INFORMATION FUSION IN TREE CLASSIFIERS

A. Senthil Kumar and K.L. Majumder
Image Processing and Products Division
Space Applications Centre
Ahmedabad-380053 INDIA.

Technical Commission I, Working Group 3

KEY WORDS : Hierarchial classification, Information Fusion, Neural Networks.

ABSTRACT

Three methods of fusing information from maximum liklihood and neural network methods for multispectral data classification are discussed in this paper. The purpose of the fusion is to enhance the interpretation of a pixel under study with the classifier that has a minimum uncertainity in assigning the pixel to one of desired classes. The classification performance with the fusion techniques is found to be superior to that of the individual classifiers.

1. INTRODUCTION

Hierarchial, decision-tree based classifiers (DTCs) are very useful for complex pattern recognition tasks involving several pattern classes and a large number of features. In remote sensing, the DTC is of great interest for classifying many earth's targets with several of their subcatagories, and for handling space-borne imaging spectrometer data with channels ranging from a few tens to a few hundreds [Kim, 1991]. There are many advantages with the DTC over an one shot classifier (OSC) for these applications. The DTCs are flexible in that new branches of the tree can be opened as and when the application demands. At each decision node of the DTC, we have a maximum of two or three classes (or groups of classes or of spectral channels), and hence the training at each node is computationally less intensive when compared to the OSC.

Despite these advantages, there are several factors that affect the classification performance of the DTC : (1). classification strategy at each decision node, (2). the design criteria of the tree; the performance depends significantly on how the given features or classes are grouped at each node, (3). its sensitivity to the noise, to mention a few. For the univariate feature cases, the classification strategy is largely restricted to simple thresholding [Sethi, 1995], while for the multivariate cases, the conventional Maximum Liklihood (ML) method is commonly employed. In the recent past, the artificial neural networks (ANNs), both supervised and unsupervised, have been explored by several research workers as a viable alternative to the conventional statistical approaches for the remote sensing data classification problems [Benediktsson, 1993, Hara, 1994]. In this paper, we are concerned with fusion of the information obtained from both the ML and the ANN classifiers in order to realize overall better classification performance for a multispectral satellite imagery data.

The motivation behind the information fusion approach (IFA) is to enhance the interpretation of a particular pixel under study with the classifier that has a minimum uncertainity in assigning the pixel to one of desired classes. It is frequently observed that the ML method works very well for some classes better than the ANN, especially in the mulispectral data classification [Bischoff, 1992]. This may be surprising, since it is now well established that the ANNs are capable of estimating the *a posteriori* conditional probabilities of all classes presented to them [Wan, 1990]. In principle, it is possible to realize these conditional probabilities with an optimum neural network architecture, provided that there is no limit on the network size and the training database is unbound. But in reality, one has to deal with only a finite set of training data and limited computational resources. A common practice is that one has to start with an *educated* guess of the network size and after training it, he has to *crossvalidate* its performance with a set of test data. If the network size chosen is less than the optimum one, it learns the training data poorly. On the other hand, if it is bigger than the optimum size, the net generalizes the test data poorly. A common practice is that we start with a set of neural networks of different sizes, train all of them before deciding the one that gives the best overall performance with both the training and the test data [Bischoff, 1992]. The penality is, however, high computation involved for training these networks, and is very cumbersome, in particular, for the DTC realization in which several decision nodes are to be trained.

In a recent study, we have reported the use of the FI for integrating an ensemble of ANNs for multispectral data classification [Kumar, 1997]. It

was shown that the FI approach gives an overall better classification when compared to that of the individual networks. We further extended this approach for combining two different information sources (the original and its smoothed version) to improve the overall classification. In this paper, we explore three methods of integrating the ML and the ANN classifiers at each decision node of the DTC, and compare their classification performance with those obtained when the classifiers are applied individually.

Earlier, Ersoy and Hong [1990] suggested a hierarchial approach for classifying airborne multispectral data. Their cascaded approach is, of course, different from the DTC in that each of the neural nets was learnt first, and classification was performed, and those misclassified pixels were allowed down in the cascade after undergoing a nonlinear tranformation. While this method works well for a low-dimensional input data, success with the high-dimensional, numerous class cases will depend heavily on how fast the nonlinear transformation can be implemented.

The rest of the paper is organised as follows. Section 2 describes the two methods of integrating the ML and the ANN classifiers. Section 3 discusses about the experimental study over a multispectral data with the design aspect of the DTC. Section 4 brings out the effect of additive noise on the classification performance of the individual and fused classifiers. Our conclusions are summarized in Sec. 5.

2. CRITERIA FOR FUSION

As mentioned above, the criteria for fusing information from different classifiers differ only by the way the information measure is defined. In the following, three methods of information fusion are discussed.

2.1. Direct Pinz Method :

Pinz and Bartl [1992] proposed earlier a method of fusing the ML and ANN methods for a one shot classifiation of the Landsat-TM multispectral data. According to this method, for each test pixel, the confidence of the ANN is first evaluated as the difference between the most activated output neurons. If this confidence is above a desired threshold, the fusion selects the ANN for classification. Else, it selects the ML classification. We have directly adapted this method for implementing it at each decision node of the DTC. This method is henceforth referred to DPM.

2.2. Modified Pinz Method :

It is clear that the DPM is biased toward the ANN as the confidence of the ML for the test pixel is never considered. This is not a desirable approach, as pointed out in the introduction, the ML performs better than the ANN. In addition, it is difficult to select an appropriate threshold since the ANNs are, in general, undertrained (ie., their learning is usually terminated after a certain number of iterations) for better generalization properties, since the overtrained network does not classify the test data as it does for the training data.

To overcome these problems, we propose here a modified version of the Pinz method described as follows: The test pixel is first subjected to both the classifiers (For ML, the search for maximum probability is not carried out). The peak difference between the two most competant outputs for the pixel is estimated for each classifier, and normalized to the maximum peak value. The pixel is assigned to the classifier for which the normalized peak difference is higher. As will be shown in the next section, this modified Pinz method (MPM) improves the overall classification accuracy when compared to that of the DPM.

2.3. The Fuzzy Integral Method:

A brief introduction on the FI is given below for the sake of completeness. For full details, the reader is referred to Kumar [1997].

The computation of the FI is as follows: Let $U=\{u_1,u_2,...,u_n\}$ be a finite set of values, and $h:U\rightarrow[0,1]$ be a function. The fuzzy integral I is evaluated from h and a parameter, the so-called fuzzy measure g, as

$$I = \max_{i=1}^{n} [\, \min \{ h(u_i) , g(A_i) \} \,] \qquad (1)$$

where $A_i = \{u_1,u_2,...,u_n\}$. The fuzzy measures, $g(A_i)$, are obtained using its additive properties in a recursive manner:

$$g(A_i) = g(u_1) = g^1,$$

$$g(A_i) = g^i + g(A_{i-1}) + \lambda g^i g(A_{i-1}), \quad \text{for } i = 2,...,n. \qquad (2)$$

The value λ is determined by solving the equation

$$\lambda + 1 = \prod_{i=1}^{n} (1+\lambda g^i), \qquad (3)$$

where $\lambda \in (-1, +\infty)$, and $\lambda \neq 0$. This is obtained by solving an (n-1)th degree polynomial equation and

finding the unique root greater than -1. The fuzzy measures g can thus be fully determined by the so-called density function g^i.

The physical interpretation of the FI can be described as follows. The density function g^i, is related to the degree of importance of the classifier u_i, towards the final evaluation. The (min) operator in Eq.(1) is interpreted as the grade of agreement between the evidence values, $h(u_i)$, and the degree of importance or expectations g, while the (max) operator does the searching process for the maximal grade of agreement between the objective evidence and the expectations. Now, let us apply these concepts for the current problem of combining different classifiers.

Consider $Y=\{C_1,C_2,...,C_n\}$ as a set of classes of interest. In hierarchial classification, each Ci may, in fact, represent a set of groups or subgroups by itself. Let $U = \{u_1,u_2,...,u_n\}$ represent the set of classifiers, and X be the pixel under consideration to be recognized. Let $h_p:U\text{-->}$ [0,1] represent the partial evaluation of the object of the pixel X for each class C_p, i.e., h_p is an indication of how certain we are in the classification of the pixel X to be in class C_p, which takes the value of unity for absolute certainity and zero when X not in C_p. Corresponding to each classifier u_i, the degree of importance, g^i, i.e., how important the classifier u_i is in the recognition of the class C_p. The classification accuracies obtained from the classifier for each class imply the degree of importance of this classifier, and hence are used here directly as the values of the density function.

3. RESULTS AND DISCUSSIONS

To validate the above methods in hierarchial classification, we have considered a multispectral data set of the IRS-1A satellite data with spatial ground resolution of 72 mts. over the north-eastern part of India. Samples of 12 prominent features were extracted visually from the data at three spectral bands, B2 (0.52-0.58 μm.), B3 (0.62 - 0.68 μm.) and B4 (0.77-0.86 μm.). The fifty percent of the samples are used for training, and the entire data set for testing the classification strategies mentioned above. Table 1 gives the classes extracted from the multispectral data with their size and their legends.

As mentioned in the introduction, another important issue is the very design of the DTC. It is, of course, essential that the groups and subgroups of the classes at each decision node must be spectrally separable. We have used the Bhattacharaya distance (BD) for clustering the classes of interest (Table 1). This distance measure is recommended as it dears a closer relationship with the classification accuracy than any other measure functions [Kim, 1991]. The binary decision tree thus obtained is shown in Fig. 1.

Table 2 summarizes the results obtained with the ML, the ANN, and the different fusion method mentioned in Sec. 2. The ANN is a multilayer perceptron network with a single hidden layer consisting of 20 hidden neurons at each decision node. The network was iteratively trained using a gradient descent algorithm till either the total squared error calculated for all the input classes and the network outputs has attained a minimm error bound (0.1) or when training has crossed 1000 iterations. The bound on the number of iterations is due to the fact that some earth's features have belongingness to more than one class, and hence the training process does not satisfy the minimum error condition. In such cases, overtraining the network does not improve the overall classification performance, even though the network would tend to memorize the training data very well, but it would generalize poorly with the rest of the samples.

As evident from the results shown in Table 2, the fusion methods described here perform better classification performances when compared to those of the individual classifiers. Both overall accuracy (i.e., the ratio of the correctly classified and the total number of samples) as well as the average of the percentage accuracies obtained for each class are given for comparison. Note that in some classes (see, for eg., sugarcane 1 and urban), the fusion methods try to obtain the balance between the ANN and the ML, while they retain the same accuracy if it is constant in both the classifiers. While, the maximum classification accuracy is obtained from the fuzzy integral fusion, the MPM edges past the DPM proposed earlier by Pinz and Bartl [1992].

4. CONCLUSION

In this paper, we have shown that by combining the maximum liklihood and the artificial neural networks, one can achieve better classification performance when compared to that of them when applied individually. Of different fusion methods, the method using the fuzzy integral is found to be the best for data classification. A detailed study is in progress for theoritical evaluation of its performance.

References

Benediktsson, J.A, Swain, P.H., and Ersoy,O.K., 1993. Conjugate-gradient neural networks classification of multisource and very high-dimensional remote sensing data. Int J. Remote Sensing, 14, pp. 2883-2903.

Bischoff, H., Schneider, W., and Pinz, A.J., 1992. Multispectral classification of Landsat Images

using neural networks. IEEE Trans Geosci Remote Sensing, **30**, pp. 482-489.

Ersoy, O.K. and Hong, D., 1990. Parallel, self-organizing, hierarchial neural networks. IEEE Trans Neural Networks, **1**, pp. 167-178.

Hara, Y., Atkins, R.G., Yueh, S.H., Shin, R.T., and Kong, J.A., 1994. Application of neural networks for radar classification. IEEE Trans Geosci Remote Sensing, **32**, pp. 100-109.

Kim, B. and Landgrebe, D.A., 1991. Hierarchial classifier design in high dimensional, numerous class cases. IEEE Trans Geosci Remote Sensing, **29**, pp. 518-528.

Kumar, A.S., Basu. S.K., and Majumder, K.L., 1997. Robust Classification of multispectral data using multiple neural networks and fuzzy integral, IEEE Trans Geosci Remote Sensing, **35 (3)**, pp. 787-790.

Pinz, A. and Bartl, R., 1992. Information fusion in image understanding: Landsat classification and ocular fundus images. In: Sensor Fusion V. The SPIE, Washington, USA, Vol. 1828, pp. 276-287.

Sethi, I., 1995. Neural implementation of tree classifiers. IEEE Trans Syst Man Cybernetics,**25**, pp. 1243-1249.

Wan, E.E., 1990. Neural network classification: a Bayesian interpretation. IEEE Trans. Neural Networks, **1**, 303-305.

Table 1. Extracted classes (legends) with corresponding number of samples

Classes (legends)	Sample size
Water (A)	451
Sugarcane 1 (B)	64
Sugarcane 2 (B)	60
Sugarcane 3 (B)	39
Wheat 1 (E)	145
Wheat 2 (F)	36
Riversand (G)	260
Fallow 1 (H)	58
Fallow 2 (I)	40
Fallow 3 (J)	31
Fallow 4 (K)	48
Urban (L)	95
Total No. of Pixels	1327

Table 2. Recognition accuracies (in %) of different classifiers (see text). Here the APA represents the average of percentage accuracies of all classes, and OA, the overall accuracy.

Class	ML	ANN	DPM	MPM	FI
A	100.0	100.0	100.0	100.0	100.0
B	40.6	78.1	46.9	64.1	58.7
C	93.3	68.3	93.3	76.7	93.3
D	23.1	41.0	23.1	28.2	41.0
E	97.9	98.6	98.6	98.6	98.6
F	42.9	28.6	42.9	37.1	28.6
G	54.2	53.1	53.5	53.5	53.5
H	67.2	89.7	67.2	82.8	67.2
I	97.5	100.0	97.5	100.0	100.0
J	67.7	67.7	67.7	67.7	67.7
K	56.3	79.2	79.2	75.0	79.2
L	86.3	57.9	73.7	79.0	83.2
APA	68.93	71.85	70.3	71.88	72.25
OA	79.03	79.56	79.2	80.24	80.47

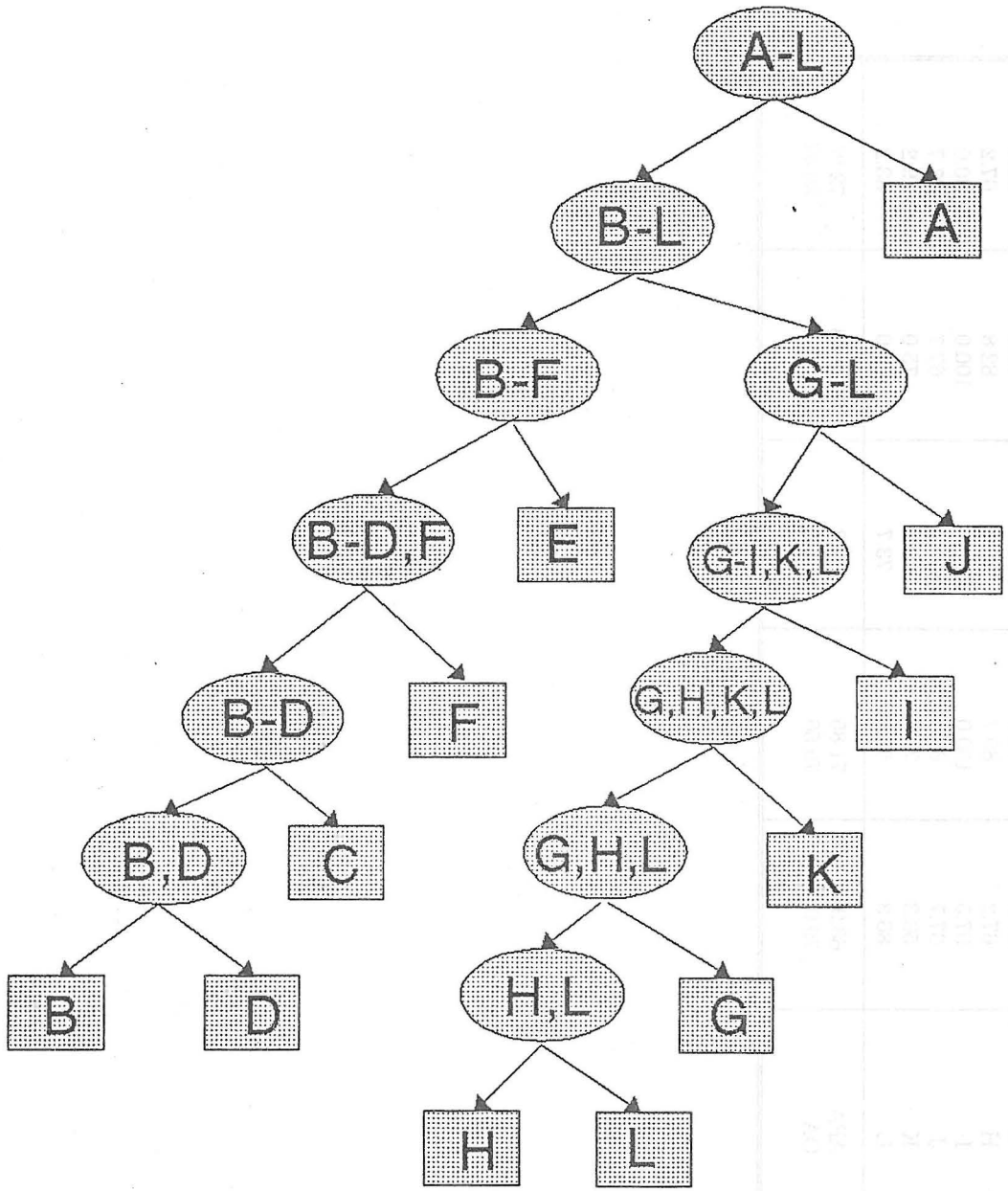

Fig.1. Decision tree obtained for classes of interest using the Bhattacharaya distance as a clustering measure.