10º Encontro de Ensino Pesquisa e Extensão

Patrocínio, MG, outubro de 2023

PREDIÇÃO DE CRISES DE EPILEPSIA UTILIZANDO APRENDIZAGEM DE MÁQUINA

Vanessa Machado Silva
Daniele Carvalho Oliveira
UFU Monte Carmelo
Modalidade: Pesquisa
Formato: Artigo Completo

Resumo:

Pessoas que possuem epilepsia sofrem de convulsões devido à perturbação da atividade das células nervosas do cérebro. Para diagnosticar a epilepsia é utilizada a eletroencefalografia, que faz o registo da atividade cerebral, amplamente usado em diagnósticos clínicos por ser realizado de forma não invasiva, além de permitir a captação e alta resolução temporal. No entanto, o eletroencefalograma(EEG) fornece dados muito extensos, uma vez que é feita durante horas de captura, consequentemente, ao realizar a análise é necessário uma extensa inspeção visual. A pesquisa propôs a criação de um sistema automatizado para realizar a classificação dos sinais do EEG. A pesquisa seguiu etapa para a criação do sistema, como a seleção da base de dados do CHB-MIT Scalp EEG Database, pré-processamento, extração de características implementação dos modelos de classificação *K-Nearest Neighbors*(KNN), *Support Vector Machine*(SVM), *Decision Trees(*DTs), *Random Forest*(RF) e *Logistic regression*(LR). Em conclusão, o objetivo é a criação de um sistema automatizado para predição de crises de epilepsia com dados coletados pelo EEG, uma vez que o EEG é composto de extensos dados, o que torna a análise manual uma tarefa árdua. A pesquisa utiliza técnicas computacionais para uma classificação de forma mais autônoma e precisa. Portanto, essa pesquisa contribui de maneira valiosa para a compreensão e o

tratamento da epilepsia e estabelece um precedente importante para futuras investigações neste domínio.

Palavras-chave: KNN; SVM; RF; EEG; Classificação, epilepsia.

Objetivo:
O objetivo dessa pesquisa é utilizar a aprendizagem de máquina para detecção de crises, por meio dos sinais do EEG. Para tanto é analisado os eventos pré ictal (evento que antecede a crise elíptica), ictal(evento que ocorre durante a crise), e pós ictal(evento que ocorre após a crise elíptica), para determinar quando ocorreram as crises. Assim, foram usados os métodos de classificação *K-Nearest Neighbors*(KNN), *Support Vector Machine*(SVM), *Decision Trees(*DTs), *Random Forest*(RF) e *Logistic regression*(LR), onde é possível comparar cada método de classificação de acordo com sua interpretação. Essa pesquisa foi utilizada no banco de dados CHB-MIT Scalp EEG Database.

Introdução

A epilepsia é uma doença que é causada por fatores genéticos ou por lesão cerebral, afeta cerca de 50 milhões de pessoas de todas as idades no mundo (Organização Mundial da Saúde, 2023). Pessoas que possuem epilepsia sofrem de convulsões devido a perturbação da atividade das células nervosas do cérebro. Existem dois tipos de epilepsia: a total e a parcial. A epilepsia total afeta toda área do cérebro, enquanto a parcial fica limitada a uma região do cérebro. Essa doença pode ser tratada de acordo com o seu tipo, pacientes que possuem a epilepsia total podem levar uma vida normal com o uso de medicamentos adequados, que oferecem aos pacientes um controle total ou quase total da crise, no entanto para a epilepsia parcial os medicamentos antiepiléticos costumam ser pouco eficazes, e por isso o ideal seria realizar a cirurgia de ressecção, que consiste em ressecar o ponto focal da epilepsia (BANERJEE; FILIPPI; HAUSER, 2009). Nesses casos, a cirurgia, envolve a ressecção das estruturas mediais do lobo temporal, incluindo a amígdala, hipocampo e o córtex entorrinal, e pode envolver também a ressecção do neocórtex temporal ou a ressecção do neocórtex no restante do cérebro. Em razão dos riscos que a cirurgia ablação parcial do lobo temporal

oferece para o paciente tais como a perda de memória, defeitos no campo visual, depressão dentre outros é vital encontrar o ponto focal com precisão.

O diagnóstico da epilepsia costuma ser realizado através de análise visual do eletroencefalograma (EEG) que contém o registro da atividade cerebral. O registro EEG pode ser obtido através de sensores colocados no escalpo e por ser um aparelho de baixo custo é amplamente usado em diagnósticos clínicos realizados de forma não invasiva, além de permitir a captação e alta resolução temporal. Nos casos em que é necessária a identificação precisa da zona epileptogênica, é indicado o uso do EEG invasivo, que coleta a atividade elétrica cerebral espontânea e desse modo é possível definir se a zona epileptógena é potencialmente ressecável, além de determinar a extensão da zona epileptogênica e a distinguir das áreas com lesões e o córtex eloquente.

O EEG fornece dados muito extensos, uma vez que é feito durante horas de captura. Consequentemente, para fazer a análise desses dados é necessária uma extensa inspeção visual devido à quantidade de materiais, aos sinais que são de baixa amplitude e aos componentes aleatórios. Assim é possível visualizar os desafios para processadores de sinais, já que ele possui baixa relação sinal-ruído, não-linearidade e não-estacionariedade (FREEMAN; QUIAN-QUIROGA, 2012). Dessa forma faz-se necessário a criação de um sistema automatizado para realizar a classificação dos sinais do EEG.

A detecção automática das crises de epilepsia é importante devido a vários fatores, como a melhoria da qualidade de vida para pacientes com epilepsia, pois ela permitirá uma resposta mais rápida e eficaz do diagnóstico, e consequentemente permitirá que o tratamento seja realizado de forma mais rápida. Além de contribuir com a comunidade médica, visto que o diagnóstico da epilepsia feita por profissionais necessita de uma grande inspeção visual o que leva tempo e gastos para realizar a análise.

Os algoritmos de aprendizado de máquina utilizam padrões de dados que permitem o treinamento das máquinas para tomada de decisões. Como a aprendizagem máquina permite a classificação e a tomada de decisões quando há uma grande quantidade de dados, além de permitir melhorar a velocidade e a eficiência, mantém um alto nível de

precisão e consistência podendo adaptar a mudanças e condições. Esta pesquisa tem como objetivo propor a utilização da aprendizagem para fazer a categorização de dados de EEG detectando os sinais característicos de uma crise de epilepsia, uma vez que este método pode alcançar maior precisão do que a classificação humana.

Metodologia

Para a realização deste trabalho utilizamos a base de dados (HOEB, 2009) . Foram selecionados os dados do primeiro paciente para a coleta dos registros do EEG para realização da análise. Pré-processamento é uma fase crucial a fim de obter dados mais limpos e com menos ruído para o seu processamento. Na etapa de extração de característica é feita a extração de informações relevantes usando de medidas estatísticas, assim conseguindo reduzir a dimensionalidade desses dados. Após isso, realiza-se a seleção de características, aqui é onde serão escolhidas as características mais relevantes. A classificação utiliza algoritmos de aprendizagem de máquina como o KNN SVM, DTs, RF e LR, para treinar modelos capazes de classificar os dados.

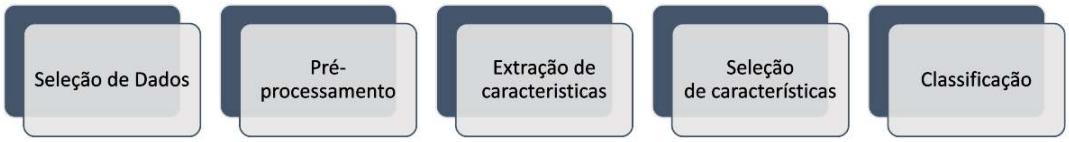

imagem1: Lista dos métodos para a realização dos resultados

Banco de dado

A base de dados usada foi a CHB-MIT Scalp EEG Database (HOEB, 2009) que contém gravações de 22 indivíduos, agrupadas em 23 casos, feitas no Hospital Infantil de Boston onde os pacientes foram monitorados após a retirada dos medicamentos anticonvulsivos. O arquivo contém o sumário com o número e nome dos canais e informações de quando se iniciou e terminou a ocorrência de crises. Essa base possui 23 canais com a frequência de 256 Hz com a duração de uma hora cada.

Pré-processamento

O pré-processamento foi feito usando a biblioteca MNE (GRAMFORT *et al*, 2013) do python para realizar a leitura e limpeza. Leitura foi utilizado o mne.io.read_raw_edf para a leitura de dados brutos. Na limpeza foram excluídos os canais T8-P8 e T7-P7, pois o primeiro contém dados duplicados e o segundo, possuir gravação oposta ao canal P7-T7. A filtragem dos dados utilizou de um filtro de 1 e 45. Por fim foi dividido os dados em épocas com duração de 15 segundos com o overlap de ⅔. O resultado desse processamento gerou uma base com 21 canais.

Extração de características

A extração de característica é usada para a identificação dos atributos mais relevantes e informativos dos dados brutos, além de reduzir a dimensionalidade desses dados. Desse modo, as bibliotecas numpy(OLIPHANT, Travis E., 2006), scipy(VIRTANEN et al., 2020) e pyeeg(BAO, 2011), a fim de realizar a extração de características, como os cálculos da média, crest, *trough*, variância, desvio padrão, *skewness*(calcula a assimetria), kurtosis, dfa, ptp, hurts, hfd, spectral, power. Assim, foi possível diminuir o conjunto de dados, melhorar a apresentação dos dados, e diminuir o tempo computacional na hora de fazer a classificação dos dados.

Seleção de características

A seleção de característica consiste em escolher um subconjunto de características relevantes dos dados e assim permitir a redução da dimensionalidade dos mesmos e um melhor desempenho computacional. Essa seleção foi feita após a extração de características com o auxílio da biblioteca sklearn (PEDREGOSA et al, 2011) que fez as importações dos seguintes módulos RFE e train_test_split. Primeiramente, foi feita a seleção de recursos com o *Feature Elimination* (RFE), onde foram selecionadas as 25 melhores características com base no modelo criado. A partir do RFE foi feita a divisão do conjunto de dados de treinamento e teste, a fim de treinar os modelos de classificação.

Classificação

A classificação é o aspecto principal da pesquisa, pois a partir dela é possível rotular o modelo em categorias com base nos padrões e características presentes nos dados os métodos de classificação usados foram *K-Nearest Neighbors*(KNN), *Support Vector*

Machine(SVM), *Decision Trees(*DTs), *Random Forest*(RF) e *Logistic Regression*(LR). A classificação foi feita a partir da extração e seleção de características e da separação dos dados do conjunto de teste e treino e avaliados usando as métricas de desempenho de acurácia, precisão, sensibilidade , *F1-score* e o suporte.

O *K-Nearest Neighbors*(KNN) é usado para classificação e regressão baseado em previsões com base no conjunto de treinamento. É um método que classifica um dado desconhecido a partir das classes mais comuns de dados mais próximos (ZHANG, 2016). Desse modo, foi considerado 7 vizinhos mais próximos para realizar as previsões, feito o treinamento e a avaliação de desempenho, onde foram calculadas a acurácia do teste, resultando na precisão de 1.0. O resultado mostrou que a precisão do conjunto de teste foi de 100%.

*Support Vector Machine(*SVM) é um algoritmo eficaz para problemas de duas classes, como no caso da pesquisa que categoriza eventos de atividade epiléptica e não epiléptica. O objetivo do SVM é encontrar um hiperplano que melhor separa os dados de diferentes classes (PANDA et al,2010). Para a aplicação dele foi importado o módulo SVC da biblioteca sklearn e feito a inicialização do modelo com o parâmetro de regularização c igual a 0.1, treinamento e a avaliação de desempenho. O algoritmo mostrou resultados semelhantes ao do KNN.

O modelo *Decision Trees(*DTs) é usado para a tarefa de classificação baseada na estrutura de uma árvore no qual cada nó representa uma decisão ou um teste em um atributo, cada ramo representa um resultado possível desse teste, cada folha apresenta uma classe ou valor de destino. Assim é feita a divisão de dados em menores, de modo que as decisões possam ser tomadas de forma sequencial. Para a aplicação foi utilizado o máximo de profundidade de 20 com o mínimo de amostras iguais a 4, após isso é realizado o treinamento do modelo, onde é mostrado o quanto a árvore se ajustou aos dados de treinamento. O resultado mostrou a acurácia do teste, resultando na precisão de 1.0.

Random Forest (RE) é usado principalmente para a tarefas de classificação e regressão, o algoritmo constrói várias árvores de decisão durante o treinamento e combina as previsões dessas árvores para obter um resultado final(WANG et al, 2019). A

implementação considerou a criação de 50 árvores, com a profundidade de no máximo 20 e o número mínimo de amostras iguais a 5, após a inicialização do modelo Random Forest foi realizado o treinamento e sua avaliação que obteve precisão da acurácia do teste de 1.0.

Logistic regression(LR) é utilizado principalmente para problemas de classificação binária, assim é possível estimar a probabilidade de um exemplo pertencer a uma das classes, utilizando funções logística para realizar a estimativa. O algoritmo foi implementado inicializando o modelo e o treinando e sua avaliação que obteve precisão da acurácia do teste de 1.0.

Resultados

Os resultados da aplicação dos métodos de classificação foram avaliados pelas métricas de desempenho de acurácia, precisão, sensibilidade , F1-score e o suporte.
Para os diferentes métodos de classificação a acurácia foi de:

- KNN: A precisão do conjunto de testes foi de 100%.
- SVM: Os resultados do SVM foram semelhantes aos do KNN, mas com parâmetros diferentes.
- Árvore de Decisão: A árvore de decisão alcançou uma precisão de 100% nos conjuntos de teste.
- Random Forest: O Random Forest obteve uma alta precisão, com 100% no conjunto de testes.
- Regressão Logística: A regressão logística alcançou uma precisão de 100% no conjunto de testes.

Estes resultados mostram que os resultados indicam um desempenho excelente na categorização das crises de epilepsia.

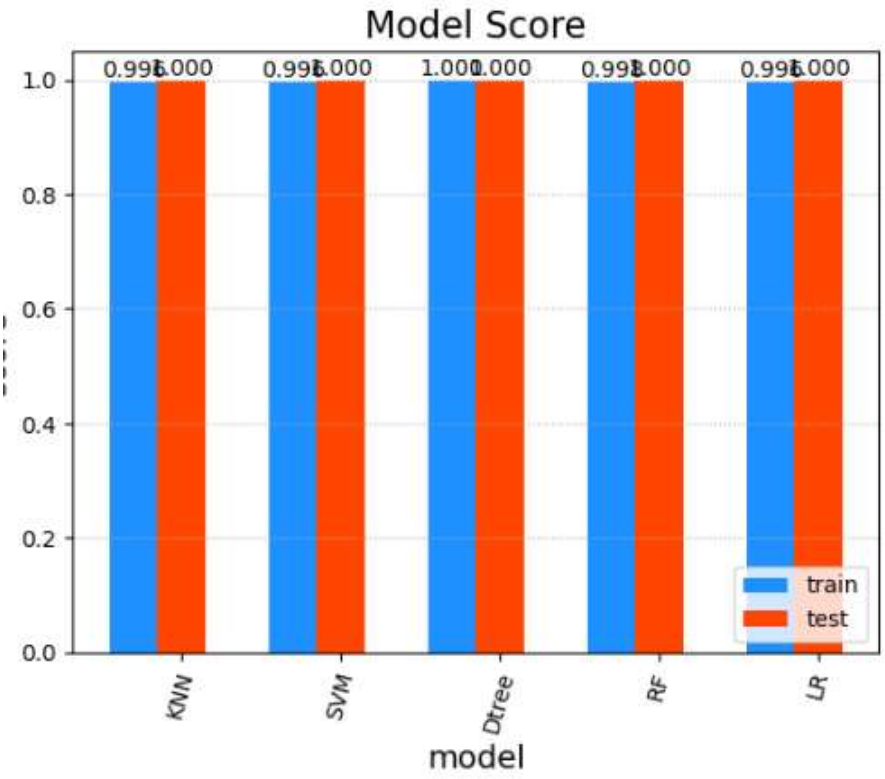

Imagem 2: Gráfico de barras que representa os escores de treinamento e teste de quatro modelos diferentes.

Foram também acrescentadas as métricas de desempenho que obtiveram o mesmo resultado para todos os modelos:

- Precisão (Precisão): A precisão para a classe 0 é de 1.00, o que significa que todas as previsões para a classe 0 foram corretas.
- Recall (Sensibilidade): A sensibilidade para a classe 0 também é de 1.00, o que significa que o modelo identificou todas as instâncias da classe 0.
- F1-score: O F1-score para a classe 0 é de 1.00, o que indica um desempenho muito bom para essa classe.
- Support: O suporte para a classe 0 é de 200, o que significa que há 200 amostras da classe 0 no conjunto de teste.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	200
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Imagem 3: Resultado da precisão, recall, f1-score e support.

Discussão

O artigo propôs a utilização de técnicas de aprendizagem de máquina, utilizando os algoritmos como KNN, SVM, Árvore de Decisão, Random Forest, a fim de automatizar a classificação dos sinais de EEG. A metodologia envolve várias etapas, como o pré-processamento, limpeza, filtragem, extração e seleção de características, todo com o intuito de realizar a redução da dimensionalidade dos dados para a realização do processamento. Todas as técnicas de classificação apresentaram resultados com altas taxas de precisão na detecção de crises de epilepsia, o que indica um desempenho excelente na categorização dos eventos.

Conclusão

Neste artigo foram apresentados os métodos de pré-processamento, extração e seleção de características, classificação e seus resultados. Desse modo, as técnicas para a classificação usadas apresentaram um bom desempenho para os dados fornecidos, com alta pontuação para os dados de treinamento e teste.

Referências

OMS. Organização Mundial da Saúde. (2023). Epilepsy. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/epilepsy>. Acesso em 01 set. 2023.

BANERJEE, P. N.; FILIPPI, D.; HAUSER, W. A. The descriptive epidemiology of epilepsy - a review. Epilepsy Research, Elsevier, v. 85, n. 1, p. 31–45, 2009.

FREEMAN, W.; QUIAN-QUIROGA, R. Imaging brain function with EEG: advanced temporal and spatial analysis of electroencephalographic signals. [S.l.]: Springer Science & Business Media, 2012.

HOEB, Ali Hossam. Application of machine learning to epileptic seizure onset detection and treatment. **Tese de Doutorado. Massachusetts Institute of Technology,** 2009.

GRAMFORT et al.MEG and EEG data analysis with MNEPython. **Front. Neurosci**., vol. 7,p. 1-13, 23 de dezembro de 2013.

OLIPHANT, Travis E. **Guide to NumPy**. Vol. 1. USA: Trelgol Publishing, 2006.

VIRTANEN et al. Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python. **Nature Methods**. Vol 17, p. 261-272. 2020.

BAO, Forrest Sheng et al. PyEEG: an open source python module for EEG/MEG feature extraction. **Computational intelligence and neuroscience**, v. 2011, 2011.

PEDREGOSA *et al.* Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, vol. 12, 2825-2830, 2011.

ZHANG, Zhongheng. Introduction to machine learning: k-nearest neighbors. **Annals of translational medicine**, v. 4, n. 11, 2016.

PANDA, R; KHOBRAGADE, P. S; JAMBHULE, P. D.; JENGTHE S. N.;PALl, P.R.; GANDHI, T. K. Classification of EEG signal using wavelet transform and support vector machine for epileptic seizure diction. **2010 International Conference on Systems in Medicine and Biology**, Kharagpur, India, pp. 405-408, 2010.

WANG, Y. CAO, J; LAI, Xi; HU, D. Epileptic State Classification for Seizure Prediction with Wavelet Packet Features and Random Forest. ***2019 Chinese Control And Decision Conference (CCDC)***. Nanchang, China. pp. 3983-3987. 2019.