Improving Error Back Propagation Algorithm by using Cross Entropy Error Function and Adaptive Learning Rate

Elsadek Hussien Ibrahim
High Institute for Engineering and
Technology Al-obour
Zagazig, Egypt

Zahraa Elsayed Mohamed
Zagazig University, Faculty of Science
Zagazig, Egypt

ABSTRACT

Improving the efficiency and convergence rate of the Multilayer Backpropagation Neural Network Algorithms is an important area of research. The last researches have witnessed an increasing attention to entropy based criteria in adaptive systems. Several principles were proposed based on the maximization or minimization of cross entropy function. One way of entropy criteria in learning systems is to minimize the entropy of the error between two variables: typically, one is the output of the learning system and the other is the target. In this paper, improving the efficiency and convergence rate of multilayer Backpropagation (BP) Neural Networks was proposed. The usual mean square error (MSE) minimization principle is substituted by the minimization of entropy error function (EEM) of the differences between the multilayer perceptions output and the desired target.

On this method improving the convergence rate of the backpropagation algorithm is also by adapting the learning rate the determined learning rate is different for each epoch and depends on the weights and gradient values of the previous one. Experimental results show that the proposed method considerably improves the convergence rates of the backpropagation algorithm.

General Terms

Artificial neural networks, error back propagation, mean square error, entropy error, learning rate.

Keywords

Artificial neural network; back propagation; mean square error; entropy error; learning rate.

1. INTRODUCTION

Artificial Neural Networks (ANNs) has been an interest research area in recent years in cognitive science, computational intelligence and intelligent information processing. They have considered as an active tool for classification. The recent vast research activities in neural classification have established that neural networks are a promising alternative to various conventional classification methods. The advantage of neural networks shows in the following theoretical aspects, they consider a self-adaptive data driven method, where they modify themselves data without any external model, they can train any function, where they use a universal approximation functions, they are more flexible in modeling real world complex relationships, because they are nonlinear models. Finally, they determine posterior probabilities, so they can find classification rule and making statistical analysis. Back-Propagation Neural Network (BPNN) consider one of the most supervised learning Artificial Neural Network (ANN) models, where its mechanism depends on calculating the errors of the output layer to find the errors in the hidden layers. This ability makes it highly suitable to be applied on problems in which no relationship is found between the output and the inputs. Due to its high rate of plasticity and learning capabilities, since, it uses gradient descent learning method which requires being careful in selection of network parameters such as network topology, initial weights and biases, activation function, performance function, learning rate, and value for the gain in the activation function. An unfortunately use of these parameters can lead to slow network convergence. Other works havet suggested some modifications to improve the training time of the network, where some of them suggested using the learning rate to speed-up the network convergence. This parameter is more effective in weight adjustments along the steepest descent. Usually error backpropagation for neural network learning use MSE as the performance function. During the learning process, the ANN goes through stages in which the reduction of the error can be extremely slow. These periods of stagnation can influence learning times. In order to resolve this problem, the MSE are replaced by entropy error function. The entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable. Entropy is a nonlinear function to represent information that can be learned from unknown data. In the learning process, some constraints on the probability distribution of the training data can be learned from their entropy. Simulation results using this error function shows a better network performance with a shorter stagnation period. Accordingly, the work dived to two parts. First, replacing cross entropy function instead of the mean square error function as a performance function. Second, adapting learning rate. This will lead to increasing convergence of error back propagation algorithm.

This paper is organized as follows. In Section 2 presenting the related work of this problem. The learning algorithms as back propagation algorithm, cross entropy function, and adaptive learning rate are provided in section 3. Experiments and results are proposed in section 4. The conclusions are also provided in section 5.

2. RELATED WORKS

There are many works for solving the problem of ANNs training and learning with different method. In [10] they modify the learning strategies to improve the convergence rates of two-term BP mode. The experiment results show that the modified two-term BP improved with a convergence rate much better when compared with standard BP. In [12] they presented a backpropagation algorithm with adaptive learning rate and momentum. Where they are adapted at each iteration to speed up the backpropagation algorithm. On the other hand, in [1] they proposed a new learning algorithm based on two-term BP method using adaptive learning rate. The proposed algorithm consists of two steps. First step, the learning rate is adjusted after each iteration. Second step, the search algorithm is refined by previous weight configurations and decreasing the global learning rate. A new dynamic learning rate in BP neural network method was proposed in [2], it depends on the change of system error to change the learning rate value. In [6] they showed the improved training algorithm of BP with self adaptive learning rate. The result of the experiment shows the effectiveness of the proposed training algorithm. Meanwhile, in [4] they proposed a differential adaptive learning rate method, where it put large learning rate at the beginning of training and gradually decreases the value of learning rate using the differential adaptive method to minimize the error and increase the convergence speed. In [11] they presented a new modified back propagation algorithm with adaptive learning rate. The method determines initial fixing of learning rate through trial and error and replaces by adaptive learning rate. In each iteration, adaptive learning rate for output and hidden layer are calculated by differential linear and nonlinear errors of output and hidden layers separately. Adaptive learning rate algorithm to train a single hidden layer neural network was proposed in [5]. The adaptive learning rate is derived by differentiating linear and nonlinear errors. The proposed algorithm converges quickly. In [16] they introduced an adaptive learning rate method by adding adjustment factor; the results showed that decreased Convergence of improved BP network. From the previous works which are mentioned above and their methods to improve the two-term BP network training and learning, there are still open fields on the enhancement of performance of BP algorithm in training and learning.

3. LEARNING ALGORITHMS

3.1 Backpropagation Algorithm

Is a gradient descent algorithm in which the network weights are adapted depending on old weight and the gradient of the performance function as shown in the following equation:

$$w(t+1) = w(t) - \eta \ \partial E / \partial w(t) \qquad (1)$$

Where η is the constant learning rate, and $\partial E \,/\, \partial W(t)$ is the derivative (slope) of the error E at time t (in epochs).

3.2 Cross Entropy Error Function

During the learning process, the ANN goes through stages in which the reduction of the error can be extremely slow. These periods of stagnation can influence learning times. In order to resolve this problem, cross entropy error function is proposed to replace the mean square error (MSE). Simulation results using this error function show a better network performance with a shorter stagnation period. The original MSE function for all training patterns is given by

$$E_m = \sum_{k=1}^{m} (t_k - y_k)^2 \qquad (2)$$

Where t_k represents the target value and y_k is the actual network value. In the backpropagation model, the error is minimized through iterative updates of weights for all training patterns. In practice, this approach enables the network to have a good performance but slow convergence to the final outcome. Therefore, in order to accelerate the BP algorithm and instead of minimizing the squares of the differences between the actual and target values summed over the output units and all cases, the following cross entropy error function is proposed to be minimized:

$$E_m = \frac{1}{m} \sum_{k=1}^{m} [t_k \ln y_k + (1 - t_k) \ln(1 - y_k)] \qquad (3)$$

To minimize the error E_m, each weight w_{jk} is updated by an amount proportional to the partial derivative of E_m with respect to the weight. Using the mean square error, the partial derivative of E_m with respect to w_{jk} is

$$\frac{\partial E_m}{\partial w_{jk}} = (y_k - t_k) y_k (1 - y_k) z_j \qquad (4)$$

Where z_j is the input of hidden layer j.

By using the cross entropy error function, the partial derivative of E_m with respect to w_{jk} becomes

$$\frac{\partial E_m}{\partial w_{jk}} = (y_k - t_k) z_j \qquad (5)$$

Thus, the error signal, propagating back from each output unit, becomes directly proportional to the difference between target value and actual value leading to a better network performance with a shorter stagnation period.

3.3 Adaptive Learning Rate Backpropagation Algorithm

With standard steepest descent, the learning rate is a constant value through the training process. The performance of the algorithm is very sensitive to the best value of the learning rate. If the learning rate is a high value, the algorithm may oscillate and don’t reach to the best convergence. If the learning rate is too small, the algorithm will take too long to converge. In fact, the optimal learning rate changes during the training process, as the algorithm moves across the performance surface. The performance of the steepest descent algorithm can be improved through allowing the learning rate to be adaptive during the training process. An adaptive learning rate will attempt to keep the learning step size as large as possible while keeping learning stable. A lot of researches used adaptive learning rate through adding or subtracting a constant value but this method doesn’t give a better convergence rates. The learning rate is made responsive to the complexity of the local error surface. An adaptive learning rate can be adaptive in the training procedure used through it. First, the initial network output and error are calculated. At each epoch, new weights and biases are calculated using the current learning rate. New outputs and errors are then calculated. If the new error exceeds the old error by more than a predefined ratio, the learning rate is decreased (typically by adding α). Otherwise, if the new error is less than the old error, the learning rate is increased (typically by subtracting α). This can be shown in the following equation:

$$\eta(t+1) = \begin{cases} \eta(t) + \alpha & if\ E(t) > E(t-1) \\ \eta(t) - \alpha & if\ E(t) < E(t-1) \\ \eta(t) & if\ E(t) = E(t-1) \end{cases} \qquad (6)$$

$$where \quad \alpha = \frac{\nabla E(t+1)}{\nabla E(t)}$$

4. PROPOSED ALGORITHM

- Determine network structure, initial weight and fixed learning rate.
- Compute the net value and output value for every hidden layer.
- Compute the net value and output for output layer.
- Calculate the error value for output layer.
- Calculate the new weight value.
- Repeat steps from 2 to 5.
- Compute adaptive learning rate.
- Update weight by adaptive learning rate.
- Stop at minimum error and minimum gradient.

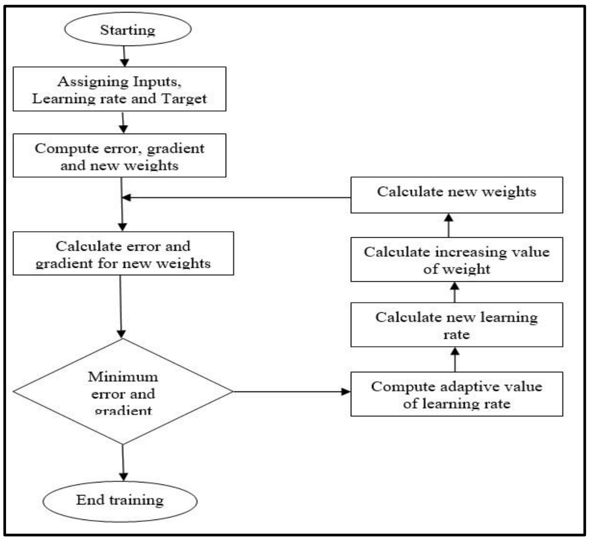
Starting
Assigning Inputs, Learning rate and Target
Compute error, gradient and new weights
Calculate new weights
Calculate error and gradient for new weights
Calculate increasing value of weight
Calculate new learning rate
Minimum error and gradient
Compute adaptive value of learning rate
End training

Fig 1: proposed algorithm flow chart

5. EXPERIMENTS AND RESULTS

In this section, testing the proposed algorithm by using two data sets from MATLAB data sets cancer data set and iris data set. The first data set contain (699) patient for every patient (9) features as input, and target values (0 or 1). The second data set contain (150) iris flowers each iris described with (4) features as input.

5.1 Cancer Data Set

First: training with (MSE) as a performance function with new adaptive learning rate. Figure (2) show that the best validation performance is 0.025606 at epoch 42.

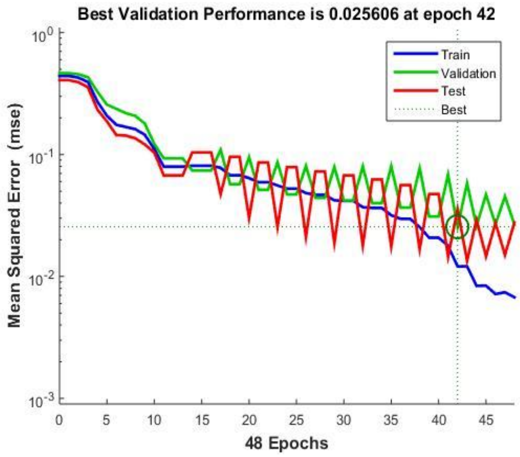

Fig 2: proposed algorithm with MSE for cancer data set

Second: training with (cross entropy) as a performance function with new adaptive learning rate. Figure (3) show that the best validation performance is 0.026298 at epoch 10.

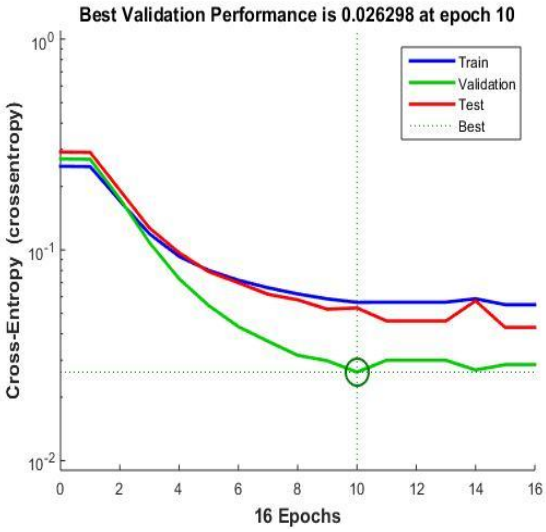

Fig 3: proposed algorithm with cross entropy error for cancer data set

From figures (2) and (3) the training curves for cancer data set oscillate to reach the error global minimum value for every epoch. Using adaptive learning rate in weight update form as proposed algorithm and using error entropy function caused reaching the global minima faster than using mean square error function.

5.2 Iris Data Set

First: training with (MSE) as a performance function with new adaptive learning rate. Figure (4) show that the best validation performance is 0.010701 at epoch 45.

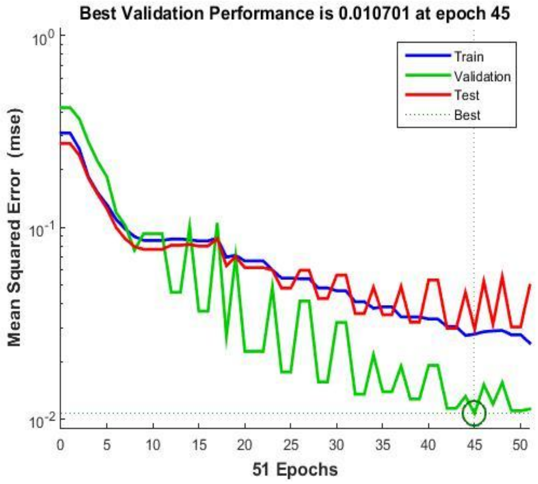

Fig 4: proposed algorithm with MSE error for iris data set

Second: training with (cross entropy) as a performance function with new adaptive learning rate. Figure (5) show that the best validation performance is 0.019627 at epoch 24.

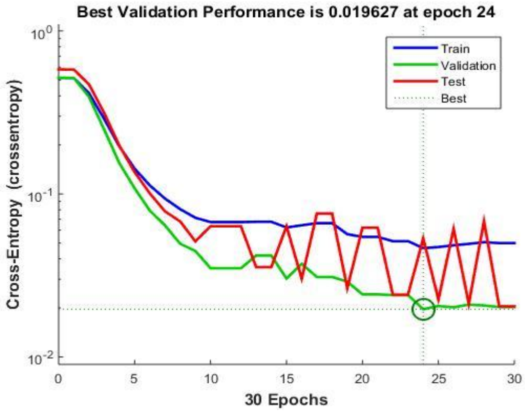

Fig 5: proposed algorithm with cross entropy error for iris data set

From figures (4) and (5) the training process for iris data set attend to reach the global minimum value for error at every epoch. Using proposed algorithm and using error entropy function lead to reach the global minima faster than using mean square error function, and give best convergence rates: best validation, and no of epochs.

6. DISCUSSION THE RESULTS

The main purpose of this research to improve the performance of BP algorithm as convergence time and convergence rates. Best validation, no of epochs, and gradient are recorded for each cancer data set and iris data set.

6.1 Cancer data set

Figure (6) shows the relation between the best validation and number of epochs by using cross entropy error function.

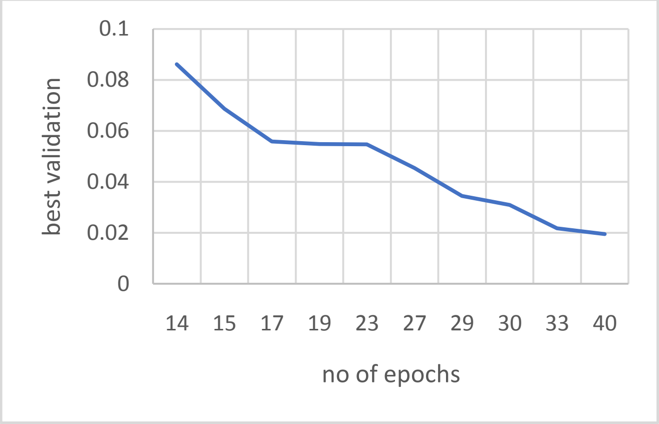

Fig 6: relation between best validation and no of epochs for cancer data set

Figure (7) shows that using of cross entropy function as error function lead to minimization of error with a small range of no of epochs although using mean square error isn't the same performance. The figure illustrates the comparison between using EEM and MSE for cancer data set.

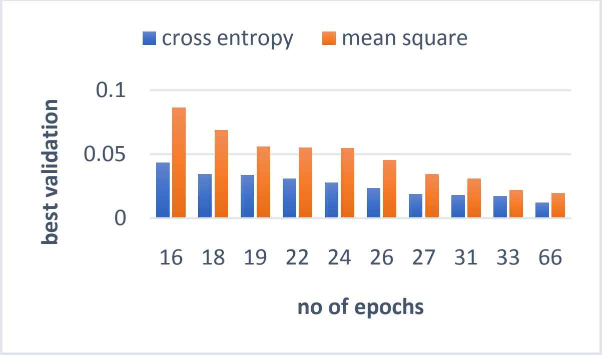

Fig 7: comparison between EEM and MSE for cancer data set

6.2 Iris data set

Figure (8) shows the relation between the best validation and number of epochs by using cross entropy error function.

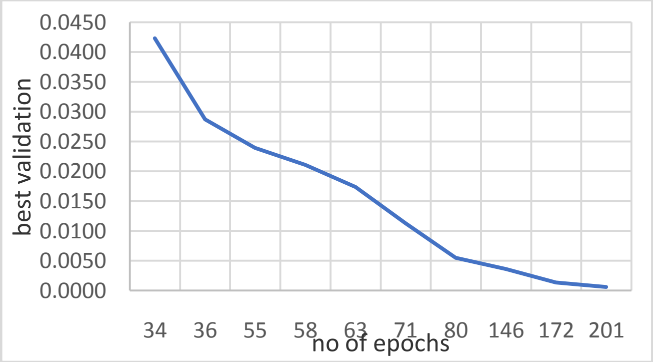

Fig 8: relation between best validation and no of epochs for iris data set

Figure (9) shows that using of cross entropy function as error function lead to minimization of error with a small range of no of epochs although using mean square error isn't the same performance. The next figure illustrates the comparison between using EEM and MSE for iris data set

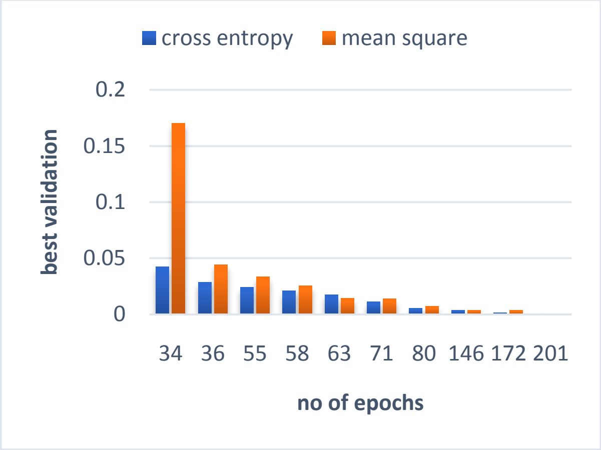

Fig 9: comparison between EEM and MSE for iris data set

7. CONCLUSION

Cross entropy function considers a binary classification error function. It leads to faster training and improved generalization. It is shown that using cross entropy function as a performance function instead of mean square error function with adaptive learning rate training improves the convergence speed of error back propagation algorithm where in the experimental results clearly show that the proposed method has improved the average number of epochs needed, and has better convergence rates. In Future, adding a new parameter for weight update form will increase the convergence rates: best validation and no of epochs.

8. REFERENCES

[1] Duffner, S., and Garcia, C. (2007). An online backpropagation algorithm with validation error-based adaptive learning rate. Artificial Neural Networks–ICANN 2007, 249-258.

[2] Guangjun, S., Jialin, Z., and Zhenlong, S. (2008). The research of dynamic change learning rate strategy in BP neural network and application in network intrusion detection. Proceedings of the 3rd International Conference on Innovative Computing Information and Control, Jun. 18-20, IEEE Xplore Press, Dalian, Liaoning, pp: 513-513.

[3] Hongmei, S., and Gaofeng, Z. (2009). A new BP algorithm with adaptive momentum for FNNs training.Proceedings of the WRI Global Congress on Intelligent Systems, (IS’ 09), IEEE Computer Society, pp: 16-20.

[4] Iranmanesh, S., and Mahdavi, M. A. (2009). A differential adaptive learning rate method for back-propagation neural networks. World Academy of Science, Engineering and Technology, 38, 289-292.

[5] Kathirvalavakumar, T., and Subavathi, S. J. (2012). Modified backpropagation algorithm with adaptive learning rate based on Citra Ramadhena et al. differential errors and differential functional constraints, Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, Mar. 21-23, IEEE Xplore Press, Salem, Tamilnadu, pp: 61-67.

[6] Li, Y., Fu, Y., Li, H., and Zhang, S.-W. (2009). the improved training algorithm of back propagation neural network with self adaptive learning rate. Proceedings of the International Conference on Computational Intelligence and Natural Computing, Jun 6-7, Wuhan, China, IEEE Computer Society, pp: 73-76.

[7] Nasr, G. E., Badr, E. A., and Joun, C. (2002). Cross entropy error in neural networks: forecasting gasoline demand. In proceedings of FLAIRS-02, 381-384: AAAI Press.

[8] Norhamreeza Abdul Hamid, N. A. H., Nazri Mohd Nawi, N. M. N., Rozaida Ghazali, R. G., and Mohd Najib Mohd Salleh, M. N. M. S. (2011). Accelerating Learning Performance of Back Propagation Algorithm by Using Adaptive Gain Together with Adaptive Momentum and Adaptive Learning Rate on Classification Problems. International Journal of Software Engineering and Its Applications, 5(4), 31-44. Network.

[9] Rady, H. (2011). Reyni’s entropy and mean square error for improving the convergence of multilayer backprobagation neural networks: a comparative study. International journal of electrical &computer sciences IJECS-IJENS vol: 11 no: 05.

[10] Shamsuddin, S. M., Sulaiman, M. N., and Darus, M. (2001). An improved error signal for the backpropagation model for classification problems. International Journal of Computer Mathematics, 76(3), 297-305.

[11] Subavathi, S. J., and Kathirvalavakumar, T. (2011). Adaptive modified backpropagation algorithm based on differential errors. International Journal of Computer Science, Engineering and Applications (IJCSEA), 1 (5), 21-34.

[12] Xiaoyuan, L., Bin, Q., and Lu, W. (2009). A new improved BP neural network algorithm. Paper presented at the 2009 2nd International Conference on Intelligent Computing Technology and Automation, ICICTA 2009, October 10, 2009 - October 11, 2009, Changsha, Hunan, China, 19-22.

[13] Yam, J. Y. F., and Chow, T. W. S. (2000). A weight initialization method for improving training speed in feed forward neural network. Neurocomputing, 30(1), 219-232.

[14] Yu, C.-C., and Liu, B.-D. (2002). A backpropagation algorithm with adaptive learning rate and momentum coefficient Proceedings of the International Joint Conference on Neural Networks, May 12-17, IEEE Xplore Press, Honolulu, HI, pp: 1218-1223.

[15] Yu, X. H., and Chen, G. A. (1997). Efficient backpropagation learning using optimal learning rate and momentum. Neural Networks, 10(3), 517-527.

[16] Zhang, X. H., Ren, F. J., and Jiang, Y. C. (2012). An Improved BP Algorithm Based on Steepness Factor and Adaptive Learning Rate Adjustment Factor. Applied Mechanics and Materials, 121, 705-709.