A constant-factor approximation algorithm for the k-median problem

By Shubham Agarwal

A Project Report Submission for:

COL870

Clustering Algorithms

Contents

List of Figures

List of Tables

1 Introduction

For the K-Median problem we are given a set of n-points and we are supposed to select a subset of k-points and assign each point in the input set to one of the k-point which is nearest. The assignment cost for this is the distance between the two points. The objective is to minimize the sum of the assignment cost. The paper proposes a LP rounding algorithm which gives an approximation factor of $6\frac{2}{3}$.

1.1 Generalized K-Median Problem

In the generalized k-median problem we are given a set of N points and a bound k for the number of centers to be selected. Further there is an assignment cost c_ij between each pair $i, j \in N$. Further there is a demand d_j for each point $j \in N$ which could be viewed as the weight assigned to each point.

1.2 LP for k-median

The Integer LP for K-Median can be written as

$$minimize \sum_{i,j \in N} d_j c_{ij} x_{ij}$$

subject to,

$$\sum_{i \in N} x_{ij} = 1 \qquad for\, each\, j \in N$$

$$x_{ij} \leq y_i \qquad for\, each\, i, j \in N$$

$$\sum_{i \in N} y_i \leq k$$

$$x_{ij} \in \{0, 1\}\, and\, y_i \in \{0, 1\}$$

where y_i represent whether the point i is selected as center and x_{ij} represent whether point j is assigned to center i.
In the relaxed version of this LP the last two condition becomes.

$$x_{ij} \geq 0\, and\, y_i \geq 0$$

Definition : For any feasible solution $(\bar{x}\ ,\ \bar{y})$ of the relaxed LP, $\bar{C}_{LP}$ denotes its objective function and $\bar{C}_j$ denotes the cost incurred in assigning point j. i.e;

$$\bar{C}_j = \sum_{i \in N} \bar{x}_{ij} c_{ij}$$

Hence, $$\bar{C}_{LP} = \textstyle\sum_{j \in N} d_j \bar{C}_j$$

2 LP Approximation Algorithm

The idea of the algorithm is to round any feasible solution $(\bar{x}\ ,\ \bar{y})$ to the relaxed LP into a feasible integer solution and such that the increase in the objective function value is bounded by a factor of $6\frac{2}{3}$.

2.1 Outline Of the Algorithm

The algorithm contains three steps. An outline of each step is given below

1. **Consolidating Locations** The intuition behind this step is to group nearby points together and then performing k-median doesn't increase cost by much. More formally,
   (a) Points with positive demands are far from each other, i.e;
   $$c_{ij} > 4max(\bar{C}_i, \bar{C}_j) \qquad \forall i, j \in N\, s.t\, d'_i, d'_j > 0$$
   (b) The cost is increased by an additive factor of $4\bar{C}_{LP}$ in this step.
2. **Consolidating Centers** The intution behind this step is that in the solution to k-median problem centers should lie far from each other. More formally,
   (a) The solution $\bar{x}, \bar{y}$ is modified to a $\frac{1}{2}$-restricted solution x', y' such that
   $$y'_j = 0 \qquad \forall j \in N\, s.t.\, d_j = 0$$
   $$y'_j \geq \frac{1}{2} \qquad \forall j \in N\, s.t.\, d_j > 0$$
   (b) The cost increased in this step is at most by a multiplicative factor of 2.
   (c) The $\frac{1}{2}$-restricted solution is further modified to a $\{\frac{1}{2}, 1\}$ integral solution, i.e;
   $$y'_j = 0 \qquad \forall j \in N\, s.t.\, d_j = 0$$
   $$y'_j = \{\frac{1}{2}, 1\} \qquad \forall j \in N\, s.t.\, d_j > 0$$
   There is no increment of cost in this step.
3. **Rounding $\{\frac{1}{2}, 1\}$ integral solution to integral solution** The cost increased in this step is at most by a multiplicative factor of $\frac{4}{3}$.

• **Claim -** For the metric k-median outlined method yields a $6\frac{2}{3}$ approximation algorithm.

Proof - Suppose the solution $(\bar{x}\ ,\ \bar{y})$ to be rounded is the optimum solution of the LP, then $\bar{C}_{LP}$ would have been the lower bound on the k-median cost. Now step 2-3 above gives a feasible integer solution to the modified input with cost at most $2\frac{2}{3}$. Now using step 1 this could be converted to an integer solution to the original problem with adding at most $4\bar{C}_{LP}$. Hence approximation factor is $4 + 2\frac{2}{3} = 6\frac{2}{3}$.

3 Consolidating Locations

In this step we simply move demands d_j for some locations j to nearby points $i \in N$ s.t; $c_{ij} \leq 4\bar{C}_j$.

Algorithm

1. Let N = {1,2,.....,n} and let the locations be arranged s.t; $\bar{C}_1 \leq \bar{C}_2 \leq .... \leq \bar{C}_n$
2. Now consider points in this order. While considering any point j, if there is a point $i < j$ such that $d'_i > 0$ and $c_{ij} \leq 4\bar{C}_j$ then
$$d'_i = d'_i + d'_j \, and \, d'_j = 0 \qquad where \, d' \, are \, modified \, demands$$
3. N' demotes the set of points with $d'_j > 0$ i.e, $N' = \{j \in N \, : \, d'_j > 0\}$.

Lemma 1 Points $i, j \in N'$ satisfies the following condition
$$c_{ij} > 4max(\bar{C}_i, \bar{C}_j) \qquad \forall i, j \in N \, s.t \, d'_i, d'_j > 0$$
The proof follows from the algorithm itself.
Lemma 2 For any feasible integer solution {x',y'} for the modified input with demands d', there is a feasible integer solution for the original input of cost at most $4\bar{C}_L P$ more than the cost of (x',y') with demands d'.
Proof Let point j's demand was moved to point j'. Therefore, $c_{jj'} \leq 4\bar{C}_j$ Now let i be a point to which j' is assigned. Hence
$$x_{ij'} = 1 \qquad since \, it's \, a \, feasible \, integer \, solution$$
Now assigning point j also to i $x_{ij} = 1$ and applying triangular inequality proves the required lemma.

4 Consolidating Centers

Lemma 1 For any feasible fractional solution $(\bar{x}, \bar{y})$, $\sum_{i:c_{ij} \leq 2\bar{C}_j} \bar{y}_i \geq \frac{1}{2}$ for each $j \in N$. **Proof**
$$\bar{C}_j = \sum_{i \in N} x_{ij} c_{ij} \qquad Hence,$$
$$\sum_{i:c_{ij} \leq 2\bar{C}_j} x_{ij} \geq \frac{1}{2}$$
Now, $x_{ij} \leq y_i \, \forall i, j \in N$ Therefore, $\sum_{i:c_{ij} \leq 2\bar{C}_j} y_i \geq \frac{1}{2}$

Claim - There is a $frac12$-restricted solution (x',y') of cost at most $2\bar{C}_{LP}$. **Proof -** Modify the solution $(\bar{x}, \bar{y})$, by moving each fractional center to a point closest to it in N'. Now for points with $y_i > 0$ and $d_i = 0$. Let $j \in N'$ be point closest to i in N'
$$y'_j = min(1, y'_i + y'_j)$$
$$y'_i = 0$$
Also
$$x'_{jj'} = min(1, x'_{jj'} + x'_{ij'}) \qquad x'_{ij'} = 0$$

Now, since $c_{jj'} \geq 4max(\bar{C}_j, \bar{C}_{j'})$. Therefore, all points i such that $c_{ij} \leq 2\bar{C}_j$ should be moved to j. Hence $y'_j \geq \frac{1}{2}$. Now using triangular inequality the claim is easily proven.

Lemma - The minimum cost of a $\frac{1}{2}$-restricted solution (x',y') is $\sum_{j \in N'} d'_j c_{s(j)j} - \sum_{j \in N'} d'_j c_{s(j)j} y'_j$.
Claim For any $\frac{1}{2}$-restricted solution (x',y') there exists a $\{\frac{1}{2}, 1\}$ integral solution of no greater cost.
Proof Using above lemma minimum cost is

$$\sum_{j \in N'} d'_j c_{s(j)j} - \sum_{j \in N'} d'_j c_{s(j)j} y'_j$$

. Hence to minimize it $y'_j = 1$ for the points for which $d'_j c_{s(j)j}$ is more. Hence set $y'_j = 1$ for first $2k - n'$ points and $y'_j = \frac{1}{2}$ for rest $2(n' - k)$ points.

5 Rounding $\{\frac{1}{2}, 1\}$-integral solution

Build a collection of trees as follows

1. For each node $i \in N$ with $y_i = \frac{1}{2}$ draw a directed edge from i to s(i), $s(i) \in N$ is the point nearest to i.
2. This can have a cycle of at most of length 2 and corresponding to closest pair of points.
3. For each such cycle let one of the node be root and delete edge from root to other node.
4. Now s(i) is parent of i if there is a directed edge from i to s(i).

5.1 An overall 8-approximation rounding

A simple rounding algorithm gives an integral solution with ratio between the cost of given $\{\frac{1}{2}, 1\}$-integral solution and the optimal integral solution is at most 2. Hence it gives an overall 8-approximation algorithm.
Simple Algorithm

1. Build a center at each node for which $y'_i = 1$.
2. Partition the nodes $\{i \in N : y'_i = 1/2\}$ into subset of even and odd levels from the collection of trees formed earlier and build a center at each node of smaller of two subsets.

Lemma- For every point $j \in N' : y'_j = 1/2$ there is a center at either j or s(j).
Proof- Point j and s(j) both belongs two different subset(odd level and even level) and there is a center build at all points in one of the subset. Hence there is a center at either j or s(j).
Claim- The above algorithm gives an integral solution with ratio between the cost of given $\{\frac{1}{2}, 1\}$-integral solution and the optimal integral solution at most 2.
Proof- Using above lemma there is center at either j or s(j). Hence j contributes at most $d'_j c_{s(j)j}$ which is double its contribution to the cost of $\{\frac{1}{2}, 1\}$-integral solution.

5.2 An overall $6\frac{2}{3}$ approximation rounding

In this a probability distribution over integral solution is constructed. The basic intuition behind this step is decomposing tree in small neighbourhoods (Step A) and create distributions over these smaller trees (Step B). Both these steps should follow the certain properties describe below.

Step A: Creating 3-level trees - From the set of trees T corresponding to $\{\frac{1}{2}, 1\}$-integral solution a set of trees T_3 are constructed which follows the following properties.

1. Each tree has at most three levels. Level of root is 0.
2. A node i is a parent of j in a tree in T_3 only if i=s(j). Thus each tree in T_3 is a subgraph of the graph defined by trees in T.
3. Each node i with $y_i = \frac{1}{2}$ belongs to tree of size 2 or more in T_3.
4. If i is the root of a tree with $y_i = \frac{1}{2}$, then the distance from i to its nearest child is at most $2c_{is(i)}$, or s(i) is a level-1 node in some other tree in T_3.
5. For each node i in a tree which is not level-0 or level-1, $c_{is(i)} \geq 2c_{s(i)s(s(i))}$.

Step A: Distribution over 3-level trees - After obtaining 3-level trees a probability distribution over integral solution is constructed and k centres are chosen such that following properties are satisfied.

1. If y_i=1, then node i is chosen.
2. If a root is not chosen, then all level-1 nodes in that tree are chosen.
3. Each root is chosen with probability at least $\frac{2}{3}$.
4. Each level-1 node is chosen with probability at least $\frac{1}{3}$.
5. Each level-2 node is chosen with probability at least $\frac{1}{2}$.
6. For each level-2 node i, conditioned on the fact that i not chosen its parent s(i) is chosen with probability at least $\frac{1}{3}$.

Lemma- If the above properties are true, then for each node j with $y_j = \frac{1}{2}$ the expected distance to its nearest centre is at most $\frac{2}{3}c_{js(j)}$. This could be proved by considering various cases for point j (whether it is a root, level 1 or level 2 node) and considering the associated properties.

Claim - The ratio between the cost of a given $\{\frac{1}{2}, 1\}$-integral solution and the optimal integral solution is at most $\frac{4}{3}$.
Proof Each node with $y_i = 1$ is chosen. For node with $y_i = \frac{1}{2}$ its expected contribution to the cost is $\frac{2}{3}c_{js(j)}$ (from above lemma) which is $\frac{4}{3}$ times its contribution to the cost of $\{\frac{1}{2}, 1\}$-integral solution.

6 Conclusion

The algorithm presented gives an overall approximation of $6\frac{2}{3}$ to metric k-median problem.

7 Reference

1. Charikar, Moses, et al. "A constant-factor approximation algorithm for the k-median problem." Proceedings of the thirty-first annual ACM symposium on Theory of computing. ACM, 1999.