Research on Randomized Greedy Algorithm for *k*-median Problem

Wang Shouqiang
Department of Information Engineering
Shandong Jiaotong University
Jinan ,China
Wangshouqiang888@163.com

Abstract--This paper presented a randomized greedy algorithm for *k*-median problem。First some facilities were drawn at random from the given facility set. Among these sample facilities, there exist *k* facilities to satisfy that the approximation ratio is at most 3 with high probability, if used to serve the whole given clients. Then, a $(3+O((\beta-1)\ln(\ln(k)/\alpha)))$ -approximation algorithm was given for this problem. At last, some datasets were used to test the valid of the greedy algorithm.

Keywords*：*Algorithm, approximation ratio, k-median; Greedy.

I. INTRODUCTION

In the *k*-median problem, we are given two subsets, the facility set F and the client set C. Let $d(f_i,c_j)\in Z^+$ denote the cost of serving client $c_j\in C$ by a facility $f_i\in F$; we will think this as the distance between client $c_j\in C$ and facility $f_i\in F$. Given integer k, the goal in this problem is to identify a subset $S\subseteq F$ of at most k facilities to serve the client set C such that the total cost of serving all clients by selecting facilities is minimal. The facility in S are said to open. In this paper, we study the *k*-median problem in metric space, which assume that distances $d(f_i,c_j)$ are symmetric and satisfy the triangle inequality.

This problem has many applications in operations research, and recently in computer science such as web service replications in a content distribution network and data mining. It is a well-known NP-Hard problem, and falls into the general class of clustering problems. Since 1982, a number of different approaches have been studied extensively from the prospective of approximation algorithms. Basically, approaches to the problem falls into these categories: LP-relaxation, filtering technique, original dual method, greedy technique and local search. Charikar et al formulated the problem as linear program and rounded the optimum fractional solution to obtain an 20/3 approximated result. This was the first constant factor approximation algorithm for *k*-median problem. Subsequently, based on ideas of cost scaling and greedy improvement, Charikar and Guha improved the 6 approximation given by Jain and Vaziran to 4. Arya et al. presented the best-known 3+ε approximation algorithms using a simple local search procedure.

Although the researches on approximation algorithm have made a great progress, the known approximation algorithm cannot give a good enough solution for larger instance in reasonable time. The drawback of almost every algorithm based on LP-relaxation is that they need to solve large linear programs and therefore have a high running time. Despite being able to obtain a high quality solution, the best-known algorithm based on local search still has an $O(nmk^2)$ running time. For large-scale instance, there are almost no algorithms that are able to find a high quality solution on a single PC in a very short time. However, in real world, many applications, such as the selection of base station in mobile communication and comparison in data mining, need a fast and effective algorithm to solve large-scale instance.

In this paper, a randomized greedy algorithm for *k*-median problem is presented。The main process of the greedy algorithm is described as follows: First, some facilities are drawn at random from the given facility set. Then, given the sampling facility set and the client set as the new input, k facilities are calculated by means of the greedy algorithm. At last, the running time and the expected approximation ration of the algorithm are analyzed in the paper.

II. GENERATING CANDIDATE FACILITY SET

Given the facility set F and the client set C, we define the *k*-median problem with balanced constraint as that each selected facility must serve at least $\alpha\frac{|C|}{k}$ clients, where $\alpha(0<\alpha\le 1)$ is a given parameter and is called as the balanced constraint parameter. Let $M=\{f_1^*,f_2^*,\ldots,f_k^*\}\subseteq F$ denote the optimal facility subset for the given instance with balanced constraint. Suppose that M divide the whole client set C into k subsets $C_1^*,C_2^*,\ldots,C_k^*$ and each size of $C_i^*(1\le i\le k)$ must be at least $|C_i^*|\ge\alpha\times n/k$ clients.

Theorem 1: Let S denote the sample set drawn at random from C, and $|S|=(2+\sqrt{3})\frac{k}{\alpha}\ln(2k)$, $n=|C|$. The probability that each C_i^* satisfies $|S\cap C_i^*|\ge 1$ must be at least 1/2.

Proof: Without loss of generality, suppose $|C_1^*|\le|C_2^*|\le\ldots|C_k^*|$. Let $n_i^*=|C_i^*|$, $S_i=S\cap C_i^*$, $n_i^s=|S_i|$. It is obvious that the probability of each point in C_i^* selected to be included in S is at least $\frac{n_i^*}{n}$. So, the expected value of random variable n_i^s must be at least $|S|\frac{n_i^*}{n}$. By Chernoff Bounds:

$\forall i$ $\Pr[\, n_i^s<\lambda|S|\frac{n_i}{n}\,]<exp(-\frac{(1-\lambda)^2}{2}|S|\frac{n_i}{n})$, where $0<\lambda<1$ is a chosen parameter that trades the sample size against the approximation factor. We would like this probability to be smaller than $\frac{1}{k}$ so that there is a constant probability that each C_i^* has at least some points in S. Let A_i denote the event that $n_i^s<\lambda|S|\frac{n_i}{n}$.

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

By the definition of α, the size of C_i* is obviously at least $\alpha\frac{n}{k}$. To ensure that S includes at least one point of each C_i*, the required sample size is $\frac{2}{(1-\lambda)^2}\frac{k}{\alpha}\log(2k)$. Set the parameter $\lambda=2-\sqrt{3}$, we can determine that $|S|=(2+\sqrt{3})\frac{k}{\alpha}\log(2k)$ is sufficient to guarantee that $\Pr[\exists i\, n_i^s<\lambda \mid S \mid \frac{n_i}{n}]<\frac{1}{2}$. It shows that $\forall i\ \Pr[n_i^s\ge\lambda \mid S \mid \frac{n_i}{n}]$ $=1\text{-}\Pr[\exists i\, n_i^s<\lambda \mid S \mid \frac{n_i}{n}]\ge\frac{1}{2}$。It is easy to see that if $k>1$, the sample set would include at least one point of C_i*.

For each point x_i in S, let $f_i\in F$ denote the closet point to x_i. Consider such facility set $R=\{f_i \mid 1\le i\le|S|\}$, it is obvious that $|R|\le|S|$. Next, we will prove that there are at most k facilities in R such that the expected total cost of these facilities is no more than 3 times the optimal cost of the original instance.

Lemma2: Denote by f_i* the optimal facility to serve C_i*. If we draw one point x from C_i* independently and uniformly, then the expected distance between x and f_i* is $E(d(f_i^*,x))=(\sum_{x\in C_i^*} d(x,f_i^*))/\mid C_i^* \mid$.

Proof: it is obvious.

Given the instance I, Suppose $R\subseteq F$. Let cost(R) denote the cost of R serving C and $OPT(I)$ represent the optimal cost of the given instance.

Theorem 3: Let R denote one of the subset of F generated by the sample set S**.** If S satisfy that $\forall i\ |S\cap C_i^*|\ge1 (1\le i\le k)$, there exists one subset $R_k\subseteq R$ such that $E(\text{cost}(R_k))\le3OPT(I)$.

Proof: Suppose that there exist k points $x_1,x_2,\ldots x_k$ in S such that $x_i\in S\cap C_i^* (1\le i\le k)$ respectively. For each x_i, since $x_i\in S\cap C_i^*$, x_i belongs to one of the points of C_i*. Denote by $f_i\in F$ the closet facility to x_i and R_k the set of these closest facilities. It is obvious that $|R_k|\le k$. Assume that $cost(R_k)$ represent the cost of R_k serving the whole client set C.

Now, we consider one special assigning method. For each subset C_i*, instead of f_i*, let f_i serve the subset C_i*. Denote by cost(R_k') the cost of this new assigning method. Since,

$$\begin{aligned}\text{cost}(R_k')&=\sum_{i=1}^{k}\sum_{y\in C_i^*} d(y,f_i)\\ &\le\sum_{i=1}^{k}\sum_{y\in C_i^*}(d(y,x_i)+d(x_i,f_i))\\ &\le\sum_{i=1}^{k}\sum_{y\in C_i^*}(d(y,f_i^*)+d(x_i,f_i^*)+d(x_i,f_i))\\ &\le\sum_{i=1}^{k}\sum_{y\in C_i^*}(d(y,f_i^*)+2d(x_i,f_i^*))\\ &=\sum_{i=1}^{k}\sum_{y\in C_i^*} d(y,f_i^*)+2\sum_{i=1}^{k}\mid C_i\mid d(x_i,f_i^*))\text{。}\end{aligned}$$

So , the expected value of cost(R_k') is:

$$\begin{aligned}E(\text{cost}(R_k'))&\le\sum_{i=1}^{k}\sum_{y\in C_i^*} d(y,f_i^*)+2\sum_{i=1}^{k}\mid C_i\mid E(d(x_i,f_i^*))\\ &=3\sum_{i=1}^{k}\sum_{y\in C_i^*} d(y,f_i^*)\quad\text{(by lemma2)}\\ &=3OPT(I)\end{aligned}$$

Because cost(R_k)≤cost(R_k'), it is easy to conclude that $E(\text{cost}(R_k))\le E(\text{cost}(R_k'))\le3OPT(I)$.

On the condition that the sample set S satisfies that $\forall i\ |S\cap C_i^*|\ge1$. By enumerating all possible k facilities in R to serve the client set C, return the minimum cost as the final solution. According to Theorem 3, the expected approximation ratio is at most 3 with probability greater 1/2. Since $|S|=(2+\sqrt{3})\frac{k}{\alpha}\ln(2k)$, it is obvious that $R\le|S|=(2+\sqrt{3})\frac{k}{\alpha}\ln(2k)$. So, the number of enumerating all possible k facilities in R is $C_{|R|}^{k}$. Owing to $C_{|R|}^{k}\le\left(\frac{|R|e}{k}\right)^k$, the number of enumerating k points is $O\left(\frac{(2+\sqrt{3})e\ln(k)}{\alpha}\right)^k$ .However, If k is enough large, the running time must be obviously very high and the algorithm has little practical value.

III. GREEDY ALGORITHM FOR *K*-MEDIAN PROBLEM

A. Algorithm

Let f denote any one facility. Define $N(f)$ as all the clients served by f. The main idea of the greedy algorithm is described as follows: First, based on the constraint parameter α, $(2+\sqrt{3})\frac{k}{\alpha}\ln(2k)$ points are drawn at random from the client set C. Denote by S the subset of these points. Then, assign each point of S to its closet facility. Suppose that R be the set of these facilities. Next, chose one facility in R that serve the minimum number of clients and delete it from R. For each point of $N(f)$, it is reassigned to the other closet facility of R. Repeat this procedure until $|R|=k$. The algorithm is described in detail below:

Input: Facility set F , Client set C, Parameter α

Output: Cost(F,C)

(1) Draw $(2+\sqrt{3})\frac{k}{\alpha}\ln(2k)$ points randomly and uniformly from C. Let S denote the set of these points.

(2) For each point x_i of S, Find the closet facility f_i . Let R denote the subset of this facility.

(3) Calculate Cost(R,C)

(4)For each facility $f_i\in R$, find all the points of $N(f_i)$.

(5)Find the facility r such that the size of $|N(r)|$ be minimum. $R=R-\{f_i\}$.

(6)For each point $c_i\in N(r)$, reassign it to the closet facility in R.

(7) Repeat (3)-(6) until $|R|=k$.

B. Complexity Analysis

Next, we will analyses both the running time and the approximate ratio of the greedy algorithm. Given instance I, step 2 finds the minimal distance from each point of S to its closest facility, and its running time is $O(|S|\times|F|)$。 The running time for computing $cost(R)$ in step 3 is $O(|C|\times|R|)$. It is obvious that step (4)-(7) runs $O(|C|\times|R|)$

times. So, the overall running time of the greedy algorithm is $O([\frac{k}{\alpha}\ln(k)]^2 n)$, where $n=|C|$, $m=|F|$.

Without loss of generality, suppose the deleted facility order from the subset R to be $r_0,\ldots r_{|R|-k+1}$. Before proving the approximation ratio, we first present some related lemmas.

Lemma 4: $1+1/2+1/3+\ldots+1/n=\ln(n)+\gamma$ (where $\gamma(<1)$ is a Euler constant)

Lemma5:$cost(R)\le cost(R\backslash\{r_0\})\le\ldots\le cost(R\backslash\{r_0,\ldots r_{|R|-k+1}\})$。

Proof: By the definition of $cost(R)$, for each client, it must be served by its closest facility. According to the step 5 above, the facility r_0 selected from R satisfies that the value of $cost(R-\{r_0\})$-$cost(R)$ is minimal. Let $N(r_0)$ denote the clients served by the facility r_0. After r_0 is deleted, each point of $N(r_0)$ has to be reassigned to one of the facility in $R\backslash\{r_0\}$ and the distance to it must be less than to r_0. So, $cost(R)\le cost(R\backslash\{r_0\})$.

Similarly, we can conclude that $cost(R\backslash\{r_0\})\le\ldots\le cost(R\backslash\{r_0,r_1,\ldots r_{|R|-k+1}\})$.

Given Instance I, define $A(I)$ to be the value obtained from the greedy algorithm and $OPT(I)$ to be the optimal value. It is obvious that $A(I)=cost(R\backslash\{r_0,r_1,\ldots r_{|R|-k+1}\})$. Among all the distances between the facility set F and client set C, define d_{max} to be the maximal and d_{min} to be the minimal. Denote by $\beta=d_{max}/d_{min}$.

Lemma 6: $cost(R)\le 3OPT(I)$

Proof: By Lemma 5 and Theorem 3, the conclusion is obvious.

Theorem7:$E(A(I))\le (3+O((\beta-1)\ln(\ln(k)/\alpha)))OPT(I)$。

Proof: given instance I, suppose r_0 to be one of the facility of R which serve the minimal clients after having run step 2-3. It is obvious that r_0 serve at most $|C|/|R|$ clients|. If r_0 being deleted, those clients served by r_0 need to be reassigned to other facilities. For each reassigned client, the varied distance is at most d_{max}-$d_{\min}$. So, we can conclude as follows:

$cost(R-\{r_0\})$-$\text{cost}(R)\le |C|/|R|(d_{max}-d_{\min})$

Similarly, we also easily conclude:

$cost(R-\{r_0,r_1\})$-$cost(R-\{r_0\})$
$\le |C|/(|R|-1)(d_{max}-d_{\min})$
……
$cost(R-\{r_0,r_1,\ldots, r_{|R|-k+1}\})$- $cost(R-\{r_0,r_1,\ldots, r_{|R|-k+1}\})\le|C|/(k+1)(cost(R-\{r_0,r_1,\ldots, r_{|R|-k+1}\})$

Summing up all the inequality above, we conclude as follows:

$\text{cost}(R-\{r_0,r_1,\ldots, r_{|R|-k+1}\})-\text{cost}(R)$
$\le(\frac{1}{|R|-k}+\frac{1}{|R|-k-1}+\ldots+\frac{1}{k+1})\times|C|\times(d_{\max}-d_{\min})$

By lemma5 , $OPT(I)\ge\text{cost}(R)/3$.

So, $E(A(I)/OPT(I))$

$=cost(R\backslash\{r_0,r_1,\ldots r_{|R|-k+1}\})/OPT(I)$
$\le 3\ cost(R\backslash\{r_0,r_1,\ldots r_{|R|-k+1}\})/\text{cost}(R)$
$\le 3(1+\frac{\ln(\frac{|R|-k}{k})\times|C|\times(d_{\max}-d_{\min})}{\text{cost}(R)}$
$\le 3(1+\frac{\ln(\frac{|R|-k}{k})\times|C|\times(d_{\max}-d_{\min})}{|C|\times d_{\min}})$
$=3(1+(\beta-1)\ln(\frac{|R|-k}{k}))$
$=3(1+(\beta-1)(\ln\frac{(2+\sqrt{3})\ln(2k)}{\alpha}-1))$

At last, we get the result of $E(A(I))\le(3+O((\beta-1)\ln(\ln(k)/\alpha)))OPT(I)$.

IV.EXPERIMENTAL RESULTS

We use some data sets from ORLIB to test the performance of the greedy algorithm. Before running the algorithm, the value of the balanced constraint parameter must be given. In this paper, we select α=0.5 to test the greedy algorithm. Owing to the randomized of this algorithm, we run this algorithm 10 times and return the average and minimal value of the returning values as the final results. Table 1 presents the result of the average sample size, average value, minimal value and approximate ratio for each selected data set respectively.

V. CONCLUSION

Based on sampling we present a greedy algorithm and prove the approximate ratio. However, not only the sampling process of this greedy algorithm, but also the expected approximate ratio has relate to the value of α. The more of the α value is, the more efficiency of this algorithm. So, this algorithm can be applied to the situation where the size of each divided subset is all most equal or have little difference. Now, we put forward one question that whether we could find an algorithm which both the sample size and the approximated ratio have no related with the parameter α. This problem needs us to be further studied.

ACKNOWLEDGMENT

This paper is supported by the Natural Science Foundation of Shandong Jiaotong University (Z201024).

REFERENCES

[1]Arora S, Raghavan P, Rao S. Approximation schemes for Euclidean k-median and related problems. Proceedings of stoc'98, 106-113, 1998.

[2]Moses Charikar et al. A constant approximation algorithm for the k-median problem. Proceedings of stoc'1999, Atlanta GA USA, 1999.

[3]Kamal Jain, Vazirani V V. Primal-dual approximation algorithms for metric facility location and k-median problems. Proceedings of focs'99, 2-13, 1999.

[4]Vijay Arya etal. Local search heurictics for k-median and facility location problems, Proceedings of stoc'2001, Hersonissons, Crete, Greece, 2001.

[6]Pan Rui, Zhu Daming. Approximated Computational Hardness and Local Search Approximated Algorithm Analysis for k-Median Problem. Journal of software,2005,16(3):392-393. (In Chinese)

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

[8] Wei-Lin Li, Peng Zhang, Da-Ming Zhu. On Constrained Facility Location Problems. Journal of computer science and technology. 2008 23(5):740-748

[9] Maria-florina Balcan, Avrim Blum, Anupam Gupta. Approximate

clustering without the approximation. Proceeding of 2009 Symposium on Discrete Algorithms - SODA2009. 1068-1077, 2009

[10] Ke Chen. On Coresets for k-Median and k-Means Clustering in Metric and Euclidean Spaces and Their Applications. Siam Journal on Computing. 2009 39(3): 923-947

[11]Amit Kumar, Yogish Sabharwal, Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimension. Jounal of the ACM-JACM. 2010, 57(2):1-32

[12] Zhang P. A new approximation algorithm for k-Facility Location problem. Theoretical Computer Science,2007,384(1):126-135.

[13] Jain K, Mahadian M, Markakais E, Saberi A, Vazirani V. Greedy facility location algorithms analyzed using dual fitting using dual fitting with factor-revealing LP. Journal of the ACM,2003,50(6):795-824.

Table 1 Result of the average sample size, average value, minimal value and approximate ratio for each selected data set respectively

Data Set	Optimal Value	Data Set Size	Sampling Size	The Average		The Minimal	
				Average Value	Approximate Ratio	Value	Approximate Ratio
pmed1	5819	100	56	6242.54	1.072786	5889	1.01203
pmed6	7824	200	67	8204.96	1.048691	7941	1.014954
pmed7	5631	200	135	6020.32	1.069139	5767	1.024152
pmed11	7696	300	73	8220.62	1.068168	7733	1.004808
pmed12	6634	300	156	6994.8	1.054386	6737	1.015526
pmed16	8162	400	76	8599.2	1.053565	8258	1.011762
pmed17	6999	400	169	7374.84	1.053699	7141	1.020289
pmed21	9138	500	77	9910.58	1.084546	9384	1.026921
pmed22	8579	500	180	9087.8	1.059308	8798	1.025527
pmed26	9917	600	78	10629.2	1.071816	10093	1.017747
pmed27	8307	600	185	8841.58	1.064353	8507	1.024076
pmed31	10086	700	79	10878.82	1.078606	10370	1.028158
pmed32	9297	700	189	9822.5	1.056524	9507	1.022588
pmed35	10400	800	79	11131.86	1.070371	10586	1.017885
pmed36	9934	800	194	10446.76	1.051617	10130	1.01973
pmed38	11060	900	80	11801.52	1.067045	11198	1.012477
pmed39	9423	900	197	9990.12	1.060185	9685	1.027804