Análise comparativa de algoritmos de árvore de decisão do sistema WEKA para classificação do uso e cobertura da terra

Luciane Yumie Sato [1]
Yosio Edemir Shimabukuro [1]
Tatiana Mora Kuplich [2]
Vitor Conrado Faria Gomes [1]

[1] Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil
{lusato, yosio}@dsr.inpe.br
vitor.gomes@inpe.br

[2] Centro Regional Sul de Pesquisas Espaciais – CRS - INPE
Caixa Postal 5021 – 97105-970 – Santa Maria - RS, Brasil
tmk@dsr.inpe.br

Abstract. In the last years, the data mining techniques are increasingly used for classification purposes, and between several techniques it is highlighted the decision tree. This tool improves the accuracy of classification, and also allows the integration of different data types in the classification. Thus, this work has as main objective to analyze and compare the best data mining algorithm of decision tree available in WEKA software to use for land use and land cover classification in the Tapajos National Forest region. For this, we used a Landsat-5/TM image, the fraction images obtained by the Linear Spectral Mixture Model, the Normalized Difference Vegetation Index, the Normalized Water Index and the Soil-Adjusted Vegetation Index as input data for the creation of the decision trees. To define the best algorithm, the total size of the decision tree, the number of leaves, the time taken for the creation of the decision tree the number of pixels correctly classified, the number of incorrectly classified pixels and Kappa were considered. The algorithm that presented the best results and that best described the classes of land use and land cover of the study area was SimpleCart algorithm, that is an implementation of the Classification and Regression Tress algorithm. The decision tree technique showed satisfactory results in the classification of the images and the results were generated quickly, showing the computational efficiency of this technique.

Palavras-chave: data mining, image classification, Tapajos National Forest, remote sensing, mineração de dados, classificação de imagens, Floresta Nacional do Tapajós, sensoriamento remoto.

1. Introdução

A Floresta Amazônica se destaca como fonte global de diversidade de fauna, flora, fonte de oxigênio e água doce. O bioma Amazônia desempenha um importante papel no balanço global do carbono e na influência sobre o sistema climático, principalmente da América do Sul, uma vez que essas florestas contribuem no direcionamento da circulação atmosférica nos trópicos (MARENGO et al., 2011).

As alterações que ocorrem na cobertura da terra nas regiões das florestas tropicais podem prejudicar severamente o funcionamento da Amazônia e, consequentemente, podem ocorrer diminuições na capacidade dessas florestas em absorver o carbono, atenuar o ciclo hidrológico regional, elevar a temperatura do solo e, gradualmente, ocasionar processos de savanização nessas áreas (MARENGO et al., 2011).

Para o acompanhamento e mapeamento da cobertura e uso da terra, diversas técnicas de mineração de dados vêm sendo utilizadas com sucesso quando aplicadas em escalas locais, regionais e globais, devido ao elevado grau de detalhes obtido com estas ferramentas de classificação (SESNIE et al., 2012; POULIOT et al., 2009; DAVRACHE et al., 2012)

Entre as técnicas de mineração de dados, a árvore de decisão é um método que se destaca no contexto de classificação de imagens, pois fornece um aumento da acurácia e permite classificar conjuntos distintos de dados (HANSEN et al., 1996; IM e JENSEN, 2005;

HANSEN et al., 2008; GONG et al., 2011). No contexto do sensoriamento remoto, voltado para o mapeamento do uso e cobertura da terra, essa última característica possibilita obter melhores resultados através da integração de diferentes produtos e de diferentes dados provenientes de múltiplos sensores.

As árvores de decisão são funções que apresentam como dados de entrada um vetor de atributos e uma decisão como valor de saída: sim ou não (BREIMAN et al., 1984). O funcionamento da árvore de decisão ocorre através da divisão de um conjunto de dados em subconjuntos de forma recursiva. A separação dos dados ocorre até que cada subconjunto esteja homogêneo, com casos de uma única classe (WITTEN et al., 2011).

Uma árvore de decisão é formada por nós, ramos e folhas. Os nós representam regiões onde são realizados testes lógicos para a separação dos dados. O primeiro nó é chamado de nó raiz e é o nó principal da árvore de decisão. Os nós que estão localizados abaixo do nó raiz são os nós filhos e esses nós estão conectados por ramos. As folhas são as regiões que estão associadas a um rótulo ou valor (FRIELD e BRODLEY, 1997; SOBRAL, 2005; POZZER, 2006; AITKENHEAD, 2008). A estrutura básica de uma árvore de decisão é ilustrada na Figura 1.

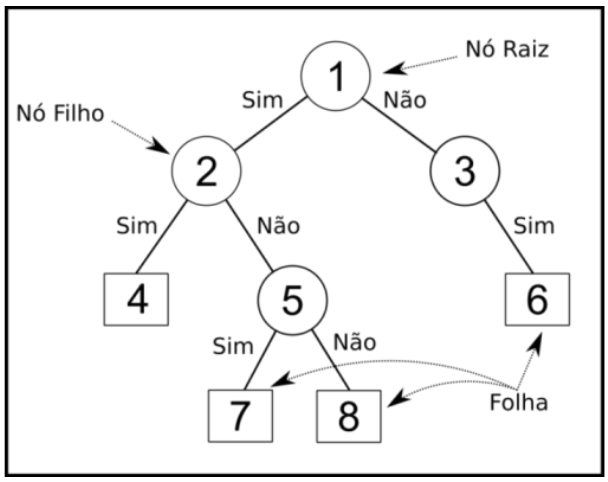

Figura 1. Estrutura geral de uma árvore de decisão.

Entre as principais vantagens do uso das árvores de decisão, se destaca a fácil interpretação dos seus resultados, pois a classificação é obtida de forma explicita, simplificando a sua interpretação. Além disso, os resultados geralmente são fornecidos rapidamente, devido à eficiência computacional apresentada por esta técnica (HANSEN et al., 1996; FRIELD; BRODLEY, 1997).

Inserido neste contexto, esse trabalho tem como principal objetivo realizar uma análise comparativa entre os algoritmos de mineração de dados que estão disponíveis no aplicativo *Waikato Environment for Knowledge Analysis* (WEKA), para a classificação do uso e cobertura da terra na região que abrange a Floresta Nacional (Flona) do Tapajós.

2. Metodologia de Trabalho

2.1 Dados de Sensoriamento Remoto

Para o treinamento e criação das árvores de decisão no WEKA, foi utilizada uma imagem do sensor *Thematic Mapper* (TM) do satélite Landsat-5 obtida no ano de 2009. Além das seis bandas espectrais dessa imagem, também foram utilizados quatro produtos, com intuito de realçar as áreas que serão classificadas e contribuir para uma melhor separabilidade das classes. A Figura 2 apresenta os dados de sensoriamento remoto e as etapas para o desenvolvimento do trabalho.

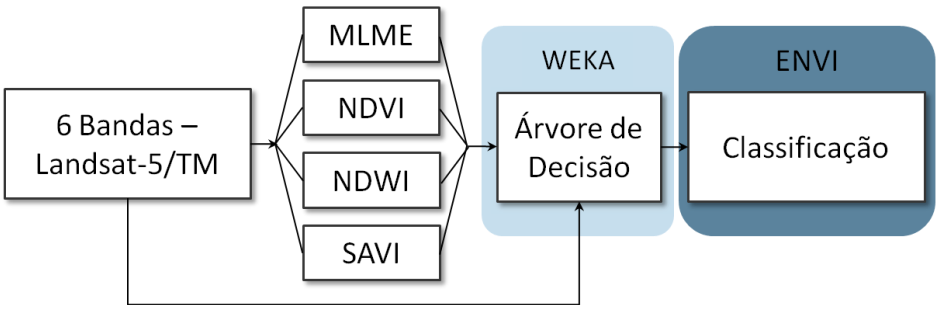

Figura 2. Etapas e os produtos de sensoriamento remoto utilizados no desenvolvimento do trabalho.

Para obter o primeiro conjunto de dados, o Modelo Linear de Mistura Espectral (MLME), foram utilizadas as bandas 1, 2, 3, 4, 5 e 7 da imagem Landsat-5/TM, conforme proposto por Shimabukuro e Smith (1991). Este modelo fornece três imagens-fração, denominadas de imagens-fração solo, sombra ou água e vegetação, e a formulação do modelo é dada pela equação

$$r_i = a * vege_i + b * solo_i + c * agua_i + e_i \quad (1)$$

onde r_i corresponde a resposta do pixel na banda i; a é a proporção de vegetação; b é a proporção de solo; c é a proporção de sombra ou água; $vege_i$ é a resposta espectral do componente vegetação na banda i; $solo_i$ é a resposta espectral do componente solo na banda i; $agua_i$ é a resposta espectral do componente sombra na banda i; e_i é o erro na banda i; sendo que i varia entre as bandas 1, 2, 3, 4, 5 e 7.

A primeira etapa para a aplicação do MLME é realizada através da determinação dos componentes puros (*endmembers*) da imagem – solo, sombra e vegetação. Nesse trabalho, para a determinação dos valores espectrais de cada componente foram selecionadas amostras de solo exposto para a componente solo, áreas em regeneração para a componente vegetação e por fim, uma amostra de água limpa para a componente sombra.

Além das imagens-fração do MLME, foram obtidos três índices de vegetação – o Índice da Diferença Normalizada (NDVI), o Índice de Umidade por Diferença Normalizada (NDWI) e o Índice de Vegetação Ajustado para o Solo (SAVI), cujas fórmulas são apresentadas na Tabela 1.

Tabela 1. Índices de vegetação obtidos a partir da imagem Landsat-5/TM.

Índice de Vegetação	Fórmula	Referência
NDVI	$NDVI = \frac{(\rho_{IVP} - \rho_V)}{(\rho_{IVP} + \rho_V)}$	Rouse et al. (1973)
NDWI	$NDWI = \frac{(\rho_{NIR} - \rho_{MIDIR})}{(\rho_{NIR} + \rho_{MIDIR})}$	Gao (1996)
SAVI	$SAVI = \frac{(\rho_{NIR} - \rho_R)}{(\rho_{NIR} + \rho_R + L)} * (1 + L)$	Huete (1988)

Na Tabela 1, ρ_{IVP} é o valor de reflectância no infravermelho próximo, ρ_V é o valor da reflectância no vermelho, ρ_{NIR} representa os valores de reflectância no infravermelho próximo, ρ_{MIDIR} representa os valores de reflectância no infravermelho médio e L é uma constante que minimiza o efeito do solo e pode variar de 0 a 1. Para esse estudo optou-se

utilizar o valor 0,5 para a constante *L*, pois no entorno da Flona do Tapajós, nem todas as áreas com vegetação são de alta densidade. Neste caso, o uso de um valor intermediário é mais adequado para toda a região de estudo.

Na Figura 3.a é apresentada uma composição colorida (R) imagem-fração solo, (G) imagem-fração vegetação e (B) imagem-fração sombra do MLME, na Figura 3.b o índice NDVI, na Figura 3.c o índice NDWI e na Figura 3.d o índice SAVI.

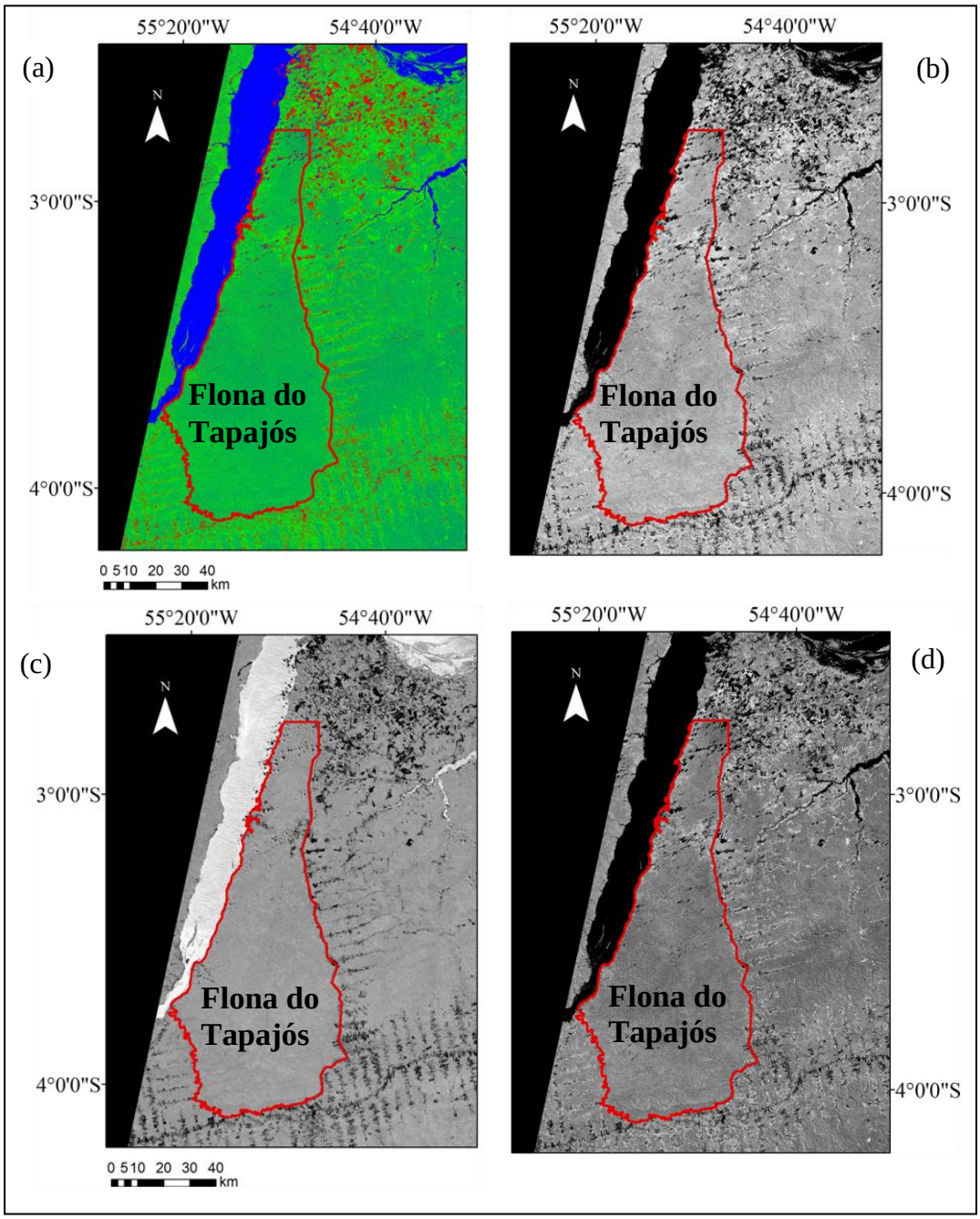

Figura 3. (a) Composição colorida (R), (G) e (B) das imagens-fração solo, vegetação e sombra, respectivamente, na (b) o índice NDVI, na (c) o índice NDWI e na (d) o índice SAVI.

2.2 Aplicativo WEKA

Para criação das árvores de decisão foi utilizado o sistema livre de mineração de dados WEKA na versão 3.6.4. Este aplicativo foi desenvolvido pela Universidade de Waikato da Nova Zelândia, apresentando interface amigável e rápido processamento dos dados. Este sistema pode ser adquirido gratuitamente em http://www.cs.waikato.ac.nz/ml/weka/.

O WEKA fornece diversos algoritmos em seu ambiente e, além disso, permite que sejam criados algoritmos de aprendizado. Os principais algoritmos que estão disponíveis são: regressão, classificação, mineração de regras de associação, seleção de atributos, etc.

Para o desenvolvimento desse trabalho foram testados os doze algoritmos de árvore de decisão que estão implementados no WEKA, sendo eles: BFTree, DecisionStump, FT, J48, J48graft, LADTree, LMT, NBTree, RandomForest, RandomTree, REPTree e SimpleCart.

O treinamento dos dados e a validação da classificação foram realizados no próprio aplicativo pelo método *Holdout*, conhecido como método Treino-Teste. Nesse método $^2/_3$ dos dados são destinados para o treinamento ou estimação dos parâmetros da árvore de decisão e $^1/_3$ dos dados são usados para validação do modelo (OLSON e DELEN, 2008).

3. Resultados e Discussão

Para avaliar o melhor algoritmo de árvore de decisão para classificação do uso e cobertura da terra da área em estudo, foram considerados o tamanho total da árvore de decisão, o número total de folhas, o tempo gasto para geração da árvore de decisão, o número de pixels corretamente classificados, o número de pixels incorretamente classificados e o valor do índice Kappa, obtido a partir da matriz de confusão. O índice Kappa utilizado para análise da classificação é fornecido pelo aplicativo WEKA após a fase de validação, motivando o uso, pois, desta forma, representa um índice equivalente para os algoritmos. É importante salientar que o número de folhas das árvores de decisão podem ser maiores que o número de classes determinados na classificação do uso e cobertura da terra, pois podem existir diversas combinações de atributos que levam a uma mesma classe.

A Tabela 2 lista as árvores de decisão testadas nesse trabalho e os parâmetros que foram utilizados para a análise comparativa. A árvore de decisão LMT apresenta um melhor índice Kappa e um menor número de nós. Apesar disso, sua estrutura é complexa e não pode ser confeccionada no aplicativo *Environment for Visualizing Images* (ENVI), o qual foi utilizado para a classificação da imagem. Por conta disso, o segundo algoritmo com melhor índice Kappa, SimpleCart, foi selecionado para gerar a classificação da imagem, estando destacado em vermelho na Tabela 2. O algoritmo SimpleCart é uma implementação do algoritmo *Classification and Regression Trees* (CART) proposto por Breiman et al. (1984).

Com relação ao número de nós e o tempo de treinamento, as árvores de decisão apresentaram grande variação no número de nós e em na maioria dos algoritmos o tempo gasto para a sua criação foi pequeno. O algoritmo que necessitou de mais tempo para a criação da árvore de decisão foi o LMT, sendo necessários aproximadamente 44 minutos. No outro extremo, está a árvore de decisão RandomTree, que levou apenas 0,68 segundos para ser confeccionada.

Entre as doze árvores de decisão testadas, onze delas foram eficientes na determinação das classes, pois apresentaram elevados valores de pixels corretamente classificados e alto valor do índice Kappa. Apenas a árvore de decisão do algoritmo DecisionStump não apresentou bons resultados, atingindo apenas 21% de acertos.

O algoritmo SimpleCart se diferenciou dos demais algoritmos devido ao tamanho da árvore, pois o número de folhas é menor em relação aos demais algoritmos, mostrando que esse método necessitou de um menor número de testes lógicos para a determinação das classes. Pode-se dizer que o cálculo do algoritmo SimpleCart foi mais eficiente do que os demais de acordo com as amostras e classes determinadas nesse trabalho.

A Figura 4 apresenta a imagem da área de estudo classificada pelo algoritmo SimpleCart gerada pelo sistema WEKA.

Tabela 2. Comparação das árvores de decisão do WEKA.

Árvore de Decisão	Tamanho	N. de Folhas	Tempo	Pixels Corretamente Classificados	Pixels Incorretamente Classificados	Kappa
BFTr	455	288	74,76s	80.8%	19,2%	0,7836
DecionStump	3	3	0,41s	21,4 %	78,6 %	0,1198
FT	427	214	83,79s	80,3%	19,7%	0,7784
J48	2.477	1.239	8,41s	79,9%	20,1%	0,7738
J48Graft	3.589	1.795	12,09s	80%	20%	0,775
LADTree	31	16	68.2s	76%	22,4%	0,7481
LMT	13	7	2.624,46s	81,7%	18,3%	0,7937
NBTree	219	110	49s	74,5%	25,5%	0,7132
RandomForest	14.475	7.238	0,99s	77,1%	22,9%	0,7425
RandomTree	14.475	7.236	0,68s	77,1%	22,9%	0,7425
RepTree	867	434	1,45s	80,1%	19,9%	0,776
SimpleCart	335	168	23s	81,2%	18,8%	0,7883

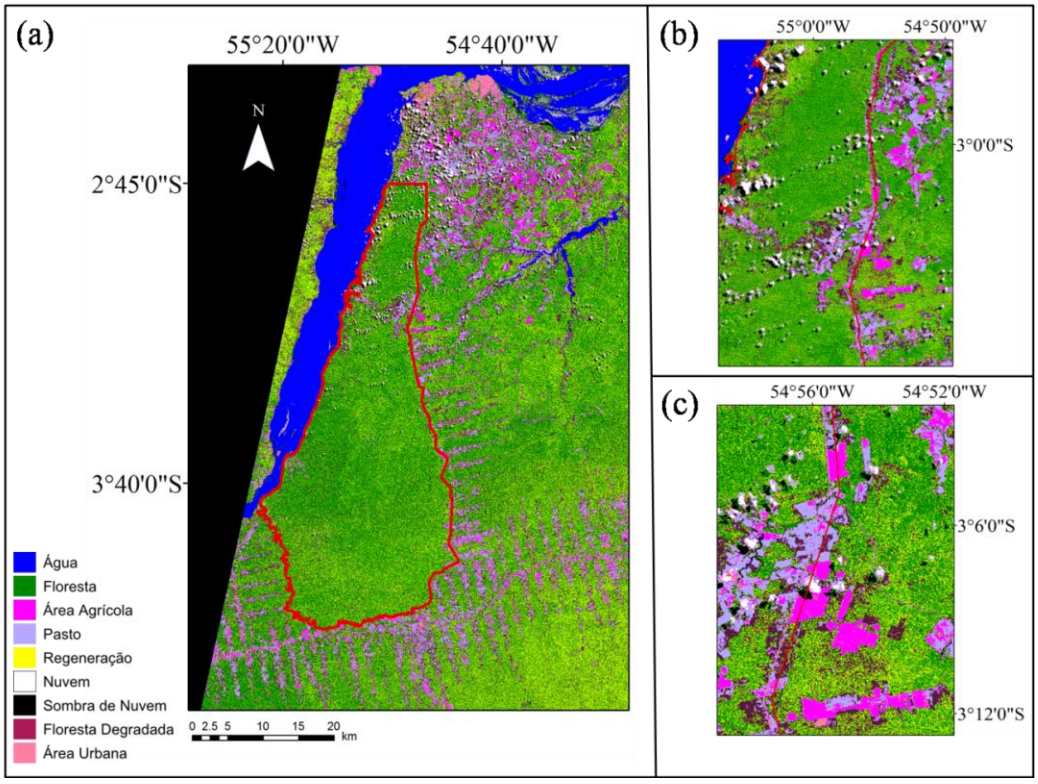

Figura 4. Imagem Landsat-5/TM de 2009 classificada pelo algoritmo SimpleCart do WEKA.

4. Conclusões

A técnica de mineração de dados pela árvore de decisão apresentou um bom desempenho na classificação do uso e cobertura da terra. Dentro dos objetivos propostos a maioria das árvores de decisão alcançaram resultados satisfatórios. Apenas uma delas – DecisionStump – não separou corretamente as classes determinadas nesse estudo.

Além disso, as árvores de decisão se mostraram computacionalmente eficientes, pois os resultados foram gerados rapidamente, mesmo com em árvores de decisão com elevado número de nós.

O algoritmo que apresentou estrutura compatível com o sistema ENVI e que melhor classificou a imagem foi o SimpleCart. Esse algoritmo destacou-se devido ao alto número de pixels corretamente classificados, ao baixo valor do número de pixels que não foram classificados corretamente e pelo maior valor do índice Kappa.

No desenvolvimento de outros trabalhos com intuito de classificar outras áreas, através do uso de árvores de decisão é interessante que sejam testados todos os algoritmos do WEKA, pois pode haver variações na qualidade da classificação com a variação dos dados de entrada para o treinamento.

Desta forma, poderá ser escolhido o algoritmo que permite a melhor separação do conjunto de dados.

5. Referências

Aitkenhead, M. J. A co-evolving decision tree classification method. **Expert Systems with Applications,** v. 34, p. 18-25, 2008.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C.J. **Classification and regression trees.** Belmont, CA: Wadsworth International, 1984. 358 p.

Davranche, A.; Lefebvre, G.; Poulin, B. Wetland monitoring using classification trees and SPOT-5 seasonal time series. **Remote Sensing of Environment**, v. 114, p. 552-562, 2010.

Friedl, M. A.; Brodley, C. E. Decision tree classification of land cover from remotely sensed data. **Remote Sensing of Environment**, v. 61, p. 399-409, 1997.

Gao, B. NDWI - A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space. **Remote Sensing of Environment**, v. 58, p. 257-266, 1996.

Gong, B.; Im, J.; Mountrakis, G. An artificial immune network approach to multi-sensor land use/land cover classification. **Remote Sensing of Environment**, v. 115, p. 600-614, 2011.

Hansen, M.; Dubayah, R.; Defries, R. Classification trees: an alternative to traditional land cover classifiers. **International Journal of Remote Sensing**, v. 17, n. 5, p. 1075-1081, 1996.

Hansen, M. C.; Roy, D. P.; Lindquist, E.; Adusei, B.; Justice, C. O.; Altstatt, A. A method for integrating MODIS and Landsat data for systematic monitoring of forest cover and change in the Congo Basin. **Remote Sensing of Environment**, v. 112, p. 2495-2513, 2008.

Huete, R. A. A soil-adjusted vegetation index (SAVI). **Remote sensing of Environment**, v. 25, n. 3, p. 295-309, 1988.

Im, J; Jensen, J. R. A change detection model based on neighborhood correlation image analysis and decision tree classification. **Remote Sensing of Environment**, v. 99, p. 326-340, 2005.

Marengo, J, Nobre, C., Chou S. C, Tomasella, J., Sampaio, G., Alves, L., Obregon, G., Soares, W., Risco das Mudanças Climáticas no Brasil, Ed. INPE, São Jose dos Campos, SP, pp.55, 2011.

Olson, D.; Delen; O. D. Advanced Data Mining Techiniques. Editora **Springer.** 2008. 180 p.

Pouliot, D.; Latifovic, R.; Fernandes, R.; Olthof, I. Evaluation of annual forest disturbance monitoring using a static decision tree approach and 250 m MODIS data. **Remote Sensing of Environment**, v. 113, p. 1749-1759, 2009.

Pozzer, C. T. **Aprendizado por árvores de decisão**. Disciplina de Programação de Jogos 3D. Universidade Federal de Santa Maria, 5p., 2006. Disponível em: <http://www-usr.inf.ufsm.br/~pozzer/disciplinas/pj3d_decision Trees.pdf>. Acesso em: 14 dez. 2010.

Rouse, J. W. Jr.; Hass, R. H.; Schell, J. A.; Deering, D. W. Monitoring vegetation systems in the great plains with Erts. In Proceedings of the Third ERTS Symposium, v. 1, p. 309-317, 1973.

Sesnie, S. E.; Gessler, P. E.; Finegan, B.; Thessler, S. Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. **Remote Sensing of Environment**, v. 112, p. 2145-2159, 2008.

Sobral, A. P. B. **Previsão de carga horária – uma nova abordagem por árvore de decisão.** 2005. 56 p. Tese de Doutorado – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2003.

Shimabukuro, Y. E.; Smith, J.A. The least-squares mixing models to generate fraction images derived from remote sensing multispectral data. **IEEE Transactions on Geoscience and Remote Sensing**, v. 29, n. 1, p. 16-20, Jan. 1991Disponível em: < http://www2.dbd.puc-rio.br/pergamum/tesesabertas/9916940_03_pretexto.pdf>. Acesso em: 15 jan. 2011.

Witten, I. H.; Frank, E.; Hall, M. A. **Data mining: practical machine learning tools and techniques.** São Francisco, CA: The Morgan Kaufmann series in data management systems, 2011. 665 p.