Kybernetika

Igor Vajda; Jiří Grim

About the maximum information and maximum likelihood principles

Kybernetika, Vol. 34 (1998), No. 4, [485]--494

Persistent URL: http://dml.cz/dmlcz/135236

KYBERNETIKA — VOLUME 34 (1998), NUMBER 4, PAGES 485-494

ABOUT THE MAXIMUM INFORMATION AND MAXIMUM LIKELIHOOD PRINCIPLES IN NEURAL NETWORKS[1]

IGOR VAJDA AND JIŘÍ GRIM

Neural networks with radial basis functions are considered, and the Shannon information in their output concerning input. The role of information-preserving input transformations is discussed when the network is specified by the maximum information principle and by the maximum likelihood principle. A transformation is found which simplifies the input structure in the sense that it minimizes the entropy in the class of all information-preserving transformations. Such transformation need not be unique - under some assumptions it may be any minimal sufficient statistics.

1. INTRODUCTION

In this paper the attention is restricted to the important class of so-called radial basis function neural networks, which are intensively studied in the recent literature. These networks were introduced by Bromhead and Lowe (4). Other contributions can be found in Specht [25], Moody and Darken [19], Lowe [18], Casdagli [5], Poggio and Girosi [23], Xu et al [33], Streit and Luginbuhl [27], Watanabe and Fukumizu [31], Ukrainec and Haykin [29] and others. A systematic treatment can be found in Chap. 7 of Haykin [11] and Chap. 30 of Devroye et al [8].

A *radial basis function network* (RBF network) consists of several layers. The *input layer* is a collection of d real data sources

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d.$$

The second *hidden layer* consists of M units. The mth unit responds by

$$\phi_m(\boldsymbol{x}) = \frac{1}{\sigma_m}\mathcal{K}\left(\frac{\boldsymbol{x} - \boldsymbol{t}_m}{\sigma_m}\right), \quad 1 \le m \le M$$

to the input $\boldsymbol{x}$. Here $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$ is a probability density on $\mathbb{R}^d$, symmetric about $0 \in \mathbb{R}^d$ and called a *symmetrical kernel*, i.e.

$$\mathcal{K}(\boldsymbol{y}) = \varphi(||\boldsymbol{y}||^2) \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^d, \ \varphi : \mathbb{R} \to \mathbb{R},$$

[1]The present work has been supported by the Grant of the Academy of Sciences of the Czech Republic No. A2075703 and partially by the Complex Research Project of the Academy of Sciences of the Czech Republic No. K1075601.

$t_m \in \mathbb{R}^d$ is a center of symmetry of ϕ_m, and $\sigma_m > 0$ characterizes a dispersion of responses around this center. All hidden layer units are supposed to be activated by the identity function with a zero threshold, i.e., $(\phi_1(x), \ldots, \phi_M(x))$ is the output of the hidden layer. The *output layer* consists of K linear units (neurons) responding by

$$\rho_k = \sum_{m=1}^{M} w_{km}\phi_m, \quad 1 \le k \le K, \tag{1.1}$$

to the output $(\phi_1, \ldots, \phi_M) = (\phi_1(x), \ldots, \phi_M(x))$ of previous layer. The network output y is either the vector $(\rho_1, \ldots, \rho_K)$ a deterministic- or stochastic one.

The RBF networks are of a great practical interest because they can easily be realized. Indeed, the mth hidden unit can formally be decomposed into a simple linear network consisting of one layer of d nods with scalar inputs x_j, $1 \le j \le d$, and constant weights $1/\sigma_m$, using the identity activation functions and thresholds $t_{mj}/\sqrt{\sigma_m}$. The output neuron has weights w_j coinciding with the outputs $(x_j - t_{mj})/\sigma_m$ of the respective nods, and an activation function equal to the above considered φ with a zero threshold.

Possible example of φ is the continuous sigmoidal function

$$\varphi(r) = \begin{cases} (2\pi)^{-\frac{d}{2}} e^{-r/2} & \text{if } r \ge 0, \\ (2\pi)^{-\frac{d}{2}}(1 + e^{r/2}) & \text{otherwise,} \end{cases}$$

leading to the Gaussian symmetric kernel

$$\mathcal{K}(y) = (2\pi)^{-\frac{d}{2}} e^{-\|y\|^2/2}.$$

The hidden nonlinear layer can thus be replaced by two linear layers, so that the whole RBF network can be realized by a three-layer perceptron.

Similar three-layer perceptron realization (see Streit and Luginbuhl [27]) applies also to the more complicated networks with *anisotropic* RBF's. These differ from the above considered *isotropic* RBF's by that the argument $(x - t_m)\sigma_m^{-1}$ of $\mathcal{K}(\cdot)$ in the definition of $\phi_m(x)$ is replaced by $(x - t_m)B_m^{-1}$, i.e. that a regular $d \times d$ *norm weighting matrix* B_m stands at the place of "norm weighting scalar" σ_m (cf. pp. 258–259 in Haykin [11]). Then

$$\phi_j(x) = \frac{1}{|\det B_j|}\mathcal{K}\left((x - t_j)B_j^{-1}\right).$$

Thus, in particular, the Gaussian kernel leads to multivariate Gaussian probability densities with means t_j and covariance matrices $\Sigma_j = B_j^T B_j$.

Early information-theoretic analyses of perceptual system have been published soon after Shannon [26]. E.g. Attneave [2] analyzed visual perception on the basis of Shannon information, Uttley [30] suggested a network for adaptive pattern recogniton using the same information, and many other simillar thoughts may be found in various sources.

More recently Linsker [16, 17] proposed a learning method based on the *principle of maximum information preservation* (the *infomax principle*). This principle consists in the maximization of the average mutual information between input and output x and y of the neural network.

The concept of mutual information has been used also by other authors. Thus Plumbley and Fallside [22] formulated the maximum information preservation principle of Linsker as a minimization of information loss. They assumed the presence of additive Gaussian noise and analyzed a single-layer network to perform the dimensionality reduction. The information loss of their scheme is upper-bounded by the entropy of the reconstruction error and, in this way, the information loss limitation problem is related to the principal component analysis. Some implications of both information principles for neural network learning algorithms has been later analyzed in more details by Plumbley [21].

Atick and Redlich [1] have investigated the principle of minimum redundancy that applies to noisy channels. A linear matrix operator is optimized to minimize a specially introduced redundancy measure. Haykin [11] has shown that, despite the differences, the principle of minimum redundancy and the principle of maximum information preservation lead to similar results.

Kay [14] considered a neural network with input vector divided into primary and contextual part. Again the relationship between the primary- and the contextual subvectors is measured by the average mutual information to analyze the underlying structural dependences.

Becker and Hinton [3] have extended the idea of maximizing mutual information to unsupervised processing of the image of a natural scene. Specifically, their unsupervised learning procedure maximizes the mutual information between higher-level outputs with adjacent receptive fields. Inspired by the work of Becker and Hinton [3], Ukrainec and Haykin [29] developed an information-theoretic model for the enhancement of radar images. For more details about information theoretic approaches to neural networks se chapter 11 in Haykin [11].

The RBF networks are usually optimized by a hybrid method (cf. Hertz et al [12]) which means that only the weights of the third layer are trained in a supervised way whereas the training of the hidden-layer components is unsupervised. This approach has a good reason since consistent estimation of all network parameters is a difficult problem. Unfortunately the global performance of the RBF neural networks strongly depends on the quality of RBF. In particular the information loss caused by improper RBF cannot be repaired by optimizing the weights of the third layer (cf. Grim [10]).

For appropriately specified weights w_{km} the responses $\rho_k(x)$ are becoming mixtures of probability densities $\phi_1(x), \ldots, \phi_M(x)$ which can be viewed as approximations to given data generating probability densities $f_k(x), 1 \leq k \leq K$.

This opens the possibility to optimize the choice of RBF's by means of the maximum likelihood principle (ML principle). Of particular interest are iterative statistical schemes leading to maximum likelihood estimates of parameters of RBF's from given parametrized families. These schemes provide the possibility of iterative learning.

Since the late sixtieth there is an iterative computational scheme called EM algorithm (cf. Dempster et al [7]) which is widely applicable to estimation of mixtures. Design of RBF networks by means of EM algorithm has been studied e. g. by Jacobs and Jordan [13], Xu and Jordan [32], Haykin [11], Palm [20], Streit and Luginbuhl [27] and Watanabe and Fukumizu [31]).

For the sake of completness, let us mention that there exist also non-parametric statistical principles leading asymptotically to optimal RBF networks (see Devroye et al [9], in particular Chapter 30, and further references there in, and also Vajda and Grim [28]).

In this paper we are interested in the infomax and ML principles. A difficulty with their application arises when the dimensionality of the input grows. The complexity of application of both the gradient ascent on the input-output Shannon information in the case of infomax, and the EM agorithm in the case of ML, grows with the dimension d of the input space $\mathbb{R}^d$. By using the idea of statistical sufficiency, we characterize a class of transformations T of the input $\boldsymbol{x}$ which preserves the input-output Shannon information $I(\boldsymbol{x}; y)$, i.e. satisfies the relation

$$I(T(\boldsymbol{x}); y) = I(\boldsymbol{x}; y),$$

and has an entropy $H(T(\boldsymbol{x}))$ not greater than $H(\boldsymbol{x})$ and also not greater then $H(U(\boldsymbol{x}))$ for any transformation U with $I(U(\boldsymbol{x}); y) = I(\boldsymbol{x}; y)$. Since the minimal entropy means a minimal source complexity (in the sense of numerical description, see Risannen [24]) the class of transformations T is an important instrument for reduction of complexity of RBF neural networks.

2. THE RESULTS

Let the weights w_{km} of the above introduced RBF network be nonnegative, and let the network output Y be random in the sense that an output neuron $Y = k$ is selected to be fired with conditional probability

$$Pr(Y = k|\boldsymbol{x}) = \frac{\rho_k(\boldsymbol{x})}{\sum_{j=1}^{K} \rho_j(\boldsymbol{x})} \triangleq p_k(\boldsymbol{x}), \quad 1 \le k \le K. \tag{2.1}$$

Further let $\boldsymbol{X}$ be a random input distributed by a probability density $f(\boldsymbol{x})$ on $\mathbb{R}^d$. Then

$$\boldsymbol{p}(\boldsymbol{x}) = (p_1(\boldsymbol{x}), \ldots, p_K(\boldsymbol{x})) \tag{2.2}$$

is the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ and

$$\boldsymbol{p} = (p_1, \ldots, p_K) = \int_{\mathbb{R}^d} \boldsymbol{p}(\boldsymbol{x}) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

is an unconditional distribution of Y. Thus

$$I(\boldsymbol{X}; Y) = \int_{\mathbb{R}^d} \left[\sum_{k=1}^{K} p_k(\boldsymbol{x}) \log \frac{p_k(\boldsymbol{x})}{p_k} \right] f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{2.3}$$

is the Shannon input-output information.

According to the infomax principle, the information (2.3) is to be maximized. Based on this principle, several iterative learning rules for centers $\boldsymbol{t}_m$ of Gaussian

RBF s and for the weight matrices $W = (w_{km})$ performing a gradient ascent on the information have been proposed and succesively applied in the literature (see e.g. Haykin [11], pp. 459–460).

Let us suppose that the information (2.3) is finite. Denote by E the expectation with respect to the distribution $f(\boldsymbol{x})$ of $\boldsymbol{X}$ on $\mathbb{R}^d$, and consider the strictly concave function

$$h_K(\boldsymbol{u}) = h_K(u_1, \dots, u_K) = -\sum_{k=1}^{K} u_k \log u_k$$

in the domain $u \geq 0$ (with $0 \log 0 = 0$). By (2.3) ,

$$I(\boldsymbol{X}; Y) = H(Y) - H(Y|\boldsymbol{X}) = h_K(\boldsymbol{p}) - Eh_K(\boldsymbol{p}(\boldsymbol{X})).$$

If $T : \mathbb{R}^d \to \mathcal{T}$ is a measurable mapping into a space with σ-algebra $\mathcal{A}$ then

$$I(T(\boldsymbol{X}); Y) = H(Y) - H(Y|T(\boldsymbol{X})) = h_K(\boldsymbol{p}) - Eh_K(E\boldsymbol{p}(\boldsymbol{X})|T^{-1}\mathcal{A})).$$

As well known, $I(T(\boldsymbol{X}); Y) \leq I(\boldsymbol{X}; Y)$ so that

$$Eh_K(E(\boldsymbol{p}(\boldsymbol{X})|T^{-1}\mathcal{A})) \geq Eh_K(\boldsymbol{p}(\boldsymbol{X})). \tag{2.4}$$

Further, by Jensen's inequality for conditional expectations (cf. (A.16) on p. 208 of Liese and Vajda [15]),

$$h_K(E(\boldsymbol{p}(\boldsymbol{X})|T^{-1}\mathcal{A})) \geq h_K(\boldsymbol{p}(\boldsymbol{X})) \quad \text{a.s.} \tag{2.5}$$

and this relation holds with $\geq$ replaced by $=$ if and only if

$$\boldsymbol{p}(\boldsymbol{X}) = E(\boldsymbol{p}(\boldsymbol{X})|T^{-1}\mathcal{A}) \quad \text{a.s.} \tag{2.6}$$

Thus

$$I(T(\boldsymbol{X}); Y) = I(\boldsymbol{X}; Y) \tag{2.7}$$

implies that the equality in (2.4) takes place, so that (2.5) holds with the sign of equality and, consequently, (2.6) is satisfied. In other words, (2.7) implies $T^{-1}\mathcal{A}$-measurability of the function $p(\boldsymbol{x})$. It follows from here in particular that if $T_0(\boldsymbol{x}) = p(\boldsymbol{x})$ is a mapping from $\mathbb{R}^d$ into the simplex $\mathcal{T}_0 \subset \mathbb{R}^K$ of stochastic K-vectors with the σ-algebra $\mathcal{A}_0$ of Borel subsets then

$$T_0^{-1}\mathcal{A}_0 \subset T^{-1}\mathcal{A}. \tag{2.8}$$

Let now E and T be defined as above and consider the strictly convex function $\psi(u) = -\log u$ in the domain $u > 0$, naturally extended to $u = 0$. If

$$E\psi(f(\boldsymbol{x})) < \infty \quad \text{and} \quad Ef(\boldsymbol{x}) < \infty \tag{2.9}$$

then we define entropy of $T(\boldsymbol{X})$ by the formula

$$H(T(\boldsymbol{X})) = E\psi(E(f|T^{-1}\mathcal{A})).$$

By the above mentioned Jensen's inequality it holds for every T

$$H(T(\boldsymbol{X})) \leq H(\boldsymbol{X}) = E\psi(f(\boldsymbol{X})). \tag{2.10}$$

and $H(T(\boldsymbol{X})) \geq 1 - Ef(\boldsymbol{X}) > -\infty$ because $\psi(u) \geq 1 - u$. By the same inequality the inclusion $\mathcal{B}_0 \subset \mathcal{B}$ implies

$$\psi(E(f|\mathcal{B}_0)) = \psi(E(E(f|\mathcal{B})|\mathcal{B}_0)) \leq E(\psi(E(f|\mathcal{B}))|\mathcal{B}_0) \text{ a.s.,}$$

which in turn implies the monotonicity relation

$$E\psi(E(f|\mathcal{B}_0)) \leq E\psi(E(f|\mathcal{B})).$$

From here and (2.8) we obtain the following.

Assertion 1. If a measurable mapping T satisfies (2.7) and the RBF network input $\boldsymbol{X}$ satisfies (2.9) then

$$H(\boldsymbol{p}(\boldsymbol{X})) \leq H(T(\boldsymbol{X})). \tag{2.11}$$

If, morover, there exists a mapping T_0 such that $p(\boldsymbol{X}) = T_0(T(\boldsymbol{X}))$ then the equality takes place in (2.11).

The first statement follows from the fact that (2.7) implies that $\mathcal{B}_0 = T_0^{-1}\mathcal{A}_0$ is included in $\mathcal{B} = T^{-1}\mathcal{A}$ and, by the monotonicity mentioned above, $H(\boldsymbol{p}(\boldsymbol{X})) = E\psi(E(f|\mathcal{B}_0))$ at most equals $H(T(\boldsymbol{X})) = E\psi(E(f|\mathcal{B}))$.

The second statement follows from the first one and from the following obvious generalization of (2.10): if T is as in (2.10) and $\tilde{T} : \mathcal{T} \to \tilde{\mathcal{T}}$ is measurable then

$$H(\tilde{T}(T(\boldsymbol{X}))) \leq H(T(\boldsymbol{X})).$$

Example 1. Let us consider isotropic Gaussian RBF's centered at $\boldsymbol{t}_m \in I\!R^d$ with variances $\sigma^2 > 0$ and let the weight matrix $W = (w_{km})$ be stochastic. Then

$$\rho_k(\boldsymbol{x}) = \sum_{m=1}^{M} w_{km}(2\pi\sigma_m^2)\exp\left(-\left[n(\bar{\boldsymbol{x}} - \boldsymbol{t}_m)^2 + (n-1)s_{\boldsymbol{x}}^2\right]/(2\sigma_m^2)\right)$$

where $\bar{\boldsymbol{x}}$ and $s_{\boldsymbol{x}}^2$ are sample mean and variance specified explicitely below and

$$f(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\rho_k(\boldsymbol{x}).$$

By the first statement, the vector $p_k(\boldsymbol{x}), 1 \leq k \leq K$ achieves the minimal entropy among all random transforms $T(\boldsymbol{X})$ preserving the information. Since the bivariate statistic

$$T^*(\boldsymbol{x}) = (T_1^*(\boldsymbol{x}), T_2^*(\boldsymbol{x})) = (\bar{\boldsymbol{x}}, s_{\boldsymbol{x}}^2) = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i, \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^2\right)$$

is sufficient for the family $(\rho_k(\boldsymbol{x}) : 1 \leq k \leq K)$, it satisfies (2.7). Thus by the second statement, $T^*(\boldsymbol{X}) = (\bar{\boldsymbol{X}}, S^2)$ achieves the minimum entropy too.

Any optimization, iterative adaptation etc. can thus be based on collections of bivariate data $(\bar{\boldsymbol{X}}, S^2)$ instead of the d -variate data $\boldsymbol{X}$.

The model of Example 1 can be generalized as follows. Let $\boldsymbol{p} = (p_1, \ldots, p_K)$ be a stochastic K-vector, $C = (c_{km})$ a stochastic $K \times M$ matrix, and let the RBF network weights be defined by

$$w_{km} = p_k c_{km}.$$

Then the formula

$$f_k(\boldsymbol{x}) = \sum_{m=1}^{M} c_{km} \phi_m(\boldsymbol{x}). \tag{2.12}$$

defines a family $\mathcal{F} = (f_k : 1 \leq k \leq K)$ of probability densities on $\mathbb{R}^d$. In this case

$$\rho_k(\boldsymbol{x}) = p_k f_k(\boldsymbol{x}).$$

Let the density of network input $\boldsymbol{X}$ be given by the formula

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \rho_k(\boldsymbol{x}) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}). \tag{2.13}$$

The conditional distribution of the network output Y given $\boldsymbol{X} = \boldsymbol{x}$ then satisfies the relation

$$p_k(\boldsymbol{x}) = \frac{p_k f_k(\boldsymbol{x})}{f(\boldsymbol{x})}.$$

and the unconditional distribution is given by the K-vector $\boldsymbol{p}$.

The distributions $\langle \boldsymbol{p}, \mathcal{F} \rangle$ define a Bayesian statistical experiment described by a random parameter Θ distributed by $\boldsymbol{p}$ and a random observation $\boldsymbol{X}$ conditionally distributed by (2.12) given $\Theta = k$. The pair $(\Theta, \boldsymbol{X})$ has the same distribution as $(\boldsymbol{X}, Y)$. Therefore

$$I(\boldsymbol{X}; \Theta) = I(\boldsymbol{X}; Y).$$

If Θ is the input of channel C then the output is the random variable Z with

$$Pr(Z = m) = \sum_{k=1}^{K} p_k c_{km} \stackrel{\Delta}{=} q_m.$$

All random variables under consideration form a Markov chain $\Theta \to Z \to X \to Y$. It follows from here (cf. Theorem 2.8.1 in Cover and Thomas [6])

$$I(\Theta; X) \leq I(\Theta, Z; X) = I(Z; X). \tag{2.14}$$

For network inputs conditionally distributed by mixtures (2.12) with anisotropic Gaussian RBF's $\phi_m(\boldsymbol{x})$, Grim [10] studied the ML estimator of weights c_{km} and parameters implicitely figuring in functions $\phi_m(\boldsymbol{x})$. He established the convergence of EM algorithm leading to iterative specification of the network, based on independent samples of data $\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn}$ distributed by f_k for $1 \leq k \leq K$. In this context an important role plays the *descriptive Bayesian experiment* $\langle \boldsymbol{q} = (q_1, \ldots, q_M), \mathcal{F}_0 = (\phi_m : 1 \leq m \leq M)\rangle$, with the unknown parameter Z, sample $\boldsymbol{X}$ distributed conditionally under $Z = m$ by ϕ_m and unconditionally by

$$\sum_{m=1}^{M} q_m \phi_m(\boldsymbol{x}) = f(\boldsymbol{x}),$$

where $f(\boldsymbol{x})$ is given by (2.13) . In (2.14) , $I(Z; \boldsymbol{X})$ is the upper bound on the information $I(\Theta; X)$ concerning the inference parameter Θ. This shows that the quality of RBF's in the family $\mathcal{F}_0$ is limiting any further decision making, i.e. that a possible information loss caused by inaccurate components ϕ_m cannot be repaired by optimizing the weights w_{km}.

Preservation of the information $I(Z; \boldsymbol{X})$ during manipulations with data, and at the same time, the need to simplify the data structure, underline the importance of the assertion that follows. In this from the formal point of view special version of Assertion 1 we consider the conditional distribution $\boldsymbol{q}(\boldsymbol{x}) = (q_1(\boldsymbol{x}), \ldots, q_M(\boldsymbol{x}))$ of Z given $\boldsymbol{X} = \boldsymbol{x}$, given by

$$q_m(\boldsymbol{x}) = \frac{q_m \phi_m(\boldsymbol{x})}{f(\boldsymbol{x})}.$$

Note that the relation

$$I(Z; \boldsymbol{X}) = I(Z; T(\boldsymbol{X})) \tag{2.15}$$

holds for every statitic $T : \mathbb{R}^d \to \mathcal{T}$ which is sufficient for $\mathcal{F}_0$. Similarly it follows from (2.12) and (2.13) , and from the convexity of logarithmic function $\varphi(u)$ and of quadratic function $\psi(u) = u^2$ figuring in (2.9), that $\boldsymbol{X}$ satisfies (2.9) if

$$-\int_{\mathbb{R}^d} \phi_{\tilde{m}}(\boldsymbol{x}) \log \phi_m(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} \phi_m(\boldsymbol{x})^2 \, \mathrm{d}\boldsymbol{x} < \infty \tag{2.16}$$

for all $1 \leq m, \tilde{m} \leq M$.

Assertion 2. If a measurable mapping T satisfies (2.15) and the RBF's satisfy (2.16) then

$$H(\boldsymbol{q}(\boldsymbol{X})) \leq H(T(\boldsymbol{X})). \tag{2.17}$$

If, morover, there exists a mapping T_0 such that $\boldsymbol{q}(\boldsymbol{X}) = T_0(T(\boldsymbol{X}))$ then the equality takes place in (2.17).

Example 2. Let $\mathcal{F}_0$ be a family of exponential densities

$$\phi_m(\boldsymbol{x}) = h(\boldsymbol{x})\, c_m \exp\left(\sum_{j=1}^{J} w_{mj} \tau_j(\boldsymbol{x})\right)$$

satisfying (2.16). Since the statistic $\boldsymbol{\tau}(\boldsymbol{x}) = (\tau_1(\boldsymbol{x}), \ldots, \tau_J(\boldsymbol{x}))$ is sufficient for $\mathcal{F}_0$, it minimizes the entropy in the class of transformations $T(\boldsymbol{x})$ satisfying (2.15).

(Received December 18, 1997.)

REFERENCES

[1] J. J. Atick and A. N. Redlich: Towards a theory of early visual processing. Neural Computation *2* (1990), 308–320.

[2] F. Attneave: Some informational aspects of visual perception. Psychological Review *61* (1954), 183–193.

[3] S. Becker and G. E. Hinton: A self-organizing neural network that discovers surfaces in random-dot stereograms. Nature (London) *355* (1992), 161–163.

[4] D. S. Bromhead and D. Lowe: Multivariate functional interpolation and adaptive networks. Complex Systems *2* (1988), 321–355.

[5] M. Casdagli: Nonlinear prediction of chaotic time-series. Physica *35D* (1989), 335–356.

[6] T. M. Cover and J. B. Thomas: Elements of Information Theory. Wiley, New York 1991.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B *39* (1977), 1–38.

[8] L. Devroye and L. Győrfi: Nonparametric Density Estimation: The L_1 View. Wiley, New York 1985.

[9] L. Devroye, L. Győrfi and G. Lugosi: A Probabilistic Theory of Pattern Recognition. Springer, New York 1996.

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

[11] S. Haykin: Neural Networks: A Comprehensive Foundation. MacMillan, New York 1994.

[12] J. Hertz, A. Krogh and R. G. Palmer: Introduction to the Theory of Neural Computation. Addison-Wesley, New York - Menlo Park - Amsterdam 1991.

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

[14] J. Kay: Feature discovery under contextual supervision using mutual information. In: International Joint Conference on Neural Networks, Baltimore MD 1992, Vol. 4, pp. 79–84.

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

[16] R. Linsker: Self-organization in perceptual network. Computer *21* (1988), 105–117.

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

[18] D. Lowe: Adaptive radial basis function nonlinearities, and the problem of generalization. In: First IEE International Conference on Artificial Neural Networks, 1989, pp. 95–99.

[19] J. Moody and C. Darken: Fast learning in locally-tuned processing units. Neural Computation *1* (1989), 281–294.

[20] H. CH. Palm: A new method for generating statistical classifiers assuming linear mixtures of Gaussiian densities. In: Proc. 12th IAPR Internat. Conference on Pattern Recognition, IEEE Computer Society Press Jerusalem 1994, Vol. II., pp. 483–486.

[21] M. D. Plumbley: A Hebbian/anti-Hebbian network which optimizes information capacity by orthonormalizing the principle subspace. In: IEE Artificial Neural Networks Conference, ANN-93, Brighton 1992, pp. 86–90.

[22] M. D. Plumbley and F. Fallside: An information-theoretic approach to unsupervised connectionist models. In: Proceedings of the 1988 Connectionist Models Summer School, (D. Touretzky, G. Hinton and T. Sejnowski, eds.), Morgan Kaufmann, San Mateo 1988, pp. 239–245.

[23] T. Poggio and F. Girosi: Regularization algorithms for learning that are equivalent to multilayer networks. Science *247* (1990), 978–982.
[24] J. Rissanen: Stochastic Complexity in Statistical Inquiry. World Scientific, NJ 1989.
[25] D. F. Specht: Probabilistic neural networks for classification, mapping or associative memory. In: Proc. of the IEEE Internat. Conference on Neural Networks, 1988, Vol. I., pp. 525–532.
[26] C. E. Shannon: A mathematical theory of communication. Bell System Tech. J. *27* (1948), 379–423, 623–656.
[27] L. R. Streit and T. E. Luginbuhl: Maximum likelihood training of probabilistic neural networks. IEEE Trans. Neural Networks *5* (1994), 5, 764–783.
[28] I. Vajda and J. Grim: Bayesian optimality of decisions is achievable by RBF neural networks. IEEE Trans. Neural Networks, submitted.
[29] A. Ukrainec and S. Haykin: A modular neural network for unhancement of errors-polar radar targets. Neural Networks *9* (1996), 141–168.
[30] A. M. Uttley: The transmission of information and the effect of local feedback in theoretical and neural networks. Brain Research *102* (1966), 23–35.
[31] S. Watanabe and K. Fukumizu: Probabilistic design of layered neural networks based on their unified framework. IEEE Trans. Neural Networks *6* (1995), 3, 691–702.
[32] L. Xu and M. I. Jordan: EM learning on a generalized finite mixture model for combining multiple classifiers. In: World Congress on Neural Networks, 1993, Vol. 4, pp. 227–230.
[33] L. Xu, A. Krzyżak and E. Oja: Rival penalized competitive learning for clustering analysis, RBF net and curve detection. IEEE Trans. Neural Networks *4* (1993), 636–649.

Ing. Igor Vajda, DrSc. and Ing. Jiří Grim, CSc., Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mails: vajda@utia.cas.cz, grim@utia.cas.cz