Korean Journal of Remote Sensing, Vol.18, No.3, 2002, pp.155~162

Study on the Effect of Discrepancy of Training Sample Population in Neural Network Classification

Sang-Hoon Lee* and Kwang-Eun Kim**

Kyungwon University*, InterSys, Inc.**

Abstract : Neural networks have been focused on as a robust classifier for the remotely sensed imagery due to its statistical independency and learning ability. Also the artificial neural networks have been reported to be more tolerant to noise and missing data. However, unlike the conventional statistical classifiers which use the statistical parameters for the classification, a neural network classifier uses individual training sample in learning stage. The training performance of a neural network is known to be very sensitive to the discrepancy of the number of the training samples of each class. In this paper, the effect of the population discrepancy of training samples of each class was analyzed with three layered feed forward network. And a method for reducing the effect was proposed and experimented with Landsat TM image. The results showed that the effect of the training sample size discrepancy should be carefully considered for faster and more accurate training of the network. Also, it was found that the proposed method which makes learning rate as a function of the number of training samples in each class resulted in faster and more accurate training of the network.

Key Words : Neural Network, Classification, Training Data Size, Learning Rate.

1. Introduction

Neural networks have been focused on as a robust classifier due to its statistical independency and learning ability(Benediktsson *et al.*, 1990, Civco, 1993). Also the artificial neural networks have been reported to be more tolerant to noise and missing data(Hepner *et al.*, 1990). There have been a lot of researches on the neural network for the supervised or unsupervised classification of remotely sensed image data(Kim and Lee, 1994, Foody *et al.*, 1995, Kim and Kim 1996). Kim and Lee(1994) applied SOM(Self Organized Mapper) for unsupervised classification of airborne remote sensing data. Kim and Kim(1996) used a neural network classifier for the mapping the trophic state of Daechung Reservoir, Korea.

Conventional statistical classifiers use the statistical parameters such as mean, standard deviation, and covariance matrix as representative signatures of a class. The basic assumption of statistical classifiers is that the training data set has a normal distribution. The population of the training samples of each class does not have any effect on the classification accuracy in the conventional statistical classification process. However,

Received 10 May 2002; Accepted 10 June 2002.

a neural network uses the individual training sample for the training of the network instead of using the statistical parameters. Therefore, the discrepancy of the number of the training samples in each class influences the training speed and accuracy.

In this paper, the effect of the discrepancy of the training sample size was analyzed and then a method for reducing this effect was proposed and experimented with Landsat TM image.

2. Neural Network Classifier

A neural network is composed of processing nodes and links. The nodes are organized into layers. The first layer is called input layer which accepts the input training sample. The last layer is the output layer where classified information is retrieved. The layer between the input and output layers is a hidden layer and the nodes in the hidden layer are known as hidden nodes. The links between nodes in successive layers are weight coefficients. The training of a neural network is the process of finding the weight coefficients which classify the input training samples accurately.

Every node in a feed forward network except the input layer performs two basic functions. The one is to collect the activation from the nodes in its previous layer. The other is to set an output activation for the next layer. The activation is defined as the output value of a neural node. For example, w_{ji} is the link between two nodes from layer i to its successive layer j. Each node, except those in the input layer, is an arithmetic unit. It takes the inputs from all the nodes of its previous layer and uses the linear combination of those input values as its net input. For a node in layer j, the net input u_j is

$$u_j = \sum_i w_{ji} a_i \tag{1}$$

where a_i is the output value of a node in layer i. The output of the node in layer j is

$$O_j = f(u_j) \tag{2}$$

where f is an activation function. Any continuous and differentiable function can be used as an activation function. Activation function often takes the form of a sigmoid function as follows.

$$f(u_j) = \frac{1}{1+\exp(-u_j)} \tag{3}$$

The calculated output value O_k of a node in output layer k will be different from the desired value T_k. The goal of training is to find the weights which minimize the difference between of O_k and Tk. Let $E(O_k, T_k)$ denotes the error function, then the weights are changed by

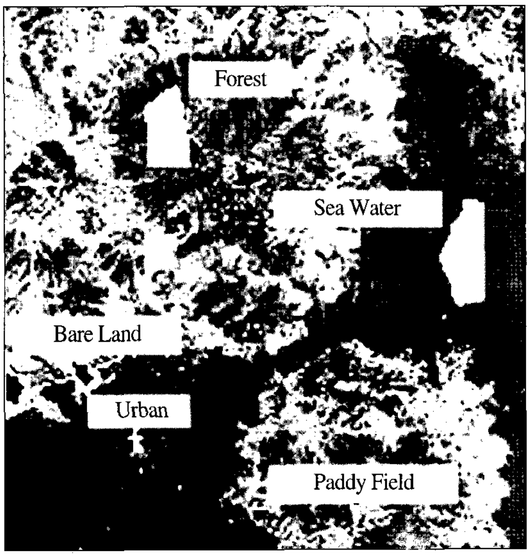
$$\Delta W_{kj} = -\eta \frac{\partial E}{\partial W_{kj}} \tag{4}$$

where η is the learning rate which is very critical for successful training. Eaton and Oliver(1992) suggested a learning rate of 1.5 divided by the square root of the sum of squared number of patterns. Heermann and Khazenie (1992) suggested a learning rate of 10 divided the product of the number of training samples and the total number of nodes in the network. These can only be used as a guideline because η also depends on the complexity of the learning task. If η is too large, the network will oscillate and will not converge. If η is too small, the network may learn very slowly or may not even converge. The Δw_{ij} can be calculated using generalized delta rule(Rumelhart *et al.*, 1986). The error in the output nodes are backward propagated during the calculation so is called back propagation learning algorithm. In data adaptive learning method, the following function is used as a error function.

$$E = \sum_k (T_k - O_k)^2 \tag{5}$$

3. Effect of Discrepancy of Training Sample Population

For the supervised classification of a remote sensing



Class	Number of Pixels
Urban	143
Sea Water	2575
Bare Land	237
Paddy Field	61
Forest	2015

Fig. 1. Landsat TM image and the training sample data for the experiments.

image, training samples for each class should be extracted which can represents the spectral signature of the class. In remote sensing image, there exist a lot of land cover types, and some of the land cover types occupies relatively very small areas compared with the other cover types. Therefore, it is very often that there exists a discrepancy in the number of selected training samples for each class. In the case of existing a large discrepancy, the neural network will learn about the spectral feature of a class which has relatively large number of training samples much more faster than about the other class which has relatively small number of training samples. In this study, these training data size discrepancy effect in a neural network classifier has been analyzed with Landsat TM image.

Fig. 1 shows the Landsat TM image with the training data. Total of 5 land cover types - Urban, Sea Water, Bare Land, Paddy Field, Forest - have been selected for the classification. Urban class has total of 143 training samples and Forest class has total of 2015 pixels of training data. Paddy Field has the smallest number of training samples - 61 pixels. Except the thermal infrared band(band 6 of TM image), total of 6 bands were used for the neural network classification.

The network model was a simple three layered feed forward network and, for the training, the generalized delta rule algorithm was employed. The number of hidden nodes were 12. As a learning rate, the value of 0.1 was found most optimum value after various experiments.

Table 1 shows the classification error matrices generated from the results of the neural network classifications by the conventional training method. After 100 iterations of training, the overall accuracy was 0.42159. The result showed that the network was trained mainly by the Forest class and the other classes was not learned by the network yet. After 500 iterations of training, the overall accuracy of classification was 0.98629. However, the producer's accuracy for the Paddy Field class was only 36.07%. Because there are only 61 samples of training data for the Paddy Field class, the classification accuracy of the Paddy Field class did affect the overall accuracy which was calculated using the total of 5031 training samples for all classes. From this analysis, it is found that when the population of the training data is very different, the higher value of

Table 1. Classification error matrices of the neural network classifications by the conventional training with learning rate of 0.1.

After 100 Iteration of Training

Class	Urban	Sea Water	Bare Land	Paddy Field	Forest	Total	Producer's Accuracy	User's Accuracy
Urban	102	0	0	0	0	102	71.33	100.0
Sea Water	0	4	0	0	0	4	0.16	100.0
Bare Land	0	0	0	0	0	0	0.0	0.0
Paddy Field	0	0	0	0	0	0	0.0	0.0
Forest	41	2571	237	61	2015	4925	100	40.91
Total	143	2575	237	61	2015	5031	Overall Accuracy	0.42159
							The KHAT Statistic	0.04700

After 500 Iteration of Training

Class	Urban	Sea Water	Bare Land	Paddy Field	Forest	Total	Producer's Accuracy	User's Accuracy
Urban	134	0	0	0	0	132	93.71	100.0
Sea Water	0	2575	0	0	0	2575	100.0	100.0
Bare Land	9	0	217	0	1	227	91.56	95.59
Paddy Field	0	0	0	22	0	22	36.07	100.0
Forest	0	0	20	39	2014	2073	99.95	97.15
Total	143	2575	237	61	2015	5031	Overall Accuracy	0.98629
							The KHAT Statistic	0.97594

After 500 Iteration of Training

Class	Urban	Sea Water	Bare Land	Paddy Field	Forest	Total	Producer's Accuracy	User's Accuracy
Urban	140	0	2	0	0	142	97.9	98.59
Sea Water	0	2575	0	0	0	2575	100.0	100.0
Bare Land	3	0	234	0	1	238	98.73	98.32
Paddy Field	0	0	0	49	0	49	80.33	100.0
Forest	0	0	1	12	2014	2027	99.95	99.36
Total	143	2575	237	61	2015	5031	Overall Accuracy	0.98629
							The KHAT Statistic	0.97594

overall accuracy of the classification does not imply the accurate classification. Because the Paddy field was not trained effectively by the network due to the small number of training samples, the producer's accuracy of the Paddy Field class was monitored during the iteration and the result was shown in Fig. 2. The overall accuracy showed the value of 0.95 after 150 iteration of training. However, the producer's accuracy of the Paddy Field class was less than 0.1 until the training is repeated 400 iterations. The accuracy of the Paddy Field class was about 0.8 after the 1000 iterations of training.

From these results, in the case of there exists a large discrepancy in the population of the training samples in each class, it was found that the classes which have very small number of training samples were trained very slowly and often unsuccessfully. It was also found that it is very dangerous to use the overall accuracy as an indicator of the classification accuracy.

4. Population Adaptive Learning Rate

In the previous section, it was shown that the discrepancy of the population of training samples in

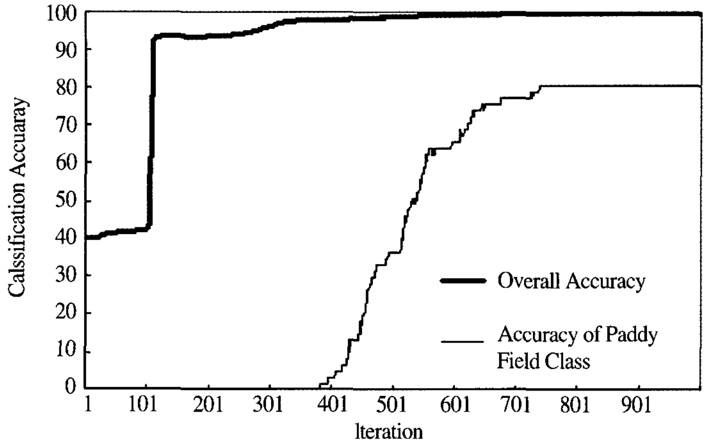


Fig. 2. Overall classification accuracy and the producer's accuracy with learning rate of 0.1 by the conventional training method.

Table 2. Classification error matrices of the neural network classifications by the proposed training with η_0 value of 0.1.

After 100 Iteration of Training

Class	Urban	Sea Water	Bare Land	Paddy Field	Forest	Total	Producer's Accuracy	User's Accuracy
Urban	133	0	6	0	0	139	93.01	95.68
Sea Water	0	2330	0	0	0	2330	90.49	100.0
Bare Land	0	0	0	0	0	0	0.0	0.0
Paddy Field	5	0	12	0	0	17	0.0	0.0
Forest	5	245	219	61	2015	2545	100	79.17
Total	143	2575	237	61	2015	5031	Overall Accuracy	0.89008
							The KHAT Statistic	0.80355

After 500 Iteration of Training

Class	Urban	Sea Water	Bare Land	Paddy Field	Forest	Total	Producer's Accuracy	User's Accuracy
Urban	133	0	0	0	0	133	93.01	100.0
Sea Water	0	2575	0	0	0	2575	100.0	100.0
Bare Land	10	0	236	0	5	251	99.58	94.02
Paddy Field	4	0	1	55	5	61	90.16	90.16
Forest	0	0	10	6	2005	2011	99.50	99.70
Total	143	2575	237	61	2015	5031	Overall Accuracy	0.99463
							The KHAT Statistic	0.99066

After 500 Iteration of Training

Class	Urban	Sea Water	Bare Land	Paddy Field	Forest	Total	Producer's Accuracy	User's Accuracy
Urban	141	1	0	0	0	142	98.60	99.30
Sea Water	0	2574	0	0	0	2574	99.96	100.0
Bare Land	2	0	235	0	3	240	99.16	97.92
Paddy Field	0	0	2	55	1	58	90.16	94.83
Forest	0	0	0	5	2013	2017	99.80	99.70
Total	143	2575	237	61	2015	5031	Overall Accuracy	0.99702
							The KHAT Statistic	0.99481

each class results in slow training. For example, the weights are updated 2575 times for Sea Water class while the weights are updated only 61 times for Paddy Field class after one complete pass of total training samples. Therefore, the neural network had much less opportunity to learn about "Paddy Field" than "Sea Water". In order to reduce this effect of population discrepancy in the training of a neural network, this study suggest to employ an individual learning rate for each class as follows.

$$\eta_i = -\eta_o \frac{P_{min}}{P_i} \quad (6)$$

η_i is the learning rate for class i and η_o is the pre-defined overall learning rate. P_{min} is the minimum number of training samples. In this study, the class which has the minimum number of training samples is the Paddy Field class and P_{min} is 61. P_i is the population of the training samples for class i.

Table 2 shows the classification error matrices generated from the results of the neural network classifications by the proposed training method. After 100 iterations of training, the overall accuracy was 0.89008. After 500 iterations of training, the overall accuracy of classification was 0.99463. Also the producer's accuracy for Paddy Field class was 90.16%. The producer's accuracy for Paddy Field class was only 36.07% in the previous section where the conventional training method was applied. After 1000 iterations of

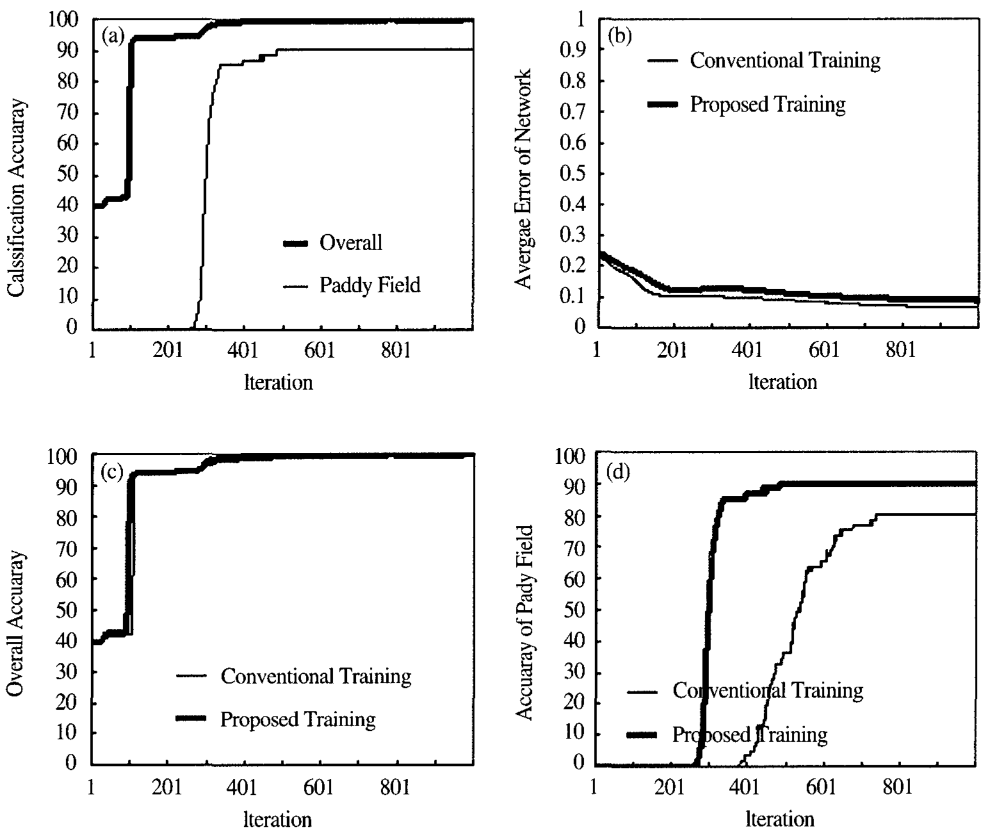


Fig. 3. (a) Overall classification accuracy and the producer's accuracy with learning rate of 0.1. (b) Average error of the network. (c) Overall classification accuracy by the conventional and the proposed training method. (d) Producer's accuracy of the Paddy Field class by the conventional and the proposed training method.

training, the overall accuracy of classification was 0.99702 and the producer's accuracy for Paddy Field class was 90.16%.

Fig. 3(a) shows the overall accuracy and the producer's accuracy of Paddy Field class. The producer's accuracy for Paddy Field class was about 0.9 after 400 iteration while it was 0.8 after 1000 iteration in the previous section. Usually, in the neural network classification the average error of the network is used as an indicator for the training performance. However, Fig. 3(b) shows that the low average error does not always imply a better ability to classify when there is a large discrepancy in the population of training samples The higher value of overall accuracy neither always imply an accurate classification as shown in Fig. 3(c) when there is a large discrepancy. In Fig. 3(d), the producer's accuracy of the class Paddy Field by the conventional training method and the proposed training method were shown together for the comparison.

In Fig. 4, the classified results by the conventional training method was shown. The classified result after 100 iterations of training was shown in (a) and the result after 500 iterations of training was shown in (b). The result shown in (c) is the classified result after the 1000 iterations of training. Fig. 5 shows the classified results by the proposed training method. The results showed that the training by the proposed method reduced the effect of the discrepancy in the population of training samples of each class.

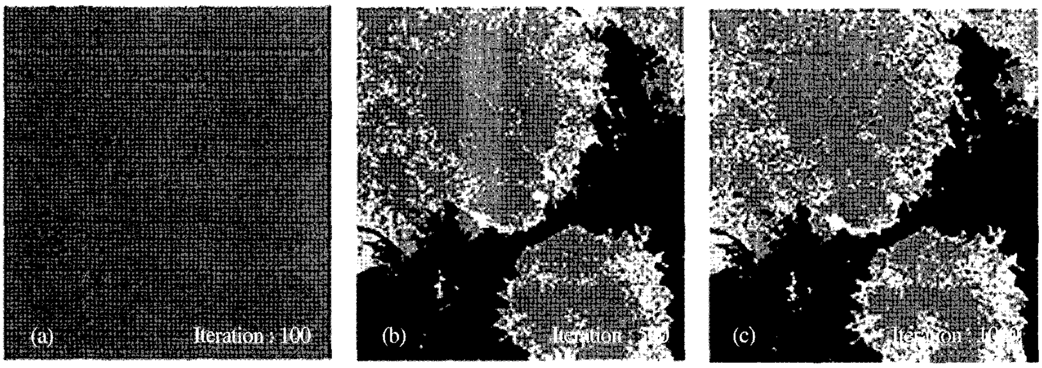


Fig. 4. Classified results by the conventional training method. (a) After 100 iterations of training. (b) After 500 iterations of training. (c) After 1000 iterations of training.

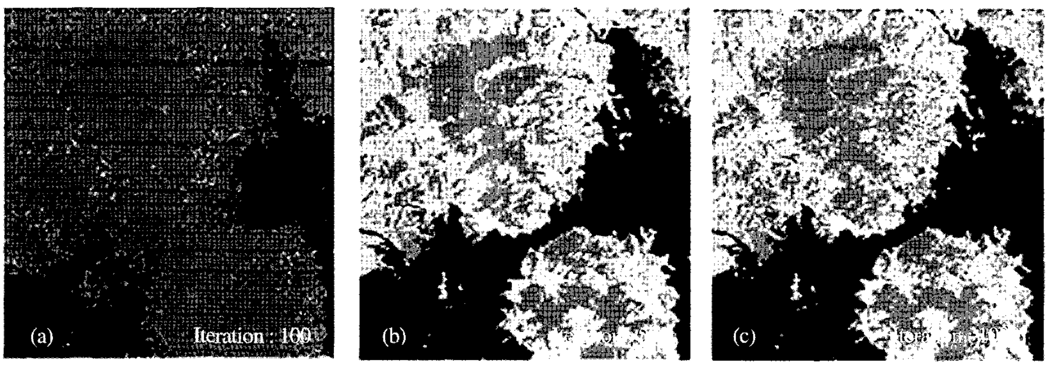


Fig. 5. Classified results by the proposed training method. (a) After 100 iterations of training. (b) After 500 iterations of training. (c) After 1000 iterations of training.

4. Conclusions

A method for reducing the effect of discrepancy in the population of the training samples in a neural network classification was proposed and experimented with Landsat TM image. It was found that the proposed method which makes learning rate as a function of the number of training samples of each class resulted in faster and more accurate training of the network. It was also found that the lower average error or the higher overall classification accuracy do not always imply the more accurate classification when there is a large discrepancy in the population of training samples in each class. The use of a dynamic adaptive learning rate which is adapted during the training according to the producer's accuracy of each class found in each iteration of training is currently under further investigation.

References

Benediktsson, J. A., Swain, P. H., and Erosy, O. K., 1990. Neural network approaches versus statistical methods in classification of multisource remote sensing, *IEEE Trans. Geosci. Remote Sensing*, 28: 540-551.

Civco, D. L., 1993. Artificial neural networks for land-cover classification and mapping, Int. J. GIS, 7: 173-186.

Eaton, H. A. C., Oliver, T. L., 1992. Learning coefficient dependence on training set size, Neural Networks, 5: 283-288.

Heermann, P. D., Khazenie, N., 1992. Classification of multispectral remote sensing data using a back-propagation neural network, *IEEE Trans. Geosci. Remote Sensing*, 30: 81-88.

Hepner, G. F., Logan T., Ritter, N., and Bryant, N., 1990. Artificial neural network classification using minimal training set: comparison to conventional supervised classification, *Photo. Eng. and Remote Sensing*, 48: 123-130.

Kim, K. E., Kim, T. K., 1996. Trophic state mapping of Daechung reservoir by neural network classification of Landsat TM images., *Korean Journal of Remote Sensing*, 12(3): 271-282.

Kim, K. E., Lee, T. S., 1994. Study on application of neural network for unsupervised training of remote sensing data, *J. GIS Assoc. Korea*, 2(2): 65-74.

Rumelhart, D. E., Hinton G. E., and Williams R. J., 1986. Learning internal representations by error propagation, *In Parallel Distributed Processig, Cambridge MA: The MIT Press*, 1: 318-362.