Proceedings of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia, June 12-14, 2006 (pp108-113)

The Optimal Multi-layer Structure of Backpropagation Networks

SONGYOT SUREERATTANAN[1], HUYNH NGOC PHIEN, NIDAPAN SUREERATTANAN, and NIKOS E. MASTORAKIS
Information Technology Department
Bank of Thailand
273 Samsen Rd., Pranakorn, Bangkok 10200
THAILAND

Abstract: - A novel algorithm obtained by using Bayesian information criterion (BIC) is presented to systematically choose the optimal multilayer network structure, via the number of hidden layers and hidden nodes of each layer, of backpropagation (BP) networks. Simulation results with daily data on stock prices in the Thai market show that the algorithm performs satisfactorily. Moreover, the proposed algorithm is also compared to Daqi-Shouyi method.

Key-Words: - Structure of backpropagation networks, Multilayer neural networks, Bayesian information criterion

1 Introduction

Among the available paradigms, backpropagation (BP) networks have the largest number of successful applications [1, 2]. In fact, they have almost become the standard for modeling, forecasting, and classification domains [3]. The BP method, discovered by Rumelhart et al. [4], is *a supervised learning technique* for training multilayer neural networks. The gradient descent (steepest descent) method is used to train BP networks by adjusting the weights in order to minimize the system error between the known output given by the user (actual output) and the output from the network (model output). To train a BP network, each input pattern is presented to the network and propagated forward layer by layer starting from the input layer until the model output is computed. An error is then determined by comparing the actual output with the model output. The error signals are used to readjust the weights starting from the output layer and backtracking layer by layer toward the input layer. This process is repeated for all training patterns until the system error converges to a minimum.

Once the number of nodes in the input and output layers have been decided, which normally depends upon the application under consideration, the important and difficult problem is how to optimally select the number of hidden layers and hidden nodes. Generally, a *trial-and-error* approach is used to determine the structure of a network in practice. Hirose et al. [5] proposed an algorithm to find the appropriate number of hidden nodes by changing the number of hidden nodes dynamically until a minimal number is found for which convergence (total mean squared error, MSE is less than a predetermined value such as 0.01) occurs. In order to compare several different models that have different numbers of parameters, a straight MSE may not be used directly [6]. Daqi and Shouyi [7] proposed an optimization method for the appropriate structures of BP networks by introducing an empirical formula for initially selecting the number of hidden nodes based on many applications in various fields and presenting a method for judging redundant hidden nodes. Although this method can perform satisfactorily to arrive at the appropriate structure, it can only apply to three-layer feedforward neural networks (one hidden layer).

In this paper, we proposed the algorithm to select the optimal network structure of BP, via the number of hidden layers and hidden nodes of each layer. In our experiment, we present the results of our simulation studies that were intended to assess the performance of the algorithm. For this purpose, we employed daily data on the stock prices in the Thai market for the comparative simulations.

2 Backpropagation Networks

Backpropagation (BP) method is a supervised learning technique for learning associations between input and output patterns as shown in Fig.1. It is a

[1] Corresponding: Email: songyots@bot.or.th; Telephone: +66-1-874-1664; Fax: +66-2-280-4590

generalization of the original two-layer perceptron (no hidden layer) introduced by Rosenblatt [8, 9], especially the version developed by Widrow and Hoff [10]. Therefore, it is also called *the generalized delta rule*. Like the delta rule, it is an optimization method based on steepest descent method that adjusts the weights to reduce the system error.

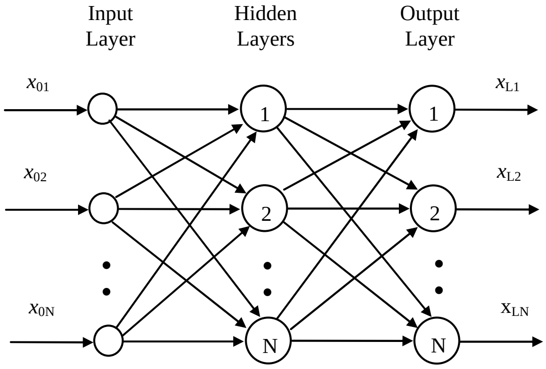

Fig.1 Backpropagation network architecture

Originally, the steepest descent method is used to train BP networks by using only the first derivatives of the error function. The error, *E*, for the network over all patterns is defined as (half) the sum of squared differences between the actual output and the model output in the output layer:

$$E = f(\boldsymbol{w}) = \sum_{p=1}^{M} E_p = \frac{1}{2}\sum_{p=1}^{M}\sum_{k=1}^{N_L}\left(o_{pk} - x_{pLk}\right)^2 \qquad (1)$$

where o_{pk} and x_{pLk} are the actual output and model output for the *k*th node in the output layer *L* and the *p*th training pattern, respectively, *M* is the number of data points, and N_L is the number of nodes in the output layer.

The goal is to evaluate the weights in all layers of the network that minimize the system error. In steepest descent, the search direction at the *t*th iteration is the negative of the gradient:

$$\boldsymbol{s}^t = -\nabla f\left(\boldsymbol{w}^t\right) \qquad (2)$$

and the weight update is

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \Delta\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \lambda\boldsymbol{s}^t = \boldsymbol{w}^t - \lambda\nabla f\left(\boldsymbol{w}^t\right) \quad (3)$$

where $\Delta\boldsymbol{w}^{t+1}$ is weight vector from $\boldsymbol{w}^t$ to $\boldsymbol{w}^{t+1}$, $\boldsymbol{s}^t$ is search direction of steepest descent, and λ is step size.

To train a BP network, each input pattern is presented to the network and propagated forward layer by layer until the output of the network is calculated, called model output. Then, the model output is compared to the actual output and an error is determined. The error signals are used to readjust the weights layer by layer in a backward direction. This process is repeated for each training pattern until the system error converges to a minimum.

3 Structure of Backpropagation Networks

Since BP training can be very costly, and the training cost increases as the network becomes more complex, the network should be kept as simple as possible. It means that few layers and nodes as needed. Generally, the number of nodes in the input and output layers depends upon the application under consideration. The number of output nodes will usually correspond to the number of different classifications needed or to the dimensions of the output vector space as required for a given mapping. It will usually be apparent from the application specification, for instance, in a time series forecasting, only one output node is required if a single value of future time is being predicted.

The number of input nodes required for the application may not be easy to determine since the nodes will usually correspond to object features, the independent variables. The features chosen should be relevant (essential) features, which best characterize the objects in the domain and hold them within the same class and discriminate well among objects belonging to different classes.

Determining the number of hidden layers and hidden nodes is more complicated than that for either input and output nodes. Therefore, for the appropriate network structure, the remaining problems are how to obtain the number of hidden layers and the number of hidden nodes for each layer.

Even though one hidden layer suffices for many applications [11, 12], two hidden layers may give better ability of generalization than one hidden layer, for more complex mappings. This is due to the fact that the nodes in one hidden layer tend to interact globally with each other, making it difficult to improve an approximation. On the other hand, in two hidden layers, the nodes in the first hidden layer can partition the input space into small regions, while the nodes in the second hidden layer can combine those outputs, giving rise to a more accurate mapping and better generalization [12].

Basically, network complexity measures are useful both to assess the relative contributions of different models and to decide when to terminate the network training. The performance measure should balance the complexity of the model with the number of training data and the reduction in the mean squared error, MSE [13].

Instead of the MSE, Akaike Information Criterion (AIC) [14] and Bayesian Information Criterion (BIC) [15, 16] can be employed to choose the best among candidate models having different numbers of parameters. While the MSE is expected to progressively improve as more parameters are added to the model, the AIC and BIC penalize the model for having more parameters and therefore tend to result in a smaller model. Both criteria can be used to assess the overall network performance, as they balance modelling error against network complexity.

The AIC, proposed by Akaike [14], has been extensively used. This criterion incorporates the parsimony criterion suggested by Box and Jenkins [17] to use a model with as few parameters as possible by penalizing the model for having a large number of parameters. The simplified and most commonly used form of the AIC is as follows:

$$\text{AIC} = M \ln(\text{MSE}) + 2P \quad (4)$$

where M is the number of data points used to train the network and P is the number of parameters involved in the model. For BP networks, the number of parameters is generally the number of weights and biases:

$$P = \sum_{i=0}^{L-1} N_{i+1}(N_i + 1) \quad (5)$$

Here N_i is the number of nodes in layer i and L denotes the output layer. MSE is defined as:

$$\text{MSE} = \text{SE} / M \quad (6)$$

where SE is the sum of squared errors.

In Eq. 4, the first term is a measure of fit and the second term is a penalty term to prevent over-fitting. When there are several competing models to choose from, select the one that gives the minimum value of the AIC.

Even if it is commonly used, when viewed as an estimator of the model order, the AIC has been found to be inconsistent [18]. Another model selection criterion, known as the Bayesian Information Criterion (BIC) or the posterior possibility criterion (PPC), was developed independently by Kashyap [15] and Schwarz [16]. The BIC can be expressed as follows:

$$\text{BIC} = M \ln(\text{MSE}) + P \ln(M) \quad [7]$$

The BIC also expresses parsimony but penalizes more heavily than the AIC models having a large number of parameters. As for the AIC, one selects the model that minimizes the BIC. It is known that the BIC gives a consistent decision rule for selecting the true model. As such, the BIC is proposed as the sole criterion for use in the determination of the optimal structure of the BP networks.

4 Proposed Algorithm

Along with the BIC criterion, we propose a new method to systematically determine the optimal network structure using a procedure that gradually increases the network complexity based on the BIC. The procedure starts with a small number of hidden nodes and trains the network until the system error is below an acceptable level. Then add a hidden node and retrain the network. This process is repeated until the current value of BIC is greater than the previous one or the decrease in BIC value becomes smaller than some small number. The proposed algorithm can apply to multilayer feedforward neural networks, and is not restricted to three-layer feedforward neural networks. It means that the algorithm can apply to networks more than one hidden layers. The algorithm can be summarized as follows:

1. Create an initial network with one hidden node and randomize the weights.
2. Train the network using with a chosen method e.g. the original BP algorithm until the system error has reached an acceptable criterion. A simple stopping rule is introduced to indicate the convergence of the algorithm. It is based upon the relative difference of the sum of squared errors (SE):
$$\left|\frac{\text{SE}(t+1) - \text{SE}(t)}{\text{SE}(t)}\right| \le \varepsilon_1 \quad [8]$$
where ε_1 is a constant that indicates the acceptable level of the algorithm and SE(t) denotes the value of SE at iteration t.
3. Check for terminating the selection of the network. A termination criterion is suggested based on the relative difference of BIC as follows:
$$\left|\frac{\text{BIC}(k+1) - \text{BIC}(k)}{\text{BIC}(k)}\right| \le \varepsilon_2 \quad [9]$$
where ε_2 is a constant that indicates the acceptable level for the structure of the network and k denotes the number of hidden nodes of the network. If the current value of BIC is greater than the previous one or the relative difference of BIC is less than or equal to ε_2, go to step 4;

otherwise add a hidden node and randomize the weights then go to step 2.

4. Reject the current network model and replace it by the previous one, then terminate the training phase.

5 Experimental Results

The stock market is an important institution serving as a channel that transforms savings into real capital formation. It will stimulate economic growth and also increases the gross national product (GNP). In this study, daily data on the stock prices and volumes in the Thai market from 1993 to 1996 were used. For the gap from Friday to Monday (weekend) and holidays when the stock exchange is closed, the data are treated as being consecutive. Three different types of common stocks; namely, Bangkok Bank Public Company Limited (BBL) in the banking sector, Shin Corporations Public Company Limited (SHIN) in the communication sector, and Land and Houses Public Company Limited (LH) in the property development sector, were chosen.

Bangkok Bank, Shin Corporations, and Land and Houses are the most important companies in their sectors. From the beginning in December 1944, Bangkok Bank has grown rapidly and is today not only the largest Thai commercial bank, but also one of the largest ones in South East Asia. Bangkok Bank has been a pioneer in the use of modern technology in its operations and in the introduction of a full range of electronic banking products to the domestic market [19]. Authorized capital is approximately 20,000 million Baht [20].

Shin Corporations Group of Companies is Thailand's leading broad-based telecommunications company, offering a comprehensive range of services including mobile phones, pagers, satellites, Internet, and data communications. Shin Corporations has established itself as a major player in the Thai telecommunications industry. At present, the Group shares 52% of the cellular market, 40% of the paging market, and is the sole operator of the national satellite network, namely THAICOM [21]. Authorized capital is approximately 5,000 million Baht [20].

Land and Houses is a leading company in property development in Thailand. The principal role of Land and Houses is the construction of high quality residential buildings for sale to customers. To maintain its high quality standards, Land and Houses has its own in-house design team (architects and engineers) and normally appoints an outside consultant as supervisor. Authorized capital is approximately 7,463.65 million Baht [20].

The data were obtained from the Stock Exchange of Thailand (SET). In each case, the data are divided into a calibration part for training and validation part for testing: 1993 to 1994 and 1995 to 1996, respectively. Before being presented to the network, the data are transformed by a linear (affine) transformation to the range [0.05, 0.95]. In this study, the input to the network may consist of the past values of stock price (P) and stock volume (V). The stock price at time t+1 is treated as a function of past values of stock price at times t, t-1, and t-2 and stock volume at times t, t-1 and t-2 as follows:

$$P(t+1) = g(P(t), P(t-1), P(t-2), V(t), V(t-1), V(t-2)) \quad [10]$$

In order to evaluate the performance of the network model, the efficiency index (EI), proposed by Nash and Sutcliffee [22], is employed to measure the performance of a given model:

$$\mathrm{EI} = \mathrm{SR} / \mathrm{ST} \quad [11]$$

$$\mathrm{SR} = \mathrm{ST} - \mathrm{SE} \quad [12]$$

$$\mathrm{ST} = \sum_{i=1}^{M} \left(y_i - \bar{y} \right)^2 \quad [13]$$

$$\mathrm{SE} = \sum_{i=1}^{M} \left(y_i - \hat{y}_i \right)^2 \quad [14]$$

$$\bar{y} = (1/M) \sum_{i=1}^{M} y_i \quad [15]$$

where SR = Variation explained by the model,
ST = Total variation,
SE = Total sum of squared errors,
y_i = Actual output, i.e. observed value at time i,
$\bar{y}$ = Mean value of the actual output,
$\hat{y}_i$ = Model output, i.e. forecast value at time i,
M = Number of data points (training patterns).

In the Daqi-Shouyi method [7], the number of hidden nodes is firstly selected to be N_1 = 5 for all stock companies; therefore the initial network structure is 6-5-1. After training of the network is completed, the eigenvalues λ of $\boldsymbol{H}^T\boldsymbol{H}$ are computed:

λ={0.0012,0.0557,0.1072,24.8287,604.453}for BBL,
λ={0.0087,0.0909,0.1270,19.9092,740.577}for SHIN,
λ={0.0025,0.0816,0.1154,24.4984,638.568}for LH.

The admissible error limit of the Daqi-Shouyi method is ε = 0.3464 for all companies. Three hidden nodes are discarded; hence the 6-2-1 network is obtained for all companies.

In our approach, the structure 6-1-1 network is determined initially for the BIC method. By using the BIC algorithm to train the network, the algorithm is stopped with the structure 6-4-1 network for all stock companies and thus the 6-3-1 network is the best, as depicted in Table 1.

The network with two hidden nodes was obtained from the Daqi-Shouyi method whereas the network with three hidden nodes was obtained from the BIC method. Although the BIC method chooses more hidden nodes, the values of the efficiency index for the network with three hidden nodes are higher than those for the network with two hidden nodes, indicating better model performance, as illustrated in Table 1.

Table 1 Computed values of efficiency index (EI) and BIC for stock prices in Thai market

Stock	6-1-1		6-2-1		**6-3-1**		6-4-1	
	EI	BIC	EI	BIC	**EI**	**BIC**	EI	BIC
BBL	0.97	-3,310.96	0.98	-3,431.71	**0.99**	**-3,593.22**	0.99	-3,579.97
SHIN	0.94	-3,057.91	0.96	-3,227.42	**0.97**	**-3,332.04**	0.97	-3,240.40
LH	0.98	-3,459.62	0.99[*]	-3,669.86	**0.99**[**]	**-3,863.04**	0.99	-3,726.92

From all experiments considered, the Daqi-Shouyi method and the BIC method represent nearly the same results to obtain the appropriate network structure. Since the network is trained only once in the Daqi-Shouyi method, it requires less computation time. However, training in the Daqi-Shouyi method does not actually apply to the discarded and selected networks (just in training for the initial network). The discarded networks may have good results with different situations such as different weight initializations. Moreover, the Daqi-Shouyi method can only apply to the networks with one hidden layer. On the other hand, the BIC method can be used for the networks with more than one hidden layer.

When there are several network models to choose from, the one that gives the minimum value of the BIC is selected to be the optimal network. In Table 2, a comparison is made between the networks with one hidden layer that were selected to be the optimal structure and the networks with two hidden layers. It is clear that the BIC method can help to eliminate the networks with two hidden layers in the data sets considered. Therefore, the optimal structures of the network with 6-3-1 for three stock companies, namely BBL, SHIN, and LH are obtained.

Table 2 Computed values of BIC for stock prices in Thai market between one and two hidden layers

Stock company	Structure			
	6-3-1	6-1-1-1	6-2-1-1	6-3-1-1
BBL	**-3,593.22**	-3,539.68	-3,519.56	-3,515.05
SHIN	**-3,332.04**	-3,283.32	-3,313.44	-3,232.35
LH	**-3,863.04**	-3,765.67	-3,818.11	-3,705.30

6 Conclusion

The proposed method using Bayesian Information Criterion (BIC) can be used to systematically determine the optimal network structure. The procedure begins with a small number of hidden nodes and gradually increases the network complexity, and then employs the BIC to terminate the network training. To compare models which have different numbers of parameters, mean squared error (MSE) cannot be used directly because normally MSE should be reduced when the number of parameters of the network increases. Instead of the MSE, the BIC can be used to choose the best model from the candidate models, having different numbers of parameters. It should be noted that while the MSE is expected to progressively improve as more parameters are added to the model, the BIC penalizes the model for having more parameters and therefore tends to result in a smaller model. As it balances modelling error against network complexity, the BIC can be used to assess the overall performance of the model.

References:

[1] N.A. Gershenfield and A.S. Weigend, *The Future of Time Series*, Technical Report, Palo Alto Research Center, 1993.

[2] H.N. Phien and J.J. Siang, Forecasting Monthly Flows of the Mekong River using Back Propagation, *Proc. IASTED Int. Conf.*, 1993, pp.17-20.

[3] C.H. Chu and D. Widjaja, Neural Network System for Forecasting Method Selection, *Decision Support Systems*, Vol.12, 1994, pp.13-24.

[4] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, Learning Internal Representations by Error Propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol.1: Foundations*, D.E.

[*] the actual value is 0.9914.

[**] the actual value is 0.9933.

Rumelhert and J.L. McClelland (eds.), MIT Press, Cambridge, Massachusetts, 1986, pp.318-362.

[5] Y. Hirose, K. Yamahsita, and S. Hijiya, Back-Propagation Algorithm which Varies the Number of Hidden Units, *Neural Networks*, Vol.4, 1991, pp.61-66.

[6] M. Brown and C. Harris, *Neurofuzzy Adaptive Modelling and Control*, Prentice Hall, UK., 1994, pp.326-327.

[7] G. Daqi and W. Shouyi, An Optimization Method for the Topological Structures of Feed-Forward Multi-Layer Neural Networks, *Pattern recognition*, Vol.31, No.9, 1998, pp.1337-1342.

[8] F. Rosenblatt, The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, *Psycho. Rev.*, Vol.65, No.6, 1958, pp.386-408.

[9] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington, D.C., 1961.

[10] B. Widrow and M.E. Hoff, Adaptive Switching Circuits, *IRE WESCON Conv. Record*, Part 4, 1960, pp.96-104.

[11] K. Hornik, M. Stinchcombe, and H. White, Multilayer Feedforward Networks are Universal Approximators, *Neural Networks*, Vol.2, 1989, pp. 359-366.

[12] D.W. Patterson, *Artificial Neural Networks: Theory and Applications*, Prentice Hall, Singapore, 1996.

[13] M. Pottmann and D.E. Seborg, Identification of Nonlinear Processes using Reciprocal Multiquadratic Functions, *J. Proc. Cont.*, Vol.2, No.4, 1992, pp.189-203.

[14] H. Akaike, A New look at the statistical model identification, *IEEE Trans. Autom. Control*, AC-19, 1974, pp.716-723.

[15] R.L. Kashyap, A Bayesian Comparison of Different Classes of Dynamic Models using Empirical Data, *IEEE Trans. Automatic Control* AC-22(5), 1977, pp.715-727.

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

[17] G.E.P. Box and G.M. Jenkins, *Time Series Analysis Forecasting and Control*. Revised Edition, Holden-Day, San Francisco, 1976.

[18] R.L. Kashyap, *Inconsistency of the AIC rule for estimating the order of autoregressive models*, Technical Report, Dep. of Electr. Eng., Purdue Univ., Lafayette, 1980.

[19] Bangkok Bank, <URL:http://www.bbl.co.th/aboutus/about_history.htm>, March 2000.

[20] The Stock Exchange of Thailand, <URL: http://www.set.or.th/cgi-bin/hsimscom.ksh>, March 2000.

[21] Shin Corporations, <URL:http://www.shinawatra.com/groupprofile/groupprofile.htm>, March 2000.

[22] J.E. Nash and J.V. Sutcliffee, River Flow Forecasting through Conceptual Models, *Journal of Hydrology*, Vol.10, 1979, pp.282-290.