The Comparison of Activation Functions for Multispectral Landsat TM Image Classification

Coşkun Özkan and Filiz Sunar Erbek

Abstract

Neural networks, recently applied to a number of image classification problems, are computational systems consisting of neurons or nodes arranged in layers with interconnecting links. Although there are a wide range of network types and possible applications in remote sensing, most attention has focused on the use of MultiLayer Perceptron (MLP) or FeedForward (FF) networks trained with a backpropagation-learning algorithm for supervised classification. One of the main characteristic elements of an artificial neural network (ANN) is the activation function. Nonlinear logistic (sigmoid and tangent hyperbolic) and linear activation functions have been used effectively with MLP networks for various purposes. The main objective of this study is to compare sigmoid, tangent hyperbolic, and linear activation functions through the one- and two-hidden layered MLP neural network structures trained with the scaled conjugate gradient learning algorithm, and to evaluate their performance on the multispectral Landsat TM imagery classification problem.

Introduction

Artificial neural networks (ANN) are computational systems based on the principles of biological neural systems. These networks have the capacity to learn, memorize, and create relationships among data. Because of having some advantages such as their non-parametric and non-linear nature, arbitrary decision boundary capabilities, easy adaptation to different types of data and input structures, and good generalization capabilities, there has, recently, been an increased interest in using these techniques for many application areas varying from military purposes to medicine, robotic, computer-vision, pattern recognition, and others. Among these advantages, the non-linear nature is introduced into ANN by its activation function, which is one of the important parameters characterizing its architecture.

Neural networks are typically organized in layers which are made up of a number of simulated nerve cells or neurons, often referred to as "processing elements" (PE) or "nodes." The inputs, received by the input layer of nodes, operate on these values and output the result. Typically, the chosen ANN topology specifies connections between the nodes of the input layer and of succeeding layers, and one of a variety of mathematical algorithms are used to determine what the weights of the interconnections should be to maximize the accuracy of the outputs produced. To be presented with new input variables and to generate predictions, neural networks are "trained," i.e., they use previous examples to learn the relationships between the input variables and the predicted variables by setting these weights. The behavior of neural networks, i.e., how they map input data to output data, is influenced primarily by the activation functions of neurons, i.e., how they are interconnected, and the weights of those interconnections.

Artificial neural network (ANN) models have been well proven as useful tools in the analysis of remotely sensed data (Krasnopolsky *et al.*, 1995; Schweiger and Key, 1997; Sunar and Özkan, 2001). Although there are a wide range of network types and possible applications in remote sensing, most attention has focused on the use of MultiLayer Perceptron (MLP) or FeedForward (FF) networks trained with a backpropagation learning algorithm for supervised classification (Atkinson and Tatnall, 1997; Day, 1997; Wilkinson, 1997; Özkan and Sunar, 1999).

In this study, sigmoid, tangent hyperbolic, and linear activation functions through one- and two-hidden layered MLP neural network structures were compared. For the training of the network, the "Scaled Conjugate Backpropagation" algorithm, effective for image classification (Özkan, 2001), was used.

Study Area and Data

The study area is the İkitelli region, a rapidly growing metropolis currently containing a wide range of land-use types located on the European side of İstanbul (Figure 1).

Satellite image data with six spectral bands from the Landsat Thematic Mapper (TM) sensor having a spatial resolution of 30 m (except for the thermal band), and acquired on 15 July 1998, was used for classification.

Methods

MultiLayer Perceptron (MLP) Network

In this study, the most widely used supervised neural classifier, the MLP network, was used. MLP networks are general-purpose, flexible, nonlinear models consisting of a number of processing elements arranged into multiple layers. Generally, they require three or more layers of processing nodes: an input layer which accepts the input variables (e.g., satellite image band values, GIS data, etc.) used in the classification procedure, one or more hidden layers which identify internal structure of the input data, and an output layer with one node per class. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer. Given enough hidden units and enough data, it has

C. Özkan is with the Remote Sensing Division, Geodesy and Photogrammetry Dept., Erciyes University, Kayseri, 38039 Turkey (cozkan@erciyes.edu.tr).

F.S. Erbek is with the Remote Sensing Division, Civil Engineering Faculty, İstanbul Technical University, İstanbul, 80626 Turkey (fsunar@ins.itu.edu.tr).

Photogrammetric Engineering & Remote Sensing
Vol. 69, No. 11, November 2003, pp. 1225–1234.

0099-1112/03/6911–1225/$3.00/0

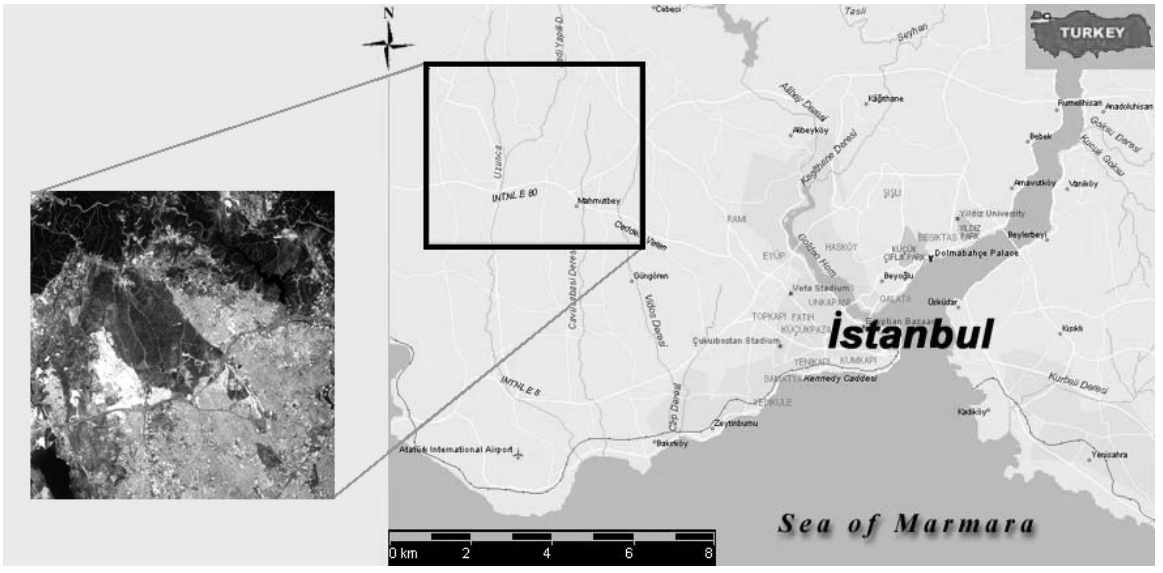
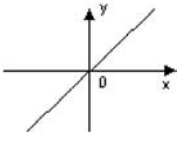
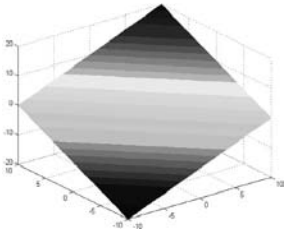
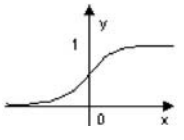
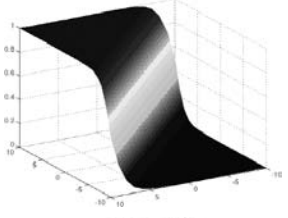
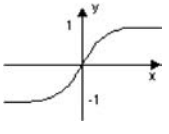
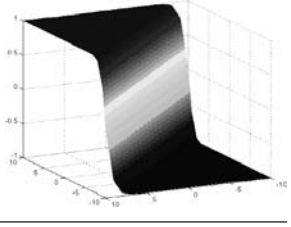


Figure 1. Landsat TM image (1998, band 3) and map of the study area.

been shown that MLPs can approximate virtually any function to any desired accuracy (Haykin, 1994).

All internode connections have associated weights, which are usually randomized at the beginning of the training. When some values pass into a node through interconnections, they are multiplied by the weight associated with these interconnections and summed. Then the activation function determines the output value of this node. Theoretically, any differentiable function can be used as an activation function. However, an activation function having a nonlinear character is so important in order to be able to discriminate the complex relationships that exist in the feature space. While linear functions are particularly used in input and output layers, non-linear activation functions can be used for hidden and output layers. The choice of activation functions may strongly influence the complexity and performance of neural networks. A large number of alternative activation functions have been described in the literature (Duch and Jankovski, 1999). The most commonly used nonlinear functional forms of activation functions satisfying the approximation conditions of ANNs are the sigmoid (logistic) and tangent hyperbolic functions (Zeng, 1999). The characteristics of sigmoid, tangent hyperbolic, and linear activation functions are given in Table 1.

TABLE 1. ACTIVATION FUNCTIONS

Activation Function	Mathematical Equation	2D Graphical Representation	3D Graphical Representation
Linear	$y = x$		
Sigmoid (logistic)	$y = \frac{1}{1 + e^{-x}}$		
Hyperbolic tangent	$y = \frac{1 - e^{-2x}}{1 + e^{2x}}$		

As a logistic function, the most widely used sigmoid function produces the output signal over the 0 to +1 closed range. The values 0 and 1 are obtained for only minus and plus infinities, respectively. As the output values come to close these limits, the derivations of this function decreases. The second most widely used activation function is the tangent hyperbolic function. The tangent hyperbolic function is a bipolar version of the sigmoid function. The tangent hyperbolic function produces the scaled output over the −1 to +1 closed range (Civco, 1993; Kaminsky *et al.*, 1997). The −1 and +1 output values are obtained for minus and plus infinity, respectively. Because the output space of the tangent hyperbolic function is broader, it may be more efficient for the classification performance of the MLP.

In the literature, it has been stated that using the tangent activation function caused faster convergence of the learning algorithms than did the sigmoid function (Bishop, 1995), and the tangent hyperbolic activation function had also been used effectively for the classification of remotely sensed multispectral images (Civco, 1993; Foody *et al.*, 1995; Kaminsky *et al.*, 1997). Haykin (1994) also stated that an asymmetric activation function such as the tangent hyperbolic function instead of the nonsymmetric sigmoid function could perform better for the MLP.

Classification

For image classification, each neuron in the input layer represents one of the input features, such as one satellite image band, while each neuron in the output layer corresponds to one of the classes. With single-date Landsat TM data, for example, there would be seven input nodes, each corresponding to a band of the Thematic Mapper sensor. Other nodes can be used for including ancillary data such as multitemporal spectral patterns, image texture, and elevation and its derivatives. Classifying multisource remotely sensed and ancillary spatial data requires the ability to match large volumes of input pattern data simultaneously to generate categorical information as output. Because the learning and recall stages depend on the linear and nonlinear combination of data patterns, instead of the statistical parameters of the input data, neural networks offer the opportunity to analyze spatial data with different origins and properties simultaneously, without *a priori* assumptions about the distribution for each data type. In fact, neural networks have the ability to learn those distributions, if they exist, from the input data. The one, two, or perhaps more hidden layers consist of a number of processing elements that enable the translation of input data into output information, which, in the present context, is the land-cover classification corresponding to an input pattern. Ideally, each data type will make a unique contribution to the discrimination of land-cover class patterns, therefore enabling the neural network to learn the spectral, spatial, and temporal signature of each class (Civco and Waug, 1994).

It has been reported by several researchers (Lippmann, 1987; Cybenko, 1989) that a single hidden layer should usually be sufficient for most problems, especially for classification tasks (Garson, 1998). Besides, it was stated that the use of a second hidden layer might bring some benefits (Chester, 1990; Hand, 1997). Instead of using a large number of hidden nodes on a single layer, it could be more suitable to use two hidden layers with a smaller number of nodes on each layer (Kavzoğlu, 2001). But it is clearly dependent on the complexity of the data used (Kanellopoulos and Wilkinson, 1997).

While the number of network inputs and outputs is dependent on the problem input and data, the number of hidden layer nodes must be specified by the user. Therefore, it is vitally important to determine the optimum number of hidden layer nodes. Although most of empirical approaches to the determination of the number of hidden layers nodes proposed in the literature (Hecht-Nielsen, 1987; Paola, 1994; Kanellopoulos and Wilkinson, 1997; Gahegan *et al.*, 1999) were a function of the numbers of input and/or output nodes, none of these suggestions has been universally accepted or used (Kavzoğlu, 2001). However, in most applications described in the literature, the results from the experience of individuals using trial-and-error methodology have been used (Bischof and Leonardis, 1998).

As with all supervised classifications, however, the quality of the training stage is of major, if not paramount, importance (Foody and Arora, 1997). In order to match input values to target output values accurately, an MLP's representation of the system is modified through an iterative learning process operating on a set of input values for which target output values are available. However, it is extremely important that a sufficient number of training samples be available to estimate the network free parameters (weights) accurately. A generally accepted guideline is to use at least five to ten times the number of training samples as free parameters (Klimasauskas, 1993; Messer and Kittler, 1998).

For the classification accuracy, the overall classification accuracies, determined from the error matrix by calculating the total percentage of pixels correctly classified and the Kappa coefficient, based on all the elements in the confusion matrix, are determined using both training and test data sets (Janssen and Vanderwel, 1994).

Application

In this study, training and test data were generated using polygon vectors designed for each land-cover type. For the selection of land-cover types, the interactive Iterative Self-Organizing DATA (ISODATA) algorithm was used and 12 classes, consisisting, in all, of 2434 training pixels and 3648 test pixels, were determined with supporting data from 1996 aerial photography (1:5,000 scale). The class codes are given in Table 2.

TABLE 2. CLASS CODES FOR IMAGE DATASET CLASSIFICATION

Class Name	Code	Explanation	# of Training Patterns	# of Test Patterns
Green area-1	1	High density vegetation (forest)	295.00	722.00
Green area-2	2	Normal density vegetation (pasture)	292.00	389.00
Green area-3	3	Low density vegetation	165.00	242.00
Soil-1	4	Bare soil	99.00	103.00
Soil-2	5	Bare soil (cultivated type)	105.00	160.00
Soil-3	6	Stone quarries	152.00	134.00
Highway-1	7	TEM highway (Asphalt)	202.00	131.00
Highway-2	8	Secondary highways (stabilized)	141.00	100.00
Urban area-1	9	Regular urban areas	232.00	323.00
Urban area-2	10	Irregular urban areas	132.00	296.00
Industrial area	11	Organized industrial areas	452.00	477.00
Water	12	Lake and dam lakes	167.00	571.00
		Total Number of Patterns	2434	3648

In the remote-sensing literature, there exist many separability measures to determine whether the signature data are a true representation of the pixels to be classified for each class. Among them, divergence (statistical distance) is based on the derivation of a measure of the difference between all pairs of classes. The divergence analysis either specifies the separability degrees for classes through the all band combination or specifies the band subsets that will maximize the classification accuracy for related classes (Jensen, 1996). However, because the effect of several well-separated classes may increase the average divergence value and make it misleading, the transformed divergence is introduced (Kavzoglu, 2000). As a general rule, if the calculated transformed divergence is greater than 1900, then the signatures can be said to be totally separable in the bands being studied; if it is between 1700 and 1900, the separation is fairly good; and below 1700, the separation is poor (Jensen, 1996). The transformed divergence values that specify the optimal band subsets for classification accuracy are given in Table 3.

In order to show the effects of these activation functions in more general terms, two distinct synthetic data sets were randomly generated. While one group of these random data sets was generated under normal distribution, the other one was generated under uniform distribution. For both synthetic data sets, training and test data were prepared separately. The statistical characteristics of these data sets, i.e., the mean and standard deviation vectors, were also set randomly. As a parallel to the original Landsat TM imagery data, the dimensions of these two groups were selected as six. The number of patterns for each class was set to 125; thus, the total number of patterns for both the normal and uniformly distributed synthetic data sets was 1500.

In order to train the MLP, the Scaled Conjugate Backpropagation (SCB) algorithm, having the advantage of being fully automated, including no user dependent parameters and avoiding a time consuming line-search, was used (Gahegan *et al.*, 1999). The detailed explanations of this learning algorithm can be found in Moller (1993).

While the topology for one-hidden layered networks was set up as 18 neurons in the hidden layer, the topology of two-hidden layered networks was set up as nine neurons in the first hidden layer and 15 neurons in the second hidden layer. In order to determine these numbers for hidden layer nodes, a trial-and-error methodology was used. Therefore, the numbers of total patterns for the one-hidden layered network and two-hidden layered network were expected to remain in the range of 1620 to 3240 and 1845 to 3690, respectively. Hence, it was anticipated that a sufficient number of patterns had been selected for a good generalization of the network.

In the application, both the homogeneous and the hybrid combinations of sigmoid, tangent hyperbolic, and linear functions were used (Table 4).

Although in the literature the 0 to +1 interval is generally used for processing efficiency in the network, in this study input data were scaled into the closed interval −1 to +1. It was thought that, if the −1 to +1 interval were used, the effect of the small differences between adjacent gray values would be diminished because the −1 to +1 interval space was twice that of the 0 to +1 interval space. The output data were coded by the "1 of C" technique, in which each output

TABLE 3. TRANSFORMED DIVERGENCE VALUES

Class	1	2	3	4	5	6	7	8	9	10	11	12
1	0.0	1992.3	1999.8	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0
2		0.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0
3			0.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0
4				0.0	2000.0	2000.0	2000.0	1953.3	2000.0	1997.7	1972.4	2000.0
5					0.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0
6						0.0	1803.5	2000.0	1938.5	1999.9	1992.8	2000.0
7							0.0	1999.7	1923.0	1999.7	2000.0	2000.0
8								0.0	1986.1	1725.7	2000.0	2000.0
9									0.0	1951.6	1995.2	2000.0
10										0.0	1997.6	2000.0
11											0.0	2000.0
12												0.0

TABLE 4. ACTIVATION FUNCTION COMBINATIONS

Number Code	Network Code	1. Hidden Layer	2. Hidden Layer	Output Layer	Type
1	t-t	Tangent Hyperbolic	—	Tangent Hyperbolic	Homogeneous
2	t-s	Tangent Hyperbolic	—	Sigmoid	Hybrid
3	s-s	Sigmoid	—	Sigmoid	Homogeneous
4	s-t	Sigmoid	—	Tangent Hyperbolic	Hybrid
5	t-p	Tangent Hyperbolic	—	Linear	Hybrid
6	s-p	Sigmoid	—	Linear	Hybrid
7	t-t-t	Tangent Hyperbolic	Tangent Hyperbolic	Tangent Hyperbolic	Homogeneous
8	t-t-s	Tangent Hyperbolic	Tangent Hyperbolic	Sigmoid	Hybrid
9	t-t-p	Tangent Hyperbolic	Tangent Hyperbolic	Linear	Hybrid
10	s-s-s	Sigmoid	Sigmoid	Sigmoid	Homogeneous
11	s-s-t	Sigmoid	Sigmoid	Tangent Hyperbolic	Hybrid
12	s-s-p	Sigmoid	Sigmoid	Linear	Hybrid
13	t-s-t	Tangent Hyperbolic	Sigmoid	Tangent Hyperbolic	Hybrid
14	t-s-s	Tangent Hyperbolic	Sigmoid	Sigmoid	Hybrid
15	t-s-p	Tangent Hyperbolic	Sigmoid	Linear	Hybrid
16	s-t-t	Sigmoid	Tangent Hyperbolic	Tangent Hyperbolic	Hybrid
17	s-t-s	Sigmoid	Tangent Hyperbolic	Sigmoid	Hybrid
18	s-t-p	Sigmoid	Tangent Hyperbolic	Linear	Hybrid

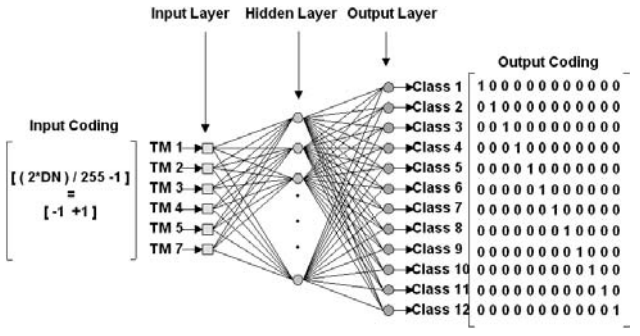


Figure 2. The input-output coding of MLP with one-hidden layer.

neuron resembles a different class (Neural Network—FAQ, ftp://ftp.sas.com/pub/neural/FAQ.html, last accessed 02 June 2003). If the neuron belongs to class 1, then the output target of this neuron will be 1 and the others will be 0. Thus, the input-output topology was constructed as shown in Figure 2.

In the training stage, the input data were entered into the network sequentially. After all input vectors were processed, the total MSE (mean square error) value was computed by comparing the network outputs with the real target values. The network parameters (weights, biases) were than updated according to this error value obtained. This type of learning is also called batch learning.

In order to prevent the overtraining (network available only for training data) or stop the training for optimum network free parameters, the overall classification accuracies for test and training data were computed at each training phase (epoch) for all network structures. Only two factors were taken into account to terminate the training phase: (1) the trend of the overall classification accuracy, and (2) the computational deficiencies of the networks. For example, when s-t and s-p reached the minimum gradient point at a particular epoch, the training stage was ended. It was shown that, although training was halted at a specific epoch, the network would be trained sufficiently for the expected maximum classification accuracy. After all these processes, the network having the test data maximum classification accuracy was selected as the optimum network for classification, and the effects of activation functions were compared through this optimum network.

To compare the activation functions more accurately, the initial free network parameters (weights and biases), initialized randomly in a small range, were set equally for all the networks, i.e., the initial parameters of one- and two-hidden layered networks were the same in their groups.

Results and Conclusions

In this study, different combinations of linear, sigmoid, and tangent hyperbolic activation functions, adapted to one- and two-hidden layered network structures, were evaluated to compare their effectiveness for classification.

For the one-hidden layered network structure, combinations of six different activation functions were tested, and test data performance graphs are given in Figure 3. Several conclusions can be drawn from the results as follows:

- As can be observed, the tangent hyperbolic activation function was better than the sigmoid function for the classification problem.
- The t-p network gave the highest test data overall accuracy in all one-hidden layered networks. The performance of this network was the highest in all applications (Figure 3d).
- In general, the sigmoid activation function gave poor results; in particular, the s-s network was the worst (Figure 3e). The s-s network producers accuracies for both the training and test data containing classes 4, 5, 7, 8, 9, and 10 were zero (Table 5). The training of this network was terminated because the minimum gradient was reached, i.e., the network was stopped before the decision boundaries were sufficiently established from the training data. The analysis of the classified image showed that class 9 pixels were misclassified as class 11. The classified images of the t-p and s-s networks are given in Figure 4.
- In general, the combination of the tangent hyperbolic and linear activation functions was found much more convenient for the classification problem. In the classification map of the t-s network, the classes 12 and 8 were misclassified and the class 10 pixels were allocated to class 12. In addition, it was observed that the linear function in the output layer had a positive effect on the performance of the network.
- On average, the convergence speed of one-hidden layered networks was better than the two-hidden layered networks. The t-t network gave the maximum convergence speed.

For the two-hidden layered network structure, 12 different activation function combinations were tested, and test data performance graphs are given in Figure 5. Several conclusions can be outlined as follows:

- The s-s-s network gave the highest test data overall accuracy for the two-hidden layered networks (Figure 5d). This result was interesting because the homogeneous combination of the sigmoid function in one-hidden layered network has undertrained because of computational deficiencies. The t-t-t network, another homogeneous combination, also gave good results. When output layers were substituted with a linear function, the performance results for both networks (s-s-p and t-t-p) were nearly equal.
- Although its performance was the highest, the classification map of the s-s-s network showed that classes 8 and 10 were incorrectly classified.
- The linear activation function had an increasing effect on the performance for the hybrid combinations of sigmoid and tangent hyperbolic functions (t-s-p, s-t-p), whereas it had a decreasing effect on the performance for the non-hybrid combinations of these functions (t-t-p, s-s-p).
- The worst performance was obtained from the t-s-s network (Figure 5h). The producers accuracy for the test data containing classes 1, 4, and 8 was zero (Table 5).
- As in the s-s network, the training of the t-s-s network had not terminated, although the results showed that the network was undertrained. This case was also seen in the classified image, i.e., most of the high dense green areas (forests) were incorrectly assigned to the water class and also classes 8 and 4 were misclassified.
- The combinations of tangent hyperbolic in the first hidden layer and sigmoid in the output layer (t-s-s, t-t-s) gave the worst results. In addition, the t-t-s classification results showed that class 12 was misclassified as class 6. The classified images of the s-s-s and t-s-s networks are shown in Figure 6.
- On average, the convergence speed of these networks was worse than the one-hidden layered networks. The s-t-s network gave the maximum convergence speed.

The overall and kappa accuracies of all networks for test and training data are given in Figure 7. According to the overall and Kappa accuracies, two-hidden layered networks gave better results than the one-hidden layered networks.

In addition the performance measures, general fitness values comparing the various networks to each other were calculated (Table 6). According to these fitness values, generally the networks numbered 2 (t-s), 4 (s-s), and 14 (t-s-s) had worse fitness values than the others. The maximum fitness value was obtained from networks 3 (s-t) and 6 (s-p). Network 10 (s-s-s), whose performance was the maximum among the two-hidden layered networks, gave the maximum fitness value with network 9 (t-t-p). In addition, network 5 (t-p), whose performance

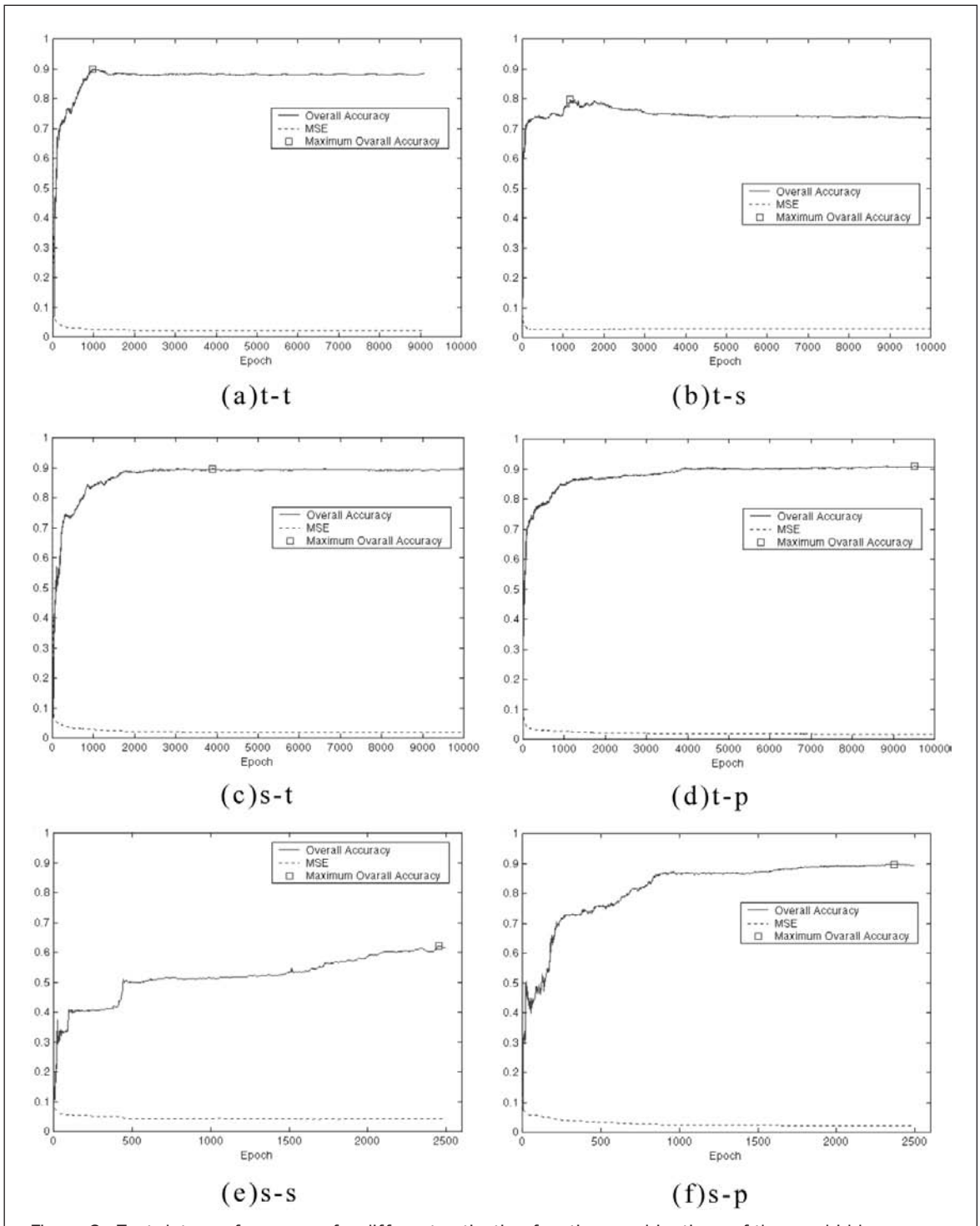


Figure 3. Test data performances for different activation function combinations of the one-hidden layered networks.

was the maximum among all the networks, gave the maximum fitness value with network 12 (s-s-p). Parallel to having higher average overall classification accuracies, the two-hidden layered network's average fitness value is also higher than the one-hidden layered networks.

The same ANN classification process was also performed for the synthetic data. The classification overall accuracies for the training and test data of the normal and uniformly distributed synthetic data are given in Table 7. From this table, the following conclusions can be drawn:

- For the normally distributed data, as expected, most of the networks gave an extremely high accuracy (between 99.8 and 100 percent). But the s-s and t-s-s networks gave worse accuracy results as in the Landsat TM imagery data application. The convergence speed was extremely high for all the networks.
- For the uniformly distributed data, the overall classification accuracies for the test data decreased. Again, the s-s and t-s-s networks gave the worst results. In particular, the s-s network result was extremely low. The s-t-s network also gave a lower classification accuracy. For one-layered networks, the linear function showed an improvement effect for overall accuracies.

TABLE 5. PRODUCERS ACCURACY OF THE TEST AND TRAINING DATA

Network	Class	1	2	3	4	5	6	7	8	9	10	11	12
t-t	Test	93.2	78.7	99.2	83.5	95.6	82.1	97.7	66	93.8	75	87.8	100
	Training	100	96.2	98.8	91.9	100	88.8	87.1	56.7	91.8	77.3	95.4	99.4
t-s	Test	94.6	50.9	0	92.2	97.5	67.9	96.9	79	91.6	79.4	79.7	100
	Training	100	70.2	0	97	100	93.4	92.1	86.5	94	97.7	97.6	100
s-t	Test	90.4	89.2	99.2	93.2	89.4	76.1	96.2	62	94.4	71.6	86	100
	Training	99.7	96.2	99.4	94.9	100	92.1	87.6	66	92.7	81.8	96.9	100
s-s	Test	93.9	90.5	89.7	0	0	96.3	0	0	0	0	66.7	100
	Training	100	99.7	99.4	0	0	98.7	0	0	0	0	44.5	100
t-p	Test	93.8	90	99.2	89.3	92.5	71.6	93.1	71	95	80.7	84.3	100
	Training	100	96.6	99.4	94.9	100	92.1	90.1	68.1	93.1	87.9	96.5	100
s-p	Test	91	86.4	95.9	93.2	93.7	72.4	93.9	64	92.3	77.7	87.2	100
	Training	99.7	96.9	99.4	93.9	100	90.1	89.6	59.6	91.8	84.1	96.2	100
t-t-t	Test	92.9	90	91.7	95.1	95.6	73.1	93.9	81	93.8	77.7	80.7	100
	Training	100	97.6	98.8	93.6	100	91.4	94.1	78.7	91.8	93.2	96	100
t-t-s	Test	90.6	84.6	94.6	0	88.1	82.8	90.8	81	93.2	85.1	87	99.5
	Training	100	99	98.8	0	100	93.4	91.1	55.3	95.7	97.7	98.2	100
t-t-p	Test	90.7	91.8	97.9	93.2	92.5	82.1	97.7	79	93.8	72.3	76.9	100
	Training	99.7	96.6	99.4	96	100	90.1	91.1	77.3	94.8	90.2	95.8	100
s-s-s	Test	98.2	90.2	88.8	90.3	91.9	75.4	95.4	83	92.9	83.4	75.1	100
	Training	100	99.3	98.8	92.9	100	94.7	90.6	88.7	95.7	99.2	99.1	100
s-s-t	Test	91.8	81.5	94.2	94.2	93.7	78.4	93.9	74	92.9	76.4	80.9	100
	Training	99.7	96.2	98.8	91.9	100	90.8	87.1	72.3	93.1	92.4	96.9	100
s-s-p	Test	92	90.5	92.1	94.2	91.9	73.9	95.4	72	94.1	73.3	83.6	100
	Training	99.7	97.9	98.2	93.9	100	91.4	89.1	80.1	94.4	93.2	96.9	100
t-s-t	Test	92.1	92.8	97.5	91.3	93.1	70.1	96.9	74	92.6	76	80.7	100
	Training	100	99	97.6	94.9	100	89.5	95	73.8	93.5	94.7	97.3	100
t-s-s	Test	0	83	93.8	0	90	68.7	90.8	0	93.8	84.5	84.3	99.8
	Training	0	99.3	99.4	0	100	96.1	98	0	98.7	99.2	99.3	100
t-s-p	Test	90.6	88.4	99.6	94.2	93.7	79.9	98.5	75	95.4	75.3	82.8	100
	Training	99.7	97.6	99.4	94.9	100	88.8	86.1	67.4	93.1	85.6	96.2	100
s-t-t	Test	93.8	88.2	90.5	94.2	92.5	70.9	96.9	82	92	78	78.2	100
	Training	100	98.6	98.2	96	100	92.1	94.6	83	94	94.7	97.8	100
s-t-s	Test	94.9	87.4	93.4	88.3	93.7	83.6	91.6	71	91	71.3	82.4	100
	Training	100	97.9	97.6	93.9	100	92.1	89.6	83	90.9	90.9	95.4	100
s-t-p	Test	91.6	88.4	98.3	84.5	93.1	76.9	99.2	72	93.8	74.3	86.4	100
	Training	99.7	97.3	98.2	92.9	100	90.8	90.1	70.2	90.5	88.6	96.5	100

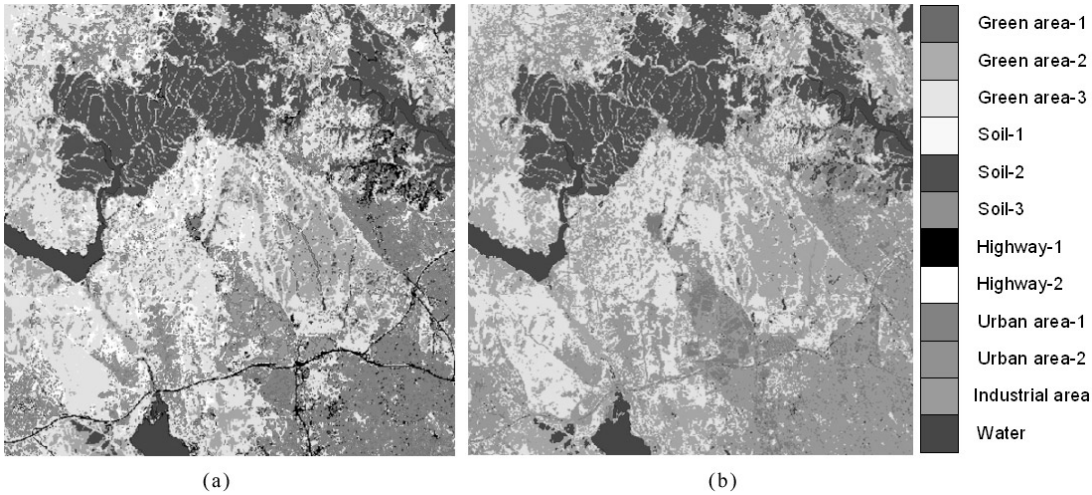


Figure 4. The classified images of (a) the t-p network and (b) the s-s network (color version at www.asprs.org).

The linear activation function in the output layer also showed an improvement effect for the hybrid combinations of sigmoid and tangent hyperbolic functions for the two-layered networks. And convergence speed for this data set was noticeably higher.

As a general conclusion from the results of the ANN classification with the original Landsat TM imagery data and the synthetic normal and, in particular, the uniformly distributed data, the selection of activation functions plays an essential role in the classification accuracy. For the one-hidden layered networks, the tangent hyperbolic activation function is superior to the sigmoid activation function. In particular, a network with the tangent hyperbolic function in the hidden layer

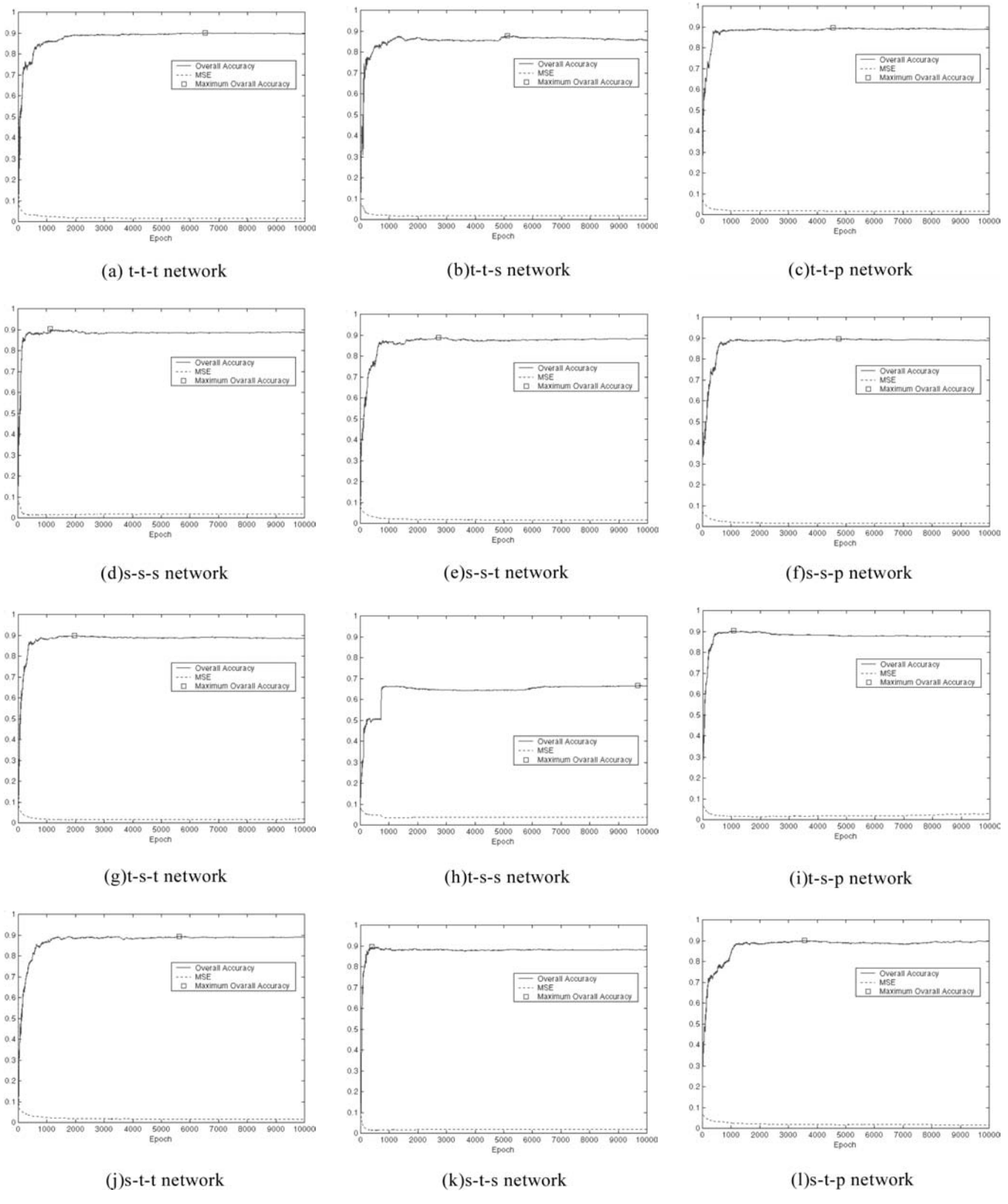


Figure 5. Test data performances for different activation function combinations of the two-hidden layered networks.

and the linear function in the output layer was seen as the optimum choice among these networks. The combination of sigmoid and linear functions also gave good accuracy. For the two-hidden layered networks, while the homogeneous combinations of sigmoid and tangent hyperbolic functions were seen as the optimum choices, the hybrid combinations of sigmoid, tangent hyperbolic, and linear functions were also optimum choices.

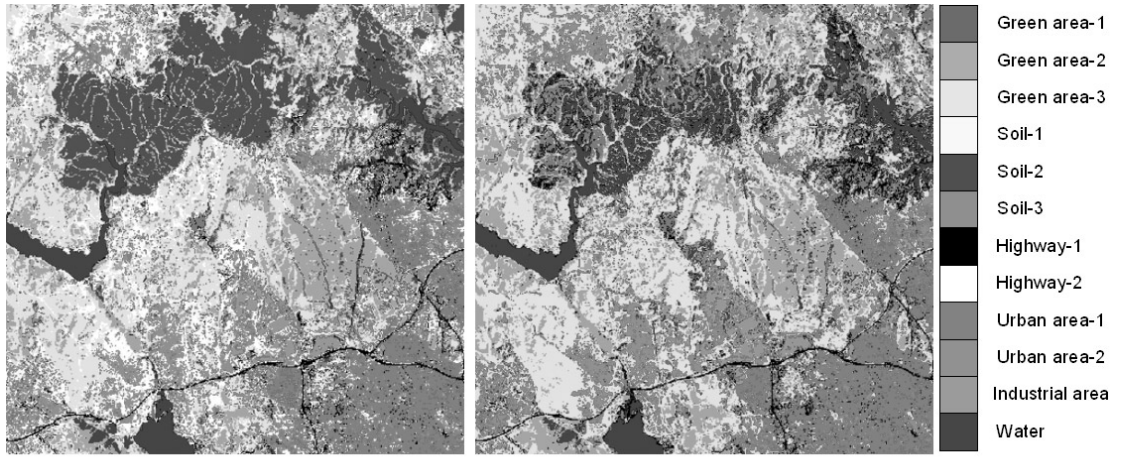


Figure 6. The classified images of (a) the s-s-s and (b) the t-s-s networks (color version at www.asprs.org).

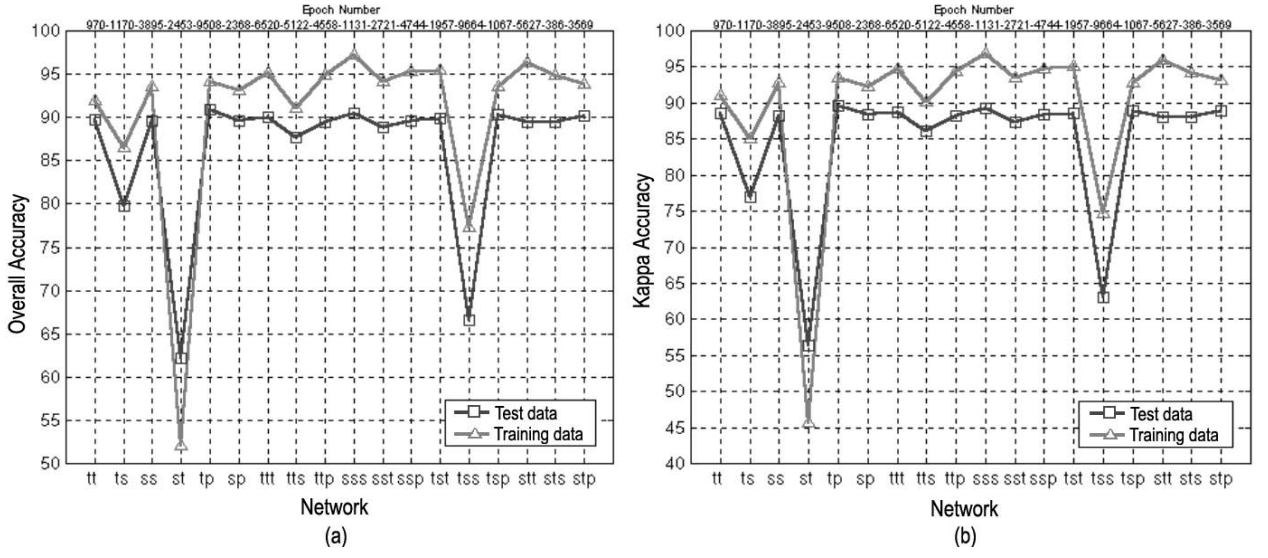


Figure 7. Classification. (a) Overall. (b) Kappa accuracy results.

TABLE 6. GENERAL FITNESS MATRIX FOR ALL NETWORKS

Network	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	100	51.9	85.9	56.4	82.5	85.0	73.9	75.4	78.4	73.6	79.1	80.4	77.8	58.6	80.4	74.1	77.3	77.8
2		100	55.3	34.2	53.4	54.3	59.7	49.6	58.0	57.0	58.3	54.2	58.3	32.2	58.5	58.3	59.5	62.5
3			100	57.2	83.4	88.1	77.7	73.7	83.2	74.0	84.3	82.6	81.7	57.2	85.8	78.1	80.4	83.8
4				100	56.8	58.6	58.4	52.3	52.5	51.9	56.3	56.8	55.2	46.9	58.9	51.6	55.6	57.1
5					100	81.0	78.4	78.7	81.1	76.6	81.2	85.3	78.9	60.5	82.7	75.0	79.5	79.2
6						100	77.8	73.2	80.9	73.2	82.2	79.7	80.5	57.9	83.8	76.6	79.7	83.0
7							100	70.5	76.4	73.9	81.0	78.4	78.0	56.1	81.3	75.9	77.6	78.9
8								100	75.0	75.4	71.1	74.8	75.0	58.0	72.5	70.2	72.3	70.3
9									100	78.2	81.9	80.2	81.8	54.1	81.5	79.8	83.6	81.7
10										100	75.6	76.6	75.8	53.7	74.7	77.5	76.6	73.1
11											100	81.9	81.4	57.7	84.8	79.9	82.4	82.8
12												100	78.9	59.6	81.2	77.9	78.3	76.9
13													100	57.0	83.6	77.6	80.0	80.7
14														100	57.3	53.2	55.3	52.4
15															100	77.0	79.8	85.8
16																100	77.4	76.9
17																	100	82.6
18																		100

TABLE 7. TRAINING AND TEST DATA OVERALL ACCURACIES FOR THE SYNTHETIC DATA

	Normally Distributed			Uniformly Distributed		
Networks	Train	Test	Epoch	Train	Test	Epoch
Sp	100	100	17	76.73	72.4	9340
Ss	0.25	0.25	17	25	24	5602
St	99.9	100	45	76.6	74.27	5641
Tp	100	100	48	76.4	74.2	1069
Ts	100	100	52	83.13	73.13	417
Tt	100	100	56	75.47	73.73	1747
Ssp	100	100	32	78.2	73.13	1062
Sss	100	100	139	82.93	73.53	333
Sst	99.9	100	49	81.6	72.33	5908
Stp	99.8	100	68	80.27	73.67	8613
Sts	100	100	94	79.13	67.67	9038
Stt	100	100	68	80.67	72.93	2339
Tsp	99.8	100	52	81.07	74.13	2563
Tss	50	50	86	84.87	65.80	9946
Tst	99.9	100	74	78.87	73.13	1565
Ttp	99.9	100	65	80.40	73.47	2853
Tts	100	100	24	83.87	73.53	345
Ttt	100	100	68	77.80	74.40	924

Although the sigmoid function is the most common function used in the processing of remotely sensed multispectral imagery, this study showed that the tangent hyperbolic function is also comparable to and can be even superior to the sigmoid function when using the Multi-Layer Perceptron network trained with the Scaled Conjugate Gradient Backpropagation learning algorithm.

References

Atkinson, P.M., and A.R.L. Tatnall, 1997. Neural networks in remote sensing, *International Journal of Remote Sensing*, 18:699–709.

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, United Kingdom, 500 p.

Bischof, H., and A. Leonardis, 1998. Finding optimal neural networks for land use classification, *IEEE Transactions on Geosciences and Remote Sensing*, 36(1):337–341.

Civco, D.L., 1993. Artificial neural networks for land-cover classification and mapping, *International Journal of Geographical Information Systems*, 7:173–186.

Civco, D., and Y. Waug, 1994. Classification of multispectral, multitemporal, multisource spatial data using artificial neural networks, *ASPRS/ACSM Annual Convention & Exposition*, 25–28 April, Reno, Nevada (American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland), pp. 123–133.

Chester, D.L., 1990. Why two hidden layers are better than one, *Proceedings of the Winter 1990 International Joint Conference on Neural Networks* (Maureen Caudill, editor), 10–16 July, Washington, D.C. (Lawrence Erlbaum Associates, Inc.), pp. 265–268.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function, *Mathematics Of Control, Signals and Systems*, 2:303–314.

Day, C., 1997. Remote sensing applications which may be addressed by neural networks using parallel processing technology, *Neurocomputation in Remote Sensing Data Analysis* (I. Kanellopoulos, G.G. Wilkinson, F. Roli, and J. Austin, editors), Springer-Verlag, Berlin, Germany, pp. 262–278.

Duch, W., and N. Jankovski, 1999. Survey of neural transfer functions, *Neural Computing Surveys*, 2:163–212.

Foody, G.M., M.B. McCullagh, and W.B. Yates, 1995. The effect of training set size and composition on artificial neural net classification, *International Journal of Remote Sensing*, 16:1707–1723.

Foody, G.M., and M.K. Arora, 1997. An evaluation of some factors affecting the accuracy of classification by an artificial neural network, *International Journal of Remote Sensing*, 18:799–810.

Garson, G.D., 1998. *Neural Networks: An Introductory Guide for Social Scientists*, SAGE Publications, London, United Kingdom, 208 p.

Gahegan, M., G. German, and G. West, 1999. Improving neural network performance on the classification of complex geographic datasets, *Geographical Systems*, 1:3–22.

Hand, D.J., 1997. *Construction and Assessment of Classification Rules*, John Wiley & Sons, New York, N.Y., 232 p.

Haykin, S., 1994. *Neural Networks: A Comprehensive Foundation*, Macmillan College Publishing Company, Inc., New York, N.Y., 81 p.

Hecht-Nielsen, R., 1987. Kolmogorov's mapping neural network existence theorem, *Proceedings of the First IEEE International Conference on Neural Networks*, 21–24 June, San Diego, California, pp. 11–14.

Janssen, L.L.F., and J.M. Vanderwel, 1994. Accuracy assessment of satellite derived land-cover data: A review, *Photogrammetric Engineering & Remote Sensing*, 60(4):419–426.

Jensen, J.R., 1996. *Introductory Digital Image Processing: A Remote Sensing Perspective, Second Edition*, Prentice Hall, Upper Saddle River, New Jersey, 225 p.

Kaminsky, E.J., H. Barad, and W. Brown, 1997. Textural neural network and version space classifiers for remote sensing, *International Journal of Remote Sensing*, 18(4):741–762.

Kanellopoulos, I., and G.G. Wilkinson, 1997. Strategies and best practice for neural network ımage classification, *International Journal of Remote Sensing*, 18(4):711–725.

Kavzoğlu, T., 2001. *An Investigation of The Design And Use of Feed-Forward Artificial Neural Networks In The Classification of Remotely Sensed Images*, Ph.D. dissertation, University of Nottingham, Nottingham, United Kingdom, 308 p.

Klimasauskas, C., 1993. Neural networks in finance & investing, *Applying Neural Networks* (R.R. Trippi and E. Turban, editors.), Probus Publishing Company, Cambridge, United Kingdom, pp. 47–72.

Krasnopolsky, V.M., L.C. Breaker, and W.H. Gemmill, 1995. A neural network as a nonlinear transfer function model for retrieving surface wind speeds from the special sensor microwave imager, *Journal of Geophysical Research*, 100:11033–11045.

Lippmann, R.P., 1987. An introduction to computing with neural nets, *IEEE ASSP Magazine*, April, 4–22.

Messer, K., and J. Kittler, 1998. Choosing an optimal neural network size to aid search through a large image database, *Proceedings of the Ninth British Machine Vision Conference (BMVC98)*, 13–16 September, University of Southampton, Southampton, United Kingdom, pp. 235–244.

Moller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning, *Neural Networks*, 6:525–533.

Özkan C., 2001. *The Classification of Satellite Imagery Data with Artificial Neural Networks*, Ph.D. dissertation, Istanbul Technical University, Istanbul, Turkey, 223 p.

Özkan, C., and F. Sunar, 1999. The use and effectiveness of artificial neural networks in forest fire classification, *RSS'99 Symposium on Earth Observation: From Data to Information*, 08–10 September, Cardiff, United Kingdom, pp. 767–772.

Paola, J.D., 1994. *Neural Network Classification of Multispectral Imagery*, MSc. dissertation, The University of Arizona, Tucson, Arizona, 169 p.

Schweiger, A.J., and J.R. Key, 1997. Estimating surface radiation fluxes in the arctic from tovs brightness temperatures, *International Journal of Remote Sensing*, 18:955–970.

Sunar, F., and C. Özkan, 2001. Forest fire analysis with remote sensing data, *International Journal of Remote Sensing*, 22(12):2265–2278.

Wilkinson, G.G., 1997. Open questions in neurocomputing for earth observation, *Neurocomputation in Remote Sensing Data Analysis* (I. Kanellopoulos, G.G. Wilkinson, F. Roli, and J. Austin, editors), Springer-Verlag, Berlin, Germany, pp. 3–13.

Zeng, L., 1999. Prediction and classification with neural network models, *Sociological Methods and Research*, 27(4):499–524.

(Received 31 October 2001; accepted 29 October 2002; revised 11 December 2002)