EMPREGO DE ÁRVORES DE DECISÃO NA CLASSIFICAÇÃO DA COBERTURA DO SOLO URBANO A PARTIR DE IMAGEM ORBITAL DE ALTA RESOLUÇÃO ESPACIAL

Mateus Henrique Pereira Rosseto
Fatec Adamantina - mateus.rosseto@fatec.sp.gov.br

Paulo Roberto da Silva Ruiz
Fatec Adamantina - paulo.ruiz2@fatec.sp.gov.br

1. Introdução

Ambientes urbanos necessitam de melhorias constantes em seu planejamento para atender as necessidades de seus cidadãos em relação a qualidade de vida e a distribuição equitativa de bens e serviços [1].

Nesse sentido, o presente trabalho objetiva explorar as potencialidades e limitações do sensor WorldView-3 (WV-3) para a classificação de alvos urbanos em dois níveis de legenda, o primeiro com 11 e o segundo com 42 classes de cobertura do solo. Para isso, foram utilizados diferentes métodos de classificação de imagens baseados em árvores de decisão. A área de estudo localiza-se no interior de São Paulo, em um setor do campus da Universidade Estadual de Campinas (UNICAMP), a qual contém grande diversidade de materiais de cobertura do solo.

1.1 Objetivos Gerais

Avaliar comparativamente classificações de cobertura do solo urbano utilizando as dezesseis bandas multiespectrais do WV-3 por meio de quatro algoritmos de indução de árvore de decisão. O objetivo é investigar a acurácia das classificações em relação a cada um dos algoritmos, bem como avaliar se há mudanças na generalização de classes abstratas e os atributos utilizados na construção das árvores de decisão.

1.2 Objetivos Específicos

Espera-se que os seguintes objetivos específicos sejam realizados:

A) Avaliar comparativamente as classificações resultantes de cada algoritmo de árvore de decisão utilizado, analisando o comportamento dos atributos selecionados na indução das árvores de decisão e identificar aqueles de melhor desempenho.

B) Avaliar transversalmente o comportamento dos algoritmos de árvore de decisão, identificando as diferenças na seleção de atributos e no desempenho da classificação frente às variações das classes abordadas.

1.3 Justificativa

A hipótese deste trabalho é a seguinte: É possível manter a mesma estrutura hierárquica de classes para a classificação de imagem WV-3 para cada um dos métodos empregados, mantendo-se o mesmo conjunto de dados de entrada?

2. Fundamentação Teórica

Altas resoluções espacial e espectral são necessárias para a discriminação de alvos urbanos [2]. Por meio da resolução espacial é possível identificar as características dos alvos como forma, textura, cor, sombra e contexto, enquanto a resolução espectral permite distinguir as classes de cobertura urbana [3].

O uso de técnicas de mineração de dados permite a execução de análises inteligentes e automatizadas a partir da descoberta de padrões ou regularidades em grandes conjuntos de dados por meio de técnicas matemáticas e diversos tipos de algoritmos, que fazem parte da Descoberta do Conhecimento em Paradigma de bancos de dados (KDD).

Nesse contexto, a utilização de algoritmos de árvores de decisão para a classificação de conjuntos de imagens orbitais torna possível processar grandes conjuntos de dados de forma rápida e eficaz. Também fornece um método intuitivo de análise dos resultados, que são apresentados em gráficos de fácil compreensão, permitindo uma análise das principais características utilizadas para identificar as classes de cobertura do solo.

3. Metodologia

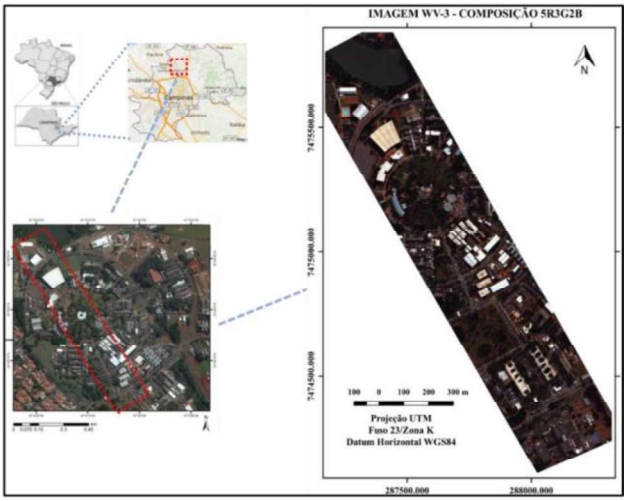
A área de estudo localiza-se na cidade de Campinas – SP, possuindo coordenada central de 22°54'3''S, 47°3'26''W e altitude média de 685 metros (Figura 1). Trata-se de um transecto do campus da Universidade Estadual de Campinas (UNICAMP). Essa área foi escolhida por possuir uma grande diversidade de alvos urbanos.

O trabalho foi realizado a partir de uma imagem orbital do WV-3, obtida em 24 de julho de 2015, com angulação de 6,52° off nadir, 40,6° de elevação solar, 0% de cobertura de nuvens, 30 cm de resolução espacial e 16 bandas multiespectrais. A Figura 2 apresenta a sequência metodológica adotada no trabalho, descrito na sequência. O pré-processamento consiste na conversão dos dados de nível de cinza para radiância, a qual consiste em uma transformação física que obtém o valor da incidência da radiação solar refletida pelos objetos para o espaço, os quais foram captados pelos sensores do satélite. A seguir é realizada a correção atmosférica, que considera o quanto a contaminação da atmosfera, com gases e partículas, afetou os dados captados pelos sensores. Dessa forma, o dado é transformado para a grandeza física de reflectância de superfície. Em seguida é realizada a fusão de bandas, a qual combina a melhor resolução espacial da banda pancromática com as bandas multiespectrais para sintetizar uma nova imagem multiespectral com melhor resolução espacial.

Neste trabalho foi adotada a classificação supervisionada por regiões. Para isso, após o pré-processamento a imagem foi segmentada por meio do algoritmo multirresolução (*Multiresolution Segmentation*), disponível no software eCognition Developer 8.7 [5]. Foram adotadas classes de cobertura urbana em dois níveis de legenda baseado em [4]. O nível de legenda 1 é composto pelas classes: coberturas

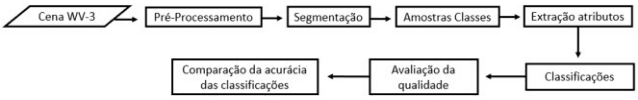
cerâmicas, coberturas metálicas, lago, materiais mistos, pavimentação não viária, pavimentação viária, piscina, solo exposto, sombra, vegetação arbórea e vegetação rasteira. Já o nível de legenda 2 é mais detalhado e possui 42 classes, sendo: acrílico, aço galvanizado brilhante, aço galvanizado com ferrugem, aço galvanizado fosco, aço galvanizado fosco manta, amianto, asfalto, asfalto pintado de branco, asfalto pintado de vermelho, bloquete, carvão coque, cerâmica escura, cerâmica iluminada, cimento pintado de cinza, cimento pintado de verde, cimento pintado de vermelho, concreto, concreto pintado de amarelo, concreto pintado de azul, concreto pintado de vermelho, cimento reforçado com fio sintético (crfs), fibra de vidro, galvalume a, galvalume b, lago, lona azul, lona verde, madeira compensada, manta asfáltica aluminizada, pastilha esmaltada, pedra mineira, pedregulho, piscina de azulejo, piscina de vinil, plástico, policarbonato, seixo argila, solo exposto, sombra, vegetação arbórea, vegetação rasteira e vidro aramado.

Figura 1 – Localização da área de estudo.



Fonte: Ruiz (2017) [4].

Figura 2 – Passos metodológicos adotados neste trabalho.



3.1 Algoritmos de Árvore de Decisão

Os algoritmos de indução de árvores de decisão utilizados neste trabalho estão implementados no software Weka 3.8.6 (*Waikato Environment for Knowledge Analysis*), trata-se de um conhecido software de aprendizado de máquina (*Machine Learning*) escrito em Java, desenvolvido pela Universidade Waikato da Nova Zelândia. Contém uma coleção de ferramentas de visualização e diversos algoritmos para solucionar problemas que demandam mineração e predição de dados [6]. A seguir serão descritos os algoritmos de árvore de decisão utilizados neste trabalho.

3.1.1 Algoritmo J48

Desenvolvido inicialmente por Ross Quinlan [7], pertence à categoria de algoritmos de classificação por árvores de decisão. Caracteriza-se com uma evolução de seus predecessores ID3, C4.5 e C5.0, todos implementados na linguagem C, o que otimiza seu desempenho. Por sua vez, o J48 é implementado na linguagem Java. Uma vantagem da utilização desse algoritmo para a tomada de decisão é que ele se mostra adequado até mesmo para os procedimentos, levando em consideração diferentes tipos de variáveis, as quais podem ser qualitativas, quantitativas, contínuas e discretas. Isso permite a criação de árvores de decisão que categorizam e destacam os atributos mais importantes no conjunto de dados para a identificação das classes definidas no conjunto de treinamento [8].

3.1.2 Algoritmo Random Tree

Random Tree é um algoritmo de classificação disponível no Weka. Em seu escopo, esse algoritmo faz uso do método estocástico, assumindo que os atributos são escolhidos aleatoriamente para cada nó antes de executar a poda para a indução da árvore. Ele também permite a estimativa de probabilidades de classe (ou a mediana de destino em uma regressão) com base em um conjunto de variáveis de retenção (*backfitting*) [9].

3.1.3 Algoritmo Rep Tree

O algoritmo de árvore de decisão *Rep Tree* (*Reduced Error Pruning tree*), em português árvore de poda de erro reduzido, tem por característica induzir uma árvore de decisão de aprendizado rápido. Em sua concepção, o método contempla o uso do ganho/variância de dados e utiliza a poda por meio do cálculo de erro reduzido nó a nó até às folhas, definidoras das classes. O *Rep Tree* constrói diversas árvores em várias iterações usando a lógica de indução regressiva. Depois disso, ele escolhe a melhor árvore para a criação do modelo de classificação. A medida utilizada nos ramos da árvore é o erro quadrático médio nas previsões feitas pelo algoritmo [10].

3.1.4 Algoritmo Random Forest

Random Forest é um algoritmo de classificação que utiliza o método de árvore de decisão de Breiman [11]. Essa técnica tem uma ideia um pouco diferente dos clássicos algoritmos de árvores de decisão, os quais têm como objetivo criar uma estrutura completa a partir de um conjunto de dados. De forma diferente, o *Random Forest* objetiva criar múltiplas árvores de decisão utilizando um subconjunto de atributos escolhidos aleatoriamente a partir do conjunto original. Essas árvores também possuem um tipo de amostragem conhecido como *bootstrap*, sendo um tipo de teste para possibilitar a busca de uma melhor precisão [12].

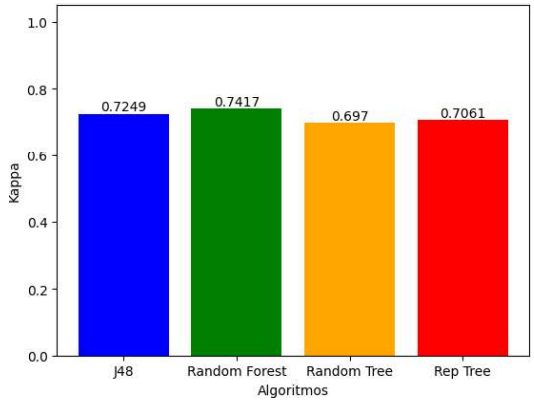
3.2 Avaliação da qualidade da classificação

A avaliação da qualidade das classificações foi realizada por meio de comparações com o mapa de verdade de solo, a partir de Ruiz [4]. Para realizar a avaliação foram gerados pontos aleatórios na área da imagem. A partir deles, foi realizado o cruzamento espacial de dados para obter a classe presente no mapa de verdade de campo e das classificações realizadas. Foram utilizados de 2500 a 3000 pontos aleatórios, variando em cada classificação devido à exclusão do retângulo envolvente da imagem. O resultado é uma tabela de classes associando cada ponto à classe presente no mapa verdade e na classificação. De posse destas tabelas, foram geradas as matrizes de confusão das classificações e calculados os coeficientes *Kappa* e as exatidões globais.

4. Resultados

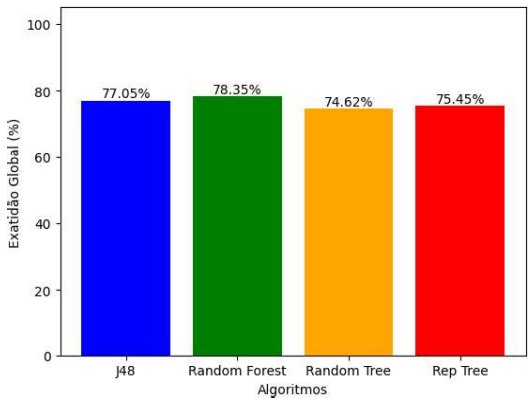
Para a realização das classificações foram extraídos 123 atributos da imagem WV-3. As Figuras 3 e 4 apresentam respectivamente os índices *Kappa* e a exatidão global alcançados pelas classificações no nível de legenda 1 a partir de cada algoritmo avaliado. É possível verificar que o *Random Forest* apresenta as maiores exatidões, seguido pelo algoritmo J48.

Figura 3 – Índices *Kappa* alcançados nas classificações do nível de legenda 1.



F**onte:** Autoria Própria (2023)

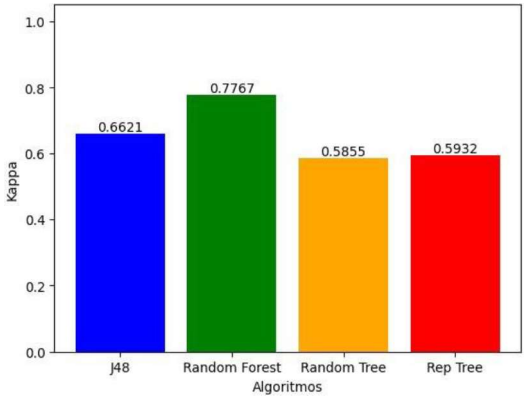
Figura 4 – Exatidões Globais alcançadas nas classificações do nível de legenda 1.



Fonte: Autoria Própria (2023)

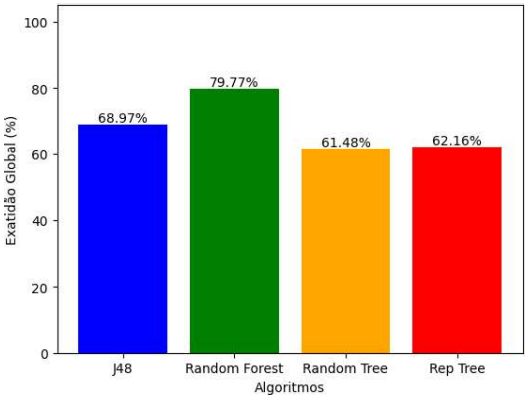
As figuras 5 e 6 apresentam os coeficientes Kappa e as exatidões globais alcançadas pelas classificações no nível de legenda 2 a partir de cada algoritmo avaliado. Da mesma forma que no nível 1, os algoritmos *Random Forest* e J48, apresentam os melhores resultados.

Figura 5 – Índices *Kappa* alcançados nas classificações do nível de legenda 2.



Fonte: Autoria Própria (2023)

Figura 6 – Exatidões Globais alcançadas nas classificações do nível de legenda 2.



Fonte: Autoria Própria (2023)

5. Discussões

Os resultados deste trabalho apresentam as classificações da imagem nos níveis de legenda 1 e 2 para os diferentes algoritmos. Por meio das Figuras 3 e 4 é possível verificar que os algoritmos *Random Forest* e J48 alcançaram os maiores índices global de acerto, respectivamente 78,35% e 77,05%. Por sua vez, os algoritmos *Random Tree* e *Rep Tree* obtiveram os piores resultados, sendo de 74,62% e 75,45% respectivamente.

Na sequência, por meio das figuras 5 e 6 é possível verificar que os mesmos algoritmos, *Random Forest* e J48, alcançaram os índices globais de acerto, respectivamente 79,77% e 68,97%. E, mais uma vez os algoritmos *Random Tree* e *Rep Tree* obtiveram os piores resultados. Cabe destacar a evidente superioridade dos resultados alcançados pelo *Random Forest*, o qual obteve exatidão global no nível 2 próxima de 80%, enquanto os demais algoritmos estão abaixo de 70%.

Os resultados revelaram dificuldades dos algoritmos em selecionar os melhores atributos para a construção das árvores. No caso do *Random Tree* é realizada uma seleção aleatória dos dados e depois utilizado um método estocástico para a realização da poda. Como trata-se da construção de apenas uma árvore, a aleatoriedade na escolha dos nós principais pode levar à seleção de atributos que não sejam os melhores para separar as classes. Já no caso do *Rep Tree*, também está presente o problema da aleatoriedade na escolha dos atributos para os nós da árvore, mas como são criadas diversas árvores e realizado um sistema de escolha, o resultado foi ligeiramente melhor do que aquele apresentado pelo *Random Tree*.

Os algoritmos *Random Tree* e *Rep Tree* demandam maiores estudos objetivando melhorar seus parâmetros para alcançar resultados mais significativos e próximos daqueles alcançados pelo J48 e *Random Forest*. Em ambos os níveis os algoritmos citados anteriormente foram os que apresentaram os menores índices de assertividade.

6. Conclusões

O presente trabalho realizou comparações de classificações da cobertura urbana a partir de um transecto de uma imagem do satélite WorldView-3 utilizando quatro algoritmos de indução de árvore de decisão, sendo eles: o J48, o *Random Forest*, o *Random Tree* e o *Rep Tree* em dois níveis de legenda, possuindo 11 e 42 classes de cobertura do solo urbano.

No nível de legenda 1, contemplando 11 classes de cobertura do solo, o algoritmo *Random Forest*, seguido pelo J48 apresentaram os melhores resultados. Já os algoritmos *Random Tree* e *Rep Tree* apresentaram resultados semelhantes entre si, mas baixos quando comparados com os dois primeiros, esse fato deve-se às suas características em selecionar atributos de modo aleatório para os nós da árvore. Já no nível de legenda 2, compreendendo 42 classes de cobertura de solo, o algoritmo *Random Forest*, novamente seguido pelo J48, apresentaram os melhores resultados. Outra vez os algoritmos *Random Tree* e *Rep Tree* apresentaram resultados muito próximos entre si e bem mais distantes dos outros dois.

Por fim, os resultados deste trabalho possibilitam responder à hipótese levantada na introdução. Sim, é possível manter a mesma estrutura hierárquica de classes para a classificação da imagem WV-3 em cada método empregado, mantendo-se o mesmo conjunto de dados de entrada. Nesse sentido, mesmo com grandes variações entre si, os resultados alcançados apresentam índices *Kappa* categorizados como bom e muito bom, colaborando com a hipótese levantada no escopo deste projeto.

5. Referências

[2] BRASIL. Estatuto da Cidade e Legislação Correlata. Lei nº 10.257, de 10 de junho de 2001 – 2. ed. – Brasília: Senado Federal, Subsecretaria de Edições Técnicas, 2002. 80 p. ISBN 85-7018-223-6.

[3] MYINT, S. W. et. al. Perpixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sensing of Environment, v. 115, n. 5, p. 1145–1161, 2011.

[4] BLASCHKE, T. et. al. Geographic object-based image analysis - towards a new paradigm. ISPRS Journal of Photogrammetry and Remote Sensing: official publication of the International Society for Photogrammetry and Remote Sensing (ISPRS), v. 87, n. 100, p. 180–191, jan. 2014.

[5] RUIZ, P. R. S. Classificação da cobertura do solo urbano usando árvores de decisão a partir de cenas WorldView-2 e WorldView-3 para diferentes níveis de legenda. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2017, 181 p. Disponível em: <http://urlib.net/8JMKD3MGP3W34P/3NJ9GU8>. Acesso em: 28 maio 2022.

[6] TRIMBLE. eCognition developer 8.7 user guide. Munich, Germany: [s.n.], 2011. 258 p. Disponível em: <http://www.ecognition.com/>. Acesso em: 28 maio 2022.

[7] KULKARNI, E. G.; KULKARNI, R. B. WEKA powerful tool in data mining. International Journal of Computer Applications (0975 – 8887). National Seminar on Recent Trends in Data Mining. 2016.

[8] QUINLAN, J. R. Induction of decision trees. Machine Learning, v.1, n. 1, p. 81 – 106, 1986.

[9] LORENZETT, C. C.; TELÖCKEN, A. V. Estudo Comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de Decisão. Revista eletrônica Unicruz, 2016.

[10] FERNANDES, A. C. et. al. Uma proposta para melhoria da previsibilidade de custo de projetos de software utilizando dados históricos de custo, a técnica de gerenciamento de valor agregado – GVA e o algoritmo Random Tree. XIII Brazilian Symposium on Information Systems, Lavras, Minas Gerais, June 5-8, 2017.

[11] WITTEN, I. H. et. al. Data mining: practical machine learning tools and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011. 664p.

[12] BREIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5-32, 2001.

[13] NETO, C. D. G. Potencial de técnicas de mineração de dados para o mapeamento de áreas cafeeiras. INPE, São José dos Campos. 2014.