FEATURE SELECTION BY USING CLASSIFICATION AND REGRESSION TREES (CART)

H. R. Bittencourt [a, *], R. T. Clarke [b]

[a] Faculty of Mathematics, Pontifícia Universidade Católica do RS, Porto Alegre, Brazil - heliorb@pucrs.br
[b] CEPSRM, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, clarke@iph.ufrgs.br

KEY WORDS: Hyper-spectral, Classification, Feature Extraction, Remote Sensing, Pattern Recognition, Agriculture

ABSTRACT:

Hyper-spectral remote sensing increases the volume of information available for research and practice, but brings with it the need for efficient statistical methods in sample spaces of many dimensions. Due to the complexity of problems in high dimensionality, several methods for dimension reduction are suggested in the literature, such as Principal Components Analysis (PCA). Although PCA can be applied to data reduction, its use for classifying images has not produced good results. In the present study, the Classification and Regression Trees technique, more widely known by the acronym CART, is used for feature selection. CART involves the identification and construction of a binary decision tree using a sample of training data for which the correct classification is known. Binary decision trees consist of repeated divisions of a feature space into two sub-spaces, with the terminal nodes associated with the classes. A desirable decision tree is one having a relatively small number of branches, a relatively small number of intermediate nodes from which these branches diverge, and high predictive power, in which entities are correctly classified at the terminal nodes. In the present study, AVIRIS digital images from agricultural fields in the USA are used. The images were automatically classified by a binary decision tree. Based on the results from the digital classification, a table showing highly discriminatory spectral bands for each kind of agricultural field was generated. Moreover, the spectral signatures of the cultures are discussed. The results show that the decision trees employ a strategy in which a complex problem is divided into simpler sub-problems, with the advantage that it becomes possible to follow the classification process through each node of the decision tree. It is emphasized that it is the computer algorithm itself which selects the bands with maximum discriminatory power, thus providing useful information to the researcher.

1. INTRODUCTION

1.1 Hyper-spectral Sensors and Dimensionality Reduction

In the last years, advances in sensor technology have made possible the acquisition of images on several hundred spectral bands. AVIRIS and HYDICE sensors are well known examples of this technology, having 224 and 210 bands, respectively. Since a great volume of data information is made available to researchers by means of hyper-spectral remote sensing, some problems can occur during the image classification process. When a parametric classifier is used, the parameters estimation becomes problematic in high dimensionality. In the traditional Gaussian Maximum Likelihood classifier, for example, the underlying probability distributions are assumed to be multivariate Normal and the number of parameters to be estimated can be very large, since with k classes, k mean vectors (of dimension px1) and k covariance matrices (dimension pxp, symmetric) are estimated (Bittencourt and Clarke, 2003a). As stated by Haertel and Landgrebe (1999), one of the most difficult problems in dealing with high dimensional data resides in the estimation of the classes' covariance matrices. Methods to solve this problem have received considerable attention from the scientific community and one way to solve this problem is the reduction of dimensionality.

There are two main reasons for keeping the dimensionality as small as possible: measurement cost and classification accuracy (Jain et al., 2000). The reduction of dimensionality is highly recommended when the number of training samples is limited, but, on the other hand, a reduction may lead to loss in the discrimination power between the classes.

1.2 Statistical Pattern Recognition and Feature Extraction

In the statistical approach to pattern recognition, each pattern is regarded as a p-dimensional random vector, where p is the number of characteristics used in classification that compose the feature space. Normally, when spectral attributes are used only, the pixels are the patterns and the p spectral bands to match up feature space. Several selection methods to determine an appropriate subspace of dimensionality m ($m < p$) in the original feature space are found in the literature. Probably the most widely-known technique is Principal Component Analysis (PCA) or Karhunen-Loève expansion, although the literature alerts to problems when PCA is used (Cheriyadat and Bruce, 2003). Although PCA is an excellent tool for data reduction, it is not necessarily an appropriate method for feature extraction when the main goal is classification, because PCA analyses a covariance matrix constructed from the entire data distribution, which does not represent the underlying class information present in the data.

This paper shows an alternative to data reduction demonstrating the ability of Classification and Regression Trees (CART) to determinate bands with highly discriminatory power between classes. This procedure is also known as feature selection.

[*] Corresponding author.

2. CLASSIFICATION AND REGRESSION TREES (CART)

As discussed in Bittencourt and Clarke (2003b), binary decision trees for classification can be viewed as a non-parametric approach to pattern recognition. A decision tree provides a hierarchical representation of the feature space in which patterns x_i are allocated to classes w_j (j=1,2,...,k) according to the result obtained by following decisions made at a sequence of nodes at which branches of the tree diverge. The type of decision tree used in this paper is discussed in detail by Breiman et al. (1984), whose contributions have been summarized by the letters CART (Classification And Regression Trees). These letters indicate that trees may be used not only to classify entities into a discrete number of groups, but also as an alternative approach to regression analysis in which the value of a response (dependent) variable is to be estimated, given the value of each variable in a set of explanatory (independent) variables.

Binary decision trees consist of repeated divisions of a feature space into two sub-spaces, with the terminal nodes associated with the classes w_j. A desirable decision tree is one having a relatively small number of branches, a relatively small number of intermediate nodes from which these branches diverge, and high predictive power, in which entities are correctly classified at the terminal nodes.

2.1 How Does CART Operate?

CART involves the identification and construction of a binary decision tree using a sample of training data for which the correct classification is known. The numbers of entities in the two sub-groups defined at each binary split, corresponding to the two branches emerging from each intermediate node, become successively smaller, so that a reasonably large training sample is required if good results are to be obtained (McLachlan , 1992).

The decision tree begins with a root node t derived from whichever variable in the feature space minimizes a measure of the impurity of the two sibling nodes. The measure of the impurity or entropy at node t, denoted by $i(t)$, is as shown in the following equation :

$$i(t) = -\sum_{j=1}^{k} p(w_j \mid t) \log p(w_j \mid t) \qquad (1)$$

where $p(w_j \mid t)$ is the proportion of patterns $\mathbf{x}_i$ allocated to class w_j at node t. Each non-terminal node is then divided into two further nodes, t_L and t_R, such that p_L , p_R are the proportions of entities passed to the new nodes t_L, t_R respectively. The best division is that which maximizes the difference given in:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \qquad (2)$$

The decision tree grows by means of the successive sub-divisions until a stage is reached in which there is no significant decrease in the measure of impurity when a further additional division s is implemented. When this stage is reached, the node t is not sub-divided further, and automatically becomes a terminal node. The class w_j associated with the terminal node t is that which maximizes the conditional probability $p(w_j \mid t)$.

2.2 CART applied to Feature Selection

The decision tree generated by CART uses only the bands that help to separate the classes, while the others are not considered. In this paper we use the tree as a feature selection method to reduce dimensionality. As an illustration, an image with only six spectral bands was classified using a decision tree; only two of the six points were identified by the CART procedure as necessary to separate three classes, as shown in Figure 1.

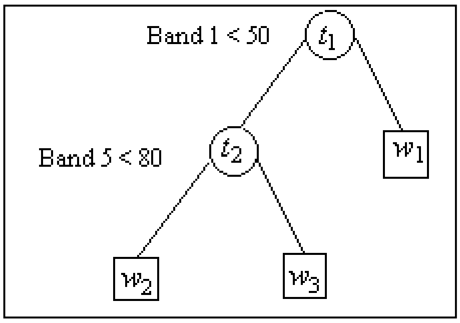

Figure 1. Example of decision tree using synthetic data

Figure 2 shows the subdivision of the feature space determined by the decision tree, where the point representing the patterns. As the tree used only two spectral bands, is possible to present the subdivision of feature space in the plane.

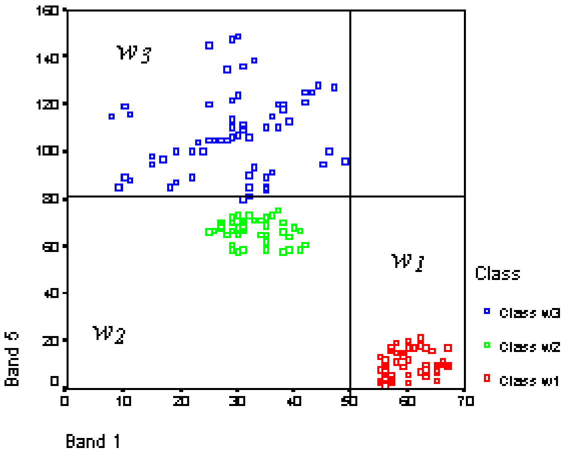

Figure 2. Subdivisions of the feature space determined by the decision tree (synthetic data)

This example is over-simplified because the solution can be presented as a two-dimensional scatter-plot. However hyperspectral images classifications generally require solutions in space of many dimensions. Another way to show the decision trees´ results is the following:

1 Band 1 > 50 Class 1 (Terminal)
Band 1 < 50 2 (3)
2 Band 5 < 80 Class 2 (Terminal)
Band 5 > 80 Class 3 (Terminal)

The form showed in (3) can be used when the classification tree contains many nodes, which complicate graphical representation. The next section gives results from applying CART to a high-dimensional classification problem with real data. The trees were constructed by software GenStat.

3. RESULTS USING HYPERSPECTRAL IMAGES

Two image segments of a rural area in the state of Indiana-USA collected by the sensor AVIRIS were analysed and classified using CART. Although the AVIRIS sensor has some 220 bands, only 195 were used because 25 bands were excluded during the pre-processing due to presence of the noise.

3.1 AVIRIS - Scene 1

In the first segment, three classes with relatively similar spectral response were considerate: w_1: woods, w_2: soybean and w_3: corn. The spectral behaviour is presented in the Figure 3.

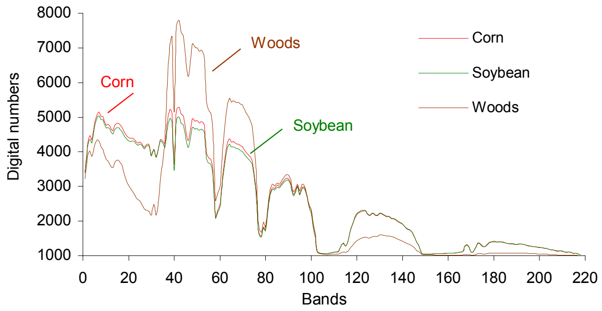

Figure 3. Mean spectral behaviour of the three classes denoted by woods, soybean and corn

Reading only the data for Woods and Soybean, a very simple tree classifies the 1747 pixels into two classes. The tree has only three nodes, two of which are terminal nodes, and the dichotomy is based solely on Band 9, as shown by Figure 4. If the count in Band 9 is less than 4204, the pixel is classed as Woods; otherwise it is Soybean. The misclassification rate is zero.

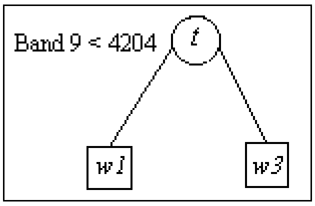

Figure 4. Decision tree to separate Woods (w_1) from Corn (w_3)

Similarly, the separation of Woods from Corn is equally simple, and depends on the value in Band 10. If the count in Band 10 is less than 3964, the pixel is classed as Woods; otherwise it is Corn. Again, the misclassification rate is zero.

The separation of Soybean from Corn is more complex, as shown by Table 1 and Appendix A. In this figure, the separation requires 117 nodes, 59 of which are terminal nodes, and the misclassification rate is 0.064 (6.4%).

Pair of Classes	*Number of spectral bands selected by the decision tree / Misclassification rate*	*Spectral bands*
Woods and Soybean	*01 / 0.0%*	*9*
Woods and Corn	*01 / 0.0%*	*10*
Soybean and Corn	*35 / 6.4%*	*1, 2, 3, 4, 5, 8, 11, 12, 14 33, 35, 39, 49, 69, 75, 77, 80, 103, 107, 114, 134, 141, 142, 144, 145, 146, 151, 152, 165, 172, 173, 179, 181, 188, 195*

Table 1. Bands selected to discriminate between pairs of classes when CART is applied and misclassification rate

CART really can operate as a data reduction technique, because is possible to identify and retain those spectral bands which result in small misclassification rates. From inspection of the mean spectral behaviour of the classes, shown in Figure 3, it can easily be seen that the separation of woods from the other classes is easer than separating corn from soybean.

3.2 AVIRIS - Scene 2

The second image considered presents three classes of corn with very similar spectral response: w_1: corn, w_2: corn minimum and w_3: corn no till. The mean spectral behaviour is presented in the Figure 5.

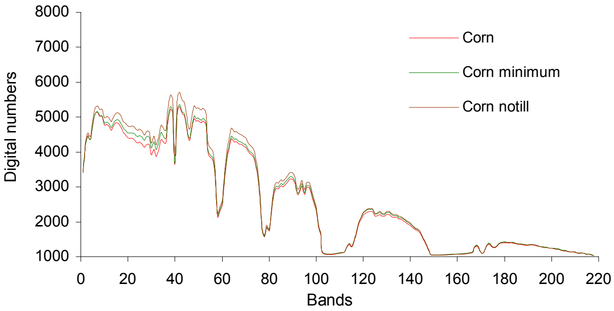

Figure 5. Mean spectral behaviour of the three classes denoted by corn, corn minimum and corn no till

Construction of decision trees for the three corn classes, taken in pairs, shows that a considerable reduction in the number of spectral bands is possible, giving misclassification rates about 1%, as shown Table 2.

Pair of Classes	*Number of spectral bands selected by the decision tree / Misclassification rate*	*Spectral bands*
Corn and Corn no till	*28 / 1.0%*	*1, 2, 3, 4, 11, 37, 38, 41 59, 103, 105, 106, 107, 110, 139, 140, 142, 143 144, 150, 153, 157, 161 173, 178, 185, 193, 194*
Corn and Corn min.	*31 / 1.1%*	*1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 23, 40, 48, 77, 78, 92, 103, 105, 106, 108, 109, 138, 142, 143, 144, 164, 183, 186, 188, 190*
Corn min and Corn no till.	*22 / 1.2%*	*1, 2, 3, 5, 6, 8, 9, 75, 93, 96, 102, 106, 124, 144, 153, 162, 172, 173, 174, 181, 190, 193*

Table 2. Bands selected to discriminate between pairs of classes when CART is applied and misclassification rate

In despite of the similar spectral response of the classes, CART can differentiate between them using fewer bands than those available in the full image.

4. FINAL REMARKS

The main conclusion of this work is that decision trees can be used to select features, even in high-dimensional space. When classes with very different spectral response are used, the data reduction is interesting because the spectral bands most useful for separating classes are identified. The results were even satisfactory for distinguishing between classes with similar spectral response, since it was possible to reduce the dimensionality considerably, whilst securing low rates of misclassification.

The results show that decision trees employ a strategy in which a complex problem is divided into simpler sub-problems, with the advantage that it becomes possible to follow the classification process through each node of the decision tree.

The software used (GenStat) can construct decision trees for separating three classes, in space of dimension $p=195$ and with numbers of pixels per class greater than 1500, in less than two minutes on a desk-top PC. This suggests that the use of the CART procedure to identify bands is a viable procedure. It must be emphasized that it is the GenStat algorithm itself that decides which spectral bands are to be retained, and which are to be discarded, when it constructs the decision tree.

References

Bittencourt, H. R., Clarke, R. T. 2003a. Logistic Discrimination Between Classes with Nearly Equal Spectral Response in High Dimensionality. *Proc. 2003 IEEE International Geoscience and Remote Sensing Symposium*, Toulouse, France, July 21-25, pp. 3748-3750.

Bittencourt, H. R., Clarke, R. T. 2003b. Use of Classification and Regression Trees (CART) to Classify Remotely-Sensed Digital Images. *Proc. 2003 IEEE International Geoscience and Remote Sensing Symposium*, Toulouse, France, July 21-25, pp. 3751-3753.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. 1984. *Classification and Regression Trees*. Wadsworth, Belmont-CA.

Cheriyadat and Bruce, 2003. Why Principal Component Analysis is not an Appropriate Feature Extraction Method for Hyperspectral Data. *Proc. 2003 IEEE International Geoscience and Remote Sensing Symposium*, Toulouse, France, July 21-25, pp. 3420-3422.

Haertel, V., Landgrebe, D., 1999. On the classification of classes with nearly equal spectral response in remote sensing hyperspectral image data. *IEEE Transactions on Geoscience and Remote Sensing*, 37 (5), pp. 2374-2386.

McLachlan, G. J., 1992. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience, New York, pp. 323-332.

Acknowledgments

The authors would like to thank Victor Haertel and Denis Altieri from Centro Estadual de Pesquisas em Sensoriamento Remoto, RS, Brazil for contributions to the work. H.R.B. extends his warmest thanks to the PUCRS, especially Prof. Alaydes Biachi, for supporting this work.

APPENDIX A

This appendix shows the GenStat output to separate Corn from Soybean.

```
***** Summary of classification tree:
      Corn and Soya

 Number of nodes:              117
 Number of terminal nodes:      59
 Misclassification rate:     0.064

 Variables in the tree:
B[3]  , B[146], B[188], B[11] , B[69] , B[2]  ,
B[49] , B[35] , B[172], B[103], B[144], B[152],
B[134], B[33] , B[145], B[77] , B[39] , B[14] ,
B[80] , B[4]  , B[1]  , B[5]  , B[181], B[173],
B[8]  , B[114], B[75] , B[107], B[165], B[142],
B[141], B[12] , B[151], B[195], B[179].

  1 B[3]<4600 2
    B[3]>4600 50
  2 B[146]<1228 3
    B[146]>1228 21
  3 B[11]<4162 4
    B[11]>4162 7
  4 B[35]<3721 5
    B[35]>3721 soya
  5 B[152]<1224 6
    B[152]>1224 corn
  6 B[14]<4120 corn
    B[14]>4120 soya
  7 B[172]<1252 8
    B[172]>1252 14
  8 B[134]<1648 corn
    B[134]>1648 9
 9 B[80]<1654 10
   B[80]>1654 13
10 B[181]<1162 11
   B[181]>1162 soya
11 B[8]<4453 corn
   B[8]>4453 12
12 B[103]<1112 soya
   B[103]>1112 corn
13 B[14]<4208 corn
   B[14]>4208 soya
14 B[33]<4088 15
   B[33]>4088 soya
15 B[39]<3917 soya
   B[39]>3917 16
16 B[4]<3902 soya
   B[4]>3902 17
17 B[107]<1290 soya
   B[107]>1290 18
18 B[141]<1116 soya
   B[141]>1116 19
19 B[151]<1324 20
   B[151]>1324 soya
20 B[3]<4061 corn
   B[3]>4061 corn
21 B[69]<4684 22
   B[69]>4684 corn
22 B[3]<4496 23
   B[3]>4496 45
23 B[145]<1328 24
   B[145]>1328 37
24 B[4]<4422 25
   B[4]>4422 corn
25 B[173]<1286 26
   B[173]>1286 corn
26 B[77]<1634 27
   B[77]>1634 31
27 B[2]<4310 28
   B[2]>4310 corn
28 B[195]<1035 29
   B[195]>1035 corn
29 B[149]<1208 soya
   B[149]>1208 30
30 B[140]<1266 soya
   B[140]>1266 soya
31 B[12]<4328 corn
   B[12]>4328 32
32 B[179]<1216 33
   B[179]>1216 corn
33 B[2]<4542 34
   B[2]>4542 corn
34 B[106]<1276 35
   B[106]>1276 corn
35 B[2]<4369 36
   B[2]>4369 soya
36 B[189]<1080 soya
   B[189]>1080 soya
37 B[1]<3190 soya
   B[1]>3190 38
38 B[8]<4644 corn
   B[8]>4644 39
39 B[165]<1328 corn
   B[165]>1328 40
40 B[103]<1112 41
   B[103]>1112 soya
41 B[142]<1074 42
   B[142]>1074 corn
42 B[86]<3269 43
   B[86]>3269 soya
43 B[195]<1042 44
   B[195]>1042 soya
44 B[2]<4385 soya
   B[2]>4385 soya
45 B[144]<1244 corn
   B[144]>1244 46
46 B[146]<1294 47
   B[146]>1294 48
47 B[114]<2382 soya
   B[114]>2382 corn
48 B[75]<3682 49
   B[75]>3682 corn
49 B[142]<1060 soya
   B[142]>1060 soya
50 B[188]<1110 51
   B[188]>1110 54
51 B[2]<4014 soya
   B[2]>4014 52
52 B[103]<1080 soya
   B[103]>1080 53
53 B[77]<1796 corn
   B[77]>1796 corn
54 B[49]<5560 55
   B[49]>5560 56
55 B[146]<1288 corn
   B[146]>1288 soya
56 B[144]<1305 57
   B[144]>1305 soya
57 B[39]<5798 soya
   B[39]>5798 58
58 B[5]<5384 corn
   B[5]>5384 corn

End *****
```