Anais X SBSR, Foz do Iguaçu, 21-26 abril 2001, INPE, p. 1353-1365, Sessão Técnica Oral - Workshops

Derivation of Land Cover Information from Remotely Sensed Data using Multiple Classifiers

CARLOS A. O. VIEIRA [1]
PAUL M. MATHER [2]

[1]Universidade Federa de Viçosa - UFV
Campus da UFV, Dep. Engenharia Civil, CEP 36.571-000, Viçosa, MG, Brazil
cvieira@mail.ufv.br

[2] The University of Nottingham – School of Geography
University Park, Nottingham NG7 2RD
Paul.Mather@nottingham.ac.uk

Abstract Due to the growing volume of image data from planned and existing sensors, new data-processing techniques are required to allow the information to be processed promptly and accurately. Although the range of image processing techniques has greatly expanded in recent years, from classical statistical approaches to neural network methods, there is no single classification algorithm capable of deriving generic products from remotely sensed data that can be used with confidence. The performance of these algorithms is strongly dependent upon the selected data set and on the efforts devoted to the design phase. In this paper, we report a more systematic investigation into the problem of combining multiple classifiers in the context of land cover mapping using remotely sensed data. Four approaches are proposed, based on voting principles, Bayesian formalism, evidential reasoning, and artificial neural networks. Preliminary results indicate that improvements can be obtained in difficult pattern recognition problems by combining or integrating the outputs of complementary multiple classifiers.

Keywords: combining classifiers, remote sensing, classification, neural network, image data, land cover mapping.

1 Introduction

In many real-world classification problems, the categories of interest are not fully separable in terms of their measured characteristics. With such problems, it is unrealistic to expect a perfect classification, with 100% accuracy. The objective of a pattern recognition system is to achieve the "best possible" performance. The obvious question that arises, of course, is how to determine the optimum classification rate.

The concept of combining the outputs of several classifiers is investigated here as an alternative to the development of new classification algorithms more complex than the present ones (Roli et al. 1997). Combined classifiers have led to improved classification performance in the context of handwriting recognition (Xu et al. 1992), signal processing (Ghosh et al. 1995), and recently in remote sensing (Roli et al. 1997, Wilkinson et al. 1995, Kanellopoulos and Wilkinson 1997). The number of classification algorithms available and the increasingly sophisticated types of features able to be used to represent and recognise patterns justify the growing interest in this topic (Xu et al. 1992). Selecting the "best" individual classifier is not

necessarily the ideal choice, since discarding the results of less-successful classifiers may waste potentially valuable discriminatory information (Tumer and Ghosh 1996). The premise of this approach is that independent classifiers may provide increased interpretation capabilities and more reliable results due to the fact that different classifiers produce different decision boundaries in feature space. However, despite the increasing body of experimental results showing classification improvements due to combining classifiers, there has been no analytical study that can quantify the achievable gains, and examine the statistical validity of these gains. This paper provides an overview of methods of combining multiple classifiers such as the voting principle, Bayesian formalism, evidential reasoning, and artificial neural networks, and also presents results that quantify the improvements due to multiple classifier combination.

2 Standard Methods for Combining the Output of Multiple Classifiers

Four techniques for combining the output of multiple classifiers (i.e., voting principle, Bayesian formalism, evidential reasoning and neural network) are reviewed in this paper. The first three are reviewed by Xu et al. (1992) in the context of handwriting recognition. These methods require that classifiers be trained independently. The fourth method, using an artificial neural network, was introduced by Wilkinson *et al.* (1995) in the context of remote sensing. The combination process begins with each classifier being individually trained, under some set of different conditions (e.g., using independent sets of discriminant variables). Each trained classifier is then presented with an identical set of inputs for which it determines its own prediction. These individual predictions combine to form the overall (final) prediction.

A significant problem when trying to combine classifiers is the different output information that various classification algorithms supply. Assume an image classification problem consisting of M mutually exclusive classes $C_1 \cup C_2 \cup \ldots \cup C_M$ with each of $C_i, \forall i \in \Lambda = \{1, 2,.., M\}$ representing a set of specified label patterns called a class (e.g., M = 10 for a problem of numeral recognition). Consider also that each class represents a set of specific patterns, and that each pattern is characterised by a feature vector $\boldsymbol{X}$. According to Xu *et al.* (1992), the output information can be divided into three levels.

- The *abstract level*, where a classifier outputs a unique label. Combinations based on this output level can be formulated as follows. Given K individual classifiers ε_k, $k = 1,..,K$ each of which assigns input pattern $\boldsymbol{X}$ to a class label j_k, i.e., produces an event $\varepsilon_k(\boldsymbol{X}) = j_k$ then the problems is to use these events to build an integrated classifier E, which gives to $\boldsymbol{X}$ a single, definitive label j, i.e., $E(\boldsymbol{X}) = j$, $j \in \Lambda \cup \{M+1\}$, where $j = M+1$ denotes that $\boldsymbol{X}$ is rejected by ε, and is left unlabelled.
- The *rank level*, where classifiers rank all possible labels in the mutually exclusive sets in a queue $L_k \subseteq \Lambda$ with the label at the top being the first choice. In this case, the problem is to use these events $\varepsilon(\boldsymbol{X}) = L_k$, $k = 1,..,K$ to build an E with $E(\boldsymbol{X}) = j$, $j \in \Lambda \cup \{M+1\}$; and,
- The *measurement level*, when a classifier attributes for each label a measurement value (e.g. probability P) to address the degree that a feature vector $\boldsymbol{X}$ belongs to that class. Thus, for an input $\boldsymbol{X}$, each ε_k produces a real vector $M_\varepsilon(k) = [P_k(1), .. ,P_k(M)]^t$ (where $P_k(i)$ denotes a kind of degree that ε_k considers that $\boldsymbol{X}$ has label i), the problem is to use these events $\varepsilon(\boldsymbol{X}) = M_\varepsilon(k)$, $k = 1,..,K$ to build an E with $E(\boldsymbol{X}) = j$, $j \in \Lambda \cup \{M+1\}$.

It is reasonable to postulate that measurement vectors of different kinds can be transformed into the same kind of measurement. The combination problem studied by Xu *et al.* (1992) and Roli *et al.* (1997) belongs to abstract level combination. In this paper, we also study the problem of combining classifiers at the measurement level.

2.1 Voting

It can be appreciated that there are areas in which different classifiers may agree on the class to be assigned. Given that there is no contradiction (or conflict) and that the independently trained classifiers agree on the outcome, there is good evidence that samples lying in such an area of feature space should be classified according to the joint agreement. However, the main problem in classification integration is then to decide what to do with those samples for which the classifiers do not agree.

Some researchers have investigated the possibility of combining classifier outputs using voting schemes (Xu *et al.* 1992, Battiti and Colla 1994, Roli *et al.* 1997). The abstract level seems to be the most common way of combining classifiers under this scheme, since any kind of classifier will at least supply output information at the abstract level.

As indicated in the preceding section, the problem is to produce a new event $E(\boldsymbol{X}) = j$ from the given events $\varepsilon_k(\boldsymbol{X}) = j_k$, $k = 1,\ldots,K$ where conflicts may exist among the decisions of K classifiers. A simple rule for resolving conflicts of this kind is *voting by majority.* The decision is made such that the label that receives more than half of the votes is taken as the final output. A *conservative voting rule* is the one that the combined classifier E decides that $\boldsymbol{X}$ comes from C_j if all K classifiers decide that $\boldsymbol{X}$ comes from C_j simultaneously, otherwise it rejects $\boldsymbol{X}$. Xu *et al.* (1992) describe a variant of the voting principle and a more general version in which, for convenience, they represent the event the event $\varepsilon_k(\boldsymbol{X}) = i$ in the form of a binary characteristic function:

$$T_k(X \in C_i) = \begin{cases} 1, & \text{when } \boldsymbol{e}_k(X) = i,\ i \in \Lambda \\ 0, & \textit{otherwise} \end{cases} \tag{1}$$

The general version is as follows:

$$E(X) = \begin{cases} j, & \textit{if } (T_E(X \in C_j)) = \max_{i \in \Lambda} T_E(X \in C_i) \geq \boldsymbol{a} * K \\ M+1, & \textit{otherwise} \end{cases} \tag{2}$$

where $T_E(X \in C_i) = \sum_{k=1}^{K} T_k(X \in C_i),\ \ i = 1,..,M$ and $0 < \alpha < 1$.

Note that voting by majority is a special case of Equation 2 with $\alpha = 0.5$. The conservative voting rule is equivalent to a special case of Equation 2 with $\alpha = 1.0$. However, in Equation 2, the threshold operation only considers that the maximal votes of the final selected label must be large enough. There may be cases in which there are more than two labels that receive the maximal vote, or the vote of maximal are not considerably larger that the vote of the second maximal. In these cases, even though the maximal vote of the selected label may be quite large, the decision still may not be reliable since there exists an opponent that may also receive a large vote. To tackle this problem, a new *comparative majority-voting* rule is proposed in Equation 3:

$$E(X) = \begin{cases} j, & \textit{if } (T_E(X \in C_j)) = \max_1, \max_1 - \max_2 \geq \boldsymbol{a} * K \\ M+1, & \textit{otherwise} \end{cases} \tag{3}$$

where $0 < \alpha < 1$. And, $\max_1 = \max_{i \in \Lambda} T_E(\boldsymbol{X} \in C_j)$ and $\max_2 = \max_{i \in \Lambda - \{j\}} T_E(\boldsymbol{X} \in C_j)$.

2.2 Bayesian Formalism

Currently, the most popular way of combining multiple classifiers using Bayesian formalism is via simple averaging of the corresponding output values (Perrone and Cooper 1993, Tumer and Ghosh 1995). For a Bayes classifier ε, its classification of an input $\boldsymbol{X}$ is actually based on a set of post-probability measurements, denoted by $P_k(\boldsymbol{X} \in C_i/\boldsymbol{X})$, $i = 1,\ldots,M$, $k = 1,\ldots,K$. This formula derives an approximation to the probabilities that $\boldsymbol{X}$ comes from each of the M classes under the condition $\boldsymbol{X}$. We can use this approximation of the post-probabilities for combining classification results on the same $\boldsymbol{X}$ by all K classifiers. One simple approach is to use the following average value as a new estimation of combined classifier E:

$$P_E(X \in C_i) = \frac{1}{K}\sum_{k=1}^{K} P_k(X \in C_i), i = 1,..,M \tag{4}$$

The final decision is given by

$$E(X) = \begin{cases} j, & if\,(P_E(X \in C_j)) = \max_{i \in \Lambda} P_E(X \in C_i) \\ M+1, & otherwise \end{cases} \tag{5}$$

that is, a Bayes decision is based on these estimated post-probabilities. We call such a combined E an *average Bayes classifier*. Moreover, we could use an additional threshold $0 \le \alpha \le 1$ in Equation 5 to take into account the trade-off between the substitution rate and the rejection rate.

An alternative way to combine classifiers using the Baysian formalism is by considering the available prior knowledge of the error of each classifier (Xu *et al.* 1992), as represented by its confusion matrix. For the k^{th} classifier ε_k, the confusion matrix can provide estimates of the conditional probabilities that propositions $\boldsymbol{X} \in C_i$, $i = 1,\ldots,M$ are true under the occurrence of the event $\varepsilon_k(\boldsymbol{X}) = j$, that is:

$$P(X \in C_i / \boldsymbol{e}_k = j_k) = \frac{n_{ij}^{(k)}}{\sum_{i=1}^{M} n_{ij}^{(k)}} \tag{6}$$

where n_{ij} are the elements in the confusion matrix, in which row i corresponds to class C_i and column j correspond to the event $\varepsilon_k(\boldsymbol{X}) = j$. On the basis of these conditional probabilities (Equation 6), the combination can be carried out using the follow belief function:

$$bel(i) = \boldsymbol{h}\prod_{k=1}^{K} P(X \in C_i / \boldsymbol{e}_k = j_k) \tag{7}$$

with η as a constant that ensures that $\sum_{i=1}^{M} bel(i) = 1$.

That is, $\frac{1}{\boldsymbol{h}} = \sum_{i=1}^{M} \prod_{k=1}^{K} P(X \in C_i / \boldsymbol{e}_k = j_k)$.

Finally, depending on these $bel(i)$ values, $\boldsymbol{X}$ can be classified according to the decision rule:

$$E(X) = \begin{cases} j, & if\,(bel(j)) = \max_{i \in \Lambda} bel(i) \ge \boldsymbol{a} \\ M+1, & otherwise \end{cases} \tag{8}$$

where $0 < \alpha < 1.0$ is a threshold.

2.3 Evidential Reasoning

Evidential reasoning allows a number of data sources to be combined to generate a joint inference concerning labelling (Lee *et al.* 1987). Details on development of the theory of evidence can be found in Shafer (1979), Garvey *et al*. (1979).

Consider the *measurement level* in which a classifier attributes for each label a measurement value (e.g. probability P) to address the degree to which that feature vector $\boldsymbol{X}$ belongs to that class. Thus, for an input $\boldsymbol{X}$, each ε_k produces the *basic probability assignments* - *bpa* - [$P_k(1)$, .. ,$P_k(M)$], where $P_k(i)$ denotes a kind of degree that ε_k considers that $\boldsymbol{X}$ has label i. These *bpa* can be multiplied by the individual classifier's confidence to generate the *mass of evidence* $m_A^k(\{P_k(1),\ldots, P_k(M), \theta\})$ for each classifier, where the symbol θ is used to represent uncertainty (1 – classifier's confidence). An estimation of the individual classifier confidences are easily obtained by testing the classifiers with an independent sample set, and they can be derived either from the overall accuracy or by computing the kappa coefficient from their respective confusion matrices.

Once two masses of evidences m_A^1 and m_A^2, representing independent opinions, are expressed relatively to the same *frame of discernment* Θ_A that corresponds to specific types of crop classes (e.g., potatoes, sugar beet), they can be combined. The combination is made based on accumulating all of the mass of evidence. The aggregation of the multiple evidences is called *Dempster's orthogonal sum*, or *rule of combination*. Dempster's rule pools masses of evidence to produce a new composite mass of evidence m_A^3 that represents the consensus of the original disparate opinions. In other words, this produces a new mass of evidence that leans toward points of agreements between the original opinions and away from points of disagreement (Garvey 1987). If $m_A^3(A_c)$ denotes the new aggregated *bpa*, the combination rule can be specified by:

$$m_A^3(A_c) = K \sum_{A_i \cap A_j = A_c} m_A^1(A_i) \bullet m_A^2(A_j) \tag{9}$$

where

$$K^{-1} = 1 - \sum_{A_i \cap A_j = f} m_A^1(A_i) \bullet m_A^2(A_j) \tag{10}$$

and ϕ means *null* intersection. The procedure continues combining the mass of evidence $m_A^3(A_c)$ with other independent classifier opinions and so on. Because Dempster's rule is both commutative and associative, multiple (independent) bodies of evidence can be combined in any order without affecting the result.

After accumulating all of the masses of evidence in $m_A^k(A_c)$ the final belief function can be computed $bel(A_i) = \sum_{A_c \subseteq A_i} m_A^k(A_c)$ and the value of $bel(\neg A_i)$, which also contain useful information for the final decision.

The combined classifier E using this theory can be defined using the following rules:

$$E(x) = \begin{cases} j, & \text{if } (bel(A_j)) = \max_{i \in \Lambda}\{bel(A_i) / \forall i, bel(A_i) \geq a_1, bel(\neg A_i) \leq a_2\} \\ M+1, & \text{otherwise} \end{cases} \tag{11}$$

where $0 < \alpha_1, \alpha_2 < 1$ are predefined thresholds that take into consideration respectively the $bel(A_i)$ and the $bel(\neg A_i)$'s.

2.4 Artificial Neural Networks (ANN)

In this section, we describe a variant of the approach suggested by Wilkinson *et al.* (1995) and Kanellopoulos and Wilkinson (1997). The different classifiers are trained using independent data sets. These classifiers are tested with another set of independent sample data. A second training set is then built up from the output test data on which the initial classifiers were tested. This new training set is then used to train a classifier in the second stage of the classification, for instance using an ANN. This ANN is implemented having M classes times K classifiers neurones in the input layer, corresponding to the outputs of all classifiers and it has one output layer with M neurones, corresponding to the M classes in study. The purpose of this approach is to be able to apply different models to the data in the classification stage and to highlight samples for which the models disagree. These patterns are then passed to the second stage of classification that has been specially trained to deal with such classes.

3 Data and Methods

A SPOT (14 June 1994) High Resolution Visible (HRV) multispectral image and two Landsat TM images (27 June and 20 July - 1994) of a region of flat agricultural land located near the village of Littleport (E. England) are used in this study. Field Data Printouts, which contain the official record of the crop being grown in each field, are also available for the summer of 1994 and were used to generate the reference image. Six crop categories were selected for these experiments (potatoes, sugar beet, wheat, fallow, onions, and peas). Registration of the image to the Ordnance Survey (GB) 1:25,000 map was performed using 17 ground control points and nearest neighbour interpolation. The RMS errors were 0.462, 0.477 and 0.438 pixels for the SPOT HRV and Landsat TM images respectively.

Since the proposed methodology requires the analysis of pixels from different geographic locations and on different dates, radiometric corrections must be applied for atmospheric variability, i.e., each satellite scene requires independent correction to give accurate surface reflectivity values. Generally, modelling radiometric correction comprises three steps: firstly, the DN are corrected to radiance, then radiance is converted to apparent reflectance (the reflectance recorded at the sensor) and finally an atmospheric correction is performed to convert apparent reflectance to surface reflectance. The first of these steps uses the calibration coefficients obtained for each spectral band from the header file supplied with the SPOT scene, while the Landsat sensor can use either fixed pair calibration coefficients for each channel, or time-variant coefficients, as indicated in (Teillet and Fedosejevs 1995). The third step uses an inversion of the 5S (Simulation of the Satellite Signal in the Solar Spectrum) model with detail of images and atmospheric conditions, in order to obtain the surface reflectances (Tanré et al. 1990).

Extensive experiments have been carried out in order to test the performance of the proposed approach for combining classifiers in section 2, and the effect of independence on the classification accuracy. These experiments can be divided into three groups as described in the following subsections.

(i) Experiment 1: *Using the same features and different classifiers trained by independent data sets.*

In this experiment the same features (i.e., the three multispectral SPOT bands) and standard classifiers (e.g., maximum likelihood, minimum distance rule and artificial neural networks) trained

by independent data sets were used. Six sample sets were selected using stratified random sampling based on the reference image (ground truth), which was generated using the same scale and projection system as the remotely sensed data. Each sample had 200 patterns per class (total 1200 pixels). Four independent sample sets were used to train the classifiers and two independent sample sets (selected at random) were reserved respectively as the testing set to check the performance, and a set to generate the prior knowledge on the error of each classifier. The four training sample sets were individually used as input to each of four classifiers: Maximum Likelihood (ML), Minimum Distance Rule (MDR) method and two variants of an Artificial Neural Network (ANN and ANNT). The Gaussian maximum likelihood and minimum distance rule methods are well-known classification algorithms (e.g., Mather 1999). Both artificial neural network architectures chosen are multilayer perceptrons using the backpropagation algorithm (Lippmann 1987, Benediktsson et al. 1990, and Civco 1993). The only difference between the models is in the input layer. The first ANN model was implemented having one neurone per spectral band in the input layer. Therefore, this neural network had three nodes in the first layer. The alternative ANNT was modelled using a 3 x 3 window of pixel data from each band of the image, giving a total of 27 nodes in the input layer, as the input (Paola and Schowengerdt, 1995). This input modification takes local texture information into account. All neural network configurations tested had an output layer with six nodes, corresponding to the six crop classes. The number of hidden layers and the number of hidden nodes were found (1 hidden layer and 10 nodes) using the Hirose et al. (1991) procedure. The learning rate and momentum were kept constant at 0.2 and 0.9, respectively.

(ii) Experiment 2: *Using different features and different classifiers trained by independent data sets.*

The previous experiment has taken a step in the direction of combining the output of different classifiers using the same features. It is possible that using qualitatively different sets of feature variables that provide the lowest correlation among classifiers may produce more promising results. Therefore, the approach outlined in this study should be replicated using an independent set of features (e.g., different multispectral bands) and on several images (e.g. Landsat). In Experiment 2, two multispectral Landsat images were split into four independent sets of three features (i.e., three multispectral bands) each. Thus, bands 1, 2 and 3 of the Landsat image (27 June 1994) were used as discriminating variables in an ML classifier. The remaining bands (i.e., 4, 5 and 7) of this image were used as input to an ANN classifier. A MDR classifier was trained using the three first bands (i.e., bands 1, 2 and 3) of the second Landsat image (20 July 1994) as features, and the three remaining bands (i.e., bands 4, 5 and 7) of this image were used as input features to train the ANNT. All the classifiers were also trained by independent data sets. Six sample sets were also selected using stratified random sampling based on the reference image. Each sample has 120 patterns per class (total 720 pixels). Four independent sample sets were used to train the classifiers and two independent sample sets (selected at random) were reserved as the testing set to check the performance and to generate the prior knowledge on the error of each classifier, respectively. The second test data set was used to train the 2ANN, before the final accuracy assessment. The four independent training sample sets were individually used as input to each of four supervised classifiers: ML, MDR, ANN, and ANNT, as described above.

(iii) Experiment 3: *Using different features and the same classifiers trained by independent data sets.* An intuitive way to improve the classification accuracy would be to combine the best individual performance classifier (i.e., the one having the high accuracy for the four independent

sets of features) from the previous experiments. For instance, in this final experiment an ANNT classifier was selected, since this algorithm had the best individual accuracy when it was applied for the four independent sets of features. Moreover, several researchers (e.g., Tumer and Ghosh 1996, Kanellopoulos and Wilkinson 1997) argue that different neural networks (of the same type) also yield different class separation surfaces, depending on the network architectures and the starting weight sets of the network concerned. Thus, the two multispectral Landsat images were again split into four independent sets of features, each with three multispectral bands, as described in the previous subsection. Then, each group of three features was used as the set of discriminant variables to be input to four independent ANNT classifiers. The ANNT classifiers were also trained by four independent sample sets, and two independent sample sets (selected at random) were used as the testing set to check the performance and to generate the prior knowledge on the error of each classifier. Each sample has also 120 patterns per class (total 720 pixels).

Standard accuracy measures derived from a confusion matrix were computed. The measures based on the confusion matrix were overall accuracy, individual class accuracy, producer's accuracy and user's accuracy. The calculations associated with these measures are described in standard textbooks (Mather, 1999). However, the major interest in these experiments is to verify the performance of individual independent classifiers and the performance of the combined classifiers, so it is the Kappa coefficient (*Ka*), variance, and test *Z* statistics that are most important, rather than the individual accuracy of the classes, which are reported here.

In addition, a pairwise test statistics for testing the significance of the classifiers (represented here by their respective confusion matrices) were performed utilising the Kappa coefficients. These results are summarised in form of a *significance matrix*, in which the major diagonal elements indicate if the respective classification result is meaningful. In this single confusion matrix case, the *Z* value can be computed using the formula:

$$Z = Ka \Big/ \sqrt{\text{var}(Ka)} \tag{12}$$

where *Z* is standardised and normally distributed and *var* is the large sample variance of the Kappa coefficient *K*. If $Z \geq Z_{\alpha/2}$ the classification is significant better than a random classification, where $\alpha/2$ is the confidence level of the two-tailed *Z* test and the degrees of freedom are assumed to be infinity. On the other hand, the off diagonal elements give an indication, again if $Z \geq Z_{\alpha/2}$, that the two independent classifiers are significantly different. The formula used to test for significance between the two independent Kappa coefficients is:

$$Z = \left| Ka_1 - Ka_2 \right| \Big/ \sqrt{\text{var}(Ka_1) + \text{var}(Ka_2)} \tag{13}$$

where the Ka_1 and Ka_2 are the two Kappa coefficients in comparison. Congalton and Green (1999) present a comprehensive review of these formulations.

4 Results

In **Table 1(a)**, the classification results are evaluated using Kappa analysis and are summarised in the form of a significance matrix for the first experiment. In this table, the Kappa and variance values for accuracy assessment were obtained using Maximum Likelihood (ML), Artificial Neural Networks (ANN), Minimum Distance Rule (MDR) and Artificial Neural Networks Texture (ANNT) classifiers, respectively. The effects of combining or integrating the outputs of

Table 1(a)

these four classifiers using conservative vote (C. Vote), majority vote (M. Vote), comparative vote (Cmp. Vote), Bayesian Formalism using average (F.B.(ave)), Bayesian Formalism using belief (F.B.(bel)), evidential reasoning (ER), and a second Artificial Neural Network (2ANN), are displayed. The first set of independent test data was used as the test set to validate the performance of the classifications. The second test data set was used to generate the prior knowledge of the error of each classifier, and for training the second Artificial Neural Network (2ANN). Independent of the method or algorithm used, the pixels received the label of the output class having the highest probability or minimum distance.

Table 1(a)

It can be seen from **Table1(a)** that the ANNT (Kappa = 0.64) gives the best performance of the four individual classifiers. The performance of the combined classifier using 2ANN (Kappa = 0.766) is superior to ANNT (and thus to all other individual classifiers) in all aspects. The major diagonal elements (representing the single error matrices) show that all the classifications are significantly better than random results, at the 95% confidence level ($Z > 1.96$, the critical value $Z_{\alpha/2}$).

It is important to determine whether combining the classifier outputs can improve on individual classification accuracies. In order to determine this, consider the pairwise test of significance utilising the Kappa analysis in the *significance matrix*. Comparing the classifier performances (off diagonal elements), as expected, there are "positive" significant improvements for the individual classifier performances (yellow classifier pairs). However, there are some combination methods that reduce the accuracy of the individual classifiers (e.g., using the conservative vote principle). The majority of experiments suggest that the use of a second artificial neural network (2ANN) strategy is able significantly to improve the classification performance when using the same set of features (i.e., three spectral bands) and independent data sets. As might be expected, the conservative vote (C. Vote) was the worst strategy, due to the fact that, using this method, there needs to be an agreement between all the classifier outputs, otherwise the pattern is rejected. In addition, the majority and comparative vote methods present some significant improvement only in relation to MDR classifier performance, which suggests that the integrated method is not guaranteed to generate improved results for all situations, especially when using the same set of discriminant variables as input into the classification process.

With the exception of the ANNT, there are significant positive improvements using evidential reasoning (ER) and Bayesian Formalism using belief (F.B.(bel)) strategies. Predictably, the combination of these independent classifiers is significantly better than the individual use of an MDR algorithm in remote sensing data.

It must be emphasised that the previous observations are obtained from a case study on 1200 patterns (pixels) using the same three features as discriminant variables and considering independent sample sets and different classifiers. The combined error rates depend not only on the error rates of individual classifiers but also on the correlation between the feature variables used. Therefore, it is necessary to perform the whole process for different feature sets that would provide the lowest correlation between classifiers and on several images (e.g., Landsat).

In order to determine whether combining the classifier outputs using independent sets of features as discriminant variables can improve on individual classification accuracies (Experiment 2), a pairwise test of significance utilising Kappa analysis was performed and the

results are presented in the **Table 1(b).** This table shows that the ANNT (Kappa = 0.725) gives the best individual performance of the four individual classifiers. The major diagonal elements show that all the classifications are significantly better than a random allocation, at the 95 percent confidence level (computed $Z_{diagonal} > 1.96$).

Table 1(b)

From the evaluation of the results, it is noteworthy that the accuracy of the combined classification in relation to individual performance levels was improved significantly. Since each classifier uses different sets of feature variables, the correlation between them is reduced. Therefore, using the combination method of Bayesian Formalism, using both average and belief, evidential reasoning, or a second neural network is able significantly to improve the individual performances of ML, MDR, ANN or ANNT classifiers, since independent features are used as discriminant variables. Thus, clearly, the combined use of four different mathematical models in the classification process yields useful dividends, as proposed by Wilkinson *et al.* (1995).

The combination rule of majority vote can also be used to significantly improve the individual performance of ML, MDR, or ANN, but this fusion method gives a lower accuracy than the one achieved by the performance of an individual ANNT classifier. There are also some accuracy improvements in using the comparative vote combination rule in relation to individual performances (e.g., ML and MDR). However, Equation 3 was set up with an arbitrary α value of 0.5; it is possible that if an appropriate value for α is used, the results could be improved for the comparative vote rule. It is important to mention that the determination of an "appropriate" value of α is case dependent. Therefore, this value must be determined experimentally. For instance, Xu *et al.* (1992) found that α equal to 0.51 gives better results for the comparative vote rule when applied in the context of handwriting recognition. However, for remotely sensed data as used in this experiment, α should be 0.1 or 0.2 in order to ensure higher accuracy with a lower rate of unrecognised pixels.

Finally, the ANNT classifier was selected to perform the third experiment in order to examine whether classification accuracy can be improved by using the best individual performance classifier for the four independent sets of features. The results are presented in **Table 1(c).** There are some improvements on the individual classifier accuracy using this data set, as expected. However, some interesting results are exhibited by the choice of classification fusion (or combination) procedure. Using the combination method of majority vote, Bayesian Formalism - using both average and belief - evidential reasoning or a second neural network is able to significantly improve the performance of individual classifiers.

Table 1(c)

5 Conclusions

This article reviews the standard techniques of combining the outputs from different classifiers, and summarises recent results that statistically quantify such improvements. Preliminary results indicate that significant improvement can be obtained in difficult pattern recognition problems, such as those that involve a large amount of noise, limited number of training data pixels, or unusually high dimensional patterns (Tumer and Ghosh 1996).

On balance, combining different types of classifiers using an independent training data set can provide modest improvements in classification accuracy for remotely sensed data. However, combining the classifier outputs using independent sets of features as discriminant variables can significantly improve on individual classification accuracies. Moreover, high levels of accuracy can be reached when classifiers like ANNT are combined using qualitatively different feature sets and independent data sets, which provide the lowest correlation among classifiers. The

ANNT classifier has the advantage of implicitly incorporating local spatial variance into the classification process. In addition, a variant of the 2ANN approach has proven to be effective in combining the output of several classifiers before making the classification decision for all the experiments carried out in this research. The 2ANN fusion method is motivated by its ability to extract the valuable amount of information from individual classifiers.

This study has taken a step in the direction of combining the output of independent classifiers in the context of remote sensing. This is an area that would merit further research in order to determine its reliability and optimise its use. There are also many other problems to study, as pointed out by Xu et al. (1992), for example: how to generalise these approaches to combine dependent classifiers? How many classifiers are appropriate for a special problem? How to distribute a given number of feature variables to each classifier?

It is possible that by tackling these questions, new insights into the subject of pattern recognition and remote sensing could be added to the literature. Previously, the main effort focused on the design of one good classifier and the reduction of the high-dimensional feature vector in order to obtain high classification accuracy. Now, the focus can be changed and the variety of classifiers developed up to date should be seen not as competitors but as vital and complementary methods (Wilkinson et al. 1995). Nevertheless, a number of classifiers can be designed, which use low dimension feature vectors of different and complementary types, as argued by Xu et al. (1992). Although each classifier may not have an optimal performance, the appropriate combination of these individual classifiers may produce a high quality overall performance.

References

Battiti, R., and Colla, A.M. Democracy in neural nets: Voting schemes for classification. **Neural Networks**, 7, 1994, pp. 691-709.

Benediktsson, J. A., Swain, P. H., and Ersoy, O. K. Neural network approaches versus statistical methods in classification of multisource remote sensing data. **IEEE Transactions on Geoscience and Remote Sensing**, 28, 1990, pp. 540-552.

Civco, D. L. Artificial neural networks for land-cover classification and mapping. **International Journal of Geographic Information Systems**, 7, 1993, pp. 173-183.

Congalton, R. G., and Green, K. **Assessing the Accuracy of Remotely Sensed Data: Principles and Practices**. New York: Lewis Publishers, 1999.

Garvey, T., Lowrance, J., And Fischler, M. An inference technique for integrating knowledge from disparate sources. In: **Proceedings of the Seventh International Joint Conference on Artificial Intelligence**, Vancouver, British Columbia, 1979, pp. 319-325.

Garvey, T. Evidential reasoning for geographic evaluation for helicopter route planning. *IEEE Transactions on Geoscience and Remote Sensing*, GE-25, 1987, pp. 294-304.

Ghosh, J., Tumer, K., Beck, S., and Deser, L. Integration of neural classifiers for passive sonar signals. In: **DSP Theory and Applications**, C. T. Leondes, editor, Academic Press, 1995.

Hirose, Y., Yamashita, K., And Hijiya, S. Back-propagation algorithm which varies the number of hidden units. **Neural Networks**, 4, 1991, pp. 61-66.

Kanellopoulos, I., and Wilkinson, G. G. Strategies and best practice for neural network image classification, **International Journal of Remote Sensing**, 18, 1997, pp. 711-725.

Lee, T., Richards, J.A., And Swain P.H., Probabilistic and evidential approaches for multisource data analysis. **IEEE Transactions on Geoscience and Remote Sensing**, GE-25, 1987, 283-293.

Lippmann, R. P. An introduction to computing with neural nets. **IEEE ASSP Magazine**, April 1987, pp. 4-22.

Mather, P. M. **Computer Processing of Remotely-Sensed Images: An Introduction**. Second Edition, Chichester: John Wiley and Sons, 1999.

Paola, J. D., and Schowengerdt, R. A. A review and analysis of backpropagation neural networks for classification of remotely sensed multispectral imagery. **International Journal of Remote Sensing**, 16, 1995, pp. 3033-3058.

Perrone, M. P., And Cooper, L. N. When networks disagree: Ensemble methods for hybrid neural networks. In: **Neural Networks for Speech and Image Processing**, R. S. Mammone, editor, Chapter 10, Chapmann-Hall, 1993.

Roli, F., Giacinto, G., And Vernazza, G. Comparison and combination of statistical and neural networks algorithms for remote-sensing image classification, In: **Neurocomputation in Remote Sensing Data Analysis**. Berlin: Springer-Verlag, 1997, pp. 117-124.

Shafer, G. **A Mathematical Theory of Evidence**, Princeton, NJ: Princeton University Press, 1979.

Tanré, D., Deroo, C., Duhaut, P., Herman, M., Morcrette, J. J., Perbos, J., and Deschamps, P. Y. Description of a computer code to simulate the satellite signal in the solar spectrum: the 5S code. **International Journal of Remote Sensing**, 11(4), 1990, pp. 659-668.

Teillet, P. M., and Fedosejevs G. On the dark target approach to atmospheric correction of remotely sensed data. **Canadian Journal of Remote Sensing**, 21(4), 1995, pp. 374-387.

Tumer, K., and Ghosh, J. Boundary variance reduction for improved classification through hybrid networks (invited paper), In: Applications and Science of artificial Neural Networks, **Proceedings of the SPIE**, 2492, 1995, pp 573-58.

Tumer, K., and Ghosh, J. Theoretical foundation of linear and order statistics combiners for neural pattern classifiers, *IEEE Transactions on neural networks*, 1996, (URL http://www.lans.ece.u-texas.edu/~kagan/publications.html).

Xu, L., Krzyzak, A., and Suen, C. Y. Methods for combining multiple classifiers and their application to handwriting recognition, **IEEE Transactions on Systems, Man and Cybernetics**, 22, 1992, pp. 418-435.

Wilkinson, G. G., Fierens, F., and Kanellopoulos, I. Integration of neural and statistical approaches in spatial data classification, **Geographical System**, 2, 1995, 1-20.

Classifiers	ML	ANN	MDR	ANNT	C. Vote	M. Vote	Cmp. Vote	F. B. (AVE)	F.B.(BEL)	ER	2ANN
KAPPA	0.578	0.568	0.428	0.64	0.379	0.615	0.563	0.619	0.647	0.623	0.766
VAR	0.000268	0.000271	0.000293	0.000251	0.000175	0.000261	0.000219	0.000257	0.000247	0.000225	0.000188
ML	**35.31**										
ANN	0.43	**34.50**									
MDR	6.33	5.90	**25.00**								
ANNT	2.72	3.15	9.09	**40.40**							
C. Vote	9.45	8.95	2.27	12.65	**28.65**						
M. Vote	1.61	2.04	**7.94**	1.10	11.30	**38.07**					
Cmp. Vote	0.68	0.23	**5.97**	3.55	9.27	2.37	**38.04**				
F. B. (AVE)	1.79	**2.22**	**8.14**	0.93	11.55	0.18	2.57	**38.61**			
F.B.(BEL)	**3.04**	**3.47**	**9.42**	0.31	13.05	1.42	3.89	1.25	**41.17**		
ER	**2.03**	**2.47**	**8.57**	0.78	12.20	0.36	2.85	0.18	1.10	**41.53**	
2ANN	**8.80**	**9.24**	**15.41**	**6.01**	20.31	7.13	10.06	6.97	5.71	7.04	**55.87**

a) Significance Matrix for Experiment 1.

Classifiers	ML	ANN	MDR	ANNT	C. Vote	M. Vote	Cmp. Vote	F.B.(ave)	F.B.(bel)	ER	2ANN
KAPPA	0.462	0.587	0.417	0.725	0.163	0.721	0.536	0.812	0.833	0.808	0.833
VAR	0.000472	0.000451	0.000491	0.000352	0.000174	0.000332	0.000325	0.000264	0.000239	0.000268	0.000239
ML	**21.27**										
ANN	4.11	**27.64**									
MDR	1.45	5.54	**18.82**								
ANNT	9.16	4.87	10.61	**38.64**							
C. Vote	11.76	16.96	9.85	24.50	**12.36**						
M. Vote	**9.13**	**4.79**	**10.60**	0.15	24.81	**39.57**					
Cmp. Vote	**2.62**	1.83	**4.17**	7.26	16.70	7.22	**29.73**				
F.B.(ave)	**12.90**	**8.42**	**14.38**	**3.51**	31.01	3.73	11.37	**49.98**			
F.B.(bel)	**13.91**	**9.37**	**15.40**	**4.44**	32.97	4.69	12.51	0.94	**53.88**		
ER	**12.72**	**8.24**	**14.19**	**3.33**	30.68	3.55	11.17	0.17	1.11	**49.36**	
2ANN	**13.91**	**9.37**	**15.40**	**4.44**	32.97	4.69	12.51	0.94	0.00	1.11	**53.88**

b) Significance Matrix for Experiment 2.

Classifiers	ANNT(1)	ANNT(2)	ANNT(3)	ANNT(4)	C. Vote	M. Vote	Cmp. Vote	F.B.(AVE)	F.B.(BEL)	ER	2ANN
KAPPA	0.62	0.667	0.695	0.728	0.36	0.799	0.773	0.842	0.863	0.853	0.878
VAR	0.00043	0.0004	0.000378	0.000349	0.000285	0.00027	0.000275	0.000298	0.000202	0.000214	0.000182
ANNT(1)	**29.90**										
ANNT(2)	1.63	**33.35**									
ANNT(3)	2.64	1.00	**35.75**								
ANNT(4)	3.87	2.23	1.22	**38.97**							
C. Vote	9.72	11.73	13.01	14.62	**21.33**						
M. Vote	**6.77**	**5.10**	**4.09**	**2.85**	18.64	**48.63**					
Cmp. Vote	**5.76**	**4.08**	**3.05**	1.80	17.45	1.11	**46.61**				
F.B.(AVE)	**8.23**	**6.62**	**5.65**	**4.48**	19.96	1.80	2.88	**48.78**			
F.B.(BEL)	**9.67**	**7.99**	**6.98**	**5.75**	22.79	2.95	4.12	0.94	**60.72**		
ER	**9.18**	**7.51**	**6.49**	**5.27**	22.07	2.46	3.62	0.49	0.49	**58.31**	
2ANN	**10.43**	**8.75**	**7.73**	**6.51**	23.97	3.72	4.91	1.64	0.77	1.26	**65.08**

c) Significance Matrix for Experiment 3.

Table 1: Significance matrices for comparison of classifiers using Kappa Analysis. The tables also present the Kappa coefficients and variance of each classifier. The Z values along the major diagonal and the other Z values (off diagonal elements) were computed using the Equations 12 and 13 respectively. Painted classifier pairs indicate significant improvements in the performance of the classifiers at 95% confidence level (Z critical value = 1.96).