Una soluzione semplice per i *big data*

Su Science un algoritmo SISSA semplifica la categorizzazione dei dati

27 giugno 2014

Categorizzare e rappresentare in maniera sintetica enormi quantità di dati (si parla di peta, o addirittura esa-byte di informazione) è la sfida del futuro. Una ricerca della Scuola Internazionale Superiore di Studi Avanzati (SISSA) di Trieste, pubblicata oggi sulla rivista *Science*, propone una procedura efficiente per affrontare questa sfida.

Gli addetti ai lavori usano l'espressione *big data* per indicare grandissime quantità di informazioni, come quelle (foto, video, testi, ma anche altri tipi di dati più tecnici) che vengono condivise da miliardi di persone in ogni momento attraverso computer, smartphone e altri dispositivi elettronici. Quello attuale è uno scenario che offre prospettive senza precedenti: tracciare le epidemie di influenza, per esempio, o monitorare il traffico stradale in tempo reale, o ancora

gestire l'emergenza in caso di catastrofi naturali. Per usare questa enorme mole di dati, però bisogna capirli, e prima ancora bisogna categorizzarli in maniera efficace, veloce e automatizzata. Uno dei sistemi più usati è un insieme di tecniche statistiche chiamate *Cluster Analysis (CA)*, in grado di raggruppare i set di dati secondo la loro "somiglianza". Due ricercatori della SISSA hanno ideato un tipo di CA basato su principi semplici e potenti, che si è dimostrata molto efficiente e in grado di risolvere alcuni dei problemi più tipici in questo ambito di analisi.

Gli insiemi di dati possono essere immaginati come una "nuvola" di punti in uno spazio a più dimensioni. Questi punti sono in genere dispersi in modi diversi: più rarefatti in una zona, più densi in un'altra. La CA serve a individuare in modo efficiente le zone più dense, raggruppando con questo criterio i dati in un certo numero di sottoinsiemi significativi. Ogni sottoinsieme corrisponde a una categoria.

"Pensate a un database di fotografie di volti", spiega Alessandro Laio, professore di Fisica e Statistica Biologica della SISSA. "Nell'archivio ci possono essere più fotografie di una stessa persona, la CA serve a raggruppare tutti gli scatti relativi allo stesso individuo. Questo tipo di analisi viene fatto per esempio dai sistemi automatici di riconoscimento dei volti".
"Noi abbiamo cercato di ideare un algoritmo più efficiente di quelli attualmente usati, che risolva alcuni dei problemi classici della CA", continua Laio.

Più nel dettaglio…

"Il nostro approccio si basa su un modo nuovo di individuare il centro dei *cluster*, cioè i sottoinsiemi" spiega Alex Rodrigez, autore insieme a Laio della ricerca. "Provate a immaginare di dover individuare tutte le città del mondo, senza avere a disposizione una mappa. Un compito immane", spiega Rodriguez. "Abbiamo perciò individuato un'euristica, cioè una regola semplice, una sorta di scorciatoia per ottenere il risultato".

Per scoprire se un luogo è una città infatti possiamo chiedere a ogni abitante di contare quanti "vicini" ha, ovvero quante persone vivono nel raggio di cento metri da casa sua. Una volta ottenuto questo numero, troviamo, per ogni abitante, la distanza minima a cui vive un altro abitante che ha più vicini di lui. "Questi due dati insieme", spiega Laio, "ci dicono quanto densamente è abitata la zona in cui vive un individuo e quanto distanti sono fra loro i cittadini che vantano il numero maggiore di vicini. Incrociando in maniera automatizzata queste informazioni, per tutta la popolazione mondiale, troveremo gli individui che rappresentano i centri dei cluster, che corrispondono alle varie città". " Il nostro algoritmo fa proprio questo tipo di calcolo, e può essere applicato in molti ambiti diversi", aggiunge Rodriguez.

La performance della procedura si è rivelata ottimale: "abbiamo testato il nostro modello matematico sull'Olivetti Face Database, un archivio di ritratti fotografici, ottenendo risultati molto soddisfacenti. Il sistema riconosce correttamente la maggior parte degli individui, e non ha mai dato 'falsi positivi'" commenta Rodriguez. "Questo significa che in qualche caso non ha riconosciuto un soggetto, ma non ha mai confuso un individuo con un altro. Rispetto ad altri metodi simili il nostro si è rivelato particolarmente efficace nell'eliminare gli *outlier*, cioè quei punti molto diversi dagli altri che tendono a sballare l'analisi".

LINK UTILI:

- **Abstract dell'articolo originale su *Science* (http://goo.gl/MfxAMw)**

IMMAGINI:

- **Crediti: SISSA**

Contatti:

Ufficio comunicazione:
pressroom@sissa.it
Tel: (+39) 040 3787557 | (+39) 340-5473118, (+39) 333-5275592
via Bonomea, 265
34136 Trieste

Maggiori informazioni sulla SISSA: www.sissa.it