Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Jiří Grim
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic

MAXIMUM-LIKELIHOOD STRUCTURING OF PROBABILISTIC NEURAL NETWORKS

No. 1894 Prague, May 1997

ÚTIA AV ČR, P.O. Box 18, 182 08 Prague, Czech Republic
E-mail: grim@utia.cas.cz
Fax: (+420)(2)688 3031

MAXIMUM-LIKELIHOOD STRUCTURING OF PROBABILISTIC NEURAL NETWORKS

Jiří Grim

Department of Pattern Recognition
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic

Abstract: *We propose a maximum-likelihood method to optimize multilayer neural networks in the framework of statistical decision-making. The method is based on approximating unknown probability distributions by finite mixtures and transforming variables between consecutive layers without information loss. Both the parameters and the structure of neural networks can be optimized simultaneously by means of EM algorithm. The structural optimization actually follows from a special subspace approach which computes statistically correct a posteriori probabilities on different subspaces. At a very general level the coordinate functions of the transform can be interpreted in terms of functional properties of neurons. In this sense the proposed method could be helpful to better understanding of biological neural systems.*

Keywords: *Probabilistic neural networks, Statistical decision-making, Structure optimization, Subspace approach, Finite mixtures, EM algorithm, Information preserving transform, Hebbian learning*

1 Introduction

Speaking about probabilistic neural networks, we refer to several papers based on approximation of probability density functions by finite mixtures in the framework of statistical decision-making (cf. Specht, 1988; Haykin, 1993; Xu & Jordan, 1993, 1996; Palm, 1994; Watanabe & Fukumizu, 1995; Streit & Luginbuhl, 1994; Bishop 1995; Grim, 1996, 1996a).

Unlike usual approaches the class-conditional densities are approximated by finite mixtures with components from a common pool of probability density functions. Thus the component density functions corresponding to neurons may be shared by all class-conditional mixtures without any structural limitations. The resulting probabilistic model can be interpreted as a three-layer feedforward neural network with the first layer of input variables, the second "hidden" layer of shared component densities and the third

[0] Supported by the Grant of the Academy of Sciences No. A2075703, by the Grant of the Grant Agency of CR, No. 402/97/1242, by the Grant of the Ministry of Education No. VS 96063 and partially by the Complex research project of the Academy of Sciences No. K1075601 of the Czech Republic.

layer of a posteriori probabilities of classes. The shared components naturally define an additional "descriptive" decision problem which can be estimated by EM algorithm in unsupervised way. The descriptive classes may correspond e.g. to some elementary situations on input.

A weak point of the probabilistic neural networks is the tacitly assumed complete interconnection of component densities with all input variables (or neurons of the preceding layer). This property follows from the fundamental fact that all component densities of a mixture must be defined on the same space and therefore they have to depend on the same set of variables. Thus the complete interconnection property of probabilistic neural networks arises from the very basic paradigma of probabilistic description. On the other hand, such a structural "rigidity" is unnatural from the point of view of biological neural systems.

In the present paper we suggest a new approach to structural optimization of the probabilistic neural networks by means of maximum-likelihood criterion - without leaving the exact framework of probability theory. The method makes use of an idea originally designed for multivariate pattern recognition (cf. Grim, 1986). It is based on finite mixtures with factorizable components including binary structural parameters. By means of a special "background" substitution technique the computation of statistically correct a posteriori probabilities can be reduced to different subspaces and, for the same reason, the receptive fields of corresponding neurons can be confined to arbitrary subsets of input variables.

In literature there is a similarly motivated subspace approach based on projecting input data vectors into class-specific subspaces. It appears that the original idea of Watanabe (1967) has been generalized to mixtures of subspaces (cf. e.g. Kohonen *et al.*, 1979; Hinton *et al.*, 1995 and others) and implemented by neural networks (cf. e.g. Oja, 1983; Oja & Kohonen, 1988 and others). Subspace-projection methods are computationally feasible, but they do not provide statistically correct decision models.

The concept of descriptive decision problem suggests theoretically well justified possibility for a sequential design of multilayer structures. We make use of the fact (cf. Grim, 1996a, Vajda & Grim, 1996) that the Shannon information about the descriptive decision problem is automatically preserved by a special class of transforms defined in terms of a posteriori probabilities. The optimization procedure starting with unsupervised estimation of the distribution mixture and including information preserving transformation of the descriptive decision problem can be applied repeatedly to optimize multilayer neural networks. Only the estimation of class-conditional component weights at the highest level has to be supervised.

The information preserving transform can be well interpreted in terms of functional properties of neurons. Particularly, by expanding the components of the transform, we obtain terms responsible for spontaneous activity of neurons, for contributions from synapses of other neurons and for lateral inhibition. The properties of the formally deduced synaptic weights justify the classical Hebb's postulate of learning.

2 Method of Mixtures

First we recall briefly solution of a statistical decision problem $\{\mathcal{X}, P(\cdot|\omega)p(\omega), \omega \in \Omega\}$ based on approximating class-conditional probability distributions by finite mixtures. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ be a finite set of classes with a priori probabilities $p(\omega)$ and $P(\boldsymbol{x}|\omega)$ be the corresponding class-conditional probability density functions on a real space $\mathcal{X} = R^N$. All statistical information about the set of classes Ω given some observation $\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}$ 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 joint probability density of $\boldsymbol{x}$. The posterior distribution $p(\omega|\boldsymbol{x})$ may further be used to define the final decision.

We assume that the conditional densities $P(\boldsymbol{x}|\omega)$ can be approximated by finite mixtures. However, unlike usual approaches, we use component densities from a common pool (cf. Bishop, 1995; Grim, 1996). In particular, we assume that there is a finite set $\mathcal{F} = \{F(\cdot|m), m \in \mathcal{M}\}$ of probability density functions on $\mathcal{X}$ such that each conditional density $P(\boldsymbol{x}|\omega)$ may be expressed as a convex combination of densities from $\mathcal{F}$

$$P(\boldsymbol{x}|\omega) = \sum_{m \in \mathcal{M}} F(\boldsymbol{x}|m)f(m|\omega), \quad \omega \in \Omega, \quad \boldsymbol{x} \in \mathcal{X}, \quad \mathcal{M} = \{1, \ldots, M\}. \tag{3}$$

Here $f(m|\omega) \geq 0$ are some conditional probabilistic weights and the components $F(\boldsymbol{x}|m)$ may be shared by all class-conditional densities $P(\boldsymbol{x}|\omega)$.

The statistical model (1) - (3) can be interpreted as a three-layer feed-forward neural network: the first layer is represented by the input variables $x_1, x_2, \ldots, x_N$, the shared component densities $F(\boldsymbol{x}|m)$ represent the second "hidden" layer of neurons and the third layer corresponds to the a posteriori probabilities $p(\omega|\boldsymbol{x})$ of classes.

As it can be seen the concept of shared components avoids structural limitations at the final level of statistical decision-making. Similar schemes have also been proposed for radial basis functions (RBF) neural networks (cf. Jacobs & Jordan, 1991; Haykin, 1993). However, instead of usual multivariate interpolation or approximation of output variables (cf. e.g. Poggio & Girosi, 1990; Powel, 1992), the purpose of RBF's in the probabilistic framework is to model the components of the underlying statistical decision problem.

By using substitution (3) we can rewrite the joint density $P(\boldsymbol{x})$ in the form

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} F(\boldsymbol{x}|m)f(m), \quad \boldsymbol{x} \in \mathcal{X} \tag{4}$$

where

$$f(m) = \sum_{\omega \in \Omega} f(m|\omega)p(\omega). \tag{5}$$

The set of shared component densities $F(\boldsymbol{x}|m)$ naturally introduces an additional descriptive decision problem $\{\mathcal{X}, F(\cdot|m)f(m), m \in \mathcal{M}\}$ with a priori probabilities $f(m)$ whereby each component in the mixture (4) may correspond e.g. to an elementary situation on the input. Given an observation $\boldsymbol{x} \in \mathcal{X}$, the a posteriori probabilities

$$f(m|\boldsymbol{x}) = \frac{F(\boldsymbol{x}|m)f(m)}{P(\boldsymbol{x})}, \quad m \in \mathcal{M}, \quad \boldsymbol{x} \in \mathcal{X} \tag{6}$$

may be interpreted as a measure of presence of different elementary situations. Simultaneously, there is a simple relation between a posteriori probabilities of classes and of the elementary situations (cf. (1), (3))

$$p(\omega|\boldsymbol{x}) = \sum_{m \in \mathcal{M}} p(\omega|m)f(m|\boldsymbol{x}), \quad p(\omega|m) = \frac{f(m|\omega)p(\omega)}{f(m)}, \quad \omega \in \Omega, \quad \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

Let us remark in this connection that in Sec.7 the output variable of the m-th neuron is defined as $\log f(m|\boldsymbol{x})$ to make the corresponding transform information preserving. Also the subspace approach of Sec.4 makes use of the fact that the computation of the a posteriori probabilities $f(m|\boldsymbol{x})$ can be confined to different subspaces. It is therefore rather important that the final solution $p(\omega|\boldsymbol{x})$ of the original decision problem can be expressed in the form of a linear combination of the a posteriori probabilities $f(m|\boldsymbol{x})$.

Let us recall that, here and in the following we consider continuous variables characterized by probability density functions. However, most of the results of Sec.3–7 apply to discrete distributions as well.

It should also be emphasized that, by introducing the concept of shared components in Eq. (3), we make an implicit assumption

$$F(\cdot|m,\omega) = F(\cdot|m), \quad \text{for all} \;\; \omega \in \Omega, \quad m \in \mathcal{M}. \tag{8}$$

In terms of Shannon information this assumption implies (cf. Grim, 1996)

$$I(\mathcal{X}, \mathcal{M} \times \Omega) = I(\mathcal{X}, \mathcal{M}), \quad I(\mathcal{X}, \Omega) \leq I(\mathcal{X}, \mathcal{M}). \tag{9}$$

The inequalities (9) clarify the meaning of the descriptive decision problem since, as the decision information $I(\mathcal{X}, \Omega)$ is bounded by the "descriptive" information $I(\mathcal{X}, \mathcal{M})$, possible information loss caused by inaccurately estimated components $F(\boldsymbol{x}|m)$ may become irreparable.

3 Hybrid Estimation Scheme

Numerically the finite mixture model of Sec.2 can be optimized by means of EM algorithm (cf. Grim, 1982, 1996). In view of the importance of this remarkable computational scheme in the following we recall that, as it appears, the first proof of its monotonous convergence is due to Schlesinger (1968). The result of Schlesinger has been further discussed in a textbook of Ajvazjan *et al.* (1974) and in a survey paper of Isaenko &

Urbakh (1976). At present the standard reference is the apparently independent paper of Dempster *et al.* (1977) who introduced the name EM algorithm and demonstrated its applicability in different fields.

Suppose now that for each $\omega \in \Omega$ there is a nonempty set $\mathcal{S}_\omega$ of independent observations identically distributed according to some unknown probability density $P(\boldsymbol{x}|\omega)$:

$$\mathcal{S}_\omega = \{\boldsymbol{x}_1^{(\omega)}, \dots, \boldsymbol{x}_{K_\omega}^{(\omega)}\}, \quad \boldsymbol{x}_k^{(\omega)} \in \mathcal{X}, \quad \mathcal{S} = \bigcup_{\omega \in \Omega} \mathcal{S}_\omega.$$

Let us note first that we can estimate the component densities $F(\boldsymbol{x}|m)$ in unsupervised way, irrespectively of the classes $\omega \in \Omega$, by maximizing the log-likelihood function

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log[\sum_{m \in \mathcal{M}} F(\boldsymbol{x}|m) f(m)], \tag{10}$$

where $|\mathcal{S}|$ denotes the number of elements in the set $\mathcal{S}$. The m.- l. estimates of parameters can be computed by the iterative equations of EM algorithm (cf. Grim, 1982, 1996):

E-step: $\quad (m \in \mathcal{M},\ \boldsymbol{x} \in \mathcal{S},\ \ t = 0, 1, \dots)$

$$q^{(t)}(m|\boldsymbol{x}) = \frac{F^{(t)}(\boldsymbol{x}|m) f^{(t)}(m)}{\sum_{j \in \mathcal{M}} F^{(t)}(\boldsymbol{x}|j) f^{(t)}(j)}, \tag{11}$$

M-step: $\quad (m \in \mathcal{M})$

$$f^{(t+1)}(m) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q^{(t)}(m|\boldsymbol{x}), \tag{12}$$

$$F^{(t+1)}(\cdot|m) = \arg \max_{F(\cdot|m)} \{ \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q^{(t)}(m|\boldsymbol{x}) \log F(\boldsymbol{x}|m) \}. \tag{13}$$

The last implicit equation can be expressed in a closed form in most cases of practical importance (cf. e.g. Grim, 1996).

Having obtained the unsupervised estimates of the components $F(\boldsymbol{x}|m)$ we may confine the supervised estimation to the conditional weights $f(m|\omega)$ ($\approx$ hybrid scheme) by using the global log-likelihood function (cf. Grim, 1996)

$$L_G = \frac{1}{|\mathcal{S}|} \sum_{\omega \in \Omega} \sum_{x \in \mathcal{S}_\omega} \log\, [P(\boldsymbol{x}|\omega) p(\omega)]. \tag{14}$$

Making substitution (3) and setting $p(\omega) = |\mathcal{S}_\omega|/|\mathcal{S}|$ we obtain the criterion

$$\bar{L}_G = \frac{1}{|\mathcal{S}|} \sum_{\omega \in \Omega} \sum_{x \in \mathcal{S}_\omega} \log\, [\sum_{m \in \mathcal{M}} F(\boldsymbol{x}|m) f(m|\omega)] \tag{15}$$

and the corresponding iteration equations:

E-step: $\quad (m \in \mathcal{M},\ \boldsymbol{x} \in \mathcal{S},\ \ \omega \in \Omega,\ \ t = 0, 1, \dots)$

$$q^{(t)}(m|\boldsymbol{x}, \omega) = \frac{F^{(t)}(\boldsymbol{x}|m) f^{(t)}(m|\omega)}{\sum_{j \in \mathcal{M}} F^{(t)}(\boldsymbol{x}|j) f^{(t)}(j|\omega)}, \tag{16}$$

M-step: $(m \in \mathcal{M}, \quad \omega \in \Omega)$

$$f^{(t+1)}(m|\omega) = \frac{1}{|\mathcal{S}_\omega|} \sum_{x \in \mathcal{S}_\omega} q^{(t)}(m|\boldsymbol{x},\omega), \tag{17}$$

However, the component densities $F(\boldsymbol{x}|m)$ can be included into the supervised estimation by means of Eq.

$$F^{(t+1)}(\cdot|m) = \arg\max_{F(\cdot|m)} \{\frac{1}{|\mathcal{S}|} \sum_{\omega \in \Omega} \sum_{x \in \mathcal{S}_\omega} q^{(t)}(m|\boldsymbol{x},\omega) \log F(\boldsymbol{x}|m)\}. \tag{18}$$

Thus, to optimize the component densities $F(\boldsymbol{x}|m)$, we can use two obviously different iteration schemes, the unsupervised (11) - (13) and the supervised one (16) - (18). The two schemes may coincide only asymptotically, provided that the concept of shared components is applicable and all the mixtures are uniquely identifiable.

This fact could be of practical importance since, in a particular case, the supervised estimates of components could yield higher classification accuracy and, on the other hand, the unsupervised scheme could be less sensitive to small data sets.

In the present paper we prefer the “hybrid” estimation procedure (16) - (17) based on unsupervised estimates of the component densities $F(\boldsymbol{x}|m)$ because, as it will be shown in Sec.7, it is suitable for a sequential design of multilayer networks by means of information preserving transforms.

4 Subspace Approach

Let us recall that one of the most natural features of neural networks is the possibility to connect any particular neuron with nearly arbitrary subset of input variables. Unfortunately, in probabilistic neural networks this simple possibility is usually not compatible with a statistically correct decision-making. If we assume that each layer of a neural network is described by a mixture of component densities corresponding to neurons, then all the components must be defined on the same input space to satisfy formal properties of probability density functions. Thus, all neurons must be connected with all input variables and, in this sense, the complete interconnection is a direct consequence of the probabilistic description.

The problem has a counterpart in pattern recognition. It is well known that in high dimensional spaces feature selection methods may appear too restrictive or inefficient if for individual classes the subsets of informative variables (features) are essentially different. Thus e.g. we would use different subsets of binary rastr fields to estimate conditional distributions of handwritten digits properly and economically. However, instead of class-specific subspaces, we have to use one and the same input subspace (possibly of a high dimension) to compute desirable a posteriori probabilities of classes since otherwise the underlying statistical model would be incorrect.

In the following sections we shall suppose that the conditional probability density functions $F(\cdot|m) \in \mathcal{F}$ are factorizable, i.e. that we can write

$$F(\boldsymbol{x}|m) = \prod_{n\in\mathcal{N}} f_n(x_n|m), \quad \boldsymbol{x} \in \mathcal{X} \tag{19}$$

where $\mathcal{N} = \{1, 2, \ldots, N\}$ is the index set and $f_n(x_n|m)$ are univariate conditional densities. It can be seen that, from the theoretical point of view, this assumption is not restrictive. In discrete case, the class of finite mixtures

$$P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} f(m) \prod_{n\in\mathcal{N}} f_n(x_n|m), \quad \boldsymbol{x} \in \mathcal{X}, \tag{20}$$

is complete in the sense that any discrete probability distribution on $\mathcal{X}$ can be expressed in the form (20), for M sufficiently large (Grim, 1996). In case of continuous variables we may refer to similar asymptotic properties of nonparametric (kernel type) density estimates (see e.g. Parzen, 1962).

It can be seen that, by Eq. (20), the variables $x_1, x_2, \ldots, x_N$ are conditionally independent with respect to the index variable m. In practical problems the conditionally independent model (20) may become less efficient e.g. in case of 'elongated clusters' of real data vectors with highly correlated components. On the other hand, we have good experience with the conditionally independent models of discrete data. Let us recall also that the distribution mixture (20) is closely related to the concept of latent classes introduced by Lazarsfeld (1966) for binary variables. In the present paper the most important argument for the conditionally independent model (20) is the possibility to avoid the necessity of fully interconnected units in probabilistic neural networks - without leaving the exact probabilistic framework.

The structural approach to probabilistic neural networks makes use of an idea originally designed for multivariate statistical pattern recognition (cf. Grim, 1986). By means of a special "background" substitution technique the computation of a posteriori probabilities $f(m|\boldsymbol{x})$ may be reduced to subsets of informative variables and, in this way, we can optimize both the structure and parameters of neural networks simultaneously.

Making substitution

$$F(\boldsymbol{x}|m) = F(\boldsymbol{x}|0)G(\boldsymbol{x}|m, \phi_m), \quad m \in \mathcal{M} \tag{21}$$

in (4), we introduce a modified mixture of densities

$$P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} F(\boldsymbol{x}|0)G(\boldsymbol{x}|m, \phi_m)f(m) \tag{22}$$

where

$$F(\boldsymbol{x}|0) = \prod_{n\in\mathcal{N}} f_n(x_n|0) \tag{23}$$

is a nonzero "background" probability density 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 $\phi_{mn} \in \{0, 1\}$:

$$G(\boldsymbol{x}|m, \phi_m) = \prod_{n\in\mathcal{N}} \left[\frac{f_n(x_n|m)}{f_n(x_n|0)}\right]^{\phi_{mn}}, \quad \phi_m = (\phi_{m1}, \ldots, \phi_{mN}) \in \{0, 1\}^N. \tag{24}$$

We can see that, by setting the structural parameter $\phi_{mn} = 0$, any component-specific density function $f_n(x_n|m)$ can be substituted by the respective univariate background density $f_n(x_n|0)$, i.e.

$$F(\boldsymbol{x}|m) = \prod_{n\in\mathcal{N}} f_n(x_n|m)^{\phi_{mn}} f_n(x_n|0)^{1-\phi_{mn}}. \tag{25}$$

In this way the component functions $G(\boldsymbol{x}|m, \phi_m)$ may be defined on different subspaces and the complexity and "structure" of the finite mixture (22) can be controlled by means of the binary parameters ϕ_{mn}. Simultaneously, the number of involved parameters is reduced whenever $\phi_{mn} = 0$ and therefore the structured (incompletely interconnected) model would be less susceptible to possible "overfitting".

It is an important aspect of the model (22) that the background probability density $F(\boldsymbol{x}|0)$ cancels in the formula (6)

$$f(m|\boldsymbol{x}) = \frac{G(\boldsymbol{x}|m, \phi_m) f(m)}{\sum_{j\in\mathcal{M}} G(\boldsymbol{x}|j, \phi_j) f(j)} \tag{26}$$

and therefore the computation may be confined only to the relevant variables. For the same reason the input connections of a single neuron can be confined to any subset of variables (neurons) by means of the binary parameters ϕ_{mn}, as it will be shown in Sec.7.

According to our best knowledge in literature there is only one similarly motivated subspace approach. It can be traced back to an early paper of Watanabe (1967) (see also e.g. Watanabe & Pakvasa, 1973, Oja, 1983) who proposed classification rule based on projecting input data vectors into class-specific subspaces spanned by groups of basis vectors, usually by principal components. The primary model for a class is 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. Prakash & Murty, 1997). The second limitation has been avoided by considering mixtures of linear models (cf. e.g. Kohonen *et al.*, 1979; Hinton *et al.*, 1995; Bregler & Omohundro, 1995; Prakash & Murty, 1997). It appears that Oja and others proposed neural network implementation of subspace methods (cf. e.g. Oja, 1989; Oja & Kohonen, 1988; Prakash & Murty, 1997; Hinton *et al.*, 1995, 1997). The subspace projection methods are computationally advantageous but they do not provide statistically correct decision models because they are not properly normalizable (cf. Hinton *et al.*, 1997).

5 Estimation of Structured Models

We assume first that the structural parameters ϕ_{mn} are a priori known and fixed. In order to estimate the unknown densities $f_n(\cdot|m)$ and component weights $f(m)$ we maximize the unsupervised log-likelihood function (10), (cf. (21))

$$L = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \log\,[\sum_{m\in\mathcal{M}} F(\boldsymbol{x}|0) G(\boldsymbol{x}|\, m, \phi_m) f(m)]. \tag{27}$$

by means of EM algorithm. The related iteration equations can be rewritten as follows (cf. (11) - (13)):

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

$$q^{(t)}(m|\,\boldsymbol{x}) = \frac{G^{(t)}(\boldsymbol{x}|\,m, \phi_m) f^{(t)}(m)}{\sum_{j \in \mathcal{M}} G^{(t)}(\boldsymbol{x}|j, \phi_j) f^{(t)}(j)}, \tag{28}$$

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

$$f^{(t+1)}(m) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q^{(t)}(m|\,\boldsymbol{x}), \tag{29}$$

$$f_n^{(t+1)}(\cdot|m) = \arg \max_{f_n(\cdot|m)} \{ \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q^{(t)}(m|\,\boldsymbol{x}) \log f_n(\boldsymbol{x}|m) \}. \tag{30}$$

Again, relation (30) can usually be expressed in explicit form (cf. e.g. Grim, 1996).

Following the original idea of Schlesinger (1968) (see also Grim, 1982) we can derive the important problem-dependent iteration equations directly from the monotonous convergence condition. Using the above notation we can write for any two finite values $L^{(t+1)}, L^{(t)}$

$$L^{(t+1)} - L^{(t)} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \frac{P^{(t+1)}(\boldsymbol{x})}{P^{(t)}(\boldsymbol{x})}.$$

We modify the last expression by adding and subtracting the term

$$\sum_{x \in \mathcal{S}} \sum_{m \in \mathcal{M}} \frac{q^{(t)}(m|\,\boldsymbol{x})}{|\mathcal{S}|} \log \left[\frac{G^{(t+1)}(\boldsymbol{x}|\,m, \phi_m) f^{(t+1)}(m)}{G^{(t)}(\boldsymbol{x}|\,m, \phi_m) f^{(t)}(m)} \right]$$

which can be shown to be finite in most cases of practical importance, e.g. for $f_n(\cdot|m)$ normal. (Note that $f^{(t)}(m) = 0$ implies $q^{(t)}(m|\boldsymbol{x}) = 0$ and $f^{(t+1)}(m) = 0$ in view of Eqs. (28) and (29).) By using substitutions (28) and (29) we obtain

$$L^{(t+1)} - L^{(t)} = \sum_{m \in \mathcal{M}} f^{(t+1)}(m) \log \frac{f^{(t+1)}(m)}{f^{(t)}(m)} + \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \sum_{m \in \mathcal{M}} q^{(t)}(m|\boldsymbol{x}) \log \frac{q^{(t)}(m|\boldsymbol{x})}{q^{(t+1)}(m|\boldsymbol{x})} + \tag{31}$$
$$+ \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q^{(t)}(m|\,\boldsymbol{x}) \log \left[\frac{G^{(t+1)}(\boldsymbol{x}|m, \phi_m)}{G^{(t)}(\boldsymbol{x}|m, \phi_m)} \right]$$

where the first and second terms represent the nonnegative Kullback-Leibler information divergences (cf. e.g. Vajda, 1992) which is nonnegative for any two discrete distributions and equals zero if and only if the two distributions are equal. Making substitution (24), we can write

$$L^{(t+1)} - L^{(t)} = I(f^{(t+1)}||f^{(t)}) + \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} I(q^{(t)}(\cdot|\boldsymbol{x})||q^{(t+1)}(\cdot|\boldsymbol{x})) + \tag{32}$$

$$+\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\frac{\phi_{mn}}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}q^{(t)}(m|\boldsymbol{x})\log\left[\frac{f^{(t+1)}(x_n|m)}{f^{(t)}(x_n|m)}\right]\geq 0.$$

The inequality holds because the implicite relation (30) guarantees the last term in (32) to be nonnegative, too.

Let us note that the inequality (32) implies all the most important properties of EM algorithm (cf. Grim, 1982). First, the sequence of values $\{L^{(t)}\}$ is always nondecreasing with equality ocurring only at stationary points of the algorithm. If the sequence converges then the necessary condition of convergence

$$\lim_{t\to 0}\,(L^{(t+1)}-L^{(t)})=0$$

implies analogous conditions for the sequences $\{f^{(t)}\}$ and $\{q^{(t)}(\cdot|\boldsymbol{x})\}$, i.e.

$$\Rightarrow\ \lim_{t\to 0}\ ||f^{(t+1)}-f^{(t)}||=0,$$

$$\Rightarrow\ \lim_{t\to 0}\ ||q^{(t+1)}(\cdot|\boldsymbol{x})-q^{(t)}(\cdot|\boldsymbol{x})||=0,\quad \boldsymbol{x}\in\mathcal{X}.$$

A difficult point in application of EM algorithm is to specify the number of components of the estimated mixture and to choose their initial parameters. According to our practical experience this problem can be solved by successive adding of components: for a given M we iterate the EM algorithm until reasonable convergence and then add a new sufficiently "flat" and randomly placed component having a relatively high initial weight (e.g. $w_{M+1}=0.5$). Continuing computation we obtain again a monotonously converging sequence $L^{(t)}$ for the new enlarged mixture. In this way there is a chance to find out the data regions not sufficiently covered by the previous set of component densities. The increased initial weight w_{M+1} helps the new component to "survive" in competition with the old well "fitted" components. Note that, again, any interrupt of regular iterations may disturb the monotonous convergency. The adding of components may be continued until the weight of the new component is repeatedly suppressed despite the increased initial value.

6 Maximum-Likelihood Structuring

Unlike the preceding section let us consider now variable structural parameters $\phi_{mn}^{(t)}$. Consequently, we have to modify the difference in (31) as follows

$$L^{(t+1)}-L^{(t)}=I(f^{(t+1)}||f^{(t)})+\frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}I(q^{(t)}(\cdot|\boldsymbol{x})||q^{(t+1)}(\cdot|\boldsymbol{x}))+ \tag{33}$$

$$+\sum_{m\in\mathcal{M}}\frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}q^{(t)}(m|\,\boldsymbol{x})\log\left[\frac{G^{(t+1)}(\boldsymbol{x}|m,\phi_m^{(t+1)})}{G^{(t)}(\boldsymbol{x}|m,\phi_m^{(t)})}\right].$$

Further, introducing notation

$$\gamma_{mn}^{(t+1)}=\frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}q^{(t)}(m|\boldsymbol{x})\log\frac{f_n^{(t+1)}(x_n|m)}{f_n(x_n|0)}, \tag{34}$$

we can rearrange Eq. (33) as follows

$$L^{(t+1)} - L^{(t)} = I(f^{(t+1)}||f^{(t)}) + \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} I(q^{(t)}(\cdot|\boldsymbol{x})||q^{(t+1)}(\cdot|\boldsymbol{x}))+$$

$$+\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}(\phi_{mn}^{(t+1)} - \phi_{mn}^{(t)})\gamma_{mn}^{(t+1)}+ \tag{35}$$

$$+\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\frac{\phi_{mn}^{(t)}}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} q^{(t)}(m|\boldsymbol{x})\log\frac{f_n^{(t+1)}(x_n|m)}{f_n^{(t)}(x_n|m)}.$$

Obviously, the last sum in the expression (35) is nonnegative for any parameters $\phi_{mn}^{(t)}$ by Eq. (30). The preceding sum in (35) is maximized by setting

$$\phi_{mn}^{(t+1)} = \begin{cases} 1, & \gamma_{mn}^{(t+1)} > 0 \\ 0, & \gamma_{mn}^{(t+1)} \le 0 \end{cases} \tag{36}$$

and it is zero if $\phi_{mn}^{(t+1)} = \phi_{mn}^{(t)}$ for all $m \in \mathcal{M},\ n \in \mathcal{N}$. If the number of nonzero structural parameters ϕ_{mn} is fixed (or bounded), e.g.

$$\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}\phi_{mn} \le d$$

then the optimal subset of nonzero parameters $\phi_{mn}^{(t+1)}$ is defined by the d highest values $\gamma_{mn}^{(t+1)} > 0$.

Again, the iterative equations (28) - (30) with the additional step (34), (36) generate a nondecreasing sequence $\{L^{(t+1)}\}$ converging to a possibly local maximum of the log-likelihood function (27). In this sense, the corresponding parameters represent the (locally) optimal m.-l. approximation of the unknown probability density $P(\boldsymbol{x})$ (cf. Grim, 1986).

Remark. Let us note that in case of discrete distributions $f_n(x_n|m)$ we can write

$$f_n^{(t+1)}(\xi|m) = \frac{1}{|\mathcal{S}|f^{(t+1)}(m)}\sum_{x\in\mathcal{S}}\delta(\xi, x_n)q^{(t)}(m|\boldsymbol{x})$$

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

and the structural criterion $\gamma_{mn}^{(t+1)}$ can be expressed in terms of Kullback-Leibler information divergence

$$\gamma_{mn}^{(t+1)} = f(m)^{(t+1)} I(f_n^{(t+1)}(\cdot|m), f_n(\cdot|0)), \tag{37}$$

of the conditional distribution $f_n^{(t+1)}(x_n|m)$ with respect to the corresponding univariate “background” distribution $f_n(x_n|0)$. In this sense m.-l. structuring of discrete neural networks is naturally related to the information measure (37). In view of the above equations, the mixture $P^{(t+1)}$ includes only conditional densities $f_n^{(t+1)}(\cdot|m)$ which are specific and most “informative” with respect to the background marginals $f_n(\cdot|0)$.

In the original paper Grim (1986) the subspace approach of Sections 4 – 6 has been applied to an artificial decision problem to classify 16-dimensional binary vectors from two equiprobable populations ω_1, ω_2 described by Bernoulli mixtures. The parameters of both mixtures (each of them with three components) were chosen randomly (cf. Grim, 1986, p. 152, Table 1). Next, two data sets $\mathcal{S}_{\omega_1}, \mathcal{S}_{\omega_2}$ were generated according to the respective mixtures ($|\mathcal{S}_{\omega_1}| = |\mathcal{S}_{\omega_2}| = 6400$) and one half of each set was used to re-estimate the original parameters (p. 153, Table 2). The remaining 3200 binary vectors from each class were used to test the classification error independently. The theoretical classification error for the re-estimated parameters ($P_E = 0.072$, total number of parameters $r = 100$) was compared with the decision-making on the optimally chosen three-dimensional subspace ($P_E = 0.221, r = 36$) and with the structured model involving the same number of parameters ($P_E = 0.139, r = 36$). As it can be seen the structured model clearly outperforms the standard feature selection method. The last result has been still improved by optimizing the background distribution ($P_E = 0.111, r = 36$).

The subspace approach has been applied further to recognition of hand-written stylized numerals on a 32x32 binary rastr (Grim, 1986a). Again 400 numerals were used to estimate each of the ten class-conditional distributions on the original 1024-dimensional binary space and the remaining 400 numerals were used to test the classification performance. The class-conditional distributions were approximated by finite mixtures of the form (22) with only one component. The total number of parameters was chosen $r = 1000$, i.e. less then 10% of the corresponding full model ($r = 10240$) based on the assumption of conditional independence of variables. The classification error $P_E = 0.011$ obtained by using independent test sets was even better than that of the full model ($P_E = 0.012$).

Recently, the subspace approach has been applied to feature selection from multimodal data (cf. Novovičová *et al.*, 1996; Pudil *et al.*, 1995; Grim, 1986, Remark 4.2).

7 Information Preserving Transforms

In the context of multilayer probabilistic neural networks the transformation of signals between subsequent layers is of fundamental meaning. Let us recall that maximization of information transmission is one of the most widely used principles in artificial neural networks (cf. e.g. “infomax” method of Linsker, 1990). However, instead of maximizing information transmission between the input and output variables, we make use of the fact (cf. Grim, 1996a, Vajda & Grim, 1996) that the Shannon information $I(\mathcal{X}, \mathcal{M})$ about the descriptive decision problem on $\mathcal{X}$

$$I(\mathcal{X}, \mathcal{M}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{M}), \tag{38}$$

$$H(\mathcal{X}) = \sum_{x \in \mathcal{X}} -P(\boldsymbol{x}) \log P(\boldsymbol{x}), \quad H(\mathcal{X}|\mathcal{M}) = \sum_{m \in \mathcal{M}} f(m) \sum_{x \in \mathcal{X}} -F(\boldsymbol{x}|m) \log F(\boldsymbol{x}|m)$$

is automatically preserved by a special class of vector transforms $\boldsymbol{T} : \mathcal{X} \to \mathcal{Y},\ \mathcal{Y} \subset R^M$,

$$\boldsymbol{T}(\boldsymbol{x}) = (\, T_1(\boldsymbol{x}), T_2(\boldsymbol{x}), \ldots, T_M(\boldsymbol{x})) \in \mathcal{Y} \tag{39}$$

and particularly by the coordinate functions

$$y_m = T_m(\boldsymbol{x}) = \log f(m|\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}, \quad m \in \mathcal{M}. \tag{40}$$

Here $f(m|\boldsymbol{x})$, (cf. (6)) define the posterior distribution on $\mathcal{M}$ given $\boldsymbol{x} \in \mathcal{X}$. In other words we can write

$$I(\mathcal{X}, \mathcal{M}) = I(\mathcal{Y}, \mathcal{M}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{M}) \tag{41}$$

where $H(\mathcal{Y})$, $H(\mathcal{Y}|\mathcal{M})$ are the corresponding unconditional and conditional entropies of the transformed distributions

$$Q(\boldsymbol{y}) = P(\boldsymbol{T}^{-1}(\boldsymbol{y})), \quad Q(\boldsymbol{y}|m) = F(\boldsymbol{T}^{-1}(\boldsymbol{y})|m).$$

Simultaneously, the entropy $H(\mathcal{Y})$ of the transformed distribution $Q(\boldsymbol{y})$ is minimized on the class of all information preserving transforms.

Note that the information preserving transform actually "unifies" the points $\boldsymbol{x} \in \mathcal{X}$ with identical posterior distributions. Instead of logarithm we could use any bijective function but the logarithmic coordinate function makes the contributions from different input variables additive. In view of the formula (26) we can write (cf. (40))

$$\begin{aligned} \mathrm{T}_m(\boldsymbol{x}) = \log f(m|\boldsymbol{x}) = \log[G(\boldsymbol{x}|m, \phi_m) f(m)] - \\ - \log[\sum_{j \in \mathcal{M}} G(\boldsymbol{x}|j, \phi_j) f(j)], \quad m \in \mathcal{M} \end{aligned} \tag{42}$$

and further, making substitution (24), we obtain

$$y_m = \mathrm{T}_m(\boldsymbol{x}) = \log f(m) + \sum_{n \in \mathcal{N}} \phi_{mn} \log \frac{f_n(x_n|m)}{f_n(x_n|0)} - \log[\sum_{j \in \mathcal{M}} G(\boldsymbol{x}|j, \phi_j) f(j)]. \tag{43}$$

It is obvious that the inputs of the function $\mathrm{T}_m(\boldsymbol{x})$ (corresponding to the m-th neuron) can be confined to an arbitrary subset of variables x_n by means of the binary structural parameters ϕ_{mn}. As the optimal choice of these structural parameters can also be included into the EM algorithm (cf. (34), (36)), we can speak about maximum-likelihood structuring of neural networks.

Let us recall that, as it could be supposed in view of Eq. (7), the transform $\boldsymbol{T}$ preserves the original decision information $I(\mathcal{X}, \Omega)$, too (cf. Grim, 1996a). Thus, we can write $I(\mathcal{X}, \Omega) = I(\mathcal{Y}, \Omega)$ in analogy with (41). Simultaneously the information preserving transform minimizes the entropy of the output space $\mathcal{Y}$ and therefore it may be expected to simplify the underlying statistical decision problem.

In view of these arguments the repeated application of the procedure including estimation and information preserving transformation of the descriptive decision problem $\{\mathcal{X}, F(\cdot|m)f(m), m \in \mathcal{M}\}$ appears to be a reasonable method for a sequential design of multilayer optimally structured feedforward neural networks. In this sense the information preserving transform correspond to one layer of neural network.

A hidden layer of the probabilistic neural network transforms the descriptive decision problem without information loss. A multilayer feedforward neural network can be designed sequentially by repeated application of a procedure including (a) unsupervised m.-l. estimation of the descriptive distribution mixture and (b) information preserving transform of the descriptive decision problem.

8 Neurophysiological Aspects

Since very beginnings the research of neural networks is motivated by wonderful performance of biological neural networks like e.g. mammalian central nervous system. On the other hand, its elementary units - neurons are known to be relatively unreliable and inaccurate. For this reason it is assumed that the excellent properties of nervous systems have to be based on some very efficient and robust principles. From this point of view, one of the most interesting results of the present paper is the possibility to design generally structured neural networks and to obtain theoretical interpretation of some basic properties of neurons in the framework of a general statistical decision problem in a mathematically correct way - without making some arbitrary heuristical assumptions.

Let us recall that, in the present paper, we actually apply method of mixtures to a statistical decision problem. The transformation of signals between layers is information preserving and the final decision-making is based on Bayes formula. The idea of shared components and the use of logarithmic coordinate function are the only arbitrary steps motivated by neural networks. In view of these facts the possibility of neurophysiological interpretation of the information preserving transform is a strong argument for the proposed structural approach to be helpful to better understanding of biological neural systems.

From the neurophysiological point of view the probability $f(m|\boldsymbol{x})$ can be naturally interpreted as a measure of excitation or probability of firing of the m-th neuron given the input pattern $\boldsymbol{x} \in \mathcal{X}$. The output signal of the m-th neuron y_m is defined as logarithm of the excitation $f(m|\boldsymbol{x})$ and therefore logarithm plays the role of activation function or response curve.

Making use of Eq. (22) we can rewrite the formula (43) as follows

$$y_m = \mathrm{T}_m(\boldsymbol{x}) = \log f(m) + \sum_{n \in \mathcal{N}} \phi_{mn} \log \frac{f_n(x_n|m)}{f_n(x_n|0)} - \log[P(\boldsymbol{x})/\prod_{n \in \mathcal{N}} P_n(x_n)]. \tag{44}$$

Consequently, we may assume the first term in (44) to be responsible for spontaneous activity of the m-th neuron. This is well compatible with the EM algorithm which computes a priori probability $f(m)$ of “firing” of the m-th neuron by Eq. (29) as a mean value of $f(m|\boldsymbol{x})$, i.e. as a mean excitation.

The second term in Eq. (44) summarizes the contributions of the connected input neurons ($\phi_{mn} = 1$). In this sense, the term

$$\log \frac{f_n(x_n|m)}{f_n(x_n|0)} = \log f_n(x_n|m) - \log f_n(x_n|0) \tag{45}$$

can be viewed as the current synaptic weight of the n-th neuron at input of the m-th neuron - as a function of the input value x_n. Let us note that the “synaptic weight” in the formula (45) depends on the probability $f_n(x_n|m)$ and not directly on the variable x_n, i.e. it is defined as a composite function of x_n.

The effectiveness of the synaptic transmission, as expressed by the formula (45), combines the statistical properties of the input variable x_n with the activity of the "postsynaptic" neuron "m". In words, the synaptic weight (45) is high when the input signal x_n frequently takes part in excitation of the m-th neuron and, in turn, it is low when the input signal x_n usually doesn't contribute to the excitation of the m-th neuron. This formulation resembles the classical Hebb's postulate of learning (cf. Hebb, 1949, p.62):

When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic changes take place in one or both cells such that A's efficiency as one of the cells firing B, is increased.

The structural optimization of Sec.8 seems to have no efficient counterpart in biological neural systems. The structure of interconnections of biological neural networks can be assumed to be given and essentially fixed. The adaptive properties of neural systems are mainly enabled by the plasticity of existing synapses. This type of learning corresponds to the otimization of structured networks of Sec.7. On the other hand, the structural properties of biological neural systems has been optimized probably during the long process of phylogenetic evolution. From this point of view the structural optimization proposed in Sec.8 can be viewed as a short-cut alternative to obtain practical solutions of high quality.

Let us recall further that the norming included in computation of $f(m|\boldsymbol{x})$ corresponds well with the competitive behaviour of neuron assemblies. The last term in (44) includes the norming coefficient responsible for competitive properties of neurons and therefore it can be interpreted as a cumulative effect of lateral inhibition. This term is identical for all neurons and therefore, at the highest level, the classification tasks are not influenced by its accuracy.

Let us also note that mathematical expectation of the last "norming" term can be expressed by means of information-divergence

$$I(P(\cdot)||F(\cdot|0)) = E_P\{\log[P(\boldsymbol{x})/\prod_{n\in\mathcal{N}} P_n(x_n)]\} \tag{46}$$

which can be viewed as a measure of dependence of the involved variables. For independent variables the expression (46) is zero.

9 CONCLUDING REMARKS

In the present paper we propose a consistent probabilistic approach to optimize multilayer neural network structures. The method is based on repeated application of a procedure including (a) unsupervised m.-l. estimation of so called descriptive distribution mixture and (b) information preserving transformation of the descriptive decision problem. Only the last decision-oriented layer is assumed to be estimated in a supervised way. The structure of the network is optimized by m.-l. estimating the involved structural parameters. An application of the proposed method to recognition of binarized nonstylized numerals will be subject of a forthcoming paper.

Throughout the paper the optimization techniques are based exclusively on EM algorithm. Though all the design problems are posed in off-line form, there is a straightforward connection to learning procedures via sequential modifications of the EM algorithm (cf. e.g. Titterington et al., 1994). Let us remark in this connection that the assumed measure of excitation of the m-th neuron $f(m|\boldsymbol{x})$ plays a central role in the EM algorithm.

Let us recall finally that, by expanding the components of the information preserving transform, we obtained terms which may correspond to some basic functional properties of neurons like spontaneous activity, adaptivity of synaptic weights, Hebbian learning and lateral inhibition. In this sense the probabilistic neural networks could contribute to better understanding of biological neural systems.

Reference

[1] Ajvazjan, S.A., Bezhaeva, Z.I., & Staroverov, O.V. (1974). *Classification of Multivariate Observations,* (in Russian). Moscow: Statistika.

[2] Bishop, C.M. (1995). Neural Networks for Pattern Recognition. New York: Oxford University Press.

[3] 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.

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

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

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

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

[8] Grim, J. (1996). 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. (1996a). 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] Haykin, S. (1993). *Neural Networks: a comprehensive foundation.* San Mateo CA: Morgan Kaufman.

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

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

[13] 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.

[14] Isaenko, O.K., & Urbakh, K.I. (1976). Decomposition of probability distribution mixtures into their components (in Russian). In *Theory of probability, mathematical statistics and theoretical cybernetics* **13**, Moscow: VINITI.

[15] 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.

[16] 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.

[17] Lazarsfeld, P.F. (1966). Latent structure analysis. In Stouffer, Guttman, Slachman, Lazarsfeld, Star and Clausen (Eds.), *Measurement and Prediction.* New York: Wiley.

[18] Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annual Review of Neuroscience*, **13**, 257-281.

[19] Novovičová, J., Pudil, P., & Kittler, J. (1996). Divergence based feature selection for multimodal class densities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 2, 218-223.

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

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

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

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

[24] Pearson C. (1894). Contributions to the mathematical theory of evolution. 1. Dissection of frequency curves. *Philosophical Transactions of the Royal Society of London* **185**, 71-110.

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

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

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

[28] 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), 1389-1398.

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

[30] 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* **I** (pp. 525-532).

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

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

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

[34] Vajda, I., & Grim, J. (1996). About optimality of probabilistic basis function neural networks. Research Report UTIA, AS CR, No. 1887.

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

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

[37] Watanabe, S., & Pakvasa, N. (1973). Subspace method in pattern recognition. In *Proceeding International Joint Conference on Pattern Recognition*, (pp. 25-32).

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

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

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