Analisi Database Crime Boston (sito kaggle)

Il database contiene 327.821 record e 17 colonne.

Colonna delle UCR offence definitions: tipologie di crimini che in america vengono divise in categorie. Diviso in 3 gradi di gravità

UCR Part I sono i crimini più pesanti come omicidio, stupro, ecc (<https://www.ucrdatatool.gov/offenses.cfm>) (<https://en.wikipedia.org/wiki/Uniform_Crime_Reports#UCR_crime_categories>)

Indicazioni pratiche:

Utilizzo la colonna del **offense\_code\_group** per identificare le tipologie, sono 67 categorie!

Part One: 9 categorie

Con le colonne **lat** e **long** ho la posizione degli eventi e con **occurred\_on\_date** ho la data e ora.

PROBLEMI RISCONTRATI:

nei dati latitudine e longitudine mancano dei record, in particolare o non c’è proprio il dato o segnala -100.000.000. TOLTI (nel database ufficiale)

Dati sono discontinui nel tempo nel senso che nel 2015 sono presenti solo i mesi di luglio e agosto. Nel 2016 e 2017 mancano dei mesi. Il database si ferma a ottobre 2018.

Nel 2016 manca: marzo, aprile, maggio, giugno, novembre, dicembre

Nel 2017 manca: febbraio, marzo

Nel 2018 manca: novembre e dicembre perché si interrompe prima

* Vi sono dei record che hanno stessa location, stesso time, diverso tipo di evento e conseguente stesso incident number, (credo che identifichino uno stesso evento ma è stato inserito più volte lo stesso elemento)

Struttura della mia tabulazione:

0.Incident\_Number, 1.Offence\_code\_group, 2.date, 3.UCR\_part, 4.lat, 5.long

Per neighborhood:

ASSUNZIONI IMPORTANTI:

parser di Location e Time basate sulle stringhe particolari in input che ho (i vari caratteri della stringa), se si cambia zona di riferimento non funzionano!! (dovrei fare un altro tipo di parser)

ho reso intere le coordinate (10 cifre senza virgola) sottratte tra di esse e valutate il raggio, (raggio bene tenerlo sulle 7 cifre di range)

tempo ho trasformato in giorni e sottratto (attraverso la classe datatime)

Creato altra funzione che calcola meglio la distanza (file “calculate\_Location”) però non so se vale la pena utilizzarlo

25/03/19

Per implementare il “Set of instances” utilizzo la libreria set di python, ogni elemento è dato dal “incident number” del mio database inserita o no in base all’algoritmo

Ho impostato come calcolare il set of instance ma non sono riuscito a concluderlo

27/03/19

Ho risolto parte del problema di “Set of instances”, adesso credo funzioni ma da problemi con i doppioni in quanto “Incident number” non è sempre univoco ha delle ambiguità. L’algoritmo ci mette sempre un tempo elevato.

Es. [Larceny, Homicide] ci mette 48 secondi circa

Possibile soluzione: creare un dataset in qualche modo tridimensionale con latitudine, longitudine e tempo come assi. In modo che valuto solo la zona di spazio interessata senza iterare tutti gli altri eventi inutilmente. Non è implementabile come matrice in quanto ci sarebbero tante celle vuote perché soprattutto per il tempo non sono così frequenti questo tipo di eventi

Altra possibile soluzione: salvo i dati in un database e faccio in modo di interrlogarlo in modo efficiente con sql (possibile collaborazione con Maurino)

27/03/19

COLLOQUIO PROF – cose da dire:

* Raccontare come ho calcolato la distanza spaziale e temporale
* Neighborhood e Set of instances
* Spiegare problem del dataset:
  + Grosse dimensioni
  + Selezione dei dati utili
  + Concentrato sul 2018
  + Mancanza di dati
  + Errori dei dati
* Spiegare problema dei tempi con calcolo set instances, chiedere suggerimenti
* Possibile collaborazione con progetto Maurino
* Dire cosa pensi di fare ora

CONSIGLI PROF:

ok maurino – progettone di complementi

capire se mantenere la struttura albero consigliata o utilizzare qualcosa di già presente che mi faciliti il problema! Non stravolgere il database ma magari concentrarmi solo su un sottoinsieme di record significativi

Cerca altre strutture dati possibili (red-black trees)

Cerca algoritmo a priori (struttura ad albero)

Association rule mining, in questo tipo di algoritmi vi è:

Ricerca dei candidati utilizzabili

Vedere se rispettano una certa regola per capire se sono significativi

(nota: distanza spaziale forse tramite maps per le strade ecc)

Cosa ho fatto:

* Preparato la funzione per il calcolo del PR e successivo calcolo PI

COSE PER DOMANI:

cercare strutture dati alternative (vedi sopra)

provare a implementare la struttura dati del paper usando le classi python e collegamenti con liste ecc

MAURINO:

* Implementa il csv in base di dati relazionale con indici per velocizzare la ricerca all’interno della stessa
* Cerca se vi sono indici in due dimensioni per mySQL
* postgres, mongDB come soluzioni per rappresentare i punti nello spazio
* vortex (per rappresentare i dati in 3 dimensioni)
* grafo delle distanze per rendere più efficiente il tutto
* pensare a grafo delle distanze viarie per tenere conto di quanto tempo si impiega da raggiungere un luogo a un altro

31/03/19

Ho creato il database mySQL di 10.000 records con incident number univoco e si chiama ‘crimedata2018’

Gli altri:

* test2018: contene parte dei dati del 2018 ma è inconsistente rispetto la dimensione delle colonne
* newtable: prova di upload di csv
* crimedataset: contiene tutti i dati del 2018 con UCR-type=One ma con i doppioni sul incident number

PER i miei test posso utilizzare il **‘crimedata2018’**

03/04/19

Sono riuscito a far interrogare a python il database e creato il file partRatioSQL.py dove ho riprogettato tutto in modo che interroghi il database mysql piuttosto che il file csv.

La velocità mi sembra migliorata ma non considerevolmente.

PER MIGLIORARE: credo che dovrei rimodellare il database in modo da fargli calcolare la distanza già prima della query così da migliorare le tempistiche (credo) per questa cosa però dovrei strutturare in modo diverso i dati, in particolare la data e la location

Potrei provare con solo 1000 record così riesco a fare test più significativi

! DA FARE: strutturare meglio i file di test e di codice

! DA FARE: fare in modo che il mysql.cursor sia passato come parametro, in questo modo lo rendo più scalabile

Performance:

* SQL [Aggravated Assault, Residential Burglary] – 3 min

INFLUENZA la performance:

Ordine all’interno della sequenza di tipi. La cardinalità del primo tipo di evento. Il numero di record. Il numero delle classi di eventi

Per ora non mi metto più a ottimizzare la base di dati. Prendo meno record ma significativi quindi:

NUOVI DATASET:

* Dataset2018\_2: preso ogni 7 un record da dataset2018, eliminato righe che hanno part = other (11.497)
* Dataset2018\_2\_one: preso i record di tipo one da dataset2018\_2 (2132) – 9 tipi di eventi
* Dataset2018Small: preso ogni 60 record uno da dataset2018 (1150)

NUOVO DATABASE:

crimedata2018small: preso i dati da dataset2018small

aggiunta della chiave primaria, indice su offence\_code\_group, occurred\_on\_date, latitude, longitude

tempi delle queys: accettabili, si rallenta se uso Larceny o aggravated assault, in quanto sono presenti in numero molto maggiore rispetto alle altre categorie

fatto la funzione che calcola il PI, lo fa in tempi ragionevoli per cui mi fermo qui con l’ottimizzazione (per ora)

08/04/19

Sono stato dal prof e mi ha detto:

* Knn: studiare come funziona per capire come ottimizzare l’algoritmo di ricerca del vicinato e set of instances
* Provare a implementare la struttura dati del paper in modo da suddividerla in levels. Per cercare di renderla il più efficiente possibile

Iniziato a implementare SPTree. L’idea è fare una list per ogni layer e mettere queste list in un vettore in modo che so a quale livello sto lavorando e posso salire e scendere dai vari livelli.

DA FARE: vai avanti con SPTree

18/04/19 (e giorni precedenti)

Ho implementato SPTree con una logica diversa rispetto alla precedente, ogni nodo ha:

* Parent1 = punta al nodo parent1
* Parent2 = punta al nodo parent2
* Value = valore del carattere che sto considerando
* Children = lista di puntatori a nodi

NB: questi puntatori (gestiti da python) puntano direttamente agli oggetti nodo originali, per cui se questi vengono modificati la modifica si vede anche attraverso l’albero!

Ho pensato a questa soluzione in quanto mi sembra la più semplice. Ora dell’albero ho bisogno solo del puntatore alla root poi posso navigarlo scendendo attraverso la lista children!!

Caso base (root e prima riga dell’albero):

root: Node(None, None, None, Lista primi figli che sono tutti i tipi di eventi)

tipi: Node(root, root, value, children)

ho dovuto settare come second parent la root per far funzionare ricorsivamente i second parent sotto. In quanto per il second parent la sua assegnazione è fatta partendo dal nodo da cui dovrà nascere il figlio (vedi dopo)

Inserimento:

PREMESSA: io inserisco le sequenze dalle più brevi alle più lunghe quindi parto da quelle lunghe 2 (quelle iniziali le inserisco nell’inizializzazione dell’albero) poi quelle da 3 e via di seguito.

Quindi se ho A🡪B e A🡪C non posso inserire A🡪B🡪C in quanto B🡪C non esiste!!!!

Assumo quindi che della sequenza che devo inserire in sostanza devo aggiungere solo l’ultimo carattere, il resto devo averlo già inserito, quindi:

itero sulla sequenza fino al penultimo carattere

itero sui figli del nodo corrente (partendo dalla root)

aggiungo il nuovo nodo con (parent1: nodo corrente, parent2: searchParent2, value, children:None)

searchParent2:

parto dal nodoCurr.parent2

cerco il nodo che ha value == nuovo value e prendo questo nodo come second parent

funzione search Sequenza:

scendo nell’albero attraverso i children seguendo la sequenza in input. Return il nodo finale di arrivo

24/04/19

Oggi mi sono concentrato sulla parte di algoritmo della Generate Candidates

Ho definito la funzione che la calcola. Ho dovuto aggiungere una nuova funzionalità al SPTree.

Ho aggiunto un attributo chiamato **candidates**. **Candidates** è una lista che contiene tutte le sequenze presenti nell’albero, in questo modo mi è più facile ritrovarle senza dover passare tutto l’albero. Nel caso in cui io elimini una parte dell’albero triggero un metodo refreshCandidates che riaggiorna la lista in modo da eliminare le sequenze che ho tagliato fuori (grazie a searchSeq)

Inoltre ho migliorato un po’ la stampa e aggiunto diversi test! (testPaper, TestDataset su tutti)

Cose da fare!

* Capire come inserire in automatico le sequenze di lunghezza superiore [vedi algo 6]
* Capire se calcolarsi il neighborhood di pari passo come nell’algoritmo del paper (quindi salvarsi il setInstances in ogni nodo oppure boh)!!
* Sistemare meglio il codice in modo che sia correlato al paper