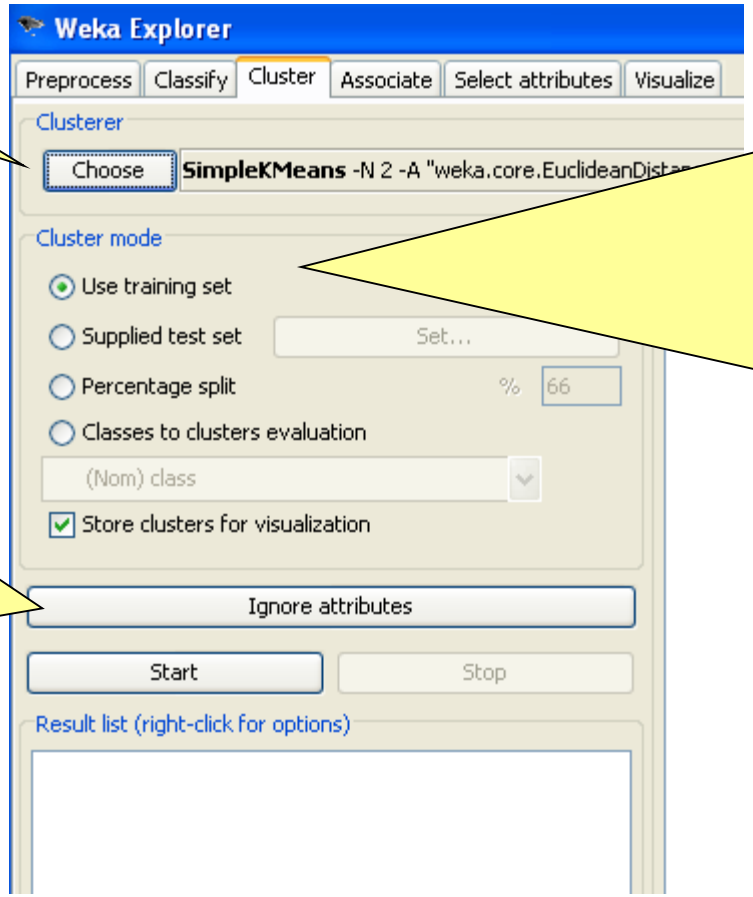
Clustering con Weka
Testo degli esercizi

Prof. Matteo Golfarelli
Alma Mater Studiorum - Università di Bologna

L'interfaccia

Algoritmo utilizzato per il clustering



E' possibile escludere un sottoinsieme degli attributi dal calcolo delle distanze

Modalità di verifica dei risultati: indica il dataset su cui sono calcolati gli indici statistici che può essere diverso da quello in base al quale sono effettivamente costruiti i cluster (es. centroidi di kMeans)

In alternativa è possibile utilizzare un attributo classe per verificare la corrispondenza tra cluster e classe (se questa è nota)

Il data set Iris

- Il data set Iris modella le caratteristiche di una famiglia di piante
  - ✓ 150 istanze
  - ✓ Nessun dato mancante

Attributo	Descrizione
SepalLength	Lunghezza del sepalo
SepalWidth	Larghezza del sepalo
PetalLength	Lunghezza del petalo
PetalWidth	Larghezza del petalo

Pre-processing

- Gli algoritmi di clustering necessitano di una misura di distanza, nei casi che vedremo la distanza euclidea.

- Nel caso in cui gli attributi coinvolti abbiano range di valore diversi è sempre necessario normalizzare tali range in modo che ognuno di essi abbia la stessa influenza nel calcolo del risultato
  - ✓ Normalizzare gli attributi numerici utilizzando il filtro Unsupervised → Attribute→Normalize

Simple K-means: i parametri

- DisplayStdDev: mostra la deviazione standard delle distanze dei singoli punti rispetto al centro del cluster. La misura è riportata separatamente per ogni attributo
  - ✓ Minore la StdDev maggiore la coesione del cluster rispetto all'attributo.
  - ✓ Permette di scegliere quali attributi utilizzare nel calcolo della similarità.
- Distance function: funzione distanza utilizzata nel calcolo
- MaxIteration: numero massimo di iterazioni per ottenere la convergenza
- NumCluster:valore di k
- Seed: valore random per la scelta dei centroidi iniziali
  - ✓ Cambiandolo cambia il loro posizionamento iniziale

Simple K-means: i risultati

- Eseguire l'algoritmo ponendo DisplayStdDev=true e NumCluster=3

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 6.9981140048267605
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute      Full Data         0          1          2
                 (150)        (61)       (50)       (39)
=============================================================
sepallength        0.4287     0.4413     0.1961     0.7073
                  +/-0.23  +/-0.1246  +/-0.0979  +/-0.1396

sepalwidth         0.4392     0.3074     0.5908     0.4509
                 +/-0.1807  +/-0.1222  +/-0.1588  +/-0.1166

petallength        0.4676     0.5757     0.0786      0.797
                 +/-0.2991  +/-0.0893  +/-0.0294   +/-0.088

petalwidth         0.4578     0.5492       0.06     0.8248
                  +/-0.318  +/-0.1135  +/-0.0447  +/-0.1171



Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

#iterazioni per la convergenza

SSE media per i punti dei cluster

Posizione del centroide per il cluster 2 sulla coordinata sepallength

DevStd dei punti del cluster 2 sulla coordinata sepalwidth rispetto alla coordinata del centroide

Dati per il centroide del clustering

Dimensione dei cluster

Simple K-means: i risultati

- Rieseguire l'algoritmo selezionando Classes to cluster evaluation

```
Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 47  0  3 | Iris-versicolor
 14  0 36 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :       17.0      11.3333 %
```

Matrice di confusione

Numero e percentuale degli errori commessi in base alla corrispondenza cluster-classi

Corrispondenza tra cluster e classi determinata in base al numero di elementi del cluster che appartengono alla classe

K-means: analisi del risultato

- Visualizzare il risultato del clustering per le diverse coppie di attributi e discutere il risultato in base al posizionamento dei centroidi e alla dispersione dei punti. Come è possibile migliorare il risultato?

Il Data set FoodNutrients

- Contiene le informazioni nutrizionali di 25 alimenti
  - ✓ Caricare il file FoodNutrients.arff

Attributo	Descrizione
EnergyCal	Calorie per 100 gr
ProteinGram	Proteine per 100 gr
FatGram	Grassi per 100gr
CalciumMG	Calcio in milligrammi per 100 gr
IronMG	Ferro in milligrammi per 100gr

- Normalizzare i dati e clusterizzarli utilizzando k-means per valori crescenti di k [2,6]
- Analizzare i risultati facendo ipotesi sul significato delle classi in base alle caratteristiche dei centroide e alle StdDev dei cluster

Il Data set Coordinates

- Contiene le coordinate geografiche di 480 punti
  - ✓ Caricare il file Coordinates.arff
- Classificare i dati utilizzando k-means con un numero di cluster compreso tra 2 e 6
  - ✓ Come varia SSE?
  - ✓ A partire da quale valore di k SSE si stabilizza?
  - ✓ K-means è in grado di catturare i cluster naturali?
    - Perche?

Coordinates con DBSCAN

- Valutare il risultato della classificazione con DBSCAN
- Identificare i corretti valori per epsilon e minpoints