MULTITEMPORAL REMOTE SENSING DATA CLASSIFICATION USING NEURAL NETWORK

G.Pasquariello & P.Blonda

Istituto Elaborazione Segnali ed Immagini - C.N.R.
Via Amendola, 166/5 - 70126 - Bari - ITALY
Fax: + 39 80 484311, E-Mail: Pasquariello@astrba.ba.cnr.it

ABSTRACT

In this work a three-layered feedforwrd Neural Netork (N-N), trained with the *backpropagation* algorithm, has been used for classifying a multitemporal Thematic Mapper image. The analisys has been extended to the case where the input data are obtained integrating the satellite image data-set with non Remote Sensed data, as digital elevation data. The aim of the research is to evaluate the effectiveness of a neural network approach with respect to a *Maximum Likelihood* (M-L) statistical one: in order to achieve this goal the overall classification accuracy has been evaluated both for N-N and M-L, comparing the difference in performance on training and test data-sets. Moreover the results obtained with the two different approaches have been related to the statistical measure of the separability on the input training data.

KEYWORDS: Maximum-Likelihood classification, Neural Network, Multitemporal data-set, Digital Elevation Model.

1. INTRODUCTION

1.1 Purpose

The objective of the present work has been to state the applicability of a Neural Network approach to the analysis of Multitemporal Remote Sensed Images integrated with ancillary data. In order to achieve this goal, it has been constructed an integrated data-set, composed by three Thematic Mapper *(TM)* geocoded images and the corresponding Digital Elevation Model (*DEM*). From this data-set, the points belonging to 25 homogeneous fields of known ground truth have been extracted; some of them have been used as training and some as test set. The selected points have been analyzed using both a Maximum-Likelihood classification algorithm and a neural network based approach. The results obtained (preliminary results was presented in (Pasquariello, 1992)) refer to the following experiments: a)classification of each single TM image; b) classification of the multitemporal TM data-set and c) classification of one TM image integrated with ancillary data. In a) the points belonging to each image have been classified separately; in b) the spectral values of all the images have been merged and then classified; in c) the points of the worst classified image have been merged with the elevation values and then classified. In the fourth section the results of the comparison of the two approaches will be presented; in section three the analogies between the N-N and the statistical pattern recognition tools will be shortly recalled together with a description of the N-N architecture used. In the following we review the previous applications of N-N to the Remote Sensing.

1.2 Related Works

The use of Neural Networks for classifying Remote Sensed images, has been recently investigated by several authors. Some of them have discussed the advantages and limits of this technique when compared to conventional statistical methodologies. Evaluations of the classification performance are generally based on parameters such as different numbers of features, different numbers of training samples per class, CPU time and overall accuracy for training and test data. Benediktsson et al (Benediktsson, 1990a,b) have produced experimental results of classification using both Neural Network models and statistical methods. In the first work these authors have analysed Landsat and topographic data, while in the second they have used subsets of very high dimensional simulated HIRIS data. In the experimentation, they have evidenced the major limitations of the conventional Bayesian classification method: the need of having specific assumptions about the probability density functions of the pattern data and the nature of covariance matrix that could be singular in classification of very high dimensional data involving limited training samples. Therefore the authors have proposed three Neural Network models, considering them more appropriate for classification of multisource and multidimensional data, because of their intrinsic non parametric and distribution free nature. The coniugate-gradient linear classifier (CGLC) and the coniugate-gradient backpropagation classifier (CGBP), are modified versions of the conventional delta rule and backpropagation methods. These two models, in fact, are derived using the coniugate-gradient optimization approach for the minimization of the cost function, instead of the most commonly used gradient descent approach. The third proposed model is a hierarchical neural network (PSHNN), involving self-organizing number of stages (SNN), that could be considered as a single particular networks. Similar to a multilayer neural network, exept that in training error-detection phase, this model has shown the best performance among the three neural networks. These three models have been introduced by the authors to overcome the two experimented limitations of conventional Neural Network models: they are timely expensive in training phase and their performance is strongly dependent on the heuristic choice of input parameters. The results of the authors confirm that Neural Networks compared well to statical methods, however, statistical methods have the best performance in terms of speed, accuracy and generalization for classification of the data sets considered in their work.

After comparing the network techniques with several statistical methods, Mulder et al.(Mulder, 1991) have evidenced that training time changes widely on the difficulty of the analyzed problem, on the choice of the training pattern set size and on the choice of input parameters. For this reason Kahny et al. (Kahny, 1991)have focused their attention on the selection of optimal input feature to improve the output performance of a three layered Neural Network applied to the classification of multifrequency polarimetric Sar-Data.

Fortuna et al.(Fortuna, 1991a,b) have applied a Neural Network, based on the Backpropagation Algorithm to classify seismic events and to model the influence of noise sources on same type of geophysical signals. In the first work, they have illustrated some expedients to speed up the training phase of the network.

Lee et al. (Lee, 1990), and Key et al. (Key, 1990) have explored the suitability of a Neural Network technique for the classification of cloud classes. They have demonstrated that very high cloud classification accuracy can be attained with the introduction of spatial information, in the form of textural indices, in conjunction to a Neural Network architecture. The four layer Neural Network classifier used a single-channel multitemporal Landsat-Mss data set, outperformed other non parametric statistical methods applied to the same data, with a smaller number of training data. It also provided important information concerning the significance of each feature vector to the classification of the selected classes.

Also Bishof et al. (Bishoff, 1991) have discussed the utility of including texture information and knowledge based methods for the improvement of classification Neural Networks performance. They have essentially proposed an hybrid system for remote sensing classification problems, having first demonstrated that Neural Networks performs better than a Bayesian classifier in the interpretation of Landsat images.

The interest for the usefulness of Neural Network in the analysis of remotely sensed data have produced the development of new and improved models in more recent works. Kanellopoulos et al(Kanellopulos, 1991) have proposed a hierarchical multiple net system for the classification of two date Spot images in 20 cover classes. However they have verified that the performance of the system is almost the same of that obtained with a single multilayer net, exept for the overall training time, that resulted almost halved.

Kwok et al(Kwok, 1991) have investigated an unsupervised Neural Network model, considered more useful than a supervised one, for automatic and real-time classification of sea ice SAR images.

Benediktsson et al(Benediktsson, 1991) and Tenorio et al (Tenorio, 1990) have proposed new Neural Network architectures for classification of multisource data, based respectively on the statistical consensus theory and on a Self Organizing Structure Algorithm.

2. BACKGROUND

2.1 N-N for Pattern Classification

In many problems Neural Networks seem to offer an interesting alternative approach to traditional statistical ones. In particular, *multilayer perceptron*, trained with the well known *backpropagation* training rule, has been widely used in alternative to Bayes classifiers in evaluating a posteriori class probabilities for classifying stochastic patterns. One of the main problems in Pattern Recognition can be summarized as follows: given an input data X, where X is a N_b-dimensional vector $X \epsilon \Re^{N_b}$, assign it to one of N_c classes $\omega_i\{i = 1, \ldots, N_c\}$ of interest. From a statistical point of view, the best decision rule assigns an input observation X to the most probable class. If we define $P(\omega_i \mid X)$ the probability of the class i given a pixel X (*a posteriori* probability), a Bayes classifier simply implements the following decision:

$$X \,\epsilon\, \omega_i \iff P(\omega_i \mid X) > P(\omega_j \mid X) \quad \forall\, j \neq i,\ j = 1, \cdots N_c \qquad (1)$$

The probability for the i-th class of originating a pixel P of value X is expressed by the *probability distribution function (pdf)* $P(X \mid \omega_i)$ times the *a priori* probability $P(\omega_i)$ of occurrence for that class. Applying Bayes theorem we obtain:

$$P(\omega_i \mid X) \quad = \quad \frac{P(X \mid \omega_i)P(\omega_i)}{P(X)} \qquad (2)$$

Substituting Eq. 2 in Eq. 1, we obtain the well known Maximum-likelihood classification rule:

$$X \,\epsilon\, \omega_i \iff P(X \mid \omega_i)P(\omega_i) > P(X \mid \omega_j)P(\omega_j) \\ \forall\, j \neq i,\ j = 1, \cdots N_c \qquad (3)$$

In the analysis of multiband Remote Sensed images the *pdf* associated to a class is normally assumed to be a multivariate normal density function: then a set of point representative of each class (training set) is used to calculate the N_b values of the mean and the $N_b(N_b+1)/2$ elements of the covariance matrix for that class. These values are used in Eq. 3 to classify each unknown input vector X.

A feed-forward neural network with N_b input units and N_c output neurons implements a mapping $\mathcal{F}$: $\Re^{N_b} \longrightarrow \Re^{N_c}$ by means of a set of P training examples $\{(X_1, Y_1), \ldots, (X_P, Y_P)\}$, where X_i and Y_i are N_b-ples and N_c-ples of values, respectively. The mapping function found by the network minimizes the quantity $E(\mid Y(X) - \mathcal{F}(X) \mid^2)$, where $E(f)$ is the expectation value of f; from this point of view, it is possible to say that the network could be used to find out an estimation of $E(Y \mid X)$, i.e. an estimate of the a posteriori probability of Eq. 1.

2.2 Neural Architecture

The architecture studied in this work is a *feed-forward NN* in which units are arranged in layers: all connections have the same direction and are allowed only between contiguous layers which are, in general, fully connected. Each neuron k, belonging to the layer j, receives as input the outputs of all the neurons l in the $j-1$ layer, which it is connected to by a synaptic strength represented by a real number $W_{lk}^{(j)}$. On the other hand, the neuron k is also connected, with strengths $W_{ki}^{(j+1)}$, to all the neurons i of the layer $(j+1)$, to which it sends the outputs

$$Y_k^{(j)} = \quad g(X_k^j, \theta_k^j) \qquad (4)$$

where $g(x, \theta)$ is the transfer function, usually defined as

$$g(x, \theta) \quad = \quad \frac{1}{1 + \exp(-\beta(x - \theta))} \; . \qquad (5)$$

θ_k^j is the threshold associated with the neuron k in the layer j , β is a parameter (*gain*) describing the slope of g,

$$X_k^j \quad = \quad \sum_l W_{lk}^{(j)} Y_l^{(j-1)} \; . \qquad (6)$$

This sum runs over the neurons in the layer $(j-1)$. A *N-N* with N_i input and N_o output neurons is trained with a supervised learning algorithm in which the net outputs $Y(X)$ is compared to the known expected answers $T(X)$. The **back-propagation algorithm** is a supervised learning algorithm, based on minimizing an error function E, by using a gradient descent method. In particular, if $(T_1^\mu, \ldots, T_{N_o}^\mu)$ is the desired output pattern, where $\mu = 1, \ldots, P$ and P is the total number of examples in the training set, $(Y_1^\mu, \ldots, Y_{N_o}^\mu)$ is the actual *NN* output in reply to the input pattern $(X_1^\mu, \ldots, X_{N_i}^\mu)$ and $\mathbf{W}$ is the matrix of the interconnection weights, the *error function* can be written as:

$$E[\mathrm{w}] = \frac{1}{2} \sum_\mu^P \sum_i^{N_o} (Y_i^\mu - T_i^\mu)^2 \; ; \qquad (7)$$

if the weights are updated after all patterns have been presented to the *NN* inputs (*batch learning*), or

$$E[\mathrm{w}] = \frac{1}{2} \sum_i (Y_i^\mu - T_i^\mu)^2 \qquad (8)$$

if the weights are updated after each pattern has been presented to the *NN* inputs (*incremental learning*).

In both cases, the gradient descent algorithm consists of changing each W_{ij}^m by an amount ΔW_{ij}^m proportional to the gradient of $E[\mathrm{w}]$, so as to slide downhill on the surface defined by the error function:

$$W_{ik} \quad \rightarrow \quad W_{ik} \; + \; \Delta W_{ik} \; , \qquad (9)$$

with

$$\Delta W_{ik} \quad = \quad -\eta \frac{\partial E[\mathrm{w}]}{\partial W_{ik}} \; + \; \alpha \Delta W_{ik}^{(old)} \; . \qquad (10)$$

The *momentum term* (Rumelhart, 1986) $\alpha \Delta W_{ik}^{(old)}$ is introduced to give each connection W_{ij} some momentum so that it tends to change in the *average downhill* direction, avoiding sudden oscillations.

For the simulations described in the next section, We have always used one hidden layer with $N_h = 2N_i + 1$ which should be sufficient according to a general theorem (Kolmogorov, 1957). The number N_i of input units corresponds to the number of bands N_b considered for each experiment, whereas the number of neurons in the output layer is given by the number of the selected classes (in all experiments $N_o = 7$). Fig. 1

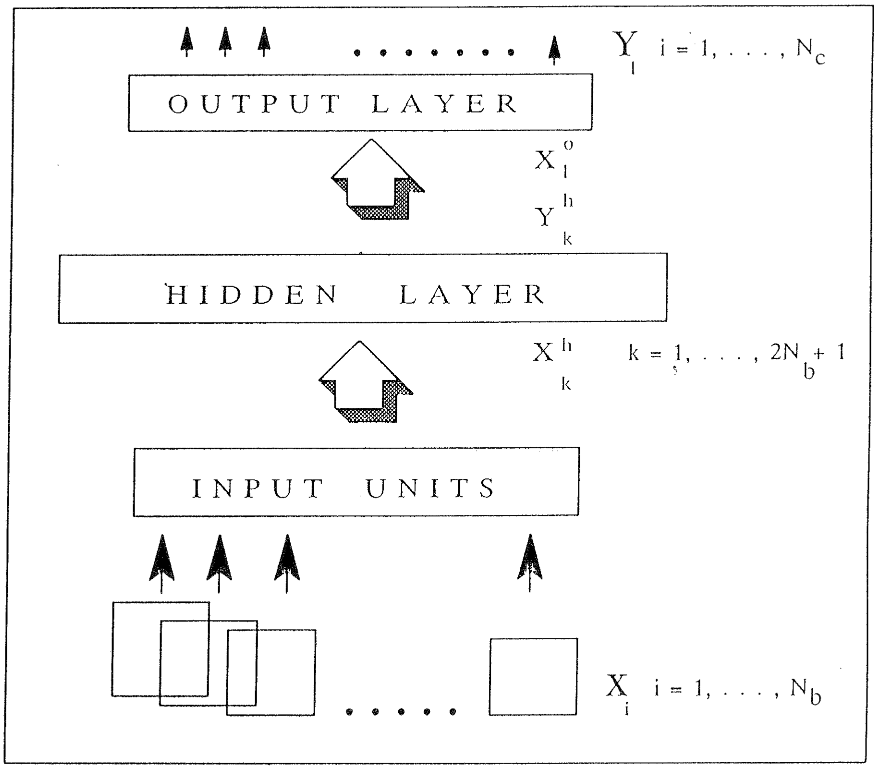
OUTPUT LAYER
Y_i $i = 1, \ldots, N_c$
X_i^o
Y_k^h
HIDDEN LAYER
X_k^h $k = 1, \ldots, 2N_b + 1$
INPUT UNITS
X_i $i = 1, \ldots, N_b$

Figure 1: Three-layer feedforward Neural Network architecture

gives a graphical representation of the used network. In our simulations we have used an incremental learning strategy with a *momentum parameter* $\alpha = 0.9$, a *strength parameter* $\eta = 1$ and a *gain factor* $\beta = 1$.

3. DATA ANALISYS

3.1 Input Data

As input data we have considered three Landsat-TM images and a Digital Elevation Model, obtained digitizing contour lines form 1 : 50,000 topographic map, with 25 m contour interval (the height error, computed as rms error on a set of fiducial points, is about 15 m). The study area is a region of Southern Italy, fully covered by Landsat-d 31.188 imagery (for the characteristics of that area see (Blonda, 1991): the spring image was acquired on April, the summer on July and the autumn one on October. Only the six bands at higher spatial resolution have been considered, for each image. The class selection was based upon information derived by visual interpretation of aerial photographs and by local inspection. The classes selected for the analysis are the following: 1)*Bare soil*; 2)*Urban areas*; 3)*Pasture*; 4)*Coniferous reafforestation*; 5)*Olive groves*; 6)*Vineyards*; 7)*Cropland.* In order to perform the comparison between the two approaches, a set of near 8,000 pixels has been extracted from the data-set, corresponding to 25 homogenous fields of known land coverage. The percentage of occurrence of the selected classes are 3.5; 4.1; 22.2; 2.1; 30.7; 8.7 and 28.7 for class from 1) to 7), respectively. As training set we used 800 pixels random selected (corresponding to the 10% of the whole data set); the remaining 90% of points has been used for testing. The same points have been used both for training Neural Network synaptic strengths and to extract the statistical features (mean vector and covariance matrix) associated to each ground class.

3.2 Results

The obtained results refer to the following experiments:

a) : The points of each image have been separately analyzed. The network for each image is composed of a 6 − 13 − 7 neurons. The overall percentage of correct classification P is reported in table 1. Figs. 2, 3, 4 show P versus iteration number for April, July and October image,

	Overall Classification accuracy (%)			
Date	*training*		*test*	
	M-L	N-N	M-L	N-N
April	91.0	95.0	89.0	90.0
July	93.5	98.0	91.3	95.3
October	91.0	96.4	88.4	90.0

Table 1: Results of the *Single Image* experiment. The ratio between test and training points is $9 : 1$. The N-N consists of a $6-13-7$ architecture.

	Overall Classification accuracy (%)			
	training		*test*	
	M-L	N-N	M-L	N-N
3-Date	98.9	99.8	95.5	98.0
Oct+Dem	92.7	98.3	91.0	94.5

Table 2: Results of *Multiple Image* and multisource *Single image plus DEM* experiments.

respectively (in the figures, the constant values refer to the M-L performances).

b) : Each point is considered as a 18-dimensional vector, i.e. the three images are classified together. In this case for the N-N we consider a $18-37-7$ architecture. The comparison between statistical and connectionistic approach is reported in raw 1 of table 2.

c) : The spectral values of the October image are merged with the value of DEM. The results of the $7-15-7$ N-N are shown in table 2. Figs. shows P versus iteration number for this data-set.

The statistical separability of each input training data set has been evaluated by computing the average *Jeffries-Matusita (JM)* distance (Swain, 1078):

$$JM_{ave} = \sum_{i=1}^{N_c} \sum_{j \neq i}^{N_c} P(\omega_i) P(\omega_j) JM_{ij} \qquad (11)$$

where JM_{ij} is the distance between the class i and j, defined as follows:

$$JM_{ij} = \left\{ \int_X \left[\sqrt{P(X \mid \omega_i)} - \sqrt{P(X \mid \omega_j)} \right]^2 dX \right\}^{1/2} \qquad (12)$$

Data-set	J_{ave}	$P_E(\%)$
April	1.348	95.4
July	1.370	96.9
October	1.284	91.2
3-Date	1.413	99.9
Oct+Dem	1.340	94.9

Table 3: JM_{ave} distances and corresponding expected probability of correct classification . The values are referred to training data.

The value JM_{ave} gives an estimate of the lower bound of the expected probability (P_E) of correct classification for the corresponding data-set. Table 3 shows the value of JM_{ave} and the corresponding P_E lower bound for each experimental analyzed training set. Comparing the values of P_E with the results obtained applying M-L to training data (second column in table 1 and table 2), the actual performance is slightly worse then theoretical one, for all experiments. These differences give a measure of the error in estimating the *pdf* associated to each class, when using a multivariate normal density function assumption. From a different point of view, they give a measure of the improvement expected when using a N-N approach.

3.3 Conclusions

The results shown in the previous section confirm that the performances of the Neural Network approach are slightly better if compared with a statistical approach. This conclusion is easily explained because N-N overcome the main limitation of the Bayesian approach, i.e. the need of having a specific probabilistic model (the *a priori* assumptions) for describing input data, whereas a N-N computes directly the *a posteriori* probabilities by means of a least squares approach on the input data.

ACKNOWLEDGMENT

The experimental part of the research described in this paper has been carried out at the Remote Sensing Laboratory, Physics Department, Bari University, directed by prof F.Posa, under contract with the Italian Space Agency. The authors would like to thank Dr. N.Cufaro Petroni for his cooperation, prof. C.De Marzo and prof. G.Nardulli for their helpful comments.

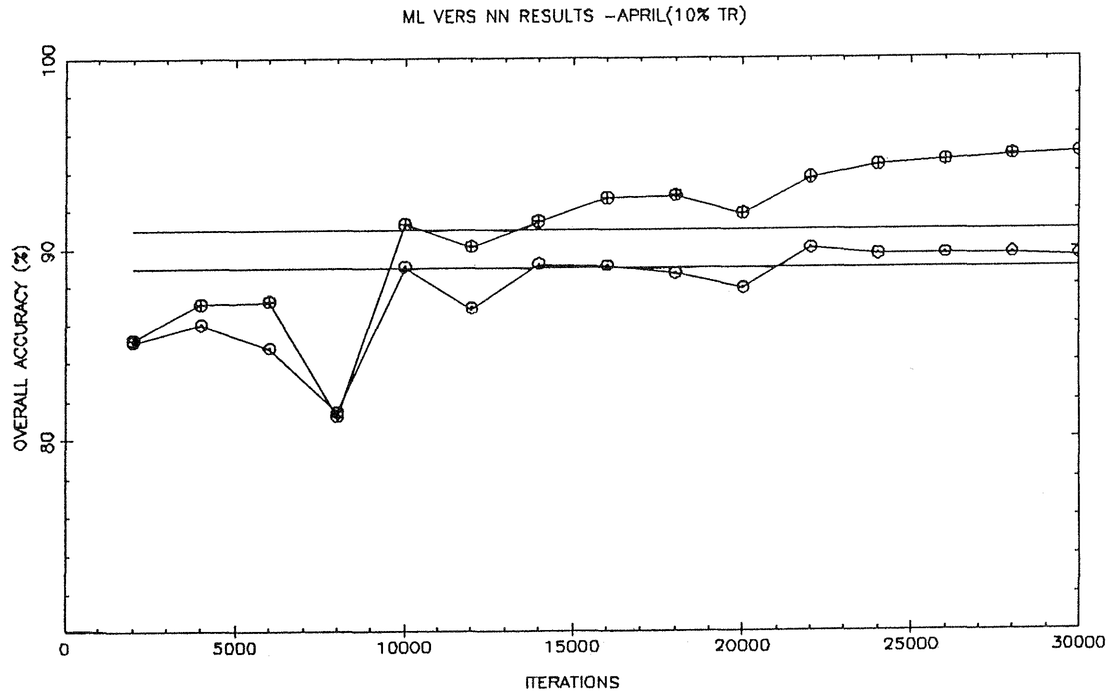

Figure 2: April Image. Classification accuracies (%) using N-N and M-L classifiers. ⊙ Training data value ⊕ Test data values

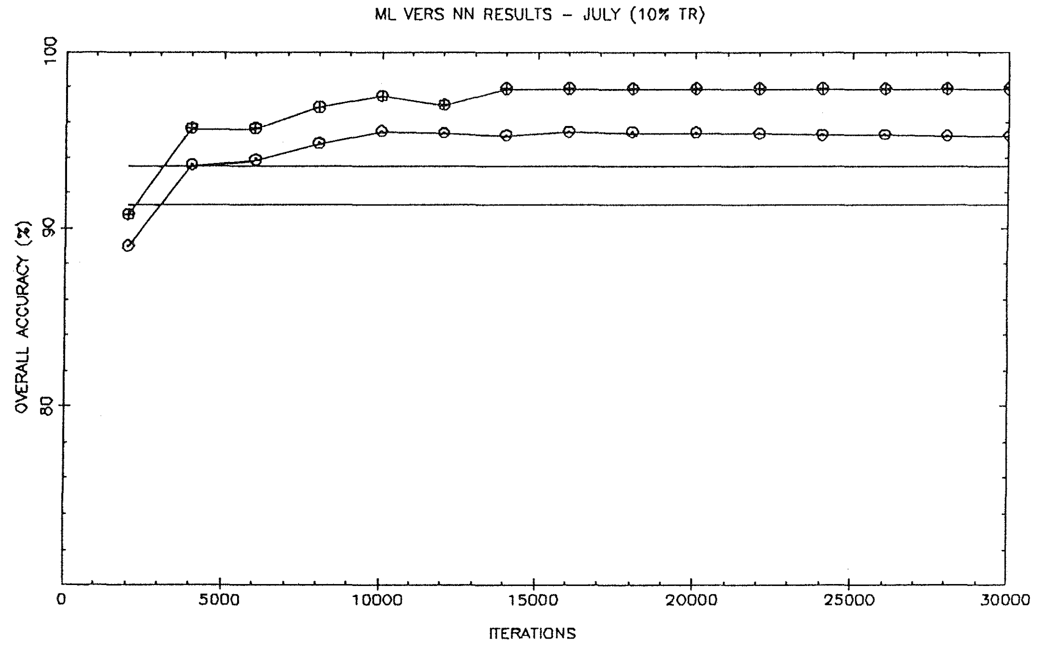

Figure 3: July Image. Classification accuracies (%) using N-N and M-L classifiers. ⊙ Training data value ⊕ Test data values

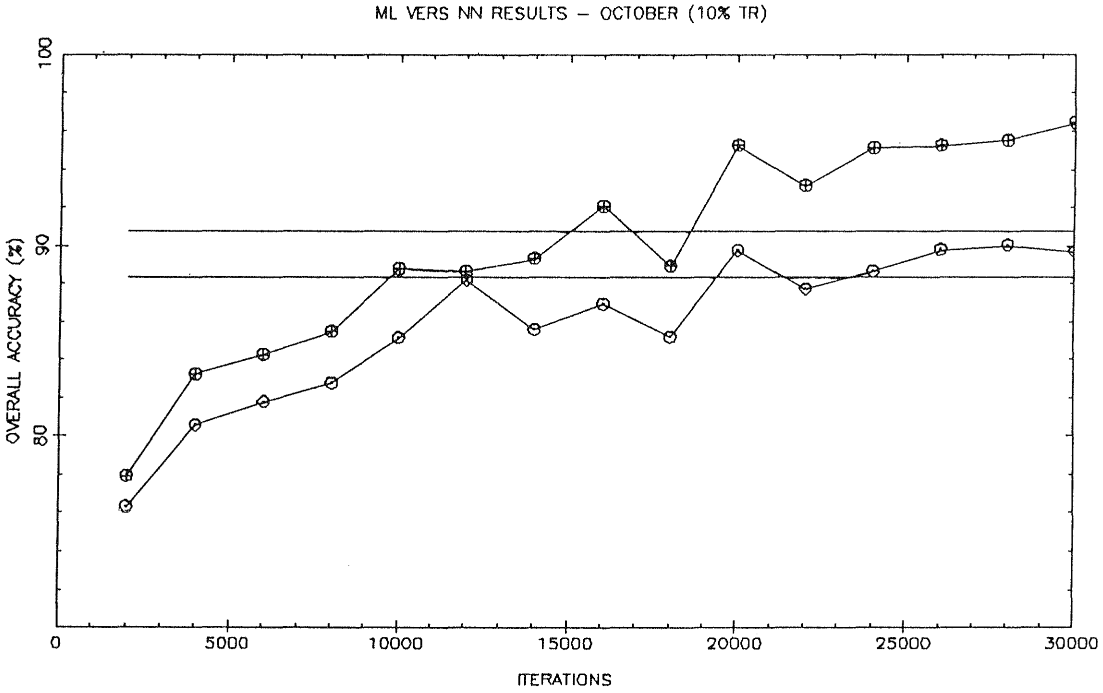

Figure 4: October Image. Classification accuracies (%) using N-N and M-L classifiers. ⊙ Training data value ⊕ Test data values

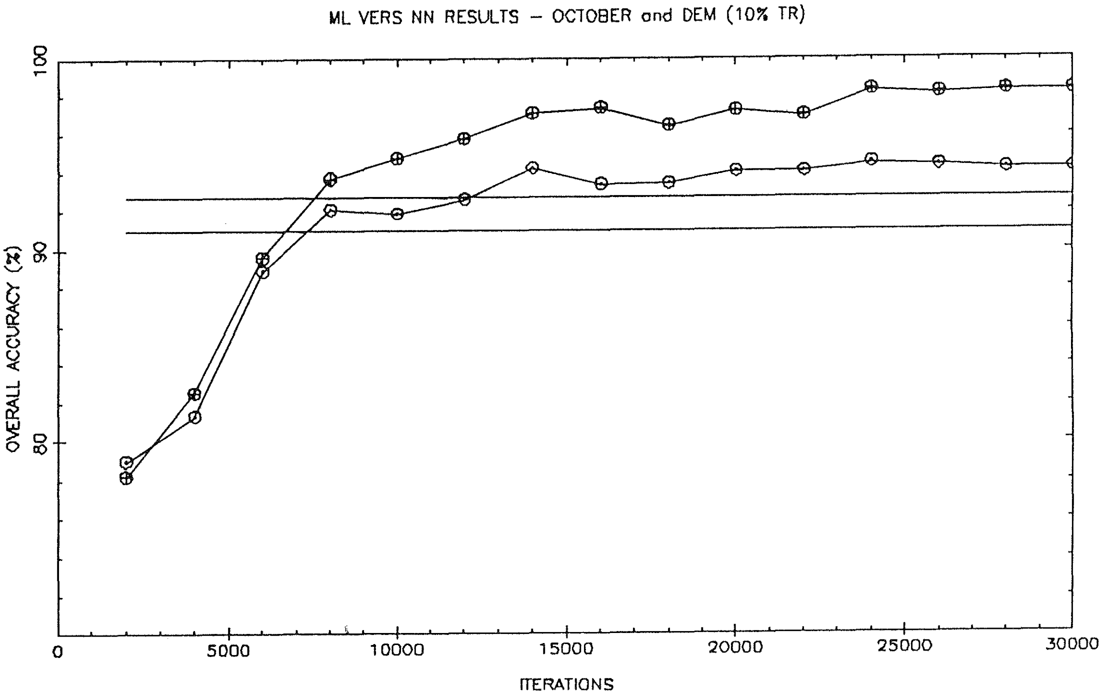

Figure 5: October TM Image and Digital Elevation Model. Classification accuracies (%) using N-N and M-L classifiers. ⊙ Training data value ⊕ Test data values

References

Benediktsson, J.A. & al, 1990a. Neural Network approaches versus statistical methods in classification of multisource Remote Sensing Data. IEEE Trans. on Geosc. and Rem. Sens., 28(4):540-551.

Benediktsson J.A. & al, 1990b. Classification of very high dimensional data using Neural Networks. In: Proc 10th IGARSS Symp, Maryland, Whashington, IEEE 90CH2825-8, 1269-1272.

Benediktsson J.A. & al, 1991. A consensual Neural Network. In: Proc 11th IGARSS Symp, Helsinky, Finland, IEEE 91CH2971-0, 2219-2222.

Bishof H. & al 1991. AI methods for Remote Sensing: Neural Networks and Knowledge-based vision. Proc 11th Earsel Symp, Gratz, Austria, ISBN 900989, 14-22.

Blonda P.N., Pasquariello G. & al 1991. An experiment for the interpretation of multitemporal remotely sensed images based on a fuzzy logic approach. Int. Journ. of Remote Sensing, 12(3):463-476.

Fortuna L. & al 1991a. A Neural Network for seismic events classification. Proc 11th IGARSS Symp, Helsinky, Finland, IEEE 91CH2971-0, 1663-1666.

Fortuna L. & al, 1991b. Multi-layer perceptrons in filtering geophysical signals. Proc 11th IGARSS Symp, Helsinky, Finland, IEEE 91CH2971-0, 1691-1694.

Kahny D. & Wiesbeck W. 1991. Optimum input parameters for classification of multifrequency polarimetric SAR data using Neural Networks. Proc 11th IGARSS Symp, Helsinky, Finland, IEEE 91CH2971-0, 2157-2160.

Kanellopoulos I. & al 1991. Neural Network classification of multi-date satellite imagery. Proc 11th IGARSS, Helsinky, Finland, IEEE 91CH2971-0, 2215-2218.

Key J. & al 1990. Neural Network Vs. Maximum Likelihood classifications of spectral and textural features in visible, Thermal, and passive microwave data. Proc 10th IGARSS Symp, Maryland, Whashington, IEEE 90CH2825-8, 1277-1280.

Kolmogorov A.N. 1957. In: Dokl Akad Nauk USSR, 114:953-956. For a discussion about the validity of the Kolmogorov's theorem appleid to multilayer neural netoworks, see, for example: R.Hecht-Nielsen, 1987. Proc. 1987 IEEE Intern. Conf. on Neural Networks, III, 563-605, New-York, IEEE Press.

Kwok R. & al 1991. Application of Neural Networks to sea ice classification using polarimetric SAR images. Proc 11th IGARSS Symp, Helsinky, Finland, IEEE 91CH2971-0, 85-88.

Lee J. & al 1990. A Neural Network approach to cloud classification. IEEE Trans. on Geosc. and Rem. Sens., 28(5):846-855.

Mulder N.J. & Spreeuwers L. 1991. Neural Networks applied to the classification of remotely sensed data. Proc 11th IGARSS Symp, Helsinky, Finland, IEEE 91CH2971-0, 2211-2213.

Pasquariello G. & Blonda P.N. 1992. Land use mapping from multitemporal satellite images and ancillary data. In: ISY conference, Munich, March.

Rumelhart D.E., Hinton G.E., Williams R.J., 1986. Learning Internal Representation by Error Propagation. Parallel Distribuited Processing, vol. I, Rumelhart D.E., McClelland J.L.. Cambridge: MIT Press.

Swain P.H. & Davis S.M. (Ed.) 1978. Remote Sensing the quantitative approach. MaGraw-Hill, New York, 170-174.

Tenorio M.F. & al 1990. Multisource remote sensing data analysis with self-organizing Neural Network. Proc 10th IGARSS Symp, Maryland, Whashington, IEEE 90CH2825-8, 1289-1292.