Arriu Simone 65170, Serra Sergio 65165

**Documentazione EEG data for Mental Attention State Detection**

Dataset: https://www.kaggle.com/inancigdem/eeg-data-for-mental-attention-state-detection

Il progetto EEG Data For Mental Attention State Detection consiste in una serie di 34 record, ognuno contenente informazioni relative al monitoraggio dello stato di attenzione di un certo utente.

Da questi record sono stati estratti solamente i dati effettivamente utili ai fini della rilevazione, cioè quei dati che sono stati rilevati dai 7 canali specificati nell'elaborato: F7, F3, P7, O1, O2, P8, AF4.

Per ogni canale, è stato calcolato lo spettrogramma considerando un campionamento di frequenza pari a 128 Hz e una finestra di Blackman con dimensione 1920 definita dal prodotto tra il campionamento di frequenza e $\Delta T = 15\ s$, ovvero la porzione di segnale EEG considerata. Seguendo il procedimento specificato nell'elaborato, è stato applicato un padding di 1024 punti, ognuno dei quali rappresenta un coefficiente della DFT. L' STFT è stata calcolata considerando un time step di 1 s.

Il passo successivo è stato produrre N feature vectors di dimensionalità 252 considerando una finestra di 15 s che scorre lungo lo spettrogramma calcolato sui canali di ogni record. Per fare ciò sono stati seguiti i seguenti passi:

1. **Bin:** le 513 frequenze ottenute dopo l'applicazione dello spettrogramma sono state raggruppate in bande da 0.5 Hz ciascuna, valutando la media della potenza spettrale delle frequenze raggruppate in ogni banda.
   Per calcolare la potenza spettrale media di ciascuna banda, è stato necessario applicare la regola di Simpson (metodo simps) che, presa in input la potenza spettrale definita su un intervallo specifico e la risoluzione di frequenza, approssima la media dell'area definita su questo intervallo.
2. È stato poi ristretto il range delle frequenze a quelle che andavano da 0 a 18 Hz ottenendo così solamente 36 frequenze.
3. **Smooth:** una volta raggruppate e ristretto il range delle frequenze, su di queste è stato applicato lo smooth tramite una media scorrevole di 15 s, applicando un'operazione matematica di convoluzione. Il procedimento consiste nel far scorrere una finestra lungo i dati di input e calcolarne la media tramite combinazioni lineari.

Questi passi sono stati eseguiti per ogni canale di ciascun record. Il risultato finale ha prodotto 34 file contenti N feature vector ciascuno, ognuno relativo ad una sessione di registrazione EEG. Non tutte queste registrazioni hanno la stessa durata, pertanto si può dire che in generale N non è uguale per tutti i file. Questo però non si può dire nel nostro caso che è più specifico in quanto abbiamo considerato solo i primi 30 minuti per ogni esperimento, e per cui, quindi, N sarà uguale per tutti i feature vector. In questi feature vector è presente anche la label di classe (come ultima componente del vettore) etichettata secondi i seguenti criteri:

- 1 se le feature sono state estratte nei primi 10 minuti dell'esperimento

- 2 se le feature sono state estratte tra i 10 e i 20 minuti dell'esperimento
- 3 se le feature sono state estratte tra i 20 e i 30 minuti dell'esperimento

Non sono stati considerati i dati oltre i 30 minuti in quanto avremmo avuto uno sbilanciamento tra le classi.

Durante il calcolo dei feature vector è stato necessario standardizzarne i valori in maniera tale che non ci fossero dei dati che, essendo troppo grandi o troppi piccoli, avrebbero influenzato le fasi successive di calcolo. La standardizzazione è stata eseguita calcolando la 0-mean e la unit variance.

L'ultima fase è stata l'applicazione dei classificatori, ovvero l'SVM, il Random Forest, la Bayesian Network, il kNN, e il Logistic Regression. Per quanto riguarda la suddivisione dei dati e delle label di classe negli insiemi di training e di test, è stata utilizzata la Cross Validation che permette di eliminare il rischio di overfitting nell'insieme di training.

Abbiamo optato per due differenti Cross Validation:

- La prima è stata implementata seguendo l'elaborato. È stato settato k = 5, così da suddividere il training set in 5 parti di uguale dimensione e considerando solo 1/k come test set e le restanti k-1/k come training set. Il procedimento è stato ripetuto k volte così da selezionare ogni volta un sub-validation set differente (metodo kfold.split). Per selezionare i feature vector da usare negli insiemi di training e test è stato utilizzato un approccio random (parametro shuffle in kfold).

  Per ogni fold sono stati applicati i cinque classificatori, utilizzando per l'SVM un kernel lineare, per il Random Forest 100 alberi decisionali, per il kNN 5 neighbors, e lasciando i valori di default per i restanti due classificatori.

Successivamente, per ogni classificatore sono state calcolate le seguenti metriche riportate in tabella:

| | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| **SVM** | **69.83%** | **71.69%** | **69.9%** | **69.45%** |
| **Random Forest** | **92.62%** | **92.68%** | **92.65%** | **92.6%** |
| **Bayesian Network** | **61.19%** | **65.18%** | **61.29%** | **59.06%** |
| **kNN** | **82.58%** | **82.86%** | **82.63%** | **82.49%** |
| **Logistic Regression** | **70.06%** | **70.41%** | **70.18%** | **69.75%** |

Le Confusion Matrix sono state ottenute sommando le Confusion Matrix prodotte da ogni fold per ogni record considerato, ottenendo così una Confusion Matrix complessiva per ogni classificatore.

**Confusion Matrix SVM**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 14765 | 3444 | 1715 |
| 2 | 2502 | 12938 | 4960 |
| 3 | 1369 | 4273 | 14586 |

**Confusion Matrix Random Forest**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 19010 | 691 | 223 |
| 2 | 744 | 18280 | 1376 |
| 3 | 225 | 1210 | 18793 |

**Confusion Matrix Bayesian Network**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 11492 | 5975 | 2457 |
| 2 | 2628 | 11462 | 6310 |
| 3 | 1466 | 4626 | 14136 |

**Confusion Matrix kNN**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 17474 | 1735 | 715 |
| 2 | 1571 | 16052 | 2777 |
| 3 | 781 | 2967 | 16480 |

**Confusion Matrix Logistic Regression**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 15410 | 2872 | 1642 |
| 2 | 3111 | 12015 | 5274 |
| 3 | 1376 | 3845 | 15007 |

I risultati ottenuti evidenziano che il Random Forest e il kNN ottengono dei risultati migliori rispetto ai restanti classificatori. I risultati sembrano buoni, ma questo potrebbe essere dato dal fatto che, usando una finestra scorrevole con un time step di 1 s due feature vector contigui siano molto simili tra di loro. Ci potrebbe essere quindi il caso che, tramite lo split random della Cross Validation, questi due feature vector molto simili vadano a finire uno nell'insieme di training e uno nell'insieme di test portando a una classificazione ovviamente corretta.

- La seconda Cross Validation è stata implementata per ovviare al problema appena descritto. È stato eliminato il parametro random e sono state considerate una Cross Validation con 34 fold (il totale dei file) in maniera tale che ad ogni fold un solo record facesse da test e i restanti 33 da training.

| | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| **SVM** | **37.71%** | **40.64%** | **37.73%** | **35.43%** |
| **Random Forest** | **42.7%** | **43.42%** | **42.75%** | **40.89%** |

| | | | | |
|---|---|---|---|---|
| **Bayesian Network** | **37.49%** | **40.64%** | **37.32%** | **31.69%** |
| **kNN** | **37.81%** | **37.98%** | **37.83%** | **37.23%** |
| **Logistic Regression** | **38.43%** | **37.85%** | **38.5%** | **35.56%** |

**Confusion Matrix SVM**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 6515 | 5860 | 7549 |
| 2 | 2466 | 7649 | 10285 |
| 3 | 2259 | 9299 | 8670 |

**Confusion Matrix Random Forest**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 9768 | 4989 | 5167 |
| 2 | 5313 | 6141 | 8946 |
| 3 | 3571 | 6688 | 9969 |

**Confusion Matrix Bayesian Network**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 3256 | 5703 | 10965 |
| 2 | 1465 | 4718 | 14217 |
| 3 | 1371 | 4135 | 14722 |

**Confusion Matrix kNN**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 7712 | 6558 | 5654 |
| 2 | 5963 | 7522 | 6915 |
| 3 | 5228 | 7329 | 7671 |

**Confusion Matrix Logistic Regression**

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 9047 | 5514 | 5363 |
| 2 | 5338 | 7838 | 7224 |
| 3 | 4885 | 8950 | 6393 |

Come previsto, i risultati sono inferiori rispetto alla Cross Validation iniziale. Il Random Forest si conferma il classificatore migliore confronto agli altri. Inoltre, l'SVM è stato il classificatore che ha impiegato più tempo a classificare i dati.

Per dare la possibilità all'utente di confrontare i risultati delle due Cross Validation, è stato prevista la possibilità di far scegliere all'utente quale delle due Cross Validation eseguire tramite un input iniziale.

Infine è stata prevista una fase di post-processing sulle predizione per cercare di migliore le metriche sulla 34 fold. Sono stati eseguiti i seguenti passi:

1. Generazione dei 34 file con due classi: la classe 2 (unfocused) con valori tra 10 e 20 minuti non è stata considerata perché ritenuta confusionaria. È stata quindi considerata la classe focus e drowsing rispettivamente con valore 1 e 3.
2. Sono stati applicati i vari classificatori sui 34 file considerando una Cross Validation leave-one-out. È stato previsto un ulteriore file in cui sono state salvate le predizioni calcolate da ogni classificatore.
3. Una volta generato il file è stato effettivamente applicato il post-proccessing sulle predizioni utilizzando una finestra scorrevole di 1 minuto senza sovrapposizioni. Dal momento che il processing dei dati genera 34 predizioni (date dalle 34 fold) per ogni classificatore, questa finestra scorrevole è stata applicata prendendo per ogni fold un intervallo di 1 minuto e andando poi a vedere quale fosse la classe prevalente.
4. Una volta che la classe prevalente è stata identificata, le predizioni di quel minuto sono state classificate con quella classe.
5. Sono state nuovamente calcolate le metriche (Accuracy, Precision, Recall, F1 Score) con le nuove predizioni.

Di seguito i risultati senza e con post-processing:

- Senza post-processing

| | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| SVM | 58.1% | 59.48% | 58.22% | 56.27% |
| Random Forest | 66.14% | 67.89% | 66.1% | 64.51% |
| Bayesian Network | 54.72% | 59.99% | 54.88% | 45.64% |
| kNN | 57.86% | 58.69% | 57.91% | 56.7% |
| Logistic Regression | 57.95% | 59.12% | 58.07% | 56.2% |

Confusion Matrix SVM

| | 1 | 2 |
|---|---|---|
| 1 | 9981 | 9943 |
| 2 | 6643 | 13075 |

Confusion Matrix Random Forest

| | 1 | 2 |
|---|---|---|
| 1 | 11827 | 8097 |
| 2 | 5310 | 14408 |

Confusion Matrix Bayesian Network

| | 1 | 2 |
|---|---|---|
| 1 | 3508 | 16416 |
| 2 | 1538 | 18180 |

Confusion Matrix kNN

| | 1 | 2 |
|---|---|---|
| 1 | 9619 | 10305 |
| 2 | 6389 | 13329 |

Confusion Matrix Logistic Regression

| | 1 | 2 |
|---|---|---|
| 1 | 9785 | 10139 |
| 2 | 6508 | 13210 |

- Con post-processing

| | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| SVM | 70.43% | 81.43% | 70.43% | 67.61% |
| Random Forest | 85.81% | 88.95% | 85.81% | 85.52% |
| Bayesian Network | 49.91% | 24.91% | 49.91% | 33.24% |
| kNN | 65.30% | 79.53% | 65.30% | 60.58% |
| Logistic Regression | 70.43% | 81.43% | 70.43% | 67.61% |

Confusion Matrix SVM

| | 1 | 2 |
|---|---|---|
| 1 | 240 | 346 |
| 2 | 0 | 584 |

Confusion Matrix Random Forest

| | 1 | 2 |
|---|---|---|
| 1 | 420 | 166 |
| 2 | 0 | 584 |

Confusion Matrix Bayesian Network

| | 1 | 2 |
|---|---|---|
| 1 | 0 | 586 |
| 2 | 0 | 584 |

Confusion Matrix kNN

| | 1 | 2 |
|---|---|---|
| 1 | 180 | 406 |
| 2 | 0 | 584 |

Confusion Matrix Logistic Regression

| | 1 | 2 |
|---|---|---|
| 1 | 240 | 346 |
| 2 | 0 | 584 |

Si può notare come, dopo aver effettuato le operazioni di post-processing i valori relativi alle metriche siano aumentati significativamente per ogni classificatore tranne che per il Bayesian Network che ha prodotto dei risultati leggermente inferiori.

**Personal EEG Concentration Task**

Dataset: https://www.kaggle.com/dqmonn/personal-eeg-tasks

Una volta completate le sperimentazioni sul dataset precedente, si è cercato di replicare il procedimento utilizzato prendendo in considerazione un dataset simile, ma con qualche differenza:

- I dati includono canali Alpha, Beta, Gamma, Delta e Theta provenienti da 4 elettrodi differenti
- Ad ogni record è stato associata una label di "concentrazione" con valori da 0 a 1, dove i valori bassi indicano una bassa concentrazione e valori altri un'alta concentrazione
- Ogni file compreso nel dataset aveva solamente dati di un particolare stato (stessa label di concentrazione)

Viste le differenze riscontrate rispetto al dataset precedente, è stato necessario cambiare alcune fasi dell'approccio utilizzato nel dataset precedente. In particolare, il problema non è stato più affrontato come un problema di classificazione, ma come un problema di regressione.

1. Nel calcolo dello spettrogramma non è stata più utilizzata una finestra da 1920 punti, ma una finestra ridotta con soli 32 punti. Questo perché i file che venivano esaminati singolarmente molto spesso non avevano abbastanza record e duravano meno rispetto agli esperimenti precedenti la cui durata era pari a 30 minuti. Invece, il campionamento di frequenza è rimasto inalterato.
2. Nella creazione dei feature vectors dove veniva utilizzata una finestra scorrevole di 15s sullo spettrogramma, è stato necessario in alcuni casi ridurre la dimensione di questa finestra sempre dovuto al fatto della limitata durata di alcuni record.
3. Nel calcolo dell'Average Spectral Power in alcuni casi non era possibile applicare il metodo Simps. In particolare, nei casi in cui non stavamo analizzando un intervallo di valori ma un valore singolo, non sarebbe stato corretto applicare una media. Abbiamo perciò lasciato il valore inalterato.
4. Dal calcolo dei feature vectors, è emerso che alcuni di essi fossero composti da un solo record. Ciò non permetteva di applicare i metodi di regressione in maniera corretta. Pertanto, si è pensato di prendere quei record e unirli in un unico feature vector. Se nel caso precedente è stata applicata una Cross-Validation con 34 fold (una fold per ogni file), in questo caso visti gli 80 file ridotti a 57 per via del problema appena descritto, abbiamo applicato una Cross-Validation con 57 fold.
5. Come detto precedentemente, i metodi di classificazione sono stati sostituiti da quelli di regressione (Linear SVR, Random Forest Regressor, Bayesian Ridge, Kn Regressor, Linear Regression). Mentre nell'altro dataset andavamo a prendere tutti i dati che andavano dalla colonna 0 alla 251, con la 252 come classe, in questo caso abbiamo preso i dati dalla colonna 0 alla 300 con la 301 come classe per via della composizione del dataset e della finestra dello spettrogramma applicata che risulta differente confronto al precedente insieme di dati.
6. A questo punto, le metriche utilizzate precedentemente non erano più calcolabili. Root Mean Squared Error, Mean Squared Error, R2 Score sono state le nuove metriche calcolate.

I risultati ottenuti sono i seguenti:

| | Linear SVR | RF Regressor | Bayesian Ridge | Kn Regressor | Linear Regressor |
|---|---|---|---|---|---|
| **RMSE** | 30.38% | 30.37% | 28.66% | 31.17% | 35.99% |
| **MSE** | 12.2% | 14.27% | 11.09% | 17.56% | 43.28% |

| | | | | | |
|---|---|---|---|---|---|
| R2 | 0.01 | 0.07 | 0.0 | 0.06 | 0.0 |

Una volta completato l'esperimento con il Personal EEG Concentration Task, è stato ripreso il dataset **EEG data for Mental Attention State Detection** e abbiamo applicato i metodi di regressione sopra citati sia con 2 classi che con 3 classi. In particolare, sono state usate le etichette <0, 1> nel caso di due classi e <0, 0.5, 1> nel caso di tre classi. I risultati ottenuti sono i seguenti:

| 2 classi | Linear SVR | RF Regressor | Bayesian Ridge | Kn Regressor | Linear Regressor |
|---|---|---|---|---|---|
| RMSE | 60,68% | 47.32% | 53.2% | 55.02% | 56.36% |
| MSE | 39.27% | 23.27% | 29.05% | 30.78% | 34.02% |
| R2 | 0.07 | 0.21 | 0.05 | 0.07 | 0.04 |

| 3 classi | Linear SVR | RF Regressor | Bayesian Ridge | Kn Regressor | Linear Regressor |
|---|---|---|---|---|---|
| RMSE | 40,71% | 39.48% | 42.68% | 45.15% | 44.67% |
| MSE | 16.58% | 15.85% | 18.52% | 20.51% | 20.79% |
| R2 | 0.02 | 0.14 | 0.04 | 0.04 | 0.03 |

Infine, è stato effettuato un ulteriore esperimento con l'obiettivo di osservare un possibile miglioramento dei dati. Sempre con il dataset **EEG data for Mental Attention State Detection** sono stati usati i metodi di Regressione, andando però ad assegnare nella fase di training 2 classi, (0,1) e nella fase di test tutte e tre le classi (0,0.5,1). I risultati ottenuti sono i seguenti:

| | Linear SVR | RF Regressor | Bayesian Ridge | Kn Regressor | Linear Regressor |
|---|---|---|---|---|---|
| RMSE | 52.52% | 41.23% | 43.88% | 49.11% | 47.25% |
| MSE | 29.58% | 17.54% | 19.85% | 24.4% | 24.35% |
| R2 | 0.07 | 0.15 | 0.04 | 0.04 | 0.03 |

Osservando i dati si può notare che, nonostante il nuovo approccio, per quanto riguarda la tabella dei risultati con 3 classi, i dati sono peggiorati se pur non in modo significativo, mentre considerando quella con due classi, i dati migliorano.