MASTER ADABI Analisi Dati per la Business Intelligence e Data Science 2023/2024

Cristina Pironi

Il deep learning per la previsione di ossigeno disciolto: un approccio LSTM

Durante il mio tirocinio presso l'INRiM, ho collaborato con un team impegnato nella costruzione del gemello digitale della Baia di Santa Teresa di Lerici, nell'ambito del progetto europeo EDTO (European Digital Twin Ocean). Il mio compito principale è stato valutare l'applicabilità di algoritmi di machine learning e deep learning per la previsione a breve termine (24 ore successive) dei livelli di ossigeno disciolto nella baia. In particolare, mi sono concentrata sulle reti neurali LSTM, poiché queste, essendo in grado di modellare relazioni complesse su diverse scale temporali, si prestano particolarmente bene all'analisi della serie temporale dell'ossigeno disciolto, caratterizzata da dinamiche non lineari, tendenze stagionali e una marcata stagionalità giornaliera. L'LSTM è in grado di elaborare sequenze multivariate e sfrutta la sua memoria a lungo e breve termine per mappare le relazioni temporali tra input e output.

Per la previsione multi-step, è stata adottata una strategia ricorsiva, validata attraverso la Walk-Forward Validation. L'approccio ricorsivo si basa su due principi chiave: l'uso del valore previsto della variabile target come input per la previsione successiva e l'utilizzo delle features esogene per il passo temporale successivo. Le features per l'addestramento includono i valori passati della variabile target e di variabili ambientali e meteorologiche, consentendo di prevedere l'andamento dell'ossigeno disciolto sulla base delle condizioni registrate negli ultimi tre giorni e delle variabili ambientali previste per le ore successive.
La Walk-Forward Validation è un metodo standard nella modellazione predittiva, in cui il modello viene addestrato su un sottoinsieme del dataset (training set), testato su un periodo successivo (test set) e poi aggiornato includendo nuove osservazioni, ripetendo il processo per diverse iterazioni. Questo consente di valutare il modello su dati futuri e testarne la capacità di generalizzazione.

Dopo un'analisi esplorativa dei dati, che ha incluso lo studio di trend, stagionalità, ciclicità e l'analisi di ACF (Autocorrelation Function), PACF (Partial Autocorrelation Function) e correlazioni tra variabili, si è passati alla fase di pre-processing. Questa fase è stata guidata dai risultati dell'analisi e dalle specifiche dell'architettura LSTM. L'analisi ha confermato la presenza di trend stagionali, una marcata ciclicità giornaliera e significative oscillazioni dovute a variabili esogene.

Le principali operazioni di pre-processing hanno incluso la normalizzazione delle variabili, per garantire un corretto apprendimento dell'LSTM, la creazione delle finestre di input di 72 ore (3 giorni) tramite il metodo della sliding window, e la suddivisione del dataset in training e test set, senza la creazione di un validation set, in quanto non è stato eseguito un tuning degli iperparametri. L'addestramento del modello è stato condotto su due anni e otto mesi di dati storici, e la stessa finestra temporale è stata mantenuta nella Walk-Forward Validation. L'LSTM è stato configurato per la previsione single-step, con iterazioni successive per ottenere una previsione multi-step.

I risultati della Walk-Forward Validation consistono in tabelle delle metriche di performance e visualizzazioni dei valori previsti per due modelli LSTM e il modello benchmark basato sulla dipendenza dal ciclo giornaliero precedente. Per ciascuna delle 5 iterazioni, i valori di MAE (Mean Absolute Error) sono stati: 2.03, 8.70, 7.63, 3.60 e 5.86 μmol/l per il modello LSTM, rispetto a

9.25, 13.43, 14.22, 7.60 e 10.83 µmol/l del benchmark. I risultati evidenziano un miglioramento significativo rispetto al benchmark, confermato anche dall'analisi visiva, che mostra la capacità dell'LSTM di seguire l'andamento della serie. Tuttavia, sono stati riscontrati episodi di sottostima e sovrastima e una struttura temporale nei residui, suggerendo una dipendenza da relazioni a breve termine.
Nella configurazione single-step, l'LSTM enfatizza la persistenza dei dati, affidandosi quasi esclusivamente al valore immediatamente precedente senza cogliere dinamiche più complesse. Nel modello multi-step, il comportamento evidenzia una forte dipendenza dalla stagionalità giornaliera, con una tendenza a replicare il pattern ciclico, piuttosto che adattarsi ad altre dinamiche e alle variazioni improvvise.

Il modello non è stato sottoposto a un tuning sistematico degli iperparametri. Tuttavia, in una sperimentazione successiva, aumentando il numero di unità LSTM per ciascuno strato, il MAE dell'LSTM è sceso a 5.04 µmol/l rispetto al valore iniziale di 8.16 µmol/l, riducendo l'errore medio di circa il 50%. Ciò dimostra che un tuning mirato potrebbe migliorare ulteriormente le performance e mitigare le problematiche di sovrastima e sottostima legate alla ripetizione della struttura stagionale, permettendo alla rete di apprendere più accuratamente le dinamiche di base.

Questo suggerisce che studi futuri, sempre nell'ambito della previsione ricorsiva, potrebbero considerare l'uso congiunto di LSTM ottimizzato e SARIMAX, per catturare le variazioni residue. Inoltre, l'integrazione di un meccanismo di attenzione potrebbe migliorare la capacità del modello di concentrarsi sui passi temporali più rilevanti, ottimizzando la cattura sia della componente stagionale che delle relazioni di breve periodo. L'attenzione consente alla rete di assegnare maggiore peso ai passi temporali più utili alla previsione, riducendo l'eccessiva dipendenza da quelli meno significativi in quel contesto. Un'architettura che include meccanismi di self-attention è il Transformer. Tuttavia, non è chiaro se le architetture LSTM e Transformer siano in grado di distinguere il contributo singolare di ciascuna feature o se, per ogni passo temporale, combinano i valori delle diverse features per estrarre un'informazione rappresentativa.
Da considerare, indipendentemente dall'architettura scelta, è l'inserimento come variabile della portata d'acqua dolce del fiume Magra, che ha un'influenza diretta e importante sul livello di salinità in baia.