DESIGN OF MULTILAYER NEURAL NETWORKS BY INFORMATION PRESERVING TRANSFORMS [1]

Jiří Grim

Institute of Information Theory and Automation, Czech Academy of Sciences
CZ - 18208 PRAGUE 8, P.O. BOX 18, Czech Republic
E-mail: grim@utia.cas.cz

Abstract: Information preserving transform based on finite mixture model is suggested to design multilayer neural networks. The information preserving transform minimizes the entropy of the output space and therefore simplifies the underlying statistical decision problem. In the framework of statistical decision-making the problem reduces to repeated m. - l. estimation of finite distribution mixtures. Formally the method can be interpreted as a theoretically well based approach to optimize radial basis function (RBF) neural networks in full generality.

1. INTRODUCTION

It appears that Broomhead and Lowe [1988] first suggested the use of radial basis functions (RBF) for the design of layered feed-forward neural networks. In its basic form the construction of RBF networks involves three different layers: the input layer of real sources $\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X},\ \mathcal{X} = R^N,$ the second "hidden" layer of radial basis functions $F(\boldsymbol{x}|m),\ \boldsymbol{x} \in \mathcal{X},\quad m \in \mathcal{M},\quad \mathcal{M} = \{1, 2, \ldots, M\}$ and the output layer performing a linear transformation from the hidden-unit space to the output space. The output units y_j are usually expressed as a weighted sum of RBF

$$y_j(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} w_{jm} F(\boldsymbol{x}|m),\ \boldsymbol{x} \in \mathcal{X},\quad j \in \mathcal{K},\quad \mathcal{K} = \{1, 2, \ldots, K\}. \tag{1}$$

Jacobs and Jordan [1991] associated with the output units a posteriori probabilities defined by a formula based on normal densities

$$y_j(\boldsymbol{x}) = \frac{w_j \exp\{-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{c}_j\|^2/2\sigma_j^2\}}{\sum_{m=1}^{M} w_m \exp\{-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{c}_m\|^2/2\sigma_m^2\}},\quad \boldsymbol{x} \in R^N,\quad j \in \mathcal{K}, \tag{2}$$

where w_j are nonnegative weights. A similar form of activation functions has been considered also by other authors (see e.g. Haykin [1994] for extensive references).

In the present paper we show that the heuristic formula (2) has an information-theoretic justification and can be used for stepwise optimal design of multilayer neural networks. In particular we introduce a special class of transforms which are information preserving provided that the underlying RBF define the true probability distribution on the input space $\mathcal{X}$. For this purpose the RBF may be estimated e.g. by EM algorithm (cf. Grim [1996]).

[1] Early version of the paper: Grim J., "Design of multilayer neural networks by information preserving transforms". In: Third European Congress on Systems Science. (Pessa E., Penna M. P., Montesanto A. eds.). Edizioni Kappa, Roma 1996, pp. 977-982.

1. STATISTICAL DECISION PROBLEM FOR NEURAL NETWORKS

We assume that on the N-dimensional real input space $\mathcal{X}$ there is a statistical decision problem $\{\mathcal{X}, P(.|\omega),\ \omega \in \Omega\}$ defined by a finite set of classes $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ with a priori probabilities $p(\omega)$, by the corresponding set of conditional probability density functions $\{P(.|\omega),\ \omega \in \Omega\}$ and by the unconditional density

$$P(\boldsymbol{x}) = \sum_{\omega \in \Omega} P(\boldsymbol{x}|\omega)p(\omega), \quad \boldsymbol{x} \in \mathcal{X}. \tag{3}$$

Solution of the statistical decision problem is assumed to be given by a posteriori probabilities

$$p(\omega|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega)p(\omega)}{P(\boldsymbol{x})}, \quad \omega \in \Omega, \quad \boldsymbol{x} \in \mathcal{X}, \quad (P(\boldsymbol{x}) > 0) \tag{4}$$

which may be used, if necessary, to optimize more complex decisions. For the sake of simplicity, we define $p(\omega|\boldsymbol{x}) = p(\omega)$ for $P(\boldsymbol{x}) = 0$.

Considering RBF networks we confine ourselves to approximating the unknown conditional densities $P(\boldsymbol{x}|\omega)$ by finite distribution mixtures (cf. Grim [1982,1996])

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

Here the components $F(\boldsymbol{x}|m)$ corresponding to RBF may be arbitrary probability density functions on $\mathcal{X}$. Taking in account that

$$f(m, \omega) = f(m|\omega)p(\omega), \quad f(m) = \sum_{\omega \in \Omega} f(m, \omega), \quad p(\omega) = \sum_{m \in \mathcal{M}} f(m, \omega), \tag{6}$$

we can write

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} F(\boldsymbol{x}|m)f(m), \quad \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

Let us note that the RBF defined by the mixture model (7) may correspond e.g. to some elementary properties or features. Thus, at a detailed level of description, the component densities $F(\boldsymbol{x}|m)$ naturally introduce an intermediate "descriptive" decision problem $\{\mathcal{X}, F(.|m), m \in \mathcal{M}\}$ with a priori probabilities $f(m)$. The occurrence of elementary properties or features can be "measured" by the a posteriori probabilities

$$f(m|\boldsymbol{x}) = \frac{F(\boldsymbol{x}|m)f(m)}{P(\boldsymbol{x})}, \quad m \in \mathcal{M}, \quad \boldsymbol{x} \in \mathcal{X}, \quad (P(\boldsymbol{x}) > 0). \tag{8}$$

which also imply the following solution of the primary decision problem:

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

A specific feature of the above scheme is the fact that the finite mixtures $P(\boldsymbol{x}|\omega)$ are defined over the same set of density functions, i.e. the component densities $F(\boldsymbol{x}|m)$ may be shared by the conditional distributions $P(\boldsymbol{x}|\omega)$. This property corresponds well with the structure of ascending neural pathways characterized by a rich branching of axons and their convergence on the neurons of subsequent layers. A simple consequence of shared components is the fact that the information $I(\mathcal{X}, \Omega)$ about Ω contained in $\mathcal{X}$ is bounded by the descriptive information

$I(\mathcal{X}, \mathcal{M})$ (cf. Grim [1996]). In view of the above mentioned aspects the information preserving transforms in the next section will be constructed with respect to the descriptive decision problem $\{\mathcal{X}, F(.|m),\ m \in \mathcal{M}\}$.

3. INFORMATION PRESERVING TRANSFORMS

In a recent paper Linsker [1989] proposed a learning method based on the principle of maximum information preservation. The fundamental idea of his influential work is to maximize the average mutual information between the input vector $\boldsymbol{x}$ and an output vector $\boldsymbol{y}$ of the neural network or, as redefined by Plumbley and Fallside [1988], to minimize possible information loss. [2] The "infomax" principle of Linsker has been used repeatedly to optimize layered neural networks (cf. Haykin [1994]) and many similar information-theoretic ideas have been published earlier.

Considering the framework of statistical decision-making we show that, in certain sense, the decision information contained in the input space $\mathcal{X}$ can be automatically preserved by a suitably chosen transform. The proof of the following theorem will be restricted to discrete variables, a generalization to continuous case (cf. Vajda and Grim [1996]) will be subject of a forthcoming paper.

Thus, in this section, we assume that $\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}$ is a vector of discrete finite valued variables $x_n \in \mathcal{X}_n, \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N$ and both the componets $F(\boldsymbol{x}|m)$ and the mixture $P(\boldsymbol{x})$ are discrete probability distributions. We consider a transform

$$\mathrm{T} : \mathcal{X} \to \mathcal{Y},\ \mathcal{Y} \subset R^M,\ \ \mathrm{T}(\boldsymbol{x}) = (\mathrm{T}_1(\boldsymbol{x}), \mathrm{T}_2(\boldsymbol{x}), \ldots, \mathrm{T}_M(\boldsymbol{x})) \in \mathcal{Y} \tag{10}$$

defined by the formula

$$\mathrm{T}_m(\boldsymbol{x}) = \varphi_m(F(m|\boldsymbol{x})),\ \ \boldsymbol{x} \in \mathcal{X},\ \ m \in \mathcal{M} \tag{11}$$

where $f(m|\boldsymbol{x})$, (cf. (8)) define the posterior distribution on M and φ_m are arbitrary one-to-one mappings of the closed interval $< 0, 1 >$ into the real line R. The transform (11) generates a partition of the space $\mathcal{X}$

$$\mathcal{S} = \{S_{\boldsymbol{y}}, \boldsymbol{y} \in \mathcal{Y}\},\ \ S_{\boldsymbol{y}} = \{\boldsymbol{x} \in \mathcal{X} : \mathrm{T}(\boldsymbol{x}) = \boldsymbol{y}\} \tag{12}$$

and defines new distributions on $\mathcal{Y}$

$$G(\boldsymbol{y}|m) = \sum_{\boldsymbol{x} \in S_y} F(\boldsymbol{x}|m), \quad Q(\boldsymbol{y}|\omega) = \sum_{\boldsymbol{x} \in S_y} P(\boldsymbol{x}|\omega), \tag{13}$$

$$Q(\boldsymbol{y}) = \sum_{\omega \in \Omega} Q(\boldsymbol{y}|\omega) p(\omega) = \sum_{m \in \mathcal{M}} G(\boldsymbol{y}|m) f(m) = \sum_{\boldsymbol{x} \in S_y} P(\boldsymbol{x}) = P(S_{\boldsymbol{y}}). \tag{14}$$

We prove the following assertion:

Theorem 3.1
The transformation (11) preserves the information in the sense that

$$I(\mathcal{X}, \mathcal{M}) = I(\mathcal{Y}, \mathcal{M}) \tag{15}$$

[2] For the information preserving property and its information theoretic characterization we refer to Pardo and Vajda [1996] and Vajda [1989].

and minimizes the output entropy

$$H(\mathcal{Y}) = \sum_{\boldsymbol{y}\in\mathcal{Y}} -Q(\boldsymbol{y}) \log Q(\boldsymbol{y}) \tag{16}$$

at the class of all transforms $\mathrm{T} : \mathcal{X} \to \mathcal{Y}$ satisfying (15).

Proof. In view of definitions (11),(12) we can write for any $m \in \mathcal{M}$ and $\boldsymbol{y} \in \mathcal{Y}$:

$$g(m|\boldsymbol{y}) = \frac{G(\boldsymbol{y}|m)f(m)}{Q(\boldsymbol{y})} = \sum_{\boldsymbol{x}\in S_y} \frac{P(\boldsymbol{x})}{P(S_y)} f(m|\boldsymbol{x}) = f(m|\boldsymbol{x}), \quad \text{for all} \quad \boldsymbol{x} \in S_y \tag{17}$$

Consequently, in the Jensen's inequality

$$\sum_{\boldsymbol{x}\in S_y} \frac{P(\boldsymbol{x})}{Q(\boldsymbol{y})} [-f(m|\boldsymbol{x}) \log f(m|\boldsymbol{x})] \leq -g(m|\boldsymbol{y}) \log g(m|\boldsymbol{y}), \quad m \in \mathcal{M}, \quad \boldsymbol{y} \in \mathcal{Y} \tag{18}$$

the equality takes place and, by summing over $m \in \mathcal{M}$ and $\boldsymbol{y} \in \mathcal{Y}$, we obtain

$$H(\mathcal{M}|\mathcal{X}) = \sum_{\boldsymbol{x}\in\mathcal{X}} P(\boldsymbol{x}) H_x(\mathcal{M}) = \sum_{\boldsymbol{y}\in\mathcal{Y}} Q(\boldsymbol{y}) H_y(\mathcal{M}) = H(\mathcal{M}|\mathcal{Y}). \tag{19}$$

The last equation proves the assertion (15) since we have

$$I(\mathcal{X}, \mathcal{M}) - I(\mathcal{Y}, \mathcal{M}) = H(\mathcal{M}|\mathcal{Y}) - H(\mathcal{M}|\mathcal{X}) = 0. \tag{20}$$

To prove the second part of the theorem we show that any information preserving transform U generates a partition $\mathcal{S}_{\mathrm{U}}$ of $\mathcal{X}$ which is identical with $\mathcal{S}_{\mathrm{T}}$ or is a refinement of $\mathcal{S}_{\mathrm{T}}$ - except for points $\boldsymbol{x} \in \mathcal{X}$ of zero probability ($P(\boldsymbol{x}) = 0$). Note that, in this way, any information preserving transform would satisfiy the desired inequality

$$H(\mathcal{Y}) = \sum_{\boldsymbol{y}\in\mathcal{Y}} -P(S_{\boldsymbol{y}}) \log P(S_{\boldsymbol{y}}) = H(\mathcal{S}_{\mathrm{T}}) \leq H(\mathcal{S}_{\mathrm{U}}). \tag{21}$$

Assume by contradiction that an information preserving transform U generates a partition $\mathcal{S}_{\mathrm{U}}$ which is not a refinement of $\mathcal{S}_{\mathrm{T}}$ in the sense, that for some set $\tilde{S} \in \mathcal{S}_{\mathrm{U}}$ and two different sets $S, S' \in \mathcal{S}_{\mathrm{T}}$ it holds $P(S \cap \tilde{S}) > 0, \quad P(S' \cap \tilde{S}) > 0$. In view of definitions (11) and (12) there are at least two points $\boldsymbol{x}, \boldsymbol{x}' \in \tilde{S}$ such that for some $m \in \mathcal{M}$ it holds

$$\boldsymbol{x} \in S \cap \tilde{S}, \quad P(\boldsymbol{x}) > 0, \qquad \boldsymbol{x}' \in S' \cap \tilde{S}, \quad P(\boldsymbol{x}') > 0, \quad f(m|\boldsymbol{x}) \neq f(m|\boldsymbol{x}'). \tag{22}$$

Consequently, in the relation (18) we obtain strict inequality for $\mathrm{T} \approx \mathrm{U}$ and $S_y \approx \tilde{S}$. Further, relation analogous to (19) holds with the strict inequality again and therefore Eq. (20) is not satisfied for the transformation U. This contradiction completes the proof.

4. CONCLUSION

The information preserving transform minimizes the entropy of the output space $\mathcal{Y}$ and therefore simplifies the underlying statistical decision problem. Roughly speaking, the transform unifies the points $\boldsymbol{x} \in \mathcal{X}$ with the identical posterior distribution $f(.|\boldsymbol{x})$. It is also easily verified that the transform (11) preserves the decision information $I(\mathcal{X}, \Omega)$, i.e., in analogy with (15) we can write $I(\mathcal{X}, \Omega) = I(\mathcal{Y}, \Omega)$.

It should be emphasized that, using mixtures, we have a theoretically well based possibility to optimize the involved RBF. Instead of maximizing complex information criteria we only need to compute m. - l. estimates of the component distributions $F(\boldsymbol{x}|m)$ by means of EM algorithm whereby the optimality of the resulting transform is automatically satisfied. The procedure may be applied repeatedly to design multilayer networks.

Reference

[1] Bromhead D.S., Lowe D., [1988], "Multivariate functional interpolation and adaptive networks". *Complex Systems*, Vol. 2., pp. 321–355

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

[3] Grim J. [1996], "Maximum Likelihood Design of Layered Neural Networks." To be presented at the *13th International Conference on Pattern Recognition*, Vienna, Austria, August 25–30, 1996.

[4] Haykin S. [1994], *Neural Networks: a comprehensive foundation.* Macmillan: New York.

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

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

[7] Linsker R. [1989], "How to generate ordered maps by maximizing the mutual information between input and output signals". *Neural Computation*, Vol. 1, pp. 402–411.

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

[9] Plumbley M.D. and Fallside F. [1988], "An information-theoretic approach to unsupervised connectionist models". In: *Proceedings of the 1988 Connectionist Models Summer School*, Eds. D. Touretzky et al., San Mateo, CA: Morgan Kaufmann, pp. 239–245.

[10] Vajda I. [1989], *Theory of Statistical Inference and Information.* Kluwer: Boston.

[11] Vajda I. and Grim J. [1996], "On information theoretic optimality of radial basis function neural networks". Research Report UTIA, AS CR, No. 1864.