Tesina di

Elaborato di Impianti di elaborazione - PCA e Clustering

Corso di Laurea

in

Ingegneria Informatica

By

Luca Antonio Scolletta

Giacomo Lisita

UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

January, 2025

INDICE

L'obbiettivo di questo elaborato è quello di creare un **workload sintetico** usando le tecniche statistiche di PCA e Clustering; lo scopo di questo workload è quello di, dato un dataset di partenza, descrivere le caratteristiche statistiche principali in termini di varianza con un numero ridotto di informazioni.

0.1 Creazione del dataset

Il primo passo effettuato è stato quello di ottenere un dataset di partenza; si è deciso di creare un dataset che contenesse informazioni relative a un dato processore, messo sotto sforzo utilizzando lo script **nbody**. In particolare, la macchina su cui è stato fatto eseguire questo script ha le seguenti caratteristiche:

Mentre lo script utilizzato, **nbody**, simula l'evoluzione di un sistema fisico di, come suggerisce il nome, N corpi sotto l'influenza della forza di gravità. È un ottimo script poiché mette sotto sforzo le chiamate ricorsive e i sottosistemi per i calcoli in virgola mobile. Si è fatto eseguire lo script per ore con i seguenti parametri:

```
./launch_nbody.sh -r 1 -n 10
```

Listing 1: Esecuzione del comando nbody

Vediamo cosa indicano i parametri specificati nel comando:

- **-r** indica il numero di volte in cui lo script deve essere eseguito;
- **-n** indica il numero di corpi utilizzati nella simulazione.

Si è poi utilizzato uno script che raccogliesse i dati del processore e della simulazione e li salvasse in un file; ecco un'immagine di esempio del dataset:

Processore(_Total)\% Tempo processore	Processore(_Total)\% Tempo privilegiato	Memoria\MByte disponibili	Memoria\% Byte vincolati in uso
10,0073124	10,594882732	13887	81,867509913
11,936631259	10,572758591	13411	83,225666052
11,081004614	10,408152853	12950	84,571309058
10,05639358	9,7513721971	12484	85,910599419
13,200083171	11,615872419	12008	87,282286046
9,6000082401	9,413980993	11540	88,627962952
10,236569029	10,030870142	11075	89,967140274
14,538586921	12,245601235	10617	91,255123376
23,587157193	24,063007504	10355	92,485921293
13,338738053	12,176617148	10061	93,323521757
9,3701682231	8,9859576973	10061	93,332451627
9,3180370591	9,1066727373	10062	93,332406411
19,926537936	19,279325131	10288	93,350752255
14,347475624	14,062358558	10405	93,355251102
20,193058085	19,53406838	10624	93,366260848
21,514744985	19,108336589	10770	93,364994855
22,494525952	20,246858092	11009	93,377711459
13,348724985	12,984832841	10975	93,678851983
22,078833774	20,855747605	10774	94,955067987
13,134842302	13,215823521	10374	96,287519624
9,4043052702	8,9825730054	9914	97,635457268
26,508932106	12,848084929	9463	98,987984215
29,632229857	14,303437463	9304	96,122486888
10,241172684	9,8012479572	8870	97,416981658
15,362519536	13,700770274	9378	89,627105275
19,88461995	18,21979451	10348	89,57603194

Figure 1: Dataset iniziale

0.2 Analisi preliminare del dataset

Lo strumento per fare un'analisi statistica dei dati è stato il software **JMP**. Come primo passo, prima di effettuare l'analisi a componenti principali, si è deciso di studiare superficialmente i dati per verificare se si potessero rimuovere informazioni che non fossero utili ai fini dello studio del dataset; in particolare, si è scelto di rimuovere la prima colonna relativa ai timestamp delle misurazioni poiché era solo un'informazione che andava ad indicare quando quest'ultime fossero state effettuate. Successivamente sono stati creati gli **istogrammi frequenziali** relativi alle colonne del dataset per valutare l'eventuale presenza di colonne che potessero essere statisticamente irrilevanti e che avrebbero potuto quindi compromettere le analisi successive.

Inserire istogrammi frequenziali

Dopo un'analisi degli istogrammi sono state trovate alcune colonne aventi varianza nulla:

- Interfaccia di rete: Errori su pacchetti in uscita;
- Interfaccia di rete: Errori su pacchetti ricevuti;

Queste colonne sono state quindi valutate come statisticamente irrilevanti e non sono state considerate nelle analisi successive; possiamo immaginare che, essendo le colonne relative alle prestazioni di rete del calcolatore, lo script nbody non metta

sotto sforzo questi sistemi e che quindi le loro prestazioni siano pressoché sempre le stesse.

0.3 Analisi a componenti principali e clustering

L'obiettivo quindi dell'analisi a componenti principali è andare a fare una riduzione dello spazio delle feature che descrivono il dataset, andando ad effettuare un base e trovando un nuovo insieme di variabili che sono una combinazione lineare di quelle iniziali e rimuovendo quelle meno impattanti in termini di varianza conservata. Per fare ciò, quello che si fa è andare a moltiplicare il vettore delle variabili iniziali normalizzate attraverso **z-score** con la *matrice degli autovettori* della matrice di covarianza:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \frac{x_1 - \bar{x}_1}{s_{x_1}} \\ \frac{x_2 - \bar{x}_2}{s_{x_2}} \end{bmatrix} \tag{1}$$

Esempio cambio di base con due features

Dal punto di vista operativo si è proceduto a fare PCA e Clustering sempre utilizzando JMP

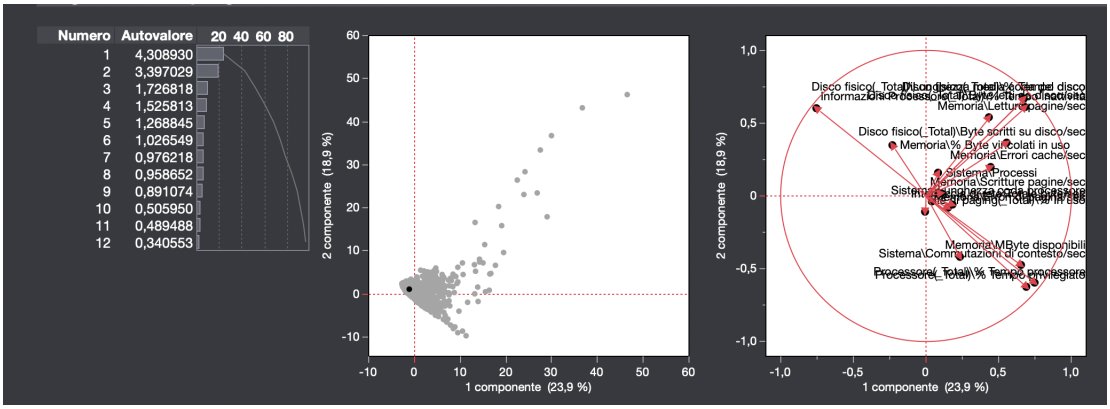

Figure 2

È stata poi fatta l'analisi degli **autovalori della matrice di covarianza** e attraverso il **test di Barlett** si sono ottenute le componenti principali e le relative **percentuali comulative** di varianza.

Numero	Autovalore	Percentuale	20 40 60 80	Percentuale cumulativa	Chi-quadrato	DF	Prob>ChiQu
1	4,308930	23,939		23,939	115859	151,882	<,0001*
2	3,397029	18,872		42,811	99887,8	141,968	<,0001*
3	1,726818	9,593		52,404	84393,1	131,081	<,0001*
4	1,525813	8,477		60,881	78640,8	117,340	<,0001*
5	1,268845	7,049		67,930	72900,9	104,081	<,0001*
6	1,026549	5,703		73,633	67903,4	91,393	<,0001*
7	0,976218	5,423		79,057	63935,1	78,984	<,0001*
8	0,958652	5,326		84,383	59095,8	67,427	<,0001*
9	0,891074	4,950		89,333	52244,5	56,934	<,0001*
10	0,505950	2,811		92,144	42688,1	47,465	<,0001*
11	0,489488	2,719		94,863	38078,1	38,534	<,0001*
12	0,340553	1,892		96,755	30740,2	30,203	<,0001*
13	0,235203	1,307		98,062	24385,4	22,911	<,0001*
14	0,160819	0,893		98,955	18759,8	16,870	<,0001*
15	0,107275	0,596		99,551	13411,9	11,808	<,0001*
16	0,080696	0,448		100,000	7327,64	7,272	<,0001*
17	0,000087	0,000		100,000	0,000	3,500	1,0000

Figure 3

Si sono selezionati quindi i componenti principali da mantenere in relazione alla varianza spiegata:

PCA	Varianza mantenuta
7	0.79
8	0.84
9	0.89
10	0.92
11	0.95
12	0.97
13	0.98

Table 1: Tabella dei dati della PCA e della varianza mantenuta

Per ogni numero di componenti principali, avendo come obiettivo quello di ridurre sia righe che colonne del dataset si è analizzata la variazione di devianza andando a scegliere diversi numeri di cluster, in particolare per **5**, **10**, **15** e **25** cluster.

Per la clusterizzazione si è scelto il **metodo di Ward**, mentre per calcolare le devianze intra e inter cluster si è proceduto prima calcolando il centroide del dataset, per poi calcolare il centroide di ogni cluster e andando così ad ottenere la devianza inter cluster; per quella intra cluster si è calcolata prima la devianza totale e poi si è sottratta quella inter cluster.

Di seguito i risultati ottenuti:

	Intra Cluster Deviance	Inter Cluster Deviance	Devianza Persa
PCA: 7, Cluster: 5	0.6983	0.3017	0.7617
PCA: 7, Cluster: 10	0.4253	0.5747	0.5459
PCA: 7, Cluster: 15	0.3133	0.6867	0.4575
PCA: 7, Cluster: 25	0.2273	0.7727	0.3896
PCA: 8, Cluster: 5	0.6935	0.3065	0.7425
PCA: 8, Cluster: 10	0.4177	0.5823	0.5108
PCA: 8, Cluster: 15	0.3061	0.6939	0.4171
PCA: 8, Cluster: 25	0.2329	0.7671	0.3556
PCA: 9, Cluster: 5	0.6670	0.3330	0.7036
PCA: 9, Cluster: 10	0.4537	0.5463	0.5138
PCA: 9, Cluster: 15	0.3356	0.6644	0.4086
PCA: 9, Cluster: 25	0.2519	0.7481	0.3342
PCA: 10, Cluster: 5	0.6864	0.3136	0.7115
PCA: 10, Cluster: 10	0.5008	0.4992	0.5407
PCA: 10, Cluster: 15	0.3676	0.6324	0.4182
PCA: 10, Cluster: 25	0.2617	0.7383	0.3207
PCA: 11, Cluster: 5	0.6952	0.3048	0.7104
PCA: 11, Cluster: 10	0.5395	0.4605	0.5625
PCA: 11, Cluster: 15	0.3958	0.6042	0.4260
PCA: 11, Cluster: 25	0.2957	0.7043	0.3309
PCA: 12, Cluster: 5	0.7426	0.2574	0.7504
PCA: 12, Cluster: 10	0.5841	0.4159	0.5966
PCA: 12, Cluster: 15	0.4194	0.5806	0.4368
PCA: 12, Cluster: 25	0.3070	0.6930	0.3278
PCA: 13, Cluster: 5	0.7362	0.2638	0.7414
PCA: 13, Cluster: 10	0.5592	0.4408	0.5680
PCA: 13, Cluster: 15	0.4330	0.5670	0.4444
PCA: 13, Cluster: 25	0.3382	0.6618	0.3514

Table 2: Tabella delle devianze intra e inter-cluster, e devianza persa per diverse combinazioni di PCA e cluster

Vediamo in seguito alcuni grafici:

Qual è la scelta ottimale di PCA e cluster per caratterizzare il workload? La risposta non è ovviamente univoca e dipende dal contesto applicativo; possiamo ragionare in termini di obiettivo desiderando una combinazione dei valori che **massimizzi** la devianza inter-cluster e **minimizzi** quella intra-cluster (valori dello stesso cluster molto simili, mentre valori di cluster diversi molto diversi tra loro). A tal proposito notiamo sicuramente che scegliere **5** o **10** cluster rappresentativi del

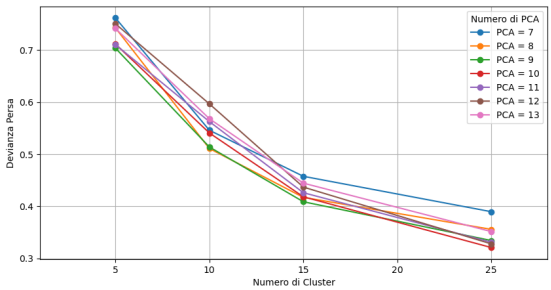

Figure 4: Enter Caption

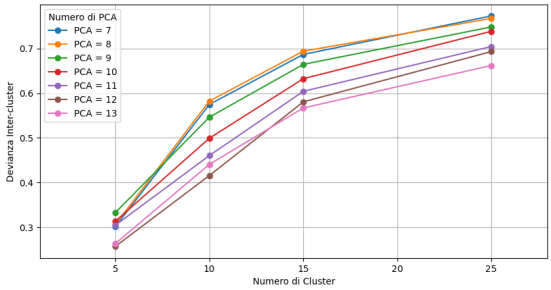

Figure 5: Enter Caption

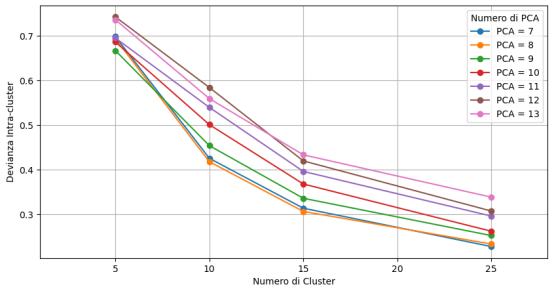

Figure 6: Enter Caption

dataset non è una scelta ottimale poiché la devianza inter cluster risulta maggiore di quella intra cluster, cioè i cluster risultano molto disomogenei e poco rappresentativi dei dati al suo interno; inoltre, la devianza persa è notevole (dal 70% al 50%). Questi valori migliorano andando a selezionare un numero superiore di cluster, con **25** cluster la scelta migliore in termini di varianza persa è andare ad utilizzare **10** componenti principali che descrivano le colonne del dataset originale, mentre fissando il vincolo sui **15** cluster si potrebbe optare per andare a utilizzare solo **9** componenti principali. Non conviene andare a selezionare un numero superiore di componenti principali poiché andremmo a perdere l'obbiettivo di ridurre la dimensionalità del dataset, ed inoltre non miglioriamo in termini di varianza persa (scegliendo un numero superiore di 10 di componenti principali la varianza persa, seppur di poco, aumenta). Non c'è una scelta migliore dell'altra e molto dipende dal dominio applicativo:

- Se vogliamo perdere meno varianza possibile e non abbiamo vincoli di costo sul numero di cluster potremmo andare a selezionare **25 cluster** e **10 componenti principali**, andando però a selezionare un numero di colonne maggiore a quelle delle soluzioni successive. In questo caso si manterrà il 55% delle colonne originali e il dataset verrà descritto con il 4% delle righe;
- Se vogliamo descrivere il dataset con meno colonne possibili e non perdere troppa varianza allora una buona scelta è **15 cluster** e **8-9 componenti principali**. Qui si mantiene dal 44% al 50% delle colonne originali e il dataset verrà descritto con il 7% delle righe.

La scelta della configurazione prevede sempre un trade-off tra perdita di devianza e budget a disposizione, per esempio considerando limiti di costo sul numero di cluster.

Si mostrano di seguito alcuni **dendrogrammi** del clustering post-PCA:

0.4 Caratterizzazione senza PCA

Si è successivamente deciso di confrontare i risultati ottenuti andando a ripetere il processo di caratterizzazione sintetica senza effettuare l'analisi a componenti principali preliminare; questo viene fatto per studiare quanto la riduzione della dimensionalità infici sulla perdita di devianza complessiva dei dati. Il numero di cluster scelti è lo stesso, ovvero **5**, **10**, **15** e **25**; si riportano quindi i risultati ottenuti:

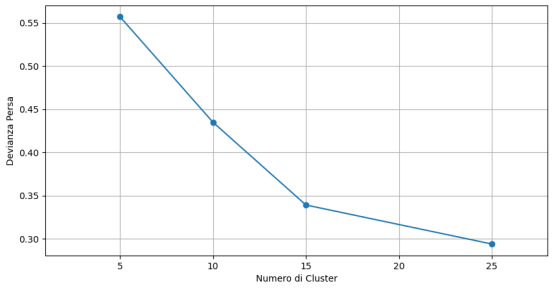

Figure 7: Enter Caption

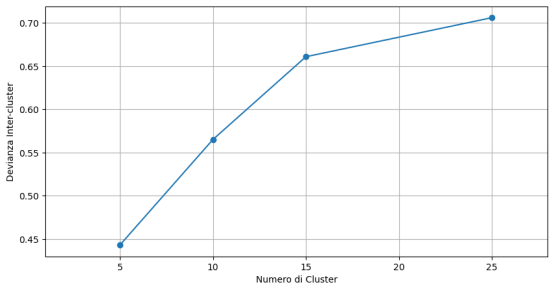

Figure 8: Enter Caption

	Intra Cluster Deviance	**Inter Cluster Deviance**	**Devianza Persa**
Cluster: 5	0.5571	0.4429	0.5571
Cluster: 10	0.4349	0.5651	0.4349
Cluster: 15	0.3390	0.6610	0.3390
Cluster: 25	0.2938	0.7062	0.2938

Table 3: Tabella delle devianze intra e inter-cluster, e devianza persa per diverse configurazioni di cluster

Si riportano in seguito anche per quest'analisi i grafici trovati:

Si noti infatti che l'analisi a componenti principali porta con sé una varianza persa maggiore rispetto a un clustering del dataset originale; la scelta sull'effettuarla o meno è sempre un trade-off, infatti se non si effettua una PCA preliminare nonostante si perda meno varianza si dovranno mantenere **tutte le colonne del dataset**

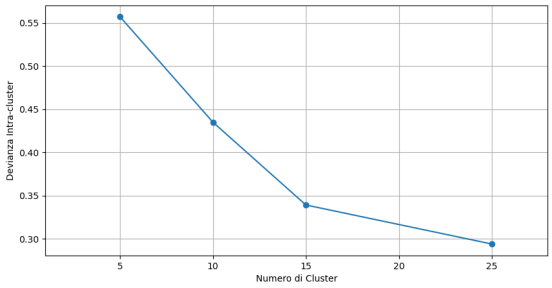

Figure 9: Enter Caption

con le eventuali correlazioni. Notiamo ovviamente che questo non è sempre un problema, per esempio nel caso in cui le colonne siano effettivamente indipendenti. Nel nostro caso potrebbe non essere una perdita troppo significativa poiché andiamo a ridurre il numero di colonne solo del 50%, ma se la riduzione delle colonne fosse considerevole allora si potrebbe optare per effettuare una PCA.