DECISION TREE DATA MINING IN OBJECT ORIENTED CLASSIFICATION FOR SUGAR CANE HARVEST TYPES

Elizabeth Goltz[1], Gustavo F. B. Arcoverde[1], Daniel A. de Aguiar[1], Bernardo F. T. Rudorff[1], Eduardo E. Maeda[2]

[1]National Institute for Space Research – INPE, [2]University of Helsinki – Department of Geography

goltz@dsr.inpe.br, gustavo@dsr.inpe.br, daniel@dsr.inpe.br, bernardo@dsr.inpe.br, eduardo.maeda@helsinki.fi

The increasing worldwide concern regarding environmental issues, in particular after the consistent reports published by the Intergovernmental Panel on Climate Change – IPCC, has significantly stimulated the production of biofuels. Among these biofuels, the ethanol produced from sugar cane is considered to be of particular importance. Currently, Brazil is the world's largest producer of sugar cane with approximately 8.7 million ha of cultivated area [1]. Traditionally, in Brazil, great part of the harvest of sugar cane is performed after the burning of leaves and straw, which can causes severe environmental problems, such as CO_2 emissions, soil degradation [2] and health problems within the population surrounding the croplands.

In this context, the Secretary of the Environment of the State of São Paulo and the sugar cane producers signed a protocol of intentions, in June 2007, anticipating the deadline limits and predicting the elimination of the use of fire for the sugar cane harvest by 2014 [3]. In order to achieve such objective, the use of remote sensing images offers a great tool in the surveillance of harvest activities and in the assessment of the compliance with the agro-environmental protocol. Currently, the Canasat Project [1] carries out the identification of these areas by visual interpretation of satellite images from the TM/Landsat and CCD/CBERS sensors acquired from April to December of each year [4]. This approach requires substantial amount of time and trained interpreters and does not allow a near real-time monitoring. Thus, other mapping methods should be investigated in order to optimize this work.

The object oriented analysis of satellite imagery holds an important position in this framework, due to the fact that in many occasions the information is not present in a single pixel, but in a group of those, with inherent context and neighbourhood relations, which can show more coherent information with agricultural lands. The use of objects can aggregate many attributes, as different metrics can be calculated by intrinsic or extrinsic statistics from the segments. Therefore, data mining techniques can help in the selection of information relevant to different branches of knowledge [6]. Compared to other data mining techniques, algorithms to generate decision trees are computationally fast, can attain easily interpretable rules and have an intrinsic ability for feature selection [7].

In order to improve the mapping conducted by the Canasat Project, the objective of this study was to analyze the generation of decision trees using multi-attributes extracted from objects in TM/Landsat sensor images aiming the classification of types of sugar cane harvesting under different soil types.

For this study, two groups of municipalities were chosen inside the state of São Paulo. The first group consisted of 4 municipalities, with a total of 70,417 hectares cultivated with sugar cane, planted predominantly on dystrophic Oxisol (dark soil). The second group consisted of 18 municipalities, with 53,961 hectares of sugar cane with a predominance of Distrofic Red-Yellow Argisoil (light soil). For both areas, images from the months of April/08, May/08, July/08 and August/08 were used in the analysis. These images were converted to apparent reflectance and georeferenced.

The attributes' extraction was performed using the software Definiens, and the selected samples were classified according to the harvest type, burnt or non-burnt, in each month, and the samples where no harvest was observed, classified as "bisada" areas. A total of 72 samples and 88 attributes were collected. Among the attributes, the following typologies were considered: pixel level, neighbourhood relation, shapes and Haralick's textures. For the data mining processing and classification, the present work made use of the C 4.5 algorithm, available in the Weka software. For the pruning and selection of the decision trees, the gradual increase of the minimum number of instances per leaf was monitored. The evaluation of decision trees was performed using the kappa index, generated by cross validation.

Regarding the evaluation of the decision trees using cross validation, a kappa index of 0.82 was found for the first group, while the second group achieved 0.69. Due to this significant difference, it is possible to assume that the landscape characteristics in which the sugar cane was cultivated has influence in the harvest type classification. However, for both groups the biggest omission and commission errors were observed in the same classes, namely the non-burnt and burnt, respectively. Despite the great diversity of attributes, 3 attributes were predominant in both groups, all from the pixel level typology: mean value, minimum pixel value and maximum pixel value. Although both trees presented the same types of attributes, the nodes positions in which these attributes were defined occurred in an unequal way. Also, it was observed that the samples in the first months of harvest were identified in the first nodes of both trees, while the samples from last months and "bisada" areas were situated to be in the last nodes. Moreover, it was found that the use of decision trees are sensitive to changes in soil, but not in a significant way, and the class representing harvest in burnt areas had higher sensitivity to variations in soil.

References

[1] Canasat Project. Available at: <http://www.dsr.inpe.br/mapdsr/eng>. Accessed on Out 2008.

[2] SASSO, C. G. "Burning of the Sugarcane – Crop: Biomass Aspects, Soil, Fertility and CO_2 Emission in Atmosphere". Thesis (Agronomy). Oeste Paulista University, Presidente Prudente, SP. 2007.

[3] AGUIAR, D. A.; RUDORFF, B. F. T.; SILVA, W. F.. "Monitoramento do modo de colheita da cana-de-açúcar no Estado de São Paulo - Brasil por meio de imagens de sensores orbitais em dois anos-safra". In: XIII SELPER, Havana, 2008.

[4] AGUIAR, D. A.; SILVA, W. F.; Feitosa, F. F.; Gonçalves, F. G;RIZZI, R.; RUDORFF, B. F. T. . "Análise espacial da colheita da cana-de-açúcar no Estado de São Paulo: a influência da precipitação". In: XIII Brazilian Remote Sensing Symposium (SBSR - 2007, Florianópolis). INPE, 2007.

[5] PINHO, C. M. D. "Object-Oriented Analysis of High Spatial Resolution Sensor Images for Intra-Urban Land Cover Classification: The Case of São José dos Campos-SP, Brazil". 180 p. (INPE-14183-TDI/1095). Thesis (Remote Sensing) - INPE, São José dos Campos, SP. 2005.

[6] WITTEN, I. H.; FRANK, E. Data Mining: Practical machine learning tools and technique. 2 ed. 3.4.12. São Francisco, EUA: Morgan Kaufmann, 558 p. 2005.

[7] YANG, C.; PRASHER, S. O.; ENRIGHT, P.; Madramootoo, C.; Burgess, M.; Goel P. K.; Callum I. "Application of Decision Tree Technology for Image Classification using Remote Sensing Data". Agricultural Systems, v. 76, p. 1101-1117, 2003.