TÉCNICAS DE MINERAÇÃO DE DADOS PARA IDENTIFICAÇÃO DE ÁREAS COM CANA-DE-AÇÚCAR EM IMAGENS LANDSAT 5

ROBSON T. NONATO[1], STANLEY R. DE M. OLIVEIRA[2]

RESUMO: Neste trabalho, verificou-se a aderência de técnicas de mineração de dados voltadas para problemas de classificação de dados na identificação automatizada de áreas cultivadas com cana-de-açúcar, em imagens do satélite Landsat 5/TM. Para essa verificação, foram estudadas imagens de áreas cultivadas com cana-de-açúcar em três fases fenológicas diferentes. Os pixels foram convertidos em valores de refletância de superfície, nas vizinhanças das cidades de Araras, São Carlos e Araraquara, no Estado de São Paulo. Foram gerados cinco modelos de árvores de decisão binária, induzidos pelo algoritmo C4.5, em que todos produziram taxas de acerto superiores a 90%. A introdução de atributos de textura trouxe ganhos significativos na acurácia do modelo de classificação e contribuiu para melhorar a distinção de áreas cultivadas com cana-de-açúcar em meio a tipos diversos de cobertura do solo, como solo exposto, área urbana, lagos e rios. Os índices de vegetação mostraram-se relevantes na distinção da fase e do estado fenológico das culturas. Os resultados reforçam o potencial forte das árvores de decisão no processo de classificação e identificação de áreas cultivadas com cana-de-açúcar, em diferentes cidades produtoras, no Estado de São Paulo.

PALAVRAS-CHAVE: mapeamento agrícola, classificação de imagens, árvore de decisão, sensoriamento remoto.

DATA MINING TECHNIQUES FOR IDENTIFICATION OF SUGARCANE CROP AREAS IN IMAGES OF LANDSAT 5

ABSTRACT: This work investigated the adherence of data mining techniques oriented to data classification problems in the identification of sugarcane crop areas in Landsat 5/TM images. To do so, pixels of images having sugarcane crop areas were studied in three different phenological phases. Such pixels were converted into surface reflectance values in neighborhood of the towns Araras, Araraquara and São Carlos in São Paulo State. It were generated five decision tree models using the algorithm C4.5 and all of them produced accuracy rates above 90%. The introduction of texture attributes brought significant gains in accuracy of the classification model and helped improve the model of distinction of areas cultivated with sugarcane in the midst of various types of land cover, such as bare soil, urban areas, lakes and rivers. The vegetation indices were relevant in distinguishing phenological phases. The results support the potential of decision trees in process of classification and identification of areas cultivated with sugarcane in different cities inside São Paulo state.

KEYWORDS: agricultural mapping, image classification, decision tree, remote sensing.

INTRODUÇÃO

O surgimento e a evolução das geotecnologias na área de sensoriamento remoto, mais especificamente a utilização de sensores situados na órbita do planeta Terra, têm permitido o desenvolvimento de metodologias inovadoras para o mapeamento da cobertura do solo. Segundo YI et al. (2007), uma das principais aplicações dos dados de sensoriamento remoto tem sido, historicamente, o mapeamento das áreas agrícolas. Este mapeamento é um procedimento essencial

[1] Mestre em Engenharia Agrícola, Universidade Estadual de Campinas, tavares.robson@gmail.com.
[2] Doutor em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Professor do Programa de Pós-Graduação da Feagri/UNICAMP, stanley.oliveira@embrapa.br.
Recebido pelo Conselho Editorial em: 4-11-2010
Aprovado pelo Conselho Editorial em: 15-5-2013

em estudos ambientais, em avaliações da biodiversidade, no monitoramento agrícola, no apoio às decisões de ações sociais, políticas e econômicas, e na estimativa da produtividade agrícola e da previsão de safras.

As estimativas da safra agrícola de um país e o conhecimento da sua distribuição no espaço geográfico são de extrema importância para o planejamento estratégico do Estado, no que concerne à formulação de políticas públicas, à logística e à segurança alimentar, além de atuar como elemento importante na formação de preços, tanto no mercado interno, como no externo (FIGUEIREDO, 2005).

Um sistema de previsão de safras eficiente é uma ferramenta importante e indispensável para qualquer país que dependa diretamente da agricultura, e a dificuldade no desenvolvimento de sistemas desse tipo deve-se à complexidade do problema de previsão de safras em si e a avanços de pesquisa necessários para que um sistema desse tipo seja concebido. Entre estes avanços, podem-se citar (CRÓSTA, 2002; ASSAD et al., 2007):

1. Melhoria da qualidade do imageamento orbital através da redução ou da eliminação total das distorções ópticas e radiométricas típicas deste processo.

2. Melhoria da qualidade da identificação automatizada das áreas de interesse contidas na imagem, visto que a estimativa correta da produção depende da estimativa correta da área ocupada pela cultura.

3. Identificação do estado ou da fase fenológica da cultura, pois, em uma imagem de satélite, podem existir áreas de uma mesma cultura em fases e estados fenológicos distintos, e estas áreas devem ser tratadas por modelos diferentes de estimativa de produção (GROHS et al., 2009).

4. Melhorias dos modelos de desenvolvimento da cultura que, em geral, incorporam variáveis climáticas, séries históricas e modelos agro-meteorológicos complexos.

Da lista acima, a identificação automática de áreas cultivadas de uma dada cultura agrícola merece destaque especial e constitui uma das etapas mais importantes no processo de previsão de safras baseada em imagens de sensoriamento remoto. Segundo ASSAD et al. (2007), a melhoria dos resultados do processo de classificação digital de regiões em imagens de satélite impacta diretamente o resultado da previsão de uma dada safra agrícola, uma vez que a produção agrícola é função da área cultivada.

Amplamente utilizadas na resolução de problemas de classificação automatizada de dados das mais diversas áreas do conhecimento, as técnicas de mineração de dados, mais precisamente as árvores de decisão, apresentam-se como alternativa promissora na resolução de problemas de identificação e classificação de regiões cultivadas com cana-de-açúcar.

O objetivo deste trabalho foi avaliar técnicas de mineração de dados voltadas para classificação de dados, dentre as quais os métodos de seleção de atributos e a técnica de árvore de decisão binária na identificação de áreas cultivadas com cana-de-açúcar, no Estado de São Paulo, em imagens com correção atmosférica do sensor TM a bordo do satélite Landsat 5. Também foi investigado o resultado da inserção de atributos de textura e de índices de vegetação com o objetivo de melhorar os resultados da identificação e a classificação de áreas cultivadas com cana-de-açúcar.

MATERIAL E MÉTODOS

Área estudada

As áreas de estudo compreendidas neste trabalho, bem como as coordenadas geográficas (Datum SAD69) estão listadas na Tabela 1. Estes municípios foram selecionados aleatoriamente da lista dos 30 municípios maiores produtores de cana-de-açúcar do Estado de São Paulo, segundo levantamento de 2007/2008 do projeto Canasat (CANASAT, 2009)

TABELA 1. Áreas sob estudo. **Areas under study**.

Cidades	Coordenadas Geográficas das Áreas de Estudo
Araras Rio Claro Leme	Lat. -22º 15' 47,87'' a -22º 30' 47,12'' Lon. -47º 34' 10,61'' a -47º 08' 15,03''
Araraquara Ibaté São Carlos	Lat. -21º 36' 47,99'' a -22º 06' 35,66'' Lon. -48º 17' 19,90'' a -47º 43' 33,94''
Jaboticabal Sertãozinho Guariba	Lat. -21º 09' 33,97'' a -21º 18' 26,06'' Lon. -48º 22' 44,96'' a -48º 09' 23,77''

Para a realização deste estudo, foram utilizadas imagens sem cobertura de nuvens na órbita 220, ponto 75, correspondentes aos dias 11-09-2008, 11-02-2009 e 24-05-2009 com seus respectivos dias julianos (DJ), em relação ao primeiro dia de seu respectivo ano: 252; 31 e 144. Estas imagens Landsat 5/TM foram adquiridas do repositório de imagens do Instituto Nacional de Pesquisas Espaciais (INPE), instituição governamental sob a responsabilidade do Ministério da Ciência e Tecnologia.

Seleção das coberturas de solo utilizadas no treinamento do classificador

Para o treinamento do classificador, foram utilizados cinco tipos de alvos diferentes: solo exposto, áreas urbanas, corpos de água (rios e lagos), florestas e áreas cultivadas com cana-de--açúcar, em três fases fenológicas diferentes.

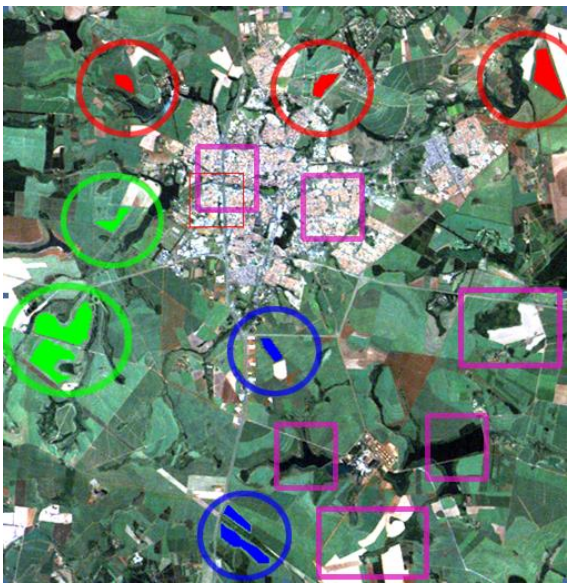

FIGURA 1. Seleção dos pixels relativos aos alvos a serem classificados: em vermelho, regiões contendo cana-de-açúcar na fase de crescimento; em verde, cana-de-açúcar na fase de perfilhamento; em azul, cana-de-açúcar na fase de maturação; por fim, na cor magenta, estão as regiões contendo outros alvos que não cana, como solo exposto, água e área urbana. **Pixels selection concerning the targets to be classified: in red, regions containing sugarcane in the growth phase; in green, sugarcane in the tillering phase; in blue, sugarcane in the maturation phase; in magenta, bare soil, urban areas, lakes and rivers**.

Na Figura 1, está ilustrada a seleção de três áreas cultivadas com cana-de-açúcar. As áreas circundadas com a cor verde representam culturas de cana-de-açúcar na fase de crescimento, correspondendo a culturas com idade em torno de sete a oito meses. As áreas circundadas com a cor vermelha correspondem a áreas com cultura de cana-de-açúcar com idade máxima entre três e quatro meses, relativas à fase de perfilhamento. Por fim, estão representadas em cor azul áreas cultivadas com cana-de-açúcar com oito ou mais meses de idade que correspondem à fase de maturação. As demais áreas destacadas com quadrados na cor violeta representam regiões na imagem contendo áreas urbanas, solo exposto, rios, lagos e outros tipos de vegetação, como florestas.

Imagens e dados auxiliares utilizados no estudo

O sensor TM capta a radiância espectral dos alvos e registra-os em pixels de imagens digitais. Cada pixel possui um valor numérico chamado nível de cinza (ND), cujos valores variam de 0 a 255 (8 bits). Na Tabela 2, verificam-se os coeficientes de calibração para conversão dos números digitais em valores de radiância propostos por CHANDER & MARKHAM (2003).

TABELA 2. Bandas espectrais do sensor TM e coeficientes de calibração. **Spectral bands of TM sensor and calibration coefficients**.

Bandas	Comprimento de Onda (µm)	Coeficientes de Calibração (Wm^{-2} sr^{1} μm^{-1})		T_{OA}
		L_{min}	L_{max}	Wm^{-2} m^{-1}
B1(azul)	0,45 - 0,52	-1,52	193	1967
B2 (verde)	0,52 - 0,60	-2,84	365	1826
B3 (vermelho)	0,63 - 0,69	-1,17	264	154
B4 (infravermelho próximo)	0,76 - 0,79	-1,51	221	1036
B5 (infravermelho médio	1,55 - 1,75	-0,37	30,2	215
B6 (infravermelho termal	10,4 - 12,5	1,2378	15,303	
B7 (infravermelho distante	2,08 - 2,35	-0,15	16,5	80,67

Fonte: CHANDER & MARKHAM (2003).

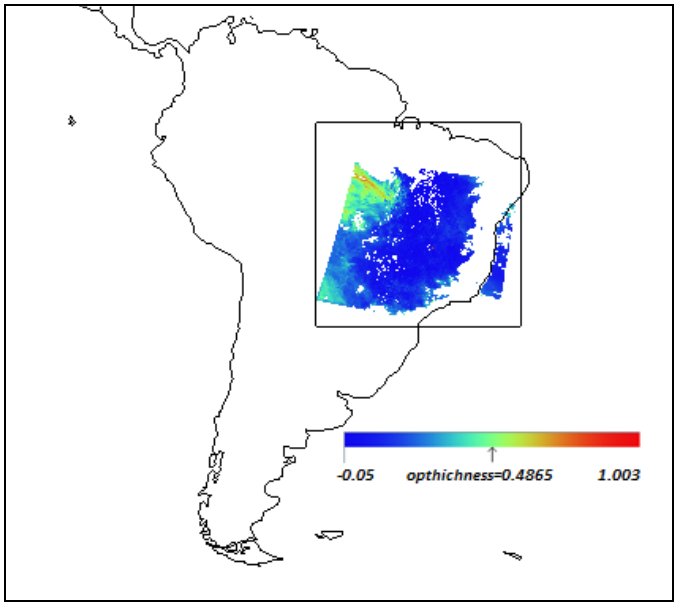

FIGURA 2. Imagem MODIS/TERRA de 23 de maio de 2009, relativa à espessura óptica dos aerossóis. Valor mínimo observado -0.05, valor máximo 1.03 e média igual a 0.4765. **Image MODIS/TERRA on May 23, 2009 related to aerosol optical thickness. Minimum value observed was -0.05, maximum value was 1.03 and the average was 0.4765**.

Segundo OLIVEIRA et al. (2009), alguns sensores orbitais vêm produzindo dados que também permitem caracterizar a atmosfera, como os fornecidos pelo sensor MODIS. Na Figura 2, está ilustrada a imagem relativa ao produto MOD04, correspondente à espessura óptica dos aerossóis no dia anterior à passagem do satélite Landsat 5.

Conversão dos números digitais em valores de radiância

Os números digitais em imagens de satélite não possuem valor físico. Para a conversão dos números digitais em valores físicos de radiância ($L_{\lambda i}$), foram utilizados os coeficientes de calibração listados na Tabela 2. O procedimento foi repetido para as seis bandas utilizadas no estudo. A equação de conversão de números digitais em valores de radiância está ilustrada a seguir (LIU, 2006):

$$L_{\lambda i} = Lmin + \frac{ND * (Lmáx - Lmín)}{255} \quad (1)$$

em que,

Lmín e *Lmáx* = radiâncias espectrais mínimas e máximas (Wm^{-2} sr^{1} nm^{1});

ND = intensidade do pixel (número inteiro entre *0* a *255*);

i corresponde às bandas (*1*,*2*,*3*,*4*,*5* e *7*) do satélite Landsat 5/TM.

Conversão da radiância em valores de refletância

Após a conversão dos NDs registrados nos pixels das imagens em valores de radiância, o próximo passo foi a obtenção dos valores de refletâncias monocromáticas de cada banda ($\rho_{\lambda i}$) que pode ser definida como a razão entre o fluxo de radiação refletida e o fluxo de radiação incidente, que é obtido através da equação (LIU, 2006):

$$\rho_{\lambda i} = \frac{\pi . L_{\lambda i}}{K_{\lambda i} . cosZ . dr} \quad (2)$$

em que,

$L_{\lambda i}$ é radiância espectral de cada banda;

$K_{\lambda i}$ é a irradiância solar espectral de cada banda no topo de atmosfera (Wm^{-2} sr^{1} μm^{-1});

Z é o ângulo zenital solar;

dr é o inverso do quadrado da distância percorrida pela radiação eletromagnética proveniente do Sol até a Terra (em UA - unidades astronômicas);

E *dr* pode ser obtido através da equação (LIU, 2006):

$$dr = 1 + 0{,}033 \cos\left(\frac{DJ . 2\pi}{365}\right) \quad (3)$$

O ângulo zenital solar foi obtido por meio da seguinte equação (LIU, 2006):

$$cosZ = \cos\left(\frac{\pi}{2} - E\right) \quad (4)$$

em que,

E é o ângulo de elevação do sol.

Este último parâmetro da equação pode ser encontrado no arquivo de metadados que vem associado a cada uma das imagens obtidas no repositório de imagens do INPE.

Correção atmosférica e refletância de superfície

Os dados de refletância obtidos através das equações anteriores correspondem a valores captados no topo da atmosfera terrestre e por isso foi necessária a realização de procedimento de correção atmosférica, utilizando o modelo de transferência radiativa denominado MODerate

Resolution Atmosferic TRANsmission (MODTRAN), proposto por BERK et al. (1998). Para a obtenção dos parâmetros relativos à caracterização da atmosfera foram utilizados os produtos MOD04, MOD05 e MOD07 da plataforma de imageamento orbital MODIS/TERRA, obtidas gratuitamente no sítio eletrônico da NASA. Para a extração dos parâmetros atmosféricos das imagens MODIS, foi adotada a metodologia proposta por OLIVEIRA et al. (2009).

Índices de vegetação

A fim de ampliar e de enriquecer o espaço de atributos-base para o treinamento do modelo de classificação, foram utilizados cinco diferentes índices de vegetação baseados em combinações lineares e não lineares das bandas espectrais do sensor TM (TANAJURA et al., 2005). Estes índices são o Normalized Difference Vegetation Index (NDVI), O Enhanced Vegetation Index (EVI), o Perpendicular Vegetation Index (PVI), o Soil Adjusted Vegetation Index (SAVI) e Ratio Vegetation Index (RVI). Os índices e suas respectivas equações estão representados na Tabela 3.

TABELA 3. Definição dos índices de vegetação utilizados no estudo. **Definition of the vegetation indices used in this study**.

Índice de Vegetação	Equação
NDVI	$(B4 - B3) / (B4 + B3)$
EVI	$G * (B4 - B3)/(k + B4 + C_1 * B3 - C_2 * B1)$
PVI	$aB4 - bB3$
SAVI	$((1 + L) * (B4_{solo} - B3_{solo}))/(B4_{solo} + B3_{solo} + L)$
RVI	B4 / B3

Na Tabela 3, B1, B3 e B4 são as bandas espectrais do sensor Thematic Mapper do satélite Landsat 5, convertidas em refletância de superfície e correspondentes às faixas espectrais do azul, vermelho e infravermelho próximo; *k* é o fator de ajuste para solo, com valor constante igual a 1; C_1 e C_2 são coeficientes de ajuste para efeito de aerossóis da atmosfera, com valores constantes iguais a 6 e 7,5, respectivamente; *G* o fator de ganho, com valor igual a 2,5; $B3_{solo}$ e $B4_{solo}$ são as médias dos valores dos pixels de solo exposto para as bandas 3 e 4 do sensor TM, respectivamente, e *L* é uma constante igual a 0,5. Os parâmetros *a* e *b* são o intercepto e o coeficiente angular da linha dos solos, respectivamente.

O NDVI pode variar de -1 a +1. Os valores negativos representam as nuvens; os valores ao redor de zero representam solo nu ou sem vegetação, e os valores maiores que zero representam a vegetação. Quanto maior o valor do NDVI, maior o vigor de crescimento da cultura (LIU, 2006).

Atributos de textura das regiões estudadas

HARALICK (1973) propõe uma metodologia para extração de texturas com base em estatística de segunda ordem, em que são definidas as características provenientes do cálculo de matrizes denominadas "matrizes de coocorrência", que consistem em uma contagem de quantas combinações diferentes de níveis de cinza ocorrem em uma imagem, em uma determinada direção. Para obtenção de tais matrizes, considera-se a variação da distância e da direção (d, θ) entre pixels vizinhos. Normalmente, são utilizados quatro direcionamentos: 0º, 45º, 90º e 135º, conforme ilustra a Figura 3.

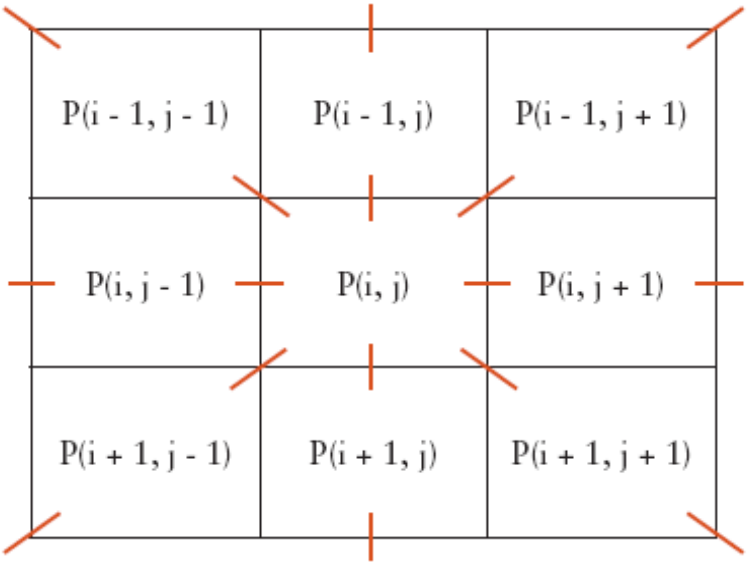

FIGURA 3. Configuração típica da vizinhança de um pixel de imagem digital. Fonte (ALVES et al., 2006). **Typical configuration of the neighborhood of a digital image pixel**.

As matrizes de coocorrência formam a base para elaboração de diversas medidas estatísticas conhecidas como descritores (HARALICK, 1973). Para cada pixel *P(i,j)* processado na imagem, há uma janela em torno dele, com distância *d=1* nas quatro direções *θ*. Utilizaram-se, neste trabalho, oito descritores, conforme ilustrado na Tabela 4.

TABELA 4. Definição dos atributos de textura utilizados neste trabalho. **Definition of the texture attributes used in this study. Source: (HARALICK, 1973)**

Nome do atributo	Descrição da finalidade do atributo de textura	Definição numérica dos atributos de textura
Média	Média da matriz de coocorrência	$MED(d, \theta) = \sum_{i-1}^{Ng} \sum_{j-1}^{Ng} i \cdot P(i, j)$
Variância	Corresponde à heterogeneidade da matriz de coocorrência em forma de desvio dos valores P da matriz.	$VAR(d, \theta) = \sum_{i-1}^{Ng} \sum_{j-1}^{Ng} (i - j)^2 \cdot P(i, j)$
Homogeneidade	Também chamada de momento da diferença, denota a homogeneidade da matriz de coocorrência.	$MDI(d, \theta) = \sum_{i-1}^{Ng} \sum_{j-1}^{Ng} \frac{1}{1 + (i - j)^2} P(i, j)]$
Contraste	Reflete a quantidade de variação local de níveis de cinza em uma imagem	$Cont(d, \theta) = \sum_{i-1}^{Ng} \sum_{j-1}^{Ng} (i - j)^2 P(i, j)$
Dissimilaridade	Mede o desvio dos valores da combinação de pares de pixels diagonais, em que apenas a contribuição do desvio é considerada	$DIS(d, \theta) = \sum_{i-1}^{Ng} \sum_{j-1}^{Ng} \lvert(i - j)\rvert \cdot P(i, j)$
Entropia	A entropia fornece o grau de dispersão de níveis de cinza de uma imagem.	$ENT(d, \theta) = -\sum_{i-1}^{Ng} \sum_{j-1}^{Ng} (i - j) \log_2[P(i, j)]$
Segundo momento	Fornece a medida de homogeneidade local dos níveis de cinza em uma imagem	$SMA(d, \theta) = \sum_{i-1}^{Ng} \sum_{j-1}^{Ng} [P(i, j)]^2$
Correlação	Mede a dependência linear dos níveis de cinza nas combinações dos pares de pixels em uma determinada direção x e y.	$COR(d, \theta) = \sum_{i-1}^{Ng} \sum_{j-1}^{Ng} \frac{ijP(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$

Na Tabela 4, *i* e *j* correspondem à posição na linha e na coluna, respectivamente; N_g é o tamanho em número de pixels da janela correspondente à região selecionada da imagem; μ_x e μ_y correspondem às médias dos valores de refletância para as linhas e colunas da janela relativa à região de processamento. Para cada uma das seis bandas utilizadas neste trabalho, foram aplicados oito filtros para a extração dos oito atributos de textura listados na Tabela 4.

Mineração de dados

Segundo FAYYAD et al. (1996), Knowledge Discovery in Database (KDD), refere-se ao processo global de descoberta de conhecimento a partir de dados, enquanto a mineração de dados é a fase principal desse processo. Dentro desse contexto, a mineração de dados deve ser entendida como a aplicação de algoritmos específicos para extrair padrões dos dados. As demais fases do processo de KDD também são importantes, pois garantem a qualidade e a utilidade do conhecimento adquirido através dos dados.

Dentre as técnicas de classificação de dados consideradas na mineração de dados, destacam-se as árvores de decisão, que são constituídas de nodos, que representam os atributos, de arcos provenientes destes nodos, que recebem os valores possíveis para estes atributos, e de nodos-folha, que representam as classes distintas de um conjunto de treinamento (HAN & KAMBER, 2006).

Para a classificação dos dados, foi utilizado o método de árvore de decisão binária. O algoritmo de indução utilizado foi o J48, uma modificação do amplamente conhecido algoritmo C4.5. Foram verificadas configurações diferentes, em que cada uma delas resultou em taxas diferentes de acerto do modelo e conjuntos de regras com cardinalidade diferente.

Avaliação do modelo de classificação

Segundo HAN & KAMBER (2006), uma das formas mais utilizadas para contornar o ajuste específico (*overfitting* em inglês) é dividir, aleatoriamente, os exemplos em dois conjuntos independentes: um de treinamento (dois terços dos dados) e o outro de teste (um terço dos dados).

Uma vez definidos o conjunto de treinamento e o conjunto de teste, a próxima fase da avaliação é a aplicação do modelo ao conjunto de testes selecionado. Como resultado, o analista obtém a conhecida matriz de erros ou matriz de confusão (Figura 4), amplamente utilizada em análise estatística de concordância (HAN & KAMBER, 2006).

		PREDITO	
		Classe A	**Classe B**
VERDADEIRO	**Classe A**	VP	FN
	Classe B	FP	VN

FIGURA 4. Matriz de confusão de dimensão 2 x 2. **2 x 2 Confusion matrix**.

Para descrever a intensidade da concordância entre dois ou mais juízes, ou entre dois métodos de classificação, utiliza-se a medida *Kappa,* que é baseada no número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre os juízes. O *Kappa* é uma medida de concordância e mede o grau de acurácia, além do que seria esperado tão somente pelo acaso. Seus valores variam de 0 a 1, representando resultados de classificação ruins e excelentes, respectivamente. O coeficiente *Kappa* pode ser definido pela seguinte equação (WITTEN et al., 2011):

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (5)$$

em que,

Pr(a) é a concordância relativa observada para uma dada classe na matriz de confusão;
Pr(e) é a probabilidade de concordância esperada para esta mesma classe.

O coeficiente *Kappa* é calculado levando-se em consideração todas as classes.

Conjunto de dados utilizados no trabalho

O conjunto de dados originais era composto por 60 atributos (59 atributos preditivos e um atributo-meta). O atributo-meta refere-se à cobertura do solo e é o alvo da classificação. Os 59 atributos preditivos eram formados por seis bandas do satélite Landsat 5, cinco índices de vegetação e 48 atributos de textura (8 descritores de textura para cada banda).

Seleção de atributos

O melhor conjunto de atributos foi selecionado, utilizando-se de um método de busca exaustiva, cujo critério final se baseia na escolha do subconjunto gerador da melhor taxa de acerto para o algoritmo de aprendizagem escolhido. Este método é conhecido como Wrapper (OLIVEIRA et al., 2007).

RESULTADOS E DISCUSSÃO

O potencial do modelo de árvore de decisão binária foi avaliado em relação à distinção de áreas cultivadas com cana-de-açúcar, em diferentes fases fenológicas e em meio a outros tipos de alvo na superfície terrestre. Este cenário reproduz, em parte, um cenário típico encontrado em uma imagem de satélite, onde pode ser de interesse identificar áreas cultivadas com cana-de-açúcar com um nível melhor de detalhes.

Após a aplicação de métodos de seleção de atributos do conjunto de dados, foram obtidos os resultados listados na Tabela 5. Nota-se que o melhor conjunto de atributos, excluindo o conjunto formado por todos os 60 atributos originais, foi conseguido com um subconjunto composto por apenas 10. O arquivo foi reduzido a 26% de seu tamanho original em kilobytes. Não houve redução do número de registros, mas sim de atributos.

TABELA 5. Resultados da classificação para diferentes conjuntos de atributos. **Classification results for different attribute sets**.

Conjuntos utilizados freqüentemente	Taxa de Acerto do modelo	Estatística KAPPA	Número de Regras
Somente Bandas (B1, B2, B3, B4, B5 e B7)	95,07%	0,93	1285
Seleção feita pelo Wrapper B5, B3_MEDIA, B3_ENTROPIA, B4_MEDIA, B4_CORRELACAO, B5_MEDIA, B5_CONTRASTE, EVI, NDVI,B3	96,68%	0,95	737
Bandas e Índices de vegetação (B1,B2,B3,B4,B5,B7, NDVI, RVI,PVI,SAVI,EVI)	94,98%	0,93	1177
Melhor conjunto entre bandas e índices de vegetação B2,B3,B5,NDVI	95,68%	0,94	1112
Todos os atributos	97,21%	0,96	654

Da análise da Tabela 5, verifica-se que o conjunto completo de atributos apresentou uma taxa de acerto de 97,21%, contra uma taxa de acerto de 95,07% do subconjunto de atributos relativo somente às bandas. Este resultado revela que a introdução de índices de vegetação e de atributos de textura trouxe um ganho de 2,14% na predição do atributo-meta. O subconjunto de atributos que produziu o melhor resultado de classificação, dentre todos os 60 atributos iniciais, apresentou taxa de acerto de 96,68% e foi composto pelas bandas B3, B4 e B5, e de suas transformações e combinações entres si. Isto indica que apenas nessas três bandas, devidamente combinadas, há informação suficiente para se alcançar um resultado de classificação muito bom.

Foram utilizados 66.714 registros relativos a pixels puros de regiões cultivadas com cana-de-açúcar em fases fenológicas diferentes e, também, de regiões contendo outros tipos de cobertura do solo. Desses registros, 44.031 foram utilizados para o treinamento e 22.683 para a avaliação do modelo. Conforme observado na Tabela 6, as maiores taxas de falsos positivos foram verificadas na classificação de solo exposto em áreas urbanas, resultado que é coerente, tendo em vista que em áreas urbanas podem existir áreas com solo exposto e cobertura do solo com refletância semelhante. Taxas maiores de falsos positivos também foram observadas na classificação das diferentes fases fenológicas da cultura de cana-de-açúcar. Este resultado também é coerente, uma vez é muito mais difícil para o classificador distinguir áreas cultivadas com cana-de-açúcar em fases fenológicas diferentes.

TABELA 6. Matriz de confusão resultante do processo de classificação utilizando método Wrapper. **Confusion matrix from the classification process by using the Wrapper method**.

	Classificado como:						
Classe	Área urbana	Lagos e rios	Florestas	Solo exposto	Perfilhamento	Crescimento	Maturação
Área urbana	10098	0	9	97	0	0	0
Lagos e rios	1	1344	0	3	0	0	0
Florestas	4	0	3434	0	0	0	0
Solo exposto	208	0	0	3506	0	0	0
Perfilhamento	0	0	0	0	1273	110	37
Crescimento	1	0	0	0	100	1239	79
Maturação	0	0	0	0	24	78	1038

A partir da Tabela 6, verifica-se boa distinção de pixels puros nas áreas cultivadas com cana-de-açúcar (Perfilhamento, Crescimento e Maturação), contra os pixels de outros tipos de cobertura do solo, como florestas, áreas urbanas e solo exposto. Outro resultado interessante, extraído da análise das métricas de qualidade do modelo (Tabela 7), é que a precisão observada para a classificação de pixels relativos a lagos e rios (corpos d'água), foi de 100%, isto é, precisão 1 (valor máximo).

TABELA 7. Métricas da qualidade do modelo extraídas da matriz de confusão. **Quality metrics of the model extracted from the confusion matrix**.

	Sensitividade (Recall)	Especificidade	Confiabilidade Positiva (Precisão)	Confiabilidade Negativa	F-measure
Área urbana	0,99	0,96	0,98	0,97	0,98
Lagos e rios	1	1	1	1	1
Florestas	0,99	0,99	0,99	0,99	0,99
Solo exposto	0,94	0,98	0,97	0,97	0,95
Perfilamento	0,89	0,99	0,91	0,99	0,9
Crescimento	0,87	0,99	0,86	0,99	0,87
Maturacao	0,91	0,99	0,89	0,99	0,9

O valor de sensitividade mais baixo foi observado para a classificação de pixels relativos à cana-de-açúcar, na fase de crescimento 0,87 (Tabela 7). A menor precisão (0,86) foi observada para a classificação de pixels relativos às regiões cultivadas com cana-de-açúcar na fase fenológica de crescimento. Estes resultados revelam que o classificador teve mais dificuldade para distinguir a fase fenológica da cultura, quando comparado com os outros alvos.

A fim de minimizar os efeitos do ajuste específico (*overfitting*), foi aplicado um procedimento de poda no modelo de árvore de decisão selecionado e, também, para deixar o modelo de aprendizagem mais compreensível através da diminuição do número de regras da árvore gerada. Na Tabela 8, são apresentados a Taxa de Acerto, a estatística Kappa e o número de regras geradas para diferentes níveis de pré-poda.

TABELA 8. Taxas de acerto, a estatística Kappa e o número de regras para diferentes níveis de pré-poda para o método Wrapper. **Accuracy rates, Kappa statistic and the number of rules for different levels of pre-pruning by using the Wrapper method**.

Número Mínimo de Pixels por Folha	Taxa de Acerto(%)	Estatística Kappa	Regras
2	96,85	0,95	737
5	96,55	0,95	465
10	96,27	0,94	299
20	95,99	0,94	168
30	95,71	0,94	127
40	95,4	0,93	108
50	95,39	0,93	88
60	95,34	0,93	76
70	95,31	0,93	57
80	95,26	0,93	51
90	95,24	0,93	48
100	95,14	0,93	45
150	94,79	0,92	31
200	94,03	0,91	28
300	93,16	0,9	25
400	92,84	0,9	23
500	91,9	0,88	19
600	91,95	0,88	18
700	91,83	0,88	17
800	91,74	0,88	11

Após a análise destes dados da Tabela 8, verifica-se que o modelo de árvore de decisão, gerado com atributos selecionados pelo método Wrapper, mantém a taxa de acerto superior a 91% para um nível de pré-poda menor ou a igual a 500, isto é, cada nó da folha da árvore poderia conter, no máximo, 500 pixels, isto é, pelo menos 500 instâncias do conjunto-teste são avaliadas em cada nó da árvore. Verifica-se também que a estatística Kappa se mantém acima de 0,88 mesmo para um número mínimo de pixels por folha maior que 500.

Outra informação relevante da análise da Tabela 8 é que ocorre diminuição drástica do número de regras para níveis de pré-poda superiores a 40 pixels por folha. Para um ponto de corte igual a 25 pixels por folha, observa-se que o número de regras cai para um patamar de 18% em

relação ao número total de regras para o modelo mais acurado e, ainda assim, a acurácia do modelo é maior que 90%.

CONCLUSÕES

As técnicas de mineração de dados (seleção de atributos e classificação de dados) permitiram o desenvolvimento de um modelo de identificação de áreas cultivadas com cana-de-açúcar, cuja taxa de acerto foi de 96,68% dos casos, utilizando o método Wrapper de seleção de atributos, com apenas 10 dos 60 atributos disponíveis.

O índice de vegetação NDVI esteve na composição de conjuntos de atributos mais adequados para modelos de distinção das áreas cultivadas com cana-de-açúcar, reafirmando seu potencial de distinção também em modelos dirigidos à descoberta de conhecimento em bancos de dados.

A menor precisão apresentada pelo classificador, Kappa igual 0,86, ocorreu na distinção de pixels relativos a áreas cultivadas com cana-de-açúcar na fase de crescimento. Para identificação de áreas cultivadas com cana, nas diferentes fases fenológicas, foi observada menor precisão do modelo, uma vez que existem semelhanças fortes entre os pixels dessas fases.

REFERÊNCIAS

ALVES, W. A. L.; ARAÚJO, S. A.; LIBRANTZ, A. F. H. Reconhecimento de padrões de texturas em imagens digitais usando uma rede neural artificial híbrida. *Exacta*, São Paulo, v. 4, n. 2, p.325-332, 2006.

ASSAD, E.D.; MARIN, R.M.; EVANGELISTA, S. R.; PILAU, F.G.; FARIA, J.R.B.; PINTO, H.S.; ZULLO JÚNIOR, J.; Sistema de previsão de safra de soja para o Brasil. *Pesquisa Agropecuária Brasileiras*. Brasília, v.42, n.5, p.615-625, 2007.

BERK, A.; BERNSTEIN, L.S.; RICHTSMEIER, S.; ACHARYA, P.K.; ATTHEW, M.W.; ANDERSON, G.P.; ALLRED, C.L.; ADLER-GOLDEN, S.; JEONG, L.S.; CHETWYND, J.H. Flaash a MODTRAN4 Atmospheric correction package for hyperspectral dataretrievals and simulations. In: ANNUAL JPL AIRBORNE EARTH SCIENCE WORKSHOP, 7., 1998. Pasadena, CA. *Proceedings...* v.7, p. 9-14.

CANASAT. *Mapeamento da cana via imagens de satélite e observação da Terra*. Disponível em: <http://www.dsr.inpe.br/mapdsr/intro.htm>. Acesso em: 13 nov. 2009.

CHANDER, G.; MARKHAM B. Revised Landsat-5 TM radiometric calibration procedures and postcalibration dynamic ranges. IEEE *Transactions on Geoscience and Remote Sensing*, New York, v.41, p.2.674–2.677, 2003.

CRÓSTA, A. P. *Processamento Digital de imagens de sensoriamento remoto*. 4. ed. Campinas: Editora da Unicamp, 2002.

FAYYAD, U.; DJORGOVSKI, G.; WEIR, N. Automating The Analysis And Cataloging of Sky Surveys. In: FAYYAD, U.; DJORGOVSKI, G.; WEIR, N. *Advances in knowledge discovery and data mining* , 3. ed. Sacramento: Association of Artificial Intelligence menlo Park, 1996. p. 471-493.

FIGUEIREDO, D.C. Projeto GeoSafras - aperfeiçoamento do sistema de previsão de safras da CONAB. *Revista de Política Agrícola*, Brasília, v.14, p.110-120, 2005.

GROHS, D.S.; BREDEMEIER, C.; MUNDSTOCK, C.M.; POLETTO, N. Modelo para estimativa do potencial produtivo em trigo e cevada por meio do sensor GreenSeeker. *Revista Engenharia Agrícola*, Jaboticabal, v.29, n.1, 2009.

HAN, J.; KAMBER, M. *Data mining*: concepts and techniques. San Francisco: Morgan Kaufmann Publishers, 2006. 770p.

HARALICK, R.M; SHUNMUGAN, K; DINSTEIN, I. Texture feature for image classification. *IEEE Transactions Systems, Man and Cybernetics*, New York, v. 3, n. 6, p.610-621, 1973.

LIU, W.T.H. *Aplicações de sensoriamento remoto*. Campo Grande: UNIDERP, 2006. 908 p.

OLIVEIRA. J. A.; DUTRA. L. V.; RENNÓ. C. D. Aplicação de Métodos de Extração e Seleção de Atributos para Classificação de Regiões. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 13., 2007, Florianópolis. *Anais...* Florianópolis: Instituto Nacional de Pesquisas Espaciais, 2007. CD-ROM.

OLIVEIRA, L.G.L.; PONZONI, F. J.; MORAES, E.C., Conversão de dados radiométricos orbitais por diferentes metodologias de caracterização atmosférica. *Revista Brasileira de Geofísica*, São Paulo, v. 27, n.1, 2009

TANAJURA, E. L. X.; ANTUNES, M. A.; UBERTI, M. A. Avaliação de Índices de Vegetação para a Discriminação de Alvos Agrícolas em Imagens de Satélites. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12., 2005, Goiânia. *Anais...* Goiânia: Instituto Nacional de Pesquisas Espaciais, 2005, CD-ROM.

YI, J.L.R.; SHIMABUKURO, Y.E; QUINTANILHA, J.A. Identificação e mapeamento de áreas de milho na região sul do Brasil utilizando imagens MODIS. *Engenharia Agrícola*, Jaboticabal, v.27, n.3, p. 753-763, 2007.

WITTEN, I.H.; FRANK, E.; HALL, M.A. *Data mining*: practical machine learning tools and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011. 629p.