Cluster

L’analisi dei cluster è una tecnica matematica usata in informatica e altre discipline, essa si basa sul considerare diversi tipi di dati (numerici, persone, misure) ed unirli in gruppi che contengono tutti elementi che hanno somiglianze tra di loro. La creazione dei cluster può essere effettuata in diversi metodi, ma tutte le tecniche hanno in comune lo scopo di rendere quanto più possibili omogenei gli elementi all’interno di un gruppo e rendere quanto il più eterogenei i gruppi così che il grado di associazione si alto tra membri dello stesso gruppo e basso tra membri di gruppi diversi.

Le tecniche di raggruppamento tendono ad unire quei dati che sono tra di loro simili e svolgono questo lavoro basandosi sul concetto che ogni elemento di un certo insieme di dati ha delle caratteristiche osservabili, possono essere il colore degli occhi per le persone, o possono essere le denunce al numero verde 1522 fatte di anno in anno per una regione ed è per questo che si usano le funzioni distanza tra i vettori delle caratteristiche che servono a calcolare le misure metriche di somiglianza.

La misura di distanza utilizzata nel progetto è la metrica Euclidea così definita:

Dove è il valore della k-esima caratteristica dell’individuo I.

Ottenuta la distanza dei vari elementi è necessario raggrupparli in cluster. Esistono due tipologie per creare i cluster

* Metodi gerarchici: mirano a costruire gerarchie di cluster; si dividono in due tipologie di approcci diversi: L’approccio agglomerativo è un approccio “bottom-up”, si parte dall’inserire ogni elemento in un singolo cluster e si procede ad accorparli due a due; l’approccio divisivo è un approccio “top-down” che da un singolo cluster che comprende tutti gli elementi viene diviso in tanti sotto cluster.

Tutti i metodi gerarchici producono una struttura ad albero chiamata “dendogramma”.

* Metodi non gerarchici: permettono di riposizionare elementi di un cluster qualora venga notato che un elemento piazzato in cluster conviene spostarlo in un altro, di questo metodo fa parte l’algoritmo k-means.

Tra i metodi gerarchici possiamo trovare

* Metodo del legame singolo: Presa la matrice delle distanze, si parte dalla distanza 0; si trovano gli elementi che hanno la distanza minore, si rimuovono quegli elementi e si cercano i primi elementi che hanno la distanza minore. Se ci sono più coppie di elementi che hanno la distanza minore nella matrice se ne sceglie uno arbitrariamente.
* Metodo del legame completo: La distanza tra due gruppi g1 e g2, con n1 e n2 individui, è definita come la massima tra tutte le distanze di n1 e n2, questo metodo privilegia la differenza tra i gruppi piuttosto che l’omogeneità del gruppo stesso.
* Metodo del legame medio: nel metodo del legame medio si considera, come distanza tra due gruppi, la media di tutte le distanze calcolate a due a due tra tutti gli elementi dei due gruppi
* Metodo del centroide: la distanza tra i gruppi g1 e g2 è calcolata sulle medie campionarie dei due gruppi. La particolarità di questo metodo è che tende ad avere un effetto gravitazionale: I gruppi più grandi tendono ad assorbire i gruppi più piccoli.
* Metodo della mediana: il metodo è simile a quello del centroide, ma non è dipendente dalla numerosità del gruppo. Quando due gruppi si uniscono, il nuovo centroide è calcolato come la semisomma dei due gruppi precedenti.

Tra i metodi non gerarchici, il metodo usato nel progetto è stato “k-means”, l’algoritmo funziona in diversi step:

1. Si fissano a priori il numero dei cluster scegliendo però elementi che hanno determinate caratteristiche
2. Si considerano tutti gli elementi e si attribuisce ad ognuno un cluster basandosi sulla distanza minore dal punto di riferimento scelto per ogni cluster
3. Si ricalcolano i centroidi dei k gruppi costituendo il nuovo punto di riferimento per i cluster così ottenuti
4. Si rivalutano le distanze per ogni unità rispetto ai centroidi dei vari cluster. Se un elemento x ha una distanza minore ad un altro centroide rispetto a quello del proprio cluster, si riposiziona l’elemento.
5. Si ricalcolano i centroidi
6. Si ripete dallo step 4, se si arriva ad un punto in cui non ci sono stati spostamenti tra elementi dei cluster, l’algoritmo si conclude.

parlare dei legami, solo le scritte a parole, formula distanza euclidea, resto a parole tutto quanto