Clustering oli extravergine d'Oliva

Relazione di Francesco Malferrari

Il set di dati preso in analisi è preveniente da oli extravergine d'oliva di 9 categorie diverse.

Le variabili sono 7 e rappresentano la concentrazione di sette acidi grassi:

- Palmitico $CH_3(CH_2)_{14}COOH$
- Palmitoleico $C_{16}H_{30}O_2$
- Stearico $C_{18}H_{36}O_2$
- Oleico $CH_3(CH_2)_7CHCH(CH_2)_7COOH$
- Linoleico $C_{18}H_{32}O_2$
- Eicosanoico $C_{20}H_{40}O_2$
- Linolenico $C_{18}H_{30}O_2$

La tabella che colleziona queste concentrazioni per ciascun olio è composta da 7 variabili e 572 campioni.

Quello che si vuole fare con questo dataset è il raggruppamento tramite vari algoritmi di clustering e l'individuazione della migliore metodologia confrontando il grafico PCA a due componenti principali con quello dei dati normali. I dati sono stati preprocessati con "Autoscale" e lo strumento utilizzato è Matlab con il Machine Learning Toolbox.

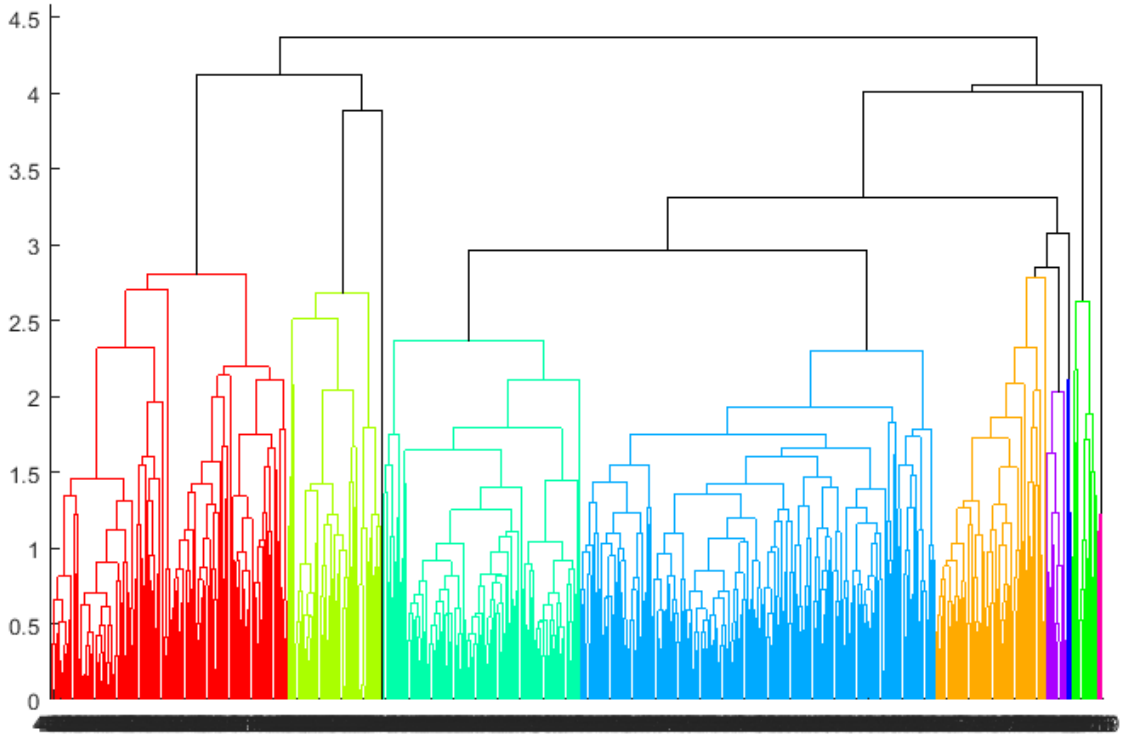
Sono stati provati e confrontati i seguenti linkage per il clustering gerarchico agglomerativo:

- Single linkage;
- Average linkage;
- Centroid linkage;
- Complete linkage;
- Ward's method.

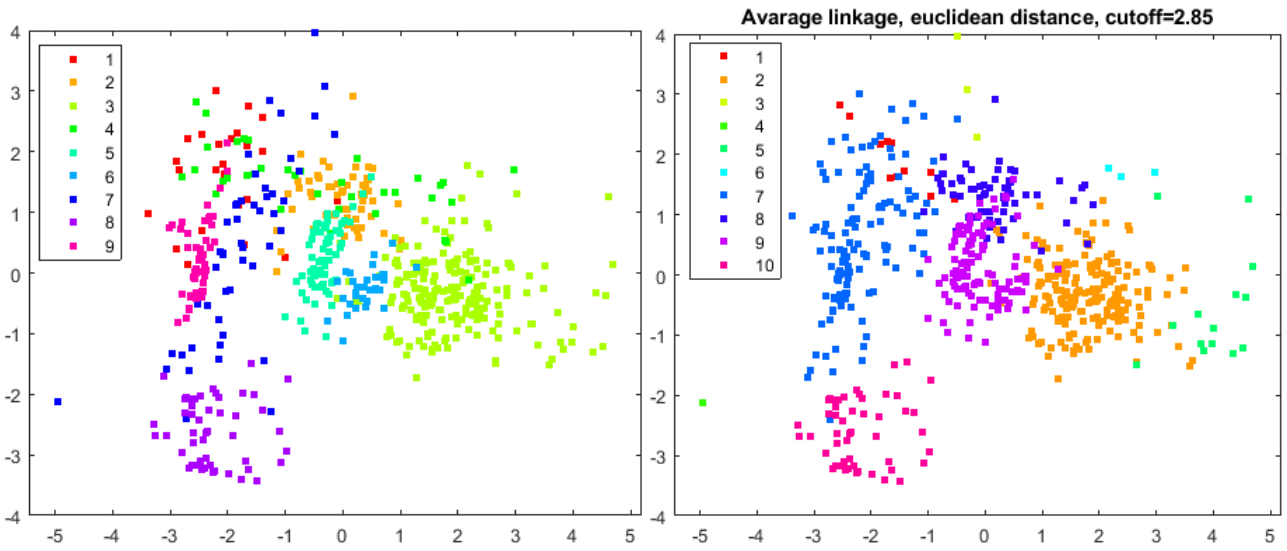
Ciascuno testato con la distanza Euclidea e con quella di Mahalanobis.

Una prima osservazione è che la distanza di Mahalanobis non è efficace su questo tipo di dati, in quanto il grafico PCA risulta avere i punti troppo sparsi rispetto ai raggruppamenti. Mentre con la distanza Euclidea mostra dei buoni risultati, in particolare con Ward e Average Linkage rispettivamente con cutoff 10.35 e 2.85.

Sono molto simili tra di loro, ma si ritiene di poco migliore il metodo Average Linkage. Qua il dendograma:

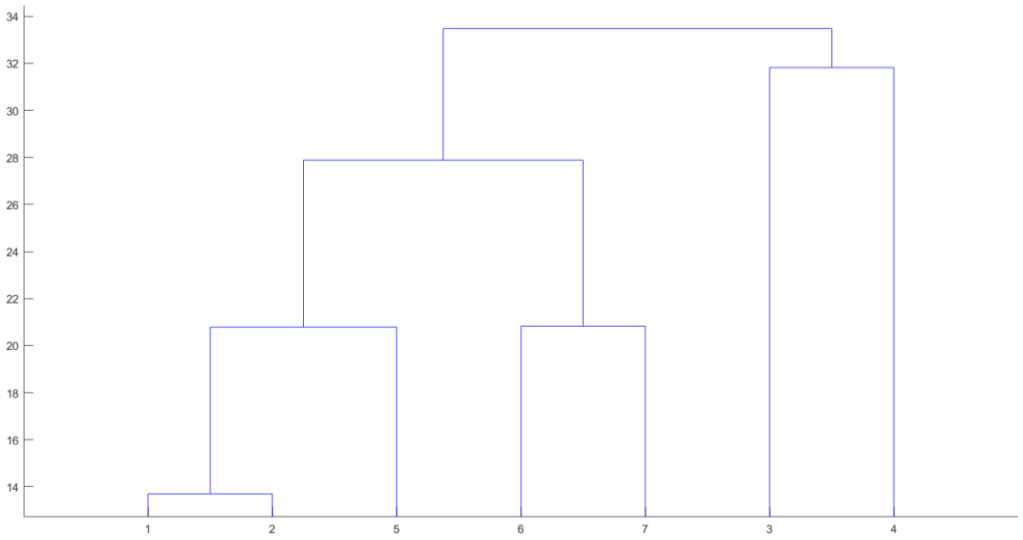


Di seguito vi è a sinistra il PCA sui dati solo autoscalati e a destra il plot col metodo appena descritto:

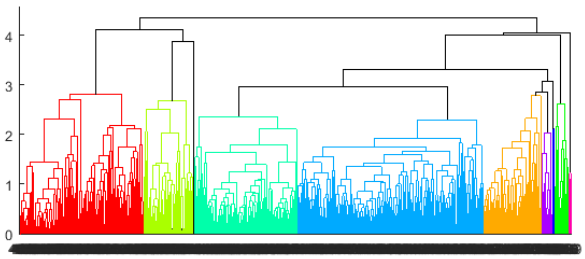


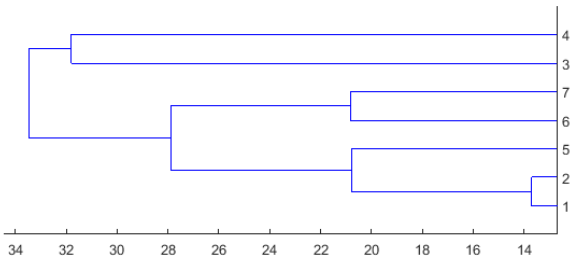
L'algoritmo riesce a dividere abbastanza bene tranne le categorie originali 4-5 e 9-7.

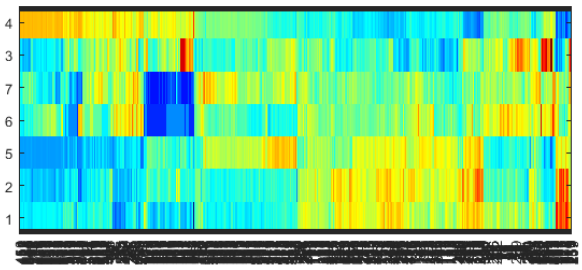
È stato poi fatto il clustering delle variabili usando come distanza la correlazione per vedere dei possibili collegamenti tra loro:



L'immagine che segue in basso a destra consiste in una figura con imagesc dei dati originali standardizzati ordinati secondo l'ordine del dendogramma dei campioni sulle x e di quello delle variabili sulle y.



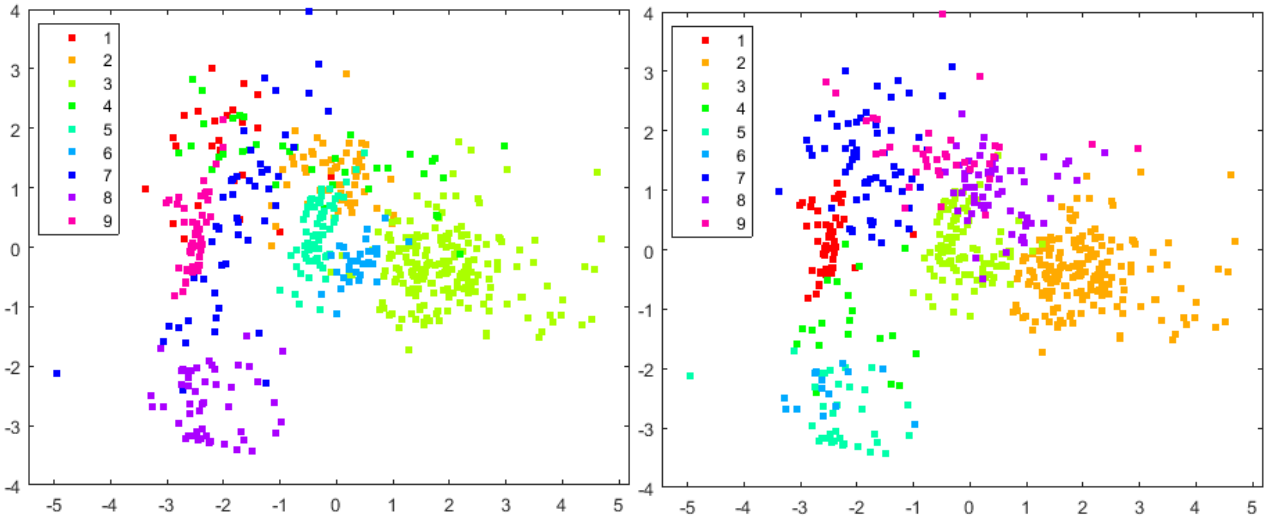




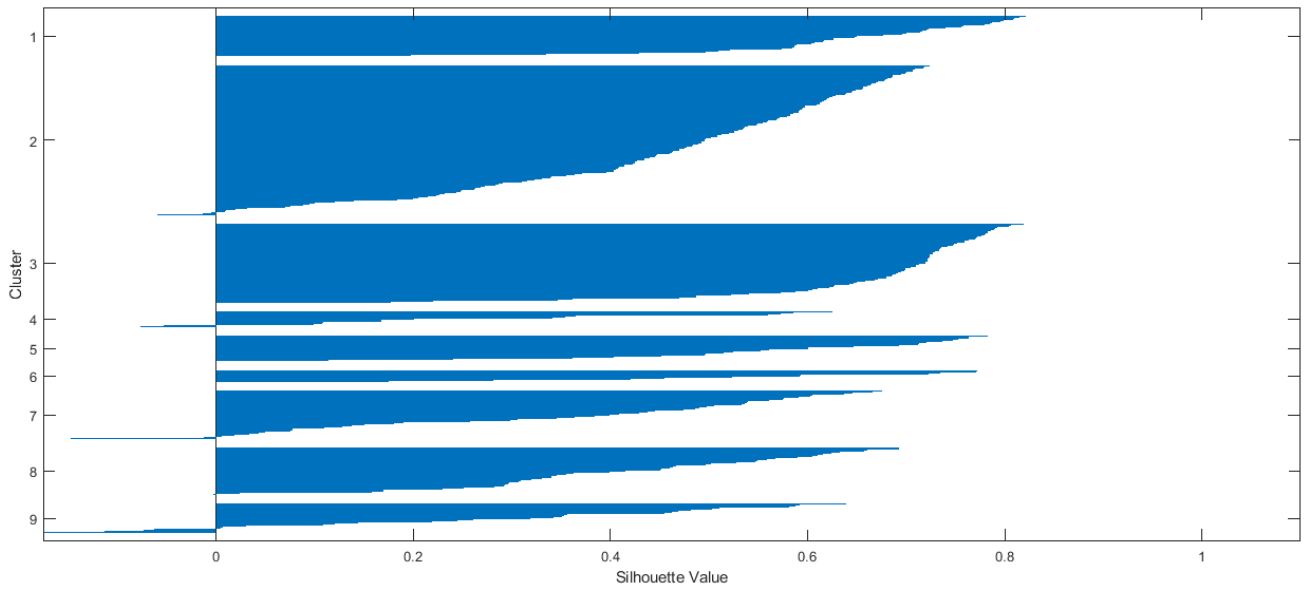
Da quanto risulta si può notare che le variabili 6 e 7 tendono ad avere valori più bassi nel secondo cluster e più alti del settimo. La variabile 5 ha valori più bassi per i primi due cluster, mentre 1 e 2 nei primi tre e più alti nell'ottavo e nel nono. La variabile 3 ha valori più alti nel sesto e settimo cluster, mentre la variabile 4 tende ad avere valori più bassi nel quarto, ottavo e nono cluster.

Dopo questi si è provato a vedere se la situazione può ulteriormente migliorare con altri algoritmi di clustering.

Il primo è l'algoritmo di partitioning K-means e di seguito si può notare a destra il plot di PCA di questo e a sinistra il plot originale (rimesso per comodità).

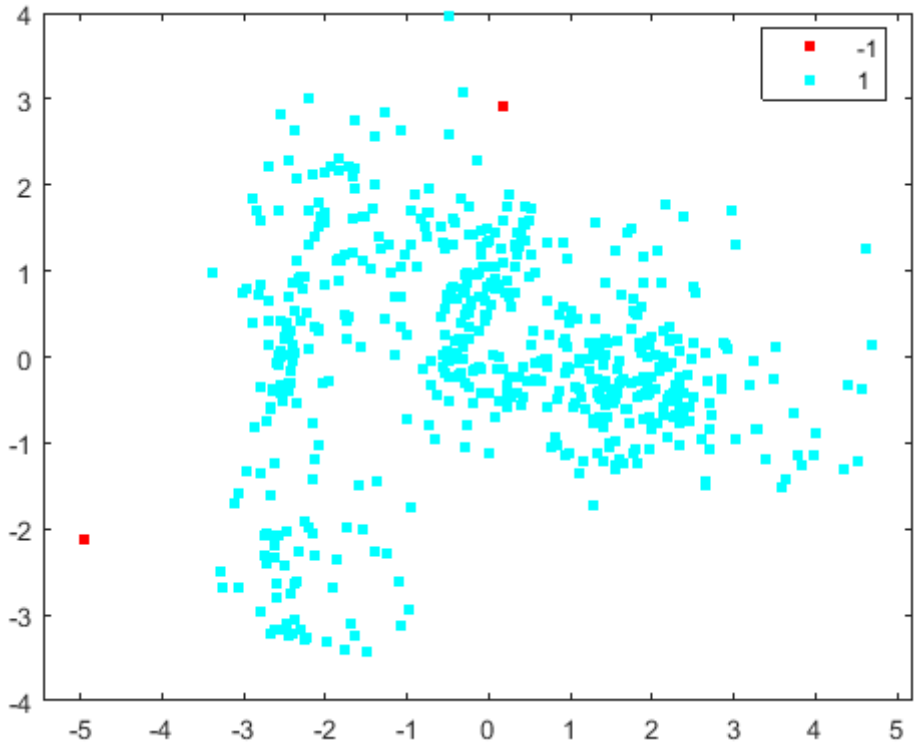


Il K-means non riesce ancora a distinguere tra le categorie 5 e 6, ma migliora rispetto alla distinzione tra le categorie 9 e 7.



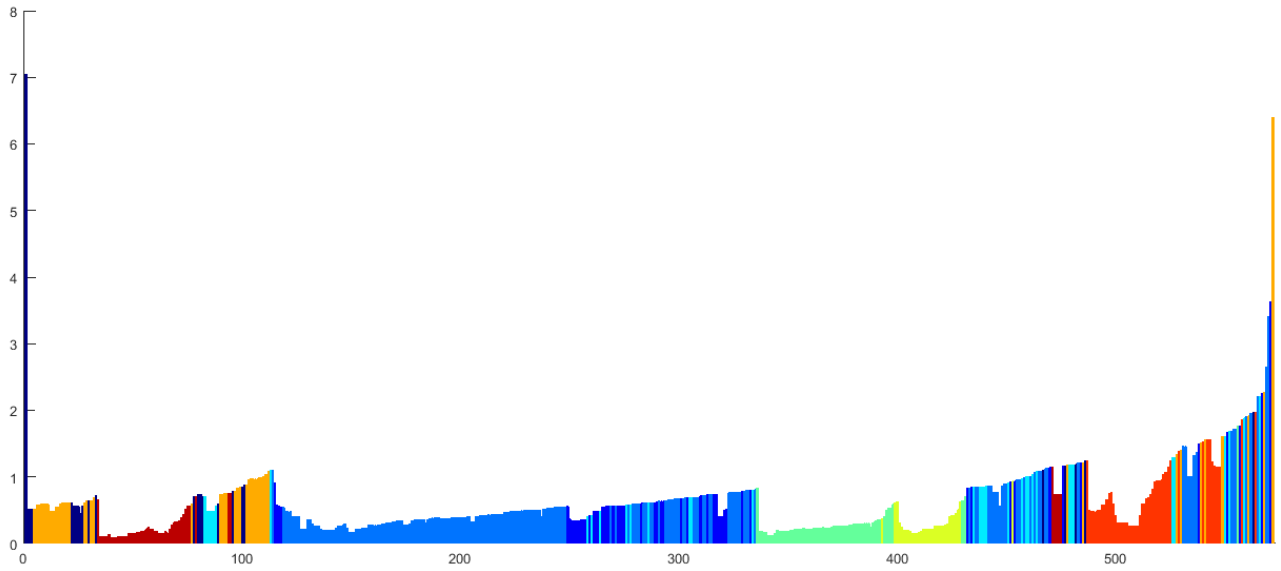
Dal grafico degli indici di Silhoutte si può notare come la maggioranza degli elementi dei cluster 3, 5 e 6 hanno valori maggiori o uguali a 0.6 il che li rende buoni. Gli altri cluster tendono ad avere in media valori più bassi (anche negativi) ma è abbastanza comprensibile, dato che il plot mostra come tutti i raggruppamenti sono vicini tra loro.

Un altro algoritmo testato è basato sulla densità ed è DBSCAN



I dati vengono divisi in due categorie di cui una è composta da solo due punti, indipendentemente dal numero minimo di punti presi e dal raggio, a causa del fatto che i punti sono troppo agglomerati. Quindi non è una soluzione efficace in questo contesto.

L'ultimo algoritmo testato sempre basato sulla densità è OPTICS



Dall'istogramma ordinato si possono notare cinque principali cluster, tutti con i colori abbastanza uniformi. Quindi si può affermare che riconosce meno cluster, ma questi sono abbastanza accurati.

In conclusione, l'algoritmo che presenta un'accuratezza migliore nel raggruppamento delle categorie è il K-means. Altri metodi che hanno mostrato buoni risultati sono:

- Average linkage: molto simile al K-means ma fa più fatica a distinguere le categorie 9 e 7;
- OPTICS: individua pochi grandi cluster (5 rispetto ai 9 dei dati originali), ma sono molto separati tra loro.