Alessandro Piscopo

Titolo in Italiano:
ESPLORAZIONE DI METODI TIME SERIES PER ANOMALY DETECTION: VALUTAZIONE E ANALISI

Titolo in Inglese:
EXPLORING TIME SERIES METHODS FOR ANOMALY DETECTION: EVALUATION AND ANALYSIS

Relatore:
Tommaso Zoppi

Email del relatore:
tommaso.zoppi@unifi.it

Email del candidato:
alessandro.piscopo1@edu.unifi.it

Il lavoro di Tesi si è concentrato sull'analisi e la valutazione di metodi time series nel campo dell'anomaly detection. Si sono valutati quali siano i benefici e i miglioramenti, in termini di performance, derivanti dall'impiego di un approccio time series, rispetto ad un approccio classico all'analisi dei dati. In particolare, è stato preso come caso studio un problema di classificazione binaria e di rilevazione di anomalie. Gli algoritmi di machine learning utilizzati sono stati: Logistic Regression, Linear Discriminant Analysis, Random Forest, XGBoost. L'analisi sperimentale è stata svolta su dei dataset ottenuti da un dispositivo chiamato ARANCINO, che è il nome commerciale per una famiglia di schede IoT e embedded che risiedono sull'omonima architettura. In istanti di tempo casuali, il dispositivo è stato sottoposto a delle *error injections*, classificate come anomalie. Dai dataset ottenuti si sono addestrati dei modelli di ML in grado di classificare le osservazioni come normali o anomale (*anomaly detectors*). Inizialmente, quindi, sono stati addestrati i modelli con un approccio classico, per poi procedere all'addestramento dei modelli con un approccio time series. In particolare, abbiamo analizzato due approcci time series, uno basato sulla media mobile e l'altro sulle differenze. Entrambi gli approcci time series sono stati ottenuti modificando i dataset, andando ad aggiungere delle features che contenessero delle informazioni time series, ovvero informazioni su un certo numero di istanti di tempo precedenti. Si sono misurate per tutti i modelli le seguenti metriche: MCC, Accuracy, Error Rate. Inoltre è stata creata appositamente per la valutazione dei modelli presi in considerazione una nuova metrica chiamata Speed Score, per misurare la velocità di rilevazione delle anomalie. I modelli sono stati addestrati con finestre temporali differenti e senza rimescolamento dei dati in fase di training. I risultati ottenuti

comparando le performance dei modelli hanno dimostrato come un approccio time series, in generale, porti a notevoli incrementi nelle prestazioni dei modelli stessi, in particolare in termini di MCC e error rate. È stata anche valutata la capacità di generalizzare dei modelli rispetto ad istanze diverse del problema, risultando discretamente buona. Per concludere è stata analizzata la differenza in termini di performance tra modelli addestrati con e senza rimescolamento dei dati. Dai risultati ottenuti si è ipotizzato che i modelli addestrati senza rimescolamento dei dati dessero troppa importanza a delle features inaffidabili. L'ipotesi è stata dimostrata eliminando le feature inaffidabili e addestrando nuovamente i modelli, osservando un netto miglioramento delle performance.