6 Clustering utenti utilizzatori

Abbiamo applicato un modello di clustering basato su algoritmo di K-Means, attraverso il quale si segmentano gli utenti utilizzatori, sulla base della loro professione, età, genere e numero di richieste di prestiti di libri cartacei effettuati.

La cluster/segmentation analysis è un insieme usate di tecniche per raggruppare oggetti in classi tra loro omogenee, ossia con caratteristiche simili.
Queste tecniche prendono in input un insieme di elementi da dividere in cluster e un numero di cluster.
In output, determinano gli insiemi di elementi che compongono ogni cluster.

Abbiamo utilizzato l'algoritmo di clusterizzazione k-means, che richiede l'indicazione a priori del numero di cluster. I dati da classificare sono attributi con valori reali, nel caso si trattasse di attributi testuali sarebbe necessaria una riconversione del dominio in valori reali.
Questo algoritmo iterativo si basa sul concetto di distanza tra elementi, per ogni cluster si definisce un centroide, ossia un punto (immaginario o reale) al centro di un cluster, e itera 3 passi:

1. Inizializzazione: si definiscono i parametri di input per eseguire l'algoritmo;
2. Assegnazione del cluster: ogni data points viene assegnato al cluster (o centroide) più vicino;
3. Aggiornamento della posizione del centroide: ricalcola il punto esatto del centroide e di conseguenza ne modifica la sua posizione.

Per utilizzare questo algoritmo caricato i dati del nostro database sul software Weka.
Abbiamo sopposto a WEKA il seguente arff file:

```
% 4 attributes
% 13 instances

@RELATION UTENTI_UTILIZZATORI

@ATTRIBUTE Email STRING
@ATTRIBUTE Eta NUMERIC
@ATTRIBUTE Sesso {M, F}
@ATTRIBUTE NumPrenotazioni NUMERIC

@DATA
carla@gmail.com,26,F,0
franco@gmail.com,35,M,2
gino@gmail.com,35,M,2
giovanna@gmail.com,27,F,0
luigi@gmail.com,24,M,0
marco@gmail.com,53,M,2
matteo@gmail.com,23,M,0
mauro@gmail.com,30,M,0
melissa@gmail.com,28,F,1
michele@gmail.com,22,M,2
piero@gmail.com,25,M,1
tiziano@gmail.com,60,M,5
vanessa@gmail.com,29,F,2
```

datiCluster.arff

Abbiamo svolto tre analisi di clustering attraverso l'algoritmo K-Means, per individuare quale causasse l'errore (*Within cluster sum of squared errors*) minore:

- Con due cluster → Cluster 0, Cluster 1
- Con tre cluster → Cluster 0, Cluster 1, Cluster 2

- Con quattro cluster → Cluster 0, Cluster 1, Cluster 2, Cluster 3

In tutte le nostre analisi abbiamo ignorato l'attributo email perché univoco per ogni istanza e non significativo.

6.1 K-Means con due cluster

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-
    pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -V -N 2 -A "weka.core.
    EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     UTENTI_UTILIZZATORI
Instances:    13
Attributes:   4
              Eta
              Sesso
              NumPrenotazioni
Ignored:
              Email
Test mode:    evaluate on training data


=== Clustering model (full training set) ===



kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 1.9617482302246845

Initial starting points (random):

Cluster 0: 35,M,2
Cluster 1: 26,F,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute          Full Data          0          1
                      (13.0)      (9.0)      (4.0)
=============================================================
Eta                   32.0769    34.1111       27.5
                    +/-11.6438 +/-13.6971   +/-1.291

Sesso                        M          M          F
  M                  9.0 ( 69%) 9.0 (100%) 0.0 (  0%)
  F                  4.0 ( 30%) 0.0 (  0%) 4.0 (100%)

NumPrenotazioni       1.3077     1.5556       0.75
                    +/-1.4367  +/-1.5899  +/-0.9574



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

```

```
0        9 ( 69%)
1        4 ( 31%)
```

Risultati con due cluster

Abbiamo anche fatto mostrare media e deviazione standard per ogni attributo nei vari cluster. L'errore di questa analisi è ~ 1.96 .

Di seguito l'elenco degli utenti che appartengono a ciascun cluster

```
0    carla@gmail.com,26,F,0       cluster1
1    franco@gmail.com,35,M,2      cluster0
2    gino@gmail.com,35,M,2        cluster0
3    giovanna@gmail.com,27,F,0    cluster1
4    luigi@gmail.com,24,M,0       cluster0
5    marco@gmail.com,53,M,2       cluster0
6    matteo@gmail.com,23,M,0      cluster0
7    mauro@gmail.com,30,M,0       cluster0
8    melissa@gmail.com,28,F,1     cluster1
9    michele@gmail.com,22,M,2     cluster0
10   piero@gmail.com,25,M,1       cluster0
11   tiziano@gmail.com,60,M,5     cluster0
12   vanessa@gmail.com,29,F,2     cluster1
```

Assegnazioni con due cluster

6.2 K-Means con tre cluster

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-
    pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -V -N 3 -A "weka.core.
    EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     UTENTI_UTILIZZATORI
Instances:    13
Attributes:   4
              Eta
              Sesso
              NumPrenotazioni
Ignored:
              Email
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 0.8754168975069254

Initial starting points (random):

Cluster 0: 35,M,2
Cluster 1: 26,F,0
Cluster 2: 25,M,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute           Full Data          0          1          2
```

```
                          (13.0)       (4.0)       (4.0)       (5.0)
==========================================================================
Eta                      32.0769       45.75       27.5        24.8
                        +/-11.6438 +/-12.7377    +/-1.291   +/-3.1145

Sesso                         M           M           F           M
  M                     9.0 ( 69%)  4.0 (100%)  0.0 (  0%)  5.0 (100%)
  F                     4.0 ( 30%)  0.0 (  0%)  4.0 (100%)  0.0 (  0%)

NumPrenotazioni          1.3077        2.75        0.75        0.6
                        +/-1.4367     +/-1.5     +/-0.9574   +/-0.8944



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       4 ( 31%)
1       4 ( 31%)
2       5 ( 38%)
```

Risultati con tre cluster

L'errore di questa analisi è ~ 0.875.

Di seguito l'elenco degli utenti che appartengono a ciascun cluster

```
0   carla@gmail.com,26,F,0      cluster1
1   franco@gmail.com,35,M,2     cluster0
2   gino@gmail.com,35,M,2       cluster0
3   giovanna@gmail.com,27,F,0   cluster1
4   luigi@gmail.com,24,M,0      cluster2
5   marco@gmail.com,53,M,2      cluster0
6   matteo@gmail.com,23,M,0     cluster2
7   mauro@gmail.com,30,M,0      cluster2
8   melissa@gmail.com,28,F,1    cluster1
9   michele@gmail.com,22,M,2    cluster2
10  piero@gmail.com,25,M,1      cluster2
11  tiziano@gmail.com,60,M,5    cluster0
12  vanessa@gmail.com,29,F,2    cluster1
```

Assegnazioni con tre cluster

6.3 K-Means con quattro cluster

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-
    pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -V -N 4 -A "weka.core.
    EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     UTENTI_UTILIZZATORI
Instances:    13
Attributes:   4
              Eta
              Sesso
              NumPrenotazioni
Ignored:
              Email
Test mode:    evaluate on training data
```

```
13
14
15 === Clustering model (full training set) ===
16
17
18 kMeans
19 ======
20
21 Number of iterations: 2
22 Within cluster sum of squared errors: 0.4179168975069252
23
24 Initial starting points (random):
25
26 Cluster 0: 35,M,2
27 Cluster 1: 26,F,0
28 Cluster 2: 25,M,1
29 Cluster 3: 60,M,5
30
31 Missing values globally replaced with mean/mode
32
33 Final cluster centroids:
34                                Cluster#
35 Attribute          Full Data          0          1          2          3
36                      (13.0)      (3.0)      (4.0)      (5.0)      (1.0)
37 ============================================================================
38 Eta                  32.0769         41       27.5       24.8         60
39                    +/-11.6438 +/-10.3923  +/-1.291  +/-3.1145     +/-NaN
40
41 Sesso                     M          M          F          M          M
42   M                9.0 ( 69%) 3.0 (100%) 0.0 (  0%) 5.0 (100%) 1.0 (100%)
43   F                4.0 ( 30%) 0.0 (  0%) 4.0 (100%) 0.0 (  0%) 0.0 (  0%)
44
45 NumPrenotazioni      1.3077          2       0.75        0.6          5
46                    +/-1.4367       +/-0  +/-0.9574  +/-0.8944     +/-NaN
47
48
49
50
51
52 Time taken to build model (full training data) : 0 seconds
53
54 === Model and evaluation on training set ===
55
56 Clustered Instances
57
58 0       3 ( 23%)
59 1       4 ( 31%)
60 2       5 ( 38%)
61 3       1 (  8%)
```

Risultati con quattro cluster

L'errore di questa analisi è ~ 0.417.

Di seguito l'elenco degli utenti che appartengono a ciascun cluster

```
1 0    carla@gmail.com,26,F,0       cluster1
2 1    franco@gmail.com,35,M,2      cluster0
3 2    gino@gmail.com,35,M,2        cluster0
4 3    giovanna@gmail.com,27,F,0    cluster1
5 4    luigi@gmail.com,24,M,0       cluster2
6 5    marco@gmail.com,53,M,2       cluster0
7 6    matteo@gmail.com,23,M,0      cluster2
8 7    mauro@gmail.com,30,M,0       cluster2
9 8    melissa@gmail.com,28,F,1     cluster1
10 9   michele@gmail.com,22,M,2     cluster2
```

```
10  piero@gmail.com,25,M,1      cluster2
11  tiziano@gmail.com,60,M,5    cluster3
12  vanessa@gmail.com,29,F,2    cluster1
```

Assegnazioni con quattro cluster

6.4 Conclusioni sulla clusterizzazione

Quindi, dato che la differenza degli errori tra l'uso di tre cluster e di quattro è molto evidente, secondo la nostra analisi sarebbe preferbile implementare un algoritmo di clustering K-Means con quattro cluster.