Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 7871-7876.

Classificação da cobertura do solo urbano inserindo árvores de decisão a rede hierárquica

Tessio Novack[1]
Hermann Kux[1]

[1]Instituto Nacional de Pesquisas Espaciais – INPE
Caixa Postal 515 – 12245-970 – São José dos Campos – SP, Brasil
{tessio, hermann}@dsr.inpe.br

Abstract. This paper proposes the use of the C4.5 decision tree classifier integrated with hierarquical nets for the purpose of urban land cover classification using an image from the QuickBird II sensor. For the separation of the classes on each level of the class-hierarquy a decision tree was created. This approach makes the definition of features and thresholds to be applied on every class description totally automatic and without the intervention from the analyst. The only subjectivities of the proposed methodology are the creation of the class-hierarquy itself and the sample selection. Different class-hierarquies, feature selection methods and minimal number of instances on the decision tree leafs were tried. At total, eighteen classifications were generated. These were evaluated quantitatively by the number of leafs, number of nodes, number of features and Kappa index value criteria. After discussions, a final classification was elected ranking these criteria. It was learned that the FCBF feature selection method makes the decision trees simpler and with fewer features at the same time that it enhances the efficiency of the classification in terms of Kappa index calculated by cross-validation. A visual inspection over the image classified by this elected model attested the success of the proposed methodology.

Palavras-chave: urban land cover, object-oriented image analysis, feature selection, decision tree, cobertura do solo urbano, análise orientada a objeto, seleção de atributos, árvore de decisão.

1. Introdução

O sensoriamento remoto, devido à crescente disponibilidade de sensores de alta resolução espacial e ao surgimento de novas metodologias para a extração de informação destas imagens, tem sido uma tecnologia cada vez mais utilizada na tomada de decisão a respeito de muitos aspectos do planejamento urbano (Blaschke e Kux, 2007; Yang, 2003). Um dos principais esforços por parte da comunidade de pesquisadores em sensoriamento remoto tem sido atualmente a concepção e avaliação de metodologias de classificação automática da cobertura do solo urbano (Pinho e Kux, 2005; Araujo e Kux, 2006; Costa et al, 2007; Darwish et al., 2003; Sande et al., 2003). Entre os objetivos de grande parte destes trabalhos está a busca por automação e, conseqüentemente, por possibilidades de exportação de tais metodologias para outras áreas de estudo e outras datas. Muitas vezes, no entanto, o potencial de automação inerente a estas metodologias é fragilizado pelo fato dos atributos e limiares estabelecidos sob as áreas testes não gerarem resultados tão satisfatórios em outras áreas, datas ou imagens de sensores com características similares. Neste sentido, a inserção de técnicas de mineração de dados pode tornar a escolha de atributos e limiarização (processos estes que tomam demasiado tempo) muito mais rápida e igualmente adaptável para qualquer área em que se queira classificar a cobertura do solo urbano. A inserção de técnicas de mineração de dados no contexto de classificação orientada a objetos deixa a cargo do analista apenas os processos de elaboração da rede hierárquica, segmentação e coleta de amostras. Esta proposta é ainda mais pertinente nos casos em que dispomos de uma gama muito grande de atributos tanto espectrais quanto de textura e forma como nos sistemas eCognition e Definiens Developer (Definiens, 2008).

1.1 Objetivo

Este trabalho propõe a inserção de árvores de decisão nos níveis de uma rede hierárquica de classes como forma de definir automaticamente atributos e limiares na classificação da cobertura do solo urbano a partir de imagem de alta resolução.

2. Metodologia

2.1 Segmentações da imagem

Esta etapa teve como finalidade a geração de segmentos de imagem que representassem o melhor possível os objetos de cobertura do solo em nossa área de estudo. O algoritmo de segmentação utilizado foi o de Baatz e Schäpe (2000). Os parâmetros deste segmentador usados em outros estudos similares ao nosso foram considerados como base e ponto inicial para a experimentação e descoberta dos parâmetros mais adequados para as classes de cobertura do solo presentes em nossa área de estudo (Darwish et al., 2003; Sande et al., 2003; Araujo e Kux, 2006; Tian e Chen, 2007). A Tabela 1 mostra os parâmetros utilizados para a segmentação da imagem após o processo de experimentação por tentativa e erro. Mais de um valor foi utilizado para cada parâmetro, pois a classificação por processo (sessão 2.6) permite re-segmentações da imagem concomitantemente a classificação hierárquica da cena.

Tabela 1. Parâmetros utilizados nas segmentações da imagem.

Classe	P_{forma}	$P_{compacidade}$	***Fator de escala***
Vegetação Geral Vegetação Arbórea Vegetação Rasteira Piscinas	0.5	0.1	25
Outras Classes	0.7	0.8	40

2.2 Coleta de amostras e exportação de atributos

Na etapa de coleta de amostras, cuidou-se para que toda a variabilidade de feições das classes fosse considerada. Exatamente trinta amostras (segmentos) foram coletadas para cada classe cobrindo toda a heterogeneidade destas quanto à forma, comportamento espectral e textura. Em seguida todos os atributos nestas três categorias disponíveis no sistema Definiens Professional, além dos atributos customizados divisão da banda do infravermelho próximo pela banda do vermelho e divisão da banda do vermelho pela banda do azul foram exportados para cada uma das classes. Ao todo, trezentos e cinqüenta e cinco atributos foram exportados.

2.3 Geração de redes hierárquica

As redes hierárquicas devem ser entendidas aqui como formas de resolução do problema de classificação da cobertura do solo na área de estudo. As redes hierárquicas foram geradas utilizando a estratégia de alocar nos níveis superiores das redes as classes de maior separabilidade, ou seja, as classes mais fáceis de serem extraídas e em que há menor confusão com as classes do mesmo nível. Os objetos foram divididos, a *priori*, como pertencentes ou não às classes de maior facilidade de separação. A hierarquia se desenvolveu em decorrência da não-associação dos objetos às classes superiores, estratégia esta utilizada por Araujo e Kux (2006) e Pinho e Kux (2005). Três redes hierárquicas foram geradas procurando-se descobrir a melhor estruturação em níveis possível para as classes consideradas neste estudo (Figura 1).

2.4 Seleção de atributos

O principal propósito da etapa de seleção de atributos é a identificação dos atributos mais relevantes e eliminação de atributos redundantes. Entende-se por atributos relevantes aqueles que apresentam alta correlação com as classes e baixa correlação com outros atributos. Além disto, a seleção de atributos tem os efeitos desejáveis de redução do tamanho das árvores de decisão, diminuição do tempo de geração e simplificação das árvores e, em muitos casos, aumento da acurácia da classificação. Neste trabalho, três métodos de seleção de atributos foram testados em conjunto com três avaliadores de atributo, sendo eles: (1) *Correlation-*

based feature selection (CFS) com o avaliador *Best First* (Hall, 1999); (2) RELIEF com o avaliador *Ranker* (Kira e Rendell, 1992) e (3) *Fast Correlation-based Feature Selection* (FCBF) com o avaliador *Assymetric Subset Evaluator* (Yu e Liu, 2003). Estes três algoritmos estão disponíveis no aplicativo Weka 3.5.8 disponível para download na Internet.

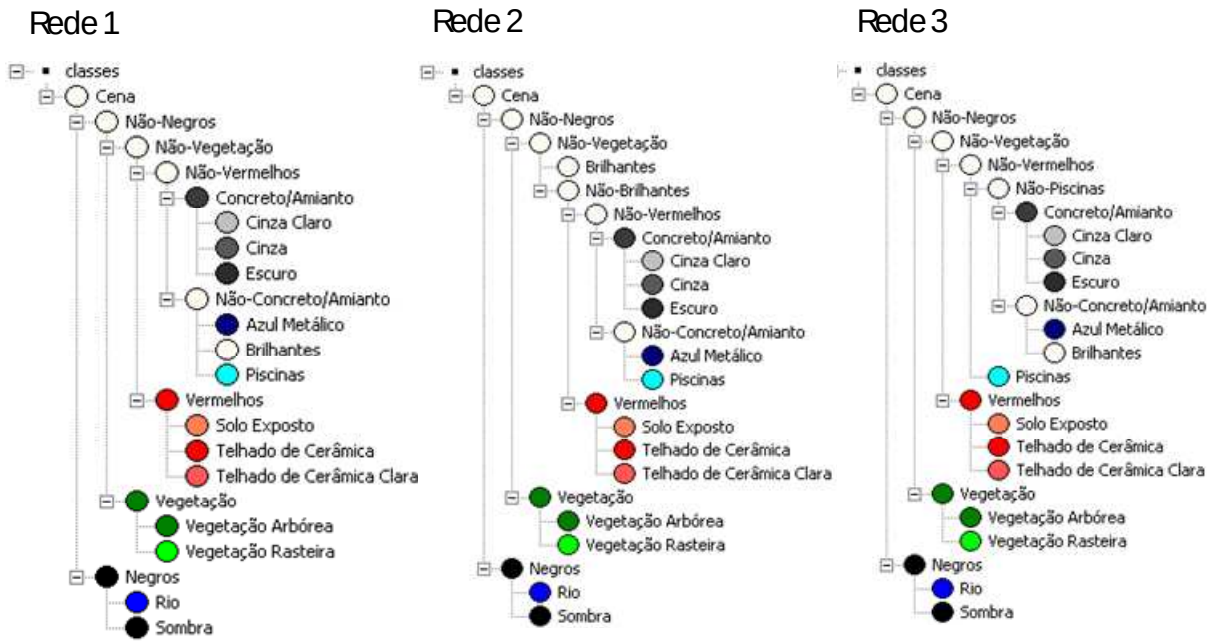

Figura 1. Redes hierárquicas de classes testadas

2.5 Geração das árvores de decisão

Algoritmos de árvore de decisão estão entre os métodos mais conhecidos e explorados em aprendizado por máquina. Um dos mais populares e eficientes destes algoritmos é o C4.5 (Quinlan, 1993), utilizado neste trabalho. O C4.5 utiliza uma abordagem recursiva de particionamento dos dados, ou seja, identifica o atributo e o ponto de separação deste atributo que melhor separa ou discrimina as classes. A mesma lógica é empreendida aos subconjuntos até que todas as classes sejam inteiramente separadas. Para cada nível de separação de classes das redes hierárquicas uma árvore de decisão foi gerada utilizando previamente os diferentes métodos de seleção de atributos e estabelecendo um número mínimo de instâncias por folha da árvore. Ao todo, dezoito modelos de classificação foram criados utilizando os diferentes métodos de seleção de atributos e número mínimo de instâncias por folha em cada uma das três redes (Tabela 2).

2.6 Classificação por processo

Assim como na classificação por rede hierárquica, a abordagem por processo classifica a imagem de cima para baixo (*top-down*), ou seja, separa as classes mais gerais ou de maior separabilidade para posteriormente separar as subclasses ou classes de maior incerteza, classificando partes da imagem seqüencialmente. A vantagem possibilitada pela abordagem por processo é que podemos re-segmentar os objetos da imagem independente da etapa ou contexto da classificação. Isto foi explorado e deu maior controle e flexibilidade à metodologia. Assim, as redes hierárquicas foram, na prática, simuladas no processo de classificação já que diferentes classes tiveram seus objetos gerados por diferentes segmentações.

Tabela 2. Rede hierárquica, método de seleção e avaliação de atributos e número mínimo de instâncias por folha nas árvores de decisão das dezoito classificações empreendidas.

Rede Hierárquica	**Método de seleção de atributos**	**Avaliador de atributos**	**Nº min. de amostras por folha**
1	CFS	Best First	2
			5
	RELIEF	Ranker	2
			5
	FCBF	Assymetric Subset Evaluator	2
			5
2	CFS	Best First	2
			5
	RELIEF	Ranker	2
			5
	FCBF	Assymetric Subset Evaluator	2
			5
3	CFS	Best First	2
			5
	RELIEF	Ranker	2
			5
	FCBF	Assymetric Subset Evaluator	2
			5

Tabela 3. Classificações empreendidas e critérios de avaliação.

Classificação			**Critérios**				
Seleção de atributos	Rede	Num. min. de instâncias por folha	Num. de nós	Num. de folhas	Num. de atributos	Num. de atributos diferentes	Ind. Kappa
FCBF	2	5	25	16	7	7	0.9452
CFS	2	5	25	16	8	7	0.9372
RELIEF	2	5	25	16	9	7	0.926
FCBF	3	5	27	17	8	6	0.9556
CFS	3	5	27	17	10	7	0.9326
RELIEF	3	5	27	17	10	7	0.9325
CFS	1	2	28	17	10	8	0.9441
CFS	1	5	28	17	10	8	0.9441
FCBF	1	5	28	17	9	7	0.9434
RELIEF	1	2	28	17	11	10	0.9391
RELIEF	1	5	28	17	11	10	0.9353
FCBF	3	2	31	19	9	7	0.9579
CFS	3	2	33	20	11	8	0.9347
RELIEF	3	2	33	20	12	9	0.9327
FCBF	1	2	34	20	10	8	0.9546
FCBF	2	2	35	21	10	9	0.9519
CFS	2	2	35	21	13	11	0.9398
RELIEF	2	2	35	21	13	10	0.9254

3. Resultados e discussões

A avaliação das classificações por diferentes redes hierárquicas, métodos de seleção de atributos e número mínimo de instâncias por folha das árvores de decisão foi feita sob os critérios de tamanho total das árvores de decisão dentro das redes (número de nós de todas as árvores de decisão dentro da rede hierárquica), número total de folhas das árvores de decisão dentro de cada rede, número de atributos usados na classificação, número de atributos diferentes usados na classificação e índice Kappa calculado na validação cruzada das

amostras. A Tabela 3 apresenta as classificações ranqueadas pelo número de nós de todas as árvores de decisão inseridas na rede. Percebe-se que a forma de classificação mais simplificada é aquela estruturada pela rede 2 e que usa o método FCBF de seleção de atributos e cujas árvores de decisão têm no mínimo cinco instâncias em cada folha. Esta classificação apresentou também o menor número de atributos, o menor número total de folhas e um índice Kappa calculado sob a validação cruzada das amostras dos mais altos (0.9452). Assim, este modelo de classificação foi empreendido sobre a imagem como um todo e o resultado foi avaliado por inspeção visual como satisfatório (Figura 2), o que corrobora a aplicabilidade da metodologia.

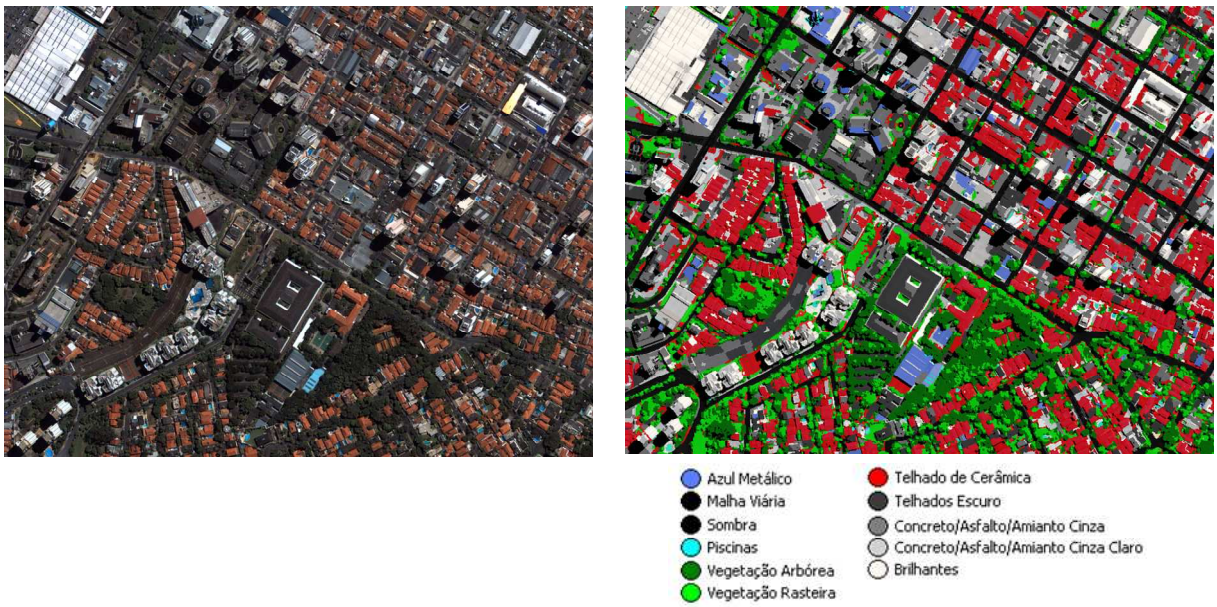

Figura 2. Classificação empreendida com a rede hierárquica 2, método de pré-seleção de atributos FCBF e número mínimo de cinco instâncias por folha da árvore de decisão.

Pode-se notar também pela Tabela 3 que os cinco mais altos índices Kappa calculados na validação cruzada das amostras pertencem às classificações em que o método de seleção de atributos FCBF foi usado. O próprio método e heurística do FCBF fazem com que ele selecione apenas os atributos mais relevantes e menos redundantes entre si, o que faz com que o modelo de classificação seja mais compacto e, por isso, mais eficiente e preciso. Outras classificações que também se mostraram compactas quanto ao número de atributos e nós das árvores de decisão e que também apresentaram bons índices Kappa foram empreendidas sobre a imagem como um todo. No entanto, o resultado não nos pareceu tão preciso quanto o modelo de classificação melhor ranqueado por estes critérios.

4. Conclusões

Este trabalho procurou mostrar que a combinação entre rede hierárquica, como modelo de solução do problema, e árvores de decisão, como forma de definição de atributos e limiares, é uma alternativa viável, simples e que surti bons resultados na classificação da cobertura do solo urbano a partir de imagens de alta resolução. A metodologia aqui proposta pode ser facilmente exportada para outras áreas já que ficará a cargo do aprendizado por máquina (algoritmo C4.5) definir atributos e limiares mais adequados para a separação das amostras. Nesta abordagem, fica a cargo do especialista apenas a criação da rede hierárquica e a coleta de amostras.

Referências

Araújo, E.H.G. **Análise multi-temporal de cenas do satélite QuickBird usando um novo paradigma de classificação de imagens e inferências espaciais: estudo de caso Belo Horizonte (MG)**, 2006, 159 p. Dissertação (Mestrado em Sensoriamento Remoto), INPE, São José dos Campos, 2006.

Baatz, M.; Schäpe, A. Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation. In: STROBL, J., BLASCHKE, T.Angewandte Geographische Informationsverarbeitung XII. Beiträge zum AGITSymposium Salzburg 2000. Karlsruhe. Herbert Wichmann Verlag, 2000. p. 12 – 23. Disponível em: <http://www.definiens.com/pdf/publications/baatz_FP_12.pdf>. Acesso em: 23 maio 2003.

Blaschke, T.; Kux, H. (editores) **Sensoriamento remoto e SIG avançados: novos sistemas e sensores inovadores**, 2ª. Edição, São Paulo: Oficina de Textos, 2007, 304 p., 2007.

Costa, G. A. O. P., Pinho, C. M. D., Feitosa, R. Q., Almeida, C. M., Kux, H. J. H., Fonseca, L. M. G., Oliveira, D. A. B. InterImage: na open source platform for automatic image interpretation. II Simpósio Brasileiro de Geomática/V Colóquio Brasileiro de Ciências Geodésicas. **Anais...** Presidente Prudente, 2007.

Darwish, A., Leukert, K., Reinhardt W., Image segmentation for the purpose of object-based classification. **IEEE**, v.2, n.3, p. 2035-2046, 2003.

Definiens: Disponível em: <http://www.definiens-imaging.com/down/ecognition>. Acesso em: 28 de fev. 2008.

Hall, M. A., Correlation-based feature selection for machine learning. Tese de Doutorado em Ciências da Computação, Universidade de Waykato, Nova Zelândia, 1999. Disponível em: Acesso em: 18 set. 2008.

Kira K., Rendell, L. A. A practical approach to feature selection. Ninth International Conference on Machine Learning. **Anais**… San Francisco, CA: Morgan Kaufmann, pp. 249–256, 1992.

Pinho, C.M.D. **Análise orientada a objetos de imagens de satélite de alta resolução espacial aplicada à classificação de cobertura do solo no espaço intra-urbano: o caso de São José dos Campos-SP, São José dos Campos**, 2005, 181 p. 2005 (INPE-14183-TDI/1095) Dissertação (Mestrado em Sensoriamento Remoto, INPE, 2005.

Quinlan, J. R., **Programs for Machine Learning**. San Mateo: Morgan Kaufmann,1993.

Sande, C. J., Jong, S. M., Roo, A. P. J., A segmentation and classification approach of IKONOS-2 imagery for land cover mapping to assist flood risk and flood damage assessment. **International Journal of Applied Earth Observation and Geoinformation** v.4 n.2, p.217-229, 2003.

Tian, J. and Chen, D. -M. Optimization in multi-scale segmentation of high-resolution satellite images for artificial feature recognition, **International Journal of Remote Sensing**, v.28, n.20, p.4625 – 4644, 2007.

Yang, X.; Remote sensing and GIS for urban analysis: an introduction. **Photogrammetric Engineering & Remote Sensing**, v. 69, n. 9, p. 937-939, 2003.

Yu, L., Liu, Y., Feature selection for high-dimensional data: a fast correlation-based filter solution. Twentieth International Conference on Machine Learning. **Anais**… Washington DC, 2003.