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 insignificativi;
- ***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.

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.
Facendo riferimento alla figura si può osservare che la prima componente del workload conserva il x% della varianza, quindi si è scelto di conservare l'x% di varianza saranno considerate le prima 6 componenti principali.
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:***
- ***Principale 2:***
- ***Principale 3:***
- ***Principale 4:***
- ***Principale 5:***
- ***Principale 6:***

1.2.3 Clustering

In questa fase è stato effettuate un operazione di clustering, per poter ridurre il numero di esperimenti, sul risultato ottenuto nello step precedente.
La tecnica di clasterizzazione scelta è di tipo gerarchico agglomerativo, in particolare è stata utilizzata la metrica di work per la creazione dei cluster.
In figura è riportato il dendogramma e la gerarchia di clusterizzazione.
Facendo riferimento alla figura precedente, si possono 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.
Sulla base dei criteri sopra descritti si è scelto di considerare 6 cluster, per ridurre il workload si è scelto di estrarre randomicamente un esperimento da ogni cluster.

1.3 Conclusioni

Dagli step descritti abbiamo ottenuto da un Workload reale uno sintetico, ma non abbiamo preservato la varianza.
Per calcolare la varianza quanta varianza abbiamo conservato bisogna calcolare quanta ne abbiamo conservato in ogni ogni step.
Per la PCA la varianza conservata è x%, valore ottenuto da JMP, per il clustering non si può fare un ragionamento basato sulla 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, e poi è stata calcolata la 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.