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, un interessante utilizzo è quello della **generalizzazione** delle parole. [SPIEGARE GENERALZZAZIONE]

I metodi implementati e testati sono quattro:

- Clustering **gerarchico**
  1. Aggregativo (o bottom-up)
  2. Divisivo (o top-down)
- Clustering **partizionale**
  1. K-Means
  2. EM

1.1 Outline

2 Related Works

3 Clustering

4 Sessione Sperimentale

5 Conclusioni