relying on unsupervised estimation of $p(\mathbf{x} \mid \omega_i)$ (15 hidden sigmoid units for the "male" class, and 16 such units for the "female" class; sigmoid activation in the output unit, all activation functions having a smoothness set to 0.4 and layer-by-layer adaptive amplitude). Training these P-ANNs required 300 epochs only, with learning rates $\eta = 0.1$. The next row of the table shows the results yielded by the supervised estimation of $P(\omega_i \mid \mathbf{x})$ via MLP. The latter has 16 hidden sigmoid units and a sigmoid

Table 1. Sex recognition rate using the cranium contour

Model	Accuracy (%)
Linear discriminant	53.80
k-NN	68.25
Parzen Window	70.75
$\tilde{P}(\omega_i \mid \mathbf{x})$	79.25
$\hat{P}(\omega_i \mid \mathbf{x})$	80.25
Pseudo-joint probability	82.00
Maximum confidence	81.50
Minimum expectation	81.25
Average	82.00
Weighted average	82.75
Rejection on ξ_1	83.00
Rejection on ξ_2	81.75
Mixture of experts	83.50

output, all having smoothness 1.25. 20000 epochs of BP with learning rate $\eta = 0.1$ were applied. Accuracies turn out to outperform the statistical techniques. The P-ANN performance is even surprisingly higher than the traditional PW, and close to the supervised, discriminative MLP. The combination techniques proposed in Section 3 are reported in the next rows of the table. The mixture of experts relies on a gating MLP with 9-hidden sigmoid units and a sigmoid output (all smoothnesses set to 1), and neuron-by-neuron adaptive amplitudes. Training required 150 epochs with a learning rate set to 0.02. It is seen that all the combination methods are effective (although, with a certain variance in terms of relative performance), showing that the difference in the information conveyed by the connectionist models involved are complementary to some extent and can be exploited jointly in order to come up with a more robust classifier. Letting the machine discover the most suitable combination law (relying on the committee machine) yields higher recognition rates than fixed (albeit plausible) mixing choices. In the best case scenario (i.e., mixture of experts) a relative error rate reduction of 16.46% is gained w.r.t the best single-model classifier. Results are of the utmost significance in an application oriented perspective, if compared with the expected recognition rate ($\sim 80\%$) by human experts [11], as well as similar classification experiments carried out using statistical approaches in the forensic sciences [5].

5 Conclusions

The paper faced a difficult, real-world classification task having the utmost relevance to anthropology and the forensic sciences, namely sex determination from CT-scan images of human skulls. Experiments were accomplished over an

original, large-scale dataset collected on the field and involving 1400 patients. Statistical and connectionist approaches were considered. In particular, neural networks having a probabilistic interpretation of their outputs were reviewed. The two paradigms can be mixed in a variety of natural, sound ways on the basis of the probabilistic meaning of their outputs. Several combination techniques were considered and compared on the field. Results are noticeable in an application perspective, turning out to be higher than the expected correctness of prediction by human experts, as well as w.r.t. statistical approaches previously investigated in the literature on forensic sciences. In particular, combination based on committee machines do particularly fit the task. Finally, P-ANNs proved themselves to be more effective than traditional statistical techniques over the multivariate density estimation task at hand.

References

1. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
2. Brasili, P., Toselli, S., Facchini, F.: Methodological aspects of the diagnosis of sex based on cranial metric traits. Homo. 51, 68–80 (2000)
3. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
4. Haykin, S.: Neural Networks. A Comprehensive Foundation. Macmillan, New York (1994)
5. Hsiao, T.H., Chang, H.P., Liu, K.M.: Sex determination by discriminant function analysis of lateral radiographic cephalometry. Journal of Forensic Sciences 41(5), 792 (1996)
6. Nixon, M., Aguado, A.S.: Feature Extraction & Image Processing, 2nd edn. Academic Press (2008)
7. Novotny, V., Iscan, M., Loth, S.: Morphologic and osteometric assessment of age, sex, and race from the skull. In: Iscan, M.Y., Helmer, R.P. (eds.) Forensic Analysis of the Skull, pp. 71–88. Wiley-Liss, New York (1993)
8. Rsing, F.W., Graw, M., Marr, B., Ritz-Timme, S., Rothschild, M.A., Rzscher, K., Schmeling, A., Schrder, I., Geserick, G.: Recommendations for the forensic diagnosis of sex and age from skeletons. HOMO - Journal of Comparative Human Biology 58(1), 75–89 (2007)
9. Trentin, E.: Networks with trainable amplitude of activation functions. Neural Networks 14(4-5), 471–493 (2001)
10. Trentin, E.: Simple and Effective Connectionist Nonparametric Estimation of Probability Density Functions. In: Schwenker, F., Marinai, S. (eds.) ANNPR 2006. LNCS (LNAI), vol. 4087, pp. 1–10. Springer, Heidelberg (2006)
11. Walrath, D.E., Turner, P., Bruzek, J.: Reliability test of the visual assessment of cranial traits for sex determination. American Journal of Physical Anthropology 125(2), 132–137 (2004)
12. Zhang, D., Lu, G.: A Comparative Study on Shape Retrieval Using Fourier Descriptors with Different Shape Signatures. Journal of Visual Communication and Image Representation 14(1), 41–60 (2003)