Business Intelligence per i Big Data

Esercitazione di laboratorio N.4

Gli obiettivi dell'esercitazione sono:

- **applicare i principali algoritmi di clustering disponibili in RapidMiner per segmentare gli utenti della campagna in funzione delle loro caratteristiche anagrafiche e lavorative e i testi in base alla similarità dei termini che contengono.**

Dati strutturati

Il dataset denominato UsersSmall (UsersSmall.xls) raccoglie dati anagrafici e lavorativi relativi a circa 300 persone contattate da un'azienda per proporgli l'iscrizione ad un loro servizio. Per tali utenti è noto se, dopo essere stati contattati, si sono iscritti al servizio proposto oppure no (valore del campo Response).

La lista completa degli attributi del dataset a disposizione (UsersSmall.xls) è riportata di seguito.

(1) Age
(2) Workclass
(3) Education record
(4) Marital status
(5) Occupation
(6) Relationship
(7) Race
(8) Sex
(9) Native country
(10) Response.

Dati testuali

Il dataset denominato Wikipedia contiene una collezione di 12 articoli di Wikipedia, appartenenti a 3 differenti categorie. In particolare, i documenti appartengono ai seguenti argomenti: matematica, cibo, sport.

Clustering di dati strutturati

L'obiettivo dell'analisi è di visualizzare e validare il risultato del processo di clustering svolto nel laboratorio precedente (Lab N.3) e applicare tecniche di riduzione della dimensionalità.

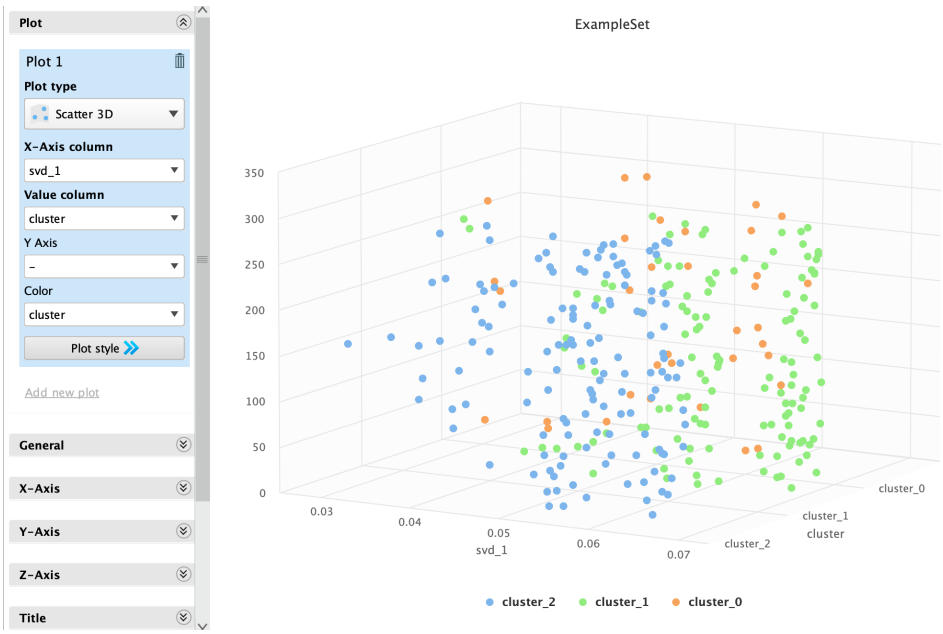
Per svolgere questo laboratorio si devono prima effettuare gli Obiettivi da 1 a 3 del Laboratorio N.3.

In questo laboratorio l'algoritmo di **clustering** in analisi è il **k-medoids.**

Obiettivo 1 - Visualizzazione/validazione del risultato di un processo di clustering tramite l'uso di tecniche di riduzione delle dimensioni dei dati – Principal Component Analysis attraverso SVD (Singular Value Decomposition) e Linear Discriminant Analysis (LDA).

1. Analizzare la qualità del clustering generato mediante una tecnica di riduzione della dimensionalità dei dati, nota come **Principal Component Analysis**. La più nota tecnica algebrica per effettuare la PCA è attraverso la decomposizione matriciale ***Singular Value Decomposition*** (SVD). La SVD trasforma i dati linearmente. Scegliendo le prime K componenti (dimensioni) ottenute dalla SVD, si possono proiettare i dati, originalmente a N-dimensioni in uno spazio a K-dimensioni, con K scelto dall'utente e minore di N.

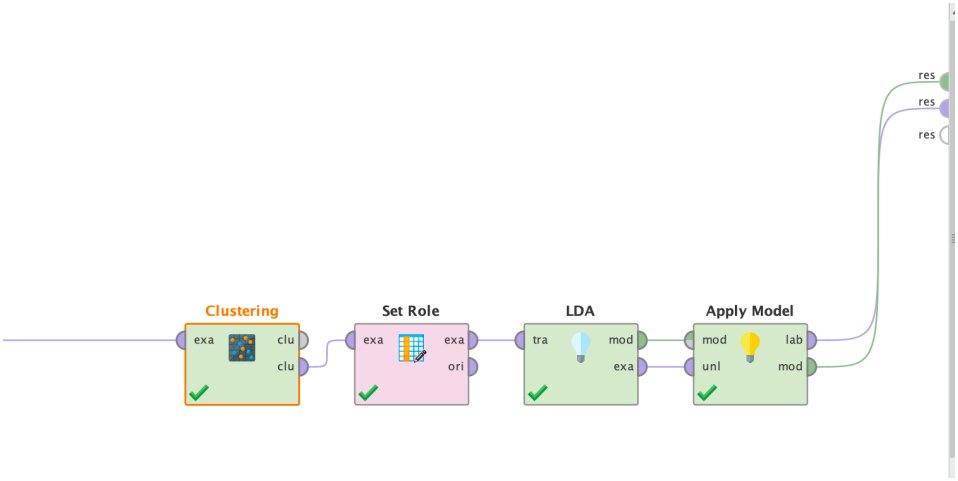
    a. Applicate l'operatore *SVD (Singular Value Decomposition)* sul dataset generato dal processo di clustering realizzato nel laboratorio precedente. Eseguire il processo impostando **K=3** e visualizzare su un grafico di tipo *scatter* i dati rispetto alle tre dimensioni individuate dall'operatore SVD. Usare l'attributo **cluster** come attributo per la **scelta dei colori dei punti**. I cluster sono ben definiti?



2. Analizzare la qualità del clustering tramite la **Linear Discriminant Analysis (LDA)**: LDA cerca di trovare un sottoinsieme di variabili (componenti) che rappresentano al meglio le differenze tra le classi dei dati massimizzando il rapporto tra la varianza tra le classi e la varianza all'interno delle classi. A differenza della Principal Component Analysis (PCA), che cerca di massimizzare la varianza totale dei dati senza tener conto delle classi, la LDA considera esplicitamente la struttura delle classi durante la riduzione della dimensionalità, rendendola particolarmente adatta quando l'obiettivo è la classificazione dei dati.

    a. Per applicare l'operatore LDA (Linear Discriminant Analysis) dovete esplicitare il ruolo del **cluster** come **label**. (Suggerimento: Utilizzare il blocco *Set Role*)

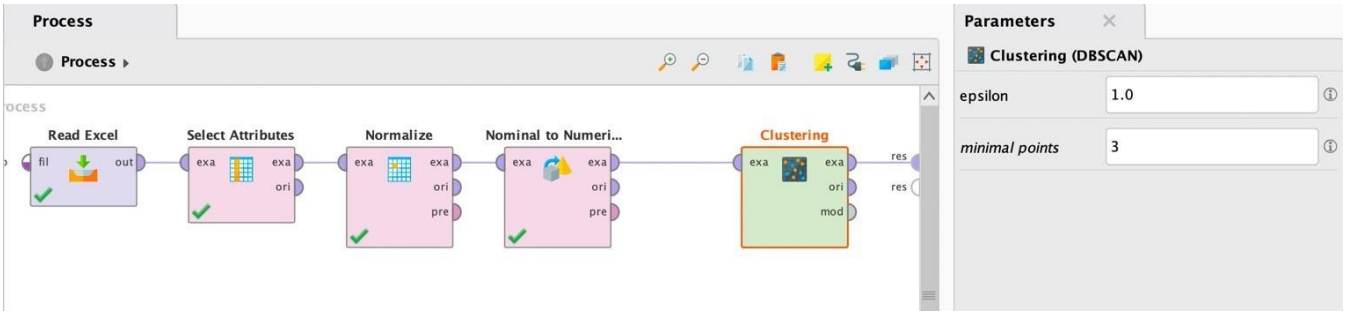
b. Applicate l'operatore **LDA** e dopo l'operatore **Apply Model.**



c. Visualizzare il risultato ottenuto con un grafico di tipo *scatter*. Ponete sull'asse delle X la colonna Age e **Value Column = cluster**. Ci sono differenze tra i due approcci? Riesce la LDA a separare correttamente i dati?

Obiettivo 2 - DBSCAN

- Applicare l'algoritmo di clustering DBScan settando min points = 3. Esegui il processo e analizza i risultati.



Valutazione oggettiva dei cluster generati per l'algoritmo DBSCAN

- Quanti cluster vengono identificati dall'algoritmo DBSCAN? Qual è il cluster che contiene il numero maggiore di elementi?
- Provare a ripetere l'esperimento cambiando il parametro *minimal points* (provare 2 e 4 come valori). Come cambia il numero di cluster individuati al variare di questo parametro? Questo comportamento è in linea con quanto ti aspettavi?
- Settare nuovamente il parametro *minimal points=3.* Come bisogna cambiare il parametro *eps* per ottenere meno punti etichettati come rumore? Perché? Prova a ripetere l'algoritmo con diversi valori di *eps* e guarda come cambia la cardinalità del cluster contenente punti rumorosi

Clustering di dati testuali

La seconda parte di questa esercitazione prevede l'analisi attraverso l'algoritmo K-Means della collezione di articoli denominata *Wikipedia*. Scompatta la cartella presente sul sito del corso ed esegui i passi seguenti.

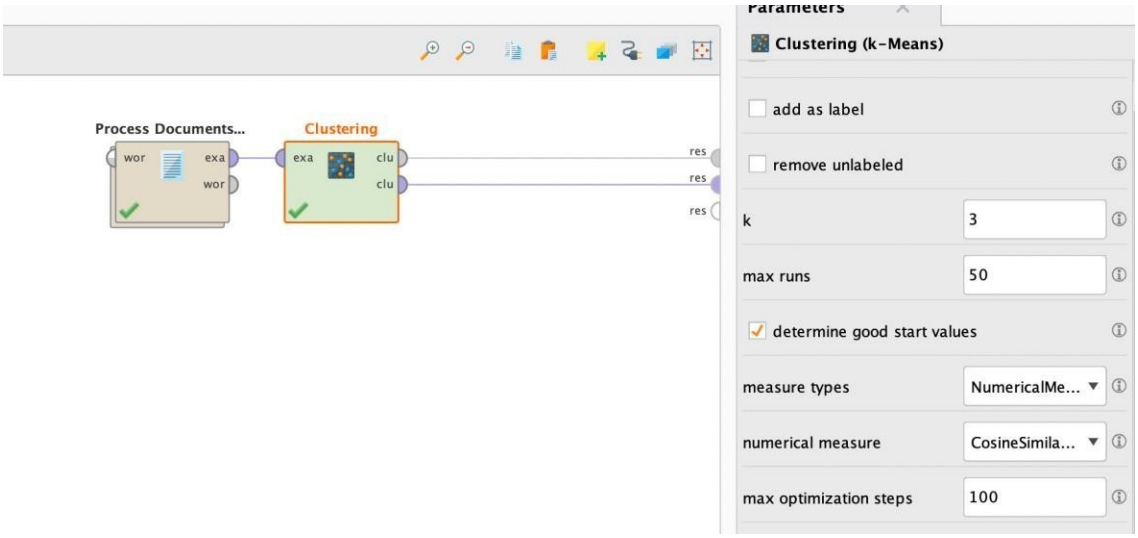
Obiettivo 1 - Import e preprocessing dei dati

- Trasforma la collezione di documenti nella matrice document*term. Per eseguire questa trasformazione, eseguire i diversi step di preprocessing, come imparato nella prima esercitazione.

Obiettivo 2 - Clustering dei dati

- Utilizzare l'algoritmo di **K-Means** per dividere la collezione in gruppi omogenei di documenti che parlino di uno stesso topic. Per i dati testuali la misura per calcolare la distanza tra punti (in questo caso tra due documenti) è la **CosineSimilarity.**

- Impostare K=3 e max_runs=50: come vengono suddivisi i 12 testi iniziali tra i 3 clusters? Selezionare l'opzione "keep_text" all'interno dell'operatore "Process Documents from Files" per tenere traccia del testo originale.



- Identificare all'interno di ciascun cluster le 3 parole che hanno un'importanza maggiore. (SUGGERIMENTO: andare nella sezione "Centroid Table")
- Provare infine a visualizzare i cluster identificati attraverso la tecnica SVD (3 dimensioni). I 3 clusters identificati sono ben distinti nello spazio tridimensionale?