Scientia Iranica, Vol. 15, No. 6, pp. 558–567

SCIENTIA
IRANICA

Research Note

Optimum Learning Rate in Back-Propagation Neural Network for Classification of Satellite Images (IRS-1D)

J. Amini[1]

Remote sensing data are essentially used for land cover and vegetation classification. However, classes of interest are often imperfectly separable in the feature space provided by the spectral data. Application of Neural Networks (NN) to the classification of satellite images is increasingly emerging. Without any assumption about the probabilistic model to be made, the networks are capable of forming highly non-linear decision boundaries in the feature space. Training has an important role in the NN. There are several algorithms for training and the Variable Learning Rate (VLR) is one of the fastest. In this paper, a network that focuses on the determination of an optimum learning rate is proposed for the classification of satellite images. Different networks with the same conditions are used for this and the results showed that a network with one hidden layer with 20 neurons is suitable for the classification of IRS-1D satellite images. An optimum learning rate between the ranges of 0.001-0.006 was determined for training the VLR algorithm. This range can be used for training algorithms in which the learning rate is constant.

INTRODUCTION

The maximum likelihood algorithm with Gaussian probability distribution functions is considered the best classifier, in the sense of obtaining the optimal classification rate. However, the application of a neural network to the classification of a satellite image is increasingly emerging. Without any assumption about the probabilistic model to be made, neural networks are capable of forming highly non-linear decision boundaries in the feature space. Also, they have the potential of outperforming a parametric Bayes classifier when the feature statistics deviate significantly from the assumed Gaussian statistics.

Bendiktsoon et al. [1] compared neural networks and statistical approaches, together, with a multispectral data classification. They noted that conventional multivariate classification methods cannot be used in processing multisource spatial data. This is due to different distribution properties and measurement scales. Heermann and Khazenie [2] compared neural networks with more classical statistical methods. Heerman and Khazenie's study emphasized the analysis of larger data sets with back propagation methods, in which error is distributed throughout the network. They concluded that the back propagation network could be easily modified to accommodate more features or to include spatial and temporal information. Hepner et al. [3] compared the use of neural network back propagation with a supervised maximum likelihood classification method, using a minimum training set. The results showed that a single training site per class of neural network classification was comparable to a four training site per class of conventional classification. The result demonstrated that the neural network method offered a potentially more robust approach to land cover classification than conventional image classification methods.

In this paper, a Multi Layer Perceptron (MLP) network with a Back Propagation (BP) algorithm is used for the classification of IRS-1D satellite images. A network with an optimum learning rate is proposed. The MLP consists of neurons that are arranged in multiple layers with connections only between nodes in the adjacent layers by weights. The layer where the

1. *Department of Surveying Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran. Email: jamini@ut.ac.ir*

input information is presented is known as the input layer and the layer where the processed information is retrieved is called the output layer. All layers between the input and output layers are known as hidden layers.

For all neurons in the network, except the input layer neurons, the total input of each neuron is the sum of the weighted outputs of the neurons in the previous layer. Each neuron is activated with input to the neuron and by the activation function of the neuron [4].

The input and output of the neuron, i, (except for the input layer) in a MLP mode, according to the BP algorithm [5], are:

$$\text{Input} \qquad X_i = \sum W_{ij} O_j + b_i, \tag{1}$$

$$\text{Output} \qquad O_i = f(X_i), \tag{2}$$

where W_{ij} is the weight of the connection from neuron i to node j, b_i is the numerical value called the bias and f is the activation function.

The sum in Equation 1 is over all neurons, j, in the previous layer. The output function is a nonlinear function, which allows a network to solve problems that a linear network cannot [2]. In this study, the tan-sigmoid and linear transfer functions are used to determine the output.

A Back-Propagation (BP) algorithm is designed to reduce error between the actual output and the desired output of the network in a gradient descent manner. The Mean Square Error (MSE) is defined as:

$$\text{MSE} = \frac{1}{2}\left(\sum_p \sum_i O_{pi} - T_{pi}\right)^2, \tag{3}$$

where p indexes the all training patterns and i indexes the output neurons of the network. O_{pi} and T_{pi} denote the actual output and the desired output of the neuron, respectively, when the input vector, p, is applied to the network.

A set of representative input and output patterns is selected to train the network. The connection weights, W_{ij}, are adjusted when each input pattern is presented. All the patterns are repeatedly presented to the network until the MSE function is minimized and the network "learns" the input patterns. Applications of the gradient descent method [6] yield the following iterative weight update rule:

$$\Delta W_{ij}(n+1) = \eta(\delta_i O_i + \alpha \Delta W_{ij}(n)), \tag{4}$$

where Δ is the learning factor and α is the momentum factor. δ_i, the neuron error, for output neuron i is, then, given as follows:

$$\delta_i = (T_i - O_i) O_i (1 - O_i). \tag{5}$$

The neuron error at an arbitrary hidden neuron is:

$$\delta_i = O_i(1 - O_i) \sum_k \delta_k W_{ki}. \tag{6}$$

In the rest of this paper, first, a fast training algorithm Variable Learning Rate (VLR) is discussed. Then, methodology and experimental results, to determine the interest network and optimum learning rate, are depicted later and finally the conclusion is presented.

FAST TRAINING ALGORITHM

Before training the network, the weights and biases must be initialized. Here, these values are randomly selected between 0 to 1. Now, the network is ready for training. The training process requires a set of training sites. During training, the weights and biases of the network are iteratively adjusted to minimize the network performance function. Here, the performance function for the network is the mean square error between the network outputs and the target outputs.

There are several training algorithms for MLP [7], some of which are fast algorithms. The faster algorithms fall into two main categories. The first category uses heuristic techniques, which were developed from an analysis of the performance of the standard steepest descent algorithm. The second category uses standard numerical optimization techniques. Variable Learning Rate (VLR) back propagation is in the first category used in this paper. In the standard steepest descent, which was discussed previously, the learning rate is held constant throughout the training. The performance of the algorithm is very sensitive to the proper setting of the learning rate. If the learning rate is set too high, the algorithm may oscillate and become unstable and, if the learning rate is too small, the algorithm will take too long to converge. It is not practical to determine the optimal setting for the learning rate before training and, 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, if we allow the learning rate to change during the training process. An adaptive learning rate will attempt to keep the learning step size as large as possible, while keeping the learning stable. The learning rate is made responsive to the complexity of the local error surface. In the adaptive learning rate, first, the initial network output and error are calculated. At each epoch, new weights and biases are calculated, using the current learning rate, and new outputs and errors are then calculated. In this algorithm, there are five training parameters: Epoch, goal, time, min-grad and lr. The learning rate, lr, is successively multiplied, with the negative of the

gradient, to determine the changes of the weights and biases. If the learning rate is made too large, the algorithm becomes unstable and, if the learning rate is set too small, the algorithm takes a long time to converge. The other parameters determine when the training stops. The training stops if the number of iterations exceed epoch, if the performance function drops below goal, if the magnitude of the gradient is less than min-grad or if the training time is longer than time seconds.

METHODOLOGY AND EXPERIMENTAL RESULTS

As seen in the previous section, the aim of this paper is the investigation of the variation learning rate, in order to determine the optimum network for classification of IRS-1D satellite images, with a neural network. For simplicity, the intensity values for three bands of an IRD-1D image are used as inputs in the input layer. Two classes; urban and suburb, are considered in the output layer.

The following parameters are used for all networks:

Number of epochs: epochs = 500,

Goal of performance function: goal = 0,

Initial learning rate: lr = 0.0010,

Increased learning rate: lr_inc = 1.0500,

Decreased learning rate: lr_dec = 0.7000,

Magnitude of gradient: min_grad = 1.0000e-008 and

Training time: time= Inf.

Two types of network, with one hidden layer and two hidden layers, respectively, are used for investigation of the learning rate. Also, two activation functions, i.e. tan-sigmois and linear, are used in the hidden and output layers, respectively. The hidden layer is responsible for internal representation of data and the information transformation between input and output layers (i.e., the learning) [8]. If there are too few neurons in the hidden layer, the network may not contain sufficient degrees of freedom to form a representation (i.e., insufficient learning capacity). If too many neurons are defined, the network may become over trained (i.e., they classify training patterns well but lack the ability to generalize other independent data) [2]. Therefore, an interesting design for the number of neurons in the hidden layer, to determine the optimum network for classification of IRS-1D images, will be important. So, here, different forms in the hidden layers are considered for each type of network.

Data for training the networks were acquired through interactive pixel sampling of an IRS-1D image. To avoid spatial autocorrelation and neighbouring pixel influences, each of the sample pixels was selected individually.

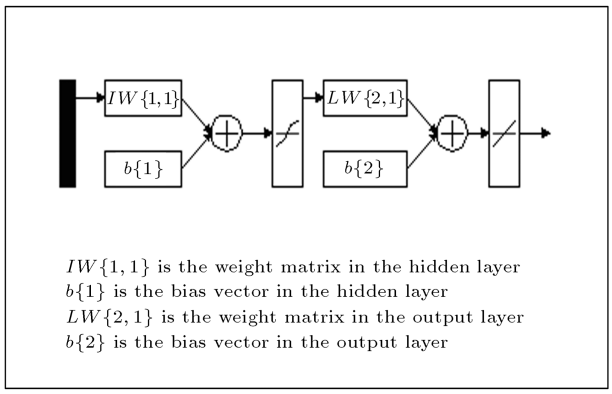

Figure 1. The network in type I.

Figure 1 shows the first network (type I) with one hidden layer.

Number of neurons in the hidden layer are varied for NN = 3, 5, 7, $\cdots$, 40, according to column 1 in Table 1. So, in this type, there are nine networks.

Now, each network is trained to determine the optimum learning rate.

Table 1 shows the performance (MSE) and optimum learning rate in columns 2 and 3, respectively, for each case.

Figure 2 shows the second network (type II) with two hidden layers.

In this type, the networks: 3-NN1-NN2-2 for NN1=3 and NN2 = 3, 5, 7, $\cdots$, 25, and 3–NN1-NN2-2 for NN1 = 5 and NN2 = 3, 5, 7, $\cdots$, 25 are considered, where NN1 is the number of neurons in the first hidden layer and NN2 is the number of neurons in the second hidden layer. Each network is trained with the same parameters as in type I to determine the optimum learning rate. Tables 2 and 3 show the performance (MSE) and optimum learning rate in columns 2 and 3, respectively, for each network.

Table 1. Optimum learning rate (for type I) for NN = 3, 5, 7, $\cdots$, 40.

3-NN-2	Performance	Optimum Learning Rate ($\times 10^{-3}$)
3	0.203234	0.3-0.6
5	0.260324	0.1-0.6
7	0.260485	0.1-0.6
10	0.251011	0.1-0.6
15	0.0685703	0.1-0.6
20	0.0424186	0.1-0.6
25	0.104335	0.1-0.6
30	0.03589203	0.1-0.6
40	0.0629649	0.1-0.6

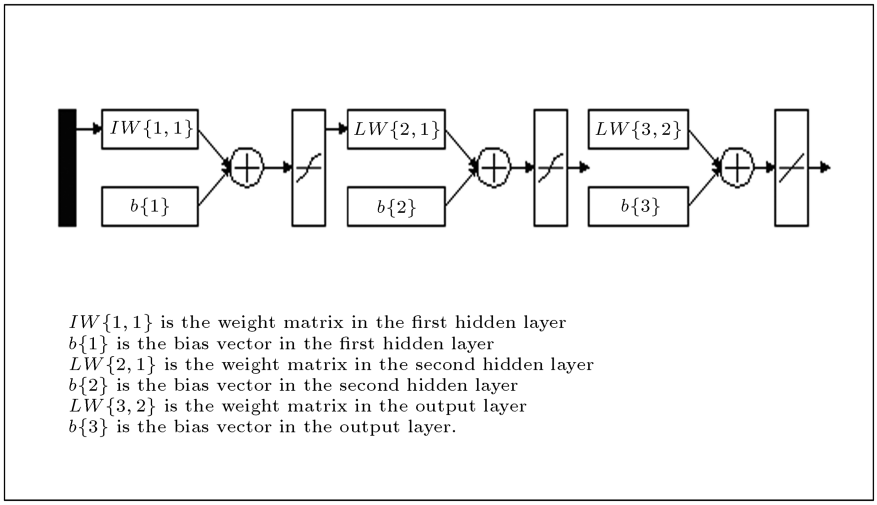

Figure 2. The network in type II.

Table 2. Optimum learning rate in type II for NN1 = 3 and NN2 = 3, 5, 7, $\cdots$, 25.

3-NN1-NN2-2	Performance	Optimum Learning Rate ($\times 10^{-3}$)
3-3	0.258988	0-0.18
3-5	0.499086	0-0.25
3-7	0.250612	0-0.25
3-10	0.869511	0.1-0.8
3-15	0.305026	0.1-0.8
3-20	0.249963	0.1-0.6
3-25	0.249962	0.1-0.65

Table 3. Optimum learning rate in type II for NN1 = 5 and NN2 = 3, 5, 7, $\cdots$, 25.

3-NN1-NN2-2	Performance	Optimum Learning Rate ($\times 10^{-3}$)
5-3	0.358587	0-0.3
5-5	0.145718	0.2-0.6
5-7	0.100179	0.15-0.6
5-10	0.216234	0.15-0.6
5-15	0.268019	0-0.77
5-20	1.07268	0.1-0.6
5-25	0.253335	0-0.35

Table 1 presents that the performance (MSE = 0.04) of the network 3-20-2 is better than other networks of type I. Also, if the performance values in Tables 2 and 3 for type II are compared, it is concluded that the performance value (MSE=0.1) for network 3-5-7-2 is better. As a final result, network 3-20-2 has minimum performance and can be used as a network for the classification of IRS-1D images.

Figures 3 and 4 (see end of the paper) show the graphs of the performance and the variable learning rate for all networks in type I and type II.

In this paper, it is shown that the network 3-20-2 with MSE = 0.04 is the best network for classification of IRS-1D satellite images. Also, the learning rate between ranges 0.001-0.006 will be a suitable rate for any training algorithm for classification of these images. To classify an image with this network, an IRS-1D image, from an area of Iran located in Ghazvin, was selected. Figure 5 shows the original image in RGB color and Figure 6 shows the classified original image in two classes: Urban and suburb with the network.

The result of this network was comprised by the Maximum Likelihood (ML) classification method. Figure 7 is resulted from ML classification with the same classes.

For investigation of an accuracy assessment, 20 random points are selected visually in the image. Table 4 shows the accuracy of NL and NN methods in this study.

The accuracy of the network can be increased if the number of neurons in the input layer are varied, i.e. adding image texture and statistical parameters to the input layer (see [6] for more information). It must

Table 4. Showing the accuracy of the methods.

Method	Overall Accuracy (%)	Overall Kappa
Neural network	79.33	0.6920
Maximum likelihood	74.33	0.6266

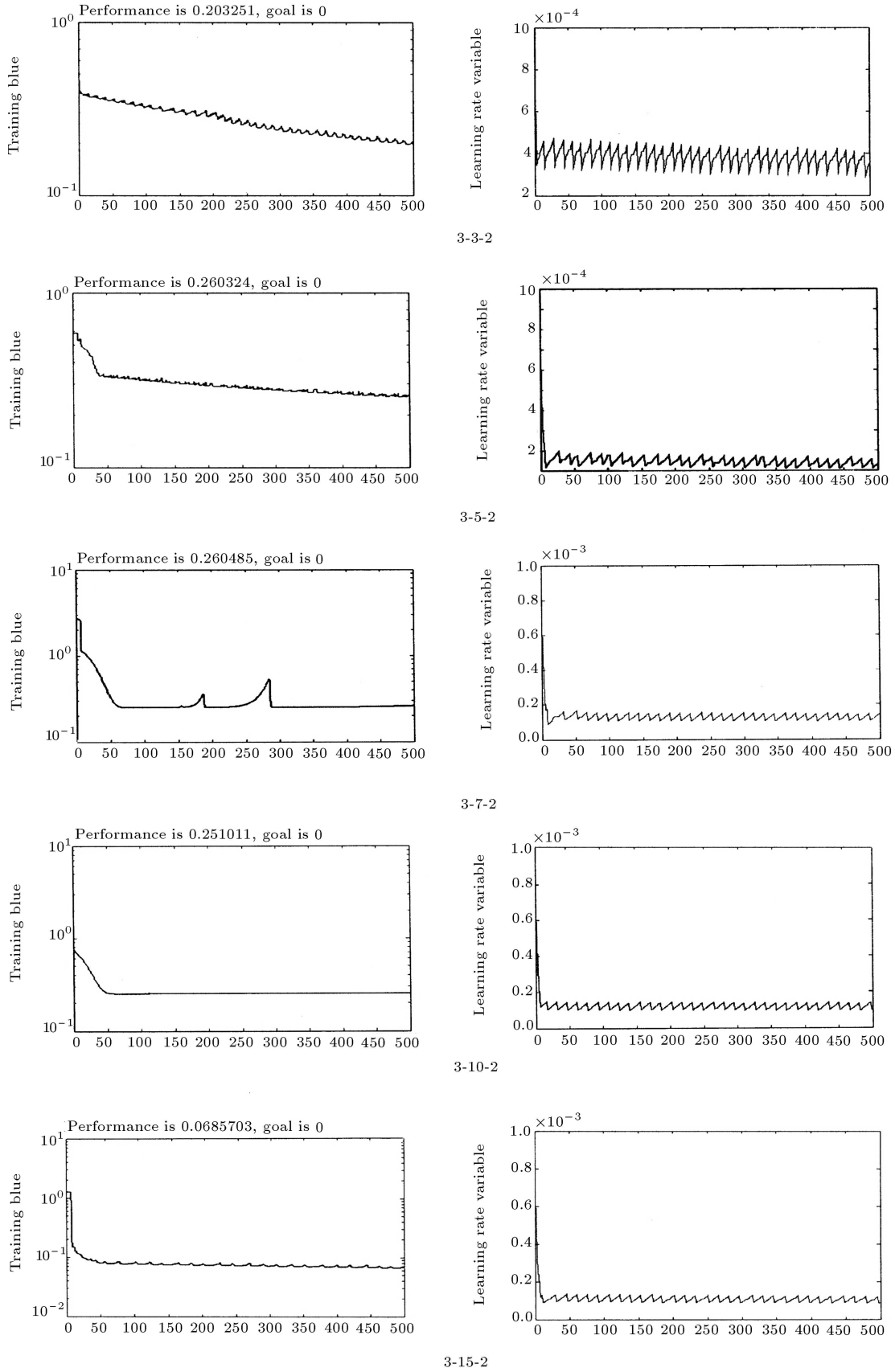

Figure 3. The plots for the performance (left-hand graphs) and the variation of learning rate (right-hand graphs) for the networks in type I.

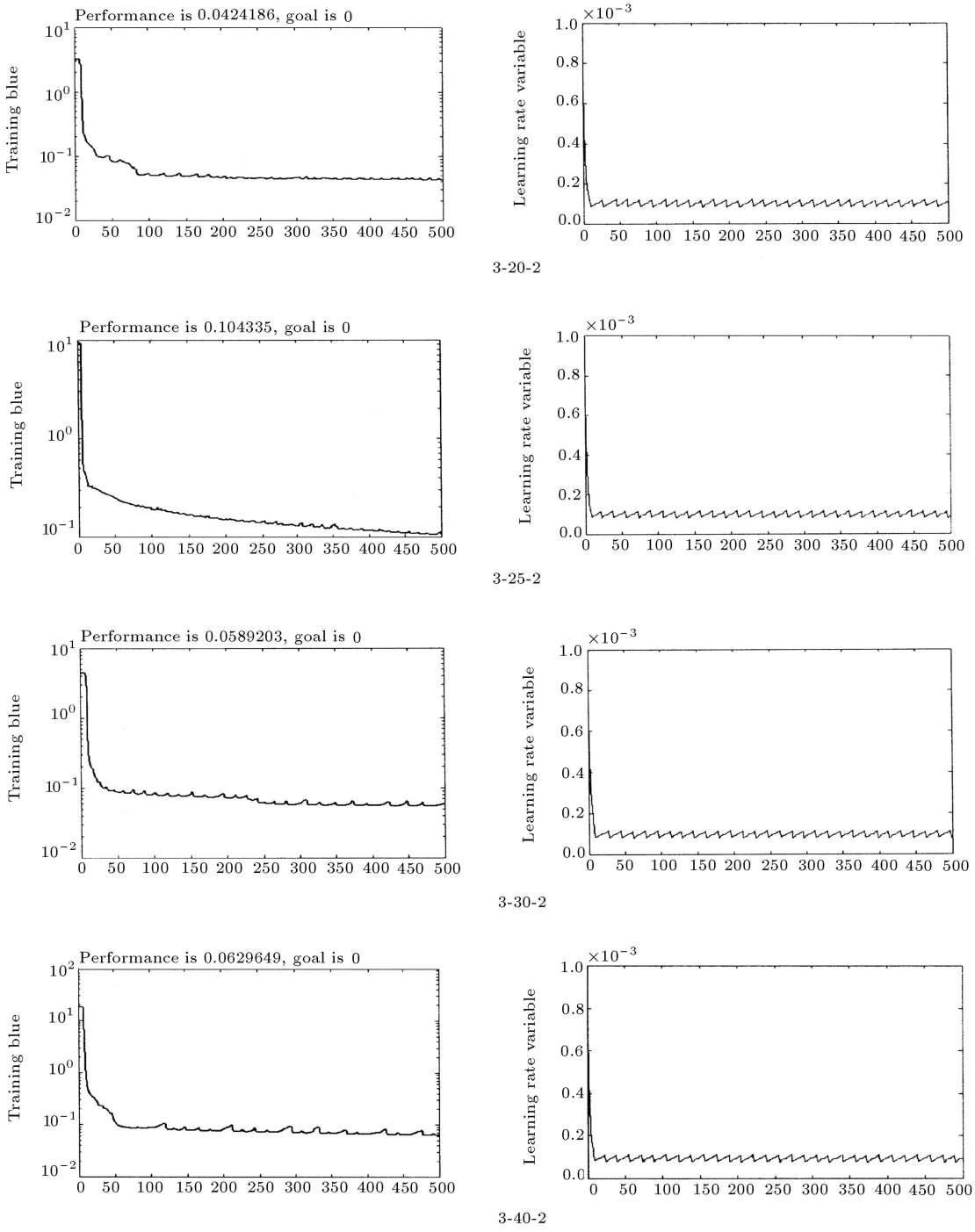

Figure 3. Continued.

be noted that the arrangement of the neural network depends on the topography and the application. The suggestions in this paper can be used in land-use/caver applications.

CONCLUSIONS

One important advantage of neural networks is their ability to learn internal information in the data and recall the knowledge acquired at the learning stage to conduct the classification. In this paper, one of the fast learning algorithms, VLR, was used for training the network. Different networks were used and the optimal learning rate for each network was determined. A network, with one hidden layer consisting of 20 neurons, was obtained as an optimal network for the classification of IRS-1D satellite images. Also, a range for the learning rate, between 0.001-0.006, was

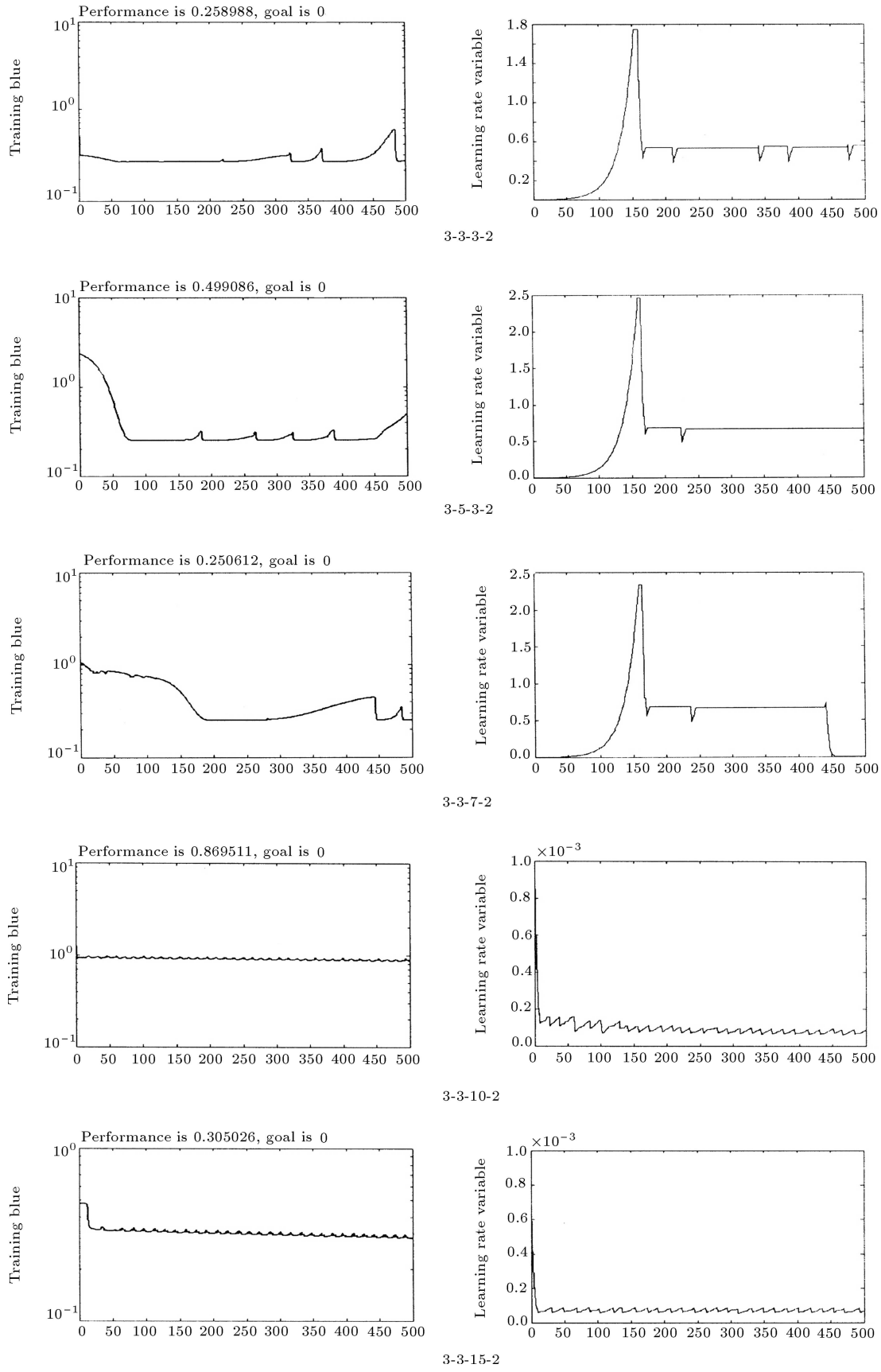

Figure 4. The plots for the performance (left-hand graphs) and the variation of learning rate (right-hand graphs) for the networks in type II.

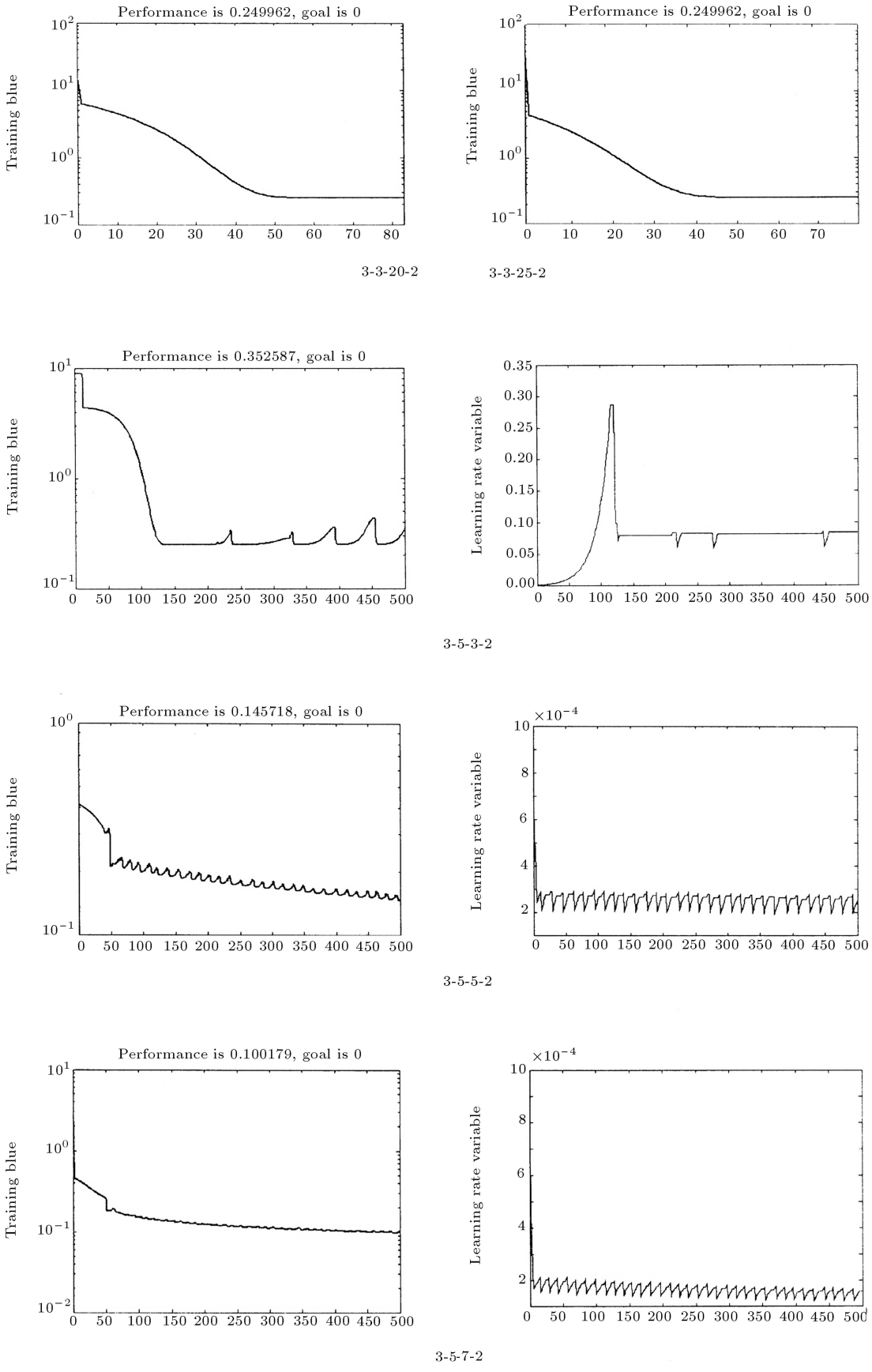

Figure 4. Continued.

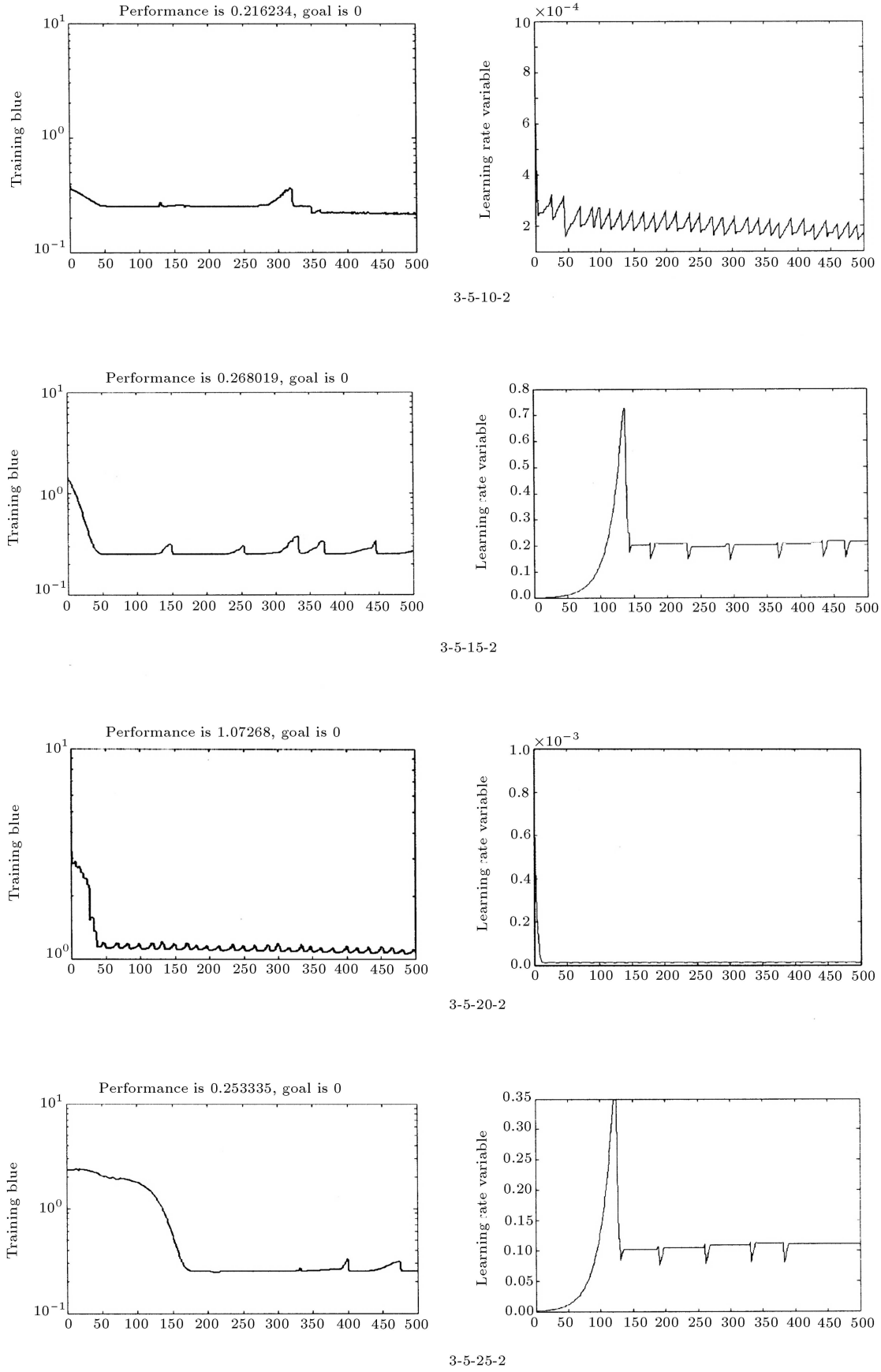

Figure 4. Continued.

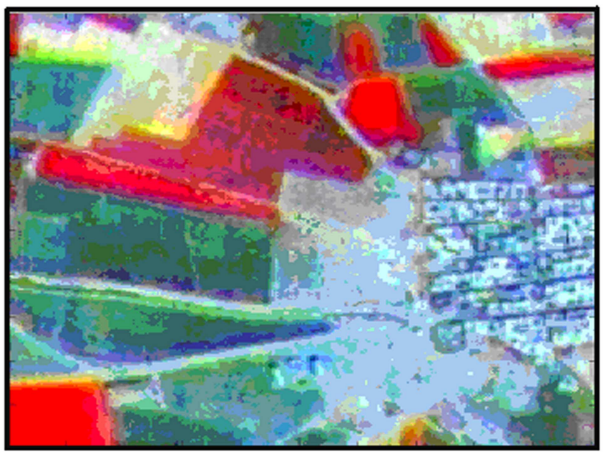

Figure 5. The original image.

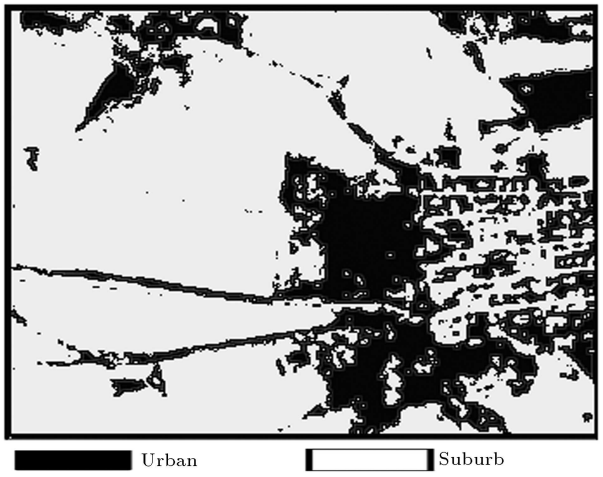

Figure 6. Classified image with the proposed network.

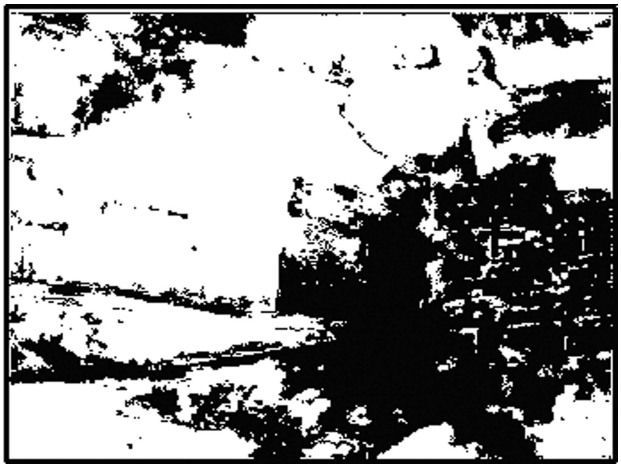

Figure 7. Maximum likelihood classification.

proposed for networks, in which the learning rate is constant during the training process.

ACKNOWLEDGMENT

This research was done within a project, grant no. 621/3/817, in 2003. The Vice Chancellor of Research at the University of Tehran supported this research.

REFERENCES

1. Bendiktsson, J.A., Swain, P.H. and Ersoy, O.K. "Neural network approaches versus statistical methods on classification of multisource remote sensing data", *IEEE Trans. Geosci. Remote Sensing*, **28**, pp. 540-552 (July 1990).

2. Heermann, P. and Khazenie, N. "Classification of multispectral remote sensing data using a backpropagation neural network", *IEEE Trans. on Geoscience and Remote Sensing*, **30**, pp. 81-88 (1992).

3. Hepner, G., Logon, T., Rittner, N. and Bryant, N. "Artificial neural network classification using a minimum training set", *Photogrammetry Engineering and Remote Sensing*, **56**(4), pp. 469-471 (1990).

4. McCelland, J.L. and Rumelhar, D.E., *Parallel Distributed Processing*, **1**, MIT Press, Cambridge, MA, USA (1986).

5. Pao, Y.H. *Adaptive Pattern Recognition and Neural Network*, Addison-Wesley Publishing Company, Inc. (1989).

6. Hosani Aria, E., Amini, J. and Saradjian, M.R. "Back propagation neural network for classification of IRS-1D satellite images", *Joint ISPRS/EARSEL Workshop High Resolution Mapping from Space*, Germany (Oct. 6-8, 2003).

7. Topouzelis, K., Karathanassi, V., Pavlakis, P. and Rokos, D. "A neural network approach to oil spill detection using SAR data", *54th International Astronautical Congress*, Bremen, Germany (29th Sept.-3rd Oct. 2003)

8. Bischof, H., Schnider, W. and Pinz, A. "Multispectral classification of landsat images using neural networks", *IEEE Trans. on Geoscience and Remote Sensing*, **30**(3), pp. 482-490 (1992).