Intraurban land cover classification using IKONOS II images and data mining techniques: a comparative analysis

Vanessa da Silva Brum Bastos
Remote Sensing Division (DSR)
National Institute for Space Research (INPE)
São José dos Campos, Brazil
vsbb@dsr.inpe.br

Leila Maria Garcia Fonseca
Image Processing Division (DPI)
National Institute for Space Research (INPE)
São José dos Campos, Brazil
leila@dpi.inpe.br

Thales Sehn Korting
Image Processing Division (DPI)
National Institute for Space Research (INPE)
São José dos Campos, Brazil
tkorting@dpi.inpe.br

Carolina Moutinho Duque Pinho
Centro de política e economia do setor público (CEPESP)
Fundação Getúlio Vargas (FGV-SP)
São Paulo, Brazil
cmdpinho@gmail.com

Rafael Duarte Coelho dos Santos
Applied Computing Division (CAP)
National Institute for Space Research (INPE)
São José dos Campos, Brazil
rafael.santos@inpe.br

Abstract— **High spatial resolution image analysis acquired over urban areas has been performed with success using Geographic Object Based Analysis (GEOBIA). However, it was observed that the use of data mining techniques in the image analysis procedures can speed up the processing time by selecting the most appropriate parameters for classification process without decreasing the classification accuracy. Therefore, this work aims at comparing some algorithms for classifying intra-urban land cover using IKONOS II images and data mining techniques. Three classification algorithms, KNN, MLP and C4.5 were analyzed.**

I. Introduction

Recent developments in remote sensing technologies include the construction of sensors with high spatial, spectral, and radiometric resolution. As a result, it has been possible to discriminate sub-metric objects, which can benefit urban studies from remote sensing technologies [6][10][13].

Regarding image analysis methods, Geographic Object Based Image Analysis (GEOBIA) has been consolidated as an adequate and efficient methodology for urban land cover monitoring [3] [13]. Despite continuous advances in the image analysis area, some problems related to the feature selection and large amount of data are still troublesome [13]. Through data mining it is possible to discover useful knowledge that is implicit in the data and then to extract patterns from large amounts of data [4] [5]. As the usage of data mining methods has showed a good potential for image analysis [2][7], this study aims at analyzing the performance of three classification methods to classify intra-urban land cover using an IKONOS image and data mining techniques. We used the three most commonly used classifiers that are implemented in WEKA (Waikato Environment for Knowledge Analysis): KNN (K Nearest Neighbor), C4.5, and MLP (Multi Layer Perceptron). WEKA is free software that contains several tools for data pre-processing and visualization as well as algorithms for regression, classification, grouping, mining association rules, and attributes selection [11].

In the next section, we give a brief review about the classifiers used in our experiments. In Section III, we present the proposed methodology and results. Finally, Section IV presents the conclusions.

II. Classification methods

The classifier KNN is an instance-based learning algorithm in which specific instances are used in the training phase to do the forecast without the need of maintaining a derived data model [9]. The KNN classifier uses the Euclidean distance as a proximity metric for determining the similarity between nearest neighbor instances and their classes [8]. The parameter K determines the number of nearest neighbors to be used in predicting the classification. Thus, the algorithm determines a group with K training objects that are closest to a test object, and labels it according to the predominant class in this group or region [8][11].

The lower the K value, the more sensitive the classifier is to noise. Conversely, higher K values indicate a greater possibility of inclusion points belonging to another class. Both situations can lead to incorrect predictions [8][11]. KNN algorithm is not a didactic classifier because the distance metrics combinations used to sculpt the boundaries in the attributes space are not displayed, and thus the knowledge representation is not explicit [11].

The C4.5 classifier generates a decision tree with a structure in which a leaf can represent a class or a decision node. Each class or decision node indicates some test to be performed on an attribute value, with a branch or sub-tree for each test result [2][8]. In the decision tree, thresholds are applied to the object's features. Observations satisfying the thresholds are assigned to the left branch; otherwise, they are applied to the right branch. In the final step, classes are assigned to the terminal nodes (or leaves) of the tree. This classifier is a fairly didactic method because it allows visualization of decision rules [11].

The classifier Multilayer Perceptron is an artificial neural network that uses a supervised error-correction learning back propagation to classify instances [2][11]. Neural networks simulate the human brain structure and it consists of an interconnected set of nodes and direct links [8].

III. Methodology and experimental results

The experiment was performed in the same study area used by Pinho et al. [6]. The area covers a region of 12-km^2 in a southern sector of São José dos Campos municipality, which has a variety of occupation and land use patterns and therefore has a wide range of intra-urban targets (Figure 1).

978-1-4799-0211-8/13/$31.00 ©2013 IEEE

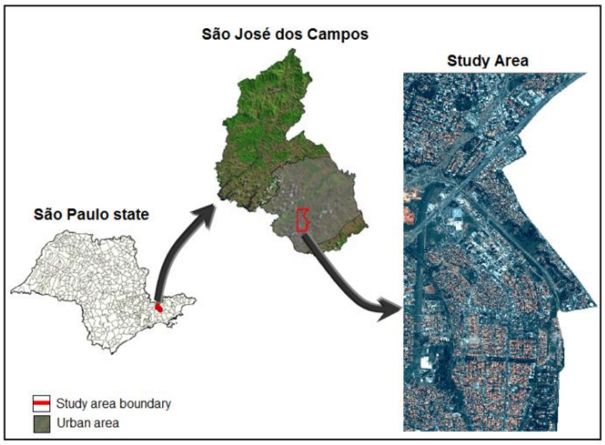

Fig. 1 - Study area location, adapted from [6]

We used an IKONOS image II, acquired on March 13, 2001: a 1.0-m panchromatic band and 4.0-m multispectral bands (blue, green, red and near-infrared). The images are standard geometrically corrected and have an 11-bit radiometric resolution and 4.85° incidence angle. The images were fused and segmented into five levels so that a specific target could be discriminated at each level.

The segmentation and feature extraction processes were performed in the Definiens software. In the training step, some segments associated with their classes were selected. This set of segments was divided into two groups: (1) a training set with 1630 instances, and (2) an evaluation set with 819 instances. Additionally, each segment was characterized by 524 attributes. These data were exported from Definiens to WEKA 3.6.0.

To evaluate the classifiers, we performed a Monte Carlo simulation. For each classifier, we ran 30 iterations for each GVP (Growing Variable Parameter), represented by the Minimum number of objects per leaf (MOL), K and Number of Hidden Neurons (NHN), respectively, for C4.5, KNN, and MLP algorithms. Figure 2 shows the Monte Carlo results. Following, we present experimental results.

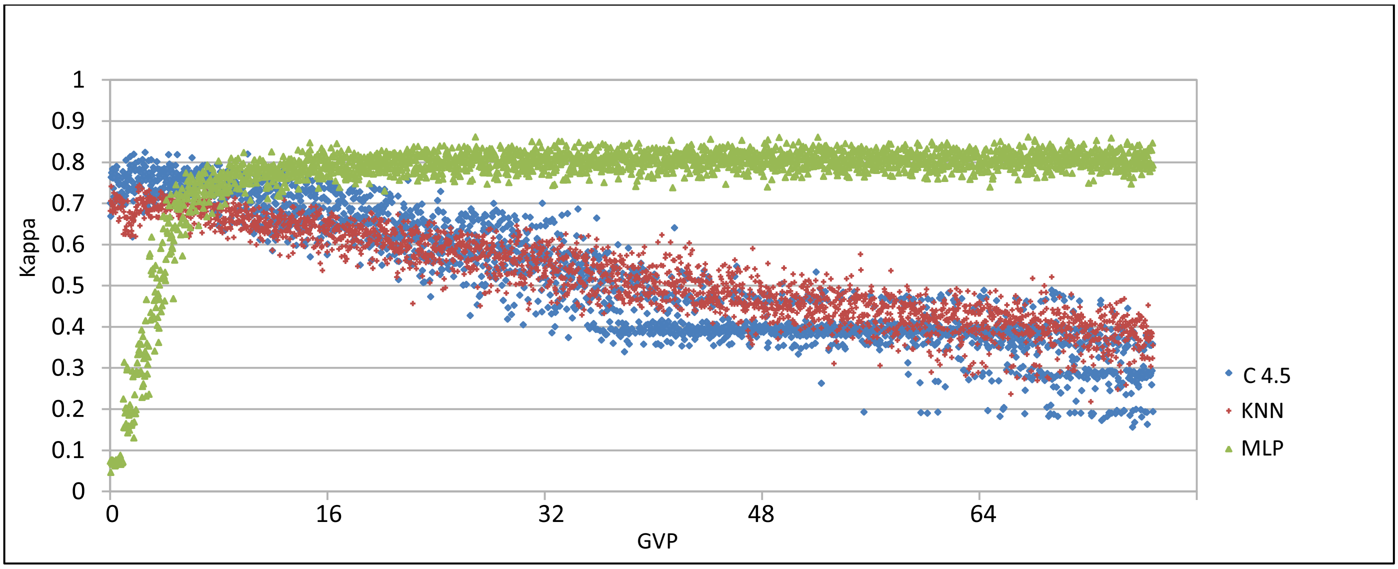

Fig. 2 - Kappa versus GVP (MOL, K, NHN) values for each classifier.

A. *C4.5 classifier*

For C4.5 algorithm, as MOL value increases the decision tree complexity and the classification accuracy decrease, as we can observe in Table 1. This is due the fact that the tree becomes more generalist. On the other hand, the processing is less computationally expensive and, consequently, faster. Taking this account, we performed the Wilcoxon Signed Rank Test, which is a non-parametric statistical hypothesis test [13], to verify if the Kappa decreasing is significant.

Thus, we consider the following hypothesis:

$\boldsymbol{H_0}$ =There is not difference among best Kappas and the others provided by each GVP value.
$\boldsymbol{H_1}$ =There is difference among best Kappas and the others provided by each GVP value.

Table 1 - Leafs and tree size for each MOL value.

MOL	Leafs	Tree size
2	24	47
5	25	49
10	22	43
15	20	39
20	16	31
25	15	29
30	14	27
35	13	25
40	12	23

Figure 3 shows the p-values for each pair of GVP values.

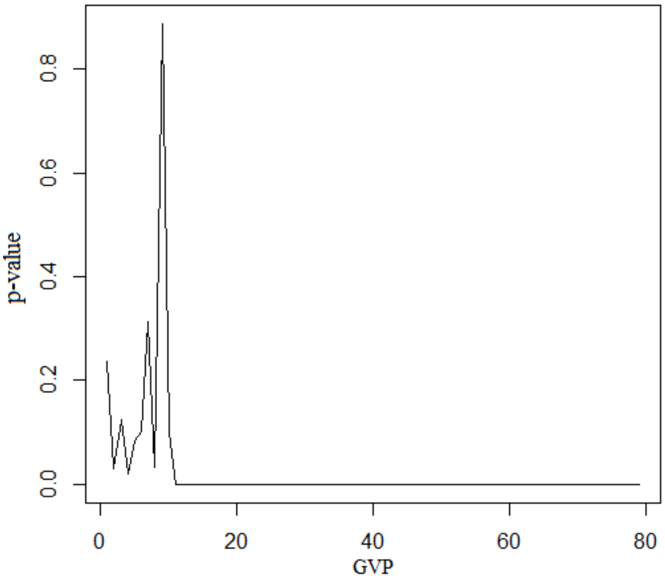

Fig. 3 – C4.5 algorithm: p-value for each GVP tested with the best Kappa values.

B. Multi Layer Percepctron classifier

The MLP classifier was tested with different NHN to evaluate its behavior in relation to this parameter. We can observe, in general, that when NHN increases so do the Kappa values. However, the greater NHN the longer is the processing time. The same hypothesis test was applied to this algorithm, as shown in Figure 4.

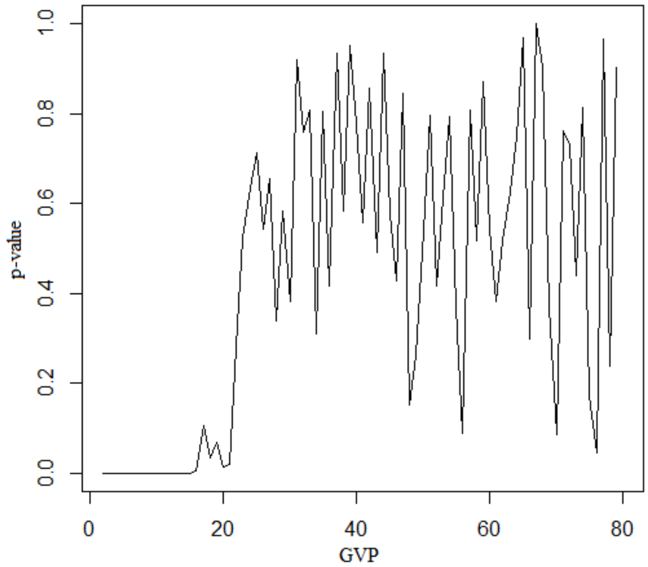

Fig. 4 – MLP algorithm: p-value to each GVP tested with the best Kappa values .

C. K – Nearest Neighbor classifier

When the K value increases the Kappa values decrease. Furthermore, the computational cost is high and increases as K increase. The same hypothesis test was applied to this algorithm, as shown in Figure 5.

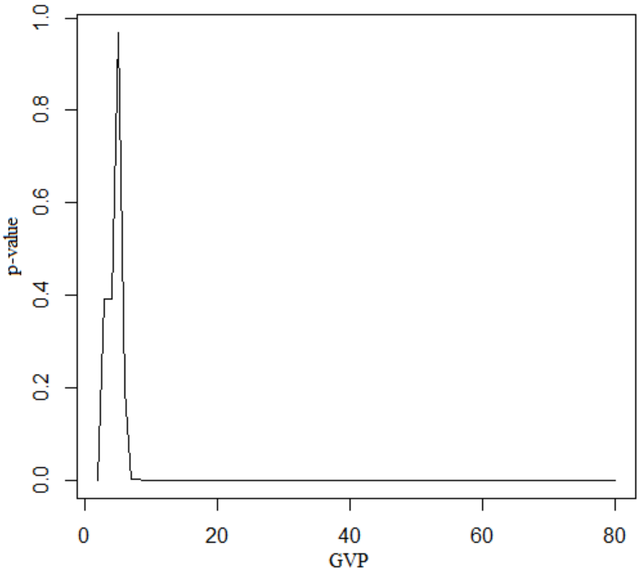

Fig. 5 – KNN algorithm: p-value to each GVP tested with the best Kappa values.

D. Comparing the classification algorithms

For C4.5 algorithm, we observed that for MOL > 9, with confidence level of 95%, H_0 hypothesis is rejected. Then, for MOL ≤ 9 the Kappa decreasing is not significant (Figure 3).
For MLP algorithm, for NHN < 20, with confidence level of 95%,, H_0 hypothesis is rejected, as we can see in Figure 4. Then, for NHN ≥ 20 the Kappa increasing is not significant.
Finally, for KNN algorithm, we observed in Figure 5 that for K > 7, with confidence level of 95%, H_0 hypothesis is rejected. Then, for K ≤ 7 the Kappa decreasing is not significant.
Therefore, give the best GVP values for each classification algorithm, we chose their respective kappa values to perform Wilcoxon Signed Rank Test among them, as presented in Table 2.

Table 2 - Optimum GVP values and their respective Kappa.

	MOL=9	K=7	NHN=20		MOL=9	K=7	NHN=20
	C4.5	**KNN**	**MLP**		**C4.5**	**KNN**	**MLP**
Kappa 1	0.6975	0.6729	0.8173	Kappa 16	0.7379	0.7259	0.7858
Kappa2	0.6821	0.619	0.7576	Kappa 17	0.7488	0.6892	0.7797
Kappa 3	0.6843	0.6435	0.7657	Kappa 18	0.7279	0.6801	0.7936
Kappa 4	0.7263	0.6845	0.7846	Kappa 19	0.7108	0.7088	0.7751
Kappa 5	0.7544	0.6273	0.7856	Kappa 20	0.7692	0.6909	0.794
Kappa 6	0.7305	0.6647	0.7981	Kappa 21	0.7235	0.6917	0.7728
Kappa 7	0.7939	0.6383	0.8127	Kappa 22	0.6527	0.7037	0.7905
Kappa 8	0.6851	0.6405	0.7758	Kappa 23	0.7317	0.6907	0.8126
Kappa 9	0.6463	0.6817	0.7953	Kappa 24	0.6339	0.649	0.8069
Kappa 10	0.7257	0.6854	0.7974	Kappa 25	0.7269	0.7045	0.7988
Kappa 11	0.6713	0.6763	0.7828	Kappa 26	0.7256	0.6696	0.7742
Kappa 12	0.7934	0.688	0.7885	Kappa 27	0.7761	0.689	0.7482
Kappa 13	0.7224	0.673	0.7882	Kappa 28	0.7135	0.695	0.7783
Kappa 14	0.7767	0.651	0.8043	Kappa 29	0.7528	0.6764	0.8181
Kappa 15	0.7085	0.6876	0.8126	Kappa 30	0.7585	0.6665	0.8172

Thus, we have the following hypothesis:

$\boldsymbol{H_0}$ = There is not difference among X and Y
$\boldsymbol{H_I}$ = There is a difference between X and Y classifier, and X is better than Y.

As we observe in Table 3, H_0 was rejected for all tests, which implies that MLP algorithm is better than the others.

Table 3 - Hypothesis tests.

X	Y	p-value
C4.5	KNN	3.46E-06
MLP	C4.5	1.30E-08
MLP	KNN	9.31E-10

III. Conclusion

For a large amount of data, the execution of manual tasks is difficult and time-consuming. In this sense, the use of data mining techniques can be used to reduce the processing time as well as to select the most appropriate parameters for the classification.

The experimental results showed that the optimum parameters for each algorithm are: MOL = 9 (C4.5), NHN = 20 (MLP), and K=7 (KNN). This MOL value represents a more generalized and small tree. These characteristics facilitate the rules creation in GEOBIA systems as eCognition [15], for example. Furthermore, generalized trees allow us to quickly observe the best attributes to distinguish the targets.

In relation to the MLP algorithm, the classification accuracy was similar for both NHN=80 and NHN=20. However, for NHN=80 the MLP algorithm takes in average 22 minutes, and for NHN=20 it spends 5 minutes in average.

In our experiments, we found that the MLP algorithm obtained better classification accuracy than the KNN and C4.5, while C4.5 was better than KNN. However, it is very difficult to translate the MLP structure into semantic networks. On the other hand, although the C4.5 algorithm have had worse classification accuracy than MLP, it is and a good alternative to be adapted to semantic networks. Besides, there are free GEOBIA softwares such as InterImage [16] and GeoDMA [7] that support semantic network implementation.

Acknowledgment

This work was supported by a fellowship from CNPq (National Council of Technological and Scientific Development), Brazil. We are very grateful for the contributions and suggestions provided by Dr. Camilo Rennó, Márcio P. Mello, and Dr. Thiago S. F. Silva from INPE, during the project development.

IV. References

[1] HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update. SIGKDD Explorations, v. 11, n. 1, p. 10-18, 2009.

[2] DEUS, L. R. Classificação da cobertura do solo urbano utilizando imagens IKONOS II e dados LiDAR. São José dos Campos: INPE, 2010.

[3] PINHO, C. M. D. Análise orientada a objetos de imagens de satélites de alta resolução espacial aplicada à classificação de cobertura do solo no espaço intra-urbano: o caso de São José dos Campos. 2005. 180 p. (INPE- 14183-TDI/1095). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2005.

[4] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence Magazine. p. 37-54, 1996.

[5] HAN, Jiawei; KAMBER, Micheline. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publisher, 2006.

[6] PINHO, C. M. D. ; FONSECA, L. M. G. ; KORTING, T. S.; ALMEIDA, C. M.; KUX,H. J. H. Land cover classification of an intra-urban scene using high-resolution images and object-based image analysis. International Journal of Remote Sensing, 2012.

[7] KORTING, T. S.; FONSECA, L. M. G.; ESCADA, M. I. S.; CÂMARA, G. GeoDMA: um sistema para mineração de dados de sensoriamento remoto. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14, 2009, Natal. Anais... São José dos Campos: INPE, 2009.

[8] TAN, Pang-Ning; STEINBACH, Michael. kNN: k-Nearest Neighbors. In: WU, Xindong; KUMAR, Vipin (Ed.). The Top Ten Algorithms in Data Mining. Boca Raton: Chapman and Hall/CRC, Taylor and Francis Group, 2009. cap. 8, p. 151-161

[9] AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-Based Learning Algorithms. Machine Learning, v. 6, p. 37-66, 1991.

[10] LEONARDI, F. Abordagens cognitivas e mineração de dados aplicadas a dados ópticos orbitais e de laser para a classificação de cobertura do solo urbano. 2010. 162 p. (sid.inpe.br/mtc-m19@80/2010/03.17.11.42-TDI). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2010.

[11] WITTEN, I. H. ; FRANK, E.;The WEKA Data Mining Software: An Update. SIGKDD Explorations, v. 11, n. 1, p. 10-18, 2009.

[12] RAMAKRISHNAN, Naren. C4.5. In: WU, Xindong; KUMAR, Vipin (Ed.). The Top Ten Algorithms in Data Mining. Boca Raton: Chapman and Hall/CRC, Taylor and Francis Group, 2009. cap. 8, p. 151-161.

[13] PINHO, C. M. D. ; UMMUS, M. E. ; NOVACK, T. Extração de feições urbanas em imagens de alta resolução espacial a partir do estudo do comportamento espectral dos alvos. RBC. Revista Brasileira de Cartografia,2011.

[14] WONNACOTT, T.H.; WONNACOTT. Introdução à Estatística. Livros Técnicos e Científicos Editora S.A., R.J .1980.

[15] LANG, S. ; D. TIEDE. Definiens Developer. GIS Business 9 2007; 34-37.

[16] G. A. O. P. COSTA, R. Q. FEITOSA, L. M. G. FONSECA, D. A. B. OLIVEIRA, R. S. FERREIRA, E. F. CASTEJON, "Knowledge-based interpretation of remote sensing data with the InterIMAGE system: major characteristics and recent developments". In: Proceedings of the 3rd GEOBIA 2010, Ghent, Belgium.