Physica A 200 (1993) 655–661
North-Holland

SDI: 0378-4371(93)E0156-9

PHYSICA A

On the combination of supervised and unsupervised learning

Nathan Intrator
School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel

The bias/variance dilemma is addressed in the context of neural networks. A bias constraint based on prior knowledge about the underlying distribution of the data is discussed as a means for reducing the overall error measure of a classifier.

1. Introduction

The problem of optimal learning in artificial neural networks is approached through a minimization of some averaged distance between the estimator and the target, on a training sample set. Under the commonly used mean squared error (MSE) distance the error can be decomposed into two terms: bias and variance (see below).

Unfortunately, although the bias does go down through this minimization procedure, the variance may go up, thus reducing the overall performance of the estimator. This is due to the fact that the observations contain noise either in the input space X or in the regression space Y, and to the fact that only the empirical risk is minimized based on a small sample set. The problem pointed above should not be confused with the fact that the estimator based on empirical risk minimization is consistent, namely, in the limit when sample size becomes infinitely large, the estimator is unbiased. We are concerned with the problem of optimal estimation using a finite fixed sample size.

Methods for controlling the variance of estimators are many. They can roughly be divided into two categories: The first contains those that are based on general principles or assumptions on the functional form of the desired estimator. They do not depend directly on the (unknown) data distribution. In the neural network framework they include methods such as weight decay and magnitude control of the weights [1,2], network pruning via weight elimination based on a simple threshold [3,4] or based on the Hessian matrix [5]. A different approach for reducing the effective number of weights is weight sharing, in which a single weight is shared among many connections in the network [6,7]. An extension of this idea is the "soft weight sharing" which favors irregularities in the weight distribution in the form of multimodality [8]. All these methods make explicit assumptions about the structure of the weight space, but with no regard to the structure of the input space. The second category contains methods that have general assumptions about the underlying distribution and important structure in the data. Such methods include principal component constraints, in which one seeks an estimator which is decomposed out of functions biased towards projections onto the principal components of the covariance matrix of the data. Similarly, constraints can be added for seeking projections that maximize entropy [9]. The difference between the two categories may be easier to see if one observes that in the first category, if the bias minimization (through mean squared error or alike) is turned off, then the additional constraints will not find any meaningful information (projections) while in the second category, if the bias minimization is turned off, then meaningful directions such as the first principal components of the data can still be found.

Bias constraints are at the heart of parametric estimation methods. For example, linear or logistic regression biases the dependency between the covariates and the data to be linear. Similarly, bias is introduced when a certain network architecture is used. In this paper we introduce bias constraints into a given feed-forward network architecture. Based on prior knowledge about the underlying distribution of the data, a specific bias constraint is discussed as a mean for reducing generalization error for classification. The application to neural networks of the general statistical framework from which the bias constraints are drawn – the exploratory projection pursuit framework [10] – is discussed in more detail in [11].

2. The bias/variance dilemma

In this section we present the bias/variance decomposition of a non-parametric estimator. For a thorough discussion of this problem in the context of neural networks, see Geman et al. [12].

The regression or classification problem is to estimate a function $f_{\mathcal{D}}(x)$ based on a fixed training set $\mathcal{D} = \{(x^1, y^1), \ldots, (x^L, y^L)\}$, using some measure of the estimation error on the training set. A good estimator will perform well not only on the training set but will also achieve good *generalization* properties, namely it will achieve small error on observation which was not included in the training set.

Evaluation of the performance of the estimator is commonly done via the mean squared error distance (MSE) by taking the expectation with respect to the (unknown) probability distribution P of y:

$$E[(y-f_{\mathscr{D}}(x))^2 \mid x, \mathscr{D}] .$$

This can easily be decomposed into

$$E[(y-f_{\mathscr{D}}(x))^2 \mid x, \mathscr{D}] = E[(y-E[y \mid x])^2 \mid x, \mathscr{D}] + (f_{\mathscr{D}}(x) - E[y \mid x])^2 .$$

The first term does not depend on the training data $\mathscr{D}$ or on the estimator $f_{\mathscr{D}}(x)$; it measures the amount of noise or variability of y given x. Hence f can be evaluated using

$$(f_{\mathscr{D}}(x) - E[y \mid x])^2 .$$

The empirical mean squared error of f is then given by

$$E_{\mathscr{D}}[(f_{\mathscr{D}}(x) - E[y \mid x])^2] ,$$

where $E_{\mathscr{D}}$ represents expectation with respect to all possible training sets $\mathscr{D}$ of fixed size L.

To further see the performance under MSE we decompose the error to bias and variance (see for example Geman et al. [12]) to get

$$\begin{aligned} E_{\mathscr{D}}[(f_{\mathscr{D}}(x) - E[y \mid x])^2] = {} & (E_{\mathscr{D}}[f_{\mathscr{D}}(x)] - E[y \mid x])^2 \\ & + E_{\mathscr{D}}[(f_{\mathscr{D}}(x) - E_{\mathscr{D}}[f_{\mathscr{D}}(x)])^2] . \end{aligned}$$

The first rhs term is called the bias of the estimator and the second term is called variance. When training on the fixed training set $\mathscr{D}$, reducing the bias with respect to this set increases the variance of the estimator thus contributing to poor generalization performance. There is often a tradeoff between variance and bias. Typically variance is reduced by smoothing, however this may introduce bias since it may blur sharp peaks, etc. Bias is reduced by prior knowledge, and when this prior knowledge is also contributing to smoothing it is likely to reduce the overall MSE of the estimator.

In the next section we discuss a general form for introducing bias into a fixed-architecture neural network. In addition we discuss a specific bias constraint useful for classification.

2.1. Adding bias constraints to a back-propagation network

As can be seen in fig. 1, a penalty term may be added to the energy functional minimized by error back propagation, for the purpose of measuring

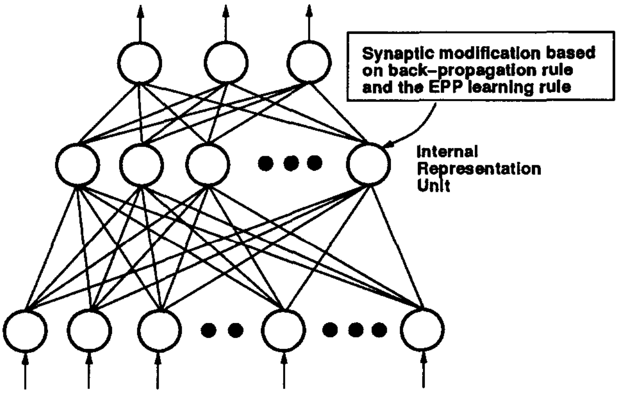
Synaptic modification based on back-propagation rule and the EPP learning rule
Internal Representation Unit

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

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, in a feed-forward network in which the number of hidden units is *much smaller* than the number of input units. Therefore, it is sufficient to add the bias penalty term only to the hidden layer (see fig. 1). We call this network exploratory projection pursuit network, to stress the fact that the bias constraints are motivated by the exploratory projection pursuit framework [10,13]. The synaptic modification equations of the hidden units' weights become

$$\frac{\partial w_{ij}}{\partial t} = -\epsilon\left(\frac{\partial \mathscr{E}(w, x)}{\partial w_{ij}} + \frac{\partial \rho(w_1, \ldots, w_n)}{\partial w_{ij}}\right.$$
$$\left. + \text{(contribution of cost} - \text{complexity terms)}\right).$$

3. Projection index for classification: the unsupervised BCM neuron

In this section we briefly describe an unsupervised learning algorithm which searches for multimodality in the projection space. As described in the previous section, the hybrid unsupervised/supervised learning network then combines the differential equations governing the unsupervised rule with those governing the supervised rule to minimize a combination of two costs: the one that comes from the teacher, and one that comes from the unsupervised (bias) constraint.

It is known from exploratory projection pursuit theory that search for structure in input space can be approached by a search for deviation from normal distribution of the projected space (the space generated by hidden unit activity in a feed-forward network). Furthermore, when input space is clustered, a search for deviation from normality can take the form of search for multi-modality, since when clustered data is projected in a direction that separates at least two clusters, it generates multi-modal projected distributions.

It has been recently shown that a variant of the Bienenstock, Cooper and Munro neuron [14] performs exploratory projection pursuit using a projection index that measures multi-modality. Such neuron allows modeling and theoretical analysis of various visual deprivation experiments [15], and is in agreement with the vast experimental results on visual cortical plasticity [16]. A network implementation which 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 [17,18]. An approach of this type has been used in image compression, with a penalty aimed at minimizing the entropy of the projected distribution [9]. This penalty certainly measures deviation from normality, since entropy is maximized for a Gaussian distribution.

The neuronal activity (in the linear region) is given by $c = m \cdot d$, where d is the input vector and m is the synaptic weight vector (including a bias). The essential properties of the BCM neuron are determined by a modification threshold Θ_{m} (which is a nonlinear function of the history of activity of the neuron) and a ϕ function that determines the sign and amount of modification and depends on the current activity and the threshold Θ_{m}. The synaptic modification equations are given by

$$\frac{\mathrm{d}m_i}{\mathrm{d}t} = \mu\phi(c, \Theta_{\mathrm{m}})\,\mathrm{d}_i\,,$$

where in a simple form $\Theta_{\mathrm{m}} = E[(m \cdot d)^2]$ and $\phi(c, \Theta_{\mathrm{m}}) = c(c - \Theta_{\mathrm{m}})$.

In the lateral inhibition network of nonlinear neurons (fig. 1) the activity of neuron k is given by $c_k = m_k \cdot d$, where m_k is the synaptic weight vector of neuron k. The *inhibited* activity and threshold of the kth neuron is given by $\tilde{c}_k = \sigma(c_k - \eta \sum_{j \neq k} c_j)$ and $\tilde{\Theta}^k_{\mathrm{m}} = E[\tilde{c}^2_k]$, for a monotone saturating function σ.

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

$$R(w_k) = -\{\tfrac{1}{3}E[\tilde{c}^3_k] - \tfrac{1}{4}E^2[\tilde{c}^2_k]\}\,.$$

The total risk is the sum of each local risk. The resulting stochastic modification equations for a synaptic vector m_k (the negative gradient of the risk) in the network are given by

$$\dot{m}_k = \mu \left(\phi(\tilde{c}_k, \tilde{\Theta}_m^k)\, \sigma'(\tilde{c}_k) - \eta \sum_{j \neq k} \phi(\tilde{c}_j, \tilde{\Theta}_m^j)\, \sigma'(\tilde{c}_j) \right) d \,.$$

This network is actually a first order approximation to a lateral inhibition network (using a single step relaxation). Its properties and connection to a lateral inhibition network are discussed in [15]. In the context of the hybrid network, this is an additional penalty to the energy minimization of the supervised network.

Some related statistical and computational issues of this projection index as well as some applications are discussed in [18].

4. Summary

A framework for introducing additional bias into a neural network architecture was presented. It is based on a penalty that allows the incorporation of additional prior information regarding the underlying data distribution, or model. The general statistical framework of exploratory projection pursuit was mentioned as the underlying framework for the procedure, and a specific bias aimed at improving classification performance was presented.

References

[1] D.C. Plaut, S.J. Nowlan and G.E. Hinton, Experiments on learning by back-propagation, Technical Report CMU-CS-86-126 (Carnegie-Mellon University, 1986).

[2] M.C. Mozer and P. Smolensky, Connection Sci. 1 (1989) 3.

[3] Y. Le Cun, J. Denker and S. Solla, Optimal brain damage, in: Advances in Neural Information Processing Systems, vol. 2, Denver, 1989, D. Touretzky ed. (Morgan Kaufmann, San Mateo, CA, 1990) pp. 598–605.

[4] A.S. Weigend, D.E. Rumelhart and B.A. Huberman, 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. (Morgan Kaufmann, San Mateo, CA, 1991) pp. 87–882.

[5] B. Hassibi and D.G. Stock, Second order derivatives for network pruning: Optimal brain surgeon, in: Advances in Neural Information Processing Systems. vol. 5, C.L. Giles, S.J. Hanson and J.D. Cowan, eds. (Morgan Kaufmann, San Mateo, CA, 1993).

[6] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang, IEEE Trans. ASSP 37 (1989) 328.

[7] Y. Le Cun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard and L. Jackel, Neural Comput. 1 (1989) 541.

[8] S.J. Nowlan and G.E. Hinton, Neural Comput. 4 (1992) 473.

[9] M. Bichsel and P. Seitz, Neural Networks 2 (1989) 133.

[10] J.H. Friedman, J. Am. Stat. Assoc. 82 (1987) 249.

[11] N. Intrator, Neural Comput. 5 (1993) 509.

[12] S. Geman, E. Bienenstock and R. Doursat, Neural Comput. 4 (1992) 1.
[13] P.J. Huber, Ann. Stat. 13 (1985) 435.
[14] E.L. Bienenstock, L.N. Cooper and P.W. Munro, J. Neurosci. 2 (1982) 32.
[15] N. Intrator and L.N. Cooper, Neural Networks 5 (1992) 3.
[16] E.E. Clothiaux, L.N. Cooper and M.F. Bear, J. Neurophysiol. 66 (1991) 1785.
[17] N. Intrator, J.I. Gold, H.H. Bülthoff and S. Edelman, Three-dimensional object recognition using an unsupervised neural network: Understanding the distinguishing features, in: Proc. 8th Israeli Conf. on AICV, Y. Feldman and A. Bruckstein, eds. (Elsevier, Amsterdam, 1991) pp. 113–123.
[18] N. Intrator, Neural Comput. 4 (1992) 98.