⊙ 研究論文

Time Series Prediction Using a Multi - layer Neural Network with Low Pass Filter Characteristics

Min - Ho Lee

저주파 필터 특성을 갖는 다층 구조 신경망을 이용한 시계열 데이터 예측

이 민 호

Key words : Time series prediction, Neural Network, Generalization, Low pass filter

Abstract

In this paper a new learning algorithm for curvature smoothing and improved generalization for multi-layer neural networks is proposed. To enhance the generalization ability a constraint term of hidden neuron activations is added to the conventional output error, which gives the curvature smoothing characteristics to multi-layer neural networks. When the total cost consisted of the output error and hidden error is minimized by gradient-descent methods, the additional descent term gives not only the Hebbian learning but also the synaptic weight decay. Therefore it incorporates error back-propagation, Hebbian, and weight decay, and additional computational requirements to the standard error back-propagation is negligible. From the computer simulation of the time series prediction with Santafe competition data it is shown that the proposed learning algorithm gives much better generalization performance.

Ⅰ. Introduction

Multi - layer neural networks have been successfully used for complicated pattern classification and function approximation problems. The issue of generalization is usually addressed by over constraining the neural network. A lower bound on the number of training samples required for generalization by a feed - forward network with fixed number of hidden units has been asymptotically estimated using saturation property of Vapnik's and Chervonenkis's growth function[1]. However, it is still open problem to match a network's size and architecture to a given training set for good network generalization. It is necessary to compromise

* 한국해양대학교 전기공학과

among the training data size, the underlying problem complexity, and the network complexity[1]. While smaller networks are not capable of representing the problems accurately, networks with too many synaptic weights actually suffer from overfitting and result in poor generalization for test data. Although many efforts have been reported to avoid the overfitting, they usually achieve the goal by reducing the network complexity with pruning algorithms [2], and are not explicitly designed for better generalization performance. Popular algorithms for synaptic weight elimination[3], weight decay[4 - 6], and weight sharing[7 - 11] all belong to this approach.

On the other hand the problem complexity may be increased to avoid overfitting. Noise injection on training data[12] and additional cost definitions[13 - 15] belong to this approach. The additional cost definitions should be carefully designed to describe appropriate additional requirements of the problem. For many applications the neural network classifier or function approximator is proceeded by pre - processors and followed by post - processors, and it is natural to incorporate these pre - and post - processing functions in the neural networks itself. Low input - to - output sensitivity and curvature smoothing are good examples. Recently we reported a new hybrid learning algorithm based on the hidden - neuron activations[16, 17], which successfully reduced the input - to - output sensitivity for improved generalization and fault - tolerance ability. Only slight modifications are needed for the conventional error back - propagation algorithm, and additional computational requirements are almost negligible.

In this paper, a better cost function with curvature smoothing is developed for function approximator with good generalization characteristic and tested with the time series prediction problem using Santafe competition data[18]. A hybrid learning algorithm with weight decay constraint is naturally derived by the steepest descent algorithm for minimizing the proposed cost function without much increasing of computation requirement. Additional cost definitions at hidden - layer neurons is explained in the next section, and followed by computer simulation results with the time series prediction of Santafe prediction of Santafe competition data. Conclusions and discussions will be made in section IV.

Ⅱ. Additional Cost Definitions at Hidden - Layer Neurons

For many supervised learning algorithms the cost function is defined as a sum - of - squares of output errors. By minimizing this cost function the neural network learns input - to - output mapping defined by training data. In addition to this simple mapping we would like to take into account of derivatives of the mapping function. The first derivatives are sensitivities of the mapping and give better robustness and fault - tolerance ability of pattern classification problem[16,17]. In this paper, the second derivatives of mapping functions are introduced to get the better generalization performance of time series prediction problem, which are related to the frequency.

For many prediction applications one is not interested in detail high - frequency behavior of the output values, and curvature smoothing or low - pass filtering is usually followed at the post - processing stage. However, this low - pass filter characteristics may also be incorporated in the neural network as additional cost function.

The frequency square of a signal $y(x)$ is regarded as a negative ratio of its second derivatives to itself, i.e. $-(d^2y/dx^2)/y$. By applying chain rule

for the feed-forward neural network, one obtains the second derivatives as

$$\frac{d^2y_i}{dx^2_k} = \sum_j W^{(2)}_{ij} f(\hat{h}_j) f(\hat{h}_j) (W^{(1)}_{jk})^2 \qquad (1)$$

where one hidden-layer networks with linear output are considered for simplicity, x_k and y_i denote and k th element of the input vector and i th element of the output vector, respectively, and $W(1)_{ij}$ is synaptic interconnection for the l th layer. the f is derivative of Sigmoid function f and $f = -ff$ is used for the bipolar hyperbolic tangent Sigmoid function. The hat represents post-synaptic neural activation before Sigmoid squashing. By comparing Eq.(1) with $y_i = \sum_j W^{(2)}_{ij} f(\hat{h}_j)$, one way notice that $f(\hat{h}_j)(W^{(1)}_{jk})^2$ is related to the frequency square.

Instead of the standard cost function a new cost function is defined as

$$E = E_0 + \gamma E_h = \frac{1}{2N_0} \sum_i^{N_0} (t_i - y_i)^2 + \gamma E_h \qquad (2)$$

where E_0 is the normalized output error and the new hidden layer penalty term is defined as

$$E_h = \frac{1}{2N} \sum_j f(\hat{h}_j) \sum_k (W^{(1)}_{jk})^2 = \frac{1}{N_h} \sum_j f(\hat{h}_j) e_j, \qquad (3)$$

where $e_j = \sum_k (W^{(1)}_{jk})^2/2$ represents the synaptic weight energy for the first layer. It is worth noting that both the f and e_j are positive. Here N_h is the number of hidden layer neurons and γ represents relative significance of the hidden layer error over the output error. Using the steepest-descent algorithm with the cost function in Eq.(2) the weight update for the layer from now becomes

$$\Delta W^{(1)}_{jk} = \eta f(\hat{h}_j) [x_k \sum_k \delta_i W^{(2)}_{ij} + \gamma (x_k h_j e_j - W^{(1)}_{jk})], \qquad (4)$$

where η is learning rate. The first term in the brackets denotes the usual back-propagated error. The second term represents new gradient from the additional cost term and is composed of the Hebbian learning term and weight decay term. Eqs.(2) and (4) show one example to incorporate frequency in neural networks.

The beauty of the proposed learning algorithm may reside in its simpleness and straightforward easy interpretation in terms of frequency. It is worth noting that the second term in Eq.(4) is the only modification from the standard back-propagation algorithm, and additional computation requirements are almost negligible.

Ⅲ. Time Series Prediction with the Proposed Learning Algorithm

The performance of our proposed learning algorithm is illustrated with the time series prediction of Santafe data set A which is the chaotic intensity pulsations of an NH3 laser. Only 1,000 samples of the sequence were provided, and the goal was to predict the next 100 samples. In this case, a single step prediction is used for comparison of generalization performance between the multi-layer neural network with proposed learning algorithm and that with standard error back propagation learning algorithm[18]. We use two layer feed-forward neural network with 25 inputs, 40 hidden neurons, and 1 output.

Fig. 1 shows the 1100 data points of the chaotic laser data. The 100 single step prediction achieved using standard multi-layer neural network is shown Fig. 2, and the proposed method with the constant hidden constraint value γ is shown in Fig. 3. Fig. 4 is the case that the γ is changed during the learning process from large initial value(γ=0.1) to small value(γ=0.00001). As shown in Figs. 2, 3, and 4, the overfitting of the prediction data is much avoided

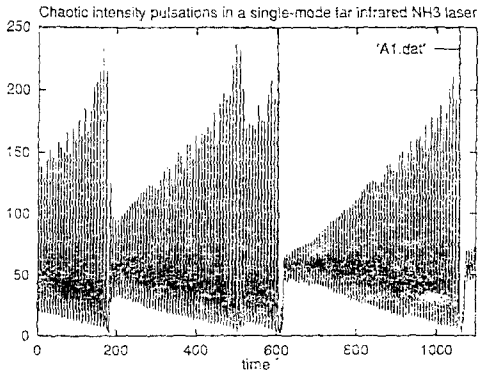

Fig. 1 Time points of chaotic laser data.

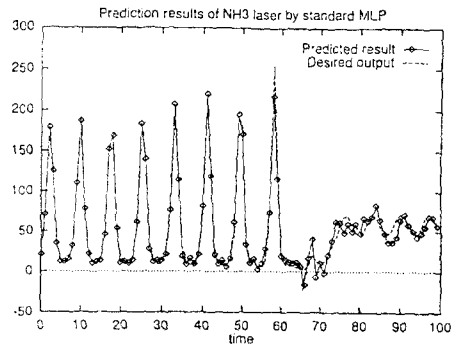

Fig. 2 Single step time series prediction using the standard error back propagation learning algorithm.

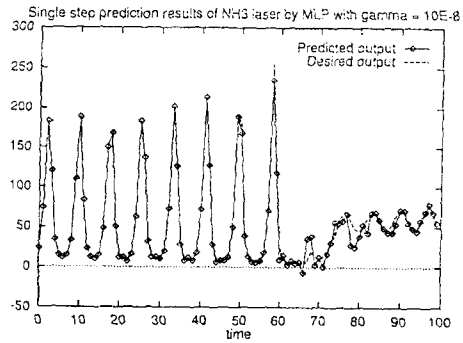

Fig. 3 Single step time series prediction using the proposed learning algorithm.

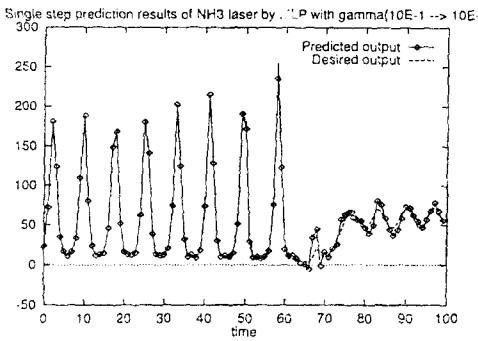

Fig. 4 Single step time series prediction using the proposed learning algorithm with gamma = 0.1 → 0.00001.

Table 1. Normalized sum squared error measures(1000 - 1100 point single step prediction).

Standard algorithm	Proposed algorithm(gamma = 10E - 5)	Proposed algorithm(gamma = 10E - 1 → 10E - 5)	Weigend
0.0365	0.0223	0.0172	0.0198

by the proposed learning algorithm. In Table 1 normalized mean square errors(NMSEs) of the predicted results are compared with other results[18].

Ⅳ. Conclusion

In this paper a hybrid learning algorithm with back - propagation, Hebbian, and weight decay is proposed for curvature smoothing and low - pass filtering. This additional functions are coded as additional weak constraint into the cost function, and the gradient - descent learning algorithm incorporates both the error back - propagation and Hebbian learning rules, and weight decay is also naturally incorporated. Only slight modification is required for the standard back - propagation code, and additional computation requirements are almost negligible. The increased generalization performance is demonstrated with the time series prediction using Santafe competition data set A. A new method for adaptively changing the hidden constraint term γ is necessary and under

investigation.

References

1) Baum, E., and Haussler, D.(1989), Neural computation, 1, 151.
2) Reed, R.(1994). IEEE Transaction on Neural Networks 2, 740 - 747.
3) LeCun, Y., Denker, J. S., and Solla, S. A.(1990). In D. Touretzky(Ed.), Advances in Neural Information Processing Systems 2(pp. 598 - 605). Morgan Kaufmann.
4) Ishikawa, M.(1994). Proc. International Conference on Fuzzy Logic, Neural Nets, and Soft Computing, 37 - 44. Iizuka, Japan.
5) Krogh, A., and Hertz, J. A.(1992). In D. Touretzky(Ed.), Advances in Information Processing Systems 4(pp. 950 957). Morgan Kaumann.
6) Weigend, A., Rumelhart, D., and Huberman, B. (1991). Neural Information Processing Systems 3, 875 - 882.
7) Fukushima, K.(1989). Neural Networks 1, 119 - 130.
8) Fukushima, K.(1993). Proc. International Joint Conference on Neural Networks, 2049 - 2054, Nagoya, Japan.
9) LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E. Hubbard, W., and Jackel, L. D. (1989). Neural Computation, 1, 541 - 551.
10) Waibel, A., Hanazawa, T., Hintion, G., Shikano, K., and Lan, K.(1989). IEEE Trans. Acoustics, Speech, Signal Processing 37, 328 - 339.
11) Nowlan, S. J., and Hinton, G. E.(1992). Neural Computation 4, 473 - 493.
12) Sietsma, J., and Dow, R. J. F.(1991). Neural Networks 4, 67 - 79.
13) Drucker, H., and LeCun, Y.(1992). IEEE Trans. Neural Networks 3, 991 - 997.\
14) Bishop, C. M.(1993). IEEE Transaction on Neural Networks 4, 882 - 884.
15) Simard, P., Victorri, B., LeCun, Y., and Deuker, J.(1992). Neural Information Processing Systems 4, 895 - 903.
16) Lee, S. Y., and Jeong, D. G.(1994a). Proc. World Congress on Neural Networks, Ⅲ, 325 - 330.
17) Lee, S. Y., and Jeong, D. G.(1994b). Proc. International Conference on Neural Information Processing, I, 189 - 194, Seoul, Korea.
18) Weigend, A., and Gershenfeld, N. A.(1992). Proc. of the NATO Advanced Research Workshop on Comparative Time Series Analysis, New Mexico, USA.Figure Captions