Técnicas de Mineração de Dados e Aprendizado de Máquina Aplicadas à Dados Ambientais de Uso e Cobertura de Solo

Mariana Albuquerque Reynaud Schaefer[1], Carlos H. T. Brumatti[1], Gustavo V. Veloso[2], Jugurta Lisboa Filho[1], Elpídio Inácio Fernandes Filho[2], Julio C. S. Reis[1]

[1]Departamento de Informática – Universidade Federal de Viçosa (UFV) – Brasil

[2]Departamento de Solos – Universidade Federal de Viçosa (UFV) – Brasil

{mariana.schaefer,carlos.h.tavares}@ufv.br,gustavo.v.veloso@gmail.com,

{jugurta,elpidio,jreis}@ufv.br

Resumo. *Este trabalho descreve o processo de uma Revisão Sistemática da Literatura (RSL) realizada na área de mineração de dados e aprendizado de máquina aplicados à dados ambientais e geográficos com foco em cobertura e uso de solo. As publicações mais relevantes obtidas ao final deste processo são detalhadas, buscando-se elencar direções futuras para pesquisas neste contexto.*

Abstract. *This work describes the Systematic Literature Review (SLR) process in data mining and machine learning applied on environmental and geographic data focused in land use and land cover. The most relevant publications obtained at the end of the review are detailed, in order to list future directions in research.*

1. Introdução

É cada vez mais expressiva a utilização de técnicas de mineração de dados e aprendizado de máquina em aplicações ambientais, como por exemplo, para estudo de características do solo em diferentes ecossistemas [Langford et al. 2019]. O aumento do acesso de usuários de diversas áreas a estas abordagens e a facilidade de acesso à diferentes tipos de dados são fatores que impactam diretamente neste crescimento. Além disso, o poder computacional cada vez maior e a viabilização da análise de grandes quantidades de dados por meio de ferramentas cada vez mais robustas levaram a uma redução significativa do tempo e custo, se comparado a métodos tradicionais de coleta e análise de dados explorados anterioremente [Molinaro and Leal 2018].

Diante deste cenário, cada técnica explorada pode apresentar vantagens e desvantagens comumente associadas ao contexto de aplicação, e estes aspectos devem ser considerados para a escolha da abordagem mais adequada. Ademais, existem outras questões importantes que devem ser avaliadas pelos usuários para aplicação de uma técnica de interesse, como por exemplo, a escolha da métrica adequada, bem como decisões relativas ao processo de obtenção e tratamento dos dados. Considerando dados ambientais e/ou geográficos, por exemplo, existem diversas publicações recentes que apresentam resultados promissores, em diferentes contextos, obtidos a partir da aplicação de técnicas de mineração de dados e aprendizado de máquina. Por exemplo, uma pesquisa recente aplicou uma técnica de aprendizado profundo (do inglês, *deep learning*), em imagens de satélite para criar um mapa global em alta resolução da altura da vegetação, que pode

Tabela 1. *Strings* utilizadas para realização das buscas em cada base de dados.

Base	*String* utilizada
Scopus	(ABS(“environmental”OR “land cover”OR “land-cover”) AND TITLE (“data mining”OR “machine learning”))
ACM	[[Abstract: ”environmental”] OR [Abstract: ”land cover”] OR [Abstract: ”land-cover”]] AND [[Title: ”machine learning”] OR [Title: ”data mining”]]
IEEE Xplore	(“Abstract”:“environmental”OR “Abstract”:“land cover”OR “Abstract”:“land-cover”) AND (“Document Title”:“machine learning”OR “Document Title”:“data mining”)

ser utilizado para mapeamento de desmatamento [Lang et al. 2022], o que ressalta a importância do uso dessas técnicas para a proposição de abordagens computacionais que sejam úteis para resolução de problemas que afetam a sociedade atual.

Assim, neste trabalho, é realizada uma Revisão Sistemática da Literatura (RSL), baseada no protocolo definido por [Kitchenham et al. 2009], com objetivo de identificar as principais publicações com foco na aplicação de técnicas de mineração de dados e aprendizado de máquina em dados do contexto ambiental e/ou geográfico e revelar potenciais lacunas e/ou oportunidades de pesquisas futuras dentro deste contexto. Em suma, a Seção 2 apresenta detalhes relativos à RSL, enquanto a Seção 3 relaciona os principais resultados obtidos a partir das publicações selecionadas, incluindo uma breve descrição das oportunidades de pesquisa identificadas. Por fim, a Seção 4 conclui este trabalho.

2. Metodologia

A RSL proposta por [Kitchenham et al. 2009] é um relatório técnico para guiar pesquisas bibliográficas e vem sendo utilizado em diversas áreas, incluindo ciência da computação [Seufitelli et al. 2021]. Para este trabalho, as etapas executadas a partir de uma adaptação do referido protocolo, são detalhadas a seguir.

Etapa 1: Definição das questões de pesquisa (QPs). Nesta etapa foram definidas quatro QPs exploratórias, relacionadas ao contexto de aplicação de técnicas de mineração de dados e aprendizado de máquina aplicadas à dados ambientais com foco em uso e cobertura do solo: (*i*) Quais técnicas e tipos de dados são explorados nos trabalhos identificados?; (*ii*) Qual o formato mais comum dos dados utilizados como entrada para os modelos?; (*iii*) Quais as principais métricas utilizadas para avaliação do desempenho das abordagens propostas?; e por fim (*iv*) Quais são os principais desafios e oportunidades de pesquisas na área?

Etapa 2: Definição das palavras-chave. Com base nas QPs, foram definidas as palavras-chave descritas a seguir: *data mining*, *machine learning*, *environmental*, *land cover*.

Etapa 3: Definição da *string* de busca. A partir das palavras-chave e das QPs definidas na Etapa 1, a *string* de busca definida inicialmente foi: “(*data mining OR machine learning*) AND (*environmental OR land cover*)”. Após pesquisa preliminar realizada nas bases de busca de interesse, a *string* inicialmente definida foi refinada. A Tabela 1 relaciona a base de dados e as *strings* utilizadas em cada uma delas.

Etapa 4: Definição das bases de busca. Como o foco principal da pesquisa é a área de Ciência da Computação, as bases definidas para pesquisa foram Scopus, ACM e IEEE Xplore[1]. No total, durante esta etapa, 3.315 publicações foram obtidas, sendo 2.886, 399 e 30 oriundas das bases Scopus, IEEE Xplore e ACM, respectivamente.

[1]www.scopus.com, dl.acm.org, ieeexplore.ieee.org

Tabela 2. Classificação das publicações selecionadas.

Classificação				
Referência	**Técnicas utilizadas**	**Validação/métricas**	**Foco principal**	**Fonte de dados**
[Vasilakos et al. 2020]	SVM (*Support Vector Machine*), RF (*Random Forest*), ANN (*Artificial Neural Networks*), DT (*Decision Tree*), LDA (*Linear Discriminant Analysis*), KNN (*K-Nearest Neighbors*)	Kappa, F1, MCC (Coeficiente de Correlação de Matthews)	Cobertura do solo	Sensoriamento remoto
[Liu and Li 2021]	SVM, RF, *Deep Learning* (VGGNET-16 (*Very Deep Convolutional Networks*), RESNET-18 (*Residual Neural Network*))	Kappa, acurácia	Cobertura do solo	Sensoriamento remoto
[Gibril et al. 2018]	DT, RF, SVM	Acurácia	Cobertura e uso de solo	Sensoriamento remoto
[Jamali 2021]	SVM, RF, *Deep Learning* (GAMLP (*Genetic Algorithm Multi-Layer Perceptron*) e FSMLP (*Derivative-free Function Multi-Layer Perceptron*))	Kappa, acurácia	Cobertura e uso de solo	Sensoriamento remoto
[Jamali 2019]	RF, DT, DTNB (*Decision Table with Naive Bayes*), J48, Lazy IBK, *Multi-layer Perceptron* (MLP), NN ge (*Non-nested Generalized Exemplars*)	Acurácia, MAE, RMSE	Cobertura e uso de solo	Sensoriamento remoto
[Keshtkar et al. 2017]	RF, DT, SVM	Kappa, acurácia	Cobertura do solo e mudança espaço-temporal da cobertura	Sensoriamento remoto
Regressão				
Referência	**Técnicas utilizadas**	**Validação/métricas**	**Foco principal**	**Dados**
[Matinfar et al. 2021]	PLSR (*Partial Least Square Regression*), RF, QRF (*Quantile Regression Forest*), CB (*Cubist*), *Fuzzy Logic*, RF-OK (*Ordinary Kriging*), QRF-OK	R2, RMSE	Predição de carbono do solo (SOC)	146 amostras de solo coletadas; Sensoriamento remoto; Dados geomorfométricos
[Naimi et al. 2021]	CB, RF	RMSE, RPIQ, R2	Predição de carbono do solo (SOC), distribuição de partículas do solo e carbonato de cálcio (CCE)	300 amostras de solo coletadas; Imagens sintéticas de solos (SySI); Sensoriamento remoto
[Rostaminia et al. 2021]	RF, CB, RF-OK, CB-OK	R2, RMSE	Predição de carbono do solo (SOC)	80 amostras de solo coletadas; Sensoriamento remoto
Classificação e Regressão				
Referência	**Técnicas utilizadas**	**Validação/métricas**	**Foco principal**	**Dados**
[Arancibia et al. 2021]	ANN, SVM, RF	Acurácia, MAE	Predição de conteúdo de barragem de rejeitos	70 amostras de solo coletadas; Sensoriamento remoto

Etapa 5: Definição de critérios de inclusão e exclusão de dados. Foram excluídas as publicações cujo título não estava de acordo com o contexto e as QPs definidas na Etapa 1. Depois deste filtro, foram mantidos apenas os artigos completos publicados em conferências (*full papers*) e periódicos que estivessem acessíveis, e com data de publicação a partir de 2017. Ao final desta etapa foram mantidas 103 publicações.

Etapa 6: Seleção de publicações. Com base nos primeiros resultados encontrados, os resumos de cada publicação foram analisados para verificar a compatibilidade com o tema explorado, além de acesso ao texto completo destas, sendo excluídas 90 publicações consideradas fora do escopo. Após este processo, foi realizada uma leitura dinâmica das publicações restantes (ou seja, 13), identificando-se alguns temas predominantes, definidos em consenso pelos autores deste trabalho.

Etapa 7: Classificação das publicações. As 13 publicações foram lidas e classificadas. Foram observadas 3 classes principais: Técnicas de Classificação, Técnicas de Regressão e *Surveys*/Estudos. Para a análise proposta neste trabalho, os *surveys*/estudos foram excluídos, por não apresentarem aplicações de métodos, restando 10 publicações, resumidas na Tabela 2, e disponibilizada na seguinte planilha: `https://bityli.com/rIUjMIq`.

3. Análise e Discussões

A partir da análise das publicações relevantes obtidas ao final da RSL, foi possível responder às questões exploratórias de pesquisa (QPs) definidas na Etapa 1.

(*i*) Quais técnicas de mineração de dados e aprendizado de máquina estão sendo aplicadas nos trabalhos identificados? As técnicas mais utilizadas são *Random Forest* (RF), método que foi predominante, seguido de *Support Vector Machine* (SVM) e *Decision Tree*

(DT), aplicadas na maioria das publicações envolvendo problemas de classificação, e *Cubist* (CB), que também é um algoritmo baseado em árvores de decisão, aplicada em todas as publicações com foco em regressão. Variações de algumas destas técnicas, como RF-OK ou CB-OK (*Ordinary Kriging*), também foram aplicadas em parte das publicações [Matinfar et al. 2021, Rostaminia et al. 2021].

Por meio de uma análise mais minuciosa das publicações, alguns pontos em comum são notados, como o bom desempenho geral da técnica SVM, principalmente em [Vasilakos et al. 2020, Jamali 2021, Keshtkar et al. 2017, Arancibia et al. 2021], cujos objetivos são classificação de uso e cobertura de solo. Algumas publicações aplicam técnicas de aprendizado profundo, como *Multi-Layer Perceptron* (MLP) e Redes Neurais Convolucionais (CNN), entre elas VGGNET-16 e RESNET-18 [Liu and Li 2021, Jamali 2021], demonstrando melhores resultados em comparação a resultados obtidos com a aplicação de técnicas convencionais de aprendizado de máquina neste contexto. Esta é uma área a ser melhor explorada na atualidade. Além disso, percebe-se que a aplicação da combinação de vários algoritmos (*ensemble*) é promissora [Vasilakos et al. 2020] e, também, ainda pouco explorada. Logo, investigar a combinação destas estratégias pode ser uma direção promissora para trabalhos futuros.

(*ii*) Qual o formato mais comum dos dados utilizados como entrada para os modelos? Todas as publicações utilizam dados espectrais de sensoriamento remoto, que são dados obtidos em comprimentos de onda de luz visível e infravermelho próximo, que representam características da vegetação, solos, clima e outras. As publicações que utilizam dados de sensoriamento remoto de diferentes janelas de tempo trazem um aprimoramento, pois é possível mesclar os dados espectrais, considerando assim as mudanças temporais que naturalmente ocorrem em determinadas classes [Liu and Li 2021], além da possibilidade de análise das mudanças resultantes de atividades antrópicas [Keshtkar et al. 2017], sendo ainda uma abordagem pouco explorada. Além de mesclar dados de diferentes períodos, também é possível mesclar dados de fontes diferentes, por exemplo, imagens de mais de um satélite, cujos resultados demonstram um aumento da acurácia em relação a dados não otimizados [Liu and Li 2021, Gibril et al. 2018]. Um ponto importante é a análise da diferença entre abordagens baseadas em pixel e baseadas em objeto. Na técnica pixel a pixel, cada pixel da imagem é comparado às classes definidas. Já na técnica de Análise de Imagens Baseadas em Objetos (OBIA), as imagens são segmentadas, sendo descritas por suas características (espectrais, de textura, espaciais, entre outras). Em [Keshtkar et al. 2017], compara-se as duas abordagens, e a técnica baseada em objeto demonstra melhores resultados em comparação à baseada em pixel.

Nas publicações onde foram aplicadas técnicas de regressão, a maioria dos trabalhos utiliza, além das variáveis ambientais (e.g., textura, rugosidade, vegetação, etc), dados de sensoriamento remoto, demostrando que a combinação destes dados produz resultados promissores, com um aumento considerável na acurácia e economia de tempo e custo de pesquisa [Matinfar et al. 2021, Rostaminia et al. 2021, Naimi et al. 2021]. É possível analisar ainda a combinação de outros tipos de dados na aplicação destas técnicas, como, por exemplo, variáveis derivadas do relevo (morfometria), tipos de solos, geologia e clima. Dentro deste contexto, uma área de estudo que se mostra importante é a seleção de variáveis, que é um processo para reduzir a quantidade de variáveis de entrada, com objetivo de reduzir o custo computacional e aumentar o desempenho dos

modelos, e notou-se que algumas técnicas podem ser melhor abordadas e exploradas. Estas publicações exploram diferentes abordagens, como método PCA (Análise de Componentes Principais) [Matinfar et al. 2021] para redução de dimensionalidade, métodos de remoção por variância [Rostaminia et al. 2021] ou por correlação [Naimi et al. 2021].

(*iii*) Quais as principais métricas utilizadas para avaliação do desempenho das abordagens propostas? Nas publicações que exploram tarefas de classificação, a principal métrica utilizada é a acurácia [Liu and Li 2021], seguida de Kappa e F1-*score*. Já nas publicações com técnicas de regressão, as métricas mais utilizadas são R2 (coeficiente de determinação), MAE (erro médio absoluto) e RMSE (raiz quadrada do erro médio). Em [Vasilakos et al. 2020], por exemplo, como as classes são desbalanceadas, o uso da acurácia como métrica foi descartado. Métricas como Kappa, também são utilizadas como complemento à acurácia em algumas publicações [Keshtkar et al. 2017, Jamali 2021]. Na maior parte das publicações, não se aprofunda no critério de escolha da métrica, e este é um ponto que pode ser melhor explorado no futuro.

(*iv*) Quais são as oportunidades de pesquisa na área? Com base na análise das publicações, percebe-se que técnicas mais sofisticadas ainda são pouco exploradas, como aprendizado profundo, ativo (*active learning*) e por reforço. Um ponto importante é que o usuário destas aplicações comumente é um especialista da área ambiental, portanto investigar aspectos da explicabilidade dos modelos pode ser extremamente útil e promissor neste contexto. Além disso, tanto em tarefas de classificação quanto regressão, utiliza-se principalmente dados de sensoriamento remoto. Isto abre possibilidades para explorar outros tipos de dados e técnicas mais sofisticadas, bem como para explorar a aplicação de técnicas no processamento dos dados, antes da aplicação dos modelos, principalmente na fase de segmentação das imagens. O uso de dados geofísicos (gamaespectrometria e magnetometria), dados gerados por sensores do tipo radar e dados termográficos são exemplos de dados poucos explorados na área ambiental. Percebe-se ainda que cada técnica aplicada pode apresentar resultados distintos com diferentes fontes de dados, não sendo possível afirmar a existência de um método estado da arte. Isso abre uma oportunidade de investigação de adequabilidade de métodos para determinadas aplicações e/ou dados.

4. Considerações Finais

Este trabalho conduziu uma RSL baseada no protocolo definido em [Kitchenham et al. 2009], para buscar publicações na área de mineração de dados e aprendizado de máquina aplicados a dados ambientais no contexto de cobertura e uso do solo. Ao final, foram selecionadas 10 publicações, que foram analisadas, buscando-se revelar lacunas que podem ser exploradas em trabalhos futuros. Evoluir neste contexto, tanto explorando novas técnicas, como aperfeiçoando as aplicações existentes, é de extrema importância, pois estas informações podem e devem ser utilizadas para apoiar decisões que impactam diretamente a vida das pessoas e o meio ambiente.

Agradecimentos. Este trabalho foi parcialmente financiado pela FAPEMIG e Funarbe.

Referências

Arancibia, G. V., Bustamante, O. P., Vigneau, G. H., Allende-Cid, H., Fuentelaba, G. S., and Nieto, V. A. (2021). Estimation of moisture content in thickened tailings dams: Machine learning techniques applied to remote sensing images. *IEEE Access*, 9:16988–16998.

Gibril, M. B. A., Idrees, M. O., Yao, K., and Shafri, H. Z. M. (2018). Integrative image segmentation optimization and machine learning approach for high quality land-use and land-cover mapping using multisource remote sensing data. *Journal of Applied Remote Sensing*, 12.

Jamali, A. (2019). Evaluation and comparison of eight machine learning models in land use/land cover mapping using landsat 8 oli: a case study of the northern region of iran. *SN Applied Sciences*, 1.

Jamali, A. (2021). Land use land cover mapping using advanced machine learning classifiers. *Ekologia Bratislava*, 40:286–300.

Keshtkar, H., Voigt, W., and Alizadeh, E. (2017). Land-cover classification and analysis of change using machine-learning classifiers and multi-temporal remote sensing imagery. *Arabian Journal of Geosciences*, 10.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15.

Lang, N., Jetz, W., Schindler, K., and Wegner, J. D. (2022). A high-resolution canopy height model of the earth. *arXiv preprint arXiv:2204.08322*.

Langford, Z. L., Kumar, J., Hoffman, F. M., Breen, A. L., and Iversen, C. M. (2019). Arctic vegetation mapping using unsupervised training datasets and convolutional neural networks. *Remote Sensing*, 11(1):69.

Liu, X. and Li, Y. (2021). Research on classification method of medium resolution remote sensing image based on machine learning. *Lecture Notes in Computer Science*, 12753 LNCS:164–173. deep learning.

Matinfar, H. R., Maghsodi, Z., Mousavi, S. R., and Rahmani, A. (2021). Evaluation and prediction of topsoil organic carbon using machine learning and hybrid models at a field-scale. *Catena*, 202.

Molinaro, C. A. and Leal, A. A. F. (2018). Big data, machine learning and environmental preservation: Technological instruments in defense of the environment. *VEREDAS DO DIREITO*, 15(31):201–224.

Naimi, S., Ayoubi, S., Demattê, J. A. M., Zeraatpisheh, M., Amorim, M. T. A., and Mello, F. (2021). Spatial prediction of soil surface properties in an arid region using synthetic soil image and machine learning. *Geocarto International*.

Rostaminia, M., Rahmani, A., Mousavi, S. R., Taghizadeh-Mehrjardi, R., and Maghsodi, Z. (2021). Spatial prediction of soil organic carbon stocks in an arid rangeland using machine lefarning algorithms. *Environmental Monitoring and Assessment*, 193.

Seufitelli, D. B., Moura, A. F. C., Fernandes, A. C., Siqueira, K. M., Brandão, M. A., and Moro, M. M. (2021). Forense digital e bancos de dados: um survey. In *Simpósio Brasileiro de Bancos de Dados (SBBD)*, pages 307–312. SBC.

Vasilakos, C., Kavroudakis, D., and Georganta, A. (2020). Machine learning classification ensemble of multitemporal sentinel-2 images: The case of a mixed mediterranean ecosystem. *Remote Sensing*, 12.