Uma avaliação de algoritmos de regressão para predição de volume de chuva

Guilherme S. E. Ferreira e Dianne S. V. Medeiros

***Resumo*— A ocorrência de eventos extremos que provocam catástrofes naturais é de difícil previsão devido à natureza caótica desses eventos. Uma primeira etapa para prever esses eventos com precisão no tempo é construir modelos capazes de prever o nível de chuva em janelas de tempo curtas. Este artigo avalia o desempenho de 10 algoritmos de regressão a fim de identificar aqueles que geram modelos assertivos para prever o nível de precipitação a cada hora. São analisados o coeficiente de determinação e a raiz do erro quadrático médio, calculados para cada modelo obtido. Os resultados mostram que os algoritmos floresta aleatória e regressão polinomial de grau 3 alcançam os menores erros médios, sendo candidatos potenciais para previsão de chuvas.**

***Palavras-Chave*— Análise de desempenho, regressão, predição de chuva.**

***Abstract*— The occurrence of extreme events that cause natural catastrophes is difficult to predict due to the chaotic nature of these events. A first step to accurately predict these events in time is to build models capable of predicting the rainfall level in a short time-window. This paper evaluates the performance of 10 regression algorithms in order to identify those that generate assertive models to predict the level of precipitation at each hour. We analyze the coefficient of determination and the root mean square error, calculated for each model. The results show that the random forest and degree 3 polynomial regression algorithms achieve the lowest average errors, being potential candidates for rainfall forecasting.**

***Keywords*— Performance analysis, regression, rainfall forecast.**

I. Introdução

As frequentes tragédias causadas pelas fortes chuvas trazem prejuízos econômicos e de vidas humanas à nação. Em Petrópolis, Rio de Janeiro, por exemplo, as fortes chuvas ocorridas em fevereiro de 2022 deixaram ao menos 241 mortos[1]. Os efeitos dessas tragédias podem ser mitigados por um sistema de alerta antecipado eficiente. No entanto, esses sistemas são complexos de serem desenvolvidos porque eventos extremos que provocam catástrofes naturais, como tempestades que inundam cidades, têm uma natureza caótica [1]. A previsão numérica de tempo é o método tradicional para prever chuvas fortes. Os modelos matemáticos simulam a atmosfera e preveem o estado futuro do clima. Esse método, no entanto, não é apropriado para previsão regional em futuro próximo (*nowcasting*) devido à baixa resolução temporal e espacial [2]. Sistemas de alerta antecipado são dependentes de conhecer essa previsão para o futuro próximo para que os alertas sejam gerados. Dessa forma, uma abordagem alternativa é necessária para complementar as predições numéricas em uma escala temporal e espacial menor. Esse complemento pode ser feito por classificadores que preveem a ocorrência de chuva e regressores que preveem o volume de chuva [3].

Diversos trabalhos focam na predição de chuvas utilizando algoritmos de aprendizado profundo, obtendo modelos assertivos com variadas estruturas e tipos de redes neurais [4]. Normalmente, redes neurais apresentam melhor desempenho quando são usadas para prever o nível de precipitação anual [5]. Outros trabalhos utilizam algoritmos de aprendizado de máquina para realizar a predição, como a regressão linear [6]. Cramer et al. [7] comparam o desempenho de um algoritmo de estado-da-arte, a Cadeia de Markov estendida para predição de chuva, com o desempenho de seis algoritmos de inteligência artificial, incluindo aprendizado profundo. Os algoritmos de inteligência artificial superam o desempenho da cadeia de Markov estendida, indicando que os sistemas de predição de chuva podem se beneficiar desses algoritmos.

Nesse contexto, este artigo tem como objetivo avaliar diversos algoritmos de regressão para previsão do volume de chuva. Diferentemente dos trabalhos citados, a ideia é construir modelos capazes de prever o volume de chuva em escala temporal pequena, a cada hora. Assim, 10 algoritmos de regressão são aplicados a um conjunto de dados público[2]. A avaliação de desempenho compara os coeficientes de determinação, R^2, e a raiz do erro quadrático médio (*Root Mean Square Error* – RMSE) obtidos. Os resultados mostram que os algoritmos Floresta Aleatória (*Random Forest* – RF) e Polinomial de grau 3 são candidatos promissores para compor um sistema inteligente de alerta e previsão de chuvas e como contribuição para trabalhos futuros foi visto que os algoritmos de regressão linear não servem para o nosso propósito.

II. Metodologia

Utiliza-se a biblioteca *Sci-kit Learn* do Python para treinar e testar os algoritmos de regressão Linear, Poisson,Normal, Lasso, Ridge, Polinomial com diferentes graus, K-Vizinhos Mais Próximos, *K-Nearest Neighbors* – KNN), RF e Árvore de Decisão (*Decision Tree* – DT). São selecionadas 5 cidades do conjunto de dados, com características climáticas distintas, e os modelos são treinados e testados individualmente para cada cidade utilizando uma proporção de dados de treino/teste de 80%/20%. As cidades selecionadas para a avaliação são São Paulo, Belém, Curitiba, Goiânia e Salvador. Utiliza-se um

Guilherme Ferreira e Dianne Medeiros, Departamento de Engenharia de Telecomunicações, Universidade Federal Fluminense, Niterói-RJ, e-mail: {guilhermesef,diannescherly}@id.uff.br. Este trabalho é parcialmente financiado por CNPq, CAPES, FAPERJ, FAPESP (2018/23062-5), Prefeitura de Niterói/FEC/UFF (Edital PDPA 2020) e RNP.

[1] https://g1.globo.com/profissao-reporter/noticia/2022/05/11/profissao-reporter-acompanha-familias-que-perderam-casas-e-parentes-na-tragedia-de-petropolis.ghtml

[2] https://portal.inmet.gov.br/dadoshistoricos

espaço temporal de 13 anos, com registros de informações climáticas a cada hora. As variáveis independentes usadas são pressão atmosférica ao nível da estação, pressão atmosférica máxima na Hora Anterior (H.A.), pressão atmosférica mínima H.A., temperatura do ar - bulbo seco, temperatura do ponto de orvalho, temperatura máxima H.A., temperatura mínima H.A., temperatura de orvalho máximo H.A., temperatura de orvalho mínimo H.A., umidade relativa máxima H.A., umidade relativa mínima H.A., umidade relativa do ar, direção do vento, rajada máxima do vento e velocidade do vento. Remove-se os registros com dados incompletos e com valores incoerentes, como -9999 em diversas variáveis. A variável dependente considerada é a precipitação total. O desempenho de cada modelo é avaliado por meio do RMSE e do R^2 obtidos para o conjunto de teste. Utiliza-se o RMSE por prover informações sobre o desempenho a curto prazo de um modelo, possibilitando uma comparação entre o valor real e o previsto. Ademais, erros grandes na predição do volume de chuva podem ser graves e o RMSE se caracteriza por penalizar mais esses erros.

III. Resultados

A definição do parâmetro k para o KNN é feita pelo método do cotovelo (*elbow method*), variando-se o valor de k de 2 a 10. A Figura 1 mostra que os valores de k ótimos obtidos para as cidades estão em torno de 3. O valor de k usado foi 3, exceto em Goiás que foi utilizado k igual a 4. As Figuras 2 e 3 mostram os valores médios de R^2 e RMSE obtidos, com um intervalo de confiança de 95%. Os valores médios de R^2 são menores que 0,62, indicando que a variável dependente não tem relação linear com as variáveis independentes. Já os valores médios RMSE são baixos, sendo limitados a menos de 2 mm de volume de chuva em uma janela de 1 hora. Os menores erros são encontrados para os algoritmos RF e Polinomial de Grau 3 (Poly 3). O algoritmo RF durante o treino apresenta R^2 de aproximadamente 0,9 para todas as cidades, um bom resultado que não é observado no teste. Isso indica a possibilidade de sobreajuste aos dados para esse modelo.

Destaca-se que, apesar de o algoritmo RF apresentar um bom resultado para o RMSE, esse algoritmo não é capaz de extrapolar valores. Dessa forma, o intervalo de valores preditos é limitado pelos valores mais altos e mais baixos nos dados de treinamento, o que pode limitar a previsão de chuvas mais intensas do que as já vistas pelo modelo no treinamento.

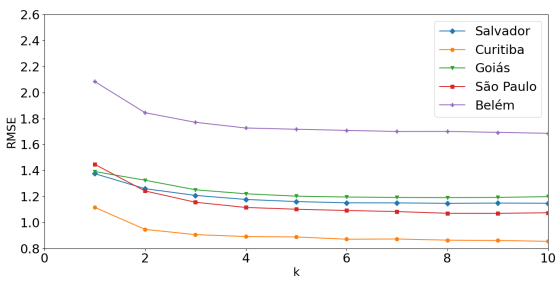

Fig. 1. Método do cotovelo para determinar o k ótimo para o algoritmo KNN aplicado a cada cidade.

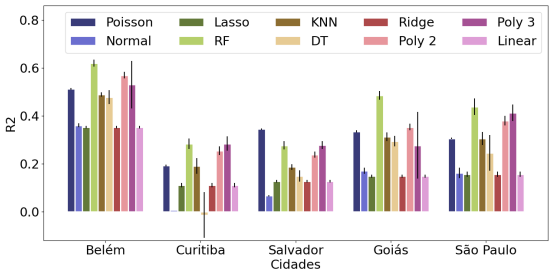

Fig. 2. Média do coeficiente de determinação, R^2, obtido para os modelos gerados pelos algoritmos para cada cidade.

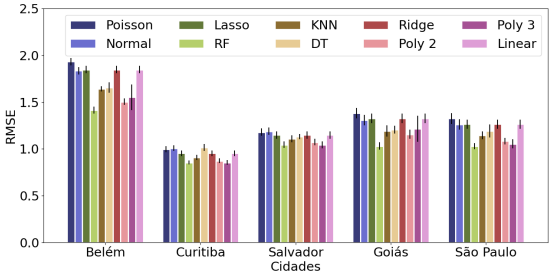

Fig. 3. Média da raiz do erro quadrático médio, RMSE, obtido para os modelos gerados pelos algoritmos para cada cidade.

IV. Conclusões

Este artigo avaliou 10 algoritmos de regressão visando gerar modelos para previsão de volume de chuva em uma janela de tempo de 1 hora. Os algoritmos foram aplicados a dados de 5 cidades que possuem diferentes comportamentos climáticos. Os resultados mostraram que os algoritmos RF e Poly 3 apresentaram as menores médias de RMSE para todas as cidades e há indícios de sobreajuste para o modelo RF obtido. Como trabalhos futuros, pretende-se estender a base de dados e refinar os modelos, investigando a importância das variáveis independentes. Vislumbra-se avaliar outros algoritmos não lineares e séries temporais para prever chuva intensa.

Referências

[1] S.-H. Moon, Y.-H. Kim, Y. H. Lee, and B.-R. Moon, "Application of machine learning to an early warning system for very short-term heavy rainfall," *Journal of Hydrology*, vol. 568, pp. 1042–1054, 2019.

[2] A. Kumar, T. Islam, Y. Sekimoto, C. Mattmann, and B. Wilson, "Convcast: An embedded convolutional lstm based architecture for precipitation nowcasting using satellite data," *PLOS ONE*, vol. 15, no. 3, pp. 1–18, 03 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0230114

[3] J. A. Weyn, D. R. Durran, and R. Caruana, "Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data," *Journal of Advances in Modeling Earth Systems*, vol. 11, no. 8, pp. 2680–2693, 2019.

[4] S. Aswin, P. Geetha, and R. Vinayakumar, "Deep learning models for the prediction of rainfall," pp. 0657–0661, 2018.

[5] M. P. Darji, V. K. Dabhi, and H. B. Prajapati, "Rainfall forecasting using neural network: A survey," pp. 706–713, 2015.

[6] S. K. Mohapatra, A. Upadhyay, and C. Gola, "Rainfall prediction based on 100 years of meteorological data," in *2017 Int. Conf. on Comp. and Comm. Tech. for Smart Nation (IC3TSN)*, 2017, pp. 162–166.

[7] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," *Expert Systems with Applications*, vol. 85, pp. 169–181, 2017.