

# Wavelet Packet Multi-layer Perceptron for Chaotic Time Series Prediction: Effects of Weight Initialization

Kok Keong Teo, Lipo Wang,\* and Zhiping Lin

School of Electrical and Electronic Engineering  
Nanyang Technological University  
Block S2, Nanyang Avenue  
Singapore 639798

<http://www.ntu.edu.sg/home/elpwang>  
{p7309881e, elpwang, ezplin}@ntu.edu.sg

\*To whom all correspondence should be addressed

**Abstract.** We train the wavelet packet multi-layer perceptron neural network (WP-MLP) by backpropagation for time series prediction. Weights in the backpropagation algorithm are usually initialized with small random values. If the random initial weights happen to be far from a good solution or they are near a poor local optimum, training may take a long time or get trap in the local optimum. Proper weights initialization will place the weights close to a good solution with reduced training time and increased the possibility of reaching a good solution. In this paper, we investigate the effect of weight initialization on WP-MLP using two clustering algorithms. We test the initialization methods on WP-MLP with the sunspots and Mackey-Glass benchmark time series. We show that with proper weight initialization, better prediction performance can be attained.

## 1 Introduction

Neural networks have demonstrated great potential for time-series prediction where system dynamics is nonlinear. Lapedes and Farber [1] first proposed to use a multi-layer perceptron neural network (MLP) for nonlinear time series prediction. Neural networks are developed to emulate the human brain that is powerful, flexible and efficient. However, conventional neural networks process signals only on their finest resolutions. The introduction of wavelet decomposition [2]-[6] provides a new tool for approximation. It produces a good local representation of the signal in both the time and the frequency domains. Inspired by both the MLP and wavelet decomposition, Zhang and Benveniste [7] proposed wavelet network. This has led to rapid development of neural network models integrated with wavelets. Most researchers used wavelets as basis functions that allow for hierarchical, multiresolution learning of input-output maps from data. The wavelet packet multi-layer perceptron (WP-MLP) neural network is an MLP with the wavelet packet as a feature extraction method to obtain time-frequency information. The WP-MLP had a lot of success in classification applications in acoustic, biomedical, image and speech.

Kolen and Pollack [8] shown that feedforward network with the backpropagation technique is very sensitive to the initial weight selection. Prototype pattern [18] and the orthogonal least square algorithm [19] can be used to initialize the weights. Geva [12] proposed to initialize the weights by clustering algorithm that is based on mean local density (MLD).

In this paper, we apply the WP-MLP neural network for temporal sequence prediction and study the effects of weight initialization on the performance of the WP-MLP. The paper is organized as follows. In Section 2, we review some background on WP-MLP. In Section 3, we describe the initialization methods. In Section 4, we present the simulation results of time series prediction for two benchmark cases, i.e. the sunspots and Mackey-Glass time series. Finally, Section 5 presents the conclusion.

## 2 Wavelet Packet MLP

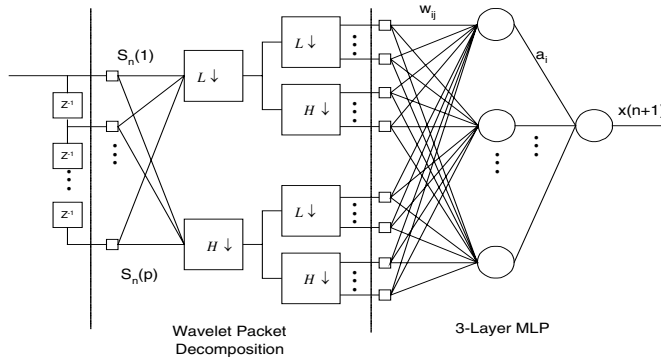


Fig. 1. Model of WP-MLP

Fig. 1 shows the WP-MLP used in this paper. It consists of three stages, i.e., input with a tapped delay line, wavelet packet decomposition, and an MLP. The output  $\hat{x}(n + 1)$  is the value of the time series at time  $n + 1$  and is a function of the values of the time series at previous time steps:

$$s_n = [x(n), x(n - 1), \dots, x(n - p)]^T \tag{1}$$

Coifman, Meyer and Wickerhauser [9] introduced wavelet packets by relating multiresolution approximations with wavelets. Recursive splitting of vector spaces is represented by a binary tree. We employed the wavelet packet transformation to produce time-frequency atoms. These atoms provide us with both time and frequency information with varying resolution through out the time-frequency plane. The time-frequency atoms can be expanded in a tree-like structure to create arbitrary tiling, which is useful for signals with complex structure. A tree algorithm [10] can be employed for computing wavelet packet transform by using the wavelet coefficients as filter coefficients. The next sample is predicted using a conventional MLP with  $k$  input units, one hidden layer with  $m$  sigmoid neurons and one linear output neuron, as

shown in Fig. 1. The architecture of the WP-MLP is defined as,  $[p : l(wlet) : h]$  where  $p$  is the number of tapped delays,  $l$  is the number of decomposition level,  $wlet$  is the type of wavelet packet used, and  $h$  is the number of hidden neurons. In the present work, the MLP is trained by the backpropagation algorithm using the Levenberg-Marquadt method [11].

### 3 Weight Initialization by Clustering Algorithms

The initialization of the weights and biases has a great impact on the network training time and generalization performance. Usually, the weights and biases are initialized with small random values. If the random initial weights happen to be far from a good solution or they are near a poor local optimum, training may take a long time or get trap in the local optimum [8]. Proper weights initialization will place the weights close to a good solution, which reduces training time and increases the possibility of reaching a good solution. In this section, we describe the method of weights initialization by clustering algorithms.

The time-frequency event matrix is defined by combining the time-frequency patterns and its respective targeted outputs of the training data. Each column is made up of time-frequency pattern and its respective target that is used for the clustering algorithm. The clustering analysis [13] is the organization of a collection of patterns into clusters based on similarity. In this way, the natural grouping of the time-frequency events is revealed. Member event within a valid cluster are more similar to each other than they are to an event belonging to a different cluster. The number of clusters is then chosen to be the number of neurons in the hidden layer of the WP-MLP. The number of clusters may be chosen before clustering, or may be determined by the clustering algorithm. In the present work, the hierarchical clustering algorithm [14] and the counterpropagation network [15] are tested. Hierarchical clustering algorithm consist of three simple steps:

- Compute the Euclidean distance between every pair of events in the data set.
- Group the events into a binary, hierarchical cluster tree by linking together pairs of events that are in close proximity. As the events are paired into binary clusters, the newly formed clusters are grouped into larger cluster until a hierarchical tree is formed.
- Cut the hierarchical tree to form the clusters according to the number of hidden neurons chosen based on a prior knowledge acquired.

Counterpropagation network is a combination of a competitive layer and another layer employing Grossberg learning. For the purpose for this paper, the forward-only variant of the counterpropagation network shall be used. Let us consider a competitive learning network consisting of a layer of  $N_n$  competitive neurons and a layer of  $N+1$  input nodes,  $N$  being the dimension of the input pattern.

$$h_k = \sum_j^{N+1} g_{kj} x_j \quad (2)$$

where  $g_{kj}$  is the weight connecting  $k$  neuron to all the inputs,  $h_k$  is the total input into neuron  $k$  and  $x_j$  is  $j$  element of the input pattern. If neuron  $k$  has the largest total input, it will win the competition and becoming the sole neuron to response to the input pattern. The input patterns that win the competition in the same neuron are considered to be in the same cluster. The neuron that win the competition will have it weight adjust according with the equation shows below.

$$w_{kj}^{new} = (1 - \alpha(t))w_{kj}^{old} + \alpha(t)x_j \quad (3)$$

where learning constant  $\alpha(t)$  is a function of time. The other neurons do not adjust their weights. Learning constant will change according to (13) to ensure that each training pattern has equal statistical importance and independence of presentation order [16] [17].

$$\alpha(\tau_k) = \frac{\alpha(1)}{1 + (\tau_k - 1)\alpha(1)} \quad (4)$$

where  $\tau_k \geq 1$  is the number of times neuron  $k$  has modified its weights including the current update. At the end of the training, those neurons that never had it weights modified are discarded. In this way, we had found the number of clusters and the number of hidden neurons of the WP-MLP. Both clustering algorithms had formed clusters and we can proceed to initialize the WP-MLP. The weights of the hidden neuron are assigned to be the centroid of each cluster.

## 4 Simulation Results

We have chosen two time series often found in the literature as a benchmark, i.e., the Mackey-Glass delay-differential equation and the yearly sunspot reading. For both data sets are divided into three sections for training, validation and testing. When there is an increase in the validation error, training stops. In order to have a fair comparison, simulation is carried out for each network over 100 trials (weight initialization can be different from trial to trial, since the clustering result can depend on the random initial starting point for a cluster).

### 4.1 Mackey-Glass Chaotic Time Series

The Mackey-Glass time-delay differential equation is defined by

$$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1+x(t-\tau)^{10}} - bx(t) \quad (5)$$

where  $x(t)$  represents the concentration of blood at time  $t$  when the blood is produced. In patients with different pathologies, such as leukemia, the delay time  $\tau$  may become very large, and the concentration of blood will oscillate, becoming chaotic when  $\tau \geq 17$ . In order to test the learning ability of our algorithm, we choose the value  $\tau=17$  so that the time series is chaotic. The values of  $a$  and  $b$  are chosen 0.2 and 0.1

respectively. The training data sequences are of length 400, followed by validation and testing data sequences of length 100 and 500 respectively. Tests were performed on random initialization, hierarchical clustering algorithm and counterpropagation network on varying network architecture. Eight neurons in the hidden layer are chosen and this is a prior knowledge required by hierarchical clustering algorithm to cut the tree i.e. there are eight clusters. The counterpropagation network is given ten neurons to process the data to search for the optimal clustering.

**Table 1.** Validation and test MSE of WP-MLP with random initialization on Mackey-Glass chaotic time series

Network Architecture	Validation MSE			Test MSE		
	Mean	Std	Min	Mean	Std	Min
[14:1(Db2): 8]	$4.17 \times 10^{-4}$	$1.89 \times 10^{-4}$	$2.36 \times 10^{-4}$	$1.80 \times 10^{-5}$	$8.13 \times 10^{-6}$	$1.02 \times 10^{-5}$
[14:2(Db2): 8]	$7.86 \times 10^{-5}$	$3.71 \times 10^{-4}$	$2.94 \times 10^{-5}$	$3.73 \times 10^{-6}$	$1.86 \times 10^{-5}$	$1.25 \times 10^{-6}$
[14:3(Db2): 8]	$3.62 \times 10^{-4}$	$3.62 \times 10^{-4}$	$2.24 \times 10^{-4}$	$1.56 \times 10^{-5}$	$4.51 \times 10^{-6}$	$9.65 \times 10^{-6}$
*[14: 3(Db2):8]	$1.15 \times 10^{-6}$	$2.94 \times 10^{-6}$	$1.15 \times 10^{-6}$	$1.12 \times 10^{-5}$	$3.83 \times 10^{-6}$	$1.49 \times 10^{-6}$

\*Wavelet Decomposition

**Table 2.** Validation and test MSE of WP-MLP initialized with hierarchical clustering algorithm on Mackey-Glass chaotic time series

Network Architecture	Validation MSE			Test MSE		
	Mean	Std	Min.	Mean	Std	Min.
[14: 1(Db2): 8]	$5.66 \times 10^{-6}$	$1.42 \times 10^{-5}$	$2.16 \times 10^{-6}$	$5.65 \times 10^{-6}$	$1.42 \times 10^{-5}$	$2.16 \times 10^{-6}$
[14:2(Db2): 8]	$2.74 \times 10^{-6}$	$2.16 \times 10^{-7}$	$2.31 \times 10^{-6}$	$2.75 \times 10^{-6}$	$3.24 \times 10^{-7}$	$2.31 \times 10^{-6}$
[14:3(Db2): 8]	$3.46 \times 10^{-6}$	$1.50 \times 10^{-6}$	$1.53 \times 10^{-6}$	$3.45 \times 10^{-6}$	$1.65 \times 10^{-6}$	$1.53 \times 10^{-6}$
*[14: 3(Db2):8]	$2.45 \times 10^{-5}$	$5.54 \times 10^{-5}$	$9.06 \times 10^{-6}$	$2.81 \times 10^{-5}$	$5.54 \times 10^{-5}$	$1.38 \times 10^{-5}$

\*Wavelet Decomposition

**Table 3.** Validation and test MSE of WP-MLP initialized with counterpropagation clustering algorithm on Mackey-Glass chaotic time series

Network Architecture	Validation MSE			Test MSE		
	Mean	Std	Min.	Mean	Std	Min.
[14: 1(Db2):10]	$1.935 \times 10^{-5}$	$3.69 \times 10^{-5}$	$1.78 \times 10^{-7}$	$2.33 \times 10^{-5}$	$4.30 \times 10^{-5}$	$3.54 \times 10^{-7}$
[14:2(Db2):10]	$1.02 \times 10^{-5}$	$1.06 \times 10^{-5}$	$1.73 \times 10^{-7}$	$1.22 \times 10^{-5}$	$1.21 \times 10^{-5}$	$2.42 \times 10^{-7}$
[14:3(Db2):10]	$2.303 \times 10^{-5}$	$4.21 \times 10^{-5}$	$4.62 \times 10^{-7}$	$2.96 \times 10^{-5}$	$3.52 \times 10^{-5}$	$9.82 \times 10^{-7}$
*[14: 3(Db2):10]	$1.96 \times 10^{-5}$	$1.95 \times 10^{-5}$	$6.20 \times 10^{-6}$	$2.34 \times 10^{-5}$	$2.24 \times 10^{-5}$	$7.86 \times 10^{-6}$

\*Wavelet Decomposition

Table 1 shows the results on prediction errors for the Mackey-Glass time series using different architecture of the WP-MLP, i.e. the mean, the standard deviation, and the minimum Mean Squared Error (MSE) of the prediction errors over 100 simulations. Table 2 and 3 show the result on prediction error for network initialized by hierarchical and counterpropagation respectively.

Hierarchical clustering algorithm provides consistent performance for the various network architecture except the network with wavelet decomposition. Counterpropagation network managed to attain the lowest minimum MSE with various network architecture. Thus we can say that it had indeed found the best clustering to place the weights near the optimum solution. The higher mean, larger standard deviation and lower minimum suggested that the probability of getting the optimal clustering is low. Therefore, the counterpropagation network needs more tuning to find the optimal clustering frequently.

## 4.2 Sunspots Time Series

Sunspots are large blotches on the sun that is often larger in diameter than the earth. The yearly average of sunspot areas has been recorded since 1700. The sunspots of years 1700 to 1920 are chosen to be the training set, 1921 to 1955 as the validation set, while the test set is taken from 1956 to 1979. Tests were performed on random initialization, hierarchical clustering algorithm and counterpropagation network on varying network architecture. Eight neurons in the hidden layer are chosen and this is a prior knowledge required by hierarchical clustering algorithm to cut the tree i.e. there are eight clusters. The counterpropagation network is given ten neurons to process the data to search for the optimal clustering.

**Table 4.** Validation and test NMSE of WP-MLP with random initialization on Sunspots time series

Network Architecture	Validation NMSE			Test NMSE		
	Mean	Std	Min	Mean	Std	Min
[12: 2(Db1): 8]	0.0879	0.300	0.0486	0.3831	0.2643	0.1249
[12:1(Db2):8]	0.0876	0.0182	0.059	0.4599	0.7785	0.1256
[12:2(Db2):8]	0.0876	0.0190	0.0515	1.0586	6.2196	0.1708
*[12:3(Db2):8]	0.0876	0.0182	0.0590	0.5063	1.0504	0.1303

\*Wavelet Decomposition

**Table 5.** Validation and test NMSE of WP-MLP with initialized with hierarchical clustering algorithm on Sunspots time series

Network Architecture	Validation NMSE			Test NMSE		
	Mean	Std	Min	Mean	Std	Min
[12: 2(Db1): 8]	0.0674	0.0123	0.0516	0.2682	0.1716	0.1305
[12:1(Db2):8]	0.0714	0.0103	0.0499	0.2906	0.1942	0.1426
[12:2(Db2):8]	0.0629	0.0100	0.0508	0.2126	0.0624	0.1565
*[12:3(Db2):8]	0.0691	0.0069	0.0523	0.2147	0.0585	0.1430

\*Wavelet Decomposition

**Table 6.** Validation and test NMSE of WP-MLP with counterpropagation clustering algorithm on Sunspots time series

Network Architecture	Validation NMSE			Test NMSE		
	Mean	Std	Min	Mean	Std	Min
[12: 2(Db1): 10]	0.079	0.0121	0.050	0.277	0.0819	0.1383
[12:1(Db2):10]	0.078	0.012	0.047	0.271	0.0880	0.1451
[12:2(Db2):10]	0.078	0.0130	0.0485	0.2722	0.0722	0.1442
*[12:3(Db2):10]	0.071	0.011	0.048	0.252	0.0631	0.1324

\*Wavelet Decomposition

Table 5 shows the results on prediction errors for the Sunspot time series using different architecture of the WP-MLP, i.e. the mean, the standard deviation, and the minimum Normalized Mean Squared Error (NMSE) of the prediction errors over 100 simulations. Table 6 and 7 show the result on prediction error for network initialized by hierarchical and counterpropagation respectively. Both initialization methods shown superior performance in all aspect compared to the WP-MLP that is initialized randomly. Thus we can say that for the problem of sunspots, the initialization methods place the weights close to a good solution frequently.

## 5 Conclusion

In this paper, we used hierarchical and counterpropagation clustering algorithms to initialize WP-MLP for time series prediction. Usually the weights are initialized with small random values that maybe far from or close to a good solution. Therefore the probability of being located near a good solution depends on the complexity of the solution space. In a relatively simpler problem such as Mackey-Glass time series, it is expected to have less poor local optimum. Thus the chance of being placed near a good solution is higher and initialization of weights may not be important. When dealing with difficult problem like Sunspot time series, it is expected to have a complex solution space that has lot of local optima. Thus random initialization cannot provide the consistent performance attained by the both initialization methods. Therefore, the weights of the WP-MLP must be initialized to achieve better performance.

## REFERENCE

1. Lapedes, A., Farber, R.: Nonlinear Signal Processing Using Neural Network: Prediction and System Modeling. Los Alamos Nat. Lab Tech. Rep, LA-UR-872662 (1987)
2. Mallat, S.G.: Multifrequency channel decompositions of images and wavelet models. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, No. 12 (1989) 2091 – 2110
3. Mallat, S.G.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Transactions on Pattern Recognition, Vol. 11, No. 7 (1989) 674 – 693

4. Vetterli, M., Herley, C.: Wavelets and filter banks: theory and design. *IEEE Transactions on Signal Processing*, Vol. 40, No.9 (1992) 2207–2232
5. Mallat, S.G.: *A Wavelet Tour of Signal Processing*. Academic Press (1998)
6. Strang, G., Nguyen, T.: *Wavelet and Filter Banks*. Wellesly-Cambridge Press (1996)
7. Zhang, Q., Benveniste, A.: Wavelet Networks. *IEEE Transactions on Neural Networks*, Vol. 3, No. 6. (1992) 889 -898
8. Kolen, J.F., Pollack, J.B.: Back Propagation is Sensitive to Initial Conditions. Tech. Rep. TR 90-JK-BPSIC, Laboratory for Artificial Intelligence Research, Computer and Information Science Department, (1990)
9. Coifman, R.R., Meyer, Y., Wickerhauser, M.V.: Entropy-Based Algorithm for Best Basis Selection. *IEEE Transactions on Information Theory*, Vol. 38, No. 2 Part 2 (1992) 713-718
10. Vetterli, M., Herley, C.: Wavelets and filter banks: theory and design. *IEEE Transactions on Signal Processing*, Vol. 40, No.9 (1992) 2207–2232
11. Hagan, M.T., Menhaj, M.B.: Training Feedforward Networks with the Marquardt Algorithm. *IEEE Transactions on Neural Networks*, Vol. 5, No. 6 (1994) 989 -993
12. Geva, A.B.: ScaleNet-Multiscale Neural-Network Architecture for Time Series Prediction. *IEEE Transactions on Neural Networks*, Vol. 9, No. 9 (1998) 1471 –1482
13. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. *ACM Computing Surveys*, Vol. 31, No. 3 (1999) 264-323
14. Aldenderfer, M.S., Blashfield, R.K.: *Cluster Analysis*. Sage Publications (1984)
15. Nielsen, R.H.: *Neurocomputing*. Addison-Wesley Publishing Company. (1990)
16. Wang, L.: On competitive learning. *IEEE Transaction on Neural Network*, Vol. 8, No. 5 (1997) 1214-1217
17. Wang, L.: Oscillatory and Chaotic dynamic in neural networks under varying operating conditions. *IEEE Transactions on Neural Network*, Vol. 7, No. 6 (1996) 1382-1388
18. Denoeux, T., Lengelle, R.: Initializing back propagation network with prototypes. *Neural Computation*, Vol. 6 (1993) 351-363
19. Lehtokangas, M., Saarinen, J., Kaski, K., Huuhtanen, P.: Initialization weights of a multiplayer perceptron by using the orthogonal least squares algorithm. *Neural Comput.*, Vol. 7, (1995) 982-999