RECOGNITION OF HANDWRITTEN NUMERALS BY STRUCTURAL PROBABILISTIC NEURAL NETWORKS

Jiří Grim, Pavel Pudil, Petr Somol
Institute of Information Theory and Automation
P.O. BOX 18, CZ-18208 Prague, Czech Republic
E-mail: grim@utia.cas.cz, pudil@utia.cas.cz, somol@utia.cas.cz

31. prosince 2008

Abstract. *The well known "beauty defect" of probabilistic neural networks is the biologically unnatural complete interconnection of neurons with all input variables. Despite of deep formal reasons of this undesirable property, it can be removed by a special subspace approach without leaving the exact framework of Bayesian decision-making. As shown in a recent paper the related structural optimization based on EM algorithm is controlled by an information criterion. In the present paper the method has been applied to recognize unconstrained handwritten numerals from the database of Concordia University, Montreal, Canada. The obtained recognition accuracy is comparable with the previously published results though it has been achieved without any preceding feature extraction.*

Keywords: *Probabilistic neural networks, Statistical decision-making, Finite mixtures, EM algorithm, Structural optimization, Recognition of numerals.*

1 INTRODUCTION

The probabilistic approach to neural networks naturally evolves from the general framework of statistical classification. The basic idea of probabilistic neural networks (PNN) is to approximate the class-conditional probability distributions by means of a kernel estimate (cf. [38]) or by a distribution mixture (cf. [8], [9], [13], [31], [39]) whereby the components of mixtures or kernels correspond to formal neurons.

There is a similarity between PNN and the radial basis functions (RBF's) neural networks (cf. e.g. [13], [34]). However, the RBF's are usually optimized for the sake of a multivariate interpolation or approximation of some output variables whereas, on the other hand, the purpose of estimating distribution mixtures is the Bayesian classification of observations. Also the simplifying assumption of radial symmetry is not necessary in case of mixture components since the EM algorithm (cf. [37], [3], [41], [4]) as an optimization tool is usually applicable in full generality.

A week point of the probabilistic approach to neural networks is the tacitly assumed complete interconnection of component distributions (neurons) with all input variables. This property follows from the fundamental fact that all component distributions of a mixture must be defined on the same space and therefore they have to depend on the same set of variables.

Recently a new approach to structural optimization of probabilistic neural networks has been proposed [11] making use of an idea originally designed for multivariate pattern recognition. It is based on finite mixtures including binary structural parameters. By means of a special "background" substitution technique the evaluation of components can be confined to "relevant" subspaces only. In this way the receptive fields of neurons can be reduced to arbitrary subsets of input variables. The optimal choice of input variables is controlled by an information criterion.

In the present paper we apply the method of structural optimization to recognition of totally unconstrained handwritten numerals from the database of Concordia University in Montreal. The problem was solved in the original space of non-reduced dimension N=1024 (binary 32x32 raster). Unlike similar published solutions we didn't use any prece-

[0]Early version of the paper: Grim J., Pudil P., Somol P. (2000) "Recognition of handwritten numerals by structural probabilistic neural networks". In: Proc. of the Second ICSC Symposium on Neural Computation, Berlin, 2000. Bothe H., Rojas R. (Eds.). ICSC, Wetaskiwin, pp. 528-534.

ding feature extraction or feature selection which may essentially improve the classification accuracy. In computational experiments with randomly initialized mixture models we obtained repeatedly recognition accuracy which is comparable with the results reported in literature.

2 RECOGNITION BASED ON MIXTURES

Let us suppose that some observations

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}$$

from an N-dimensional discrete space $\mathcal{X}$ are to be classified into one of a finite set of mutually exclusive classes

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}.$$

Considering a statistical problem of pattern recognition we assume that the random occurrence of observations $\boldsymbol{x} \in \mathcal{X}$ is characterized by class-conditional probability distributions $P(\boldsymbol{x}|\omega)$ and by the related a priori probabilities $p(\omega), \omega \in \Omega$. All statistical information about the set of classes Ω, given some observation $\boldsymbol{x} \in \mathcal{X}$, is expressed by the Bayes formula for a posteriori probabilities

$$p(\omega|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega)p(\omega)}{P(\boldsymbol{x})} \quad \omega \in \Omega \tag{1}$$

where

$$P(\boldsymbol{x}) = \sum_{\omega \in \Omega} P(\boldsymbol{x}|\omega)p(\omega) \tag{2}$$

is the unconditional probability distribution of $\boldsymbol{x}$. The posterior distribution $p(\omega|\boldsymbol{x})$ may be used to define a unique final decision or to evaluate some more complex decisions including e.g. a loss function.

In view of the Bayes formula (1) the decision problem can be solved by estimating the unknown probabilistic description of classes. In the present paper we assume that the conditional distributions $P(\boldsymbol{x}|\omega)$ can be approximated by finite mixtures of the form

$$P(\boldsymbol{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} F(\boldsymbol{x}|m, \omega) f(m|\omega), \tag{3}$$

$$\boldsymbol{x} \in \mathcal{X}, \quad \sum_{m \in \mathcal{M}_\omega} f(m|\omega) = 1, \quad \omega \in \Omega$$

where $f(m|\omega) \geq 0$ are some conditional probabilistic weights, $F(\boldsymbol{x}|m, \omega)$ the component distributions and $\mathcal{M}_\omega$ the index set. For the sake of a simple notation we introduce consecutive indexing of components. We denote $\mathcal{M}_{\omega_k}$ the index set of the class $\omega_k \in \Omega$:

$$\mathcal{M}_{\omega_k} = \{M_{\omega_{k-1}} + 1, M_{\omega_{k-1}} + 2, \ldots, M_{\omega_k}\}, \tag{4}$$

$$M_{\omega_{k-1}} < M_{\omega_k}, \quad M_{\omega_0} = 0, \quad k = 1, 2, \ldots, K,$$

i.e. the number of components of the mixture $P(\boldsymbol{x}|\omega_k)$ is $|\mathcal{M}_{\omega_k}| = (M_{\omega_k} - M_{\omega_{k-1}})$. In this way the component index m uniquely identifies the class $\omega \in \Omega$ and therefore the parameter ω can be partly omitted in Eq. (3), i.e. we can write

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} F(\boldsymbol{x}|m) f(m), \quad f(m) = f(m|\omega)p(\omega). \tag{5}$$

As already mentioned in Introduction, the basic idea of PNN is to view the component distributions in Eq. (5) as formal neurons. In other words, the output of the m-th neuron as a function of $\boldsymbol{x}$ is defined by the component $F(\boldsymbol{x}|m)$. Consequently, for each $\omega \in \Omega$ the a posteriori probability $p(\omega|\boldsymbol{x})$ is proportional to a weighted sum of output variables of neurons from $\mathcal{M}_\omega$ (cf. (1), (5)).

An important feature of PNN is the possibility to optimize the multivariate components by means of EM algorithm (cf. e.g. [8], [9]). As it will be shown in the following, the EM algorithm can be modified to include the structural optimization of PNN.

3 STRUCTURAL MODEL

One of the most natural features of multilayer neural networks is the possibility to connect any particular neuron with arbitrary subset of nodes of input layer. Unfortunately, in probabilistic neural networks this structural freedom is not compatible with a statistically correct Bayesian decision-making. For example, if we assume that each layer of a neural network is described by a mixture of component distributions corresponding to neurons, then all the components must be defined on the same input space to satisfy the norming property

$$\sum_{x \in \mathcal{X}} P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} f(m) \sum_{x \in \mathcal{X}} F(\boldsymbol{x}|m) = 1.$$

Obviously, any component $F(\boldsymbol{x}|m)$ defined on a subspace of $\mathcal{X}$ (i.e. normed to 1 on a subspace of $\mathcal{X}$) would disturb the above norming condition. For this reason all the neurons must be connected with all the input variables and, in this sense, the complete interconnection property of probabilistic neural networks is enforced by the very basic paradigm of probabilistic description. On the other hand, such a structural "rigidity" is unnatural from the point of view of biological neural systems. It should be also emphasized that optimization of the completely interconnected models may cause computational difficulties because of a large number of involved parameters.

To avoid the undesirable complete interconnection property we apply the structural approach to probabilistic neural networks [6],[11]. Making substitution

$$F(\boldsymbol{x}|m) = F(\boldsymbol{x}|0)G(\boldsymbol{x}|m,\phi_m),$$

we introduce a modified mixture of distributions

$$P(\boldsymbol{x}|\omega) = \sum_{m\in\mathcal{M}_\omega} F(\boldsymbol{x}|0)G(\boldsymbol{x}|m,\phi_m)f(m) \qquad (6)$$

where

$$F(\boldsymbol{x}|0) = \prod_{n\in\mathcal{N}} f_n(x_n|0), \quad \mathcal{N} = \{1,2,\ldots,N\} \qquad (7)$$

is a nonzero "background" probability distribution common to all classes $\omega \in \Omega$. The background distribution is usually defined as a product of marginals, i.e. $f_n(x_n|0) = P_n(x_n)$. The component functions $G(\boldsymbol{x}|m,\phi_m)$ include additional binary structural parameters $\phi_{mn} \in \{0,1\}$:

$$G(\boldsymbol{x}|m,\phi_m) = \prod_{n\in\mathcal{N}} \left[\frac{f_n(x_n|m)}{f_n(x_n|0)}\right]^{\phi_{mn}}, \qquad (8)$$

$$\phi_m = (\phi_{m1},\ldots,\phi_{mN}) \in \{0,1\}^N.$$

We can see that, by setting $\phi_{mn} = 0$, any component-specific distribution $f_n(x_n|m)$ can be substituted by the respective (nonspecific) univariate background distribution $f_n(x_n|0)$, i.e. we can write equivalently

$$F(\boldsymbol{x}|m) = \prod_{n\in\mathcal{N}} f_n(x_n|m)^{\phi_{mn}} f_n(x_n|0)^{1-\phi_{mn}}. \qquad (9)$$

It can be seen that the component functions $G(\boldsymbol{x}|m,\phi_m)$ may be defined on different subspaces and the complexity and "structure" of the finite mixture (6) can be controlled by means of the binary parameters ϕ_{mn}.

It is an important aspect of the model (6) that the background probability distribution $F(\boldsymbol{x}|0)$ can be canceled in the Bayes formula (1), i.e. we can write

$$p(\omega|\boldsymbol{x}) = \frac{\sum_{m\in\mathcal{M}_\omega} G(\boldsymbol{x}|m,\phi_m)f(m)}{\sum_{\omega\in\Omega}\sum_{j\in\mathcal{M}_\omega} G(\boldsymbol{x}|j,\phi_j)f(j)}. \qquad (10)$$

Therefore, the a posteriori probability $p(\omega|\boldsymbol{x})$ is proportional to weighted sum of the component functions $G(\boldsymbol{x}|m,\phi_m)$ which can be defined on different subspaces:

$$p(\omega|\boldsymbol{x}) \approx \sum_{m\in\mathcal{M}_\omega} G(\boldsymbol{x}|m,\phi_m)f(m). \qquad (11)$$

As it will be shown in the next section, the optimal choice of structural parameters ϕ_{mn} can be included into the EM algorithm (cf. [11]).

According to our best knowledge, in literature there is no similar statistically correct subspace approach to Bayesian decision-making. The only related method can be traced back to an early paper of Watanabe [44] (see also e.g. [28], [46]) who proposed a classification rule based on projecting input data vectors into class-specific subspaces spanned by subsets of basis vectors, usually by subsets of principal components. The primary model for a class is then a linear subspace (linear manifold) of the Euclidean pattern space and the input vector $\boldsymbol{x} \in \mathcal{X}$ is classified according to its largest projection. In view of the typical properties of subspace methods (a) the classification of a pattern $\boldsymbol{x} \in \mathcal{X}$ is based solely on its direction and does not depend on the magnitude of $\boldsymbol{x}$ and (b) the decision surfaces are quadratic (cf. [35]). The second limitation has been avoided by considering mixtures of linear models (cf. e.g. [20], [16], [1]). It appears that Oja and others proposed neural network implementation of subspace methods (cf. e.g. [29], [30], [35], [15]). The subspace projection methods are computationally simple but they do not provide statistically correct decision models because they are not properly normalizable (cf. [15]).

4 ESTIMATION OF STRUCTURAL MODELS

Given a set of independent observations

$$\mathcal{S}_\omega = \{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(K)}\}, \quad \boldsymbol{x}^{(k)} \in \mathcal{X},$$

we can compute maximum-likelihood estimate of the mixture (6) by maximizing the log-likelihood criterion

$$L = \frac{1}{|\mathcal{S}_\omega|}\sum_{x\in\mathcal{S}_\omega} \log\Big[\sum_{m\in\mathcal{M}_\omega} F(\boldsymbol{x}|0)G(\boldsymbol{x}|\,m,\phi_m)f(m|\omega)\Big] \qquad (12)$$

by means of EM algorithm (cf. e.g. [3], [4], [49]). In our case we can write the iterative equations of EM algorithm in the form (cf. [11]):

E-Step: $(m \in \mathcal{M}_\omega,\ \boldsymbol{x} \in \mathcal{S}_\omega,\ t = 0,1,2,\ldots)$

$$q^{(t)}(m|\boldsymbol{x}) = \frac{G^{(t)}(\boldsymbol{x}|\,m,\phi_m^{(t)})f^{(t)}(m|\omega)}{\sum_{j\in\mathcal{M}_\omega} G^{(t)}(\boldsymbol{x}|j,\phi_j^{(t)})f^{(t)}(j|\omega)}, \qquad (13)$$

M-Step: $(m \in \mathcal{M}_\omega, n \in \mathcal{N})$

$$f^{(t+1)}(m|\omega) = \frac{1}{|\mathcal{S}_\omega|}\sum_{x\in\mathcal{S}_\omega} q^{(t)}(m|\,\boldsymbol{x}), \qquad (14)$$

$$f_n^{(t+1)}(\xi|m) =$$

$$= \frac{1}{|\mathcal{S}_\omega| f^{(t+1)}(m)} \sum_{x \in \mathcal{S}_\omega} \delta(\xi, x_n) q^{(t)}(m|\boldsymbol{x}), \quad \xi \in \mathcal{X}_n \tag{15}$$

$$\gamma_{mn}^{(t+1)} = \frac{1}{|\mathcal{S}_\omega|} \sum_{x \in \mathcal{S}_\omega} q^{(t)}(m|\boldsymbol{x}) \log \frac{f_n^{(t+1)}(x_n|m)}{f_n(x_n|0)} \tag{16}$$

$$\phi_{mn}^{(t+1)} = \begin{cases} 1, & \gamma_{mn}^{(t+1)} \in \Gamma^{(t+1)}, \\ 0, & \gamma_{mn}^{(t+1)} \notin \Gamma^{(t+1)}, \end{cases} , \tag{17}$$

where $\Gamma^{(t+1)}$ is the set of r highest quantities $\gamma_{mn}^{(t+1)}$:

$$\Gamma^{(t+1)} \subset \{\gamma_{mn}^{(t+1)}\}_{m \in \mathcal{M}_\omega \ n \in \mathcal{N}}, \quad |\Gamma^{(t+1)}| = r. \tag{18}$$

The iterative equations of EM algorithm generate a nondecreasing sequence $\{L^{(t)}\}_0^\infty$ converging to a possibly local maximum of the log-likelihood function (12), (cf. [11]).

Let us note that Eq. (16) can be rearranged by using equation (15):

$$\gamma_{mn}^{(t+1)} = f^{(t+1)}(m|\omega) \sum_{\xi \in \mathcal{X}_n} f_n^{(t+1)}(\xi|m) \log \frac{f_n^{(t+1)}(\xi|m)}{f_n(\xi|0)} = f^{(t+1)}(m|\omega) I(f_n^{(t+1)}(\cdot|m)||f_n(\cdot|0)) \tag{19}$$

and the structural criterion $\gamma_{mn}^{(t+1)}$ can be expressed in terms of Kullback-Leibler discrimination information (see e.g. [44]) $I(f_n^{(t+1)}(\cdot|m)||f_n(\cdot|0))$ between the conditional component-specific distribution $f_n^{(t+1)}(x_n|m)$ and the corresponding univariate "background" distribution $f_n(x_n|0)$. In this sense at each iteration the r-tuple of the most informative conditional distributions $f_n^{(t+1)}(\cdot|m)$ is included in the structural mixture model at each iteration.

Remark 4.1. In the standard form, the EM algorithm is an off-line estimation method. However, there is a straightforward connection to learning procedures via a sequential modification of the EM algorithm which can be interpreted from the neurophysiological point of view (cf. [12]).

5 COMPUTATIONAL EXPERIMENTS

The numeral database of Concordia University in Montreal, Canada was used repeatedly by different authors to test and compare various classification methods. The totally unconstrained handwritten numerals were collected from so called "dead-letter" envelopes by the U.S. Postal Service at different locations in the United States and digitized in bilevel on a 64x224 grid of 0.153 mm square raster fields. This corresponds to a resolution of approximately 166 PPI (cf. [2]).

The numerals show many different styles as well as sizes. For this reason the numerals were size-normalized probably in all the published experiments. Most authors have followed suggestion of the original documentation to use 4000 specified numerals for training of classifiers (400 per class) and 2000 numerals (200 per class) for independent testing. In most cases also different feature extraction methods were used in the preprocessing phase.

In the present paper the training- and testing sets were used as proposed in documentation. In the preprocessing phase all numerals were normalized to the size 32x32 in a simple way, by periodical deleting or doubling the rows and/or columns. No special feature extraction method was used, however, in order to decrease positional dependencies, the training data set was extended by 5 horizontal and 5 vertical shifts $(-2, -1, 0, +1, +2)$ with the resulting number of 100000, (= 5x5x4000) training numerals. This idea can be viewed as an analogy of the well known microscopic movements of human eye observing a fixed object.

The class-conditional distributions were approximated in the original 1024-dimensional space by the structural distribution mixtures (6), i.e. in the form

$$P(\boldsymbol{x}|\omega) = F(\boldsymbol{x}|0) \sum_{m \in \mathcal{M}_\omega} f(m|\omega) G(\boldsymbol{x}|m, \phi_m). \tag{20}$$

Since in our case the variables are binary: $x_n \in \{0, 1\}$, we can write

$$\theta_{nm} = f_n(1|m), \quad n \in \mathcal{N}, \quad m \in \mathcal{M}, \tag{21}$$

$$f_n(x_n|m) = \theta_{nm}^{x_n}(1 - \theta_{nm})^{1-x_n}, \quad x_n \in \{0, 1\} \tag{22}$$

and further

$$F(\boldsymbol{x}|0) = \prod_{n \in \mathcal{N}} \theta_{n0}^{x_n}(1 - \theta_{n0})^{1-x_n}, \tag{23}$$

$$G(\boldsymbol{x}|m, \phi_m) = \prod_{n \in \mathcal{N}} \left[\frac{\theta_{nm}^{x_n}(1 - \theta_{nm})^{1-x_n}}{\theta_{n0}^{x_n}(1 - \theta_{n0})^{1-x_n}} \right]^{\phi_{mn}}. \tag{24}$$

The parameters $f(m|\omega), \theta_{mn}$ and ϕ_{mn} were estimated by means of the EM algorithm of Section 4. In repeated computations the iterative procedure (13)-(17) was started randomly with identical number of components $|\mathcal{M}_\omega| = 35$. From the computational point of view the number of components appears to be rather unessential parameter since it is spontaneously suppressed in the course of EM iterations.

The total number of nonzero parameters ϕ_{mn} was set in different experiments to different values between 2000 and 7000. At the beginning the number

Tabulka 1: Recognition of numerals from the database of Concordia University, Montreal. Classification accuracy (class-conditional and global) of 8 independent randomly initialized solutions as verified by independent test set of 2000 numerals. The first 10 columns represent recognition accuracy of the classes "0", "1", ..., "9" respectively, the last column contains the global (average) accuracy.

Class:	0	1	2	3	4	5	6	7	8	9	Average
Solution 1	0.920	0.855	0.945	0.810	0.815	0.825	0.935	0.895	0.850	0.905	0.8755
Solution 2	0.810	0.810	0.920	0.820	0.830	0.805	0.930	0.895	0.840	0.900	0.8560
Solution 3	0.925	0.860	0.955	0.825	0.875	0.845	0.955	0.860	0.895	0.900	0.8895
Solution 4	0.935	0.905	0.935	0.820	0.835	0.825	0.960	0.850	0.905	0.910	0.8880
Solution 5	0.940	0.900	0.910	0.830	0.875	0.815	0.945	0.880	0.905	0.875	0.8875
Solution 6	0.885	0.905	0.895	0.810	0.865	0.780	0.950	0.865	0.845	0.870	0.8670
Solution 7	0.905	0.855	0.940	0.795	0.825	0.805	0.935	0.805	0.860	0.900	0.8625
Solution 8	0.900	0.865	0.925	0.820	0.845	0.870	0.940	0.855	0.905	0.910	0.8835

of component specific parameters θ_{mn}(characterized by $\phi_{mn} = 1$) was identical in all components with the initial position and value chosen randomly. In the course of iterations we observed strong differentiation. There was a clear tendency to accumulate the specific parameters θ_{mn} at a small number of significant components. Simultaneously, the components containing only nonspecific parameters θ_{0n} (i.e. with only zero structural parameters $\phi_{mn} = 0$) can be replaced by a single component weighted by the corresponding sum of weights. In this way the EM iteration process repeatedly resulted in a small number of components (10 - 20) with a relatively high number of component specific parameters (300 - 500) and one component without specific parameters. The weight of components is generally increasing with the number of specific parameters but this dependence doesn't hold strictly. By displaying the location of the chosen specific parameters at the raster we can see that the components roughly correspond to different variants of the considered numeral in the database (cf. Fig. 1).

The class-conditional probability distributions were estimated in 8 independent randomly initialized computational experiments. We needed several tens of iterations of EM algorithm to achieve the ultimate classification accuracy. In all experiments we obtained recognition accuracy between 85% and 89%, as shown in the Table 1.

6 RESULT COMPARISON

Table 2 shows some results relating to the same data and published in literature. For the sake of comparison we confined ourselves to formally identical experiments only with the recommended training- and test sets. Also, to keep the comparison simple, we ignore the reject option considered by several authors.

Let us recall that, as it appears, in the published experiments the numerals were size-normalized and, unlike our solutions, transformed to a relatively small number of highly informative features. Thus, Kim & Lee [19] and Cho [2] used so called Kirsch masks to compute directional features. Hwang & Bang [17] extracted features called "peripheral directional contributivity", Lam & Suen [22] and Legault & Suen [24] used structural approaches to extract features. The feature extraction methods often make use of some informal a priori knowledge and may essentially improve the final recognition quality.

7 CONCLUDING REMARK

In the present paper we show that the biologically unnatural complete interconnection property of probabilistic neural networks can be removed in a statistically correct way without leaving the exact framework of Bayesian decision-making. The method is based on distribution mixtures with product components including structural parameters.

The present application of the structural approach corresponds to a three-layer neural network including the input layer, the second layer of structural component functions and the third layer of output nodes corresponding to a posteriori probabilities.

Let us remark that the method proposed in the paper [11] can be used to design multilayer neural networks by applying the structural optimization repeatedly, layer by layer. This is partly enabled by the previously introduced concept of information preserving transform of the decision problem (cf. [9]) and also by the binary approximation of PNN (cf. [12]).

Tabulka 2: Comparison of published results on recognition of numerals from the database of Concordia University, Montreal. Only experiments using the recommended training- and test sets are included.

Author	year	accuracy
Lam & Suen [22]	(1988)	0.9310
Legault & Suen [24]	(1989)	0.9390
Krzyzak et al. [21]	(1990)	0.8640
Krzyzak et al. [21]	(1990)	0.9485
Mai & Suen [25]	(1990)	0.9295
Nadal & Suen [26]	(1990)	0.8605
Suen et al. [40]	(1990)	0.9305
Kim & Lee [19]	(1994)	0.9540
Kim & Lee [19]	(1994)	0.9585
Lee [23]	(1995)	0.9780
Hwang & Bang [17]	(1996)	0.9785
Cho [2]	(1997)	0.9605

Reference

[1] Bregler, C. & Omohundro, S.M., (1995): Nonlinear image interpolation using manifold learning. In: G. Tesauro, D.S. Touretzky,& T.K. Leen (Eds.), *Advances in Neural Information Processing Systems* **7** (pp. 971-980), Cambridge, MA.: MIT Press.

[2] Cho, S.-B. (1997): Neural network classifiers for recognizing totally unconstrained handwritten numerals. *IEEE Trans. on Neural Networks*, **8**, (pp. 43-53).

[3] Dempster, A.P., Laird, N.M., & Rubin, D.B., (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Stat. Soc.*, **B 39**, pp. 1-38.

[4] Grim J., (1982a): On numerical evaluation of maximum likelihood estimates for finite mixtures of distributions. *Kybernetika*, **18**, pp. 173-190.

[5] Grim J., (1982b): Design and optimization of multilevel homogeneous structures for multivariate pattern recognition. In: *Fourth FORMATOR Symposium 1982*, Academia, Prague 1982, pp. 233-240.

[6] Grim J., (1986a): Multivariate statistical pattern recognition with nonreduced dimensionality. *Kybernetika*, **22**, pp. 142-157.

[7] Grim, J., (1986b): Sequential decision-making in pattern recognition based on the method of independent subspaces. In: F. Zitek (Ed.), *Proc. of the DIANA II Conf. on Discriminant Analysis*, (pp. 139-149), Prague: Mathem. Institute of the AS CR.

[8] Grim, J., (1996a): Maximum Likelihood Design of Layered Neural Networks. In: *Proceedings of the 13th International Conference on Pattern Recognition* **IV** (pp. 85-89), Los Alamitos: IEEE Computer Society Press.

[9] Grim, J., (1996b): Design of multilayer neural networks by information preserving transforms. In: E. Pessa, M.P. Penna, A. Montesanto (Eds.), *Proceedings of the Third European Congress on System Science* (pp. 977-982), Roma: Edizzioni Kappa.

[10] Grim J. (1999a): A sequential modification of EM algorithm. In *Studies in Classification, Data Analysis and Knowledge Organization*, Gaul W., Locarek-Junge H., (Eds.), pp. 163-170, Springer, 1999.

[11] Grim J. (1999b): Information approach to structural optimization of probabilistic neural networks. In proceedings of: *4th System Science European Congress*, L. Ferrer, A. Caselles, R. Beneyto, R. Pla, I. Martinez, V. Rossi, J. Martinez, J.R. Hernandez (Eds.), (pp. 527-540), Valencia: Sociedad Espanola de Sistemas Generales, 1999.

[12] Grim J. & Pudil P., (1998b): On virtually binary nature of probabilistic neural networks. In: *Advances in Pattern Recognition*, (Proceedings of the IAPR International Workshops SSPR'98 and SPR'98), Sydney, August 11-13, 1998), A. Amin, D. Dori, P. Pudil, H. Freeman (Eds.), (pp. 765-774), Springer: New York, Berlin, 1998.

[13] Haykin, S., (1993): *Neural Networks: a comprehensive foundation.* San Mateo CA: Morgan Kaufman.

[14] Hebb, D.O., (1949): *The Organization of Behavior: A Neuropsychological Theory.* New York: Wiley.

[15] Hinton, G.E., Dayan, P., & Revow, M., (1997): Modeling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks*, **8**, 65-74.

[16] Hinton, G.E., Revow, M., & Dayan, P., (1995): Recognizing handwritten digits using mixtures of linear models. In: G. Tesauro, D.S. Touretzky & T.K. Leen (Eds.), *Advances in Neural Information Processing Systems* **7**, pp. 1015-1022. Cambridge, MA.: MIT Press.

[17] Hwang, Y.-S. & Bang, S.-Y., (1996): An efficient method to construct a radial basis function neural network classifier and its application to unconstrained handwritten digit recognition. In: *Proceedings of the 13th International Conference on Pattern Recognition* (pp. 640-644), Los Alamitos: IEEE Computer Society Press.

[18] Jacobs, R.A., & Jordan, M.I., (1991): A competitive modular connectionist architecture. In: R.P. Lippmann, J.E. Moody & D.J. Touretzky (Eds.), *Advances in Neural Information Processing Systems* **3** (pp. 767-773), San Mateo CA: Morgan Kaufman.

[19] Kim, Y.J. & Lee, S.W., (1994): Off-line recognition of unconstrained handwritten digits using multilayer backpropagation neural network combined with genetic algorithm. (in Korean) In *Proc. 6th Workshop on Image Processing and Understanding*, (pp. 186-193).

[20] Kohonen, T., Nemeth, G., Bry, K., Jalanko, M., & Makisara, K. (1979). Spectral classification of phonemes by learning subspace methods. In *Proceeding IEEE International Conference on Acoustics, Speach and Signal Processing* (pp. 97-100), Washington D.C.

[21] Krzyzak A., Dai W. & Suen C.Y., (1990): Unconstrained handwritten character classification using modified backpropagation model. In *Proc. 1st Int. Workshop Frontiers of Handwriting Recognition and Understanding*, Montreal, Canada, 1990, (pp. 155-166).

[22] Lam L. & Suen C.Y., (1988): Structural classification and relaxation matching of totally unconstrained ZIP-code numbers. *Pattern Recognition*, **21**(1), (pp. 19-31).

[23] Lee, S.-W., (1995): Multilayer cluster neural network for totally unconstrained handwritten numeral recognition. *Neural Networks*, **8**, (5), pp. 783-792, 1995.

[24] Legault R., Suen C.Y., (1989): Contour tracing and parametric approximation for digitized patterns. In *Computer Vision and Shape Recognition*, Krzyzak A., Kasvand T., & Suen C.Y. (Eds.), Singapore: World Scientific (1989), (pp. 225-240).

[25] Mai T. & Suen C.Y., (1990): A generalized knowledge-based system for the recognition of unconstrained handwritten numerals. *IEEE Trans. on Syst. Man, Cybern.*, **20**, (pp. 835-848).

[26] Nadal C. & Suen C.Y., (1990): Recognition of totally unconstrained handwritten digits by decomposition and vectorization. Concordia University, Montreal, Tech. Rep., (1988).

[27] Novovičová, J., Pudil, P. & Kittler, J., (1996): Divergence based feature selection for multimodal class densities. *IEEE Trans. on PAMI*, **18**, 2, pp. 218-223.

[28] Oja, E., (1983): *Subspace Methods of Pattern Recognition.* Letchworth, U.K.: Research Studies Press, 1983.

[29] Oja, E., (1989): Neural networks, principal components and subspaces. *International Journal of Neural Systems*, **1**, 61-68.

[30] Oja, E., & Kohonen, T., (1988): The subspace learning algorithm as a formalism for pattern recognition and neural networks. In: *Proceeding 1988 IEEE International Conf. on Neural Networks* (pp. 277-284), San Diego, CA.

[31] Palm, H.Ch., (1994): A new method for generating statistical classifiers assuming linear mixtures of Gaussian densities. In: *Proc. of the 12th IAPR Int. Conf. on Pattern Recognition, Jerusalem, 1994*, **II**, (pp. 483-486), Los Alamitos: IEEE Computer Soc. Press.

[32] Parzen, E., (1962): On estimation of a probability density function and its mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.

[33] Poggio, T., & Girosi, F., (1990): Networks for approximation and learning. *Proceedings of the IEEE*, **78**, 1481-1497.

[34] Powell, M.J.D., (1992): The theory of radial basis function approximation. In: *Advances in Num. Analysis II.* Oxford: Clarendon Press.

[35] Prakash, M., & Murty, M.N., (1997): Growing subspace pattern recognition methods and their neural-network models. *IEEE Trans. on Neural Networks*, **8**, pp. 161-168.

[36] Pudil, P., Novovičová, J., Choakjarernwanit, N., & Kittler, J., (1995): Feature selection based on the approximation of class densities by finite mixtures of special type. *Pattern Recognition* **28**(9), pp. 1389-1398.

[37] Schlesinger, M.I., (1968): Relation between learning and self-learning in pattern recognition. (in Russian), *Kibernetika*, (Kiev), No. 2, pp. 81-88.

[38] Specht, D.F., (1988): Probabilistic neural networks for classification, mapping or associative memory. In: *Proceeding of the IEEE International Conference on Neural Networks, July 1988*, Vol. I, (pp. 525-532).

[39] Streit, L.R., & Luginbuhl, T.E., (1994): Maximum likelihood training of probabilistic neural networks. *IEEE Trans. on Neural Networks*, **5**, pp. 764-783.

[40] Suen C.Y., Nadal C., Mai T., Legault R. & Lam L., (1990): Recognition of handwritten numerals based on the concept of multiple experts. In *Proc. 1st Int. Workshop Frontiers of Handwriting Recognition*, Montreal, Canada, 1990, (pp. 131-144).

[41] Titterington, D.M., Smith, A.F.M., & Makov, U.E., (1985). *Statistical analysis of finite mixture distributions.* New York: John Wiley & Sons.

[42] Vajda, I., (1992): *Theory of Statistical Inference and Information.* Boston: Kluwer.

[43] Vajda I. & Grim J., (1998): About the maximum information and maximum likelihood principles in neural networks. *Kybernetika*, Vol. 34, No. 4, pp. 485-494.

[44] Watanabe, S., (1967): Karhunen-Loeve expansion and factor analysis. In: *Trans. of the Fourth Prague Conf. on Information Theory*, (pp. 635-660), Prague: Academia.

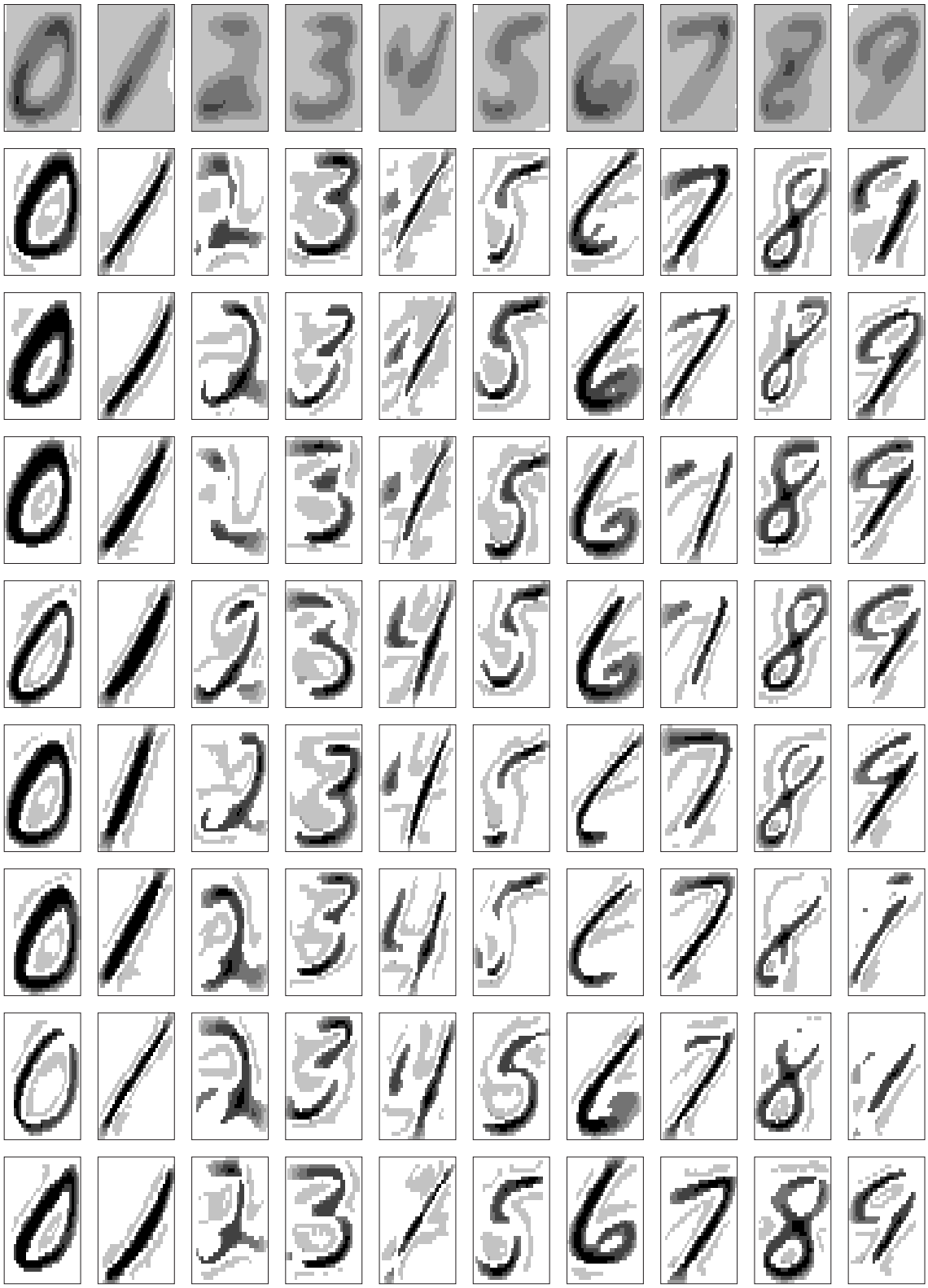
[45] Watanabe, S., & Fukumizu, K., (1995): Probabilistic design of layered neural networks based on their unified framework. *IEEE Trans. on Neural Networks*, **6**, 691-702.

[46] Watanabe, S., & Pakvasa, N., (1973): Subspace method in pattern recognition. In: *Proc. Int. Joint Conf. on Pattern Recognition*, (pp. 25-32).

[47] Wu, C.F.J., (1983): On the convergence properties of the EM algorithm, *Annals of Statistics*, **11**, 95-103.

[48] Xu L. & Jordan M.I., (1993): EM learning on a generalized finite mixture model for combining multiple classifiers, In: *World Cong. on Neural Networks.* **4** (pp. 227-230). Portland OR.

[49] Xu L. & Jordan, M.I., (1996): On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, **8**, 129-151.



Obrázek 1: The first row of the figure shows the marginal probabilities for all classes (i.e. "mean digits"). The gray-levels reflect the increasing values of the respective parameters θ_{nm}. The next eight rows show "receptive fields" of the first-layer neurons as defined by the structural parameters $\phi_{nm} = 1$. The white raster fields correspond to the zero values $\phi_{nm} = 0$ (i.e. unused inputs). The first eight components of each class-conditional mixture are shown respectively.