DMC_x^2 e aprendizado de máquina aplicados à análise de séries temporais de dados meteorológicos

Apresentação do andamento da pesquisa

Discente: Fernando Ferraz Ribeiro
Orientador: Prof. Dr.Gilney Figueira Zebende
Coorientador: Prof. Dr. Juan Alberto Leyva Cruz

18/06/2023

UEFS PPGM - Feira de Santana, BA

Sumário

Introdução

Este conjunto amplo de fenômenos é comumente identificado e agrupado por algumas de suas características: são formados pela contribuição de um conjunto (geralmente grande) de componentes (muitas vezes simples) que, interagindo, estruturam-se de forma auto-organizada, gerando resultados inesperados, que não podem ser previstos pelos estudos estatísticos e/ou matemáticos tradicionais dos elementos formadores do sistema.

Reconhecimento

Em 2021, a Academia Real das Ciências da Suécia concedeu metade do Prêmio Nobel de Física para Syukuro Manabe e Klaus Hasselmann, cujos estudos apresentam modelos complexos para a análise do clima. Em particular apontam uma correlação entre as emissões de dióxido de carbono e as mudanças climáticas.

Muitos fenômenos complexos são investigados pela análise de grandes conjuntos de dados. É notável a velocidade e quantidade de dados que são gerados e armazenados pela humanidade atualmente. A aquisição, manipulação, gestão, armazenamento e criação de valor a partir de dados, através de ambientes computacionais, tem-se apresentado como um novo paradigma tecnológico. Um campo do conhecimento que recebeu a denominação de Ciência de Dados, conceito que envelopa alguns termos frequentemente associados à inovação científica, técnica e social como *Big Data*, mineração de dados, *Business Intelligence* internet das coisas, inteligência artificial e aprendizado de máquina(AM), dentre outros (EMC EDUCATION SERVICE, 2015, p. 12-13).

As séries temporais são definidas como um conjunto de observações (numéricas ou categóricas) ordenado no tempo. Embora muitos dos dados que descrevem as dinâmicas espaciais podem ser registrados na forma de sérias temporais (abastecimento de água nas tubulações, consumo de energia elétrica nos imóveis, fluxos de pessoas e veículos pela cidade, casos de uma doença por dia, etc.), contudo as técnicas de medição de correlações, bem como a devida exploração destas para inferir novos conhecimentos, permanecem como perguntas abertas em muitas sub-áreas das ciências ambientais(Bermudez-Edo; BARNAGHI; MOESSNER, 2018).

Proposta

Esta pesquisa propõe-se um estudo de dois conjuntos de variáveis meteorológicas, utilizando o DMC_x^2 como ferramenta de medição das correlações entre múltiplas variáveis. Após a avaliação dos resultados destes estudos, propõe-se a criação de um modelo preditivo utilizando ferramentas de AM e o DMC_x^2 .

Objetivo Principal

O objetivo principal desta pesquisa é: investigar as correlações entre as variáveis meteorológicas de diferentes localidades através do coeficiente DMC_x^2 e utilizar o conhecimento destas correlações para alimentar um modelo preditivo de condições meteorológicas.

Objetivos Gerais

1. Implementar um algoritmo computacional geral para calcular o DMC_x^2 para qualquer número de séries temporais.
2. Analisar um conjunto de dados climáticos contendo medições meteorológicas de todas as capitais brasileiras.
3. Analisar um conjunto de dados meteorológicos sobre radiação solar com estações locadas em diversas partes do globo.
4. Desenvolver e implementar um algoritmo de predição baseado em aprendizado de máquina e redes neurais artificias agregados com o coeficiente DMC_x^2 .

Premissas

1. Os fenômenos climáticos estão relacionados de forma complexa. Por exemplo: massas de ar percorrem distâncias na atmosfera e influenciam uma série de variáveis climáticas nas localidades por onde passam, mas que também são influenciadas, em seu percurso ou sua dissolução pelas mesmas variáveis.
2. O DMC_x^2 , pelas características de análise do método, pode ajudar a entender estas correlações.
3. O DMC_x^2 é uma generalização do método ρ_{DCCA} para múltiplas séries temporais.
4. O ρ_{DCCA} , em determinadas condições testadas, apresentou resultados mais interessantes (como melhor descrição dos fenômenos) que os apresentados pelo coeficiente de Pearson quando aplicado à séries temporais (WANG et al., 2013).

Questões

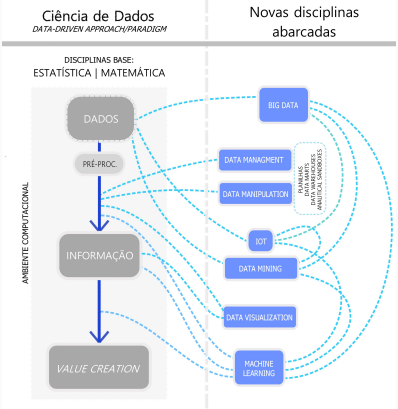
1. É possível estabelecer e medir correlações entre variáveis meteorológicas de uma determinada localidade e um conjunto de outras localidades?
2. Em caso de resposta positiva, seria possível utilizar essas correlações para melhorar modelos meteorológicos preditivos?

Hipóteses

1. Um método baseado no DMC_x^2 seria um ferramental importante no estudo de correlações de variáveis climáticas envolvendo um grande número de localidades.
2. É possível criar uma modelo preditivo para séries temporais de aprendizado de Máquina eficiente baseado no DMC_x^2 .

Metodologia

Figure 1: Diagrama conceitual - AM



Fonte: Elaborada pelos autores

Princípios norteadores

- Conheça os dados
- Entenda os algoritmos
- Desconfie dos resultados

Etapas

Com os dois conjuntos de dados organizados, para cada uma das análises seguiremos as seguintes etapas.

1. Aquisição dos dados
2. Análise exploratória
3. Pré-processamento
4. Seleção de variáveis
5. Aplicação do DMC_x^2
6. Visualização e análise dos resultados
7. Seleção de variáveis para aplicação do modelo preditivo
8. Validação e ajustes do modelo
9. Visualização e análise dos resultados

Fontes de dados

Instituto Nacional de Meteorologia (INMET) ⟨https://portal.inmet.gov.br/⟩. Do massivo conjunto de dados disponível, foram baixados apenas os registros das capitais.

Baseline Surface Radiation Network (BSRN) ⟨https://bsrn.awi.de/⟩. Uma rede de medições meteorológicas de alta precisão com estações filiadas no mundo inteiro.

Técnicas de aquisição de dados

Para grandes conjuntos de dados, faz-se necessário a utilização de técnicas raspagem de dados (data scraping) para adquirir e organizar os conjuntos de dados. Para os dados do INMET será usado o pacote Bealtiful Soap do python para baixar os arquivos.

Para os dados do BSRN será usado um pacote próprio (pvlib) para acessar os dados.

Análise exploratória

A análise exploratória dos dados tem por objetivo entender características, potenciais, limitações e possíveis erros na coleta dos dados de cada um dos conjuntos de dados.

Pré-processamento

Etapa que visa preparar o conjunto de dados para a aplicação do algoritmo. Aspectos da análise exploratória devem ser considerados e estratégias devem ser definidas para: eliminar valores com erros de medição, tratar valores faltantes, dependendo do método aplicado, normalizar as medições.

No caso da base do BSRN, é preciso considerar que a incidência solar é distribuída ao longo do globo terrestre de acordo com a longitude. reorganizar as séries para os horários locais de cada estação é uma tarefa necessária para as análises que levem em conta variáveis relacionadas com a radiação.

Seleção de atributos

Em ciência de dados, a seleção de atributos é a escolha de um subconjunto de atributos ou variáveis são selecionados para uma determinada análise ou para a criação de um modelo.

Nesta pesquisa utilizaremos alguns algoritmos de seleção de atributos para comparar os resultados com as correlações do DMC_x^2 , a saber:

- Time Series Feature Importance
- Mutual Information
- Autoencoder
- Random Forest Importance(?)

Aplicação do DMC_x^2

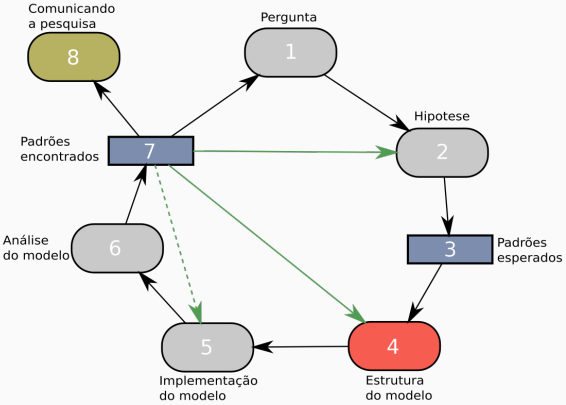
Selecionar variáveis de acordo com os algoritmos de seleção de atributos.

Utilizar estratégias de visualização de dados e comparar os resultados do DMC_x^2 com as seleções de variáveis

Seleção de variáveis para aplicação do modelo preditivo

Baseado nas análises anteriores, escolher um conjunto de variáveis para iniciar a implementação do modelo.

Figure 2: Diagrama de Grimm e Railsback



Fonte: Elaborada pelos autores

Questionamentos de pesquisadores sobre riscos da I.A.

⟨https://www.newyorker.com/humor/daily-shouts/another-warning-letter-from-ai-researchers-and-executives⟩

Fundamentação Teórica

DFA - (PENG et al., 1994)

1. Pegando a série temporal $\{x_i\}$ com i variando de 1 à N, a série integrada X_k é calculada por $X_k = \sum_{i=1}^{k} [x_i - \langle x \rangle]$ com k também variando entre 1 e N;
2. A série X_k é dividida em $N - n$ caixas de tamanho n (escala temporal), cada caixa contendo $n + 1$ observações, iniciando em i até $i + n$;
3. Para cada caixa um polinômio (geralmente de grau 1) é ajustado, gerando $\widetilde{X}_{k,i}$ com $i \leq k \leq (i + n)$ eliminando assim a tendência (detrended values);
4. para cada caixa é calculado: $f^2_{DFA}(n, i) = \frac{1}{1+n} \sum_{k=i}^{i+n} (X_k - \widetilde{X}_{k,i})^2$
5. Para todas as caixas de uma escala temporal o DFA é calculado como: $F_{DFA}(n) = \sqrt{\frac{1}{N-n} \sum_{i=1}^{N-n} f^2_{DFA}(n, i)}$;
6. Para um número de diferentes escalas temporais (n), com valores possíveis entre $4 \leq n \leq \frac{N}{4}$, a função F_{DFA} é calculada para encontrar a relação entre $F_{DFA} \times n$

DCCA - (PODOBNIK; STANLEY, 2008)

1. Para duas séries temporais $\{x_i\}$ e $\{x_i\}$ com i variando de 1 à N, as séries integradas X_k e Y_k são calculadas por $X_k = \sum_{i=1}^{k} [x_i - \langle x \rangle]$ e $Y_k = \sum_{i=1}^{k} [y_i - \langle x \rangle]$ com k também variando entre 1 e N;
2. As séries X_k e X_k são divididas em $N - n$ caixas de tamanho n (escala temporal), cada caixa contendo $n + 1$ observações, iniciando em i até $i + n$;
3. Para cada caixa um polinômio (geralmente de grau 1) é ajustado, gerando $\widetilde{X}_{k,i}$ para a primeira série e $\widetilde{Y}_{k,i}$ para a segunda com $i \leq k \leq (i + n)$ eliminando assim a tendência (detrended values);
4. Para cada caixa é calculado: $f^2_{DCCA}(n, i) = \frac{1}{1+n} \sum_{k=i}^{i+n} (X_k - \widetilde{X}_{k,i})(Y_k - \widetilde{Y}_{k,i})$
5. Para todas as caixas de uma escala temporal a função $F^2_{DCCA}(n)$ é calculada por: $F^2_{DCCA}(n) = \frac{1}{N-n} \sum_{i=1}^{N-n} f^2_{DCCA}(n, i)$;
6. Para um número de diferentes escalas temporais (n), com valores possíveis entre $4 \leq n \leq \frac{N}{4}$, a função $F^2_{DCCA}(n)$ é calculada para encontrar a relação entre $F^2_{DCCA}(n) \times n$.

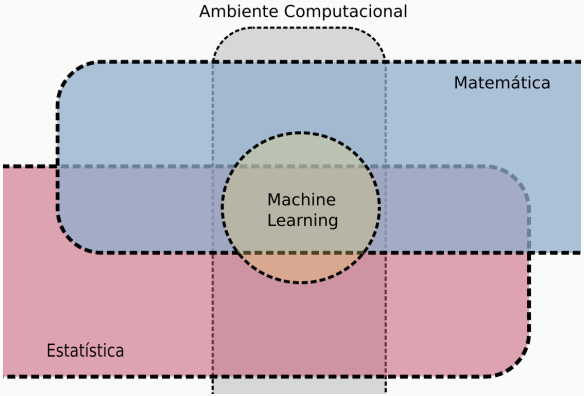
ρ_{DCCA} - (ZEBENDE, 2011)

$$\rho_{DCCA}(n) = \frac{F^2_{DCCA}(n)}{F_{DFA1}(n)F_{DFA2}(n)} \tag{1}$$

$$DMC_x^2 \equiv \rho_{y,x_i}(n)^T \rho^{-1}(n) \rho_{y,x_i}(n) \tag{2}$$

$$\rho^{-1}(n) = \begin{pmatrix} 1 & \rho_{x_1,x_2}(n) & \rho_{x_1,x_3}(n) & \dots & \rho_{x_1,x_i}(n) \\ \rho_{x_2,x_1}(n) & 1 & \rho_{x_2,x_3}(n) & \dots & \rho_{x_2,x_i}(n) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho_{x_i,x_1}(n) & \rho_{x_i,x_2}(n) & \rho_{x_i,x_3}(n) & \dots & 1 \end{pmatrix}^{-1} \tag{3}$$

Figure 3: Conceituação - AM



Fonte: Elaborada pelos autores

Resultados

Produtos da Tese

- Artigo: revisão de literatura.
- Aplicativo de Cálculo do DMC_x^2 .
- Pacote Python para cálculo do DFA, DCCA, ρ_{DCCA} , e DMC_x^2 .
- Artigo: revista Software X.
- Artigo: análise de dados meteorológicos com DMC_x^2
- Implementação do modelo de AM usando DMC_x^2 .
- Artigo: validação do modelo AM.
- Aplicação Artigo AM / dados meteorológicos

Figure 4: (FILHO et al., 2023)

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Statistical study of the EEG in motor tasks (real and imaginary)

F.M. Oliveira Filho [a,b], F.F. Ribeiro [a,e], J.A. Leyva Cruz [c], A.P. Nunes de Castro [d], G.F. Zebende [c,*]

[a] *Earth Sciences and Environment Modeling Program, State University of Feira de Santana, Bahia, Brazil*
[b] *SENAI CIMATEC University Center, Salvador, Bahia, Brazil*
[c] *State University of Feira de Santana, Bahia, Brazil*
[d] *Jorge Amado University Center, Salvador, Bahia, Brazil*
[e] *Federal University of Bahia, Salvador, Bahia, Brazil*

Referências

Bermudez-Edo, M.; BARNAGHI, P.; MOESSNER, K. Analysing real world data streams with spatio-temporal correlations: Entropy vs. Pearson correlation. *Automation in Construction*, v. 88, n. May 2017, p. 87–100, 2018. ISSN 09265805. 7

EMC EDUCATION SERVICE. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis, Indiana, USA: JOHN WILEY \& SONS, 2015. ISBN 978-1-118-87605-3 1-118-87605-9 978-1-118-87613-8 1-118-87613-X 978-1-118-87622-0 1-118-87622-9.
6

FILHO, F. O. et al. Statistical study of the EEG in motor tasks (real and imaginary). Physica A: Statistical Mechanics and its Applications, v. 622, p. 128802, jul. 2023. ISSN 03784371.
35

PENG, C.-K. et al. Mosaic Organization of DNA Nucleotides. v. 49, n. 2, p. 1685–1689, 1994.
28

PODOBNIK, B.; STANLEY, H. E. Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Physical Review Letters*, v. 100, n. 8, 2008. ISSN 00319007.
29

WANG, G. J. et al. Random matrix theory analysis of cross-correlations in the US stock market: Evidence from Pearson's correlation coefficient and detrended cross-correlation coefficient. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 392, n. 17, p. 3715–3730, 2013. ISSN 03784371.
11

ZEBENDE, G. F. DCCA cross-correlation coefficient: Quantifying level of cross-correlation. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 390, n. 4, p. 614–618, 2011. ISSN 03784371.
30

ZEBENDE, G. F.; SILVA, A. M. Detrended Multiple Cross-Correlation Coefficient. *Physica A*, Elsevier B.V., v. 510, p. 91–97, 2018. ISSN 0378-4371.
31

Obrigado