Algoritmo para Otimização de Categorização

Carlos Alberto Pires de Castro Filho[1], Corina da Costa Freitas[2], Sidnei João Siqueira Sant´Anna[2]

[1]Programa de Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

[2]Departamento de Processamento de Imagens – DPI
Instituto Nacional de Pesquisas Espaciais – INPE

{pires,corina,sidnei}@dpi.inpe.br

Abstract. *One of the steps in the construction of thematic maps is the categorization. This process is done by partitioning the range of values of a particular attribute in several subintervals and associating a new nominal value to all instances that are in that subinterval. This work aims at presenting a method of categorization that seeks to maximize the Kappa coefficient of agreement, found in the classification process that models the construction of the thematic map. The search method proposed is the hill climbing greedy, having the ability to perform floating steps. The tests were conducted having the objective function as the Kappa coefficient obtained by the validation over training data and by the cross-validation leave-one-out method. As a result, it was observed that the Kappa value increased together with the number of steps performed by the algorithm developed and that new tests are still necessary to validate it.*

Resumo. *Uma das etapas na construção de mapas temáticos é a de categorização. Este processo é feito particionando o intervalo de valores de um determinado atributo em vários subintervalos e associando um novo valor, neste caso ordinal, a todas as instâncias que se encontram naquele subintervalo. O presente trabalho tem por objetivo apresentar um método de categorização que busque maximizar o coeficiente de concordância Kappa encontrado no processo de classificação que modele a construção do mapa temático.O método de busca proposto é o de hill climbing greedy, possuindo a capacidade de realizar passos maiores, chamados de passos flutuantes. Os testes foram realizados tendo como função objetivo o coeficiente Kappa obtido a partir tanto da validação sobre os dados de treinamento como pelo método de validação cruzada leave-one-out. Como resultado observou-se que o valor de Kappa aumentou com o número de passos realizados pelo algoritmo desenvolvido e que novos testes ainda são necessários para validá-lo.*

Palavras-chave: *categorização, otimização, mapa temático, hill climbing.*

1. Introdução

Em um problema computacional, os atributos a serem processados podem ser divididos em grupos de acordo com as características de seus dados. Conforme Yang e Webb [2009] atributos podem ser *qualitativos* ou *quantitativos*.

Atributos qualitativos possuem dados que não admitem operações aritméticas. Bussab e Morettin [2002] afirmam que um atributo qualitativo é nominal quando não existe nenhuma ordenação nas possíveis realizações e ordinal quando essas realizações podem ser ordenadas.

Atributos quantitativos admitem operações aritméticas e podem ser discretos, quando oriundos de uma contagem, ou contínuos, quando oriundos de uma medição [Yang e Webb 2009]. Bussab e Morettin [2002] complementam ainda que os atributos discretos são aqueles que possuem os valores possíveis dentro de um conjunto finito. Por outro lado, os atributos contínuos possuem os valores possíveis pertencentes a um intervalo de números reais.

Os processos de transformação entre diferentes tipos de atributos possuem nomes específicos na literatura. A transformação de atributos quantitativos contínuos em quantitativos discretos é chamada de discretização. A discretização é feita particionando o intervalo de valores de um determinado atributo em vários subintervalos e associando um novo valor discreto específico a todas as instâncias que pertencem a cada subintervalo. Sendo A um atributo que possua valores ordenados $(x_1,...,x_i,...,x_j,...,x_m,...,x_N)$, $1 < i < j < m < N$, referentes às N instâncias e K o número de subintervalos, o processo de discretização D pode ser representado pela Equação 1.

$$D(A): \left\{ [x_1; x_i]_1, [x_{i+1}; x_j]_2, ..., [x_m; x_N]_K \right\} \qquad \text{Equação 1}$$

Além de viabilizar a execução de alguns algoritmos referentes aos métodos de seleção de atributos e de classificação de dados, outros motivos para a discretização são o aumento na velocidade computacional e da interpretabilidade dos modelos de classificação gerados [Liu et al. 2002]. Ambos os benefícios costumam ser observados, por exemplo, nos métodos de classificação por árvores de decisões que processam dados previamente discretizados.

Outro processo de transformação entre diferentes tipos de atributos é o de categorização. Para efeito do presente trabalho a categorização é definida como o processo que gera atributos qualitativos ordinais, podendo ser estes oriundos de atributos quantitativos discretos ou contínuos. O processo de categorização é semelhante ao de discretização, havendo somente a diferença de que todas as instâncias pertencentes a cada subintervalo receberão um valor categórico específico e não mais um valor numérico. Juntamente a isto, os valores categóricos recebidos deverão ser possíveis de ordenamento, pertencentes todos a um mesmo tema.

A categorização de um determinado atributo possui vasta aplicação na geração de mapas temáticos. Tais produtos cartográficos necessitam que o atributo quantitativo que será utilizado como tema seja categorizado justamente para que as categorias geradas possuam diferentes representações cartográficas. Para efeito deste trabalho, o atributo que será categorizado para ser utilizado como tema será chamado de atributo-tema.

Como exemplo de produto que passou pelo processo de categorização visando a geração de mapas temáticos, pode-se citar o Mapa Temático de Estimativa de Prevalência de Esquistossomose desenvolvido por Martins-Bedê et al. [2009]. Neste caso, o atributo-tema possui valores quantitativos contínuos percentuais os quais foram

associados a uma das três categorias de prevalência Baixa (0%,5%], Média (5%,15%] ou Alta (15%,100%].

A categorização pode, ainda, ser realizada em dois momentos da construção do mapa temático: na etapa final, onde o mapa já se encontra pronto possuindo todos os valores quantitativos de cada área ou *pixel*; ou na etapa prévia à geração do modelo de classificação do mapa. Quando a categorização é realizada na etapa final, ela primeiramente captura o intervalo de valores do atributo-tema através da análise de todos os valores dos elementos existentes no mapa já construído. Após isto, são construídos os subintervalos aos quais cada elemento será associado. Finalmente a cada subintervalo será designado um nome categórico. Este processo será chamado de *pós-categorização*. Por outro lado, a categorização pode ser realizada na etapa prévia à geração do modelo de classificação do mapa. Neste caso os elementos do mapa temático ainda não possuem valores, sendo necessário definir um modelo de classificação a partir de amostras de treinamento. É sobre estas amostras que é realizado o processo de categorização, passando o mesmo a ser chamado de *pré-categorização*.

O objetivo deste trabalho é de propor um processo de pré-categorização que realize uma busca para maximizar a qualidade na etapa de classificação. Esta busca será feita através da variação nos limites das categorias geradas sobre o atributo-tema. Tal objetivo se mostra de grande importância visto que a qualidade observada no resultado de um processo de classificação influencia diretamente a qualidade temática de mapas que utilizem este tipo de dado na sua construção.

As seções deste trabalho estão assim divididas para facilitar o entendimento: na seção 2 serão apresentados os principais métodos de pré-categorização descritos na literatura; na seção 3 encontra-se descrito o algoritmo do método proposto; na seção 4 está descrita a metodologia que será utilizada para realização dos testes; na seção 5 estão os resultados obtidos com os testes realizados; e na seção 6 estão algumas conclusões e sugestões para trabalhos futuros.

2. Principais Métodos de Pré-categorização

Conforme Dent et al. [2008], os métodos de categorização mais tradicionais são os de *intervalos iguais* e *quantil*. Em Liu et al. [2002], são descritos métodos não supervisionados para categorização já que levam em conta somente os valores do atributo-tema e não levam em consideração os valores dos demais atributos.

No método dos intervalos iguais a categorização é realizada dividindo-se os valores constantes da faixa de domínio do atributo-tema pelo número de classes de interesse. Neste caso serão obtidos subintervalos, os quais serão as categorias, de tamanhos iguais. Sendo K o número de categorias definidas pelo usuário e x_{min} e x_{max}, respectivamente, os valores mínimos e máximos observados no atributo-tema, então o método define categorias com larguras iguais a calculada pela Equação 2.

$$\delta=(x_{max} - x_{min}) / K \qquad \text{Equação 2}$$

Portanto, a categorização pelo método dos intervalos iguais $Cat_{IntIguais}$ sobre o atributo-tema A_{tema} irá gerar o conjunto de categorias conforme a Equação 3.

$$Cat_{IntIguais}(A_{tema}):\{[x_{min}, x_{min}+\delta] ,]x_{min}+\delta , x_{min}+2\delta], \dots ,]x_{min}+(K-1)\delta , x_{max}]\} \qquad \text{Equação 3}$$

No método do quantil, a categorização é realizada dividindo-se o número total

de instâncias *N* pelo número de categorias de interesse *K*, obtendo-se $n = \lfloor N/K \rfloor$ instâncias em cada categoria onde $\lfloor N/K \rfloor$ representa o maior inteiro menor ou igual a *N*/*K*. Portanto, a categorização pelo método do quantil $Cat_{Quantil}$ sobre o atributo-tema A_{tema} cujos valores mínimos e máximos sejam, respectivamente, x_{min} e x_{max} irá gerar o conjunto de categorias conforme a Equação 4.

$$Cat_{Quantil}(A_{tema}):\{[x_{min}, x_n] ,]x_n , x_{2n}], \ldots ,]x_{(K-1)n} , x_{max}]\} \qquad \text{Equação 4}$$

Ao analisar os *divisores de categorias* gerados por ambos os métodos, observa-se que no método dos intervalos iguais estes valores serão os que irão definir os diferentes subintervalos e são calculados por $x_{min}+(j-1)\delta$ com *j* variando de 1 a *K*. Por outro lado, no método do quantil os divisores de categorias serão calculados pela média aritmética entre os valores das instâncias adjacentes e pertencentes a diferentes subintervalos, obtendo-se o valor $x_{(j-1)n}$ onde *j*, mais uma vez, varia de 1 a *K*.

3. Otimizador de Categorização

O algoritmo proposto no presente trabalho tem o nome de Otimizador de Categorização. Sua primeira etapa é de ordenamento do atributo de interesse, isto é, do atributo-tema. Observa-se que o ordenamento ocorrido nesta etapa, dependendo do algoritmo utilizado, poderá gerar custos computacionais desnecessários ao processo de categorização. Algoritmos de ordenamento computacionalmente eficientes, como o *Quicksort* [Theodoridis e Koutroumbas 2006], possuem complexidade de O(*N*log*N*), sendo *N* o número total de instâncias de um atributo-tema.

Após o ordenamento das instâncias conforme os valores do atributo-tema, inicia-se o processo de busca do algoritmo de categorização proposto. Neste caso irá ocorrer uma variação nos valores dos divisores de categorias de tal forma que diferentes subintervalos poderão ser gerados, sendo cada conjunto de subintervalos considerado um *estado de busca* [Russel e Norvig 2004].

Novos estados de busca são então definidos fazendo com que, caso uma instância esteja no limite entre duas categorias, esta instância seja retirada da atual e inserida na adjacente. Outra forma de analisar este mesmo processo é fazendo com que cada um dos (*K*-1) divisores de categorias, sendo *K* o número de categorias, seja deslocado para a esquerda ou para a direita, variando os limites entre categorias adjacentes. A partir do *estado atual* de busca do algoritmo é possível então gerar 2(*K*-1) diferentes novos estados a serem analisados.

O método de busca proposto para otimização do processo de categorização é o de *hill climbing*, utilizando um algoritmo *greedy*. Tal método é vastamente utilizado na área de Inteligência Artificial para buscar valores ótimos locais. No método de *hill climbing* primeiramente é definida uma *função objetivo*, a qual se quer maximizar (ou, dependendo do caso, minimizar). Após esta definição, é verificado o valor desta função no *estado atual*, onde a busca se encontra, e nos *estados vizinhos*. Caso o valor da função objetivo em um dos estados vizinhos seja superior ao do estado atual, o algoritmo realiza um *passo* para o estado vizinho, passando o mesmo a ser o estado atual e reiniciando todo o processo de busca [Russel e Norvig 2004].

Sendo o algoritmo *greedy*, os passos realizados durante a busca sempre serão para o estado vizinho que tiver o maior valor de função objetivo, ignorando os demais estados. Além disso, o algoritmo não realiza mais qualquer passo caso nenhum dos estados vizinhos tenha um valor superior ao do estado atual, encerrando o processo de busca [Theodoridis e Koutroumbas, 2006].

Conforme Russel e Norvig [2004] descrevem, o método de *hill climbing* possui como desvantagem o fato de ficar "preso" em máximos locais e platôs, não conseguindo atingir o máximo global. Entretanto, visando contornar este problema, foi implementado no método proposto o *passo flutuante*, o qual terá o tamanho definido pelo usuário. Tal passo flutuante somente será realizado pelo algoritmo caso o mesmo atinja um valor de máximo buscando verificar se estados mais distantes possuem melhores resultados. Caso o passo flutuante seja realizado, o algoritmo de *hill climbing* volta a realizar passos simples em seguida buscando maximizar a função objetivo.

A função objetivo definida para o algoritmo proposto é o coeficiente de concordância Kappa obtido através da validação do resultado de uma classificação. Observa-se que, como consequência desta função objetivo, para cada estado de busca será realizado um processo de categorização, classificação e validação da classificação para que se obtenha o valor de Kappa. Para maiores detalhes sobre o cálculo para obtenção do coeficiente Kappa sugere-se consultar Congalton e Green [1999].

Como parâmetros de entrada do algoritmo desenvolvido é necessário que o usuário informe: o número de categorias K que deseja por ocasião da busca da categorização ótima; o número mínimo de instâncias por cada categoria; o valor γ do passo flutuante da busca; o endereço da planilha de dados estruturados; se deseja que a busca seja realizada sobre os dados normalizados ou sobre os dados originais; se o estado inicial da busca será obtido pelo método dos intervalos iguais ou do quantil; e se na etapa de validação da classificação o valor do coeficiente de concordância Kappa será calculado baseado nos próprios dados de treinamento, em um processo de validação cruzada 10-*fold* [Theodoridis e Koutroumbas 2006] ou em um processo de validação cruzada 5x2-*fold* [Dietterich 1997].

Observa-se ainda que o número mínimo de instâncias por categoria irá limitar a geração de novos estados a serem analisados. Caso o número de instâncias em uma determinada categoria seja inferior ao informado pelo usuário, aquele todo aquele estado será considerado inválido pelo algoritmo.

O algoritmo de otimização de categorização foi desenvolvido na linguagem de programação JAVA com o objetivo de utilizar as diversas funções disponíveis no programa de mineração de dados WEKA [2010].

Analisando as complexidades de todas as funções envolvidas no algoritmo proposto, chega-se a conclusão de que a complexidade do algoritmo, já com as devidas simplificações, é $O(K^2N^2)$. Neste caso considerou-se que o número de flutuações será proporcional ao número de categorias K.

4. Resultados

Os resultados atingidos pelo algoritmo proposto foram obtidos através de testes utilizando uma tabela estruturada que integra atributos extraídos de dados de sensoriamento remoto do Projeto Radiografia da Amazônia [DSG 2008] com valores de biomassa obtidos por inventários florestais cedidos pelo Instituto Nacional de Pesquisas da Amazônia (INPA). Utilizando estes dados, pretende-se otimizar o processo de categorização necessário à construção de um mapa temático de estimativa de biomassa.

O conjunto de dados de sensoriamento remoto é composto por 140 atributos extraídos de produtos fornecidos pelo Projeto Radiografia da Amazônia. Maiores detalhes encontram-se descritos por Casto-Filho e Santos [2010]. O inventário florestal

foi realizado pelo Instituto Nacional de Pesquisas da Amazônia (INPA) em 2011 no município de São Gabriel da Cachoeira onde 29 parcelas de floresta primária foram inventariadas. Através dos procedimentos descritos por Lima et al. [2012] foi aplicado um modelo alométrico transformando os dados dendrométricos coletados em campo em estimativas de biomassa.

Os parâmetros de entrada do programa são: uma tabela estruturada de dados composta por 140 atributos e 29 instâncias; estados iniciais por intervalos iguais e por quantil; número de categorias igual a 3; número mínimo de instâncias igual a 2; valor de passo flutuante igual a 4 e; métodos de avaliação com amostras de treinamento. Além disto, o método de classificação utilizado foi o de árvore de decisões univariada (ou OBCT - *Ordinary Binary Classification Tree*) conforme descreve Quinlan [1993] para o algoritmo conhecido como C4.5 e que está disponível no WEKA [2010] com o nome de J4.8. A quantidade limitada de testes se deve à limitação do tamanho do artigo.

Visando analisar a capacidade de busca do método proposto, foi também desenvolvido um algoritmo de busca exaustiva o qual avaliará todas as possibilidades do espaço de estados. Através desta busca será possível encontrar o estado de categorização com máximo valor do coeficiente Kappa e verificar a capacidade de busca do algoritmo do método proposto.

A Figura 1 ilustra os resultados obtidos pelo algoritmo de Otimização de Categorização, validada com amostras de treinamento e com estado inicial obtido pelo (a) método do quantil e (b) dos intervalos iguais.

Comparando os resultados obtidos, observa-se que o algoritmo de Otimização de Categorização obteve valor de Kappa máximo (igual a 1,00) somente para o caso em que o estado inicial foi o obtido através do método do quantil. Quando a busca iniciou pelo estado inicial obtido pelo método dos intervalos iguais, apesar de ter realizado 2 passos, não conseguiu atingir o valor máximo de Kappa, atingindo somente um máximo local.

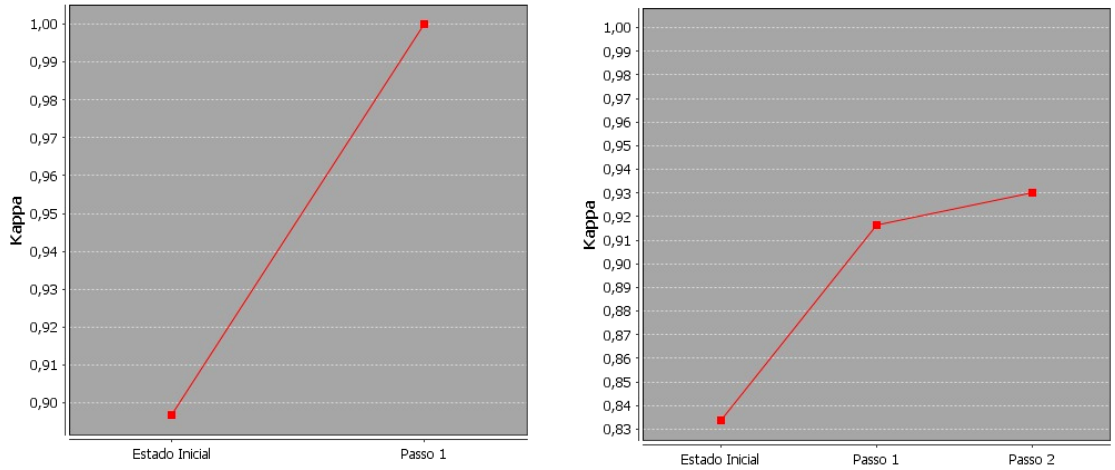

Figura 1. Busca com algoritmo de Otimização de Categorização validado através das próprias amostras de treinamento e tendo como estado inicial o obtido pelo método do (a) Quantil e dos (b) Intervalos Iguais.

Outra comparação realizada foi a referente à busca pelo método exaustivo. Neste caso o algoritmo realizou 300 passos, onde para cada passo foi realizada uma

categorização, uma classificação e um cálculo de Kappa na etapa de avaliação. Embora tenha apresentado alto custo computacional, o resultado obtido pelo método exaustivo obteve o valor de Kappa igual a 1,00.

No processo de Otimização de Categorização para 3 categorias com estado inicial obtido pelo método do quantil, foi encontrado o estado referente aos índices das instâncias de

$$Cat_{OtimizaçãoCategorização}(A_{tema}):\{[1 , 8] , [9 , 18], [19 , 29]\}.$$

O resultado obtido pelo processo de Otimização de Categorização apresentou ainda os valores dos divisores de categorias entre 8 e 9 de 1,01ton/*pixel* e de 18 e 19 de 1,36ton/*pixel*. Tais valores de divisores são essenciais na construção da legenda do mapa temático de estimativa de biomassa que segue a categorização calculada pelo algoritmo proposto.

O modelo de classificação a ser utilizado no mapa temático será o de árvore de decisão OBDT ilustrado na Figura 2. Neste caso as três categorias geradas poderão assumir, posteriormente, outro valor categórico referente à quantidade de biomassa estimada *Baixa*, *Média* e *Alta*.

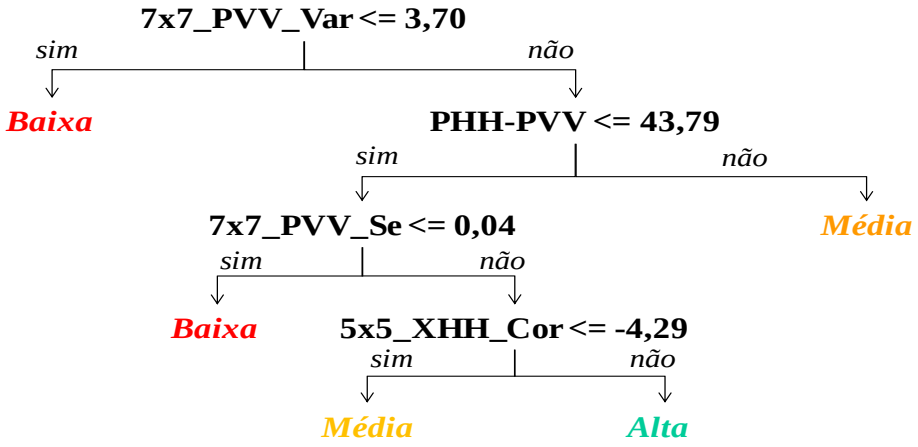

Figura 2. Modelo de classificação por árvore de decisão OBDT.

5. Conclusões

De acordo com os testes realizados, o algoritmo proposto de Otimização de Categorização demonstrou capacidade de realizar passos no espaço de estados obtendo subintervalos de categorias que aumentaram o índice de concordância Kappa. O destaque ocorreu por conta dos testes que obtiveram o estado inicial através do método do quantil que conseguiram resultados iguais aos obtidos pelo algoritmo de busca exaustiva. Entretanto, novos testes são necessários para avaliar os benefícios em iniciar a busca com o estado inicial obtido pelo método do quantil e dos intervalos iguais.

Em trabalhos futuros pretende-se realizar o desenvolvimento de algoritmos para realizar testes de hipótese visando verificar se os valores do coeficiente de concordância Kappa são estatisticamente diferentes e se o método de busca pela Otimização de Categorização realmente apresenta ganho.

Buscando validar o método proposto para construção de diversos tipos de mapas temáticos, futuros testes com outras bases de dados ainda são necessários.

Referências

Bussab, W. O.; Morettin, P. A. (2002), Estatística Básica. 5ª Edição. Editora Saraiva. 526 p. CDD -519.5

Castro-Filho, C. A. P.; Santos, J. R. (2010), Classificação de Imagens PolInSAR Utilizando Técnicas de Mineração de Dados. In Proceedings of IX Seminário de Atualização em Sensoriamento Remoto e Sistemas de Informações Geográficas Aplicados à Engenharia Florestal, Curitiba. IX Seminário de Atualização em Sensoriamento Remoto e Sistemas de Informações Geográficas Aplicados à Engenharia Florestal, 2010.

Congalton, R. G.; Green, K. (1999), Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers. Denver. EUA. 180 p. ISBN 0873719867.

Dent, B.; Torquson, J.; Hodler, T. (2008), Cartography: Thematic Map Design. 6th ed. McGraw-Hill Science. 368 p. ISBN 0072943823.

Dietterich, T. G. (1997), Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Science, p. 1-24,

Diretoria de Serviço Geográfico (DSG). (2008), Infra-estrutura de dados espaciais para a Amazônia. São José dos Campos. Palestra realizada no Encontro de Usuários de Sensoriamento Remoto das Forças Armadas (SERFA), em 2008.

Lima, A. J. N.; Suwa, R.; Henrique, G., et al. (2012), Forest Ecology and Management Allometric models for estimating above- and below-ground biomass in Amazonian forests at São Gabriel da Cachoeira in the upper Rio Negro , Brazil. Forest Ecology and Management, v. 277, p. 163-172. doi: 10.1016/j.foreco.2012.04.028,

Liu, H.; Hussain, F.; Tan, C.; Dash, M. (2002), Discretization: an enabling technique. Data Mining and Knowledge Discovery 6(4), 393–423.

Martins-Bedê, F. T.; Freitas, C. D. C.; Dutra, L. V., et al. (2009), Risk Mapping of Schistosomiasis in Minas Socioeconomic Spatial Data. IEEE Transactions on Geoscience and Remote Sensing, v. 47, n. 11, p. 3899-3908.

Quinlan, J.R**. (**1993), C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

Russel, S.; Norvig, P. (2004), Inteligência Artificial. Rio de Janeiro : Campus, 2004. ISBN 85-352-1177-2.

Theodoridis, S.; Koutroumbas, K. (2006), Pattern Recognition. 3rd ed. Academic Press. 856 p. ISBN 0123695317.

WEKA. Waikato Environment for Knowledge. (2010). Disponível em < http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em 11 Fev 2010.

Yang, Y.; Webb, G. I. (2009), Discretization for naive-Bayes learning : managing discretization bias and variance. Machine Learning, p. 39-74. doi: 10.1007/s10994-008-5083-5.