OPTIMIZING THE MULTILAYER PERCEPTRON - PROBLEMS, TOOLS AND STRATEGIES

Dr. Matthias Kerling
Deutscher Investment-Trust, Mainzer Landstr. 11-13, 60329 Frankfurt
Tel.: 069/263-14292, Fax: (069) 263-14803, email: kerlingm@dit.de

ABSTRACT: Neural networks are a powerful and flexible statistical tool but there are several problems in making use of them. The central problem of neural network cited in literature is the overfitting effect which has its origin in the well known bias-variance-dilemma and in the disturbances which overlay economic relationships in the form of noise. Especially for the Multilayer Perceptron, numerous optimization strategies have been developed in the last decade to overcome these problems. But if we take into account the findings of ordinary statistics some of these procedures are not correctly designed, some are used in the wrong way, and some are misused. The following contribution presents the background of individual procedures and discusses how to use and combine them in a flexible optimization strategy.

KEYWORDS: Multilayer Perceptron, Nonlinearity, Bias-Variance-Dilemma, Overfitting, Crossvalidation Techniques, Regularisation, Pruning, Construction, Statistical Optimization, Optimization Strategies

THESIS

- The Multilayer Perceptron is nothing else but an extremly flexible statistical tool and its capability and power is in direct relation to the complexity and parametrization of the network structure in use. Depending on this parametrization the MLP pass a smooth transition from parametric to nonparametric models.
- Making use of parametric models with small network structures, one has to formulate concrete assumptions about the economic relationship of interest. If such models are not specified correctly, they have a bias. Making use of nonparametric models with complex network structures, one abandons those assumptions. The form of the modeled relationship is directly related to the sample that is used for estimation. Therefore, nonparametric models suffer from a high variance.
- On the one hand, economic relationships are complex and not well understood. On the other hand, the MLP is able to model any relationship as long as it is sufficiently parametrisized (see Hornik et al. 1989). Consequently at first glance, the use of highly parametrisized models seems to be a good strategy.
- But the often cited overfitting problem of the MLP has its origin in the well known bias-variance-dilemma (see Geman et al., 1992) described above. It occurs if the model has too many degrees of freedom.
- Due to this bias-variance-dilemma we should favour small models with an ‘optimal’ network structure. This is a structure which is just able to model the unknown economic relationship and keeps the variance of the model as small as possible (see Anders, 1995, p. 5, 1997).
- Furthermore, from a economic point of view, the extraction of optimal structures appears important since it allows the identification of relevant factors. This is the basis for analysing cause and effect in economic relationships. The following discussion gives an overview over a wide range of optimization tools, their advantages and shortcomings and how to use and combine them in a correct manner.

POTENTIAL DETERMINANTS OF OVERFITTING

- The actual cause of the overfitting problem are the disturbances which overlay available structures in the form of noise. Without noise there will be no overfitting. The degree of noise is the first and most relevant determinant of overfitting.
- In the classical literature of neural networks, in general two determinants are made responsible for overfitting: Because the approximation ability of the MLP is determined by the number of its free parameters/weights, both its capability and the danger of the overfitting is in direct connection. Since the parameter of the MLP are not analytically exact definable but determined iteratively with gradient descent procedures, the number of training cycles is often interpreted as further determinant.
- In fact both are directly related: A highly parametrisized network loses its generalization ability after a few training cycles. On the other hand, a appropriately sized net requires considerably more training cycles in order to capture any available structures at all.
- However, the following discussion of alternative optimization strategies however, will show that the attempt to avoid overfitting by the number of training cycles (the well known ‘stopped training’) is methodically extremely problematic.

THE SHORTCOMINGS OF STOPPED TRAINING

The primary objective of crossvalidation techniques is the appraisal of the generalization accuracy of a statistical model. In addition they offer support in judging the parametrisation of a model. Within the framework of neural networks they are often misused in order to limit the number of training cycles of the MLP via stopped training (see Weigend et al. 1990).

- Crossvalidation gives information on overfitting. Instead of measuring the possible degree of overfitting (which may give an indication about the parametrization of a given network structure), making use of stopped training the parameter estimation will be stopped just in the moment when overfitting occurs. The only information you get about overfitting is that it just happens (at following see Kerling, 1996, pp. 410 and Kerling, 1998, pp. 329).
- Because structural information of the crossvalidation sample is used for this procedure the crossvalidation error gives no valid information about the generalization accurracy of the network. It has to be tested with a second crossvalidation sample. This worsens the omnipresent problem of small data sets in economic applications.
- But even if a second crossvalidation sample is in use the weights of the network are not determined and information about the generalization accuracy are liable to random influences (e.g. the composition of the crossvalidation sample, the training algorithm, the step width, etc.).
- A second but even more important point of criticism against stopped training refers to the bias-variance-dilemma discussed already. The emphasis of optimizing the MLP has to be put towards the network structure. This is the only way to solve the structural problem of highly parametrisized models and to limit the model variance. Stopped training per se doesn’t solve the problem of overparametrisation.
- Furthermore statistical tests which represent the most powerful tools in order to optimize the network structure of the MLP do not work with stopped training. Using stopped training a network does not reach a minimum in weight space. If the weights of the network are not determined they are no reliable basis for statistical tests. If statistical optimization strategies use this unreliable information they may give misleading signals.
- In contrast to stopped training a kind of ‘forced overfitting’ would give information about the possible degree of overfitting. At the same time the weights of the network are a certain basis for statistical tests. Using these tests iteratively the fitting potential of a network is restricted step by step. In conjunction with ‘forced overfitting’ the reduction of the overfitting effect is measurable. During this iterative optimization the error on the training sample will increase because of the restricted potential of the network. At the same time the error on the crossvalidation sample will decrease because of the reduced tendency to overfit the data. In this way the potential of the network will be restricted in an orderly manner.
- Finally it has to be commented that in ordinary statistcs the question of interrupted parameter estimation has never arisen because the parameters of linear models are in general analytically exact definable. Even in nonlinear

regression models whose parameters must also be determined by gradient descent procedures the unambiguity of a global/local minimum and hence the identification of a model in question is of special importance. This is the basis for statements about expected parameter values, estimation errors and tests concerning their statistical significance. The attempt to make a virtue out of the necessity of iterative parameter estimation appears as deadend street for statistical tests.

VALIDATION IN NONLINEAR MODELS

Crossvalidation techniques like simple crossvalidation, bootstrapping, jackknifing, etc. are well known in ordinary statistics but they do not take into account the special requirements of nonlinear error surfaces. The following briefly discusses the shortcomings of linear crossvalidation techniques.

SHORTCOMINGS OF SIMPLE CROSSVALIDATION

- In economic applications the problem of small data sets is omnipresent. Dividing those data sets into two subsets which should be both representative for an unknown totality seems to be quite difficult. Selecting data points for crossvalidation by chance leads to random results. Information about the generalization accuracy of the forecasting model is far away from being reliable (see Poddig, 1994a, p. 334, Poddig, 1994b, 1996.
- Furthermore, picking out a certain fraction of the available data for crossvalidation reduces the number of training patterns and worsens the problem of analyzing small data sets with a high parametric forecasting model. In particular it has to be taken into account that the parameters are subject to additional distortions because of the reduction of the training set.

ADVANTAGES AND SHORTCOMINGS OF MULTIPLE CROSSVALIDATION TECHNIQUES

- Simple crossvalidation is just one simple method in order to get information about the validity of a specific forecasting model. In contrast methods like jackknifing (see Mosteller/Tukey, 1968), bootstrapping (see Efron, 1979, 1983), leave-one-out- or v-fold-crossvalidation (see Geisser, 1975) provide a more reliable basis.
- The focus of multiple crossvalidation techniques is a different one. In a first step the parameters of a model are estimated using the whole data set. In a second step the model's generalization accuracy will be estimated via numerous simple crossvalidations based on slightly modified data sets.
- This has several advantages. The parameter of the model are estimated by the whole data set. They are not subject to additional distortions. The information about the generalization accuracy of the model is by far more reliable since a considerably higher volume of data can be employed for crossvalidation. Besides one obtains information about the stability or the estimating error of the parameters which results from the lacking of larger amounts of data. Some statistical tests like jackknifing make use of this information in order to eliminate non significant parameters.
- But these multiple crossvalidation techniques do not take into account that there may be numerous minima in the error surfaces of nonlinear models. Random initialization and/or different training sets for individual crossvalidations may result in not comparable models that take quite differing positions in the weight space. In this case one gets the crossvalidation errors of different models instead of measuring the generalization accuracy of one particular model in question. Furthermore, there is no guarantee that individual weights are carrying out the same function over all models (for example there may be network symmetries). Statistical tests based on not comparable models may cause misleading results.

NONLINEAR CROSSVALIDATION

- Moody/Utans (1995, pp. 288 -289) have suggested a sequential two step nonlinear crossvalidation procedure. At the first step a single network has to be trained into a local or global minimum. The obtained network with weight matrix $\boldsymbol{w}$ will be subject to further examination in the second step. In order to get information about the prediction risk of

this network it has to be **retrained** *v* times using ***w*** as starting point and holding out just one subset of the original training sample for crossvalidation.

- The assumption of nonlinear crossvalidation is that retraining of a specific network on a reduced data set does not lead to a large difference in the locally optimal weights but causes only a slightly different error surface. Thus, starting form ***w*** assures that the multiple crossvalidation estimates the generalization accuracy of a particular model in question.
- This nonlinear crossvalidation ensures that the weights of the network in question are a certain basis for statistical tests because all *v*+1 modells are trained into the same minimum of the error surface and the weights of the different models are comparable. There is no chance for network symmetries.

OPTIMIZING THE INTERNAL STRUCTURE

So far it seems to be a good strategy to prefer the nonlinear crossvalidation of Moody/Utans (1995) and to refuse the traditional stopped training. But crossvalidation is just a supporting tool and as mentioned earlier the main emphasis has to be put towards the optimization of the network structure of the MLP. The following gives a brief summery of some more or less well known techniques to optimize the internal network structure between the input and the hidden layer of the MLP. The destination is to show connections between individual procedures and to discuss more sophisticated techniques designed to overcome the shortcomings of simpler strategies.

REGULARISATION - WEIGHT DECAY AND PENALTY TERM

- Simple regularisation techniques (see Hertz et al., 1991, pp.157) try to avoid overfitting by keeping the weights small. Towards the original training step Δw_i they press the weights into the direction of zero 'by hand' or ad a penalty term $P(\mathbf{w})$ to the error function E.

(I) weight decay: $\Delta w_i = -h \frac{\partial E}{\partial w_i} - e_i w_i \qquad e_i = \frac{const}{\left(1 + w_i^2\right)^2}$

(II) penalty term: $P(\mathbf{w}) = +g \frac{1}{2} \sum_{i=1}^{n} w_i^2 \qquad P(\mathbf{w}) = +g \frac{1}{2} \sum_{i=1}^{n} \frac{w_i^2}{1 + w_i^2}$

- As in the case of stopped training, the net should be impeded to fit itself too strongly to the data. Therefore there are similar points of criticism. The model may be restricted to quasi linear models based on information that has no link to the original application. So the error function minimized doesn't represent the structure of the problem to be solved. Furthermore the optimization of the decay rate (e, g) represents a further degree of freedom and the parameter estimation of the network will be extended.

TESTING SIGNIFICANCE

While regularisation techniques restrict the potential of the MLP by keeping its parameter small, statistical tests try to identify parameters whose expected values do not differ significant form zero. These weights should be removed since even their sign can not be predicted with accurate safety. Because statistical tests are based on the deviation of the expected parameter values from zero they should not combined with stopped training. Both regularisation techniques and stopped training prevent the parameters from reaching their real expected values by holding them close to zero in an artificial manner. Kerling (1998, pp. 350) shows that the following tests may lead to misleading signals and suboptimal network architectures if used in combination with stopped training.

Pruning small weights

(III) Small weights: $test(w_i) = |w_i|$ or $test(w_i) = w_i^2$

- The elimination of small weights is only the simplest way in order to slim down the internal network structure of the MLP. But eliminating small weights is not a test in the statistical sense. It does not note that every parameter estimation is subject to estimation errors. The expected value of a parameter **and** its estimation error are the basis for statistical tests.

The Test of Finnoff/Zimmermann (FZ-Test)

- ... use information of a special gradient descent procedure, the simple delta-rule, in order to estimate the variance of each parameter and to construct a kind of t-test. The estimation of the parameter variance is based on the parameter fluctuations caused by individual data vectors p. Looking back at the dicussion above concerning stopped training the following formula presupposes that the network is in a minimum of the error surface.

(IV) Finnoff/Zimmermann (1992): $$test(w_i) = \frac{|w_i|}{\sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(w_{i,p} - w_i\right)^2}}$$

Jackknifing

- While the FZ-Test is based on one specific minimizing procedure and the parameter fluctuations caused by individual data vectors, Jackknifing makes use of the parameter fluctuations $w_i^{\circ(l)}$ caused by *k* varying compositions of the training sample. In the framework of the discussion so far, Jackknifing provides several advantages. Originally designed to handle small data sets via undistorted parameter estimators w_i°, Jackknifing seems to fit perfectly into the omnipresent problem of small data sets in economic applications. Multiple crossvalidation techniques like the nonlinear crossvalidation of Moody/Utans (1995) have just been recommended. So the *k* re-estimations of a model in question do not represent additional effort. The possibility to employ this advanced statistical test is just an free and efficient by-product of nonlinear crossvalidation.

(V) Poddig (1994a): $$test(w_i) = \frac{|w_i^{\circ}|}{\sqrt{\frac{\frac{1}{k-1}\sum_{l=1}^{k}(w_i^{\circ(l)} - w_i^{\circ})^2}{k}}}$$

The Wald-Test

- ... is a standard procedure in ordinary statistic. White (1992b) shows that the parameters of the MLP are asymptotically normally distributed if they are at least locally identified. The real covariance matrix of the parameters is indeed not computable, but White provides a weakly consistent estimator and makes the Wald-Test applicable in the framework of the MLP.

(VI) White (1992b): $$test(w_i) = \frac{w_i^2}{s_{w_i}^2} \qquad s_{w_i}^2 = \frac{1}{P}\mathbf{C}_{ii} \qquad \frac{1}{P}\mathbf{C} = \frac{1}{P}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$$

$$\mathbf{A} = \frac{1}{P}\sum_{p=1}^{P}\frac{\partial E_p}{\partial w \partial w'} \qquad \mathbf{B} = \frac{1}{P}\sum_{p=1}^{P} e_p^2\left(\frac{\partial E_p}{\partial w}\right)\left(\frac{\partial E_p}{\partial w}\right)'$$

For all these procedures is valid that eliminated weights may be animated again by some test modifications or by their first derivative of the error function. So at first glance these procedures seem to be the first choice in order to optimize the internal structure of the MLP. However, it is precondition for their correct use that the parameters are at least identified locally. This should have been guaranteed by forced overfitting. But because of the hierarchical structure of the MLP this assumption may be violated, particular if a model in question contains irrelevant hidden units. But this problem has to be solved by optimizing the number of hidden units.

TESTING RELEVANCE

Statistical tests just take care about the statistical significance of the parameters. They do not consider the contribution of individual parameters for the error minimization. Certainly the following procedures can not replace the statistical tests discussed above but they supply additional information in order to decide about the elimination of individual parameters.

Simple Sensitivity Analysis

- The emphasis of sensitivity analysis is to measure the relevance of specific parameters. Parameters which do not add value to minimize the error function may be eliminated. This procedure seems to be intuitive but there are also some points of criticisme: In order to get the test value of individual weights one has to calculate the error of n+1 models (the number of current weights and the 'parent model' itself). There is no critical value in order to eliminate one specific weight. Sensitivity analysis suppose that the parameters are independent. If they are not, the parameters have to be reestimated in order to get reliable test values. This makes sensitivity analysis quite time consuming. Furthermore, simple sensitivity analysis do not take into account that smaller models with fewer parameters in general result in higher error rates.

(VII) $$test(w_i) = E(w_i = 0) - E$$

Information and Selection Criteria

- Information and selection criteria deal with the last point of criticism. They adjust the resulting error of a specific model to the number of parameters in use. The following are only a selection of current available measures. The advantage of these criteria is that they take into account the dimension of alternative (hierarchical) models. With test(w_i) = 0 they also provide a natural limit to eliminate or to keep one specific parameter alive. But they are time consument as well. Finally they also suppose parameter independence and therefore do require reestimation after each test calculation.

(VIII) Akaike (1974): $$AIC = \frac{1}{P} SSE(\mathbf{w}) + s^2 \frac{2K}{P}$$

(IX) Schwarz (1978): $$SIC = \frac{1}{P} SSE(\mathbf{w}) + \frac{K \ln(P)}{P}$$

(X) Murata et al. (1994): $$NIC = \frac{1}{P} SSE(\mathbf{w}) + \frac{1}{P} tr[\mathbf{BA}^{-1}]$$

(XI) $$test(w_i) = IC(w_i = 0) - IC$$

Optimal Brain Damage

- ... does focus on the time consum of the criteria discussed so far. Under the assumption that a model in question is in a minimum of the error surface optimal brain damage approximates the additional error which would result if w_i will

be set equal to zero. Early Brain Damage suggested by Tresp et al. (1996) represents an alternative designed for the combination with 'stopped training'. But as mentioned earlier the identification of the model parameters is of importance for subsequent statistical tests. Furthermore the question of a critical value and the problem of parameter reestimation is solved neither by Optimal nor by Early Brain Damage.

(XII) Le Cun et al. (1990): $test(w_i) = \frac{1}{2} \frac{\partial^2 E}{\partial w_i^2} w_i^2$

Optimal Brain Surgeon

- ... takes care about this 'last problem' of sensitivity analysis using the Hesse Matrix.

(XIII) Hassibi/Storck (1993): $test(w_i) = \frac{w_i^2}{2[\mathbf{H}^{-1}]_{ii}} \quad \Rightarrow \quad \Delta w_j = -\frac{w_i^2}{[\mathbf{H}^{-1}]_{ii}} \mathbf{H}^{-1}\mathbf{I}_i$

OPTIMIZING THE HIDDEN LAYER

- The optimization of the number of hidden units is of special importance for the MLP. Eliminating or adding further hidden units into an existing network strcuture is an serious intervention. The procedures discussed so far only concern individual weights while a hidden unit stands for a group of weights which represents a closed functional unit. So the number of hidden units sets the framework for the desired capacity, the undesired flexibility (overfitting) and the varity of models a specific network structure is able to represent.
- In general, techniques like regularisation and testing the relevance could be modified in order to eliminate hidden units. However, the discussion so far shows that statistical tests are more stringent and should be also mighty tools for this task. But as already mentioned there may be some problems making use of statistical tests in the framework of the MLP. These problems have there origin in the hierarchical structure of the MLP.
- The test statistics already discussed presuppose that the network parameters to be tested are (at least locally) identifiable. If we are testing parameters of the internal network structure we implicitly assume that the parameters between hidden and output layer differ from zero. Otherwise the internal parameters to be tested may take any value without effect. They would be not identifiable. If we turn procedure upside-down and want to test the significance of a hidden unit (its parameter to the output) we have to presuppose that the parameters connecting the hidden unit with the input layer differ from zero. So it's a classical test of a joint hypothesis. The following test statistics show ways out of this dilemma.

WHITE'S NEURAL NETWORK TEST

- The basic consideration of White's lagrange multiplier test is the notion that if a model is correct specified there should be no correlations between the residuals u_t^* of the model in question and any transformation of the input vectors $\mathbf{x}_t$. If there are correlations there may be some kind of neglected nonlinearities. In the framework of the MLP the transformation of the input vectors is performed by the output function y (e.g. logistic) of an additional hidden unit. In order to avoid the parameter estimation for reasons discussed above, White engages the weight vectors $\mathbf{w}_i$ of p additional phantom hidden units with random numbers and tests the following hypothesis.

(XIV) White (1989):
$$H_0: E[y(\mathbf{x}_t'\mathbf{w}_i)u_t | \mathbf{w}_i] = 0$$
$$H_a: E[y(\mathbf{x}_t'\mathbf{w}_i)u_t | \mathbf{w}_i] \neq 0 \quad \text{für } i = q+1, \ldots, q+p$$

- Because of the random initialization it is recommended to add p phantom hidden units in one step. As a consequence the probability to discover correlations will increase but the power of the test will suffer under rising p. Therefore White suggests to limit the number of phantom hidden units via principal component analysis. Finally it has to be mentioned that his test is not consistent. If the test is not able to detect correlations one must not conclude that H_0 is valid. From this point of view the following test of Teräsvirta et al. (1993) has some advantages which were

confirmed in some time series analysis (see Teräsvirta et al., 1993).

THE LAGRANGE MULTIPLYER TEST OF TERÄSVIRTA/LIN/GRANGER

- White implicitly tests whether the weight which would connect an additional hidden unit with the output is significant under the assumption that the input weights of this unit are already 'determined' by random numbers. In contrast the test of Teräsvirta/Lin/Granger is based on a Taylor Approximation which simulates the functionality of an additional hidden unit. So there is no need to estimate hierarchical arranged parameters. For their purposes they use a slightly modified output function y :

(XV) Teräsvirta et al. (1993): $y(\mathbf{x}_t'\mathbf{w}) = \frac{1}{1+\exp(-\mathbf{x}_t'\mathbf{w})} - \frac{1}{2}$ with

$$\frac{\partial y(0)}{\partial w_i} = \frac{1}{2}\mathbf{x}_{t,j} \;,\; \frac{\partial y(0)}{\partial w_i \partial w_j} = 0 \;,\; \frac{\partial y(0)}{\partial w_i \partial w_j \partial w_k} = \frac{1}{16}\mathbf{x}_{t,i}\mathbf{x}_{t,j}\mathbf{x}_{t,k}$$

- Adding the taylor approximation of a first hidden unit to a linear model they get the following network functionality, where $\mathbf{x}_t'\mathbf{w}_0$ represents the part of the original linear model and $\mathbf{w}_1$ is the weight vector of one first hidden unit.

(XVI) Teräsvirta et al. (1993): $f(\mathbf{x}_t, \mathbf{W}) = \mathbf{x}_t'\mathbf{w}_0 + \sum_{i=0}^{I}\sum_{j=i}^{I}\sum_{k=j}^{I} d_{ijk}\mathbf{x}_{t,i}\mathbf{x}_{t,j}\mathbf{x}_{t,k}$ with

$$d_{ijk} = b_1 d_{ijk} w_{1,i} w_{1,j} w_{1,k}$$

- With this in mind they test the following hypothesis:

(XVII) Teräsvirta et al.(1993): $H_0: d_{ijk} = 0 \;\forall\; i, j, k \; mit \; i = 0, \ldots, I;\; j = i, \ldots, I;\; k = j, \ldots, I$

$$H_a: d_{ijk} \neq 0$$

- But the taylor approximation results in numerous third order polynomials which may also correlate with each other. In order to increase the power of the test it is also recommended to reduce the polynomials via principal component analysis. Both test statistics are defined in the framework of neural networks. In ordinary statistics there are also further tests to discover nonlinearities (see Kerling, 1998, pp.393).

OPTIMIZING THE INPUT UNITS/FACTORS

- As in linear models the statement 'garbage in, garbage out' is valid. So also in the framework in neural networks the number and selection of exogenous factors is the most important task. Already at the beginning of the process of model building, the number of potential factors should be kept as small as possible because irrelevant factors act as additional noise and increase the danger of overfitting. High correlated input factors result in redundant information and distorted parameter estimators. Finally the interpretation of a model containing irrelevant input factors will be complicated.
- For these reasons the biggest effort in selecting the relevant factors should be spent before starting the modelling process. In addition to the traditional correlation analysis there are several nonlinear test procedures offered by ordinary statistics. Examples are the Nonlinear Granger Causality Test or the Delta-Test of Pi (1993).
- In principle the techniques to optimize the internal network structure are able to eliminate remaining irrelevant input units. Nevertheless it should be proven that there are no more irrelevant input units after this optimization. In essence the sensitivity analysis and the Wald Test are available for this purpose. The latter one is able to test the significance of a group of parameters which connect one specific input unit with the hidden layer (see White, 1992b, p. 108).

CONCLUSION

- The findings so far can be summerized as follows. Neural networks are no intelligent systems but a flexible statistical tool. For this reason especially the MLP is at least subject to the same assumptions as linear regression analysis and because of the omnipresent bias-variance-dilemma tiny models should be preferred.
- At the beginnig of modelling an unknown economic relationship, a wide range of statistical tests (based on fundamental considerations) should be performed in order to discriminate relevant from irrelevant factors. The emphasis of neural networks is to model nonlinear relationships. Therefore these investigations must not be limited to linear tests.
- Crossvalidation techniques enable the monitoring of an iterative parameter estimation but this information should not be used for stopped training. The available information about the overfitting effect would be cut off. Information about the generalization accuracy would be liable to random influences and the 'estimated' parameters would be no basis for statistical tests.
- There are several precedures available in order to optimize the internal network structure but in principle advanced statistical techniques should be prefered. As in ordinary statistics their results should be supplemented by sensitivity analysis and/or selection criteria.
- However these test are only applicable if the network parameters are identified. But redundant input and irrelevant hidden units may result in non unique error minima. Therefore at a first step the number of necessary hidden units should be determined by building up the hidden layer step by step. The test statistics of White (1989) and Teräsvirta et al. (1993) seem to be adequate for this task. Due to the complexity of neural networks and the resluting problems in real life applications the linear regression model should be the starting point. Only the proof of neglected nonlinearities should justify the use of neural networks or more complex network structures containing further hidden units.
- There are interdependences between the parameters of the different layers because of the hierarchical structure of the MLP. Therefore constructing the hidden layer and optimizing the internal network structure should be carried out in alternation. Finally it should be proven that the resulting model does not contain remaining irrelevant inputs.

CONCURRENT AND SUPPORTING CONSIDERATIONS

Especially in economic applications there is a misproportion between signal information and signal noise. There are two possibilities to reduce this disparity: Increase signal information and/or reduce signal noise. As mentioned above signal noise is the origin of the overfitting problem. Because noise of certain degree would mask any signal the use of noise filters like smoothing time series is usual. But if the relationship in question is a multivariate one it may be destroyed by univariate filters. Therefore multivariate techniques like clustering or topoligical maps (Kohonen, 1981) have to be taken into account (for discussion see Kerling, 1988, pp. 411). Weigend et al. (1996) suggest a procedure called 'clearning' which combines the parameter estimation of the MLP and the adjustment of the training data in order to reduce the noise level. In contrast 'increase signal information' simply states that the more information is available about an optimization problem the more information should be implemented into the network structure a priori. Zimmermann/Weigend (1996) suggest an unconventional but powerful procedure. They enlarge the output layer in order to model some variables that are in direct link to the variable of interest simultanously. The interaction of this multivariate output is modelled a priori via an interaction layer. This procedure is another kind of regularisation but it uses useful information about the problem structure.

Akaike, H. 1974, “A New Look at the Statistical Model Identification“, IEEE Transactions on Automatic Control, Vol. AC-19, No.6, pp 716 -723.

Anders, U., 1995,. “Neuronale Netze in der Ökonometrie- Die Entmythologisierung ihrer Anwendungen“, Zentrum für Europäische Wirtschaftsforschung GmbH, Discussion Paper No. 95-26, Mannheim.

Anders, U., 1997, Statistische Neuronale Netze, München.

Efron, B., 1979, “Bootstrap Methods: Another Look at the Jackknife“, The Annals of Statistics, Vol. 7, No.1, pp. 1 - 26.

Efron, B., 1983, “Estimating the error rate of a prediction rule: improvements on cross-validations“, Journal of The American Statistical Association, Vol. 78, No. 382, pp. 316 -331.

Finnoff, W., Zimmermann, H.G., 1992, “Detecting Structure in Small Datasets by Network Fitting under Complexitiy Constraints“, Arbeitspapier, Siemens AG, Zentrale Forschungs- und Entwicklungsabteilung, München.

Geisser, S., 1975, “The Predictive Sample Reuse Method with Applications“, Journal of The American Statistical Association, Vol. 70, No. 350, pp. 320 - 328.

Geman, S., Bienenstock, E., Doursat, R., 1992, “Neural Networks and the Bias/Variance Dilemma“, Neural Computation, 4, pp. 1 - 58.

Hassibi, B., Storck, D.G., 1993, “Second order derivatives for network pruning: Optimal Brain Surgeon“, Advances in Neural Information Processing Systems 5, San Mateo, CA, pp.1164 - 171.

Hertz J., Krogh, A., Palmer, R.G., 1991, Introduction to the Theory of Neural Computation, Reading, MA.

Hornik, K., Stinchcombe, M., White, H., 1989, “Multilayer Feedforward Networks are Universal Approximators“, Neural Networks, 2:5, pp. 359 - 368.

Kerling, M., 1996, “Corporate Distress Diagnosis - An International Comparison“, Neural Networks in Financial Engineering, Singapore, pp. 407 - 422.

Kerling, M., 1998, Moderne Konzepte der Finanzwirtschaft, Bad Soden/Ts.

Kohonen, T., 1981 “Automatic Formation of Topological Maps in a Self-Organizing System“, Proceedings of the 2nd Scandinavian Conference on Image Analysis, pp. 214 - 220.

Le Cun, Y., Denker, J., Solla, S., 1990, “Optimal Brain Damage“, Advances in Neural Information Processing Systems 2, San Mateo, CA, pp. 598 - 605.

Moody, J.E., Utans, J., 1995, “Architecture Selection Strategies for Neural Networks: Application to Corporate Bond Rating Prediction“, Neural Networks in the Capital Markets, Chichester, pp. 276 - 300.

Mosteller, F., Tukey, J.W., 1968, “Data Analysis, Including Statistics“, The Handbook of Social Psychology, Reading, MA, pp. 80 - 203.

Murata, N., Yoshizawa, S., Amari, S., 1994, “Network Information Criterion - Determining the Number of Hidden Units for an Artificial Neural Network Model“, IEEE Transaction on Neural Networks, Vol. 5, No. 6, pp. 865 - 872.

Pi, H., 1993, “Dependency Analysis and Neural Network Modeling of Currency Exchange Rates“, Proc. of the first International Workshop on Neural Networks in the Capital Markets, London.

Poddig, Th., 1994a, “Ein Jackknife-Ansatz zur Strukturextraktion in Multilayer-Perceptrons bei kleinen Datenmengen“, Künstliche Intelligenz in der Finanzberatung Grundlagen - Konzepte - Anwendungen, Wiesbaden, pp. 333 - 345.

Poddig, Th., 1994b, “Mittelfristige Zinsprognose mittels KNN und ökonometrischer Verfahren“, Neuronale Netze in der Ökonomie, München, pp. 209 - 289.

Poddig, Th., 1996, Analyse und Prognose von Finanzmärkten, Bad Soden/Ts.

Schwarz, G., 1978, “Estimating the Dimension of a Model“, The Annals of Statistics, Vol. 6, No. 2, pp. 461 - 464.

Teräsvirta, T., Lin, C.-F., Granger, C.W.J., 1993, “Power of the neural network linearity test“, Journal of Time Series Analysis, Vol. 14, No. 2, pp. 209 - 220.

Tresp, V., Neuneier, R., Zimmermann, H.G., 1996, “Early Brain Damage“, Arbeitspapier, Siemens AG, Zentrale Forschungs- und Entwicklungsabteilung, München.

Weigend, A.S., Huberman, B.A., Rumelhart, D.E., 1990, “Predicting the Future: A Connectionist Approach“, Standford

Univeristy, Standford, CA.

Weigend, A.S., Zimmermann, H.G., Neuneier, R., 1996, “Clearning“, Neural Networks in Financial Engineering, Singapore, pp. 511 - 521.

White, H., 1989, “An additional hidden unit test for neglected nonlinearity in multlayer feedforward networks“, Proc. of the International Joint Congerence on Neural Networks, Washington, DC, Vol. II, pp 451 - 455.

White, H. ,1992, “Learning in Artifficial Neural Networks: A Statistical Perspective“, Artificial Neural Networks: Approximation and Learning Theory, Cambridge, MA, pp. 90 - 131, original in Neural Computation, 1, 1989, pp. 425 - 464.

Zimmermann, H.G., Weigend, A.S., 1996, “Finding Nonlinear Structure Using An Interaction Layer“, Arbeitspapier, Siemens AG, Zentrale Forschungs- und Entwicklungsabteilung, München.