Università di Pisa	A.A. 2007- 2008

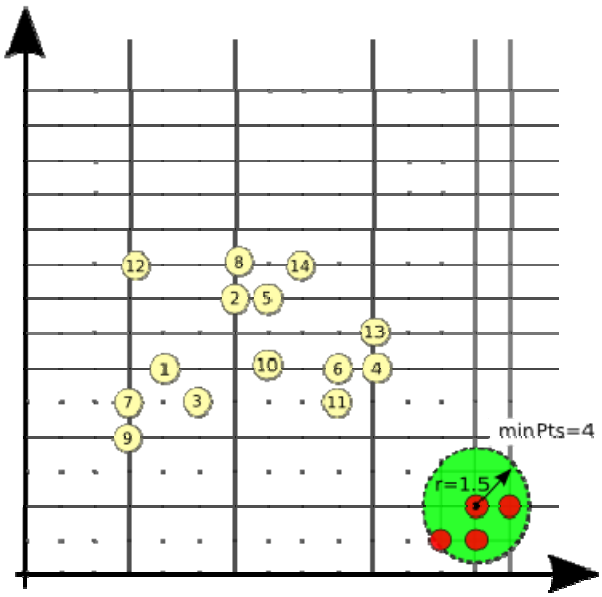
Tecniche Data Mining - *Corsi di Laurea Specialistica in Informatica e Tecnologie Informatiche*

Verifica 1 : Clustering, Pattern Frequenti, Regole Associative

Verifica del 6 Aprile 2009

Esercizio 1 - Clustering (11 punti)

In riferimento al dataset a lato, si determini il numero e la composizione dei cluster che si otterrebbero applicando un algoritmo **density-based** come DBSCAN, con parametri MinPts=3 (incluso il punto centrale) ed epsilon=1.5 (raggio dell'intorno). In particolare, elencare i punti che sono *rumore*, i punti *core* e i punti *border*. Si confronti il clustering ottenuto con quello ottenibile con **K-means**, per un numero K di clusters pari a quello trovato in precedenza.



Esercizio 2 – Clustering (11 punti)

Si esegua l'algoritmo di clustering agglomerativo gerarchico **min-link** in riferimento ad un dataset caratterizzato dalla seguente matrice di **similarità**. Mostrare i risultati ottenuti disegnando il dendrogramma, e proporre uno specifico clustering selezionando un adeguato taglio del dendrogramma ottenuto.

	P1	P2	P3	P4	P5	P6
P1	0.00					
P2	0.24	0.00				
P3	0.22	0.15				
P4	0.90	0.20	0.15	0.00		
P5	0.66	0.14	0.28	0.33	0.00	
P6	0.23	0.25	0.11	0.22	0.50	0.00

Esercizio 3 – Pattern frequenti e regole associative (**11** punti)

Si consideri il seguente insieme di transazioni:

Transazioni	Item Acquistati
1	{A, B, D}
2	{A, C, E}
3	{A, D }
4	{C, E }
5	{B, D, F}
6	{A, C, D, E}
7	{C, D, E}
8	{B, D}
9	{A, C, D, E}
10	{B, F}

a) Eseguire l'algoritmo *Apriori* per l'estrazione di itemset frequenti con min_sup = 30%, mostrando le varie fasi dell'algoritmo.
b) Si determinino le regole associative con confidenza minima dell'80%, che abbiamo almeno due item nella premessa.
c) Esprimere, inoltre: la percentuale di itemset frequenti trovati (rispetto a tutti gli itemset che si potrebbero generare) e la percentuale di pruning dell'algoritmo (definita come la percentuale di itemset che non sono stati considerati candidati perché (i) non sono stati generati durante la fase di generazione (ii) o sono stati tagliati dalla fase di pruning).
d) Elencare quali, fra gli itemset frequenti trovati, sono (i) massimali e/o (ii) chiusi.

Esercizio 4 – Pattern frequenti e regole associative (**15** punti)

a) Qual è il supporto dell'itemset vuoto, denotato {} ?
b) Qual è la confidenza di una regola che ha come premessa l'itemset vuoto?
c) Che relazione c'è fra la confidenza di $X \rightarrow Y$ e $\{\} \rightarrow X,Y$?