An Efficient Local Search Algorithm for k-Median Problem

RUI PAN
Shandong University
School of Computer Sci. and Tech.
No. 73 Jing Shi Road, Ji'nan 250061
P.R.China

DAMING ZHU
Shandong University
School of Computer Sci. and Tech.
No. 73 Jing Shi Road, Ji'nan 250061
P.R.China

Abstract: The *k*-median problem is one of the NP-hard combinatorial optimization problems. It falls into the general class of clustering problem and has application in the field of classification and data mining. One has confirmed that local search technique is the most effective and simplest method for designing the algorithms for *k*-median problem, and has been looking for the more efficient algorithms which can simplify the search space of the problem to solve the large-scale instance and obtain the high quality solution. In this paper, we first analyze the search space of the problem by making use of fitness distance correlation method and reveal the relation between local minima and global minima, and then we propose a more effective and efficient algorithm which gradually scales down the size of the instance based on the intersection of local minima so that the original search space is simplified and the better solution is found. Finally, elaborate experimental results attest the efficiency and computational effect of the algorithm.

Key–Words: k-Median, NP-hard, Combinatorial optimization, Local search, Approximation algorithm

1 Introduction

The *k*-median problem is one of the most famous combinatorial optimization problems. It can be stated as: given two sets of *F* and *C*, which contain *n* facilities and *m* cities respectively. For every facility $i \in F$ and city $j \in C$, there is a connection cost between facility *i* and city *j*. The objective is to open a subset $S \subseteq F$, $|S| \leq k$, where *k* is given as the input parameter, and connect every city $j \in C$ to a facility in *S* so that the total connection cost is minimized.

The *k*-median problem has occupied the central place in operations research and computer science. It falls into the general class of clustering problem and has application such as classification and ranking, data mining and web information retrieving. It is known to be the NP-hard [11], and can not be solved in polynomial time, unless P=NP, and therefore is well studied in the field of approximation algorithm [6, 10, 5, 1, 2, 13] in which a number of different algorithms have been proposed for this problem using a variety of techniques , such as LP rounding, primal-dual method, local search and combinations of these methods. However, the local search method is the most effective and simplest one among different methods. The best-known approximation algorithm for this problem proposed by Arya [1] is based on a simple local search procedure, and achieves $3 + \epsilon$ approximation ratio. Local search method can be best understood by interpreting the feasible solutions as discrete points in a high-dimensional hilly landscape, and the quality to be optimized as the corresponding potential energy. The algorithm based on local search proceeds iteratively, improving the solution by small modifications step by step. The neighborhood of current point, defined by the set of permitted modifications of the solution, is searched for points of low energy. If such a point is found, it is substituted for the current point, and a new search is started. Otherwise, the process stops because a local minimum has been reached. The computational complexity of a local search procedure depends on the size of the neighborhood and the time needed to evaluate a modification. In general, the larger the neighborhood, the more the time one needs to search and the better the local minima.

A crucial problem in local search method is the local minima. Although these solutions maybe are good enough, they are not necessarily optimal. Usually, there is almost no chance to find the optimal solution as the size of problem rises. Furthermore, if local search procedure finds a local minimum, there is no obvious way to proceed any further toward solutions of better quality. Metaheuristics [12, 15] try to remedy that. One of the methods is Repeated Local Search where local search is restarted from a new arbitrary solution every time it reaches a local minimum until a number of restarts are completed. The

best local minimum found over the many runs is returned as an approximation of the optimal solution. Modern metaheuristics tend to be much more sophisticated than repeated local search pursuing a range of objectives that go beyond simply escaping from local minima. Members of this class include Tabu Search [9], Genetic Algorithms [16, 7], Guided Local Search [17] and others.

In this paper, we first analyze the relation between local minima and global minima by experiments on instances of *k*-median problem in ORLIB [3]. The results show that the intersection of some local minima contains facilities in the corresponding optimal solution with a great probability, and the size of intersection of several local minima is usually larger than $3k/4$. Based on these results, we devise a new multilevel reduction algorithm for *k*-median problem, short for MRA. In MRA, we extract the intersection of some local minima and delete some cities connected to a facility in the intersection, then construct a new smaller instance and call the simple local search procedure to solve it, as the new instance of *k*-median problem is smaller, the local search procedure is more efficient, finally we merge the solution to the smaller instance into the intersection extracted before. Hence, the solution found by MRA is improved. The procedure proceeds iteratively until the stop criterion is satisfied. We compare MRA with Arya's algorithm [1], the best-known algorithm for *k*-median problem. The experimental results demonstrate the advantage of MRA over Arya's algorithm.

The content of this paper is organized as follows. In Section 2, we introduce a simple local search procedure for *k*-median problem, which is taken as the subroutine in our final algorithm, and instances which we used to obtain experimental results. In Section 3, we make use of the method of fitness distance correlation to analyze the experimental results obtained by solving instances in ORLIB [3] using the simple local search procedure in Section 2, and illustrate the relation between local minima and global minima for *k*-median problem. In Section 4, Multilevel reduction algorithm for *k*-median problem based on the important conclusion in Section 3 is presented. In Section 5, elaborate experimental results are listed. The concluding remarks appear in Section 6.

2 Local Search and Instances

For the convenience of description, first the definition of *k*-median problem is given formally.

Definition 1 *In k-median problem, there are a set F of n facilities and a set C of m cities, as well as an integer k,* $0 < k \leq |F|$. *There is a connection cost* $c_{ij} > 0$ *between facility* $i \in F$ *and city* $j \in C$. *The objective is to identify a subset* $S \subseteq F$, $|S| \leq k$, *to serve cities in C so that* $\sum_{j \in C} c_{ij}$ *is minimized, where* $i \in S$.

2.1 Local Search Procedure with Swaps

The simple local search procedure (LSS for abbreviation) used to solve *k*-median problem is described in Fig.1.

An operation *op* is called admissible for S if $cost(op(S)) < cost(S)$. The only operation permitted in LSS is a swap. A swap is effected by deleting a facility $s \in S$ and adding a facility $s' \in F - S$ to S. We start with an arbitrary set of *k* facilities and keep improving our solution with such swaps till it is possible to do so. At any execution of the step 2 of the LSS in Fig 1, there will be at most a polynomial number of swaps to be checked for admissibility, and each swap will take a polynomial time, therefore LSS will terminate in polynomial time.

The swap is defined formally as,

$$op(S) = S - s + s' \quad \text{for } s \in S \text{ and } s' \in F - S.$$

and denoted by $\langle s, s' \rangle$.

2.2 Testing Instances

We select three classes of instances, ORLIB [3], SL [14] and GR [8], as the test bed, which help us produce all experimental results and then draw the related conclusions.

ORLIB was introduced by Beasley [3] and contains a collection of test data sets for variety of problems in operations research. There are 40 instances for *k*-median problem in ORLIB, and each instance is a plane graph in which every vertex is a city as well as a facility and the connection cost between a facility and a city, which needs to be computed beforehand by making use of the Shortest-Path algorithm, is the

Local Search Procedure

1. $S \leftarrow$ an arbitrary solution with k facilities.
2. While $\exists$ an operation *op* such that,
   $cost(op(S)) < cost(S)$,
   do $S \leftarrow op(S)$.
3. return S.

Figure 1: Local search procedure for *k*-median problem.

length of the shortest path between corresponding vertices. The number of vertices in these instances varies from 100 to 900, and the value of *k* from 5 to 200.

SL is a slight extension to ORLIB proposed by Senne and Lorena [14]. It contains three new instances, sl700, sl800 and sl900, which are based on instances pmed34, pmed37 and pmed40 in ORLIB, respectively, except that the values of *k* are 233, 267 and 300.

GR was introduced by Galvao and Revelle [8] and first used for *k*-median problem by Senne and Lorena [14], and contains two plane graphs with 100 and 150 vertices, respectively. Every vertex is equally a city as well as a facility. The connection cost of a city to a facility has been included in the data file. The value of *k* in these instances varies from 5 to 60.

3 Analysis of Local Minima and Global Minima

Fitness landscapes have been introduced to describe the dynamics of evolutionary adaptation in nature [18] and have become a powerful concept in evolutionary theory. They are equally suited to describe the behavior of heuristic search methods in optimization. Based on the analysis of fitness landscape, researchers often identify the structure of a given problem and design highly effective search algorithms. Many properties of fitness landscapes have strong influence on heuristic search, such as the fitness differences between neighboring points in the landscape, the distribution of the local minima in the search space. Thus many methods have also been proposed to measure these properties. The fitness distance correlation (FDC) is one of the methods, and is proposed to identify the correlation between the fitness of the local minimum and the distance of the local minimum to the global minimum.

In *k*-median problem we denote the value of a solution, i.e. the total connection cost, as the fitness of the solution, and the number of different facilities between a solution and a corresponding global minimum as the distance of the local minimum to the global minimum.

To study the relation between local minima and global minima, we first conduct some experiments on instances from ORLIB. The experimental results reveal the following facts.

1) For a given instance of *k*-median problem, the distance between local minima and global minima has little relation with the fitness.

2) The local minima are found in a small fraction of the search space, and they appear to be relatively close together. The distance between a local minimum and a global minimum is nearly $k/9$, where k is the input parameter in an instance. This means 90% facilities on average in a local minimum are in the global minimum of corresponding instance. A similar result is shown in the literature [4], in which this structure is called "big valley" structure.

The above experimental conclusions motivate us to study further the relation between the intersection of some local minima and the global minimum. We first test the sizes of the intersections of local minima on some instances from ORLIB. The intersections are based on *p* different local minima, where $2 \leq p \leq 12$. For any *p*, we test 50 times and compute the mean size. Each time we use the LSS in Fig.1 to get *p* different local minima and compute the number of the common facilities in these *p* local minima. Fig.2 shows the results. The *x* axis is *p*, the number of local minima which are used to get the intersection. The *y* axis is the ratio of size of intersections to the input parameter *k* in corresponding instances. From Fig.2, we can find that the size of intersection of local minima decreases as the number of the local minima increases, and the size of the intersections is larger than $3k/4$ at $p \leq 3$.

Subsequently, we analyze the relation between the intersection of local minima and global minima. Based on the previous experimental data, we compute the percentage of optimal facilities in the intersection, i.e. those facilities in global minima. The results are shown in Fig.3. The *x* axis is *p*, the number of local minima which are used to get the intersection. The *y* axis is the percentage of optimal facilities in the intersections. Clearly, Fig.3 shows that the percentage increases and approaches 1 as the number of local minima increases.

Based on above experimental results, we get the

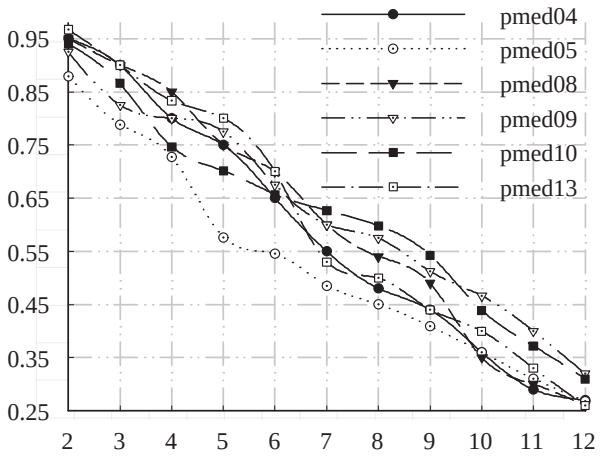

Figure 2: Size of intersection of local minima.

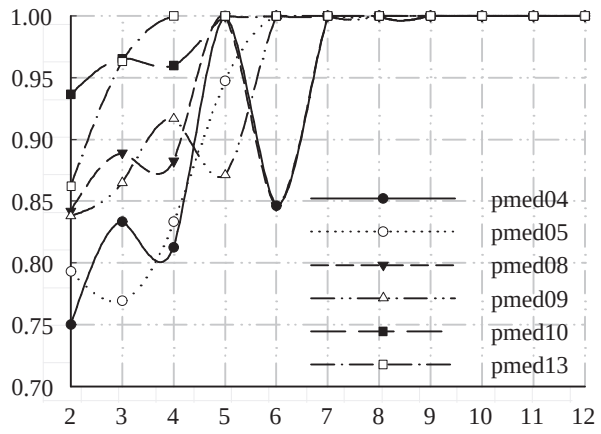

Figure 3: Percentage of optimal facilities in intersections.

fact that the intersection of some different local minima is approximately optimal and when the number of the local minima which are used to get the intersection is less than 4, the size of the intersection is very large compared with the input parameter k in corresponding instance.

4 Multilevel reduction algorithm for k-median problem

Based on the analysis in the above section, we propose a new multilevel reduction algorithm for k-median problem. In MRA we first get a subset S_p of facilities which is the intersection of p local minima of the given instance, where $p > 1$, and then construct a new instance by deleting those facilities in S_p from original set F of facilities and cities which are connected to facilities in S_p from original set C of cities. Thus we simplify the search space and are able to get a better solution S' to the small instance by using LSS. Finally we merge S_p and S' to get a feasible solution S^* to the original instance. MRA is outlined in Fig.4.

In MRA it is obvious that any partial solution S_p will not be a feasible solution to instance I of k-median problem (Strictly speaking, any subset of facilities with not more than k facilities is a solution to the given instance, but customarily one only thinks of the subset containing k facilities as the feasible solution.), however, according to the conclusions of Section 3, the facilities in S_p belong to S_{opt} with a very high probability, where S_{opt} is the optimal solution to I. Thus we denote S_p as an optimally partial solution.

When instance I is very large, the search space of I is also so rugged that a long jump in LSS can not escape from the local minimum, and therefore the final solution found is sometimes not good enough. However, the size of S_p is so large at $p \leq 3$ that the new instance I' constructed in step 2.3 in MRA is small and the search space of I' is smooth, which let the LSS be easy to escape from a local minimum and more efficient, thus finding the optimal solution S' to I' with a very high probability. Since the feasible solution S^* produced by merging S_p and S' is not necessarily a local minimum to I, therefore the LSS must be called on it again.

Multilevel Reduction Algorithm

1. Get a local minimum S_0 to instance $I = (F, C, k)$ using LSS.
2. Repeat following procedure t times, where t is specified in advance.
   - 2.1 Get $p - 1$ local minima, $S_1, S_2, \ldots, S_{p-1}$, to instance I using LSS.
   - 2.2 Get the intersection S_p, $S_p = S_1 \cap S_2 \cap \ldots \cap S_{p-1}$.
   - 2.3 Construct a new small instance $I' = (F', C', k')$ by following procedure.
     - i. $F' \leftarrow F - S_p$.
     - ii. $C' \leftarrow C - C_p$, C_p contains cities which are connected to facilities in S_p in the case of S_0.
     - iii. $k' \leftarrow k - |S_p|$.
   - 2.4 Get a local minimu S' to $I'=(F', C', k')$ using LSS.
   - 2.5 Merge S_p and S' to obtain a feasible solution $S^* = S_p \cup S'$ to the original instance I.
   - 2.6 Run LSS again to improve S^*.
   - 2.7 If S^* is better than S_0, then $S_0 \leftarrow S^*$.
3. Return S_0 as the final solution to I.

Figure 4: Multilevel reduction algorithm for k-median problem.

The theoretical analysis in this section shows that the final solution obtained by MRA is better than that of the only LSS.

5 Elaborate Experiments

To verify the effectiveness of MRA, we conduct a series of experiments on instances from ORLIB, SL and GR. The results presented in the section attempt to provide a comprehensive picture of the performance of MRA on k-median problem.

We define the percentage of error of algorithm A as follows.

$$A\_Err\%(I) = \frac{A(I) - OPT(I)}{OPT(I)} \times 100\%,$$

where I is an instance of k-median problem, and $A(I)$

and $OPT(I)$ are the value of the solution obtained by algorithm *A* and the corresponding optimal value, respectively. The improvement percentage of algorithm *A* over algorithm *B* is defined as:

$$A\_IP\%(I) = (B\_Err\%(I) - A\_Err\%(I)) \times 100\%.$$

The bigger $A\_IP\%(I)$ is, the more remarkable the performance improvement of algorithm *A* is.

According to the conclusions in Section 3, the parameter *p*, which is the number of local minima used to obtain the intersection in MRA, is very important, and will affect the quality of final solution. To guarantee the intersection S_p contains more facilities in the optimal solution with a high probability and the new instance I' constructed in MRA is small enough, we take $p = 3$. For $p > 3$, the running time for getting the intersection S_p is too long so that we do not consider that case.

We also implement the best-known algorithm for *k*-median problem, Arya's local search algorithm [1], short for LSA, to compare with MRA and demonstrate the effectiveness of our algorithmic idea.

All instances are tested on a Personal Computer with Pentium IV-2.8 Ghz processor and 1G RAM. The codes of MRA and LSA have been written in Java and compiled with the version of JDK 1.5.0_03-b07.

The elaborate experimental results are listed in Table 1, in which there are three parts. First part successively indicates the name of instance, the numbers of facilities and cities, *k* and optimal value. The second part indicates the mean value of the solution and the mean percentage of error on 50 times test for each instance, obtained by LSA and MRA, respectively. The third part is improvement percentage of MRA over LSA.

The experimental results are remarkable, and demonstrate that MRA has a great performance improvement compared with LSA and significantly outperforms LSA. The mean percentage of improvement of MRA is 4.511%.

6 Conclusion

In this paper for *k*-median problem we first study the relation between local minima found by a simple local search procedure and global minima by using the FDC analysis method and corresponding experimental results. Based on the conclusions, we propose a new effective multilevel reduction algorithm for *k*-median problem. The elaborate experiments on instances from three famous libraries, ORLIB, SL and GR, also further confirm our algorithmic idea, and demonstrate that our algorithm is superior to the best-known algorithm proposed by Arya [1].

Acknowledgements: The research was supported by the Natural Science Foundation of China (grant No. 60273032, 60573024).

References:

[1] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala and V. Pandit, Local Search Heuristics for k-Median and Facility Location Problems, In: *Proc. of the 33rd Annual ACM Symp. on Theory of Computing.* 2001, pp. 21–29.

[2] M. Badoiu, S. Har-Peled and P. Indyk, Approximate Clustering via Core-Sets, In: *Proc. of the 34th Annual ACM Symp. on Theory of Computing.* 2002, pp. 250–257.

[3] J.E. Beasley, A Note on Solving Large p-Median Problems, *European Journal of Operational Research.* 21, 1985, pp. 270–273.

[4] K. Boese, Cost Versus Distance in the Traveling Salesman Problem, *Technical Report, TR-950018, UCLA CS Department.* 1995.

[5] M. Charikar, S. Guha, Improved combinatorial algorithms for the facility location and k-Median problems, In: *Proc. of the 40th Annual Symp. on Foundations of Computer Science.* 1999, pp. 378–388.

[6] M. Charikar, S. Guha, E. Tardos and D. Shmoys, A Constant-Factor Approximation Algorithm for the k-Median Problem (Extended Abstract), In: *Proc. of the 31th Annual ACM Symp. on Theory of Computing.* 1998, pp. 1–10.

[7] B. Freisleben, P. Merz, New Genetic Local Search Operators for the Traveling Salesman Problem, In: *Proc. the 4th International Conference on Parallel Problem Solving from Nature-PPSN IV.* 1996, pp. 890–900.

[8] R.D. Galvao, C.S. Revelle, A Lagrangean Heuristic for the Maximal Covering Problem, *European Journal of Operational Research.* 18, 1996, pp. 114–123.

[9] F. Glover, Tabu Search Part II. ORSA, *Journal of Computing.* 2, 1990, pp. 4–32.

[10] K Jain, V. Vazirani, Primal-Dual Approximation Algorithms for Metric Facility Lcation and k-Median Problems, In: *Proc. of the 40th Annual Symp. on Foundations of Computer Science.* 1999, pp. 2–13.

[11] O. Kariv, L. Hakimi, An Algorithmic Approach to Network Location Problem, part II: the p-medians, *SIAM Journal of Applied Mathematics.* 37, 1979, pp. 539–560.

Table 1: The performance comparisons between MRA and LSA

Instances				Mean Value of Sol.		Mean Err%		MRA_IP%
Name	Nos. of F&C	k	OPT	LSA	MRA	LSA_Err%	MRA_Err%	
pmed02	100	10	4093	4276	4093	4.471%	0.000%	4.471%
pmed03	100	10	4250	4474	4250	5.271%	0.000%	5.271%
pmed04	100	20	3034	3286	3034	8.306%	0.000%	8.306%
pmed05	100	33	1355	1453	1358	7.232%	0.221%	7.011%
pmed07	200	10	5631	6103	5633	8.382%	0.036%	8.347%
pmed08	200	20	4445	4776	4445	7.447%	0.000%	7.447%
pmed09	200	40	2734	2900	2740	6.072%	0.219%	5.852%
pmed10	200	67	1255	1387	1256	10.518%	0.080%	10.438%
pmed12	300	10	6634	6849	6634	3.241%	0.000%	3.241%
pmed13	300	30	4374	4728	4377	8.093%	0.069%	8.025%
pmed14	300	60	2968	3251	2970	9.535%	0.067%	9.468%
pmed15	300	100	1729	1870	1733	8.155%	0.231%	7.924%
pmed17	400	10	6999	7289	7004	4.143%	0.071%	4.072%
pmed18	400	40	4809	5164	4809	7.382%	0.000%	7.382%
pmed19	400	80	2845	3083	2853	8.366%	0.281%	8.084%
pmed20	400	133	1789	1919	1792	7.267%	0.168%	7.099%
gr100	100	15	3893	3933	3902	1.027%	0.231%	0.796%
	100	20	3565	3606	3569	1.150%	0.112%	1.038%
	100	25	3291	3325	3294	1.033%	0.091%	0.942%
	100	30	3032	3064	3034	1.055%	0.066%	0.989%
	100	35	2784	2807	2784	0.826%	0.000%	0.826%
	100	40	2542	2557	2544	0.590%	0.079%	0.511%
gr150	150	15	7390	7467	7406	1.042%	0.217%	0.825%
	150	20	6454	6628	6471	2.696%	0.263%	2.433%
	150	25	5875	6098	5903	3.796%	0.477%	3.319%
	150	35	5192	5311	5195	2.292%	0.058%	2.234%
	150	45	4636	4722	4646	1.855%	0.216%	1.639%
	150	50	4374	4440	4381	1.509%	0.160%	1.349%
	150	60	3873	3904	3875	0.800%	0.052%	0.749%
sl700	700	233	1847	1933	1895	4.656%	2.599%	2.057%
sl800	800	267	2026	2176	2068	7.404%	2.073%	5.331%
sl900	900	300	2106	2290	2145	8.737%	1.852%	6.885%
								Ave.4.511%

[12] I.H. Osman, *An Introduction to Meta-heuristics. Operational Research Tutorial Papers*, Operational Research Society Press, Birming-ham, UK, 1995, pp. 92–122.

[13] R. Pan, D.M. Zhu, S.H. Ma and J.J. Xiao, Approximated Computational Hardness and Local Search Approximated Algorithm Analysis for k-Median Problem, *Journal of Software.* 16, 2005, pp. 392–399.

[14] E.L.F. Senne, L.A.N. Lorena, Langrangean/Surrogate Heuristics for p-Median Problems, In: *Computing Tools for Modeling, Optimization and Simulation: Interfaces in Computer Science and Operation Research.* 2000, pp. 115–130.

[15] E.G. Talbi, A Taxonomy of Hybrid Metaheuristics, *Journal of Heuristics.* 8, 5, 1995, pp. 541–564.

[16] H.K. Tsai, J.M. Yang and C.Y. Kao, Solving Traveling Salesman Problems by Combining Global and Local Search Mechanisms, In: *Proc. the Congress on Evolutionary Computation.* 2002, pp. 1290–1295.

[17] C. Voudouris, E. Tsang, Guided Local Search and its Application to the Traveling Salesman Problem, *European Journal of Operational Research.* 113, 1999, pp. 469–499.

[18] S. Wright. The Roles of Mutation, Inbreeding, Cross-breeding, and Selection in Evolution, In: *Proc. the 6th Congress on Genetics.* 1932, pp. 356–366.