

# A SOLUTION TO THE PROBLEM OF LOCAL MINIMA IN BACKPROPAGATION ALGORITHM

Aziz Can Yüçeturk<sup>1</sup>, Amaç Herdağdelen<sup>2</sup>, Kıvanç Uyanık<sup>2</sup>

<sup>1</sup>Computer Engineering Dept., Ege University, Izmir, Turkey

<sup>2</sup>Izmir Fen Lisesi, Izmir, Turkey

Email: yuceturk@staff.ege.edu.tr, amac.herdagdelen@service.raksnet.com.tr

## Abstract

The backpropagation algorithm is successfully used in many applications in multilayered network as a supervised learning technique. But the weak side of the algorithm that impacts the performance of the method is the presence of local minima. In this paper a new technique has been developed to overcome this problem. During the training, a mutation operation is applied to the algorithm, so if the algorithm gets stuck in local minima, it has a chance to jump out of local minima and continue to training process.

**Keywords:** Artificial Neural Network, Backpropagation, Local Minima

## 1. Introduction

The backpropagation algorithm is central to much current work on learning in neural network. The development of backpropagation was reported by Rumelhart, Hinton and Williams in 1986 [1]. This algorithm is based on the error correction learning rule. The very general nature of the backpropagation training method means that a backpropagation net can be used to solve problems in many areas [2, 3].

In spite of these important properties, a major criticism is commonly moved against backpropagation algorithm. Backpropagation algorithm is actually a gradient method and therefore there is no guarantee at all that the global minimum of error surface can be reached. It is useful to have a simple geometrical picture of the error minimization process, which can be obtained by viewing error function  $E(w)$  as an error surface sitting above weight space, as shown in Figure 1 [4].

The minimum for which the value of the error function is smallest is called the global minimum, while other minima are called local minima. Local minimum and global minimum points are illustrated in Figure 1 as A and B, respectively.

A more theoretical definition of local and global minimum is as follows [5]. A vector  $w^*$  is said to be a local minimum of an input-output function  $F$  if it is no worse than its neighbors, that is, if there exists an  $\xi$  such that

$$F(w^+) \leq F(w) \quad \text{for all } w \text{ with } \|w - w^*\| < \xi$$

The vector  $w^+$  is said to be a global minimum of the function  $F$  if it is no worse than all other vectors, that is,

$$F(w^*) \leq F(w) \quad \text{for all } w \in \mathcal{R}^n$$

Where  $n$  is the dimension of  $w$ .

The only general analysis concerning the problem of local minima in multilayer perceptrons has been published by Baldi and Hornik [6]. Under the hypothesis that the neurons are linear, it has been proved that only one minimum exists, the other points where the gradient is null being saddle points.

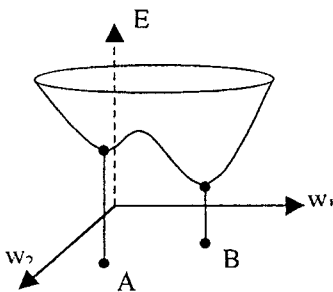


Figure 1. Error Function  $E(w)$

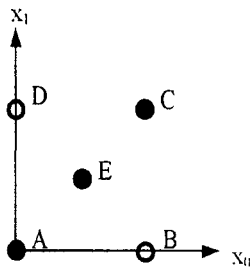


Figure 2. Training Example

## 2. Proposed Approach

As defined earlier, the point of local minimum is closest point to the solution compared to its all neighbors. But there must be a global minimum far away from this point. As the system falls into a local minima, the small changes in the weight vector do not take out the system from this hole and bring it to global minimum. Therefore we need radical changes in weight vector so that the system gets out of this local minimum.

The mutation operation that is commonly used in genetic algorithms is proposed as a solution to this problem. Each time we apply the input-output training set to the system, we also apply a mutation operation to the original synaptic weights in the system. In this way, we obtain both the trained weights and the mutant weights. As we use the mutant weights with mutant weights and continue with the training procedure. If the error function of the trained weights is less than the mutant weights, that means there is no need for a mutation operation. Then we continue to training phase with trained weights. The mutation operation is applied to the weights upon a probability function. The mutant weight is a function of original weight with a probability of  $p$ . Also a mutation constant determines the effect of mutation operation to the original weights.

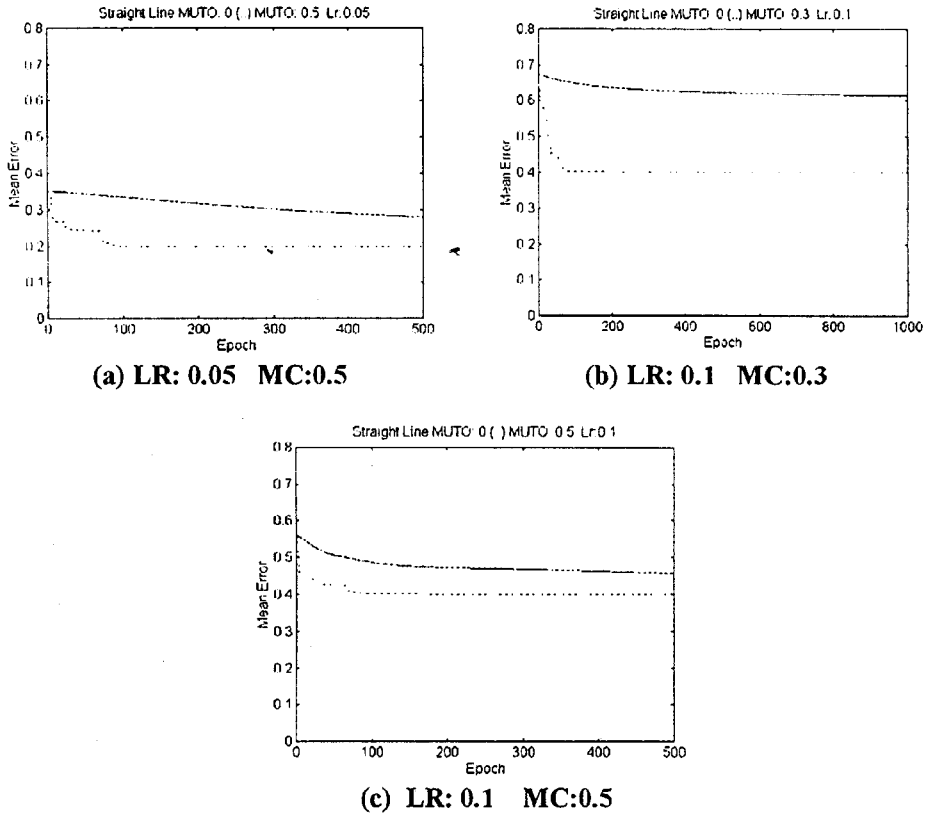
In this study we used the simple example that is proposed by Gori and Tesi [7]. This is an XOR-like function with an additional point 'E' as given in Figure 2. Practical results of the proposed method are given in Figure 3. These figures are representing Mean Error-Epoch graphs. The mean error is calculated as the mean of absolute difference between target values and actual outputs. In each figure there are straight lines and dashed lines. Straight lines are representing the error of traditional backpropagation algorithm and the dashed lines are representing the error of the new approach proposed in this study.

## 3. Conclusion

In this study we analyzed the problem of local minima in backpropagation learning and developed a new method so that the backpropagation converges to optimal solution.

If we analyze the graphs given in Figure 3 we conclude that if we take small learning rates, the new method gives much more successful results. If the learning rate gets higher, then the performance of the new approach decreases. The reason is that the deviation of weight vector that is caused by the mutation effect becomes negligible with higher learning rates. In Figure 3, the dashed lines were illustrating the error of new

approach. The stepwise structures of these dashed lines are the results of mutation operation applied to the weight vectür.



**Figure 3. Error Graphs with Different Learning Rates (LR) and Mutation Constants (MC) (Dashed Lines: New Approach).**

## References

- [1] Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986, "Learning Representations by Backpropagation Error", *Nature*, 323, pp. 533-536.
- [2] Elman, L., Zipser, D., 1988, "Learning the Hidden Structure of the Speech", *Journal of Acoustic Society of America*, Vol. 83, No. 4.
- [3] Bourslard, H., Wellekens, C., 1989, "Speech Pattern Discrimination and Multilayered Perceptrons", *Computer Speech and Language*, Vol. 3, pp. 1-19.
- [4] Bishop, C.M., 1995, *Neural Networks for Pattern Recognition*, Oxford University Press, New York.
- [5] Bertsekas, D.P., 1995, *Nonlinear Programming*, Belmont, MA: Athena Scientific.
- [6] Baldi P., Hornik, K., 1989, "Neural Networks and Principal Component Analysis: Learning from Examples and Local Minima", *Neural Networks*, Vol. 2, pp. 53-58.
- [7] Gori, M., Tesi, A., 1992, "On the Problem of Local Minima in Backpropagation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, pp. 76-86.