MINERAÇÃO DE DADOS PARA CLASSIFICAÇÃO DAS FASES FENOLÓGICAS DA CULTURA DA CANA-DE-AÇÚCAR UTILIZANDO DADOS DO SENSOR MODIS E DE PRECIPITAÇÃO

JOÃO FRANCISCO GONÇALVES ANTUNES [1]
STANLEY ROBSON DE MEDEIROS OLIVEIRA [2]
LUIZ HENRIQUE ANTUNES RODRIGUES [3]

RESUMO: Os dados do sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*) fornecem coberturas de áreas com grande extensão e alta periodicidade, características fundamentais que possibilitam o monitoramento de culturas agrícolas estratégicas para o Brasil, como a da cana-de-açúcar. A mineração de dados é uma abordagem promissora para melhorar a análise de dados de sensoriamento remoto. O objetivo deste trabalho foi aplicar técnicas de mineração de dados para classificação das fases fenológicas da cana-de-açúcar no Estado de São Paulo, utilizando dados MODIS e, também, de precipitação que auxiliam na caracterização do ciclo de desenvolvimento da cultura. As abordagens de seleção de atributos mostraram que todos os atributos do conjunto de dados foram considerados relevantes para a classificação. O balanceamento de classes pelo método de amostragem foi fundamental para melhorar a acurácia do modelo de classificação gerado pelo algoritmo J48. A descoberta do conhecimento pode ser feita através de regras de decisão relevantes para especialistas, revelando a aderência de técnicas de mineração de dados em problemas de classificação de imagens de satélite.

PALAVRAS-CHAVE: sensoriamento remoto, seleção de atributos, balanceamento de classes, árvore de decisão, regras de decisão, descoberta do conhecimento.

DATA MINING FOR CLASSIFICATION OF PHENOLOGICAL STAGES OF SUGARCANE CROP USING MODIS AND RAINFALL DATA

ABSTRACT: MODIS (Moderate Resolution Imaging Spectroradiometer) data provide coverage of large areas and high periodicity. These characteristics are fundamental for monitoring strategic agricultural crops in Brazil, such as sugarcane. Data mining is a promising approach to improve remote sensing data analysis. The objective of this work was to apply data mining techniques to classify phenological stages of sugarcane crop in São Paulo, Brazil, by using MODIS as well as rainfall data that assist in the characterization of the crop development cycle. The feature selection approaches showed that all attributes of the dataset were considered relevant for the classification problem. Balancing classes by sampling method was essential to improve the accuracy of the classification model generated by the algorithm J48. The knowledge discovery was performed through relevant decision rules from the specialists´ point of view. The results revealed the adherence of data mining techniques to satellite image classification problems.

KEYWORDS: remote sensing, feature selection, balancing classes, decision tree, decision rules, knowledge discovery.

---

[1] Matemática Aplicada, Embrapa Informática Agropecuária, Doutorando Feagri/Unicamp, E-mail: joaof@cnptia.embrapa.br
[2] Ciência da Computação, Embrapa Informática Agropecuária, E-mail: stanley@cnptia.embrapa.br
[3] Engenharia Agrícola, Feagri/Unicamp, E-mail: lique@feagri.unicamp.br

1. INTRODUÇÃO

O setor agrícola brasileiro está sendo marcado por um novo ciclo no plantio da cana-de-açúcar. O Brasil é hoje o maior produtor de cana-de-açúcar e exportador de açúcar do mundo, gerando mais de dois bilhões de dólares por ano na balança comercial. O Estado de São Paulo é o maior produtor nacional, com grandes áreas de plantio e várias usinas instaladas.

A cana-de-açúcar é uma poácea semiperene, pois permite cerca de cinco cortes anuais para posterior reforma do canavial. Possui um rápido crescimento, reprodução abundante e o aproveitamento econômico de grande parte da planta. O ciclo fenológico é composto pelas fases de brotação, perfilhamento, crescimento e maturação que pode ser afetado pelo regime de chuvas ao longo do desenvolvimento da planta. No Estado de São Paulo, do plantio até o primeiro corte recebe o nome de cana-planta, cujo ciclo tem duração de 18 meses. Após o primeiro corte, a rebrota passa a ter um ciclo de 12 meses, sendo denominada cana-soca. O plantio da cana-de-açúcar é feito de janeiro a março e a colheita de abril a novembro.

Os satélites ambientais provêm uma visão sinóptica e sistemática de grandes áreas, gerando imagens que possibilitam a análise do ciclo fenológico de culturas agrícolas. O sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*), a bordo das plataformas orbitais EOS (*Earth Observing System*), liderado pela NASA (*National Aeronautics and Space Administration*), tem gerado dados processados para estudos da vegetação. O satélite Terra foi lançado em dezembro de 1999 e tem passagem pelo Equador às 10h30 (horário local), em órbita descendente. Os dados MODIS de moderada resolução espacial, elevada repetitividade temporal, boa qualidade radiométrica, alta precisão geométrica, com correção atmosférica e de distribuição gratuita, possuem um grande potencial de aplicação no monitoramento da cultura da cana-de-açúcar, a partir da classificação de uma série temporal de imagens.

A principal etapa do processo de descoberta do conhecimento é a mineração de dados que consiste em extrair padrões relevantes que estão implícitos nos dados, podendo ser utilizada para aumentar o potencial de aplicações em dados de sensoriamento remoto. A técnica de árvore de decisão tem sido cada vez mais utilizada na classificação de imagens, tanto pelos melhores resultados obtidos, como pela visualização de padrões através de regras de decisão.

Nesse contexto, o objetivo deste trabalho foi aplicar técnicas de mineração de dados para classificação das fases fenológicas da cultura da cana-de-açúcar no Estado de São Paulo, utilizando dados MODIS e de precipitação.

2. MATERIAL E MÉTODOS

As áreas de cultivo da cana-de-açúcar foram obtidas a partir de um mapa temático da região nordeste do Estado de São Paulo, que é representativa das fases fenológicas da cultura durante a safra. As fronteiras entre as fases fenológicas da cana-de-açúcar foram definidas com base em Fernandes (2009): fase 1 - perfilhamento, entre primeiro decêndio de agosto e terceiro de novembro; fase 2 - crescimento rápido, entre primeiro decêndio de dezembro e terceiro de janeiro; fase 3 - crescimento lento, entre primeiro decêndio de fevereiro e terceiro de março; fase 4 - maturação, entre primeiro decêndio de abril e terceiro de junho. A fase de brotação não foi definida no trabalho porque é um estágio prematuro com predomínio de solo exposto.

Para este estudo foi utilizado o produto Índice de Vegetação MOD13Q1 do MODIS/Terra na resolução espacial de 250 m, obtidos gratuitamente do LP-DAAC (*Land Processes Distributed Active Archive Center*), em forma de quadrantes, no formato HDF (*Hierarchical Data Format*) e na projeção cartográfica sinusoidal. A manipulação dos dados foi realizada automaticamente por meio de rotinas em linguagem IDL (*Interactive Data Language*) do software ENVI (*The Environment for Visualizing Images*), com a execução de programas do pacote computacional gratuito MRT (MODIS *Reprojection Tools*).

O produto MOD13Q1 consiste de composições dos pixels de alta qualidade radiométrica, melhor geometria de observação, mínima presença de nuvens e aerossóis, selecionados das imagens diárias durante o período de 16 dias (LATORRE et al., 2007). Os dados extraídos do produto MOD13Q1 utilizados neste trabalho foram os de refletância das bandas espectrais listadas na **Tabela 1** e dos dois índices de vegetação, descritos a seguir.

Tabela 1: Características das bandas espectrais do sensor MODIS.

Banda	Região do Espectro	Faixa Espectral (nm)
1	Azul - AZU	459 a 479
2	Vermelho - VER	620 a 670
3	Infravermelho próximo - IVP	841 a 876
7	Infravermelho médio - IVM	2105 a 2155

O NDVI (*Normalized Difference Vegetation Index*), calculado com dados das bandas VER e IVP, tem como característica diminuir a interferência do solo na resposta espectral da vegetação e, assim pode ser utilizado na avaliação das mudanças do vigor vegetativo pela correlação com a biomassa (ROUSE et al., 1973). O EVI (*Enhanced Vegetation Index*), calculado com dados das bandas AZU, VER e IVP, melhora a sensibilidade em regiões de maior densidade de biomassa, além de propiciar o monitoramento da vegetação através da redução dos efeitos do dossel e de influências atmosféricas (HUETE et al., 2002). Para auxiliar na caracterização das fases fenológicas da cana-de-açúcar, foram utilizados os dados de precipitação acumulada de 16 dias calculada por interpolação de 195 estações pluviométricas no Estado de São Paulo, obtidos do Agritempo (Sistema de Monitoramento Agrometeorológico), da Embrapa Informática Agropecuária e do Cepagri/Unicamp.

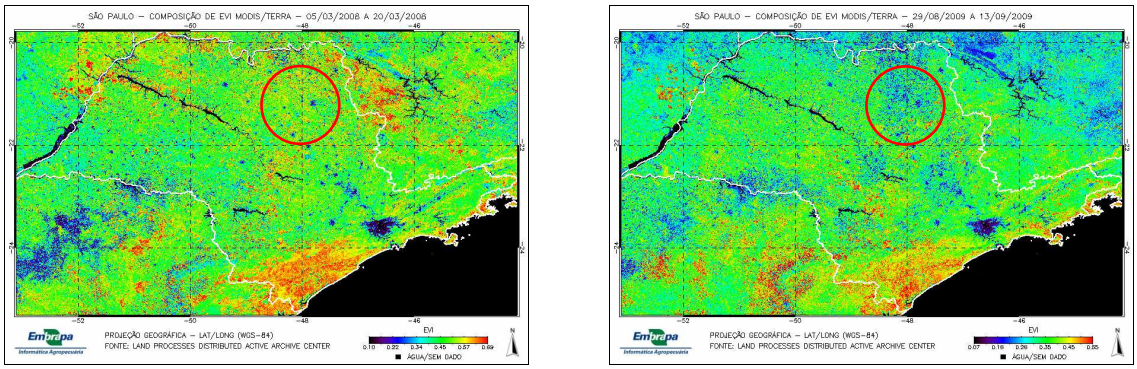
A seleção de atributos é realizada para escolher os atributos que melhor distinguem as classes com o propósito de aumentar a acurácia da classificação e reduzir o tempo de processamento. Os métodos fundamentam-se em técnicas estatísticas e matemáticas, selecionando os atributos que contém informações relevantes para a separabilidade das classes. Neste trabalho foram aplicados os métodos de seleção Qui-quadrado (χ^2) e InfoGain (*Information Gain*). A técnica de classificação utilizada foi a árvore de decisão com o algoritmo J48 (C4.5). Mais detalhes podem ser encontrados em Han e Kamber (2006).

A etapa de preparação de dados onde são realizadas as tarefas de seleção, transformação e limpeza dos dados e, também, a etapa de modelagem em que a classificação é executada para busca de padrões nos dados, foram realizadas com o software de domínio público Weka (*Waikato Environment for Knowledge Analysis*), da Universidade de Waikato, Nova Zelândia, que consiste de uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados, implementados em Java (WITTEN; FRANK, 2005).

3. RESULTADOS E DISCUSSÃO

O processamento das imagens MOD13Q1 consistiu no mosaico dos quadrantes H13V10 e H13V11, a transformação para a projeção cartográfica geográfica, o recorte para os limites do Estado de São Paulo, a extração dos dados MODIS e a gravação no formato GeoTIFF. Os dados de refletância AZU, VER, IVP, IVM e os dois índices de vegetação NDVI e EVI foram obtidos em composições de 16 dias, de janeiro a dezembro de 2005 a 2009, sendo 23 imagens para cada ano, totalizando 115 imagens de toda a série temporal. Em cada composição foram coletados os dados MODIS de 565 pixels nas áreas de cultivo da cana-de-açúcar.

A **Figura 1** apresenta duas composições de EVI, de 05 a 20 de março de 2008 e de 29 de agosto a 13 de setembro de 2009, onde, numa visão geral, colorações mais alaranjadas representam maior biomassa, colorações esverdeadas baixa biomassa, colorações azuladas solo exposto e preto corpos d'água.



março/2008 setembro/2009

Figura 1: Composições de 16 dias de EVI - safra 2008/2009 - São Paulo.

As classificações foram realizadas num conjunto de dados com os sete atributos preditores AZU, VER, IVP, IVM, NDVI, EVI e Precipitação, com o atributo meta "Fase" definindo as classes de interesse, sendo: 11.300 registros - "Crescimento rápido", 11.300 registros - "Crescimento lento", 19.775 registro - "Maturação" e 22.600 registros - "Perfilhamento", totalizando 64.975 instâncias. Após a transformação dos dados, foi aplicada a seleção de atributos para identificar os atributos irrelevantes, em que ambos os métodos não fizeram descarte de atributos. O ordenamento dos atributos em termos da sua contribuição preditiva decrescente para geração do modelo foi: Precipitação > IVP > EVI > AZU > NDVI > IVM > VER. A classificação das fases fenológicas da cana-de-açúcar foi executada com o algoritmo J48 no Weka, em validação cruzada de 10 partes, obtendo-se uma baixa acurácia de 56,99%.
Porém, observa-se um desbalanceamento entre as classes, o que pode comprometer a acurácia do classificador devido a um possível enviesamento do modelo. Para contornar esse problema, foi utilizado o método estatístico de amostragem com reposição para balancear o número de registros, com a aplicação no Weka do filtro supervisionado *Resample*. Após isso, o conjunto de dados ficou: 16.295 registros - "Crescimento rápido", 16.148 registros - "Crescimento lento", 16.197 registro - "Maturação" e 16.335 registros - "Perfilhamento", totalizando 64.975 instâncias. Com isso, a classificação foi executada novamente com o algoritmo J48 no Weka, em validação cruzada de 10 partes, obtendo-se uma boa acurácia de 74,63%, ou seja, um aumento significativo de 30% após o balanceamento das classes.
A descoberta do conhecimento pelo modelo de classificação gerado pelo algoritmo J48 pode ser feita pela visualização de padrões relevantes através de regras de decisão. A **Figura 2** mostra a acurácia das árvores de decisão geradas para diferentes níveis de pré-poda, de modo a generalizar a árvore para evitar o sobreajuste do classificador e reduzir o número de regras.

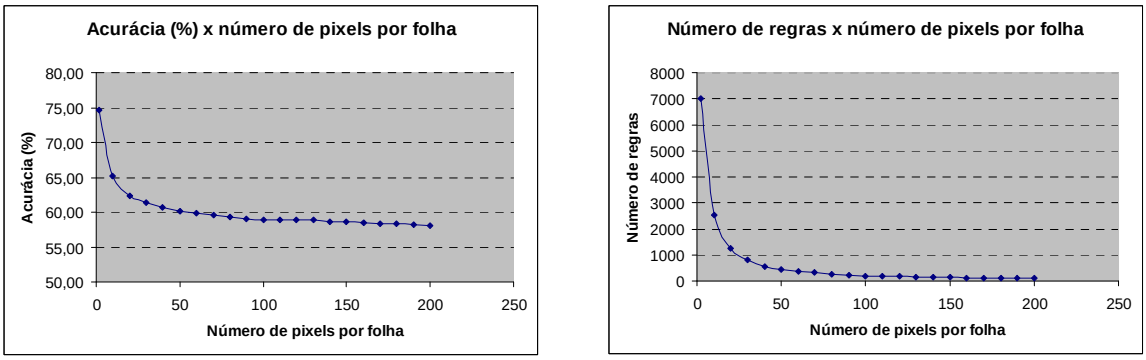


Figura 2: Acurácia e número de regras para diferentes níveis de pré-poda.

Pode-se verificar que a árvore de decisão mantém a acurácia superior a 60% para um nível de pré-poda menor ou igual a 50 pixels por folha, diminuindo o número de regras para 6% do

modelo com mais acurácia. Pode ser vantajoso escolher modelos que geram um número menor de regras, mesmo não tendo uma boa acurácia. A **Figura 3** ilustra uma árvore de decisão gerada com o nível de pré-poda de 2.400 pixels por folha, com acurácia de 54,86%, onde foram destacadas quatro regras compactas que são representativas para especialistas.

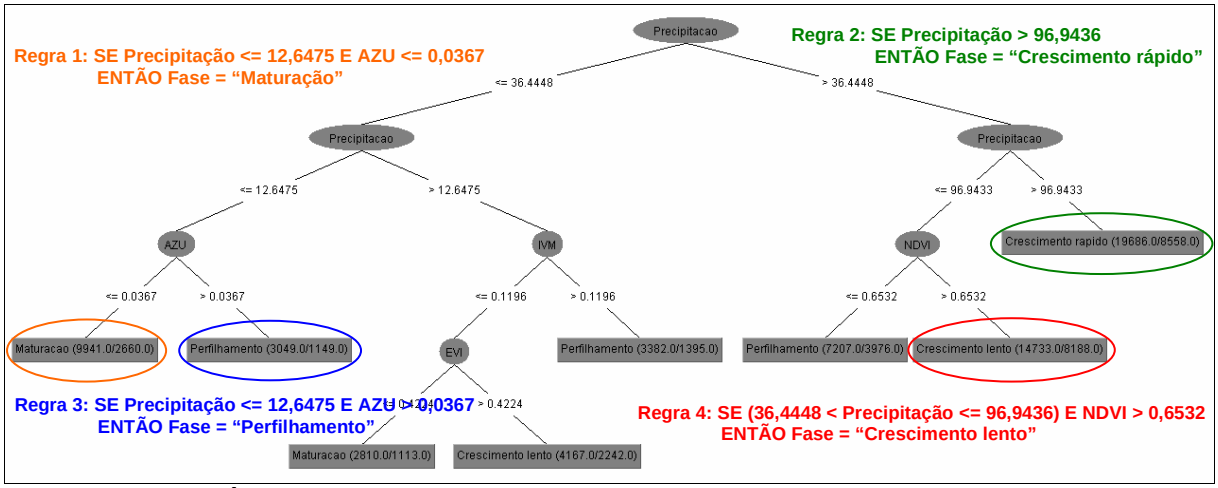


Figura 3: Árvore de decisão para um nível de pré-poda de 2.400 pixels por folha.

4. CONCLUSÕES

Os métodos de seleção de atributos χ^2 e InfoGain mostraram que todos os atributos do conjunto de dados foram considerados relevantes para a classificação.

O balanceamento de classes realizado pelo método de amostragem foi fundamental para melhorar a acurácia do modelo de classificação gerado pelo algoritmo J48.

A descoberta do conhecimento pode ser feita através de regras de decisão relevantes para especialistas, evidenciando a aderência das técnicas de mineração de dados para classificação das fases fenológicas da cultura da cana-de-açúcar.

5. REFERÊNCIAS

FERNANDES, J. L. **Monitoramento da cultura de cana-de-açúcar no Estado de São Paulo por meio de imagens Spot Vegetation e dados meteorológicos.** 97 p. (Dissertação de Mestrado). Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas. 2009.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques.** 2nd ed., 770 p., San Francisco: Morgan Kaufmann, 2006.

HUETE, A.; DIDAN, K.; MIURA, T.; RODRIGUEZ, E. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. (Special Issue). **Remote Sensing of Environment**, v. 83, n.1-2, p. 195-213, 2002.

LATORRE, M. L.; SHIMABUKURO, Y. E.; ANDERSON, L. O. **Sensor MODIS: Produtos para Ecossistemas Terrestres - MODLAND.** In: O Sensor Modis e suas aplicações ambientais no Brasil - Shimabukuro. Y. E.; Rudorff, B. F. T.; Ceballos, J. C. (Coords). São José dos Campos: Editora Parêntese, SP, Brasil, 2007.

ROUSE, J. W.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the great plains with ERTS. In: ERTS-1 SYMPOSIUM, 3., **Proceedings…** Washington, D. C.: NASA, v. 1, p. 309-317, 1973.

WITTEN, I. H.; FRANK, E. **Data mining: pratical machine learning tools and techniques.** 2nd ed., 525 p., San Francisco: Morgan Kaufmann, 2005.