



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

Apresentação do andamento da pesquisa

Fernando Ferraz Ribeiro

15/05/2023

UEFS PPGM - Feira de Santana, BA

1. Introdução
2. Aspectos Metodológicos
3. Fundamentação Teórica
4. Resultados
5. Considerações Finais
6. Referências

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.

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 séries 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).

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 artificiais 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).

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?

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 .

Aspectos Metodológicos

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

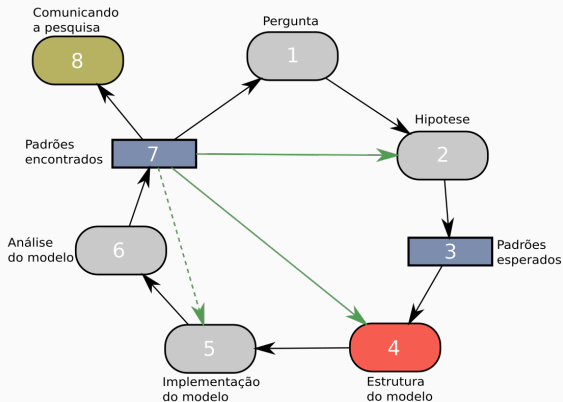
Baseline Surface Radiation Network (BSRN) (<https://bsrn.awi.de/>).
Uma rede de medições meteorológicas de alta precisão com estações filiadas no mundo inteiro.

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

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

Validação do Algoritmo de AM

Figure 1: Diagrama de Grimm e Railsback



Fonte: Elaborada Pelo Autor

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 $\tilde{X}_{k,i}$ com $i \leq k \leq (i + n)$ eliminando assim a tendência (detrended values);
4. para cada caixa é calculado: $f_{DFA}^2(n, i) = \frac{1}{1+n} \sum_{k=i}^{i+n} (X_k - \tilde{X}_{k,i})^2$
5. Para todas as caixas de uma escala temporal o DFA é calculado como: $F_{DFA}(n) = \sqrt{\frac{1}{N-n} \sum_{i=1}^{N-n} f_{DFA}^2(n, i)}$;
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 $F_{DFA} \times n$

1. Para duas séries temporais $\{x_i\}$ e $\{y_i\}$ 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 $Y_k = \sum_{i=1}^k [y_i - \langle y \rangle]$ com k também variando entre 1 e N ;
2. As séries X_k e Y_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 $\tilde{X}_{k,i}$ para a primeira série e $\tilde{Y}_{k,i}$ para a segunda com $i \leq k \leq (i + n)$ eliminando assim a tendência (detrended values);
4. Para cada caixa é calculado:
$$f_{DCCA}^2(n, i) = \frac{1}{1+n} \sum_{k=i}^{i+n} (X_k - \tilde{X}_{k,i})(Y_k - \tilde{Y}_{k,i})$$
5. Para todas as caixas de uma escala temporal a função $F_{DCCA}^2(n)$ é calculada por: $F_{DCCA}^2(n) = \frac{1}{N-n} \sum_{i=1}^{N-n} f_{DCCA}^2(n, i)$;
6. Para um número de diferentes escalas temporais (n), com valores possíveis entre $4 \leq n \leq \frac{N}{4}$, a função $F_{DCCA}^2(n)$ é calculada para encontrar a relação entre $F_{DCCA}^2(n) \times n$.

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


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

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

Resultados


Considerações Finais

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

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


19

 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.


20

 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.

21

 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.

22