Federated DBSCAN

Tesi di Laurea in

Ingegneria Informatica

Candidato
Gabriele Marino

Relatori
Prof. Francesco Marcelloni

UNIVERSITÀ DI PISA

Introduzione e Problema

- Per raggiungere la massiva quantità di dati richiesti dai moderni algoritmi di intelligenza artificiale, è spesso necessario che più utenti o organizzazioni combinino i propri dati. Lo scambio diretto dei dati è spesso però impossibile per motivi di sicurezza, o a causa delle norme vigenti sulla privacy.
- Problema
  - Come è possibile, per più utenti o organizzazioni (client), aggregare i propri dati per applicare l'algoritmo di raggruppamento DBSCAN in queste condizioni limitanti?

Soluzioni

- Federated DBSCAN: i client condividono informazioni di carattere generale con un server fidato che, dopo averle processate con una variante del DBSCAN, aggiorna i client con i risultati ottenuti.
  - Dataset partizionati orizzontalmente (diverse entità, stessi attributi): Horizontal Federated DBSCAN
    - ⇨ I client ripartiscono il proprio dataset tramite una griglia e condividono solo il numero di oggetti in ciascuna cella di essa.
  - Dataset partizionati verticalmente (stesse entità, diversi attributi): Vertical Federated DBSCAN
    - ⇨ I client condividono solo una matrice che esprime, per ogni punto del dataset, i punti ad esso vicini, data una certa metrica.

Risultati

- Le versioni federated del DBSCAN sono state implementate in Python, testate partizionando tra i client due diversi dataset per ciascuna versione, e valutate secondo diverse metriche: punteggi ARI e AMI, purezza, precisione e richiamo BCubed.
  - In ciascuna delle metriche, entrambi gli algoritmi hanno riportato ottimi risultati (>0.9), non lontani da quelli ottenuti dall'applicazione del DBSCAN sugli stessi dataset e con analoghi parametri di ingresso.
  - La versione orizzontale ha dimostrato di conservare buoni risultati nella valutazione delle metriche anche escludendo dal processo di apprendimento fino al 20% dei client.