Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 7813-7820.

GeoDMA – Um sistema para mineração de dados de sensoriamento remoto

Thales Sehn Korting[1]
Leila Maria Garcia Fonseca[1]
Maria Isabel Sobral Escada[1]
Gilberto Câmara[1]

[1] Instituto Nacional de Pesquisas Espaciais – INPE
Divisão de Processamento de Imagens Caixa Postal 515 – 12245-970
São José dos Campos – SP, Brasil
{tkorting,leila,isabel,gcamara}@dpi.inpe.br

Abstract. Although a huge amount of remote sensing data has been provided by Earth observation satellites, a few data manipulation and information extraction techniques based on data mining have been developed and made available for remote sensing users. Besides, the actual implementations of geographical data mining systems are generally proprietary software, which prejudices the academic research. In this context, the present paper presents a new system for remote sensing data mining – GeoDMA (Geographical Data Mining Analyst). It implements all processing steps involved in the image analysis process: segmentation, attributes extraction and selection, classification and exploratory data analysis tools. The proposed system deals with remote sensing imagery and objects obtained by segmentation. Taking into account the images and segments the system is able to extract spatial and spectral attributes, which are used in the classification stage. GeoDMA works as an add-on for TerraView, which means that it uses all the geographical database manipulation and visualization structure provided by TerraView. Both systems are developed using TerraLib library and are free software. In this paper we describe the principal modules implemented in the proposed system. An example for urban application is presented to better understand all processing steps involved in the data mining process. For the future, the proposed system will include temporal analysis tools to study the objects evolution, record its evolution history as well as to perform scenario simulations.

Keywords: data mining, image processing, software mineração de dados, processamento de imagens, sistema computacional.

1. Introdução

Com a constelação de satélites em crescimento, aumenta também o volume de dados disponíveis. Com isso, torna-se muito importante o desenvolvimento de sistemas capazes de manipular esses dados de modo a convertê-los em informações que possam ser utilizadas pelas diversas esferas da sociedade. Um exemplo de banco de dados geográficos com alto volume de informação é o do projeto PRODES (CÂMARA; VALERIANO; SOARES, 2006), que contém dados atualizados sobre o desmatamento da Amazônia.

Ferramentas de mineração de dados podem aumentar o potencial da análise e aplicações de dados de sensoriamento remoto. Entretanto, poucos sistemas de mineração de dados estão disponíveis para os usuários. Um exemplo de sistema de mineração de dados com aplicação em sensoriamento remoto é o desenvolvido por (SILVA et al., 2005). Silva propõe um método para identificar padrões de desmatamento da Amazônia baseado em atributos geométricos dos objetos. O protótipo desenvolvido por eles utiliza as funcionalidades de outros sistemas, como SPRING (CÂMARA et al., 1996), Fragstats (MCGARIGAL; MARKS, 1995), e WEKA (WITTEN; FRANK, 2000) em algumas fases do processamento, o que dificulta seu uso.

Portanto, o objetivo deste trabalho é apresentar um sistema para mineração de dados, chamado GeoDMA, (*Geographical Data Mining Analyst*), cujo funcionamento é acoplado ao sistema TerraView, disponível livremente no site `http://www.dpi.inpe.br/terraview/`. TerraView é um sistema computacional capaz de lidar com bancos de dados espaciais, imagens e regiões resultantes do processo de segmentação, além de outros dados geográficos, como por exemplo, mapas de estradas, dados cadastrais, etc.

O GeoDMA realiza todas as fases de processamento necessárias para manipular dados de sensoriamento remoto, incluindo os processos de segmentação, extração e seleção de atributos, treinamento, classificação e análise exploratória dos dados.

2. Mineração de dados espaciais em imagens de sensoriamento remoto

Os sistemas de mineração de dados são usados para manipular uma grande quantidade de informação utilizando as técnicas de aprendizado por máquina para extrair ou ajudar a evidenciar padrões nestes dados, e auxiliando na descoberta de conhecimento. Exemplos de sistemas deste tipo são: VISIMINE (AKSOY et al., 2004), ADam (RUSHING et al., 2005), KIM (SCHRODER et al., 2000) e PattFinder (SILVA et al., 2005), que estão mais focados em métodos de agrupamento que operam somente no espaço de atributos espectrais.

O método de mineração proposto por Silva et al. (2005) foi desenvolvido para uma aplicação específica de identificação de padrões de desmatamento. Para o desenvolvimento da metodologia, os seguintes passos foram implementados:

- Definição de uma tipologia de padrões, de acordo com a aplicação desejada (Um exemplo de tipologia para o desmatamento é mostrado na Figura 1);
- Construção de um conjunto de referência com padrões espaciais. Este conjunto de referência é construído por meio de um conjunto de imagens protótipo. Objetos da paisagem são então identificados e rotulados, num processo supervisionado. Esta identificação implica na segmentação da imagem e sua rotulação é realizada de acordo com a tipologia espacial anteriormente definida (Figura 2);
- Mineração do banco de dados por meio de um classificador estrutural (guiado pelo domínio da aplicação), associando o conjunto de referência aos padrões espaciais identificados nas imagens, assim revelando as configurações espaciais presentes em cada imagem.

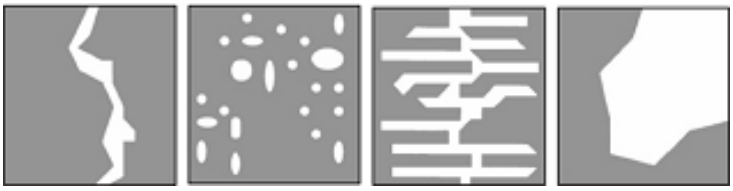


Figura 1: Padrões espaciais do desmatamento tropical. Da esquerda para a direita: corredor, difuso, espinha de peixe e geométrico. (LAMBIN; GEIST; LEPERS, 2003).

O sistema GeoDMA surgiu a partir das idéias propostas por Silva et al. (2005), mas que ao longo do seu desenvolvimento se tornou mais completo, podendo ser usado em diversas aplicações, tais como em agricultura e áreas urbanas. O GeoDMA também inclui métricas espectrais, que podem ser extraídas das imagens de entrada, e outros classificadores baseados nos Mapas Auto-Organizáveis (SOM – *Self Organizing Maps*) (KOHONEN, 2001), e em Redes Neurais.

3. Descrição do Sistema

O GeoDMA funciona como um *plugin* para o sistema TerraView. Isto significa que toda a estrutura que manipula e visualiza bancos de dados geográficos é proporcionada pelo TerraView. O *plugin* é executado em conjunto com o TerraView, e produz os resultados que são exibidos

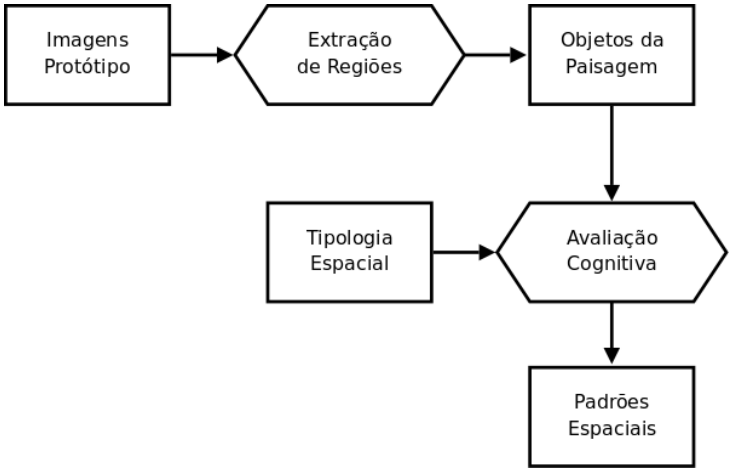
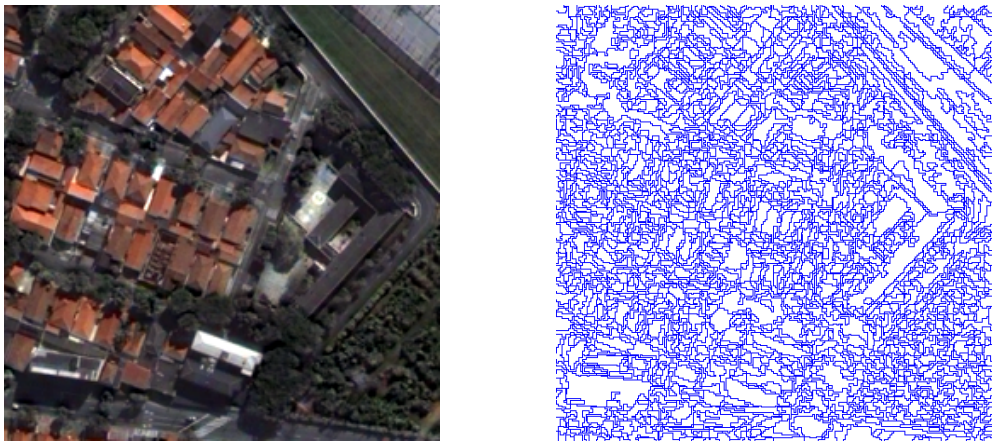


Figura 2: Construção do conjunto de padrões espaciais de referência.

na sua tela principal. O sistema foi desenvolvido em C++, usando a biblioteca TerraLib para as operações geográficas e de processamento de imagens, e QT para a construção da interface com o usuário.

A entrada do sistema GeoDMA é composta de imagens e objetos (segmentos) resultantes da segmentação, como mostra a Figura 3. Neste sentido é possível integrar a informação espacial, contida nos segmentos, e a informação espectral extraída das bandas de entrada, proporcionando ao usuário uma vasta gama de atributos para serem utilizados na classificação e reconhecimento dos padrões presentes em seu banco de dados.



(a) (b)

Figura 3: Dados de entrada: a) Imagem original e b) resultado da segmentação de (a).

A Figura 4 apresenta um diagrama que descreve todas as fases de processamento do sistema GeoDMA. Deve-se considerar que as bases de dados e atributos ficam armazenadas na estruturas de dados do TerraView, bem como a visualização dos resultados, através de mapas temáticos.

A seguir, apresenta-se uma descrição de todos os módulos de processamento do sistema proposto:

Extração de Atributos: este módulo realiza a extração de atributos, considerando imagens e objetos como entrada. Assim, características espectrais como a média por banda, variância de cada banda e textura são extraídas. Atributos espaciais incluem área,

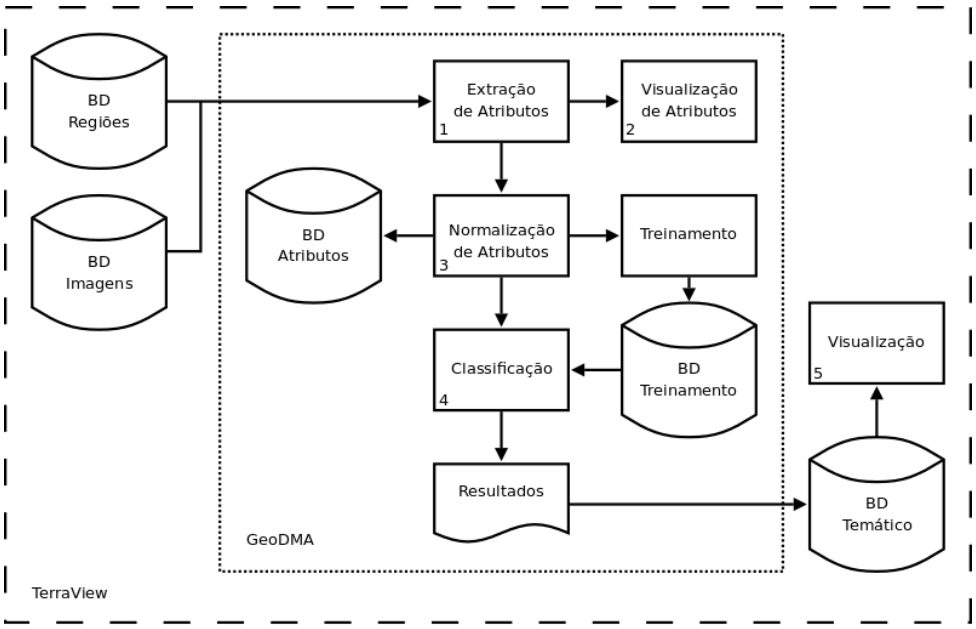


Figura 4: Diagrama do sistema GeoDMA – partindo de imagens e regiões até os mapas temáticos.

perímetro, dimensão fractal, grau de retangularidade e ângulo principal, que são calculados e armazenados nas tabelas de referência do sistema TerraView;

Visualização dos atributos: dado o conjunto de atributos, o usuário possui uma ferramenta muito importante para realizar a análise exploratória, através da visualização do espaço de atributos. Dois atributos são selecionados e um gráfico de dispersão mostra a separabilidade dos dados. A Figura 5 mostra um exemplo com os atributos "relação perímetro-área" e "média dos *pixels* na banda 1";

Normalização de atributos: o estágio de normalização é fundamental para a análise exploratória, visto que os dados possuem diferentes escalas, e que podem mascarar determinados atributos, e destacar outros, quando na verdade todos deveriam possuir o mesmo grau de importância no início do processo de mineração de dados;

Classificação: dois algoritmos para classificação estão disponíveis na versão atual: o algoritmo supervisionado de árvore de decisão versão C4.5 (QUINLAN, 1993), e o algoritmo não-supervisionado de Mapas Auto-Organizáveis (SOM);

Visualização: a interface do TerraView proporciona a visualização dos dados em uma estrutura onde a saída é dividida em diferentes classes, ou temas, de acordo com os resultados da classificação.

4. Operação do Sistema: Um Exemplo de Aplicação

Esta seção mostra um exemplo de aplicação para um dado de área urbana. Neste exemplo os dados de entrada são a imagem multiespectral QuickBird de uma região de São José dos Campos, e os segmentos obtidos pelo segmentador por crescimento de regiões (BINS et al., 1996). A próxima fase de processamento é a extração de atributos. Neste caso, todas as métricas disponíveis no sistema foram utilizadas. Os valores dos atributos podem ser normalizados de

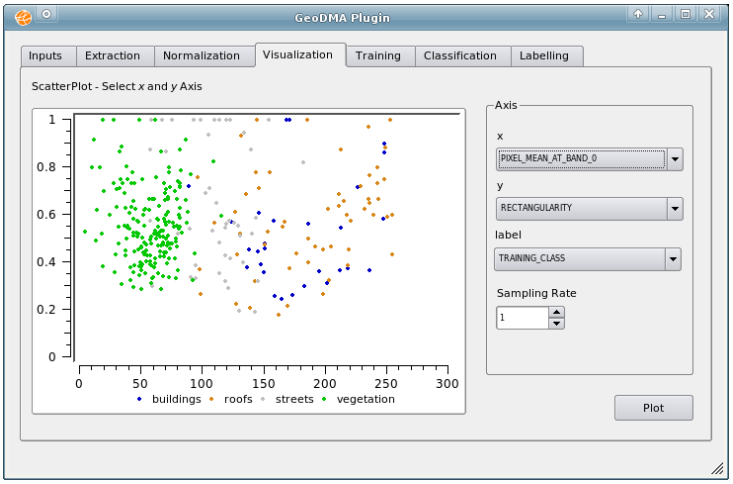


Figura 5: Gráficos de dispersão utilizando atributos espectrais e espaciais.

acordo com a opção do usuário. Os atributos disponíveis para extração são apresentados em uma janela (Figura 6), onde pode-se selecionar aqueles que serão usados.

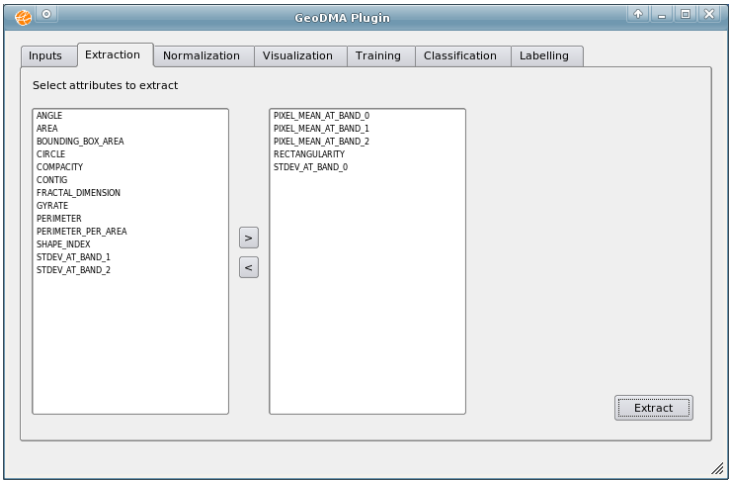


Figura 6: Atributos disponíveis para extração.

A Figura 7 mostra a fase de treinamento, onde o usuário seleciona regiões amostrais para cada classe (prédios, ruas, telhados e vegetação). Ou seja, cada região é rotulada com o identificador de cada classe, de acordo com a escolha do usuário. Essas regiões amostrais são utilizadas para projetar o classificador, que pode ser qualquer um dos implementados no sistema. Neste caso foi utilizado o classificador por árvore de decisão. Nesta fase pode-se visualizar a distribuição dos dados no espaço de atributos para analisar a correlação entre os dados para discriminar determinadas classes, similar ao gráfico mostrado na Figura 5.

Após a análise dos dados, todas as regiões são classificadas. No presente exemplo, o classificador por árvores de decisão foi o escolhido. A árvore resultante é mostrada ao usuário para que o mesmo possa validar seu modelo, e entender a contribuição de cada atributo no resultado final. Neste exemplo, a árvore gerada é mostrada na Figura 8. O sistema mostra o mapa temático diretamente na tela principal do TerraView (Figura 9). Cada classe é mapeada para o tema definido na etapa de treinamento. Desta maneira, o usuário pode mostrar ou ocultar classes específicas, de acordo com a aplicação desejada.

Uma vantagem da utilização do GeoDMA é a facilidade nos processos de extração,

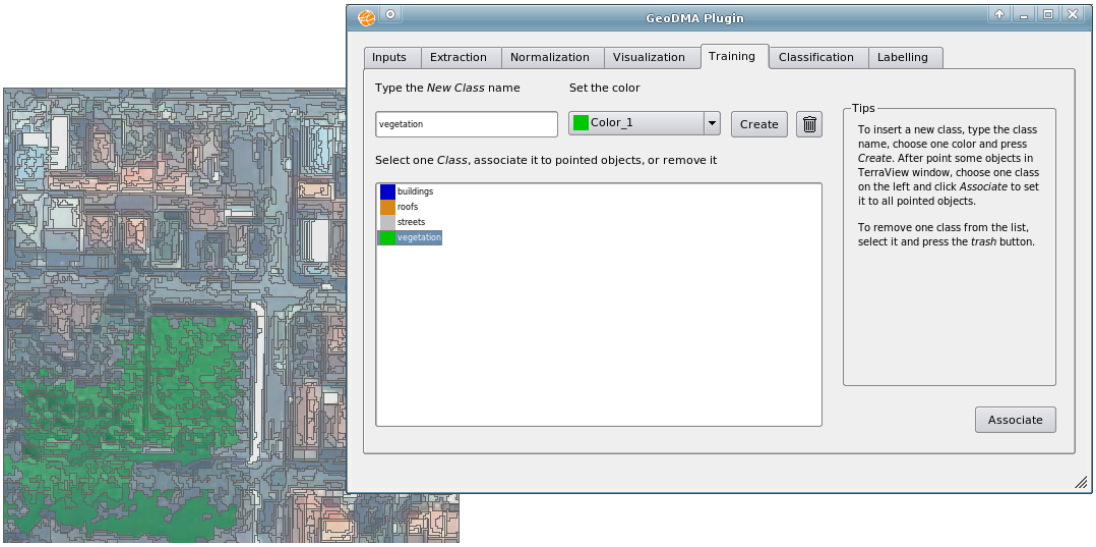


Figura 7: Etapa de treinamento na interface do GeoDMA.

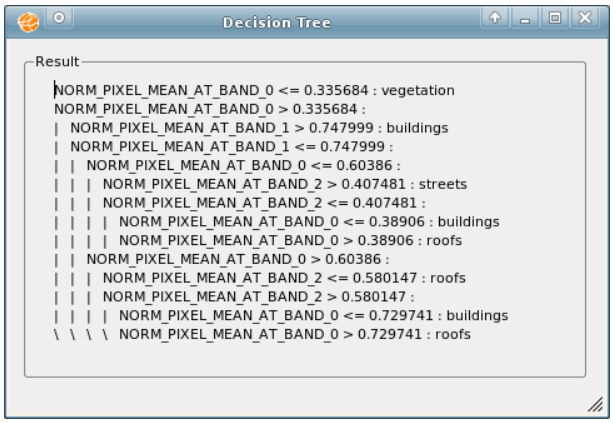


Figura 8: Árvore de Decisão resultante.

normalização e armazenamento dos atributos dos dados, que podem ser utilizados posteriormente em outras análises. Em grandes bancos de dados, esta ferramenta é muito útil, especialmente quando são consideradas a carga de trabalho e o tempo de processamento. Além disso, o sistema torna-se atrativo pelo uso da interface do TerraView, na qual diferentes temas (ruas, telhados, etc.) podem ser exibidos separadamente. A interface do GeoDMA é amigável, pois as fases de processamento seguem uma sequência lógica similar aos procedimentos de análise realizados pelo usuário.

5. Conclusões

Um sistema de mineração de dados de sensoriamento remoto foi apresentado neste trabalho. Este sistema é baseado no protótipo proposto por Silva et al. (2005), onde foram adicionados mais funcionalidades e integradas no mesmo ambiente computacional. O sistema foi desenvolvido como um *plugin* do TerraView, aproveitando os seus recursos de gerenciamento de banco de dados geográficos para facilitar a visualização dos dados. O código foi desenvolvido utilizando a biblioteca TerraLib, disponível livremente através do endereço `http://www.terralib.org/`. Mais informações sobre o sistema podem ser encontradas no endereço `http://www.dpi.inpe.br/geodma/`.

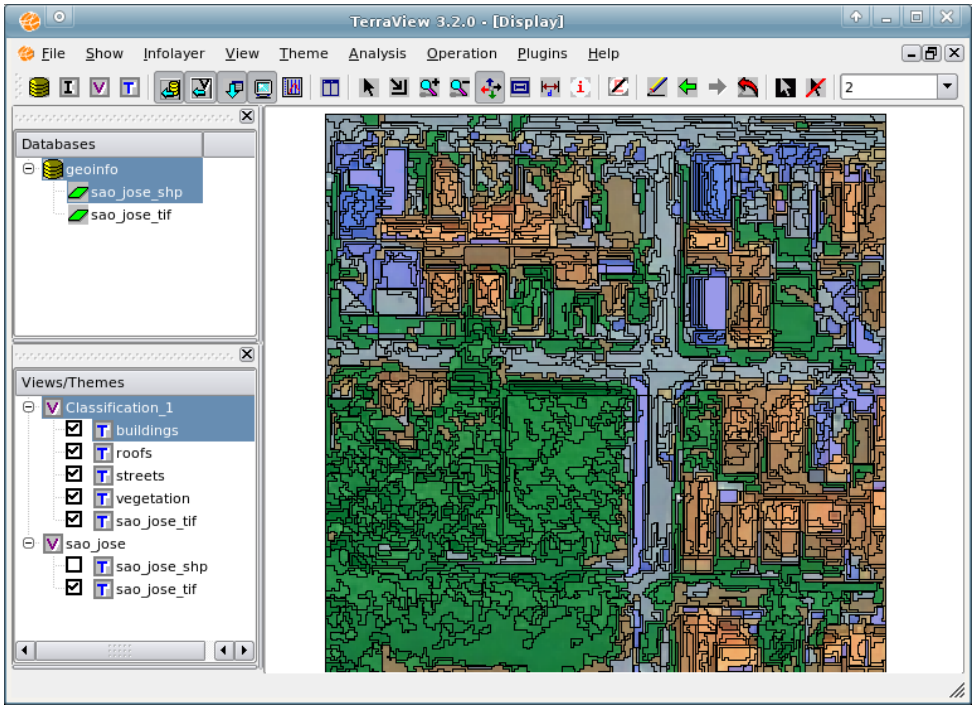


Figura 9: Mapa temático resultante.

Trabalhos futuros incluem a otimização dos algoritmos incluídos no sistema, de modo a torná-lo mais eficiente. Outras modificações, tais como a inclusão de ferramentas para seleção de atributos, de acordo com métricas propostas em (OLIVEIRA; DUTRA; RENNÓ, 2005; FAYYAD; IRANI, 1992) e a manipulação de dados temporais para analisar a evolução dos padrões, devem ser implementadas.

Referências

AKSOY, S. et al. Interactive training of advanced classifiers for mining remote sensing image archives. In: ACM NEW YORK, NY, USA. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2004. p. 773–782.

BINS, L. et al. Satellite imagery segmentation: a region growing approach. *SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO*, INPE, v. 8, 1996.

CÂMARA, G. et al. Spring: Integrating remote sensing and gis by object oriented data modeling. *Computers and Graphics*, v. 20, p. 3, 1996.

CÂMARA, G.; VALERIANO, D. de M.; SOARES, J. V. *Metodologia para o Cálculo da Taxa Anual de Desmatamento na Amazônia Legal*. São José dos Campos, 2006. 24 p.

FAYYAD, U.; IRANI, K. The Attribute Selection Problem in Decision Tree Generation. In: AMER ASSN FOR ARTIFICIAL. *AAAI-92: Proceedings Tenth National Conference on Artificial Intelligence/July 12-16, 1992*. [S.l.], 1992. v. 1001, p. 48109.

KOHONEN, T. *Self-Organizing Maps*. [S.l.]: Springer, 2001.

LAMBIN, E.; GEIST, H.; LEPERS, E. Dynamics of land-use and land-cover change in tropical regions. *Annual Review of Environment and Resources*, v. 28, p. 205–241, 2003.

MCGARIGAL, K.; MARKS, B. *FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure*. [S.l.]: US Dept. of Agriculture, Forest Service, Pacific Northwest Research Station, 1995.

OLIVEIRA, J.; DUTRA, L.; RENNÓ, C. Aplicação de Métodos de Extração e Seleção de Atributos para Classificação de Regiões. *Anais XII Simpósio Brasileiro de Sensoriamento Remoto*, p. 4201–4208, 2005.

QUINLAN, J. *C4. 5: Programs for Machine Learning*. [S.l.]: Morgan Kaufmann, 1993.

RUSHING, J. et al. ADaM: a data mining toolkit for scientists and engineers. *Computers and Geosciences*, Elsevier, v. 31, n. 5, p. 607–618, 2005.

SCHRODER, M. et al. Interactive learning and probabilistic retrieval in remote sensingimage archives. *Geoscience and Remote Sensing, IEEE Transactions on*, v. 38, n. 5 Part 1, p. 2288–2298, 2000.

SILVA, M. P. S. et al. Mining patterns of change in remote sensing image databases. In: *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2005. p. 362–369. ISBN 0-7695-2278-5.

WITTEN, I.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. [S.l.]: Morgan Kaufmann, 2000.