Communicated by Steven J. Nowlan

Combining Exploratory Projection Pursuit and Projection Pursuit Regression with Application to Neural Networks

Nathan Intrator*
Institute for Brain and Neural Systems, Brown University, Box 1843, Providence, RI 02912 USA

We present a novel classification and regression method that combines exploratory projection pursuit (unsupervised training) with projection pursuit regression (supervised training), to yield a new family of cost/complexity penalty terms. Some improved generalization properties are demonstrated on real-world problems.

1 Introduction

Parameter estimation becomes difficult in high-dimensional spaces due to the increasing sparseness of the data. Therefore, when a low-dimensional representation is embedded in the data, dimensionality reduction methods become useful. One such method—projection pursuit regression (Friedman and Stuetzle 1981) (PPR)—is capable of performing dimensionality reduction by composition, namely, it constructs an approximation to the desired response function using a composition of lower dimensional smooth functions. These functions depend on low-dimensional projections through the data.

When the dimensionality of the problem is in the thousands, even projection pursuit methods are almost always overparameterized, therefore, additional smoothing is needed for low variance estimation. Exploratory projection pursuit (Friedman and Tukey 1974; Friedman 1987) (EPP) may be useful in these cases. It searches in a high-dimensional space for structure in the form of (semi)linear projections with constraints characterized by a projection index. The projection index may be considered as a universal prior for a large class of problems, or may be tailored to a specific problem based on prior knowledge.

In this paper, the general form of exploratory projection pursuit is formulated to be an additional constraint for projection pursuit regression. In particular, a hybrid combination of supervised and unsupervised artificial neural network (ANN) is described as a special case. In addition, a specific projection index that is particularly useful for classification (Intrator 1990; Intrator and Cooper 1992) is introduced in this context.

*Present address: Computer Science Department, Tel-Aviv University, Ramat-Aviv, 69978 Israel.

There have been many other attempts to combine unsupervised with supervised learning (Yamac 1969; Gutfinger and Sklansky 1991; Bridle and MacKay 1992). The formulation discussed below is based on projection pursuit ideas that generalize many of the classical statistical methods, and in our case, suggests a well-defined statistical framework, that allows formulation and comparison between these methods.

2 Brief Description of Projection Pursuit Regression

Let (X, Y) be a pair of random variables, $X \in R^d$, and $Y \in R$. The problem is to approximate the d-dimensional surface

$$f(x) = E[Y \mid X = x]$$

from n observations $(x_1, y_1), \ldots, (x_n, y_n)$.

PPR tries to approximate a function f by a sum of ridge functions (functions that are constant along lines)

$$f(x) \simeq \sum_{j=1}^{m} g_j(a_j^T x)$$

The fitting procedure alternates between an estimation of a direction $\hat{a}$ and an estimation of a smooth function $\hat{g}$, such that at iteration j, the square average of the residuals

$$r_{ij}(x_i) = r_{ij-1} - \hat{g}_j(\hat{a}_j^T x_i)$$

is minimized. This process is initialized by setting $r_{i0} = y_i$. Usually, the initial values of a_j are taken to be the first few principal components of the data.

Estimation of the ridge functions can be achieved by various nonparametric smoothing techniques such as locally linear functions (Friedman and Stuetzle 1981), k-nearest neighbors (Hall 1989b), splines, or variable degree polynomials. The smoothness constraint imposed on g implies that the actual projection pursuit is achieved by minimizing at iteration j, the sum

$$\sum_{i=1}^{n} r_{ij}^2(x_i) + C(\hat{g}_j)$$

for some smoothness measure C.

Due to the fact that the estimation of the nonparametric ridge functions is not decoupled from the estimation of the projections, overfitting is very likely to occur in one of the low-order $\hat{g}_j$, thereby invalidating subsequent estimations. Obviously, if $\hat{g}$ is not well estimated, the search for optimal projection direction will not yield good results.

Several alternatives have been considered in addressing this problem:

- Choose the ridge functions $\{g_j\}$ from a very small family of functions, for example, sigmoidals with a variable threshold. This eliminates the need to estimate the nonparametric ridge function, but increases the complexity of the architecture. This approach is widely used in artificial neural networks, and may partially explain their success.

- Estimate a fixed number of ridge functions and projections concurrently (as opposed to sequential estimation) provided that the ridge functions are taken from a very limited set of functions. Again this is used in the context of neural networks, due to the relatively small additional computational burden.

Additionally, one may attempt to

- Partially decouple the estimation of the response function, or the estimation of each of the ridge regression functions from the estimation of the projections.

Ultimately, it is reasonable to combine all of the above. One such implementation is presented in the following sections. First, the issue of decoupling the estimation of the ridge functions from the estimation of the projections is discussed.

3 Estimating the Projections Using Exploratory Projection Pursuit

Exploratory projection pursuit is based on seeking *interesting* projections of high-dimensional data points (Switzer 1970; Kruskal 1969, 1972; Friedman and Tukey 1974; Friedman 1987; Jones and Sibson 1987; Hall 1988; Huber 1985, for review). The notion of interesting projections is motivated by an observation that for most high-dimensional data clouds, most low-dimensional projections are approximately normal (Diaconis and Freedman 1984). This finding suggests that the important information in the data is conveyed in those directions whose single dimensional projected distribution is far from gaussian. Various projection indices (measures for the goodness of a projection) differ on the assumptions about the nature of deviation from normality, and in their computational efficiency. They can be considered as different priors motivated by specific assumptions on the underlying model.

To partially decouple the search for a projection vector from the search for a nonparametric ridge function, we propose to add a penalty term, which is based on a projection index, to the energy minimization associated with the estimation of the ridge functions and the projections. Specifically, let $\rho(a)$ be a projection index that is minimized for projections with a certain deviation from normality. At the jth iteration, we minimize the sum

$$\sum_i r_j^2(x_i) + C(g_j) + \rho(a_j)$$

When a concurrent minimization over several projections/functions is practical, we get a penalty term of the form

$$B(\hat{f}) = \sum_j [C(g_j) + \rho(a_j)]$$

Since C and ρ may not be linear, the more general measure that does not assume a stepwise approach, but instead seeks l projections and ridge functions concurrently, is given by

$$B(\hat{f}) = C(g_1, \ldots, g_l) + \rho(a_1, \ldots, a_l)$$

In practice, ρ depends implicitly on the training data (the empirical density) and is therefore replaced by its empirical measure $\hat{\rho}$.

3.1 Some Possible Measures. Some applicable projection indices have been discussed (Huber 1985; Jones and Sibson 1987; Friedman 1987; Hall 1989a; Intrator 1990). Probably, all the possible measures should emphasize some form of deviation from normality but the specific type may depend on the problem at hand. For example, a measure based on the Karhunen Loève expansion (Mougeot *et al.* 1991) may be useful for image compression with autoassociative networks, since in this case one is interested in minimizing the L^2 norm of the distance between the reconstructed image and the original one, and under mild conditions, the Karhunen Loève expansion gives the optimal solution.

A different type of prior knowledge is required for classification problems. The underlying assumption then is that the data are clustered (when projecting in the right directions) and that the classification may be achieved by some (nonlinear) mapping of these clusters. In such a case, the projection index should emphasize multimodality as a specific deviation from normality. A projection index that emphasizes multimodalities in the projected distribution (without relying on the class labels) has recently been introduced (Intrator 1990) and implemented efficiently using a variant of a biologically motivated unsupervised network (Intrator and Cooper 1992). Its integration into a backpropagation classifier will be discussed below.

4 A Variant of Projection Pursuit Regression: Backpropagation Network

In this section, we consider a parametric approach—the backpropagation network—as a variant of PPR. In this context the addition of an exploratory projection index is discussed.

Backpropagation (Werbos 1974; Le Cun 1985; Rumelhart *et al.* 1986) has been chosen as a possible representative for the first two alternatives presented in Section 2, since it has become a useful tool for solving complicated pattern recognition tasks such as speech recognition (Lippmann 1989), and since the class of functions that can be approximated by a backpropagation type network is very large. This architecture (with an unlimited number of projections) can uniformly approximate arbitrary continuous functions on compact sets (Cybenko 1989; Hornik *et al.* 1989) as well as their derivatives (Hornik *et al.* 1990), and do so efficiently. Related results can be found (Carroll and Dickinson 1989; Funahashi 1989; Hecht-Nielsen 1989; Hornik 1991; Ito 1991).

In this method, the error is efficiently propagated backward to the previous layer for modification of their synaptic weights (projections). The single hidden layer architecture is of the form

$$f(x) = \sum_j \beta_j \sigma \left(\sum_{k=1}^{d} \omega_{jk} x_k + w_{j,0} \right)$$

where σ is an arbitrary (fixed) bounded monotone function. The form

$$f(x) = \sigma \left[\sum_j \beta_j \sigma \left(\sum_{k=1}^{d} \omega_{jk} x_k + w_{j,0} \right) \right]$$

is more suitable for classification tasks.

Since this method can approximate any continuous function, great care should be taken so that the variance of the estimator is not large, namely, that the model does not "overfit" the training data (Wahba 1990; Geman *et al.* 1992, for discussion). This can be done using some form of complexity regularization (Barron and Barron 1988; Barron 1989; White 1990; Moody 1991) or by weight elimination penalties that aim to reduce the effective number of parameters in the model (Plaut *et al.* 1986; Mozer and Smolensky 1989; Le Cun *et al.* 1990; Weigend *et al.* 1991).

The performance of the network is measured using a loss criterion, for example, mean squared error between the output and the target of the network (the class label). The estimation of the weights is done by minimizing the empirical average of the error via gradient descent of the form: $\partial w_{ij}/\partial t = -\partial \mathcal{E}/\partial w_{ij}$, where $\mathcal{E} = E_x[\mathcal{E}(x,\omega)]$, is the average contribution to the loss criterion of each of the random inputs x.

4.1 Adding EPP Constraints to Backpropagation Network. One way of adding some prior knowledge into the architecture is by minimizing the effective number of parameters using weight sharing, in which a single weight is shared among many connections in the network (Waibel *et al.* 1989; Le Cun *et al.* 1989). An extension of this idea is the "soft weight sharing," which favors irregularities in the weight distribution in the form of multimodality (Nowlan and Hinton 1992). This penalty

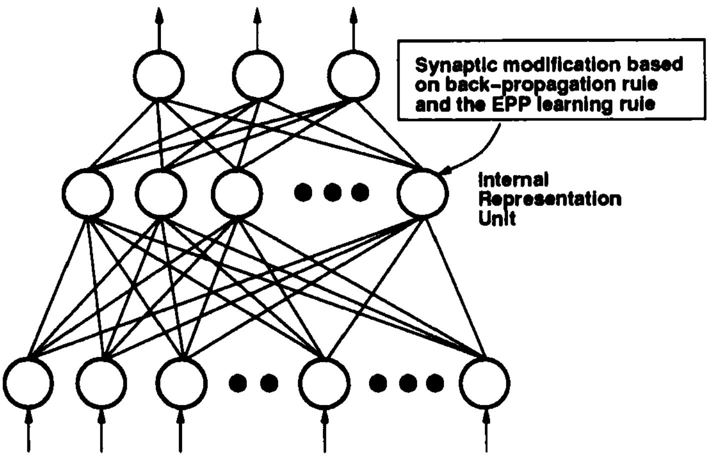

Figure 1: A hybrid EPP/PPR neural network (EPPNN).

improved generalization results obtained by weight elimination penalty. Both these methods make an explicit assumption about the structure of the weight space, but with no regard to the structure of the input space.

As described in the context of projection pursuit regression, a penalty term may be added to the energy functional minimized by error back-propagation, for the purpose of measuring directly the goodness of the projections sought by the network. Since our main interest is in reducing overfitting for high-dimensional problems, our underlying assumption is that the surface function to be estimated can be faithfully represented using a low-dimensional composition of sigmoidal functions, namely, using a backpropagation network in which the number of hidden units is *much smaller* than the number of input units. Therefore, the penalty term may be added only to the hidden layer (see Fig. 1). The synaptic modification equations of the hidden units' weights become

$$\frac{\partial w_{ij}}{\partial t} = -\epsilon\Bigg[\frac{\partial\mathcal{E}(w,x)}{\partial w_{ij}} + \frac{\partial\rho(w_1,\ldots,w_n)}{\partial w_{ij}} + (\text{contribution of cost/complexity terms})\Bigg]$$

An approach of this type has been used in image compression, with a penalty aimed at minimizing the entropy of the projected distribution (Bichsel and Seitz 1989). This penalty certainly measures deviation from normality, since entropy is maximized for a gaussian distribution.

5 Projection Index for Classification: The Unsupervised BCM Neuron

Intrator (1990) has recently shown that a variant of the Bienenstock, Cooper, and Munro neuron (BCM) (Bienenstock *et al.* 1982) performs exploratory projection pursuit using a projection index that measures multimodality. This neuron version allows theoretical analysis of some visual deprivation experiments (Intrator and Cooper 1992), and is in agreement with the vast experimental results on visual cortical plasticity (Clothiaux *et al.* 1991). A network implementation that can find several projections in parallel while retaining its computational efficiency, was found to be applicable for extracting features from very high-dimensional vector spaces (Intrator and Gold 1992; Intrator *et al.* 1991; Intrator 1992).

The activity of neuron k in the network is $c_k = \sum_i x_i w_{ik} + w_{0k}$. The *inhibited* activity and threshold of the kth neuron is given by

$$\tilde{c}_k = \sigma\left(c_k - \eta \sum_{j \neq k} c_j\right), \qquad \tilde{\Theta}_m^k = E[\tilde{c}_k^2]$$

The threshold $\tilde{\Theta}_m^k$ is the point at which the modification function ϕ changes sign (see Intrator and Cooper 1992 for further details). The function ϕ is given by

$$\phi(c, \Theta_m) = c(c - \Theta_m)$$

The risk (projection index) for a single neuron is given by

$$R(w_k) = -\left\{\frac{1}{3}E[\tilde{c}_k^3] - \frac{1}{4}E^2[\tilde{c}_k^2]\right\}$$

The total risk is the sum of each local risk. The negative gradient of the risk that leads to the synaptic modification equations is given by

$$\frac{\partial w_{ij}}{\partial t} = E\left[\phi(\tilde{c}_j, \tilde{\Theta}_m^j)\sigma'(\tilde{c}_j)x_i - \eta \sum_{k \neq j} \phi(\tilde{c}_k, \tilde{\Theta}_m^k)\sigma'(\tilde{c}_k)x_i\right]$$

This last equation is an additional penalty to the energy minimization of the supervised network. Note that there is an interaction between adjacent neurons in the hidden layer. In practice, the stochastic version of the differential equation can be used as the learning rule.

5.1 Some Related Statistical and Computational Issues of This Projection Index. This section discusses some commonly asked questions regarding the connection of the above projection index to previous work in pattern recognition and statistics.

Although the projection index is motivated by the desire to search for clusters in the high-dimensional data, the resulting feature extraction method is quite different from other pattern recognition methods that search for clusters. Since the class labels are not used in the search, the projection pursuit is not biased to the class labels. This is in contrast with classical methods such as discriminant analysis (Fisher 1936; Sebestyen 1962, and numerous recent publications).

The projection index concentrates on projections that allow discrimination between clusters and not faithful representation of the data. This is in contrast to principal components analysis, or factor analysis, which tend to combine features that have high correlation (see review in Harman 1967). The method differs from cluster analysis by the fact that it searches for clusters in the low-dimensional projection space, thus avoiding the inherent sparsity of the high-dimensional space.

The projection index uses low-order polynomial moments, which are computationally efficient, yet it does not suffer from the main drawback of polynomial moments—sensitivity to outliers. It naturally extends to multidimensional projection pursuit using the feedforward inhibition network. The number of calculations of the gradient grows linearly with the dimensionality and *linearly* with the number of projections sought.

6 Applications

We have applied this hybrid classification method to various speech and image recognition problems in high-dimensional space. In one speech application we used voiceless stop consonants extracted from the TIMIT database as training tokens (Intrator and Tajchman 1991). A detailed biologically motivated speech representation was produced by Lyon's cochlear model (Lyon 1982; Slaney 1988). This representation produced 5040 dimensions (84 channels $\times$ 60 time slices). In addition to an initial voiceless stop, each token contained a final vowel from the set [aa, ao, er, iy]. Classification of the voiceless stop consonants using a test set that included 7 vowels [uh, ih, eh, ae, ah, uw, ow] produced an average error of 18.8% while on the same task classification using backpropagation network produced an average error of 20.9% (a significant difference, $p < 0.0013$). Additional experiments on vowel tokens appear in Tajchman and Intrator (1992).

Another application is in the area of face recognition from gray level pixels (Intrator *et al.* 1992). After aligning and normalizing the images, the input was set to 37×62 pixels (total of 2294 dimensions). The recognition performance was tested on a subset of the MIT Media Lab database of face images made available by Turk and Pentland (1991) which contained 27 face images of each of 16 different persons. The images were taken under varying illumination and camera location. Of the 27 images available, 17 randomly chosen ones served for training and the remaining 10 were used for testing. Using an ensemble average of hybrid networks (Lincoln and Skrzypek 1990; Pearlmutter and Rosenfeld 1991; Perrone and Cooper 1992) we obtained an error rate of 0.62% as opposed to 1.2% using a similar ensemble of backpropagation networks. A single backpropagation network achieves an error between 2.5 and 6% on these data. The experiments were done using 8 hidden units.

7 Summary

A penalty that allows the incorporation of additional prior information on the underlying model was presented. This prior was introduced in the context of projection pursuit regression, classification, and in the context of backpropagation network. It achieves partial decoupling of estimation of the ridge functions (in PPR) or the regression function in backpropagation net from the estimation of the projections. Thus it is potentially useful in reducing problems associated with overfitting, which are more pronounced in high-dimensional data.

Some possible projection indices were discussed and a specific projection index that is particularly useful for classification was presented in this context. This measure that emphasizes multimodality in the projected distribution was found useful in several very high-dimensional problems.

Acknowledgments

I wish to thank Leon Cooper, Stu Geman, and Michael Perrone for many fruitful conversations and the referee for helpful comments. The speech experiments were performed using the computational facilities of the Cognitive Science Department at Brown University. Research was supported by the National Science Foundation, the Army Research Office, and the Office of Naval Research.

References

Barron, A. R. 1989. Statistical properties of artificial neural networks. In *Proc. IEEE Conf. on Decision and Control*, pp. 280–285. IEEE Press, New York.

Barron, A. R., and Barron, R. L. 1988. Statistical learning networks: A unifying view. In *Computing Science and Statistics: Proc. 20th Symp. Interface*, E. Wegman, ed., pp. 192–203. American Statistical Association, Washington, DC.

Bichsel, M., and Seitz, P. 1989. Minimum class entropy: A maximum information approach to layered networks. *Neural Networks* **2**, 133–141.

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. 1982. Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *J. Neurosci.* **2**, 32–48.

Bridle, J. S., and MacKay, D. J. C. 1992. Unsupervised classifiers, mutual information and 'Phantom Targets'. In *Advances in Neural Information Processing*

Systems, Vol. 4, J. Moody, S. Hanson, and R. Lippmann, eds., pp. 1096–1101. Morgan Kaufmann, San Mateo, CA.

Carroll, S. M., and Dickinson, B. W. 1989. Construction of neural net using the radon transform. In *International Joint Conference on Neural Networks*, Vol. 1, pp. 607–611. IEEE Press, New York.

Clothiaux, E. E., Cooper, L. N., and Bear, M. F. 1991. Synaptic plasticity in visual cortex: Comparison of theory with experiment. *Journal of Neurophysiology* **66**, 1785–1804.

Cybenko, G. 1989. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* **2**, 303–314.

Diaconis, P., and Freedman, D. 1984. Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188.

Friedman, J. H. 1987. Exploratory projection pursuit. *J. Am. Statist. Assoc.* **82**, 249–266.

Friedman, J. H., and Stuetzle, W. 1981. Projection pursuit regression. *J. Am. Statist. Assoc.* **76**, 817–823.

Friedman, J. H., and Tukey, J. W. 1974. A projection pursuit algorithm for exploratory data analysis. *IEEE Transact. Computers* **C**(23), 881–889.

Funahashi, K. 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks* **2**, 183–192.

Geman, S., Bienenstock, E., and Doursat, R. 1992. Neural networks and the bias-variance dilemma. *Neural Comp.* **4**, 1–58.

Gutfinger, D. and Sklansky, J. 1991. Robust classifiers by mixed adaptation. *IEEE Transact. Pattern Anal. Machine Intelligence* **13**, 552–567.

Hall, P. 1988. Estimating the direction in which data set is most interesting. *Probab. Theory Rel. Fields* **80**, 51–78.

Hall, P. 1989a. On polynomial-based projection indices for exploratory projection pursuit. *Ann. Statist.* **17**, 589–605.

Hall, P. 1989b. On projection pursuit regression. *Ann. Statist.* **17**, 573–588.

Harman, H. H. 1967. *Modern Factor Analysis*, 2nd ed. University of Chicago Press, Chicago.

Hecht-Nielsen, R. 1989. Theory of the backpropagation neural network. In *International Joint Conference on Neural Networks*, Vol. 1, pp. 593–606. IEEE Press, New York.

Hornik, K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4**, 251–257.

Hornik, K., Stinchombe, M., and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* **2**, 359–366.

Hornik, K., Stinchombe, M., and White, H. 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks* **3**, 551–560.

Huber, P. J. 1985. Projection pursuit. (with discussion). *Ann. Statist.* **13**, 435–475.

Intrator, N. 1990. Feature extraction using an unsupervised neural network. In *Proceedings of the 1990 Connectionist Models Summer School*, D. S. Touretzky,

J. L. Ellman, T. J. Sejnowski, and G. E. Hinton, eds., pp. 310–318. Morgan Kaufmann, San Mateo, CA.

Intrator, N. 1992. Feature extraction using an unsupervised neural network. *Neural Comp.* **4**, 98–107.

Intrator, N., and Cooper, L. N. 1992. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks* **5**, 3–17.

Intrator, N., and Gold, J. I. 1992. Three-dimensional object recognition of gray level images: The usefulness of distinguishing features. *Neural Comp.* **5**, 61–74.

Intrator, N., and Tajchman, G. 1991. Supervised and unsupervised feature extraction from a cochlear model for speech recognition. In *Neural Networks for Signal Processing – Proceedings of the 1991 IEEE Workshop*, B. H. Juang, S. Y. Kung, and C. A. Kamm, eds., pp. 460–469. IEEE Press, New York.

Intrator, N., Gold, J. I., Bülthoff, H. H., and Edelman, S. 1991. Three-dimensional object recognition using an unsupervised neural network: Understanding the distinguishing features. In *Proceedings of the 8th Israeli Conference on AICV*, Y. Feldman and A. Bruckstein, eds., pp. 113–123. Elsevier, Amsterdam.

Intrator, N., Reisfeld, D., and Yeshurun, Y. 1992. Face recognition using a hybrid supervised/unsupervised neural network. Preprint.

Ito, Y. 1991. Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. *Neural Networks* **4**, 385–394.

Jones, M. C. and Sibson, R. 1987. What is projection pursuit? (with discussion). *J. R. Statist. Soc. Ser.* **A**(150), 1–36.

Kruskal, J. B. 1969. Toward a practical method which helps uncover the structure of the set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In *Statistical Computation*, R. C. Milton and J. A. Nelder, eds. Academic Press, New York.

Kruskal, J. B. 1972. Linear transformation of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Application in the Behavioral Sciences, I, Theory*, R. N. Shepard, A. K. Romney, and S. B. Nerlove, eds., pp. 179–191. Seminar Press, New York.

Le Cun, Y. 1985. Une procédure d'apprentissage pour réseau à seuil assymétrique. In *Cognitiva 85: A la Frontière de l'Intelligence Artificielle des Sciences de la Connaissance des Neurosciences*, pp. 599–604, Paris. (Paris 1985), CESTA.

Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comp.* **1**, 541–551.

Le Cun, Y., Denker, J., and Solla, S. 1990. Optimal brain damage. In *Advances in Neural Information Processing Systems*, Vol. 2, D. Touretzky, ed., pp. 598–605. Morgan Kaufmann, San Mateo, CA.

Lincoln, W. P., and Skrzypek, J. 1990. Synergy of clustering multiple backpropagation networks. In *Advances in Neural Information Processing Systems*, Vol. 2, D. S. Touretzky and R. P. Lippmann, eds., pp. 650–657. Morgan Kaufmann, San Mateo, CA.

Lippmann, R. P. 1989. Review of neural networks for speech recognition. *Neural Comp.* **1**(1), 1–38.

Lyon, R. F. 1982. A computational model of filtering, detection, and compression in the cochlea. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France.

Moody, J. E. 1991. Note on generalization, regularization and architecture selection in nonlinear learning systems. In *Neural Networks for Signal Processing—Proceedings of the 1991 IEEE Workshop*, B. H. Juang, S. Y. Kung, and C. A. Kamm, eds., pp. 1–10.

Mougeot, M., Azencott, R., and Angeniol, B. 1991. Image compression with back propagation: Improvement of the visual restoration using different cost functions. *Neural Networks* **4**, 467–476.

Mozer, M. C., and Smolensky, P. 1989. Using relevance to reduce network size automatically. *Connection Sci.* **1**(1), 3–16.

Nowlan, S. J. and Hinton, G. E. 1992. Simplifying neural networks by soft weight-sharing. *Neural Comp.* **4**, 473–493.

Pearlmutter, B. A., and Rosenfeld, R. 1991. Chaitin-Kolmogorov complexity and generalization in neural networks. In *Advances in Neural Information Processing Systems*, Vol. 3, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, eds., pp. 925–931. Morgan Kaufmann, San Mateo, CA.

Perrone, M. P., and Cooper, L. N. 1992. Improving network performance: Using averaging to construct hybrid networks. *Proceedings of the CAIP Conference*, Rutgers University, October.

Plaut, D. C., Nowlan, S. J., and Hinton, G. E. 1986. *Experiments on learning by back-propagation*. Tech. Rep. CMU-CS-86-126, Carnegie-Mellon University.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing*, Vol. 1, D. E. Rumelhart and J. L. McClelland, eds., pp. 318–362. MIT Press, Cambridge, MA.

Sebestyen, G. 1962. *Decision Making Processes in Pattern Recognition*. Macmillan, New York.

Slaney, M. 1988. *Lyon's cochlear model*. Tech. Rep., Apple Corporate Library, Cupertino, CA 95014.

Switzer, P. 1970. Numerical classification. In *Geostatistics*, V. Barnett, ed. Plenum Press, New York.

Tajchman, G. N., and Intrator, N. 1992. Phonetic classification of TIMIT segments preprocessed with Lyon's cochlear model using a supervised/unsupervised hybrid neural network. In *Proceedings International Conference on Spoken Language Processing*, Banff, Alberta, Canada.

Turk, M., and Pentland, A. 1991. Eigenfaces for recognition. *J. Cog. Neurosc.* **3**, 71–86.

Wahba, G. 1990. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59. SIAM, Philadelphia.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. 1989. Phoneme recognition using time-delay neural networks. *IEEE Transact. ASSP* **37**, 328–339.

Weigend, A. S., Rumelhart, D. E., and Huberman, B. A. 1991. Generalization by weight-elimination with application to forecasting. In *Advances in Neural Information Processing Systems*, Vol. 3, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, eds., pp. 875–882. Morgan Kaufmann, San Mateo, CA.

Werbos, P. 1974. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. dissertation, Harvard University.

White, H. 1990. Connectionists nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks* **3**, 535–549.

Yamac, M. 1969. Can we do better by combining 'supervised' and 'nonsupervised' machine learning for pattern analysis. Ph.D. dissertation, Brown University.

Received 26 June 1992; accepted 26 October 1992.

This article has been cited by:

1. Shimon Edelman, Sharon Duvdevani-Bar. 1997. Similarity, Connectionism, and the Problem of Representation in VisionSimilarity, Connectionism, and the Problem of Representation in Vision. *Neural Computation* **9**:4, 701-720. [Abstract] [PDF] [PDF Plus]
2. Tin-Yau Kwok, Dit-Yan Yeung. 1997. Constructive algorithms for structure learning in feedforward neural networks for regression problems. *IEEE Transactions on Neural Networks* **8**:3, 630-645. [CrossRef]
3. David J. Field . 1994. What Is the Goal of Sensory Coding?What Is the Goal of Sensory Coding?. *Neural Computation* **6**:4, 559-601. [Abstract] [PDF] [PDF Plus]