META-HEURISTIC BASED CLUSTERING OF TWO-DIMENSIONAL DATA USING NEIGHBOURHOOD SEARCH WITH DATA MINING TECHNIQUE AS AN APPLICATION OF P-MEDIAN PROBLEM

1D.Srinivas Reddy, 2Dr A.Govardhan, 3S.S.V.N Sharma
1Vaageswari College of Engineering, Karimnagar, India
2JNTUH University, Director of Evaluation, Hyderabad, India
3Director of Vaagdevi College of Engineering, India

ABSTRACT

The p-median difficulty is the well known facility allocation essence. The neighborhood search (NS) algorithm provides a solution to that problem. This NS algorithm is hybridized with data mining (DM) technique to attain a solution that is better than NS approach [1]. Two new Metaheuristic clustering algorithms HDMNS (Hybrid Data Mining based Neighborhood Search) and HMDMNS (Hybrid Multiple Data Mining based Neighborhood Search) are proposed schemes which are hybrid versions of NS algorithm [10, 11 and 17]. The clustering technique is the process of dividing the points into related groups. The proposed NS with DM technique method can also be used as a clustering algorithm based on the nature of the p-median problem. The proposed algorithm has two phases; First phase in both algorithms constructs the basic solution using the NS approach. In HDMNS approach DM technique is used to locate the improved solution which gives the better clustering of two-dimensional data space. In HMDMNS the second phase is also uses the DM technique to get the best quality cluster based on the idea to get better results with mining if exists and the results are compared with the well known K-means clustering algorithm, results proves that both the methods out performs k-means algorithm.

Index Terms- HDMNS (Hybrid Data Mining based Neighborhood Search), HMDMNS (Hybrid Multiple Data Mining based Neighborhood Search, neighborhood search (NS), Data mining (DM), Clustering
I. INTRODUCTION

The p-median problem which is a combinatorial optimization problem in natural world is NP-Hard that realizes facilitators that serve the maximum locations [7, 8]. The p-median problem will be practical in several applications such as growing marketing strategies in the realm of management sciences and locating server positions in computer networks (CNs) [12]. The use the p-median problem as a clustering technique. In this work a metaheuristic method k-Mean-GRASP is proposed to solve the p-median problem which is discussed in detail. The arrangement of similar objects into different groups is 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 [15]. 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. Many optimization problems are successfully applied to solve the GRASP (Greedy Randomized Adaptive Search Procedures) metaheuristic. 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 neighbourhood is determined by the enhancement phase to find an improved one. The outcome is the paramount solution found over all iterations.

Procedure-Mean-GRASP()

1. Initialize best_sol as ∅
2. repeat
3. sol βk-Means(data points);
4. best_sol βEnhancement (sol);
5. if cost(sol) > cost(best_sol)
6. best_sol βsol;
7. end if
8. until Termination criterion;
9. return best_sol;

Procedure-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 1: K-Mean GRASP Procedure

Figure 2: K-Means algorithm

The NS technique reveals all possible combinations with the elements in the neighbourhood of individual elements in the solution and determine the optimal solution i.e., which serves the maximum locations so that the sum of the total distance from the each element to the facilities is minimized. The NS Approach Metaheuristic is an iterative one which contains two phases [16]. Construction phase is the first phase that structure the initial solution and based on the initial solution the second phase gawk at for the optimal solution based on NS approach and then the feasible solution space is computed to obtain the optimal
solution. The NS Approach mechanism is illustrated in the Figure 3. It is having two phases: the construction phase and NB Search phase.

**Procedure NSApproach ()**
1. optml_sol ← ∅
2. sol ← Construction(data points);
3. best_sol ← NBSearch(sol);
4. if cost(sol) > cost(optml_sol)
5. optml_sol ← sol;
6. end if
7. until Termination criterion;
8. return optml_sol;

**Procedure HDMNS ()**
1. Initialize sol, optimal_sol, Mined_sol ← ∅
2. Initialize sol_space ← ∅
3. Initialize USS, FIS, UFIS ← ∅
4. Read list, p
5. Sol←NSApproach(list, p)
6. USS←Update_sol_space (sol)
7. FIS←Generate_frequent_items(USS)
8. UFIS←Update_supportcount(FIS, USS)
9. Mined _sol←Generate _mined _solution(UFIS)
10. Update optimal_sol

**Figure 3: NS Approach procedure**

**Figure 4: Hybridized Data Mining NS algorithm**

## II. HYBRIDIZED DATA MINING NEIGHBOURHOOD SEARCH APPROACH AND HYBRID MULTIPLE DATA MINING BASED NEIGHBORHOOD SEARCH

The HDMNS () procedure shown in the Figure 4 consists of three phases. The first phase NS Approach (), computes the initial solution using NS approach by considering the given list and user specified ‘P’ [18]. The second phase is the application of DM technique to the basic realistic solution. The obtained solution in first phase is input for the second phase which is used to generate the updated solution space (USS) that consists of all the probable solutions generated using the result obtained in the NSApproach (). Based on USS, the Frequent item set (FIS) are generated which consists the set of all individual items that are present in the USS [4, 5 and 9]. Currently, support count is calculated and updated for each item in the FIS. After updating support count, sort the frequent items in the decreasing order of support count and then the mined solution is constructed by deliberating the items with high support count until the size of the solution or the number of items exactly equals to ‘P’. The final phase is used to obtain the optimal solution. The optimal solution is updated using enhancement phase described in NSApproach which examines the global optimal solution that optimizes the objective of the given p-median problem [13, 14]. HMDMNS described in the Figure 5, is the Metaheuristic based on Neighbourhood Search (NS) hybridized with Data Mining Technique (HDMNS) multiple times iteratively with Frequent Mining to provide an updated solution to p-median problem if exits in each iteration [6]. The resulting local optimal solution from NS method serves as a basis for identification of practical solution space that holds different possible solutions of similar size and by the application of frequent mining technique on it results in identification of frequent items. Basing on the support count, most practical solution is identified. The frequent mining technique is applied only one time in HDMNS and yields a better solution. In HMDMNS it is applied constantly several times based on the inspiration that on applying the mining technique one time derives a better solution. Then applying the same over and over again will definitely yield optimal one.
III. HDMNS AND HDMDNS APPROACHES AS CLUSTERING METHODS, K-MEANS

Of the most popular heuristics for solving clustering problems by k-means clustering Algorithm (CA) which is simple and widely used [2]. This algorithm segregates the data into k disjoint clusters. The center 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. The k-means algorithm is described in the Figure 6. The cluster quality computed for k-means is same as p-median problem objective function value. So, the solution approaches proposed for p-median can also be considered as new clustering mechanisms which are efficient than k-means. And the cluster quality estimation for the k-means algorithm is the squared error criterion is given by $E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$ where E is the sum of square-error for all objects in the database, ‘P’ is the point in the space representing the given object, and $m_i$ is the mean of the cluster $C_i$. This criterion makes the resulting k clusters as solid.

IV. 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 Taillard, 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 Figure 7 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 Figure 8 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.
The experimental results obtained for NS approach, HDMNS and HMDMNS are analyzed in this section, and the results are evaluated on the basis of quality of the solution against ‘P’. Experiments are carried out on data sets with 15, 25, 50 and 75 points. Results are tabularized and graphs are outlined. The origin of data sets under study is acquired from the web site of Professor Eric Taillard, Kent University of Applied Sciences of Western Switzerland. The associated website for p-median problem instances is http://mistic.heig-vd.ch/taillard/problemes.dir/location.html. In Graph-1 the cluster quality is compared for algorithms HDMNS, HMDMNS and k-means algorithms with intervals 5 for p and n=25. It is identified that HMDMNS is resulting with good quality cluster than the other two.
The same is observed in the remaining graphs plotted for cluster quality with n=50, 15 and 75 with p=15, 3 and 20 respectively in Graphs:2, 3 and 4. Execution times are also compared for all the three algorithms for p=5, 15 and 3 and identified that HMDMNS can also be used as a better clustering algorithms than existing k-means algorithm. It is depicted in Graphs 5 and 6.

V. CONCLUSION

It is observed that in all the test cases, Hybrid Multiple Data Mining Neighbourhood Search (HMDMNS) Metaheuristic performs much better as a clustering algorithm (CA) when compared with the efficient existing method k-means. So instead of using k-means; HMDMNS can be used as a new clustering mechanism.

REFERENCES


