Progetto Elaborazione Linguaggio Naturale: Tecniche di Clustering

Giuseppe De Palma

Alma Mater Studiorum - Università di Bologna
giuseppe.depalma@studio.unibo.it
Matricola: 854846

Sommario Ciaone

1 Introduzione

Il *clustering* (o analisi dei gruppi) è una forma di *machine learning* non supervisionato che permette di raggruppare in *clusters* elementi non annotati dati in input. Un cluster è una collezione di oggetti "simili" tra loro che sono "dissimili" rispetto agli oggetti degli altri cluster. Questo tipo di machine learning è ottimo per partizionare un insieme di dati in diverse "categorie", quindi poter eseguire diverse analisi ed ottenere nuove informazioni. Applicazioni tipiche in cui il clustering viene molto usato è il riconoscimento di email di spam (le email a scopi pubblicitari o di frode), oppure per l'aggregazione di notizie (vedasi Google News per un esempio).

Il clustering trova possibili applicazioni anche nel campo dell'elaborazione del linguaggio naturale. Oltre alle nuove possibili analisi sui corpora ed al fornire una visualizzazione pittografica delle parole raggrupate, un interessante utilizzo è quello della **generalizzazione** delle parole.

Possiamo considerare i vari clusters delle classi di equivalenza. Per questo motivo, se avessimo un dataset su cui comporre i clusters fatto di frasi e parole, allora si potrebbe assumere che una qualche parola che compare in una frase può essere sostituita con un'altra dello stesso cluster lasciando intatta la correttezza della frase. Ad esempio, se avessimo nel nostro dataset "per Lunedì", "per Martedì", "per Mercoledì", "per Sabato", "per Domenica", senza avere "per Giovedì" e "per Vernedì", e avessimo un cluster in cui i giorni della

settimana sono raggruppati insieme, allora potremmo generalizzare l'utilizzo della preposizione "per" con Giovedì e Vernedì.

Il clustering, quindi, può essere molto utile anche nell'elaborazione del linguaggio naturale. Nel progetto in studio vengono testate le capacità di alcune tecniche di clustering da cui si derivano dei risultati per mostrarne le differenze, i pregi e i difetti. I dati utilizzati negli esperimenti, comunque, non sono parti di testo, ma semplici dataset di vettori numerici 2D in modo tale da poter facilmente visualizzare i grafici relativi ai clusters e determinare le caratteristiche di ogni tecnica.

1.1 Outline

[SCRIVERE OUTLINE]

2 Clustering

Ci sono numerosi algoritmi per effettuare clustering, ma essi sono classificabili in poche tipologie: il clustering gerarchico e il clustering partizionale. Clustering partizionale consiste nel ottenere dei clusters, di solito iterativamente, ma spesso senza determinare la relazione tra gli elementi. Si inizia con un insieme di clusters iniziale ed iterativamente si riassegnano gli oggetti nei clusters corretti. Il clustering gerarchico, invece, forma un albero (la gerarchia) degli elementi dove un nodo rappresenta un sotto-cluster del nodo padre e le foglie sono i singoli oggetti.

Un'altra importante distinzione tra gli algoritmi di clustering è il *soft clustering* e *hard clustering*. Nel primo caso, ogni oggetto può essere assegnato a più clusters secondo un qualche grado di appartenza, mentre nel secondo caso ogni oggetto è assegnato ad un unico cluster. In questo progetto vedremo quattro diversi algoritmi, due della classe di clustering gerarchico, due del clustering partizionale. I primi tre eseguono hard clustering mentre l'ultimo soft clustering.

Di seguito sono elencati i metodi implementati e testati:

- Clustering **gerarchico**
  1. Aggregativo (o bottom-up)

2. Divisivo (o top-down)

– Clustering **partizionale**
  1. K-Means
  2. EM (soft clustering)

2.1 Aggregativo

2.2 Divisivo

2.3 K-Means

2.4 EM

3 Sessione Sperimentale

4 Conclusioni