

Soluzioni Esercitazione Data Mining

21/11/2005

Francesco Folino

Soluzione Esercizio 1

1. Single-Link (Min)

Il valore più piccolo è 0.11 tra p3 e p6. Fondiamo questi in un'unico cluster e ricalcoliamo la matrice di dissimilarità:

	P1	P2	P3UP6	P4	P5
P1	0.00	0.24	0.22	0.37	0.34
P2	0.24	0.00	0.15	0.20	0.14
P3UP6	0.22	0.15	0.00	0.25	0.28
P4	0.37	0.20	0.15	0.00	0.29
P5	0.34	0.14	0.28	0.29	0.00

Essendo l'approccio di tipo Single-link, la prossimità fra cluster viene così calcolata:

$$\text{dist}(P1, P3UP6) = \min(\text{dist}(1,3), \text{dist}(1,6)) = \min(0.22, 0.23) = 0.22$$

$$\text{dist}(P2, P3UP6) = \min(\text{dist}(2,3), \text{dist}(2,6)) = 0.15$$

$$\text{dist}(P3UP6, P4) = \min(\text{dist}(P3, P4), \text{dist}(P6, P4)) = 0.15$$

$$\text{dist}(P3UP6, P5) = \min(\text{dist}(P3, P5), \text{dist}(P6, P5)) = 0.28$$

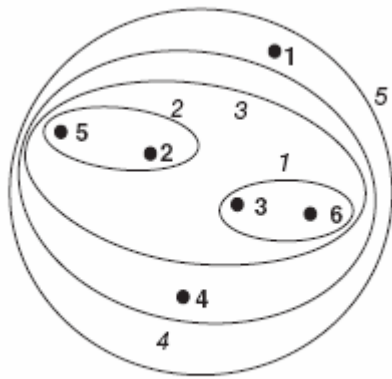
Fondiamo P2 e P5 insieme e riaggiorniamo la matrice di dissimilarità:

	P1	P2UP5	P3UP6	P4
P1	0.00	0.24	0.22	0.37
P2UP5	0.24	0.00	0.15	0.20
P3UP6	0.22	0.15	0.00	0.25
P4	0.37	0.20	0.25	0.00

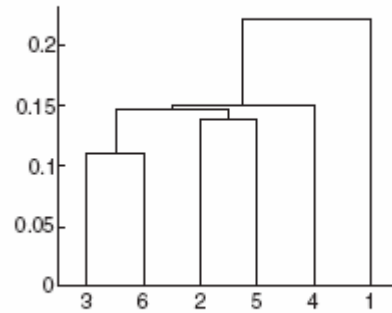
Questa volta fondiamo (P2UP5) e (P3UP6) e ricalcoliamo la matrice di dissimilarità:

	P1	P2UP3UP5UP6	P4
P1	0.00	0.22	0.37
P2UP3UP5UP6	0.22	0.00	0.15
P4	0.37	0.15	0.00

Fondiamo (P2UP3UP5UP6) e P4. Nell'ultimo passo fondiamo P1 e (P2UP3UP5UP6). Avremo così, il seguente dendrogramma:



(a) Single link clustering.



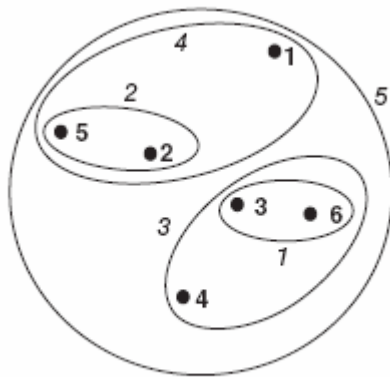
(b) Single link dendrogram.

2. Complete-Link (Max)

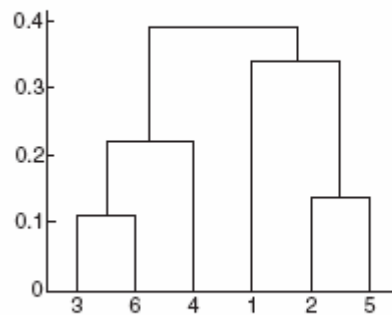
Il metodo è del tutto analogo, l'unica differenza è nel modo in cui viene calcolata la prossimità fra 2 punti. In questo approccio, sappiamo che:

$$dist(C_1, C_2) = \max_{x \in C_1, y \in C_2} dist(x, y)$$

Fonderemo nell'ordine P3 e P6. Poi, P2 e P5. Quindi P4 e (P3UP6). Successivamente P1 e (P2UP5). Infine (P3UP6UP4) e (P1UP2UP5). Avremo così il seguente dendrogramma:



(a) Complete link clustering.



(b) Complete link dendrogram.

Soluzione Esercizio 2

stato iniziale:

C1 = {x1, x2, x3}, centroide c1 = (4.00, 6.33)
C2 = {x4, x5, x6}, centroide c2 = (6.00, 5.67)
C3 = {x7, x8}, centroide_c3 = (2.50, 5.50)

iterazione 1: per ogni punto, cerchiamo il suo centroide più vicino

	C1	C2	C3
x1	5.67	8.33	5.00*
x2	3.33	4.67	1.00*
x3	6.33	3.67*	7.00
x4	2.67*	3.33	5.00
x5	4.33	1.67*	5.00
x6	4.33	1.67*	5.00
x7	7.33	8.67	5.00*
x8	2.67*	5.33	5.00

NOTA: i numeri marcati con * rappresentano le distanze minime.

Riassegniamo i punti ai cluster "più vicini"

C1 = {x4, x8}, centroide c1 = (4.50, 8.50)
C2 = {x3, x5, x6}, centroide c2 = (7.00, 4.33)
C3 = {x1, x2, x7}, centroide_c3 = (1.67, 5.67)

iterazione 2:

=====

	C1	C2	C3
x1	4.00*	10.67	4.67
x2	6.00	5.67	1.00*
x3	8.00	1.33*	8.00
x4	1.00*	5.67	5.67
x5	6.00	0.67*	6.00
x6	6.00	1.33*	6.00
x7	10.00	8.33	4.33*
x8	1.00*	7.67	5.67

Riassegniamo i punti ai cluster "più vicini"

C1 = {x1, x4, x8}, centroide c1 = (3.67, 9.00)
C2 = {x3, x5, x6}, centroide c2 = (7.00, 4.33)
C3 = {x2, x7}, centroide_c3 = {1.50, 3.50}

iterazione 3:

=====

	C1	C2	C3
x1	2.67*	10.67	7.00
x2	5.67	5.67	2.00*
x3	9.33	1.33*	7.00
x4	2.33*	5.67	8.00
x5	7.33	0.67*	7.00
x6	7.33	1.33*	5.00
x7	9.67	8.33	2.00*
x8	0.33*	7.67	8.00

Riassegniamo i punti ai cluster "più vicini"

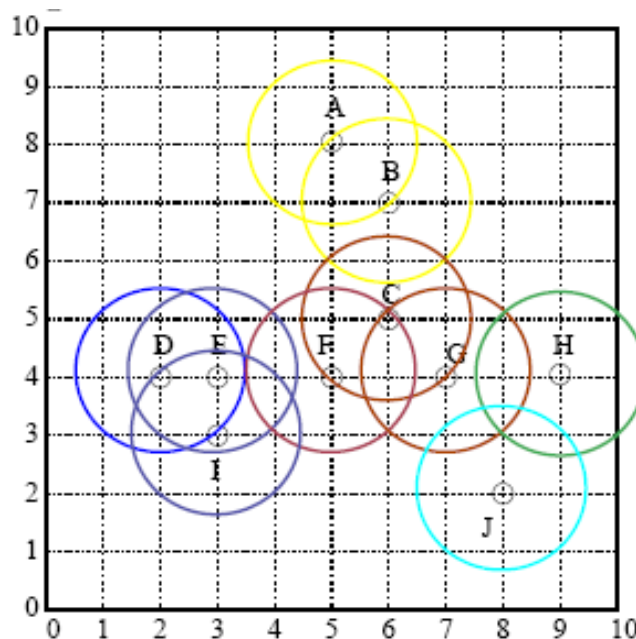
$C1 = \{x1, x4, x8\}$, centroide $c1 = (3.67, 9.00)$

$C2 = \{x3, x5, x6\}$, centroide $c2 = (7.00, 4.33)$

$C3 = \{x2, x7\}$, centroide $c3 = (1.50, 3.50)$

Non ci sono più riassegnamenti, per cui l'algoritmo si ferma!

Soluzione Esercizio 3



I cluster saranno allora:

$C1 = \{A,B\}$, $C2 = \{D,E,I\}$, $C3 = \{C,F,G\}$, $C4=\{H\}$, $C5=\{J\}$ dove H e J sono pertanto outliers.

Soluzione Esercizio 4

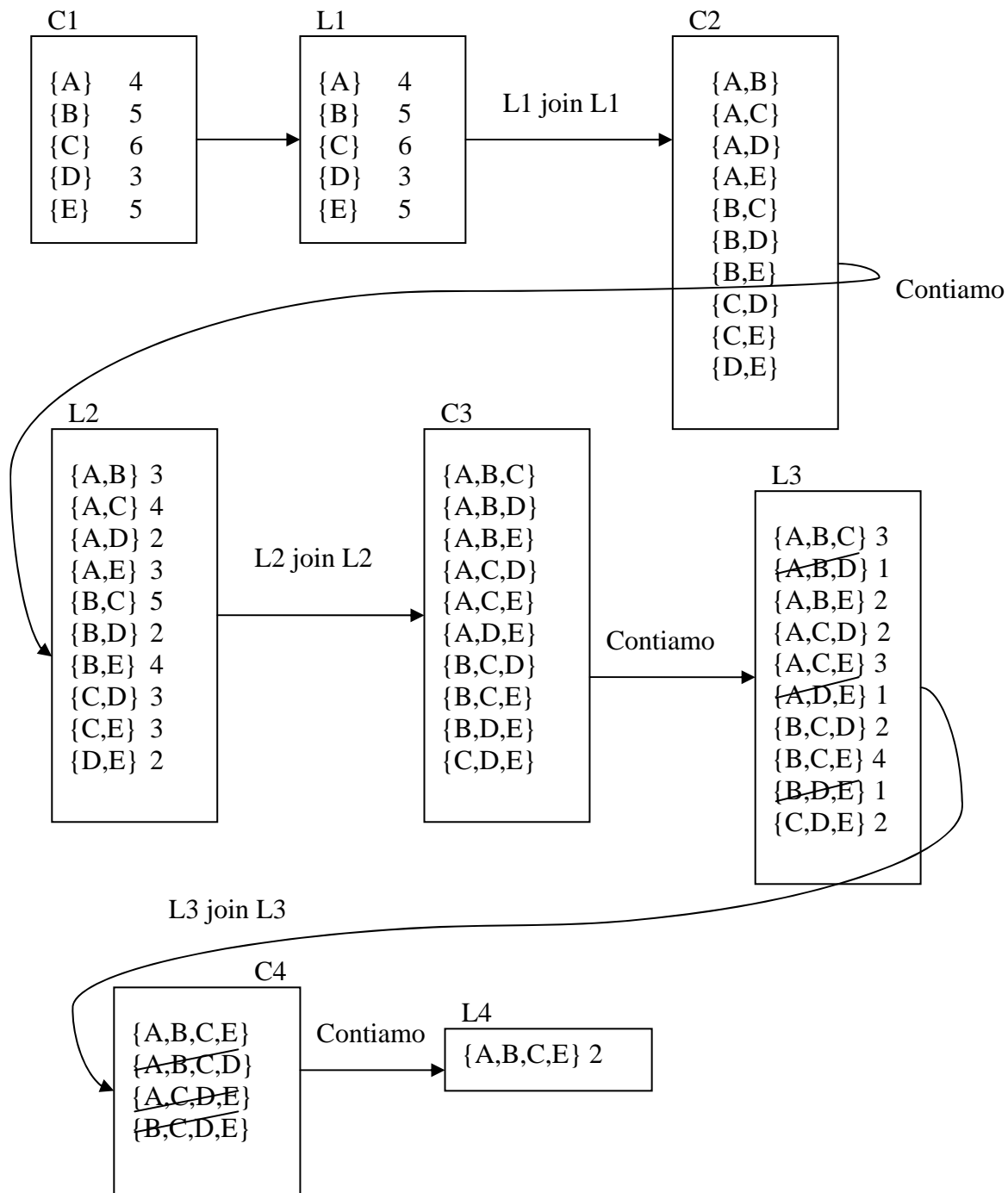
a) Troviamo i Frequent Itemsets con Apriori supponendo che $\text{minsup}=2$. L'intero procedimento per la determinazione degli itemsets frequenti è riportato in basso. È importante osservare come la *proprietà Apriori* sia importante per ridurre il numero dei candidati per cui effettuare il conteggio.

In particolare, consideriamo l'insieme $C4 = \{\{A,B,C,E\}, \{A,B,C,D\}, \{A,C,D,E\}, \{B,C,D,E\}\}$.

$\{A,B,C,D\}$ ha come sotto-itemset $\{A,B,D\}$ che è infrequente, pertanto non può essere frequente e può essere eliminato. Analogamente $\{A,C,D,E\}$ che contiene l'itemset infrequente $\{C,D,E\}$ e $\{B,C,D,E\}$ che contiene l'itemset infrequente $\{B,D,E\}$.

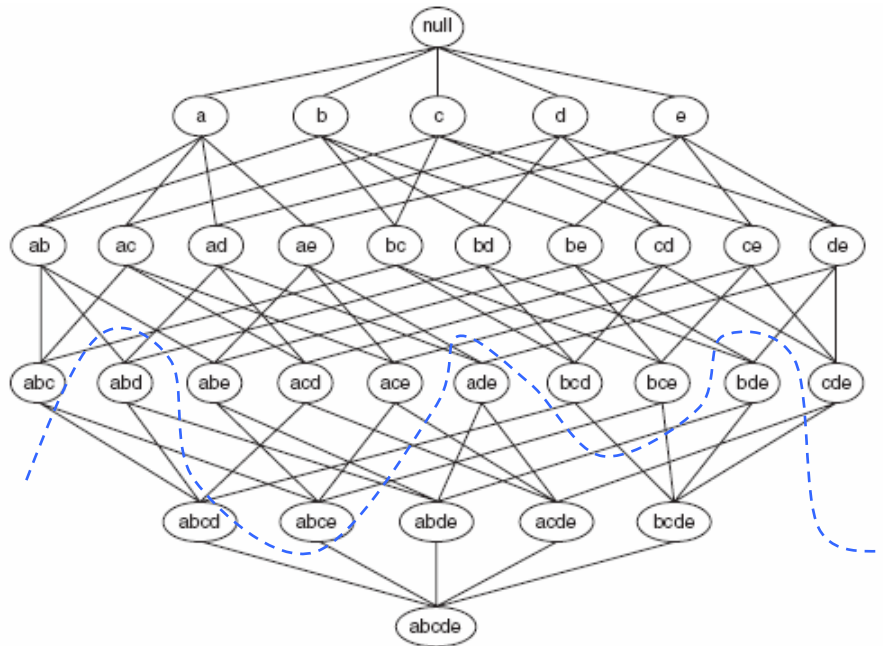
Pertanto otteniamo che l'insieme F è così fatto:

$F = \{A,B,C,D,E, AB, AC, AD, AE, BC, BD, BE, CD, CE, DE, ABC, ABE, ACD, ACE, BCD, BCE, CDE, ABCE\}$



b) Un Maximal Frequent Itemset è un frequent itemset per il quale nessuno dei suoi immediati super-sets è frequente. I MFI danno una rappresentazione compatta degli itemsets frequenti. Infatti gli MFI sono il più piccolo insieme di itemsets dai quali tutti gli itemsets frequenti possono essere derivati.

Al fine di determinare l'insieme M, costruiamoci il reticolo di tutti gli itemsets possibili e tracciamo la curva che separa gli itemsets frequenti (al di sopra della curva) da quelli infrequenti (al di sotto della curva). Saranno MFI quegli itemsets posti in prossimità del bordo e che hanno tutti i loro immediati super-sets infrequenti.



Avremo perciò che:

$$M = \{\{A,C,D\}, \{B,C,D\}, \{C,D,E\}, \{A,B,C,E\}\}.$$

c) Per generare tutte le regole associative che vengono fuori dall'itemset frequente $\{B,C,E\}$, elenchiamo tutti i sottoinsiemi non vuoti di $l = \{B,C,E\} = \{BC, BE, CE, B,C,E\}$.

Per ogni sottoinsieme di s di l , generiamo la regola $s \rightarrow (l-s)$. Pertanto:

Rule	Support	Confidence
$B \rightarrow CE$	4	$P(BCE)/P(B)=4/5$
$C \rightarrow BE$	4	$P(BCE)/P(C)=4/6$
$E \rightarrow BC$	4	$P(BCE)/P(E)=4/5$
$BC \rightarrow E$	4	$P(BCE)/P(BC)=4/5$
$BE \rightarrow C$	4	$P(BCE)/P(BE)=4/4$
$CE \rightarrow B$	4	$P(BCE)/P(CE)=4/5$

Ricordiamo a tal proposito che:

$$\text{conf}(A \rightarrow B) = P(B|A) = \text{supp}(A \cup B) / \text{supp}(A).$$