Capitolo 1

Introduzione

1.1 Inquadramento generale

L'analisi e la comprensione di grosse quantità di dati in questi ultimi anni è diventata una pratica fondamentale per comprendere i fenomeni che ci circondano. In particolare quando questi fenomeni sono descritti da una posizione geografica e da un momento temporale preciso ci permettono di realizzare delle deduzioni che altrimenti sarebbero impensabili.

In questo ramo della ricerca si inserisce il *Data Mining*, ovvero l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di informazioni utili da grandi quantità di dati attraverso metodi automatici o semi-automatici (*machine learning*).
Nel lavoro svolto in particolare viene effettuata una *analisi delle associazioni*, una tecnica particolare di data mining utilizzata per la ricerca di connessioni di eventi con determinate caratteristiche.

1.2 Breve descrizione del lavoro

L'algoritmo preso in esame è lo *Spatio-Temporal Breath-first Miner (STB-FM)* definito da Piotr S. Maciag and Robert Bembenik. Questo particolare algoritmo permette di definire delle sequenze di tipologie di eventi connesse nello spazio e nel tempo. Viene definito un vicinato basato su un raggio spaziale e un intervallo di tempo dal quale valutare se il singolo evento è "vicino" ad altri. Questa valutazione viene fatta per tutti gli eventi di uno stesso tipo

rispetto a tutti gli eventi di un altro tipo. Da queste valutazioni si ricava un valore di *connessione* tra tipi compreso tra $[0, 1]$, più questo valore tende a 1 più i rispettivi tipi sono associati.
Questo lavoro viene fatto su diverse combinazioni di tipi, queste combinazioni vengono unite in sequenze e a ogni sequenza viene associato il valore di associazione.
es.

$[A, B, C]$ - 0.8
$[B, D]$ - 0.5

Tramite questi valori dovremmo essere in grado di capire quanto degli eventi sono associati e, in questi casi, cercare di prevenire/supportare (a seconda dei casi) l'evento di tipologia successiva.

Il caso preso in esame è quello dei crimini avvenuti a Boston, utilizzando il database fornito dal Boston Police Department (BPD) in cui sono registrati tutti i crimini avvenuti a Boston dal 2015, etichettandoli con la tipologia (di crimine), le coordinate GPS dell'evento e il momento in cui avvengono, con il giorno e l'orario.

es.

Offence Code	Date	Lat	Long
LarcFromMotVehic	01/01/2018 00:00	4.235.314.550	-7.107.763.936
ResidentialBurglary	01/01/2018 00:00	4.229.755.533	-7.105.970.910
AggravatedAssault	01/01/2018 02:23	4.235.040.583	-7.106.512.526

Tabella 1.1: esempio eventi

Esso rispetta tutti i vincoli di applicazione di questo algoritmo, vi è un gran numero di eventi etichettati per tipologia, geolocalizzati spazialmente e temporalmente, pertanto è stato scelto per l'applicazione pratica.

1.3 Scopo e prospettive

Lo scopo di questo lavoro è quindi quello di implementare il l'algoritmo *Spatio-Temporal Breath-First Miner (STBFS)* per capire le sue applicazioni

a casi concreti come quello dei crimini di Boston e analizzarne l'efficacia anche in termini di tempi di computazione.
Esso si apre a possibili sviluppi futuri anche in contesti completamente diversi rispetto a quello preso in esame, come ad esempio l'analisi dell'incidenza di epidemie.

1.4 Struttura della tesi

La tesi è strutturata nel seguente modo:

- Nel **capitolo due** si parla della base teorica su cui si basa l'algoritmo, in particolare i calcoli che si effettuano e la struttura dati utilizzata nel paper (anche possibili alternative come algoritmo apriori?)
- Nel **capitolo tre** si analizza in modo più approfondito il dataset utilizzato e le varie considerazioni fatte
- Nel **capitolo quattro** si parla dell'implementazione effettuata
- Nel **capitolo cinque** si analizzano i risultati ottenuti sia in termini di tempi di computazione che in termini di significato degli stessi
- **Conclusioni** e prospettive future

Capitolo 2

Base teorica

2.1 Paper

Come precedentemente anticipato l'algoritmo oggetto di questo lavoro è
A Novel Breadth-first Strategy Algorithm for Discovering Sequential Patterns from Spatio-temporal Data
di Piotr S. Maciag e Robert Bembenik del *Instituite of Computer Science, Warsaw University of Tecnology, Nowowiejska 15/19, 00-665, Warsaw, Poland.*
Di seguito vi è la trascritta la base teorica su cui si fonda l'algoritmo e le strutture dati utilizzate per la sua realizzazione.

2.2 Vicinato

Il problema che si considera è quello della scoperta di pattern da un certo dataset di istanze di eventi, i quali sono di una certa tipologia, definiamo quindi:

$D \rightarrow$ dataset di istanze di eventi
$F \rightarrow$ insieme di tipologie di eventi

Ogni istanza $e \in D$ ha:

- chiave di identificazione (unica)
- location spaziale (es. coordinate geografiche)
- istante temporale

- tipologia $f \in F$

La sequenza di eventi (pattern) è così definita:

$\vec{s} = f_{i_1} \rightarrow f_{i_2} \rightarrow ... \rightarrow f_{i_n}$, dove $f_{i_1}, f_{i_2}, ..., f_{i_n} \in F$

Quindi per ogni due tipologie di eventi consecutive in una sequenza $f_{i_{j-1}} \rightarrow f_{i_j}$, le istanze dell'evento $j-1$ sono connesse con con il tipo successivo spazialmente e temporalmente.

L'insieme di eventi collegati in questo modo a una determinata istanza viene definito **neighborhood** o vicinato.

Esempio

Consideriamo una situazione come quella in Fig. 2.1, dove:

$D = \{a1, a2, b1, b2, b3, b4, b5, b6, b7, b8, c1, c2, c3, c4\}$

$F = \{A, B, C\}$

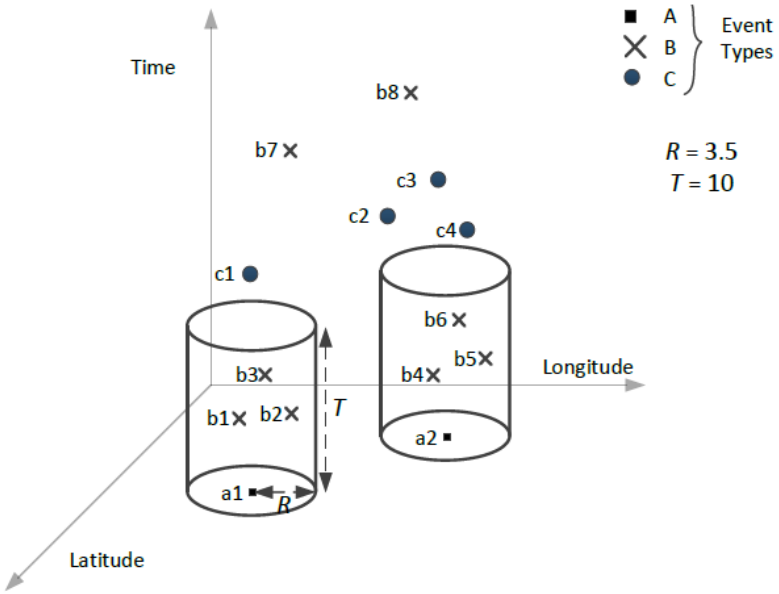

Figura 2.1: esempio di istanze con vicinato degli eventi di tipo A

Una sequenza significativa per esempio potrebbe essere $\vec{s} = A \rightarrow B \rightarrow C$.

Per valutare la connessione $A \rightarrow B$ bisogna considerare il *neighborhood* tra le loro istanze. Come si nota dalla figura è stato scelto un raggio spaziale pari a $R = 3.5$ e un intervallo temporale pari a $T = 10$ per la dimensione del vicinato.

2.3 Nozioni di base

Dopo aver inquadrato graficamente il problema ora definiamo formalmente i concetti base usati per il calcolo di tutte le sequenze pattern usati nell'algoritmo.

Spazio Neighborhood Con $V_{N(e)}$ denotiamo lo spazio di neighborhood (vicinato) dell'istanza e. Questo spazio si basa su tre dimensioni, che sono le due dimensioni spaziali - latitudine e longitudine - e la dimensione temporale. Graficamente ne risulta un cilindro, con parametri R che denota il raggio spaziale e T l'intervallo temporale.

In Figura 2.2 vengono mostrati i due neighborhood tratti dall'*esempio* della Figura 2.1: $V_{N(a1)}$ e $V_{N(a2)}$.

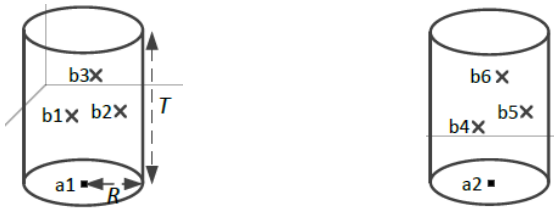

Figura 2.2: $V_{N(a1)}$ *e* $V_{N(a2)}$

Neighborhood rispetto a una tipologia di evento Data una certa istanza e, il *neighborhood* di e è definito nel modo seguente:

$$N_f(e) = \{e | p \in D(f) \\ \wedge\ distance(p.location, e.location) \leq R \\ \wedge\ (p.time - e.time) \in [0, T]\}$$

dove R e T sono i parametri dello spazio di vicinato $V_{N(e)}$ e $D(f)$ è l'insieme di istanze degli eventi di tipo f nel dataset D.
Nota si considerano solo gli eventi che si susseguono dal punto di vista temporale, in quanto è poco significativo considerare gli eventi passati dell'istanza nella ricerca di sequenze.

Riassumento con $N_f(e)$ denoto l'insieme di istanze di tipo f contenute all'interno dello spazio $V_{N(e)}$.
Nel nostro *esempio* della Figura 2.2:
$N_B(a1) = \{b1, b2, b3\}$ e $N_B(a2) = \{b4, b5, b6\}$

Set di istanze Per una sequenza di tipi di eventi $\vec{s} = \vec{s}[1] \rightarrow \vec{s}[2] \rightarrow ... \rightarrow \vec{s}[m]$ di lunghezza m, gli insiemi (set) di istanze $I(\vec{s}[1]), I(\vec{s}[2]), ..., I(\vec{s}[m])$ che sono inclusi nella sequenza $\vec{s}$ sono definiti come segue:

1. Per un tipo di evento $\vec{s}[1]$, il set di istanze $I(\vec{s}[1])$ è definito come:

   $$I(\vec{s}[1]) = D(\vec{s}[1])$$

2. Per i tipi $\vec{s}[2] \rightarrow ... \rightarrow \vec{s}[m]$ con $i = 2, 3, ..., m$, gli insiemi di istanze $I(\vec{s}[i]$ sono definiti così:

   $$I(\vec{s}[i]) = distinct(\bigcup_{e \in I(\vec{s}[i-1])} N_{\vec{s}[i]}(e))$$

In pratica per il primo tipo di evento (d'ora in poi nominato solo "tipo") che partecipa alla sequenza $\vec{s}$, il set di istanze $I(\vec{s}[1])$ corrisponde al set di istanze di tipo $\vec{s}[1]$ in D, ovvero $D(\vec{s}[1])$.
Per i tipi successivi di $\vec{s}$, i set $I(\vec{s}[i])$ sono definiti come insiemi di istanze contenute nei vicinati di istanze a partire da $I(\vec{s}[i-1])$.
Seguendo questo meccanismo si valuta tutta la sequenza e tendendo in considerazione l'insieme di istanze calcolato al passaggio precedente.

Consideriamo la sequenza $\vec{s} = A \rightarrow B$ dal dataset dell'*esempio* in Figura 2.1. In questo caso avremmo i seguenti set di istanze:
$I(\vec{s}[1]) = \{a1, a2\}$
$I(\vec{s}[2]) = \{b1, b2, b3, b4, b5, b6\}$

Participation Ratio

Participation Index

2.4 Parametri di threshold

2.5 Struttura ad albero

2.6 Alternativa - Algoritmo apriori