DLS-ICBP Neural Networks with Applications in Time Series Prediction*

Songcan Chen Qun Dai
(Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, P.R.China, 210016)
Corresponding author: s.chen@nuaa.edu.cn daiqun@nuaa.edu.cn
Tel: +86-25-8489-2805, Fax: +86-25-84498069

Abstract: As a generalization to multi-layer perceptron (MLP), circular back-propagation neural network (CBP) possesses better adaptability. On CBP basis, an improved CBP (ICBP) in this paper is presented. Although having less adjustable weights, ICBP has better adaptability than CBP, which tallies quite the famous Occam's razor principle for model selection. In its application to time series, considering both the structural changes and correlations of time series itself, we introduce the principle of the discounted least squares (DLS) to CBP and ICBP, respectively, and investigate their predicting capability further. Introduction of DLS improves the predicting performance of both on benchmark time series dataset. Finally, the comparison of experiment results shows that ICBP with DLS (DLS-ICBP) has better predicting performance than DLS-CBP.

Keywords: Discounted Least Squares (DLS); (Improved) Circular Back-propagation Network; Time Series Predicting or Forecasting, Occam's Razor Principle

I. Introduction

EBP (Error Back-propagation) is probably the most popular learning algorithm in the study of artificial forward neural networks by which multiple-layer perceptron (MLP)[1] has most widely received attention in both theory and applications due to its excellent properties like universal approximating ability to arbitrary continuous functions. On its basis, Sandro Ridella *et al* proposed a circular BP neural network (CBP) [2,3,4] through adding or augmenting an extra node to the original BP input layer and taking the sum of all squared components of an input vector presented to the network as a incoming signal of the added node. The authors proven that CBP possesses favorable capabilities in generalization and adaptability compared to the MLP model [2,3,4]. Under the CBP framework, both the vector quantization (VQ)[4,5] and the radial basis function (RBF) networks[3] can respectively be constructed , and hence CBP shows great flexibility. However, there also exist several deficiencies in it: 1) The incoming signal of the added node only is an isotropic, i.e., an equally-weighted sum of all squared component values, thus it lacks anisotropy among different components for an input vector; 2) Due to such an isotropy, it cannot simulate the famous Bayesian classifier in a more direct way; 3) It requires probably more hidden nodes to approximate any continuous function to arbitrary precision. As a result, redundant parameters may lead to over-fitting, which will lower the generalization capability[2].

The goal of this paper aims at obtaining a general improved network model for CBP, for short, ICBP. Actually, ICBP is similar in structure to CBP, but there are two major changes made: a) the

* Supported by Natural Science (grant No.BK2002092) and "QingLan" project foundations of Jiangsu province and Returnee foundation of China.

incoming signal of the added node is not an isotropic sum of all squared input components as in CBP instead of their anisotropic sum; b) more importantly, the weight values between the added node and all hidden nodes are set to a special common value (in our case, all 1 or all -1)[6, 7] rather than usually adjustable parameters as in CBP so that the total number of the adjustable weights in ICBP is probably reduced. Our motivation of the alterations to CBP is to control the network complexity and to avoid possible over-fitting in this way while still make the ICBP structure as simple as possible. The newly-constructed model has following characteristics: Firstly, besides inheriting those CBP characteristic of constructive equivalence to VQ and RBF [5, 6], ICBP can also model the famous Bayesian classifier in a direct constructive way [10]. Secondly, although having less adaptable weights than CBP, ICBP has better generalization and adaptation [6]. Thirdly, it can still adopt the BP learning algorithm to perform training with the learning complexity equal to that of CBP. Naturally, various existing improved algorithms to BP can also be applied to upgrade performances of CBP and ICBP. In addition, due to assigning special constant values either +1s or −1s to all the weights connecting the added node and all the hidden nodes to respectively form ICBP+1 or ICBP-1 networks, consequently ICBPs have less adjustable weights but better generalization and adaptability than CBP [6,7]. This indeed demonstrates rationality of the famous Occam's razor principle, i.e., network with simple structure but just good training performance is generally better generalization than the one with slightly better training performance but more complex structure.

Time series prediction is one of the active applied areas for neural networks. As a proven effective nonlinear predicting tool, neural networks are successful in industry and financial areas. However, during predicting process, merely simply using the original ICBP to directly predict time series will result in, to some degree, neglect for the inherent structural changes and time correlation in time series itself. Intuitively, a predicting point has stronger correlation to observations closer to it and weaker one to those far away from it. Therefore in training process samples in time window impose different influences on network weights: the nearer is the observation from the predicting point, the greater is the influence. Moreover the idea of discounted least square formulates and reflects exactly this influence [8]. To make ICBP embody the above characteristic, we bring forward a DLS-ICBP, based upon DLS and oriented to time series prediction, introducing DLS to its cost function. DLS cost function biases learning towards most recent observations in a time series but without ignoring long term effects. The experiments of Benchmark chaotic time series and certain city's water consumed quantity time series prediction indicate that DLS improves ICBP performance. Simulation results also show that DLS-ICBP performance is much better than DLS-CBP, which prove the superiority of ICBP again.

II. ICBP Network

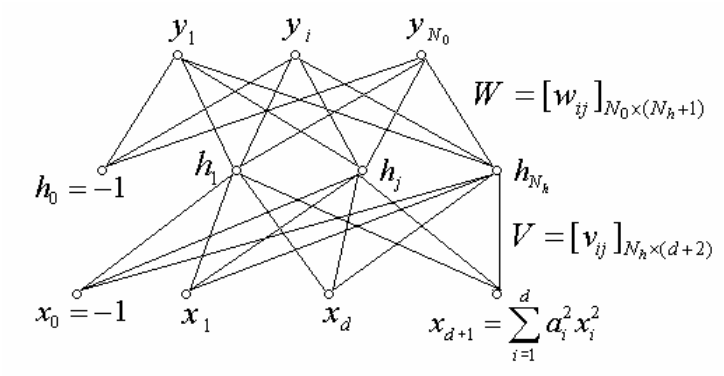

Fig. 1 ICBP three-layer network model

Fig.1 shows a three-layer ICBP network with N_O output nodes, N_h hidden nodes, *d* input nodes with respect to *d* dimensional input pattern or vector and an extra input node with an input $x_{d+1}=\sum_{i=1}^{d} a_i^2 x_i^2$. In particular, when all a_i are taken equally, ICBP reduces to the CBP. And at the same time, ICBP weights, $v_{j(d+1)}(j=1\ldots N_h)$, connecting the extra or added node to the *j*th node of the hidden layer differ from CBP corresponding ones: in ICBP, $v_{j(d+1)}(j=1\ldots N_h)$ take a common constant directly, while the counterparts in CBP are adaptable parameters. Consequently, the difference of the number of adaptable parameters for these two models is $|N_h-d|$. In general, the number of hidden nodes in function approximation is larger than that of input nodes due to such a fact that the forward multi-layer networks with sufficient hidden node number can approximate any continuous function to arbitrary precision [1]. Therefore, the adjustable parameters of ICBP are often less than those of CBP. Now let the expected outputs of the network be $o_i(i=1.....N_O)$, and thus a corresponding sum-of-squares error function *E* is defined as:

$$E=\frac{1}{2}\sum_i (o_i-y_i)^2 \qquad (1)$$

where y_i is actual output of the *i*th output node $(i=1.....N_O)$. Adopting the known-as error back-propagation learning algorithm (in fact, any other improved algorithms can be applied) and after algebra manipulation, the weight (w_{ij}) adjustments between output and hidden layer are easily derived as follows:

$$\Delta w_{ij}(t)=-\eta\frac{\partial E_p}{\partial w_{ij}}=\eta[o_i(t)-y_i(t)]\cdot h_j(t),\quad i=1.....N_o, j=0.....N_h \qquad (2)$$

The weight (v_{jk}) adjustments between hidden layer and input layer are:

$$\Delta v_{jk}(t)=\eta[\sum_{l=1}^{N_o}(o_l(t)-y_l(t))w_{lj}]\cdot h_j(t)\cdot(1-h_j(t))x_k(t) \qquad j=1.....N_h, k=0.....d \qquad (3)$$

Adjusting formula for $a_k, (k=1\ldots d)$ is:

$$\Delta a_k = 2\eta \sum_{j=1}^{N_h} \left\{ \sum_{i=1}^{N_o} \left[(o_i - y_i) w_{ij} \right] \cdot h_j (1 - h_j) v_{j,d+1} \right\} \cdot a_k x_k^2, \quad k = 1.....d \tag{4}$$

III. DLS-ICBP Neural Network

A. DLS error back-propagation algorithm

Learning by the error backpropagation is an error minimization procedure which uses the gradient descent to weight error space to minimize a quadratic measure of total error. The most commonly used error measure is the Ordinary Least-Square criterion (OLS) as shown in (5).

$$E_{LS} = \frac{1}{2N} \sum_{p=1}^{N} (\vec{o}_p - \vec{y}_p)^2 \tag{5}$$

where N is the total number of observations in the sample, $\vec{o}_p$ is the desired response and $\vec{y}_p$ is the observed response, with p=N being the most recent observation. LS measures give equal weight to all observations in the sample (training set). In time series analysis with structural changes, it is often desirable to overweight more recent observations for the reasons discussed above. The idea of DLS arises from here. The cumulative error calculated by the DLS procedure is given by

$$E_{DLS} = \frac{1}{2N} \sum_{p=1}^{N} w(p) (\vec{o}_p - \vec{y}_p)^2 \tag{6}$$

N, $\vec{o}_p$, $\vec{y}_p$ are defined as above, and $w(p)$ is an adjustment of the contribution of observation p to the overall error [8] or discounted rate. In general, there are many different ways of biasing the cost function through $w(p)$ (such as linear, exponential, etc.) to differentially weight the contribution of each observation towards the total error. In the paper, we examine a simple sigmoidal decay as follows

$$w(p) = \frac{1}{1 + e^{(a - bp)}} \qquad \text{where } b=2a/N \tag{7}$$

The parameters a and b are used to scale and offset the sigmoid. The DLS cost function is asymptotically invariant with respect to the sample size (N). Since b in (7) is derived from a and N, the only control parameter is the discount rate a. The learning rule is derived in the usual way by repeatedly changing the weights by an amount proportional to

$$\frac{\partial E_{DLS}}{\partial W} = \frac{1}{1 + e^{(a - bp)}} \frac{\partial E_{LS}}{\partial W} \tag{8}$$

B. DLS-ICBP neural network

In this paper DLS cost function is incorporated into ICBP network. A group of biased weights modifying formulae are derived out accordingly from (8), (2), (3) and (4). Again after complicated algebra manipulation, the weight (w_{ij}) adjustments between the output and hidden layer for DLS-ICBP are respectively modified to:

$$\Delta w_{ij}(t) = -\eta \frac{\partial E_p}{\partial w_{ij}} = \frac{\eta}{1 + e^{(a - bp)}} [o_i(t) - y_i(t)] \cdot h_j(t), \quad i = 1.....N_o, j = 0.....N_h \tag{9}$$

Weights (v_{jk}) adjustments between hidden layer and input layer are:

$$\Delta v_{jk}(t) = \frac{\eta}{1+e^{(a-bp)}}[\sum_{l=1}^{N_o}(o_l(t)-y_l(t))w_{lj}]\cdot h_j(t)\cdot(1-h_j(t))x_k(t)$$
$$j = 1.....N_h, k = 0.....d \qquad (10)$$

Adjusting formula for $a_k, (k = 1...d)$ is:

$$\Delta a_k = \frac{2\eta}{1+e^{(a-bp)}}\sum_{j=1}^{N_h}\left\{\sum_{i=1}^{N_o}[(o_i-y_i)w_{ij}]\cdot h_j(1-h_j)v_{j,d+1}\right\}\cdot a_k x_k^2, \quad k = 1.....d \qquad (11)$$

In the DLS-ICBP, learning is biased toward more recent observations with long term effects experiencing exponential decay through time. This is particularly important in systems in which the structural relationship between input and response vectors changes gradually over time but certain elements of long-term memory are still retained. Experiments results show that DLS-ICBP achieves better predictive effects than ICBP in both single step and multiple steps prediction.

IV. Experiment Results

A. Chaotic time series prediction [11]

Time series produced by iterating the logistic map

$f(x) = \alpha x(1-x)$ (continuous) or $x(n+1) = \alpha x(n)(1-x(n))$ (discrete)

is probably the simplest system capable of displaying deterministic chaos. This first-order difference equation, also known as the Feigenbaum equation, has been extensively studied as a model of biological populations with non-overlapping generations, where $x(n)$ represents the normalized population of n-th generation and α is a parameter that determines the dynamics of the population. The behavior of the time series depends critically on the value of the bifurcation parameter α. When α reaches a value of about 3.56, the output becomes chaotic. In this contrastive experiment, we set $\alpha = 3.56$ and produce 100 elements sequence orderly. First, we take the data pairs of $(x_t, x_{t+1}), (1 \le t < 50)$ as training set, and then we take 50 data points equally spaced in [0.5, 0.99] as the test data to do 50 times of single step predictions. The experimental results are averaged from 50 times of experiments. Fig.2 shows the contrastive experimental results of CBP, DLS-CBP, ICBP-1 and DLS-ICBP-1 respectively when N_h=7. MVAR, a 50 time average of the sum of squared difference between the 50 predicting results and targets, is chosen as the performance measure index as shown in Table.1.

$$MVAR = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{P}\sum_{k=1}^{N_o}[o_i(t)-y_i(t)]^2 \qquad (12)$$

where N represents the number of repeated experiments; P represents P predicting points; N_o is the number of output nodes. Here N=50, P=50, N_o =1. From the obtained MVARs in the experiments, DLS-ICBP-1 has less MVAR value than both ICBP-1 and DLS-CBP, meaning that it has is better predicting performance than ICBP-1. In order to further verify effectiveness of the

proposed model, we also adopt the Normalized Mean Square Error (NMSE) for the test set (TS) as another comparison index as defined in (13):

$$NMSE = \frac{\sum_{t \in TS} (y_t - \hat{y}_t)^2}{\sum_{t \in TS} (y_t - \bar{y})^2} \tag{13}$$

where y_t denotes the target value of predicting point t, $\hat{y}_t$ represents the predictive value of point t, and $\bar{y}$ denotes the actual mean value of the whole set. From Fig. 2, it seems that the predicting curves of all methods are, visually, all similar. However respectively from the MVARs and NMSEs in both Table 1 and Table 2, we can observe that though the DLS-CBP yields the worst predicting performance in all compared models, even inferior to these models without using DLS technique, DLS-ICBP-1 predictive quality is *just* slightly lower than the best CBP in the MVAR but conversely slightly higher than CBP in the NMSE and seats No. 1 of predicting performance. The interpretation of why these models yield such diverse results seems to attribute to such a fact that the chaotic time series possess their own regularities. Chaos do not mean orderless[11] and can be generated from simple definite system. In chaos there exist attractors which possess attractability and stability to small disturbance. This experiment indicates that BP and CBP networks both indeed have strong approximating capability to chaotic system, while DLS method slightly weakens the regularity in chaotic system and probably impairs the network predicting performance, instead.

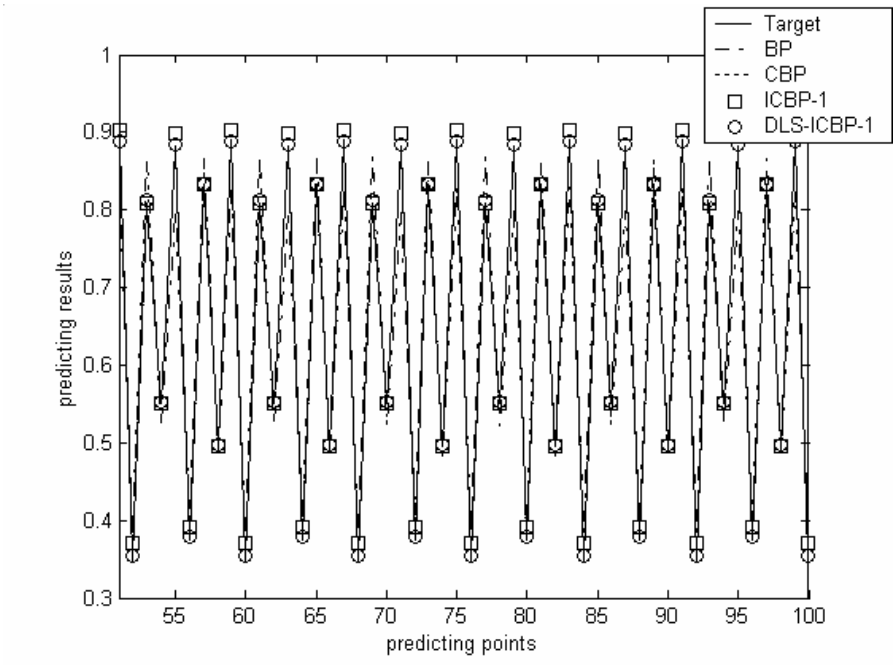

Fig.2 Chaotic time series single step prediction results on [0.5, 0.99], compared with the targets, of the respective network model

B. Applications to city daily life water consumed quantity

Predicting city daily life water consumed quantity can help to lay a productive course,

economize energy sources and boost production benefits. It is practically valuable for civil life and manufacture. In terms of the historical data provided, we predict the water consumed quantity one month to one quarter ahead for the use of water supply department. According to the practical condition of this experiment, we choose the dynamic prediction way. Generally, the dynamic predicting model can be described as follows:

$$y_{i,t} = F[W_{i,t}, y_{i,t-1}, \ldots\ldots, y_{i,t-m}, y_{i-1,t}, \ldots\ldots, y_{i-1,t-m}, \ldots\ldots\ldots\ldots, y_{i-n,t}, \ldots\ldots y_{i-n,t-m}],$$

That is to say, the information of the preceding m months, the month of the preceding n years and the preceding m months of the preceding n years are taken into the network inputs in prediction. Not supplied with enough data, we take n as zero here. We use the four network models of CBP, DLS-CBP, ICBP-1, DLS-ICBP-1, respectively to process single step prediction for the recent 12 months. The network inputs include the year, the month, the planned consuming quantity of the month and the actual consumed quantity of the preceding m months. Because the values of the monthly planned consuming quantity and the monthly actual consumed quantity are huge, they are mapped into the district of (0, 1). Let M be the set of the monthly planned consuming quantity or the monthly actual consumed quantity; let IN_i be the i-th monthly planned consuming quantity or the *i*-th monthly actual consumed quantity and in_i be the i-th actual input to the network; then

$$in_i = \frac{IN_i - \min(M)}{\max(M) - \min(M)} 0.8 + 0.1$$

Fig. 3 shows the predicting results of corresponding network models in contrast to the actual consumed quantity. As seen from Fig.3 and Table.1 and Table 2, the models with DLS method improve largely the performance of the ones without it. Moreover, the performances of both ICBP-1 and DLS-ICBP-1 are better than CBP and DLS-CBP, respectively.

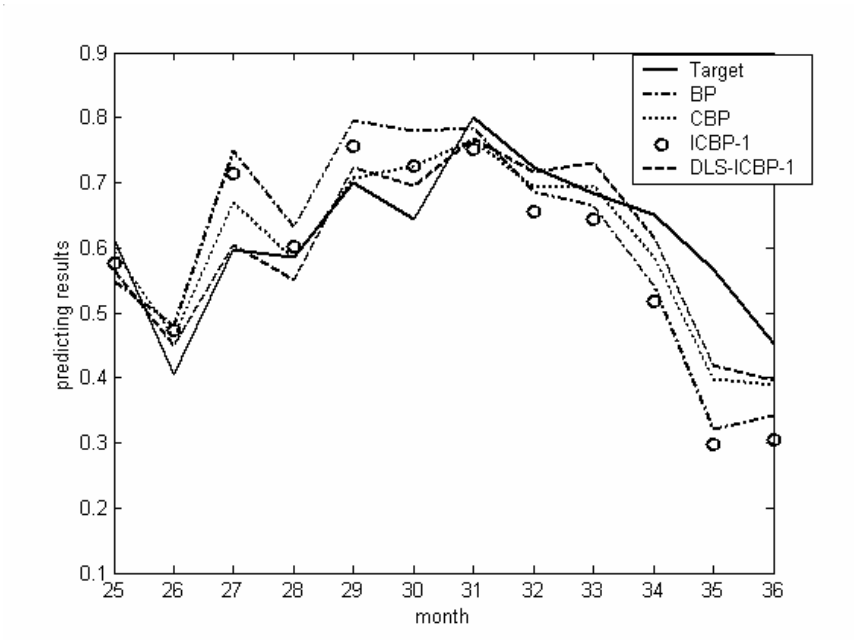

Fig.3 Twelve months water consumed quantity prediction results, compared with the actual

consumed quantity, of the respective network model

C. Non-stationary variance time series prediction [8]

To compare the performance characteristics of CBP and DLS-CBP, ICBP and DLS-ICBP, we carry out a controlled predictive experiment using a simple sinusoid function. In the experiment the variance of the time series is changing through time. The non-stationary variance time series is built from:

$$y = a(n+x)*\sin(2\pi x)+b,\ \ x\in[0,1],\ \ n\in\{0,\ldots\ldots,7\}$$

With the parameters set to a=3, b=[2a]/N. A serial of 800 elements is generated from the above formula. With the time window set to four, we take the data pairs of $\{(y_t, y_{t+1}, y_{t+2}, y_{t+3}), y_{t+4}\}, (1 \le t \le 600)$ as the training data, and then take the data of $(y_t, y_{t+1}, y_{t+2}, y_{t+3}), (604 \le t \le 703)$ as the test data. The experimental results are shown in Fig. 4, Fig. 5 and Table.1, respectively. From them, we can observe that the predicting performances of both DLS-CBP and DLS-ICBP-1 are better than CBP and ICBP-1 respectively. Moreover, DLS-ICBP-1 performs much better than DLS-CBP, which also illustrates the improvement of ICBP to CBP.

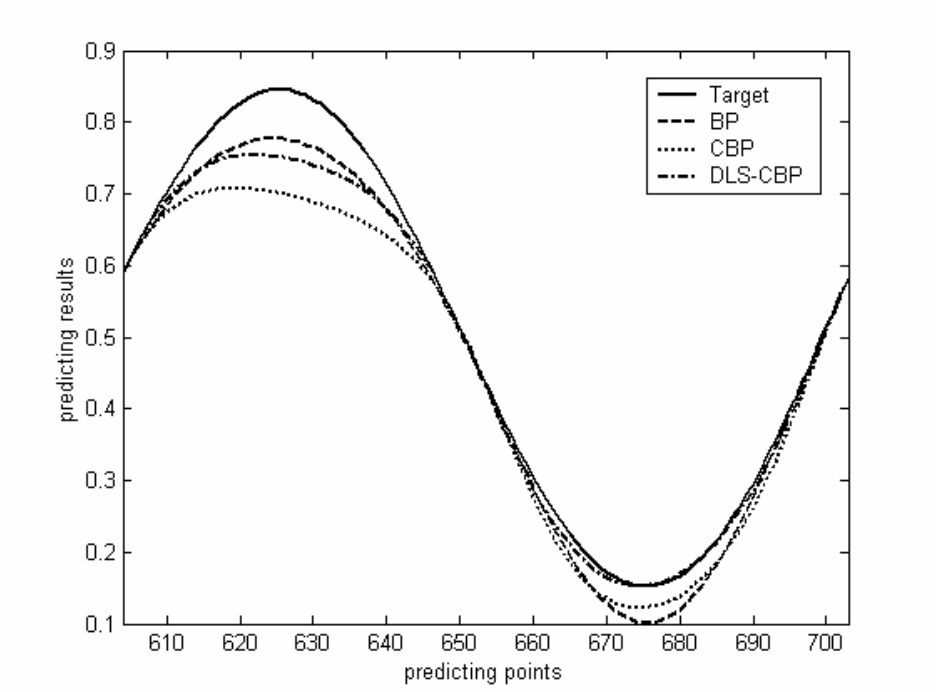

Fig.4 Non-stationary covariance time series prediction results on [604,703], compared with the targes, of BP, CBP and DLS-CBP

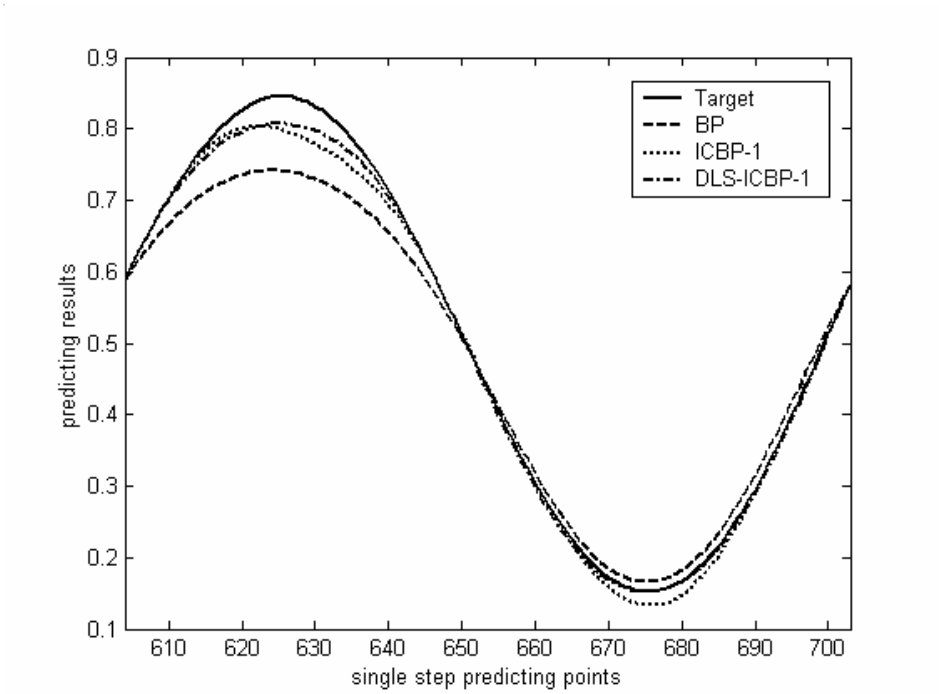

Fig. 5 Non-stationary covariance time series prediction results on [604,703], compared with the targes, of BP, ICBP-1 and DLS-ICBP-1

Table.1 The comparative MVAR measurements of the experiments of respective models

Time series	BP	CBP	DLS-CBP	ICBP-1	DLS-ICBP-1
Chaotic	0.1140	3.2201e-004	0.1487	0.0075	6.6657e-004
City monthly water consumed quantity	0.1492	0.0969	0.0752	0.0575	0.0382
Non-stationary covariance time series	0.2434	0.2180	0.1972	0.1643	0.0328

Table.2 The NMSE measurements of the comparative simulations of the respective networks

Time series	BP	CBP	DLS-CBP	ICBP-1	DLS-ICBP-1
Chaotic	0.0480	2.8098e-004	0.0627	0.0032	1.3574e-004
City monthly water consumed quantity	1.1041	0.4048	0.5569	0.4415	0.2828
Non-stationary covariance time series	0.0028	0.0047	0.0018	0.0012	5.5172e-004

V. Conclusion

It has experienced a fairly long history since the investigation of time series prediction. The traditional statistical methods can hardly solve non-linear time series predicting problems. Predicting using neural networks is an effective way. BP, CBP and ICBP, etc, possess good performance in experiments. However, all of them neglect inherent structural changes and time correlation in time series itself. Intuitively, predicting point has stronger correlation to

observations closer to it and weaker one to those far away from it. Therefore in training process samples in time window impose different influences on network weights: the nearer is the observation from the predicting point, the greater is the influence. Moreover the idea of discounted least square formulates exactly this influence. The forecasting experiments on benchmark non-stationary time series and the data sets of daily life water consumed quantity have proved that the method of DLS boosts the predictive performance of ICBP and CBP by 60% and 20% respectively. And at the same time, the comparisons of experimental results show clearly that DLS-ICBP predicts better than DLS-CBP, which proves again that the generalization of ICBP to CBP is valuable. The principle of DLS fits the prediction in economical area especially. In that area, the structural changes of time series take place slowly, along with the time going and the economical environment changing. When the inherent economic laws are reacting on the economical data, short-term trends can have the opposite effects on them. The selection of the parameters a and b of $w(p)$ in DLS is very important. Experiments show that as long as we select these parameters conservatively, the two types of problems of whether the recent training samples are over or insufficiently stressed are not crucial [8].

Acknowledgement: We are very grateful for the reviewers' constructive comments which improve the presentation of this paper.

References

[1] Simon Haykin, (1999) Neural Networks: A Comprehensive Foundation, Second Edition, Prentice-Hall, Inc..

[2] S. Ridolla, S. Rovetta and R. Zunino (1997) Circular back-propagation networks for classification. IEEE Trans. Neural Networks, 8(1):84-97..

[3] S. Ridolla, S. Rovetta and R. Zunino (1997) CBP networks as a generalized neural model, In: Proceedings of Int. Conf. On Neural Networks, Houston, USA, June, pp.210-214.

[4] S. Ridolla, S. Rovetta and R. Zunino (1999) Circular Backpropagation Networks Embed Vector Quantization, IEEE Trans. Neural Networks, 10(4): 972-975.

[5] Zhang Benzhu, Chen Songcan (2001) Equivalence between vector quantization and ICBP networks, J. of Data Acquisition and Processing (in Chinese), 16(3):291-294.

[6] Zhang Benzhu (2001) The research on the performance and applications of improved BP neural networks, Thesis of Master degree (in Chinese), Nanjing University of Aeronautics and Astronautics.

[7] Dai Qun, Chen Songcan, Zhang Benzhu (2003) Improved CBP neural networks with its applications in time series prediction, Neural Proc. Lett., 18(3):217-231.

[8] A.-Paul Refenes, Y. Bentz and D. W. Bunn (1997). Financial time series modelling with discounted least squares backpropagation. Neurocomputing 14: 123-138.

[9] Vapnik,V. N. (1995) The nature of statistical learning theory, Springer, Berlin.

[10] Zhang Benzhu, Chen Songcan (2001) The equivalence between ICBP and the Bayesian classifier, Technical Report, Nanjing University of Aeronautics & Astronautics..

[11] Chen C.L.Philip and Wan John Z. (1999) A Rapid Learning and Dynamic Stepwise Updating Algorithm for Flat Neural Networks and the Application to Time-Series Prediction. IEEE Trans. Syst., Man and Cybern.-part B, 29(1):62-72.