



MODELO DE APRENDIZADO DE MÁQUINA BASEADO NO ρ_{DCCA} e DMC_x^2

Seminário de Qualificação

Discente: Fernando Ferraz Ribeiro

Orientador: Prof. Dr. Gilney Figueira Zebende

Coorientador: Prof. Dr. Juan Alberto Leyva Cruz

08/11/2024

PPGM UEFS - Feira de Santana, BA

Sumário

1. Introdução
2. Metodologia
3. Fundamentação Teórica
4. Resultados
5. Aprendizado de máquina
6. Referências

Introdução

Sistemas Complexos

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.

Sistemas Complexos e Ciência de Dados

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).

Séries Temporais

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).

Proposta

Esta pesquisa propõe a exploração das potencialidades dos coeficientes ρ_{DCCA} e DMC_x^2 como componentes de **Modelos de aprendizado de máquina(AM)**

Objetivo Principal

O objetivo principal desta pesquisa é: investigar as potencialidades dos coeficientes ρ_{DCCA} e DMC_x^2 como componentes na geração de **um modelo de aprendizado de máquina (AM)**.

Objetivos Gerais

1. Implementar um algoritmo computacional geral para calcular o DMC_x^2 para qualquer número de séries temporais.
2. desenvolver um pacote Python **modular** que permita **flexibilidade** e interoperabilidade com o ecossistema de ferramentas de **AM** disponíveis para a linguagem;
3. 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. As ferramentas de **AM** são extremamente competentes na resolução de problemas complexos, baseados em grandes conjuntos de dados;
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).

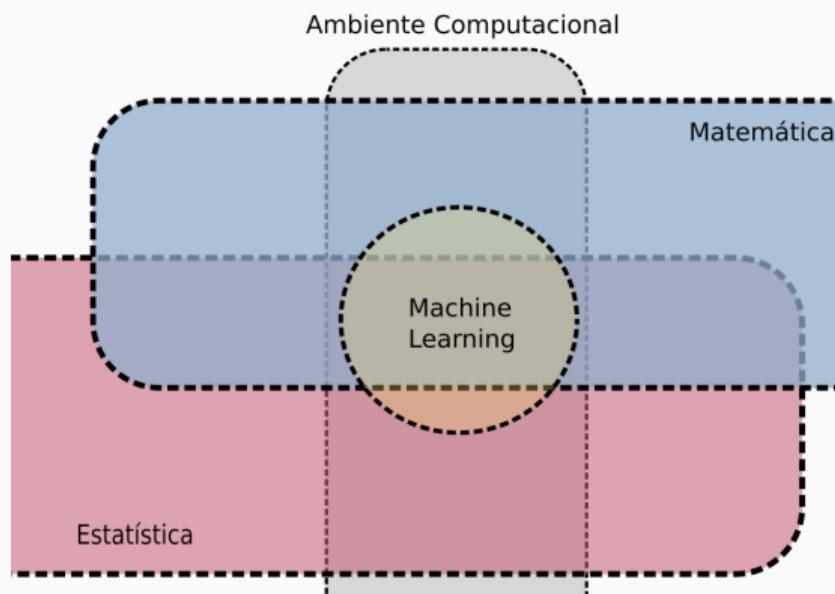
Hipótese

1. É possível criar uma modelo preditivo de **AM** que se beneficie dos coeficientes ρ_{DCCA} e DMC_x^2 .

Metodología

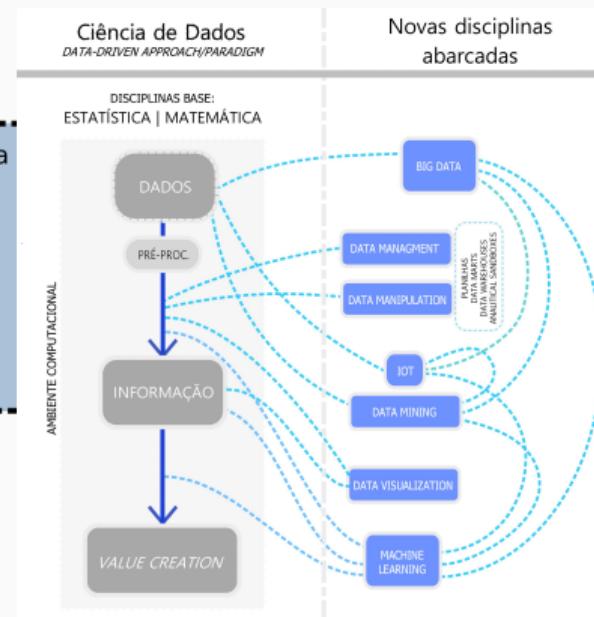
Aprendizado de Maquina - Machine Learning

Figure 1: Conceituação - AM



Fonte: Elaborada pelos autores

Figure 2: Diagrama conceitual - AM



Fonte: Elaborada pelos autores

Aprendizado de Maquina - caracterização

- P (performance)- Desempenho do algoritmo...
- T (task) - na execução de uma tarefa...
- E (experience) - através da experiência (dados).

(MITCHELL, 1997)

[⟨https://www.cs.cmu.edu/~ninemf/courses/601sp15/index.html⟩](https://www.cs.cmu.edu/~ninemf/courses/601sp15/index.html)

Princípios norteadores

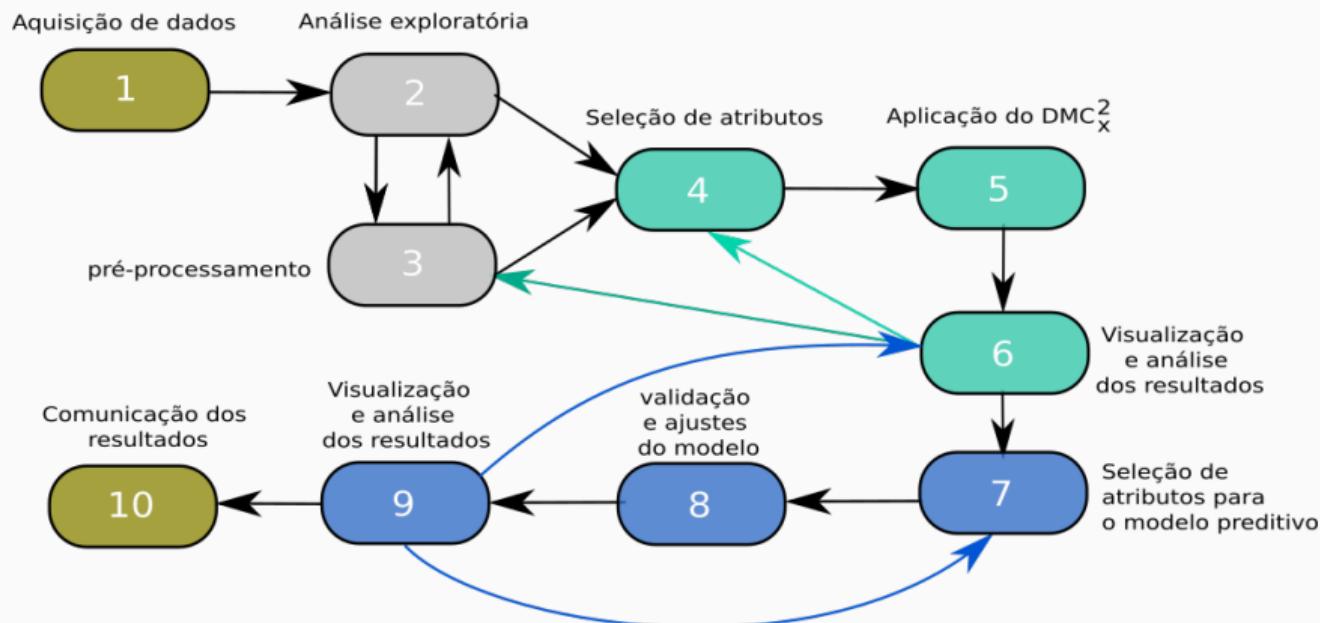
- Conheça os dados
- Entenda os algoritmos
- Desconfie dos resultados

Etapas

1. Implementação do algoritmo do DMC_x^2 ;
2. desenvolvimento da biblioteca;
3. elaboração da arquitetura do modelo;
4. Seleção de atributos;
5. Validação e ajustes do modelo;
6. Visualização e análise dos resultados.

Metodologia - Fluxograma

Figure 3: Fluxograma metodológico



Fonte: Elaborada pelos autores

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.

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.

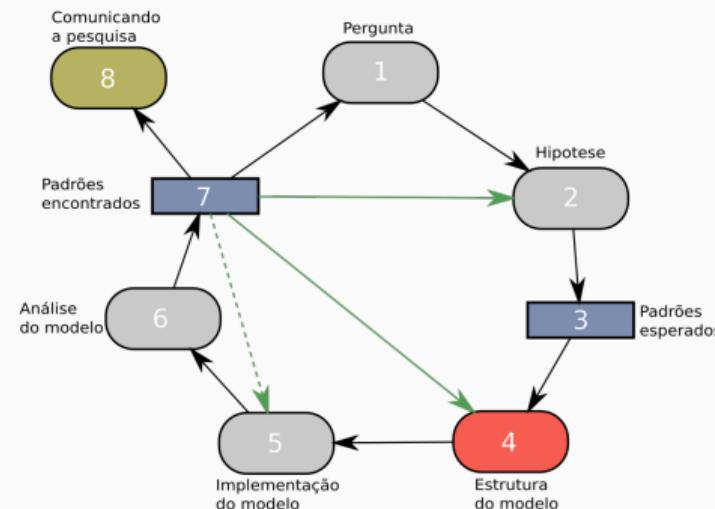
Utilizar estratégias de visualização de dados e comparar os resultados do DMC_x^2 com as seleções de variáveis

Seleção de variáveis para aplicação do modelo preditivo

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

Validação do Algoritmo de AM

Figure 4: Diagrama de Grimm e Railsback



Fonte: Elaborada pelos autores

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 $F_{DFA} \times n$

DCCA - (PODOBNIK; STANLEY, 2008)

1. Para duas séries temporais $\{x_{\alpha,i}\}$ e $\{x_{\beta,i}\}$ com i variando de 1 à N , as séries integradas $X_{\alpha,k}$ e $X_{\beta,k}$ são calculadas por $X_k = \sum_{i=1}^k [x_i - \langle x \rangle]$ com k também variando entre 1 e N ;
2. As séries $X_{\alpha,k}$ e $X_{\beta,k}$ 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 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 ;
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 - (ZEBENDE; SILVA, 2018)

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

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

$$\rho_{Y,X_i}(n)^T = [\rho_{Y,X_1}(n), \rho_{Y,X_2}(n), \dots, \rho_{Y,X_j}(n)] \quad (4)$$

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 *Journal of Statistical Software*.
- Implementação do modelo de AM usando DMC_x^2 .
- Artigo: validação do modelo AM.
- Artigo de aplicação.

Figure 5: (FILHO et al., 2023)

 ELSEVIER

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

Algoritmo registrado

Figure 6: Registro de Software



Algoritmo registrado - 01

$$\begin{aligned} DMC_x^2 &= \left(\rho_{X_2, X_3}^2 \times \rho_{Y, X_1}^2 - \rho_{Y, X_1}^2 + \rho_{X_1, X_3}^2 \times \rho_{Y, X_2}^2 - \rho_{Y, X_2}^2 \right. \\ &\quad + 2 \times \rho_{X_1, X_2} \times \rho_{Y, X_1} \times \rho_{Y, X_2} - 2 \times \rho_{X_1, X_3} \times \rho_{X_2, X_3} \times \rho_{Y, X_1} \\ &\quad + \rho_{X_1, X_2}^2 \times \rho_{Y, X_3}^2 - \rho_{Y, X_3}^2 + 2 \times \rho_{X_1, X_3} \times \rho_{Y, X_1} \times \rho_{Y, X_3} \\ &\quad - 2 \times \rho_{X_1, X_2} \times \rho_{X_2, X_3} \times \rho_{Y, X_1} \times \rho_{Y, X_3} \\ &\quad - 2 \times \rho_{X_1, X_2} \times \rho_{X_1, X_3} \times \rho_{Y, X_2} \times \rho_{Y, X_3} \\ &\quad \left. + 2 \times \rho_{X_2, X_3} \times \rho_{Y, X_2} \times \rho_{Y, X_3} \right) \Bigg/ \\ &\quad \left(\rho_{X_1, X_2}^2 + \rho_{X_1, X_3}^2 + \rho_{X_2, X_3}^2 - 2 \times \rho_{X_1, X_2} \times \rho_{X_1, X_3} \times \rho_{X_2, X_3}^{-1} \right) \end{aligned} \tag{5}$$

Multi Cross-correlation Analysis in a Multi-channel EEG applied in Motor Activity (Real/Imaginary)

[⟨https://255ribeiro.github.io/Multi_Cross-correlation_EEG/⟩](https://255ribeiro.github.io/Multi_Cross-correlation_EEG/)

Artigo 02

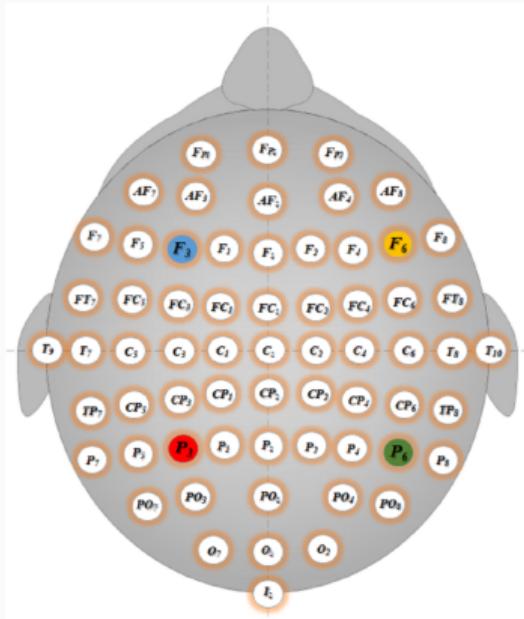


Figure 7: Posição dos canais de EEG e canais utilizados

Artigo 02

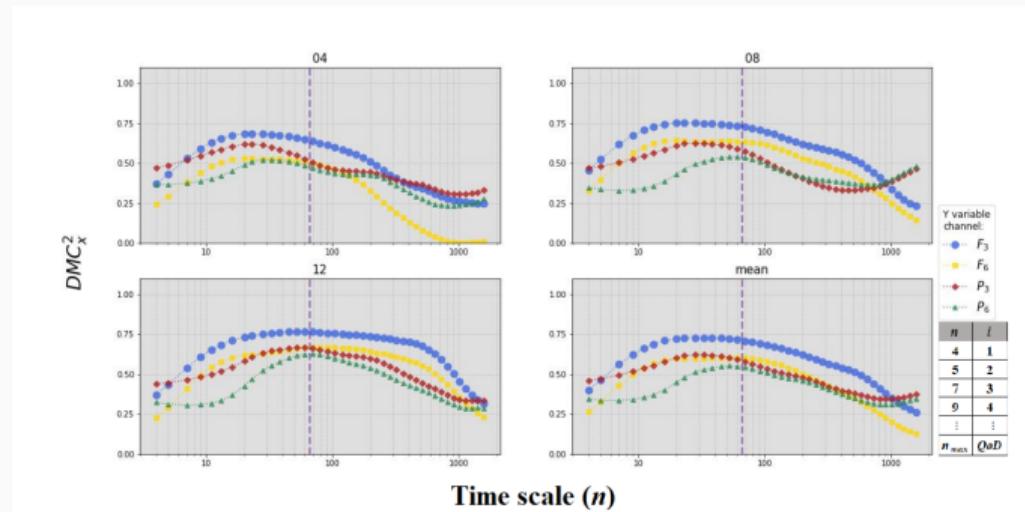


Figure 8: DMC_x^2 as a function of time scale n . Here is showing the results for subject S014 recordings for Task 2, presenting experiments 04, 08, 12 and the mean values for these experiments. The vertical line represents $n = 67$ and QoD is the total amount of time scales involved in DMC_x^2 calculations.

Artigo 02

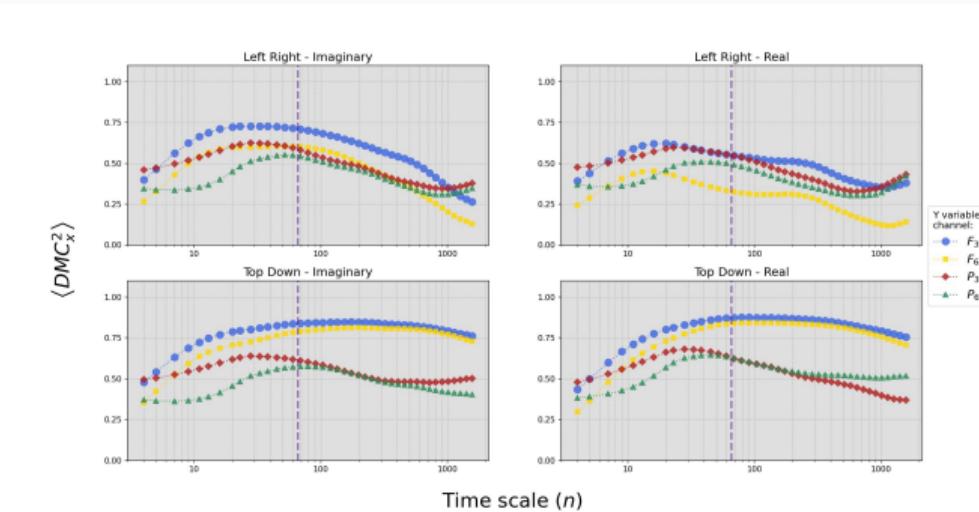


Figure 9: Mean values of $DMC_x^2 \times n$ for all Tasks: Left/Right (Imaginary), Left/Right (Real), Top/Down (Imaginary), and Top/Down (Real) for the S014 subject.

Artigo 02

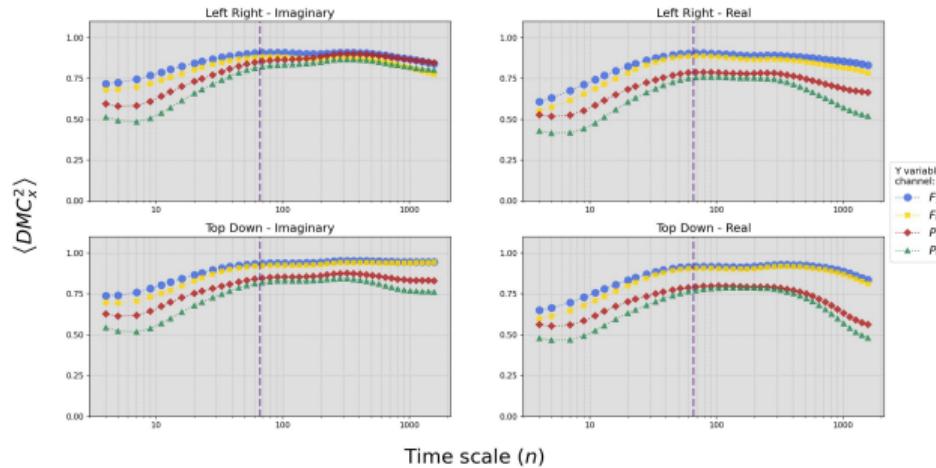


Figure 10: Mean values of $DMC_x^2 \times n$ for all Tasks: Left/Right (Imaginary), Left/Right (Real), Top/Down (Imaginary), and Top/Down (Real) for the S036 subject.

Artigo 02

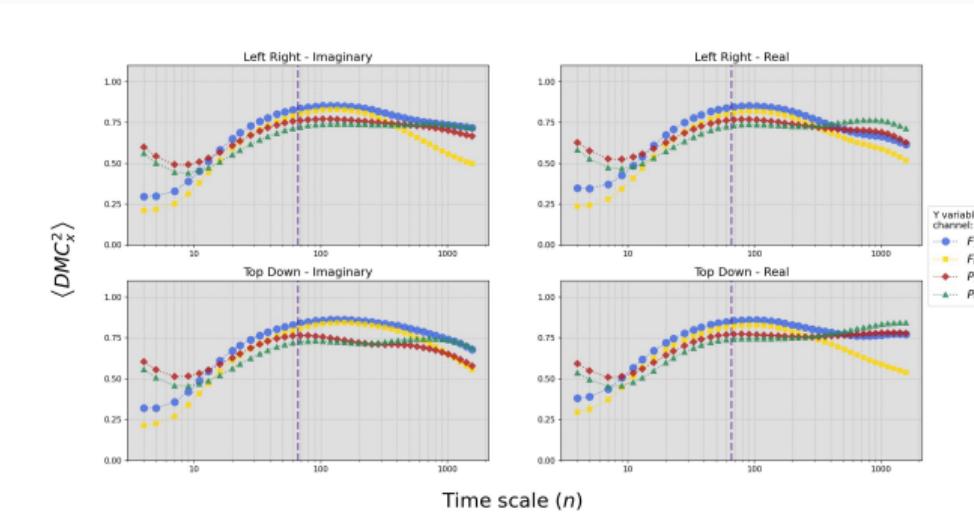


Figure 11: Mean values of $DMC_x^2 \times n$ for all Tasks: Left/Right (Imaginary), Left/Right (Real), Top/Down (Imaginary), and Top/Down (Real) for the S039 subject.

Artigo 02

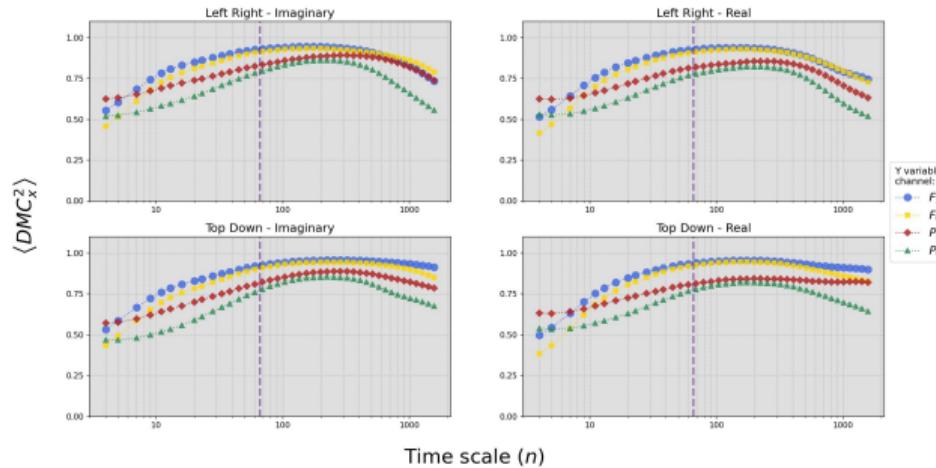


Figure 12: Mean values of $DMC_x^2 \times n$ for all Tasks: Left/Right (Imaginary), Left/Right (Real), Top/Down (Imaginary), and Top/Down (Real) for the S078 subject.

Artigo 02

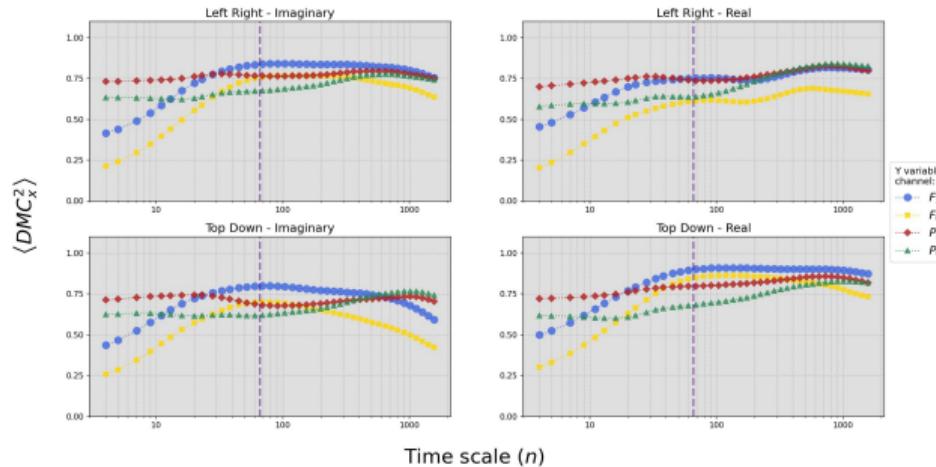


Figure 13: Mean values of $DMC_x^2 \times n$ for all Tasks: Left/Right (Imaginary), Left/Right (Real), Top/Down (Imaginary), and Top/Down (Real) for the S099 subject.

Artigo 02

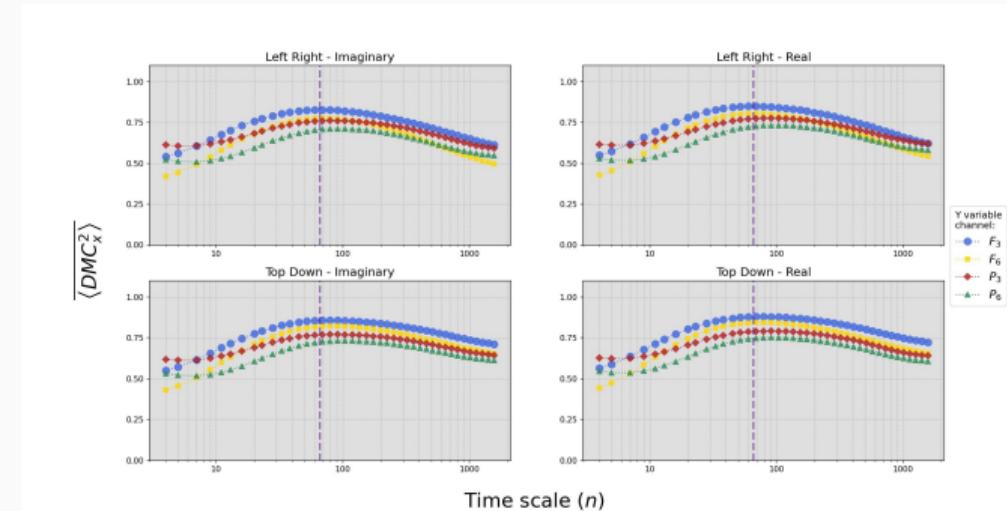


Figure 14: $DMC_x^2 \times n$ mean global for all Subjects and Tasks: Left/Right (Imaginary), Left/Right (Real), Top/Down (Imaginary), and Top/Down (Real).

Artigo 02

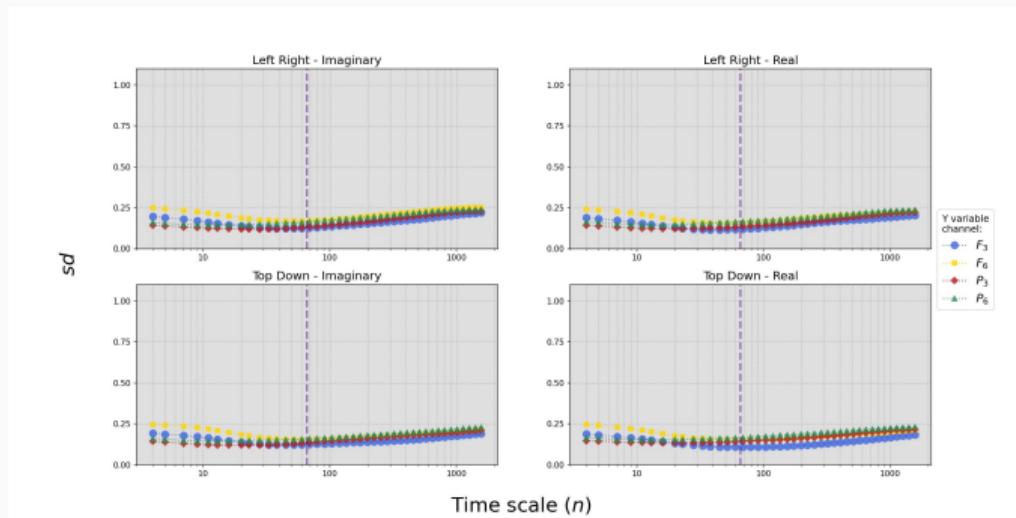


Figure 15: Standard deviation, *sd*, of the global mean for all Subjects and Tasks.

Artigo 02

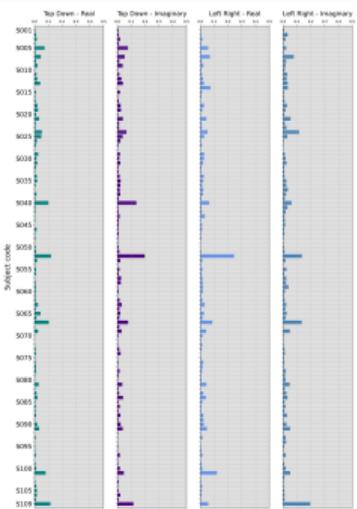


Figure 16: MSE
for the Channel F_3 .

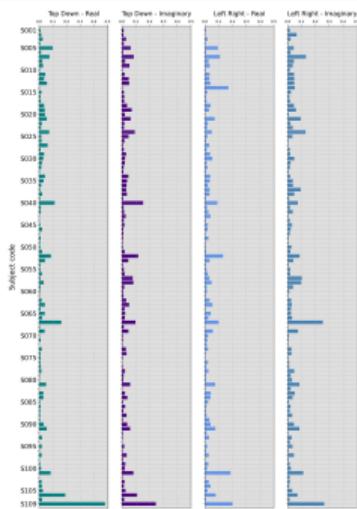


Figure 17: MSE
for the Channel F_6 .

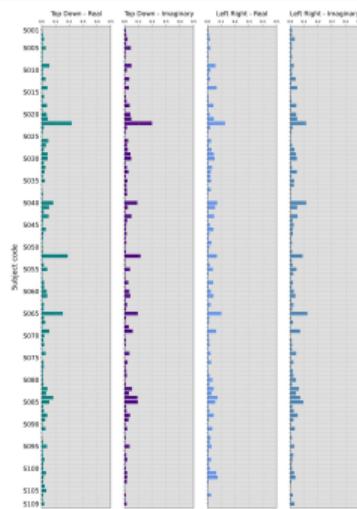


Figure 18: MSE
for the Channel P_3 .

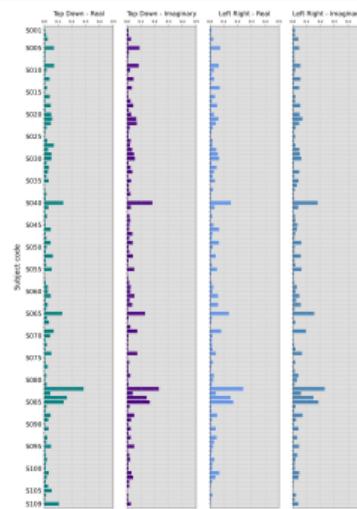


Figure 19: MSE
for the Channel P_6 .

Algoritmo - biblioteca Python/Zig

```
pip install zebende  
(https://pypi.org/project/zebende/)
```

Algoritmo - biblioteca Python/Zig

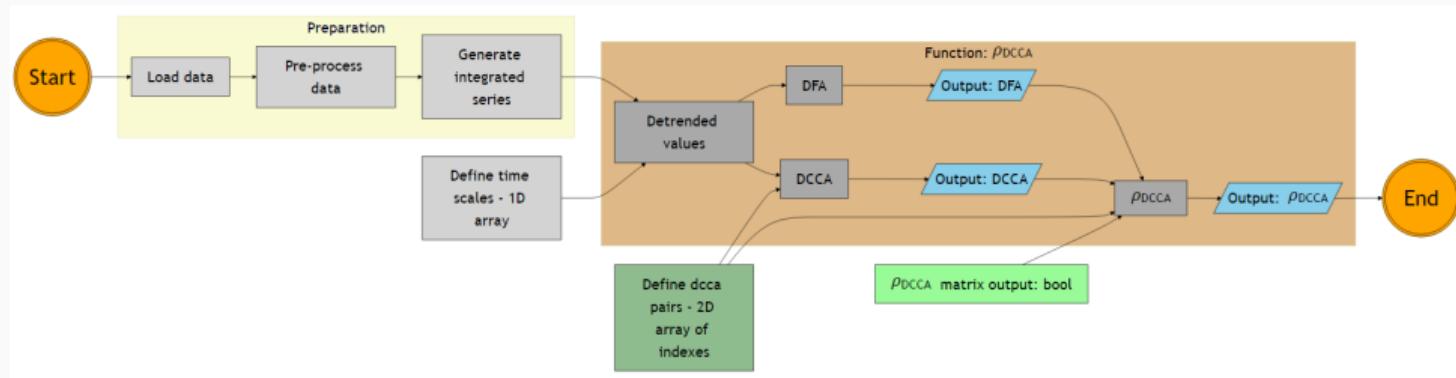


Figure 20: ρ_{DCCA} flowchart.

Algoritmo - biblioteca Python/Zig

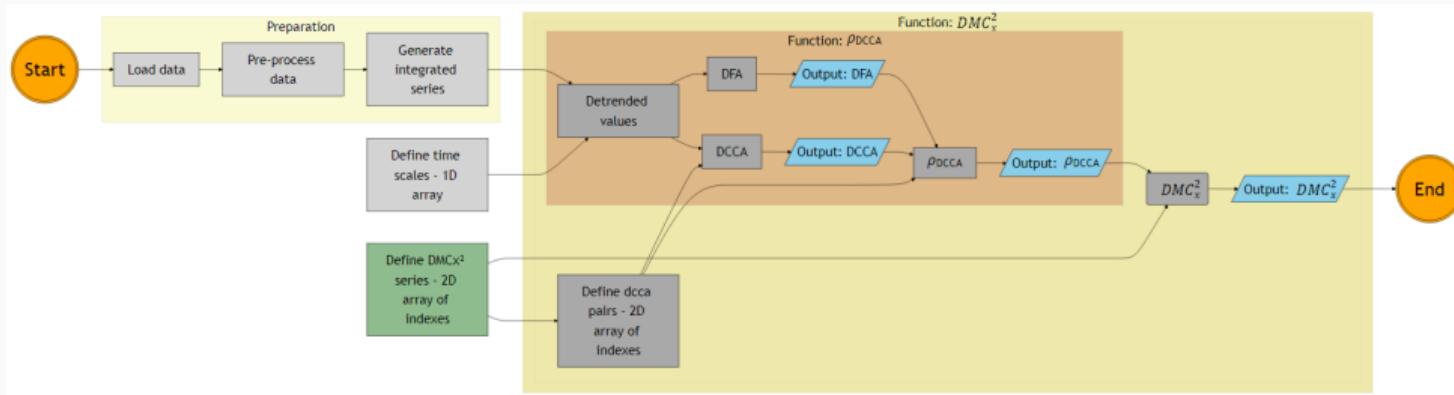


Figure 21: DMC_x^2 flowchart.

Aprendizado de máquina

Seleção de variáveis

Seleção de variáveis

A seleção de características é o processo de reduzir o número de variáveis de entrada em um conjunto de dados para melhorar o desempenho e a eficiência de um modelo de aprendizado de máquina. É essencial selecionar as características mais relevantes, pois características excessivas ou irrelevantes podem levar ao *overfitting*, à diminuição da precisão do modelo e ao aumento dos custos computacionais.

Inspiração - ganho de informação

$$H(Y) = - \sum_{i=1}^k p_i \log_2 p_i \quad (6)$$

$$IG(A, Y) = H(Y) - \sum_{v \in values(A)} \frac{|N_v|}{|N|} H(Y|A=v) \quad (7)$$

Proposta

$$\max_{(k=1,n)} = |\rho_{DCCA}(Y, x_k)| - DMC_x^2(x_k : X_{i \neq k}) \quad (8)$$

$$\max_{(k=1,n-1)} = DMC_x^2(Y : x_{k1}, x_{k2}) - DMC_x^2(x_{k1} : X_{i \neq \{k1,k2\}}) \quad (9)$$

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.

8

 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.

7

 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.

32

 MITCHELL, T. M. *Machine Learning*. [S.I.]: McGraw-Hill, 1997. ISBN 978-0-07-042807-2.

16

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

26

 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.

27

 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.

12

 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.

28

 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.

29

Obrigado
