Capitolo 1

Clusters e clustering

Con il nome di *clustering* si definisce la tecnica non supervisionata utile al raggruppamento di oggetti in insiemi secondo determinate regole. Gli elementi appartenenti allo stesso *cluster* sono più simili tra loro rispetto che a quelli contenuti negli altri gruppi. La *cluster analysis* è diventata una componente molto importante nell'analisi dei dati, sia in campo scientifico che industriale. Quando un insieme di dati deve essere organizzato, questa tecnica assegna ogni punto ad un gruppo utilizzando strutture presenti nei *row data*.

L'apprendimento non supervisionato è una tecnica che consiste nel fornire al sistema una serie di dati in input che, successivamente, riclassificherà ed organizzerà sulla base di caratteristiche che questi oggetti hanno in comune senza avere bisogno di ulteriori informazioni.

Queste regole tendono a minimizzare la *distanza* interna a ciascun gruppo ed a massimizzare quella tra i gruppi stessi.

La distanza (o *metrica*) è una qualsiasi funzione $d : X \times X \to \mathbb{R}$ che soddisfa [1]:

$$
\begin{aligned}
&d(x, y) \geq 0 \\
&d(x, y) = 0 \iff x = y \\
&d(x, y) = d(y, x) \\
&d(x, y) \leq d(x, z) + d(z, y)
\end{aligned}
$$

La scelta della metrica influenza drasticamente la forma dei cluster poiché i rapporti tra le varie distanze possono cambiare totalmente. Tra le metriche più comuni sono presenti la *distanza euclidea* e la *distanza di Manhattan*.

Non esiste un algoritmo in grado di raggruppare qualsiasi dato in modo più corretto rispetto ad un altro, infatti esistono tecniche specifiche per ogni tipologia di dato, da quello statistico a quello sociale.

Ci si può ispirare a molteplici nozioni per il raggruppamento: dalla creazione di gruppi caratterizzati da una piccola distanza tra i singoli membri, all'utilizzo di particolari distribuzioni statistiche.

Molti degli algoritmi usati per fare clustering godono di alcuni tratti in comune che ci portano a definire i *modelli*.

I modelli di clustering tipici sono:

- Partizionale
- Gerarchico
- Spectral
- Altri modelli

1.1 Clustering partizionale

Gli algoritmi di clustering di questa famiglia creano una partizione delle osservazioni minimizzando la funzione di costo:

$$\sum_{i=1}^{k} E(C_i)$$

ove k è il numero dei cluster richiesti in output, C_i è l'i-esimo cluster e $E : C \to R^+$ è la funzione di costo associata al singolo cluster.

Questa funzione viene spesso tradotta in [2]:

$$E(C_i) = \sum_{j=1}^{|C_i|} dist(x_j, center(i))$$

dove $|C_i|$ è il numero di oggetti presenti nel i-esimo cluster e $dist(x_j, center(i))$ è una funzione che calcola la distanza tra il punto x_j ed il centro del cluster i-esimo.

Questa tipologia di algoritmi solitamente richiede di specificare il numero di cluster distinti che si vogliono raggiungere a processo terminato e mira ad identificare i gruppi naturali presenti nel dataset, generando una partizione composta da cluster disgiunti la cui unione forma il dataset originale.

Questo significa che ogni cluster ha almeno un elemento e che un elemento appartiene ad un solo cluster.

Per segmentare l'insieme in sottogruppi si utilizza il concetto di centri: inizialmente sono posizionati in modo casuale, o secondo un qualsivoglia algoritmo, e iterativamente mossi fino a che questi non raggiungano uno stato di fermo. Quando ciò accade si è in grado di definire a quale centro appartiene il singolo punto e, di conseguenza, la struttura dei cluster calcolati.

Gli algoritmi più famosi appartenenti questa categoria sono:

- K-means
- K-medoids
- Affinity Propagation

1.2 Clustering gerarchico

Gli algoritmi di clustering gerarchico, invece, creano una rappresentazione gerarchia ad albero dei cluster. Le strategie sono tipicamente di due tipi:

- Metodo agglomerativo *(bottom-up)*
  Con un cluster per ogni oggetto si procede con l'unione di questi, basando la selezione degli insiemi da unire su una *funzione di similarità*.
- Metodo divisivo *(top-down)*
  Partendo da un unico cluster contenente tutti li oggetti, lo si divide basando la scelta della selezione dell'insieme da dividere su una *funzione di similarità*.

Questa tipologia di algoritmi necessita anche di alcuni criteri di collegamento che specifica la dissimilarità di due insiemi utilizzando la distanza valutabile tra gli stessi insiemi. Questi criteri possono essere:

- Complete linkage: distanza massima tra elementi appartenenti a due cluster
$$\max \{ d(a,b) : a \in C_1, b \in C_2 \}$$
- Minimum o single-linkage: distanza minima tra elementi appartenenti a cluster diversi
$$\min \{ d(a,b) : a \in C_1, b \in C_2 \}$$
- Average linkage: la media delle distanze dei singoli elementi
$$\frac{1}{|C_1||C_2|} \sum_{a \in C_1} \sum_{b \in C_2} d(a,b).$$

con C_1 e C_2 i due cluster da unire e d la metrica prescelta.

Gli algoritmi più famosi appartenenti questa categoria sono SLINK (single-linkage) e CLINK (complete-linkage) [3].

1.3 Clustering spectral

Questa tipologia nasce dal bisogno di eliminare delle mancanze presenti negli altri metodi.

I metodi partizionali riescono a creare cluster solamente di forma sferica, basandosi, come abbiamo già visto, sul concetto di centro; al contrario i metodi density-based hanno un approccio troppo sensibile ai parametri dati.

Per risolvere questi problemi si utilizza un grafo di similarità che ha come vertici le componenti da clusterizzare e, come valore degli archi, la similarità delle componenti tra cui questo è sotteso.

Una volta creato il grafo, si procede con una serie di tagli minimi che possono essere trovati in questo modo:

$$NormCut(C_1, C_2) = \frac{Cut(C_1, C_2)}{Vol(C_1)} + \frac{Cut(C_1, C_2)}{Vol(C_2)}$$

dove C_1, C_2 sono due gruppi di nodi, w_{ij} il peso dell'arco sotteso tra i e j, Cut e Vol definiti come segue:

$$Cut(C_1, C_2) = \sum_{i \in C_1, j \in C_2} w_{ij}$$

$$Vol(C) = \sum_{i \in C, j \in V} w_{ij}$$

Uno degli algoritmi più famosi appartenente a questa categoria è senz'altro Spectral.

1.4 Altri modelli di clutering

Come già detto, non esiste un solo modo, o algoritmo che sia, per riuscire a raggruppare degli oggetti. Qualsiasi intuizione potrebbe infatti generare un nuovo modello di pensiero, come successo per il *density-based* e per il *distribution-based*.

- Density-based: Negli algoritmi di clustering density-based, il raggruppamento avviene analizzando l'intorno di ogni punto dello spazio, connettendo regioni di punti con densità sufficientemente alta ed eliminando il rumore, ovvero gli elementi appartenenti a regioni con bassa densità [4].
  L'algoritmo più famoso appartenente a questa categoria è senz'altro DBSCAN.

- Distribution-based: Questa tipologia di algoritmi è quella che si avvicina di più allo studio statistico. I cluster possono essere definiti come insiemi di oggetti che appartengono, probabilmente, alla stessa distribuzione [5].
  L'algoritmo che definisce al meglio questa tipologia è il Gaussian Mixture Models.

Bibliografia

[1] Wikipedia. Definizione di distanza. http://it.wikipedia.org/w/index.php?title=Distanza_(matematica)&oldid=61709316.

[2] springerreference. Partitional clustering. http://www.springerreference.com/docs/html/chapterdbid/179343.html.

[3] Matteo Matteucci. Hierarchical clustering algorithms. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html.

[4] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

[5] Jia Li. Mixture models. http://sites.stat.psu.edu/~jiali/course/stat597e/notes2/mix.pdf.