Definizione e calibrazione di una procedura per la quantificazione di interazione tra eventi monodimensionali

Matteo Citterio

Introduzione

Il report ha come fine definire una procedura che sia in grado di quantificare l'eventuale interazione che intercorre tra eventi appartenenti ad un intervallo monodimensionale ed esporne limiti e potenzialità, approfondendo l'eventuale rapporto che sussiste tra la natura dell'interazione e una diversa distribuzione dei dati. Per raggiungere tale obbiettivo vengono utilizzati dataset prodotti sia da simulazioni[1] sia da dati reali inerenti a diversi campi di studio[2].

Definizione della procedura

La procedura immaginata, schematizzata in **Figura 1**, si divide in tre parti principali:

1. Inizialmente si **calcola il coefficiente di Hopkins**[3]: uno strumento statistico in grado di misurare la tendenza di un dataset a formare cluster che funge da test di ipotesi nel quale H_0 è che i dati siano derivanti da un processo poissoniano e quindi distribuiti uniformemente. Il calcolo produce un valore $h \in [0, 1]$, il quale tende a 1 per dati che clusterizzano, a 0 per dati equispaziati e a 0.5 per dati distribuiti uniformemente[4]. Il calcolo viene ripetuto per m iterazioni, numero che viene ottenuto con simulazioni al variare della taglia del campione e che permette di avere un errore relativo[5] sulla media inferiore all'2,5%.

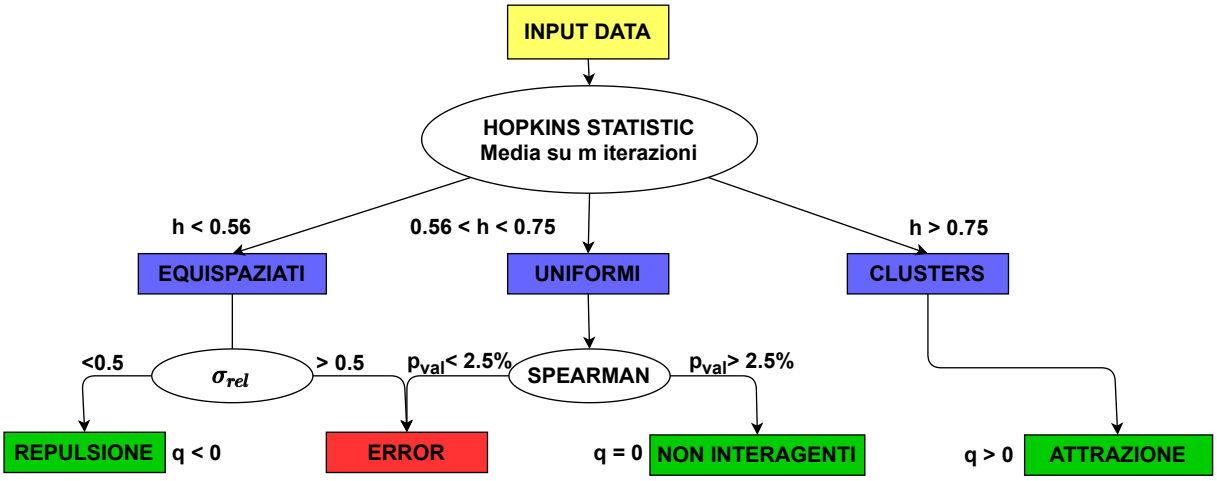

Figura 1 : Schema riassuntivo della procedura. Sono distinguibili tre diversi passaggi: calcolo del coefficiente di Hopkins, secondo controllo sull'ipotesi nulla assunta a partire dalla valutazione di h ed infine computo del quantificatore q definito in (1) come valore di output dell'algoritmo. E' possibile osservare come nel caso di $\sigma_{rel} > 0.5$ o $p_{val} < 2.5\%$, essa non sia in grado di stabilire l'interazione che intercorre tra i dati; questo particolare punto viene approfondito in *Validazione e limiti mediante dati reali.*

2. Una volta definite delle opportune soglie S_h per h (si veda sezione *Calibrazione soglie*), è possibile distinguere i tre diversi casi in cui si presentano i dati e applicare un **secondo controllo** che assuma come ipotesi nulla la distribuzione individuata con il calcolo di h. Come secondo controllo sugli equispaziati si definisce un valore massimo dell'osservabile σ_{rel}[6] accettato, al quale (secondo modalità chiarite in *Calibrazione soglie*) è connessa

una determinata significatività statistica. Per dati distribuiti uniformemente, invece, dal momento che indipendenza implica incorrelazione, si calcola il coefficiente di correlazione di Spearman e con il relativo p_{value} si è in grado di quantificare con che significatività i dati risultano incorrelati o meno (H_0)[7]. Nel caso invece di clustering, le simulazioni con diverso numero di cluster, taglia del campione e larghezza dei cluster mostrano come il coefficiente di Hopkins sia da solo un test soddisfacente.

3. Infine si **definisce una osservabile q** $\in (-1, 1)$ che quantifica l'interazione dei dati del campione:

$$q := \begin{cases} < 0, \, repulsione & (\sigma_{rel} \cdot 2) - 1 \\ = 0, \, non\, interagenti & \\ > 0, \, attrazione & (h - 0.75) \cdot (1/0.25) \end{cases} \tag{1}$$

Calibrazione soglie

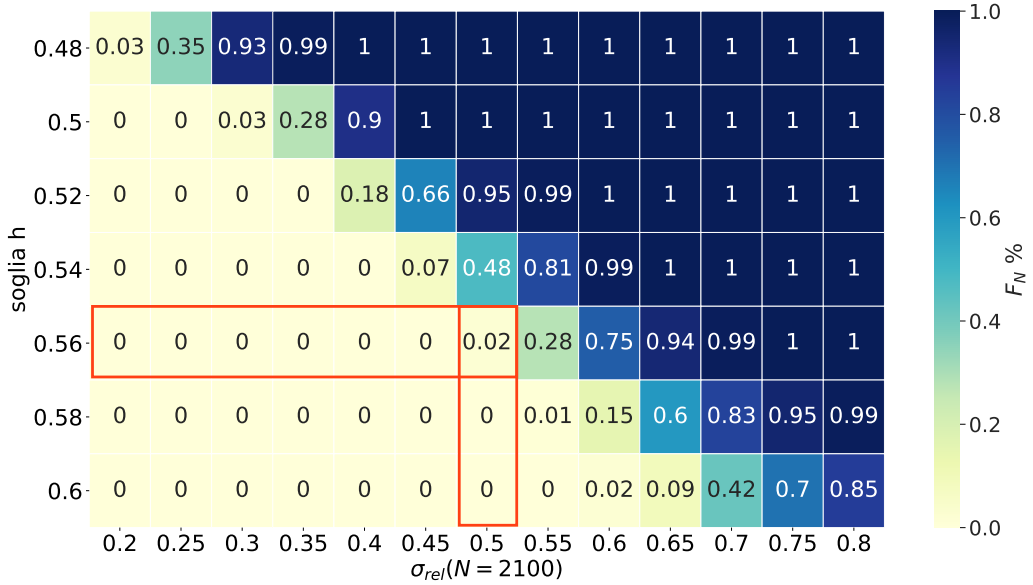

Figura 2 : Sono riportati i risultati della simulazione con taglia del campione N=2100 eseguita al fine di decidere S_h per la distinzione tra equispaziati e processo di Poisson. Le colonne riportano la percentuale di falsi negativi[8] per fissata larghezza della distribuzione σ_{rel}. Le righe invece sono le percetuali di falsi negativi che si avrebbero adottando una particolare S_h al variare della larghezza massima accettata della distribuzione dei Δx. In rosso viene evidenziata la particolare soglia adottata nel corso della procedura.

La scelta di S_h che distingue il caso *uniforme* da quello *equispaziato*, viene fatta sulla base di una simulazione che a partire da dati equispaziati introduce un 'rumore' in grado di riprodurre una distribuzione degli intervalli Δx tra eventi consecutivi con una σ_{rel} desiderata. Generato il dataset, è possibile calcolare la percetuale di falsi negativi[9] al variare della soglia; a priori questo procedimento andrebbe fatto in funzione della taglia del campione, tuttavia le simulazioni mostrano come, a partire da 2000 eventi[10] per fisso rumore, h sia stabile[11]. A questo punto è necessario scegliere arbitrariamente una larghezza limite σ_{rel} della distribuzione dei Δx oltre la quale gli eventi si definiscono non interagenti e distribuiti uniformemente. Per i fini del report si è scelta come larghezza $\sigma_{rel} = 0.5$, alla quale corrisponde una $S_h = 0.56$ con un annessa percentuale di falsi negativi $F_N \leq 2\%$.

Per la distinzione tra il caso di clustering e quello di distribuzione uniforme, le simulazioni mostrano come il coefficiente di Hopkins sia in grado di distinguere i due diversi fenomeni, nelle forme più disparate, con una F_N minima per una soglia di h=0.75. Per quantificare invece

quanto la correlazione sia significativamente diversa da 0, si adotta un livello di significatività $\alpha = 2.5\%$ nel test di Spearman.

Validazione e limiti mediante dati reali

L'applicazione della procedura ai dati reali[12] ha prodotto i risultati riportati nella tabella di **Figura 3**. In particolare l'analisi del genoma di escherichia coli ha fatto emergere i limiti di quest'ultima: con un errore percentuale su $h_{best} < 2.5\%$[13], si ha che il valore di h_{best} dotato di errore è comunque superiore a S_h per gli equispaziati e anche il controllo su σ_{rel} fallisce. Le simulazioni mostrano come nel caso di taglia N=4600 [14], prendendo come larghezza $\sigma_{rel} = 0.570$ e come S_h di distinzione tra equispaziati e uniforme il valore $h_{best} + \sigma_h = 0.565$, si avrebbe associata una $F_N = 18.0\%$. In queste condizioni il risultato sarebbe di repusione con $q = -0.14$. Se ne conclude che in questo caso la procedura risulti 'indecisa' tra i due diversi comportamenti e dunque inaffidabile. Test più approfonditi andrebbero fatti per stabilire la natura di un'eventuale interazione tra dati.

	h_{best}	σ_h	Second control	Result	q
Terremoti	0.703	± 0.009	$r = 0.12, p = 0.67$	Indipendenti	0
Genoma	0.558	± 0.006	$\sigma_{rel} = 0.570$	?	0 / -0.14
aletoledo	0.9954	± 0.0003	/	Attrazione	0.98
mexicodoug	0.976	± 0.001	/	Attrazione	0.9
nixonrichard	0.931	± 0.003	/	Attrazione	0.72
maestro, duca, scoglio	0.855	± 0.007	/	Attrazione	0.42
Cristo, donna, Beatrice	0.902	± 0.003	/	Attrazione	0.61
e	0.604	± 0.0002	$r = 0.03, p = 0.15$	Indipendenti	0

Figura 3 : Ogni riga della tabella corrisponde ad un dataset: *aletoledo*, *mexicodoug*, e *nixonrichard* sono i nomi dei tre utenti reddit i cui post sono stati analizzati e per questi, dalla definizione della procedura, la colonna *Second control* risulta vuota. Le ultime tre righe riportano i gruppi di[15] o le singole parole scelte per i test sulla *Divina Commedia*.

Riguardo ai terremoti è emerso che i diversi eventi non risultano dipendenti: questo potrebbe essere dovuto al fatto che sono stati presi in considerazione solo i terremoti significativi su scala globale; una stessa analisi fatta a livello locale potrebbe dare risultati diversi e far emergere una dipendenza dovuta al movimento delle singole placche. Anche per gli utenti di Reddit la procedura è andata a buon fine, mostrando come i commenti nel tempo tendano a clusterizzare. Ciò potrebbe essere spiegato immaginando che un utente può risultare più attivo una volta iniziata una particolare discussione e che si possa presentare un periodo di inattività qualora questa terminasse.

L'analisi sulla Divina Commedia ha invece mostrato come le *stopwords*[16], in un testo sufficientemente lungo, tendono a disporsi indipendentemente dalla occorrenza della precedente, mentre vocaboli specifici della vicenda narrata nel testo tendono invece a formare dei cluster, detti *semantic clusters*. Nello specifico la parola *e* si distribuisce uniformemente nell'arco del testo ($h = 0.604$) mentre l'unione di *maestro*, *duca* e *scoglio*, così come quella di *Cristo*, *donna* e *Beatrice* tendono ad attrarsi [17] con q di 0.42 e 0.61 rispettivamente.

In **conclusione** la procedura è risultata efficacie per i diversi dataset ma può incontarare delle difficoltà per eventi che mostrano una leggera repulsione. Raffinamenti ulteriori andrebbero fatti in questa direzione.

Notes

[1]Repository Github con codice: `https://github.com/matteocitterio/Laboratorio_Computazionale`

[2]

- Dataset **terremoti significativi** ultimi 4000 anni. Vengono presi in considerazione solo gli eventi che riportano nella data l'anno, il mese e il giorno. `www.ngdc.noaa.gov/hazel/view/hazards/earthquake/search`, ogni terremoto è visto come 'evento'.
- **Genoma di Escherichia-coli** `www.ncbi.nlm.nih.gov/pmc/articles/PMC56896/`, la base azotata di avvio del gene viene vista come 'evento'.
- Commenti scritti su **Reddit** da diversi utenti nell'arco di poco meno di 10 anni `drive.google.com/file/d/1fhVeVZSqmVjVvoLpuNb7nHMjfoYSwgte`, la data del commento viene presa come 'evento'.
- **Divina Commedia** `https://raw.githubusercontent.com/dlang/druntime/master/benchmark/extra-files/dante.txt`, la posizione di un determinato vocabolo rispetto a tutte le parole del testo viene preso come 'evento'.

[3]`https://github.com/romusters/hopkins`

[4]`https://en.wikipedia.org/wiki/Hopkins_statistic`

[5]$\sigma^x_{rel} = \sigma_x / x_{best}$

[6]see note 5, si tratta dell'errore relativo sulla lunghezza degli intervalli ed esprime di fatto la larghezza della distribuzione dei Δx.

[7]`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html`

[8]i.e. dato un numero N di iterazioni del processo, il numero di occorrenze in cui si ha un h superiore alla soglia (corrispondenti ad un test negativo per gli equispaziati) diviso N.

[9]see note 8

[10]ovvero l'ordine di grandezza dei dataset analizzati in *Validazione e limiti mediante dati reali.* Nella simulazione riportata si è utilizzato N=2100.

[11]si veda nella repository di note 1 la simulazione a supporto di questa affermazione.

[12]see note 2

[13]see Sec. *Definizione della procedura*, 1.

[14]ovvero quella del campione preso in considerazione, Genoma di Escherichia Coli.

[15]Per 'evento' qui si intende l'occorrenza di una delle parole appartenenti al gruppo

[16]congiunzioni, preposizioni e articoli; si veda `https://amslaurea.unibo.it/3101/1/baruffaldi_federico_tesi.pdf`

[17]Nello specifico è possibile vedere come clusterizzino in parti specifiche del testo: rispettivamente le cantiche *Inferno* e *Paradiso*; nella repository di note 1 sono presenti grafici a sostegno di queste affermazioni.