Ufficio Accordi e Relazioni Internazionali

C N R

Oggetto: Relazione Finale sull'attività svolta presso il Department of Computer Science, University of California (UCLA), dal 23 Maggio 2013 al 15 giugno 2013.

L'attività svolta nel periodo di stage ha previsto una prima fase di studio dello stato dell'arte e poi una successiva fase di definizione di algoritmi efficienti per il data streaming di grosse moli di dati con il supporto di piattaforme di Cloud Computing.

Nella prima fase la ricerca è stata incentrata sullo studio delle caratteristiche che gli algoritmi di data mining su streaming devono supportare e sulle caratteristiche che i tool per la loro progettazione e implementazione devono avere. In particolare, sono stati analizzati i tool principali che permettono un'efficiente implementazione di algoritmi distribuiti di data mining su piattaforme di tipo Cloud.

Fra i tool analizzati, la scelta è ricaduta su Spark e D-Spark, sviluppati entrambi dall'Università di Berkeley e che supportano la distribuzione efficiente e operazioni avanzate sui dati, estendo in maniera interessante il paradigma Map-Reduce. Infatti, Spark estende il paradigma classico con la possibilità di introdurre dei cicli e quindi supportare algoritmi differenti dalle soluzioni classiche Map-Reduce, quali ad esempio il sistema Hadoop; inoltre con l'introduzione del supporto alla persistenza dei dati, esso garantisce anche una maggiore efficienza, rispetto ai sistemi esistenti. D-Spark estende Spark, con la possibilità di trattare computazioni deterministiche di tipo batch definite su piccoli intervalli di tempo e, quindi di fatto, supportando il trattamento di dati di tipo streaming.

Utilizzando questi tool, è stata progettata un'architettura a più livelli per la definizione e implementazione di algoritmi di data mining in generale (e di data streaming in particolare) che possano sfruttare le piattaforme di cloud computing, fornendo anche una stima dell'errore dell'algoritmo e dei tempi di esecuzione su Cloud e dei relativi costi sulla base di un'analisi delle performance dell'algoritmo stesso su un piccolo campione dei dati originali. Sarà così possibile prima, stimare i costi, i tempi e l'errore di un algoritmo utilizzando un cluster o una macchina locale o comunque limitando i costi utilizzando solo un numero limitato dei nodi di una piattaforma Cloud e dopo questa fase di analisi, si potrà decidere di investire di più lanciando l'algoritmo su un numero elevato di nodi.

L'architettura definita prevede al livello più basso, il sistema Mesos per la gestione di cluster, quindi i tool Hadoop, Spark e D-Spark per il supporto alla scrittura degli algoritmi di data mining e per la distribuzione dei dati; al livello superiore, è previsto un tool per la stima dei costi e dei tempi di esecuzione su Cloud e un tool per la stima dell'errore basato sull'esecuzione degli algoritmi utilizzando un sampling dei dati. Questo tool è in fase di implementazione e utilizzerà un algoritmo basato sulla ben conosciuta tecnica del bootstrapping per la stima dell'errore, estesa per poter operare anche su un piccolo campione dei dati originali.

Rende (CS), 03/07/2013

Il sottoscritto

(Gianluigi Folino)