COMBINING BINARY CLASSIFIERS TO IMPROVE TREE SPECIES DISCRIMINATION AT LEAF LEVEL

Xolani Dastile[1], Gunther Jäger[2], Pravesh Debba[3], and Moses Cho[4]

[1]Anti-Corruption and Security, South African Revenue Services (SARS), xdastile@yahoo.com
[2]Department of Statistics, Applied University of Stralsund, gunther.jager@gmail.com
[3]Built Environment, Council for Scientific and Industrial Research (CSIR), pdebba@csir.co.za
[4]Natural Resource Environment (NRE), Council for Scientific and Industrial Research (CSIR), mcho@csir.co.za

ABSTRACT

This paper focuses on the discrimination of seven different savannah tree species at leaf level using hyperspectral data. The data is small in size, high-dimensional and shows large within-species variability combined with small between species variability which makes discrimination between the tree species (hereafter referred to as classes) challenging. We focus on two classification methods: K-nearest neighbour and feed-forward neural networks for the discrimination of the classes. For both methods, direct 7-class prediction results in high misclassification rates. We therefore construct binary classifiers for all possible binary classification problems and combine them using Error Correcting Output Codes (ECOC) to form a 7-class predictor. ECOC with 1-nearest neighbour binary classifiers result in no improvement compared to a 1-nearest neighbour 7-class predictor whereas ECOC with neural networks binary classifiers improve accuracy by 10% compared to neural networks 7-class predictor, and error rates become acceptable.

1. INTRODUCTION

Hyperspectral remote sensors record data (hereafter referred to as hyperspectral data) of objects using many contiguous wavelength bands of the electromagnetic spectrum. The hyperspectral data is obtained remotely either by mounting hyperspectral sensors on airplanes and satellites or by using hand-held devices. Possible applications of hyperspectral data include, but are not restricted to, mineral identification and mapping, urban settlement classification, plant species identification and forest monitoring (Schmidt & Skidmore, 2003).

The hyperspectral data that we were using was challenging for the following three reasons: (1) it was high dimensional, (2) it was small in size and (3) it showed high within-class variability and small between-class variability. For reason (1), usually some kind of "dimension reduction" is performed for hyperspectral data prior to classification. In this study we used sequential selection of bands, for example, we chose every 10th (20th or 30th) band, as a simple form of dimension reduction. We chose these bands because our experiments showed that these bands produced reasonable error probability estimates. For reason (2), it is difficult to make good classifiers and obtain at the same time reliable estimates of the error probabilities. Since the data set was small, we could not afford an independent test set but we needed all the data for constructing a good classifier. In order to obtain a reliable estimate of the error probability of a classifier, all the data in combination with K-fold cross-validation was used. For reason (3) we decided to use nearest-neighbour classifiers and neural networks since they do not rely on normality assumptions of the distributions. Both classifiers have good theoretical properties (Cybenko, 1989; Duda et al., 2001) and are often used in applications (O'Farrell et al., 2005). We showed in this paper that although classification of our tree species data was challenging, it was nevertheless possible and resulted in reasonable classification accuracies. Our results in using 7-class K-nearest neighbour and neural networks classifiers were unsatisfactory. However, binary classification where we, for example, distinguished one class versus the rest, showed acceptable misclassification rates. Hence, we used binary classifiers and combined their outputs (Lorena et al., 2008) for 7-class predictions. There are different ways of combining binary classifiers to obtain a multiclass classifier. Often majority votes have been suggested (Krebel, 1999). In 1995 a new approach for combining binary classifiers, called "Error Correcting Output Codes", was suggested in Dietterich & Bakiri (1995) and shown to be successful (Dietterich & Bakiri, 1995; Ghani, 2000). We therefore used this combining method as a possible solution to our problem.

The aim of this study is to improve the classification accuracy of the classifiers that are applied to several spectrally similar tree species data which has high dimensionality resulting in multicollinearity, and small number of observations within each class.

2. METHODS

We employed two classifiers, the K-nearest neighbour classifier and the Neural Networks classifier. For both classifiers we needed to determine optimal parameters. For K-nearest neighbour classifier, we determined an optimal number of nearest neighbours K, and for Neural Networks classifiers we determined (i) an optimal number of hidden layers and (ii) an optimal number of hidden neurons. We obtained this by splitting the data set randomly into a 70% training set (to construct the classifiers) and a 30% independent test set (to estimate the error probability of the classifiers). The optimal parameters are those parameters of the classifiers which resulted in small error probability estimates on the test set. For example, it was noticed in the plot on the left of Figure 1 that the number of nearest neighbours that gave a small error probability estimate on the test set was one, thus this value was chosen as the optimal value for nearest neighbours. Again, it was noticed in the plot on the right of Figure 1 that the optimal number of hidden

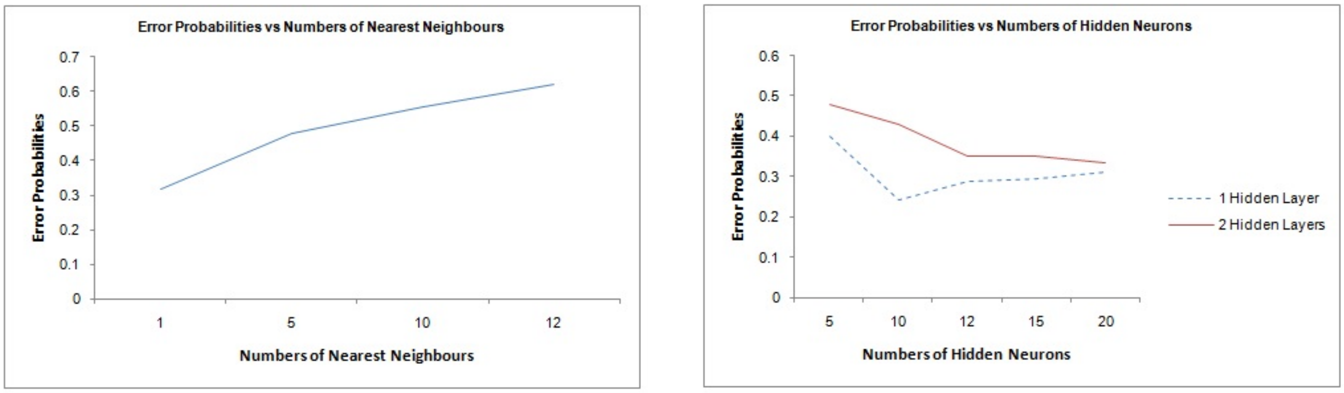


Fig. 1: Error probabilities vs Numbers of Nearest Neighbours (on the left) and Error probabilities vs Numbers of Hidden Neurons (on the right) using all the bands (i.e. wavelengths)

layers was one, since the one hidden layer produced small error probability estimates compared to two hidden layers, and 10 was the optimal number of hidden neurons when one hidden layer was used. Having found these optimal parameters, cross-validation was then used to obtain error probability estimates of the classifiers using these optimal parameters.

3. DATA SET

A hand-held Analytical Spectral Device (ASD) spectrometer (FieldSpec3 Pro FR) was used to record hyperspectral measurements of leaf samples taken from different savannah tree species in the Kruger National Park in South Africa, in an attempt to assess tree species diversity in the park. The study area is located in the “lowveld” savanna *biome* in the northeast of South Africa. The data was collected in May 2008. Seven common plant tree species in the area were considered: *Combretum apiculatum (CA), Combretum heroense (CH), Terminalia sericea (TS), Gymnosporia senegalensis (GS), Lonchocarpus capassa (LC), Gymnospora buxifolia (GB), and Combretum zeyheri (CZ)*. The number of observations for each species were 23, 20, 22, 18, 25, 21, and 19, respectively. The total data set therefore had 148 observations for the species measurements. The wavelength range of the data was $400\,$nm to $2500\,$nm at a spectral resolution of 1 *nm*. The dimension of the data was $2101 \times n$ (where 2101 was the number of wavelength bands which were treated as variables and n was the number of samples in each class). Thus our data set had high dimensionality.

Figure 2 (on the left) shows reflectance spectra of *Combretum apiculatum.* This reflectance spectra shows a large within-species variability. If we superimpose reflectance spectra of other species, there will be a strong overlap of the spectra. This is underlined in Figure 2 (on the right) which shows the mean reflectances of two classes (i.e. CA and TS species) together with the 2-σ bands (σ is the empirical standard deviation). Hence, this shows a small between-species variability, which enhances the problem

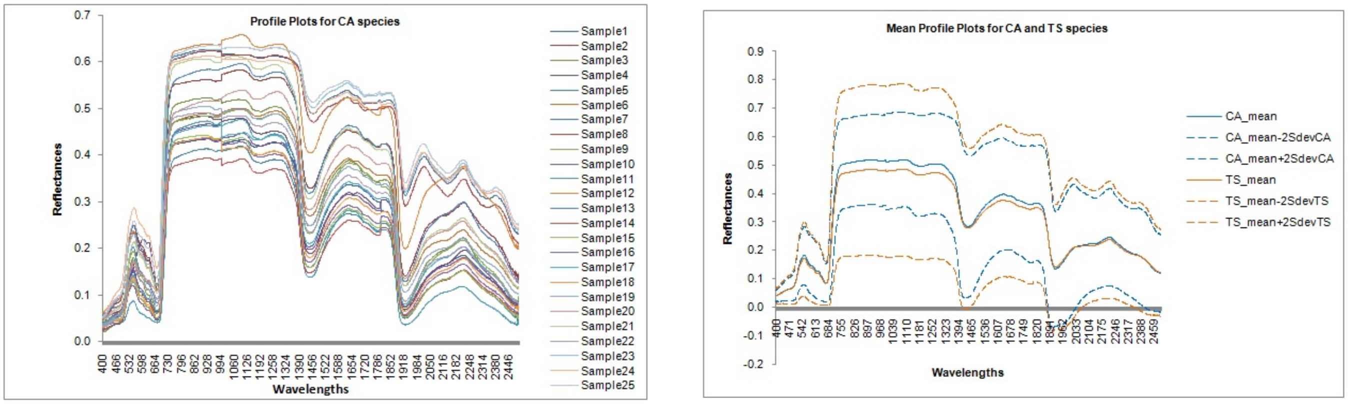


Fig. 2: 25 samples reflectance curves of Combretum Apiculatum (on the left) and Mean reflectances of CA and TS species with 2 σ bands (on the right)

of separating these species.

3.1. Classifiers

3.1.1. K-nearest neighbour classifier

The K-nearest neighbour classifier is a model-free classifier that provides good performance for optimal values of K. In the K-nearest neighbour decision rule, a test sample x is assigned to the class that is most frequently represented among the K closest training samples to x. Hence the K-nearest neighbour classifier is based on the distances from the sample x to the training samples.

3.1.2. Neural networks classifier

A feed-forward neural networks consists of layers with neurons. The first layer is the input layer, the following layers are hidden layers and the last layer is the output layer. In each neuron, for an input vector $(x_1, x_2, \ldots, x_N)'$ an output $y = f\left(b + \sum \nu_i x_i\right)$ is computed. Here, b and ν_i are the bias and the weights of the neuron, respectively, and f is a transfer function. This output, y, is either an output of the neural networks classifier or an input for the next layer of neurons. Hence, the data flow from the input layer to the hidden layer, from there to the next hidden layer and so on up to the output layer. We define for the output $(y_1, y_2, \ldots, y_P)'$ of the output layer, and a target output $(t_1, t_2, \ldots, t_P)'$, the error function $E = \sum_{i=1}^{P} (t_i - y_i)^2$. A feed forward neural networks classifier can be trained by adjusting the weights and biases to minimize E. In this study we used the *resilient back-propagation algorithm* (Riedmiller & Braun, 1993) for training. The neural networks toolbox (version 5.0.2 (R2007a))) of MATLAB was used. The training parameter “goal” was set to 0.03. Training is stopped if the error function falls below this value. The training data was presented to the neural networks for 1000 epochs. Our experiments showed that these values for epochs and goal parameter were optimal for training our neural networks.

3.2. Combining Binary Classifiers

3.2.1. Binary classifiers

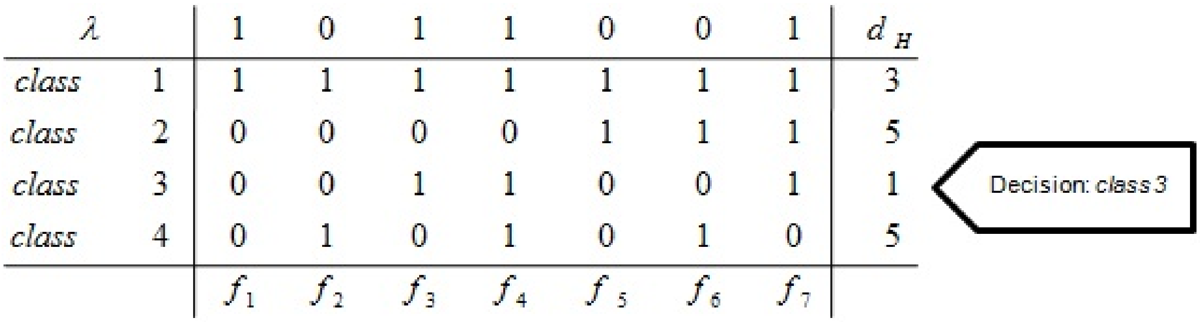
A classification problem which involves only two classes is called a binary classification problem. The classifiers for binary classification problems are called binary classifiers.

3.2.2. Combining binary classifiers to solve a multiclass problem: Error Correcting Output Codes (ECOC)

To combine the binary classifiers using ECOC (Jolliffe, 1970) we define a code matrix $M \in \{0,1\}^{c \times l}$, where c denotes the number of classes and l denotes the number of binary classifiers. If $\mathcal{C} = \{\omega_1, \omega_2, \ldots, \omega_c\}$ is the set of class labels, where ω_i denotes class i for $i = 1, 2, \ldots, c$, then each column of M represents a partition of $\mathcal{C}$ into metaclasses C^+ and C^- coded by 1 if $\omega_i \in C^+$ and 0 if $\omega_i \in C^-$, then each partition corresponds to exactly one binary classifier of length c. For a $c-$class classification problem, the number of binary classification problems that can be formed in this way is $2^{c-1} - 1$. If all binary classifiers predict the correct metaclasses, then each row of M can be interpreted as a class.

The output coding for a binary classifier f_i when evaluating an unknown object with feature vector x (hereafter referred to as object x) having class label $\omega(x)$ is $f_i(x) = 1$ if $\omega(x) \in C^+$ and $f_i(x) = 0$ if $\omega(x) \in C^-$ for $i = 1, 2, 3, \ldots, l = 2^{c-1} - 1$. The classification of a new object x using ECOC results in binary vector $\lambda = (f_1(x), f_2(x), \ldots, f_l(x))$ which is then compared to each row of M using some distance measure. One of the possible distance measures which is used is the Hamming distance (Diamond & Kloeden, 1994), d_H, i.e. the number of bit positions where two binary vectors differ. If the i^{th} row of M has the smallest Hamming distance to λ, class ω_i will be assigned as class of x. ECOC can only correct up to $h = \left\lfloor \frac{d_{H_{min}} - 1}{2} \right\rfloor$ wrong bits where $d_{H_{min}}$ is the minimum Hamming distance between the rows of M (Dietterich & Bakiri, 1995). Figure 3 shows an example of a code matrix for a 4-class problem with seven binary classifiers and classification of a new sample.

For a good ECOC, the rows and the columns of M must be well separated according to the Hamming distance (Dietterich & Bakiri, 1995). It is therefore suggested in Dietterich & Bakiri (1995) to use exhaustive codes for classification problems with $3 \leq c \leq 7$. An optimal exhaustive code is constructed in the following way (see Dietterich & Bakiri (1995)): Row 1 of M consists of all ones; row 2 consists of 2^{c-2} zeros followed by $2^{c-2} - 1$ ones; row 3 consists of 2^{c-3} zeros, followed by 2^{c-3} ones, followed by 2^{c-3} zeros, followed by $2^{c-3} - 1$ ones; in general, in row i, there are alternating runs of 2^{c-i} zeros and ones. This is illustrated in Figure 3 for a 4-class classification problem. Note that for our tree species hyperspectral data set, the values of c and l are seven (7) and sixty-three (63) respectively.



λ		1	0	1	1	0	0	1	d_H
class	1	1	1	1	1	1	1	1	3
class	2	0	0	0	0	1	1	1	5
class	3	0	0	1	1	0	0	1	1
class	4	0	1	0	1	0	1	0	5
		f_1	f_2	f_3	f_4	f_5	f_6	f_7	

Decision: *class 3*

Fig. 3: The binary vector λ for a new sample has the smallest Hamming distance to *class* 3, hence the new sample is assigned to *class* 3.

4. RESULTS

4.1. The 7-class predictors versus ECOC combiners

Table 1 shows that the neural networks (7-class) classifier outperforms the 1-nearest neighbour (7-class) classifier by roughly 7%. However, even neural networks classifiers in Table 1 are not good enough for practical use. We see that 1-nearest neighbour ECOC combiner and 1-nearest neighbour (7-class) classifier give identical error probability estimates. It is not difficult to show theoretically that this will always be the case for ECOC combination of 1-nearest neighbour classifiers. The neural networks ECOC combiner improves the misclassification rate by about 10% for all selected bands.

Table 1: Averages of the error probability estimates on the test set for seven class classifiers and ecoc, the averages of the standard errors of the error probability estimates in brackets ().

	Bands			
Classifiers	**all bands**	**every 10th**	**every 20th**	**every 30th**
1-Nearest_Neigh (7-class)	0.3356 (0.0391)	0.3409 (0.0392)	0.3340 (0.0390)	0.3353 (0.0390)
Neural_Net (7-class)	0.2772 (0.0370)	0.2401 (0.0352)	0.2582 (0.0359)	0.2528 (0.0358)
1-Nearest Neigh ECOC	0.3356 (0.0391)	0.3409 (0.0392)	0.3340 (0.0390)	0.3353 (0.0390)
Neural_Net ECOC	0.1480 (0.0294)	0.1561 (0.0299)	0.1415 (0.0286)	0.1492 (0.0294)

5. DISCUSSION AND CONCLUSION

Hyperspectral data of seven different tree species in Kruger National Park were used to construct 7-class predictors. The hyperspectral data showed large within-class variability and small between-class variability. This study has shown that the 7-class classifiers (1-nearest neighbour classifiers and neural networks classifiers) performed poor on the test sets resulting in error probability estimates of approximately 33% for the 1-nearest neighbour classifier and 26% for the neural networks classifier. As a possible solution, we decomposed the 7-class classification problem into a set of binary classification problems. For each of these binary classification problems, a binary classifier was constructed and combined using Error Correcting Output Codes (ECOC) to form a 7-class predictor. We found that the ECOC predictor which was formed by combining 1-nearest neighbour binary classifiers, classified exactly in the same way as the 7-class 1-nearest neighbour classifier. The ECOC predictor which was formed by combining neural networks binary classifiers showed improvements of about 10% in the misclassification rate.

One major limitation of the study is the small sample size (i.e. 148 samples with 7 classes). This limitation results in not having a good classifier and a good estimate of its error probability at the same time. Although collecting hyperspectral field samples at leaf level is very time consuming and expensive, this type of research will typically lead to proper analysis of hyperspectral satellite data. With satellite data, although it is still expensive, one could analyse large areas with much more samples. So this is a preliminary study if such hyperspectral satellite data can be used to discriminate between spectrally similar species. Another problem of this study is the small between-class variability combined with a high within-class variability. Despite these limitations, this study shows that tree species discrimination using hyperspectral data is feasible. For future study, we suggest feature extraction methods (e.g. Principal Components Analysis, see Jolliffe (1986)) or feature selection methods (e.g. classification trees, see Breiman et al. (1998)) to reduce dimensionality of the hyperspectral data. An example of a systematic way for selecting "useful bands" is presented in Mutanga & Adam (2009) for the discrimination between papyrus vegetation species.

REFERENCES

Breiman, L., Friedman, J. H., & Olshen, R. A. (1998). *Classification and Regression Trees*. Chapman & Hall/CRC.

Cybenko, G. (1989). Approximation by superpositions of sigmoidal functions. *Math. of Control, Signal and Systems*, 2, 303–314.

Diamond, P. & Kloeden, P. (1994). *Metric Spaces of Fuzzy Sets Theory and Applications*. World Scientific Publishing Co. Pty. Ltd.

Dietterich, T. G. & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2, 263–286.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons, Inc., New York., 2nd edition.

Ghani, R. (2000). Using error-correcting codes for text classification. In *Proceedings of the Seventeenth International Conference on Machine Learning*: Morgan Kaufmann Publishers Inc., San Fransisco.

Jolliffe, I. T. (1970). *An Introduction to Error-Correcting Codes*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Jolliffe, I. T. (1986). *Principal Components Analysis*. Springer-Verlag, New York Inc.

Krebel, U. H. G. (1999). Pairwise classification and support vector machines. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods: Support vector learning* (pp. 255–268).: MIT Press, Cambridge, MA.

Lorena, A. C., Carvalho, A. C., & Gama, J. M. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, 30, 19–37.

Mutanga, O. & Adam, E. (2009). Spectral discrimination of papyrus vegetation (cyperus papyrus l.) in swamp wetlands using field spectrometry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 6, 612–620.

O'Farrell, M., Lewis, E., Flanagan, C., Lyons, W., & Jackman, N. (2005). Comparison of k-nn and neural network methods in the classification of spectral data from an optical fibre-based sensor system used for quality control in the food industry. *Sensors and Actuators B*, 111(112), 354–362.

Riedmiller, M. & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *IEEE International Conference On Neural Networks*, (pp. 586–591).

Schmidt, K. S. & Skidmore, A. K. (2003). Spectral discrimination of vegetation types in coastal wetland. *Remote Sensing of Environment*, 26, 92–108.