IMPROVING THE GENERALIZATION OF NEURAL NETWORKS BY CHANGING THE STRUCTURE OF ARTIFICIAL NEURON

***Mohammad Reza Daliri*[1], *Mehdi Fatan*[2]**
[1]Biomedical Engineering Department and Iran Neural Technology Center,
Faculty of Electrical Engineering, Iran University of Science and Technology (IUST),
Narmak, 16846-13114 Tehran, Iran (Email: daliri@iust.ac.ir)
[2]Mechatronics Group, Faculty of Electrical Engineering,
Qazvin Islamic Azad University, Qazvin, Iran (Email: Mehdi_iji@yahoo.com)
Corresponding author: M.R. Daliri, Email: daliri@iust.ac.ir,

ABSTRACT

This paper introduces a change in the structure of an artificial neuron (McCulloch and Pitts), to improve the performance of the feed forward artificial neural networks like the multi-layer perceptron networks. Results on function approximation task and three pattern recognition problems show that the performance of a neural network can be improved by a simple change in its traditional structure. The first problem is about approximation of a complicated function and the other tasks are three pattern classification problems which we have considered the digit, face and 3D object recognition experiments for evaluation. The results of the experiments confirm the improvement of the generalization of the proposed method in compared to the traditional neural network structure.

Keywords: Improve Generalization of MLP; Artificial Neuron; Function Approximation; Digit Recognition; Face Recognition; 3D Object Recognition.

1. Introduction

An artificial neural network (ANN) is a biologically driven computational model which composed of some processing elements, called neurons. The structure of a network is created by connecting the neurons with some specific weights bound to the connections. There are several training and recall algorithms attached to the structure of a neural network. The connections and their relevant weights play an important role in the structure of a neural network, for this reason they are sometimes called connectionist models to show the importance of this role. The connection weights are considered as the memory of the system [1].

Artificial neural networks have taken their structures from the brain, so they have similarities to the biological neural networks. Even though they might be considered as a model of the brain, but they are a very simplified version of it. Actually the aim of this modeling is more to use them as a tool for solving problems similar to what the brain does and not to model the brain. There are many higher-order cognitive functions in the brain which make the structure very complex and some of these functions have not been discovered yet. This makes the problem of the modeling of the brain more difficult. By developing the knowledge of the structure and the function of the brain, better computational models can be constructed and can be used for solving more complex problems [2].

ANNs have been used in solving many engineering problems including the pattern recognition and function approximation problems. The most important property of a NN is its generalization ability in solving these problems. The generalization is referred to as the ability of a network in extending its responses to new data or noisy data or the behavior of network in the new situations.

Here to show the advantages of the new proposed structure for an artificial neuron, we focus on a class of neural networks commonly used for classification problems namely the feed forward neural networks (FFNN), especially the multi-layer perceptron networks.

There are several ways for improving the functionality and performance of FFNN and data fitting is the most important ability of it. In the next section we explain some approaches that make the generalization ability of feed forward networks better and then we propose a new method that changes the structure of a classic artificial neuron (McCulloch and Pitts [16]) which is a very common neuron model used in artificial neural networks. This paper shows that by a small change in the structure of a primitive neuron, we can improve the nonlinearity and the generalization ability of the

MLP neural networks. The paper is organized in the following form: The next section describes some ways to improve the generalization of a neural network and introduce the new approach for this aim. Section 3 gives a mathematical analysis for proving less sensitivity of new neuron to input signal variations and the 4th section shows the results of a function approximation task and some pattern recognition problems for showing the advantages of the MLP networks made with this kind of changed artificial neuron. The last section concludes the paper.

2. Background and Introduction To the New Method

The generalization of a MLP network can be improved by using as much small networks as possible. This is a traditional way which has the best generalization and smoothness on test set [3]. There are several strategies that can be used in the training stage which can improve the generalization ability of a network. One approach is using the cross validation method [4]. This method continues the training until the error decreases for both train and test data and as soon as the test error increases the training stops. The training process is repeated for several times then the network with the best result on test data is selected as the optimal network. This can be computationally very expensive especially in complicated problems because the training process must be repeated for several times [3, 5].

Another way which can be used for improving the generalization of a neural network is multi-objective optimization. In this method, several cost functions are used instead of considering just the sum of the squared error (mse) of the training data. The approach balances these cost functions [6].

To improve the generalization of a network we can modify the performance function by using an additional term to the cost function of mse . This term consists of the mean of the sum of the squares of the network weights and biases [7]:

$$mserg = \gamma \ mse + (1\text{-}\gamma)\ msw \tag{1}$$

where $mserg$ is the new performance function and γ is a weight factor showing the importance of each term. In this equation the mse and the msw are defined as following:

$$mse = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)^2 \tag{2}$$

$$msw = \frac{1}{n}\sum_{i=1}^{n} w_i^2 \tag{3}$$

This new function will create the network with smaller weights and biases. This causes the network response to be smoother and so the probability of over fit to the training data decreases [7].

There are relationships between the generalization ability of a neural network and the sensitivity of it to input noise. Experimentally it has been shown that there is a kind of correlation between the generalization and the stability to input noise [8].

Another approach which can help to improve the generalization ability of a neural network is to remove those connections that are inactive in the network or to prune those connections with small values. Sometimes creating connections from one layer to the next layers (shortcut connections; Figure 1) also can help the network to have better generalization and this has been shown experimentally [9].

In [10] a sliding mode control and Levenberg-Marquardt algorithm was proposed for improving the generalization of MLPs. The proposed algorithm restricts the norm of the weights vector. The norm constrains control the degree of freedom of the neural networks. They consider different norm possibilities in their approach and they select the best generalization performance for the validation sets.

Siraj and Osman [11] proposed an improvement approach for the generalization of neural networks using MLP discriminant based on multiple classifier failures. In this approach multiple classifier systems were used to construct the discriminant set. Each classifier can learn a special situation, so that by managing the advantages of different classifiers the generalization of neural networks can be improved.

Hua et al. [12] suggested a new objective function for a single hidden network to improve the generalization performance of the feed-forward neural networks. The objective function in their approach was made by two information entropy terms including the cross entropy and the fuzzy entropy.

In [13] two approaches for the generalization ability enhancement of the neural networks was proposed namely the AlgoRobust and AlgoGS. The first algorithm was not so insensitive to the noises and the second approach used the Gauss-Schmidt algorithm. This algorithm was used to determine the weights that should be updated in each epoch and the other weights were unchanged in that epoch. Using this approach better generalization ability was obtained.

In some other studies, the generalization improvement of other neural network structures has been considered. For example in [14] an approach for the improvement of generalization of radial basis function neural networks has been proposed. This approach uses a statistical linear regression method for this improvement. The cross-validation approach was used in this study to find the stopping criteria for the training. Further improvement of the generalization of the network was also considered using a bootstrap method.

For more details about the factors that affect the generalization property of a neural network see [15].

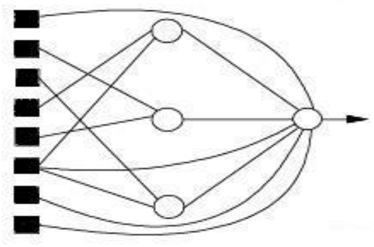

Fig .1.a neural network with shortcut connections

It has been observed that there is a nonlinear property in a neuron's structure in biological neural networks. This nonlinearity appears in several parts including in synaptic connections, in summation of input signals and also the output of a natural neuron which is referred to as the activation function. In traditional neuron models like McCulloch and Pitts [16], the nonlinear part is only expressed in the transfer function. This can create a network with a low generalization ability which can lead to lower performance in approximating nonlinear relations. Such a network composed of this type of neurons normally has many neurons causing an increase in the time of training and also the time of execution. For these reasons the process of finding the optimal network became hard.

To increase the nonlinear property of a neuron and make it more similar to a natural neuron model and solve some of the problems mentioned before, we can apply a set of nonlinear relations as net summation function or in the other parts of the architecture of a model of a neuron. A simple function which can be used for this aim is the *mean* or *average* function and in this paper, we used it as an input summation function. By this simple relation we can improve the performance of a traditional neuron structure and thus having a more powerful feed-forward network for using in different applications. The summation function of the neuron is now expressed as following:

$$NET = \frac{\sum_{i=1}^{n} w_i x_i}{N} \qquad (4)$$

where NET is the new output function of the neuron before the activation function, N is the number of inputs in a neuron architecture and the numerator is the original neuron summation function. This makes the neuron less sensitive to input changes and better generalization for unseen data. Experimental results illustrate changes in the performance of the MLP neural networks. This resulting function changes the gradient term for every neuron in the network.

$$\frac{\partial F_{new}}{\partial x} = \frac{1}{n}\frac{\partial F_{old}}{\partial x} \qquad (5)$$

Where ∂F_{new} is the new gradient term of the performance function, sum of the squares of error (SSE) and has a new term $\frac{1}{n}$, where n is the number of inputs for the neuron and this affect the learning rate for every neuron in the BP algorithm.

3. Mathematical Analysis

This section mathematically analyses the new summation function and it shows the behavior of a neuron which includes this function. These analyses are based on a single neuron and it can be generalized to a feed forward neural network like the MLP network. Such a network consists of some layers and some neurons in every layer. The results here can be extended to a multi-layer network and can be shown that it will be less sensitive to variations of input signals or noise added to them, because a noise term like β can be divided and removed several times in the neurons of a layer and in the sequences of neurons in different layers. Considering a feature vector X, we can write the following equation for the linear part of a neuron:

$$X=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \Rightarrow NET_{old} = \sum_{i=1}^{n} w_i x_i \qquad \text{And} \qquad NET_{new} = \frac{\sum_{i=1}^{n} w_i x_i}{N} \Rightarrow$$

$$\text{If } NET_{old} > 0 \Rightarrow NET_{new} < NET_{old} \Rightarrow \Phi(NET_{new}) < \Phi(NET_{old})$$

$$\text{If } NET_{old} < 0 \Rightarrow NET_{new} > NET_{old} \Rightarrow \Phi(NET_{new}) > \Phi(NET_{old})$$

where Φ is an increasing function which represents the activation function of the neuron. Now consider a new feature vector X_{new} which is a noisy version of X, we can write the following equations about it:

$$X_{new} = \begin{bmatrix} X_1 + \Delta X_1 \\ X_2 + \Delta X_2 \\ \vdots \\ X_n + \Delta X_n \end{bmatrix} \Rightarrow NET_{old}{}' = \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i \Delta x_i \Rightarrow$$

$$NET_{old}{}' = NET_{old} + \beta \qquad \text{and} \qquad NET_{new}{}' = \frac{\sum_{i=1}^{n} w_i x_i}{N} + \frac{\sum_{i=1}^{n} w_i \Delta x_i}{N} \Rightarrow$$

$$NET_{new}{}' = NET_{new} + \frac{\beta}{N}$$

where X is the vector of input signal to a neuron , NET_{old} is primitive input summation function, NET_{new} is the linear output of the proposed method, Φ is the activation function of the neuron, β is the effect of noise added to the input signal in primitive summation function and X_{new} is noisy input vector. Also NET'_{old} and NET'_{new} are the net summation functions with noisy data. The last equation shows that the neuron made of the proposed net input function is less sensitive to input variations like noise and so it has better generalization ability and is more robust to new signals.

4. Experimental Results

Here we have performed several experiments to show the ability of the proposed method in solving real problems. The experiments consist of using a neural network with the proposed structure used for function approximation, digit recognition and face recognition and 3D object recognition problems. The details of the different experiments are given in the following subsections.

4.1. Function Approximation Problem

This section illustrates the result of function approximation using the proposed structure for a feed forward neural network. The function F which was used for the experiment had the values in the range of [-5, 5] which is given by the following formula and shown in figure 2:

$$F = x^2 \mathbf{Cos}(x)^2 \mathbf{Sin}(x)^2 \tag{6}$$

Two networks are simulated using MatLab [7], one with the traditional structure and the other with the proposed structure both having two layers of neurons, one hidden layer and the output layer. We used hyperbolic tangent as the transfer function in the hidden layer and one linear neuron as the output layer. The training set consists of 100 pairs and we used another 100 pairs as the test set. This experiment was executed for 2,3,5,7 and 9 neurons in the hidden layer and repeated for 50 times. We computed the mean squared errors (MSE) for the evaluation of each network. Both networks were trained using the back propagation algorithm with a variable learning rate and a momentum term. Table 1 shows the result of the average MSE for 50 iterations. The results indicate that the proposed structure has improved the error rate of the function approximation in compared to the traditional one.

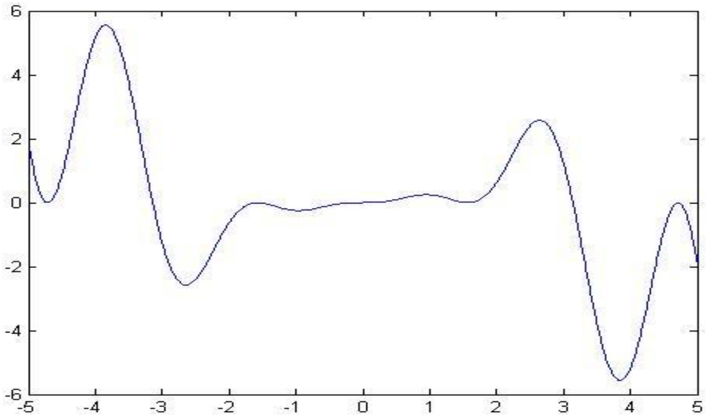

Fig.2. A plot of the function used for the experiment for approximation using neural networks.

Table 1. Result of the experiment for the function approximation.

Network Structure	**Average MSE for 50 iterations**
Traditional MLP Network	1.561
New MLP Network	0.871

4.2. Digit Recognition Problem

In this section we evaluated the ability of the proposed structure in a digit recognition problem. The data used for this experiment, was provided from a subset of the MNIST database [17]. Figure 3 represents a sample of four images for four different digits in this database. The training set was composed of 10000 hand-written digit images, each including the 724x1 vector of the data that was generated from the 28x28 original digit images. We used 5000 images of the same size as the test set. The digit images were first converted to binary images and then the raw data were changed to 724x1 vectors. The structure of the two MLP networks used for this experiment had 10 neurons in the hidden layer and the logarithm of sigmoid was used as the transfer function for hidden neurons with 10 linear neurons in the output layer. Table 2 shows the results for this experiment. For computing the recognition rate reported in the table, we considered the number of images in the test set which were classified correctly by the method divided by the total number of test images (5000). As we can see from this table, the new structure has highly improved the accuracy of the recognition.

Fig. 3. A sample of four different digit images in the MNIST database.

Table 2. Result of hand-written digit recognition experiment for the two neural networks structures.

Network Structure	**Recognition Rate**
Traditional MLP Network	80%
New MLP Network	89%

4.3. 1. Face Recognition Problem: ORL Database

In this subsection of the experimental results, we evaluated the performance of the new proposed method in the face recognition problem. The ORL face database [18] was used for this evaluation (Figure 4). There are 400 face images from 40 persons in this dataset. The data were randomly divided into train and test sets for 15 times, using 5 images per person for training and 5 images for test. So, there were in total 200 face images as a train set and 200 images as a test set. After enhancing and resizing images to 20x20 sizes, the normalized raw data of face images were used for the generation of train and test feature vectors. Two MLP networks with a two layer structure, composed of 20 hidden neurons with the logarithm of sigmoid transfer function and 40 linear output neurons, were used for this experiment. Table 3 shows the average recognition rate for the two networks. The recognition rate was measured using the number of correctly classified faces in the test set divided by the total number of faces in this set. This was measured for each randomly division of the train and test sets. In the table the average of 15 runs of this division has been reported. On average the recognition rate of the neural network with the new structure has improved by 4%.

Fig.4. A subset of ORL face database.

Table 3. Result of face recognition problem using the two neural network structures.

Network Structure	Recognition Rate
Traditional MLP Network	94%
New MLP Network	98%

4.3.2. Face Recognition Problem: Yale Database

For our evaluation for the face recognition problem, we considered another face database namely the Yale face dataset which has higher complexity than the ORL database. Yale database consists of 165 images of 15 different persons and 11 images for every person. Figure 5 shows some images of this database. Images have low quality with noisy background and variation in illumination of images. Histogram of quantized SIFT features [19] were used as feature vector for classification of images. The training set included of 90 randomly selected images and 6 images for every person. The remaining 75 images used as the test set for this experiment. This random selection was repeated for 15 times and the average of the accuracy was measured and reported here. Two MLP networks with two layers structure, composed of 35 logarithm of sigmoid hidden neurons and 15 linear output neurons were used for this experiment. Table 4 represents the results of this test. The results indicate that the proposed structure can improve the accuracy of the face recognition even in a more complex scenario. The average recognition rate for the 15 runs of the algorithm has been reported here and it has been measured using the same approach which was discussed in the previous subsection.

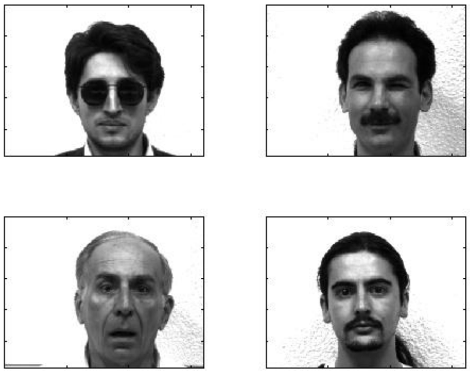

Fig.5. Some sample images from the Yale database.

Table 4. Result of face recognition with the Yale face database.

Comparision of two MLP networks	**Recognition Rate**
Traditional MLP Network	83%
New MLP Network	87%

4.4. 3D Object Recognition Task

For the last part of the experiments, 3D object recognition was evaluated for comparing the new method with the traditional one. The database of 3D objects was taken from [20], which it has two separate folders. One folder includes the images for the model and another folder for the test. These images have been taken from 11 different objects and from different views. Every folder consists of 55 images including five images for every object with different view and illumination. Similar to previous experimental results section, histogram of quantized SIFT key points [19] was used as feature vector for classification of objects. Structure of two MLP networks, composed of 25 logarithm of sigmoid hidden neurons and 11 linear output neurons were used for the experiments here. Figure 6 shows some instances of these images and Table 5 indicates the results of the recognition rates for the 3D object classification for the two networks. The results show an improvement of 3% in the recognition rate for the 3D object classification of the proposed approach in compared to the traditional method.

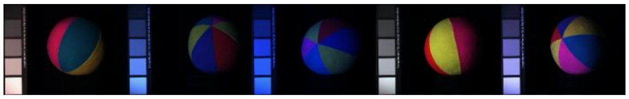

Fig. 6. Some images from the 3D Object database.

Table 5. Result of 3D object recognition.

Comparison of two MLP networks	**Recognition Rate**
Traditional MLP Network	90%
New MLP Network	93%

5. Conclusions

In this paper we have proposed a simple modification in the structure of a neuron which can improve the generalization ability of the networks which use this structure. We evaluated this structure using the feed-forward neural networks for solving several problems and compared the results with a traditional neural network. The problems in which they were used in the experiments were the function approximation, hand-written digit recognition, face and 3D object recognition. All the results indicate a better generalization property for the proposed method. This new structure sometimes concludes slower training but we can reach better decision boundaries and it makes the networks less sensitive to noisy data. Although we have shown that the modification in the artificial neuron structure can improve the performance of the feed-forward artificial neural networks like the MLP networks, but as a suggestion the proposed structure of the artificial neuron can be used in a different topology of artificial neural networks to improve their ability. This has to be tested experimentally using different neural network structures (other than the MLP which has been considered here) and the results should be compared while the same networks use the traditional neuron structure. Based on the results we have obtained here, we expect that this improves the generalization of those networks as well, because the proposed new structure has been proposed at the level of the neuron model and not at the level of the network and so should be independent of the network structure.

References:

[1] Haykin S., *Neural Networks and Learning Machines*, 3rd Edition, Prentice Hall, 2009.

[2] Kasabov N.K., *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*, The MIT Press, Cambridge, Massachusetts, London, second edition, 1996.

[3] Lawrence S., Giles C.L., Tsoi A.C., What Size Neural Network Gives Optimal Generalization? , *Technical Report*, Institute for Advanced Computer Studies, University of Maryland, 1996.

[4] Stone M., Asymptotics for and against Cross-Validation, *Biometrika*, 64 (1): 29-35, 1977.

[5] Prechelt L., Automatic Early Stopping using Cross Validation: Quantifying the Criteria, *Neural Networks*, 11 (4): 761-767, 1998.

[6] Teixeira R.A., Braga A.P., Takahashi R.H.C., Saldanha R.R., Improving Generalization of MLPs with Multi-objective Optimization, *Neurocomputing*, 35 (1-4): 189-194, 2000.

[7] Demuth H., Beale M., Neural Network Toolbox for Use with MATLAB, user's guide version 4, 2000.

[8] Bernier J.L., Ortega J., Ros E., Rojas I., Prieto A., A New Measurement of Noise Immunity and Generalization Ability for MLPs, *Int. J. Neural Syst.* , 9 (6): 511-521, 1999.

[9] Rabunal J.R., Dorado J., *Artificial Neural Networks in Real-Life Applications*, Idea Group Publishing, 2006.

[10] Costa M.A., Braga A.D.P., de Menezes B.R., Improving Generalization of MLPs with Sliding Mode Control and the Levenberg-Marquardt Algorithm, *Neurocomputing*, 70: 1342-1347, 2007.

[11] Siraj F., Osman W.R.S., Improving Generalization of Neural Networks Using MLP Discriminant Based on Multiple Classifiers Failures, *The 2nd Intl. Conf. on Computational Intelligence, Modelling and Simulation*, Bali, 2010.

[12] Hua Q., Gao Y., Wang X.-Z., Zhao B.-Y., A New Approach to Improving Generalization Ability of Feed-Forward Neural Networks, *Intl. Conf. on Machine Learning and Cybernetics (ICMLC)*, Qingdao, 2010.

[13] Wan W., Mabu S., Shimada K., Hirasawa K., Hu J., Enhancing the Generalization Ability of neural Networks Through Controlling the Hidden Layers, *Applied Soft Computing*, 9: 404-414, 2009.

[14] Lin C.L., Wang J.F., Chen C.Y., Chen C.W., Yen C.W., Improving the Generalization Performance of RBF Neural Networks Using a Linear Regression Technique, *Expert Systems with Applications*, 36: 12049-12053, 2009.

[15] Zhong S., Cherkassky V., Factors Controlling Generalization Ability of MLP Networks, *International Joint Conference on Neural Network*, Washington, DC, USA, 1999.

[16] McCulloch W. S., Pitts W., A Logical Calculus of the Ideas Immanent in Nervous Activity, Bulletin *of Mathematical Biology*, 5(4), 115-133, 1943.

[17] LeCun Y., Bottou L., Bengio Y., and Haffner P., Gradient-Based Learning Applied to Document Recognition, *Proceedings of the IEEE*, 86 (11): 2278- 2324, 1998.

[18] Samaria F., Harter A., Parameterisation of a Stochastic Model for Human Face Identification, *Proc. Of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota FL, 1994.

[19] Zhang J., Marszalek M., Lazebnik S., Schmid C., Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study, *International Journal of Computer Vision*, 73 (2): 213-238, 2007.

[20] Funt B., Barnard K., Martin L., Is Machine Colour Constancy Good Enough?, *In Proceedings of the 5th European Conference on Computer Vision*, 1998.