XXIV SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

DETECÇÃO DE SECAS E VISUALIZAÇÃO DE PADRÕES CLIMÁTICOS COM APRENDIZADO DE MÁQUINA

Taís Maria Nunes Carvalho [1] *; Francisco de Assis de Souza Filho* [2] *& Tereza Margarida Xavier de Melo Lopes* [3]

Palavras-Chave – aprendizado de máquina, detecção de secas, máquinas de vetores de suporte.

RESUMO

As mudanças climáticas têm sérias implicações para a ocorrência de eventos climáticos extremos, com efeitos sociais, econômicos e ambientais significativos. A previsão de eventos extremos de seca é fundamental para o planejamento eficiente dos recursos hídricos, assim como para a adoção de medidas preventivas. Além disso, compreender os padrões climáticos que resultam em anos secos pode ser importante para os tomadores de decisão. Nesse trabalho, propomos um modelo de classificação de secas utilizando a técnica de aprendizado de máquina de máquinas de vetores de suporte. O modelo foi aplicado para o estado do Ceará, frequentemente atingido por secas longas e severas. As variáveis explicativas consistem em dados globais de temperatura em ponto de grade, o que permitiu identificar as regiões que mais influenciam a ocorrência de secas no Ceará. O modelo apresentou uma acurácia de 77%, ou seja, dos 61 anos avaliados, 47 foram classificados corretamente como anos secos ou não secos. A avaliação dos pesos atribuídos pelo modelo confirma a influência da temperatura da superfície do Oceano Pacífico e do Atlântico Sul sobre o regime de secas no Ceará.

ABSTRACT

Climate change has serious implications for the occurrence of extreme events, with significant social, economic and environmental effects. Predicting drought events is essential for efficient planning of water resources, as well as for the adoption of preventive measures. Also, understanding the climatic patterns that result in dry years can be important for decision makers. In this study, we propose a drought classification model using the machine learning technique called support vector machine. The model was applied to the state of Ceará, which is often affected with long and severe droughts. The explanatory variables consist of global gridded temperature data, which allowed the identification of the regions that most influence the occurrence of droughts in Ceará. The model presented an accuracy of 77%, i.e., 47 out of the 61 years evaluated by the model were correctly classified as dry or non-dry years. The evaluation of the weights attributed by the model confirms the influence of the surface temperature of the Pacific Ocean and the South Atlantic on the drought regime in Ceará.

[1] Universidade Federal do Ceará, Campus do Pici, taismarianc@gmail.com
[2] Universidade Federal do Ceará, Campus do Pici, assis@ufc.br
[3] Universidade Federal do Ceará, Campus do Pici, terezamelo@alu.ufc.br

INTRODUÇÃO

As mudanças climáticas têm sérias implicações para a ocorrência de eventos climáticos extremos, com efeitos sociais, econômicos e ambientais significativos (ORLOWSKY; SENEVIRATNE, 2012). Em regiões com elevada variabilidade climática, esses efeitos podem ser ainda mais sérios, impondo um desafio para a gestão de recursos hídricos (DA SILVA *et al.*, 2021; GONDIM *et al.*, 2018). Nesse contexto, identificar e compreender a ocorrência eventos extremos de seca é fundamental para garantir a segurança hídrica e evitar a ocorrência de conflitos pelo uso da água.

A previsão de vazões a partir de variáveis climáticas é uma ferramenta útil para a identificação antecipada de secas. Entretanto, esse problema também pode ser explorado como uma tarefa de classificação (DANANDEH MEHR, 2021; KUMAR VIDYARTHI; JAIN, 2020; PAULO; PEREIRA, 2007; SIGAROODI *et al.*, 2014). Essa abordagem pode ser interessante para a extração de conhecimento sobre a formação de eventos de seca ou para a identificação de secas em cenários de mudanças climáticas.

Técnicas de aprendizado de máquina tem ganhado espaço em aplicações da área de recursos hídricos (NUNES CARVALHO; DE SOUZA FILHO; PORTO, 2021; SHEN, 2018; TAHMASEBI *et al.*, 2020; TYRALIS; PAPACHARALAMPOUS; LANGOUSIS, 2019). Um ponto negativo dessas técnicas é a sua natureza "caixa-preta", que muitas vezes dificulta a interpretação dos modelos e dos resultados obtidos a partir da sua aplicação. Entretanto, diversos estudos têm explorado esse problema, e algumas estratégias já foram propostas para superá-lo, como o uso de gráficos de dependência parcial (do inglês, partial dependence plots), gráficos de efeitos locais acumulados (do inglês, accumulated local effects) e o uso de modelos lineares generalizados para explicar a relação entre variáveis (APLEY; ZHU, 2020). Para uma revisão completa a respeito dos métodos de interpretação de modelos de aprendizado de máquina, ver Molnar (2019) e Biecek and Burzykowski (2021).

Nesse trabalho, propomos um modelo de classificação de secas usando máquinas de vetores de suporte. A previsão é feita a partir de dados de temperatura da superfície terrestre e do mar em ponto de grade, o que possibilita a análise das regiões que tem maior influência na ocorrência de secas na área de estudo, em uma abordagem semelhante à adotada em Barnes et al. (2019). A metodologia proposta foi aplicada para o estado do Ceará, historicamente atingido por secas longas e severas. O estado possui registros de seca que datam do período colonial (CAMPOS, 2015), e mais recentemente, entre 2012 e 2018, registrou a seca com o maior período de retorno médio bivariado da sua história (PONTES FILHO *et al.*, 2020).

METODOLOGIA

Dados

Os dados de temperatura da superfície e do mar foram extraídos a partir das bases em ponto de grade Berkeley Earth Surface Temperature (BEST) da Berkeley Earth (ROHDE *et al.*, 2013). Antes do século XX, a cobertura de dados era insuficiente; portanto, o período de análise foi limitado aos anos de 1957 a 2018, quando há cobertura global completa. Os eventos de seca foram identificados a partir de registros históricos das secas no nordeste semiárido brasileiro (Tabela 1).

Tabela 1 – Registros de seca no Ceará. Fonte: Adaptado de Souza Filho (2003).

	Registros de seca
Século XX	1951/1953, 1958
	1966
	1970
	1979/1984
	1993, 1997
Século XXI	2001, 2007/2008
	2012/2018

O conjunto de dados contém 61 observações e 64,419 variáveis explicativas, correspondentes aos pontos de grade que compõe a base BEST. Os dados foram padronizados para que tivessem média 0 e desvio padrão 1.

Modelo de classificação

O problema de detecção de secas foi formulado como uma tarefa de classificação binária, onde um ano pode ter (1) ou não (0) um evento de seca. As variáveis explicativas correspondem a média da temperatura dos meses de janeiro, fevereiro e março para os pontos de grade da base de dados. Esses meses correspondem ao início da quadra chuvosa no Ceará, influenciando na ocorrência ou não de um ano seco.

O modelo foi validado a partir da técnica de validação cruzada leave-one-out. Nesse método, o algoritmo de aprendizado é aplicado uma vez para cada instância do conjunto de dados: o modelo é treinado para *n-1* observações e é testado na restante, onde *n* é o número de observações. A resposta do modelo durante o teste é registrada, e o procedimento é repetido até que se tenha *n* previsões, a partir das quais se avalia o desempenho do modelo.

Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (do inglês, Support Vector Machine – SVM), são amplamente utilizadas para problemas de classificação, embora também possam ser aplicadas para regressão (BOSER; GUYON; VAPNIK, 1992). No SVM, o objetivo principal é encontrar um hiperplano que se ajuste aos dados de treinamento, minimizando a norma euclidiana do vetor de coeficientes. Este modelo usa uma função kernel para mapear dados de entrada para espaços dimensionais superiores, onde possam ser linearmente separáveis. Caso a transformação não seja necessária, utiliza-se um kernel linear. Em problemas de regressão, uma "margem" simétrica é adicionada em torno da função estimada, onde os erros absolutos devem ser iguais ou menores que o erro máximo ε (Awad e Khanna 2015).

$$min_{\beta,\beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\xi_i \quad (1)$$

Sujeito a:

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i| \quad (2)$$

Onde C é o parâmetro de custo e ξ é a variável de folga e corresponde à distância tolerável de outliers da margem. Nós testamos o modelo utilizando tanto um kernel linear quanto um kernel radial, descrito a partir da seguinte equação:

$$K_{RBF}(\boldsymbol{x}, \boldsymbol{x}') = e^{-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^2} \quad (3)$$

Onde x e x' são amostras dos dados de entrada e γ é um parâmetro relacionado à variância da função. Este parâmetro foi definido como o inverso do tamanho dos dados de treinamento. Para avaliar as regiões que tem maior influência sobre a ocorrência de secas na área de estudo, foram extraídos os pesos atribuídos aos pontos de grade pelo modelo. Essa análise foi feita para a classificação com o kernel linear.

Métricas de desempenho

Para avaliar o desempenho do modelo de classificação, além da análise da matriz de confusão, foram utilizadas as seguintes métricas: acurácia, precisão, recall e F1, conforme as equações abaixo:

$$Acurácia = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{FP} + \mathrm{FN} + \mathrm{VN}} \quad (4)$$

$$Precisão = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \quad (5)$$

$$Recall = \frac{VP}{VP + FN} \quad (6)$$

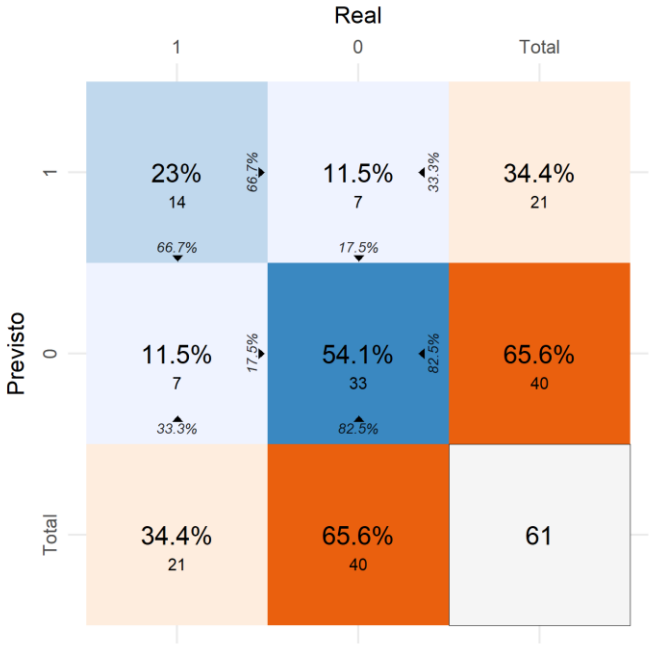
$$F1 = 2 * \frac{precisão * recall}{precisão + recall} \quad (7)$$

As métricas se baseiam na ocorrência de verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN). A métrica F1 é calculada a partir da precisão e do recall.

RESULTADOS

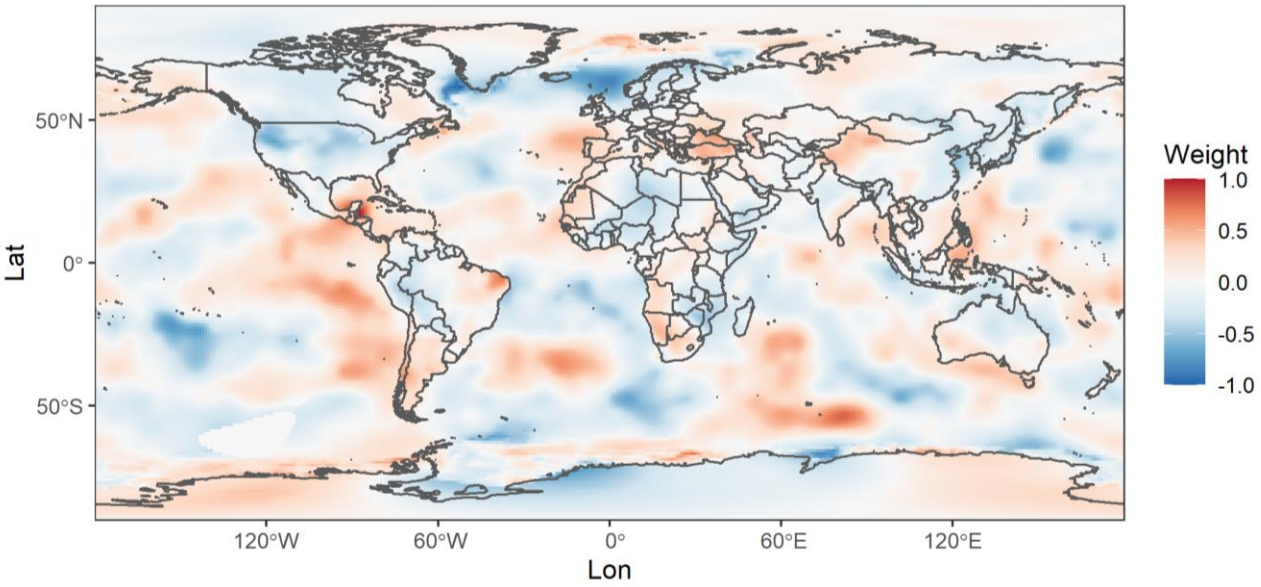
O desempenho do modelo de classificação (kernel linear) pode ser inicialmente avaliado pela matriz de confusão (Figura 1). Assim, os verdadeiros positivos ocorrem quando o modelo classifica um ano seco corretamente (canto superior esquerdo), e um verdadeiro negativo, quando classifica um ano sem evento de seca corretamente (canto inferior direito). Um falso positivo acontece quando um ano é classificado como seco quando de fato não foi (canto superior direito), e um falto negativo, quando o ano é classificado como não seco quando na realidade, foi um ano seco (canto inferior esquerdo). A partir da matriz de confusão pode-se estimar as métricas de desempenho do modelo, listadas a seguir: acurácia: 77%, precisão: 67%, recall: 50%, F1: 61%. O modelo SVM com kernel radial apresentou desempenho inferior ao modelo com kernel linear.

Figura 1 – Matriz de confusão do modelo de classificação de secas.



O desempenho do modelo foi satisfatório, a julgar pela acurácia obtida. A medida de precisão, importante quando desejamos ter certeza da classe apontada pelo modelo, indica que alguns anos que deveriam ter sido classificados como secos, não foram.

Figura 2 – Pesos atribuídos a cada pixel pelo algoritmo SVM.



A análise dos pesos atribuídos pelo modelo SVM aos pontos de grade indica que a temperatura da superfície do Oceano Pacífico e do Atlântico Sul tem forte influência sobre o regime de secas no Ceará – devido a influência do El Niño-Oscilação do Sul no regime de chuvas do Ceará. Os resultados são coerentes com outros estudos a respeito dos padrões climáticos de seca no Ceará e no Nordeste (HASTENRATH; HELLER, 1977; MARKHAM; MCLAIN, 1977). Outras regiões também tiveram pesos significativos no modelo, como o Atlântico Norte e uma porção do Oceano Índico.

CONCLUSÃO

Nesse trabalho, propomos um modelo de classificação para a identificação de anos secos no Ceará a partir de dados globais de temperatura da superfície terrestre e do mar. O modelo apresentou desempenho satisfatório, com uma acurácia de 77%. Apesar de ser desejável que um modelo de identificação de secas seja o mais preciso possível, outras variáveis explicativas precisariam ser consideradas para que o desempenho fosse melhor.

Uma vantagem do modelo apresentado é a possibilidade de identificar eventos de secas futuros apenas com projeções climáticas de temperatura. Além disso, também foi apresentada uma estratégia de interpretação do modelo desenvolvido, com a identificação dos padrões climáticos da seca no Ceará.

REFERÊNCIAS

APLEY, D. W.; ZHU, J. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, v. 82, n. 4, p. 1059–1086, 1 set. 2020. Disponível em: <https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssb.12377>. Acesso em: 30 jun. 2021.

BARNES, E. A. *et al.* Viewing Forced Climate Patterns Through an AI Lens. *Geophysical Research Letters*, v. 46, n. 22, p. 13389–13398, 28 nov. 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1029/2019GL084944>. Acesso em: 4 jul. 2020.

BIECEK, P.; BURZYKOWSKI, T. *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. [S.l.]: CRC Press, 2021.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. Training algorithm for optimal margin classifiers. 1992, [S.l.]: Publ by ACM, 1992. p. 144–152.

CAMPOS, J. N. B. Paradigms and Public Policies on Drought in Northeast Brazil: A Historical Perspective. *Environmental Management*, v. 55, n. 5, p. 1052–1063, 1 maio 2015. Disponível em: <www.brasiliana.usp.br.>. Acesso em: 30 jun. 2021.

DA SILVA, M. V. M. *et al.* Projection of Climate Change and Consumptive Demands Projections Impacts on Hydropower Generation in the São Francisco River Basin, Brazil. *Water*, v. 13, n. 3, p. 332, 29 jan. 2021. Disponível em: <https://www.mdpi.com/2073-4441/13/3/332>. Acesso em: 4 fev. 2021.

DANANDEH MEHR, A. Drought classification using gradient boosting decision tree. *Acta Geophysica*, v. 1, p. 3, 24 abr. 2021. Disponível em: <https://doi.org/10.1007/s11600-021-00584-8>. Acesso em: 30 jun. 2021.

GONDIM, R. *et al.* Climate change impacts on water demand and availability using CMIP5 models in the Jaguaribe basin, semi-arid Brazil. *Environmental Earth Sciences*, v. 77, n. 15, p. 0, 2018. Disponível em: <http://dx.doi.org/10.1007/s12665-018-7723-9>.

HASTENRATH, S.; HELLER, L. Dynamics of climatic hazards in northeast Brazil. *Quarterly Journal of the Royal Meteorological Society*, v. 103, n. 435, p. 77–92, 1 jan. 1977. Disponível em: <https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.49710343505>. Acesso em: 14 jun.

2021.

KUMAR VIDYARTHI, V.; JAIN, A. Knowledge Extraction from Trained ANN Drought Classification Model. *Journal of Hydrology*, p. 124804, 7 mar. 2020. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S002216942030264X>. Acesso em: 8 mar. 2020.

MARKHAM, C. G.; MCLAIN, D. R. *Sea surface temperature related to rain in Ceará, north-eastern Brazil [6]*. *Nature*. [S.l: s.n.]. , 1977
MOLNAR, C. *Interpretable Machine Learning*. [S.l: s.n.], 2019. Disponível em: <https://christophm.github.io/interpretable-ml-book/>. Acesso em: 30 jun. 2021.

NUNES CARVALHO, T. M.; DE SOUZA FILHO, F. DE A.; PORTO, V. C. Urban Water Demand Modeling Using Machine Learning Techniques: Case Study of Fortaleza, Brazil. *Journal of Water Resources Planning and Management*, v. 147, n. 1, p. 05020026, 31 jan. 2021. Disponível em: <https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29WR.1943-5452.0001310>. Acesso em: 19 jan. 2021.

ORLOWSKY, B.; SENEVIRATNE, S. I. Global changes in extreme events: Regional and seasonal dimension. *Climatic Change*, v. 110, n. 3–4, p. 669–696, 22 fev. 2012. Disponível em: <http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php>. Acesso em: 18 jan. 2021.

PAULO, A. A.; PEREIRA, L. S. Prediction of SPI drought class transitions using Markov chains. *Water Resources Management*, v. 21, n. 10, p. 1813–1827, 6 out. 2007. Disponível em: <https://link.springer.com/article/10.1007/s11269-006-9129-9>. Acesso em: 30 jun. 2021.

PONTES FILHO, J. D. *et al.* Copula-Based Multivariate Frequency Analysis of the 2012–2018 Drought in Northeast Brazil. *Water*, v. 12, n. 3, p. 834, 16 mar. 2020. Disponível em: <https://www.mdpi.com/2073-4441/12/3/834>. Acesso em: 23 jun. 2020.

ROHDE, R. *et al.* Geoinfor Geostat: An Overview. v. 1, p. 1, 2013. Disponível em: <http://dx.doi.org/10.4172/2327-4581.1000101>. Acesso em: 30 jun. 2021.

SHEN, C. *A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists*. *Water Resources Research*. [S.l.]: Blackwell Publishing Ltd. Disponível em: <https://doi.org/10.1029/>. Acesso em: 30 jun. 2021. , 1 nov. 2018

SIGAROODI, S. K. *et al.* Long-term precipitation forecast for drought relief using atmospheric circulation factors: A study on the Maharloo basin in Iran. *Hydrology and Earth System Sciences*, v. 18, n. 5, p. 1995–2006, 27 maio 2014.

SOUZA FILHO, F. DE A. Variabilidade e Mudança Climática nos Semi-Áridos Brasileiros. In: TUCCI, C. E. M.; BRAGA, B. (Org.). . *Clima e recursos Hídricos no Brasil*. [S.l: s.n.], 2003. .
TAHMASEBI, P. *et al. Machine learning in geo- and environmental sciences: From small to large scale*. *Advances in Water Resources*. [S.l.]: Elsevier Ltd. , 1 ago. 2020

TYRALIS, H.; PAPACHARALAMPOUS, G.; LANGOUSIS, A. *A brief review of random forests for water scientists and practitioners and their recent history in water resources*. *Water (Switzerland)*. [S.l.]: MDPI AG. Disponível em: <www.mdpi.com/journal/water>. Acesso em: 30 jun. 2021. , 1 maio 2019