Modelagem do número de novos casos confirmados por dia da COVID-19 no Brasil com uso de LSTM e predição linear

Karhyne P. Assis[1], Camila M. Silva[1], Kenji N. Filho[1], Ricardo Suyama[1], André K. Takahata[1]

[1]Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC (UFABC) - Santo André - SP - Brasil

{karhyne.assis, camila.merces, kenji.nose, ricardo.suyama, andre.t}@ufabc.edu.br

Abstract. *We analyzed the behavior of unit step predictors to predict the number of reported cases of COVID-19 per day. We investigated predictors created with the use of long short-term memory (LSTM) neural networks and we assessed their performance in comparison to linear predictors. We identified cases in which LSTM performs better, but also some challenges to make the LSTM based predictors capable of generalizing its performance.*

Resumo. *Analisamos o comportamento de modelos de predição de passo unitário para predição de número de novos casos de COVID-19 confirmados por dia. Utilizamos preditores com uso de rede neural de memória de longo e curto prazo (LSTM) em comparação com preditores lineares. Identificamos cenários em que a LSTM apresenta melhores resultados, mas que também há desafios para que a LSTM possa generalizar os seus resultados.*

1. Introdução

Em março de 2020 foi decretada pela Organização Mundial da Saúde (OMS) a pandemia global da COVID-19 causada pelo vírus SARS-CoV-2. Essa pandemia representa um enorme desafio para a construção de sistemas de prevenção e controle emergenciais em diversos países, onde sua propagação e disseminação causou enormes impactos sociais no mundo [Oliveira et al., 2020].

Utilizamos no presente trabalho o número cumulativo de casos confirmados de COVID-19 do Brasil retirados de uma base de dados elaborada pelo Centro de Ciência e Engenharia de Sistemas (CSSE) da Universidade Johns Hopkins (JHU). Essa base é atualizada diariamente e contém diversas outras informações acerca da pandemia como número cumulativo de mortes e o número cumulativo de casos de 274 países [Dong et al., 2020]. No caso do Brasil, a base de dados da JHU utiliza informações fornecidas pela Comissão Nacional de Saúde.

Em particular, o objetivo principal foi analisar a dinâmica da notificação de novos casos de COVID-19 no Brasil com uso de preditores de passo unitário, visando a construção de preditores a mais longo prazo em trabalhos futuros[1]. Para a construção do modelo, utilizamos a rede neural recorrente com uso de memória de longo e curto prazo (LSTM, *long short-term memory*) adequada para modelar séries temporais em modelos dinâmicos não lineares, levando-se em conta informações de longo e curto prazos [Sherstinsky, 2020] e como base de comparação, utilizamos um preditor linear obtido com o uso do método dos mínimos quadrados (MMQ) [Romano et al., 2011].

[1] Código-fonte disponível em: < https://github.com/karhyne/COVID-19-no-Brasil-com-uso-de-LSTM-e-predicao-linear/blob/main/COVID_19_LSTM_PredicaoLinear.ipynb>

2. Metodologia

A partir dos dados da COVID-19 da JHU [Dong et al. 2020], obtivemos os dados relativos ao número cumulativos de casos confirmados no Brasil entre os dias 26 de fevereiro de 2020 e 11 de junho de 2021, denotados por $x_c(n), n = 0, \dots, N-1$, em que $n = 0$ corresponde ao primeiro dia da série temporal analisada e N ao número total de dias considerados. Em seguida, obtivemos o número de novos casos diários, $x(n)$, fazendo-se $x(n) = x_c(n) - x_c(n-1)$, considerando $x_c(n) = 0$ para $n < 0$. Buscamos com o uso da LSTM modelar a dinâmica dos novos casos diários, com a realização da predição de passo unitário, isto é, correspondente a um dia. Para isso, utilizamos como entrada da LSTM o vetor formado por $x(n)$ e seus respectivos atrasos

$$\mathbf{x}_n = [x(n) \quad x(n-1) \quad \dots \quad x(n-L)]^T,$$

em que L é o número de atrasos considerados e $y(n) = x(n+1)$ corresponde ao valor a ser estimado pela LSTM. Assim, denotando a estimativa retornada pela LSTM como $\hat{y}(n) = f(\mathbf{x}_n)$, a raiz quadrada do erro quadrático médio (RMSE, *root-mean-square error*) e o erro percentual absoluto médio (MAPE, *Mean Absolute Percentage Error*) são dados, respectivamente, por

$$RMSE = \sqrt{EQM} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}[f(\mathbf{x}_n) - y(n)]^2},$$

$$MAPE = \frac{1}{N}\sum_{n=0}^{N-1}\left|\frac{f(\mathbf{x}_n) - y(n)}{y(n)}\right| \times 100\% .$$

Para implementação da LSTM, utilizamos o pacote Keras[2] em linguagem Python com uma rede com u unidades de LSTM variando no intervalo $6 \leq u \leq 18$, seguida por uma camada densa de 1 unidade. Como base para comparação, realizamos a predição linear, em que os valores preditos são obtidos por meio de uma função linear tal que $\hat{y}_{linear}(n) = \mathbf{w}^T\mathbf{x}_n$, onde $\mathbf{w}$ é um vetor de tamanho $L + 1 \times 1$ obtido de modo a minimizar o EQM. Essa regressão linear foi implementada por meio com uso pacote scikit-learn[3], também em linguagem Python. Para o treinamento e teste dos algoritmos foram utilizados, respectivamente, 90% e 10% dos dados, sendo que a sequência temporal dos dados foi mantida em ambos os conjuntos de dados.

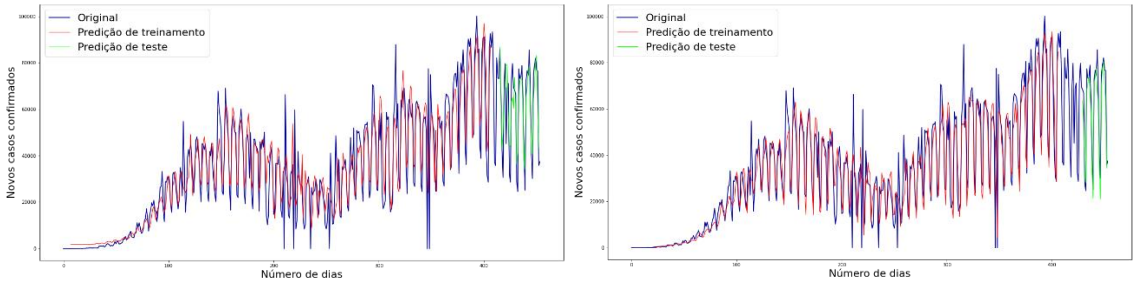
3. Resultados

Nesta seção apresentamos os resultados obtidos com a LSTM e o preditor linear para uma janela temporal de 7 e 21 dias na entrada do preditor (*L=6* e *20*, respectivamente). Para a LSTM com uma janela temporal de 7 dias, o RMSE dos dados de treinamento ficou entre 9631,40 ($u = 14$) e 10315,32 ($u = 18$) e para os dados de teste ficou entre 9854,09 ($u = 11$) e 11079,30 ($u = 18$). No caso do menor RMSE no teste ($u = 11$), observou-se um RMSE de treinamento de 9769,43, bastante próximo ao menor RMSE de treinamento observado.

Na Figura 1 é mostrado o comparativo entre o uso da LSTM ($u = 11$) e o preditor linear. É possível observar que, em geral, ambos resultados estão próximos dos dados originais, em azul, porém, o preditor linear apresentou uma discrepância maior tanto nos picos quanto nos vales da série temporal em relação à LSTM. Esse efeito foi observado tanto nos dados de treinamento, em vermelho, quanto nos dados de teste, em verde. Isso

[2] https://keras.io/
[3] https://scikit-learn.org/

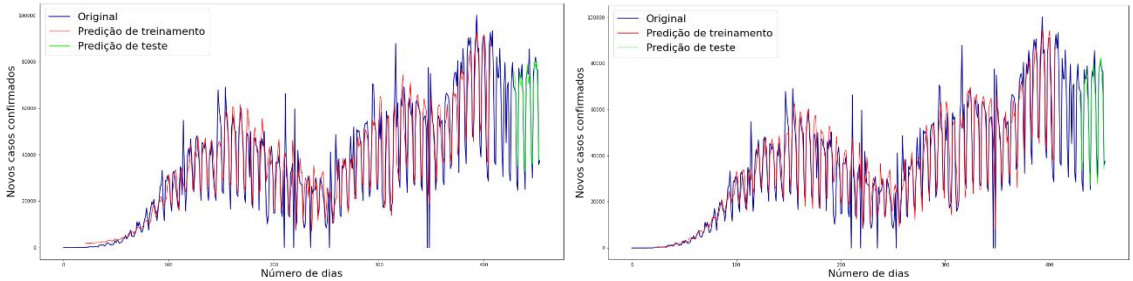
é refletido na Tabela 1(a), em que tanto nos dados de treino quanto no de teste, a LSTM teve um RMSE menor que o preditor linear.



(a) (b)

Figura 1: Predição para novos casos diários informados no Brasil com $L = 6$. (a) Preditor linear. (b) Preditor com uso de LSTM com $u = 11$.

Para a janela temporal de 21 dias com a LSTM, o RMSE de treinamento ficou entre 8198,42 ($u = 12$) e 10976,56 ($u = 6$) e o RMSE de teste entre 6575,54 ($u = 18$) e 12026,93 ($u = 14$). Na Figura 2 são mostrados os gráficos em que se pode comparar as saídas do preditor com LSTM ($u = 18$) e do preditor linear aos dados originais, em que as saídas dos dois preditores acompanham, grosso modo, as variações dos dados originais.



(a) (b)

Figura 2. Predição para novos casos diários informados no Brasil com $L = 20$. (a) Preditor linear. (b) Preditor com uso de LSTM com $u = 18$.

Tabela 1. RMSE da predição de novos casos confirmados diários de COVID-19 com predição linear e LSTM

Janela de 7 dias		
	Treinamento	Teste
LSTM ($u = 11$)	9769,43	9854,09
Pred. Linear	10553,99	10799,77

(a)

Janela de 21 dias		
	Treinamento	Teste
LSTM ($u = 18$)	8510,21	6575,54
Pred. Linear	10016,11	5579,47

(b)

Na Tabela 1(b) é possível observar que na comparação com os resultados da Tabela 1(a), percebemos uma melhoria nos resultados em todos os casos analisados com o aumento do tamanho da janela temporal. Além disso, na Tabela 1(b) podemos observar que o LSTM consegue um menor RMSE de treinamento, mas a situação se inverte no conjunto de teste. Uma vez que o número diário de novos casos confirmados apresenta uma grande variação no tempo e o RMSE penaliza mais os erros associados às maiores magnitudes, considerando-se um mesmo erro percentual, refinamos a análise com o uso do MAPE, pois essa métrica utiliza o mesmo peso para erros percentuais iguais. Assim, no caso da LSTM, o valor do MAPE calculado os dados de teste foi de aproximadamente 13,7% e 8,2%, respectivamente, para as janelas temporais de 7 dias com $u = 11$ e de 21 dias com $u = 18$. Para o preditor linear, o MAPE calculado foi de aproximadamente

18,4% e 9,5%, respectivamente para as janelas de 7 e 21 dias. Deste modo, pudemos verificar que o MAPE calculado apresentou melhor resultado para o LSTM em comparação com o preditor linear.

4. Conclusões

Para a janela de 7 dias a LSTM obteve um desempenho melhor que o preditor linear, o que era esperado devido às características da LSTM, que incluem o tratamento de efeitos de memória de longo e curto período bem como não linearidades que podem permitir uma melhor modelagem da dinâmica dos dados. Por outro lado, no caso da janela de 21 dias foi observado que a LSTM obteve um RMSE de treinamento consideravelmente melhor que o preditor linear, mas o mesmo não foi obtido no teste. Por outro lado, ao analisar o valor do MAPE, foi observado que a LSTM sempre foi superior ao preditor linear. É interessante notar que o RMSE é maior no treinamento do que no teste para as duas abordagens. Há grande chance de que isso ocorra devido ao caráter não estacionário dos dados, uma vez que a característica de propagação do vírus se altera com o passar do tempo [Sabino et al., 2021]. Como consequência, essas alterações são mais evidenciadas no dado de treinamento uma vez que representa um período maior, como pode ser visto nas Figuras 1 e 2. O melhor desempenho do LSTM nos dados de treinamento pode ser atribuído ao fato desse modelo conseguir acomodar melhor as não-estacionaridades que o modelo linear, mas há uma dificuldade de se generalizar os resultados. Assim, fatores que interferem nessa capacidade como o tamanho da janela temporal do preditor e a estrutura da rede neural devem ser investigados com mais profundidade. Como perspectivas de trabalhos futuros, visando a diminuição das métricas de erro, podemos considerar a inclusão de dados ligados à fatores que afetam a dinâmica da propagação do vírus na entrada do preditor, como informações a respeito de políticas públicas (vacinação, restrição de abertura de comércio, entre outros) no modelo. Além disso, podemos elencar o desenvolvimento futuro do algoritmo para realizar a predição de um número maior de passos.

Agradecimentos

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES e pelos comentários e leitura cuidadosa dos revisores anônimos desse trabalho.

Referências

Dong, E., Du, H. and Gardner L. (2020) “An interactive web-based dashboard to track COVID-19 in real time”, In: Lancet, 20(5), 533-534.

Oliveira, W. K., Duarte, E., França, GVA. And Garcia, LP. (2020) “How Brazil can hold back COVID-19”, Epidemiol Serv Saude; 29(2): e2020044.

Romano, J. M. T., Attux, R. R. F., Cavalcante, C.C., Suyama, R. (2011) “Unsupervised signal processing: channel equalization and source separation”. CRC Press.

Sabino, E. C., et al. (2021) “Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence”, In: Lancet, 397(10273), 452-455.

Shertinsky, A. (2020) “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network”, Physica D: Nonlinear Phenomena, v. 404, p. 132306.