CLASSIFICAÇÃO DE IMAGENS POLINSAR UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

POLINSAR IMAGE CLASSIFICATION USING DATA MINING TECHNICS

Carlos Alberto Pires de Castro Filho [1,2], João Roberto dos Santos [2]

[1] Diretoria de Serviço Geográfico - DSG, Quartel General do Exército, Bloco "F", 2º Piso, Setor Militar Urbano, 70630-901 – Brasília, DF
[2] Instituto Nacional de Pesquisas Espaciais - INPE, Av. dos Astronautas, 1758, 12.227-010 - São José dos Campos, SP, Brasil, pires@dpi.inpe.br, jroberto@dsr.inpe.br

RESUMO

A Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases*), ou KDD, objetivam gerar técnicas para análise de dados através de algoritmos de mineração. No Subprojeto Cartografia Terrestre, da Diretoria de Serviço Geográfico – DSG – está previsto o imageamento de uma área de aproximadamente 770.000 km^2 da região amazônica utilizando tecnologia de Radares de Abertura Sintética Interferométricos e Polarimétricos. O objetivo deste trabalho é de analisar o potencial de dados de SAR para classificação de uso do solo. Nesta análise foram utilizadas técnicas de mineração de dados identificando quais tipos de atributos são os mais adequados para discretizar as classes a serem definidas. Além destas técnicas foram também selecionados atributos que melhor classificaram separadamente a imagem através de uma árvore de decisões. Os resultados obtidos indicaram que a classificação com os melhores atributos obtidos separadamente nas etapas de treinamento obtiveram melhor avaliação. Conclui-se que apesar dos resultados terem sido melhores com o método proposto, a avaliação da classificação com os atributos selecionados automaticamente se aproximou bastante.

Palavras-chave: Mineração de Dados, classificação, sensoriamento remoto, radar.

ABSTRACT

Knowledge Discovery in Databases – KDD – is intended to generate new techniques to analyze data through data mining algorithms. In the brazilian Terrestrial Cartography Subproject, also known as "Amazon Radiography", of the Geographic Service of Brazilian Army (DSG), is expected the imagery of an area of approximately 770,000 km2 of the amazon region, using the Polarimetric Interferometric and Synthetic Aperture and Radar technology. The aim of this study is to examine the potential of SAR data for land use classification. Data mining techniques were used to identify the features that best discretized the classes. In addition to these techniques, features that best separately classified the image using decision tree were also selected. The results indicates that the best classification evaluation was obtained with the features with best results separately. We conclude that although the results were better with the proposed method, the evaluation of the classification with the automatically selected attributes were very close.

Keywords: Data Mining, classification, remote sensing, radar.

INTRODUÇÃO

A dinâmica do mundo moderno contribui com o aumento diário de dados e informações disponíveis na sociedade humana. Proporcionalmente ao crescimento dos dados disponíveis, faz-se necessário o desenvolvimento de técnicas e ferramentas que visem organizá-los e analisá-los de forma inteligente e automática. Enquanto os Bancos de Dados visam a organização, os estudos de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases*), ou KDD, objetivam gerar novas técnicas e

ferramentas capazes de automatizar a análise de grandes volumes de dados, utilizando-se para isto de metodologias científicas (FAYYAD *et al*, 1996).

De acordo com MAIMON e ROKACH (2005), a etapa de mineração de dados é o núcleo do processo de KDD já que envolve a inferência de algoritmos que exploram os dados através do desenvolvimento de modelos para o reconhecimento de padrões.

Atualmente observa-se que o desenvolvimento tecnológico e a necessidades de conhecimento do terreno, geraram um crescimento exponencial de sensores remotos, que produzem imensas quantidades de dados. De acordo com GE *et al*. (2008), dados geo-espaciais podem ser bastante beneficiados com técnicas de mineração, visto que estas técnicas podem ser aplicadas tanto em dimensões espaciais, espectrais ou temporais.

No Brasil, entre os projetos em execução que visam obter dados de sensores remotos para mapeamento, ou produção de conhecimento geo espacial, encontra-se em destaque o de Cartografia da Amazônia, mais especificamente o Subprojeto Cartografia Terrestre, também conhecido como Radiografia da Amazônia. Neste projeto será recoberta uma área de aproximadamente 770.000 km^2 da região amazônica, visando a futura confecção de cartas na escala 1:50.000 pela Diretoria de Serviço Geográfico (DSG) do Exército Brasileiro. Para tal, será utilizada a tecnologia de Radares de Abertura Sintética Interferométricos e Polarimétricos – PolInSAR – aerotransportados, o que irá gerar diversos dados nas bandas X e P polarimétricas, modelos digitais do terreno e de superfície e bandas de coerência interferométrica.

Estes produtos serão utilizados como dados para a confecção de diversos documentos cartográficos, dentre os quais se destaca a geração das bases cartográficas de 616 (seiscentos e dezesseis) folhas, na escala 1:100.000, jamais produzidas. A utilização da tecnologia SAR, que independe de fatores climáticos e da iluminação solar, possibilitará o mapeamento destes "vazios cartográficos" em regiões que, por possuírem constante presença de nuvens, apresentavam-se como um desafio.

Neste cenário, o objetivo deste trabalho é de analisar o potencial de dados de SAR para classificação de uso do solo. Nesta análise serão utilizadas técnicas de KDD, buscando minerar e, consequentemente, identificar quais tipos de atributos são os mais adequados para discretizar as classes a serem definidas.

DESENVOLVIMENTO

A área de trabalho foi selecionada de acordo com a disponibilidade das imagens SAR. Neste sentido foram analisadas cenas de uma região próxima a cidade de Barcelos – AM, com coordenadas variando entre de 63º00' e 63º15'w/ 0º44' e 0º58's, no sistema geodésico WGS-84, e com área de cerca de 800km^2. Tais imagens seguem, ainda, as características técnicas conforme consta da Tabela 1, sendo adquiridas através de um imageamento aerotransportado datado de 30 de março de 2009.

Visando futuras análises acerca das classes relativas ao uso do solo presentes na imagem SAR, foram selecionadas imagens óticas como dados auxiliares. As imagens óticas, por sua vez, foram selecionadas de tal forma que englobassem a mesma área da imagem SAR, que fossem de uma data próxima e que apresentassem um baixo percentual de nuvens. Portanto, foi escolhida uma cena do sensor Landsat 5 TM, nas bandas 1, 2, 3, 4, 5 e 7, de órbita-ponto 233-061 e datada de 10 de dezembro de 2008.

A partir das bandas de trabalho selecionadas iniciou-se um processo de preparo que teve as seguintes etapas:
- reamostragem espacial das bandas Landsat para que obtivessem o mesmo tamanho de pixel da imagem SAR, isto é, 5,0 metros de resolução espacial;
- utilização de filtro adaptativo GAMMA, com janela 5x5, sobre as imagens SAR visando minimizar o efeito do ruído *speckle*;
- mudança no sistema de coordenadas e sistema geodésico, fazendo com que todas as imagens estivessem em coordenadas UTM, zona 20S e sistema WGS-84;
- registro entre bandas SAR e Landsat, utilizando as primeiras como base, aplicando polinômio do 1º grau e reamostragem pelo método do vizinho mais próximo. Para cada registro foram selecionados no mínimo 20 (vinte) pontos homólogos e o maior erro médio quadrático obtido foi de 0,13 pixel; e

- corte das bandas Landsat, diminuindo a área da cena para que fosse igual a de trabalho, definida pelas imagens SAR.

Extração de Atributos

Após o preparo das imagens Landsat e SAR, pode-se dar início ao processo de KDD. A primeira etapa é a extração de novos atributos, definindo o espaço a ser trabalhado. Neste trabalho optou-se por utilizar, além das bandas polarimétricas e de coerência interferométrica SAR, também a altura interferométrica. Através da operação matemática de diminuição entre o Modelo Digital de Superfície (DSM) e do Modelo Digital do Terreno (DTM), foi gerada a banda de altura interferométrica. A altura interferométrica, em regiões de densa floresta, representa a altura das árvores, podendo ser um atributo que poderá posteriormente auxiliar na descrição das classes. Neste trabalho será usada a abreviação de Hint para este atributo.

Segundo GAMA (2007), a altura interferométrica contribui para analisar a variabilidade da estrutura florestal. Juntamente a isto, outros atributos derivados deste também podem ser representativos, como seu logaritmo e seu quadrado. Neste sentido, foram calculadas as bandas LogHint e Hint2 as quais também foram definidas como atributos.

Além dos atributos disponíveis, foram extraídos também outros parâmetros os quais são associados a diferentes tipos de vegetação. Logo, foram confeccionadas as bandas de Razão de polarização paralela (Rp), de Razão de polarização cruzada (Rc) e de Potência total (PotSpan). Juntamente a estes parâmetros, as bandas relativas aos Índices Incoerentes de Pope também foram confeccionadas: índice de biomassa (BMI), índice de estrutura do dossel (CSI) e índice de espalhamento volumétrico (VSI).

Para cada banda a ser utilizada como atributo foi dado então um código, visando facilitar sua identificação. As características do espaço de atributos a ser trabalhado encontra-se na Tabela 1.

Tabela 1. Características do Espaço de Atributos.
Table 1. Feature Space Characteristics.

Código / Banda	Discriminação
P_HH	Orto-imagem da banda P de polarização HH
P_HV	Orto-imagem da banda P de polarização HV
P_VV	Orto-imagem da banda P de polarização VV
P_Coh	Banda P de coerência interferométrica
X_HH	Orto-imagem da banda X de polarização HH
X_Coh	Banda X de coerência interferométrica
Hint	Altura interferométrica DSM-DTM
Hint2	Altura interferométrica ao quadrado
LogHint	Logaritmo de altura interferométrica
VSI	Índice de espalhamento volumétrico
CSI	Índice de estrutura do dossel
BMI	Índice de biomassa
Rp	Razão de polarização paralela
Rc	Razão de polarização cruzada
PotSpan	Potência total

Definição das Classes

Para delimitar as classes nas imagens foram utilizadas as imagens óticas Landsat, levando em consideração o formato, a resposta espectral, a textura e o contexto das feições. O uso desta imagem foi de suma importância visto que os limites entre as classes puderam ser fotointerpretados sobre imagens com data próxima à do imageamento SAR.

Buscando identificar os diferentes tipos de classes presentes na imagem, foram utilizados dados do Projeto RADAMBRASIL, de 1978. Através dele foram gerados relatórios e mapas temáticos na escala 1:1.000.000 a partir de uma base cartográfica 1.250.000. Dentre esses mapas temáticos, o fitoecológico apresenta os sistemas ecológicos integrados onde se destacam as classes relativas ao uso do solo.

Apesar da grande diferença de datas, observa-se que as classes praticamente não mudaram entre a imagem e os dados do RADAMBRASIL, havendo somente uma pequena mudança em algumas poucas áreas próximas ao rio Negro que sofreram antropização. Logo, uma das classes a serem trabalhadas neste trabalho será justamente a de áreas antropizadas. As demais áreas são as de água, campinarana arbórea, campinarana arbustiva, florestas primárias alagadas e florestas primárias de solo firme.

Seleção de instâncias e Pré-Processamento dos Dados

Visando treinar e, posteriormente à classificação, avaliar as classes a serem trabalhadas na imagem, buscou-se selecionar regiões amostrais, ou regiões de interesse (do inglês *Region of Interest* – ROI), observando as variações de nível de cinza nos diversos atributos. De acordo com SHEKHAR (2001) dados espaciais devem ser tratados de forma diferenciada aos dados com ocorrências independentes, isto é, sem influência de vizinhança. Afirma ainda que estes tipos de dados, no caso dos pixels, por possuírem forte influência dos vizinhos, podem ser manipulados pelo usuário de tal forma que os valores de seus atributos sejam mais representativos. Esta manipulação dos valores dos atributos é feita estipulando-se pesos para a influência que os valores dos atributos dos pixels vizinhos terão.

Desta forma, no presente trabalho optou-se por trabalhar com a média dos valores de atributos dos pixels pertencentes a cada um dos polígonos selecionados como ROI. Foram então selecionados 50 (cinqüenta) polígonos para serem usados como amostras de treinamento e 30 (trinta) como amostras de teste (ou validação) para cada classe. Em ambos os casos cada polígono selecionado possuía cerca de 200 pixels e foram distribuídos homogeneamente por toda a imagem. Esta seleção de polígonos foi realizada sobre a imagem ótica, utilizando técnicas de fotointerpretação as quais seriam utilizadas com maior dificuldade caso fossem aplicadas sobre as imagens de SAR. Ao término do processo, foi montada uma planilha, totalizando 300 linhas de instâncias referentes à todas as classes em trabalho, a qual foi importada para o programa Weka.

Mineração

Conforme informado no item anterior, o programa escolhido para mineração dos dados foi *o Waikato Environment for Knowledge* – Weka. Esta escolha deu-se em função de ser um programa livre, disponível através da página *http://www.cs.waikato.ac.nz/ml/weka/* da internet, e plenamente utilizado em diversos trabalhos científicos, obtendo bons resultados.

Apesar do Weka possuir diversas opções de algoritmos de mineração, optou-se pelo uso de uma "árvore de decisão". De acordo com QUINLAN (1993), o método de classificação por árvore de decisões tem como vantagem o fato de possuir natureza e propriedades não-paramétrica, podendo classificar imagens com distribuições estatísticas diferentes da gaussiana, heterogêneas e possuidoras de ruídos.

Uma das variantes mais conhecidas e usadas de árvores de decisão é a do algoritmo C4.5 (QUINLAN, 1993, citado por SANTOS, 2009). Escolheu-se então o uso da classificação pelo algoritmo J4.8, que é o nome dado à implementação feita em JAVA do C4.5 no Weka.

RESULTADOS

Mineração de Dados

Visando avaliar os atributos gerados e apresentados na Tabela 1, foi utilizada a função de seleção de atributos do Weka. Nesta função são selecionados os métodos de avaliação e de procura dos atributos desejados pelo usuário, sendo informado quais são os mais representativos para o caso de uma classificação.

No presente trabalho foram usados os métodos de avaliação por *CFS subset*, medida de Qui-quadrado, Principais Componentes e *Gain Ratio*. Já, os métodos de procura utilizados foram os de *Best First*, Exaustivo e *Ranking*. Os resultados obtidos encontram-se na Tabela 2.

Tabela 2. Métodos de seleção de atributos.
Table 2. Feature Selection Methods.

Método de avaliação	Método de procura	Atributos selecionados
CFS subset	*Best first*	P_HH, P_HV, P_VV, P_Coh, X_HH, Hint, CSI_P, Rc_P
CFS subset	Exaustivo	P_HH, P_HV, P_VV, P_Coh, X_HH, Hint, CSI_P, Rc_P
Qui-quadrado	*Ranking*	P_Coh, P_VV, P_HH, Hint, Hint2, X_HH, P_HV, BMI_P, logHint, PotSpan_P, X_Coh, Rc_P, CSI_P, Rp_P
Principais componentes	*Ranking*	P_HH, P_HV, P_VV, P_Coh, X_HH, X_Coh
Gain Ratio	*Ranking*	P_Coh, Hint, Hint2, X_HH, PotSpan_P, P_HH, P_VV, P_HV, logHint, X_Coh, Rc_P, BMI_P, CSI_P, Rp_P, VSI_P

Analisando os resultados obtidos na Tabela 4, foram definidos como os atributos mais representativos os seguintes: P_HH, P_HV, P_VV, P_Coh, X_HH, Hint, CSI_P e Rc_P. No entanto observa-se que estes atributos são os que apresentaram menor correlação entre eles, não sendo necessariamente os mesmos que melhor classificarão uma imagem.

Utilizando estes atributos selecionados pelo Weka e executando o algoritmo J4.8, foi obtida a árvore de decisões ilustrada na Figura 1. Esta árvore, com 14 (quatorze) folhas, obteve sobre as amostras de treinamento o resultado de 254 (84,6667%) instancias classificadas corretamente e valor Kappa de 0,816. Observa-se ainda que apesar de terem sido selecionados, os atributos relativos às bandas P_HH e Rc_P não foram utilizados na construção da árvore.

Visando testar a influencia de cada uma das bandas em trabalho na construção da árvore de decisões pelo algoritmo J4.8, foram realizados testes onde optou-se por executar o algoritmo diversas vezes, sendo que em cada uma das vezes seria utilizada somente uma das bandas como atributo. Os resultados encontram-se na Tabela 3.

Através da Tabela 3, nota-se que os melhores resultados (resultados com percentual de acerto acima dos 50%) em ordem decrescente foram para os atributos P_Coh, P_VV, P_HH, Hint, Hint2, X_HH, P_HV e BMI.

Buscou-se então uma comparação entre o método de seleção de atributos do Weka e o método de seleção de atributos onde se utilizou uma banda de cada vez para construção de árvores de decisões. Neste segundo caso, os atributos que foram utilizados para construir a árvore foram os que obtiveram os melhores resultados (acima de 50% de classificações corretas) apresentados na Tabela 3. A árvore de decisões construída está ilustrada na Figura 2. Neste caso o número de instancia classificadas corretamente foi de 265 (88,333%), obtendo um Kappa de 0,86, a partir de uma árvore com 12 folhas.

Comparando as matrizes de confusão da etapa de treinamento das árvores de decisão é possível verificar quais classes que mais se confundiram. Na Figura 3, em ambos os casos, observa-se uma confusão entre as classes de floresta primária alagada e de campinarana arbórea. Já, as classes de água e de campinarana arbustiva em ambos os casos ficaram bem definidas, conseguindo bons resultados.

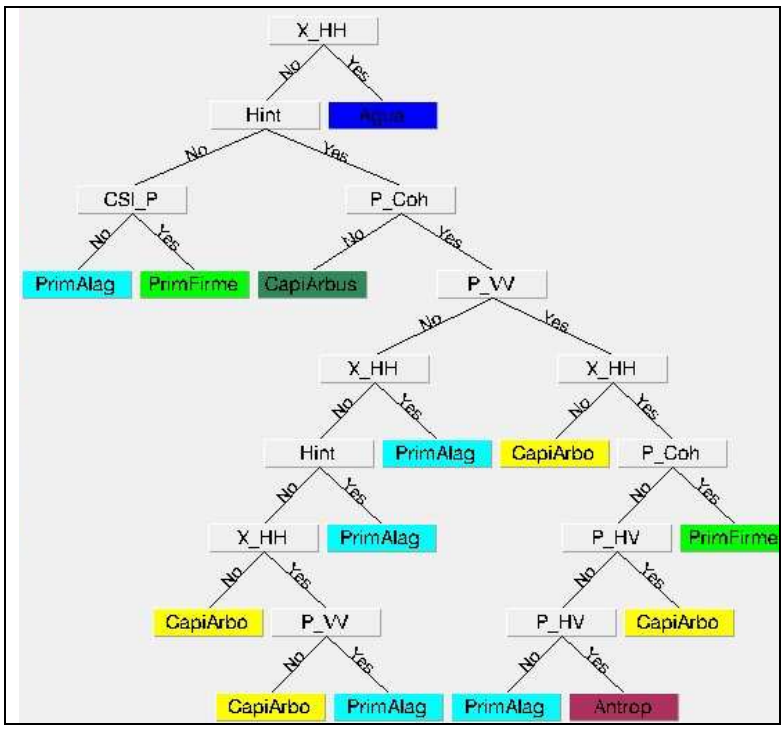


Figura 1. Árvore de decisões para atributos selecionados pelo Weka.
Figure 1. Decision tree by Weka selected features.

Tabela 3. Resultados do algoritmo J48 utilizando somente uma banda por vez.
Table 3. Results of the J48 classifier using a banda at a time.

Banda	No de Folhas	Classificações Corretas	Índice KAPPA
P_HH	10	192 (64,000%)	0,568
P_HV	9	161 (53,667%)	0,444
P_VV	11	197 (65,667%)	0,588
P_Coh	11	229 (76,333%)	0,716
X_HH	10	169 (56,333%)	0,476
X_Coh	9	135 (45,000%)	0,340
Hint	5	186 (62,000%)	0,544
Hint2	7	185 (61,667%)	0,540
LogHint	10	116 (38,667%)	0,264
VSI	1	50 (16,667%)	0,000
CSI	9	119 (39,667%)	0,276
BMI	14	160 (53,333%)	0,440
Rp	2	51 (17,000%)	0,004
Rc	6	73 (23,333%)	0,092
PotSpan	10	142 (47,333%)	0,368

Classificação da Imagem de SAR

Após a construção das árvores de decisão pelo Weka, iniciou-se o processo de classificação das bandas de SAR. Ambas as árvores das Figuras 1 e 2 foram construídas no ENVI 4.5. Visualmente os resultados iniciais obtidos apresentaram uma quantidade bastante grande de pixels isolados, resultado da classificação pixel a pixel sobre dados de SAR. Por este motivo buscou-se realizar uma pós classificação sobre estas imagens aplicando um filtro de maioria. Após diversos testes, optou-se visualmente por uma janela 11x11. Os resultados podem ser observados na Figura 4.

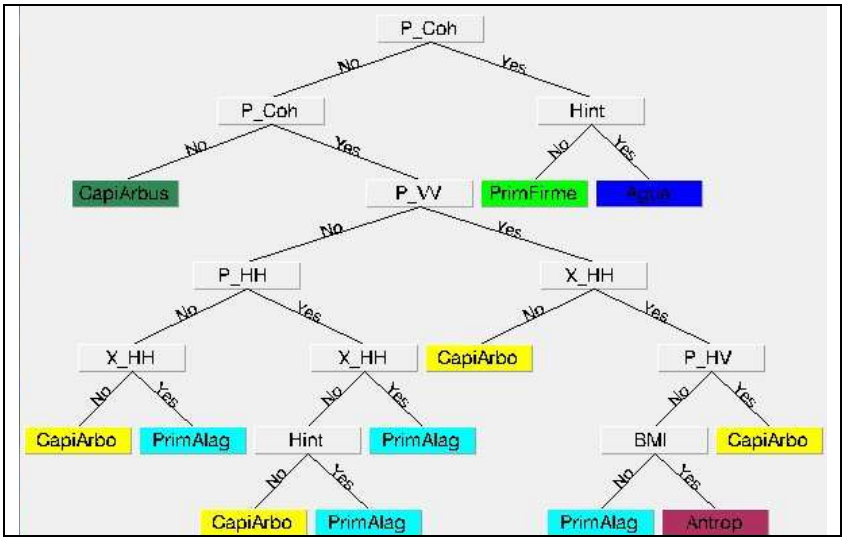


Figura 2. Árvore de decisões para atributos com melhores resultados individuais.
Figure 2. Decision tree for features with best separate results.

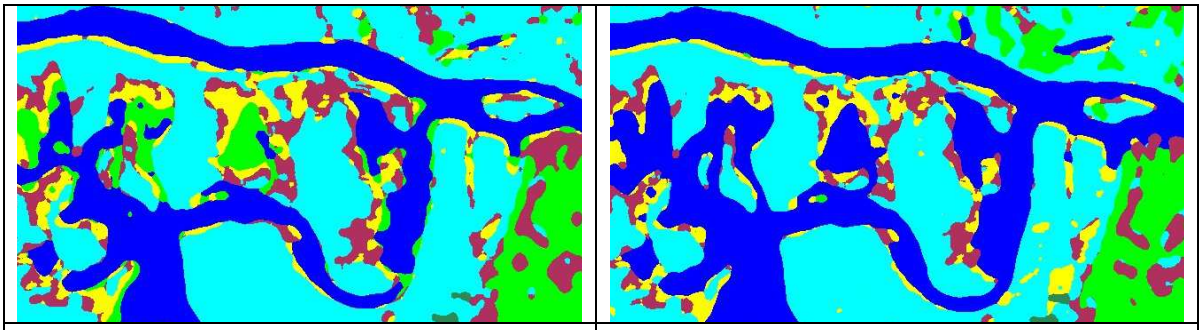
a	b	c	d	e	f	<-- classif. como
49	0	0	0	1	0	a = Água
0	38	7	0	4	1	b = Antrop
0	3	39	0	8	0	c = CampiArbo
0	0	1	48	1	0	d = CampiArbust
0	4	10	2	34	0	e = PrimAlag
0	3	0	0	1	46	f = PrimFirme

(a)

a	b	c	d	e	f	<-- classif. como
50	0	0	0	0	0	a = Água
0	46	0	0	3	1	b = Antrop
0	4	40	0	6	0	c = CampiArbo
0	0	1	48	1	0	d = CampiArbust
0	4	11	2	33	0	e = PrimAlag
0	2	0	0	0	48	f = PrimFirme

(b)

Figura 3. Matriz de confusão do treinamento relativo à árvore de decisões para atributos selecionados pelo Weka (a) e para atributos com melhores resultados individuais (b).
Figure 3. Training confusion matrix of the decision tree with features selected by Weka (a) and for best separate features (b).



(a) (b)

Figura 4. Imagens das classificações apresentadas na Figura 4 com aplicação de filtro de maioria com janela 11x11.
Figure 4. Images from the classification on Figure 4 with the usage of majority filter window 11x11.

AVALIAÇÃO E ANÁLISE DOS RESULTADOS

A avaliação das classificações foi realizada através dos polígonos de amostras de teste coletados anteriormente. Portanto, foram utilizados 30 (trinta) polígonos visando avaliar cada classe. Diferentemente do processo de aprendizagem, neste caso utilizou-se todos os pixels dentro dos polígonos, não somente o valor da média destes, totalizando cerca de 5.000 pixels para cada classe. Os resultados podem ser observados na Tabela 4.

Tabela 4. Características do Espaço de Atributos.
Table 4. Feature space characteristics.

Classificação	Acurácia	Índice Kappa
Imagem classificada utilizando a árvore de decisões obtida com atributos selecionados pelo Weka com aplicação de filtro de pós classificação de maioria com janela 11x11	89.6884%	0.8761
Imagem classificada utilizando a árvore de decisões obtida com atributos que obtiveram melhores resultados individuais com aplicação de filtro de pós classificação de maioria com janela 11x11	90.2618%	0.8830

Observa-se na Tabela 4 que a classificação realizada com os atributos que obtiveram melhores resultados individuais também obteve os melhores resultados. No entanto a diferença foi muito pequenas. Isto indica que a função de seleção de atributos do Weka auxilia positivamente nesta etapa, não sendo necessário executar os algoritmo de construção de árvore de decisão individualmente para cada atributo e analisar qual obteve resultados satisfatórios.

CONCLUSÕES

No presente trabalho foram aplicadas técnicas de auxilio a seleção de atributos e posteriormente de construção de árvores de decisões voltadas para a classificação. Ambas as técnicas obtiveram bons resultados, chegando a um percentual de acerto próximo de 90% da área classificada.

As bandas SAR se mostraram adequadas para serem usadas como atributos no processo de classificação e distinguir as classes definidas neste trabalho. Houve um destaque por parte dos atributos referentes às bandas de altura interferométrica, de coerência interferométrica P e das orto-imagens de freqüência banda P de polarização VV e HH e de freqüência X de polarização HH. No entanto, os índices incoerentes de POPE *et al*. (1994), as razões de polarização cruzadas, paralelas e a potencia total (Span) não se mostraram como atributos adequados ao processo de classificação.

As classes de água, campinarana arbustiva e de floresta primária de solo firme foram as que obtiveram os melhores resultados. Houve grande confusão entre as classes de campinarana arbórea e de floresta primária alagada. Esta confusão pode ter ocorrido em função da época em que a imagem foi obtida, podendo haver uma maior diferenciação em outras épocas. Nestes casos em que se trabalha com regiões propensas a alagamento, é importante analisar a influência da sazonalidade, visto que as respostas das classes se tornam muito variáveis.

REFERÊNCIAS

FAYYAD, U. *et al*. R. Editores. Advances in Knowledge Discovery and Data Mining. *MIT Press*. 1a Edição. 1996. ISBN 0262560976

GAMA, F. F. Estudo da interferometria e polarimetria SAR em povoamentos florestais de eucalyptus SP. 2007. 243 p. (INPE-14778-TDI/1231). Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2007. Disponível em: <http://urlib.net/sid.inpe.br/mtc-m17@80/2007/04.04.12.36>. Acesso em: 27 nov. 2009.

GE, Y. *et al*. Geo-spatial Data Analysis, Quality Assessment and Visualization. *ICCSA*, Part I, LNCS 5072, pg 258-267, 2008.

MAIMON, O.; ROKACH, L. Data Mining and Knowledge Discovery Handbook. *Springer*. 1a Edição. 2005. ISBN 0387244352.

QUINLAN. J. R. C4.5: Programs for Machine Lerning. *Morgan Kaufmann*. California, 1993.

SHEKHAR, S. *et al*. What´s Spatial About Spatial Data Mining: Three Case Studies. 2001. Disponível em <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.882>. Acesso em 11 Fev 2010.