Assegnamento 2 : Clustering

January 24, 2022

Nicolò Toscani

1 Esercizio 0

1.1 Esercizio 0a

Problema: distinguere i punti di appartenenza a:

- distribuzione gaussiana centrata in (3,3) di colore blu
- distribuzione gaussiana centrata in (7,7) di colore rosso

La coordinata (0,0) é posizionata in basso a sinistra.

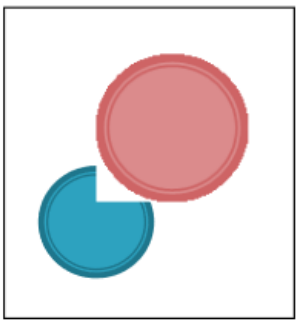


Figure 1: Regione di interesse

Viene caricato il training set *gausstrain.arff* che contiene 58 punti appartenti alla distribuzione gaussiana centrata in (7,7) e 42 punti appartenti alla distribuzione gaussiana centrata in (3,3).

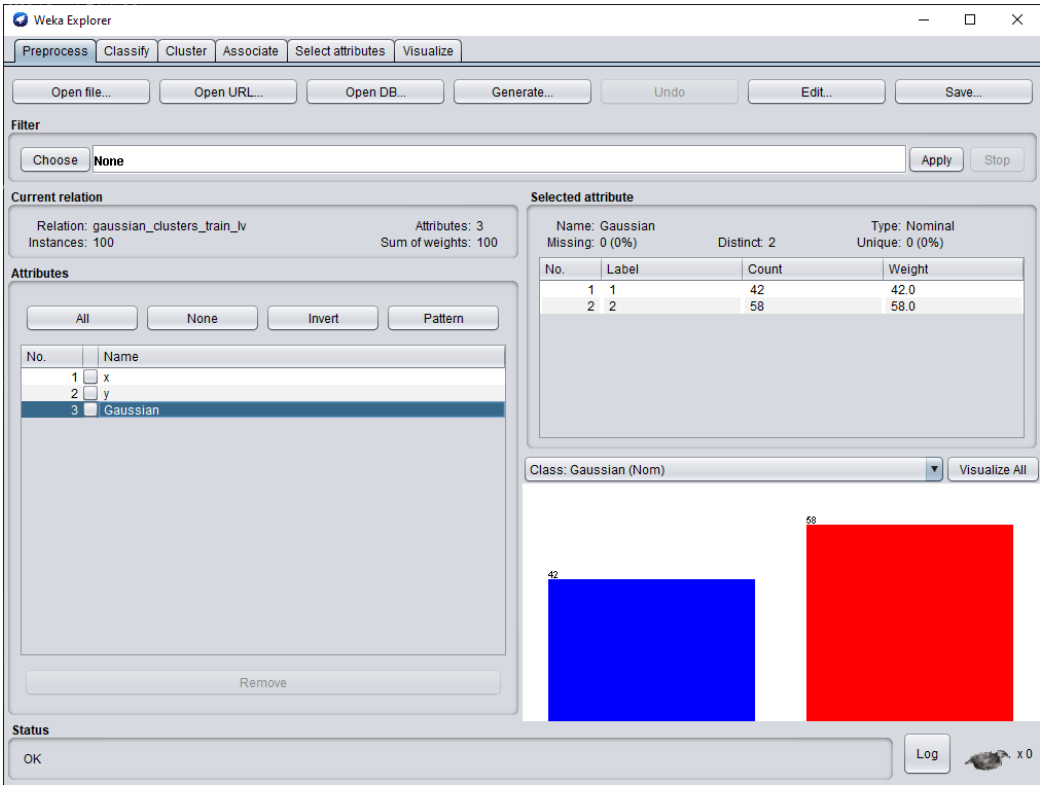


Figure 2: Training set

Questo *training set* é formato da 100 istanze contenente anche l'etichetta di appartenenza alla relativa distribuzione. Questo attributo viene ignorato per eseguire l'algoritmo *K-Means* in modo *non supervisionato*. L'algoritmo eseguito con k (numero di cluster) converge in 7 iterazioni trovando i 2 centroidi dei cluster con le seguenti coordinate:

Cluster 1 (3,3)

- x: 2.7273
- y = 3.0107

Cluster 0 (7,7)

- x: 7.2312
- y = 7.7187

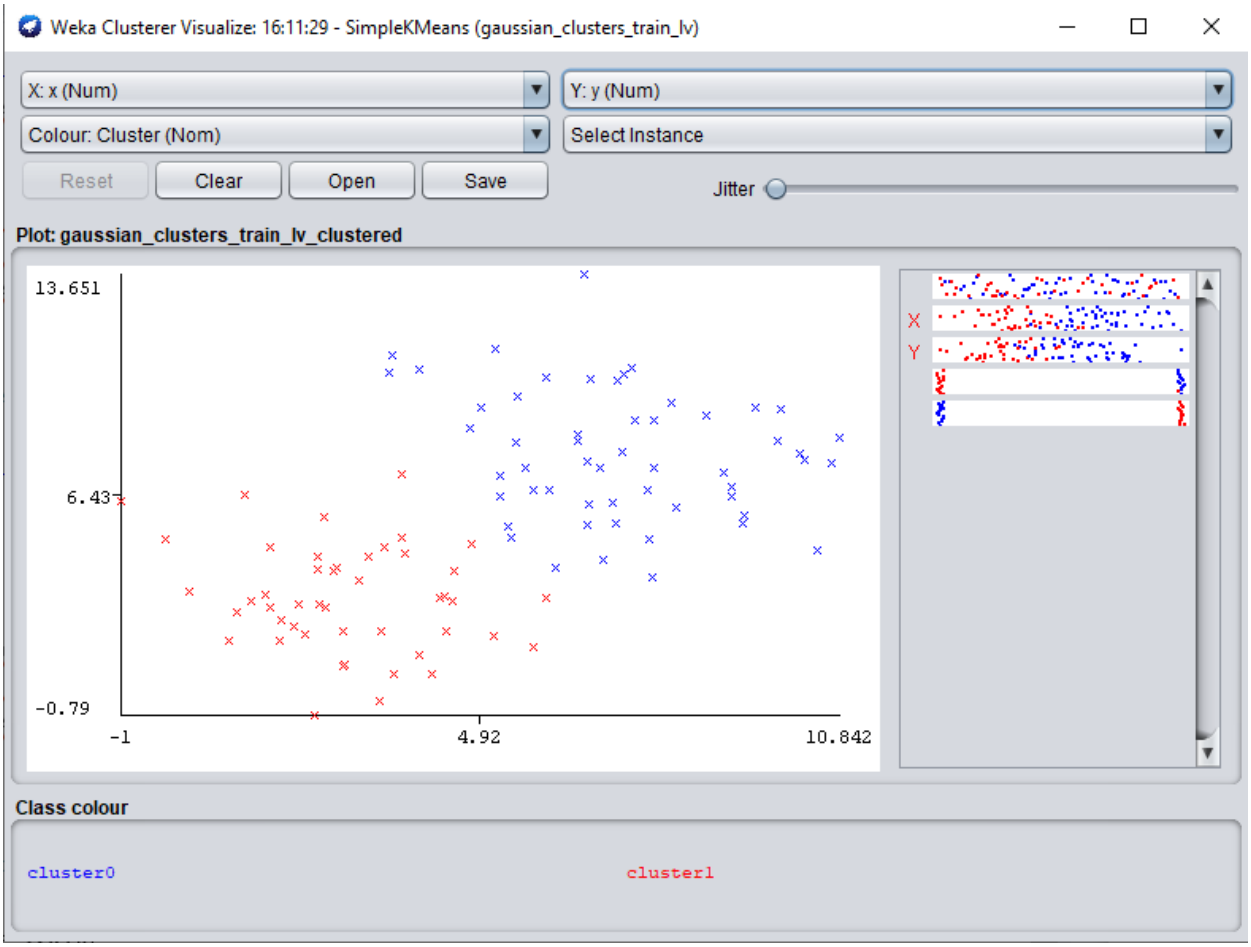


Figure 3: Cluster individuati *unsupervised learning*

1.2 Esercizio 0b

Successivamente si tiene conto dell'etichetta di appartenenza alla distribuzione gaussiana di appartenenza presente nel dataset effetuando cosi un *apprendimento supervisionato*.

Eseguendo l'algoritmo *K-Means* con gli stessi parametri dell'esercizio precedente si nota che abbiamo un 4% di errore di classificazione ovvero 4 punti in ingresso al modello non vengono correttamente classificati.

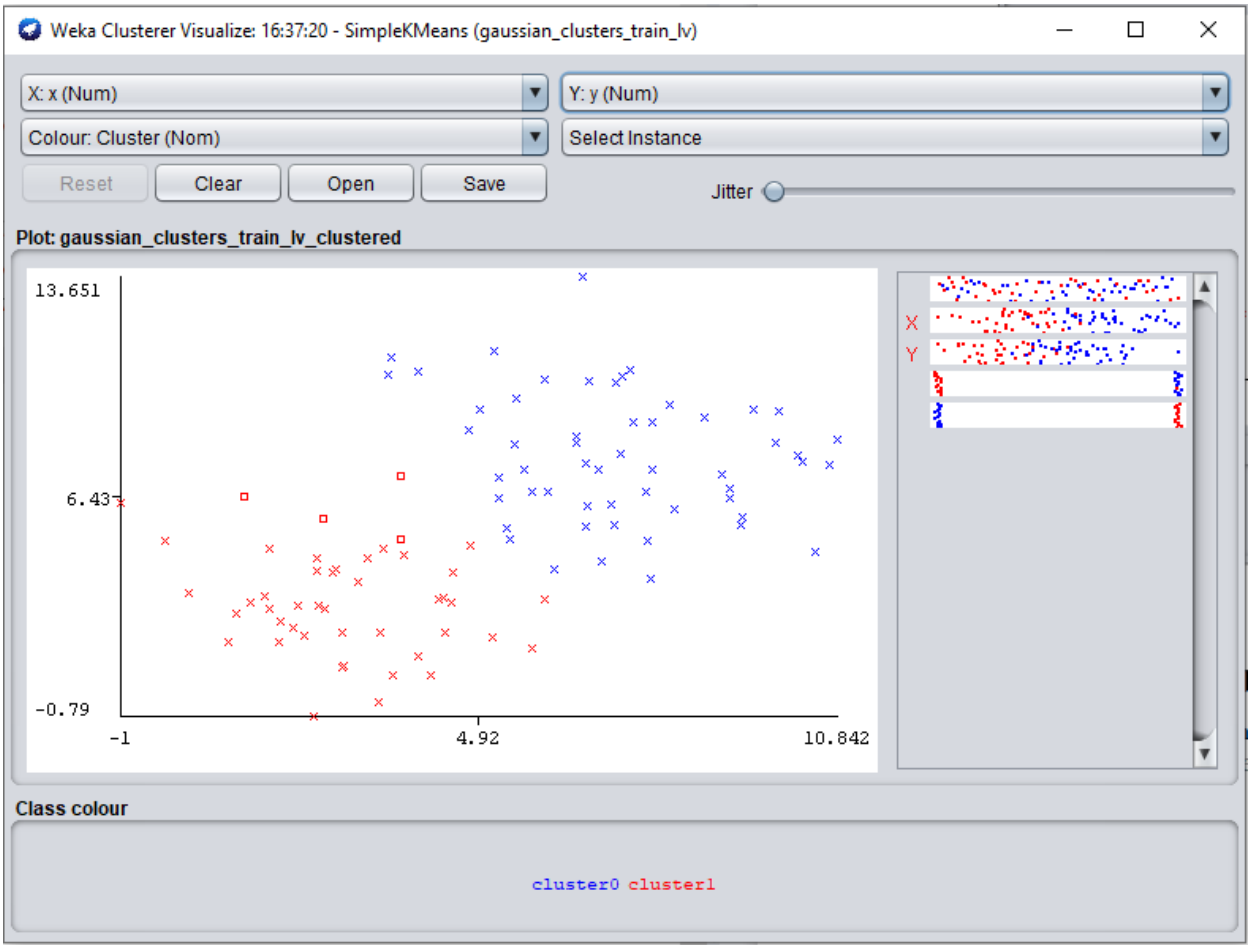


Figure 4: Cluster individuati *supervised learning*

1.3 Esercizio 0c

Per questo esercizio viene utilizzato il file *gausstrainnhv.arff* che contiene punti ad una distanza molto alta dal centro della distribuzione gaussiana.

Figure 5: Dataset gausstrainnhv.arff

Nella figura é possibile osservare che alcuni punti di una distribuzione sono molto vicini al centroide della successiva distribuzione gaussiana. Eseguendo l'algoritmo *K-Means* con gli stessi parametri dell'esercizio precedente si ottengono i seguenti centroidi per i 2 cluster:

Cluster 0 (3,3)

- x: 3.1113
- y = 2.3389

Cluster 1 (7,7)

- x: 7.1559
- y = 8.3404

I nuovi centroidi risultano leggermenti differenti rispetto ai precedenti ma comunque vicini a quelli delle distribuzioni di partenza.
Di seguito si possono notare i punti che vengono assegnati ai cluster non corretti.

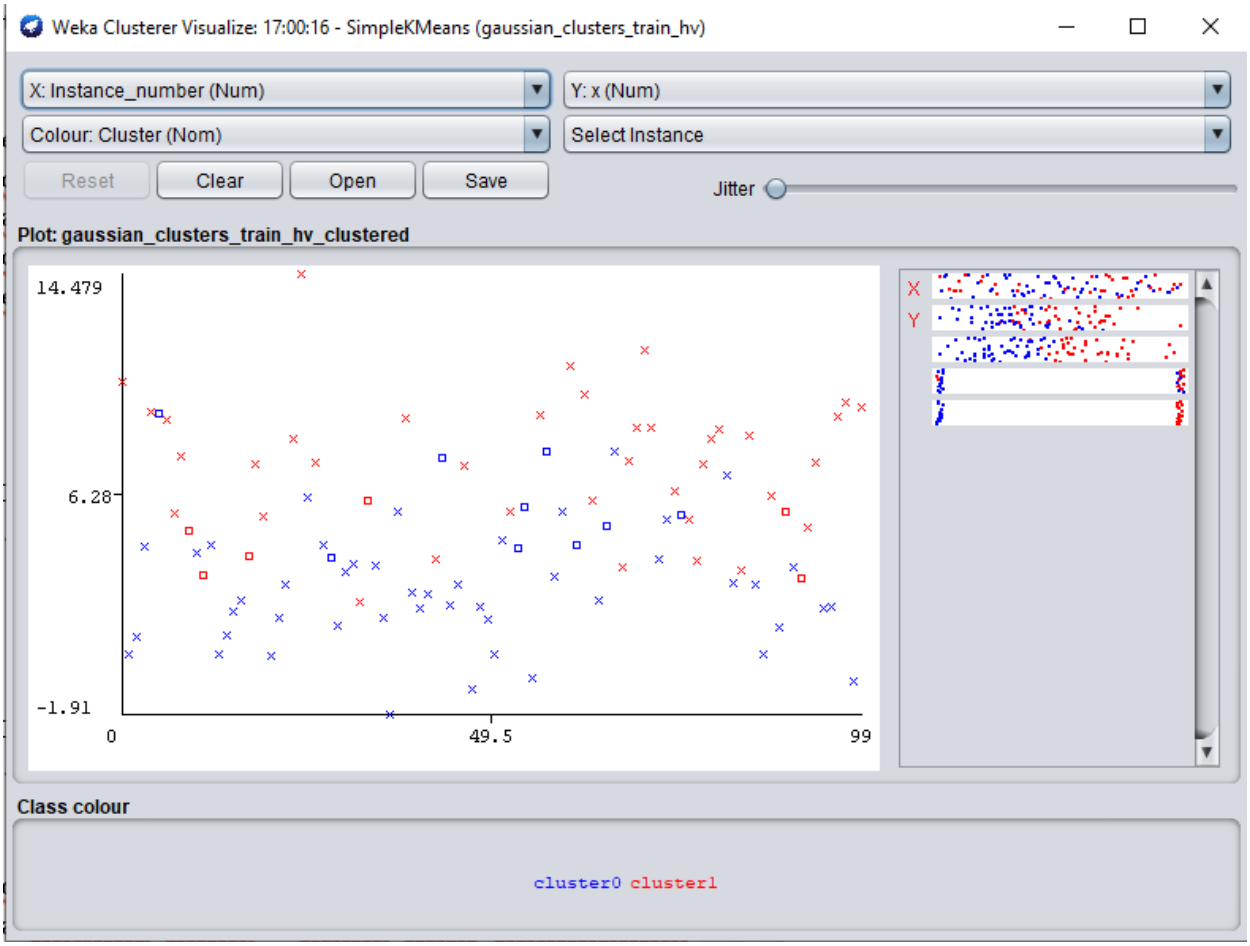


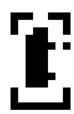
Figure 6: Errore di assegnamento ai cluster

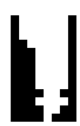
2 Esercizio 1

Viene caricato il file *Bigtest*1_104.*arff* che contiene la rappresentazione binaria delle cifre da 0 a 9 delle cifre delle targhe. L'obbiettivo é quello di riuscire ad assegnare la codifica binaria del carattere digitale alla corretta cifra.
L'immagine in arrivo viene rappresentata come un array di 104 valori binari.
Viene eseguito l'algoritmo *K-Means* utilizzando il *training set* in modalitá *non supervisionata* ed impostando a 10 il numero di cluster desiderati (cifre da 0 a 9).
Otteniamo i centroidi dei 10 cluster trovati rappresentativi delle 10 cifre.

```
0: 0,0,0,1,1,0,0,0,0,1,1,1,1,1,1,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,0,1,1,1,1,1,0,0,0,0,1,1,1,1,1,0,0,1,1,1,0,1,1,1,1,1,1,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0
1: 0,0,1,1,1,1,0,0,1,1,1,1,1,1,1,0,1,1,0,0,0,0,1,0,1,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,1,1,1,1,1,1,0,1,1,1,0,0,1,1,0,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,1,0,0,0,1,1,0,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0
2: 0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,0,0,0,1,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,0,1,1,0,0,0,0,0,1,1,1,1,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0
3: 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,1,0,0,0,0,0,0
4: 0,0,1,1,0,0,0,0,0,1,1,1,1,1,0,0,1,1,1,0,0,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,0,1,1,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0
5: 0,0,0,1,1,0,0,0,0,1,1,1,1,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,1,0,0,1,1,0,1,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,0,1,0,1,1,0,0,0,0,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,0,0,0,0,0,1,0,0,0,0
6: 0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,1,1,1,0,0,1,0,0,1,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,0,1,1,1,0,0,1,1,1,1,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,1,1,0,1,1,1,1,1,1,1,0,0,1,1,1,1,0,0
7: 0,0,0,1,1,0,0,0,0,1,1,1,1,1,1,0,0,1,1,0,0,1,1,0,1,1,0,0,0,1,1,1,1,1,0,0,0,0,1,0,1,1,0,0,0,1,1,1,1,1,0,0,0,1,1,1,1,1,0,0,0,1,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,1,1,1,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0
8: 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,0
9: 0,0,1,1,1,0,0,0,0,1,1,1,1,1,0,0,1,1,0,0,0,0,1,0,1,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,0,1,1,0,0,0,1,1,0,1,1,0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1,1,0,0,1,1,1,0,1,1,1,1,1,1,0,0,0,0,1,1,0,0,0
```

Figure 7: Centroidi





Si puó notare che non tutti i cluster individuati rappresentano correttamente le dieci cifre.
Una possibile causa potrebbe essere la presenza di rumore nei dati.

3 Esercizio 2

Utilizzando sempre il *training set* $Bigtest1\_104.arff$ utilizziamo l'algoritmo *X-Means* in modalitá supervisionata. Questo algoritmo permette di settare il numero minimo e massimo dei centroidi desiderati.
In questo modo si valuta quanti cluster vengono identificati al variare del massimo mantenendo il minimo costante a 10.
Osserviamo che il numero di cluster ottenuti varia al variare del parametro *max*. Per valori di *max* minore o uguale a 20 viene trovato un numero di cluster uguali al valore di *max*. Per valori maggiori di 20 non vengono identificati nuovi cluster ma il valore rimane sempre fisso a 20.

Considerando che solamente un cluster viene associato in automatico a ciascuna classe dobbiamo assegnare i cluster non etichettati alla corretta classe osservando quale classe di istanza contengono piú di frequente nel loro gruppo.
Viene fissato 10 come valore minimo e 20 come valore massimo.

```
Class attribute: class
Classes to Clusters:

   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  <-- assigned to cluster
   0   1   0   0   0   0 270 252   1   2   0   0   2   1   0   5   1  11  10   1 | 0
   3   0   0   1   0   0   0   0   0   2   0   0 254 265   0   0   0   0   0   0 | 1
  18   3   0   0 319 221   1   1   1   1   0   0   0   1   0   0   0   0   0   0 | 2
   0   0   0   1   0   0   1   1 425 257   0   0   2   7   4  28   0   3  10   0 | 3
   0   1   0   2   0   0   0   1   0   0 340 251   6   0   0   0   0   0   0   0 | 4
   0   0 375 217   0   0   0   1   1   1   0   0   1   0   1   1   0  20   0   1 | 5
   0   1   5   5   1   0   2   0   0   0   0   4   0   0   0   0 236 132   0 213 | 6
 249 277   0   0   0   7   0   0   0   0   0   0   0   0   0   0   0   0   0   0 | 7
   0   0   2   1   0   0   2   2   2   1   0   1   0   2   4   2  89 104 395   4 | 8
   0   0   0   0   0   0   1   1   0   0   0   1   0   2 373 291   0   2   3   0 | 9

Cluster  0 <-- No class
Cluster  1 <-- 7
Cluster  2 <-- 5
Cluster  3 <-- No class
Cluster  4 <-- 2
Cluster  5 <-- No class
Cluster  6 <-- 0
Cluster  7 <-- No class
Cluster  8 <-- 3
Cluster  9 <-- No class
Cluster 10 <-- 4
Cluster 11 <-- No class
Cluster 12 <-- No class
Cluster 13 <-- 1
Cluster 14 <-- 9
Cluster 15 <-- No class
Cluster 16 <-- 6
Cluster 17 <-- No class
Cluster 18 <-- 8
Cluster 19 <-- No class
```

Cluster	**Classe**
0	7
1	7
2	5
3	5
4	2
5	2
6	0
7	0
8	3
9	3
10	4
11	4
12	1
13	1
14	9
15	9
16	6
17	6
18	8
19	6

Table 1: Assegnamento Cluster-Classe

Di seguito i valori di accuratezza relativi alla matrice di confusione di ogni classe.

- Classe 0: 0.962
- Classe 1: 1.03
- Classe 2: 0.96
- Classe 3: 0.94
- Classe 4: 0.99
- Classe 5: 0.88

- Classe 6: 1.36
- Classe 7: 1.03
- Classe 8: 0.68
- Classe 9: 1.05

3.1 Esercizio 2b

Utilizando come *training set* $Bigtest1\_104.arff$ e impostando a 10 il valore minimo dei cluster desiderati e a 20 il valore massimo otteniamo la seguente classificazione.
Individuamo ora i centroidi dei diversi cluster trovati e cataloghiamo con le relative etichette di appartenenza.