Estudo de técnicas de mineração de dados e aprendizado de máquina aplicadas a dados ambientais

Omitido

1

Resumo. *Este artigo descreve o processo de Revisão Sistemática da Literatura realizado na área de mineração de dados e aprendizado de máquina aplicados a dados ambientais e geográficos, com foco em classificação de solos. As publicações mais relevantes obtidas ao final deste processo são detalhadas, buscando-se elencar direções futuras para pesquisas nesta área.*

1. Introdução

É cada vez mais expressiva a utilização de técnicas de mineração de dados e aprendizado de máquina em aplicações ambientais. O aumento do acesso de usuários de diversas áreas a estes métodos e a facilidade de acesso a bases de dados são fatores determinantes para este crescimento. Além disso, o aumento do poder computacional na última década e a facilidade de análise de grandes quantidades de dados levaram a uma redução significativa do tempo e custo se comparado a métodos tradicionais de coleta e análise de dados.

Cada técnica terá vantagens e desvantagens para cada caso, e deve-se pesar estas questões no momento da sua escolha. Também existem outras questões importantes que devem ser avaliadas pelos usuários, como a escolha da métrica adequada, a obtenção e a preparação dos dados. Para dados ambientais e/ou geográficos, existem diversas publicações com resultados promissores, e é possível resolver problemas de classificação ou regressão, a depender da necessidade e da técnica escolhida.

Com o objetivo de buscar lacunas e oportunidades de pesquisas futuras que tragam uma contribuição para a área da Ciência da Computação, de forma geral, e para as áreas ambientais relacionadas, foi aplicada a metodologia de Revisão Sistemática da Literatura (RSL) [Kitchenham et al. 2009], que está demonstrada neste trabalho, identificando-se as publicações mais relevantes neste contexto. Estas publicações foram, então, analisadas e descritas de forma resumida.

Este artigo está dividido em 4 seções. A seção 2 discute a metodologia da Revisão Sistemática da Literatura utilizada, descrevendo os passos para chegar até as publicações consideradas mais relevantes. A seção 3 faz uma análise dos resultados obtidos, buscando responder às questões definidas na primeira etapa da RSL, além de uma análise da metodologia e resultados alcançados nas publicações selecionadas, mapeando as oportunidades para trabalhos futuros nesta área de interesse. A seção 4, ao final, traz a conclusão.

2. Metodologia

A Revisão Sistemática da Literatura (RSL), proposta por [Kitchenham et al. 2009] é um relatório técnico para guiar pesquisas bibliográficas. Para este trabalho, as etapas descritas abaixo, adaptadas deste protocolo, foram executadas.

Tabela 1. *Strings* utilizadas

Base	*String* utilizada
Scopus	datamining AND machine learning AND (environmental OR geographic) AND (database OR data)
ACM	[Title: ”machine learning”] AND [Abstract: sig soil water mapping environment geographic] AND [Abstract: data] AND [Title: geographic]
G. Scholar	allintitle: ”machine learning”soil OR geographic OR SIG OR vegetation

2.1. Definição das questões de pesquisa

Nesta fase, foi definida a questão principal de pesquisa: ”Quais técnicas de aprendizado de máquina estão sendo aplicadas a análise de dados ambientais e/ou geográficos?”.

De modo mais geral, outras questões de pesquisa foram definidas, como:

”Quais tipos de dados são explorados e aplicados às técnicas?”

”Quais as principais métricas utilizadas para avaliação do desempenho das abordagens propostas?”

”Quais são os principais desafios e oportunidades de pesquisas na área?”

2.2. Definição das palavras-chave

Com base na questão inicialmente definida, foram selecionadas as seguintes palavras-chave: *data mining*, *machine learning*, *environmental data*, *geographic data*.

2.3. Definição da *string* de busca

A partir das palavras-chave e da questão, a *string* de busca definida inicialmente foi: (datamining AND machine learning AND (environmental OR geographic) AND (database OR data)).

Após a pesquisa nas bases de busca de interesse, a *string* inicialmente definida foi refinada. A Tabela 1 especifica a base de dados e as *strings* utilizadas em cada uma delas.

2.4. 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[1] e ACM[2]. Além disso, foi realizada uma busca no Google Scholar[3] para garantir se existia alguma publicação diferente das encontradas, dentre as mais relevantes. No total, 2001 publicações foram obtidas.

[1] www.scopus.com
[2] dl.acm.org
[3] scholar.google.com

Tabela 2. Classificação das publicações

Classificação				
Referência	**Técnicas utilizadas**	**Validação**	**Objetivo**	**Dados**
[Vasilakos et al. 2020]	SVM, RF, ANN, DT, LDA, KNN	Kappa, F1, MCC (Coeficiente de Correlação de Matthews)	Cobertura do solo	Sensoriamento remoto
[Liu and Li 2021]	SVM, RF, *deep learning* (VGGNET-16, RESNET-18)	Kappa, acurácia	Cobertura do solo	Sensoriamento remoto
[Gibril et al. 2018]	Decision Tree, RF, SVM	Acurácia	Cobertura e uso de solo	Sensoriamento remoto
[Jamali 2021]	SVM, RF, deep learning (GAMLP e FSMLP, baseados em *Multi-Layer Perceptron*)	Kappa, acurácia	Cobertura e uso de solo	Sensoriamento remoto
[Jamali 2019]	RF, DT, DTNB, J48, Lazy IBK, *Multi-layer Perceptron*, NN ge	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**	**Objetivo**	**Dados**
[Matinfar et al. 2021]	PLSR (estatístico), RF, QRF, CB, *fuzzy logic*, RF-OK, QRF-OK	R2, RMSE	Predição de carbono do solo (SOC)	146 amostras 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 coletadas; Syntetic Soil Images (SySI); Sensoriamento remoto
[Rostaminia et al. 2021]	RF, CB, RF-OK, CB-OK	R2, RMSE	Predição de carbono do solo (SOC)	80 amostras coletadas; Sensoriamento remoto

2.5. Definição de critérios de exclusão de dados

Foram excluídas as publicações cujo título não estava de acordo com as questões definidas e após este filtro, foram mantidos apenas os artigos de conferências e periódicos, com data de publicação a partir de 2015. Com isto, 100 publicações restaram.

2.6. Seleção de publicações

Com base nos primeiros resultados encontrados, os resumos de cada publicação foram analisados para verificar a compatibilidade com o tema desejado, além de acesso ao texto completo destas, sendo excluídas 88 publicações fora do escopo. Após este processo, foi feita uma leitura dinâmica das 12 publicações restantes, identificando-se alguns temas predominantes, apresentados a seguir.

2.7. Classificação das publicações

As 12 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* foram excluídos, por não serem aplicações de métodos, restando 9 publicações, resumidas na Tabela 2.

3. Análise e discussão

Para analisar as publicações obtidas ao final da Revisão Sistemática, é necessário responder às questões definidas na primeira etapa.

a. Quais técnicas de aprendizado de máquina estão sendo aplicadas à análise de dados ambientais e/ou geográficos?

As técnicas mais utilizadas nas 10 publicações são *Random Forest* (RF), aplicada em todas elas, *Support Vector Machine* (SVM) e *Decision Tree* (DT), aplicadas na maioria das publicações de classificação, e *Cubist* (CB), que também é um algoritmo do tipo árvore de decisão, aplicada em todas as publicações de regressão. Variações de algumas destas técnicas também foram aplicadas em grande parte das publicações.

Realizando uma análise mais minuciosa das publicações, alguns pontos em comum são notados, como o bom desempenho geral da técnica *Support Vector Machine* (SVM), principalmente em [Vasilakos et al. 2020, Jamali 2021, Keshtkar et al. 2017], cujos objetivos são classificação de uso e cobertura de solo (LULC). Algumas publicações aplicam técnicas de *deep learning*, como *Multi-Layer Perceptron* (MLP) e Redes Neurais Convolucionais (CNN), como VGGNET-16 e RESNET-18 [Liu and Li 2021, Jamali 2021], demonstrando melhores resultados do que as técnicas convencionais de *machine learning*. Esta é uma área a ser melhor explorada na atualidade. Além disso, a aplicação da combinação de vários algoritmos (*ensemble*) é promissora [Vasilakos et al. 2020] e, também, ainda pouco explorada. Uma combinação destas aplicações é um bom caminho para análise.

b. 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 períodos diferentes 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, topológicas, entre outras). Em [Keshtkar et al. 2017], compara-se as duas abordagens, e a técnica baseada em objeto demonstra melhores resultados que a baseada em pixel.

Nas publicações onde foram aplicadas técnicas de regressão, a maior parte utiliza, além das variáveis ambientais (como textura, rugosidade, vegetação, entre outros), dados de sensoriamento remoto, demostrando que a combinação destes dados produz resultados satisfatórios, 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 a combinação de outros tipos de dados na aplicação destas técnicas, como por exemplo variáveis derivadas do relevo (morfometria), tipos de solos, geologia,

clima e outras. Uma área de estudo que se mostra importante é a seleção de covariáveis, com algumas técnicas que podem ser melhor abordadas e exploradas. Estas publicações utilizam diferentes abordagens, como método PCA [Matinfar et al. 2021], métodos de remoção por variância [Rostaminia et al. 2021] ou por correlação [Naimi et al. 2021].

c. Quais as principais métricas utilizadas para avaliação do desempenho das abordagens propostas?

Nas publicações com técnicas de classificação, as métricas mais utilizadas são Acurácia, Kappa e F1-*score*, principalmente a primeira delas. Já nas publicações com técnicas de regressão, as métricas mais utilizadas são R2, MAE (erro médio absoluto) e RMSE (raiz quadrada do erro médio). Em [Vasilakos et al. 2020], como as classes são desbalanceadas, o uso da acurácia como métrica foi descartado. Já em [Liu and Li 2021], apesar de ocorrer o mesmo desbalanceamento entre as classes, a acurácia é a principal métrica utilizada. As outras publicações de classificação, além da acurácia, utilizam outras métricas como complemento [Jamali 2019, Jamali 2021, Keshtkar et al. 2017, Gibril et al. 2018]. Na maior parte das publicações, não se aprofunda no critério de escolha da métrica, e isto é um ponto que pode ser melhor explorado.

d. Quais são as oportunidades de pesquisa na área?

Com base na análise das publicações, percebe-se que técnicas mais sofisticadas de *machine learning* são pouco exploradas, como *deep learning*, *active learning* e aprendizado por reforço. Um ponto importante é que o usuário destas aplicações normalmente será da área ambiental, portanto a explicabilidade dos modelos é algo extremamente importante.

Além disso, os principais dados utilizados, tanto para classificação quanto para regressão, são dados de sensoriamento remoto. Isto abre possibilidades para explorar outros tipos de dados e técnicas mais sofisticadas. 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 que cada técnica aplicada pode apresentar resultados distintos com diferentes fontes de dados, não sendo possível afirmar a existência de um método melhor que outro. Isso abre uma oportunidade de investigação de adequabilidade de métodos para determinadas aplicações e/ou dados.

4. Conclusão

Este trabalho aplicou a Revisão Sistemática da Literatura [Kitchenham et al. 2009] para buscar publicações na área de mineração de dados e aprendizado de máquina aplicados a dados ambientais. Foram selecionados 9 publicações, que ao fim, são analisadas, buscando-se caminhos a seguir em trabalhos futuros. Evoluir neste contexto, tanto explorando novas técnicas, como aperfeiçoando as aplicações, é de extrema importância, pois estas informações podem e devem ser utilizadas para apoiar decisões impactam diretamente a vida das pessoas e o meio ambiente.

Referências

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. cited By 13.

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. cited By 15.

Jamali, A. (2021). Land use land cover mapping using advanced machine learning classifiers. *Ekologia Bratislava*, 40:286–300. Support Vector Machine (SVM) and Random Forest (RF)¡br/¿classificação¡br/¿algoritmos deep learning para otimização de parÂmetros.

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. cited By 36.

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.

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 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 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. cited By 3.

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*. cited By 1.

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. cited By 0.

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. cited By 12.