Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 10, n. 1, 2023.

Trabalho apresentado no XLII CNMAC, Universidade Federal de Mato Grosso do Sul - Bonito - MS, 2023

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Modelos de Aprendizado Profundo guiados por TDA para previsão de Séries Temporais: Uma aplicação do índice S&P500

Jerson L. Alves[1]
Renné Rodrigues Lima[2]
Davi W. Misturini[3]
Francisco A. dos Santos[4]
João B. Florindo[5]
IMECC, Unicamp, SP

A Análise Topológica de Dados (TDA, Topological Data Analysis) é um campo recente que fornece um conjunto de ferramentas topológicas e geométricas para inferir características relevantes para dados, dos mais simples aos mais complexos, como o Teorema da Incorporação de Takens [1] e a norma topológica (1) que são utilizados neste trabalho.

$$||\eta||_p = (\sum_{k=1}^{\infty} ||\eta_k||_p^p)^{\frac{1}{p}}, \quad \text{sendo } ||\cdot||_p \text{ a norma } L^p,\ 1 \leq \mathrm{p} \leq \infty. \tag{1}$$

De acordo com [2], os aspectos topológicos e geométricos geralmente estão associados a espaços contínuos, por isso, uma maneira natural de destacar alguma estrutura topológica dos dados é "conectar" pontos de dados próximos uns dos outros para exibir uma forma contínua global subjacente aos dados. E assim, quantificar a noção de proximidade entre os pontos de dados, o que geralmente é feito usando uma distância, e muitas vezes é conveniente em TDA considerar conjuntos de dados como espaços métricos discretos ou como amostras de espaços métricos.

Analisar uma Série Temporal é observar o comportamento de uma sequência de observações de algo ao longo do tempo. Essa análise se torna bastante intrigante com a possibilidade de predizer dados futuros utilizando dados reais passados, o que pode influenciar de forma importante na qualidade da tomada de uma decisão [2].

Os aspectos topológicos e geométricos de uma Série Temporal geralmente estão associados a espaços contínuos, e neste estudo foi utilizado o Teorema da Incorporação de Takens para a reconstrução de um espaço de fase a partir da Série Temporal dos preços de fechamento diário do índice S&P500 denotada por S_1, do dia 04/01/2010 até o dia 24/12/2019, totalizando um intervalo de tempo de 10 anos, ou 2515 dias, sem a influência da pandemia da Covid19.

A partir daí, foi aplicado o método de janelas deslizantes com intervalo 3 unidades e tamanho de passo 1 unidade para gerar subséries temporais e calculada a norma topológica em cada uma das janelas, gerando assim uma Série Temporal das normas, e esta última foi concatenada com a Série Temporal dos preços de fechamento diário do índice S&P500 e denominada S_2.

[1] j235140@dac.unicamp.br
[2] r225144@dac.unicamp.br
[3] d235799@dac.unicamp.br
[4] f219979@dac.unicamp.br
[5] florindo@unicamp.br

De posse das duas Séries Temporais, S_1 e S_2 , os dados de cada uma individualmente foram aplicados nos modelos de redes neurais Convolucional e Long Short Term Memory (LSTM), da seguinte forma: 80% dados de treino e 20% dados de teste, ou seja 2012 dados de treino e 503 dados de teste.

Comparou-se os resultados computacionais aplicando os modelos: Convolucional e LSTM. Para isso foram utilizadas as métricas Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE) para avaliação dos modelos, com base em [3]. Como pode ser visto consultando as tabelas 1 e 2.

Tabela 1: Série Temporal $S\&P500$ (S_1).

Índice	Métrica	Convolucional	LSTM
$S\&P500$	MAE	18,554	**19,097**
	MSE	692,006	**727,684**
	RMSE	26,306	**26,976**

Tabela 2: Série Temporal concatenada $S\&P500$ e norma topológica (S_2).

Índice	Métrica	Convolucional	LSTM
$S\&P500$	MAE	**18,423**	19,511
	MSE	**687,547**	760,299
	RMSE	**26,221**	27,573

Nota-se, que tanto na aplicação dos modelos a S_1 quanto a S_2, o modelo Convolucional se mostra mais eficiente na comparação com o modelo LSTM. No entanto, ao concatenar os dados das séries S_1 e S_2, há uma ligeira melhora na predição dos dados apenas no modelo Convolucional, enquanto não melhora nada no modelo LSTM. Os erros calculados MAE, MSE e RMSE reduzem 0,4%, 0,798% e 0,399%, respectivamente, no modelo Convolucional. Ao tempo que aumentam no modelo LSTM.

A simulação foi satisfatória, pode-se concluir que o modelo Convolucional mostra uma melhora significativa quando guiados por TDA, mostrando-se eficaz a utilização da norma como característica na análise dos dados. Por outro lado, o modelo LSTM não apresentou qualquer melhora ao combinar os dados com TDA.

Agradecimentos

Esse trabalho possui suporte em parte pelo Instituto Federal do Piauí - IFPI e pela Fundação de Amparo à Pesquisa do Estado do Piauí - FAPEPI.

Referências

[1] Anubha Goel, Puneet Pasricha e Aparna Mehra. “Topological data analysis in investment decisions”. Em: **Expert Systems with Applications** 147 (2020), p. 113222.

[2] Jean-Daniel Boissonnat, Frédéric Chazal e Mariette Yvinec. **Geometric and topological inference**. Vol. 57. Cambridge University Press, 2018.

[3] Aurélien Géron. “Hands-on machine learning with scikit-learn and tensorflow: Concepts”. Em: **Tools, and Techniques to build intelligent systems** (2017).