

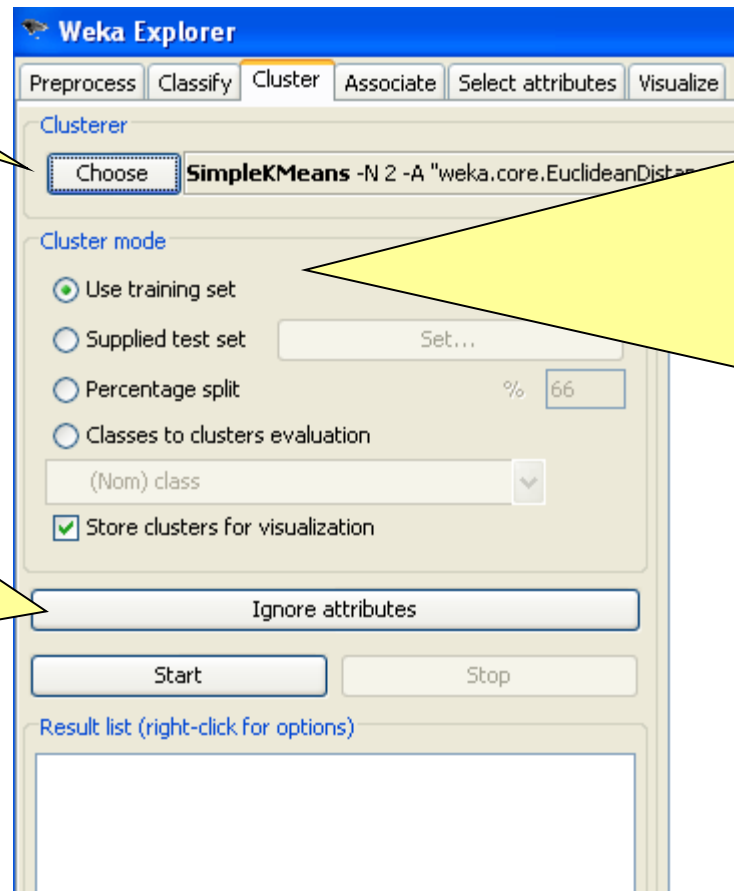
## Data Mining

### **Esercizi di clustering - Soluzioni**

# 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
<b>SepalLength</b>	Lunghezza del sepalo
<b>SepalWidth</b>	Larghezza del sepalo
<b>PetalLength</b>	Lunghezza del petalo
<b>PetalWidth</b>	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:

Attribute	Full Data (150)	Cluster# 0 (61)	1 (50)	2 (39)
sepal.length	0.4287 +/-0.23	0.4413 +/-0.1246	0.1961 +/-0.0979	0.7073 +/-0.1396
sepal.width	0.4392 +/-0.1807	0.3074 +/-0.1222	0.5908 +/-0.1588	0.4509 +/-0.1166
petal.length	0.4676 +/-0.2991	0.5757 +/-0.0893	0.0786 +/-0.0294	0.797 +/-0.088
petal.width	0.4578 +/-0.318	0.5492 +/-0.1135	0.06 +/-0.0447	0.8248 +/-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 sepal.length

DevStd dei punti del cluster 2 sulla coordinata  
sepal.width 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
```

Matrice di confusione

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

```
Incorrectly clustered instances :
```

```
17.0 11.3333 %
```

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?
- Risulta evidente che:
  - ✓ Il cluster 1 è meglio separato dagli altri visto il posizionamento dei suoi centroidi e i relativi valori di dispersione
  - ✓ I cluster 0 e 2 risultano poco separati sugli attributi sepalwidth e sepalwidth
    - $0.4413 + 0.1246 = 0.5659 \approx 0.7073 - 0.1396 = 0.5677$
    - $0.3074 + 0.1222 = 0.4296 \approx 0.4509 / 0.1166 = 0.3353$
- Per migliorare il risultato si possono eliminare gli attributi scarsamente informativi
  - ✓ Verificare l'efficacia attivando la verifica mediante le classi



# Il Data set FoodNutrients

- Contiene le informazioni nutrizionali di 25 alimenti
  - ✓ Caricare il file [FoodNutrients.arff](#)

Attributo	Descrizione
<b>EnergyCal</b>	Calorie per 100 gr
<b>ProteinGram</b>	Proteine per 100 gr
<b>FatGram</b>	Grassi per 100gr
<b>CalciumMG</b>	Calcio in milligrammi per 100 gr
<b>IronMG</b>	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

# II Data set FoodNutrients

Number of iterations: 2  
 Within cluster sum of squared errors: 5.069321339929419  
 Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Cluster#		
	Full Data (27)	0 (9)	1 (18)
EnergyCal	0.4331 +/-0.2699	0.763 +/-0.1442	0.2681 +/-0.1233
ProteinGram	0.6316 +/-0.2238	0.6316 +/-0.0912	0.6316 +/-0.2696
FatGram	0.3285 +/-0.2962	0.6988 +/-0.1701	0.1433 +/-0.108
CalciumMG	0.1076 +/-0.2156	0.0104 +/-0.0018	0.1562 +/-0.2521
IronMG	0.3421 +/-0.2657	0.3576 +/-0.0386	0.3343 +/-0.3272

Number of iterations: 3  
 Within cluster sum of squared errors: 4.077107847192327  
 Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Cluster#			
	Full Data (27)	0 (8)	1 (12)	2 (7)
EnergyCal	0.4331 +/-0.2699	0.7917 +/-0.1236	0.3367 +/-0.102	0.1886 +/-0.1376
ProteinGram	0.6316 +/-0.2238	0.6184 +/-0.0878	0.7982 +/-0.1286	0.3609 +/-0.1908
FatGram	0.3285 +/-0.2962	0.7336 +/-0.1438	0.1908 +/-0.125	0.1015 +/-0.1035
CalciumMG	0.1076 +/-0.2156	0.0104 +/-0.002	0.1192 +/-0.2862	0.1989 +/-0.1691
IronMG	0.3421 +/-0.2657	0.3523 +/-0.0376	0.3379 +/-0.2607	0.3377 +/-0.4237

- C0 è ben caratterizzato per valori elevati di EnergyCal e FatGram
- Nella soluzione a 3 cluster il C0 rimane invariato mentre la caratterizzazione tra C1 e C2 è rilevante solo per ProteinGram
- In entrambe le soluzioni IronMG è poco caratterizzante

# II Data set FoodNutrients

```
Number of iterations: 3
Within cluster sum of squared errors: 3.229030897655616
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute      Full Data      Cluster#
                (27)      0      1      2      3
                (8)      (11)      (7)      (1)
=====
EnergyCal      0.4331      0.7917      0.3345      0.1886      0.36
                +/-0.2699 +/-0.1236 +/-0.1067 +/-0.1376 +/-0
ProteinGram    0.6316      0.6184      0.799      0.3609      0.7895
                +/-0.2238 +/-0.0878 +/-0.1348 +/-0.1908 +/-0
FatGram        0.3285      0.7336      0.189      0.1015      0.2105
                +/-0.2962 +/-0.1438 +/-0.131 +/-0.1035 +/-0
CalciumMG      0.1076      0.0104      0.0392      0.1989      1
                +/-0.2156 +/-0.002 +/-0.0739 +/-0.1691 +/-0
IronMG         0.3421      0.3523      0.3355      0.3377      0.3636
                +/-0.2657 +/-0.0376 +/-0.2733 +/-0.4237 +/-0
```

- L'aggiunta di C3 permette di caratterizzare meglio la differenza tra C1 e C2 in termini di CalciumMG
- C3 è composto da un solo elemento
- IronMG rimane poco caratterizzante

# II Data set FoodNutrients

Number of iterations: 4

Within cluster sum of squared errors: 2.750432407251998

Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Cluster#					
	Full Data (27)	0 (7)	1 (8)	2 (6)	3 (1)	4 (5)
EnergyCal	0.4331 +/-0.2699	0.821 +/-0.0991	0.2883 +/-0.0781	0.1533 +/-0.1108	0.36 +/-0	0.472 +/-0.0831
ProteinGram	0.6316 +/-0.2238	0.609 +/-0.0904	0.8553 +/-0.1043	0.3421 +/-0.2018	0.7895 +/-0	0.6211 +/-0.1012
FatGram	0.3285 +/-0.2962	0.7669 +/-0.1171	0.125 +/-0.0805	0.0746 +/-0.0822	0.2105 +/-0	0.3684 +/-0.093
CalciumMG	0.1076 +/-0.2156	0.0103 +/-0.0021	0.0518 +/-0.0844	0.2279 +/-0.1651	1 +/-0	0.0105 +/-0.0092
IronMG	0.3421 +/-0.2657	0.3481 +/-0.0385	0.3545 +/-0.3115	0.3697 +/-0.4547	0.3636 +/-0	0.2764 +/-0.1462

- C4 raccoglie gli alimenti con valori medi di EnergyCal, ProteinGram, FatGram
- Per quanto riguarda CalciumMG C4 è molto simile a C0
- IronMG rimane poco caratterizzante

# II Data set FoodNutrients

```
Number of iterations: 4
Within cluster sum of squared errors: 1.5257151920333285
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute      Full Data      Cluster#
                (27)      0          1          2          3          4          5
                (27)      (7)        (8)        (2)        (1)        (5)        (4)
=====
EnergyCal      0.4331      0.821      0.2883      0.0333      0.36      0.472      0.2133
                +/-0.2699 +/-0.0991 +/-0.0781 +/-0.0471 +/-0 +/-0.0831 +/-0.073

ProteinGram    0.6316      0.609      0.8553      0.1053      0.7895      0.6211      0.4605
                +/-0.2238 +/-0.0904 +/-0.1043 +/-0.1489 +/-0 +/-0.1012 +/-0.0662

FatGram        0.3285      0.7669      0.125      0          0.2105      0.3684      0.1118
                +/-0.2962 +/-0.1171 +/-0.0805 +/-0 +/-0 +/-0.093 +/-0.0756

CalciumMG      0.1076      0.0103      0.0518      0.2017      1          0.0105      0.241
                +/-0.2156 +/-0.0021 +/-0.0844 +/-0.0156 +/-0 +/-0.0092 +/-0.2113

IronMG         0.3421      0.3481      0.3545      0.9455      0.3636      0.2764      0.0818
                +/-0.2657 +/-0.0385 +/-0.3115 +/-0.0771 +/-0 +/-0.1462 +/-0.1055
```

- Con l'aggiunta del nuovo cluster C0, C1 e C4 rimangono invariati
- Gli elementi di C5 sembrano provenire da C2 che si caratterizza ora per valori bassi per calorie, proteine e grassi e valori alti per il calcio e il ferro e ferro

# FoodNutrients: ricapitolando....

Clust	Caratterizzazione
C0	cibi grassi altamente proteici ed energetici
C1	cibi proteici ma con pochi grassi e calorie
C2	cibi leggeri ma ricchi di calcio
C3	un solo elemento
C4	cibi con apporto medio di grassi proteine e calorie
C5	cibi leggeri in termini di calorie e grassi ma ricchi di proteine, calcio e ferro

■ Verifichiamo caricando il data set FoodNutrientClassified.arff che contiene la classificazione dei cibi in Tipi e Super tipi

✓ Si attivi Classes to cluster evaluation

# FoodNutrients: ricapitolando....

Clust	Caratterizzazione	Super tipo	Tipo
C0	cibi grassi altamente proteici ed energetici	Meat	Pork
C1	cibi proteici ma con pochi grassi e calorie	Meat	Beef
C2	cibi leggeri ma ricchi di calcio	Fish	Clams
C3	un solo elemento	Fish	No class
C4	cibi con apporto medio di grassi proteine e calorie	Meat	Lamb
C5	cibi leggeri in termini di calorie e grassi ma ricchi di proteine, calcio e ferro	Fish	Fish



# II Data set Coordinates

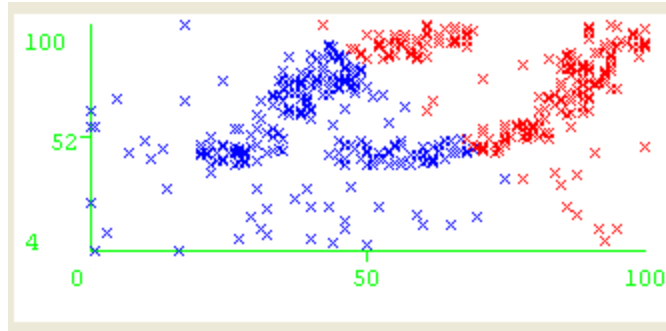
- Contiene le coordinate geografiche di 480 punti
  - ✓ Caricare il file `Coordinates.arff`
- Classificare i dati utilizzando k-means con un numero di cluster compreso tra 2 e 6
  - ✓ Come varia SSE?
  - ✓ A partire da quale valore di k SSE si stabilizza?
  - ✓ K-means è in grado di catturare i cluster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



# Coordinates con K-means

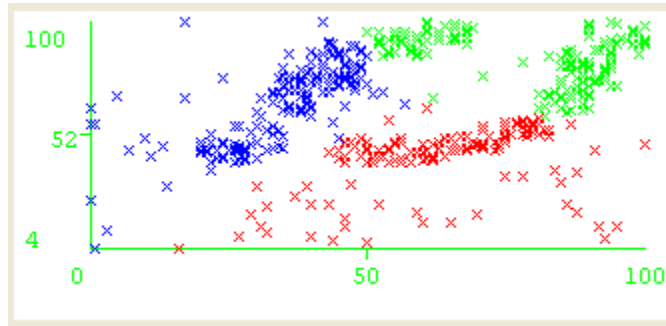
K=2

SSE=29.39



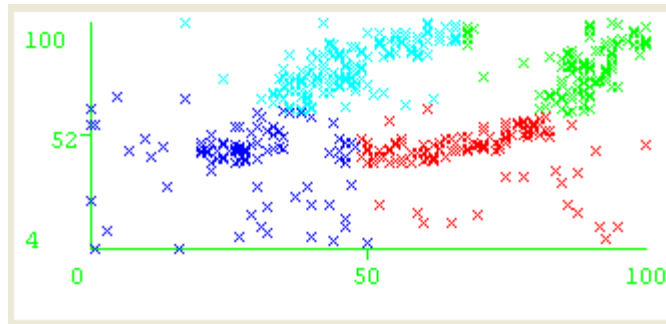
K=3

SSE=19.89



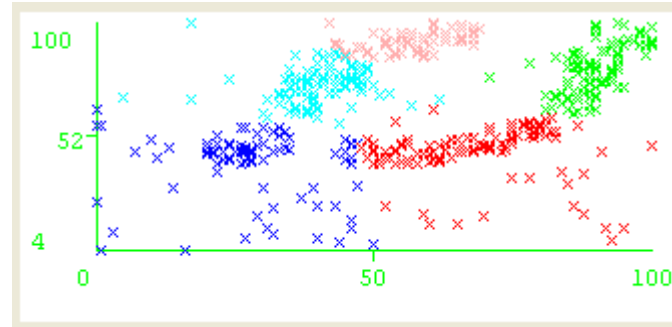
K=4

SSE=12.09

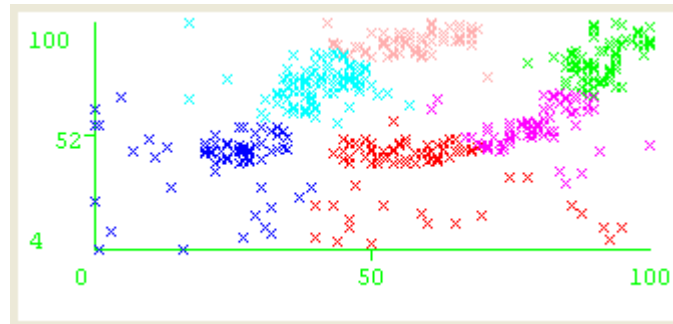


# Coordinates con K-means

K=5  
SSE=9.54



K=6  
SSE=7.87



- SSE si stabilizza con K=5 perché i cluster individuati sono tutte scomposizioni dei singoli cluster naturali
- K-means non è adatto a questo data set poiché la forma allungata dei cluster naturali non può essere catturata

**E' preferibile utilizzare DBSCAN**

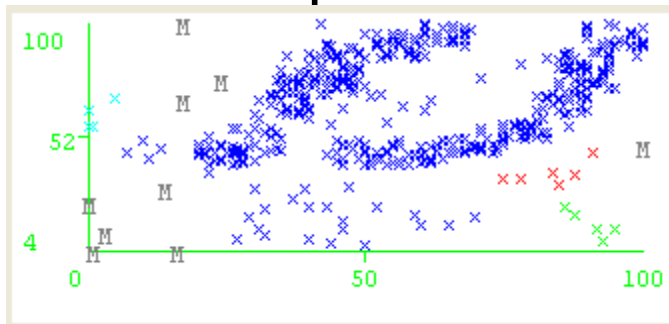
# Coordinates con DBSCAN

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

Eps=0.1

MinPts=4

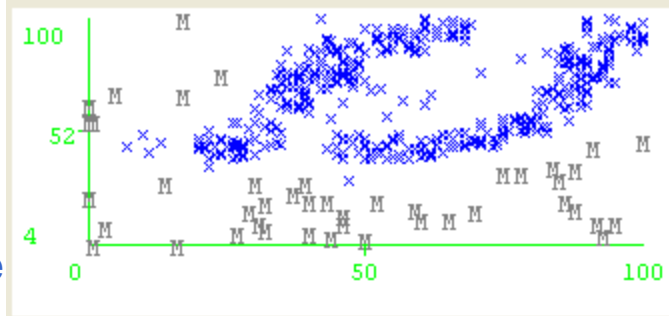
- ✓ I parametri non catturano correttamente il rumore e le separazioni tra i cluster: tutte le zone risultano dense! E' necessario ricercare aree a maggiore densità



Eps=0.1

MinPts=8

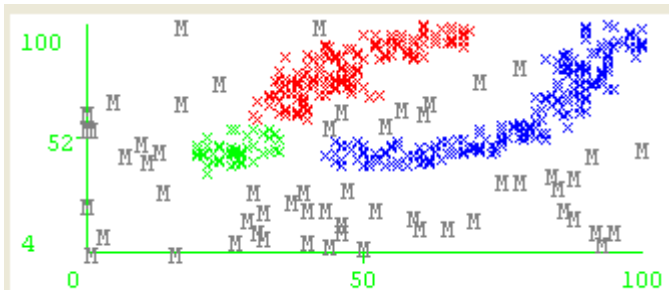
- ✓ Aumentare la densità permette di identificare meglio i punti di rumore ma non consente di differenziare i due cluster naturali



Eps=0.05

MinPts=4

- ✓ Raggio insufficiente

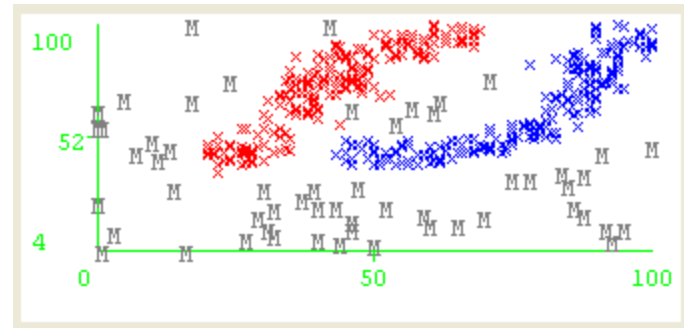


# Coordinates con DBSCAN

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

Eps=0.06

MinPts=8



Eps=0.06

MinPts=3

- ✓ Incorretta individuazione dei punti di rumore

