Spectral Attributes Selection based on Data Mining for Remote Sensing Image Classification

Raian V. Maretto[1,2], Thales S. Körting[2], Emiliano F. Castejon[2], Leila M. G. Fonseca[2], Rafael Santos[3]

[1]Fundação de Ciências, Aplicações e Tecnologias Espaciais (FUNCATE)
São José dos Campos – SP – Brazil.

[2]Image Processing Division – National Institute for Space Research (INPE)
São José dos Campos – SP – Brazil.

[3]Applied Computing and Mathematics Associated Laboratory – National Institute for Space Research (INPE)
São José dos Campos – SP – Brazil.

{raian, thales, castejon, leila}@dpi.inpe.br, rafael.santos@inpe.br

Abstract. *Remote sensing images are a rich source of information for studying large-scale geographic areas. The new satellite generations have producing huge amounts of data. Data mining techniques have been emerged last years as powerful tools to help in the analysis of these data. In the area of remote sensing image analysis, software like GeoDMA, eCognition, InterIMAGE, and others are available for end users. These software provides tools to extract several attributes of the images. These attributes are then used in image classification and analysis. When dealing with high resolution multispectral satellites, we have a large quantity of attributes. In many cases, the attributes are highly correlated, and consequently may not help to separate the classes of interest. Thus, this work shows the results of an approach to analyze the correlation of the attributes between several classes of interest, selecting those that will better distinguish them. In this way, it is possible to reduce the amount of data to be used during classification and analysis, consequently reducing the computational time for classification.*

1. Introduction

The increased accessibility of the new generation high-spatial resolution multispectral sensors has improved the level of complexity required in the analysis techniques. In particular, many traditional per-pixel analysis may not be suitable to high-spatial resolution imagery, due to its high-frequency components and the horizontal layover caused by off-nadir look angles [Im et al. 2008]. Aiming to overcome this problem, in the last decades, several approaches and platforms have been developed with algorithms that consider contextual information and pixel region properties [Körting et al. 2013; Syed et al. 2005; Walter 2004].

Current software can extract several statistical, spatial, color, texture or topological attributes. However, most of them often do not help to distinguish between the classes of interest, due to its high correlation. Thus, the attributes selection phase often relies on *ad hoc* decisions about what of them can better describe the classes. The huge number of attributes available makes a detailed exploratory time-consuming and dependent on expertise [Körting et al. 2013]. Many works have proved that data mining techniques can be useful to this purpose [Dash and Liu 1997; Kohavi and Kohavi 1997; Laliberte et al. 2012].

In this context, the main objective of this work is to analyze the correlation of the spectral attributes between a set of classes of interest, in order to verify what of them best distinguish these classes. A case study is presented over a small region of the city of São José dos Campos, using a WorldView-2 image. It is important to emphasize that although this study is in a preliminary stage, the results are promising and reached improvements in the accuracy of the classification, even as a good reduction in the computational time.

2. Spectral attributes selection

Most of attributes selection approaches focuses on a global selection, analyzing the correlation for the whole set of attributes and classes, even as its capacity to distinguish between all the classes. In this work, we propose an approach based on the analysis of the best attributes to distinguish pairs of classes. For this, we applied the C4.5 decision tree algorithm [Quinlan 1993], which constructs the classification model based on the divide and conquer strategy. It applies thresholds to the object attributes, and then, observations that are smaller than these thresholds are assigned to the left branch, otherwise to the right branch [Hastie et al. 2008; Körting et al. 2013; Ruggieri 2002].

One important feature of decision tree algorithms is that they indicate the best attributes to distinguish between the classes of interest, according to the entropy measure. However, it analyses all the classes together, choosing the best attributes to distinguish between all of them. To compare the pairs of classes, we isolate only the corresponding samples to the pair being compared and then constructed a decision tree for them. Figure 1 shows three examples of the decision trees. These trees were used in the experiment presented in Section 3, and show for example, that the attribute Band Ratio of the band 2 is the best to distinguish the classes Ceramic Roof and Bare Soil.

With the attributes indicated by this analysis, we construct a matrix as shown in Table 1, which indicates what attributes are the best to separate each pair of classes, and then, these attributes will be used to the classification process. With the selected set of attributes, is expected an increase in the accuracy of the classification, even as a decrease of the computational cost.

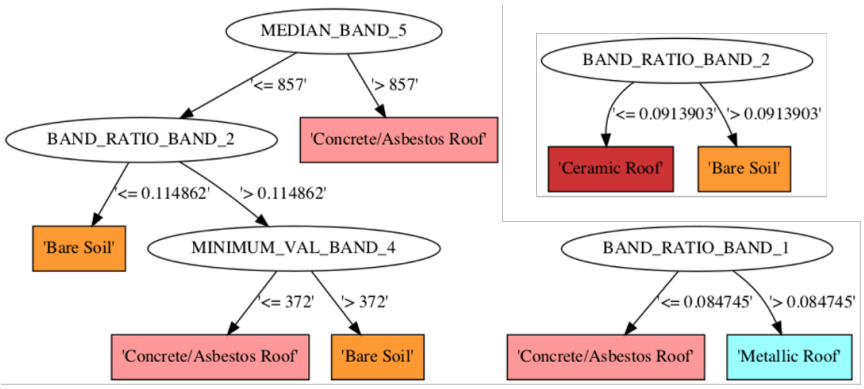

Figure 1. Examples of decision trees comparing pairs of classes.

3. Experimental Results

To evaluate the effectiveness of the approach proposed, we tested a WorldView-2 image of a small area in São José dos Campos, Brazil. The image has 8 multispectral bands with 0.5 meter of spatial resolution. It is important to keep in mind that, in this phase, the objective is not to provide the optimal classification result. The aim of this experiment is to verify the improvement in the distinction between a set of classes when using only the previously selected attributes for the classification, in comparison with the results obtained in the classification using all the spectral attributes computed for the image.

The image was segmented using the Region Growing algorithm [Bins et al. 1996], and then 19 spectral attributes were computed for each band, being 14 statistical measures (like mean, variance, standard deviation, etc) and 5 texture measures based on [Haralick et al. 1973]. Thus, we have 152 spectral attributes for each segment region. The image used and the segmentation result are presented in Figure 2.

Figure 2. Image used in the test (a) and the result of its segmentation (b).

The class typology includes roofs (ceramic, metallic, concrete and asbestos), vegetation, shadow, asphalt pavement and bare soil. Firstly, around 130 samples were collected, distributed between all the classes. Using these samples, we made two classification experiments, using 66% of them to train the decision tree algorithm, and 34% to validate the classification model.

In the first experiment, we used all the 152 attributes to build the decision tree using the C4.5 algorithm. In the second experiment, we applied the proposed approach to select the best attributes and then, we built the decision tree using only the subset of the selected attributes, comparing the validation results with the previous. Figure 1 shows some examples of the decision trees used to select the attributes, and the matrix with all the selected attributes is shown in Table 1.

Table 1. Matrix with the selected attributes for the classification.

Classes	Concrete/Asbestos Roof	Ceramic Roof	Asphalt Pavement	Vegetation	Metallic Roof	Bare Soil	Shadow
Concrete/Asbestos Roof							
Ceramic Roof	BR_B2						
Asphalt Pavement	C_B0	BR_B2					
Vegetation	MD_B4	MD_B4	MD_B1				
Metallic Roof	BR_B1	BR_B2	BR_B2	MD_B1			
Bare Soil	BR_B2 MD_B5 MIN_B4	BR_B2	BR_B1 AM_B4	ME_B4	BR_B1		
Shadow	ME_B5	ME_B5	SM_B0	BR_B4	MD_B2	MD_B5	

Where:

- AM_B4 → Amplitude on Band 4
- BR_B1 → Band Ratio of Band 1
- BR_B2 → Band Ratio of Band 2
- BR_B4 → Band Ratio of Band 4
- C_B0 → Number of valid values on Band 0
- MD_B1 → Median on Band 1
- MD_B2 → Median on Band 2
- MD_B4 → Median on Band 4
- MD_B5 → Median on Band 5
- ME_B4 → Mean on Band 4
- ME_B5 → Mean on Band 5
- MIN_B4 → Minimum Value on Band 4
- SM_B0 → Sum on Band 0

In both experiments, the results were evaluated with the 34% remaining samples (the 66% used to build the tree were not used in the validation). The decision tree built for the first experiment is shown in Figure 3. In this experiment, the classification obtained an accuracy of 63.64% in the validation, with an error of 36.36%, and the kappa value obtained was 0.57.

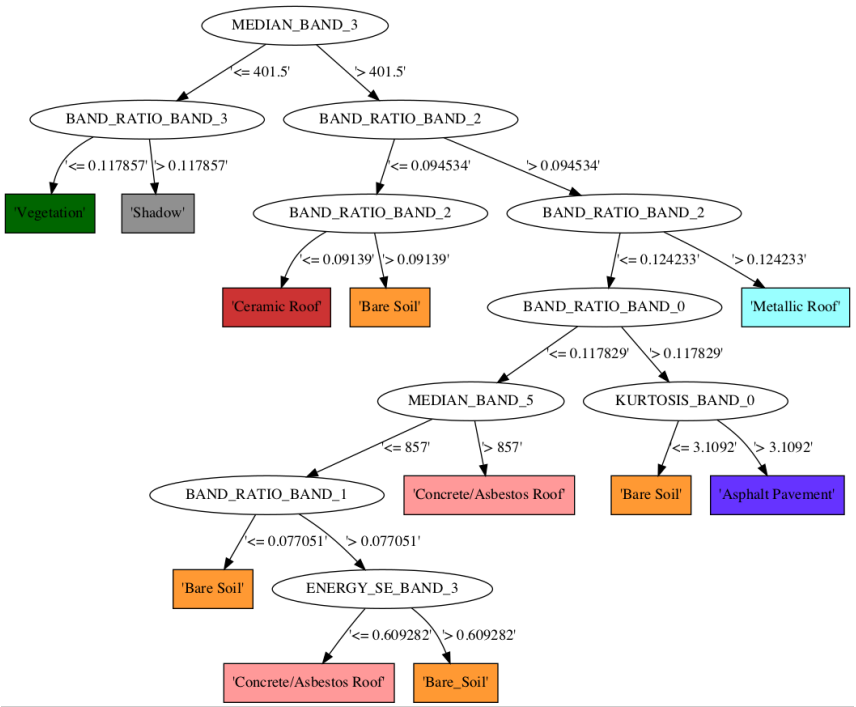

Figure 3 – Decision tree built using the whole set of attributes.

The decision tree built in the second experiment, considering only the subset of the selected attributes, is shown in Figure 4. We can see, in this second tree, several differences in the attributes used, and in the importance attached for some of them. In the decision trees algorithm, the attribute that provides the greater distinction between the classes is assigned to the root, and nodes in lower levels, receive smaller importance. In this way, the lower levels provides a finer adjustment for the classification. In this experiment, the classification obtained an accuracy of 70.45%, with an error of 29.54%, and the kappa value obtained was 0.65.

4. Concluding Remarks and Future advances

This work has shown that the attributes selection approach proposed can help for the improvement of the accuracy in the classification through Data Mining techniques. In our experiments, the results increased around 7% the accuracy when compared to the original classification. We believe that, with more adjustments in the methodology and in the models for classification, we can obtain more relevant improvements. Moreover, with the reduction on the amount of attributes used in the classification, we can also reduce the computational cost of this process.

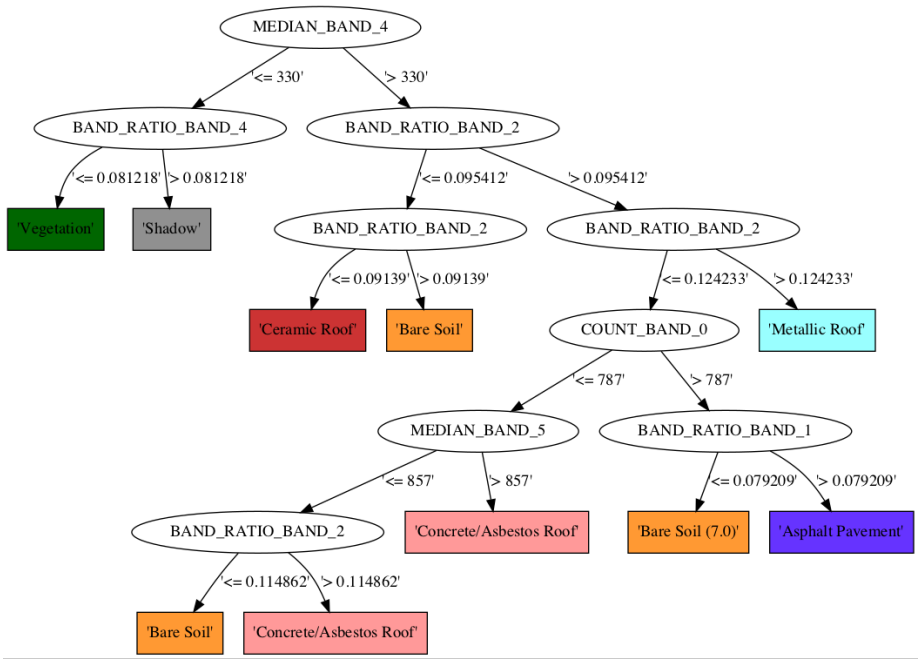

Figure 4. Decision tree built using only the subset of selected attributes.

As future steps, we must validate the results for an entire image, and then automatize the process of the comparison between the classes of interest. We aim to implement the results in GeoDMA platform and then study the improvements in the computational performance of the process, comparing with other classification processes. This work also gives way to thought about approaches using evolutionary computing or other optimization methods, aiming to improve the selection process to try to find the optimal set of attributes, in order to help the analysts to both, improve the classification results, and understand more about the data being classified.

5. References

Bins, L. S., Fonseca, L. M. G., Erthal, G. J. and Ii, F. M. (1996). Satellite Imagery Segmentation : a region growing approach. Anais VIII Simposia Brasilerio de Sensoriamenta Remoto, Salvador, Brasil, 14-19 abril 1996, INPE, p. 677–680.

Dash, M. and Liu, H. (1997). Feature Selection for Classification. Intelligent Data Analysis, v. 1, n. 97, p. 131–156.

Haralick, R., Shanmugan, K. and Dinstein, I. (1973). Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics. http://dceanalysis.bigr.nl/Haralick73-Textural features for image classification.pdf.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2008). The elements of statistical learning: data mining, inference and prediction. The Mathematical, v. 27, n. 2, p. 83–85.

Im, J., Jensen, J. R. and Tullis, J. a. (2008). Object-based change detection using correlation image analysis and image segmentation. International Journal of Remote Sensing, v. 29, n. 2, p. 399–423.

Kohavi, R. and Kohavi, R. (1997). Wrappers for feature subset selection. Artificial Intelligence, v. 97, n. 1-2, p. 273–324.

Körting, T. S., Garcia Fonseca, L. M. and Câmara, G. (2013). GeoDMA-Geographic Data Mining Analyst. Computers and Geosciences, v. 57, p. 133–145.

Laliberte, a. S., Browning, D. M. and Rango, a. (2012). A comparison of three feature selection methods for object-based classification of sub-decimeter resolution UltraCam-L imagery. International Journal of Applied Earth Observation and Geoinformation, v. 15, n. 1, p. 70–78.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. San Mateo, California: Morgan Kaufman Publishers.

Ruggieri, S. (2002). Efficient C4 . 5. IEEE Transactions on Knowledge and Data Engineering, v. 14, n. 2, p. 438–444.

Syed, S., Dare, P. and Jones, S. (2005). Automatic classification of land cover features with high resolution imagery and lidar data: an object-oriented approach. … : the national biennial Conference of the …, p. 512–522.

Walter, V. (2004). Object-based classification of remote sensing data for change detection. ISPRS Journal of Photogrammetry and Remote Sensing, v. 58, n. 3-4, p. 225–238.