2001 년 한국정보처리학회 추계 학술발표논문집 제 8 권 제 2 호

시계열예측에 대한 역전파 적용에 대한 결정적, 추계적 가상항 기법의 효과

조태호
한국과학기술정보연구원, 정보기술부
e-mail : tcjo@netian.com

The Effect of Deterministic and Stochastic VTG Schemes on the Application of Backpropagation to Multivariate Time Series Prediction

Taeho C. Jo
Dept. of Information Technology, KISTI

Abstract

Since 1990s, many literatures have shown that connectionist models, such as back propagation, recurrent network, and RBF (Radial Basis Function) outperform the traditional models, MA (Moving Average), AR (Auto Regressive), and ARIMA (Auto Regressive Integrated Moving Average) in time series prediction. Neural based approaches to time series prediction require the enough length of historical measurements to generate the enough number of training patterns. The more training patterns, the better the generalization of MLP is. The researches about the schemes of generating artificial training patterns and adding to the original ones have been progressed and gave me the motivation of developing VTG schemes in 1996. Virtual term is an estimated measurement, X(t+0.5) between X(t) and X(t+1), while the given measurements in the series are called actual terms. VTG (Virtual Term Generation) is the process of estimating of X(t+0.5), and VTG schemes are the techniques for the estimation of virtual terms. In this paper, the alternative VTG schemes to the VTG schemes proposed in 1996 will be proposed and applied to multivariate time series prediction. The VTG schemes proposed in 1996 are called deterministic VTG schemes, while the alternative ones are called stochastic VTG schemes in this paper..

1. Introduction

Time series prediction is the process of forecasting a future measurement by analyzing the pattern, the trends, and the relation of past measurements and the current measurement [1]. Time series prediction is studied in the several fields: data mining in computer science, industrial engineering, business management & administration and other fields. The domains of time series prediction are various from financial area to natural scientific area: stock price, stock price index, interest rate, exchanging rate of foreign currencies, the amount of precipitation, and so on. The traditional approaches to time series prediction are statistical models: AR (Auto Regressive) , MA (Moving Average), ARMA (Auto Regressive Moving Average), and Box-Jekins Model [1]. These models are mainly linear models and the trends of time series should be analyzed before applying them to time series prediction.

Literatures have shown that neural-based approaches, such as back propagation, RBF (Redial Basis Function), and recurrent network, outperform the traditional approaches (statistical models) in the performance of predicting future measurements. In the neural based approaches, back propagation is used most commonly; it has the ability of universal approximation [2]. In 1991, A.S. Weigend and D.E. Rumelhart proposed the first neural based approach, back propagation, to time series prediction [3]. N. Kohzadi presented that back propagation outperforms one of statistical models, ARIMA, in the performance of forecasting the price of cow and wheat flour, in 1996 [4]. D. Brownstone presented that back propagation is more excellent than multi linear regression in forecasting stock market movement [5]. M. Milliaris also presented that back propagation is over one of traditional models, Black-Scholes model, in the performance of forecasting of S&P 100 implied volatility [6]. Note that Black-Scholes model is most commonly applied to predict S&P 100 implied volatility in the statistical models [6]. A.U Levin proposed back propagation in the selection of the beneficial stocks [7]. In 1997, J. Ghosn and Y. Bengio predicted the profits of the stock using neural network [8]. So, These literatures show that neural-based approaches should replace the statistical ones to time series prediction.

The neural-based approaches require the enough length of historical data. Essentially, neural networks require many training patterns enough for the robust generalization [2]. Training patterns for the neural network, what is called time delay vectors, are generated from the time series in the training period by sliding window. The longer the historical length of time series in training period, the more the training patterns generated. The number of training patterns influences the generalization performance.

Actually, training patterns are not always given enough for the robust generalization. It is necessary to maintain the robust generalization performance, although training patterns are not enough. It is proposed that the generalization performance is improved by generating derived training patterns from the original ones and adding the derived training patterns to the original ones. Here, let's assume that the training patterns given originally are called natural training patterns, while the training patterns generated from them are called artificial training patterns. The use of both natural training patterns and artificial training patterns for training the neural network improves its generalization performance. In 1993, Abu-Mustafa proposed the use of hints, the artificial training patterns generated by the prior knowledge about the relations between input vector and output vector of the natural training patterns [9]. In 1995, Abu-Mustafa presented that hints contributed to reduce the prediction error in forecasting the exchange rate between USD (US Dollar) and DM (Deuch Mark) [10]. In 1994, D.A. Cohn, Z. Ghahramami, and M. J. Jordan proposed active learning, in which the neural network is trained by generating several artificial training patterns from each natural training pattern, simultaneously [11]. In 1995, A. Krogh and J. Vedelsby applied active learning to multiple neural networks [12]. In 1996, G. An proposed the scheme of generating artificial training patterns by adding noise to each natural training pattern and training the neural network with both artificial ones and natural ones [13]. And he validated that his scheme contributed to reduce the generalization error through the sine function approximation and digit recognition [13]. In 1997, S. Cho, M. Jang, and S. Chang proposed the scheme of training neural network with the natural training patterns and the artificial ones. The artificial training patterns are called virtual samples, in which input pattern is randomized and the output pattern is determined with the committee of neural networks [14]. D. Saad and S.A. Solla applied the combination of An's scheme and weight elimination in the process of training the neural network [15]. Y. Grandvalet, S. Canu, and S. Boucheron proved that G. An's scheme improve the generalization performance theoretically [16].

As mentioned above, Abu-Mustafa's scheme needs the prior knowledge about the natural training patterns; this scheme can not work without the prior knowledge. Actually, the prior knowledge to generate hints is not always given. In the D.A. Cohn's scheme and Cho's scheme, the generation of artificial training patterns is rule of sum and very heuristic. The effect of both schemes depends on the process of generating the artificial training patterns. Except Abu-Mustafa's scheme, almost mentioned schemes are validated through toy experiments: the function approximation [13][14] and robot arm kinematics [14].

In 1996, T.C. Jo proposed VTG (Virtual Term Generation) schemes of improving the precision of time series prediction by estimating the midterm X(t+0.5) between X(t) and X(t+1) [17]. Virtual term is the estimated value of X(t+0.5), between X(t) and X(t+1), while actual term is the given term in the time series [17]. VTG (Virtual Term Generation) means the process of estimating virtual terms, and VTG schemes are the techniques of estimating virtual terms. In 1997, T.C. Jo proposed the several schemes of estimating midterms and applied them to forecasting the annual number of sunspots [18]. All of the proposed VTG schemes contributed to reduce the predicted error [18]. In 1998, T.C. Jo applied the VTG schemes to multi-variable time series prediction: the prediction of monthly precipitation in west, middle, and east area of the State, Tennese of USA [19]. In 1999, T.C. Jo applied the VTG schemes to forecasting S&P 500 stock price index in financial area [20]. The VTG schemes in [17] and [18] are called deterministic VTG schemes in this paper. Deterministic VTG schemes means the method of estimating virtual terms with a particular equation. The deterministic VTG schemes proposed in [17] and [18] are mean method, 2nd LaGrange method, and 1st Taylor method. Mean method is the scheme of estimating the virtual term X(t+0.5) by averaging the adjacent actual terms, X(t) and X(t+1). Second LaGrange method is the scheme of estimating virtual terms with the equation derived from 2nd Lagrange interpolation. All of deterministic VTG schemes reduced the prediction error compared with the case of naïve neural-based approach: the neural-based approach to time series without VTG.

In this paper, alternative VTG schemes will be proposed and compared with the deterministic ones. These VTG schemes are stochastic ones: uniform VTG scheme, normal VTG scheme, and triangle VTG scheme. Stochastic VTG schemes are the methods of estimating virtual terms with random value, while deterministic VTG schemes do not use random values to estimate virtual terms. The estimated values are variable to each trial of VTG with same scheme in stochastic VTG schemes, while the estimated values are constant to each trial of VTG with same scheme in deterministic VTG schemes. The advantage of stochastic VTG schemes over the deterministic VTG schemes is the diversity of virtual terms with same scheme. This advantage means that the stochastic VTG schemes have the potential possibility of optimizing the estimated values of virtual terms with several trials or evolutionary computation. Another advantage over the deterministic VTG schemes is simplicity in its application to VTG, except mean method. Both second LaGrange method and first order Taylor method are more complicated than mean method or the stochastic VTG schemes. The estimated value of each virtual term is between two adjacent actual terms; the value is almost mean of them. In this paper, the basis of the stochastic VTG schemes is mean method in deterministic VTG schemes; the estimated value of each virtual term is determined by adding the mean of the adjacent actual terms and random value. The stochastic VTG schemes proposed in this paper are uniform method, normal method, and triangle method, based on the distribution for generating random values.

The model of neural network applied to time series prediction in this paper is backpropagation. The model, backpropagation, is used most commonly in the models of neural network. Although there are many models of neural networks in the world, backpropagation is applied to majority of fields in supervised learning. The reason of using backpropagation commonly is that the model is implemented most easily in the models of neural network in the world. The learning algorithm of backpropagation will be included in [2], and skipped in this paper.

In the organization of this paper, both kinds of VTG schemes will be described in the next section. In third section, two application methods of backpropagation to multivariate time series prediction, separated method and combined method, will be described [21]. In fourth section, conditions, procedure, and results of experiment to compare couple kinds of VTG schemes will be presented. The data used in the experiment is the artificial time series generated from a dynamic system, Mackay Glass equation. In the fifth section, the meaning and discussion of this studies and remaining tasks to improve the proposed scheme will be mentioned as the conclusion of this paper.

2. VTG Scheme

In this section, the schemes of estimating virtual terms will be described. There are two kinds of VTG schemes; one kind is deterministic VTG schemes proposed in [17] and [18], the other kind is stochastic VTG schemes proposed in this paper. Deterministic schemes are the schemes of estimating each virtual term with a particular equation, while stochastic schemes are the schemes of estimating each virtual term with an equation and a random value. Deterministic VTG schemes consist of mean method, second order LaGrange method, and first order Taylor method. And stochastic VTG schemes consist of uniform method, normal method, and triangle method.

Deterministic VTG schemes

This section will describe the VTG schemes proposed in 1996 and 1997: mean method, 2nd Lagrange method, and 1st Taylor method [17] [18]. In mean method, a virtual term, is estimated by averaging the two values, X(t) and X(t+0.5) like the Eq.(1).

$$\hat{X}(t+0.5) = \frac{1}{2}(X(t) + X(t+1)) - -(1)$$

In 2nd LaGrange method, two 2nd polynomials, $P_{21}(x)$ and $P_{22}(x)$ are constructed based on Lagrange interpolation like the following this, in the assumption that the given points are $(0, X(t)), (1, X(t+1)), (2, X(t+2))$ and $(0, X(t-1)), (1, X(t)), (2, X(t+1))$.

$$P_{21}(x) = \frac{1}{2}[(x - X(t+1))(x - X(t+2)) - 2(x - X(t))(x - X(t+2)) + (x - X(t))(x - X(t+1))]$$

$$P_{22}(x) = \frac{1}{2}[(x - X(t))(x - X(t+1)) - 2(x - X(t-1))(x - X(t+1)) + (x - X(t-1))(x - X(t))]$$

The virtual term, $X(t+0.5)$ is estimated by averaging values of $P_{21}(0.5)$ and $P_{22}(1.5)$, like Eq. (2).

$$X(t+0.5) \approx P_{21}(0.5) \approx P_{22}(1.5)$$

$$\hat{X}((t+0.5) = \frac{1}{2}(P_{21}(0.5) + P_{22}(1.5)) - - - (2)$$

But if is 0 or $T-1$, $X(t+0.5)$ is estimated with $P_{22}(1.5)$ either $P_{21}(0.5)$ or.

$\hat{X}((t+0.5) = P_{21}(0.5)$ if $t=0$

$\hat{X}((t+0.5) = P_{21}(1.5)$ if $t=T-1$

The nodes in the output layer generated the probability of the category given the set of words as the input pattern. Therefore, output pattern is a numerical vector consisting of normalized vales from 0 to 1, like the backpropagation. The process of computing a vector consisting of probability of each category from the set of words selected from a particular document is called generalization, and will be discussed in the subsection 2.3.

Stochastic VTG schemes

The base equation of stochastic VTG schemes is eq.(3).

$$\hat{X}(t+0.5) = \frac{1}{2}(X(t) + X(t+1)) + \varepsilon --- (3)$$

In the above equation, the estimated value of X(t+0.5) is the summation of a random value, ε , and the average of two adjacent actual terms. The consideration of the stochastic VTG schemes is the method of generating a random value, ε .

Uniform method of stochastic VTG schemes is the scheme of estimating each virtual term with the summation of the average of adjacent actual terms and a random value generated from the uniform distribution in figure 1.

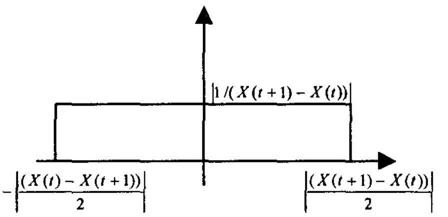

Figure 1. The uniform distribution for generating a random value

In figure 1, the estimated value of each virtual term from X(t) to X(t+1) with constant probability. The x-axis means the random value, ε , while y-axis means the probability of generating each random value ε . The probability is constant to all random values within the given range.

The second stochastic method, normal method, is the method of estimating virtual term with the summation of their average and the random value, ε , is generated based on normal distribution. In its parameters, mean is 0 and the standard deviation is $|1/2(X(t+1)-X(t))|$.

The third stochastic method, triangle method, is the scheme of estimating each virtual term with the summation of their average and the random value based on the distribution in the figure 2. Unlike the uniform distribution, this scheme has the most probability that the estimated value of a virtual term, X(t+0.5) is the average of two adjacent actual terms, X(t) and X(t+1).

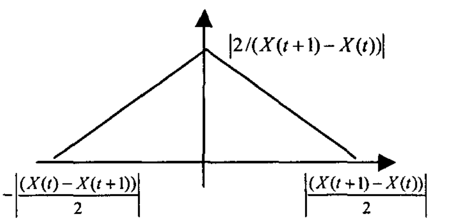

Figure 2. The triangle distribution for generating a random variable

3. Application of Back Propagation

This phase is to predict the value of future by composing training sample from a time series including virtual terms. For first, the neural approach to time series prediction is mentioned.

Without virtual terms, the time series is like the following this.

X1(1), X1(2), ... ,X1(T)
X2(1), X2(2), ... ,X2(T)
....
Xn(1), Xn(2), ... ,Xn(T)

The training pattern from the above time series is like the following this in separate model, with the shift of 1 step[10]. These patterns are composed like that from univariate time series in [1][3]. In this case, the variables belonging to the given time series are independent among them.

input: [Xk(t-d), Xk(t-d+1), ... , Xk(t-1)]
output: Xk(t)

But in combined model, the training pattern from above time series is like the following this and the variables included in the time series are influenced among them.

input:[X1(t-d), X1(t-d+1), ... ,X1(t-1), ... , Xn(t-d), Xn(t-d+1), ... ,Xn(t-1)]
output: Xk(t)

With virtual terms, the time series is like the following this.

X1(1), X1(1.5), ... ,X1(T)
X2(1), X2(1.5), ... ,X2(T)
....
Xn(1), Xn(1.5), ... ,Xn(T)

The training pattern from the above time series including virtual terms is made like the following this in separate model. This case is same to that in univariate time series presented in [2], [12], [13], and, [19].

input: [Xk(t-d), Xk(t-d+0.5), ... , Xk(t-1)]
output: Xk(t)

In combined model, the training pattern from the time series including virtual terms is like the following this.

input: [X1(t-d), X1(t-d+0.5), ... ,X1(t-1), ... , Xn(t-d), Xn(t-d+0.5), ... ,Xn(t-1)]
output: Xk(t)

4. Experiment & Results

In order to validate the insertion of this paper, conditions, procedures, and results of this experiment will be described in this section. Both kinds of VTG schemes are applied to multivariate time series prediction, and both separated model and combined model mentioned in the precious section are applied to neural-based approaches to multivariate time series prediction. The data of time series used in this experiment is artificial time series generated from the Lorenz equation, a dynamic system. The model of neural network used in this paper is back propagation, which is used most commonly in the models of neural network. In this experiment, both VTG schemes are compared with naïve neural based approach and An's scheme of generating artificial training patterns proposed in [13].

The time series data is generated from the following equation called Lorenz equation.

$$\Delta x_1(t) = \sigma(x_2(t) - x_1(t))$$
$$\Delta x_2(t) = \rho \cdot x_1(t) - x_2(t) - x_1(t)x_3(t)$$
$$\Delta x_3(t) = x_1(t)x_2(t) - bx_3(t)$$

Three variables are determined and 1000 terms in the time series are generated from the above equation. Training period is from 1 to 700 and test period is from 701 to 1000. In the above equations, σ , ρ , and b are given parameters. In this experiment, three groups of time series data are used by modifying these parameters

The neural model of this experiment is back propagation. The architecture of this model is presented like the table 1.

Table 1. The architecture of back propagation in this experiment

		Input Nodes	Hidden Nodes	Output Nodes
Separated	Naïve Approach An's Scheme	10	10	1
	Both VTG Schemes	19	10	1
Combined	Naïve Approach An's Scheme	30	10	1
	Both VTG Schemes	57	10	1

In the case of time series including virtual terms, d-1 virtual terms are included in the sliding window, if the size of sliding window is d. The total number of terms within the sliding window becomes 2d-1. The learning rate of this mode is set 0.1 and the initial weight is given at random. The training epochs is fixed to 5000 in any case. The measurement of the time series prediction is prediction error, MSE (Mean Square Error)

The results of the first group of time series data are presented like the table 2. The parameters of this data is given as $\sigma = 0.99, \rho = 2.25$, and $b = 1.775$

Table 2. The results of the first group:
$\sigma = 0.99, \rho = 2.25, b = 1.775$ (MSE $*10^{-4}$)

	1st Variable		2nd Variable		3rd Variable	
	분리	혼합	분리	혼합	분리	혼합
No VTG	4.967	1.418	4.713	1.778	2.125	1.521
An (0,0.05)	3.332	1.416	3.380	1.456	2.912	1.213
An (0,0.1)	3.313	1.319	3.295	1.434	2.918	1.092
Mean	1.085	0.9360	1.263	0.7791	1.082	0.6758
2nd Lag	1.049	0.6681	0.9181	0.9123	1.049	0.6543
1st Tayor	0.2727	0.08016	0.3318	0.4968	0.6092	0.4525
Uniform	0.9594	0.07418	1.114	0.1325	0.9645	0.8247
Gaussian	1.252	0.2341	1.198	2.919	1.055	0.9826
Triangle	0.9813	0.2312	1.462	1.578	0.9252	0.7420

The results of the second group are presented in the table 3. Its parameter of the above equation is given as $\sigma = 0.94, \rho = 2.20$, and $b = 1.650$

Table 3. The results of the second group:
$\sigma = 0.94, \rho = 2.20, b = 1.650$ (MSE $* 10^{-4}$)

	1st Variable		2nd Variable		3rd Variable	
	분리	혼합	분리	혼합	분리	혼합
No VTG	1.842	0.9126	1.890	0.9794	1.387	2.603
An (0,0.05)	1.470	0.8828	1.459	0.9426	1.023	1.425
An (0,0.1)	1.458	0.8509	1.464	0.9321	1.012	1.490
Mean	1.074	0.6785	1.058	0.7426	0.5327	0.4990
2nd Lag	0.6791	0.4998	0.6320	0.6252	0.4644	0.4472
1st Tayor	0.7668	0.2371	0.4816	0.2746	0.4709	0.2504
Uniform	0.9206	0.5911	1.354	0.6226	0.4757	0.4980
Gaussian	0.8193	0.6206	1.383	0.6644	0.6562	0.5992
Triangle	0.9417	0.6570	0.7412	0.6804	0.6035	0.5842

The results of the third group are presented in the table 4. Its parameter of Lorenz equation is given as $\sigma = 0.90, \rho = 2.30$, and $b = 1.0$

Table 4. The results of the third group:
$\sigma = 0.90, \rho = 2.30, b = 1.0$ (MSE $* 10^{-4}$)

	1st Variable		2nd Variable		3rd Variable	
	분리	혼합	분리	혼합	분리	혼합
No VTG	5.164	1.796	5.245	3.574	2.961	1.711
An (0,0.05)	3.739	1.796	3.845	2.834	2.028	1.469
An (0,0.1)	3.722	1.697	3.853	2.771	2.128	1.524
Mean	1.514	1.267	1.659	1.687	1.406	1.043
2nd Lag	1.436	0.8027	1.269	1.754	1.394	0.8479
1st Tayor	0.2587	0.1273	0.2016	0.2133	1.072	0.6377
Uniform	1.340	0.8361	1.348	1.579	1.434	0.9399
Gaussian	1.660	1.288	2.353	1.924	1.578	1.222
Triangle	1.883	1.294	2.089	1.821	1.550	1.066

From table2 to table 4, An's scheme improved the performance of time series prediction about 10-20% compared with naïve approach. Deterministic VTG schemes improved its performance even more than 50%; the prediction error is reduced less than half of the naïve approach and An's scheme. Stochastic VTG schemes improved its scheme compared with naïve neural-based approach and An's scheme outstandingly but are little inferior to the deterministic schemes except mean method.

5. Conclusion

This paper proposed the alternative VTG schemes and showed that these schemes reduced prediction error compared with naïve neural based approach and An's scheme and they are comparable with deterministic VTG schemes proposed [17] and [18]. The proposed VTG schemes have advantages over deterministic VTG schemes; stochastic VTG schemes have the diversity in the estimation of virtual terms. This provides the potentiality of the application of evolutionary computation to optimize virtual terms. The optimization of virtual terms means the maximization of prediction performance. Although the stochastic VTG schemes are little inferior to the deterministic VTG schemes in general, the optimization of virtual terms with evolutionary computation will improve the prediction performance compared with the deterministic VTG schemes.

References

[1] G. E. P. Box, G.M. Jekins, and G.C. Reisel, Time Series prediction, Prentice Hall, 1994.

[2] S. Haykine, Neural Networks, Macmillan Colledge Publishing Company, 1994.

[3] A.S. Weigend and D.E. Rumelhart, Generalization through Minimal Networks with Application to Forecasting, Interface 91, 1991, pp362-370.

[4] N. Kohzadi, M. S. Boyd, B. Kermanshahi, and I. Kaastra, A comparison of artificial neural network and time series models for forecasting commodity process, Neurocomputing No 10, 1996, pp169-181.

[5] D. Brownstone, Using percentage accuracy to measure neural network prediction on Stock Market movements, Neurocomputing No 10, 1996, pp237-250,

[6] M. Malliaris and L. Salchenberger, Using neural networks to forecast the S&P 100 implied volatility, Neurocomputing No 10, 1996, pp183-195,

[7] A.U Levin, Stock Selection via Nonlinear Multi-Layer-Factor Models, Neural Information Processing Systems Vol 8, 1996, pp966-972,

[8] J. Ghosn and Y. Bengio, Multi-Task Learning for Stock Selection, Neural Information Processing Systems Vol 9, 1997, pp947-952.

[9] Y. S. Abu-Mustafa, A Method for Learning from Hints, Advances in Neural Information Processing System Vol 5, 1993, pp73-80,

[10] Y. S. Abu-Mustafa, Financial Application of Learning from Hints, Neural Information Processing Systems Vol 7, 1995, pp411-418.

[11] D.A. Cohn, Z. Ghahramami, and M. J. Jordan, Active Learning with Statistical Models, Neural Information Processing Systems Vol 6, 1994, pp705-712.

[12] Krogh and J. Vedelsby, Neural Network Ensembles, Cross Validation, and Active Learning, Neural Information Processing Systems Vol 7, 1995, pp231-238

[13] G. An, The Effects of Adding Noise During Backpropagation Training on a Generalization Performance, Neural Computation Vol 7, 1996, pp643-674

[14] S. Cho, M. Jang, and S. Chang, Virtual Sample Generation using a Population of Network, Neural Processing Letters Vol 5 No 2, 1996, pp83-89.

[15] D. Saad and S.A. Solla, Learning with Noise and Regularizers in Multilayer Neural Networks, Neural Information Processing Systems Vol 9, 1997, pp260-266

[16] Y. Grandvalet, S. Canu, and S. Boucheron, Noise Injection: Theoretical Prospects, Neural Computation Vol 9, 1997, pp1093-1108.

[17] T.C. Jo and S. Cho, Time Series Prediction using Virtual Term Generation Scheme, The Proceedings of ITC-CSCC 96, 1996, pp1282-1285.

[18] T.C. Jo, The Prediction of the Annual Number of Sunspots with Virtual Terms Generation Schemes, The Proceeding of PACES/SPICIS 97, 1997, pp135-141.

[19] T.C. Jo, The Connectionist Approach to Multi Variables Forecasting of Precipitation with Virtual Term Generation Schemes, The Proceedings of IEEE IJCNN 98, 1998, pp2531-2534.

[20] T.C. Jo, Neural Approach to Forecasting of S&P 500 Stock Price Index with Virtual Term Generation, The Proceedings of International Conference on Artificial Intelligence 99, 1999, pp502-507.