Presentazione del Dunn Index

2) Il problema della valutazione di una o più tecniche di cluster analysis si occupa di determinare se e quanto un set di dati ha una struttura non randomica. Tale valutazione viene fatta tramite metriche di valutazione ed indici che sono tradizionalmente classificati in 2 categorie: esterni o supervisionati e interno o non supervisionati. Gli indici esterni valutano la similarità con una struttura esterna nota, quelli interni danno delle misure di coesione o separazione proprie delle strutture di ogni cluster.
Il dunn index appartiene alla categoria degli indici interni ed è definito come:

$$V_D = \min_{1 \le i \le k} \left\{ \min_{i+1 \le j \le k} \left(\frac{D(C_i, C_j)}{\max_{1 \le l \le k} diam(C_l)} \right) \right\},$$

k = il numero totale di gruppi formati dall'algoritmo di cluster

D = distanza tra due cluster → $D(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y),$

diam = diametro → $diam(C_l) = \max_{x,y \in C_l} d(x, y).$

d = $d(u, v) \triangleq \|u - v\| \triangleq \langle u - v | u - v \rangle^{1/2}$

Dunque, il Dunn Index è calcolato come il rapporto tra la minima distanza tra due diversi cluster e il più grande diametro tra i vari cluster. Valori alti dell'indice, ossia distanza minima tra cluster di gran lunga superiore al diametro massimo, significano che l'algoritmo produce CWS (compact and well-separated) clusters. Per capire il numero ottimale di cluster da utilizzare, è sufficiente vedere a quale k corrisponde il Dunn Index più alto. [1] *"Quantifica la relatività tra il grado di compattezza e il grado di separazione degli ammassi."* [3].

Gli aspetti negativi legati a questa metrica sono dovuti alla sua scarsissima scalabilità (più o meno si parla di 20 ore per dataset contenenti 2 milioni di osservazioni) e forte sensibilità agli outliers. Per questo motivo, spesso la distanza tra cluster viene approssimata alla distanza tra centroidi, definita come: $\delta ij = d\ (\ vi\ ,\ vj\ )$, *(2) con* vi *e* vj *sono rispettivamente i centroidi dei cluster* Ci *e* Cj *."*[3]

3) Articoli:

[1] https://www.crcpress.com/Data-Clustering-Algorithms-and-Applications/Aggarwal-Reddy/p/book/9781466558212

[2] Clustering validation of CLARA and K-means using Silhuette & Dunn Measures on Iris Dataset

[3] Parallel and scalable Dunn Index for the validation of big data clusters

[4] BMS: An improved Dunn index for Document Clustering validation

Articoli ambito medico:

[5] Dunn's index for cluster tendency assessment of pharmacological data sets

[6] Fuzzy and hard clustering analysis for thyroid disease

4) Pacchetti Python:

- Validclust

5) Per vedere il comportamento della metrica al variare della complessità di un dataset, la si è testata su un dataset artificiale costituito solo da 0 ed 1 che ad ogni test veniva disturbato con tuple di "disturbo" di numeri compresi tra 0 ed 1 sostituite in maniera randomica all'interno del dataset originario.
Il dataset è creato unendo 150 tuple contenenti 150 0 ed altrettante tuple contenenti solo 1, dunque con una shape finale di 300x150.
La valutazione del primo test, con dataset senza disturbo e kmeans con k =2, ha come dunn index il valore inf. Questo è il più alto valore che l'indice possa assumere.
Già al secondo test con un rumore di appena 3 tuple, il valore del dunn index cambia radicalmente passando da inf a 0.68. In questo caso, provando ad aumentare il numero di cluster da 2 a 4 si ottiene un leggero miglioramento, infatti risulta essere pari a 0.975. Per come è stato creato il test, si nota che ogni volta che si runna la cella, il valore del Dunn

Index cambia leggermente. Questo andamento è probabilmente legato alla randomicità con cui il rumore viene inserito all'interno del dataset.
Andando avanti con i test si ottengono i seguenti risultati:

N. TEST	Dimensione Rumore	N. Cluster	Valore del Dunn Index
3	6	2	0.6893
3	6	4	0.90687
3	6	6	0.91909
4	12	4	0.81404
4	12	7	0.82073
5	24	2	0.6312
5	24	8	0.8350
6	48	2	0.8762
6	48	10	0.7793
7	96	2	0.829667
7	96	8	0.73328
8	150	2	0.7914
8	150	12	0.7284
9	270	2	0.8108
9	270	7	0.675
10	300	2	0.66502
10	300	11	0.6778

Siccome le tuple di rumore sono sostituite in maniera casuale all'interno del dataset, nonostante si sia impostato il random_state del k-means in modo tale che l'inizializzazione iniziale sia sempre la stessa, la metrica cambia leggermente perchè il dataset è ogni volta leggermente diverso.

6) Si è poi passati a test su dataset di complessità reale. I 5 dataset presi in esame sono relativi a dati pubblici di cartelle cliniche e differiscono in dimensioni e complessità.
Per ogni dataset, è stato selezionato il numero di cluster tra 2 e 15 che massimizzasse il dunn index.
Nel primo caso il dataset era di dimensioni 1257 × 16 e la heatmap mostrava forte correlazioni solo tra 2 di queste 16 variabili. In questo caso, il numero di cluster che massimizzava il dunn index è 4 che corrisponde ad un dunn index pari a: 0.01889, ossia molto basso.

Nel successivo dataset si avevano dimensioni pari a 422 × 10, quindi inferiori al precedente e correlazioni mai superiori a 0.28. In questo caso, con 14 cluster, si aveva un dunn index pari a 0.0698, leggermente più alto del primo caso, ma ancora molto molto vicino allo 0.
Il terzo dataset 67 × 20 e con correlazioni di media basso valore, ma molto sparse restituisce il valore di dunn index più alto per 10 cluster: 0.346.
Il quarto dataset ha dimensioni simili al secondo 425 × 16 e correlazioni medio basse con tutte le variabili. In questo caso si ottengono 4 clusters con dunn index pari a 0.10.
L'ultimo dataset ha dimensioni 173 x 13 e basse correlazioni. Il dunn index più alto si ha in corrispondenza di 14 cluster ed è pari a: 0.196.