



Ricardo Augusto Manfredini
Geraldo Nunes Corrêa
Bruno Rodrigues de Oliveira
Suellen Teixeira Zavadzki de Pauli
Organizadores

Aplicações de Machine Learning



2021

Ricardo Augusto Manfredini
Geraldo Nunes Corrêa
Bruno Rodrigues de Oliveira
Suellen Teixeira Zavadzki de Pauli
Organizadores

Aplicações de Machine Learning



Pantanal Editora

2021

Copyright© Pantanal Editora

Editor Chefe: Prof. Dr. Alan Mario Zuffo

Editores Executivos: Prof. Dr. Jorge González Aguilera e Prof. Dr. Bruno Rodrigues de Oliveira

Diagramação: A editora. **Diagramação e Arte:** A editora. **Imagens de capa e contracapa:** Canva.com. **Revisão:** O(s) autor(es), organizador(es) e a editora.

Conselho Editorial

Grau acadêmico e Nome

Prof. Dr. Adaylson Wagner Sousa de Vasconcelos
Prof. Msc. Adriana Flávia Neu
Prof. Dra. Albys Ferrer Dubois
Prof. Dr. Antonio Gasparetto Júnior
Prof. Msc. Aris Verdecia Peña
Prof. Arisleidis Chapman Verdecia
Prof. Dr. Arinaldo Pereira da Silva
Prof. Dr. Bruno Gomes de Araújo
Prof. Dr. Caio Cesar Enside de Abreu
Prof. Dr. Carlos Nick
Prof. Dr. Claudio Silveira Maia
Prof. Dr. Cleberton Correia Santos
Prof. Dr. Cristiano Pereira da Silva
Prof. Ma. Dayse Rodrigues dos Santos
Prof. Msc. David Chacon Alvarez
Prof. Dr. Denis Silva Nogueira
Prof. Dra. Denise Silva Nogueira
Prof. Dra. Dennyura Oliveira Galvão
Prof. Dr. Elias Rocha Gonçalves
Prof. Me. Ernane Rosa Martins
Prof. Dr. Fábio Steiner
Prof. Dr. Fabiano dos Santos Souza
Prof. Dr. Gabriel Andres Tafur Gomez
Prof. Dr. Hebert Hernán Soto Gonzáles
Prof. Dr. Hudson do Vale de Oliveira
Prof. Msc. Javier Revilla Armesto
Prof. Msc. João Camilo Sevilla
Prof. Dr. José Luis Soto Gonzales
Prof. Dr. Julio Cezar Uzinski
Prof. Msc. Lucas R. Oliveira
Prof. Dra. Keyla Christina Almeida Portela
Prof. Dr. Leandris Argentele-Martínez
Prof. Msc. Lidiene Jaqueline de Souza Costa Marchesan
Prof. Dr. Marco Aurélio Kistemann
Prof. Msc. Marcos Pisarski Júnior
Prof. Dr. Marcos Pereira dos Santos
Prof. Dr. Mario Rodrigo Esparza Mantilla
Prof. Msc. Mary Jose Almeida Pereira
Prof. Msc. Núbia Flávia Oliveira Mendes
Prof. Msc. Nila Luciana Vilhena Madureira
Prof. Dra. Patrícia Maurer
Prof. Msc. Queila Pahim da Silva
Prof. Dr. Rafael Chapman Auty
Prof. Dr. Rafael Felipe Ratke
Prof. Dr. Raphael Reis da Silva
Prof. Dr. Renato Jaqueto Goes
Prof. Dr. Ricardo Alves de Araújo
Prof. Dra. Sylvana Karla da Silva de Lemos Santos
Prof. Dr. Wéverson Lima Fonseca
Prof. Msc. Wesclen Vilar Nogueira
Prof. Dra. Yilan Fung Boix
Prof. Dr. Willian Douglas Guilherme

Instituição

OAB/PB
Mun. Faxinal Soturno e Tupanciretã
UO (Cuba)
IF SUDESTE MG
Facultad de Medicina (Cuba)
ISCM (Cuba)
UFESSPA
UEA
UNEMAT
UFV
AJES
UFGD
UEMS
IFPA
UNICENTRO
IFMT
UFMG
URCA
ISEPAM-FAETEC
IFG
UEMS
UFF
(Colômbia)
UNAM (Peru)
IFRR
UCG (México)
Mun. Rio de Janeiro
UNMSM (Peru)
UFMT
Mun. de Chap. do Sul
IFPR
Tec-NM (México)
Consultório em Santa Maria
UFJF
UEG
FAQ
UNAM (Peru)
SEDUC/PA
IFB
IFPA
UNIPAMPA
IFB
UO (Cuba)
UFMS
UFPI
UFG
UEMA
IFB
UFPI
FURG
UO (Cuba)
UFT

Conselho Técnico Científico

- Esp. Joacir Mário Zuffo Júnior

- Esp. Maurício Amormino Júnior
- Esp. Tayronne de Almeida Rodrigues
- Lda. Rosalina Eufrausino Lustosa Zuffo

Ficha Catalográfica

Dados Internacionais de Catalogação na Publicação (CIP)
(eDOC BRASIL, Belo Horizonte/MG)

A642 Aplicações de machine learning [livro eletrônico] / Organizadores Ricardo Augusto Manfredini... [et al.]. – Nova Xavantina, MT: Pantanal, 2021. 55 p. : il.

Formato: PDF

Requisitos de sistema: Adobe Acrobat Reader

Modo de acesso: World Wide Web

ISBN 978-65-81460-20-4

DOI <https://doi.org/10.46420/9786581460204>

1. Inteligência artificial. 2. Redes neurais. 3. Aprendizado de máquina. I. Manfredini, Ricardo Augusto. II. Corrêa, Geraldo Nunes. III. Oliveira, Bruno Rodrigues de. IV. Pauli, Suellen Teixeira Zavadzki de.

CDD 006.3

Elaborado por Maurício Amormino Júnior – CRB6/2422



Nossos e-books são de acesso público e gratuito e seu download e compartilhamento são permitidos, mas solicitamos que sejam dados os devidos créditos à Pantanal Editora e também aos organizadores e autores. Entretanto, não é permitida a utilização dos e-books para fins comerciais, exceto com autorização expressa dos autores com a concordância da Pantanal Editora.

Pantanal Editora

Rua Abaete, 83, Sala B, Centro. CEP: 78690-000.
Nova Xavantina – Mato Grosso – Brasil.
Telefone (66) 99682-4165 (Whatsapp).
<https://www.editorapantanal.com.br>
contato@editorapantanal.com.br

APRESENTAÇÃO

Este livro aborda cinco diferentes contextos em que as técnicas de aprendizado de máquina podem ser utilizadas, servindo como referência prática em diferentes abordagens, tais como: previsão de consumo de energia elétrica, previsão do valor do preço do petróleo, classificação de arritmias cardíacas e método para a seleção automática de artigos.

Estas aplicações são explanadas pelos autores e diferentes técnicas de aprendizado de máquina são utilizadas, dentre elas: redes neurais (Dense, convolucional, recorrentes, Perceptron multicamadas, Elman e Jordan), Naïve Bayes e mineração de textos. Para a aplicação os autores empregam os softwares (línguas de programação) Python e R. Com o objetivo de apresentar aplicações de algumas das técnicas com destaque na atualidade, primeiramente há um descritivo a respeito de cada abordagem e então, são reportados os treinamentos dos modelos de aprendizado de máquina e os resultados. As técnicas incluem classificação, regressão e também mineração de texto. As possíveis abordagens para os problemas relatados não se restringem às técnicas utilizadas, mas a intenção é motivar o leitor a explorar aplicações na área de aprendizado de máquina.

O livro é destinado a profissionais, estudantes, pesquisadores e demais interessados no tema aprendizado de máquina, estatística e áreas relacionadas a fim de colaborar com a explanação de possibilidades de aplicações destas técnicas em contextos diversos. Presume-se que o leitor esteja familiarizado com os conceitos básicos de machine learning, álgebra linear, probabilidade, e análise de algoritmos. A intenção com esta obra é, primordialmente, explicar as possibilidades de aplicação dos algoritmos elencados.

Nos capítulos 1 e 2 as técnicas de redes neurais artificiais são aplicadas para a previsão de preço do valor de petróleo e consumo de energia elétrica, respectivamente. Tais técnicas tratam de modelos computacionais inspirados no sistema nervoso central de um animal. Elas são apresentadas como um sistema de neurônios interconectados, que podem computar valores de entradas, simulando o comportamento de redes neurais biológicas. Na primeira abordagem a autora utiliza as redes Perceptron multicamadas, Elman e Jordan, já na segunda o autor faz uso das redes híbridas Dense, convolucional e recorrente.


No capítulo 3 é tratada a classificação de arritmias cardíacas e, além da técnica de Naïve Bayes também é utilizada a Transformada Wavelet, que é uma transformada integral que utiliza função wavelets que são capazes de decompor determinado sinal (dado) em diferentes escalas. Além desta, também é empregada uma técnica de Ensemble, que encapsula os modelos obtidos por vários algoritmos de aprendizagem a fim de obter uma única previsão global. Por fim, no capítulo 4 é abordado um método para a seleção automática de artigos. Para isto, é utilizado mineração de texto, que trata do processo de obtenção de informações importantes de um texto.

Os organizadores

SUMÁRIO

Apresentação	4
Capítulo I.....	6
Predição diária do preço de petróleo WTI	6
Capítulo II	15
Redes Neurais Artificiais Híbridas Para a Predição de Consumo de Energia Elétrica	15
Capítulo III.....	32
Reconhecimento de padrões de arritmias cardíacas no Eletrocardiograma (ECG) empregando Transformada Wavelet e o classificador Naïve Bayes	32
Capítulo IV	45
Uso da mineração de textos na análise exploratória de artigos científicos.....	45
Índice Remissivo	54
Sobre os organizadores.....	55

Predição diária do preço de petróleo WTI

 10.46420/9786581460204cap1

Suellen Teixeira Zavadzki de Pauli^{1*} 

INTRODUÇÃO

Há situações no cotidiano em que nos deparamos com análises de dados obtidos ao longo do tempo, de forma sequencial, sendo estas informações coletadas em intervalos de horas, dias, meses ou outras medidas temporais. Alguns exemplos deste tipo de análise são a previsão de demanda semanal em uma indústria, valores máximos em ações na bolsa de valores a cada minuto, temperatura máxima do ar a cada hora, preço histórico diário do petróleo WTI.

Uma característica intrínica deste tipo de dados é que possuem dependência entre as observações. Neste sentido, há modelos estatísticos bastante sólidos na literatura que concentram-se na análise desta dependência. Tais modelos assumem que a série temporal segue um modelo estocástico com forma conhecida. Inicialmente, a classe de modelos estacionários obteve grande destaque, estes assumem que o processo permanece em equilíbrio estatístico, com propriedades probabilísticas que não mudam ao longo do tempo, variando sobre um nível médio constante fixo e com variância constante. Posteriormente, outros modelos foram propostos, dado o desafio da não estacionariedade das séries (Box et al., 2015).

Um modelo que descreve a estrutura de probabilidade de uma sequência de observações é denominado de processo estocástico. Alguns processos estocásticos estacionários são bastante tradicionais na literatura, como o autorregressivo (*autoregressive*, AR), o média móvel (*moving average*, MA) e o autorregressivo-média móvel (*autoregressive-moving average*, ARMA). Com a necessidade de modelos para procesos não estacionários, foram desenvolvidos também os autorregressivos de média móvel integrada (*autoregressive integrated moving average*, ARIMA). Além destes, há outras variações, para mais detalhes a respeito destes modelos ver (Brockwell et al., 2016; Box et al., 2015).

O mecanismo que gera a série temporal é frequentemente pensado para considerar três componentes: sazonal, tendência e um erro aleatório. A presença de tendência e variação sazonal pode ser difícil de estimar e remover, porém, isto é bastante importante nas abordagens estatísticas tradicionais pois é necessária a especificação para assumir um modelo de série temporal. O uso das redes neurais neste contexto não necessita de especificações das relações entre variáveis de entrada e de saída. As camadas ocultas de uma rede neural removem a necessidade de pré-especificar da natureza do mecanismo

¹ Doutoranda na Universidade Federal do Paraná

* Autor(a) correspondente: est.suellen@gmail.com

de geração de dados. Isso ocorre porque podem aproximar funções de decisão extremamente complexas (Lewis, 2017). Por outro lado, isto não torna menos importantes os modelos tradicionais, dado que a interpretabilidade dos parâmetros fica mais complexa com o uso das RNAs, assim, a escolha do modelo deve ser feita alinhada com o objetivo da análise.

A análise de séries temporais tem aplicação nos mais diversos campos, como ciências ambientais, medicina, ciências sociais, negócios, indústria, governo, economia, entre outros (Montgomery et al., 2008). A fim de ilustrar a variedade de aplicações de diferentes modelos de redes neurais em dados de estrutura temporal, são descritos alguns exemplos da literatura. Ali et al. (2017) aplicaram a rede neural perceptron multicamadas para a previsão de secas, enquanto Dudek (2016) aplicou tal rede para a previsão de preços de eletricidade. Devadoss (2013) fez a previsão para o mercado de ações, também utilizando a rede perceptron multicamadas. Wu et al. (2019) utilizaram as redes Elman e Jordan para previsão de casos de brucelose, uma doença infecciosa. Lee et al. (2018) aplicaram a rede neural recorrente Elman para prever e analisar uma série temporal de consumo de energia elétrica. Wang et al. (2021) aplicou a rede Elman no mercado de ações. Šestanović aplicou a rede Jordan para a previsão de inflação. Neste capítulo a aplicação utilizada é a série diária de petróleo WTI.

Para a aplicação, deseja-se um modelo capaz de realizar a previsão diária de preço de petróleo baseada em informações históricas. Algumas perguntas poderiam surgir neste momento, como qual modelo poderia ser utilizado para solução deste problema e quantos períodos de tempo seriam suficientes para realizar uma boa previsão.

Não há uma única resposta para estas questões, afinal, há inúmeros modelos que podem ser aplicado para a previsão de séries históricas, como os exemplos já comentados. É importante saber que antes de se aplicar um modelo deve-se compreender a estrutura do mesmo, há determinados modelos que não podem ser utilizados para variáveis dependentes, que é o caso de séries históricas. Para esta aplicação a técnica escolhida foi a Rede Neural Artificial (RNA), mais especificamente a Perceptron multicamadas, Elman e Jordan.

A divisão das seções está descrita a seguir. A primeira corresponde a materiais e métodos e está dividida na subseção base de dados, na qual o conjunto de dados utilizado é descrito, na subseção Redes Neurais Artificiais, em que há uma breve descrição a respeito da técnica e na Raiz do Erro Quadrático Médio, que contempla a técnica de comparação dos modelos. Na seção metodologia está descrito os detalhes da análise e, por fim, em resultados e discussão está a comparação entre as técnicas.

MATERIAL E MÉTODOS

Base de dados

A série diária de petróleo WTI foi coletada por meio do *software* R (R Core Team, 2021), com o pacote *ipeadata*, os dados são alimentados com informação do *U.S. Energy Information Administration* (EIA). As unidades estão em US\$ por barril de petróleo.

Redes Neuais Artificiais

As Redes Neurais Artificiais são técnicas de aprendizado de máquina que possuem a capacidade de aprendizagem, a qual ocorre nos processos iterativos dos ajustes dos pesos. De acordo com as características do conjunto de exemplos usados no treinamento, pode-se ter uma aprendizagem supervisionada ou não supervisionada, sendo que a primeira ocorre quando cada exemplo apresenta uma saída esperada e quando não existe uma saída esperada a aprendizagem é chamada não supervisionada (Braga et al., 2000).

Diferentes modelos de RNAs já foram desenvolvidos e para esta abordagem serão utilizadas as redes Perceptron multicamadas, Elman e Jordan, todas com aprendizagem supervisionada.

As RNAs foram desenvolvidas com inspiração na neurobiologia. O neurônio é uma unidade de processamento de informação, o qual é essencial para a operação de uma rede neural. As variáveis na rede são consideradas como um conjunto sinais de entrada, cada qual com um determinado peso. Essas informações passam por um somador, o que constitui em um combinador linear. Após isto passa por uma função de ativação, para que seja restringida a saída do neurônio (Fausett, 2006; Haykin, 1999). A Figura 1 ilustra o funcionamento de uma rede, é possível que haja mais de uma camada oculta, porém, aqui vamos nos restringir a uma somente.

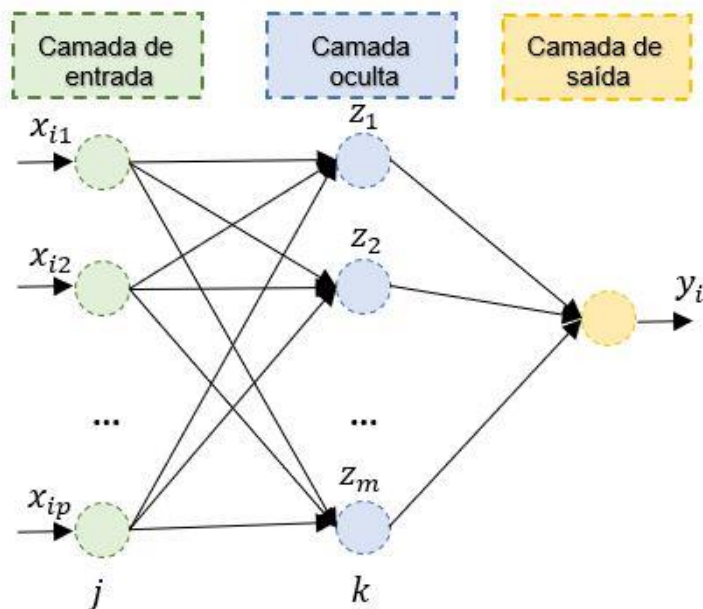


Figura 1. Rede Neural Perceptron multicamadas. Adaptado de (de Pauli et al., 2020).

Na rede neural Perceptron multicamadas basicamente há uma camada de entrada com p neurônios, k neurônios na camada oculta e o na de saída, a indexação i corresponde período da série que está passando pela rede. A saída y_i pode ser calculada conforme $y_{io} = g[\sum_{k=1}^m w_{k,o} z(\sum_{l=1}^p w_{j,k} x_{ij})]$, onde $w_{j,k}$ realiza a ponderação do neurônio de entrada j para o oculto k , considerando $j = [0, 1, \dots, p]$

e $k = [0, 1, \dots, m]$ e $w_{k,o}$ o faz de k para a saída o , x_{ij} representa o valor da variável de entrada j no período i , $z(\cdot)$ e $g(\cdot)$ são as funções de ativação da camada oculta e de saída, respectivamente. Para o ajuste dos parâmetros do modelo ocorre, em geral, busca-se minimizar a função de erro do modelo pelo método gradiente (Bishop, 1994; Haykin, 1999).

Aqui foi feita uma breve introdução para a rede Perceptron multicamadas, com a qual é possível ter-se uma ideia do funcionamento das RNAs. Além destas há outras redes, dentre as quais a Elman e a Jordan, que são redes recorrentes e são também utilizadas nesta aplicação, para maiores detalhes destas redes ver em (Elman, 1990; Jordan, 1986).

Raíz do Erro Quadrático Médio

Para a avaliação dos modelos foi utilizado a raiz do erro quadrático médio (*Root Mean Square Error*, RMSE), sendo representado por $\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$, em que y_i corresponde ao valor observado i na série e \hat{y}_i ao valor estimado pelo modelo, com $i = [0, 1, \dots, N]$, sendo N o total de observações avaliadas.

METODOLOGIA

Para a aplicação foi coletado o preço diário do período de 01 de janeiro de 2000 até 31 de março de 2020. Os dados foram divididos em três partes: treinamento, teste e validação. A primeira parte, de 01 de janeiro de 2000 até 25 de março de 2014, o que corresponde a 75% dos dados. A segunda, com 15% dos dados, inicia após o treinamento e finaliza em 23 de março de 2017. A terceira, com 15% dos dados inicia após o teste e finaliza em 31 de março de 2020. Na Figura 2 é possível visualizar a série histórica assim como as divisões descritas.

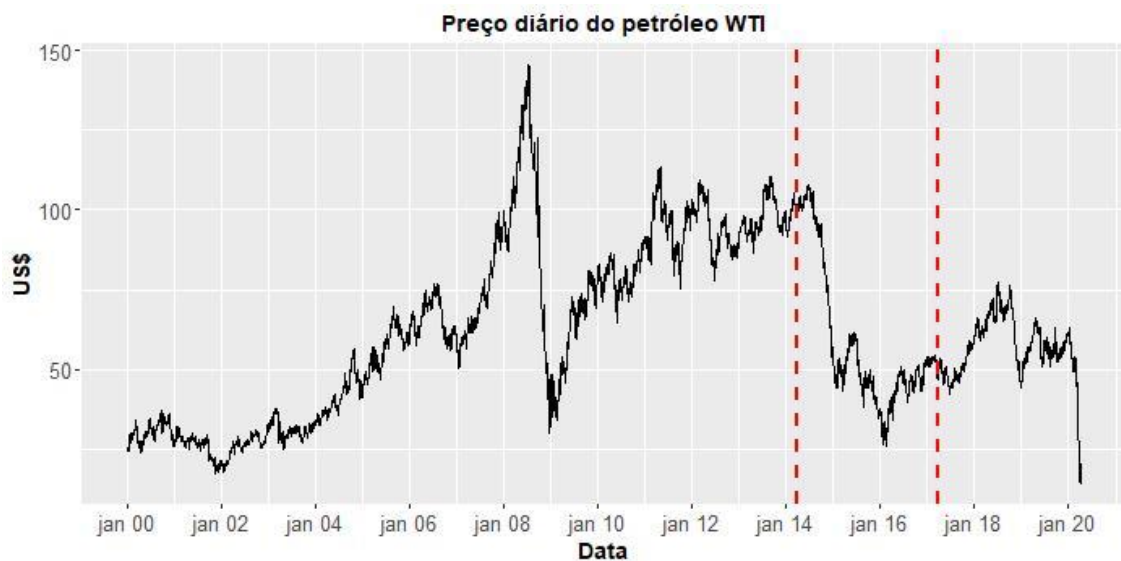


Figura 2. Série histórica de preço do petróleo WTI. Fonte: A autora.

Para esta aplicação foram consideradas cinco variáveis históricas na camada de entrada do modelo e na camada de saída 6 neurônios. Na Figura 3 é ilustrada a forma de construção da base de dados, em que com valores de preço coletados em quinze dias é possível construir cinco linhas de base, cada qual com cinco valores correspondentes a camada de entrada (t-5, t-4, t-3, t-2, t-1), os quais representam às cinco variáveis históricas e seis à de saída (t, t+1, t+2, t+3, t+4, t+5), que são correspondentes aos seis dias posteriores.

01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5				
	t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5			
		t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5		
			t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5	
				t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5

Figura 3. Construção da base histórica. Fonte: A autora.

Antes de iniciar o treinamento dos modelos, é interessante verificar as relações entre as variáveis. Neste caso, trata-se de dados históricos diários e portanto já poderia se imaginar que haveria forte relação entre as variáveis, porém, isto nem sempre é evidente. Na Figura 4 estão os valores correspondentes as correlações entre as variáveis. É fácil perceber que esta relação é bastante alta e decresce lentamente conforme os dias se afastam, mesmo observando t-5 com t+5, por exemplo, em que tem-se 10 dias de diferença a correlação ainda é bastante alta.

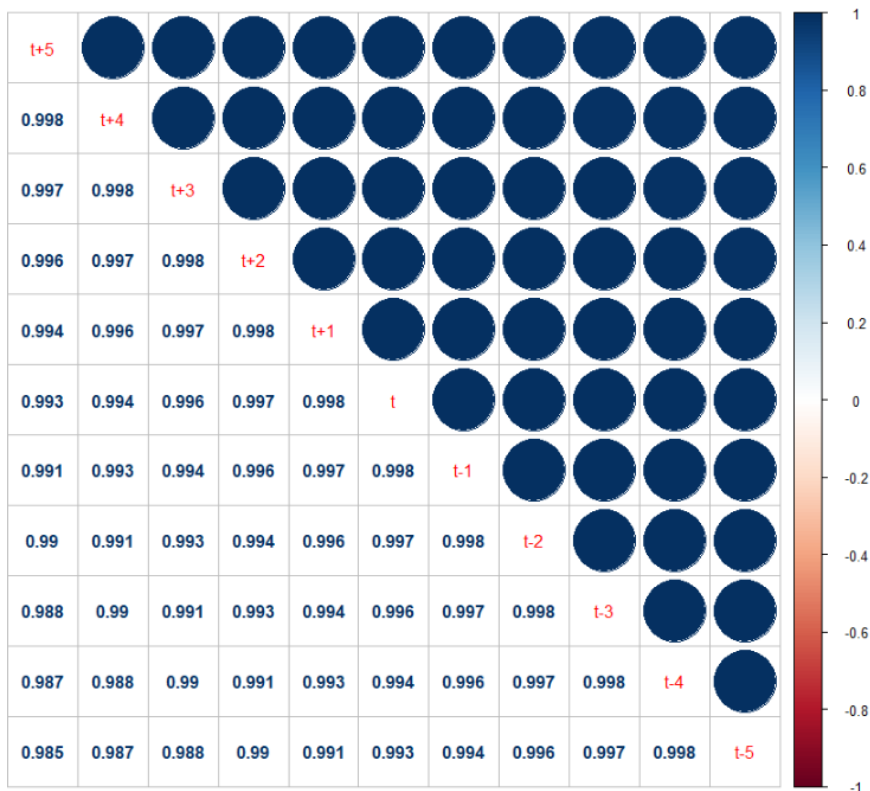


Figura 4. Correlação entre as variáveis. Fonte: A autora.

Para o treinamento, os dados foram normalizados entre 0 e 1. Para fins de comparação, taxa de aprendizagem e o número de iterações foi mantido igual para todos os modelos. As redes Elman, Jordan e Perceptron multicamadas foram treinadas com as possibilidades de 1 a 5 neurônios na camada oculta e em cada uma destas tentativas com 1 a 5 variáveis na camada de entrada. Para a rede Perceptron multicamadas foram testadas as possibilidades de função de ativação Sigmóide logística e tangente hiperbólica.

RESULTADOS E DISCUSSÃO

Após o período de treinamento, realizado para as respostas de 6 neurônios na camada de saída, os modelos foram avaliados quanto ao valor de raiz do erro quadrático médio, os resultados obtidos constam no mapa de calor da Figura 5. É possível verificar em vermelho os melhores valores obtidos, a rede Perceptron multicamadas em ambas as opções de função de ativação obtiveram melhores resultados se comparados com as redes Elman e Jordan para esta aplicação.

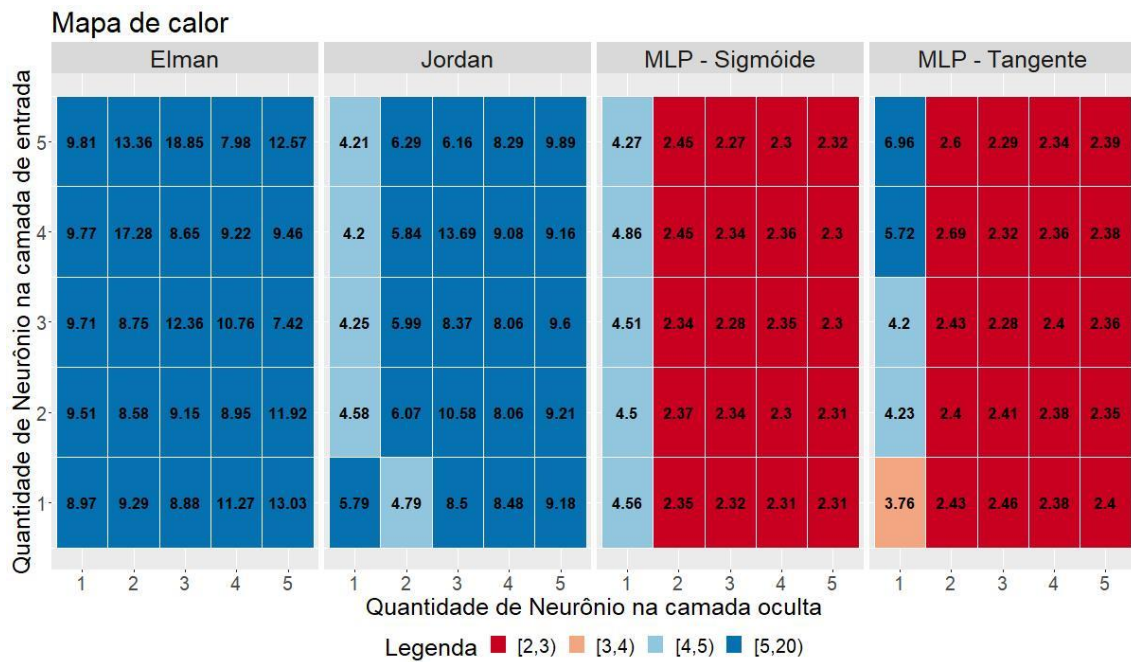


Figura 5. Valores de RMSE em cada configuração de redes no período de teste. Fonte: A autora.

Para a rede Elman o menor valor de RMSE foi de 7,42, com 3 neurônios na camada de entrada e 5 na camada oculta. A rede Jordan, com menor RMSE de 4,20, obteve este resultado com 4 neurônios na camada de entrada e 1 na camada oculta. Na rede Perceptron multicamadas, com a função de ativação Sigmóide, o melhor resultado obtido foi de 2,27, com a função de ativação Tangente foi de 2,28.

Para as possibilidades de rede avaliadas, o melhor resultado encontrado foi da rede Perceptron multicamadas com função de ativação Sigmóide. Com a Figura 6 é possível verificar a previsão para 1 a 6 dias considerando o melhor modelo encontrado em comparação ao observado.

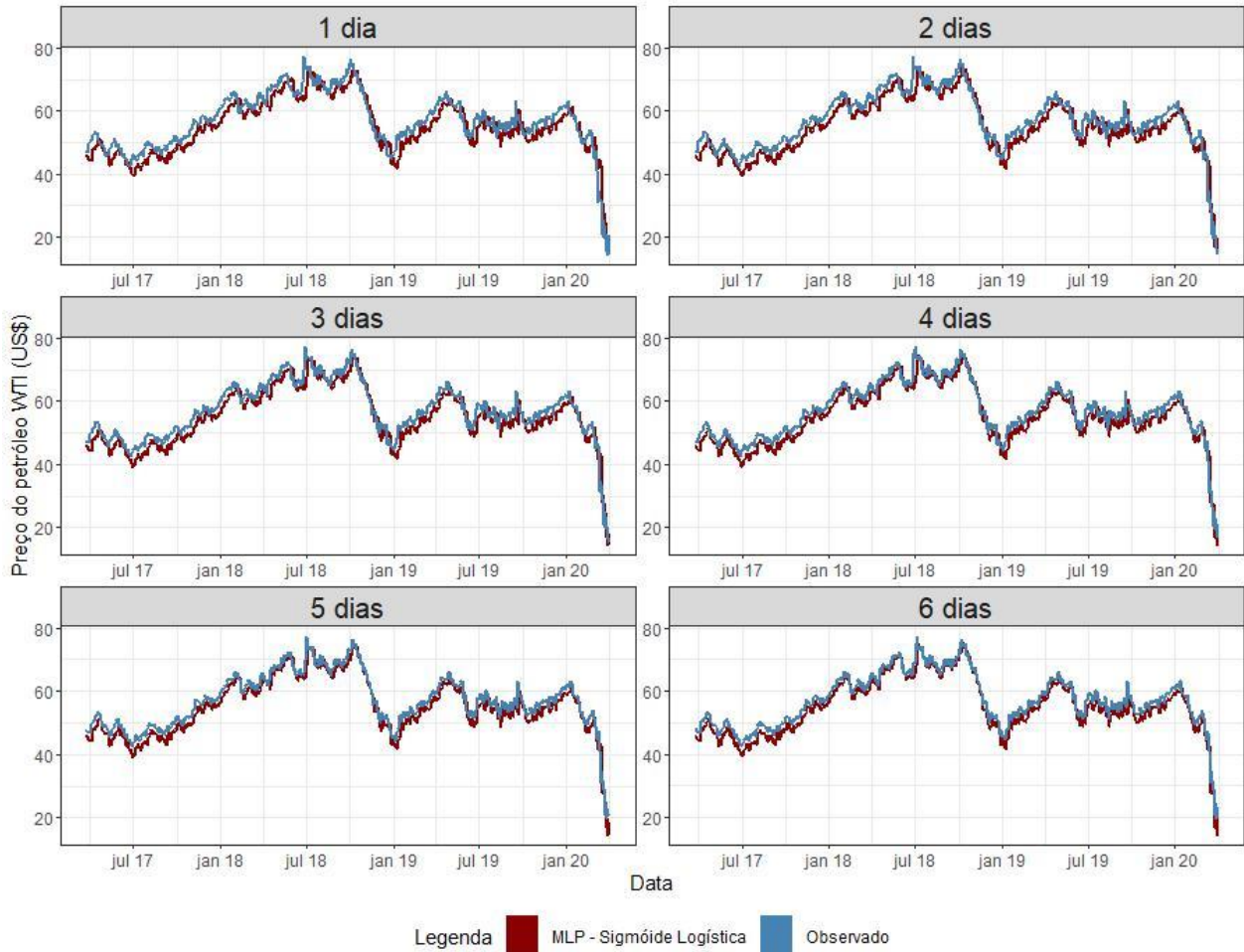


Figura 6. Previsão no período de validação. Fonte: A autora.

O modelo final foi capaz de acompanhar a série, conforme observado na Figura 6, que corresponde ao período de validação. Também é perceptível que a previsão foi adequada para todos os dias. No geral, o erro da previsão aparenta ser de subestimação do valor, pois a linha da previsão está abaixo do valor observado.

Outras técnicas de redes neurais poderiam ser utilizadas esta aplicação, uma delas é a denominada Memória longa de curto prazo (*Long short-term memory*, LSTM), a qual atualmente tem sido destacada para estruturas de dados temporais. Além disto, haveria a possibilidade de testar as redes com maior quantidade de neurônios na camada oculta ou incluindo mais variáveis na entrada. A taxa de aprendizagem e quantidade de iterações poderiam também ser modificadas. Além destas possibilidades ainda existe a opção da modificação do algoritmo de aprendizagem, para a busca de minimização do erro existe uma variedade de derivações do gradiente e além de outras.


Com esta aplicação a ideia principal foi descrever na prática uma possível utilização de uma das técnicas de aprendizado de máquina. Por se tratar de uma base de dados pública o código elaborado permite a total reprodução do que foi feito.

REFERÊNCIAS BIBLIOGRÁFICAS

- Zulifqar A et al. (2017). Forecasting drought using multilayer perceptron artificial neural network model. *Advances in Meteorology*.
- Bishop CM (1994). Neural networks and their applications. *Review of Scientific Instruments*, 65(5): 1803–1832.
- Box George EP et al. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Braga ADP et al. (2000). *Redes Neurais Artificiais: Teoria e Aplicações* 2 ed. Rio de Janeiro: LTC.
- Brockwell PJ, Davis RA (2016). *Introduction to Time Series and Forecasting*, 3 ed., Springer International Publishing.
- Devadoss AV, Ligorí TAA (2013). Forecasting of stock prices using multi layer perceptron. *International journal of computing algorithm* 2: 440-449.
- Dudek G (2016). Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting." *International Journal of Forecasting* 32(3): 1057-1060.
- Elman JL (1990). Finding structure in time. *Cogn Sci* 14(2): 179–211
- de Pauli STZ et al. (2020). Comparing Artificial Neural Network Architectures for Brazilian Stock Market Prediction. *Annals of Data Science*, 1-16.
- Fausett LV (2006). *Fundamentals of neural networks: architectures, algorithms and applications*. Pearson Education India.
- Haykin S (2006). *Neural networks: a comprehensive foundation*. Mc Millan, New Jersey.
- Jordan M (1986). Finding structure in time. In: *Proceedings of the eighth annual conference of the cognitive science society*, 531–546.
- Lee CY, Jinho K (2018) The prediction and analysis of the power energy time series by using the elman recurrent neural network. *Journal of the Society of Korea Industrial and Systems Engineering* 41 (1): 84-93.
- Lewis NC (2017). *Neural Networks for Time Series Forecasting with R: An Intuitive Step by Step Blueprint for Beginners*. AusCov.
- Montgomery CD et al. (2008). *Introduction to Time Series Analysis and Forecasting*.
- R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Šestanović T (2019). Jordan neural network for inflation forecasting. *Croatian Operational Research Review*, 23-33.
- Wang Y et al. (2021). Advantages of direct input-to-output connections in neural networks: The Elman network for stock index forecasting. *Information Sciences*, 547: 1066-1079.

Wu W et al. (2019). Time series analysis of human brucellosis in mainland China by using Elman and Jordan recurrent neural networks. *BMC infectious diseases* 19(1): 1-11.

Redes Neurais Artificiais Híbridas Para a Predição de Consumo de Energia Elétrica

 10.46420/9786581460204cap2

Ricardo Augusto Manfredini^{1*} 

INTRODUÇÃO

Nas últimas décadas, a população mundial está aumentando rapidamente e, em função desse aumento, a energia global demandada e consumida também está crescendo cada vez mais (LoBrano et al., 2014). Em relação a isso, os prédios residenciais ou comerciais, são identificados como grandes consumidores de energia em todo o mundo, representando cerca de 30% da demanda global de eletricidade relacionada ao consumo de energia no setor residencial (Lusis et al., 2017). Os prédios são responsáveis por uma significativa parte do desperdício de energia também. O desperdício de energia e as mudanças climáticas representam um desafio para a sustentabilidade, sendo crucial tornar os prédios mais eficientes (Marino et al., 2016). Portanto, o desenvolvimento e o uso de produtos limpos e energia renovável em prédios ganhou amplo interesse (Lo Brano et al., 2014). No setor residencial e comercial, os sistemas fotovoltaicos (PV) são a geração distribuída mais comum, minimizando a dependência da demanda das tradicionais usinas de energia e maximizando a autossuficiência das famílias (Lui et al., 2017).

Devido à dependência do PV das condições meteorológicas, a natureza intermitente da energia gerada traz alguma incerteza (Theocharides et al., 2017). Da mesma forma, o consumo de energia elétrica destes prédios também possuem incertezas inerentes à sazonalidade. A maneira mais fácil para gerenciar o risco da energia solar e aproveitar esse poder é prever a quantidade de energia a gerar (Mosaico; Saviozzi, 2019) bem como o consumo. Uma previsão confiável é a chave para várias aplicações de rede inteligente, como despacho, resposta ativa à demanda, regulação da rede e gestão inteligente de energia (Massucco et al., 2019).

O consumo energético de um prédio e a geração PV podem ser representados por uma série temporal com tendências e sazonalidade (Moretti; Tolo, 2006). Existem numerosos estudos de predição sobre séries temporais, desde as clássicas regressões lineares a trabalhos mais recentes utilizando algoritmos de aprendizado de máquina, que são poderosas ferramentas na previsão do consumo de energia elétrica e geração de energia fotovoltaica (Su et al., 2019). Recentemente, muitas técnicas de

¹ IFRS – Instituto Federal de Ciências e Tecnologia do Rio Grande do Sul

* Autor correspondente: ricardo.manfredini@farroupilha.ifrs.edu.br

previsão de energia fotovoltaica foram desenvolvidas, mas ainda não há uma unidade completamodelo de previsão versal e metodologia para garantir a acurácia das predições. Em relação a isso, Redes Neurais Artificiais (RNA) são algoritmos de aprendizado de máquina muito populares para previsão e classificação objetos e são baseados na abordagem clássica da rede neural *feed-forward* (Theocharides et al., 2017). RNAs são sistemas de computação inspirados nas redes neurais biológicas do cérebro, como os neurônios funcionam, passam e armazenam informações (Massucco et al., 2019; Theocharides et al., 2017).

Devido ao desenvolvimento acelerado da tecnologia de computação, a RNA forneceu uma estrutura poderosa para a aprendizagem supervisionada (Liu et al., 2017). O aprendizado profundo permite modelos compostos de várias camadas aprender representações de dados (Marino et al., 2016). Redes Neurais Profundas (DNN²) são inspiradas na estrutura dos sistemas visuais dos mamíferos e elas também são uma importante ferramenta de aprendizado de máquina que tem sido amplamente usado em muitos campos (Yi et al., 2017). DNN emprega uma arquitetura de múltiplas camadas de neurônios em uma RNA e podem representar funções com maior complexidade (Liu et al., 2017).

Este trabalho teve como objetivo a predição do consumo de energia elétrica de um prédio comercial utilização de RNA, nas suas diversas arquiteturas. Foram utilizadas e testadas diversas arquiteturas de RNAs e selecionada uma arquitetura híbrida (Densa, Convolucional e Recorrente), descrita originalmente por Lai et al. (2018) e adaptada para este estudo de caso.

As seções deste capítulo estão organizadas da seguinte forma: inicialmente é feita uma fundamentação teórica dos conceitos, técnicas e ferramentas utilizadas. Posteriormente são definidos os materiais e métodos que serão utilizados para a validação do modelo proposto. Finalmente, são apresentados os resultados do modelo proposto comparando-o com outros modelos.

FUNDAMENTAÇÃO TEÓRICA

Séries Temporais

Séries temporais são conjuntos de observações ordenadas no tempo (Moretti; Toloí, 2006). Uma série temporal pode ser definida como uma classe de fenômenos cujo processo observacional e consequente quantificação numérica geram uma sequência de observações distribuídas ao longo do tempo.

Históricos de consumo de energia elétrica ao longo do tempo são basicamente séries temporais univariadas (Spiegel, 1974) com tendências, ciclos, sazonalidades e aleatoriedades. Tendências são características de longo prazo relacionado com um intervalo de tempo. Ciclos são oscilações a longo prazo, mais ou menos regulares, em torno de uma linha ou curva de tendência. Sazonalidades são padrões

² DNN – do inglês *Deep Neural Network*

regulares observados de tempos em tempos. Finalmente, aleatoriedade são, basicamente, efeitos que ocorrem aleatoriamente e que não podem ser captados pelos ciclos, tendências e sazonalidades.

Desta forma, os modelos de predições de séries temporais mais utilizados na literatura são os de regressões linear e polinomiais. Dentre os modelos de regressões podemos citar o método SARIMAX (SARIMAX, 2021). Este modelo estatístico é uma variante do modelo autorregressivo de médias móveis (ARMA), adicionando derivações para tornar o modelo estacionário (I), adicionando sazonalidade (S) e finalmente adiciona-se o efeito de variáveis exógenas (X) ou aleatórias ao longo do tempo. Neste trabalho, utilizou-se o modelo SARIMAX como linha base de comparação dos seus resultados, da sua aplicação sobre o caso teste e os resultados obtidos de outros modelos de predição.

Redes Neurais Artificiais Convulsionais

Redes Neurais Artificiais Convulsionais (CNN³) são um tipo de DNN que, comumente, é aplicado para analisar imagens. Um dos principais atributos da CNN é conduzir diferentes camadas de processamento que geram uma representação eficaz das características das extremidades das imagens. A arquitetura da CNN permite múltiplas camadas dessas unidades de processamento a serem empilhadas, este modelo de aprendizagem profunda pode enfatizar a relevância de características em diferentes escalas (Yang et al., 2015).

A figura 1 demonstra uma arquitetura típica de uma CNN, composta de pelo menos, uma camada de convolução, uma camada *pooling*, uma camada de *flattening* e camadas densas.

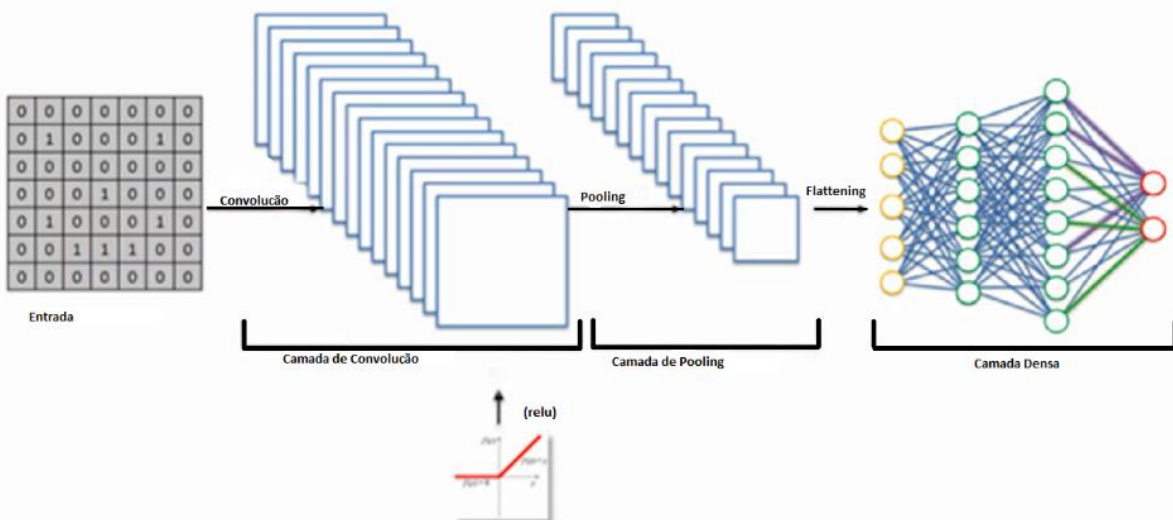


Figura 1. CNN básica. Fonte: O autor.

Na camada de covolução é aplicada um filtro (*kernel*, que também é uma matriz) à matriz de entrada visando a sua redução com a manutenção de suas caractterística mais importantes. A figura 2

³ CNN – do inglês *Convolutional Neural Network*

representa, passo a passo, a aplicação da função de convolução $g(x, y) = \omega * f(x, y) = \sum_{dx=-a}^a \sum_{dy=-b}^b \omega(dx, dy)f(x + dx, y + dy)$, onde $g(x, y)$ representa o elemento da matriz de convolução, que é o produto matricial da matriz colorido na figura pelo *kernel*, a cada passo desloca-se uma posição à direita até a última coluna da matriz de entrada, após desloca-se uma linha a baixo e continua-se o processo até percorrer toda a matriz de entrada. No exemplo da figura 2 uma matriz de entrada de 7×7 foi reduzida para uma matriz de convolução de 5×5 . Todo o processo representado na figura é repetido para cada um dos *kernels* utilizados, gerando várias matrizes de convolução.

Para a camada de *pooling*, é usual aplicar a função de ativação *relu* $f(x) = \max(0, x)$, por exemplo, gerando uma nova matriz reduzida como demonstra a figura 3.

Finalmente a camada de *flattening* nada mais é que a transformação das matrizes das camadas de *pooling* em vetores, os quais serão as entradas da camada densa.

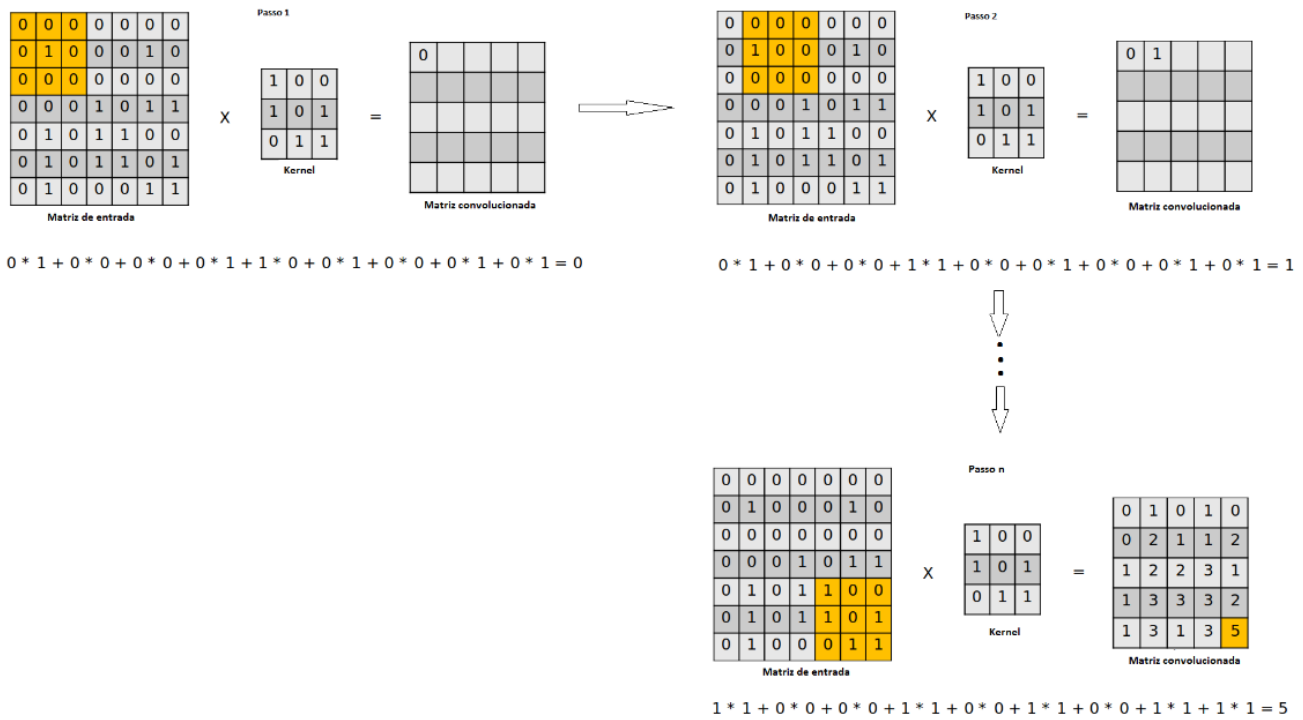


Figura 2. Processo de Convolução. Fonte: O autor.

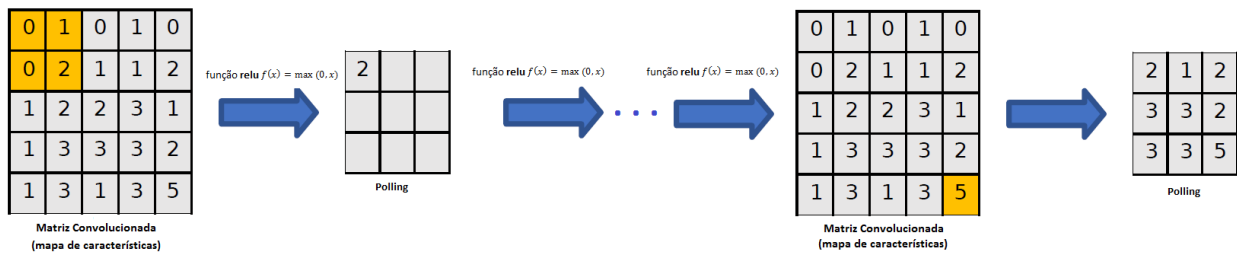


Figura 3. Processo de Pooling. Fonte o Autor.

Redes Neurais Artificiais Recorrentes

Nas RNAs tradicionais, as entradas (e saídas) são independentes umas das outras, dificultando sua utilização, por exemplos, no processamento de linguagem natural onde uma palavra na frase depende de palavras anteriores da mesma frase, ou ainda, em séries temporais que precisamos conhecer os valores ao longo do tempo para melhores projeções.

Em contraposição, as redes neurais artificiais recorrentes (RNN⁴) (Hammer, 2007) armazenam seu estado anterior e o utilizam também como entrada do estado atual para cálculos de novas saídas. Outra forma de pensar sobre as RNNs é que elas possuem uma “memória” que captura informações sobre o que foi calculado até agora. Em teoria, os RNNs podem fazer uso de informações em sequências arbitrariamente longas, mas, na prática, elas se limitam a olhar para trás apenas algumas etapas. A Figura 4 é uma representação típica de uma RNN.

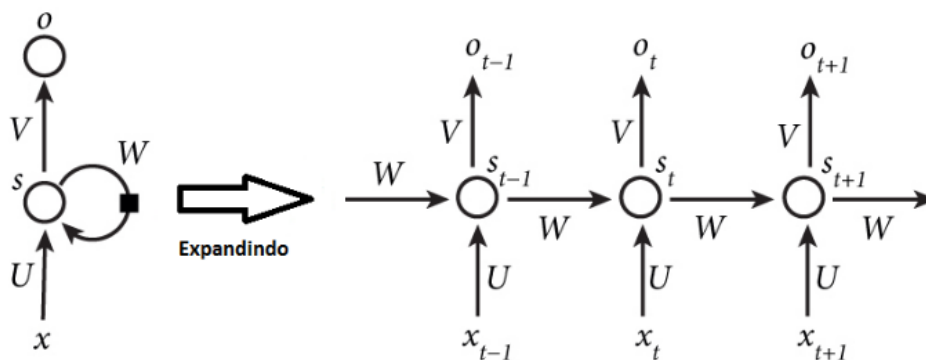


Figura 4. RNN básica. Fonte: O Autor.

A figura 4 mostra uma RNN sendo expandida em uma rede completa. Onde x_t é a entrada na etapa de tempo t . Por exemplo, x_1 poderia ser um vetor *one-hot* correspondente a segunda palavra de uma frase, s_t é o estado oculto na etapa de tempo t . É a “memória” da rede. s_t é calculado com base no estado oculto anterior e a entrada na etapa atual: $s_t = f(Ux_t + Ws_{t-1})$. A função f geralmente é uma não linearidade, como *tanh* ou *relu*. s_{-1} , que é necessário para calcular o primeiro estado oculto, normalmente é inicializado com zeros. o_t é a saída na etapa t . Por exemplo, se quiséssemos prever a próxima palavra em uma frase, seria um vetor de probabilidades em nosso vocabulário. $o_t = \text{softmax}(Vs_t)$. Por expandindo, queremos dizer simplesmente que escrevemos a rede para a sequência completa. Por exemplo, se a sequência que nos interessa é uma frase de 5 palavras, a rede seria desdobrada em uma rede neural de 5 camadas, uma camada para cada palavra.

⁴ RNN – do inglês *Recurrent Neural Network*

MATERIAL E MÉTODOS

Este trabalho foi realizado no Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development (GECAD5), centro de pesquisa localizado no Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto ISEP/IPP, Porto, Portugal. Da mesma forma que o modelo HyFIS2 (Josi et al., 2016), o modelo proposto utiliza os dados reais de consumo elétrico de setores do Prédio N do ISEP/IPP onde está localizado o GECAD. O prédio possui cinco medidores de energia que armazenam os dados de consumo de energia elétrica de setores específicos do prédio, com intervalo de tempo de 10 segundos. Estas informações, bem como dados meteorológicos são armazenados num servidor SQL automaticamente, através de agentes desenvolvidos em Java.

Para validação do modelo descrito a seguir, foram realizados testes utilizando os mesmos dados de consumo aplicados ao modelo SARIMAX e o HyFIS2. O setor dos laboratórios do Prédio N não foi computado pois possui grande variação de consumo em razão dos experimentos lá conduzidos que geram muitos *outliers* no histórico de consumo. Para os testes dos experimentos foi realizada uma média horária dos consumos armazenados a cada dez segundo, devido a necessidade de predição da próxima hora de consumo.

O Modelo Long and Short Time series Network Adapted (LSTNetA)

O modelo desenvolvido para a predição de consumo energético utilizou com base o modelo proposto por Lai et al. (2018), representado na Figura 4, constitui-se de uma RNA híbrida, com três camadas distintas, inicialmente possui uma camada convolucional para a extração de padrões de curto prazo da série temporal, tem como entrada a série temporal, a saída desta camada é a entrada da camada recorrente que memoriza informações históricas da série temporal, que por sua vez sua saída é a entrada da camada densa altamente conectada. Finalmente a saída da camada altamente conectada é combinada com a saída da regressão linear autorregressiva (ARMA) (Zhang, 2003) garantindo que a saída terá a mesma escala da entrada, compondo assim a predição.

⁵ <http://www.gecad.isep.ipp.pt/GECAD/Pages/Pubs/PublicationsPES.aspx>

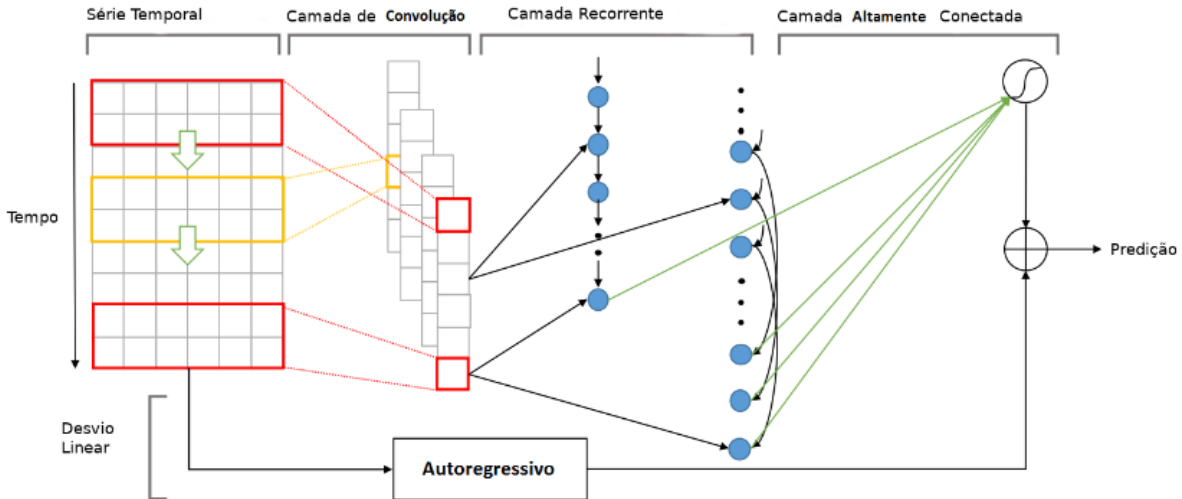


Figura 5. Arquitetura do modelo LSTNetA. Fonte: Adaptado de Lai (Lai et al., 2018).

A Figura 6 sumariza a implementação da rede LSTNetA. A camada de convolução está representada pela classe **Conv2D**, a camada recorrente está representada pelas classes **GRU**, a camada densa está representada pelas classes **Dense**, a autorregressão está representada na classe **PostARTrans**.

É importante salientar que a camada recorrente usa uma das variantes de RNN a GRU (*Gated Recurrent Unit*) (Chung et al., 2014), esse modelo de RNA assim como as LSTM (*Long Short-Term Memory*) visa resolver o problema da memória de curto prazo das RNN que, em séries longas, têm dificuldade de transportar os resultados de etapas anteriores para as posteriores.

This may be caused by multiline strings or comments not indented at the same level as the code.
Model: "model"

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 24, 24)]	0	
reshape (Reshape)	(None, 24, 24, 1)	0	input_1[0][0]
conv2d (Conv2D)	(None, 19, 1, 100)	14500	reshape[0][0]
dropout (Dropout)	(None, 19, 1, 100)	0	conv2d[0][0]
reshape_1 (Reshape)	(None, 19, 100)	0	dropout[0][0]
pre_skip_trans (PreSkipTrans)	(None, 1, 100)	0	reshape_1[0][0]
gru (GRU)	[(None, 100), (None, 60600		reshape_1[0][0]
gru_1 (GRU)	[(None, 5), (None, 5 1605		pre_skip_trans[0][0]
dropout_1 (Dropout)	(None, 100)	0	gru[0][1]
post_skip_trans (PostSkipTrans)	(None, 95)	0	gru_1[0][1] input_1[0][0]
pre_ar_trans (PreARTrans)	(None, 24)	0	input_1[0][0]
concatenate (Concatenate)	(None, 195)	0	dropout_1[0][0] post_skip_trans[0][0]
flatten_1 (Flatten)	(None, 24)	0	pre_ar_trans[0][0]
flatten (Flatten)	(None, 195)	0	concatenate[0][0]
dense_1 (Dense)	(None, 1)	25	flatten_1[0][0]
dense (Dense)	(None, 24)	4704	flatten[0][0]
post_ar_trans (PostARTrans)	(None, 24)	0	dense_1[0][0] input_1[0][0]
add (Add)	(None, 24)	0	dense[0][0] post_ar_trans[0][0]

=====
Total params: 81,434
Trainable params: 81,434
Non-trainable params: 0
=====

Figura 6. Sumário da implementação do LSTNet. Fonte: O Autor.

Na etapa de *backpropagation*, processo de aprendizagem das RNAs, as RNNs sofrem com o problema da dissipação do gradiente (*The Vanishing Gradient Problem*). Gradientes são valores usados para atualizar os pesos das redes neurais. O problema da dissipação do gradiente é quando esses propagados durante o treinamento de uma rede, vão sofrendo multiplicações por valores menores que 1 a cada camada da rede atravessada, chegando nas camadas iniciais da rede com valores ínfimos. Isso faz com que o ajuste dos pesos, calculados a cada iteração do treinamento da rede, sejam também ínfimos, isto onera o treinamento da rede.

Desta forma, nas RNNs as camadas que recebem uma pequena atualização do gradiente param de aprender, com isso as RNNs podem esquecer o que foi visto em sequências mais longas, tendo assim uma memória de curto prazo.

A Figura 7 mostra uma arquitetura típica de uma GRU. Basicamente o que a difere de uma RNN padrão são as portas de descarte (*reset gate*) e de atualização (*update gate*), que através da aplicação das funções da ativação *Sigmoid* e *tanh*, é definido se a saída anterior h_{t-1} será considerada ou descartada para o cálculo da nova saída.

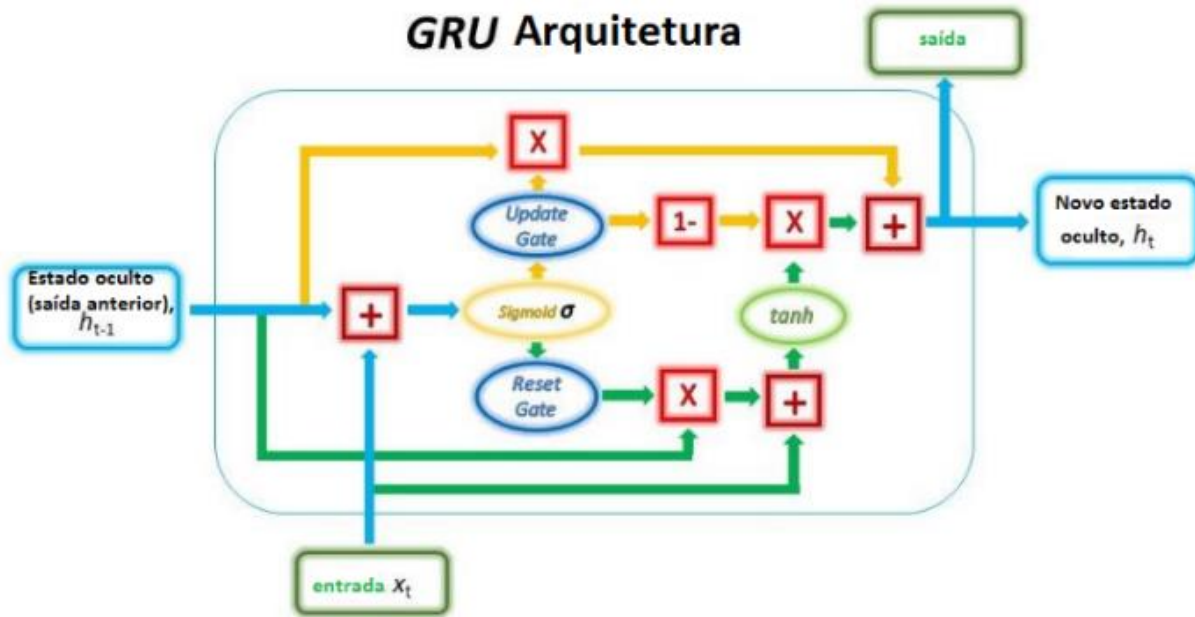


Figura 7. Arquitetura típica de uma GRU. Fonte: O Autor

O modelo LSTMNetA foi desenvolvido na linguagem de programação Python versão 3.7 (Python, 2021) utilizando a biblioteca de aprendizagem de máquina, desenvolvida pelo Google, TensorFlow versão 2.0.

TRABALHOS RELACIONADOS

A Figura 8, representa a série temporal do consumo de energia elétrica utilizada pelo modelo SARIMAX, para treinar e testar o modelo LSTMNetA e o HyFIS2. O gráfico superior representa série histórica do consumo em *watts*, que inicia às zero horas de 08/04/2019 às oito horas de 20/12/2019. O gráfico ao centro mostra a tendência calculada da série e o inferior à sua sazonalidade.

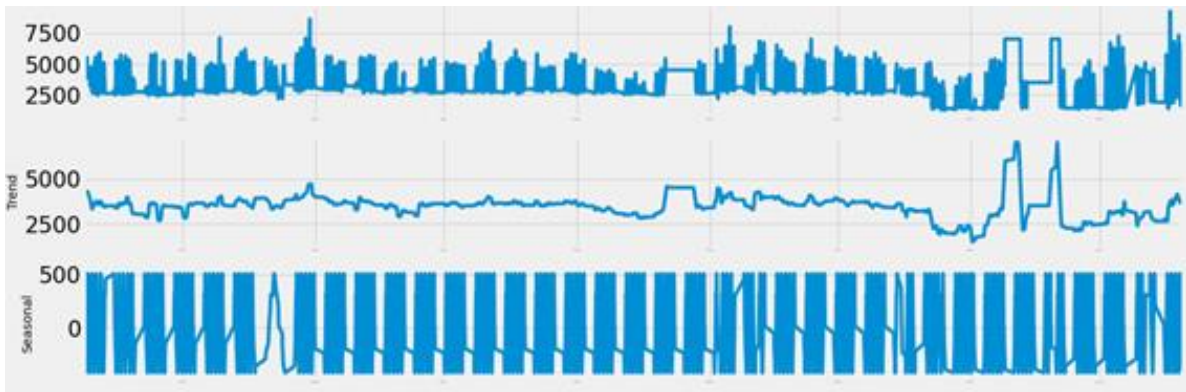


Figura 8. Série histórica de consumo. Fonte: O Autor.

SARIMAX

Como visto anteriormente, o método SARIMAX é um método estatístico de análise de séries temporais, possibilitando a predição através de regressões lineares. Desta forma, não se pode caracterizá-lo como um algoritmo de aprendizagem de máquina. No escopo deste trabalho foi aplicado para obter dados de predição de um modelo amplamente utilizado, obtendo-se resultados para comparação com o modelo proposto e com o modelo HyFIS2.

Para verificar a acurácia de todos os modelos abordados, neste trabalho, foram utilizados os últimos 120 registros, correspondentes a cinco dias de consumo, para comparação entre o consumo real e o predito, demonstrado na Figura 9. Para cálculo do erro utilizado para a verificação dos resultados deste trabalho, em todos os modelos, foi utilizado a raiz do erro quadrático médio (*Root Mean Square Error* – RMSE – Descrito no capítulo 01), demonstrado na Figura 10. A aplicação deste modelo resultou num RMSE médio de 604,72 que foi considerado como acurácia deste modelo, neste trabalho.

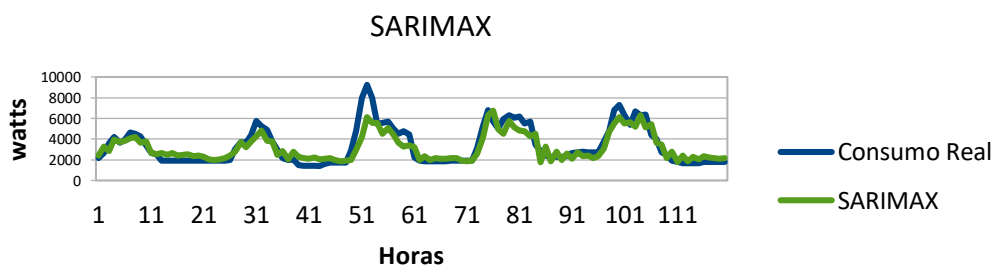


Figura 9. Comparativo Consumo Real X Sarimax. Fonte: O Autor.

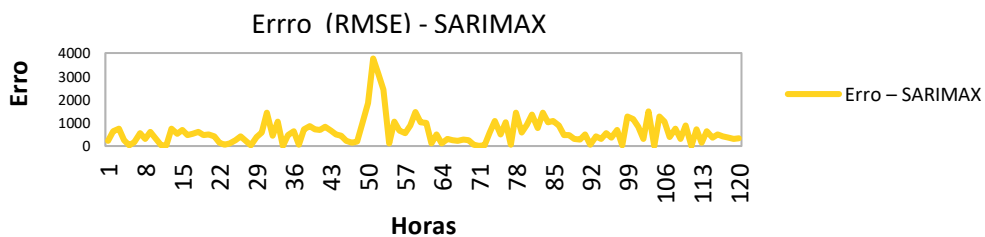


Figura 10. Erros verificados do método SARIMAX. Fonte: O Autor.

Modelo HyFIS2

O modelo HyFIS2 (*Hybrid neural Fuzzy Inference System*) utiliza uma abordagem híbrida com a combinação de RNA densa e lógica difusa (*fuzzy logic*). O sistema inclui cinco camadas, conforme mostrado na Figura 11. Na primeira camada, os nós são as entradas que transmitem os sinais para a próxima camada. Na segunda e na quarta camadas, os nós atuam como funções de pertinência para expressar as variáveis linguísticas difusas de entrada-saída. Nessas camadas, os conjuntos *fuzzy* definidos para as variáveis de entrada-saída são representados como: grande (L), médio (M) e pequeno (S). No entanto, para algumas aplicações, estes podem ser mais específicos e representados como, por exemplo, positivo grande (LP), positivo pequeno (SP), zero (ZE), negativo pequeno (SN) e negativo grande (LN). Na terceira camada, cada nó é um nó de regra e representa uma regra difusa. Os pesos de conexão entre a terceira e a quarta camada representam fatores de certeza das regras associadas, ou seja, cada regra é ativada e controlada pelos valores de peso. Por fim, a quinta camada contém o nó que representa a saída do sistema.

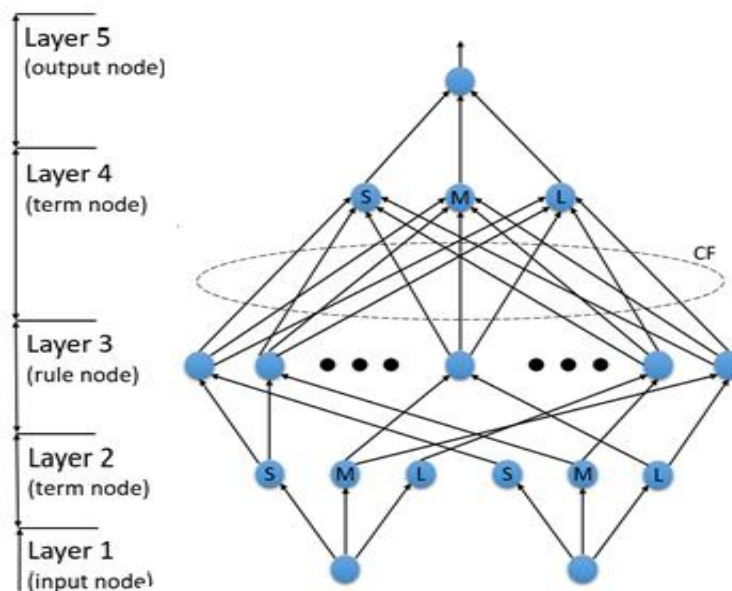


Figura 11. Estrutura Neuro-Fuzzy do modelo HyFIS2. Fonte: Jozi et al. (2016).

Para predição de consumo de eletricidade, como em todos os modelos testados, foram utilizados os últimos 120 registros dos históricos, correspondente a cinco dias de consumo. A comparação entre o consumo real e o predito, é demonstrado na Figura 12. A Figura 13 mostra os erros RMSE apurados. A aplicação deste modelo resultou num RMSE médio de 602,71 que foi considerado como acurácia deste modelo, neste trabalho.

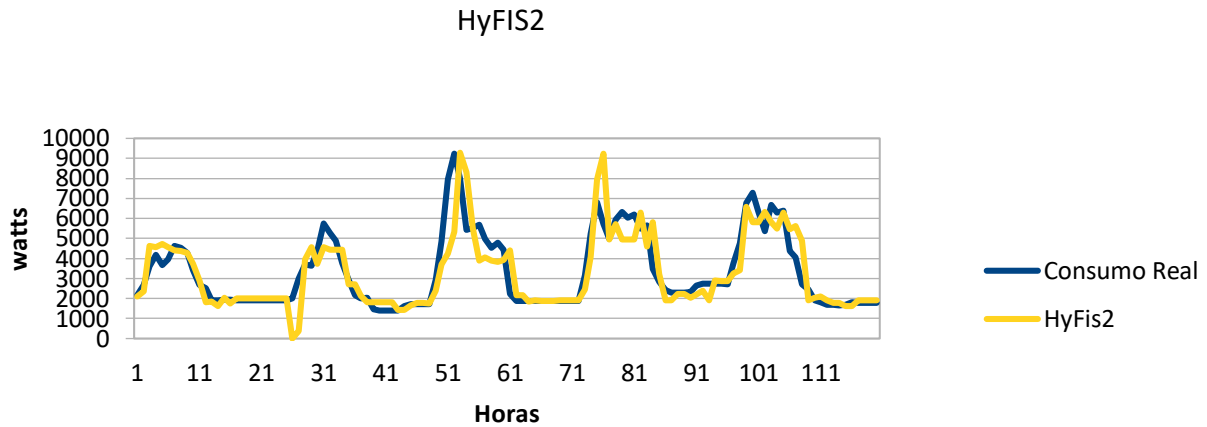


Figura 12. Comparativo Consumo Real X HyFis2. Fonte: O Autor.

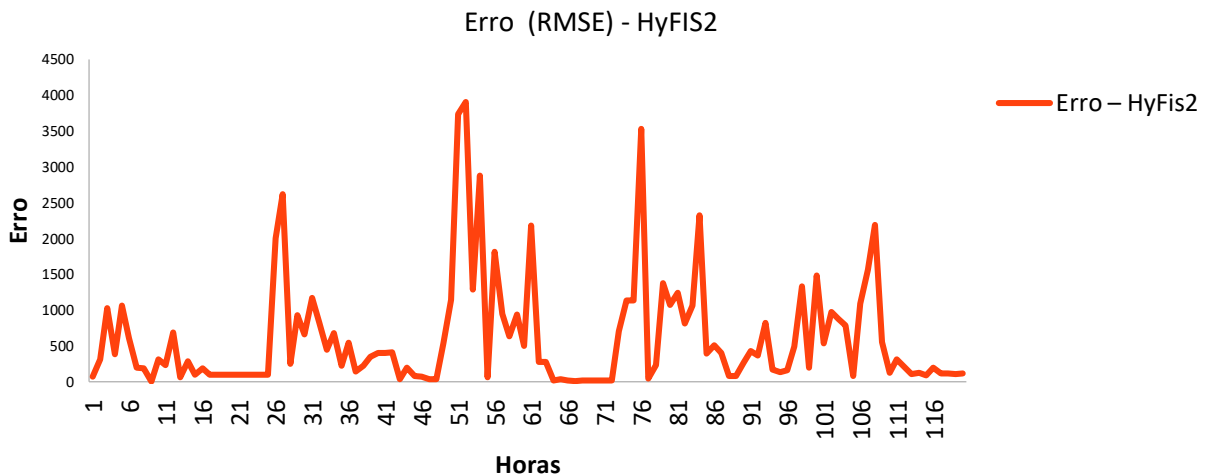


Figura 13. Erros verificados do modelo HyFIS2. Fonte: O Autor.

APLICAÇÃO DO MODELO LSTNETA

O treinamento da RNA LSTNetA foi realizado conforme descrito anteriormente, utilizando-se os dados do consumo real de energia elétrica do prédio N do ISEP/IPP onde está localizado o GECAD, excetuando-se o setor de laboratórios. A série histórica analisada foi das zero horas de 08/04/2019 às oito horas de 20/12/2019, com medições a cada dez segundos, totalizadas a cada hora, resultando 4186 registros, contendo hora e consumo. O treinamento foi realizado com taxa de aprendizagem de 0.0003, utilizando o método estocástico Adam (Kingma; Ba, 2015) de otimização da descida do gradiente para atualização dos pesos no processo de *backpropagation*. Para os pesos iniciais da RNA foi utilizado o algoritmo *VarianceScaling* (He et al., 2010) que gera pesos iniciais com valores na mesma escala das entradas. O kernel de convolução utilizado foi uma matriz identidade 6×6 e foi realizado um loop de

treinamento com 1000 épocas. Todos estes parâmetros foram obtidos de forma experimental e os de melhor resultado, selecionados.

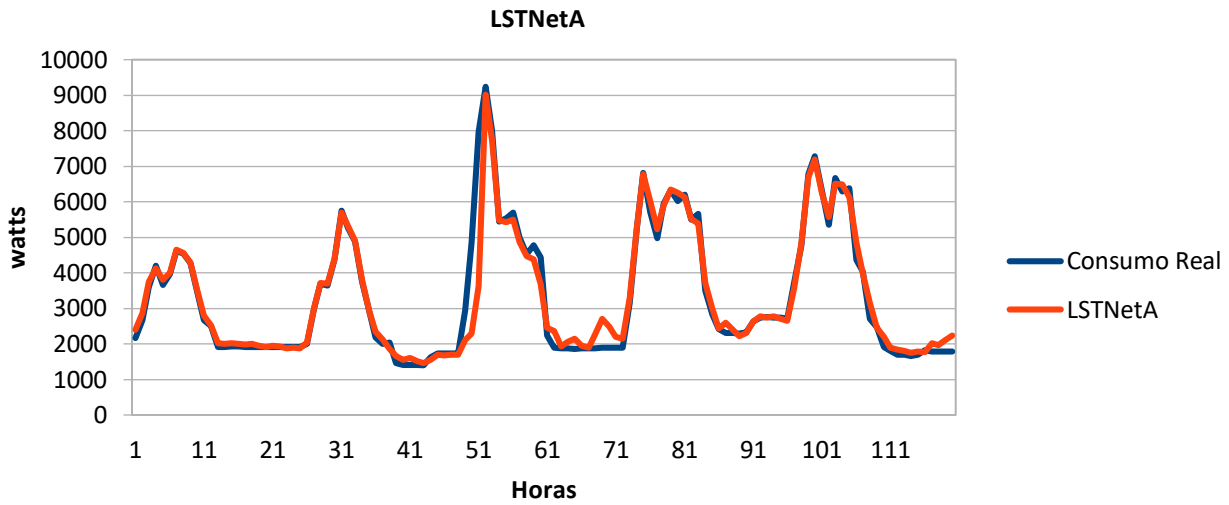


Figura 14. Comparativo Consumo Real X LSTNetA. Fonte: O Autor.

Para predição de consumo de eletricidade, como em todos os modelos testados, foram utilizados os últimos 120 registros dos históricos, correspondente a cinco dias de consumo. A comparação entre o consumo real e o predito, é demonstrado na Figura 14. A Figura 15 mostra os erros RMSE apurados. A aplicação deste modelo resultou num RMSE médio de 198,44 que foi considerado como acurácia deste modelo, neste trabalho.

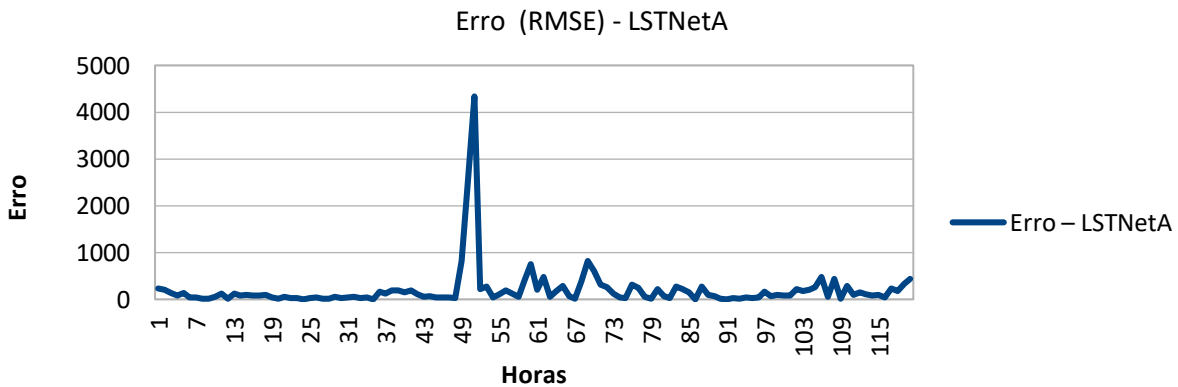


Figura 15. Erros verificados do modelo LSTNetA. Fonte: O Autor.

RESULTADOS E CONSIDERAÇÕES FINAIS

A Tabela 1 mostra um fragmento dos resultados dos três modelos, a coluna *Data e Hora*, a coluna *Real* mostrando o consumo real de eletricidade em *watts* naquela data e hora, a coluna *LSTNetA* a predição deste modelo na data e hora, a coluna *Erro – LSTNetA* o erro absoluto deste modelo na predição, a coluna *HyFIS2* a predição deste modelo na data e hora, a coluna *Erro – HyFIS2* o erro absoluto deste

modelo na predição, finalmente as colunas *SARIMAX* e *Erro – SARIMAX*, representando a predição e o erro absoluto, respectivamente, no modelo SARIMAX.

Comparando os resultados dos modelos SARIMAX, HyFIS2 e LSTNetA, pode-se observar, como demonstrado na Figura 16, que o método LSTNetA, com os dados utilizados para teste, foi o que apresentou as predições mais próximas do consumo real de energia elétrica, onde a linha vermelha, que representa as predições do modelo LSTNetA, em grande parte do período sobrepôs a linha azul que representa o consumo real. Isso demonstra uma predição muito próxima do valor real de consumo, com erros baixos.

Tabela 1. Fragmento de Predições e Erros dos 3 Modelos.

Data e Hora	Consumo Real	LSTNetA	Erro – LSTNetA	HyFis2	Erro – HyFis2	SARIMAX	Erro – SARIMAX
19/12/2019 09:00	4759,38	4824,27	64,8900	3427,13	1332,2500	4721,76	37,6190
19/12/2019 10:00	6781,51	6685,28	96,2346	6583,38	198,1300	5516,26	1265,2476
19/12/2019 11:00	7279,1	7194,26	84,8373	5798,56	1480,5400	6124,20	1154,8976
19/12/2019 12:00	6332,88	6247,08	85,8038	5798,38	534,5000	5497,10	835,7849
19/12/2019 13:00	5350,34	5569,95	219,6063	6322,98	972,6400	5653,27	302,9276
19/12/2019 14:00	6677,56	6499,50	178,0639	5798,37	879,1900	5197,56	1479,9983

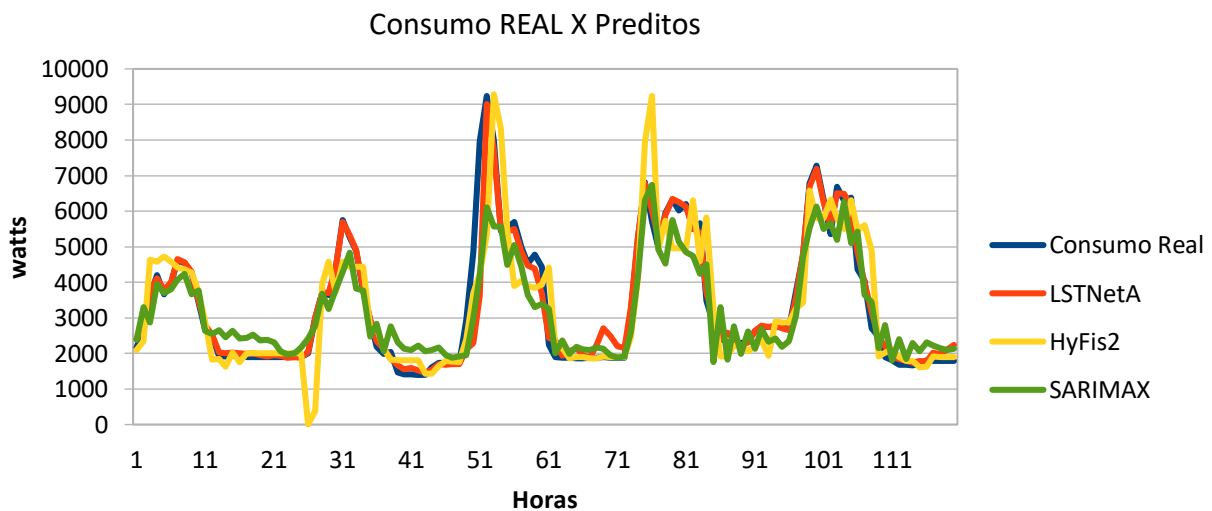


Figura 16. Comparativo Consumo Real X Modelos de Predição. Fonte: O Autor.

A Figura 17 representa os erros (RSME) dos três modelos, permitindo comparar a assertividade das predições de cada um dos métodos e, ainda, concluir que o método LSTNetA apresentou uma melhor eficácia em suas predições em comparação aos métodos SARIMAX e HyFIS2. Esta afirmação pode ser corroborada com os dados apresentados na Tabela 2, onde o erro total médio do modelo LSTNetA é significativamente menor que os demais modelos.

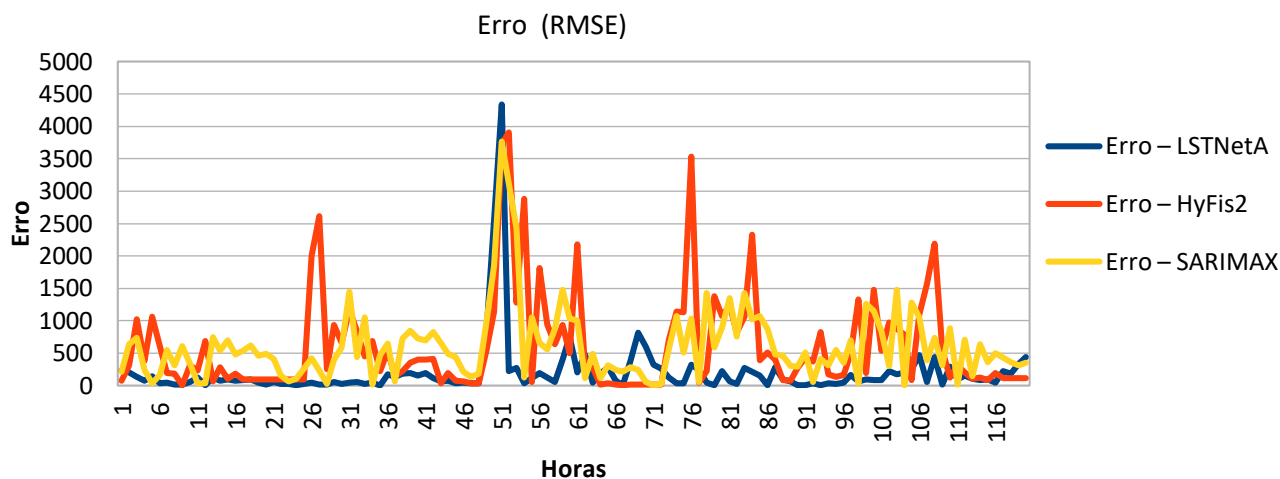


Figura 17. Comparativos de erros verificados em todos os modelos. Fonte: O Autor.

Tabela 2. RSME dos 3 Modelos Testados.

	Erro – LSTNetA	Erro – HyFis2	Erro – SARIMAX
RSME	198,4496	602,7109	604,5810


REFERÊNCIAS BIBLIOGRÁFICAS

- Chung J et al. (2014). Empirical evaluation of gated recurrent neural network on sequence modeling. in NIPS 2014 Workshop on Deep Learning, December 2014.
- Das UK et al. (2018). Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, 81: 912–928.
- He K et al. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.
- Lai G et al. (2018). Modeling long- and short-term temporal patterns with deep neural networks. 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, 95–104.
- Liu W et al. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234: 11–26.
- Lo Brano V et al. (2014). Artificial neural networks to predict the power output of a PV panel. *International Journal of Photoenergy*.
- Lusis P et al. (2017). Short-term residential load forecasting: Impact of calendar effects and forecast granularity. *Applied Energy*, 205: 654–669.
- Hammer B (2007). *Learning with Recurrent Neural Networks*. London: Springer.

- Jozi A et al. (2016). Energy Consumption Forecasting based on Hybrid Neural Fuzzy Inference System. 2016 IEEE Symposium Series on Computational Intelligence
- Kingma DP, Ba J (2016). Adam: A Method for Stochastic Optimization. Computer Science, Mathematics ICLR 2015. (SSCI). 6-9. Disponível em: <https://arxiv.org/abs/1412.6980.pdf>. Acessado em: 01/03/2021.
- Marino DL et al. (2016). Building energy load forecasting using Deep Neural Networks. IECON Proceedings (Industrial Electronics Conference), 7046–7051.
- Massucco S et al. (2019). A hybrid technique for day-ahead pv generation forecasting using clear-sky models or ensemble of artificial neural networks according to a decision tree approach. Energies, 12(7).
- Montavon G et al. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing: A Review Journal, 73: 1–15.
- Morete PA, Toloi CMC (2006). Análise de séries temporais. 2. ed. São Paulo: Egard Blucher.
- Mosaico G, Saviozzi M (2019). A hybrid methodology for the day-ahead pv forecasting exploiting a clear sky model or artificial neural networks. In IEEE EUROCON 2019 -18th International Conference on Smart Technologies, 1–6.
- Pelletier C et al. (2021). Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. MDPI - Remote Sensinf - Open Access Journal. Disponível em: <https://arxiv.org/pdf/1811.10166.pdf> Acessado em: 01/03/2021.
- Python (2021). Python is a programming language that lets you work quickly and integrate systems more effectively. Disponível em: <https://www.python.org>. Acessado em: 01/03/2021
- Reddy KS, Ranjan M (2003). Solar resource estimation using artificial neural networks and comparison with other correlation models. Energy Conversion and Management, 44(15): 2519–2530.
- SARIMAX (2021). SARIMAX: Introduction. Disponível em: https://www.statsmodels.org/dev/examples/notebooks/generated/statespace_sarimax_stata.html. Acessado em: 01/03/2021
- Spiegel MR (1974). Estatística. Brasília: McGraw-Hill do Brasil.
- Su D et al. (2019). Machine learning algorithms in forecasting of photovoltaic power generation. In 2019 International Conference on Smart Energy Systems and Technologies (SEST), 1–6.
- TensorFlow (2021). Uma plataforma completa de código aberto para machine learning. Disponível em: <https://www.tensorflow.org>. Acessado em: 01/03/2021
- Theocharides S et al. (2017). Pv Production Forecasting Model Based on Artificial Neural Networks (Ann). 33rd European Photovoltaic Solar Energy Conference, 1830 – 1894.
- Yang J et al. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. in Ijcai, 15: 3995–4001.

- Yi H et al. (2017). A study on Deep Neural Networks framework. Proceedings of 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, IMCEC 2016, 1519–1522.
- Zhang GP (2003). Time series forecasting using a hybrid arima and neural network model. Neurocomputing, 50:159–175.

Reconhecimento de padrões de arritmias cardíacas no Eletrocardiograma (ECG) empregando Transformada Wavelet e o classificador Naïve Bayes

 10.46420/9786581460204cap3

Bruno Rodrigues de Oliveira^{1*} 

INTRODUÇÃO

Uma das principais causas de óbito no mundo é ocasionada por doenças cardiovasculares, que já são consideradas de proporções epidêmicas de acordo com a Organização Mundial de Saúde. Em 2016, por consequência de tal doença, cerca de 17,9 milhões de pessoas vieram a óbito². Para se ter uma ideia dos números envolvidos, nos EUA a cada 36 segundos uma pessoa morre de doença cardiovascular e a cada 4 óbitos 1 é em decorrência de alguma doença do coração (CDC, 2020). No Brasil, a taxa de mortalidade por conta de infarto agudo do miocárdio (IAM) é de 183,3 por 100 mil habitantes (Santos, 2018) e a cada 90 segundos morre uma pessoa por doença cardiovascular de acordo com a Sociedade Brasileira de Cardiologia³.

O coração humano é composto por quatro câmaras, sendo dois átrios e dois ventrículos. As câmaras direitas recebem o sangue do sistema circulatório e o bombeia para os pulmões, já as câmaras esquerdas recebem este sangue oxigenado e o bombeia para o sistema circulatório e para os órgãos periféricos. A atividade contrátil das câmaras ocorre devido a despolarização e repolarização elétrica das células do coração, que decorrem das alterações químicas no conteúdo intracelular (Guyton; Hall, 2006; Hampton, 2014; Mohrman, 2011).

Um dos exames mais empregados para avaliar a saúde do coração é o eletrocardiograma (ECG), o qual é “considerado padrão ouro para o diagnóstico não invasivo das arritmias e distúrbios de condução, além de ser muito importante nos quadros isquêmicos coronarianos, constituindo-se em um marcador de doença do coração” (Nicolau et al., 2003). O ECG é o resultado da sobreposição da atividade elétrica das células do miocárdio, estas estão localizadas em diferentes partes do coração e cada conjunto delas tem características próprias, pois apresentam potenciais de repouso e de ação com magnitudes e durações distintas. Isso torna o ECG um sinal com representação peculiar, sendo este formado geralmente por ondas nominadas de P, Q, R, S e T, onde a onda P representa a despolarização

¹ Pantanal Editora.

* Autor correspondente: bruno@editorapantanal.com.br

² [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))

³ <http://www.cardiometro.com.br/>

dos átrios, o complexo QRS a despolarização dos ventrículos e a onda T a repolarização ventricular (Hampton, 2014).

Para registrar o ECG, diferentes configurações de eletrodos (sensores) podem ser utilizadas e estas recebem o nome de derivações. Nas derivações bipolares (DI, DII e DIII) e unipolares aumentadas (avR, avL e avF) os eletrodos são dispostos nos braços e pernas. Por outro lado, na derivação precordial, rotuladas de V1 a V6, estes são dispostos no tórax e nos membros inferiores (Hampton, 2014; Mohrman, 2011).

Normalmente o marcapasso do coração é determinado no nodo Sinoatrial (SA), localizado no átrio direito. O impulso elétrico se propaga das células adjacentes para o átrio esquerdo e para o nodo Atrioventricular (AV), que é retardado em cerca de 130 milissegundos (ms), possibilitando que a contração dos átrios se encerre antes de iniciar a dos ventrículos. O impulso elétrico continua seu trajeto e chega às fibras de Purkinge, onde a condução é rápida. Estas fibras são formadas pelo feixe de His e seus ramos esquerdo e direito, os quais conduzem o impulso elétrico a todas as células ventriculares (Mohrman, 2011). Pode ocorrer que o marcapasso cardíaco seja assumido por outras áreas do coração devido a certos bloqueios da condução do impulso elétrico, mecanismos de reentrada, doença do nodo SA, bradicardia sinusal, alimentação, uso de tabaco, dentre outros fatores (Woods et al., 2005). Nestes casos ocorrem batimentos ectópicos e mudanças no ritmo cardíaco, levando ao surgimento das arritmias cardíacas.

As arritmias cardíacas são distúrbios ocasionados por alterações na formação e/ou condução do impulso elétrico através do tecido do miocárdio, podendo, assim, modificar a origem e/ou a difusão fisiológica do estímulo elétrico do coração, motivo pelo qual têm no eletrocardiograma o método de escolha para seu estudo e diagnóstico. As alterações na velocidade da propagação do estímulo elétrico, isoladamente, levando a bloqueios dos fascículos ou ramos, não são consideradas arritmias cardíacas (Nicolau et al., 2003).

O estudo da eletrofisiologia do coração teve início com Willem Einthoven, por volta de 1885, com o desenvolvimento do galvanômetro de corda. Muitos avanços foram feitos no aperfeiçoamento do eletrocardiógrafo, em 1961 Norman Jefferis Holter descreveu os primeiros usos práticos da utilização de eletrocardiógrafos mais compactos para monitoramento ambulatorial (Dimarco; Philbrick, 1990). Hoje tais dispositivos são utilizados para registro de sinais de ECG em períodos superiores à dias de duração (Barrett et al., 2014). Algumas das vantagens da utilização do monitoramento ambulatorial são: correlação entre sintomas e arritmias, detecção de isquemia miocárdica, monitoramento após IAM, avaliação de arritmias em pacientes assintomáticos, prevalência de contração ventricular prematura, avaliação do envolvimento cardíaco na sarcoidose, diagnóstico de acidente vascular cerebral, dentre outros (Latchamsetty; Bogun, 2015; Kuchar et al., 1987; Suzuki et al., 1994; Lipski et al., 1976; Shafqat et al., 2004).

Devido à grande quantidade de dados que são gerados pelos dispositivos Holter, ferramentas computacionais são indispensáveis para os diagnósticos médicos. Muitos métodos computacionais foram

concebidos nos últimos anos para análise automática de ECG. Interessa a essa pesquisa aqueles métodos que são dedicados ao reconhecimento dos padrões arrítmicos empregando técnicas de aprendizado de máquina.

Neste capítulo é descrita uma abordagem matemático/computacional para o reconhecimento de arritmias cardíacas empregando uma das técnicas mais simples de aprendizado de máquina, denominada de Naïve Bayes. Para a extração de atributos dos sinais de ECG é utilizado o método de processamento de sinais denominado de Transformada Wavelet e, visto que são gerados muitos vetores de atributos para o mesmo sinal, várias máquinas de aprendizado foram induzidas, exigindo, portanto, o emprego da técnica de Comitê de Máquinas.

Este capítulo está assim organizado. A seção material e métodos está dividida em 2 subseções, na primeira, Base de Dados, estão discriminados os registros ECG utilizados. Na segunda, Ferramentas, são descritas a Transformada Wavelet, a abordagem Naïve Bayes e o Comitê de Máquinas (Ensemble). Na seção Metodologia Proposta é apresentada uma nova metodologia para o reconhecimento de arritmias cardíacas e, por fim, na seção Resultados e Discussões, são apresentados os resultados obtidos empregando a base de dados MIT-DB e a metodologia proposta.

MATERIAL E MÉTODOS

Base de dados

As arritmias são classificadas pela *Association for the Advancement of Medical Instrumentation* (AAMI) de acordo com o Quadro 1 (AAMI, 1987). Devido a quantidade de batimentos cardíacos da classe normal ser superior as quantidades de batimentos arrítmicos, na pesquisa ora apresentada foram consideradas apenas duas classes, a saber: normal e abnormal, sendo que a última inclui as classes supraventricular ectópico e ventricular ectópico. Os batimentos cardíacos das classes fusão e desconhecido não foram empregadas.

Quadro 1. Classes AAMI dos batimentos cardíacos e rótulos associados.

Classe AAMI	Batimento Cardíaco	Rótulo
Normal	Normal	N
	Bloqueio do ramo esquerdo	L
	Bloqueio do ramo direito	R
	Escape atrial	e
	Escape atrioventricular	j
Supraventricular Ectópico	Contração prematura atrial	A
	Contração prematura atrial aberrante	a
	Contração atrioventricular prematura	J
	Contração prematura supraventricular	S
Ventricular Ectópico	Contração prematura ventricular	V

Classe AAMI	Batimento Cardíaco	Rótulo
	Escape ventricular	E
Fusão	Fusão de normal e ventricular	F
Desconhecido	Ritmado	/
	Fusão ritmado e normal	f
	Não classificado	Q

As classificações dos batimentos cardíacos (coluna 2 do Quadro 1) são empregados pela PhysioNet (Research Resource for Complex Physiologic Signals) nos sinais de ECG que são disponibilizados em seu repositório para alguns dos seus conjuntos de dados. Nesta pesquisa foi utilizada a base de dados MIT-BIH Arrhythmia Database (MIT-DB) (Goldberger et al., 2000) que é composta por 48 registros de ECG de 47 pacientes com duração aproximada de 30 minutos cada, os quais foram obtidos entre 1975 e 1979 no Boston's Beth Israel Hospital (BIH) Arrhythmia Laboratory. Os sinais de ECG foram amostrados a taxa de 360 Hz com 11 bits de resolução sobre um intervalo de 10 microvolt (mV), todos foram previamente analisados por especialistas que providenciaram as marcações das arritmias, seguindo o esquema dos rótulos discriminado no Quadro 1.

Ferramentas

A seguir serão descritas, de forma sucinta, as ferramentas matemáticas utilizadas para resolver o problema de reconhecimento de arritmias cardíacas. Para mais detalhes sobre estas ferramentas, o leitor deve consultar as referências (Daubechies, 1992; Mallat, 2009; Soman et al., 2010).

Transformada Wavelet

A Transformada Wavelet Contínua (Continuous Wavelet Transform - CWT) é uma transformada integral multiescala que computa a similaridade entre um sinal x e uma função núcleo ψ , denominada de wavelet. A CWT tem sido aplicada nas mais diversas áreas, tais como análise do clima, séries financeiras, monitoramento cardíaco, remoção de ruído em dados sísmicos e astronômicos, caracterização de fissuras, solução rápida de equações diferenciais parciais, computação gráfica, caracterização de turbulência, dentre outras (Soman et al., 2010).

Uma função $\psi(t) \in L^2(\mathbb{R})^4$ com norma unitária é uma função wavelet se a condição de admissibilidade, $C_\psi = 2\pi \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty$, for satisfeita, onde $\Psi(\omega)$ é a Transformada de Fourier de $\psi(t)$. Essa condição implica que a função $\psi(t)$ tem média nula, isto é $\int_{-\infty}^{\infty} \psi(t) dt = 0$ (Daubechies, 1992), e isso caracteriza seu caráter ondulatório e o suporte compacto, diferenciando-a das funções núcleo do tipo senoides que oscilam indefinidamente.

⁴ Espaço das funções reais de quadrado integrável.

Em processamento de sinais a CWT é empregada, principalmente, para análise multiescala. Para tanto é necessário gerar uma família de funções wavelets, utilizando translações por τ e dilatações por $s \neq 0$ da “wavelet mãe” $\psi_{s,\tau}(t) = |s|^{-0.5}\psi(\frac{t-\tau}{s})$, onde t é um índice de tempo. Assim, a CWT de um sinal $x(t) \in \mathbb{R}^N$ é dada pela integral $CWT_x(s, \tau) = \int_{-\infty}^{\infty} x(t)\underline{\psi}_{s,\tau}(t) dt$ onde $\underline{\psi}_{s,\tau}(t)$ denota o conjugado complexo de $\psi_{s,\tau}(t)$. Portanto, a CWT transforma um sinal unidimensional de comprimento N em uma representação bidimensional $s \times N$. Quando o parâmetro de escala s está restrito ao intervalo aberto $(0, 1)$, a família de wavelets é uma versão comprimida da wavelet “mãe” $\psi_{s,\tau}(t)$. Portanto, as porções do sinal $x(t)$ com a mais alta frequência serão ressaltadas. Por outro lado, se $s > 1$, tem-se uma versão dilatada da wavelet mãe e então as oscilações de baixa frequência são destacadas na transformada (Mallat, 2009).

Existem muitas funções wavelets e cada uma possui características que a tornam mais ou menos adequada dependendo da aplicação selecionada. Neste capítulo foi empregada a wavelet complexa de Morlet pois ela possui características frequências que são úteis para a resolução do problema de reconhecimento de padrões arrítmicos. Sua expressão analítica é dada por $\psi(t) = \frac{1}{\sqrt{\pi}} \exp(-t^2)\exp(j2\pi Ct)$, onde j é a unidade imaginária, C a frequência central e $\exp()$ a função exponencial.

Aprendizado de Máquina, Naïve Bayes e Comitê de Máquinas

Dado um conjunto de dados $T = \{(\mathbf{x}_k, y_k)\}_{k=1}^K$ que possui informações de um ambiente qualquer, o objetivo do aprendizado de máquina é aprender uma função $\hat{h}(\mathbf{x}_k, \theta) = y_k$ que associa os padrões $\mathbf{x}_k \in \mathbb{R}^N$ às classes⁵ $y_k \in \{0,1\}$ ⁶, onde θ é um vetor de parâmetros da função \hat{h} a qual é uma estimativa da função real h , que por sua vez representa o processo que gera y_k a partir de \mathbf{x}_k . Os padrões $\mathbf{x}_k = [x_{k1}, x_{k2}, \dots, x_{kN}]$ são compostos das variáveis (atributos) x_{kn} ($n = 1, 2, \dots, N$) independentes, sendo que cada componente explica, ao seu modo, a variável dependente y_k . Para aprender a função \hat{h} , utiliza-se parte do conjunto T na etapa conhecida como treinamento (ou indução). Esta abordagem é denominada de aprendizado supervisionado, pois conhecemos *a priori* as classes de cada padrão. A outra parte do conjunto T , não empregada para o aprendizado, é utilizada para testar se a função aprendida é uma boa estimativa da função real, utilizando alguma(s) métrica(s) de performance.

Um dos algoritmos mais simples para aprender \hat{h} é conhecido como Naïves Bayes, que se baseia no conceito de probabilidade condicional pelo Teorema de Bayes (Kubat, 2015). Assim, para obter a probabilidade de um padrão \mathbf{x}_k pertencer à classe y_{ki} , calculamos a probabilidade condicional

⁵ Em problemas de regressão y_k é um valor real.

⁶ $\{0,1\}$ significa um problema de classificação binário. Outras designações podem ser adotadas conforme o número de classes.

$p(y_{ki}|\mathbf{x}_k) = p(\mathbf{x}_k|y_{ki})p(y_{ki})/p(\mathbf{x}_k)$. Como a verossimilhança depende de N variáveis, seu cálculo tem alto custo computacional. Para driblar esse custo, supõe-se que as variáveis independentes são também mutuamente independentes, por isso o método é denominado “*Naïve*” que significa ingênuo. Portanto, a equação anterior toma a forma $p(y_{ki}|\mathbf{x}_k) \approx p(y_{ki}) \prod_{n=1}^N p(x_{kn}|y_{ki})$, onde o denominador não foi considerado pois é uma constante para todas as classes. Embora tal suposição não se verifica em muitos problemas reais, ainda assim este método é um bom classificador, porque o valor em si da probabilidade computada é menos importante, já que ela é utilizada na comparação das probabilidades. Ou seja, faz-se o seguinte: se $p(y_{ki}|\mathbf{x}_k) > p(y_{kj}|\mathbf{x}_k)$, então \mathbf{x}_k pertence à classe i , onde $i = 0$ e $j = 1$ (ou $i = 1$ e $j = 0$). Se as variáveis são reais (valores contínuos), então deve-se utilizar alguma função de distribuição de probabilidade (Kubat, 2015). A mais utilizada é a Gaussiana, dada pela equação $p(x_{kn}|y_{ki}) = (2\pi\sigma_{n,y_i}^2)^{-0.5} \exp[-(x_{kn} - \mu_{n,y_i})/2\sigma_{n,y_i}^2]$, onde os parâmetros desvio padrão σ_{n,y_i}^2 e média μ_{n,y_i} são aprendidos na fase de treinamento para cada variável independente e classe, os quais compõem o vetor θ .

Se um certo padrão \mathbf{x}_k puder ser representado de diferentes modos, aplicando por exemplo uma ferramenta como a CWT para extrair informações latentes deste padrão, então é necessário aprender várias funções \hat{h} para cada modo, onde cada uma delas aprenderá particularidades distintas do mesmo padrão. Neste caso, tem-se um comitê de máquinas de aprendizado e a decisão final sobre a classe do padrão \mathbf{x}_k será tomada levando em conta a predição de cada uma das funções induzidas. Para tanto, o voto majoritário ponderado é uma das abordagens possíveis, e nesta considera-se que a classe predita é dada por $\hat{y}_k = \arg \max_c \{1/Q \sum_{q=1}^Q w_q p_q(c|\mathbf{x}_k)\}$, onde c é uma classe, $p_q(c|\mathbf{x}_k)$ é a probabilidade associada à função \hat{h}_q , w_q o peso dessa função, sendo estes valores associados à q -ésima função aprendida (Dietterich, 2000).

Para a determinação dos pesos w_q calcula-se $w_q = M_{\hat{h}_q} / \sum_{q'=1}^Q M_{\hat{h}_{q'}}$, onde $M_{\hat{h}_q}$ é uma métrica de performance obtida pela função estimada \hat{h}_q . Outras abordagens para calcular os pesos podem ser encontradas em Oliveira et al. (2019).

Para medir a performance dos modelos induzidos foram utilizadas três medidas: Acurácia (Acc), Precisão (Pr) e Recobrimento (Re) dadas pelas fórmulas: $Acc = \frac{TP+TN}{TP+TN+FP+FN}$, $Pr = \frac{TP}{TP+FP}$ e $Re = \frac{TP}{TP+FN}$, onde TP, TN, FP, FN são as quantidades de verdadeiros positivos e negativos, e falsos positivos e negativos, respectivamente.

METODOLOGIA PROPOSTA

Para reconhecer um batimento cardíaco como pertencente a classe normal ou abnormal (arritmico), primeiro é necessário separar cada batimento do sinal de ECG. Isso é feito detectando as

ondas R de cada batimento. Visto que na base MIT-DB a posição das ondas R já estão destacadas, estes valores serão utilizados para segmentação de cada batimento tomando um segmento correspondente a 0,8 segundos (tempo médio de um ciclo cardíaco), sendo 0,3 segundos antes e 0,5 segundos depois de cada onda R. Para cada batimento é aplicada a CWT nas escalas que possuem informações espectrais relevantes. Para descobrir quais as faixas de frequências mais importante para cada tipo de batimento e, portanto, quais as escalas mais adequadas, foi feita uma análise frequencial da média de 10 batimentos da mesma classe empregando a Transformada de Fourier de Tempo Curto (Sundararajan, 2001). Considerando como corte o valor da densidade espectral de potência $2 \times 10^{-4} \text{ mV}^2/\text{Hz}$, o Quadro 2 a seguir mostra as informações espectrais dos batimentos cardíacos considerados pela AAMI (Quadro 1).

Quadro 2. Intervalos de frequência para cada tipo de batimento cardíaco.

Batimento	Intervalo de frequência (Hz)
N	2,50 – 24,50
L	1,20 – 13,70
R	0,60 – 30,50
e	0,03 – 30,00
j	0,80 – 26,40
A	0,25 – 33,00
a	0,05 – 35,80
J	0,05 – 22,45
S	0,15 – 34,10
V	0,03 – 12,25
E	0,18 – 13,00
F	0,37 – 16,05
P	0,02 – 9,05
f	0,09 – 21,00
Mínimo – Máximo	0,02 – 35,80

Nota-se do Quadro 2 (cinco primeiras linhas) que os batimentos classificados na classe normal, estão compreendidos no intervalo de frequência [0,03 ; 30,50]Hz, enquanto que, para os batimentos das outras classes, observa-se um intervalo de [0,02 ; 35,80]Hz. Logo, apenas as informações espectrais não são suficientes para a classificação. No entanto, a CWT combina informações espectrais com informações temporais, e mesmo que o espectro Wavelet seja diferente do espectro de Fourier, uma vez que existe uma relação proporcional inversa entre escala e frequência, as informações frequenciais do Quadro 2 são úteis para que se possa fixar as escalas de análise da CWT.

Visto que os sinais de ECG foram amostrados a 360Hz, então, quando a frequência central da Wavelet Complexa de Morlet é igual a 1, na escala $s = 1$ os coeficientes wavelets representam informações em torno dos 360Hz. Portanto, considerando as frequências mínimas e máximas dos batimentos analisados, Quadro 2, as escalas escolhidas devem variar de $s = 10$ a $s = 1000$, onde na escala $s = 10$ aparecerão as componentes espectrais em torno de 36Hz e na escala $s = 1000$ em torno de 0,36Hz. No entanto, essa escala muito alta pouca informação adiciona à análise, pois, a maior parte da energia está concentrada nas frequências mais altas. Sendo assim, aqui será fixada a escala máxima igual $s = 100$, a qual conterá informações em torno de 3,6Hz. Embora nem todo o conteúdo espectral seja coberto com esta escolha, destaca-se que na aplicação da transformada wavelets existe o fenômeno denominado de escape de energia. Isso significa que outras frequências além daquelas mencionadas poderão ser cobertas, já que, as respostas em frequências das funções wavelets não são ideais. Mais detalhes podem ser consultados em Oliveira et al. (2018).

Para exemplificar como as informações espectrais variam entre escalas, na Figura 1 (a) está ilustrado um segmento inicial do sinal de ECG do registro 233 e na Figura 1 (b) o escalograma dos coeficientes wavelets deste sinal, o qual é obtido calculando o quadrado do valor absoluto desses coeficientes. As escalas variam entre 10 e 100 com um incremento de 5. Na Figura 1 (b), as cores mais claras e mais escuras representam os maiores e menores valores, respectivamente. Na Figura 1 (a) o 1º e o 3º batimentos são da classe contração prematura ventricular, enquanto os demais são da classe normal.

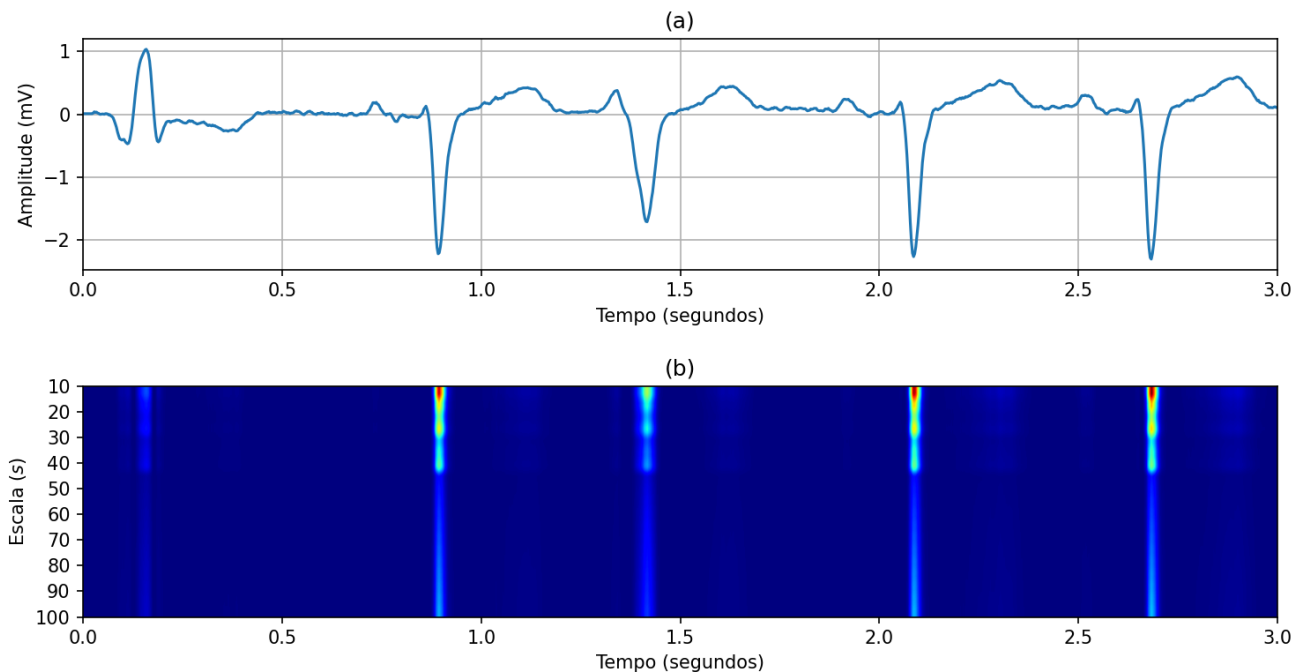


Figura 1. A curva (a) representa os 3 segundos iniciais do sinal de ECG do registro 233 e a imagem (b) o escalograma dos coeficientes wavelets.

Comparando as porções do escalograma relativas aos batimentos de classes distintas nota-se que os coeficientes wavelets refletem informações espectrais distintas para estas classes de batimentos cardíacos nas escalas selecionadas, como por exemplo: a faixa de coeficientes significantes que aparece (azul bem claro) imediatamente ao lado esquerdo, somente para os batimentos do tipo contração prematura ventricular. Além disso, é notável que há maior concentração de energia nas localizações das ondas R para os batimentos do tipo normal do que para as contrações prematuras ventriculares, já que esta classe de arritmia tem frequência máxima em torno de 12Hz, conforme Quadro 2.

Para as fases de treinamento (indução) e teste dos modelos de reconhecimento de padrões foram fixados dois conjuntos de registros de ECG, de acordo com o Quadro 3. Essa separação, com alguma permutação entre os registros, é bastante comum na literatura especializada, pois os registros selecionados para treinamento possuem características variadas possibilitando que o aprendizado seja mais diverso e com isso seja atingida uma melhor capacidade de generalização (Oliveira, 2020). Além disso, essa divisão proporciona uma distribuição adequada das classes positiva e negativa para ambas as fases.

Quadro 3. Registros de ECG utilizados nas fases de treinamento e teste.

Registros	Fase
101, 102, 104, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230, 232, 234	Treinamento
100, 103, 105, 107, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 217, 219, 221, 222, 228, 231, 233	Teste

Na abordagem proposta, os coeficientes wavelets foram empregados como vetores de atributos para o treinamento dos modelos de reconhecimento de padrões. Visto que as escalas para aplicação da CWT devem ser selecionadas no intervalo $[10, 100]$ e a quantidade destas depende do incremento utilizado para gerar os valores nesse intervalo, então, há muitos vetores de coeficientes wavelets, implicando na indução de muitas máquinas de aprendizado, para um mesmo padrão. Para simplificar o processo de aprendizagem e reduzir o custo computacional, foram realizadas três médias dos coeficientes dentro deste intervalo, e os vetores médios foram empregados para a indução de três máquinas⁷ de aprendizado. A primeira delas aprenderá os padrão médios no intervalo $[10, 40]$, a segunda máquina no intervalo $]40, 70]$ e a terceira no intervalo $]70, 100]$.

RESULTADOS E DISCUSSÃO

O emprego da abordagem proposta resulta, na fase de treinamento, em 7.715 instâncias da classe positiva e 54.469 da classe negativa, enquanto na fase de teste, 4.726 instâncias da classe positiva e 46.302

⁷ Pode-se utilizar também outras quantidades de máquinas à critério do projetista.

da classe negativa. Após induzidos os três modelos de reconhecimento de padrões, empregando as médias dos coeficientes conforme especificado na seção anterior, obtém-se os resultados da Tabela 1. Os pesos dos modelos M0, M1 e M2, obtidos a partir da acurácia, foram de 0,2052, 0,3781 e 0,4166, significando que os modelos apresentam diferentes valores de acurácia, logo sua importância geral para reconhecimento dos padrões é significativamente distinta.

Tabela 1. Resultados obtidos nas fases de treinamento e teste, para cada máquina induzida e também para o comitê.

Máquina M0			Máquina M1			Máquina M2		
Acc	Pr	Re	Acc	Pr	Re	Acc	Pr	Re
Fase de treinamento								
0,4330	0,1385	0,7448	0,7975	0,1655	0,1841	0,8788	0,4712	0,3640
Fase de teste								
0,3264	0,0774	0,5749	0,9062	0,4920	0,3912	0,9188	0,6443	0,2463
Comitê de Máquinas								
Acc			Pr			Re		
0,9199			0,5921			0,4344		

A partir dos resultados apresentados na Tabela 1, observa-se que a máquina M0, que foi induzida empregando os coeficientes nas escalas 10 a 40, gerou os piores resultados de acurácia (Acc) e precisão (Pr), enquanto que a máquina M2, que empregou as escalas mais altas, acima de 70 até 100, gerou a mais alta acurácia (Acc) e precisão (Pr) em ambas fases. Apenas na fase de treinamento esta máquina também retornou um recobrimento (Re) maior que àquele da máquina M1, a qual foi induzida com os coeficientes nas escalas acima de 40 até 70. Entretanto, as máquinas M1 e M2, em ambas fases, geraram um recobrimento (Re) menor que àquele da máquina M0.

Em linhas gerais, estes resultados significam que a utilização de escalas mais altas resulta em máquinas cujos modelos predizem melhor as instâncias da classe positiva, pois a precisão foi maior para as máquinas M1 e M2, implicando em menos Falsos Positivos (FP). Por outro lado, nas escalas mais baixas as máquinas induzidas predizem melhor as instâncias da classe negativa, menos Falsos Negativos (FN), por isso o recobrimento foi maior para a máquina M0. Esta constatação está de acordo com o que foi observado no fragmento de ECG e seu respectivo escalograma, ilustrados na Figura 1, a partir da qual podemos observar que existe diferença na distribuição de energia entre escalas para diferentes classes de batimentos cardíacos.

Deste modo, o algoritmo Naïve Bayes construiu três modelos distintos, pois foram utilizados dados distintos referentes à diferentes bandas de frequência. Como cada máquina é melhor para reconhecer determinado tipo de padrão, a utilização de um comitê de máquina é imprescindível. Sua utilidade fica evidente nos resultados obtidos, pois foi possível combinar o melhor de cada uma das máquinas. Ou seja, se utilizássemos apenas uma das máquinas, ou teríamos uma acurácia e precisão mais

altas, ou um recobrimento mais alto. Por outro lado, ao combiná-las, consegue-se obter uma acurácia maior que os resultados individuais, e uma precisão e recobrimento mais equilibrados.

Embora a precisão do comitê de máquinas seja inferior à precisão da Máquina M2 em 0,0523, o valor do recobrimento é maior para o comitê em 0,1881. Analogamente ocorre para o recobrimento na comparação do comitê de máquinas e da máquina M0. Portanto, embora haja perda de performance em relação à alguma medida, o ganho obtido ao utilizar o comitê de máquina é bastante superior.

Adicionalmente, um resultado surpreendente que observamos na Tabela 1 é que as máquinas M1 e M2 geraram resultados superiores na fase de teste em relação à fase de treinamento. É surpreendente pois na fase de treinamento os mesmos dados foram utilizados para induzir o modelo e também para predição, ao contrário do que foi implementado na fase de teste.

Em geral, a abordagem proposta é efetiva para o reconhecimento dos padrões eletrocardiográficos, sendo que a melhor acurácia obtida foi de 91,99%. Tal resultado foi conseguido empregando uma técnica de aprendizado de máquina de baixo custo computacional, a qual se baseia apenas no cálculo de probabilidade e na estimação de dois parâmetros para cada atributo, que no presente estudo foi de 288⁸ atributos por instância. Além disso, não foi empregada nenhuma fase de pré-processamento, mas utilizados os dados crus provenientes da base dados, sendo que alguns destes possuíam ruídos típicos encontrados em sinais de ECG (Oliveira, 2020).

Para concluir, destaca-se que a abordagem proposta para reconhecimento de padrões cardíacos normais ou arrítmicos ainda está em estágio inicial de desenvolvimento e, portanto, carece de melhorias e ajustes. Vários outros caminhos podem ser tentados a fim de melhorar a performance de predição, dentre eles: testar outras funções wavelets⁹ e variações de escala que possam, porventura, serem mais adequadas à análise eletrocardiográfica; verificar se outros comprimentos de janelas de análise dos sinais de ECG possam contribuir com mais informação relevante; empregar outros modelos de aprendizados de máquinas mais robustos e ampliar a base de dados empregada; verificar como o método de pré-processamento empregado pode melhorar a predição.

Os arquivos (scripts) utilizados para o desenvolvimento da pesquisa apresentada estão disponíveis neste link <https://github.com/brunobro/reconhecimento-de-arritmias-utilizando-transformada-wavelet>.

REFERÊNCIAS BIBLIOGRÁFICAS

AAMI (1987). Practice for Testing and Reporting Performance Results of Ventricular Arrhythmia Detection Algorithms. Arlington.

⁸ Esse valor é devido ao tamanho da janela de análise escolhido, que foi de 0.8 segundos.

⁹ Uma função que vem ganhando destaque na análise de ECG e que foi recentemente proposta são as Golden Wavelets Gossler et al. (2016).

- Barrett PM et al. (2014). Comparison of 24-hour Holter monitoring with 14-day novel adhesive patch electrocardiographic monitoring. *The American journal of medicine*, 127(1): 11-95. DOI: 10.1016/j.amjmed.2013.10.003.
- CDC (2020). Centers for Disease Control and Prevention. *Circulation*, 141(9): e139-e596.
- Daubechies I (1992). *Ten Lectures on Wavelets*. Auckland: Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611970104.
- Dietterich TG (2000). Ensemble methods in machine learning. Dietterich TG. *Multiple Classifier Systems*. Berlin: Springer. DOI: 10.1007/3-540-45014-9_1.
- Dimarco JP, Philbrick JT (1990). Use of ambulatory electrocardiographic (Holter) monitoring. *Annals of internal medicine*, 113(1): 53-68.
- Goldberger AL et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220.
- Gossler FE et al. (2016). A wavelet generated from Fibonacci-coefficient polynomials and its application in cardiac arrhythmia classification. In *Proc. of XIX ENMC-National Meeting on Comp. Model. and VII ECTM-Meeting on Materials Science and Tech.*
- Guyton AC, Hall JE (2006). *Tratado de fisiologia médica*. 11ª ed. Rio de Janeiro: Elsevier.
- Hampton JR (2014). *ECG Essencial*. 8ª ed. Rio de Janeiro: Elsevier.
- Kuchar DL et al. (1987). Prediction of serious arrhythmic events after myocardial infarction: signal-averaged electrocardiogram, Holter monitoring and radionuclide ventriculography. *Journal of the American College of Cardiology*, 9(3): 531-538.
- Kubat M (2015). *An Introduction to Machine Learning*. New York: Springer.
- Latchamsetty R, Bogun F (2015). Premature ventricular complexes and premature ventricular complex induced cardiomyopathy. *Current Problems in Cardiology*, 40: 379-422.
- Lipski J et al. (1976). Value of Holter monitoring in assessing cardiac arrhythmias in symptomatic patients. *The American journal of cardiology*, 37(1): 102-107.
- Mallat S (2009). *A Wavelet Tour of Signal Processing: The Sparse Way*. 3ª ed. Burlington: Elsevier.
- Mohrman DE (2011). *Fisiologia cardiovascular*. 6. ed. Porto Alegre: McGraw-Hill.
- Nicolau JC et al. (2003). Diretriz de interpretação de eletrocardiograma de repouso. *Arquivo Brasileiro de Cardiologia*, São Paulo, 80(2):1-18. DOI: 10.1590/S0066-782X2003000800001.
- Oliveira BR de et al. (2018). A wavelet-based method for power-line interference removal in ECG signals. *Research on Biomedical Engineering*, 34(1): 73-86. DOI: [10.1590/2446-4740.01817](https://doi.org/10.1590/2446-4740.01817).
- Oliveira BR de et al. (2019). Geometrical features for premature ventricular contraction recognition with analytic hierarchy process based machine learning algorithms selection. *Computer methods and programs in biomedicine*, 169: 59-69. DOI: 10.1016/j.cmpb.2018.12.028.
- Oliveira BR de (2020). Reconhecimento de Contração Prematura Ventricular utilizando separação cega de fontes e comitê de máquinas bayesianas. Programa de Pós-graduação em Engenharia Elétrica

(Tese), Ilha Solteira-SP, 151p. Disponível em <<http://hdl.handle.net/11449/194111>>. DOI: [10.13140/RG.2.2.24356.60802](https://doi.org/10.13140/RG.2.2.24356.60802).

Santos JD et al. (2018). Mortalidade por infarto agudo do miocárdio no Brasil e suas regiões geográficas: análise do efeito da idade-período-corte. *Ciência & Saúde Coletiva*, 23: 1621-1634. DOI: 10.1590/1413-81232018235.16092016.

Shafqat S et al. (2004). Holter monitoring in the diagnosis of stroke mechanism. *Internal medicine journal*, 34(6): 305-309.


Soman KP et al. (2010). *Insight Into Wavelets: From Theory to Practice*. 3ª ed. New Delhi: PHI.

Suzuki T et al. (1994). Holter monitoring as a noninvasive indicator of cardiac involvement in sarcoidosis. *Chest*, 106(4): 1021-1024.


Sundararajan D (2001). *The Discrete Fourier Transform. Theory, Algorithms and Applications*. New Jersey: World Scientific.


Woods SL et al. (2005). *Enfermagem em cardiologia*. Barueri: Manole.

Uso da mineração de textos na análise exploratória de artigos científicos

 10.46420/9786581460204cap4

Geraldo Nunes Corrêa^{1*} 

Solange Oliveira Rezende² 

Ricardo Marcondes Marcacini² 

INTRODUÇÃO

A Mineração de Textos (MT) pode ser definida como um conjunto de técnicas e processos para descoberta de conhecimento inovador a partir de dados textuais (Ebecken et al., 2003). Em um contexto no qual grande parte da informação corporativa, como e-mails, memorandos internos e blogs industriais, é registrada em linguagem natural, a MT surge como uma poderosa ferramenta para gestão do conhecimento.



Figura 1 Etapas do processo de mineração de textos (Rezende et al., 2003).

Dentro do contexto de MT, as hierarquias de tópicos desempenham um papel importante na recuperação e organização de informação, principalmente em tarefas de busca exploratória. Nesse tipo

¹ Universidade do Estado de Minas Gerais, Frutal

² Instituto de Ciências Matemáticas e Computação de São Carlos, USP

* Autor correspondente: geraldo.correa@uemg.br

de tarefa, o usuário geralmente tem pouco domínio sobre o tema de interesse, o que dificulta expressar o objetivo diretamente por meio de palavras-chave (Marchionini, 2006).

Assim, torna-se interessante disponibilizar previamente algumas opções para guiar o processo de busca da informação. Para tal, cada grupo possui um conjunto de descritores que contextualizam e indicam o significado dos documentos ali agrupados. Essa organização está relacionada com a hipótese de que se um usuário está interessado em um documento específico pertencente a um determinado tópico, deve também estar interessado em outros documentos desse tópico e de seus subtópicos (Manning et al., 2008).

Desta forma, neste trabalho a MT foi utilizada para descobrir conhecimento útil a partir de coleções textuais, o que viabiliza sobremaneira a análise exploratória de documentos científicos. Assim, durante a fase de Extração de Padrões, métodos de agrupamento de documentos foram utilizados para a organização de coleções textuais de maneira não supervisionada.

Em tarefas de agrupamento, o objetivo é organizar um conjunto de objetos em grupos, em que objetos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos objetos de outros grupos. Em outras palavras, o agrupamento é baseado no princípio de maximizar a similaridade intragrupo e minimizar a similaridade intergrupos (Everitt et al., 2001). Os métodos de agrupamento também são conhecidos como algoritmos de aprendizado por observação ou análise exploratória dos dados, pois a organização obtida é realizada por observação de regularidades nos dados, sem uso de conhecimento externo (Xu; Wunsch, 2008).

Em resumo, neste trabalho foi aplicado um método de agrupamento de documentos, visando o aprendizado não supervisionado de hierarquias de tópicos em coleções textuais envolvendo artigos científicos na área médica com o objetivo de busca exploratória sobre um determinado tema de pesquisa.

O campo de estudo escolhido para a realização deste trabalho foi um Hospital que atua no tratamento contra o Cancer e que realiza milhares de atendimentos diários nos mais diferentes tipos da doença. Tal hospital mantém um Instituto de Ensino e Pesquisa, formado por uma equipe multidisciplinar preparada para oferecer suporte aos colaboradores e alunos de mestrado e que tem por objetivo promover o desenvolvimento da pesquisa científica na instituição. O Instituto alunos de mestrado e doutorado, que atuam nas seguintes linhas de pesquisa.

1. Biologia Tumoral
2. Cuidados Paliativos e Qualidade de Vida
3. Epidemiologia Clínica e Molecular em Oncologia
4. Fatores Ambientais e Câncer
5. Cirurgia Experimental e Minimamente Invasiva

DESCRIÇÃO DO USO DE MINERAÇÃO DE TEXTOS PARA APOIAR O LEVANTAMENTO BIBLIOGRÁFICO PARA PESQUISA MÉDICA

A área de pesquisa específica foi selecionada para aplicar o processo de Mineração de Textos descrito na seção anterior. Abaixo está instanciado o processo utilizado nesta primeira atividade com o Hospital do Câncer.

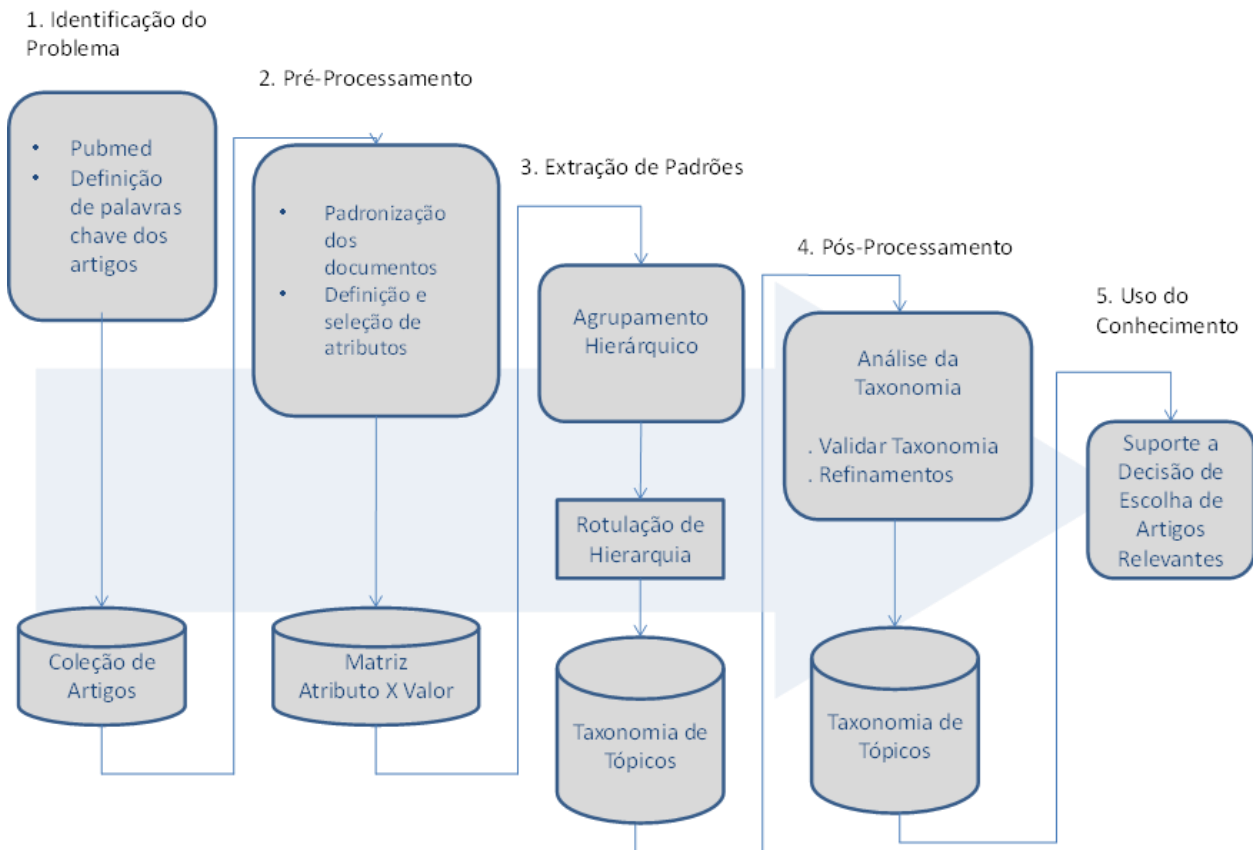


Figura 2. Processo de Mineração de Texto Aplicado no Hospital de Câncer: os autores

Identificação do Problema

Nesta etapa inicial foi definido que o tema de pesquisa a ser explorado no processo de mineração de texto é o câncer de cabeça e pescoço. Dentro deste contexto de pesquisa, existem vários artigos científicos e o problema é a identificação de artigos relevantes ao pesquisador. Neste sentido, o uso de métodos de agrupamento hierárquico é adequado para a resolução de problemas, uma vez que aborda a busca exploratória de documentos de interesse ao pesquisador.

Conforme orientação do médico pesquisador do Hospital do Câncer, foi indicada uma biblioteca digital relevante dentro da área de pesquisa, a *US National Center for Biotechnology Information*. Dentro desta biblioteca foi utilizada a base de dados Pubmed, que inclui artigos referentes à área médica. Para se ter uma ideia da dimensão da quantidade de artigos, uma simples pesquisa usando as palavras chaves *head and neck cancer* retornou mais 230.000 artigos, conforme pode ser observado na figura seguinte.

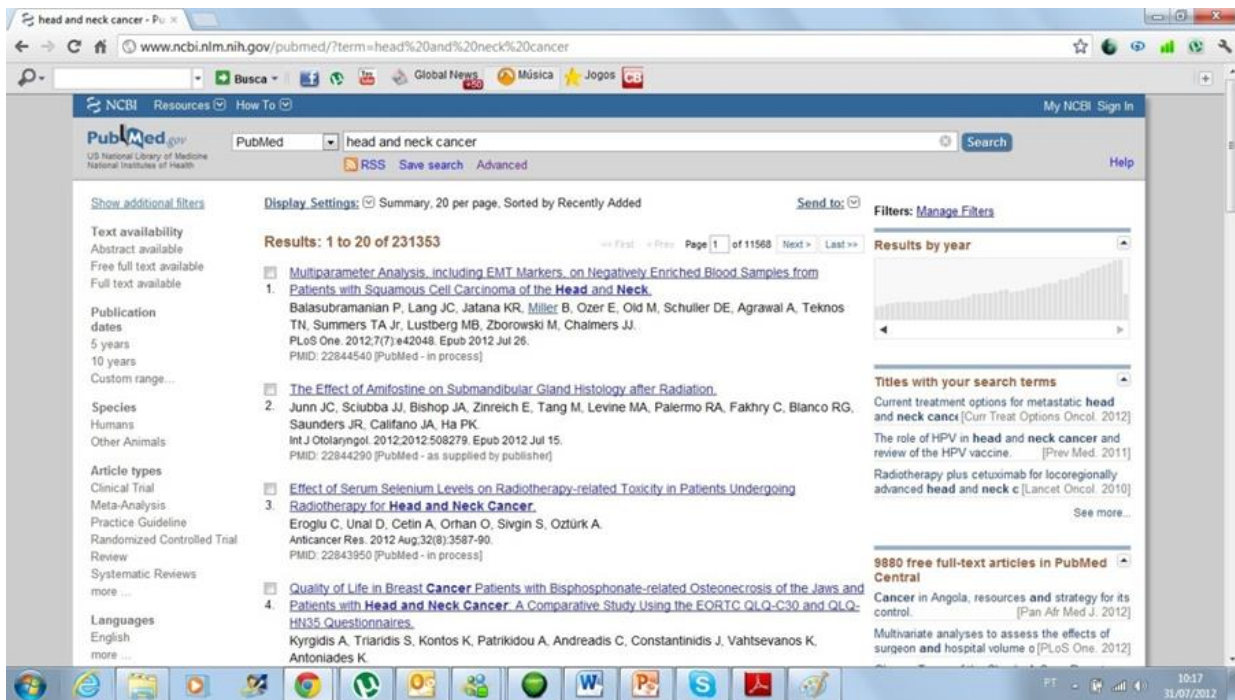


Figura 3. Pesquisa inicial de artigos sobre câncer de cabeça e pescoço: os autores

No entanto, um recurso de busca avançada da biblioteca pôde ser utilizado, reduzindo o número total de artigos de interesse. Outro recurso importante da biblioteca digital Pubmed é o mecanismo de exportação dos artigos. A partir do resultado de uma consulta, tanto o artigo completo como o resumo. Entre os formatos que podem ser escolhidos estão o pdf e xml.

Pré-Processamento

Em posse da coleção de arquivos, seja eles completos ou resumos, o passo seguinte é a padronização dos textos, ou seja, os documentos são convertidos para a forma de texto plano sem formatação. Para isso, foram desenvolvidos dois scripts, um para converter arquivos do formato pdf e outro do formato xml.

O código abaixo demonstra o parser dos arquivos xml extraídos da Pubmed.

```
<?php
$xml = simplexml_load_file("pubmed_result.xml");
foreach($xml->PubmedArticle as $PubmedArticle){

    $titulo = $PubmedArticle->MedlineCitation->Article->ArticleTitle;

    $id = $PubmedArticle->MedlineCitation->PMID;

    $abstract = $PubmedArticle->MedlineCitation->Article->AbstractText;

    $dia = $PubmedArticle->MedlineCitation->Article->ArticleDate->Day;

    $mes = $PubmedArticle->MedlineCitation->Article->ArticleDate->Month;

    $ano = $PubmedArticle->MedlineCitation->Article->ArticleDate->Year;
```

```

$data = $ano.$mes.$dia;

if(strlen($titulo) > 20 && strlen($abstract) > 100 && strlen($id) == 8 && strlen($data)==8){

    echo "Title: ".$titulo."\n";

    echo "Abstract: ".$abstract."\n";

    echo "ID: ".$id."\n";

    echo "Date: ".$data."\n";

}

}

?>

```

Extração de Padrões

Nesta etapa, o objetivo é organizar o conjunto de artigos científicos em grupos, baseado em uma medida de proximidade, na qual artigos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos artigos de outros grupos. Ainda nesta etapa, a análise de agrupamento também é conhecida como aprendizado por observação ou análise exploratória dos dados, pois a organização dos objetos em grupos é realizada apenas pela observação de regularidades nos dados, sem uso de conhecimento externo, ou seja, não supervisionada.

Esta etapa de extração de padrões é apoiada por uma ferramenta desenvolvida no Laboratório de Inteligência Computacional - Labic – da USP de São Carlos, denominada TORCH (*Topic Hierarchy*), que realiza a tarefa de agrupamento utilizando os conceitos citados.

Pós-processamento

A avaliação do conhecimento extraído é realizada de forma subjetiva, utilizando o conhecimento do pesquisador. No contexto deste trabalho, a qualidade da hierarquia de tópicos está diretamente relacionada com a qualidade do agrupamento na extração de padrões. Neste trabalho não são utilizados índices estatísticos para expressar o “mérito” das estruturas encontradas, ou seja, para quantificar alguma informação sobre a qualidade de um agrupamento. Tal tarefa é realizada pelo especialista de domínio.

Uso do Conhecimento

Na etapa de uso do conhecimento, os resultados são validados pelo médico especialista tornam-se aptos a serem utilizados para apoiar a decisão de escolha dos artigos a serem utilizados na pesquisa científica, conforme os objetivos estabelecidos na etapa de Identificação do Problema.

RESULTADOS

Com base no problema descrito, foram coletados da Pubmed, 91 artigos completos no formato pdf. Na sequência, esta coleção de artigos foram transformados para o formato txt, sendo que destes 5 não foram convertidos por problemas do OCR (*Optical Character Recognition*). Desta forma 84 artigos representaram a base inicial de documentos para a etapa de pré-processamento.

Tendo como base esta coleção de documento, foi utilizada a ferramenta TORCH para realização da etapa de pré-processamento e geração da hierarquia de tópicos. A ferramenta foi configurada para a apresentação de 7 níveis, sendo cada nível rotulado por 3 descritores conforme algoritmo para esta finalidade.

Como pode ser observado na Figura 4, a rotulação dos 7 primeiros níveis ficou da seguinte maneira:

1. *Cells, Cancers, Expressed;*
2. *Patients, Treatments, Carcinomas;*
3. *Thyroids, Patients, Cancers;*
4. *Tumors, Rnas, Cancers;*
5. *Cancers, Patients, Necks;*
6. *Patients, Cancers, Survive;*
7. *Cancers, Patients, Treaments;*

Observando a rotulação de cada nível, percebeu-se a repetição de termos em diferentes níveis. Além disso, dois níveis, 2 e 7, podem ser considerados iguais, uma vez que o termo *Carcinomas* é sinônimo do termo *Cancers*.

Partindo da identificação deste problema na geração do agrupamento, a ferramenta foi utilizada novamente utilizando-se diferentes configurações de níveis para verificar o agrupamento gerado.

Após a geração dos agrupamentos com diferentes níveis foi realizada uma reunião com o médico pesquisador. A questão da rotulação dos níveis foi discutida e um outro problema foi identificado. Como o pré-processamento foi realizado em artigos completos, um componente é comum a todos os documentos é o tópico de introdução. Em tal tópico sempre existem palavras (termos) que são comuns em vários artigos, possuindo, assim, uma alta frequência, gerando resultados distorcidos na etapa de pré-processamento.

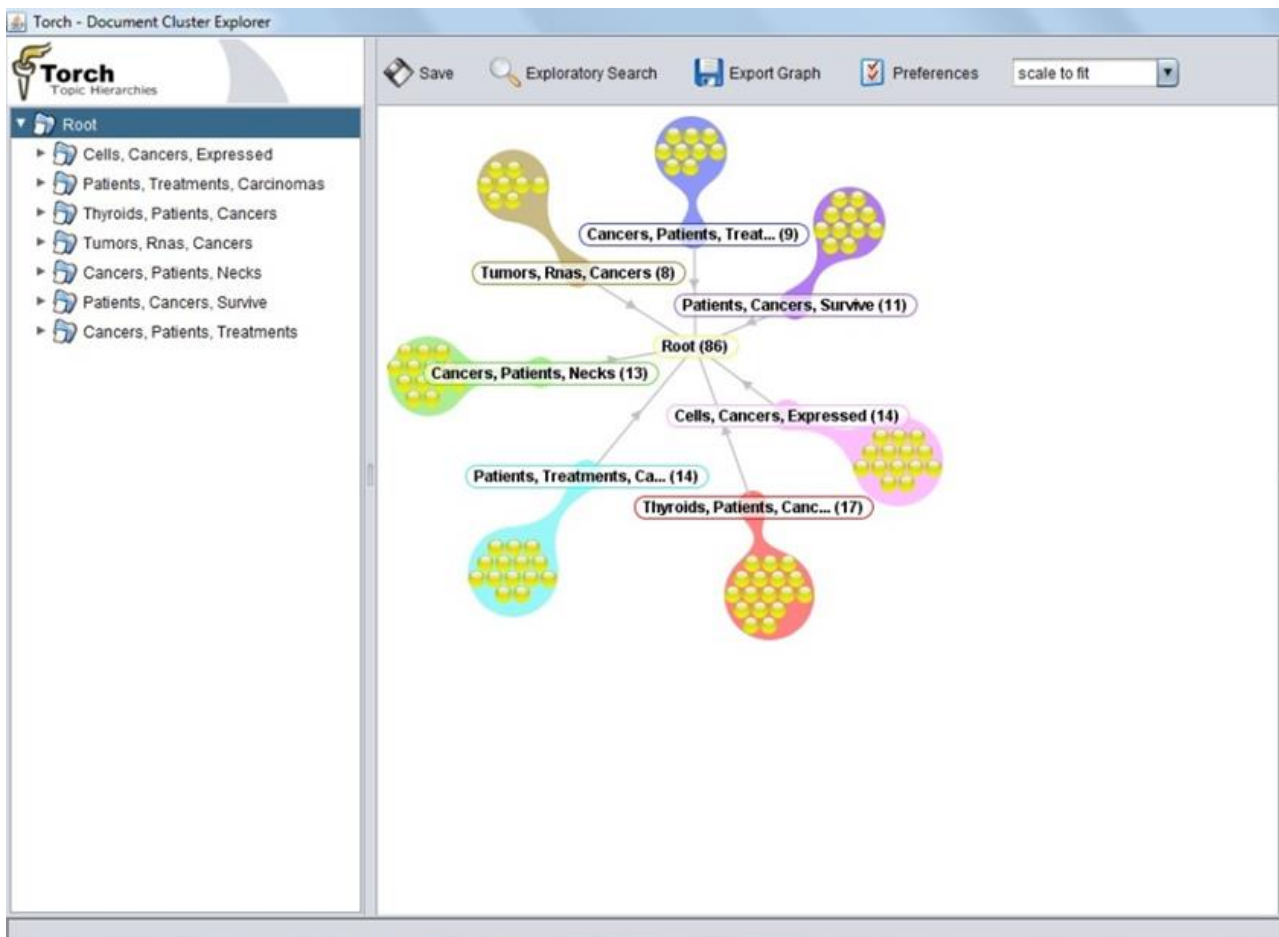


Figura 4. Geração do agrupamento dos artigos completos em 7 níveis: os autores

Desta forma, outra estratégia foi adotada para a busca exploratória na área de pesquisa selecionada neste projeto. Ao invés de artigos completos, o médico pesquisador solicitou a realização da tarefa de pré-processamento sobre os *abstracts* dos artigos, nos quais estão presentes as palavras (termos) que representam a essência da proposta do autor.

Na sequência, foram gerados em formato xml os *abstracts* dos mesmos artigos selecionados na estratégia anterior. O resultado alcançado a partir de tal coleção de documentos é apresentado na figura seguinte.

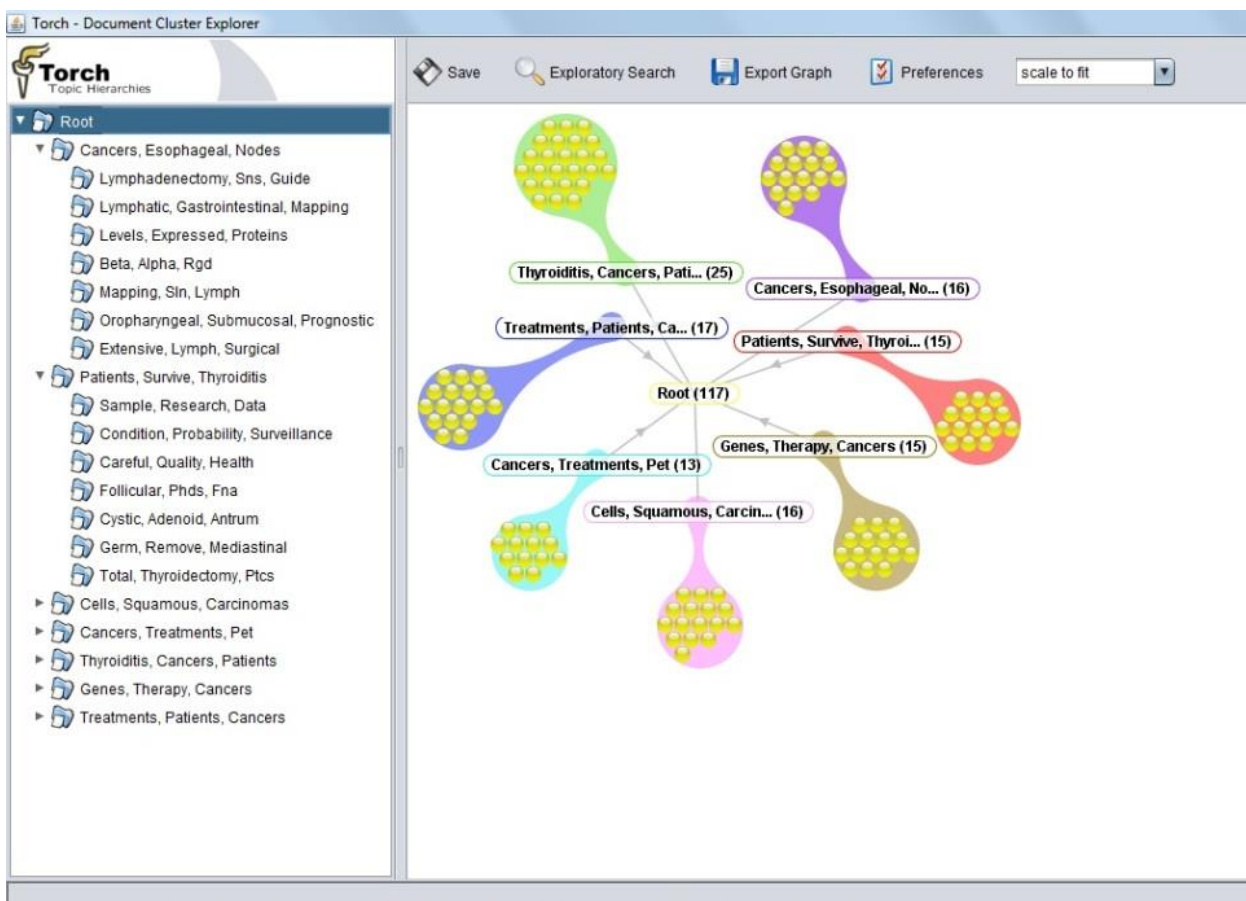


Figura 5. Geração do agrupamento a partir dos resumos dos artigos: os autores.

CONCLUSÕES

No decorrer do desenvolvimento deste trabalho, um dos objetivos principais foi a extração de hierarquias de tópicos a partir bases de artigos médicos sobre câncer. As bases de artigos científicos disponíveis na PubMed representam a evolução do conhecimento da área ao longo do tempo. Ainda, a quantidade de informação publicada na PubMed excede a capacidade humana de analisá-la manualmente, incentivando o uso de técnicas de mineração de textos.

A partir da análise das hierarquias de tópicos extraídas da base de artigos coletadas, é possível concluir que métodos não supervisionados foram eficazes para extrair tópicos mais genéricos sobre os dados (níveis mais altos da hierarquia). Estes tópicos são úteis para realizar uma primeira análise exploratória com usuários que não possuem conhecimento aprofundado sobre o assunto descrito nos artigos. Por outro lado, os tópicos mais genéricos não representam conhecimento inovador para usuários especialistas do domínio.

Assim, métodos não supervisionados tendem a selecionar os termos mais frequentes dos textos, o que leva a formação de tópicos mais genéricos. A inclusão de um dicionário ou ontologia de domínio para apoiar a seleção de termos mais específicos, bem como técnicas de aprendizado ativo que permitam a inclusão de especialistas de domínio na extração de tópicos, são de fundamental importância para

suportar a atividade humana de realizar a pesquisa científica com um grau de profundidade mais específica.

REFERÊNCIAS BIBLIOGRÁFICAS

- Ebecken NFF et al. (2003). Mineração de textos. In Rezende SO (editor), *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 13: 337-370. Manole, 1ª edição. Citado na página 3.
- Everitt BS et al. (2001). *Cluster Analysis*. Arnold Publishers. Citado nas páginas 1, 9, 11, 16.
- Liu L et al. (2005). *A comparative study on unsupervised feature selection methods for text clustering*. In NLP-KE '05. Proceedings of 2005 International Conference on Natural Language Processing and Knowledge Engineering, páginas 597_601. Citado nas páginas 8, 12.
- Luhn HP (1958). *The automatic creation of literature abstracts*. IBM Journal os Research and Development, 2(2):159_165. Citado na página 6.
- Manning CD et al. (2008). *An Introduction to Information Retrieval*. Cambridge University Press. Citado nas páginas 1, 4, 6, 15.
- Rezende SO et al. (2003). Mineração de dados. In Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 12, Manole, 1ª edição., 307-335. Citado nas páginas 1, 3, 8.
- Xu R, Wunsch D (2008). *Clustering*. Wiley-IEEE Press, IEEE Press Series on Computational Intelligence. Citado nas páginas 1, 9, 13, 16.

ÍNDICE REMISSIVO

A

algoritmo, 13, 25, 27, 42, 51
aprendizado de máquina, 4, 9, 13, 16, 17, 35,
37, 43
arritmia, 41

C

classificação, 4, 17, 37, 39
conhecimento, 46, 47, 50, 53
convolucional
rede neural, 4, 21

D

Dense
rede neural, 4, 22

E

Elman
rede neural, 4, 8, 9, 10, 12, 14, 15
energia elétrica, 4, 8, 16, 17, 21, 24, 27, 29
ensemble, 31

J

Jordan
rede neural, 4, 8, 9, 10, 12, 14, 15

M

mineração de texto, 4, 48

P

padrão, 24, 33, 37, 38, 41, 42
petróleo, 4, 7, 8, 10
predição, 4, 16, 17, 18, 21, 25, 26, 28, 29, 38, 43
pré-processamento, 43, 51, 52
Pubmed, 48, 49, 51


R


recorrente
rede neural, 4, 8, 21, 22
rede neural, 7, 8, 9, 17, 20


S


série temporal, 7, 8, 16, 17, 21, 24

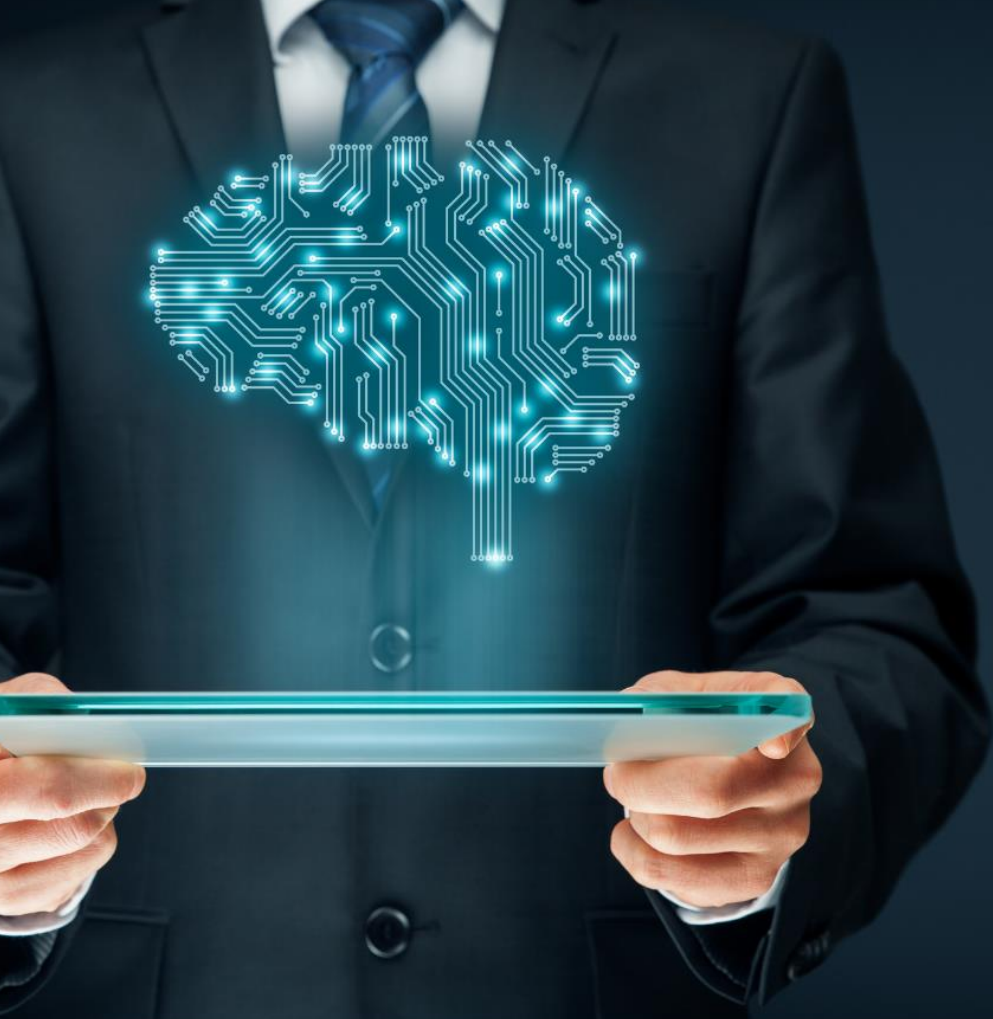
SOBRE OS ORGANIZADORES

 **Ricardo Augusto Manfredini** Possui graduação em Bacharel em Ciências da Computação pela Universidade de Caxias do Sul (1990), mestrado em Ciências da Computação pela Universidade Federal do Rio Grande do Sul (2001) , doutorado pelo Instituto de Biotecnologia da Universidade de Caxias do Sul(2015) e pós-doutorado no GECAD do Instituto Superior de Engenharia do Porto em Inteligência Artificial e IoT, também é professor do Instituto Federal de Ciências e Tecnologia do Rio Grande do Sul - campus Farroupilha. Tem experiência na área de Ciência da Computação, com ênfase em Programação, Engenharia de Software e Tolerância a Falhas, atuando principalmente nos seguintes temas: IoT & IA (2020), bioinformática, injeção de falhas, sistemas distribuídos, tolerância a falhas, metodologias de desenvolvimento de sistemas e linguagens de programação.

 **Geraldo Nunes Corrêa** Possui graduação em Ciência da Computação pela Universidade de São Paulo (1991), mestrado em Ciências da Computação (Área de Inteligência Artificial, Banco de Dados e Manufatura) pela Universidade de São Paulo (1994) e doutorado em Engenharia Mecânica (Departamento de Engenharia de Produção) pela Universidade de São Paulo (1999). Pós doutorado em Mineração de Textos no Instituto de Ciências Matemáticas e Computação da Universidade de São Paulo (2013). Consultor em soluções educacionais. Mentor de Startups. Empreendedor Digital.

 **Bruno Rodrigues de Oliveira** Graduado em Matemática pela Universidade Estadual de Mato Grosso do Sul (UEMS, 2008). Mestrado (2015) e Doutorado (2020) em Engenharia Elétrica pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP, 2015). Atualmente é Analista Judiciário no Tribunal de Justiça de Mato Grosso do Sul e professor de Matemática no Colégio Maper e Editor na Pantanal Editora. Tem experiência nos temas: Matemática, Processamento de Sinais via Transformada Wavelet, Análise Hierárquica de Processos, Teoria de Aprendizagem de Máquina e Inteligência Artificial.

 **Suellen Teixeira Zavadzki de Pauli** Doutoranda em Métodos Numéricos na Universidade Federal do Paraná, possui mestrado em Engenharia da Produção pela Universidade Federal do Paraná (2020), pós-graduação em Engenharia da Confiabilidade pela Universidade Tecnológica Federal do Paraná (2016) e graduação em Estatística pela Universidade Federal do Paraná (2013). Possui experiência com monitoramento de modelos estatísticos de crédito e análise de indicadores. Atuação no planejamento e estratégia de campanhas de marketing através da análise de produtos e de perfil de clientes, dashboards analíticos e KPIs. Geração e implementação de simulador de oferta ideal para o cliente (next best offer). Experiência na avaliação de dados e cenários, utilização de técnicas estatísticas (análises descritivas, árvores de decisão, testes estatísticos). Experiência como professor substituto na Universidade Tecnológica Federal do Paraná (2020). Experiência com pesquisa em redes neurais artificiais aplicadas em dados de séries temporais (preço de petróleo e valor de ações), modelos mistos, análise de variância, entre outras técnicas. Familiaridade com linguagem SQL e ferramentas de análises como SAS, SPSS, R, Phyton e pacotes office. Coorganizadora do RLadies Curitiba.



Pantanal Editora

Rua Abaete, 83, Sala B, Centro. CEP: 78690-000

Nova Xavantina – Mato Grosso – Brasil

Telefone (66) 99682-4165 (Whatsapp)

<https://www.editorapantanal.com.br>

contato@editorapantanal.com.br