Incorporation of Statistical Methods in Multi-step Neural Network Prediction Models

Guy-Michel Cloarec, John Ringwood *Senior Member*, IEEE

Abstract— **This paper addresses the problems associated with multi-step ahead prediction neural networks models. We will see how some concepts from the statistical theory field can be applied in various ways to improve the modelling. The generalization and error autocorrelation problems will be addressed using topological and methodological approach among which network committees, statistical bootstrap and principal component analysis will play a key role. These methods will be applied to the sunspot time series.**

Keywords— **Multi-step Ahead Prediction, Network Committees, Topology Optimization, Statistical Methods, Bootstrap, Principal Components Analysis.**

I. Introduction

THE determination of the optimal structure in neural networks modelling is a critical issue can cannot be adressed analytically. When some a priori knowledge or understanding of the problem is available or when the complexity of the function to map is low, it is possible to find a topology which is close to optimal. When the complexity of the problem increases, we have to use all the available methods, including techniques from the statistical theory field [1], [2].

Time series modelling is a class of problem where the goal is to approximate an hypothetical function describing the behaviour of a dynamical system on the basis of observations. The modelling task is then complicated by the ignorance of the optimal input space and all the model uncertainties tend to downgrade the prediction accuracy. Moreover, in applications such as predictive control that require multi-step ahead predictions, and therefore the use of past predicted values in the model, the optimality of the topology and of the network parameters are essential to ensur reliability.

It is well known that modelling methods are problem dependant and that any modelling methodology cannot pretend to be completly universal. However, it is possible to draw some general lines on the modelling methodology in the case of multi-step ahead prediction of time series with neural networks. The purpose of this paper will therefore be to demonstrate how the use of statistical methods can improve the performance of neural networks models on critical times series problem. The benchmark time series expressed on Figure 1 [3] examined here will be the sunspot series well known for its pseudochaotic dynamics and probably non-stationary behaviour.

Ing.Dip. G-M. Cloarec and Dr. J. Ringwood are with the Control Systems Group, School of Electronic Engineering, Dublin City University (DCU), Dublin, Ireland. E-mail: cloarecg, ringwood@eeng.dcu.ie.

The paper is organized as follows : Section II expresses the modelling problems associated with sunspot prediction and gives a review of the previous studies. Section III will aim at explaining the rational of the different techniques and methods used proposed our the modelling. The results of our modelling and the comparisons with other studies will be done in section IV while section V will conclude this study.

II. Sunspot Series Forecasting

The sunspot series is composed of the yearly mean of an arbitrarily defined sunspot number R_I that expresses the number of spots or group of spots on the surface of the sun [3]. Although such phenomena have been observed since 1700, there is still no physical explanation. The resulting time series is thought to be chaotic with a pseudoperiod of 12 years with some dynamical system behaviour. It has been used as a benchmark problem for various time series prediction methods [4], [5], [6].

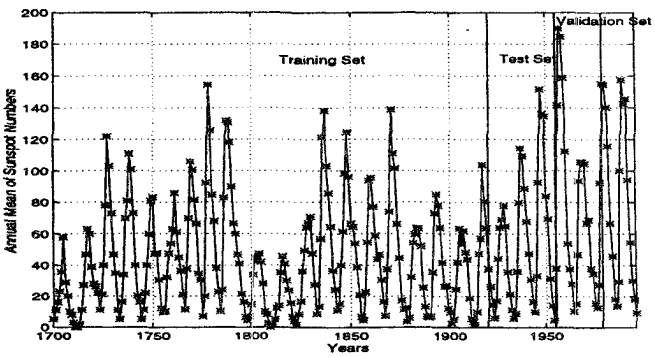

Fig. 1. Sunspot Numbers Variations.

Modelling of the sunspot series must observe several constrains. The first is the very limited number of data with regard to the 12 years pseudoperiod. Most of the statistical analysis and identification methods rely on statistical coherence of the data set, i.e. data sets on which it is possible to carry out generalization assessments. Another problem associated the limited number of data and the lack of understanding of the physical phenomena is the question of the time invariance of the sun considered as a dynamical system. The reduced observation time scale, with regard to astronomical time scales and the important variance of the pseudoperiodic signals both in amplitude and frequency (the *cycle* varies between 6 and 14 years), increases the uncertainty on the nature of the system studied [7]. The small data set leads us to consider the full set of reconstructed

0-7803-4859-1/98 $10.00©1998 IEEE

annual relative sunspot numbers and modelling methods that take advantage of the reduced data sets [3], [8].

In order to enable comparisons with previous studies, the data set obtained from the Sunspot Index Data Center [3] that covers the yearly means from 1700 to 1996, was split into three subsets expressed on Figure 1. The first one called *training set* covers values from 1700 to 1920. The *test set* covers the 1921-1955 period and the *validation set* 1956-1979. Since the number of samples available is critical for statistical coherence, used methods that enabled to use larger sets that will be refered as the *extended training set* covering the 1700-1955 period and the *extended validation set* covering the 1956-1996 period.

The modelling performed uses a feed-forward neural network that performs the mapping of the input space, the twelve previous sunspot numbers, to the output space i.e. the next sunspot number.

III. A Methods in Multi-step Neural Networks Prediction Modeling

A. Input Space Orthogonalization and Principal Component Analysis

The determination of the optimal topology is one of the most critical issues in neural network modelling [2], [9]. Only a minimal structure will enable good prediction and generalization properties. Despite attempts to estimate the problem complexity using Vapnik-Chervonenkis dimensions bounds [15], there is no theoretical result which gives the characteristics of the optimal neural network.

The curse of dimensionality [2] expresses the problem of an exponentially growing network as the input space grows. The first step in the modelling is therefore to optimize the input space. This is performed by the orthogonalization of the input space using the principal components. In this methodology new inputs are selected using linear systems model selection methods [10], [11], by the normalization of the transformed input space [12], [13] and by the selection of appropriate neural networks topology based on heuristic understanding of the problem [14].

Principal Component Analysis (PCA) is a multivariate analysis method that aims at determining correlation relationships between different variables [12], [13]. In neural network time series modelling context, these variables can be either the tapped delayed inputs or the output signals from each network layer [14]. The PCA method can be used to transform the data so they have maximum variations and are orthogonal. The crosscorrelation criteria are then used to evaluate the level of information contained within the transformed signals (the principal components scores), so that a dimensionality reduction can be performed. The orthogonalized signals can be interpreted as *extracted features* in multivariate data analysis, where the dimensionality reduction can be called compression, for example in the communications field [12], [13].

The first six principal components were selected, although only the first three explained 95 % of the data variance. The sunspot series and the first three components scores are expressed on Figure 2. The choice of the principal components is critical since the proportion of variance is not a causality criterion. We implemented linear models using principal components as inputs and statistical decision methods [11] to determine the *best* principal components in respect to prediction error. The rational for using orthogonalized inputs was to feed the network with the minimal information, or in other words to reduce the hypervolume of the input space and therefore the network size. Moreover, the use of uncorrelated inputs improves the convergence of the gradient based learning algorithms.

The orthogonalization was performed on data series made of the past 12 sunspot numbers [5]. The three first principal components, the main extracted features, can be associated respectively with a two or three years delayed signal and a nine years delayed signal, with the third component corresponding to the so-called *super-period* [3] that could also be interpreted as a signal expressing the non-stationarity nature of the system (see Figure 2).

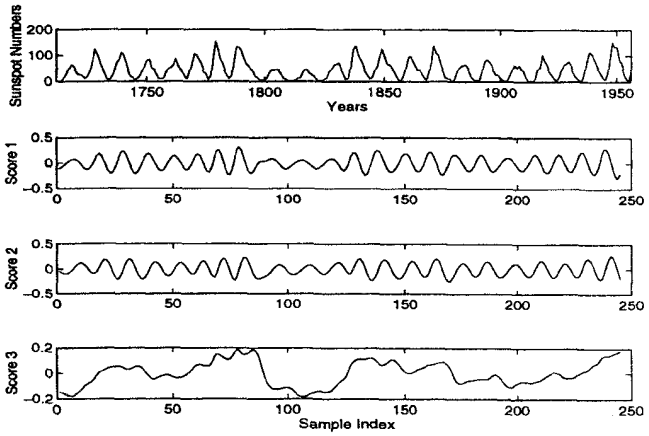

Fig. 2. Sunspot Series and Principal Components.

To determine the optimal topology, one should use the PCA on a trained network which is sufficiently large to perform off-line neuron pruning. Altough the term *sufficiently large* can be estimated using Vapnik-Chervonenkis dimensions [15], we used the results of the input space orthogonalization and of a linear modelling exercice (the complexity of the function is proportional to the number of parameters required in the linear model). Initially, a 6-6-1 MLP neural network was trained to map the transformed and reduced input space to the output space. The output signals of the input and hidden layers were analyzed using the principal component analysis method and possible redundant neurons were detected. The process was reiterated 10 times on different training subsets in order to achieve a good statistical inference on the topology determination and neuron pruning. The analysis indicated that a 4-3-1 structure would reduce correlations between layers and make the networks close to optimally regularized.

B. Statistical Bootstrap and Generalization Estimates

One of the problems main in neural network modelling is the overfitting problem [9] that leads to poor generalization. Assuming that generalization is possible, the problem is then to determine a reliable generalization error estimate that will be used for model selection and optimization. Classical methods use single-sample statistics (AIC, FPE, MDL criteria) [16], [11] that estimate the model performance on a single set (usually the training set), or the split-sample method that estimates the error on a test set which is independent of the training set.

Other methods such as cross-validation and bootstrap [17], [18], use statistical theory principles to obtain better estimates of the error without using test sets (which is an advantage with limited data sets problems). In k-folded cross validation, the training set is split into k distinct equal size sets and k-1 models are estimated using training sets composed of k-1 subsets where the one left out each time is different. The bootstrap method is used to decrease the statistical dependencies within subsets. The method consists to draw with replacement samples at random to form a given size training subsets [20]. For cross validation and bootstrap methods, the optimality hides in so called γ ratio problem [19] which relates to the choice of the number of subsets and the size of the subsets. A greater number of subsets improves the statistical inference, but may result with unacceptable small bootstrap set sizes, which do not adequately represent the coherent characteristics of the data. Small subsets may emphasize idiosyncrasies which increase the variance of the estimate error. We arbitrarily defined 36 samples (3 pseudoperiods) to be the minimal subset size although previous studies [4], [5] used smaller sets (23) for test and validation sets. The statistical bootstrap will give us statistics that will be used for inference estimation which is useful for model selection and generalization error estimation.

The training was performed on a 4-3-1 network using (successively) 3 bootstraps of 120 samples ($\gamma \approx 0.5$) and at each epoch the generalization error is estimated using 15 bootstraps of 36 samples each. Because of the hazards due to random initial conditions, the training was performed on an ensemble of 20 networks. The backpropagation algorithm was used during the first 10 epochs and was subsequently replaced by the Levenberg-Marquardt algorithm supervised by an early stop criterion.

Remedies to poor generalization are numerous and are known as regularization methods [9]. The *early stop* method, the oldest form of regularization, aims at the overfitting problem [20] and gives good results when good generalization estimates are available. It has been shown in [20] that the use of statistical bootstrap estimates of the generalization property are a *viable means for implementing an early stop rule.*

Figure 3 shows the variations of the model response MSE on the extended training and validation sets as well as the bootstrap generalization estimates when the network was trained to convergence. The early stop criterion was implemented on the basis of the minima of the generalization

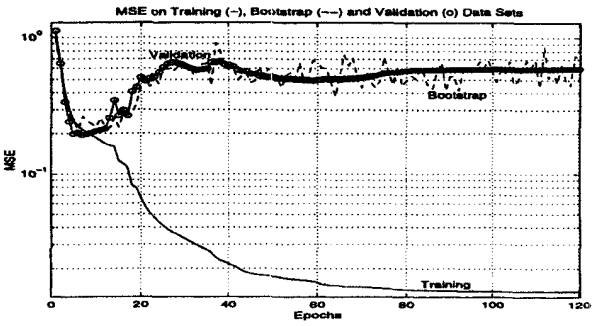

Fig. 3. Generalization Estimates in Early Stop Training

error estimates. The parameters associated with the minimum were memorized and the training stoped as soon as the MSE exceeded minimum value by 20 %. Such a measure is used to prevent premature cessation of training.

C. Network Committees

The concept of network committees or ensembles of networks [2], [22] comes from the field of statistical theory and is a natural extension of the estimates accuracy improvement through statistical data analysis. The commitees are collections of finite numbers of neural networks trained on different sample sets, with different initial conditions or learning criteria. Their used is known to improve the generalization performance by decreasing the effect of spatial and temporal dependencies.

Once trained, the network committee predictions can be used to improve the decision making process during the model structural determination. The predictions can also be combined to produce a better prediction. The network commitees were used during the network topology determination in conjunction with the prinicpal component analysis. By enabling statistical inference, the model selection process was made on the basis of confidence limits.

The committees were also used to perform improved prediction. While the committees are usually combined using linear combiners [22], [2], we choosed to use a neural network, designated as the *combining network.* The overall structure of the predictor is expressed on Figure 4 This network was designed to combine the multi-step ahead predictions of the committees to give a better prediction. Multistep ahead predictors are known to be very unstable and very sensitive to initial errors so that a neural network must be used to overcome the error propagation problem.

The networks were trained using different initial conditions on bootstrap ensembles originated from the same limited training set. It is important to obtain a *population* of networks with *spread* characteristics so that the performance analysis and the prediction combination are improved. The need for spread characteristics is also the rational for the use of the early stop training methods which gives emphasis on generalization rather than structural reduction. A drawback of spread characteristics [22] is that a portion of the committee downgrades the combination per-

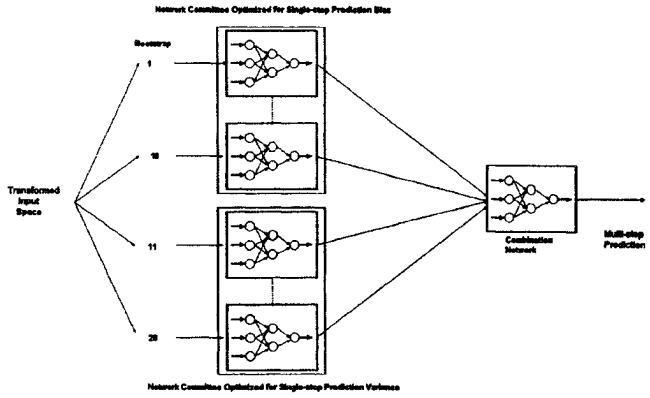

Fig. 4. Multi-step Prediction Neural Network Structure.

formance and so must be eliminated through a statistical process.

The PCA method was used to determine the initial topology of the committee of networks as well as the combination network. It has been shown in [20] that the more the networks are regularised, the better are the generalization estimates with resulting improvements with the early stop training method. The networks were trained, neurons were pruned using PCA results and the weights were subsequently put through an elimination process. The resulting committee of networks was regularized but with characteristics different enough to be combined efficiently.

D. Bias/Variance Dilemma and Trajectory Learning

The *Bias/Variance Dilemma* [1], [2] refers to the trade off between the prediction error bias and variance that the training algorithm must perform to give satisfactory generalization properties. This is usually expressed by the MSE criterion that can be decomposed into bias and variance components. The characteristics of the problem to solve may require that control of that trade off be exercised.

When committees are used, the trained networks exhibit different characteristics with regard to prediction error bias and variance. It is therefore possible to use that distribution in the combination process to influence the trade off. Moreover, it is possible to refine or to regularize further these networks with bias or variance oriented criteria.

The trade off between prediction error bias and variance is a critical issue when multi-step ahead forecasting is involved. It is well known that when past predictions are used to estimate the future values, the resulting prediction error is autocorrelated and the predictor may become unstable. A way to improve the predictor robustness and which enables a longer prediction horizon to be achieved is to emphasize the reduction of the error bias compared to the variance. However, the determination of such a trade off is difficult to estimate. One method of performing the optimal trade off is to traine the combination network with predicted inputs. The recursive nature of the learning process requires non conventional training algorithms gathered under the *trajectory learning* acronym [9].

The trajectory learning algorithm is difficult to set up and because of the autocorrelated error is prone to instability. We propose here a new method that has the advantage of performing trajectory learning but still uses classical training algorithms.

The key idea is to use multi step ahead predictions obtained by network committees trained with single step ahead prediction error criterion and optimize for error bias and variance, and then to combine them so that the overall multi step ahead prediction is improved. The combination network which has the multi step ahead predictions as inputs is trained using a classical learning algorithm. In other words, we use the statistical properties of the network commitees to improve the prediction and then increase the reliable horizon of prediction

To be efficient, the method requires *good* single-step ahead predictors so that their multi-step ahead predictions enable an efficient recombination. This was obtain by assessing the generalization properties of the networks. The robustness of the predictors was estimated for horizons of prediction varying between one and five years.

The horizon of prediction is determined on the basis of the prediction errors obtained by the committees of networks optimized for single-step ahead prediction and pruned in regard of the bias or variance criteria. Figure 7 expresses the prediction error obtained for horizon of prediction varying from one to seven years and obtained on the extended training data set.

When the horizon of prediction has been defined, it is possible to implement the weight elimination process. In order to increase the *spread*, i.e. the variations of the network population, a simple elimination without retraining algorithm supervised by prediction error bias and variance is performed on the network committee. The network committee is divided in two groups composed respectively of the networks that minimize the error bias and variance. One by one, the weights are eliminated from the network and the multi-step prediction is analysed with respect to error and variance criteria. If a small increase or a decrease of the prediction error criterion is observed, the weight is pruned. This simple algorithm, although far from being optimal, enables a drastic reduction of the number of parameters.

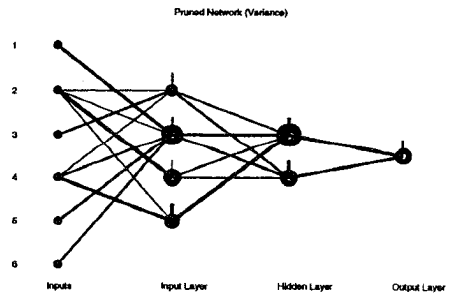

Fig. 5. Weight Elimination Optimized for Variance Criterion.

Figure 5 shows a network on which a weight elimination otpimized for minimizing the error variance has been per-

formed. The width of the weigths is proportional to their amplitude rank between the two layers. The size of the neurons is propotional to the rankink of the absolute sum of weights leaveing the neurons (neurons are ranked within layers). When color graphs are enabled, the color of the neurons depends on the rank of the absolute sum of the weights which fires the neurons. The use of such representations enables the *visualization* the network behaviour, to gives information on the *critical paths* within a network, as well as indications for supervised regularization.

The modeling procedure is outlined in Figure 9.

IV. Results

Sunspot series forecasting is a benchmark problem for nonlinear modelling [4], [16] and neural network modelling in particular [5], [6], [7]. One difficulty is the very nature of the problem : an astronomical phenomena for which only a limited number of observations (or reconstructions) are available, which is particularly small with regard to the length pseudoperiods observed, and the lack of physical understanding of the choatic sun behaviour.

[4] performed its modeling in 1980 and following studies [5], [7] evaluated their methods using the same limited sets. For example, both the test and validation sets were based on two pseudoperiods which, considering the erratic behaviour of the series, do not guarantee the significance of the performance estimates. With more data available and by using methods (such as statistical bootstrap) that enable to use larger data sets and better performance estimates, we performed the modelling using the data from 1700 to 1955 as the training set and 1956 to 1997 as the validation set.

However, to enable comparisons with previous studies, the modelling was first performed using the same data set. The results results are gathered on Table I, with the double figures indicating mean and standard deviation respectively.

Model	Training 1700-1920	Test 1921-1955	Validation 1956-1979
Tong and Lim [4]	0.097	0.097	0.28
Weigend et al. [5]	0.082	0.086	0.35
Svarer Linear [7]	0.132	0.13	0.37
Svarer Pruned [7]	$0.09^{0.001}$	$0.082^{0.007}$	$0.35^{0.05}$
Network Committee	$0.096^{0.002}$	$0.104^{0.032}$	$0.169^{0.0019}$
Hybrid Neural Network	0.09	0.085	0.157

TABLE I
Comparisons of the Normalized Error for Single-step Ahead Forecasting.

When performed on the extended training set, the single-step ahead predictions obtained on the extended validation sets are more representative of the modelling performance. The predictions and the actual sunspot numbers, along with the 95 % confidence limits, are expressed on Figure 6.

Obtaining satisfactory single-step prediction was only the first step of the multi-step ahead modelling process. Using the committee, we have performed mutli-step prediction with various horizon of prediction. The statistics

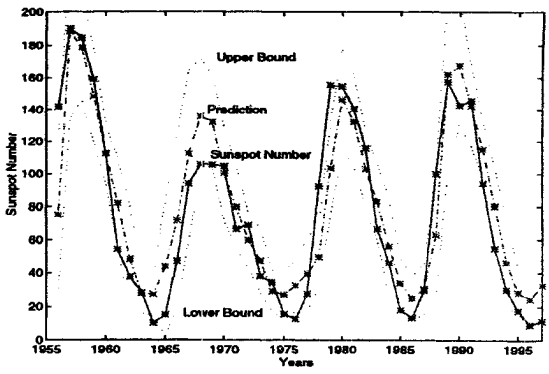

Fig. 6. One Step Ahead Forecast on the Validation Set.

of the normalized multi-step ahead prediction MSE on the extended sets are expressed in Figure 7.

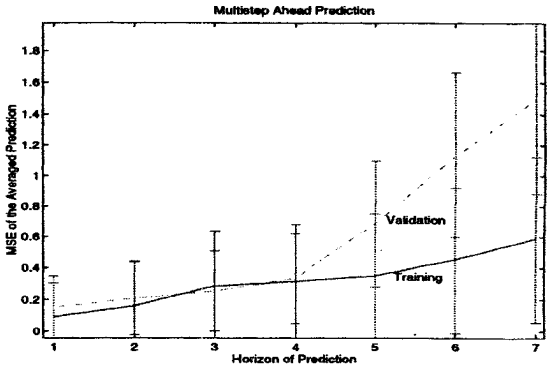

Fig. 7. Prediction Error Statistics on Extended Sets.

Figure 7 shows that beyond four-step ahead horizon, the prediction of the individual network committee members are not reliable. A combination network, that aimed at combining the four-step ahead prediction of a selected subcommittee of networks was now trained. The subcommittee members were selected in relation to their four-step ahead prediction error bias and variance. The final committee of networks was composed of 16 networks optimal for error variance and 8 networks optimal for error bias.

The prediction normalized MSE obtained on the extended training sets was 0.3119 while we obtained 0.3361 on the extended validation set for H=4. The four-steps ahead prediction is shown in Figure 8.

It is difficult to assess the method based on the sunspot series application since the non-stationarities make the evaluation not necessarily representative on small data sets.

V. Conclusions

A general methodology for multi-step neural network modelling has been presented with an emphasis on the various concepts and techniques related to the field of statistical theory. The transformation of the input space has been performed to optimize the neural network topology and an early stop training method using the bootstrap en-

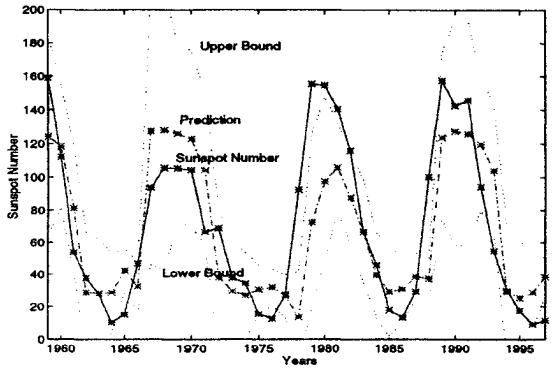

Fig. 8. Four-step Ahead Forecast on the Validation Set.

sembles, has been presented. Furthermore, network committees combined with a weight elimination method were used to deal with the bias/variance dilemma and to implement a new approach for trajectory learning. These proved to be satisfactory for the sunspot prediction.

This modelling methodology was applied to the sunspot series, given its benchmark status, but the very limited data sets and the non-stationarity of the system made modelling and model estimation more critical. This was partly overcame by the statistical methods which enabled us to use larger effective sets for both training and validation.

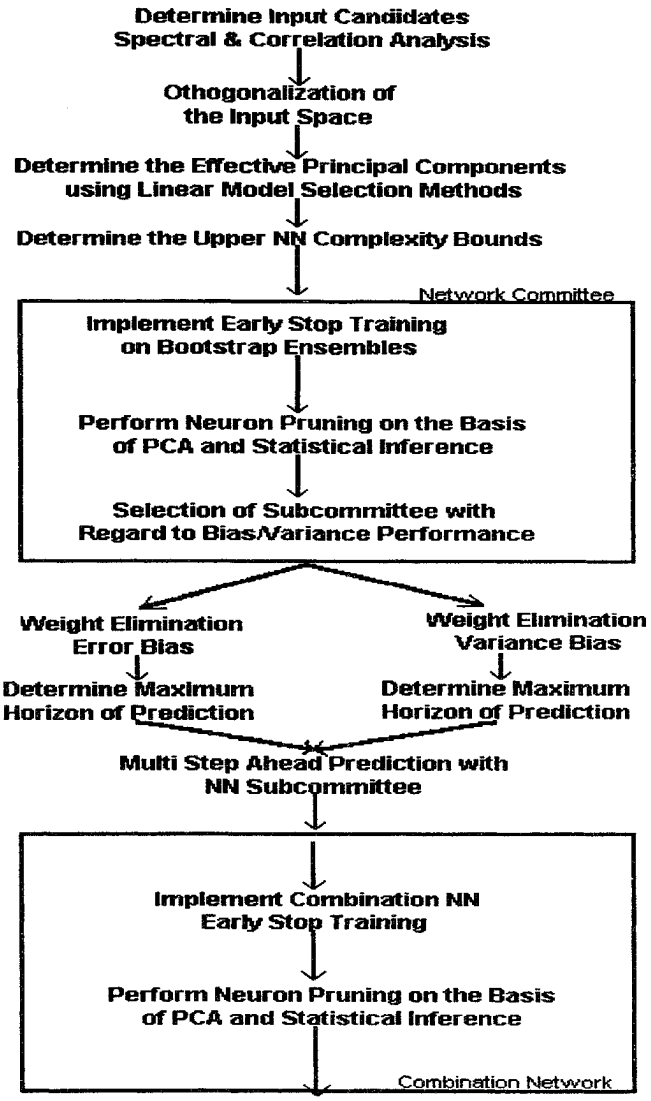

Fig. 9. Modeling Flowchart.

References

[1] B.D. Ripley, *Can Statistical Theory Help us Use Neural Networks Better ?*, 29th symposium on the Interface : Computing Science and Statistics, 1997.

[2] S. Lawrence, C.L. Giles, A.C. Tsoi, *What Size of Neural Network gives Optimal Generalization ?*, Technical Report UMIACS-TR-96-22, Institute for Advanced Computer Studies, University of Maryland MD, June 1996.

[3] P. Cugnon, *Sunspot Index Data Center*, Royal Observatory of Belgium, Av. Circulaire, 3-B-1180 Brussels.

[4] H. Tong, K.S. Lim, *Theshold Autoregression, Limit Cycles and Cyclical Data*, Journ. Roy. Stat. Soc. B, vol. 42, pp 245, 1980.

[5] A.S. Weigend, B.A. Huberman, D.E. Rumelhart, *Predicting the Future : A Connectionist Approach*, International Journal of Neural Systems, vol. 3, pp 193-209, 1990.

[6] Y.R. Park, T.J. Murray, C. Chen, *Predicting Sun Spots using a Layered Perceptron Neural Network*, IEEE Transactions on Neural Networks, vol. 7, No. 2, March 1996.

[7] C. Svarer, L.K. Hansen, J. Larsen, *On Design and Evaluation of Tapped-Delay Neural Network Architecture*, in H.R. Berendji *et al.* Proceedings of the 1993 IEEE Int. Conference on Neural Networks, IEEE Service Center, NJ, vol. 1, pp. 46-51, 1993.

[8] E.M. Azoff, *Reducing Error in Neural Networks Time Series Forecasting*, Neural Comput & Applic, vol. 1, pp 240-247, Springer-Verlag London Ltd, 1993.

[9] W. Sarle, *Comp.ai.neural-nets Frequently Asked Questions*, ftp://rtf.mit.edu/pub/usenet/news.answers/ai-faq/neural-nets.

[10] R. Johansson, *System Modeling and Identification*, Prentice Hall Information and System Sciences Series, 1993.

[11] J. Sjöberg, *Non-Linear System Identification with Neural Networks*, Ph.D. Thesis No 831, Division of Automatic control, Department of Electrical Engineering, Linköping University, Sweden, 1995.

[12] A.C. Rencher, *Methods of Multivariate Analysis*, Wiley Series in Probability and Mathematical Statistics.

[13] J.D. Jobson, *Applied Multivariate Data Analysis, Volume II : Categorical and Multivariate Methods*, Springer Texts in Statistics, 1992.

[14] S. Haykin, *Neural Networks : A Comprehensive Foundation*, Prentice Hall, 1994.

[15] A. Hole, *Vapnik-Chervonenkis Generalization Bounds for Real Valued Neural Networks*, Neural Computation 8, pp 1277-1299, MIT Press, 1996.

[16] G.E. box, G.M. Jenkins, *Time Series Analysis : Forecasting and Control*, HoldenDay, San Francisco, 1970.

[17] J.S. Hjorth, *Computer Intensive Statistical Methods : Validation, Model Selection and Bootstrap*, Chapman & Hall, 1994.

[18] B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, 1993.

[19] M. Kearn, *A Bound on the Error of Cross Validation using the Approximation and Estimation Rates, with Consequences for the Training-Test Split*, Neural Computation 9, pp 1143-1161, MIT Press, 1997.

[20] L.K. Hansen, J. Larsen, and T. Fog: *Early Stop Criterion from the Bootstrap Ensemble*, in Proceedings of IEEE Proceedings of ICASSP'97, vol. 4, pp. 3205-3208, Munich, Germany, April 1997.

[21] G-M. Cloarec, *Statistical Methods for Neural Network Prediction Models*, Research Report EE/JVR/97/2, Control Systems Group, School of Electronic Engineering, Dublin City University, 1997.

[22] P. Sollich, A. Krogh *Learning with Ensembles : How Overfitting can be Useful*, Advances in Neural Information Processing Systems, vol. 8, pp 190-196, 1995.