Data Mining e Scoperta di Conoscenza

Progetto 14

Clustering di Microarray Data

1. Si estenda Weka con le nuove classi **`weka.core.GeneInstance.java`** e **`weka.core.GeneInstances.java`** che permettano di accedere ad un dataset non solo per riga ma anche per colonna. Nel dettaglio, la seconda classe deve implementare almeno il metodo getRow() (che restituisce una **`GeneInstance`** rappresentante una colonna del dataset).
2. Si implementino le seguenti varianti algoritmiche:
   a. L'algoritmo **`weka.clusters.GeneClusterKMeans`**
   b. L'algoritmo **`weka.clusters.GeneClustersEM`**

   che clusterizzino i dati per colonne, utilizzando le classi definite al punto 1.
3. Si applichino gli algoritmi di cui al punto 2 sui dati genetici contenuti nella repository **microarray_project_data.zip** forniti con il progetto.
   a. Si valuti l'accuratezza degli algoritmi implementati confrontando i risultati con la classificazione predefinita sui dati, utilizzando le opportune metriche per la valutazione.
   b. Si combini la clusterizzazione per colonne con quella per righe (applicata utilizzando tecniche tradizionali) e si discuta come varia l'accuratezza predittiva. Nel dettaglio, si chiede di clusterizzare il dataset per righe e conseguentemente di clusterizzare ogni gruppo scoperto per colonna e valutare l'accuratezza dei cluster scoperti.

NOTE PER L'ESECUZIONE DEL PROGETTO

1. Scrivi un rapporto di circa 10 pagine in cui
   a. Descrivi analiticamente l'algoritmo che hai implementato.
   b. Commenti le parti essenziali del codice Java che hai scritto, e metti in un'appendice l'intero codice
   c. commenti e illustri graficamente e quantitativamente gli esperimenti effettuati.
2. Prepara delle slides Powerpoint (non più di 10 slides) in cui riassumi gli esiti del progetto