NEURAL NETWORKS APPLIED TO THE CLASSIFICATION OF REMOTELY SENSED DATA

N.J. Mulder
I.T.C.
P.O.Box 6
7500 AA Enschede
Netherlands

and University Twente
Dept. BSC/EL
P.O.Box 217
7500 AE ENSCHEDE
Netherlands

L. Spreeuwers
University Twente
Dept. BSC/EL
P.O.Box 217
7500 AE ENSCHEDE
Netherlands

ABSTRACT

A Neural network with topology 2-8-8 is evaluated against the standard of supervised non-parametric maximum likelihood classification. The purpose of the evaluation is to compare the performance in terms of training speed and quality of classification. Classification is done on multispectral data from the Thematic Mapper(TM3,TM4) in combination with a ground reference class map. This type of data is familiar to professionals in the field of remote sensing. This means that the position of clusters in feature space is well known and understood, and that the spatial pattern is equally well known. As a spin-off, the application of a neural net to a classical task of statistical pattern recognition helps to demystify neural networks.

neural nets, k-nearest neighbours, remote sensing, classification

INTRODUCTION

After the sobering up, out of speculations about neural networks as prevalent in the behaviouristic school of artificial intelligence, in the late nineteen-seventies, as a result of the publication on Perceptrons by Minsky (Minsky, 1969) (2), a new wave of speculations started off in the early nineteen-eighties.

Neural networks are of technical interest because of the parallel nature of the calculations. Several manufacturers are attempting to build and market hardware neural network devices now.

In case these parallel processing devices become available at the right price/performance figure, image processing of remote sensing data, with its massive data flows, could profit from these devices. For this reason it is worthwhile to investigate the performance of simulated (neural) nets, when applied to e.g. computation intensive classification tasks. Duda & Hart (Duda, 1973) (1) evaluated the performance of perceptrons for speech recognition.

The authors decided to evaluate the performance of a neural network, with topology 2-8-8 with a maximum detector/selector at the output, against the performance of the standard classification rule, for the classification of multispectral data.

The standard classification rule is the cost weighted supervised non-parametric maximum likelihood rule. It is the standard rule because it maximises economic benefit of the decision making process (ref. operations research). The selection of the non-parametric estimation of probability density functions avoids the use of wrong assumptions, such as the assumption of a Gaussian distribution (ref. standard textbooks, lecture notes).

The learning strategy of neural networks needs attention. The usual way is to use backpropagation, where the training samples are presented one by one. The weights of the decision functions are adjusted for every straining sample, and often the sample set must be cycled many (like 1000) times through the training set for the weights to stabilise. For at least linear decision rules it was known as early as 1970 that the simplex method should be used. The authors set out to investigate whether it is possible to develop learning rules which are inherently parallel rather than sequential. At the moment of finalising this paper, progress was at the point where it is recognised that the training problem is, basically, a curve (surface) fitting problem which has already been solved

DEFINING THE STANDARD

Given an area where for each area element the tupple (class, x1, x2) is known, then frequency(class,x1,x2) can be calculated. For a given observation tupple (x1,x2), the frequency of occurrence with each of the members of the set (class) is recoverable. If the cost of a wrong classification is ECU 1, and the benefit of a correct classification is also ECU 1, then the maximum benefit, minimum cost decision is to assign the class label to the sample that has the maximum frequency. With proper normalisation over class and (x1,x2) the Bayes rewriting tautology appears:

$$P(class \mid x1,x2) * P(x1,x2) = P(x1,x2 \mid class) * P(class)$$

With the assumed full knowledge of the class x1,x2 relation there is no need for the Bayes rule. There is also no need to parameterise freq(class,x1,x2). The only thing needed when the training set does not cover the whole area but is otherwise proportional, is some form of frequency smoothing. The k-nearest neighbours, k-NN, method provides such a smoothing mechanism. The k-NN method with class proportional sampling will be used for comparison.

FEATURE EXTRACTION

From a TM dataset of Biddinghuizen in the Flevopolder, Netherlands, channels TM3 and TM4 were selected as spectral features. As the reflectance model is multiplicative, the assumption of a 2-dimensional Gaussian distribution would be invalid! So it does not make sense to run a test using Gaussian maximum likelihood classification.

THE REFERENCE, CLASS MAP

For each area element a class label is known. A special class is the class of mixels which is only known for the field ownership boundaries to start with. The class of mixels has been merged with the nill class, representing the "unknown".

THE TRAININGSET

Frequency(class,x1,x2) is calculated from the class map and the two feature images x1 and x2. The size of the area is 320 x 200 scene elements, resampled to 25 m. Each file uses 70.4 kbytes. About 50 k of the reference map belong to the class "0", representing unknown/not defined/mixels.

The procedure for determining frequency without use of a full 2-dim array, as in scattergrams, is to shift bytes x2,x1,class → integer, sort the integer array and determine Frequency(class,x1,x2) from runlengths of the class,x1,x2 tupples.

THE REFERENCE CLASSIFICATION LOOK-UP TABLE

By applying a weightfactor of +1 ECU for good classification and -1 ECU for wrong classification the minimum cost rule is equal to the maximum likelihood rule. Having a complete reference map means that maximum likelihood is equivalent to maximum frequency. Placing the (class : where Frequency(class,x1,x2) is max over class) in a classification look-up table ClassLUT(x1,x2), classification is executed by:class' = ClassLUT(x1,x2).

THE CONFUSION TABLE FREQ(CLASS, CLASS'')

The confusion tables are calculated also by shifting (class,class') → integer, sort integer array, calculate runlengths. The figure of merit is defined as (benefit-cost)/scene-element. The more familiar figure is the relative error: cost/scene-element. In the figure of merit, the rows and columns for class = "0" are not included.

EXPERIMENTAL RESULTS

a) MAXFREQ CLASSIFICATION,
this is the standard. Figure 1, shows the partitioning of the TM3, TM4 feature space. Classification of 70.4 k of scene-elements produces 19.4 k of non-zero elements. Benefit = 15969 ECU. Relative benefit = 0.82 ECU / scene-element. Error rate = 0.09.

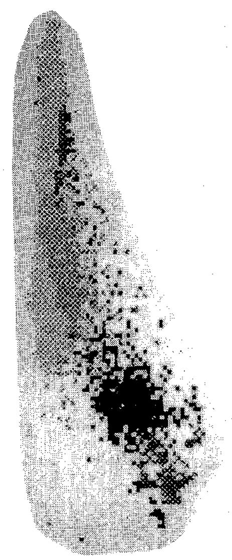

fig.1

b) k-NN, FREQUENCY SMOOTHING,
assumes a certain degree of continuity of probability density in featurespace. The classification look-up table of figure 2 is a smoothed version of the one in figure 1.

Number of non-zero elements is 19.4 k. Benefit = 15344 ECU. Relative benefit = 0.79 ECU / scene-element. Error rate = 0.105.

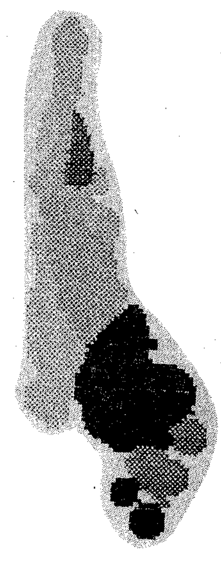

fig.2

c) NEURAL NET 2-8-8,
after about 1000 iterations of error back propagation, taking about 10 hrs. of training on a SUN workstation. This compares to 6 sec. training for maxFreq, and 12 sec. training for k-NN. The ratio of neural net training to standard non-parametric maximum likelihood training is of the order of 36000 to 12, or 3000 : 1!

Number of non-zero elements is 19.5 k. Benefit = 14214 ECU. Relative benefit = 0.73 ECU / scene-element. Error rate = 0.135.

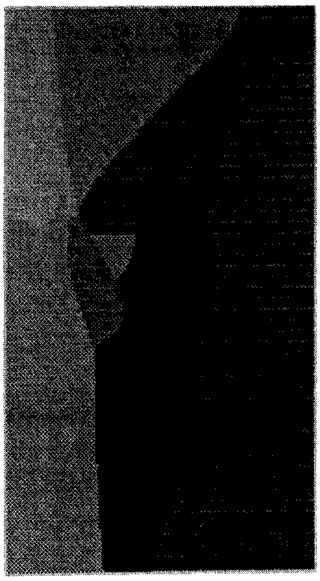

Fig.3

CONCLUSIONS

The training time using error backpropagation in the neural network with topology 2-8-8 is about 3000 times that of the standard method.

The quality of the neural network classification schema is less than that of the k-nearest neighbour classification schema. In economic terms, the benefits compare for neural net to k-NN as 0.73 to 0.79 ECU / scenel, and in terms of error rates as 0.135 to 0.105. The 3000 fold increase in training time results in an increase in error rate from 10.5% to 13.5%, which is a relative increase in error of about 30%!

RECOMMENDATIONS

The k-NN method can easily be implemented in a network on the basis of minimum distance classification for a set of subclasses.

The backpropagation schema for training "neural" nets has no reason for existance other than to contribute to the mystification of the subject. For two layer perceptrons, it is known that the weights can be found using the simplex method of operations research. It should be easy to formulate the construction of decision boundaries as a surface fitting problem, and solve it accordingly without falling back to sequential training.

REFERENCES

(1) R.O. Duda and P.E. Hart "Pattern Classification and Scene Analysis", New York: Wiley, 1973.

(2) M. Minsky and S. Papert "Perceptrons: An Introduction to Computational Geometry", Cambridge, Massachusetts; The MIT Press, 1969.