Sviluppo di procedura per valutare interazione tra eventi e applicazione a dataset multidisciplinari

Andrea Carlo Zecchinelli, 25 Aprile 2021

I. INTRODUZIONE

In questo lavoro propongo una procedura volta a stimare se gli eventi di un processo puntuale siano indipendenti tra loro o se invece tendano ad interagire, valutando nel secondo caso se l'interazione sia di tipo repulsiva o attrattiva.

Dopo la fase di preprocessamento dei dataset, che consiste nell'ordinamento crescente e normalizzazione dei dati nell'intervallo $[0, 1]$, il metodo si compone di due step:

1. Stimare se gli eventi siano compatibili con un processo indipendente di Poisson, caratterizzato da una distribuzione delle distanze tra gli eventi Δx_i esponenziale decrescente, sulla base del *p-value* ottenuto effettuando un test di bontà del fit di Kolmogorov-Smirnov, con valore threshold pari a $p = 0.05$

2. Valutare se l'interazione tra gli eventi sia repulsiva o attrattiva sulla base del valore della statistica di Hopkins ottenuta per il dataset indagato

II. PROCEDURA

Al fine di validare e calibrare la procedura ho prima di tutto costruito un modello nullo, ovvero un modello che generi dataset di eventi indipendenti, e due modelli positivi: uno per il caso di eventi repulsivi, con tendenza a equispaziarsi ed uno per il caso attrattivo, che porta quindi alla formazione di cluster.

A. Descrizione modelli

Modello nullo Produce un dataset indipendente prendendo dati generati in maniera random

Modello positivo repulsivo Tramite la creazione di uno spazio lineare di punti equidistanziati costruisce un dataset di eventi dipendenti repulsivi. Nel modello è implementato un parametro di rumore random ρ. Un valore $\rho = 0$ non addiziona alcun rumore ai dati generati, viceversa $\rho = 1$ addiziona un rumore random 10 volte superiore al Δx dei dati equispaziati, riducendo il modello al caso nullo.

Modello positivo attrattivo Genera eventi clusterizzati selezionando dati da una distribuzione gaussiana centrata in una posizione casuale all'interno dell'intervallo. Il modello presenta due parametri principali: il numero di cluster desiderati nel dataset e il parametro di rumore random ρ, definito come la frazione di punti di rumore random da inserire nel dataset rispetto al totale. Analogamente al modello repulsivo, un valore $\rho = 0$ restituisce un dataset di eventi completamente clusterizzati, $\rho = 1$ viceversa corrisponde al modello nullo.

Figura 1 Visualizzazione dei dataset generati dal modello positivo attrattivo con numero di cluster pari a 5 (sx) e dal modello positivo repulsivo (dx), in funzione del parametro di rumore ρ

B. Validazione KS

Per validare il primo stimatore ho analizzato in funzione della grandezza N del set le percentuali di falsi negativi e positivi, definiti rispettivamente come i dataset indipendenti per i quali il test ha restituito un $p\text{-}value < 0.05$, e quelli dipendenti, sia repulsivi che attrattivi, per i quali ho ottenuto $p\text{-}value > 0.05$. Per ogni valore di N ho effettuato 1000 simulazioni consecutive e i risultati ottenuti sono riportati in Figura 2. Il test risulta essere affidabile per dataset con almeno 100 eventi e si nota inoltre come la percentuale di falsi positivi ottenuta testando dati clusterizzati diminuisca all'aumentare del numero di cluster presenti.

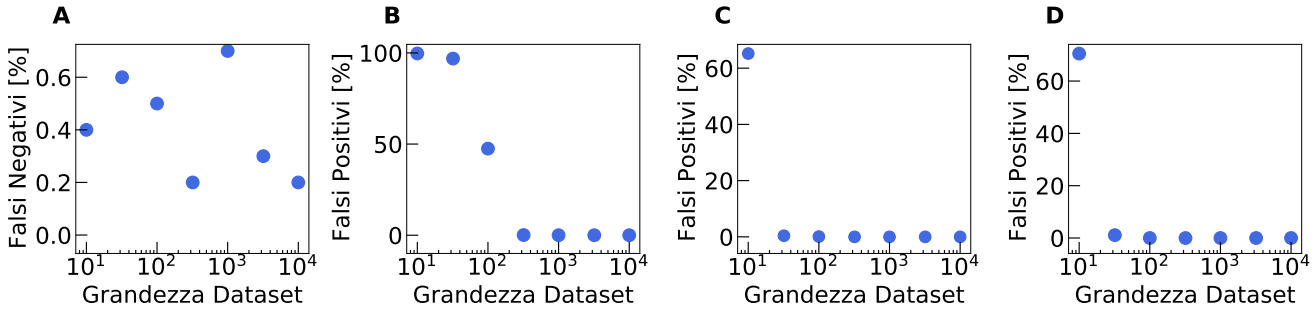

Figura 2 Lo stimatore basato sul test KS risulta affidabile per dataset caratterizzati da N > 100. (A) Percentuale di falsi negativi ottenuti per dataset ottenuti dal modello nullo; (B) percentuale falsi positivi ottenuti per dataset attrattivo con 1 cluster; (C) percentuale falsi positivi per dataset attrattivo con 5 cluster; (D) percentuale falsi positivi per dataset equidistanziato. Le grandezze testate sono: 10, 32, 100, 320, 1000, 3200 e 10000

C. Calibrazione statistica di Hopkins

La statistica di Hopkins è un metodo per misurare la tendenza a formare cluster di un dataset: dati fortemente aggregati tendono a fornire valori vicini ad 1, dati random vicini a 0.5 e infine dati perfettamente equidistanti restituiscono il valore minimo prossimo allo 0. Questo stimatore tuttavia presenta importanti caratteristiche e limitazioni di cui è doveroso tener conto: in primo luogo esso è intrensicamente aleatorio, essendo calcolato tramite il confronto di distanze prese su un sottocampione del dataset scelto casualmente con quelle di un campione totalmente random generato ad hoc [1]; in aggiunta la varianza del valore ottenuto per prove ripetute su uno stesso dataset è fortemente influenzata dalla grandezza del set. Infine il codice utilizzato [2] per computare il valore della statistica risulta sovrastimare sistematicamente i valori per il caso random e per il caso equidistante, che risultano rispettivamente prossimi a 0.66 e 0.42.

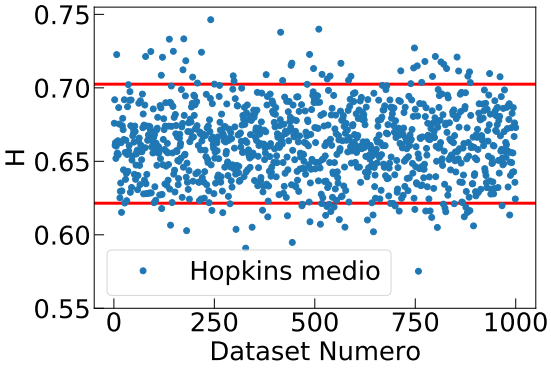

Figura 3 Distribuzione del valore della statistica di Hopkins per dataset indipendenti e valori threshold adottati (linee rosse). Ogni punto del grafico è media di 50 prove effettuate sullo stesso set. Tutti i dataset studiati sono formati da 100 eventi. I valori soglia di H definiscono il valore minimo per considerare un set di dati attrattivo e massimo per identificarlo come repulsivo, la scelta è stata effettuata in modo da ottenere un rate di falsi positivi inferiore al 5%

Ho quindi deciso, partendo dalla distribuzione dei valori di H mediati su 50 prove consecutive ed effettuate per 1000 set di dati indipendenti con $N = 100$, di definire due threshold che garantiscano un massimo di 5% di falsi positivi, ovvero dataset random identificati come repulsivi o attrattivi. I valori di H identificati sono dunque: $H_1 = 0.6215$ e $H_2 = 0.7025$, come evidenziato in Figura 3.

D. Validazione procedura

Al fine di testare la procedura ho analizzato dataset dipendenti ed indipendenti di diverse taglie (N = 100,320,1000,3200 e 10000) ottenuti tramite i modelli sopra descritti, applicando 4 livelli di rumore crescente.

Tabella I Risultati dell'applicazione della procedura a dataset di diversi ordini di grandezza e a diversi livelli di rumore applicato ρ. I dataset etichettati dalla lettera 'E' sono ottenuti tramite il modello repulsivo, i dataset 'Rand' sono completamente indipendenti e i dataset 'C' sono attrattivi. In grassetto sono evidenziati i dati valutati come indipendenti da ciascuno stimatore.

N=100				**N=1000**				**N=10000**			
Dataset	p-value	Hopkins	σ(H)	Dataset	p-value	Hopkins	σ(H)	Dataset	p-value	Hopkins	σ(H)
Eρ=0	0.0000	0.4260	0.0112	Eρ=0	0.0000	0.4277	0.0035	Eρ=0	0.0000	0.4288	0.0016
Eρ=0.25	0.0000	0.5186	0.0396	Eρ=0.25	0.0000	0.5699	0.0207	Eρ=0.25	0.0000	0.5835	0.0064
Eρ=0.5	**0.0688**	**0.6115**	0.0585	Eρ=0.5	0.0183	0.5972	0.0224	Eρ=0.5	0.0000	0.6097	0.0060
Eρ=0.75	**0.3223**	**0.6299**	0.0669	Eρ=0.75	0.0044	**0.6238**	0.0220	Eρ=0.75	0.0000	**0.6423**	0.0079
Rand	**0.7993**	**0.6756**	0.0726	Rand	**0.5435**	**0.6588**	0.0152	Rand	**0.7209**	**0.6711**	0.0082
Cρ=0.75	**0.1703**	**0.6944**	0.0949	Cρ=0.75	0.0124	0.7253	0.0203	Cρ=0.75	0.0000	0.7136	0.0083
Cρ=0.5	0.0010	0.7732	0.0663	Cρ=0.5	0.0000	0.7679	0.0231	Cρ=0.5	0.0000	0.7629	0.0090
Cρ=0.25	0.0000	0.8456	0.1327	Cρ=0.25	0.0000	0.8470	0.0255	Cρ=0.25	0.0000	0.8636	0.0100
Cρ=0	0.0000	0.9695	0.0420	Cρ=0	0.0000	0.9903	0.0101	Cρ=0	0.0000	0.9994	0.0006

Dai risultati riportati in Tabella I, si osserva come generalmente ci sia un buon accordo tra i due stimatori e come la precisione tenda ad aumentare con l'incremento della grandezza del set. Inoltre entrambi gli stimatori sembrano avere più difficoltà ad identificare correttamente i dati repulsivi rispetto a quelli attrattivi.

III. APPLICAZIONE A DATI MULTIDISCIPLINARI

Da ultimo ho applicato la procedura a dataset di eventi reali: le interazioni di 3 utenti Reddit nell'anno solare 2012 [3], il numero della base di partenza dei geni nel genoma di Escherichia-Coli [4], i principali terremoti più intensi registrati dal 1950 ad oggi [5] e la posizione di 2 specifiche parole nei primi 3 capitoli de *i Promessi Sposi (1840)* [6].

I geni tendono a repellersi L'analisi ha restituito per il test KS un *p-value* indistinguibile dallo 0, puntando dunque ad una possibile dipendenza delle posizioni di inizio di codifica dei geni. L'analisi di Hopkins ha inoltre fornito un $H = 0.552 \pm 0.009$, che coerentemente suggerisce una interazione repulsiva.

L'analisi sui terremoti è inconcludente Considerando i terremoti globali dal 1950 ad oggi ho ottenuto *p-value*$= 2 \cdot 10^{-6}$ e $H = 0.70 \pm 0.01$. Per questo dataset i due stimatori risultano in disaccordo, il primo suggerisce l'esistenza di una dipendenza tra gli eventi mentre la statistica di Hopkins si trova esattamente a ridosso del valore di soglia. Una possibile analisi più accurata dovrebbe considerare non solo i terremoti più intensi, in modo da osservare le scosse di assestamento o terremoti secondari innescati dal più violento e dovrebbe concentrarsi su regioni più localizzate.

Nei testi letterari il comportamento è diverso in funzione della parola indagata Per il testo ho analizzato due diverse parole, scelte sulla base della loro funzione logica nel testo: la congiunzione coordinante 'e' e la proposizione semplice 'di'. La preposizione risulta essere indipendente (*p-value*$= 0.08$, $H = 0.68 \pm 0.03$), mentre la congiunzione debolmente repulsiva (*p-value*$= 9 \cdot 10^{-8}$, $H = 0.57 \pm 0.01$). Questo risultato è interessante e suppongo sia dovuto al fatto che le proposizioni sono utilizzate indipendentemente dal loro ultimo impiego in base alla necessità, mentre le congiunzioni, per necessità stilistiche, (i.e. non voler creare catene di frasi coordinate, seguite da ripetizioni di frasi semplici) tendono ad essere usate in maniera equispaziata nel testo.

Le interazioni degli utenti su Reddit risultano clusterizzate Per tutti e 3 gli utenti studiati, la procedura evidenzia una forte tendenza a formare cluster, con *p-value* non distinguibili da 0 e valori di $H = 0.945 \pm 0.006$, $H = 0.980 \pm 0.004$, $H = 0.904 \pm 0.08$. Tale comportamento è banalmente spiegabile considerando in primo luogo la necessità di dormire; in secondo luogo è ragionevole pensare che anche nell'arco della giornata gli utenti tenderanno a concentrare le interazioni nei momenti "liberi". In Figura 4 è sono riportate le interazioni registrate per un utente nel mese di Maggio 2012.

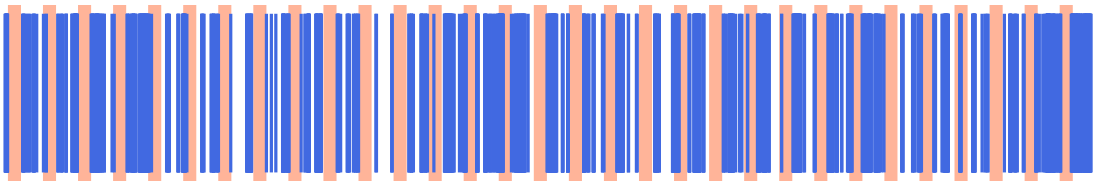

Figura 4 Visualizzazione delle interazioni durante il mese di Maggio 2012 per l'utente 'aletoledo'. I dati risultano evidentemente clusterizzati e in rosso sono evidenziati quelli che stimo essere i periodi notturni. Si nota come in questi intervalli le occorrenze sono sistematicamente minori se non assenti.

Riferimenti bibliografici

[1] https://en.wikipedia.org/wiki/Hopkins_statistic
[2] https://github.com/romusters/hopkins
[3] https://drive.google.com/file/d/1fhVeVZSqmVjVvoLpuNb7nHMjfoYSwgte
[4] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC56896/
[5] https://www.ngdc.noaa.gov/hazel/view/hazards/earthquake/search
[6] https://it.wikisource.org/wiki/I_promessi_sposi_(1840)/Capitolo_I
https://it.wikisource.org/wiki/I_promessi_sposi_(1840)/Capitolo_II
https://it.wikisource.org/wiki/I_promessi_sposi_(1840)/Capitolo_III