Impianti di Elaborazione

Andrea Scognamiglio - Mtr M63/598
Cristian Tommasino - Mtr. M63/615

Indice

IMPERATOR ROMANORVM SEMP AVGVST FRIDERICVS DEI GRATIA

Capitolo 1

PCA e Clustering

Estrapolare un Workload sintetico a partire dal workload reale riportato nel file *PCA-CLASTERING-2017.jmp*.

1.1 Obiettivo

Considerato il workload reale si vuole ottenere un workload sintetico che contenga un numero di osservazioni minori ma che conservando quanta più varianza possibile.

1.2 Estrazione del Workload Sintetico

Per l'estrazione del workload sintetico, dopo aver visionato i dati, si è scelto di seguire i seguenti step:

- Analisi del ***CV(Coefficente di Variazione)*** per eliminazione di parametri statisticamente non significativi;
- ***PCA(Principal Component Analysis)*** per la riduzione del numero di parametri e per l'eliminazione della correlazione tra essi;
- ***Clustering*** per la riduzione del numero di esperimenti.

1.2.1 Analisi del Coefficiente di Variazione

In prima istanza è stata effettuata un'analisi sul coefficiente di variazione(COV) il quale esprime quanta varianza contiene un parametro.
Quando il coefficiente di variazione è troppo piccolo il parametro corrispondente non è statisticamente significativo, quindi in questa fase si eliminano i

parametri con COV nullo.
Nella figura si nota che non ci sono colonna con coefficiente di variazione nulla, quindi tutti i parametri saranno utilizzati nelle successive analisi.

	Feature	Coefficient of Variation
1	Free_chached	16,472107348
2	Free_chachedPSS	19,304337052
3	Free_free	23,967256806
4	LostRam	41,600127016
5	TotalFree_	10,931016222
6	TotalUsed_	4,2498839746
7	Used_buffers	18,654578834
8	Used_PSS	3,9192501025
9	Used_shmem	28,103294244
10	Used_slab	0,9257198666
11	ZRAMinSWAP	0,3663004654
12	ZRAMPhysicalUsed	0,4042369077
13	reads_completed	2,9959659776
14	reads_merged	4,8637905221
15	sectors_read	1,6491891156
16	reading_time(ms)	3,9731659841
17	writes_completed	0,5820962711
18	writes_merged	0,4890506188
19	sectors_written	0,5652401651
20	writing_time(ms)	0,4614537729
21	io_in_progress	592,5230724
22	io_time	0,5363466144
23	io_weighted_time	0,5407412274
24	ReadingTime_over_ReadsCompleted	1,0237875449
25	WritingTime_over_WritesCompleted	0,1630171572

Figura 1.1: Cofficenti di Variazione(CV)

1.2.2 PCA

In questa fase è stata applicata la ***PCA(Principal Component Analysis)*** la quale trasforma un workload con parametri correlati in uno contente parametri non correlati.
L'utilizzo della PCA in questa è necessario anche per la fase successiva, in quanto il clustering funziona meglio se i parametri non sono correlati.
Per effettuare la PCA si è fatto utilizzo del tool ***JMP***, nella figura è riportato l'output.

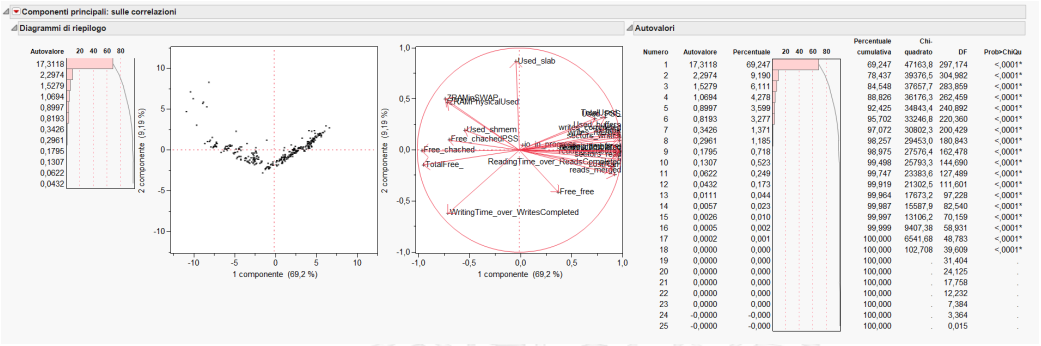


Figura 1.2: Risultato PCA

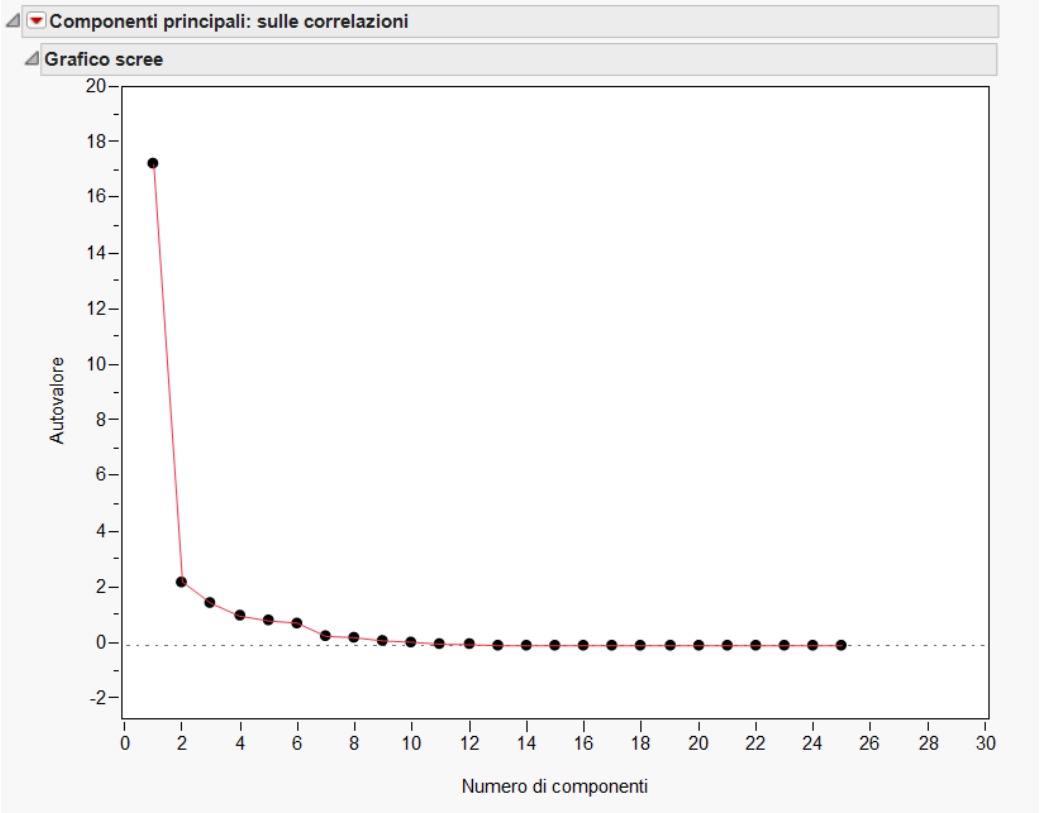


Figura 1.3: Grafico Scree

Considerando il grafico in Figura 1.3, rappresentate sull'asse delle x il numero di componenti principali e sull'asse delle y gli autovalori, per scegliere il numero di componenti principali ci si posiziona nel ginocchio della curva, poichè in quel punto ci si assicura che aggiungendo un'ulteriore componente la varianza conservata non aumenta significativamente.
Quindi si è scelto di considerare 6 componenti principali.

	Features	1 Principale	2 Principale	3 Principale	4 Principale	5 Principale	6 Principale
1	Free_chached	-0,22927	-0,00635	-0,12292	0,04946	-0,08315	0,00838
2	Free_chachedPSS	-0,16579	0,06438	0,33871	0,23843	-0,27940	-0,41395
3	Free_free	0,09168	-0,27206	0,41570	-0,34264	0,47827	0,12458
4	LostRam	0,21849	-0,12360	0,00246	-0,02381	0,04258	-0,01459
5	TotalFree_	-0,22635	-0,09523	0,21319	-0,00528	0,01743	-0,11136
6	TotalUsed_	0,20114	0,21207	-0,31258	0,02178	-0,05062	0,17114
7	Used_buffers	0,19943	0,13295	-0,19669	-0,12572	0,04117	-0,33174
8	Used_PSS	0,17717	0,20363	-0,33268	0,09068	-0,08827	0,39126
9	Used_shmem	-0,12856	0,12889	0,33863	0,39607	-0,39075	0,28984
10	Used_slab	-0,00782	0,57347	0,16880	-0,12321	0,06191	-0,30734
11	ZRAMinSWAP	-0,17430	0,33023	0,18796	-0,08362	0,20619	0,31128
12	ZRAMPhysicalUsed	-0,16907	0,30859	0,22810	-0,04592	0,18541	0,38745
13	reads_completed	0,23595	-0,01343	0,11754	0,06816	-0,05428	0,02839
14	reads_merged	0,22501	-0,15864	0,12894	0,09023	-0,06213	0,07221
15	sectors_read	0,23416	-0,05917	0,11344	0,08540	-0,07776	0,01716
16	reading_time(ms)	0,23404	-0,03321	0,12663	0,08242	-0,06297	0,06481
17	writes_completed	0,23370	0,11243	0,10922	-0,00215	0,00190	-0,05056
18	writes_merged	0,23552	0,08157	0,11093	0,01263	-0,00821	-0,03372
19	sectors_written	0,23593	0,06079	0,11656	0,01888	-0,00891	-0,02509
20	writing_time(ms)	0,23512	-0,00347	0,12626	0,03655	-0,01612	-0,01556
21	io_in_progress	0,00838	0,03132	-0,10059	0,74708	0,64033	-0,13383
22	io_time	0,23471	0,09684	0,11151	0,00594	-0,00329	-0,04031
23	io_weighted_time	0,23567	-0,00881	0,12679	0,04455	-0,02416	-0,00196
24	ReadingTime_over_ReadsCompleted	0,22213	-0,10614	0,14720	0,12574	-0,09016	0,17144
25	WritingTime_over_WritesCompleted	-0,16859	-0,41100	-0,03030	0,11119	-0,05217	0,13612

Figura 1.4: Autovettori

Nella figura sono evidenziati in rosso i parametri che hanno contribuito maggiormente, in segno positivo o negativo, alla creazione delle componenti principali scelti.
In particolare:

- ***Principale 1:*** *Free_chached, LostRam, TotalFree, reads_completed, reads_merged, sectors_read, reading_time(ms), writes_completed, writes_merged, sector_written, writing_time(ms), io_time, io_weighted_time e ReadingTime_over_ReadsCompleted*;
- ***Principale 2:*** *Used_slab e WritingTime_over_WritesCompleted*;
- ***Principale 3:*** *Free_free, TotalUsed e Used_PSS*;
- ***Principale 4:*** *io_in_progress*;

- ***Principale 5:*** *Used_ shmem*;
- ***Principale 6:*** *Free_ chachedPSS, Used_ buffers e Used_ PSS, ZRam-PhysicalUsed*;

1.2.3 Clustering

In questa fase è stato effettuate un operazione di clustering sul risultato ottenuto nello step precedente.
La tecnica di clasterizzazione scelta è di tipo gerarchico agglomerativo, in particolare è stata utilizzata la metrica di word per l'aggregazione dei cluster.
In figura è riportato il dendogramma risultante.

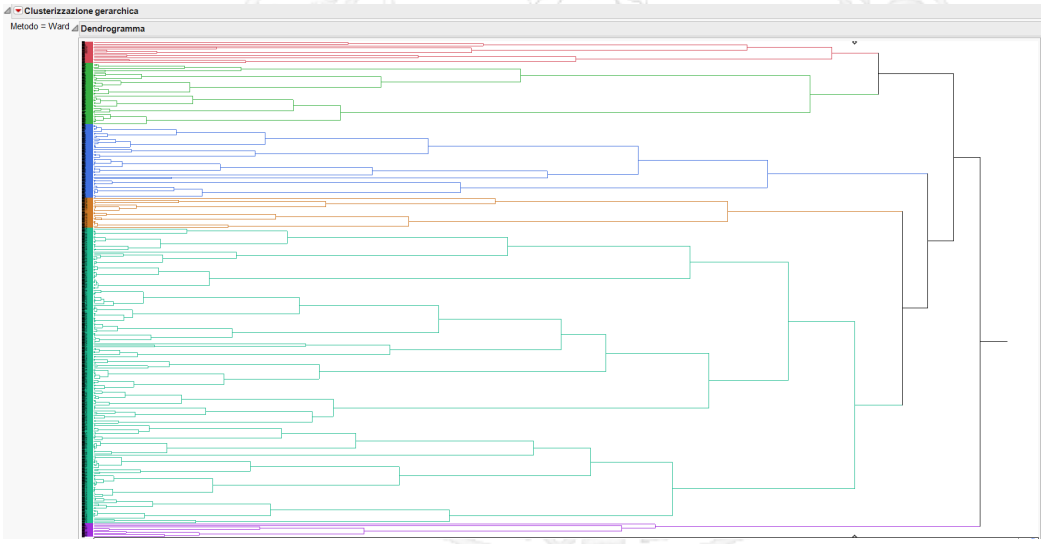
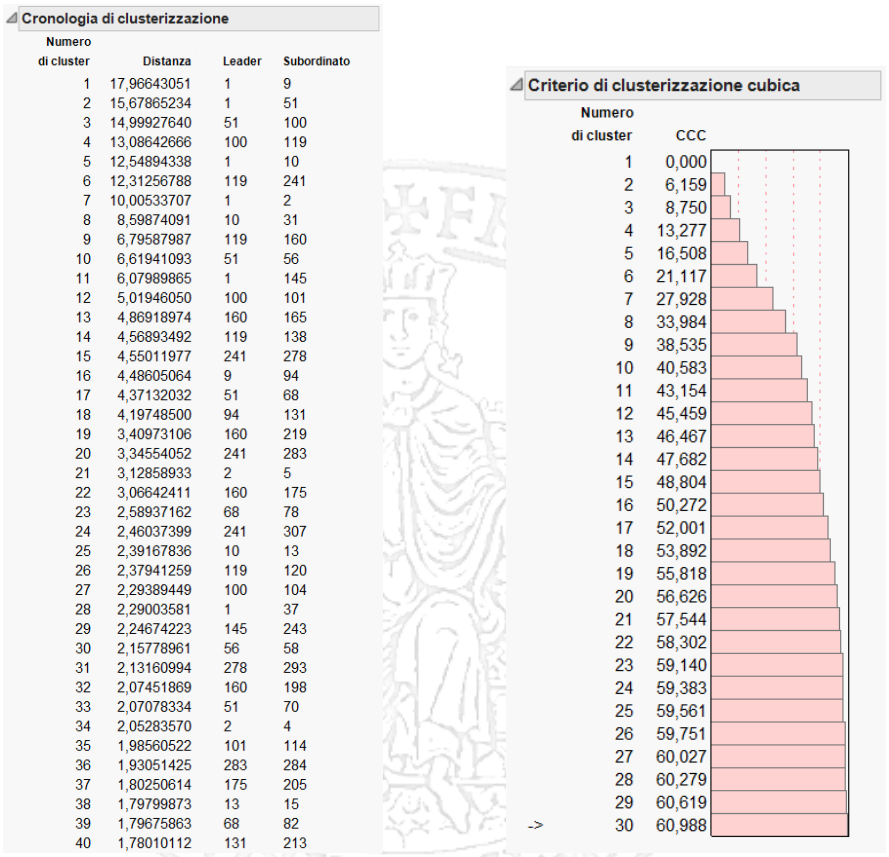


Figura 1.5: Dendogramma

Facendo riferimento alla Figura 1.5, si può scegliere il numero di cluster posizionandosi nel ginocchio della curva rappresentante le distanze tra cluster.

In maniera analoga si può scegliere il numero di cluster utilizzando il criterio clusterizzazione cubica, riportato il Figura 1.6, scegliendo il numero cluster utilizzando la regola del massimo salto.



Cronologia di clusterizzazione

Numero di cluster	Distanza	Leader	Subordinato
1	17,96643051	1	9
2	15,67865234	1	51
3	14,99927640	51	100
4	13,08642666	100	119
5	12,54894338	1	10
6	12,31256788	119	241
7	10,00533707	1	2
8	8,59874091	10	31
9	6,79587987	119	160
10	6,61941093	51	56
11	6,07989865	1	145
12	5,01946050	100	101
13	4,86918974	160	165
14	4,56893492	119	138
15	4,55011977	241	278
16	4,48605064	9	94
17	4,37132032	51	68
18	4,19748500	94	131
19	3,40973106	160	219
20	3,34554052	241	283
21	3,12858933	2	5
22	3,06642411	160	175
23	2,58937162	68	78
24	2,46037399	241	307
25	2,39167836	10	13
26	2,37941259	119	120
27	2,29389449	100	104
28	2,29003581	1	37
29	2,24674223	145	243
30	2,15778961	56	58
31	2,13160994	278	293
32	2,07451869	160	198
33	2,07078334	51	70
34	2,05283570	2	4
35	1,98560522	101	114
36	1,93051425	283	284
37	1,80250614	175	205
38	1,79799873	13	15
39	1,79675863	68	82
40	1,78010112	131	213

Criterio di clusterizzazione cubica

	Numero di cluster	CCC
	1	0,000
	2	6,159
	3	8,750
	4	13,277
	5	16,508
	6	21,117
	7	27,928
	8	33,984
	9	38,535
	10	40,583
	11	43,154
	12	45,459
	13	46,467
	14	47,682
	15	48,804
	16	50,272
	17	52,001
	18	53,892
	19	55,818
	20	56,626
	21	57,544
	22	58,302
	23	59,140
	24	59,383
	25	59,561
	26	59,751
	27	60,027
	28	60,279
	29	60,619
->	30	60,988

Figura 1.6: Gerarchia clustering e Criterio di Clasterizzazione Cubica

Sulla base dei criteri sopra descritti si è scelto di considerare 6 cluster, per la generazione del workload sitnetico si è scelto di estrarre randomicamente un esperimento da ogni cluster.

1.3 Conclusioni

Dagli step descritti abbiamo ottenuto un Workload sintetico, ma non si è detto quanta varianza si è conservato.
Per calcolare quanta varianza abbiamo conservato bisogna calcolare quanta ne abbiamo conservato in ogni ogni step.
Per la PCA la varianza conservata è 95,702%, valore ottenuto da JMP.
Per il clustering non si può utlizzare la varianza ma bisogna utilizzare la devianza poichè essa è indipendente dal grado di libertà dei cluster i quali hanno diverse dimensioni.
Per il calcolo della percentuale di devianza conservata utilizzando il clustering si è calcolata la devianza del workload sottoposto a PCA, in quanto il clustering è stato effettuato successivamente.
La devianza conservata è calcolata come la somma della devianza inter-cluster e intra-cluster, poichè utilizzando il clustering si perde varianza inter-cluster e scegliendo un campione per ogni cluster si perde varianza infra-cluster.
Nella seguente figure è riportata la devianza del workload sottoposto a PCA.

	Componenti Principali	Varianza	Devianza	Devianza Totale PCA
1	Principale1	17,311783969	5591,7062219	7727,93249
2	Principale2	2,297449611	742,07622436	7727,93249
3	Principale3	1,5278624623	493,49957531	7727,93249
4	Principale4	1,0693918983	345,41358316	7727,93249
5	Principale5	0,8996989129	290,60274886	7727,93249
6	Principale6	0,8193007317	264,63413635	7727,93249

Figura 1.7: Devianza PCA

Nella seguente figura è riportato la devianza inter-cluster.

	Componenti Principali	Varianza	Devianza	Devianza Inter-Cluster
1	Principale1	15,651572551	78,257862754	148,03870631
2	Principale2	2,736682497	13,683412485	148,03870631
3	Principale3	3,2323921252	16,161960626	148,03870631
4	Principale4	3,3767072631	16,883536315	148,03870631
5	Principale5	3,7334595869	18,667297935	148,03870631
6	Principale6	0,8769272381	4,3846361905	148,03870631

Figura 1.8: Devianza Inter-Cluster

Nella seguente figura è riportato la devianza intra-cluster.

	Componenti Principali	Cluster	Numero Campioni	Media	Dev std	Devianza	Devianza Intra-Clustering
1	Principale1	1	14	-4,3329	6,22222	503,30906659	2497,1490806
2	Principale1	2	40	-6,6239	0,93049	33,766838222	2497,1490806
3	Principale1	3	48	-3,4088	0,92830	40,501683553	2497,1490806
4	Principale1	4	20	-2,2316	0,45524	3,9376823457	2497,1490806
5	Principale1	5	193	2,7262	1,96787	743,52427803	2497,1490806
6	Principale1	6	9	0,8568	4,33550	150,37244486	2497,1490806
7	Principale2	1	14	4,1876	2,53757	83,710244157	2497,1490806
8	Principale2	2	40	0,4480	0,36416	5,1718726495	2497,1490806
9	Principale2	3	48	-0,8395	0,57131	15,340448145	2497,1490806
10	Principale2	4	20	-0,9979	0,69182	9,0936399578	2497,1490806
11	Principale2	5	193	-0,1042	1,25552	302,65562997	2497,1490806
12	Principale2	6	9	0,4250	1,37400	15,103115781	2497,1490806
13	Principale3	1	14	-2,0084	1,36899	24,363561268	2497,1490806
14	Principale3	2	40	1,2896	0,97029	36,716976198	2497,1490806
15	Principale3	3	48	-1,2750	0,80981	30,821854706	2497,1490806
16	Principale3	4	20	1,0100	1,02297	19,882904369	2497,1490806
17	Principale3	5	193	0,1332	0,80183	123,44297203	2497,1490806
18	Principale3	6	9	-0,9078	1,80319	26,012056251	2497,1490806
19	Principale4	1	14	0,0643	1,03381	13,894000981	2497,1490806
20	Principale4	2	40	-0,3163	0,83726	27,339164793	2497,1490806
21	Principale4	3	48	-0,4771	0,45321	9,6536684042	2497,1490806
22	Principale4	4	20	1,6184	0,43354	3,5711476222	2497,1490806
23	Principale4	5	193	-0,2082	0,23279	10,405070842	2497,1490806
24	Principale4	6	9	4,7192	0,73995	4,3802351577	2497,1490806
25	Principale5	1	14	-0,6070	1,10438	15,855638844	2497,1490806
26	Principale5	2	40	0,5412	0,84394	27,776899818	2497,1490806
27	Principale5	3	48	-0,0966	0,66285	20,650089608	2497,1490806
28	Principale5	4	20	-1,9106	0,66692	8,4507570869	2497,1490806
29	Principale5	5	193	-0,0048	0,30437	17,787283949	2497,1490806
30	Principale5	6	9	3,4030	0,83086	5,5226840303	2497,1490806
31	Principale6	1	14	0,8761	1,87580	45,742293625	2497,1490806
32	Principale6	2	40	0,6280	0,47087	8,6471598604	2497,1490806
33	Principale6	3	48	-1,1806	0,73787	25,589281515	2497,1490806
34	Principale6	4	20	0,2413	0,71853	9,8093238385	2497,1490806
35	Principale6	5	193	0,1051	0,61520	72,666760506	2497,1490806
36	Principale6	6	9	-0,6477	0,45831	1,6803510185	2497,1490806

Figura 1.9: Devianza Intra-Cluster

Quindi per il calcolo della percentuale di devianza conservata utilizzando la tecnica di clasterizzazione si utilizzata la seguente formula:

$$1 - \frac{devianza_{inter\_cluster} + devianza_{intra\_cluster}}{devianza_{pca}}$$

In tabella è riportato la percentuale di devianza persa/conservata utilizzando il clustering.

Devianza PCA	7727,93249
Devianza Clustering(intra-cluster+inter-cluster)	2645,1867
Percentuale devianza persa clusterizzando	34,22%
Percentuale devianza conservata clusterizzando	65,78%
Significatività PCA	95,70%

Nella seguente figure è riportata la percentuale di varianza persa/conservata del workload sintetico.

	Conservata	**Perse**
Workload Reale	100,00%	0,00%
PCA	95,70%	4,30%
Clustering	62,95%	37,05%

In conclusione utilizzando 6 componenti principali e 6 cluster si perde il 37,05% di significatività rispetto al workload reale.