Hybrid k-mean GRASP for Partition based Clustering of Two Dimensional Data Space as an Application of p-median Problem

D.Srinivas Reddy[1*], A.Govardhan[2], SSVN Sarma[3]

Abstract— The most wide spread facility location paradigm is the p-median problem. It is a paradigm known NP-Hard and combinatorial optimization problem. In different application areas the facility location modeling is extensively used. It is used in marketing to analyze customers and for network establishment in cellular tower arrangement which serves maximum clients, in computer networks and in many other areas. Metaheuristic plays an important role in many areas like Operations Research, Algorithm analysis, Data Mining etc. In this paper a new clustering algorithm k-Mean-GRASP (Greedy Randomized Adaptive Search Procedure) is proposed, which determines the number of clusters of user choice similar to k-means, and follows Metaheuristic approach. Generally, Metaheuristic is a two phase iterative method. So, the proposed algorithm is also encompassed with two phases. First phase ascertains the cluster of user specified number using k-means algorithm. At this stage the resultant cluster is considered as a best cluster. The second phase strives for the improvement of the cluster so obtained in the first phase. In the proposed work the first phase is termed as Construction phase which makes use of k-Means algorithm and the second phase as Enhancement phase. Our empirical results put forward that the proposed k-Mean-GRASP clustering algorithm outperforms the other methods. Clustering is the process of dividing the points into similar groups. The proposed method can also be used as a clustering algorithm based on the nature of the p-median problem.

Index Terms— Clustering, GRASP, Metaheuristic, Data mining, Two dimensional data space, Construction phase and Improvement phase.

1 INTRODUCTION

THE p-median problem can be described as : Given the set of facilities F and C a set of customers and d is a distance function from C x F to R which estimates the distance between a customer and a facility. For any positive integer p and number of facilities n, where p ≤ n, the p-median problem ascertains a subset R of facilities F such that |R| = p so that the sum of the distances from each customer to its adjoining facility is minimized. Here in this work we take into account of the assumption F = C, that is every customer location can be considered as a facility, and for giving equal importance to each location it is considered that $w_i = 1$. Mathematically p-median problem is stated as [1].

$$Minimize f(d,x) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i d_{ij} x_{ij} \quad (1)$$

$$subject\ to \quad \sum_{j=1}^{n} x_{ij} = 1 \qquad \forall i \quad (2)$$

$$x_{ij} \le y_i \qquad \forall i, j \quad (3)$$

$$\sum_{j=1}^{n} y_i = p \quad (4)$$

$$x_{ij} = 0 \ or \ 1 \qquad \forall i, j \quad (5)$$

$$y_i = 0 \ or \ 1 \quad \forall j \quad (6)$$

Where,

n = number of locations
x_{ij} = 1 if a location i is assigned to facility located at j,
= 0 other wise
y_i = 1 if j th location is a facility
= 0 other wise
d_{ij} = distance measured from location i to location j
p = preferred number of locations as facilities

In the above definition, objective function (1) minimizes the sum total of the distances between the customer locations and the preferred number of locations. Constraint given in (2) assures that each location is allotted to exactly one nearest facility. Constraint given in (3) forbids the assignment of a customer locations to a facility that was not selected as a desired location. The constraint (4) describes the total number of preferred locations as p and finally the Constraints (5) and (6) assures that x and y are binary valued. Since the solution of the p-median problem partitions the solution space we can classify the given space as groups and hence we use the p-median problem as a clustering technique. In this paper a metaheuristic method k-Mean-GRASP is proposed to solve the p-median problem which is discussed in detail in section 2.

The arrangement of similar objects into different groups is

1. *Asst.Prof., Dept.of Computer Science & Eng., Vaageswari College of Engineering,Karimnagar, India – 505481,PH-+919989171133. E-mail: srinivasreddydhava@gmail.com*
2. *Professor, Dept. of Computer Science & Eng., Jawaharlal Nehru Technological University Hyderabad, India, PH-+919963045551. E-mail: govardhan_cse@yahoo.co.in*
3. *Professor, Dept. of Computer Science & Eng., Vaagdevi College of Engineering, Warangal, India. E-mail: ssvn.sarma@gmail.com*

known as Clustering, i.e., the segregation of information into subsets (clusters), so that the elements of each subset possess some common mannerism. Data clustering is an important Data Mining task and is a general mechanism for analyzing statistical data, which is used in several other areas; such as mechanical process industries, machine learning, pattern recognition, image analysis and bioinformatics. Metaheuristics embody a principal category of approximate practice for decipher of hard combinatorial optimization problems, for simplification of which the use of exact methods is impractical. There are several general purpose high-level procedures that can be instantiated to explore the solution space of a specific optimization problem efficiently. Earlier, metaheuristics, like genetic algorithms, tabu search, simulated annealing, ant systems, GRASP, and others, have been introduced and are applied to real-life problems in several areas of science [2]. Many optimization problems [3] are successfully applied to solve the GRASP (Greedy Randomized Adaptive Search Procedures) metaheuristic [4, 5]. The search process for identifying the solution employed by GRASP is iterative and each pass consists of two phases: construction and enhancement phase. In the construction phase a feasible solution is built, and then its neighborhood is determined by the enhancement phase to find an improved one. The outcome is the paramount solution found over all iterations.

The paper is structured as follows: In section 2, the proposed k-Mean-GRASP algorithm and its phases – Construction and Enhancement phases are described in detail. In section 3, experimental results and comparisons of cluster quality and execution times are anticipated. Section 4 provides the conclusions.

2 THE GRASP METAHEURISTIC

Many optimization problems had utilized GRASP [6] and got fruitful results [3]. It consists of two- phase which are applied iteratively. The construction phase is the first phase of GRASP in which an absolute solution is built. Because this absolute solution is not assured to be locally optimal, a new enhancement phase is applied in the second phase. This process is repeats until a annihilation measure is reached and the best solution found over all passes is taken as final result. The k-Mean-GRASP process logic is exemplified in Figure 1. Initially, the variable to hold the best solution originate is initialized. Then the construction phase is executed and then the enhancement phase is applied to the constructed solution. The quality of the obtained solution is compared to the current best solution found and, if necessary, the best solution is updated. At last the best solution is returned.

A latest approach in metaheuristic research is the investigation of hybrid metaheuristic [7]. One such hybrid methods consequences from the amalgamation of concepts and strategies from two or more metaheuristics and another one counterparts to metaheuristics pooled with concepts and procedures from other areas responsible for performing specific tasks that can improve the original method. The hybridization of GRASP with data mining process initially proposed, introduced and adopted to the set packing problem [8,9,10,11,12,13].

```
Procedure k-Mean-GRASP()
1.  Initialize best_sol as ∅
2.  repeat
3.   sol ← k-Means(data points);
4.   best_sol ← Enhancement(sol);
5.  if cost(sol) > cost(best_sol)
6.   best_sol ← sol;
7.  end if
8.  until Termination criterion;
9.  return best_sol;
```

Figure 1. k-Mean GRASP procedure

```
Procedure k-means (data points)

1. Initialize k-points as cluster centers
2. Assign each data point to the nearest
   cluster center
3. Recompute the cluster centers for each
   cluster as the mean of the cluster.
4. Repeat steps 2 and 3 until there is not
   any more change in the value of the
   means
```

Figure 2. k-means algorithm

The logic behind GRASP construction phase is described in Figure 2, which is k-means itself in the proposed work. One of the most popular heuristics for solving clustering problems is the k-means clustering algorithm. The algorithm segregates the data into k disjoint groups (clusters). The hub of each cluster is labeled as centroid. It partitions the objects so as to minimize the sum total of the squared distances between the centroid of the clusters and their objects [14

The basic logic of the GRASP Enhancement is presented in Figure 3. At first initialize control variables. The function cost_eval() appraises the expenditure of a solution by figuring out the aggregation of the distances amongst all customers and their closest facilities. After that the neighborhood of the contemporary solution is convened and if any better one found, it befall into the current one. Then the same process repeats again, till no further improvement is made. It is iterated p times. Next the best solution found is returned. In each iteration, solitary element ri of the solution is replaced by all elements nearby to it in its partition (cluster) Pi. Here it is assumed that an element e is close by r_i in its partition Pi if the distance among e and ri is fewer or equal to the average of distances between ri and all elements in Pi.

To reduce the computational endeavor of the enhancement phase, the resolution obtained in each exchange is approximately assessed and the best one is exactly appraised. The function approx_cost_eval() estimates the expense of a solu-

tion approximately by recalculating the distances within the partition Pi only, without making this calculation inside the other partitions. Since there was a change of location it may be necessary for the exact reckoning. Then it is assessed for better solution than the current one. If there exists a better one than the current one, the new one turn out to be the current solution and the enhancement phase incepted again.

```
Procedure Enhancement (sol)
1.  best_sol ← sol;
2.  best_cost ← cost_eval( sol);
3.  repeat
4.  no_improvements ← true;
5.  for i = 1 to p
6.  app_best_sol ←∅;
7.  app_best_cost ← ∞;
8.  for each element e in Pi close to ri
9.  appsol ← exchange(best_sol, ri, e);
10. appcost ← appcosteval(app_sol);
11. if approxcost < approxbestcost  then
12.  appbestsol ← appsol;
13.  appbestcost ← appcost;
14. end if
15. end for
16. exactsolcost ← costeval(appbestsol);
17. if exactsolcost < bestcost  then
18.  best_sol ← appbestsol;
19.  bestcost ← exactsolcost;
20. noimprovements ← false;
21. end if
22. end for;
23. until noimprovements;
24. return bestsol;
```

Fig. 3. Proposed Enhancement Phase used in k-Mean-GRASP procedure

3 Experimental Results

The experimental results acquired for k-GRASP and k-MEANS are presented in this section, and the results are compared on the bases of solution quality against k. Experiments are conducted on data sets with 50, 75, 100 points. Results are tabulated and graphs are plotted. The data sets under study are taken from the web site of Professor Eric Taillarrd, University of Applied Sciences of Western Switzerland. The companion website for p-median problems instances is http://mistic.heigvd.ch/taillard/problemes.dir/location.html . Here quality of the solution (cluster) i.e., sum of the distances from each customer location to its closest facility (cluster center) is measured for both k-means algorithm and the hybridized k-mean-GRASP algorithm.

In Fig.4 solution/cluster quality is compared using both algorithms k-Means and k-Mean-GRASP for the data set of size 50 with number of facility locations (cluster centers) incremented by 5. In Fig.5 solution/cluster quality is compared using both algorithms k-Means and k-Mean-GRASP for the data set of size 75 with number of facility locations (cluster centers) incremented by 10.

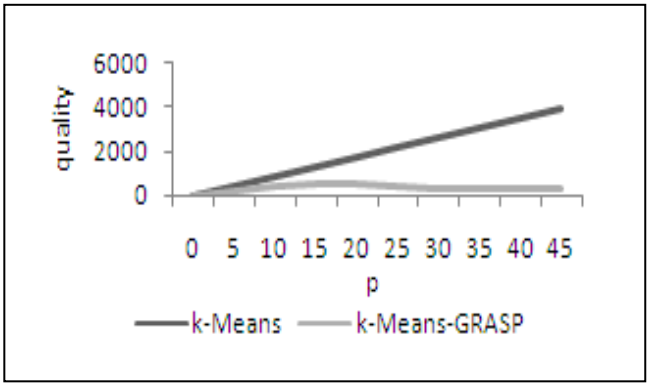

Fig.4. k-Mean Vs K-Mean GRASP

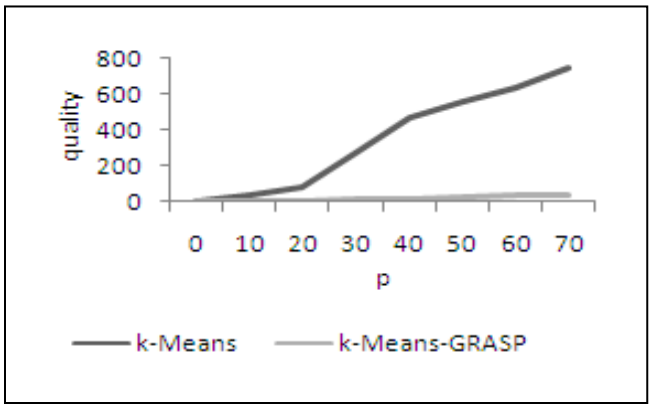

Fig.5. k-Mean Vs k-Mean GRASP

In Fig.6 solution/cluster quality is compared using both algorithms k-Means and k-Mean-GRASP for the data set of size 100 with number of facility locations (cluster centers) incremented by 10. In all cases proposed k-means-GRASP clustering algorithm outperforms k-means because of the simple and powerful enhancement phase which checks the combinations (exchanges) for optimal solution.

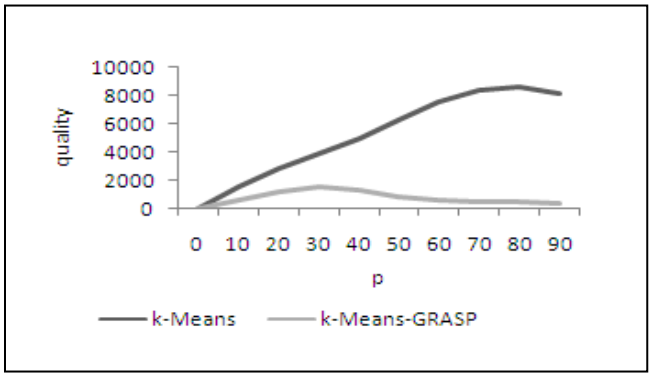

Fig.6. k-Mean Vs k-Mean GRASP

4 CONCLUSIONS

It is observed that k-Mean-GRASP outperforms k-means in quality aspect than k-means because in the Enhancement phase of k-Means-GRASP, exchanges are made for improved clusters. Since solution of p-median problem partitions the data space given to it with respect to the user specified p number of locations as facilities by minimizing aggregation of the separation from each location (customer) to its nearest facility. By assuming, each facility as a cluster center and the nearest locations as cluster elements we can apply proposed k-Mean-GRASP algorithm to simplify p-median problem as a partition based clustering algorithm for two dimensional space.

REFERENCES

[1] Moh'd Belal Al-Zoubi, Ahmed Sharieh, Nedal Al-Hanbali and Ali Al-Dahoud, "A Hybrid Heuristic Algorithm for Solving the P-Median Problem", *J.of Computer Science*(Special Issue) pp.80-83, Science Publications,2008.

[2] I. Osman and G. Laporte, "Metaheuristics: A bibliography", *Annals of Operations Research*, 63, pp. 513-623, 1996.

[3] T. A. Feo and M. G. C. Resende, *GRASP: An annotated bibliography, Essays and Surveys in Metaheuristics*, Kluwer Acadamic Publishers, 2002.

[4] T. A. Feo and M. G. C. Resende, "A probabilistic heuristic for a computationally difficult set covering problem", *Operational Reseach Letters*, vol. 8 , pp.67-71,1989.

[5] T. A. Feo and M. G. C. Resende, "Greedy randomized adaptive search procedures", *J.of Global Optimization*, vol.6, pp. 1609-1624, 1995.

[6] M. G. C. Resende and C. C. Ribeiro, *Greedy randomized adaptive search procedures, Handbook of Metaheuristics*, Kulwer Academic Publishers, 2003.

[7] E. G. Talbi, "A taxonomy of hybrid metaheuristics", *J. of Heuristics*, vol. 8, pp.541-564,2002.

[8] M. H. F. Ribeiro, V. F. Trindade , A. Plastino and S. L. Martins, "Hybradization of GRASP metaheuristic with datamining techniquess", *Proc. of the ECAI Workshop on Hybrid Metaheuristics*, pp.69-78,2004.

[9] M. H. F. Ribeiro, V. F. Trindade, A. Plastino and S. L. Martins, "Hybrazation of GRASP Metaheuristic with data mining techniques", *J. of Mathematical Modelling and Algorithms*, vol. 5, pp.23-41, 2006.

[10] Alexndre Plastino, Eric R Fonseca, Richard Fuchshuber, Simone de L Martins, Alex.A.Freitas, Martino Luis and Said Salhi, "A Hybrid Datamining Metaheuristic for the p-median problem", *Proc. of SIAM Journal* (Data Mining), 2009.

[11] L. F. Santos, M. H.F. Ribeiro, A. Plastinoand S. L. Martins, "A hybrid GRASP with data mining for the maximum diversity problem", *Proc. of the Int. Workshop on Hybrid Metaheuristics*, LNCS 3636, pp. 116-127, 2005.

[12] L.F. Santos, C.V.Albuquraque, s. L. Martins and A. Plastino, "A hybrid GRASP with data mining for efficient server replication for reliable multicast", *Proc. of the IEE GLOBECOM Conf.*, 2006.

[13] G. Grahne and J. Zhu, "Efficiently using prefix-trees in mining frequent item-sets", *Proc. of the IEEEICDM Workshop on Frequent Itemset Mining Implementations*, 2003.

[14] N. Mladenovic, J. Brimberg, P. Hansen and Jose A. Moreno-Perez, "The p-median problem: A survey of metaheuristic approaches", *European J. of Operational Research*, vol. 179 , pp.927-939,2007.