Elisa Maria Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola:869436 - 19 luglio 2024

ANALISI DEI CINQUE DATASETS

In questa parte del progetto ho analizzato cinque datasets contenenti dati reali di cartelle cliniche, calcolando il valore della gap statistic su tre diversi algoritmi: k-means con due, tre e quattro centroidi e metriche di distanza diverse, cluster gerarchico applicato con diversi metodi e distanze, dbscan passandogli diversi valori di eps e cambiando il numero minimo di punti, per poi confrontare ciascun valore e capire qual è il migliore.
Per l'ultimo caso ho preso spunto da questo articolo:
https://doi.org/10.1023/A:1009745219419
Ho deciso di utilizzare entrambe le funzioni per il calcolo della statistica gap: index.Gap e clusGap.
Per il clustering gerarchico e dbscan, clusGap non supporta questi algoritmi, quindi utilizzo solo index.Gap.
Per calcolare il k-means con la distanza Manhattan ho utilizzato la funzione KMEANS_FUNCTION, che in questo caso funziona solo con index.Gap
Quando utilizzo clusGap, applico la media dei risultati che essa produce, per ottenere così un solo risultato finale.

Per importare ciascun set di dati ho utilizzato la funzione *read.csv()* e ho notato che la maggior parte di loro contiene dei valori NaN *(Non un numero),* quindi ho deciso di utilizzare la funzione *na.omit* per rimuovere questi dati.

Primo set di dati: NEUROBLASTOMA
Il primo set di dati è composto da 169 righe e 12 colonne. Analizza i dati dei pazienti affetti da neuroblastoma, un tumore che ha origine nei tessuti nervosi, soprattutto nella cavità addominale e nei tessuti della ghiandola surrenale.

	index.Gap	clusGap
K-means con due centroidi e distanza Euclidea	0.3236445	0.2089862
K-means con due centroidi e distanza Manhattan	0.3376814	
K-means con tre centroidi e distanza Euclidea	0.2372976	0.1992339
K-means con tre centroidi	0.2717767	

e distanza Manhattan		
K-means con quattro centroidi e distanza Euclidea	0.1932953	0.1948551
K-means con quattro centroidi e distanza Manhattan	0.2014243	
Cluster gerarchico con metodo complete e distanza euclidea	-0.2063236	\
Cluster gerarchico con metodo complete e distanza manhattan	-0.2291615	\
Cluster gerarchico con metodo sigle e distanza euclidea	-0.4339513	\
Cluster gerarchico con metodo sigle e distanza manhattan	-0.4452318	\
Cluster gerarchico con metodo average e distanza euclidea	-0.4068472	\
Cluster gerarchico con metodo average e distanza manhattan	-0.4262724	\
Cluster gerarchico con metodo Ward e distanza euclidea	-0.2182547	\
Cluster gerarchico con metodo Ward e distanza manhattan	-0.2123512	\
DBSCAN	0.4283512	\
DBSCAN con $eps = numero\ di\ righe/2$	0.4255778	\
DBSCAN con eps uguale a quello precedente ma $minPts = eps + 1$	0.4215725	

Elisa Maria Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola:869436 - 19 luglio 2024

Il valore migliore della gap statistic è **0.44** in DBSCAN.

Dataset contenente pazienti affetti da NEUROBLASTOMA							
Metodo	Cluster	Distanza	Nstart	Linkage	Eps	MinPoints	Valore Gap
DBSCAN	**/**	**/**	**/**	**/**	**85**	**14**	**0.4255778**
DBSCAN	/	/	/	/	85	86	0.4215725
DBSCAN	/	/	/	/	4.9	14	0.418907
K-Means	2	Manhattan	25	/	/	/	0.3376814
K-Means	2	Euclidea	25	/	/	/	0.3236445
K-Means	3	Manhattan	25	/	/	/	0.2717767
K-Means	3	Euclidea	25	/	/	/	0.2372976
K-Means	4	Manhattan	25	/	/	/	0.2014243
K-Means	4	Euclidea	25	/	/	/	0.1932953
Cluster gerarchico	2	Euclidea	/	Complete	/	/	-0.2063236
Cluster gerarchico	2	Manhattan	/	Ward	/	/	-0.2123512
Cluster gerarchico	2	Euclidea	/	Ward	/	/	-0.2182547
Cluster gerarchico	2	Manhattan	/	Complete	/	/	-0.2291615
Cluster gerarchico	2	Euclidea	/	Average	/	/	-0.4068472
Cluster gerarchico	2	Manhattan	/	Average	/	/	-0.4262724
Cluster gerarchico	2	Euclidea	/	Single	/	/	-0.4339513
Cluster gerarchico	2	Manhattan	/	Single	/	/	-0.4452318

Elisa Maria Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola:869436 - 19 luglio 2024

Secondo set di dati: SEPSI SIRS

Il secondo set di dati è composto da 1257 righe e 16 colonne. Analizza i pazienti affetti da sepsi, un'infezione generalizzata che può colpire uno o più organi e compromettere anche la loro funzionalità. La SIRS è una risposta infiammatoria sistemica messa in atto dall'organismo.

	index.Gap	clusGap
K-means con due centroidi e distanza Euclidea	0.327228	0.2097501
K-means con due centroidi e distanza Manhattan	0.3374389	
K-means con tre centroidi e distanza Euclidea	0.2586502	0.20394
K-means con tre centroidi e distanza Manhattan	0.266701	
K-means con quattro centroidi e distanza Euclidea	0.2115479	0.197116
K-means con quattro centroidi e distanza Manhattan	0.2192956	
Cluster gerarchico con metodo complete e distanza euclidea	2.209649	\
Cluster gerarchico con metodo complete e distanza manhattan	2.225656	\
Cluster gerarchico con metodo sigle e distanza euclidea	2.220826	\
Cluster gerarchico con metodo sigle e distanza manhattan	2.224055	\
Cluster gerarchico con	2.213402	\

metodo average e distanza euclidea		
Cluster gerarchico con metodo average e distanza manhattan	2.212292	\
Cluster gerarchico con metodo Ward e distanza euclidea	2.287899	\
Cluster gerarchico con metodo Ward e distanza manhattan	2.296092	\
DBSCAN	1.854507	\
DBSCAN con $eps = numero\ di\ righe/2$	1.856206	\
DBSCAN con eps uguale a quello precedente ma $minPts = eps + 1$	1.857256	

Il valore migliore della gap statistic è **2.30** nel cluster gerarchico utilizzando il metodo Ward e la distanza Manhattan.

Dataset contenente pazienti affetti da SEPSI SIRS							
Metodo	Cluster	Distanza	Nstart	Linkage	Eps	MinPoints	Valore Gap
Cluster gerarchico	**2**	**Manhattan**	**/**	**Ward**	**/**	**/**	**2.296092**
Cluster gerarchico	2	Euclidea	/	Ward	/	/	2.287899
Cluster gerarchico	2	Manhattan	/	Complete	/	/	2.225656
Cluster gerarchico	2	Manhattan	/	Single	/	/	2.224055
Cluster gerarchico	2	Euclidea	/	Single	/	/	2.220826
Cluster gerarchico	2	Euclidea	/	Average	/	/	2.213402

Cluster gerarchico	2	Manhattan	/	Average	/	/	2.212292
Cluster gerarchico	2	Euclidea	/	Complete	/	/	2.209649
DBSCAN	/	/	/	/	629	630	1.857256
DBSCAN	/	/	/	/	629	17	1.856206
DBSCAN	/	/	/	/	78	17	1.854507
K-Means	2	Manhattan	25	/	/	/	0.3374389
K-Means	2	Euclidea	25	/	/	/	0.327228
K-Means	3	Manhattan	25	/	/	/	0.266701
K-Means	3	Euclidea	25	/	/	/	0.2586502
K-Means	4	Manhattan	25	/	/	/	0.2192956
K-Means	4	Euclidea	25	/	/	/	0.2115479

Terzo set di dati: DEPRESSIONE E SCOMPENSO CARDIACO
Il terzo set di dati è composto da 425 righe e 16 colonne. Analizza i pazienti con depressione e insufficienza cardiaca. L'insufficienza cardiaca è una compromissione cronica della funzione cardiaca e la depressione, i disturbi d'ansia sono condizioni psichiatriche in questi pazienti (*10.1097/HRP.0000000000000162*).

	index.Gap	clusGap
K-means con due centroidi e distanza Euclidea	0.3260289	0.2118088
K-means con due centroidi e distanza Manhattan	0.3371491	
K-means con tre centroidi e distanza Euclidea	0.2195197	0.2014391
K-means con tre centroidi e distanza Manhattan	0.2785578	
K-means con quattro centroidi e distanza	0.2359968	0.1905523

Euclidea		
K-means con quattro centroidi e distanza Manhattan	Non può essere calcolata, distanza = 0	
Cluster gerarchico con metodo complete e distanza euclidea	0.4865334	\
Cluster gerarchico con metodo complete e distanza manhattan	0.6060147	\
Cluster gerarchico con metodo sigle e distanza euclidea	0.4622431	\
Cluster gerarchico con metodo sigle e distanza manhattan	0.4717816	\
Cluster gerarchico con metodo average e distanza euclidea	0.491041	\
Cluster gerarchico con metodo average e distanza manhattan	0.4685552	\
Cluster gerarchico con metodo Ward e distanza euclidea	0.6135533	\
Cluster gerarchico con metodo Ward e distanza manhattan	0.6232382	\
DBSCAN	5.95631	\
DBSCAN con $eps = numero\ di\ righe/2$	-0.1769808	\
DBSCAN con eps uguale a quello precedente ma $minPts = eps + 1$	3.738619	

Il valore migliore della gap statistic è **5.96** in DBSCAN.

Elisa Maria Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola:869436 - 19 luglio 2024

Dataset contenente pazienti affetti da DEPRESSIONE e SCOMPENSO CARDIACO							
Metodo	Cluster	Distanza	Nstart	Linkage	Eps	MinPoints	Valore Gap
DBSCAN	**/**	**/**	**/**	**/**	**60**	**17**	**5.95631**
DBSCAN	/	/	/	/	70	17	3.738619
Cluster gerarchico	2	Manhattan	/	Ward	/	/	0.6232382
Cluster gerarchico	2	Euclidea	/	Ward	/	/	0.6135533
Cluster gerarchico	2	Manhattan	/	Complete	/	/	0.6060147
Cluster gerarchico	2	Euclidea	/	Average	/	/	0.491041
Cluster gerarchico	2	Euclidea	/	Complete	/	/	0.4865334
Cluster gerarchico	2	Manhattan	/	Single	/	/	0.4717816
Cluster gerarchico	2	Manhattan	/	Average	/	/	0.4685552
Cluster gerarchico	2	Euclidea	/	Single	/	/	0.4622431
K-Means	2	Manhattan	25	/	/	/	0.3371491
K-Means	2	Euclidea	25	/	/	/	0.3260289
K-Means	3	Manhattan	25	/	/	/	0.2785578
K-Means	4	Euclidea	25	/	/	/	0.2359968
K-Means	3	Euclidea	25	/	/	/	0.2195197
DBSCAN	/	/	/	/	213	17	-0.1769808
K-Means	4	Manhattan	25	/	/	/	Non può essere calcolata, distanza = 0

Elisa Maria Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola:869436 - 19 luglio 2024

Quarto set di dati: ARRESTO CARDIACO

Il quarto set di dati è composto da 416 righe e 10 colonne. Analizza i pazienti con arresto cardiaco in Spagna.

	index.Gap	clusGap
K-means con due centroidi e distanza Euclidea	0.3261278	0.2090555
K-means con due centroidi e distanza Manhattan	0.3295978	
K-means con tre centroidi e distanza Euclidea	0.2365713	0.2007318
K-means con tre centroidi e distanza Manhattan	0.2648943	
K-means con quattro centroidi e distanza Euclidea	0.2044823	0.1943299
K-means con quattro centroidi e distanza Manhattan	0.2140072	
Cluster gerarchico con metodo complete e distanza euclidea	-0.1195936	\
Cluster gerarchico con metodo complete e distanza manhattan	-0.05987362	\
Cluster gerarchico con metodo sigle e distanza euclidea	-0.1616678	\
Cluster gerarchico con metodo sigle e distanza manhattan	-0.1382712	\
Cluster gerarchico con metodo average e distanza euclidea	-0.09111444	\
Cluster gerarchico con metodo average e	-0.1337553	\

Elisa Maria Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola:869436 - 19 luglio 2024

distanza manhattan		
Cluster gerarchico con metodo Ward e distanza euclidea	-0.03565727	\
Cluster gerarchico con metodo Ward e distanza manhattan	-0.03598114	\
DBSCAN	1.048338	\
DBSCAN con $eps = numero\ di\ righe/2$	1.052062	\
DBSCAN con eps uguale a quello precedente ma $minPts = eps + 1$	1.054733	

Il valore migliore della gap statistic è **1.054** in DBSCAN con parametri modificati.

Dataset contenente pazienti affetti da ARRESTO CARDIACO							
Metodo	Cluster	Distanza	Nstart	Linkage	Eps	MinPoints	Valore Gap
DBSCAN	**/**	**/**	**/**	**/**	**208**	**209**	**1.054733**
DBSCAN	/	/	/	/	208	11	1.052062
DBSCAN	/	/	/	/	7.5	11	1.048338
K-Means	2	Manhattan	25	/	/	/	0.3295978
K-Means	2	Euclidea	25	/	/	/	0.3261278
K-Means	3	Manhattan	25	/	/	/	0.2648943
K-Means	3	Euclidea	25	/	/	/	0.2365713
K-Means	4	Manhattan	25	/	/	/	0.2140072
K-Means	4	Euclidea	25	/	/	/	0.2044823
Cluster gerarchico	2	Euclidea	/	Ward	/	/	-0.03565727
Cluster gerarchico	2	Manhattan	/	Ward	/	/	-0.03598114
Cluster	2	Manhattan	/	Complete	/	/	-0.05987362

gerarchico							
Cluster gerarchico	2	Euclidea	/	Average	/	/	-0.09111444
Cluster gerarchico	2	Euclidea	/	Complete	/	/	-0.1195936
Cluster gerarchico	2	Manhattan	/	Average	/	/	-0.1337553
Cluster gerarchico	2	Manhattan	/	Single	/	/	-0.1382712
Cluster gerarchico	2	Euclidea	/	Single	/	/	-0.1616678

Quinto set di dati: DIABETE DI TIPO 1

Il quinto set di dati è composto da 67 righe e 20 colonne. Analizza i pazienti affetti da diabete di tipo 1, una malattia in cui il livello di zucchero nel sangue è troppo alto perché il corpo non riesce a produrre un ormone chiamato insulina.

	index.Gap	clusGap
K-means con due centroidi e distanza Euclidea	0.3266966	0.2137857
K-means con due centroidi e distanza Manhattan	0.3450115	
K-means con tre centroidi e distanza Euclidea	0.2458823	0.2015388
K-means con tre centroidi e distanza Manhattan	0.2578782	
K-means con quattro centroidi e distanza Euclidea	0.2021218	0.1923128
K-means con quattro centroidi e distanza Manhattan	0.2110242	
Cluster gerarchico con metodo complete e distanza euclidea	0.9086383	\

Cluster gerarchico con metodo complete e distanza manhattan	0.8996735	\
Cluster gerarchico con metodo sigle e distanza euclidea	0.67492	\
Cluster gerarchico con metodo sigle e distanza manhattan	0.6512111	\
Cluster gerarchico con metodo average e distanza euclidea	0.6985328	\
Cluster gerarchico con metodo average e distanza manhattan	0.662888	\
Cluster gerarchico con metodo Ward e distanza euclidea	0.8926816	\
Cluster gerarchico con metodo Ward e distanza manhattan	0.9066049	\
DBSCAN	0.6442917	\
DBSCAN con $eps = numero\ di\ righe/2$	0.6384481	\
DBSCAN con eps uguale a quello precedente ma $minPts = eps + 1$	0.6496576	

Il valore migliore della gap statistic è **0.90** nel cluster gerarchico con metodo Ward e distanza euclidea.

Dataset contenente pazienti affetti da DIABETE di TIPO 1							
Metodo	Cluster	Distanza	Nstart	Linkage	Eps	MinPoints	Valore Gap
Cluster gerarchico	**2**	**Euclidea**	**/**	**Complete**	**/**	**/**	**0.9086383**
Cluster gerarchico	2	Manhattan	/	Ward	/	/	0.9066049

Cluster gerarchico	2	Manhattan	/	Complete	/	/	0.8996735
Cluster gerarchico	2	Euclidea	/	Ward	/	/	0.8926816
Cluster gerarchico	2	Euclidea	/	Average	/	/	0.6985328
Cluster gerarchico	2	Euclidea	/	Single	/	/	0.67492
Cluster gerarchico	2	Manhattan	/	Average	/	/	0.662888
Cluster gerarchico	2	Manhattan	/	Single	/	/	0.6512111
DBSCAN	/	/	/	/	34	35	0.6496576
DBSCAN	/	/	/	/	24	21	0.6442917
DBSCAN	/	/	/	/	34	21	0.6384481
K-Means	2	Manhattan	25	/	/	/	0.3450115
K-Means	2	Euclidea	25	/	/	/	0.3266966
K-Means	3	Manhattan	25	/	/	/	0.2578782
K-Means	3	Euclidea	25	/	/	/	0.2458823
K-Means	4	Manhattan	25	/	/	/	0.2110242
K-Means	4	Euclidea	25	/	/	/	0.2021218