Deep Learning sobre dados meteorológicos tabulares: Proposta e Base de Dados

Matheus de Souza Ataide 147375

Mauricio de Sousa Araujo 184477

Raysa Masson Benatti 176483

1 Introdução

Técnicas de Deep Learning têm sido crescentemente exploradas na última década em razão do aumento de poder de processamento computacional e disponibilidade de dados, o que permitiu o uso e desenvolvimento de diferentes arquiteturas de redes neurais profundas. Um dos maiores trunfos dessas ferramentas é a possibilidade de reconhecer padrões em dados não estruturados, como imagens, sinais elétricos e texto. Em razão disso, seu uso tem sido concentrado em áreas como visão computacional e processamento de linguagem natural.

Menos explorado tem sido o uso de tais técnicas sobre dados tabulares ou estruturados. Em geral, métodos tradicionais de Machine Learning performam melhor nesse universo, sendo, portanto, mais utilizados. No entanto, o interesse em experimentar ferramentas de Deep Learning sobre esse tipo de dado vem aumentando - afinal, parte considerável dos dados disponíveis em diversos domínios está em formato tabular.

Embora, em geral, não exista estrutura hierárquica a ser explorada em dados estruturados, há indícios de que técnicas de Deep Learning podem agregar vantagens ao processo - especialmente para datasets maiores -, como facilitar eventuais análises multimodais e análise de dados online, e diminuir a necessidade por *feature engineering* [1, 2]. Nesse cenário, destaca-se a arquitetura TabNet (Attentive Interpretable Tabular Learning), cuja descrição foi publicada pela Google Research em preprint de 2020 [2, 3]. Essa e outras arquiteturas de redes neurais profundas têm sido usadas sobre dados tabulares em aplicações que incluem predição de prejuízos na indústria de seguros [4], geração sintética de dados a partir de modelagem probabilística [5], uso sobre dados de busca para análise de comportamento de consumidor [6], embedding de variáveis categóricas [7] e previsão em séries temporais [7, 8, 9].

Técnicas de previsão em séries temporais, em particular, nos interessam: dados tabulares históricos são fartamente disponíveis e relativamente fáceis de obter, em diversos domínios - e em um domínio, especialmente, abarcado pela motivação de *Artificial Intelligence for Social Good* [10] e cuja importância tem crescido notavelmente: mudanças climáticas. Nesse contexto se desenha o foco deste trabalho.

2 Descrição do Problema

Mudanças climáticas e aquecimento global são realidades preocupantes: alterações nas condições climáticas - como as que têm sido observadas nas últimas décadas - podem acarretar malefícios para diversos nichos da sociedade. O clima tem impactos significativos sobre diversas atividades humanas, como agricultura, aviação e navegação, pesca, saúde, lazer, turismo e outras.

Tradicionalmente, o clima tem sido objeto de estudo de ciências como a física, a meteorologia e a geografia; recentemente, técnicas de aprendizado de máquina têm se somado às ferramentas usadas para estudá-lo, dados o desenvolvimento de tais técnicas e a insuficiência de modelos físicos para previsão climática em determinados contextos. A realidade do aquecimento global tem trazido, ainda, novas demandas na área, cujo conhecimento deve se complexificar cada vez mais [11].

Com isso em mente, o presente projeto visa a analisar dados climáticos de modo a compreender as variações do passado e prever variações futuras. A compreensão do comportamento de variáveis climáticas ao longo do tempo passa pelo reconhecimento de padrões que podem ser detectados, por exemplo, com o uso de redes neurais profundas para análise de séries temporais, conforme ilustrado por algumas das referências mencionadas.

Para tanto, a aplicação de algumas dessas técnicas será explorada sobre a base de dados descrita na seção a seguir.

3 Base de Dados

Utilizaremos a base de dados climáticos (série histórica) disponibilizada pelo INMET (Instituto Nacional de Meteorologia) via requisição. Trata-se de dados metereológicos de diferentes resoluções medidos em determinadas estações localizadas no Brasil. Tal base é constituída por dezesseis *features*, dentre as quais selecionaremos as seguintes:

- Data da medição;
- Hora da medição;
- Precipitação total;
- Pressão atmosférica;
- Temperatura do ar (bulbo seco);
- Temperatura do ar (bulbo úmido);
- Umidade relativa do ar.

Para nossa análise, consideraremos dados de resolução diária, medidos três vezes ao dia, no período de 1960 a 2020, nas seguintes estações: 83856 - São Paulo (Horto Florestal) (desativada em 1982), 83781 - São Paulo (Mir. De Santana) e 83004 - São Paulo (Iag), localizadas na cidade de São Paulo - totalizando, aproximadamente, 150 mil instâncias.

Referências

[1] Mikael Huss. *Tabular Data and Deep Learning: Where Do We Stand?* http://shorturl.at/fvzH7. 2020.

[2] Sercan O. Arik e Tomas Pfister. "TabNet: Attentive Interpretable Tabular Learning". Em: *arXiv* (2020).

[3] Mikael Huss. *Modelling tabular data with Google's TabNet*. http://shorturl.at/oAGX2. 2020.

[4] Dian Maharani, Hendri Murfi e Yudi Satria. "Performance of Deep Neural Network for Tabular Data - A Case Study of Loss Cost Prediction in Fire Insurance". Em: *International Journal of Machine Learning and Computing* 9.6 (2019), pp. 734–742.

[5] Lei Xu et al. "Modeling Tabular Data using Conditional GAN". Em: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (2019).

[6] Malay Haldar et al. "Applying Deep Learning To Airbnb Search". Em: *arXiv* (2018).

[7] Rachel Thomas. *An Introduction to Deep Learning for Tabular Data*. http://shorturl.at/tLOTV. 2018.

[8] Meng-Hua Yen et al. "Application of the deep learning for the prediction of rainfall in Southern Taiwan". Em: *Scientific Reports* 9 (2019).

[9] Tony Zhou. *Deep Learning for Time Series and why DEEP LEARNING?* http://shorturl.at/boxS1. 2020.

[10] Nenad Tomašev et al. "AI for social good: unlocking the opportunity for positive impact". Em: *Nature Communications* 11 (2020).

[11] David Rolnick et al. "Tackling Climate Change with Machine Learning". Em: *arXiv* (2019).