Correct Preprocessing

Lorenzo Monaci

14 novembre 2024

Indice

1 Impatto sulle prestazioni

Inizialmente i dati vengono normalizzati tramite

```
fit_transform()
```

prima di effettuare lo split nei dataset di training e di test. Questo approccio è rischioso perché può portare ad avere dataset di train e test affetti da errori che dipendono dalla *media* e dalla *devaizione standard* del dataset *non*-preprocessato, il che può condurre a predizioni errate e a un degrado delle prestazioni.

2 Post-split preprocessing

Per ovviare a questo problema i dati di train e test vengono normalizzati *post-split*, in questo modo la normalizzazione dipenderà solamente dalla media e dalla deviazione standard *dei dati che verranno effettivamente usati per addestrare e testare* la CNN e il RFC.

3 Confronto con il precedente approccio

Segue una rappresentazione dei risultati della CNN per i due approcci, i grafici sono relativi allo stesso soggetto, i cui timestamp sono sovrapposti di 100 unità temporali.

3.1 Prima

Accuracy: 95.693779% before

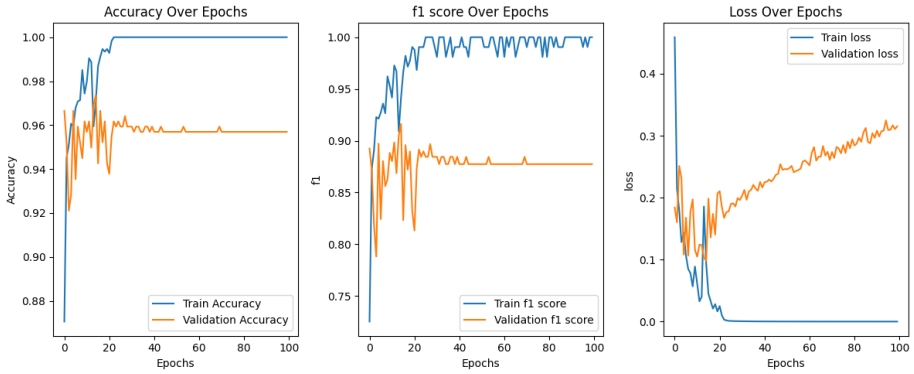

Figura 1: Normalizzazione pre-split

3.2 Dopo

Accuracy: 96.172249% after

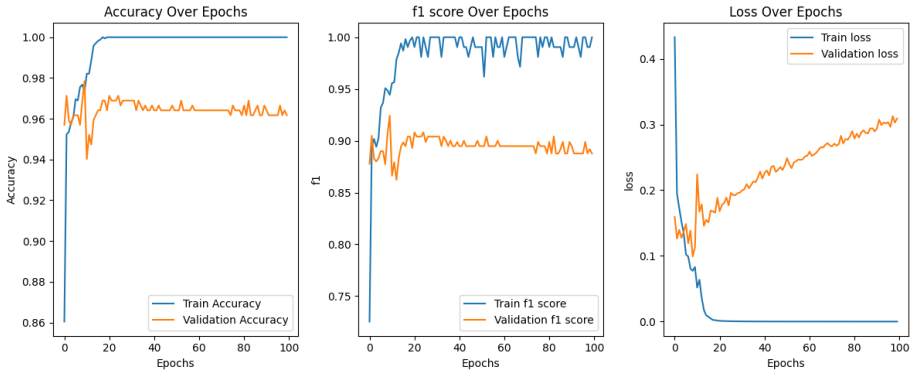

Figura 2: Normalizzazione post-split

4 Conclusioni

Dai grafici è evidente che con la normalizzazione post-split l'andamento di accuracy, f1, e loss è meno altalenante e soprattutto otteniamo una accuracy più elevata di un punto percentuale e un **f1 score** più alto di almeno 2 punti percentuali.

La modifica al nostro codice è servita per ottenere un'evoluzione della CNN più stabile e un miglioramento nelle prestazioni.