ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ
И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

Предмет: Вовед во наука на податоци

Predicting Epileptic Seizures through EEG Signal Processing

Студенти: Давид Христов, Огнен Трајковски
Индекси: 221085, 221199

Професор: Димитар Трајанов

Датум: 20.07.2025

Содржина

1 Домен на проектот и опис на проблемот

EEG, односно **електроенцефалографија** е неврофизиолошка дијагностичка метода со која се испитува мозочната активност преку добивање на сигнали од мозокот, лекарот утврдува дали мозочната активност е нормална или можеби има одредени нарушувања.

Овие сигнали се добиваат преку поставување на електроди на површината на главата. Сигналите добиени преку EEG ја рефлектираат невронската активност и се користат во широк спектар на медицински и когнитивни истражувања и дијагнози, како што се:

- Дијагноза и следење на **епилепсија**
- Анализа на **нарушувања на спиењето**
- Проценка по **мозочен удар**
- Истражувања на **ментално здравје и когнитивни состојби**

EEG сигналите се електрични шеми генерирани од невроните на мозокот кои комуницираат едни со други. Невроните комуницираат преку електрични импулси, а кога многу неврони се активираат заедно, нивната комбинирана електрична активност може да се открие со **EEG**.

Електричната активност е претставена како брановидни линии со различни фреквенции и амплитуди што одговараат на различни состојби на мозокот.

EEG сигналите се **шумливи, нелинеарни, временски зависни и полни со артефакти**, што претставува голем предизвик при нивната обработка и анализа. Тие се со ниска амплитуда (обично во μV) кои се појавуваат како бранови со различни фреквенции и амплитуди.

1.1 Видови бранови според фреквенција

Име на бран	Фреквенција	Што претставува
Delta ()	0.5 – 4 Hz	Длабок сон, несвесна состојба
Theta ()	4 – 8 Hz	Лесен сон, длабока релаксација, медитација
Alpha ()	8 – 13 Hz	Смирена будност, затворени очи, без ментален напор
Beta ()	13 – 30 Hz	Активно размислување, концентрација, алармна состојба
Gamma ()	30 – 100+ Hz	Висока когнитивна активност, свесност, обработка на сетила

Табела 1: Видови на EEG бранови според фреквенција

Во случај на епилепсија, најважно е да се следат **spikes и sharp waves** (остри врвови), **spike-and-wave комплекси** кој се типичен шаблон кај генерализирани напади.

2 Прибирање на податоци (Data Acquisition)

Користена е јавна база на податоци како што е CHB-MIT Scalp EEG Database:

- Собрана од **Children's Hospital Boston**, од страна на MIT, се состои од **EEG** снимки од испитаници со нерешливи напади
- Снимките се групирани во 23 случаи и се собрани од 22 испитаници (5 мажи, на возраст од 3 до 22 години; и 17 жени, на возраст од 1,5 до 19 години)
- Секој случај (chb01, chb02, итн.) содржи помеѓу 9 и 42 континуирани .edf датотеки од еден испитаник, каде што има и **chb_summary** каде што кажува во која снимка, во кој период пациентот имал напад
- Секоја сесија содржи и анотации за почеток и крај на епилептичен напад
- Сигналите се снимани со 256Hz во секунда со 16-битна резолуција

2.1 Значење на сигналот

Секој сигнал е разлика во напонот (во микроволти, μV) со текот на времето, обично семплирана на 256 Hz во овој збир на податоци.

Всушност, сигналот е разликата во напонот помеѓу електродите на FP1 (фронтополарна лева страна) и F7 (фронтална лева страна). Сите абнормални скокови овде укажуваат на активност во тој регион на мозокот.

2.2 EEG канали

Секој .edf фајл се состои од **повеќе канали**. Секој канал го покажува напонот помеѓу 2 електроди (Fp1-F7, F7-T7).

EEG канали: Fp1-F7, F7-T7, T7-P7, и слично – се електродни парови од 10–20 EEG системот.

Вредности: **μV (микроволти)** – електричен потенцијал меѓу двата електроди.

2.3 Поделба на EEG канали според регионите на мозокот

- FP1-F7, F7-T7, T7-P7, P7-O1 (лева хемисфера)
- FP1-F3, F3-C3, C3-P3, P3-O1 (средно лева)
- FP2-F4, F4-C4, C4-P4, P4-O2 (средно десна)
- FP2-F8, F8-T8, T8-P8, P8-O2 (десна хемисфера)

Отприлика 23 канали, понекогаш 24-26.

2.4 Снимање на податоците

Снимање на податоците значи процесот на регистрирање и зачувување на EEG сигналите од електродите на главата на пациентот.

Се снима со 256 Hz → 256 примероци во секунда, поточно 256 мерења во секунда по канал.

- 23+ канали (електроди)
- Времетраење: 1 час по фајл (3600 × 256 921,600 редови)

2.5 Значење на карактеристиката

Во обработката на биомедицински сигнали (EEG, ECG, EMG), **карактеристика** е мерка што ја извлекуваме од сигналот за да добиеме корисна информација.

Карактеристика	**Значење**
Базна линија (Baseline)	Рамна или конзистентна линија без големи осцилации – типична за опуштена состојба на мозокот
Мали осцилации	Нормални мозочни ритмови, како што се алфа (8–12 Hz), бета (13–30 Hz), тета (4–7 Hz) и делта (0.5–4 Hz) бранови
Остри спајкови или изблици	Брзи и високи скокови во сигналот. Можат да бидат индикатор за почеток на епилептичен напад
Ритмички, повторливи бранови	Постои регуларен, синхронизиран шаблон во сигналот
Ненадејни скокови или падови	Изненадни промени во сигналот. Може да се јават поради шум, артефакти

Табела 2: Карактеристики на EEG сигнали

2.6 EDF Format

Самата база на податоци е во формат **.edf (European Data Format)** кој е стандардизиран формат за складирање на биомедицински сигнали.

Структура на .edf фајл:

1. *Заглавие (header)* - ASCII текст со фиксна должина од 256 бајти
2. *Податоци (signal data)* - бинарни податоци што складираат примероци од секој канал

3 Препроцесирање на податоци за ML модели

Чистење на сигналите од шумови и непотребни компоненти:

- **Филтрирање**: Примена на Bandpass Filter (0.5 – 40 Hz) за отстранување на DC offset и високофреквентни шумови
- **Артефакт редукција**: Отстранување на артефакти од трепкање, движење на очи, мускулна активност
- **Нормализација**: Сигналите се нормализираат на нула средна вредност и единечна стандардна девијација
- **Сегментација**: EEG сигналите се делат на движечки прозорци (sliding windows) од фиксна должина

3.1 Signal Processor

Оваа класа ќе ја користиме за обработка на самите EEG сигнали. Дефинираме:

- **sampling_rate**: фреквенцијата на семплирање на самиот сигнал (256Hz)
- Листа на фреквентни рабови (frequency band edges)

Фреквентни појаси (frequency bands):

- 1-5 Hz (- delta)
- 5-10 Hz (- theta)
- 10-15 Hz (- alpha)
- 15-20 Hz (- low beta)
- 20-25 Hz (- high beta)

4 Feature Extraction

Имаме **3 функции** за екстракција на карактеристики од EEG сигналите:

4.1 Temporal features (темпорални/временски) карактеристики

4.2 Spectral features (спектрални) карактеристики

Овој вид на карактеристики ни служи за вадење на **фреквенциски карактеристики** од EEG сигналот. Користиме **Welch-ов метод** за да добиеме **Power Spectral Density (PSD)**.

Welch-ов метод:

1. Сегментирање на сигналот

Карактеристика	**Објаснување**
mean	Просек на амплитудата на сигналот
variance	Колку сигналот варира (нестабилност)
skewness	Асиметрија на сигналот
kurtosis	Острината на пиковите во сигналот
rms	Root Mean Square – мерка на енергијата
zero_crossings	Број на премини преку нулата
peak_amp	Највисока апсолутна вредност
peak_count	Број на локални максимуми

Табела 3: Темпорални карактеристики

2. Windowing
3. Fourier трансформација
4. Пресметување на Power Spectral Density
5. Просек

4.3 Non-linear features (нелинеарни) карактеристики

- **Hjorth параметри** - активност, комплексност, мобилност
- **Fractal dimension** - мери грубоста или комплексноста
- **Hurst експонент** - мерка за долгорочна корелација

5 Вчитување на податоците и филтрирање

Ги вчитуваме податоците од EDF фајл и применуваме филтрирање на сигналот:

1. **Вчитување на податоци** - преку read_raw_edf функцијата
2. **Филтрирање на сигналот** - bandpass filter (0.25-25Hz)
3. **Процесирање на снимката** - делење на временски сегменти
4. **Пресметување на епохи** - преклопувачки временски прозорци
5. **Batch обработка за меморија**
6. **Извлекување на карактеристики**
7. **Лабелирање на seizure колоната**

6 Спојување на сите csv фајлови

Направивме скрипта за спојување на сите процесирани csv фајлови во еден **subjects.csv** фајл со:

- Сите екстрактнати карактеристики
- "seizure"колона (1 или 0)
- "subject"колона
- "start_time"колона

7 Тренирање на моделите

Скриптата **classifier.py** имплементира комплетен pipeline за класификација со 5 различни алгоритми:

- **MLP (Multi-Layer Perceptron)** - невронска мрежа
- **SVM (Support Vector Machine)** - со RBF kernel
- **Random Forest** - ensemble од дрва за одлучување
- **AdaBoost** - адаптивен boosting алгоритам
- **KNN (K-Nearest Neighbors)** - класификација врз база на најблиски соседи

8 Оптимизација

8.1 Batch обработка

Batch значи **групно процесирање на повеќе канали одеднаш**, наместо еден по еден. Користиме batch обработка бидејќи:

- Драстично ја подобрува брзината (2.3x побрзо)
- Организиран код
- Скалабилност на повеќе канали
- Можности за паралелизација

8.2 Векторизирани операции

Векторизацијата е техника каде наместо да обработуваме елементи еден по еден, ги обработуваме сите одеднаш користејќи специјализирани процесорски инструкции.

8.3 Паралелна обработка

Користиме **multiprocessing** за паралелна обработка со:

- Максимум 4 workers или колку CPU cores имаме
- Подготвени аргументи за секој фајл

9 Резултати

Model	CV Acc	CV TPR	CV FPR	Test Acc	Test TPR	Test FPR	Time
MLP	0.998	0.721	0.0002	0.996	0.441	0.0004	350.5
SVM	0.988	0.959	0.011	0.990	0.948	0.010	37928
Random Forest	0.999	0.839	0.00005	0.999	0.889	0.00005	589.3
AdaBoost	0.998	0.712	0.0005	0.998	0.730	0.0005	1397.
KNN	0.999	0.820	0.00002	0.999	0.867	0.000	175.4

Табела 4: Резултати од сите алгоритми

Користиме повеќе видови на пристапи за машинско учење:

- Невронски мрежи (MLP)
- Кернел методи (SVM)
- Ensemble методи (Random Forest)
- Boosting (AdaBoost)
- Instance-based learning (KNN)

10 Заклучок

Проектот успешно демонстрира примена на машинско учење за предвидување на епилептични напади преку обработка на EEG сигнали. Најдобри резултати се постигнати со Random Forest и KNN алгоритмите, додека SVM покажа највисока чувствителност но со значително подолго време на тренирање.