Send Orders for Reprints to reprints@benthamscience.ae

Metric *k*-median Problem and Its Application in Reverse Greedy Randomized Algorithm

Shouqiang Wang* and Sheng Zhang

School of Information Science and Electric Engineering, Shandong Jiaotong University, Jinan 250357, China

Abstract: The k-median problem has been widely applied in many research fields such as clustering, logistic center etc. Its approximated algorithm has been interested by many computer theory scientists. In 2006, a reverse greedy algorithm for the metric k-median problem has been proposed by Chrobak and the approximative ratio is proved between Ω(lg(n)/lg(lg(n))) and Ω(lg(n)). In this paper, we present an improved version for the algorithm. In our improved algorithm, there are two central ideas, which include are randomized sample and reverse greedy. We proved the expected approximation ratio of the improved algorithm is $2\left(\ln\left(\frac{(2+\sqrt{3})\ln(2k)}{\alpha}-1\right)\right)+2$ and its running time $[\frac{k}{\alpha}(\ln(k))]2n$, where n represents the size of the given point set and α denotes the balanced parameter of the given point set.

Keywords: *k*-median, randomized algorithm, reverse greedy, approximation ratio.

1. INTRODUCTION

The metric *k*-median problem is described as follows: given a set of *n* points, for each two points *i*, *j* in the given points set, define *c*(*i, j*) as the cost between *i* and *j*. The goal is to select at most *k* points from the given points set so as to minimize the sum of the assignment costs.

This problem has been proved to be a NP-Hard problem, and has been applied many fields such as clustering, operation research, web service replications in a content distribution network and logistics center selection etc. Since 1982, it attracted many computer theory scientists to study its approximation algorithms. In 1966, Balinski proposed a LP-relaxation approach to research its approximated algorithm. After that, a lot of algorithms were proposed by many computer scientists. These algorithms can be classified as filtering technique [2], original dual method [3], greedy technique [4] and local search [5-7] etc. The first constant factor approximation algorithm for the *k*-median problem was given by Charikar [8]. The main idea of this algorithm is linear program and rounded technique, and the approximate ratio is proved to be 20/3. In 1999, a 6-approximation algorithm for the *k*-median problem was presented by Jain and Vaziran [3]. Charikar and Guha [8] improved the 6 approximation to 4. The main strategy of this improved algorithm is cost scaling and greedy improvement. In 2001, a local search algorithm was proposed by Arya [5] and its approximate ratio is $3+\varepsilon$. In recent years, the approximation ratio was improved by a series of papers [5, 9-11] and the current best result is $1+\sqrt{3}+\varepsilon$ for any $\varepsilon>0$ via pseudo approximation [12].

Define *P* as the given point set, for the metric *k*-median problem, Amos Fiat presented a reverse greedy algorithm [13]. The algorithm initially set $P_n=P$, and it repeat *n-k* steps,at each step let $P_{k-1}= P_k-\{r_k\}$, where r_k in P_k is chosen so that the cost(P_{k-1})-is minimized. Fiat wondered whether this algorithm is *O*(1)-approximation algorithm. In 2006, Marek Chrobak [13] presented a nearly tight analysis of this algorithm by showing that its approximation ratio is between Ω(lg(*n*)/lg(lg(*n*))) and Ω(lg(*n*)).

In this paper, we proposed a randomized reverse greedy algorithm for the metric *k*-median problem with the minimum subset size constraint: given a finite point set *P* in a metric space and parameters *k* and α, where each subset size of the solution has at least $\frac{n\alpha}{k}$ points, select *k* center points such that minimize the sum of the assignment costs. We proved that the randomized algorithm expected approximate ratio is $2\left(\ln\left(\frac{(2+\sqrt{3})\ln(2k)}{\alpha}-1\right)\right)+2$ with high probability and the running time is $[\frac{k}{\alpha}(\ln(k))]^2 n$.

We organized this paper as follows: Section 2 presented an randomized (*2, O*(*ln*(*k*)/α)-approximation for the metric *k*-median problem, section 3 discussed the improved reverse greedy algorithm and section 4 concluded this algorithm

2. AN (2,O(LN(*K*)/α)-APPROXIMATION ALGORITHM

In general, an (α,β)-approximation for *k*-median problem guarantees a cost of at most α *OPT* and uses at most βk medians. Given point set *P*, we assume the *k* optimal subsets to be $P_1^*, P_2^*,\ldots,P_k^*$ and each optimal center of P_i^* is correspondingly to be defined as $f_i^*(i=1,2,\ldots,k)$. For each subset P_i^*, suppose its size to be at least $\frac{|P|\alpha}{k}$, where $0<\alpha\leq 1$ and we call it as the balanced parameter.

***Theorem 1*:** Given point set *P*, Denote *S* as the point set drawn uniformly at random from *P*. If the size *S* is greater than $(2+\sqrt{3})\frac{k}{\alpha}\ln(2k)$, then, $Pr(|S\cap P_i^*| \geq 1) \geq 1/2$, which is

1874-4443/14

mean that the probability of S include at least one point of each P_i^* is at least 1/2.

Proof: Assume the size of P is n, i.e., $n=|P|$, without loss of generality, suppose $|P_1^*| \le |P_2^*| \le$, …, $\le |P_k^*|$. Let $n_i^* = |P_i^*|$, $S_i = S \cap P_i^*, n_i^s = |S_i|$. The probability of each point in P_i^* included in S is obviously at least $\frac{n_i^*}{n}$. So, the expected value of the random variable n_i^s must be at least $\frac{n_i^*}{n}|S|$.

Based on Chernoff Bounds, we have $\forall i \Pr[n_i^s < \lambda|S|\frac{n_i^*}{n}] < e^{\left(-\frac{(1-\lambda)^2}{2}|S|\frac{n_i^*}{n}\right)}$, where $\lambda(0<\lambda<1)$ is selected as a parameter to trades the size of S against the probability. In order to let P_i^* include some points of S with a constant probability, we attempt to make this probability be smaller than $\frac{1}{k}$. Let A_i denote the event that $n_i^s < \lambda|S|\frac{n_i^*}{n}$. Then,

$$Pr[\exists i, n_i^s \le \lambda|S|\frac{n_i^*}{n}] = \Pr\,[\cup_{i=1}^k A_i] \le \sum_{i=1}^k \Pr\,(A_i)$$

According to the definition of α, it is obvious that the size of P_i^* is at least $\frac{n\alpha}{k}$. To ensure each P_i^* is contained in S at least one point of P_i^*, Set $|S|=\frac{2}{(1-\lambda)^2}\frac{k}{\alpha}\log(2k), \lambda = 2-\sqrt{3}$, we can determine that $|S| = (2+\sqrt{3})\frac{k}{\alpha}\log(2k)$ is sufficient to guarantee that $Pr[\exists i, n_i^s \le \lambda|S|\frac{n_i^*}{n}] < \frac{1}{2}$. It shows that $Pr[\forall i, n_i^s > \lambda|S|\frac{n_i^*}{n}] = 1 - Pr\,[\exists i, n_i^s \le \lambda|S|\frac{n_i^*}{n}] \le \frac{1}{2}$. On the other hand, since $n_i^s \ge \lambda|S|\frac{n_i^*}{n} = \ln(2k)$ and k is obviously greater than 1, So the probability that S would include at least one point for each P_i^* is at least $\frac{1}{2}$.

On condition that S contains at least one point of each P_i^*, next, we will prove the theorem that at most k centers in S such that the expected total cost is less 2 times than the optimal cost of the given instance.

Denote by f_i^* the optimal center to serve P_i^*. If we draw one point x at random from P_i^* independently and uniformly, then the expected distance between x and f_i^* is $E(d(f_i^*, x))=\frac{\sum_{x\in P_i^*} d(x,f_i^*)}{|P_i^*|}$. Let $d(x, y)$ denote the distance between any two points x, y in P. Given a point set $F\subseteq P$, the cost of F is defined by $cost(F)=\sum_{x\in P} d(x,F)$, where $d(x, F)$ denotes the minimum distance between x and the point in F. Our goal is to find a k-element set $F\subseteq P$ that minimizes the $cost(F)$. Let OPT denote the optimal cost.

***Theorem 2*:** If S satisfy that $\forall i\ |S\cap P_i^*| \ge 1 (1 \le i \le k)$, there exists one subset $F\subseteq S$ such that $E(\text{cost}(F)) \le 2OPT$.

***Proof*:** Without lost of generality, Define by $F=\{f_1, f_2, \ldots, f_k\}$ as the k points in S such that $f_i \in S\cap P_i^* (1\le i\le k)$ respectively, i.e , f_i belongs to one point of P_i^*. Let $F^*=\{f_1^*, f_2^*, \ldots\ldots, f_k^*\}$ denote the optimal solution. Now, one special assigning point method is considered. For each subset P_i^*, instead of f_i^*, let f_i serve the subset P_i^*. Denote by $\text{cost}(F_k')$ as the cost of this new assigning method.

$$\text{cost}(F_k') = \sum_{i=1}^k \sum_{y\in F_i^*} (d(y, f_i))$$

$$\le \sum_{i=1}^k \sum_{y\in F_i^*} (d(y, f_i^*) + d(f_i^*, f_i))$$

$$= \sum_{i=1}^k \sum_{y\in F_i^*} d(y, f_i^*) + \sum_{i=1}^k |F_i^*| d(f_i, f_i^*)$$

So, the expected value of $\text{cost}(F_k')$ is:

$$E(cost(F_k')) \le \sum_{i=1}^k \sum_{y\in F_i^*} d(y, f_i^*) + \sum_{i=1}^k |F_i^*| E(d(f_i^*, f_i))$$

$$=2\sum_{i=1}^k \sum_{y\in F_i^*} d(y, f_i^*)$$

$$=2OPT$$

For each optimal subset P_i^*, if the point x in P_i^* is assigned to f_j other than f_i, where f_j is the closed point in F, then we have $d(x, f_j)\le d(x, f_i)$. Further, we conclude that $cost(F)\le cost(F_k')$ and $E(cost(F))\le E(cost(F_k')) \le 2OPT$.

On condition that $\forall i\ |S\cap P_i^*|\ge 1$, to enumerate each possible k elements from S to serve P and calculate its cost, let the minimum value of all possible cost as the algorithm final solution, According to Theorem 2, the expected approximation ratio is at most 2 with probability greater than $\frac{1}{2}$. Since $|S| = (2+\sqrt{3})\frac{k}{\alpha}\log(2k)$ and the number of all possible k elements in S is $C_{|S|}^k$, owing to $C_{|S|}^k \le \left(\frac{|S|e}{k}\right)^k$ [14], so, the number of $C_{|S|}^k$ can be concluded as $O\left(\frac{(2+\sqrt{3})e\ln(k)}{\alpha}\right)^k$. It is obvious that if k is enough large, the running time must be very high and the algorithm has little practical value.

3. IMPROVED ALGORITHM

For each input instance for the metric k-median problem, the following algorithm produces an $O(\ln(\ln(k)/\alpha)$ approximation with high probability.

1) Construct subset S by drawing from the given point set P uniformly, where $|S| = (2+\sqrt{3})\frac{k}{\alpha}\log(2k)$.

2) For each point in P, compute the distance to the nearest point in S, let $R=S$.

3) Repeat the following process until $|R|=k$

Set $R=R-\{r_k\}$, where $r_k \in R$ is chosen so that the $cost(R)$ is minimized.

Before giving the approximate ratio, we first analyze the running time of this algorithm. In step 2, calculating the closest center to each point of S requires running $O(|S|\times|P|)$ times. Meanwhile, the number of iterations of step 3 is $|S|-k$, and the running time for computing $cost(S)$ is $O(|P|\times|S|)$. In each iteration, the size of S decreases by 1. So, the running time of step 3 is $O([|S|+(|S|-1)+,\ldots,+k])$. Based on the analysis above, we conclude that the overall time complexity of the algorithm is $O(|S|\times|P|)+ O([|S|+(|S|-1)+,\ldots,+k]\times|P|)$. Let n denote the size of the set P. Then, the time complexity can be simplified as $O([\frac{k}{\alpha}\ln(k)]^2 n)$.

Without loss of generality, suppose the deleted center order from R to be $r_0, r_1, \ldots\ldots, r_{|R|-k+1}$. Before presenting reverse greedy algorithm approximation ratio, we first given some related lemmas

Lemma 3: $1+1/2+1/3+\ldots+1/n=\ln(n)+\gamma$ (where $\gamma(<1)$ is a Euler constant)

Lemma 4: $cost(R)\leq cost(R\backslash\{r_0\})\leq\ldots\leq cost(R\backslash\{r_0,\ldots,r_{|R|-k+1}\})$。

Proof: For each center r, define $N(r)$ as the clients served by r. According to the definition of cost(R), each point in P must be assigned to its closest center in R. Based on step 3, the selected center r_0 from R satisfies that the condition that the value of $cost(R-\{r_0\})$-$cost(R)$ is minimal. If r_0 is deleted, each point in $N(r_0)$ has to be reassigned to one of the center in R-$\{r_0\}$, and the distance from the point to the center in R-$\{r_0\}$ is greater than r_0. So, $cost(R)\leq cost(R-\{r_0\})$. Similarly, we can conclude that $cost(R)\leq cost(R-\{r_0\})\leq,\ldots,\leq cost(R-\{r_0,\ldots,r_{|S|-k+1}\}$.

Let $F^*=\{f_1^*, f_2^*,\ldots\ldots, f_k^*\}$ denote the optimal solution. For each optimal center f_i^*, let f_i' denote the closet center in R to f_i^* and define $F'=\{f_1', f_2', \ldots, f_k'\}$.

Lemma 5: $\sum_{r\in R\backslash F'}[cost(R\backslash\{r\}-cost(R)]\leq cost(F')-cost(R)$

Proof: Let $N(r)$ denote the points of P that is closest to r than any other point in S, i.e., $N(r)=\{x|x\in P\wedge r\in R\wedge d(x,r)=d(x, S)\}$. Now, we consider the left side of the inequality:

$$\sum_{r\in R\backslash F'}[cost(R\backslash\{r\}-cost(R)]$$

$$=\sum_{r\in R\backslash F'}\sum_{x\in N(r)}[d(x,R\backslash\{r\})-d(x,R)] \quad [1]$$

For the right side of the inequality:

$$cost(F')\text{-}cost(R)$$

$$=\sum_{x\in P}[d(x,F')\text{-}d(x,R)]$$

$$+\sum_{r\in F'}\sum_{x\in N(r)}[d(x,F')-d(x,R)]$$

$$=\sum_{r\in F'}\sum_{x\in N(r)}[d(x,F')-d(x,R)]$$

$$+\sum_{r\in R\backslash F'}\sum_{x\in N(r)}[d(x,F')-d(x,R)] \quad [2]$$

On the expression $\sum_{r\in F'}\sum_{x\in N(r)}[d(x,F')-d(x,R)]$, since $r\in F'$ *and* $F'\subseteq R$, it follows that $d(x,F')=d(x,R)$, further, $\sum_{r\in F'}\sum_{x\in N(r)}[d(x,F')-d(x,R)]=0$ So, the expression of (2) can be abbreviated as cost(F')-cost(R) $=\sum_{r\in R\backslash F'}\sum_{x\in N(r)}[d(x,F')-d(x,R)]$.

Based on the above analysis, the expression of the lemma 5 can be rewritten as:

$$\sum_{r\in R\backslash F'}\sum_{x\in N(r)}[d(x,R\backslash\{r\})-d(x,R)]$$
$$\leq\sum_{r\in R\backslash F'}\sum_{x\in N(r)}[d(x,F')-d(x,R)]$$

If $F'\subseteq R$, it is obvious that $d(x, F')\geq d(x,R)$. For each center $r\in R\backslash F'$, since $F'\subseteq R$ and $r\notin F'$, it is easy to see that $F\subseteq R\backslash\{r\}$. Further, for the point $x\in N(r)$. we conclude $d(x, S\backslash\{r\})\leq d(x,F')$, So:

$d(x, S\backslash\{r\})$-$d(x, S)\leq d(x, F')$-$d(x, S)$.

Consider all x and sum up the above inequality, we come to the conclusion of this lemma.

Lemma 6: $cost(F')$-$cost(R)\leq 2cost(F^*)$

Proof **:** For each point $x\in P$, define $f(x)$ as the closet center in R to x, $f^*(x)$ as the closet center in F^* to x, and $f'(x)$ as the closest center in F' to x. By the definition of F', we conclude that $d(f'(x), f^*(x))\leq d(f(x), f^*(x))$. On the other hand, we have the following expression:

$$d(x,f'(x))\leq d(x,\ f^*(x))+d(f'(x),f^*(x))$$
$$\leq d(x,\ f^*(x))+d(f'(x),f^*(x))$$
$$\leq d(x,\ f^*(x))+d(x,f^*(x))+d(x,f(x))$$
$$\leq 2d(x,f^*(x))+d(x,f(x))$$

It follows that $d(x,f'(x))-d(x,f(x))\leq 2d(x,f^*(x))$ Sum up over all x, we come to the conclusion.

Theorem **7:** Given point set P, with high probability, the expected approximation ratio of this reverse algorithm is $2\left(\ln\left(\frac{(2+\sqrt{3})\ln(2k))}{\alpha}-1\right)\right)+2$, where α denotes the balanced parameter.

Proof**:** After running the process of R=R-$\{r_0\}$ in step 3 many times, the size of R decreases to j. Let R_j denote the subset of R whose size is j. While R_j becomes R_{j-1}, by lemma 2, the cost should be increase simultaneously. We first estimate the incremental cost in iteration j of step 3.

$$cost(R_{j-1})\text{-}cost(R_j)$$
$$\leq min_{r\in R_j\backslash F'}(cost(R_j\backslash\{r\})\text{-}cost(R_j))$$
$$\leq\frac{1}{|R_j\backslash F'|}\sum_{r\in R_j\backslash F'}(cost(R_j\backslash\{r\})-cost(R_j))$$
$$\leq\frac{1}{j-k}\sum_{r\in R_j\backslash F'}(cost(R_j\backslash\{r\})-cost(R_j))$$
$$\leq\frac{1}{j-k}(cost(F')-cost(R_j))$$
$$\leq\frac{2}{j-k}cost(F^*)$$

The first inequality is based on the definition of R_{j-1}, the second one is estimated the fact that the minimum is less than the average, and the third one follows from $F'\leq k$.

Summing up over j=$|S|$,$|S|$-1,...,k , by lemma 1, we obtain the following upper bound:

$$cost(R_k)\text{-}cost(R_{|S|})$$
$$\leq(\frac{2}{|S|-k}+\frac{2}{|S|-k-1}+\cdots+\frac{2}{k+1})cost(F^*)$$
$$\leq 2\ln(\frac{|S|}{k}-1)cost(F^*)$$

By lemma 2 and theorem 3, we conclude that $E(cost(R_{|S|}))\leq 2cost(F^*)$. If we apply the expected cost of $cost(R_{|S|})$ to the expression above, the following upper bound of $cost(R_k)$ is:

$$E(cost(R_k)) \leq \left[2\ln\left(\frac{|S|}{k}-1\right)+2\right] cost(F^*)$$

$$=[2\ln\left(\frac{(2+\sqrt{3})\ln(2k)}{\alpha}-1\right)+2]\, cost(F^*)$$

By theorem 1, for each i, the probability of $|S\cap P_i^*| \geq 1$ is at least $\frac{1}{2}$. So, if running the reverse algorithm once, we obtain the expected approximation ratio with probability greater than $\frac{1}{2}$. Further, if we run the algorithm $\lceil \log_2 n \rceil$ times and select the best result of them as the final result, we can obtain the result with probability greater than $1-\frac{1}{n}$.

Based on the result of the Theorem 3, Compared to the algorithm presented by M.Chrobak, whose approximate ratio is $\log n$ and the running time is $O(n^3)$, we get good approximate results. If the balanced parameter α is big enough, for example, close to 1, i.e., the size of the k optimal subsets is close to equal, the expected approximation ratio is less than $2\ln\left(\left(2+\sqrt{3}\right)\ln(2k)-1\right)+2$. And if k is less than 10, the expected approximate result may be close to the local search algorithm with single swap. Run several times, the minimum approximate ratio may be better than the local search. Meanwhile, the running time of the reverse greedy algorithm is lower than the local search.

CONCLUSION

In this paper, we presented a reverse greedy randomized approximate algorithm for the metric k-median problem. The main idea of the randomized algorithm is to draw one subset at random, which includes at least one point of each optimal client subset with high probability. Based on this sampling subset, we invoke reverse greedy algorithm to find k centers to serve the given point set. We presented this improved algorithm approximate ratio and concluded that if each size is close to equal it may get better approximate ratio and running less time than the algorithm put forward by M.Chrobak. Of course, there are also some disadvantages for this algorithm. For example, not only the sampling process, but also the expected approximate ratio relies on the balanced parameter. If the parameter α is enough big, the algorithm will run more time. So, only to the situation where the size of each divided subset is all most equal or have little difference does this algorithm suit.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors are grateful to the anonymous referees for their valuable comments and suggestions to improve the presentation of this paper.

FUNDING

This work was supported by the National Natural Science Foundation of China (GrantNo.61103022), the Project of Shandong Jiaotong University Program (GrantNo.Z201124). And this article content has no conflict of interest

REFERENCES

[1] M. L. Balinski, "On finding integer solutions to linear programs", *Proceedings of IBM scientific computing symposium on combinatorial problems*, pp. 225-248, 1966.

[2] J. H. Lin and J. S. Vitter, $\boldsymbol{\varepsilon}$-approximation with minimum constraint violation, *Proceedings of 24th ACM Symposium on Theory of Computing*, pp. 771-782, 1992.

[3] K. Jain and V. V. Vazirani, "Primal-dual approximation algorithms for metric facility location and k -median problems", *Proceedings of 31th ACM Symposium on Theory of Computing*, pp.2-13, 1999.

[4] K. Jain, M. Mahdian, and A.Saberi, "A new greedy approach for facility location problems", *Proceeding of 34th ACM Symposium on Theory of Computing*, pp. 731-740, 2002.

[5] V. Arya, "Local search heurictics for k-median and facility location problems", *Proceedings of the 33th Annual ACM Symposium on Theory of Computing*, pp. 21-29, 2001.

[6] D. S. Hochbaum, "Heuristics for the fixed cost median problem", *Mathematical Programming*, vol. 22, pp. 148-162, 1982.

[7] S. Arora, P. Raghavan and S. Rao, "Approximation schemes for Euclidean k-median and related problems", *Proceedings of the 30th Annual ACM Symposium on Theory of Computing,* pp.106-113, 1998.

[8] M. Charikar and S. Guha, "Improved combinatorial algorithms for the facility location and k-median problems", *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pp.1-10, 1999.

[9] M. Charikar and S. Guha, "Improved combinatorial algorithms for facility location problems", *SIAM Journal on Computing*, vol. 34, no. 4, pp. 803-824, 2005.

[10] K. Jain, M. Mahdian, E. Markakis, A. Saberi, and V. V. Vazirani, "Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP", *Journal of the ACM*, vol. 50, no. 6, pp. 795-824, 2003.

[11] M. Charkar and S. Li, "A dependent LP-rounding approach for the k-median problem, *The 39th International Colloquium on Automata, Languages and Programming*, pp. 194-205, 2012.

[12] H. Li and O. Svensson, "Approximating k-median via pseudo-approximation", *In Proceedings of the 45th annual ACM Symposium on Symposium on theory of computing*, STOC '13, pp. 901-910, 2013.

[13] M. Chrobak, C. Kenyon, and N. Young, "The reverse greedy algorithm for the metric k-median problem", *Information Processing Letters*, vol. 97, pp. 68-72, 2006.

[14] R. Motawani and P. Raghavan, "*Randomized Algorithms*", Cambridge University Press, Cambridge, UK, 1995.

Received: September 22, 2014 Revised: November 03, 2014 Accepted: November 06, 2014