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 Nozioni base

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, definendo una sorta di *vicinato* comune tra istanze.

QUA FAI L'ESEMPIO DEL PAPER PER SPiEGARE COSA SI INTENDE (QUESTO È IL CAPITOLO DEL VICINATO???)

2.3 Vicinato

2.4 Parametri di threshold

2.5 Struttura ad albero

2.6 Alternativa - Algoritmo apriori