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

Vũ Bảo Quốc

Trường Đại học Công nghệ Thông tin, ĐHQG – TP. HCM

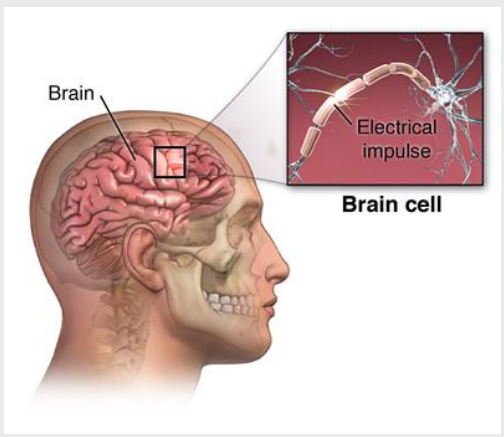
MỤC TIÊU

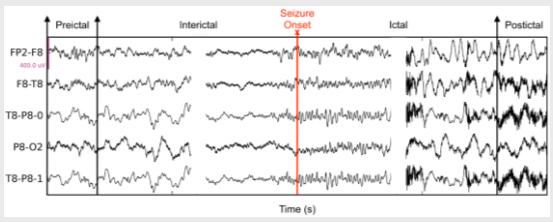
- Xây dựng một mô hình phân loại động kinh chính xác dựa trên dữ liệu EEG sử dụng thuật toán Time series forest (TSF).
- Tối ưu hóa các tham số và siêu tham số của thuật toán TSF để đạt hiệu suất tốt nhất trong phân loại động kinh.
- Đánh giá hiệu quả của mô hình phân loại được xây dựng trên dữ liệu thực tế và so sánh với các phương pháp k-Nearest Neighbor (kNN).

ĐẶT VẤN ĐỀ

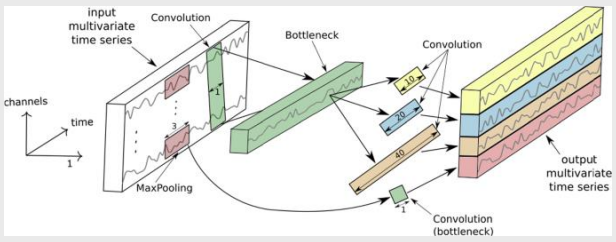
Động kinh là một chứng rối loạn não mãn tính không lây nhiễm, gây chấn thương và nguy hiểm tính mạng nếu không chẩn đoán và can thiệp kịp thời. Việc chẩn đoán và phân loại dựa vào quan sát và đánh giá tín hiệu não. Đánh giá điện não đồ thủ công phức tạp, không nhất quán kết quả. Do đó, trích xuất và phân loại tín hiệu điện não đồ bằng mô hình học máy cần thiết. TSF sử dụng Random forest khắc phục hạn chế phân loại dữ liệu EEG của kNN. Nghiên cứu này đề xuất sử dụng TSF phân loại cơn động kinh dựa trên dữ liệu EEG.

GIỚI THIỆU





Time series classifier



Classes:

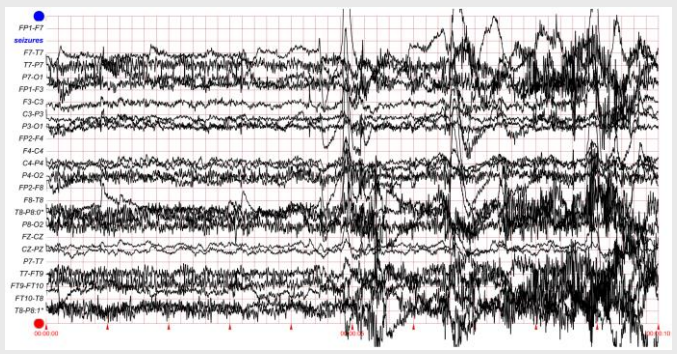
1. Generalized seizures
- Absence seizures EEG
- Tonic-clonic seizures EEG

2. Focal seizures
- Simple focal seizures EEG
- Complex focal seizures EEG
- Secondary generalized seizures EEG

MÔ TẢ

1. Thu thập dữ liệu EEG

Bộ dữ liệu của Bệnh viện Nhi đồng Boston có sẵn trên trang web Physionet. Bộ dữ liệu bao gồm 23 bệnh nhân (5 nam, 17 nữ) trong độ tuổi từ 3 đến 22 tuổi. Các tệp .edf chứa các bản ghi EEG, với mỗi trường hợp chứa từ 9 đến 42 tệp .edf và thông tin về giới tính và độ tuổi. Thời lượng các tệp .edf từ 1 giờ đến 4 giờ và tần số lấy mẫu là 256 Hz

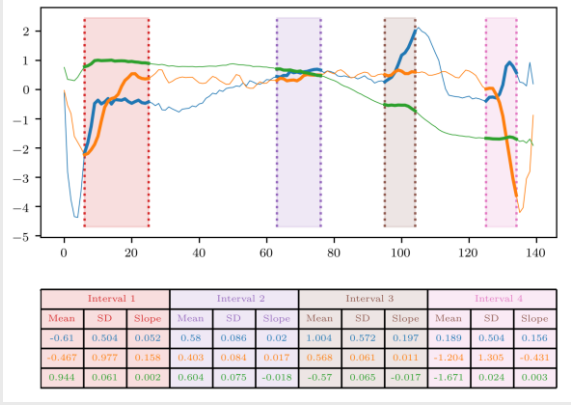


Hình 1. *Tín hiệu EEG của bệnh nhân mã chb03 có cơn động kinh được lưu trữ trên Physionet (chb03/chb03_03.edf)*

2. Trích xuất đặc trưng

TSF là thuật toán xem xét thông tin từ chuỗi con của chuỗi thời gian. Với siêu tham số là độ dài tối thiểu cho chuỗi con, khoảng ngẫu nhiên được tạo với chỉ số bắt đầu, chỉ số kết thúc và độ dài của các khoảng

Chuỗi con tương ứng với một interval là tập hợp các giá trị từ chuỗi thời gian nằm trong interval. Từ mỗi chuỗi con, ba đặc trưng được trích xuất: giá trị trung bình, độ lệch chuẩn và độ dốc (**Hình 2**).



Interval 1			Interval 2			Interval 3			Interval 4		
Mean	SD	Slope	Mean	SD	Slope	Mean	SD	Slope	Mean	SD	Slope
-0.61	0.504	0.052	0.58	0.086	0.02	1.004	0.572	0.197	0.189	0.504	0.156
-0.467	0.977	0.158	0.403	0.084	0.017	0.568	0.061	0.011	-1.204	1.305	-0.431
0.944	0.061	0.002	0.604	0.075	-0.018	-0.57	0.065	-0.017	-1.671	0.024	0.003

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 TSF

- Sử dụng thuật toán Time Series Forest, xây dựng mô hình phân loại EEG với quyết định cây ngẫu nhiên.
- Tối ưu hóa tham số và siêu tham số để đạt hiệu suất cao nhất.
- Train và test mô hình trên tập dữ liệu, đánh giá độ chính xác

4. Đánh giá và so sánh

- Sử 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 TSF 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 TSF, như ensemble learning, feature selection, hoặc mô hình học sâu khác, để nâng cao tính chất phân loại và khả năng dự đoán.

NII

Vũ Bảo Quốc – Trường Đại học Công nghệ Thông tin, ĐHQG – TP. HCM

TEL: 0348726502 Email: quocvb.17@grad.uit.edu.vn