Data Mining

Scritto del 26/06/2017
(SOLUZIONI degli ESERCIZI)

Problema. Sia D un dataset di N transazioni su un insieme di item I, e sia $\epsilon > 0$ un parametro. Supponiamo che per ogni itemset X sia nota una stima s_X del suo *supporto vero* (i.e., $\text{Supp}_D(X)$), tale che

$$\text{Supp}_D(X) - \epsilon \leq s_X \leq \text{Supp}_D(X) + \epsilon$$

Definito un ordine $X_1, X_2, X_3, \ldots$ per tutti gli itemset, tale che $s_{X_1} \geq s_{X_2} \geq s_{X_3} \ldots$, siano $K < K'$ due indici positivi tali che $s_{X_K} > s_{X_{K'}} + 2\epsilon$.

1. Dimostrare che per ogni coppia di indici $i, j \geq 1$, con $i \leq K < K' \leq j$, si ha $\text{Supp}_D(X_i) > \text{Supp}_D(X_j)$
2. Usando il punto precedente, provare che l'insieme $\{X_1, X_2, \ldots, X_{K'}\}$ contiene i Top-K frequent itemset rispetto al supporto vero (ovvero i K itemset con supporto vero più alto e quelli con lo stesso supporto del K-esimo).

Soluzione.

1. From the property of the estimate and from the chosen ordering we conclude that

$$\begin{aligned} \text{Supp}_D(X_i) &\geq s_{X_i} - \epsilon \\ &\geq s_{X_K} - \epsilon \\ &> s_{X_{K'}} + \epsilon \\ &\geq \text{Supp}_D(X_j). \end{aligned}$$

2. From the previous point we know that for each X_j with $j > K'$ there are at least K itemsets whose support is strictly greater than $\text{Supp}(X_j)$, namely $X_1, X_2, \ldots, X_K$. Therefore, X_j cannot belong to the Top-K frequent itemsets.

□

Problema. Sia P un insieme di N punti in uno spazio metrico (M, d), e sia $\mathcal{C} = (C_1, C_2, \ldots, C_k; c_1, c_2, \ldots, c_k)$ un k-clustering di P, con $k < N$. Ogni punto $q \in P$ è rappresentato da una coppia $(\text{ID}(q), (q, c(q)))$, dove $\text{ID}(q)$ è una chiave distinta in $[0, N-1]$ e $c(q) \in \{c_1, \ldots, c_k\}$ è il centro del cluster a cui q appartiene. Progettare un algoritmo MapReduce efficiente per determinare, per ogni cluster C_i, i due punti $q_1(i), q_2(i) \in C_i$ più distanti dal centro c_i, determinando il numero di round, lo spazio locale e lo spazio aggregato richiesti dall'algoritmo. (Si assuma che per ogni i sia $|C_i| > 2$ e non esistano due punti equidistanti dal centro.)

Soluzione. The idea is the following. In the first round we partition the points arbitrarily into $\sqrt{N}$ subsets of equal size, namely S_j with $0 \leq j < \sqrt{N}$. Then, in each subset S_j we select the two points most distant from c_i, for every cluster C_i. If a cluster C_i has less than two points (i.e., either 0 or 1) in S_j we select all points from $S_j \cap C_i$. In the second round, for each cluster C_i, we determine the two points most distant from c_i, among the at most $2\sqrt{N}$ ones selected in the first round, which will coincide with $q_1(i)$ and $q_2(i)$. A detailed specification of the algorithm is as follows:

Round 1

- **Map phase.** Each pair $(\mathrm{ID}(q), (q, c(q)))$ is mapped into the intermediate pair $(j, (q, c(q)))$, with $j = \mathrm{ID}(q) \bmod \sqrt{N}$.
- **Reduce phase.** For each $0 \leq j < \sqrt{N}$ independently gather the set S_j of points represented by intermediate pairs with key j and select, for each C_i, the $\min\{2, |C_i \cap S_j|\}$ points of $C_i \cap S_j$ with largest distance from c_i. Represent each selected point q as a new pair $(c(q), q)$, where $c(q)$ is the key.
  Observation: At the end of the round there will be at most $2\sqrt{N}$ pairs with the same key $c(q)$.

Round 2

- **Map phase.** Identity
- **Reduce phase.** For each center c_i independently, gather all pairs (c_i, q) (where $c_i = c(q)$) produced in the previous orund, and return the two pairs with the largest values of $d(c_i, q)$. These will be the pairs $(c_i, q_1(i))$ and $(c_i, q_2(i))$

In order to prove correctness it is sufficient to observe that for every $1 \leq i \leq k$, if after the map phase of the first round $q_1(i)$ (resp., $q_2(i)$) is in S_j, then $q_1(i)$ (resp., $q_2(i)$) is surely among the $\min\{2, |C_i \cap S_j|\}$ points of $C_i \cap S_j$ with largest distance from c_i. As for the performance, we have that

- The number of rounds is 2.
- Since each S_j has size $\Theta\left(\sqrt{N}\right)$ and, in the first round, at most $2\sqrt{N}$ points are selected for each cluster, the local space is $M_L = \Theta\left(\sqrt{N}\right)$
- Since each point is represented at most twice, once as an input and once as a result of the selection done in the first round, the aggregate space is $M_A = \Theta(N)$.

□