Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 997-1004.

O efeito da utilização de diferentes índices de vegetação na classificação de imagens digitais pela aprendizagem por árvore de decisão

Robson Tavares Nonato [1]
Luiz Henrique Antunes Rodrigues [1]

[1] Universidade Estadual de Campinas - UNICAMP
Caixa Postal 6011 - 13083-875 - Campinas - SP, Brasil
{robson.nonato, lique}@agr.unicamp.br

Abstract. The right choice of the set of data attributes is very important in remote sensing data classification based in the decison tree learning method. Vegetation indexes are an alternative way to increase the set of data attributes because they incorporate important spectral information. The objective of this work were investigate the effect of inclusion of vegetation indexes NDVI, EVI, PVI, SAVI and RVI in the remote sensing data classification process based in decision tree learning methods.

Palavras-chave: image classification, decision trees, vegetation index, data mining, feature selection, classificação de imagens, árvores de decisão, mineração de dados, seleção de atributos, índices de vegetação.

1. Introdução

Técnicas de mineração de dados como as árvores de decisão tem sido cada vez mais utilizadas no processo de classificação de imagens, tanto pela acurácia dos resultados obtidos, como pela capacidade de propiciar ao analista a visualização imediata de padrões através de regras explícitas de decisão. A fim de ampliar o espaço de atributos e com o objetivo de aumentar a quantidade de informação disponível para a obtenção de uma classificação mais acurada é comum a inclusão de novas informações, como os índices de vegetação, além das bandas espectrais, no espaço dos atributos que serão utilizados na construção do modelo de árvores de decisão.

Pode-se pensar que quanto maior o número de atributos para representar um padrão(classe), maior o poder discriminatório do classificador. Porém, nem sempre isso é verdade. Na prática, o que acontece é uma degradação na acurácia dos resultados da classificação com o aumento da dimensionalidade dos dados mantendo-se constante o número de amostras de treinamento (Oliveira et al., 2007). Segundo Jain et al. (2000), existem duas razões para reduzir esta dimensionalidade: diminuir o custo de processamento e aumentar a acurácia da classificação.

No caso da inclusão de índices de vegetação no espaço de atributos é possível ainda estar introduzindo, como efeito colateral, redundância de informação visto que esses índices são funções explícitas de atributos já inclusos no modelo. E por serem funções dessas bandas, os índices de vegetação não somente estão correlacionados, de forma não linear, com alguns atributos já incluídos no modelo como é possível informar ao analista a função exata que traduz esse relacionamento.

Nesse contexto, esse trabalho teve como objetivo analisar a contribuição da inclusão de diferentes índices de vegetação como Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Percentage Vegetation Index (PVI), Soil Advanced Vegetation Index (SAVI) e Ratio Vegetation Index (RVI) no processo de classificação de imagens Landsat 5 TM pela aprendizagem por árvores de decisão.

2. Índices de Vegetação

A vegetação em geral é predominante na maioria das porções continentais ao longo da Terra, além lhe proporcionar um dos mais importantes ecossistemas. Por essa razão, estudos das mais diversas modalidades são voltados direta ou indiretamente para as áreas com vegetação (Tanajura et al, 2005).

O desenvolvimento de relações funcionais entre as características da vegetação e dados coletados remotamente tem sido meta de muitos estudos, principalmente aplicados aos setores agrícola e florestal. Para minimizar a variabilidade causada por fatores externos como solo, atmosfera e geometria de aquisição dos dados, a reflectância espectral tem sido transformada e combinada em vários índices de vegetação. Esta abordagem para a determinação de parâmetros da vegetação é referida na literatura como abordagem empírica (Hall et al, 1995). Os mais comumente empregados utilizam a informação contida nas reflectâncias dos dosséis referente às regiões do vermelho e infravermelho próximo, as quais são resultado de combinações de duas ou mais bandas espectrais através da soma, da diferença, da razão entre as bandas ou qualquer combinação (Wiegand et al, 1991).

Conforme Chen et al (1986) e Vygodskaia et al (1989), o emprego dos índices de vegetação, para caracterizar e quantificar determinado parâmetro biofísico de culturas agrícolas tem-se duas grandes vantagens: a) Permite reduzir a dimensão das informações multiespectrais através de um simples número além de minimizar o impacto das condições de iluminação e visada; b) fornece um número altamente correlacionado aos parâmetros agronômicos.

No entanto, os vários índices de vegetação podem ser diferentemente afetados pelas características de iluminação e visada, pela arquitetura do dossel e pelo substrato abaixo do dossel, justificando assim um estudo para avaliar o tipo de índice de vegetação mais adequado para cada aplicação (Tanajura et al, 2005).

Uma avaliação destes índices permitirá uma melhor escolha entre os diversos índices disponíveis na literatura. Os índices em uso na literatura são basicamente de dois tipos (Baret e Guyot, 1991): os índices baseados inclinação como o NDVI e o SAVI e aqueles baseados em distância, como o PVI.

3. Mineração de Dados e Árvores de Decisão

Com a melhoria dos recursos computacionais nos últimos anos e o surgimento de uma vasta quantidade de dados de diversos sensores orbitais, tem sido cada vez mais necessário a utilização de técnicas automatizadas de aquisição de conhecimento e uma das ferramentas de mineração comumente utilizadas na tarefa de classificar regiões a partir de conhecimento contido em imagens são as árvores de decisão.

As árvores de decisão são constituídas de nodos que representam os atributos; de arcos, provenientes destes nodos e que recebem os valores possíveis para estes atributos; e de nodos folha, que representam as diferentes classes de um conjunto de treinamento (Ingargiola, 1996). Classificação, neste caso, é a construção de uma estrutura de árvore, que pode ser usada para classificar corretamente todos os objetos do conjunto de dados da entrada (Brazdil, 1999).

Após a construção de uma árvore de decisão é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados no treinamento. Esta estratégia permite estimar como a árvore generaliza os dados e se adapta a novas situações, podendo, também, se estimar a proporção de erros e acertos ocorridos na construção da árvore (Brazdil, 1999).

Como a árvore de decisão é um modelo não paramétrico, esse se ajusta mais precisamente aos dados que compõe a amostra, assim é necessário manipular o número de nodos e a altura da árvore obtida a fim de se conseguir um modelo capaz de classificar adequadamente uma

imagem, de maneira que a construção de um modelo de árvore de decisão é um processo constante de análises e refinamento.

4. Materiais e Métodos

4.1 Materiais

O seguintes materiais foram utilizados nesse trabalho:

- Imagens Landsat 5 TM(*Thematic Mapper*), órbita 220, ponto 75 feitas no dia 5 de março de 2007.
- Software ENVI versão 4.2 para processamento e extração dos atributos das imagens
- Software Waikato Environment for Knowledge Analysis (WEKA) versão 3.4.13
- Arquivo vetorial referente a base municipal IBGE(1997) em formato *.shp referente a região de Araras, São Paulo

4.2 Área de estudo

A região de estudo encontra-se localizada a leste do estado de São Paulo, contida no município de Araras com uma superfície total de 348,86 Km^2, compreendida entre as latitudes 22^o 11'a 22^o 28' S e longitudes 47^o 28' a 47^o 10' W.

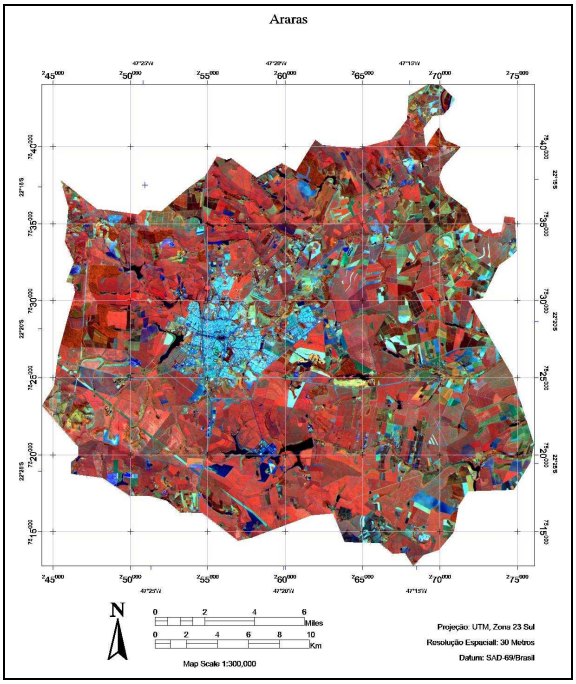

Figura 1 – Mapa contendo a região de estudo.

4.3 Processamento de imagens e preparação dos dados

Foram carregadas no sofware ENVI imagens Landsat relativas as áreas de estudo nas Bandas TM1(Azul), Banda TM2(Verde), Banda TM3(Vermelho), Banda TM4(NIR), Banda TM5(MIR) e Banda TM7(FIR) relativas a região de estudo. A partir dessas imagens foram construídas imagens índices de vegetação que são constituídas por combinações dessas bandas no software WEKA.

Os índices de vegetação utilizados no estudo foram:

- NDVI – Normalized Difference Vegetation Index
- EVI – Enhanced Vegetation Index
- PVI – Proportion Vegetation Index
- SAVI – Soil Adjusted Vegetation Index
- RVI – Ratio Vegetation Index

Tabela 1. Índices de vegetação, fórmulas e referências.

Índice de Vegetação	**Fórmula**	**Referência**
NDVI	$(TM4 - TM3) / (TM4 + TM3)$	(Rouse et al,1973)
EVI	$G * (TM4 - TM3)/(k + TM4 + C_1 * TM3 - C_2 * TM1)$	(Justice *et al*., 1998).
PVI	$((TM3 - TM3_{solo})^2 + (TM4 - TM4_{solo})^2)^{1/2}$	(Richardson e Wiegand, 1977)
SAVI	$((1 + L) * (TM4 - TM3))/(TM4 + TM3 + L)$	(Huete, 1988)
RVI	$TM4 / TM3$	(Jordan,1969)

Em que: TM1, TM3, TM4 são as bandas espctrais do sensor Termal Mapper do satélite Landsat 5; 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; $TM3_{solo}$ e $TM4_{solo}$ são as médias dos valores dos píxels de solo exposto para as bandas 3 e 4 respectivamente e L é uma constante igual a 0,5.

Uma região menor dentro da área de estudo, composta por diferentes tipos de alvos a fim de propor um conjunto de teste bem diversificado, foi escolhida. A área escolhida continha os seguintes alvos:

- Cultura de cana-de açúcar
- Solo exposto
- Área urbana
- Floresta Nativa

Através da ferramenta *ROI Tool* do software ENVI foram marcadas regiões relativas às áreas contendo o atributo alvo, nesse caso as regiões cultivadas com cana-de-açucar, dentro da imagem em estudo. Foram exportadas para arquivos no formato ASCII, contendo as coordenadas geográficas de cada pixel, as imagens em todas as bandas e também as imagens referentes aos índices de vegetação. Os índices de vegetação não obtidos diretamente através

do software ENVI, foram obtidos através de manipulações algébricas entre colunas no software WEKA.

Os arquivos ASCII foram processados resultando num arquivo de dados do WEKA (*Arfff ile).* Nesse arquivo foram inseridos todos os atributos referentes às bandas e imagens índice de vegetação. Cada observação na tabela corresponde ao Diginal Number(DN), variável que assume valores de 0 a 255, representando a refletância de um alvo e registrada em um pixel da imagem feita pelo sensor. Este registro é feito para cada uma de suas bandas espectrais assim como também são feitos para os valores de DNs referentes aos índices de vegetação.

O arquivo foi preparado para utilização pelo software WEKA onde foi feita uma busca por valores faltantes e valores duplicados. Uma vez concluída a etapa de preparação dos dados o arquivo foi carregado no software WEKA para análise.

4.4 Seleção de atributos

A fim de selecionar um conjunto menor de atributos para a classificação das imagens foi utilizada a análise de componentes principais para identificação e seleção de atributos explicando a maior porção de variabilidade no conjunto de dados.

4.5 Análise e refinamento dos modelos de árvores de decisão gerados

O algorítimo utilizado na modelagem foi o J4.8, implementado em Java no Weka que é uma pequena variação do algorítimo original cujo nome é C4.5. Foram feitas análises das regras de classificação obtidas da modelagem e foram feitas buscas por regras com maior precisão. Também foram construídos modelos para diferentes conjuntos de atributos a fim de identificar conjuntos pequenos porém capazes de gerar modelos com altas taxas de acerto e acurácia.

4.6 Validação do modelo

A estratégia utilizada para a validação do modelo foi o particionamento do conjunto de dados em duas amostras. Uma amostra de aprendizado contendo 66% das observações e uma amostra de validação contendo 34% das observações no conjunto de dados. Para medir a taxa de acerto foi utilizada a estatística Kappa que é extraída da matriz de confusão gerada na etapa de validação do modelo.

5. Resultados e Discussões

Na tabela 2 é possível visualizar os resultados da classificação feita pelo modelo de árvore de decisão com diferentes índices de vegetação. Ali também é possível visualizar os diferentes conjuntos de atributos investigados. Na tabela estão listados somente os conjuntos de bandas e índices de vegetação que propiciaram taxas de acerto acima de 90%.

A estatística ou coeficiente de determinação Kappa é uma importante medida da qualidade do modelo, incorpora em seu cálculo tanto as taxas de acerto como as taxas de erro, e pode assumir os valores expressos na tabela 4. Essas taxas de acerto e de erro são extraídas da matriz de confusão geradas na etapa de validação do modelo. A tabela 3 corresponde a matriz de confusão para classificação realizada com todas as bandas e índices de vegetação.

Tabela 2: Matriz de confusão para o conjunto composto por todas as bandas e índices de vegetação

Classes	**Cana-de-açúcar**	**Não cana**	**Total**
Cana-de-açúcar	146	26	172
Não cana	41	737	778
Total	187	763	950
% de acerto	78,0	96,5	92,9

Tabela 3. Taxas de acerto para diferentes conjuntos de atributos e qualidade do modelo através da estatística Kappa.

Conjunto de atributos	**Taxa de Acerto**	**Estatística Kappa**
TM1, TM2, TM3, TM4, TM5, TM7, NDVI, EVI ,PVI, SAVI, RVI	92,94%	0,77
TM1 ,TM2, TM3, TM4	91,68%	0,73
TM1, TM2, NDVI	90,73%	0,71
TM1, TM2, EVI	91,15%	0,72
TM1, TM2, PVI	91.68%	0,74
TM1, TM2, SAVI	90,94%	0,71
TM1, TM2, RVI	90,73%	0,71
TM2 ,TM3, NDVI	90,84%	0,70
B1, B7, NDVI Conjunto obtido após análise de componente principais	92,84%	0,75
B1, B7, SAVI	93,26	0,78
B1, B7, EVI	93,47%	0,78
B1, B7, PVI	93,47%	0,78
B1, B7, RVI	92,84%	0,75

Após a análise por componentes principais foi verificado que as bandas: TM1, TM7 e NDVI, continham boa parte da variabilidade dos dados, devendo ser incluídas em todos os modelos de subseqüentes.

Tabela 4. Possíveis valores da estatística Kappa

Estatística Kappa	Qualidade
0,00	Péssima
0,21 – 0,40	Ruim
0,21 – 0,40	Razoável
0,41 – 0,60	Boa
0,61 – 0,80	Muito Boa
0,81 – 1,00	Excelente

6. Conclusões

Os resultados mostraram que o modelo conseguiu bons resultados quando alimentado por diferentes conjuntos de atributos. Assim com base nos resultados da investigação, concluímos que:

- a inclusão de mais de um índice de vegetação no espaço de atributos não traz melhorias significativas ao modelo e ao resultado final da classificação.
- que para classificação de culturas como cana-de-açúcar índices de vegetação como EVI e PVI propiciaram melhores resultados na classificação por aprendizado por árvore de decisão.
- devido à presença de diversos tipos de alvos na imagem investigada a inclusão da Banda do azul (TM1) trouxe ganho de informação para o modelo e melhora significativa nos resultados da classificação.
- que os índices de vegetação conseguem incorporar grande parte da informação contida nas bandas do vermelho TM3 e NIR TM4, podendo ser incorporados às bandas TM1 E TM7 trazendo bons resultados para classificação do tipo de alvo estudado nesse trabalho.

Por fim, concluímos que é possível obter acurácia no resultado final da classificação e melhora no tempo de processamento dos arquivos, utilizando-se um conjunto reduzido de atributos.

7. Referências

Brazdil, P. Construção de Modelos de Decisão a partir de Dados. Disponível por WWW em: http://www.ncc.up.pt/~pbrazdil/Ensino/ML/DecTrees.html, 1999.

Chen, C.S.; Tardin, A,T.; Batista, G.T. **Índices de Vegetação e suas aplicações na agricultura.** São José dos Campos, SP: Instituto Nacional de Pesquisas Espaciais, INPE, 1986. 24p. (INPE-3912-MD/030).

Clevers, J.G.P.W. The derivation of a simplified reflectance model for the estimation of leaf area index, **Remote Sensing of Environnment,** v.1, n.25, p 53-70, 1988.

Hall, F. G., Townshend, J. R., Engman, E. T., Status of remote sensing algorithms for estimation of land surface state parameters, **Remote Sensing of Environnment,** 51:138-156, 1995.

Huete, A.R. A soil adjusted vegetation index (SAVI), **Remote Sensing of Environnment,** v.2, n.25: p.295-309, 1988.

Ingargiola, Giorgio. Building Classification Models: ID3 and C4.5. Disponível por WWW em: http://www.cis.temple.edu/~ingargio/cis587/ readings/id3-c45.html, 1996.

Jain. A. K., Robert P.W. Moa D., Moa. J. Statistical Pattern Recognition: A Review. **IEEE Trans. On Pattern Analysis And Machine Intelligence**, Vol. 22, Nº.1, 2000.

Oliveira. J. A., Dutra. L. V., Rennó. C. D. Aplicação de Métodos de Extração e Seleção de Atributos para Classificação de Regiões. **In: XIII Simpósio Brasileiro de Sensoriamento Remoto**. [CD-ROM]. Instituto Nacional de Pesquisas Espaciais. Florianópolis, Brasil. Abril, 2007.

Richardson, A.J. and Wiegand, C.L. Distinguishing vegetation from soil background information. **Photogrammetric Engineering and Remote Sensing,** v.1 n.43, p.1541-1552, 1977.

Rouse, J.W., R.H. HAAS, J.A. SCHELL, D.W. DEERING, J.C. HARLAN. **Monitoring the vernal advancement of retrogradation (greenwave effect) of natural vegetation**. NASA/GSFC, Type III, Final Report, Greenbelt, MD, 1974, 371 p.

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: XII Simpósio Brasileiro de Sensoriamento Remoto**. [CD-ROM]. Instituto Nacional de Pesquisas Espaciais. Goiânia, Brasil. Abril, 2005.

Wiegand, C.L.; Gansman, HW.; Cuellar, J.A.; Gergberman, A.H.; Rinchardson, A.J. Vegetation density as deduced from ERTS-1MSS response. In: **Proceedings ERTS SIMPOSIUM**, 3,1974, Washington. Proceedings. Washington,DC:[sn], 1974. v1, p93-116.