Complementary Approaches for a Clustering Problem

Edson L.F. Senne*, Luiz A.N. Lorena**

*FEG/UNESP - Universidade Estadual Paulista
Faculdade de Engenharia - Departamento de Matemática
12500-000 Guaratinguetá, SP - Brazil
e-mail: elfsenne@feg.unesp.br
**LAC/INPE - Instituto Nacional de Pesquisas Espaciais
Av. dos Astronautas, 1758 - Caixa Postal 515
12227-010 São José dos Campos, SP - Brazil
e-mail: lorena@lac.inpe.br

Abstract

The lagrangean/surrogate relaxation has been explored as a faster computational alternative to traditional lagrangean heuristics. This paper discusses two approaches for using lagrangean/surrogate heuristics to a classical clustering problem: the p-median problem, that is, how to locate p facilities (medians) on a network such as the sum of all the distances from each vertex to its nearest facility is minimized.

Key words: Linear Programming, Networks, Optimization Techniques, Heuristics.

1. Introduction

The search for p-median nodes on a network is a classical clustering problem. The objective is to locate p facilities (medians) such as the sum of the distances from each demand point to its nearest facility is minimized. The problem is well known to be NP-hard (Garey and Johnson, 1979) and several heuristics have been developed for p-median problems. The combined used of lagrangean relaxation and subgradient optimization in a primal-dual viewpoint was found to be a good solution approach to the problem (Beasley, 1985). By other hand, column generation is a powerful tool for solving large scale linear programming problems and has been a natural choice in several applications, such as the well-known cutting-stock problem, vehicle routing and crew scheduling (Gilmore and Gomory, 1961, 1963; Desrochers and Soumis, 1989; Desrochers *et al*., 1992; Vance *et al*., 1994). However, the use of column generation to solve p-median problems was not sufficiently explored. The initial attempts appear to be the ones of Garfinkel *et al*. (1974) and Swain (1974), that report convergence problems, even for small instances, when the number of medians is small compared to the number of candidate points in the network. This observation was also confirmed later by Galvão (1981). The solution of large-scale instances using a stabilized approach is reported by du Merle *et al*. (1999).

The lagrangean/surrogate relaxation has been explored recently as a faster computational alternative to traditional lagrangean heuristics. This new relaxation has been used in a number of applications, such as multidimensional knapsack problems (Freville *et al*., 1990), set covering problems (Lorena and Lopes, 1994), generalized assignment problems (Lorena and Narciso, 1996), and also p-median problems (Senne and Lorena, 2000).

The lagrangean/surrogate relaxation and the traditional column generation approach can be combined to accelerate and stabilize primal and dual bounds obtained using the reduced cost selection. It is well known the equivalencies of the Dantzig-Wolfe decomposition (Dantzig and Wolfe, 1960), column generation and lagrangean relaxation optimization. Solving a linear programming by Dantzig-Wolfe decomposition is the same as solving the lagrangean by Kelley's cutting plane method (Kelley, 1960). However, as observed before, in many cases a straightforward application of column generation may result in slow convergence. The lagrangean/surrogate relaxation can be used to stabilize and accelerate the column generation process, providing the selection of new productive columns.

This paper discusses these two approaches of using lagrangean/surrogate relaxation to solve p-median problems: (i) combined with subgradient optimization in a primal-dual viewpoint, and (ii) combined with the column generation process for linear programming problems. The paper is organized as follows. Section 2 presents the lagrangean/surrogate relaxation combined with subgradient optimization heuristic. Section 3 discusses the combination of lagrangean/surrogate relaxation and column generation process. The computational tests, conducted with problems from the literature, are presented in the Section 4. We conclude that these two approaches of using lagrangean/surrogate relaxation to solve p-median problems are complementary and can be put to work together.

2. The Lagrangean/Surrogate Relaxation and Subgradient Optimization

The p-median problem can be modeled as the following binary integer programming problem:

$$v(P) = \min \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ij}$$

$$\text{subject to} \quad \sum_{i=1}^{n} x_{ij} = 1; \; j \in N \qquad (1)$$

$$\sum_{i=1}^{n} x_{ii} = p \qquad (2)$$

$$x_{ij} \le x_{ii}; \; i, j \in N \qquad (3)$$

$$x_{ij} \in \{0,1\}; \; i, j \in N \qquad (4)$$

where: $[d_{ij}]_{nxn}$ is a symmetric cost (distance) matrix, with $d_{ii} = 0$, $\forall i$; $[x_{ij}]_{nxn}$ is the allocation matrix, with $x_{ij} = 1$ if node i is allocated to node j, and $x_{ij} = 0$, otherwise; $x_{ii} = 1$ if node i is a median and $x_{ii} = 0$, otherwise; p is the number of facilities (medians) to be located; n is the number of nodes in the network, and N = {1, ..., n}. Constraints (1) and (3) ensure that each node j is allocated to only one node i, which must be a median. Constraint (2) determines the exact number of medians to be located (p), and (4) gives the integer conditions.

For a given $\lambda \in R_{+}^{n}$ and $t \ge 0$, the lagrangean/surrrogate relaxation of (P) can be defined by:

$$v(L_tSP^{\lambda}) = \min \sum_{j=1}^{n}\sum_{i=1}^{n}(d_{ij} - t\lambda_j)x_{ij} + t\sum_{j=1}^{n}\lambda_j$$

subject to (2), (3) and (4).

The problem (L_tSP^{λ}) is solved considering implicitly constraint (2) and decomposing for index i, obtaining the following n problems:

$$\min \sum_{j=1}^{n}(d_{ij} - t\lambda_j)x_{ij}$$

subject to (3) and (4).

Each problem is easily solved by letting: $\beta_i = \sum_{j=1}^{n}\{\min(0, d_{ij} - t\lambda_j)\}$, and choosing I as the index set of the p smallest β_i (here constraint (2) is implicitly considered). Then, a solution x_{ij}^{λ} to the problem (L_tSP^{λ}) is:

$$x_{ii}^{\lambda} = \begin{cases} 1, & \text{if } i \in I \\ 0, & \text{otherwise} \end{cases}$$

and for all $i \neq j$,

$$x_{ij}^{\lambda} = \begin{cases} 1, & \text{if } i \in I \text{ and } d_{ij} - t\lambda_j < 0 \\ 0, & \text{otherwise} \end{cases}$$

The lagrangean/surrogate solution is given by: $v(L_tSP^{\lambda}) = \sum_{i=1}^{n}\beta_i x_{ii} + t\sum_{j=1}^{n}\lambda_j$. The interesting characteristic of relaxation (L_tSP^{λ}), is that for t = 1 we have the usual lagrangean relaxation using the multiplier λ. For a fixed multiplier λ, the best value for t can result of a lagrangean dual:

$$v(D_t^{\lambda}) = \max_{t \geq 0} v(L_tSP^{\lambda}).$$

A search procedure for finding an approximate value for t is shown in (Senne and Lorena, 2000).

The combined use of lagrangean/surrogate and subgradient optimization is given by the following algorithm:

Algorithm LSSH (lagrangean/surrogate subgradient heuristic)

Given $\lambda \geq 0$, $\lambda \neq 0$;

Set lb = - ∞, ub = + ∞;

Repeat

Solve relaxation (L_tSP^λ) obtaining x^λ and v(L_tSP^λ);

Obtain a feasible solution x_f and update v_f accordingly;

Update lb = max [lb, v(L_tSP^λ)];

Update ub = min [ub, v_f];

Set $g_i^\lambda = 1 - \Sigma x_{ij}^\lambda$, i ∈ N;

Update the step size θ;

Set $\lambda_i = \max \{ 0, \lambda_i + \theta . g_i^\lambda \}$, i ∈ N;

Until (stopping tests).

The initial λ used is $\lambda_i = \min_{j\in N}\{d_{ij}\}$, i ∈ N. The step sizes used are: $\theta = \pi(ub - lb)/\|g^\lambda\|^2$. The control of parameter π is the Held and Karp (1971) classical control. It makes 0 ≤ π ≤ 2, beginning with π = 2 and halving π whenever lb not increases for 30 successive iterations. The stopping tests used are: number of iterations greater than 1000, π ≤ 0.005, and (ub – lb) < 1.

Solution x^λ is not necessarily feasible to (P), but the set I identify median nodes that can be used to produce feasible solutions to (P). Two heuristics are used to make x^λ primal feasible. The first calculates the upper bound at each iteration of LSSH while π is not halved. This heuristic simply makes: $v_f = \sum_{i=1}^{n}(\min_{i\in I} d_{ij})$. The second, as suggested by Beasley (1993), is an interchange heuristic which is used when π is updated to π/2.

3. The Lagrangean/Surrogate Relaxation and Column Generation

The p-median problem can be also modeled as the following set partition problem:

$$v(SPP) = \text{Min } \sum_{j=1}^{m} c_j x_j$$

subject to $$\sum_{j=1}^{m} A_j x_j = 1 \qquad (5)$$

$$\sum_{j=1}^{m} x_j = p \qquad (6)$$

$$x_j \in \{0,1\}$$

where: $S = \{S_1, S_2, ..., S_m\}$ is a set of subsets of N; A = $[a_{ij}]_{nxm}$ is a matrix with a_{ij} = 1 if $i \in S_j$, and a_{ij} = 0, otherwise; and $c_j = \underset{i \in S_j}{\text{Min}} \left(\sum_{k \in S_j} d_{ik} \right)$. This formulation is found in Minoux (1987). The same formulation can be obtained from the problem (P) applying the Dantzig-Wolfe decomposition considered by Garfinkel *et al*. (1974) and Swain (1974). If S is the set of all subsets of N, the formulation can give an optimal solution to the p-median problem. But the number of subsets can be huge, and a partial set of columns should be considered. Problem (SPP) is also known as the restricted master problem in the column generation context (Barnhart *et al*., 1998).

In this paper we consider the following linear programming set covering relaxation of (SPP):

$$v(\text{SCP}) = \text{Min} \sum_{j=1}^{m} c_j x_j$$

subject to

$$\sum_{j=1}^{m} A_j x_j \geq 1 \qquad (7)$$

$$\sum_{j=1}^{m} x_j = p$$

$$x_j \in [0,1].$$

Observe that $d_{ij} \geq 0, \forall i,j$ and (5) can be replaced with (7) in the linear model.

The lagrangean/surrogate relaxation is integrated to the column generation process transferring the multipliers π_j (j = 1,...,n) of problem (SCP) to the problem $\underset{t \geq 0}{\text{Max}}\ v(LS_t P^{\pi})$. The median with smallest contribution on $v[\underset{t \geq 0}{\text{Max}}\ v(LS_t P^{\pi})]$ (and allocated non-medians) results to be the one selected to produce the incoming column on the sub-problem:

$$v(\text{Sub}_t P) = \underset{j \in N}{\text{Min}} \left[\underset{y_{ij} \in \{0,1\}}{\text{Min}} \sum_{i=1}^{n} \left(d_{ij} - t.\pi_j \right) y_{ij} \right].$$

The reduced cost (for t = 1) is rc = v(Sub_1-P) - α and rc < 0 is the condition for incoming columns, but it well known (Barnhart *et al.*, 1998) that, for j = 1,..., n, all the corresponding columns $\begin{bmatrix} \frac{y_j}{1} \end{bmatrix}$ satisfying: $\left[\underset{y_{ij} \in \{0,1\}}{\text{Min}} \sum_{i=1}^{n} \left(d_{ij} - \pi_j \right) y_{ij} \right] < |\alpha|$, can be added to the pool of columns, accelerating the column generation process.

The combined use of lagrangean/surrogate and column generation process is given by the following algorithm:

<u>Algorithm LSCG</u> (lagrangean/surrogate column generation heuristic)

1. Set an initial pool of columns to (SCP);
2. Solve (SCP) obtaining the duals prices π_j, j = 1,...,n, and α;
3. Solve approximately a local lagrangean/surrogate dual $\underset{t \geq 0}{\text{Max}}\ v(LS_t P^{\pi})$, returning the corresponding columns of (Sub_tP);
4. Append to (SCP) the columns $\begin{bmatrix} \frac{y_j}{1} \end{bmatrix}$ satisfying $\left[\underset{y_{ij} \in \{0,1\}}{\text{Min}} \sum_{i=1}^{n} \left(d_{ij} - \pi_j \right) y_{ij} \right] < |\alpha|$;
5. If no columns are found in step 4 then stop;
6. Perform tests to remove columns and return to step 2.

Note that, by setting t = 1, the algorithm LSCG gives the traditional column generation process.

4. Computational Tests

The approaches discussed above were programmed in C and run on a Sun Ultra 30 workstation. First, we have considered instances drawn from OR-Library (Beasley, 1990) for which optimal solutions are known. The results are reported in the table below. Table 1 reports the results for LSSH and LSCG algorithms and contains:

– the number of nodes in the network and the number of medians to be located;
– the optimal integer solution for the instance;

– p_gap = 100 * | (v(SCP) – optimal) | / optimal, that is, the percentage deviation from optimal to the best primal solution value v(SCP) found;

– d_gap = 100 * (optimal – $v(LS_t P^{\pi})$) / optimal, that is, the percentage deviation from optimal to the best relaxation value $v(LS_t P^{\pi})$ found;

– the total computational time (in seconds).

Table 1: Computational results for OR-Library instances

			LSSH			**LSCG**		
n	p	optimal solution	p_gap	d_gap	total time	p_gap	d_gap	total time
100	5	5819	-	-	0.76	-	-	36.35
100	33	1355	-	-	1.14	-	-	0.37
200	5	7824	-	0.523	5.61	-	-	902.77
200	10	5631	-	-	5.87	-	-	996.00
200	67	1255	-	-	12.33	-	-	1.29
300	5	7696	-	0.046	12.94	0.246	1.796	17889.12
300	10	6634	-	0.131	14.43	-	-	10749.91
300	30	4374	-	-	15.74	-	-	831.22
300	100	1729	-	-	57.73	0.116	0.058	4.55
400	5	8162	-	0.866	18.48	0.686	1.662	52807.93
400	10	6999	-	0.440	24.23	-	-	36829.25
400	40	4809	-	-	46.62	-	-	1055.20
400	133	1789	-	-	231.51	0.112	0.950	6.21
500	167	1828	-	-	377.14	0.055	0.310	11.00
600	200	1989	-	-	879.95	0.302	0.285	15.81
700	233	1847	-	-	494.52	0.081	0.379	21.50
800	267	2026	-	-	1360.49	0.518	0.346	26.14
900	300	2106	0.047	0.004	2994.11	0.518	0.827	33.37

Table 1 shows that the combined use of lagrangean/surrogate and column generation can be very interesting, especially for large-scale problems. The results of Table 1 also show that for a given number of nodes, smaller the number of medians in the instance, harder is the problem to be solved using the column generation approach. The opposite occurs for lagrangean/surrogate approach combined with subgradient search methods, i.e., the instances for which the number of medians is about a third of the number the nodes seem to be easy to LSCG and hard to LSSH.

The computational tests for a large-scale instance drawn from the TSPLIB, compiled by Reinelt (1998), confirm these conjectures. Table 2 shows the results for the Pcb3038 instance

(3038 nodes). We can note that as the number of median increases, the performance of LSCG improves in such a way that, for p = 500, it is better than LSSH. In this table p_gap and d_gap are calculated as follows:

– p_gap = 100 * | (v(SCP) – best known solution) | / best known solution

– d_gap = 100 * (best known solution – $v(LS_t P^\pi)$) / best known solution

Table 2: Computational results for OR-Library instances

		LSSH			LSCG		
p	best known solution	p_gap	d_gap	total time	p_gap	d_gap	total time
300	187723.46	1.305	0.056	719.04	0.043	0.044	22235.02
350	170973.34	2.067	0.050	731.50	0.044	0.045	10505.93
400	157030.46	1.630	0.012	919.79	0.008	0.008	4686.27
450	145422.94	1612	0.056	745.86	0.052	0.053	1915.84
500	135467.85	2.344	0.040	684.82	0.036	0.036	597.86

5. Conclusion

The combined use of lagrangean/surrogate relaxation with sugbradient optimization (LSSH) and with column generation (LSCG) seen to be a good approach to solve p-median clustering problems. In terms of computational performance we have noted that, for a given number of nodes (n), the most time-consuming problems for LSSH correspond to p = n/3. By other hand, for a given number of nodes, the computational performance of LSCG improves as the number of median increases. So, these two approaches have complementary computational behavior and can be put to work together. That will be important to develop decision support systems for large-scale data obtained from geographical information systems.

Acknowledgments: The authors acknowledge Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (process 99/06954-7) and Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (processes 302408/88-6 and 300837/89-5) for partial financial research support.

References

Barnhart, C.; Johnson, E.L.; Nemhauser, G.L.; Savelsbergh, M.W.P.; Vance, P.H. (1998) Branch-and-Price: Column Generation for Solving Huge Integer Programs, Operations Research, 46 : 316-329.

Beasley, J.E. (1985) A note on solving large p-median problems, European Journal of Operational Research, 21 : 270-273.

Beasley, J.E. (1990) OR-Library: Distributing test problems by electronic mail. Journal Operational Research Society, 41 : 1069-1072.

Beasley, J.E. (1993) Lagrangean heuristics for location problems, European Journal of Operational Research, 65 : 383-399.

Dantzig, G.B.; Wolfe, P. (1960) Decomposition principle for linear programs. Operations Research, 8 : 101-111.

Desrochers, M.; Desrosiers, J.; Solomon, M. (1992) A New Optimization Algorithm for the Vehicle Routing Problem with Time Windows, Operations Research, 40 : 342-354.

Desrochers, M.; Soumis, F. (1989) A Column Generation Approach to the Urban Transit Crew Scheduling Problem, Transportation Science, 23 : 1-13.

du Merle, O.; Villeneuve, D.; Desrosiers, J.; Hansen, P. (1999) Stabilized column generation. Discrete Mathematics, 194 : 229-237.

Freville, A.; Lorena, L.A.N.; Plateau, G. (1990) New subgradient algorithms for the 0-1 multiknapsack Lagrangean and surrogate duals, Pre-publication # 90. LIPN-Universite Paris Nord.

Galvão, R.D. (1981) A Note on Garfinkel, Neebe and Rao's LP Decomposition for the p-Median Problem. Transportation Science, 15 (3) : 175-182.

Garey, M.R.; Johnson, D.S. (1979) Computers and intractability: a guide to the theory of NP-completeness, W.H. Freeman and Co., San Francisco.

Garfinkel, R.S.; Neebe, W.; Rao, M.R. (1974) An Algorithm for the M-median Location Problem. Transportation Science, 8 : 217-236.

Gilmore, P.C.; Gomory, R.E. (1961) A linear programming approach to the cutting stock problem. Operations Research, 9 : 849-859.

Gilmore, P.C.; Gomory, R.E. (1963) A linear programming approach to the cutting stock problem - part II. Operations Research, 11 : 863-888.

Held, M.; Karp, R.M. (1971) The traveling-salesman problem and minimum spanning trees: part II, Mathematical Programming, 1 : 6-25.

Kelley, J.E. (1960) The Cutting Plane Method for Solving Convex Programs, Journal of the SIAM, 8 : 703-712.

Lorena, L.A.N.; Lopes, F.B. (1994) A surrogate heuristic for set covering problems, European Journal of Operational Research, 79/1 : 138-150.

Lorena, L.A.N.; Narciso, M.G. (1996) Relaxation Heuristics for Generalized Assignment Problem. European Journal of Operational Research, 91 : 600-610.

Minoux, M. (1987) A Class of Combinatorial Problems with Polynomially Solvable Large Scale Set Covering/Set Partitioning Relaxations. RAIRO, 21 (2) : 105–136.

Reinelt, G. (1998) www.iwr.uni-heidelberg.de/iwr/comopt/soft/TSPLIB95/TSPLIB.html.

Senne, E.L.F.; Lorena, L.A.N. (2000) Lagrangean/Surrogate Heuristics for p-Median Problems. In Computing Tools for Modeling, Optimization and Simulation: Interfaces in Computer Science and Operations Research, M. Laguna and J.L. Gonzalez-Velarde (eds.) Kluwer Academic Publishers, 115-130.

Swain, R.W. (1974) A Parametric Decomposition Approach for the Solution of Uncapacitated Location Problems. Management Science, 21 : 955-961.

Vance, P.H.; Barnhart, C.; Johnson, E.L.; Nemhauser, G.L. (1994) Solving Binary Cutting Stock Problems by Column Generation and Branch-and-Bound, Computational Optimization and Applications, 3 : 111-130.