OPTIMIZATION OF TRAINING DATA REQUIRED FOR NEURO-CLASSIFICATION

Xin Zhuang, Graduate Research Assistant
Agricultural Engineering Department
and
Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, IN 47907-1146
(317)494-1187
zhuang@ecn.purdue.edu

D. Fabián Lozano-Garcia, Remote Sensing Application Manager
Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, IN 47907

Bernard A. Engel, Associate Professor
Agricultural Engineering Department
and
Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, IN 47907-1146

R. Norberto Fernández, Manager
Global Resources Information Database, United Nations Environment Programme
Nairobi, Kenya

Chris J. Johannsen, Director
Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, IN 47907

Commission III: Mathematical Analysis of Data

ABSTRACT:

Classification of remotely sensed data with artificial neural networks is called *neuro-classification*. Artificial neural networks have shown great potential in classification of remotely sensed data. The amount of data used for training a neural network affects accuracy and efficiency of the neural network classifier. A neural network was trained separately with 5%, 10%, 15%, and 20% image data from a LANDSAT Thematic Mapper scene, which was acquired 29 July 1987. At a risk level of 5%, the results showed that (a) classifiers NN-5% (neuro-classification with 5% of the image data used for training), NN-10%, and NN-15% did not differ from one another, (b) classifiers NN-15% and NN-20% did not differ from each other, but (c) classifiers NN-5% and NN-10% differed from classifier NN-20%. The training rates were reduced by more than 10 seconds/cycle as we increased the percentage of the image data for training a neural network. Ten percent image data are needed to adequately train a neural network classifier, the classifier provides satisfactory performance.

KEY WORDS: Neuro-Classification, Artificial Neural Networks, Image Processing.

1. INTRODUCTION

Artificial neural networks have been used for image processing and have shown great potential in classification of remotely sensed data. However, the amount of data necessary for training a neural network has not been addressed. Benediktsson *et al.* (1990) classified an image (135 x 131 pixels) using a neural network with the back-propagation learning algorithm. They trained with approximately seven percent of the image data and obtained a training accuracy of 93%. Hepner *et al.* (1990) performed a neuro-classification of a four-band (bands 1, 2, 3 and 4) LANDSAT Thematic Mapper (TM) image (459 x 368 pixels) with four land-cover categories (water, urban, forest and grass). They used 100 (10 x 10) pixels per category for training the neural network classifier. Two LANDSAT TM images were enhanced with a digital land-ownership data and then classified for crop residues (Zhuang *et al.*, 1991; Zhuang, 1990). The neural network classifiers were trained with approximately ten percent of the TM data, and an overall accuracy of more than 90% was obtained for each classification. From these neuro-classifications, one to ten percent of image data were used for the training of the neural networks. Therefore, the amount of data used for the training needs to be investigated.

The objective of this study was to investigate the amount of image data necessary for training a neural network classifier. A LANDSAT TM image was classified with the classifier, and 5%, 10%, 15%, and 20% of the TM data were used for the training.

2. MATERIALS AND METHODS

2.1 LANDSAT TM Data

The LANDSAT TM scene used in this project was acquired 29 July 1987. The scene covered an approximately 10.36 km^2 area (107 x 107 pixels), including sections 3, 4, 9, and 10 located in T28N, R5E of Richland township, Miami County, Indiana, U.S.A. Seven categories of land cover for these sections included *corn, soybeans, forest, pasture, bare soil*, and *river*. The ground observation data were provided for section 9. Aerial photographs from 1987 were available for this study area. The U.S. Geological Survey 1:24,000 topographic map of the Roann, Indiana Quadrangle was also used as a reference.

2.2 Neural Network

The neural network used in this study was configured as a three-layer back-propagation network, including input, hidden and output layers. Adjacent layers were fully interconnected. The input layer was composed of an Nx8 array of binary-coded units, corresponding to N bands (N = 7 in this study) of the 8-bit LANDSAT TM data. Twenty units were assigned to the hidden layer, and six thermometer-coded units in the output layer referred to the six categories of land cover. With thermometer coding, for example, category 4 of the six categories would be represented as 1 in four most-significant bits and 0 in the remaining two bits (4 = 1 1 1 1 0 0).

For the training of a neural network, the TM data were fed to the input layer and propagated through the hidden layer to the output layer, and then the differences between the computed outputs and the desired outputs were calculated and fed backward to adjust the network connections (weights). This process continued until the maximum of the differences was less than or equal to the desired error. Additional details of the network are given in Zhuang (1990).

The neural networks simulator was NASA NETS (Baffes, 1989), which runs on a variety of machines including workstations and PCs. The simulator provides a flexible system for manipulating a variety of configurations of neural networks and uses the learning algorithm of the generalized delta back-propagation. The NETS software was run on SUN SPARC workstations for image classification. Interface routines were developed to make NETS suitable for image classification (Zhuang, 1990).

2.3 Neuro-Classifications

The neural network classified an unknown pixel based on the knowledge learned from a training data set. We trained a neural network separately with 5%, 10%, 15%, and 20% TM data. Therefore, four neural networks with the same configuration were separately trained corresponding to the various percentages of training data. These four neural network classifiers were named *NN-5%, NN-10%, NN-15%*, and *NN-20%*. For the study area, training samples were selected for six land-cover categories based on the corresponding reference information, including the ground observation data, the aerial photographs, and spectral features from individual categories. The training data for category *river* were obtained by an unsupervised classification (clustering) of the portion of the image containing the river.

2.4 Normalization of Classification Results

With the iterative proportional fitting procedure, a contingency table can be standardized to have uniform margins for both rows and columns in order to examine the association or interaction of the table (Fienberg, 1971). The classification results were summarized as a confusion matrix for each classifier. Individual entries of the confusion matrix were divided by the table total, and the result produced a contingency table. The contingency table was normalized with the iterative proportional fitting procedure. The procedure made the row and column margins consecutively equal one. A standard function from SAS software (SAS Institute, 1988a) was used to implement the procedure on contingency tables. Before implementing the iterative proportional fitting procedure, we eliminated zero counts in a contingency table using the method of smoothing with pseudo-counts (Fienberg and Holland, 1970).

2.5 Evaluation of Classifications

Multiple comparisons were made to evaluate the four classifiers, including NN-5%, NN-10%, NN-15%, and NN-20%. By extracting the correct percentages of each classification category from Tables 1 through 4, we produced a performance summary table of classifiers (Table 5). The Tukey multiple comparison method was used for the evaluation of these four classifiers. Any two population means of classifiers will be judged to be different from each other if the difference of the corresponding sample means is greater than the Tukey distance (Mendenhall and Sincich, 1989). The Tukey multiple comparison method is supported by SAS software (SAS Institute, 1988b).

3. RESULTS

The results of the Tukey multiple comparisons (Table 6) provided the overall classification accuracies for the classifiers of NN-5%, NN-10%, NN-15%, and NN-20%. At a risk level of 5%, the results showed that (a) classifiers NN-5%, NN-10%, and NN-15% did not differ from one another, (b) classifiers NN-15% and NN-20% did not differ from each other, but (c) classifiers NN-5% and NN-10% differed from classifier NN-20%. The training rates of the neural network classifiers are illustrated in Figure 1.

4. DISCUSSION

As shown in table 8, we could train the neural network with either 5%, 10%, or 15% TM data because the statistical evaluation showed no significant differences among the three corresponding classifiers at a 5% risk level. The evaluation was done based on individual category accuracies highlighted in Tables 1 through 4. However, interpretations of classified images (Figure 2) show that the two classification results of NN-10% and NN-15% were more uniform than the result of NN-5%. We could not interpret that the classification result of NN-20% differed from the results of NN-10% and NN-15%.

As we increased the percentage of the TM data for training, the corresponding training rate also increased (Figure 1). When we increased 5% to 10% for training, the training rate was decreased by 2 seconds/cycle. When we increased 10% to 15% or 20%, the training rate was reduced by 11 or 16 seconds/cycle, respectively. In this project, the training periods ranged from 100 to 200 cycles.

CONCLUSIONS

Considering the evaluation results and the training rates, we recommand using around 10% TM data to train a neural network, and the performance of a neural network classifier is satisfactory for this level of training.

5. ACKNOWLEDGEMENTS

This research was supported by NASA Research Grant NAGW-1472 and the Laboratory for Applications of Remote Sensing at Purdue University.

6. REFERENCE

Baffes, P.T., 1989. *Nets User's Manual. Version 2.0.* AIS at NASA/JSC, Athens, Geogia, U.S.A., 76 p.

Benediktsson J.A., P.H. Swain, and O.K. Ersoy, 1990. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Trans. Geoscience and Remote Sensing,* GE-28(4):540-552.

Fienberg, S.E., 1971. A statistical technique for historians: standardizing tables of counts. *Journal of Interdisciplinary History*, Vol. 1, pp. 305-315.

Fienberg, S.E., and P.W. Holland, 1970. Methods for Eliminating Zero Counts in the Contingency Tables. *Random Counts in Scientific Work* (G.P. Patil, editor). Pennsylvania State University Press, University Park, Pennsylvania, U.S.A., Vol. 1, pp. 233-260.

Hepner, G.F., T. Logan, N. Ritter, and N. Bryant, 1990. Artificial neural network classification using a minimal training set: comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing,* 56(4):496-473.

Mendenhall, W., and T. Sincich, 1989. *A Second Course in Business Statistics: Regression Analysis.* Dellen Publishing Company, San Francisco, California, U.S.A., 864 p.

SAS Institute, 1988a. *SAS/IML User's Guide, Release 6.03 Edition.* SAS Institute Inc., Cary, North Carolina, U.S.A., 357 p.

SAS Institute, 1988b. *SAS/STAT user's Guide, Release 6.03 Edition.* SAS Institute Inc., Cary, North Carolina, U.S.A., 549 p.

Zhuang, X., B.A. Engel, M.F. Baumgardner, and P.H. Swain, 1991. Improving Classification of Crop Residues Using Digital Land Ownership Data and LANDSAT TM Imagery. *Photogrammetry Engineering and Remote Sensing*, 57(11):1487-1492.

Zhuang, X., 1990. *Determining Crop Residue Type and Class Using Satellite Acquired Data,* M.S.E. Thesis. Department of Agricultural Engineering, Purdue University, West Lafayette, Indiana, U.S.A., 129p.

Table 1. Normalized results for the classification results obtained with the neural network algorithm for the classification using 5% of data for training.

Classification categories	Reference categories					
	Corn	Soybeans	Forest	Pasture	Bare Soil	River
Corn	**0.9465**	0.0026	0.0317	0.0125	0.0011	0.0055
Soybeans	0.0044	**0.9275**	0.0080	0.0454	0.0025	0.0123
Forest	0.0065	0.0442	**0.9389**	0.0102	0.0001	0.0002
Pasture	0.0012	0.0221	0.0006	**0.8970**	0.0789	0.0002
Bare Soil	0.0297	0.0009	0.0010	0.0410	**0.9271**	0.0003
River	0.0002	0.0001	0.0078	0.0001	0.0000	**0.9918**

Table 2. Normalized results for the classification results obtained with the neural network algorithm for the classification using 10% of data for training.

Classification categories	Reference categories					
	Corn	Soybeans	Forest	Pasture	Bare Soil	River
Corn	**0.9528**	0.0080	0.0122	0.0200	0.0006	0.0064
Soybeans	0.0035	**0.9121**	0.0321	0.0412	0.0001	0.0111
Forest	0.0185	0.0169	**0.9408**	0.0234	0.0001	0.0003
Pasture	0.0007	0.0414	0.0029	**0.9145**	0.0404	0.0001
Bare Soil	0.0019	0.0195	0.0005	0.0008	**0.9770**	0.0003
River	0.0082	0.0001	0.0035	0.0001	0.0000	**0.9880**

Table 3. Normalized results for the classification results obtained with the neural network algorithm for the classification using 15% of data for training.

Classification categories	Reference categories					
	Corn	Soybeans	Forest	Pasture	Bare Soil	River
Corn	**0.9607**	0.0059	0.0072	0.0182	0.0012	0.0068
Soybeans	0.0031	**0.9619**	0.0272	0.0075	0.0001	0.0002
Forest	0.0146	0.0064	**0.9561**	0.0222	0.0003	0.0004
Pasture	0.0005	0.0160	0.0010	**0.9116**	0.0586	0.0123
Bare Soil	0.0009	0.0006	0.0052	0.0445	**0.9486**	0.0002
River	0.0077	0.0001	0.0000	0.0001	0.0000	**0.9920**

Table 4. Normalized results for the classification results obtained with the neural network algorithm for the classification using 20% of data for training.

Classification categories	Reference categories					
	Corn	Soybeans	Forest	Pasture	Bare Soil	River
Corn	**0.9692**	0.0043	0.0108	0.0082	0.0002	0.0073
Soybeans	0.0026	**0.9570**	0.0066	0.0212	0.0028	0.0098
Forest	0.0087	0.0016	**0.9875**	0.0013	0.0005	0.0004
Pasture	0.0004	0.0226	0.0009	**0.9505**	0.0255	0.0001
Bare Soil	0.0067	0.0062	0.0001	0.0122	**0.9748**	0.0001
River	0.0001	0.0001	0.0026	0.0001	0.0000	**0.9970**

Table 5. Performance summary of the neural network classifiers corresponding to the trainings with 5%, 10%, 15%, and 20% TM data.

Classification categories	Classifiers			
	NN-5%	NN-10%	NN-15%	NN-20%
Corn	0.9465	0.9528	0.9607	0.9692
Soybeans	0.9275	0.9121	0.9619	0.9570
Forest	0.9389	0.9408	0.9561	0.9875
Pasture	0.8970	0.9145	0.9116	0.9505
Bare Soil	0.9271	0.9770	0.9486	0.9748
River	0.9918	0.9880	0.9920	0.9970

Table 6. SAS output from the multiple comparisons of the classifiers corresponding to the trainings with 5%, 10%, 15%, and 20% TM data.

General Linear Models Procedure
Tukey's Studentized Range (HSD) Test for variable: Y
NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.
Alpha= 0.05 df= 14 MSE= 0.00017
Critical Value of Studentized Range= 4.111
Minimum Significant Difference= 0.0219
Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	METHOD
A	0.9727	6	NN-20%
A			
B A	0.9552	6	NN-15%
B			
B	0.9475	6	NN-10%
B			
B	0.9381	6	NN-5%

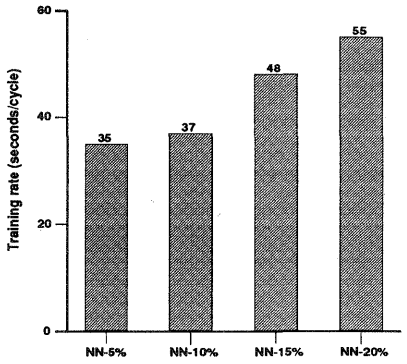

Figure 1. Training rates of the four neural network classifiers.

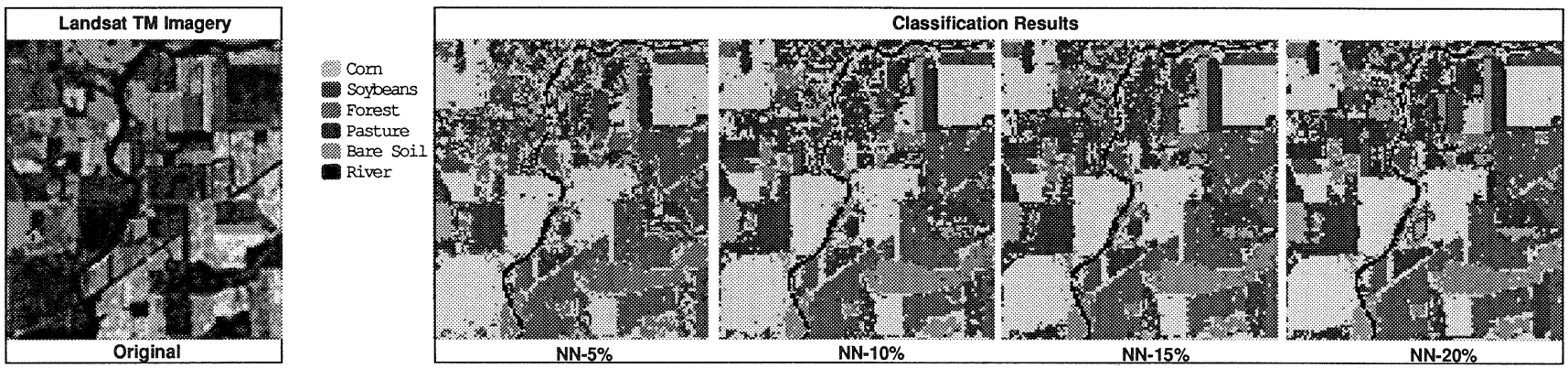

Figure 2. LANDSAT imagery and the classification results obtained with NN-5%, NN-10%, NN-15%, and NN-20%.
(Legends are associated with the classification results.)