INDUSTRY LAB
STUDIO DI ANOMALY DETECTION: IL CASO FASTWEB

Silvia Bordogna
736610
Università degli studi di Milano Bicocca
s.bordogna2@campus.unimib.it

Stefano Daraio
718443
Università degli studi di Milano Bicocca
s.daraio@campus.unimib.it

Andrea Armando Tinella
771399
Università degli studi di Milano Bicocca
a.tinella@campus.unimib.it

July 13, 2019

Contents

ABSTRACT

Il progetto si pone come obiettivo di risolvere un problema di anomaly detection in ambito TLC, relativo a un caso di studio dell'operatore Fastweb per il mercato italiano.

Dopo una parte iniziale dedicata alla comprensione del problema e a una prima analisi di tipo qualitativo-descrittiva, il problema si è spostato sulla scelta dei modelli più consoni da applicare al caso in oggetto e sulla valutazione dei risultati ottenuti.

1 Introduzione

L'anomaly detection è quell'area del machine learning deputata all'individuazione, trattamento e analisi di comportamenti anomali e di difficile spiegazione in cui può incorrere un processo che faccia largo uso di big data.

Inizialmente, questa tipologia di problemi era legata soprattutto all'ambito industriale e alla produzione su larga scala, essendo le modalità di raccolta e analisi dei dati costose e complesse da realizzare.

A partire dall'inizio degli anni 2000, il crescente interesse e sviluppo di fenomeni diversi (cloud, banda larga, AI, industria 4.0, big data) ha permesso di ampliare le applicazioni dell'anomaly detection e di rendere gli algoritmi connessi a essa sempre più raffinati, efficienti e a buon mercato, riuscendo a rispondere in modo più rapido e dettagliato a problemi che in passato avevano scarse possibilità di successo.

2 Contesto

Fastweb è un operatore italiano di telecomunicazioni che ha come servizi principali la connessione a banda larga e la telefonia mobile e fissa, fornendo i suoi servizi a clienti sia privati che enterprise.

Una delle caratteristiche che ha contraddistinto il suo successo a livello nazionale è da sempre stata la sua attenzione a fornire ai propri clienti le tecnologie più moderne e innovative, investendo una quota considerevole del suo bilancio in R&D ed essendo stata tra i primi operatori sul mercato italiano a fornire servizi quali ad esempio la fibra ottica e la connessione 5G.

Per mantenere elevati i suoi standard di affidabilità e qualità del servizio, uno dei focus aziendali è l'acquisizione, archiviazione e misurazione dello stato della propria rete in tempo reale, in modo da intervenire tempestivamente e precisamente su eventuali problemi a livello tecnico che potrebbero inficiare la fruizione del servizio da parte degli utenti.

3 Analisi esplorativa

Il dataset fornito per la soluzione di questo problema è già suddiviso in training e test, con 16M di righe ciascuno (800 MB). Dal momento che nei dati indicati ad uso di test non è presente alcuna anomalia, considerando la grandezza già significativa dei dati di training e la tematica della stagionalità tra i due dataset, si è scelto di proseguire l'analisi utilizzando solo quest'ultimo set di dati.

Si è scelto di utilizzare come strumento di analisi il linguaggio di programmazione Python, utilizzato all'interno dell'ambiente di sviluppo Google Colab, avente le stesse funzioni di un Notebook Jupyter.

Questa scelta ha permesso di avere a disposizione un potere computazionale maggiore rispetto all'utilizzo delle macchine a disposizione del team di progetto, rendendo il processo di analisi più snello ed evitando i probabili colli di bottiglia dati dalle macchine fisiche rispetto a quelle cloud di Colab.

Le sei variabili del dataset sono le seguenti:

- *TS* - l'intervallo di tempo in cui è stata effettuata l'osservazione. L'intervallo di tempo tra un'osservazione e l'altra è di 5 minuti. La variabile è stata convertita in formato datetime tramite libreria pandas, in modo da poter essere letta come variabile di tempo e non come stringa.
- *USAGE* - Il livello di utilizzo del singolo kit all'istante di tempo considerato, espresso come somma della banda di rete utilizzata da tutti gli utenti associati a un determinato kit.
- *KIT_ID* - Il codice identificativo del kit. Per kit si intende ogni apparato di proprietà di Fastweb che instrada la connessione per rifornire gli utilizzatori finali, che possono essere identificati come un gruppo di unità abitative.

- *AVG_SPEED_DW* - La velocità media in download disponibile per ogni utente connesso a un determinato kit. Esso è un valore predeterminato per ogni kit; probabilmente vengono effettuate delle medie a cadenze di tempo regolari per avere espressa direttamente la media di connessione.

- *NUM_CLI* - La percentuale del numero di clienti che utilizza un determinato kit, rispetto a quella massima possibile dal sistema. Non avendo accesso al numero dei clienti, non è possibile ricavare dai dati se i kit riforniscono un numero di clienti omogeneo o esso vari.

- *VAR_CLASS* - La variabile target, con cui si definisce se in un particolare momento un determinato kit ha avuto un'anomalia.

Successivamente, si è svolta la fase di esplorazione dei dati. Di seguito sono riportate alcune caratteristiche emerse che si considerano importanti ai fini della scelta del metodo di risoluzione del problema:

- Il dataset è sbilanciato (solo 508 osservazioni classificate come anomalia su 16M di osservazioni totali)

- Sono presenti 1.977 kit, di cui solo 3 presentano casi di anomalia.

- Il periodo di osservazione è di un solo mese, Novembre 2018. I kit hanno però periodi di osservazioni diversi. La maggior parte dei kit ha registrato 8357 osservazioni delle 8636 possibili (29 giorni sul 30). Dei 3 kit che presentano l'anomalia però, uno è stato osservato per solo 13 giorni.

- I tre periodi di anomalia osservati occorrono in diversi giorni della settimana, a diverse ore e persistono per un periodo differente su ognuno dei kit osservati.

- Il dataset è esente da presenza di valori nulli

- La velocità media del kit ha una distribuzione normale, con media pari a 94.322. Il valore assunto dai kit con presenza di anomalie è però inferiore (intorno al 37 percentile).

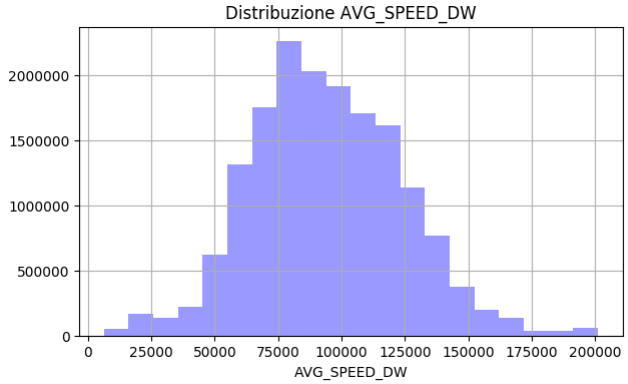

Figure 1: Distribuzione velocità media

- Più della metà dei kit ha una percentuale di utilizzo inferiore al 50%. Solo due dei kit con anomalie assumono un valore prossimo alla media, mentre un kit ha una percentuale di utilizzo del 3%.

Come già accennato si rileva un problema significativo di sbilanciamento delle classi. Inoltre, dato che i tre casi di anomalia risultano diversi tra loro, il campione appare molto piccolo ma molto variabile. Dunque si ritiene che difficilmente si possano considerare questi esempi rappresentativi di un evento generalizzabile tramite la procedura di training/test.

Sulla base di queste osservazioni si è ipotizzato di confrontare ciascun kit con se stesso, in una finestra di attività normale, rimuovendo il più possibile variazioni dovute a fattori esterni (es. stagionalità, percentuale di utilizzo del kit, media velocità, ecc.)

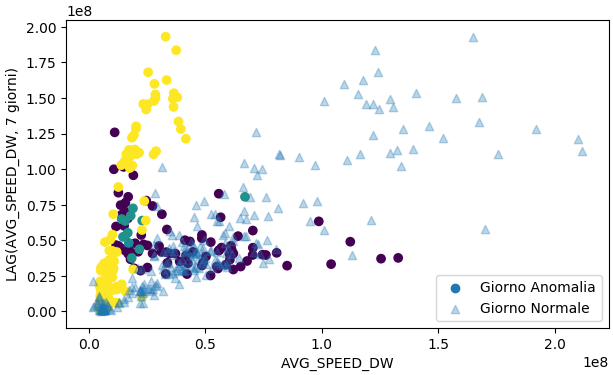

Figure 2: Scatter plot di *USAGE* - giorno X vs. giorno X - 7 giorni

Si nota dal grafico che i punti gialli, corrispondenti al periodo dell'anomalia per una delle due finestre, sono in relazione diversa con i valori della finestra di sette giorni prima rispetto invece al tipo di correlazione presente quando l'utilizzo del kit è normale.

Questa osservazione ha guidato l'approccio adottato in seguito.

4 Costruzione e implementazione dei modelli

4.1 Approccio metodologico

Uno dei metodi più largamente utilizzati per individuare efficacemente le anomalie è attraverso la costruzione e la definizione di una soglia che permetta di stabilire con sufficiente precisione in quali momenti è più probabile che si sia verificata un'anomalia. La definizione della soglia si basa solitamente sulla distribuzione di probabilità della variabile monitorata: quando i valori osservati sono troppo estremi si decide di dare un *alert*. Il rischio è che vengano dati troppi falsi allarmi. Per questo la soglia dovrebbe essere individuata anche sulla base di un analisi costi/benefici di un falso positivo rispetto al mancato riconoscimento di una vera situazione di malfunzionamento. In questo caso specifico non sono noti questi dettagli e per questo si è optato per riconoscere il maggior numero di anomalie.

Esistono vari metodi per costruire la soglia, in questo caso si è optato per costruire la soglia sulla distribuzione della distanza tra la serie durante due finestre temporali diverse. La distanza utilizzata è quella di Mahalanobis, che essendo basata sulle correlazioni tra variabili, dovrebbe intercettare la relazione osservata in fase di analisi esplorativa.

Denominatore comune dei tre metodi è l'aver ricavato la distribuzione prendendo in considerazione solo i dati "sani", cioè etichettati come non anomali, e, una volta costruita la distribuzione, applicare il metodo considerando come punti i dati con presenza di anomalie. In caso in cui i punti fossero distanti oltre una certa soglia, ottenuta sulla base della deviazione standard della distribuzione, si avrà che l'anomalia è stata effettivamente osservata e quindi il modello distingue correttamente dati anomali e non anomali.

Dopo aver definito i dati da utilizzare nell'analisi, essi sono stati dapprima normalizzati secondo la funzione MinMaxScaler. Su questa è stata calcolata la matrice delle covarianze che ci ha permesso Successivamente di calcolare la distanza di Mahalanobis, e definire gli outlier con un fattore k, definito come la media moltiplicata per 3.

4.2 Descrizione affinamenti del processo

Inizialmente si è deciso di utilizzare un approccio quanto più generico possibile, in modo da usare i risultati ottenuti come *baseline* per successivi miglioramenti.

Nei diversi passaggi nel modello sono stati trovati e valutate nuove peculiarità che permettessero un miglior grado di interpretazione delle osservazioni.

Si è deciso di riportare nel presente elaborato solo i tentativi ritenuti più significativi per la comprensione e interpretazione nell'approccio alla tematica, tralasciando quelli meno rilevanti.

4.2.1 Primo approccio

Inizialmente abbiamo considerato come training set la totalità dei KIT_ID che non rappresentavano anomalie. Consideriamo il primo giorno di rilevazioni di ogni kit_id non anomalo e lo confrontiamo con lo stesso periodo di 24 ore una settimana dopo.

Nel Test Set è stato scelto di prendere le osservazioni di tutti e 3 i KIT_ID contenenti anomalie dall'inizio dell'anomalia per 24 ore, equiparando le anomalie potenziali a quelle effettive.

Così facendo otteniamo i seguenti valori in termini di *precision* e *recall* del modello

	Modello 1
Precision	18,72%
Recall	90,35%

Vediamo che abbiamo un basso livello di *precision* in quanto essendoci un problema di *class imbalance* sono presenti numerose osservazioni classificate erroneamente come anomalie.

D'altro canto abbiamo una buon grado di *recall* poichè il modello riesce a classificare correttamente i valori effettivamente anomali in più del 90% dei casi.

E' necessario quindi riuscire ad ottenere risultati migliori sulla *precision* evitando la classificazione erronea delle osservazioni non anomale.

Sono stati individuate tre aree critiche su cui intervenire per raggiungere un miglior risultato :

- Non siamo in grado di individuare anomalie ricorrenti in periodi con cali fisiologici del livello di "USAGE" come ad esempio durante l'orario notturno.
- Tutte le anomalie per i diversi KIT_ID sono considerate della stessa natura nonostante possano esserci diverse motivazioni derivanti, inoltre non viene considerato il fattore della stagionalità caratterizzanti le serie storiche.
- Si comparano KIT_ID con caratteristiche diverse, ad esempio il numero di osservazioni ed il diverso ordine di grandezza dell' AVG_SPEED_DW e del USAGE.

4.2.2 Secondo approccio

Per ovviare alle problematiche emerse in precedenza si è creato un insieme omogeneo di 50 KIT_ID non anomali da usare per costruire la distribuzione: si sono perciò scelti kit con caratteristiche abbastanza simili al kit anomalo, utilizzando come variabili di confronto la percentuale di utilizzo, la velocità media e un numero sufficiente di osservazioni. Successivamente, si è fatta la media dell'utilizzo della banda del *subset* ottenuto precedentemente per ogni periodo di tempo, a distanza di cinque minuti.

Si è costruito il *train set* partendo da un giorno prima l'inizio dell'anomalia fino all'istante prima dell'inizio di essa.

Il test set, d'altro canto, è stato costruito partendo dall'inizio dell'anomalia fino ad un giorno dopo.

Otteniamo quindi i seguenti valori di *precision* e *recall* per ogni KIT_ID usato in fase di test

	KIT_1	KIT_2	KIT_3
Precision	23,43%	25,37%	27,67%
Recall	91,45%	90,15%	92,34%

Una volta applicata la distanza di Mahalanobis, i risultati sono stati i seguenti:

Di seguito sono riportati i risultati nell'applicazione di un modello su un KIT_ID e nel primo grafico vediamo la distribuzione delle distanze che porta all'individuazione di un valore critico di 2.54 che viene successivamente utilizzato come *threshold* nell'individuazione delle osservazioni anomale.

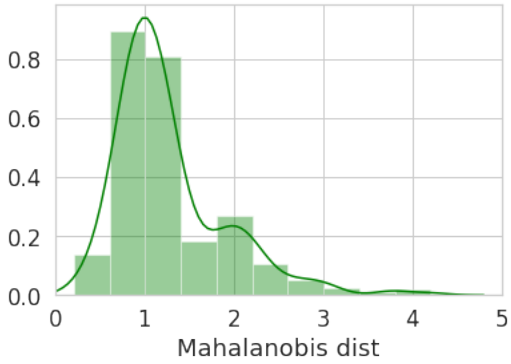

Figure 3: Distribuzione della distanza di Mahalanobis

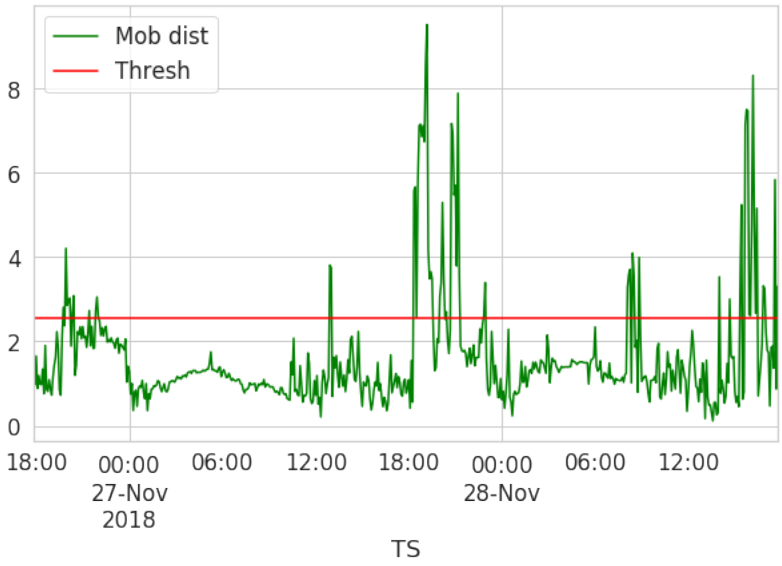

Figure 4: Serie storica della distanza di Mahalanobis con valore soglia

In generale possiamo notare un miglioramento in termini di *precision* grazie soprattutto alla maggior completezza dei KIT_ID selezionati e a dei range di valori dei parametri studiati comparabili per ogni KIT_ID anomalo successivamente valutato nel test set. In particolare per il KIT_ID 3 l'anomalia avviene durante la fascia serale/notturna.

Questo approccio rischia però di perdere la componente di variabilità del modello, andando a considerare la media di 50 KIT_ID.

La *precision* inoltre soffre dei problemi già visti nel paragrafo precedente, si potrebbe considerare come effettiva anomalia solo un periodo prolungato (ad esempio 1 ora) di periodi aventi predominanza di anomalie. Questo avrebbe il vantaggio di evitare ulteriori false anomalie ma porterebbe ad un dilatarsi del periodo di tempo per l'identificazione della problematica.

E' inoltre evidente dai dati come ci siano situazioni di anomalie (ad esempio periodi prolungati con variabile USAGE uguale a 0) e questo pensiamo possa essere imputabile a dati non correttamente valorizzati oppure ad anomalie verificatesi ma non segnalate.

4.2.3 Terzo approccio

Considerati i limiti espressi nel secondo approccio, si è ipotizzata una terza via, che non si è però validata sulla popolazione intera per via della complessità richiesta nell'automatizzare il processo.

In questa versione si confronta ogni serie con ciascun singolo kit dei 50 selezionati e non con la loro media, ottenendo così molteplici previsioni di anomalia. Questa variazione del processo è fondata sul principio sottostante le tecniche

di *ensemble learning*, coinvolgendo diversi kit di caratteristiche simili nel calcolo della distanza per ottenere una valutazione di anomalia da ciascuno e, infine, sintetizzare le diverse opinioni in un grado di certezza dell'anomalia.

In questo caso inoltre si preferisce non inserire variazioni basate sulla stagionalità, confrontando ad esempio giorni della settimana diversi, in quanto a nostro avviso si comporterebbe l'aggiunta di un livello di variabilità molto complesso.

Inoltre, avendo notato come talvolta questa metodologia indichi come anomali dei punti isolati, della durata inferiore all'ora, si intende aggiungere un ulteriore livello di conferma dell'*alert*, dato proprio dalla persistenza del segnale. Questo approccio ridurrebbe il numero di falsi positivi a scapito di un accertamento ritardato delle reali anomalie. Si ribadisce l'importanza di conoscere precisi valori di costo di ciascuna casistica per variare questi parametri e minimizzare le perdite.

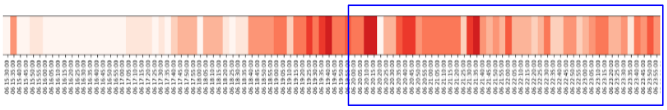

Figure 5: Alert con colorazione in base al grado di confidenza dell'anomalia

In figura 5 si riporta un esempio grafico del risultato ottenuto da questo approccio. La gradazione di rosso rispecchia il grado di certezza della previsione, dato proprio dalla percentuale di volte in cui l'osservazione è stata indicata come anomala. Rosso intenso si verifica quando il confronto con la maggior parte dei kit simili restituisce una previsione di *alert*, mentre quando il colore tende al bianco si avrà che la maggior parte delle soglie ha indicato l'intervallo di tempo come non anomalo. Il vantaggio di questa metodologia si esemplifica ad esempio nella figura riportata. All'inizio della barra si notano infatti alcune segnalazioni di colore rosso intenso che però durano a volte per pochi minuti. Si ipotizza che questi *alert* non siano da considerarsi tali in quanto il segnale non persiste. Al contrario quando l'anomalia è reale (riquadro blu in figura 5) si nota che il colore rosso intenso persiste per un lasso di tempo prolungato. Si evince dall'esempio che in questo modo sarebbe possibile filtrare una quota di falsi positivi, aumentando così la textitprecision della previsione.

5 Considerazioni finali e conclusioni

L'obiettivo iniziale del progetto si può ritenere raggiunto: partendo da una baseline metodologica e andando a raffinare il processo analitico gradualmente, si è riusciti ad individuare chiaramente se i casi segnalati come anomali lo fossero anche in pratica.

Purtroppo, non è stato possibile ottenere con la medesima efficacia una generalizzazione del modello sui casi di non anomalia per verificare se ci fossero state delle anomalie non ravvisate dal sistema. L'eccessivo sbilanciamento dei dati, infatti, ha portato o a un overfitting non desiderabile anche applicando metodi di under-sampling, o ad avere una presenza troppo massiccia di falsi positivi, non riuscendo ad analizzare meglio i casi considerati e rendendo il modello poco consigliato da utilizzare in un ambiente di produzione.

Possibili miglioramenti della qualità del dataset potrebbero riguardare l'arricchimento di informazioni sia riguardo agli utenti (integrando ad esempio informazioni di carattere socio-demografico) sia riguardo ai kit stessi (ad esempio, localizzazione dei kit, periodo passato dalla loro attivazione, eventuali differenze di modelli, registrazione di più parametri ricavati dal sistema).

Un aspetto da tenere in considerazione, e che utilizzando le informazioni in possesso del gruppo di lavoro non è stato possibile ricavare, è che probabilmente molti dati utili al miglioramento del modello vengono secretati anche alla stessa Fastweb, considerando che si appoggia alla rete TIM per quasi tutto il percordo della banda, rendendo perciò faticoso un eventuale miglioramento delle performance generali dell'infrastruttura, in quanto bisognerebbe trovare un compromesso tra due operatori aventi obiettivi e interessi differenti.

References

[1] Percorso Dataset di partenza

[2] Documentazione sul funzionamento rete Fastweb

[3] Sito ufficiale di Google Colab

[4] Sito ufficiale di Python

[5] Articolo sull'anomaly detection