POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Informatica

Riassunto di Tesi

An associative classification algorithm based on temporal association rules

Candidato

Andrea Settimo

Relatore

Prof. Paolo Garza

Introduzione

Questa tesi ha come obiettivo la progettazione, la realizzazione e la validazione di un classificatore basato su regole di associazione per la classificazione di dati che contengo tra i lori attributi il concetto di tempo. La scelta di un classificatore associativo deriva dalla capacità ereditata dalle regole di associazione, di estrapolare informazioni di alta importanza, all'interno di un dataset.

Per validare la tesi si prende in analisi un log file di una compagnia di bike sharing di Barcellona. Tale file di log contiene le informazioni sullo stato di occupazione delle stazioni del servizio di bike sharing ogni 2 minuti. In particolare, ogni dato del file di log rappresenta in un determinato istante temporale qual è la situazione di occupazione (numero di biciclette presenti e numero di baie disponibili) di una specifica stazione di biciclette.

L'obiettivo della tesi è di identificare, in base alla conoscenza pregressa dello stato delle stazioni, quale possa essere lo stato futuro di una stazione utilizzando le regole di associazione. Al fine di realizzare una soluzione scalabile, si è deciso di utilizzare un sistema per l'analisi dei Big Data. Il framework per Big Data utilizzato è Spark, per via delle sue note qualità di scalabilità.

Per il gestore del sistema di bike sharing analizzato, la predizione dello stato di una stazione in base alla situazione pregressa comporterebbe un miglioramento del servizio di distribuzione delle biciclette. Il dato predittivo può fornire informazioni tali da identificare le stazioni che in futuro saranno sprovviste di biciclette e rifornirle, rendendole nuovamente disponibili all'utenza con un notevole ritorno economico.

Metodologia e Implementazione

Il dataset considerato è contenuto all'interno di un file "csv", con la seguente struttura:

$$timestamp, StationId, used, free.$$

I campi used e free identificano, rispettivamente, il numero di slot di biciclette utilizzati e liberi. Questi due campi devono essere utilizzati in modo opportuno, poiché permettono di definire qual è lo stato della stazione in un determinato timestamp. Se free è uguale a 0, la stazione si trova nello stato Full (la stazione è completa di biciclette), invece se il campo used è uguale a 0, la stazione si trova nello stato di Empty (la stazione è sprovvista di biciclette). In tutti gli altri casi, la stazione si trova nello stato normale (Normal). Nel caso in cui si imposti un valore di soglia (numBikes_Th) che permetta di distinguere meglio il caso in cui la stazione si trovi nella situazione di "quasi pieno" o di "quasi vuoto", si possono attribuire rispettivamente gli stati AlmostFull (se free è minore o uguale a numBikes_Th) e AlmostEmpty (se used è minore o uguale a numBikes_Th).

Una volta attribuito lo stato ad ogni stazione presente nei record, si esegue una trasformazione del dataset in una nuova rappresentazione, chiamata sequenza temporale, che tiene conto del concetto di tempo e del susseguirsi di finestre temporali in cui ricadono le stazioni con i relativi stati. La sequenza temporale è descritta nel seguente modo:

$$0\_\{StationID - State, \ldots\} \rightarrow 1\_\{StationID - State, \ldots\} \rightarrow$$

$$\ldots \rightarrow (window - 1)\_\{StationID - state, \ldots\}.$$

Ogni sezione contenuta tra le frecce viene chiamata finestra temporale. La lunghezza di una sequenza è definita dal parametro di window, la frequenza di campionamento degli stati delle stazioni e i relativi stati vengono definiti dal parametro granularity. Generate le sequenze di ingresso, queste sono suddivise per ottenere il training e il test set. Il primo insieme è usato per addestrare il modello predittivo, il secondo per validare la qualità del modello di classificazione.

Il classificatore proposto si basa sull'utilizzo delle regole di associazione. Per estrarre le regole di associazione dal training set viene utilizzato l'algoritmo FP-Growth presente in Spark. Le sequenze devono essere manipolate, trasformandole in una lista di finestre. Ogni elemento della lista viene visto dall'algoritmo come un item. Una volta estratte le regole, i seguenti vincoli devono essere soddisfatti:

- Ogni delta temporale presente nell'antecedete (corpo della regola) sia minore del delta presente nel conseguente (testa della regola);
- All'interno dell'antecedente sia presente la finestra temporale con il delta pari a zero.

Durante la predizione di una stazione, all'interno dell'ultima finestra temporale di una sequenza di test, bisogna selezionare tutte le regole che soddisfano le seguenti condizioni: all'interno della testa della regola deve essere presente la stazione interessata e il corpo della regola deve coincidere con le finestre temporali di test precedenti all'ultima in cui ricade la stazione che si vuole predire. Le regole selezionate devono essere ordinate in modo decrescente in base ai seguenti campi: confidenza, lift, lunghezza e ordine lessico grafico della sequenza. L'attribuzione dello stato può avvenire in vari modi, in base alla configurazione del classificatore:

- Il parametro K definisce il numero massimo di regole da selezionare che diano il loro contributo nella classificazione:

  - Se K è uguale a 1, si prende la prima regola presente nella partizione e lo stato presente, si associa quello stato alla stazione considerata;
  - Se K è maggiore di 1, possiamo definire delle metodologie per attribuire lo stato:
    * Effettuare un voto di maggioranza: lo stato più frequente viene scelto come predizione;
    * Effettuare una selezione dello stato associato ad uno score che viene creato in modo pesato in accordo con l'importanza delle regole che hanno quella stazione in un determinato stato.

- Il parametro booleano K_restriction definisce un vincolo relativo al numero di regole selezionate:
  - Se ha valore "true" e se il numero di regole selezionate non è pari a K, allora la stazione relativa viene classificata nello stato Normal;
  - Se ha valore "false" e se il numero di regole selezionate non è pari a K, allora si effettua la classificazione in modo normale in base ad altri parametri passati.

Un'altra possibilità di classificazione può essere quella di suddividere i dati di training e di test in fasce temporali. Questa modalità permette ad un dato di test di selezionare solo quelle regole che ricadono nella stessa finestra temporale, così da poter effettuare la classificazione come definita in precedenza.

Risultati e Conclusioni

Studiando il dataset si evince che gli stati delle stazioni sono caratterizzati da un forte sbilancio, poiché gli stati di Full e Empty compaiono in minor frequenza rispetto allo stato Normal. Questo soprattutto comporta il richiamo basso per gli stati Full ed Empty, derivanti dall'output di classificazione.

Per selezionare i valori degli hyperparmaters del classificatore e quali parametri utilizzare per la trasformazione del dataset, viene utilizzata la tecnica del grid search. Dai risultati ottenuti si nota, che la classificazione con la suddivisione in fasce orarie porta a prestazioni migliori rispetto al caso di classificazione normale, raggiungendo un valore di accuratezza pari a 85.20%.

Un possibile sviluppo futuro potrebbe essere quello di utilizzare un'altra tipologia di dataset con una dimensione maggiore, per poter definire, in aggiunta, un validation set per compiere la validazione dei parametri. Un altro possibile sviluppo potrebbe essere quello di trovare una modalità di ripiego per gestire i salti temporali che vi sono all'interno del dataset e gestire i valori di used e free di una relativa stazione non coerenti al variare del tempo. Un'implementazione futura riguardante la classificazione può essere quella di introdurre il model ensemble, che permette di avere più classificatori in parallelo, attribuendo ad ognuno di essi un set di regole di associazione, create su fasce orarie diverse. La predizione finale può essere data da un voto di maggioranza o da un voto pesato, in base alla fascia oraria su cui si vuole predire lo stato della stazione.