DISCRETIZATION OF PROBABILISTIC NEURAL NETWORKS WITH BOUNDED INFORMATION LOSS

Jiří Grim

Institute of Information Theory and Automation
Pod vodárenskou věží 4, Prague

Abstract

In the framework of statistical classification the design of multilayer neural networks can be posed as a problem of approximating class conditional distributions by finite mixtures. The approach is based on the fact that at each layer the mixture components corresponding to units (neurons) define information preserving transform between consecutive layers. In this way the classification problem is transformed without information loss and the class-conditional distributions can be estimated at each layer again. In the present paper we prove that a logarithmic version of the transform is fault tolerant in the sense that a bounded inaccuracy of the estimated components may cause only a bounded information loss. Using this result we define discrete transforms with bounded information loss with the aim to reduce the complexity of estimating of the involved multidimensional probability density functions.

1 Introduction

The design of artificial neural networks is usually based on the concept of a formal neuron which is defined most frequently by means of a response function (unit step function or sigmoidal function) depending on a weighted sum of input variables. The properties of artificial neural networks can be optimized for some purpose in terms of the involved parameters.

[0] Early version of the paper: Grim J., "Discretization of probabilistic neural networks with bounded information loss." In: Computer-Intensive Methods in Control and Data Processing, pp. 205–210, (Preprints of the 3rd European IEEE Workshop CMP'98, Prague, September 7 – 9, 1998), J. Rojicek et al., eds., UTIA AV CR: Prague (1998)

In the framework of statistical classification there is an alternative probabilistic approach to optimization of neural networks. We approximate the unknown class-conditional probability density functions by finite mixtures and identify the component densities of mixtures with neurons. In this way we obtain the optimal parameters of neurons by maximum-likelihood (m.-l.) estimation of mixtures based on EM algorithm (cf. [3], [4]).

The possibility of a general design of layered neural networks by means of m.-l. estimation of distribution mixtures has been studied by many authors (cf. e.g. [5], [9], [15], [19], [20]). A recent reference to probabilistic neural networks is the paper of Streit and Luginbuhl [16] who proposed m.-l. training of a feed-forward layered neural network to recognize samples from different classes. They assume approximation of class-conditional distributions by mixtures of normal components - for each class separately. In this sense the class-conditional mixtures include different sets of components. Unfortunately, this approach would mean that, at higher levels of neural network, there are different (disjunct) sets of neurons for each class. In biological terms, this approach would correspond to non-overlapping receptive fields of output neurons.

To obtain a more realistic model of biological neural systems we suggested the concept of shared components with the corresponding modification of EM algorithm (cf. [6]). In this way the component distributions may be shared by all class-conditional mixtures and the corresponding connections to output units are not restricted.

The statistical classification based on class-conditional mixtures with shared components introduces naturally an additional decision problem which could be called descriptive. Each of the shared component distributions can be assumed to define an individual elementary class corresponding e.g. to some elementary situation on input. The secondary "descriptive" classes can be identified in usual way by a posteriori probabilities which are simply related to the primary classification problem.

In connection with optimization of multilayer neural networks there is a long history of information-theoretic approaches (cf. [9], [12]). In the context of probabilistic neural networks it has been shown (cf. [6], [18]) that the Shannon information contained in the descriptive decision problem can be automatically preserved by a special class of transforms based on mixtures. Roughly speaking, the information preserving transform defined by a vector of coordinate functions "unifies" the points of the input space having identical a posteriori probability distributions. Simultaneously, the transform minimizes the entropy of the transformed space.

It will be shown in the present paper that by choosing logarithmic coordinate functions, we obtain a transform which is fault tolerant in the sense that a bounded inaccuracy of the estimated component distributions may cause only bounded information loss. Using this result we show that the coordinate functions may be discretized while keeping the resulting information loss under an arbitrarily chosen positive bound.

We summarize first the basic properties of probabilistic neural networks (cf. Sec. 2-4). Then we prove the fault-tolerant property of information preserving transforms and define the discrete transform as a direct consequence. Loosely speaking, the information loss caused by discretization can be done arbitrarily small by increasing the number of discrete values.

2 Statistical classification based on finite mixtures

Assume that there is a finite set of K classes with a priori probabilities $q(\omega), \omega \in \Omega$ characterized by a corresponding finite set of conditional probability density functions $\{p(.|\omega),\ \omega \in \Omega\}$ on Ndimensional real space $\mathcal{X} = R^N$. We denote

$$\{\mathcal{X}, p(\cdot|\omega)q(\omega), \omega \in \Omega\}, \quad \Omega = \{\omega_1, \ldots, \omega_K\} \tag{1}$$

the related statistical classification problem on $\mathcal{X}$. In other words, any given point $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathcal{X}$ is to be classified with respect to the classes Ω.

Denoting $p(\boldsymbol{x})$ the unconditional probability density function

$$p(\boldsymbol{x}) = \sum_{\omega \in \Omega} p(\boldsymbol{x}|\omega)q(\omega), \quad \boldsymbol{x} = (x_1, \ldots, x_N) \in \mathcal{X}, \tag{2}$$

$$x_n \in \mathcal{X}_n = R, \quad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \mathcal{X}_N$$

we can write the formula for a posteriori probabilities

$$q(\omega|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega)q(\omega)}{p(\boldsymbol{x})}, \quad \omega \in \Omega. \tag{3}$$

They contain all statistical information about Ω given an input vector $\boldsymbol{x} \in \mathcal{X}$ and can be used to define a unique decision, if desirable.

In the framework of probabilistic neural networks (cf. [6], [7]) we assume that each conditional distribution $p(\boldsymbol{x}|\omega)$ can be approximated by a finite mixture with components from a common finite set of M probability density functions

$$\mathcal{F} = \{f(.|m), m \in \mathcal{M}\}, \quad \mathcal{M} = \{1, 2, \ldots, M\}. \tag{4}$$

where $\mathcal{M}$ denotes the index set. In particular we assume

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

$$\boldsymbol{x} \in \mathcal{X}, \quad \omega \in \Omega, \quad f(.|m) \in \mathcal{F}$$

where $w(m|\omega)$ are some conditional weights:

$$w(m|\omega) \geq 0, \quad \sum_{m \in \mathcal{M}} w(m|\omega) = 1,$$

$$w(m) = \sum_{\omega \in \Omega} w(m|\omega)q(\omega). \tag{6}$$

In this way the component densities $f(\boldsymbol{x}|m)$ may be "shared" by all class-conditional probability densities $p(\boldsymbol{x}|\omega)$.

The concept of shared components naturally corresponds with the structural properties of biological neural pathways since there is a rich branching of axons into multiple endings and, on the other hand, a large number of different axons converging at a single neuron. In view of these well known properties of biological neural networks any structural limitations implied by separate estimation of class conditional densities would be unrealistic.

By substituting (5), Eq. (2) can be rewritten in the form

$$p(\boldsymbol{x}) = \sum_{\omega\in\Omega} q(\omega) \left[\sum_{m\in\mathcal{M}} f(\boldsymbol{x}|m) w(m|\omega) \right] = \tag{7}$$

$$= \sum_{m\in\mathcal{M}} f(\boldsymbol{x}|m) w(m), \quad \boldsymbol{x} \in \mathcal{X}$$

where $w(m)$ is defined in (6). Let us note that the set $\mathcal{F}$ of component densities introduces an additional secondary "descriptive" decision problem with a priori probabilities $w(m)$. In this sense each component $f(\boldsymbol{x}|m)$ of the mixture (7) may correspond to an elementary situation on input. Given a vector $\boldsymbol{x} \in \mathcal{X}$, the presence or relevance of elementary input situations can be characterized by the posterior distribution on $\mathcal{M}$:

$$w(m|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|m)w(m)}{p(\boldsymbol{x})}, \quad m \in \mathcal{M}, \quad \boldsymbol{x} \in \mathcal{X}. \tag{8}$$

It is easy to see that there is a simple relationship between the a posteriori probabilities obtained for the primary and descriptive decision problem:

$$q(\omega|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega)q(\omega)}{p(\boldsymbol{x})} = \sum_{m\in\mathcal{M}} q(\omega|m) w(m|\boldsymbol{x}), \tag{9}$$

$$q(\omega|m) = \frac{w(m|\omega)q(\omega)}{w(m)}.$$

In view of the Bayes formula (9) the solution of practical decision problems can be confined to estimating the unknown component densities $f(\boldsymbol{x}|m)$ and the conditional weights $w(m|\omega)$. The procedure starting with estimation of a new distribution mixture on the transformed space can be repeated several times to optimize multilayer networks. The final decision-making based on identification of elementary input situations can be performed by linear combinations of a posteriori probabilities at the highest level.

Let us remark that, by introducing the concept of shared components in Eq. (5), we make an implicit assumption

$$f(\cdot|m,\omega) = f(\cdot|m), \quad \omega \in \Omega, \quad m \in \mathcal{M}. \tag{10}$$

Considering this identity we can easily verify (cf. [7]) that the Shannon information $I(\mathcal{X},\mathcal{M})$ about the set $\mathcal{M}$ contained in the space $\mathcal{X}$ is the same as the information about the direct product $\Omega \times \mathcal{M}$:

$$I(\mathcal{X}, \mathcal{M}\times\Omega) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{M}\times\Omega) \tag{11}$$

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

Here $H(\mathcal{X})$ and $H(\mathcal{X}|\mathcal{M})$ denote the related Shannon entropies. In other words the information $I(\mathcal{X},\mathcal{M})$ cannot be increased by the conditional weighting of components (5). Also from Eq. (5) it follows (cf. [6]) that the decision information $I(\mathcal{X},\Omega)$ is bounded by the "descriptive" information $I(\mathcal{X},\mathcal{M})$:

$$I(\mathcal{X},\Omega) \leq I(\mathcal{X},\mathcal{M}). \tag{12}$$

3 Hybrid estimation scheme

Numerically the finite mixture model of Sec. 2 can be optimized by means of EM algorithm (cf. [3], [4], [6]). 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 $p(\boldsymbol{x}|\omega)$:

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

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

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \left[\sum_{m \in \mathcal{M}} f(\boldsymbol{x}|m) w(m) \right], \tag{13}$$

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:

E-step: $(m \in \mathcal{M}, \ \boldsymbol{x} \in \mathcal{S})$

$$w(m|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|m) w(m)}{\sum_{j \in \mathcal{M}} f(\boldsymbol{x}|j) w(j)}, \tag{14}$$

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

$$w^{'}(m) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} w(m|\boldsymbol{x}), \tag{15}$$

$$f^{'}(\cdot|m) =$$

$$= \arg \max_{f(\cdot|m)} \{ \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} w(m|\boldsymbol{x}) \log f(\boldsymbol{x}|m) \}. \tag{16}$$

(Here apostrophe denotes the new values.)

The last implicit equation can be solved in a closed form in many cases of practical importance (cf. e.g. [4], [6]).

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

$$L_G = \frac{1}{|\mathcal{S}|} \sum_{\omega \in \Omega} \sum_{x \in \mathcal{S}_\omega} \log \ [p(\boldsymbol{x}|\omega) q(\omega)]. \tag{17}$$

We can avoid estimating $q(\omega)$ by setting $q(\omega) = |\mathcal{S}_\omega|/|\mathcal{S}|$ and further, making substitution (5), we obtain the criterion

$$\bar{L}_G = \frac{1}{|\mathcal{S}|} \sum_{\omega \in \Omega} \sum_{x \in \mathcal{S}_\omega} \log \left[\sum_{m \in \mathcal{M}} f(\boldsymbol{x}|m) w(m|\omega) \right] \tag{18}$$

and the corresponding iteration equations:

E-step: $(m \in \mathcal{M},\ \boldsymbol{x} \in \mathcal{S}_\omega,\ \ \omega \in \Omega)$

$$w(m|\boldsymbol{x},\omega) = \frac{f(\boldsymbol{x}|m)w(m|\omega)}{\sum_{j\in\mathcal{M}} f(\boldsymbol{x}|j)w(j|\omega)}, \tag{19}$$

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

$$w^{'}(m|\omega) = \frac{1}{|\mathcal{S}_\omega|}\sum_{x\in\mathcal{S}_\omega} w(m|\boldsymbol{x},\omega), \tag{20}$$

Here the component densities $f(\boldsymbol{x}|m)$ in (19)-(20) are assumed to be known from the unsupervised scheme (14)-(16) but they can be included into the supervised estimation by means of the equation

$$f^{'}(\cdot|m) = \tag{21}$$

$$= \arg\max_{f(\cdot|m)} \left\{ \frac{1}{|\mathcal{S}|}\sum_{\omega\in\Omega}\sum_{x\in\mathcal{S}_\omega} w(m|\boldsymbol{x},\omega)\log f(\boldsymbol{x}|m) \right\}.$$

Thus, to optimize the component densities $f(\boldsymbol{x}|m)$, we can use two obviously different iteration schemes, the unsupervised (14) - (16) and the supervised one (19) - (21). As it can be seen the two schemes yield identical results only asymptotically, provided that the concept of shared components is applicable and all the mixtures are uniquely identifiable.

In the present paper we prefer the “hybrid” estimation procedure (19) - (20) based on unsupervised estimates of the component densities $f(\boldsymbol{x}|m)$ as it is suitable for a sequential design of multilayer networks by means of information preserving transforms.

4 Information preserving transforms

A general functional principle to be modeled by artificial neural networks can be characterized as transformation of signals. It can be understood that in this connection many information theoretic ideas have been suggested and analyzed to optimize neural networks (cf. e.g. [1], [9], [10], [12]).

One of the most influential method is due to Linsker [12] who proposed a learning algorithm based on maximization of the average mutual information between the input and output vector of the neural network. The final “infomax” learning rule is derived to perform gradient ascent on the Shannon information criterion. Also additive Gaussian noise is to be assumed to avoid singular situations.

The “infomax” approach is well applicable in case of approximation problems when there is some known functional relation to be approximated or reproduced by output values. It is, however, less useful when the output values should correspond to some conditional probabilities.

In the framework of statistical decision problem essential property of any transform is to preserve the original decision information contained in the input space. [1] It has

[1] For the information preserving property and its information theoretic characterization we refer to Vajda [17].

been shown (cf. [5], [7]) that the Shannon information $I(\mathcal{X},\mathcal{M})$ about the descriptive decision problem on $\mathcal{X}$ is automatically preserved by any transform

$$\mathbf{T}:\mathcal{X}\to\mathcal{Y}, \quad \mathcal{Y}\subset R^M, \tag{22}$$

$$\mathbf{T}(\boldsymbol{x}) = (\mathrm{T}_1(\boldsymbol{x}),\mathrm{T}_2(\boldsymbol{x}),\ldots,\mathrm{T}_M(\boldsymbol{x}))\in\mathcal{Y}$$

defined by Eqs.

$$\mathrm{T}_m(\boldsymbol{x}) = \varphi_m(w(m|\boldsymbol{x})), \quad \boldsymbol{x}\in\mathcal{X}, \quad m\in\mathcal{M} \tag{23}$$

where $w(.|\boldsymbol{x})$, (cf. (8)) is the posterior distribution on $\mathcal{M}$ given $\boldsymbol{x}\in\mathcal{X}$ and φ_m are one-to-one mappings of the closed interval $< 0,1 >$ on a subset of the real line R. The transformation (23) preserves Shannon information in the sense that

$$I(\mathcal{X},\mathcal{M}) = I(\mathcal{Y},\mathcal{M}) \tag{24}$$

and simultaneously the Shannon entropy $H(\mathcal{Y})$ of the transformed distribution is minimized at the class of all transforms $\mathbf{T}$ satisfying (24). Loosely speaking, the transform "unifies" the points $\boldsymbol{x}\in\mathcal{X}$ with the identical posterior distribution $w(.|\boldsymbol{x})$. As the information preserving transform minimizes the entropy of the output space $\mathcal{Y}$, it may be expected to simplify solution of the transformed decision problem too. Recall that minimum entropy means a minimal source complexity, also in the sense of numerical description (cf. [14]).

For a discrete input space $\mathcal{X}$ the proof of the above assertions has already been published in [5]. In the paper [7] the theorem is modified by introducing the concept of shared components. A generalization to continuous input space can be found in [18]. It is also easily verified (cf. [7]) that the transform (23) preserves the decision information $I(\mathcal{X},\Omega)$, i.e. we have $I(\mathcal{X},\Omega) = I(\mathcal{Y},\Omega)$ in analogy with (24).

5 Fault-tolerant property

We prove first that, choosing logarithmic coordinate functions ϕ_m, we obtain an information preserving transform $\mathbf{T}$ which is fault-tolerant in the sense that a bounded approximation inaccuracy may cause only bounded information loss. The following theorem is a generalization of a similar assertion proved earlier (cf. [5]).

Theorem 5.1
Let $\mathcal{F} = \{F(.|m),\ m\in\mathcal{M}\}$ be a finite set of arbitrary probability distributions on a σ-algebra $\mathcal{A}$ of subsets of $\mathcal{X}$ with a priori probabilities $w(m), m\in\mathcal{M}$. We denote by $P(\cdot)$ the corresponding distribution mixture

$$P(\cdot) = \sum_{m\in\mathcal{M}} F(\cdot|m)w(m). \tag{25}$$

Further let $\mathbf{T}:\mathcal{X}\to\mathcal{Y},\ \mathcal{Y}\subset R^M$ be a measurable transform of the space $\mathcal{X}$

$$\boldsymbol{y} = \mathbf{T}(\boldsymbol{x}) = (\mathrm{T}_1(\boldsymbol{x}),\mathrm{T}_2(\boldsymbol{x}),\ldots,\mathrm{T}_M(\boldsymbol{x}))\in\mathcal{Y} \tag{26}$$

and let the coordinate functions T_m satisfy for some positive $\delta > 0,\ \epsilon > 0$ the inequality

$$|\mathrm{T}_m(\boldsymbol{x}) - \ln[w(m|\boldsymbol{x}) + w(m)\delta]| < \epsilon, \quad \boldsymbol{x}\in\mathcal{X} \tag{27}$$

where $w(\cdot|\boldsymbol{x})$ is the posterior distribution on $\mathcal{M}$ given $\boldsymbol{x} \in \mathcal{X}$. Then the information loss accompanying the transformation $\mathbf{T}$ is bounded by the inequality

$$I(\mathcal{X},\mathcal{M}) - I(\mathcal{Y},\mathcal{M}) =$$

$$= H(\mathcal{M}|\mathcal{Y}) - H(\mathcal{M}|\mathcal{X}) < \delta + 2\epsilon \tag{28}$$

where $H(\mathcal{M}|\mathcal{X}), H(\mathcal{M}|\mathcal{Y})$ are the corresponding conditional Shannon entropies.

Proof. We denote by $\mathcal{S}_T$ the partition of $\mathcal{X}$ induced by $\mathbf{T}$

$$\mathcal{S}_T = \{S_{\boldsymbol{y}} : \boldsymbol{y} \in \mathcal{Y}\},$$

$$S_{\boldsymbol{y}} = \mathbf{T}^{-1}(\boldsymbol{y}) = \{\boldsymbol{x} \in \mathcal{X} : \mathbf{T}(\boldsymbol{x}) = \boldsymbol{y}\} \tag{29}$$

and by $Q(\boldsymbol{y})$ and $G(\boldsymbol{y}|m), m \in \mathcal{M}$ the new distributions on a σ-algebra $\mathcal{B}$ of subsets of $\mathcal{Y}$ defined by $\mathbf{T}$:

$$G(\boldsymbol{y}|m) = F(\mathbf{T}^{-1}(\boldsymbol{y})|m), \quad \boldsymbol{y} = (y_1, \ldots, y_m) \in \mathcal{Y},$$

$$Q(\boldsymbol{y}) = P(\mathbf{T}^{-1}(\boldsymbol{y})) = \sum_{m \in \mathcal{M}} G(\boldsymbol{y}|m)w(m). \tag{30}$$

Since $y_m = \mathrm{T}_m(\boldsymbol{x})$ for all $\boldsymbol{x} \in S_{\boldsymbol{y}}$ we may rewrite the inequality (27) as follows (cf. (29))

$$y_m - \epsilon < \ln[w(m|\boldsymbol{x}) + w(m)\delta] \leq y_m + \epsilon, \tag{31}$$

$$\boldsymbol{x} \in S_{\boldsymbol{y}},\ \boldsymbol{y} \in \mathcal{Y},\ m \in \mathcal{M}.$$

The left-hand side of (31) may be further rewritten as follows

$$\exp\{y_m - \epsilon\} < w(m|\boldsymbol{x}) + w(m)\delta, \tag{32}$$

$$\boldsymbol{x} \in S_{\boldsymbol{y}},\ \boldsymbol{y} \in \mathcal{Y},\ m \in \mathcal{M}.$$

Conditional expectation $E_P\{.|S_{\boldsymbol{y}}\}$ of this inequality yields

$$\exp\{y_m - \epsilon\} < u(m|\boldsymbol{y}) + w(m)\delta, \tag{33}$$

for $\boldsymbol{y} \in \mathcal{Y},\ m \in \mathcal{M}$ where (cf. (8))

$$u(m|\boldsymbol{y}) = E_P\{w(m|\boldsymbol{x})|S_{\boldsymbol{y}}\} = \frac{G(\boldsymbol{y}|m)w(m)}{Q(\boldsymbol{y})}, \tag{34}$$

$$m \in \mathcal{M}, \quad \boldsymbol{y} \in \mathcal{Y}$$

is the posterior distribution on $\mathcal{M}$ given $\boldsymbol{y} \in \mathcal{Y}$. The inequality (33) can be rewritten in the form

$$y_m < \ln[u(m|\boldsymbol{y}) + w(m)\delta] + \epsilon \tag{35}$$

and further, taking in account the right-hand part of (31), we obtain

$$\ln w(m|\boldsymbol{x}) < \ln[u(m|\boldsymbol{y}) + w(m)\delta] + 2\epsilon \tag{36}$$

$$\boldsymbol{x} \in S_{\boldsymbol{y}},\ \boldsymbol{y} \in \mathcal{Y},\ m \in \mathcal{M}.$$

Again, multiplying the last inequality by $w(m|\boldsymbol{x})$, using the Jensen's inequality for the convex function $t \ln t$ and making the conditional expectation $E_P\{.|S_{\boldsymbol{y}}\}$, we obtain (cf. (34))

$$E_P\{w(m|\boldsymbol{x}) \ln w(m|\boldsymbol{x})|S_{\boldsymbol{y}}\} < \tag{37}$$

$$< u(m|\boldsymbol{y})\{\ln [u(m|\boldsymbol{y}) + w(m)\delta] + 2\epsilon\}.$$

Further, summing over $m \in \mathcal{M}$, we can write

$$-E_P\{H_{\boldsymbol{x}}(\mathcal{M})|S_{\boldsymbol{y}}\} < \tag{38}$$

$$< -H_{\boldsymbol{y}}(\mathcal{M}) + \sum_{m\in\mathcal{M}} u(m|\boldsymbol{y}) \ln[1 + \frac{w(m)\delta}{u(m|\boldsymbol{y})}] + 2\epsilon$$

where

$$H_{\boldsymbol{x}}(\mathcal{M}) = \sum_{m\in\mathcal{M}} -w(m|\boldsymbol{x}) \ln w(m|\boldsymbol{x}), \tag{39}$$

$$H_{\boldsymbol{y}}(\mathcal{M}) = \sum_{m\in\mathcal{M}} -u(m|\boldsymbol{y}) \ln u(m|\boldsymbol{y}).$$

In view of the inequality $\ln(1 + \xi) < \xi$, we obtain

$$H_{\boldsymbol{y}}(\mathcal{M}) - E_P\{H_{\boldsymbol{x}}(\mathcal{M})|S_{\boldsymbol{y}}\} < \sum_{m\in\mathcal{M}} w(m)\delta + 2\epsilon.$$

Finally, expectation E_Q of the last inequality can be rewritten in the form

$$E_Q\{H_{\boldsymbol{y}}(\mathcal{M})\} - E_P\{H_{\boldsymbol{x}}(\mathcal{M})\} =$$

$$= H(\mathcal{M}|\mathcal{Y}) - H(\mathcal{M}|\mathcal{X}) < \delta + 2\epsilon$$

which completes the proof of the assertion (28).

It can be seen that for ϵ and δ approaching zero the theorem implies information preserving property of the transform $\mathrm{T}_m(\boldsymbol{x}) = \ln w(m|\boldsymbol{x}), m \in \mathcal{M}$, as it has been discussed in Sec. 4 above. Let us note also that the added positive constant $w(m)\delta$ in (27) avoids possible singularity which could be caused by infinite values of logarithm near the zero point $w(m|\boldsymbol{x}) = 0$.

6 Discrete transforms with bounded information loss

Let us recall that the transformed space $\mathcal{Y}$ is a subset of M-dimensional real space R^M the dimension of which could be very high ($M \approx 10^1 - 10^2$). Consequently, in case of multilayer networks, the arising problem of estimation of probability density functions on the space $\mathcal{Y}$ could tend to be ill posed.

Note that by each new component distribution added to the model we increase the number of independent parameters and therefore, for a fixed sample size, there is an increasing risk of "overfitting" of the estimated distribution. One way how to avoid this problem of high dimensionality and to improve practical applicability of the transform $\mathbf{T}$, is to reduce the cardinality of the transformed space $\mathcal{Y}$, e.g. by making the coordinate functions ϕ_m discrete. In this connection we can use the fault tolerant property to discretize the probabilistic neural networks with a reasonably bounded information loss. The following Corollary is an immediate consequence of Theorem 5.1.

Corollary 6.1

Let $\mathbf{T} : \mathcal{X} \to \mathcal{Y}$ be a transform of the space $\mathcal{X}$ defined by the coordinate functions

$$y_m = \mathrm{T}_m(\boldsymbol{x}) = \psi_m(w(m|\boldsymbol{x})), \quad \boldsymbol{x} \in \mathcal{X} \tag{40}$$

where $\psi_m(\xi),\ m \in \mathcal{M}$ are stepwise approximations of the functions $\ln[\xi + w(m)\delta]$ with the positive parameters $\epsilon > 0, \delta > 0$. In particular, let

$$J_m = \{|\frac{1}{\epsilon}\ln[1 + \frac{1}{w(m)\delta}]|\} \tag{41}$$

be the number of discrete values of the function ψ_m defined by the equation

$$\psi_m(\xi) = \ln[w(m)\delta] + j\epsilon, \qquad 1 \le j \le J_m \tag{42}$$

for

$$w(m)\delta[e^{(j-1)\epsilon} - 1] \le \xi < w(m)\delta[e^{j\epsilon} - 1],$$

and by the equation

$$\psi_m(\xi) = \ln[1 + w(m)\delta],$$

for

$$w(m)\delta[e^{J_m\epsilon} - 1] \le \xi \le 1.$$

whereby notation $\{|..|\}$ stands for the integer part of the parenthesized expression. Then the transform (40) satisfies the condition (27) and therefore the corresponding information loss is bounded by $\delta + 2\epsilon$.

7 Concluding remarks

The present probabilistic approach to multilayer neural networks is characterized by information preserving transform of the underlying classification problem between consecutive layers. At each layer the class-conditional probability density functions are approximated by finite mixtures whereby the component densities (corresponding to neurons) define the coordinate functions of the transform. The dimensionality of the estimated densities is therefore given by the number of mixture components (neurons) at the preceding layer and can be relatively high ($M \approx 10^2$). By introducing discrete transforms we avoid the difficult problem of estimating multivariate (possibly ill posed) density functions and simplify the approximation problem. Instead of normal mixtures we can use more simple models, e.g. general discrete mixtures with product components.

Let us remark, however, that discretization based on the Corollary 6.1 is not suitable in case of binary variables ($J_m = 2$) since we could be forced to choose relatively large values of δ and ϵ (cf. (41)) with undesirable consequences. In view of its theoretical and practical relevance the problem of binary approximating will be subject of a particular paper [8].

References

[1] J.J. Atick, “Could information theory provide an ecological theory of sensory processing ?” *Network*, Vol. 3, pp. 213-251, 1992.

[2] J.S. Bridle, “Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters”. In: *Advances in Neural Information Processing Systems 2*, Ed.: D.S. Touretzky, San Mateo, CA: Morgan Kaufmann, pp. 211-217, 1990.

[3] A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm.” *J.Roy.Statist.Soc.* , **B 39**, pp.1-38, 1977.

[4] J. Grim, “On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions.” *Kybernetika*,**18**(3), pp.173–190, 1982.

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

[6] J. Grim, “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, 1996.

[7] J. Grim, “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*, 977-982, Roma: Edizzioni Kappa, 1996.

[8] J. Grim, P. Pudil, “On virtually binary nature of probabilistic neural networks.” To be presented at the *2nd International Workshop on Statistical Techniques in Pattern Rrecognition*, Sydney, August 11–13, 1998.

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

[10] J. Kay, “Feature discovery under contextual supervision using mutual information.” *International Joint Conference on Neural Networks*, Baltimore MD, Vol. 4, pp. 79-84, 1992.

[11] F. Liese and I. Vajda, *Convex Statistical Distances.* Teubner: Leipzig, 1987.

[12] R. Linsker, “Perceptual neural organization: Some approaches based on network models and information theory.” *Annual Review of Neuroscience*, Vol. 13, pp. 257-281, 1990.

[13] M.C. Pardo and I. Vajda, “Distances of probability distributions satisfying the information processing theorem of information theory”. *Trans. of IEEE on Information Theory*, (submitted) 1996.

[14] J. Rissanen, “Stochastic Complexity in Statistical Inquiry.” World Scientific: New Jersey, 1989.

[15] D.F. Specht, “Probabilistic neural networks for classification, mapping or associative memory”. In: *Proc. of the IEEE Int. Conference on Neural Networks*, July 1988, Vol. I., pp. 525-532, 1988.

[16] L.R. Streit and T.E. Luginbuhl, “Maximum likelihood training of probabilistic neural networks”. *IEEE Trans. on Neural Networks*, Vol. 5., No.5, pp. 764-783, 1994.

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

[18] I. Vajda and J. Grim, “About the maximum information and maximum likelihood principles”. *Kybernetika*, 1997, (to appear).

[19] S. Watanabe and K. Fukumizu, “Probabilistic design of layered neural networks based on their unified framework”. *IEEE Trans. on Neural Networks*, Vol. 6., No. 3, pp. 691-702, 1995.

[20] L. Xu and M.I. Jordan, “ EM learning on a generalized finite mixture model for combining multiple classifiers”, *World Congress on Neural Networks.*, Vol. 4, pp. 227-230, 1993.