CMPUT 675: Approximation Algorithms **Winter 2018**

Lecture 8 (Feb 1, 2018):Local Search for k-Median

Lecturer: Mohammad R. Salavatipour *Scribe: Based on older notes*

8.1 k-median problem

k-median is an important clustering problem that has similarities to both k-center and facility location problem. An instance of this problem is similar to k-center: given a metric $d(.,.)$ and a integer k. We have a set F of facilities/centres and the goal is to select a set $F' \subseteq F$ with $|F'| = k$ minimizing the sum of distances of all the points to the nearest centre: $\min \sum_j d(j, F') = kmed(F')$.

Without loss of generality, we can assume that $|F'| = k$. As in the k-center problem, we assume that the distance matrix is symmetric, satisfies triangle inequality, and has zeros on the diagonal. We present a local search algorithm for k-median problem with good approximation ratio. For every subset $F' \subseteq F$ we use $kmed(F')$ to denote the cost of the solution if set F' is chosen.

Local search algorithm

1. Start from an arbitrary F' with $|F'| = k$
2. On each iteration see see if swapping a facility in F' with one in $F - F'$ improves the solution
3. Iterate until no single swap yields a better solution

Figure 8.1: Local search algorithm for k-median problem

Theorem 8.1 *If F' is a local optimum and F^* is a global optimum, then* $\text{kmed}(F') \leq 5\text{kmed}(F^*)$

Proof: Our proof is based on "Simpler Analyses of Local Search Algorithms for Facility Location" by Gupta and Tangwongsan (arXiv:0809.255). The proof will focus on constructing a set of special swaps S. These swaps will all be constructed by swapping into the solution location i^* in F^* and swapping out of the solution one location i' in F'. Each $i^* \in S$ will participate in exactly one of these k swaps, and each $i' \in F'$ will participate in at most 2 of these k swaps. We will allow the possibility that $i^* = i'$, and hence the swap move is degenerate, but clearly such a "change" would also not improve the objective function of the current solution, even if we change the corresponding assignment. Let $\phi : F^* \to F$ be a mapping that maps each $f^* \in F^*$ to the nearest facility in F, i.e. $d(f^*, \phi(f^*)) \leq d(f^*, f)$ for all $f \in F'$.

Let $R \subseteq F'$ be those that have at most one $f^* \in F^*$ mapped to them. Now we define a set of k pairs of potential swaps: $S = \{(v, f^*) \subseteq R \times F^*\}$ such that:

1. $\forall f^* \in F^*$, it appears in exactly one pair $(v, f^*) \in S$.
2. each node $r \in R$ with $\phi^{-1}(r) =$ appears in at most two swaps.
3. each nodr $r \in R$ with $\phi^{-1}(r) = f^*$ appears only in one swap.

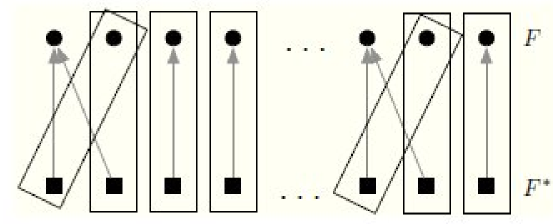

Figure 8.2: An example of mapping $\phi : F^* \to F$

How to build this set S? for each $r \in R$ with in-degree 1 we add pairs $(r, \phi^{-1}(r))$ to S. Let F_1^* be those of F^* that are matched this way. Other facilities in R have in-degree zero; let us call this set R_0. Note that

$$|F^* \setminus F_1| \leq 2|R_0|.$$

Now we can add other pairs by arbitrarily matching each node of R_0 with at most two in $F^* \setminus F_1^*$.

Observation: For any pair $(r, f^*) \in S$ and $\tilde{f}^* \in F^*$ with $\tilde{f}^* \neq f$: $\phi(\tilde{f}^*) \neq r$.

We use the fact that none of these potential swaps (in S) are improving to derive a bound on the cost of local optimum. Suppose that $\sigma : D \to F'$ and $\sigma^* : D \to F^*$ are mappings of clients to facilities in the local optimum and global optimum, respectively. For each $j \in D$, let $O_j = d(j, F^*) = d(j, \sigma^*(j))$ be the cost of connecting j in the optimum solution and $A_j = d(j, F') = d(j, \sigma(j))$ be its cost in the local optimum. We use $N^*(f^*) = \{j | \sigma^*(j) = f^*, f^* \in F^*\}$ to denote those assigned to f^* in the optimum solution and $N(f) = \{j | \sigma(j) = f, f \in F'\}$ to denote those assigned to f in the local optimum.

Lemma 8.2 *For each swap $(r, f^*) \in S$:*

$$\text{kmed}(F' + f^* - r) - \text{kmed}(F') \leq \sum_{j \in N^*(f^*)} (O_j - A_j) + \sum_{j \in N(r)} 2O_j.$$

Proof: Suppose we do the swap (r, f^*) and let's see how much the cost increases (note that since we are at a local optimum, this must be the case). We can upper bound this by giving a specific assignment of clients to facilities. Clearly the optimum assignment of clients to facilities cannot cost more than this:

- each client of $N^*(f^*)$ is assigned to f^*
- each client $j \in N(r) \setminus N^*(f^*)$ is assigned by the following rule: suppose $\tilde{f}^* = \sigma^*(j)$; we assign j to $\tilde{f} = \phi(\tilde{f}^*)$. Note that $\tilde{f} \neq r$.
- the assignment of all other clients remain unchanged.

For each $j \in N^*(f^*)$ the change in cost is exactly $O_j - A_j$; summing this over all $j \in N^*(f^*)$ gives the first term on the RHS. For $j \in N(r) \setminus N^*(f^*)$, the change in cost is:

$$\begin{aligned}
d(j, \tilde{f}) - d(j, r) &\leq d(j, \tilde{f}^*) + d(\tilde{f}^*, \tilde{f}) - d(j, r) && \text{using triangle inequality} \\
&\leq d(j, \tilde{f}^*) + d(\tilde{f}^*, r) - d(j, r) && \text{since } \tilde{f} \text{ is closest to } \tilde{f}^* \\
&\leq d(j, \tilde{f}^*) + d(j, \tilde{f}^*) && \text{using triangle inequality} \\
&= 2O_j
\end{aligned}$$

Thus, summing up the total change for all these clients is at most: $\sum_{j \in N(r) \setminus N^*(f^*)} 2O_j \leq \sum_{j \in N(r)} 2O_j$. ■

Now we use this lemma and sum over all pairs $(r, f^*) \in S$. Note that each $f^* \in F^*$ appears exactly once and each $r \in R \subseteq F'$ appears at most twice. Therefore:

$$\begin{aligned} \sum_{(r,f^*) \in S} (\text{kmed}(F' + f^* - r) - \text{kmed}(F')) &\leq \sum_{f^* \in F^*} \sum_{j \in N^*(f^*)} (O_j - A_j) + 2 \sum_{r \in R} \sum_{j \in N(r)} 2O_j \\ &\leq \text{kmed}(F^*) - \text{kmed}(F') + 4\text{kmed}(F^*) \end{aligned}$$

This implies that $\text{cost}(F') \leq 5\text{cost}(F^*)$.

■

Note that the running time of this algorithm is not necessarily polynomial. To get polynomial time algorithm we only consider swaps which improve the cost by a factor of at least $(1 + \delta)$ for some $\delta > 0$. So when the algorithm stops we are in an almost locally optimum solution, i.e. each potential swap can only improve by a factor of smaller than $1 + \delta$. Then the statement of lemma 8.2 would change to:

$$\text{kmed}(F' + f^* - r) - (1 - \delta)\text{kmed}(F') \leq \sum_{j \in N^*(f^*)} (O_j - A_j) + \sum_{j \in N(r)} 2O_j.$$

And then essentially the same analysis shows that the approximation ratio of the algorithm is at most $5(1 + \delta)$ which is $5 + \epsilon$ for sufficiently small $\epsilon > 0$. Since at each step of the local search, the value of the solution goes down by at least a constant factor, with M being the total sum of all edges as an upper bound for the value of the initial solution, it takes at most $O(\log_{1+\delta} M)$ steps to arrive at a locally optimum solution which is polynomial.

Improvment using t-swaps: A similar anaylsis shows that if one considered all t-swaps (instead of just 1-swaps) for a constant value of t at each step then the local search has a ratio of $3 + \frac{2}{t}$. More specifically, the algorithm starts with an arbitrary set F' of size k and in each iteration it checks whether swapping up to t centres in F' with those in $F - F'$ improves the solution or not.

Theorem 8.3 *t-swap local search for k-median has approximation ratio $3 + \frac{2}{t}$.*

As before let σ, σ be the mapping of clients to centres in F' and F^*, respectively. Similarly $\phi : F^* \to F'$ maps each $f^* \in F^*$ to nearest centre in F'. We give a partition of F' to $\{R_i\}_{i=1}^r$ and F^* into $\{F_i^*\}_{i=1}^r$. For each element $f \in F'$ let $deg(f) = |\phi^{-1}(f)|$.

> $i = 1$
> While there is an element $f \in F'$ with degree > 0 do
> $R_i \leftarrow f+$ any set of size $deg(f) - 1$ of elements of F' with degree 0.
> $F_i^* \leftarrow \phi^{-1}(R_i)$
> $F' \leftarrow F - R_i$; $F^* \leftarrow F^* - F_i^*$; $i \leftarrow i + 1$.

Here are some facts about the sets R_i and F_i^*'s:

- $|R_i| = |F_i^*|$ for all i.

- Each set R_i has exactly one element with degree > 0.
- For $j \in R_i$, if $\sigma^*(j) \notin F_i^*$ then $\phi(\sigma^*(j)) \notin R_i$.

Lemma 8.4 *If $|R_i| = |F_i^*| \leq t$ then*

$$\text{kmed}((F' \setminus R_i) \cup F_i^*) - \text{kmed}(F') \leq \sum_{j \in N^*(F_i^*)} (O_j - A_j) + \sum_{j \in N(R_i)} 2O_j.$$

The proof is similar to that of 8.2; assign each $j \in N^*(F_i^*)$ to $\sigma^*(j)$ and each $j \in N(R_i) \setminus N^*(F_i^*)$ to $\phi(\sigma^*(j))$. For the case that $|R_i| = |F_i^*| = s > t$ let $\tilde{R}_i$ be the degree 0 elements of R_i. We consider all pairs of swaps of the form $(r, f^*) \in \tilde{R}_i \times F_i^*$.

Lemma 8.5 *If $|R_i| = |F_i^*| = s > t$ then:*

$$\frac{1}{s-1} \sum_{(r,f^*) \in \tilde{R}_i \times F_i^*} [\text{kmed}(F' + f^* - r) - \text{kmed}(F')] \leq \sum_{j \in N^*(F_i^*)} (O_j - A_j) + \sum_{j \in N(R_i)} 2O_j.$$

Proof: Consider each swap $(r, f^*) \in \tilde{R}_i \times F_i^*$. An argument similar to proof of Lemma 8.2 shows: $\text{kmed}(F' - r + f^*) - \text{kmed}(F') \leq \sum_{j \in N^*(f^*)} (O_j - A_j) + \sum_{j \in N(r)} 2O_j$. Now suppose $j \in N^*(F_i^*)$. Then any $f^* \in F_i^*$ appears in $s-1$ pairs of swaps in $\tilde{R}_i \times F_i^*$. So summing over all of these and noting the $\frac{1}{s-1}$ factor we get the first term on the right hand side. For $j \in N(\tilde{R}_i)$, it appears in s pairs of $R_i \times F_i^*$. Since $\frac{s}{s-1} \leq 1 + \frac{1}{t}$, summing over all these we get the bound $\sum_{j \in N(R_i)} 2(1 + \frac{1}{t})O_j$. ■

Thus:

$$\begin{aligned}
0 \leq \sum_{i:|R_i| \leq t} (\text{kmed}((F' \setminus R_i) \cup F_i^*) - \text{kmed}(F')) + \sum_{i:|R_i|>t} \frac{1}{|R_i|-1} \sum_{(r,f^*) \in \tilde{R}_i \times F_i^*} (\text{kmed}(F' - r + f^*) - \text{kmed}(F')) \\
\leq \sum_i \Big(\sum_{j \in N^*(F_i^*)} (O_j - A_j) + \sum_{j \in N(R_i)} 2(1 + \frac{1}{t})O_j \Big) \\
= \text{kmed}(F^*) - \text{kmed}(F') + 2(1 + \frac{1}{t})\text{kmed}(F^*)
\end{aligned}$$