

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học

**CS2205 - PHƯƠNG PHÁP LUẬN
NGHIÊN CỨU KHOA HỌC**

Lớp học

CS2205.APR2023

Giảng viên

PGS.TS. LÊ ĐÌNH DUY


Thời gian

04/2023 - 06/2023

----- *Trang này cố tình để trống* -----

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/F6ZWkMUeDTU>
- Link slides (dạng .pdf đặt trên Github của nhóm):
●
(ví dụ: <https://github.com/mynameuit/CS519.M1.KHCL/TenDeTai.pdf>)
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">● Họ và Tên: Vũ Bảo Quốc● MSSV: 220101027 	<ul style="list-style-type: none">● Lớp: CS2205.APR2023● Tự đánh giá (điểm tổng kết môn): 7.5/10● Số buổi vắng: 1● Số câu hỏi QT cá nhân: 3● Số câu hỏi QT của cả nhóm: 15● Link Github: https://github.com/baoquocvu/CS519.M1.KHCL/
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÂN LOẠI ĐỘNG KINH DỰA TRÊN DATA EEG SỬ DỤNG THUẬT TOÁN TIME SERIES FOREST

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

SEIZURE CLASSIFICATION BASED ON EEG DATA USING THE TIME SERIES FOREST ALGORITHM

TÓM TẮT (Tối đa 400 từ)

Động kinh là một chứng rối loạn não mãn tính không lây nhiễm, nghiêm trọng trên toàn cầu, có khả năng gây chấn thương và nguy hiểm tính mạng nếu không được chẩn đoán và can thiệp kịp thời. Việc chẩn đoán và phân loại động kinh thường dựa vào việc quan sát và đánh giá các bất thường của hệ thần kinh thông qua các sơ đồ tín hiệu não. Tuy nhiên, việc đánh giá điện não đồ một cách thủ công rất phức tạp, tốn nhiều thời gian mà độ nhất quán trong kết quả chẩn đoán thấp ngay cả đối với các chuyên gia. Do đó, để đạt được độ nhất quán và chính xác cao, các đặc điểm của tín hiệu điện não đồ cần được trích xuất và phân loại bằng sự hỗ trợ của các mô hình học máy. Time series forest sử dụng thuật toán Random forest khắc phục được những hạn chế trong việc phân loại dữ liệu EEG của k-Nearest Neighbor nói riêng và các mô hình phân loại khác nói chung. Vì vậy, nghiên cứu này đề xuất sử dụng time series forest cho phân loại các cơn động kinh dựa trên dữ liệu EEG của bệnh nhân. Nghiên cứu dự kiến thực hiện trích xuất đặc trưng, xây dựng, đánh giá và tối ưu mô hình phân loại time series forest trên bộ dữ liệu của Bệnh viện Nhi đồng Boston có sẵn và được lưu trữ trên trang web máy chủ Physionet (<https://physionet.org/content/chbmit/1.0.0/>). Kết quả mong đợi của nghiên cứu này là xây dựng được mô hình phân loại time series forest có độ chính xác trên 90% trong phân loại động kinh, đồng thời có khả năng phân loại đa dạng các loại động kinh và đưa ra dự đoán tin cậy về loại động kinh. Qua quá trình tối ưu hóa, mô hình sẽ vượt trội và chính xác hơn mô hình chưa được tối ưu hóa, đồng thời cải thiện tính ổn định. Mô hình phân loại sử dụng time series forest được kỳ vọng sẽ có hiệu suất vượt trội so với phương pháp k-Nearest Neighbor, mang lại độ chính xác và độ phân loại cao hơn, đảm bảo khả năng phân loại chính xác trên cả dữ liệu huấn luyện và dữ liệu mới.

GIỚI THIỆU

Động kinh là một trong những chứng rối loạn não mãn tính không lây nhiễm, nghiêm trọng trên toàn cầu [1]. Khi một cơn động kinh xảy ra, nó có thể gây ra một số chấn thương hoặc thậm chí gây nguy hiểm đến tính mạng cho bệnh nhân. Gần 1% dân số thế giới mắc bệnh động kinh, khoảng 80 trên 100.000 trường hợp động kinh mới được chẩn đoán hàng năm ở các nước phát triển, nó xuất hiện chủ yếu ở trẻ em và người lớn tuổi [1, 2]. Bệnh động kinh

không thể chữa khỏi, nhưng chứng rối loạn này có thể được kiểm soát bằng thuốc và các chiến lược khác [3, 4]. Để giảm tác động của các cơn co giật và nâng cao chất lượng sống của bệnh nhân, việc chẩn đoán và phân loại bệnh động kinh kịp thời là vô cùng quan trọng.

Động kinh là một tình trạng rối loạn chức năng của hệ thần kinh, do sự phóng điện bất thường, quá mức và đồng bộ của một nhóm tế bào thần kinh, dẫn đến các cơn động kinh không chủ ý [5, 6]. Các cơn động kinh có thể được phát hiện thông qua phân tích các tín hiệu não được tạo ra bởi mạng lưới tế bào thần kinh [7]. Mạng lưới này có cấu trúc phức tạp và kết nối mật thiết nhau tạo ra các con đường truyền tín hiệu [8]. Việc phát hiện những bất thường của hệ thần kinh thường được thực hiện bằng cách theo dõi các tín hiệu não sử dụng phương pháp điện não đồ (Electroencephalogram – EEG) [9]. Tuy nhiên, những tín hiệu này chứa nhiều, phi tuyến tính, không ổn định và mang tính phức tạp cao, gây ra một lượng dữ liệu lớn cần được xử lý [8, 9]. Do đó, đội ngũ y, bác sĩ cần được trang bị kiến thức đầy đủ để quan sát và đánh giá dữ liệu này một cách chính xác.

EEG là một kỹ thuật phổ biến được sử dụng để kiểm tra hoạt động và phân tích các bất thường của não thông qua các tình trạng co giật [10]. Tuy nhiên, việc phát hiện cơn co giật và đánh giá thời gian co giật bằng cách thủ công của các chuyên gia trong quá trình ghi điện não đồ là rất phức tạp và tốn thời gian [10, 11]. Thường mất hàng giờ đến hàng ngày để xem xét các bản ghi điện não đồ cho một bệnh nhân động kinh. Độ chính xác của chẩn đoán cũng không nhất quán và tính chủ quan cao ngay cả với các chuyên gia có kinh nghiệm lâu năm [11]. Do đó, để đạt được độ nhất quán và chính xác cao, các đặc điểm của tín hiệu điện não đồ cần được trích xuất và phân loại bằng sự hỗ trợ của các mô hình học máy.

Phân loại dữ liệu dạng time series đã đóng vai trò quan trọng trong nhiều lĩnh vực như tài chính và y học [12, 13]. Những nghiên cứu trước đây đã chỉ ra rằng, trong các thuật toán phân loại dữ liệu time series như Support vector machine, Logistic regression, Naïve Bayes, k-Nearest Neighbor (kNN) và Random forest thì kNN đem lại hiệu quả cao nhất [1, 14]. Tuy nhiên, kNN gặp hạn chế khi sử dụng khoảng cách Euclidean, vì nó chỉ áp dụng cho các chuỗi có cùng độ dài trong khi dữ liệu EEG thường có độ dài chuỗi khác nhau [15]. Hơn nữa, khoảng cách Euclidean so sánh giá trị tại từng thời điểm một cách độc lập, trong khi các giá trị chuỗi của EEG có mối tương quan với nhau [15]. Vì những lý do này, kNN không phải là một thuật toán tiềm năng trong việc phân loại EEG.

Thuật toán Time series forest (TSF) là một trong những thuật toán phân loại được đề xuất dựa trên ý tưởng random forest [16]. TSF xem xét thông tin từ các chuỗi con của chuỗi thời gian và trích xuất ba đặc điểm chính: Trung bình, độ lệch chuẩn và độ dốc từ mỗi chuỗi con [15, 17]. Việc này giúp khắc phục vấn đề về không gian xử lý lớn trong EEG [15]. TSF sử dụng random forest để học và phân loại dựa trên các đặc điểm được trích xuất từ các chuỗi con [15]. So với kNN, TSF khắc phục được hạn chế của kNN trong việc xử lý dữ liệu chuỗi thời gian với độ dài khác nhau và tận dụng được mối tương quan giữa các giá trị chuỗi thời gian, làm cho nó trở thành một lựa chọn phù hợp hơn cho việc phân loại dữ liệu EEG.

Vì vậy, nghiên cứu này đề xuất sử dụng TSF cho phân loại các cơn động kinh dựa trên dữ liệu EEG của bệnh nhân. Với:

- Input của mô hình là các tín hiệu EEG được ghi lại từ hoạt động não của bệnh nhân đang có cơn động kinh.

- Output là kết quả phân loại cơn động kinh của bệnh nhân, bao gồm các loại cơn động kinh như Generalized seizures (Absence seizures hoặc Tonic-clonic seizures) và Focal seizures (Simple focal seizures hoặc Complex focal seizures hay Secondary generalized seizures).

MỤC TIÊU

- Xây dựng một mô hình phân loại động kinh chính xác dựa trên dữ liệu EEG sử dụng thuật toán TSF.

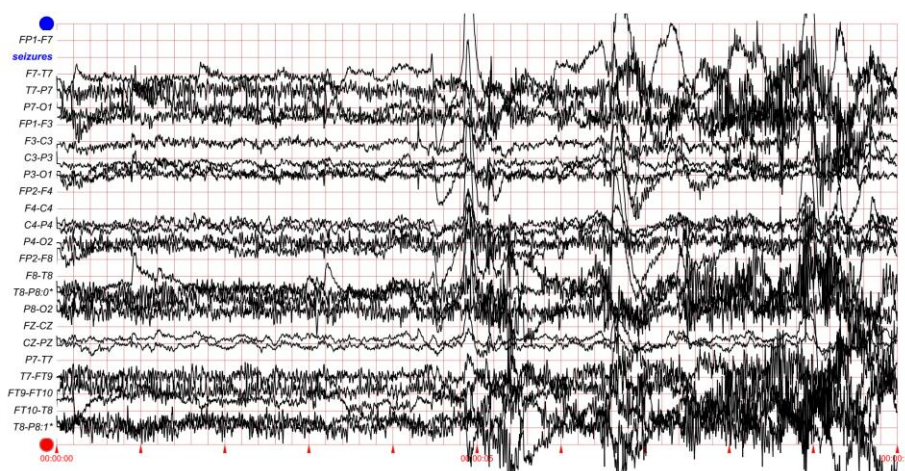
- Nghiên cứu và tối ưu hóa các tham số và siêu tham số của thuật toán TSF để đạt hiệu suất tốt nhất trong phân loại động kinh.

- Đánh giá hiệu quả của mô hình phân loại được xây dựng trên dữ liệu thực tế và so sánh với các phương pháp kNN.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Thu thập dữ liệu EEG:

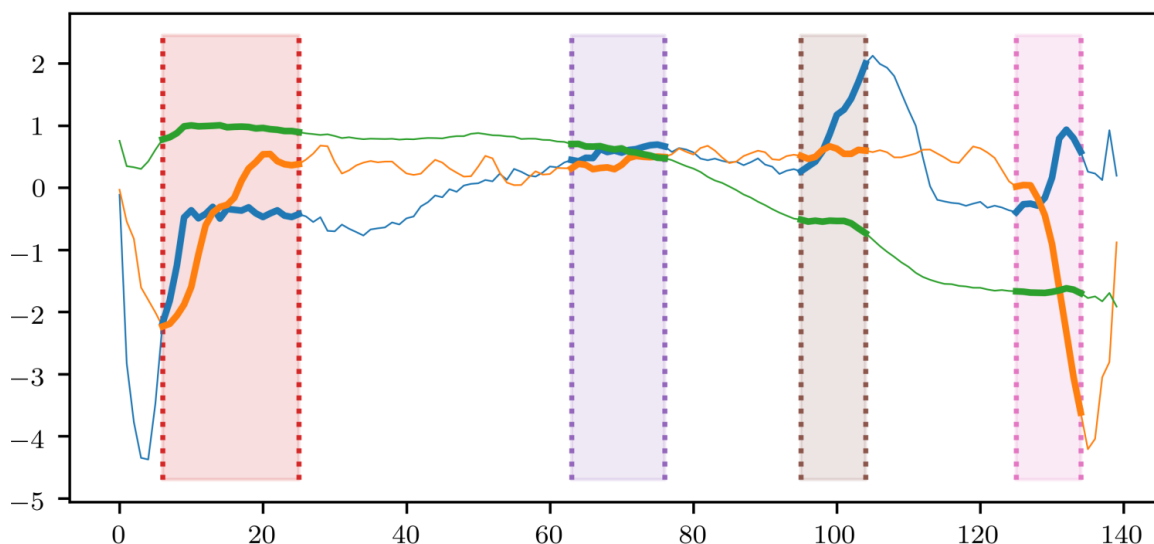
Bộ dữ liệu của Bệnh viện Nhi đồng Boston có sẵn và được lưu trữ trên trang web máy chủ Physionet (<https://physionet.org/content/chbmit/1.0.0/>) [18]. 23 bệnh nhân, 5 nam từ 3-22 tuổi, 17 nữ từ 1 đến 19 tuổi, bao gồm các tệp .edf chứa các bản ghi EEG. Mỗi trường hợp chứa từ 9 đến 42 tệp .edf và có thông tin về giới tính và độ tuổi của người tham gia. Các tệp .edf có thời lượng từ 1 giờ đến 4 giờ, và các tín hiệu được lấy mẫu với tần số 256 Hz. Bộ dữ liệu cũng cung cấp thông tin về cơn động kinh trong các tệp được chú thích và các thông tin khác như vị trí điện cực và thời gian bắt đầu/kết thúc của từng tệp.



Hình 1. Tín hiệu EEG của bệnh nhân mã chb03 có cơn động kinh được lưu trữ trên Physionet (chb03/chb03_03.edf)

2. Trích xuất đặc trưng:

Thuật toán TSF xem xét thông tin từ các chuỗi con của chuỗi thời gian. Cho độ dài tối thiểu cho các chuỗi con, là một siêu tham số, các khoảng ngẫu nhiên được tạo, với các chỉ số bắt đầu, chỉ số kết thúc và độ dài của tất cả các khoảng được tạo ngẫu nhiên. Đối với một time series nhất định và một interval nhất định, chuỗi con tương ứng là tập hợp các giá trị có thứ tự từ time series thuộc về interval. Từ mỗi dãy con, ba đặc trưng được trích xuất bao gồm: giá trị trung bình, độ lệch chuẩn và độ dốc (**Hình 2**)



Hình 2. Các intervals ngẫu nhiên được tạo ra và các chuỗi con tương ứng từ mỗi time series được trích xuất

3. Xây dựng mô hình phân loại TSF:

- Sử dụng thuật toán Time Series Forest: Xây dựng mô hình phân loại dựa trên thuật toán Time Series Forest. Áp dụng thuật toán random để tạo ra một decision Tree tối ưu hóa việc phân loại dữ liệu EEG.

- Tối ưu hóa tham số và siêu tham số: Điều chỉnh các tham số và siêu tham số của thuật toán Time Series Forest, bao gồm số lượng cây, độ sâu cây, các ngưỡng phân loại, để đạt hiệu suất phân loại tốt nhất trên dữ liệu EEG.

- Train và test mô hình: Sử dụng tập dữ liệu được chia thành tập train (80%) và tập test (20%) để huấn luyện và kiểm tra mô hình phân loại. Đánh giá độ chính xác, độ phân loại và các độ đo đánh giá khác của mô hình.

4. Đánh giá và so sánh:

- Sử dụng phương pháp cross-validation: Áp dụng phương pháp cross-validation để đánh giá hiệu suất của mô hình phân loại trên các tập dữ liệu khác nhau và đảm bảo tính khái quát của mô hình.

- So sánh kết quả của mô hình phân loại sử dụng thuật toán Time Series Forest với phương pháp kNN và đánh giá tính hiệu quả và hiệu suất của mô hình phân loại đề xuất.

5. Tối ưu hóa và cải tiến:

- Dựa trên kết quả đánh giá và so sánh, tiến hành tối ưu hóa mô hình phân loại động kinh dựa trên dữ liệu EEG. Điều chỉnh các tham số và siêu tham số của mô hình để cải thiện hiệu suất và độ chính xác của nó.

- Nghiên cứu và áp dụng các phương pháp mở rộng và cải tiến cho thuật toán Time Series Forest, như ensemble learning, feature selection, hoặc mô hình học sâu khác, để nâng cao tính chất phân loại và khả năng dự đoán.

KẾT QUẢ MONG ĐỢI

Mô hình TSF được kỳ vọng đạt độ chính xác cao trên 90% trong việc phân loại các loại động kinh dựa trên dữ liệu EEG. Ngoài ra, TSF có khả năng phân loại đa dạng các loại động kinh và đưa ra các dự đoán tin cậy về loại động kinh mà bệnh nhân đang gặp phải.

Qua quá trình tối ưu hóa, mô hình có thể đạt hiệu suất tốt nhất trong phân loại động kinh. Điều này đồng nghĩa với việc mô hình có khả năng phân loại vượt trội và đạt độ chính xác cao hơn so với mô hình chưa được tối ưu hóa. Đồng thời, tính ổn định của mô hình cũng được cải thiện nhờ việc điều chỉnh các tham số và siêu tham số.

Mô hình phân loại sử dụng TSF được dự kiến có hiệu suất vượt trội so với các phương pháp phân loại sử dụng kNN. Độ chính xác và độ phân loại mong đợi cao hơn và có sự cải thiện so với phương pháp kNN. Mô hình cũng được đánh giá có thể đảm bảo khả năng phân loại chính xác trên cả dữ liệu huấn luyện và dữ liệu mới.

TÀI LIỆU THAM KHẢO

- [1]. Milind Natu, Mrinal Bachute, Shilpa Gite, Ketan Kotecha, and Ankit Vidyarthi: Review on epileptic seizure prediction: machine learning and deep learning approaches. *Computational and Mathematical Methods in Medicine* 2022 (2022).
- [2]. Behnaz Akbarian and Abbas Erfanian: Automatic seizure detection based on nonlinear dynamical analysis of EEG signals and mutual information. *Basic Clin Neurosci.* 9(4): 227 (2018).
- [3]. Syed Muhammad Usman, Shehzad Khalid, Muhammad Haseeb Aslam: Epileptic Seizures Prediction Using Deep Learning Techniques. *IEEE Access* 8: 39998-40007 (2020)
- [4]. Syed Muhammad Usman, Shehzad Khalid, and Zafar Bashir: Epileptic seizure prediction using scalp electroencephalogram signals. *BBE.* 41(1): 211-220 (2021).
- [5]. Alison M. Pack: Epilepsy Overview and Revised Classification of Seizures and Epilepsies. *Continuum (Minneapolis, Minn.)*. 25(2) (2019).
- [6]. Robert S. Fisher: The New Classification of Seizures by the International League Against Epilepsy 2017. *Curr Neurol Neurosci Rep.* 17(6):48 (2017).
- [7]. Sai Manohar Beeraka, Abhash Kumar, Mustafa Sameer, Sanchita Ghosh, and Bharat Gupta: Accuracy enhancement of epileptic seizure detection: a deep learning approach with hardware realization of STFT. *Circuits Syst. Signal Process.* 41(1): 461–484 (2022).

- [8]. Anand Shankar, Hnin Kay Khaing, Samarendra Dandapat, and Shovan Barma: Analysis of epileptic seizures based on EEG using recurrence plot images and deep learning. *Biomed. Signal Process. Control.* 69: 102854 (2021).
- [9]. Azmi Shawkat Abdulbaqi, Muhanad Tahrir Younis, Younus Tahreer Younus, and Ahmed J. Obaid: A hybrid technique for EEG signals evaluation and classification as a step towards to neurological and cerebral disorders diagnosis. *Int. J. Nonlinear Anal. Appl.* 13(1): 773-781 (2022).
- [10]. Raafat Hammad Seroor Jadah: Basic electroencephalogram and its common clinical applications in children. *Electroencephalography-From Basic Research to Clinical Applications* (2020).
- [11]. Catalina Gómez, Pablo Arbeláez, Miguel Navarrete, Catalina Alvarado-Rojas, Michel Le Van Quyen, and Mario Valderrama: Automatic seizure detection based on imaged-EEG signals through fully convolutional networks. *Sci Rep.* (10): 21833 (2020).
- [12]. Guangyi Zhang, Vanda Davoodnia, Alireza Sepas-Moghaddam, Yaoxue Zhang, S. Ali Etemad: Classification of Hand Movements from EEG using a Deep Attention-based LSTM Network. *CoRR* abs/1908.02252 (2019)
- [13]. Sarah J. MacEachern and Nils D. Forkert: Machine learning for precision medicine. *Genome.* 64(4): 416-425 (2021).
- [14]. Zoltan Geler, Vladimir Kurbalija, Mirjana Ivanovic, Milos Radovanovic: Weighted kNN and constrained elastic distances for time-series classification. *Expert Syst. Appl.* 162: 113829 (2020)
- [15]. Johann Faouzi: Time series classification: A review of algorithms and implementations. *Machine Learning (Emerging Trends and Applications)* (2022).
- [16]. Houtao Deng, George C. Runger, Eugene Tuv, Vladimir Martynov: A time series forest for classification and feature extraction. *Inf. Sci.* 239: 142-153 (2013)
- [17]. Matthew Middlehurst, James Large, Anthony J. Bagnall: The Canonical Interval Forest (CIF) Classifier for Time Series Classification. *IEEE BigData 2020*: 188-195
- [18]. Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Cir.* 101(23): e215-e220, (2000).

----- *Trang này cố tình để trống* -----