ASSESSMENT OF CLASSIFIERS THROUGH DECISION TREE AND REGRESSION TREE ALGORITHMS IN URBAN AREA USING *WORLDVIEW-2* IMAGE

Bruna Maria Pechini Bento [1], *Hermann Johann Heinrich Kux* [2], *Thales Sehn Körting* [3]

[1] Norwegian School of Economics, brunapechini@gmail.com; [2] Instituto Nacional de Pesquisas Espaciais, hermann@dsr.inpe.br; [3] Instituto Nacional de Pesquisas Espaciais, tkorting@dpi.inpe.br

ABSTRACT

Geographic Object-Based Image Analysis allows the simulation from the view of a human interpreter using knowledge models expressed by semantic networks. Data mining techniques have been widely used as a support tool for the construction of the semantic network. In this sense, the aim of this study is to analyze the performance of the CART and C4.5 algorithms, which use decision trees, to classify urban land cover. A *WorldView-2* image was used for this analysis. Both algorithms presented good accuracy. The C4.5 algorithm accuracy presented average values slightly higher than the CART algorithm. C4.5 was supported by other software for the execution of the analyses. This posed a challenge to the researchers for data integration, data format conversion and also file replication. Differently, the CART algorithm tested is part of an integrated GEOBIA platform, which benefits the user reducing the time spent to execute all the image analysis steps.

Key words *— WorldView-2, GEOBIA, Data mining, C4.5 Algorithm, CART Algorithm.*

1. INTRODUCTION

The third generation of imaging satellites provides data with higher spatial, spectral, radiometric and temporal resolutions, opening new possibilities to explore the spatial complexity of the urban phenomenon [1, 2]. The increase of informative content from such images requires the search for new methodologies and tools to analyze them, since the analysis of these images by conventional methods, such as pixel by pixel classifiers and by region, resulted in limitations for detailing classes and consequently with low accuracies [4, 5].

The Geographic Object-Based Image Analysis (GEOBIA) paradigm represents a solution to overcome these limitations. This approach allows the simulation of the contextual view from a human interpreter, using knowledge models expressed by semantic networks and multiple levels of interconnected classification [3]. However, the construction of knowledge models is a complex task, requiring the knowledge of the scene beforehand and demands a long period for its realization [2, 6]. In this sense, data mining techniques have been used as a support tool for the construction of semantic networks. Decision tree algorithms have been widely used in data mining tasks. At the classification of orbital images, these algorithms select automatically the most appropriate attributes for the characterization of classes to be discriminated. The result is represented by a decision tree which is later adapted into a semantic network [7, 8, 6].

Thus, this study proposes a methodology to use jointly cognitive and data mining approaches aiming to analyze the performance of both decision and regression tree algorithms for the classification of urban land cover generated from images with high spatial resolution. In order to verify if these trees are equivalent in performance, the following specific objectives are proposed: to compare the structure of the decision trees and the classifications generated by data miners C4.5 [9] and CART (Classification and Regression Trees) [10]. For this analysis, a *WorldView-2* scene from São José dos Campos city, São Paulo State (Brazil) was used.

2. METHODOLOGY

Totally 12 types of different materials were identified in the area under study (Table 1).

Table 1. Urban land cover classes discriminated by hierarchy classes of segmentation.

Level 1	Level 2	Level 3	Level 4
Blocks	Impervious covers	Ceramic	Orange colored ceramic
			Grey ceramic
			Weathered red ceramic
			New red ceramic
		Metallic Covers	Bright metallic
			Metallic with grey painting
		Derived from cement	New cement Weathered cement
		Non-road paving	Quartzite
	Pervious covers	Water bodies	Swimming pool
		Bare soil	Bare soil
		Vegetation	Arboreal vegetation

Moreover, two segmenters were considered at eCognition software. Initially, the Multi-resolution Segmentation algorithm was used [11]. For the hierarchical network of objects, a Top-Down strategy was adopted. The first level was created to separate Blocks from Streets. The blocks were classified by the predominance of targets, such as: Vegetation, Buildings, Houses, Metallic Covers and Mixed Areas. Thus, in the lower levels of segmentation each block type received different parameters in the segmentation criteria.

The second segmentation level sought to separate pervious from impervious areas; the third level aimed to characterize urban targets; the 4th level was created to discriminate smaller targets, causing a super segmentation of the image. Due to that, the 5th segmentation level was created, using the Spectral Difference segmentation algorithm, which allows to group statistically similar contiguous segments with a Bottom-Up strategy.

Aiming to increase the diversification of the input dataset for object classification from the image, customized attributes were generated, based on mathematical operations applied on image objects.

In order to make the classification models equivalent to C4.5 and CART algorithms, the same values of algorithm parameters were defined for both tests, namely the Type of Test and the Minimum number of samples for Decision Tree Leaf (M) to generate a rule. The test option chosen was of type Cross Validation, totaling 10 validations. As for parameter M, 14 tests were done, varying the parameter from 2 to 15 samples for each algorithm, totaling 28 models of trees tested.

The decision on the number of validation samples was made based on the size of the area of each class for each test. The larger the area occupied by a class, the greater the number of validation samples collected. In total, 455 validation samples were collected at the 5th segmentation level.

Due to the large number of tests made at both experiments, preliminary evaluation of accuracy was needed, to have a first overview of how the accuracies would look like. So the sampling set for the first accuracy evaluation presented 65 samples, obtained at segmentation level 5.

The evaluation of decision trees was done based on its complexity. For this task, the depth value of each tree was analyzed. This was done through counting all nodes on it, i.e. its leaves and decisions.

3. RESULTS

The results of preliminary accuracies obtained with tests, are presented on Figure 1.

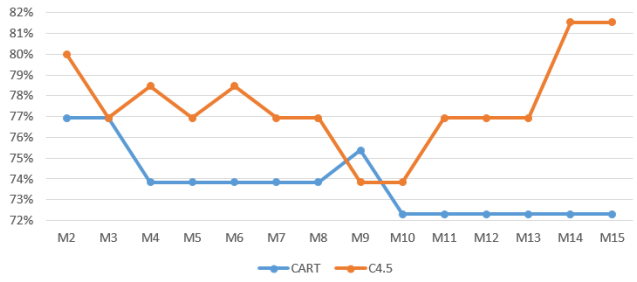

Figure1. Comparison of preliminary accuracies obtained at classifications generated by algorithms CART and C4.5. The minimum value of instances per leaf is represented by M following the value of this parameter per test.

The analysis of graph at Figure 1 indicates that both algorithms present a good preliminary accuracy. As for the CART algorithm, the preliminary accuracies concentrate between 72.3% and 76.9%, while at the C4.5 algorithm these accuracies are relatively higher, between 73.8% and 81.5% which corresponds to an average of 3.8% above the accuracies obtained by the CART classifier. Besides that, both algorithms remain stable with the increase of the M value, varying 4.6% in the tests with CART, 7.7% in those with algorithm C4.5 and 9.2% when comparing among them.

The depth values from each tree are presented on the graph below (Figure 2).

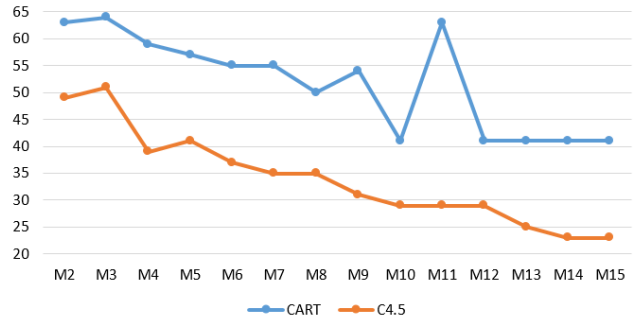

Figure 2. Comparison of depth from decision trees generated by algorithms CART and C4.5. The minimum value of instances per leaf is represented by M followed by the value of this parameter per test.

Referring to the tests performed with the CART algorithm, the variation between the largest and the smallest tree was 22 nodes, while with the C4.5, the tree sizes varied in 28 nodes. When compared, the C4.5 algorithm presents a mean difference of 18 nodes less than the trees generated by the CART algorithm, which means that the algorithm C4.5 was able to generate more generic trees.

In the search for a decision tree with both a good generalization ability and accuracy, decision trees were found where the model used value 10 for parameter M. Its decision trees are presented below (Figures 3 and 4).

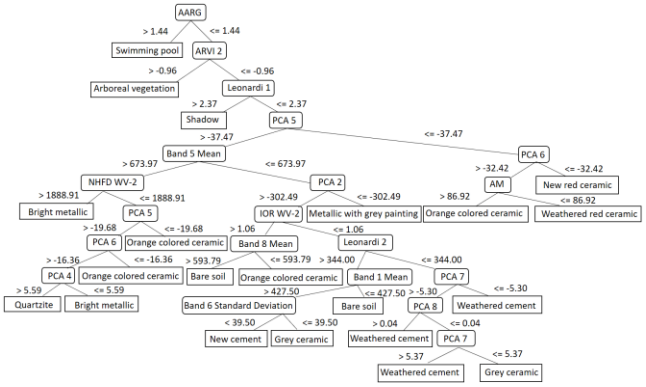

Figure 3. Decision tree generated by algorithm CART with M=10.

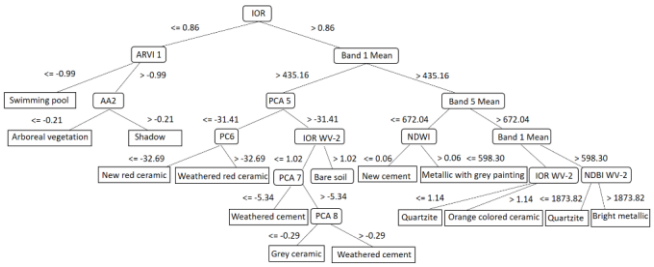

Figure 4. Decision tree generated by algorithm C4.5 with M=10.

The preliminary accuracy of the model generated by the CART algorithms in the chosen model presented an accuracy corresponding of 72.3%, while the algorithm C4.5 presented a slightly higher accuracy, of 73.8%, which corresponds to a difference of 1.5%.

As for the depth of the decision trees, the tree obtained by the CART model presented 41 nodes while that of the C4.5 model 29 nodes, which corresponds to a difference of 12 nodes between them. Thus, it can be stated that the C4.5 tree reached a higher level of generalization.

Considering the most suitable attributes for the characterization of each soil cover class and, therefore, the classification model generated, the WorldView-2 image was classified according to Figures 5 and 6. In order to check the class given to each object of the image, its edges were maintained.

Figure 5. Classification of urban land cover, using algorithm CART with M=10.

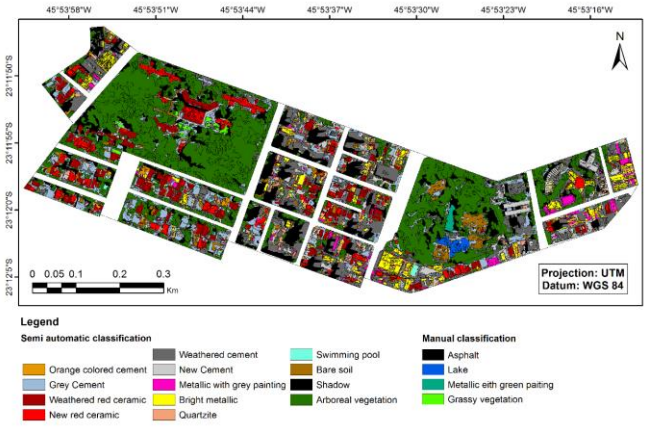

Figure 6. Classification of urban land cover using algorithm C4.5 with M=10.

Finalizing, the accuracy of the urban land cover classifications generated by the CART and C4.5 algorithms with an M value of 10 are presented. The results are shown in the confusion matrices (Table 2 and Table 3).

Table 2. Confusion matrix of classification with algorithm CART, with M=10.

Classif. / Ref.	Oran. col. Ceram.	Grey ceram.	Weath. red ceram.	New red ceram.	Weath. Cem.	New cem.	Metal. w. grey p.	Bright metal.	Quart.	Swim. pool	Bare soil	Shadow	Arb. Veg.	Total	User accuracy
Oran. col. Ceram.	3	2	0	0	0	1	0	0	0	0	0	0	0	6	0.07
Grey ceram.	1	12	0	0	1	0	0	0	0	0	0	1	0	15	0.24
Weath. red ceram.	19	5	25	2	0	0	0	0	0	0	0	0	0	51	1.00
New red ceram.	12	0	0	28	0	0	0	0	0	0	2	0	0	42	0.93
Weath. Cem.	0	20	0	0	52	5	2	1	0	0	1	3	0	84	0.87
New cem.	3	2	0	0	0	27	0	1	5	0	1	0	0	39	0.68
Metal. w. grey p.	0	1	0	0	0	0	6	0	0	0	1	0	0	8	0.60
Bright metal.	5	4	0	0	0	3	1	31	6	1	0	0	0	51	0.89
Quart.	0	1	0	0	0	3	0	2	4	0	1	0	0	11	0.27
Swim. pool	0	0	0	0	0	0	0	0	0	3	0	0	0	3	0.60
Bare soil	1	0	0	0	1	1	0	0	0	0	15	0	0	18	0.75
Shadow	1	2	0	0	5	0	1	0	0	0	0	51	0	60	0.93
Arb. Veg.	0	1	0	0	1	0	0	0	0	0	0	0	65	67	1.00
Total	45	50	25	30	60	40	10	35	15	5	20	55	65	455	
Producer Accuracy	0.5	0.8	0.49	0.67	0.62	0.69	0.75	0.61	0.36	1	0.83	0.85	0.97		
Global accuracy	0.7														

Table 3. Confusion matrix of classification with algorithm C4.5, with M=10.

Classi / Ref.	Oran. col. Ceram.	Grey ceram.	Weath. red ceram.	New red ceram.	Weath. Cem.	New cem.	Metal. w. grey p.	Bright metal.	Quart.	Swim. pool	Bare soil	Shadow	Arb. Veg.	Total	User accuracy
Oran. col. Ceram.	3	0	1	0	0	0	0	0	0	0	3	0	0	7	0.60
Grey ceram.	0	7	1	0	2	2	0	0	2	0	0	0	0	14	0.15
Weath. red ceram.	0	2	35	0	0	0	0	0	0	0	6	0	0	43	0.78
New red ceram.	0	0	5	25	0	1	0	0	0	0	0	0	0	31	1.00
Weath. Cem.	0	20	1	0	41	9	6	0	2	4	2	0	0	85	0.65
New cem.	1	1	0	0	1	23	1	1	2	0	0	0	0	30	0.58
Metal. w. grey p.	0	0	0	0	0	0	3	0	0	0	0	0	0	3	0.20
Bright metal.	0	0	0	0	0	4	2	33	6	0	0	0	0	45	0.94
Quart.	0	0	0	0	0	1	1	1	8	0	0	0	0	11	0.40
Swim. pool	0	0	0	0	0	0	1	0	0	6	0	0	0	7	0.60
Bare soil	1	5	0	0	0	0	0	0	0	0	15	0	0	21	0.50
Shadow	0	8	2	0	14	0	1	0	0	0	2	54	3	84	0.98
Arb. Veg.	0	4	0	0	5	0	0	0	0	0	2	1	62	74	0.95
Total	5	47	45	25	63	40	15	35	20	10	30	55	65	455	
Producer Accuracy	0.43	0.5	0.81	0.81	0.48	0.77	1	0.73	0.73	0.86	0.71	0.64	0.84		
Global accuracy	0.69														

In general, the classifications obtained a similar global accuracy, and the classification generated with the CART algorithm presented a slightly higher value. The graph allowed to observe that both algorithms obtained low accuracy in the classification of the classes Orange Ceramic and Weathered Cement. In the first case the classification was included by the CART algorithm in the Grey Ceramic classes and New Cement. At algorithm C4.5 the result was Weathered Red Ceramic and Bare Soil. In the second case,

the confusion occurred with classes Grey Ceramic, New Cement, Metallic with Grey paint and Bare Soil in both models. The CART algorithm also included it in the Shadow class and the algorithm C4.5 in the Quartzite and Swimming Pool.

The CART algorithm presented still a low accuracy in the classification for classes Weathered Red Ceramic and Quartzite. The first was classified as Orange Ceramic, Grey Ceramic and New Red Ceramic while Quartzite was included in the Bright Metallic, Grey Ceramic and Bare Soil. The C4.5 algorithm presented low accuracy for the classification of the Grey Ceramic and Shade classes. The first one was included in Weathered Red Ceramic and Quartzite and the second one in the classes Weathered Cement, Shade, Grey Ceramic, Arboreal Vegetation, Weathered Red Ceramic and Metallic with Grey Paint.

4. CONCLUSION

As shown, the performance of decision tree algorithms for urban land cover classification from high spatial resolution orbital images was analyzed using Data Mining algorithms C4.5, available in both WEKA and eCognition platforms. Regarding its accuracy, although both algorithms presented good results, C4.5 obtained, in general, better results than CART. As for the complexity of the decision tree models obtained in the experiments, C4.5 presented a higher generalization capacity on the formulation of the rules by attribute. As presented, the test using C4.5 required the migration from the GEOBIA environment to Data Mining and its return, after the definition of the classification model. This poses challenges to researchers such as format conversion and integration of the data and knowledge of the software to be used.

Consequently, the possibility to use a platform that integrates all image analysis tasks, is a benefit for the user, regarding the time and cost savings to carry out all these steps. This scenario summarizes the test performed on the eCognition platform with the CART algorithm. In spite of the advantages described, the high licensing cost may be a barrier to its use. As an alternative it is suggested to look for computational environments using the free software policy.

5. REFERENCES

[1] Souza, I.M.; Pereira, M.N.; Fonseca, L.M.G.; Kurkdjian, M.L.N.O. "Mapeamento do uso do solo urbano através da classificação por regiões baseada em medidas texturais". In: Simpósio Brasileiro de Sensoriamento Remoto, 11. (SBSR), 2003, Belo Horizonte. Anais... São José dos Campos: INPE, 2003. pp. 1967-1968. CD-ROM. ISBN 85-17-00017-X. (INPE-16180-PRE/10783).

[2] Almeida, C.M. "Aplicação dos sistemas de sensoriamento remoto por imagens e o planejamento urbano e regional". Arq.Urb – Revista Eletrônica de Arquitetura e Urbanismo (USJT), n. 3, pp. 98-123, 2010.

[3] Bento, B.M.P. "Avaliação de classificadores por árvore de decisão e árvore de regressão em cenas urbanas do sensor Worldview-2". Dissertação (Mestrado em Sensoriamento Remoto) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, pp.140. 2016.

[4] Hay, G; Castilla, G. "Geographic Object-Based Image Analysis (GEOBIA): A new name for a new discipline". In: BLASCHKE, T.; LANG, S.; HAY, G. Object-based image analysis. Spatial concepts for knowledge-driven remote sensing applications. Berlin, Heidelberg: Springer-Verlag, 2008. cap. 1.4, pp. 75-89.

[5] Francisco, C.N.; Almeida, C.M. "Avaliação de desempenho de atributos estatísticos e texturais em uma classificação de cobertura da terra baseada em objeto". Bol. Ciênc. Geod., sec. Artigos, Curitiba, v. 18, n. 2, pp. 302-326, Abril-Junho, 2012.

[6] Pinho, C.M.D.; Silva, F. C.; Fonseca, L.M.G.; Monteiro, A.M.V. "Urban land cover classification from high-resolution images using the C4.5 algorithm". The International Archives of the Photogrammetry, Remote sensing and Spatial Information Sciences. vol. XXXVII, part. B7, pp. 695-700, Peking, 2008.

[7] Aksoy, S.; Koperski, K.; Tusk, C.; Marchisio, G. "Interactive Training of Advanced Classifiers for Mining Remote Sensing Image Archives". ACMInternational Conference on Knowledge Discovery and Data Mining. Seattle, pp. 773-782, 2004.

[8] Korting, T.S.; Fonseca, L.M.G.; Câmara, G. "GeoDMA – Geographic Data Mining Analyst". Computer & Geosciences, v. 57, pp. 133-145, 2013.

[9] Quinlan, J.R. "C4.5: Programs for Machine Learning". San Mateo: Morgan Kaufmann Publishers, 1993. 302 pp. ISBN(1-55860-238-0).

[10] Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. "Classification and Regression Trees". Boca Raton: Chapman & Hall/CRC, 1984. 293 pp. ISBN(0-412-04841-8).

[11] Baatz, M.; Schäpe, A. "Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation". In: Angewandte Geographische Informationsverarbeitung XII. Beiträge zum Agit Symposium Salzburg, 12., 2000, Karlsruhe. Proceedings… Karlsruhe Herbert Wichmann Verlag, pp. 12-23, 2000.

Acknowledgements
The authors thank CAPES – (Coordination for the Improvement of High Level Personnel) for the financial support to the execution of this work and company DIGITALGLOBE, for the provision of the *WorldView-2* image.