HEURISTIC METHODS FOR LARGE CLUSTERING PROBLEMS

Éric D. Taillard,

IDSIA (Lugano)

PAREO, Versailles, July 1998.

Centroid clustering

Given

n entities with weights

dissimilarities between entities

$p < n$

Find

A location for p centres

An allocation of each entity to the centres

Such that

Sum of weighted dissimilarities between entities and centres is minimized

p-median (p-MP)

Centres can be placed only on entities

Minimum sum-of-squares clustering

m measures for the entities, dissimilarity: squared distance in $\mathbb{R}^m$

Multi-source Weber

Entities: points on $\mathbb{R}^2$, dissimilarity: Euclidean distance

Clustering : usage

Classification

Centres: ideal individual

Statistical software

Parallel processing

Create sub-problem containing entities closely related

Location

Warehouses

Emitter antenna

Emergency services

Airports

Switching centres

Public facilities

ALTERNATE PROCEDURE (ALT)

For the MWP : proposed by Cooper (1963)

```
0) Input : Set of entities with weight and dis-
   similarity measure, problem specific allo-
   cation and location procedures.
1) Choose an initial position for each centre.
2) Repeat the following steps while the loca-
   tion of the centres varies :
   2a) Allocate the entities given the centre
       locations.
   2b) Given the allocation made at step 2a,
       locate each centre optimally.
```

Complexity

Average (empirical) complexity : $\hat{O}(.)$

Step 1) $O(pn)$

Step 2a) $O(pn)$

Step 2b) $O(n)$ (MWP, SSC), $O(n^2/p)$ (p-MP)

Number of repetition of loop 2: $\hat{O}(p^a n^b)$

Total complexity:

$\hat{O}(p^{\alpha} n^{\beta})$ for large p: $\alpha \cong 0.8$, $\beta \cong 1.2$. For $p = O(n) \Rightarrow \sim\hat{O}(n^2)$

QUALITY OF GREEDY LOCAL SEARCH

p-MP instances of Beasley

p \ n	100	200	300	400	500	600	700	800	900
5	0	0	0	0,03	0	0	0,00	0,06	0,12
10	0,04	0,08	0	0,14	0,25	0	0,02	0,13	0
$n/10$	0,04	0,00	0,01	0,10	0,10	0,05	0,01	0,07	0,06
$n/5$	0,04	0,15	0,04	0,05	0,05	0,04	0,06	—	—
$n/3$	0,04	0,05	0,06	0,09	0,14	0,14	—	—	—

Average quality (% above optimum)

p \ n	100	200	300	400	500	600	700	800	900
5	0	0	0	0,26	0	0,07	0,01	1,16	0,40
10	0,29	0,32	0	0,54	1,01	0	0,04	0,55	0
$n/10$	0,87	0,07	0,11	0,29	0,21	0,16	0,09	0,26	0,21
$n/5$	0,39	0,51	0,13	0,17	0,24	0,20	0,23	—	—
$n/3$	0,22	0,56	0,17	0,39	0,49	0,30	—	—	—

Worst solution produced over 20 runs (% above optimum)

p \ n	100	200	300	400	500	600	700	800	900
5	1,0	1,0	1,0	0,9	1,0	0,7	0,6	0,95	0,7
10	0,85	0,5	1,0	0,3	0,75	1,0	0,55	0,55	1,0
$n/10$	0,95	0,95	0,95	0,2	0,3	0,4	0,55	0,2	0
$n/5$	0,9	0,2	0,45	0,3	0,7	0,4	0,2	—	—
$n/3$	0,8	0,85	0,25	0,4	0,25	0,05	—	—	—

Proportion of runs finding optimum

Candidate list search (CLS)

1) Input: initial solution $\boldsymbol{s_1}$ (location of the p centres), parameter q.
2) Generate π, a random permutation of the elements $\{1, \ldots, p\}$ and μ, a random permutation of the elements $\{1, \ldots, n\}$. Set $i = 0$, $j = 0$.
3) For $k = 1$ to q repeat:
   3a) $i = (i \text{ modulo } n) + 1$; $j = (j \text{ modulo } p) + 1$.
   3b) Compute d_{max}, the distance of the most (weighted) distant entity.
   3c) While the weighted distance from entity μ_i to the nearest centre is lower than $(d_{max} - f(\boldsymbol{s_k})/n)/k + f(\boldsymbol{s_k})/n$ do: $i = (i \text{ modulo } n) + 1$.
   3d) Close centre π_j and open a new one located at entity μ_i to obtain a perturbed solution $\boldsymbol{s_k}'$
   3e) Improve $\boldsymbol{s_k}'$ with ALT to obtain $\boldsymbol{s_k}''$
   3f) If $f(\boldsymbol{s_k}'') < f(\boldsymbol{s_k})$ then $\boldsymbol{s_{k+1}} = \boldsymbol{s_k}''$, else $\boldsymbol{s_{k+1}} = \boldsymbol{s_k}$.
   3g) If $j = p$, generate a new random permutation π.

LOCAL OPTIMIZATION META-HEURISTIC (LOPT)

0) Input : Solution s composed of parts $s_1, \ldots, s_p$, dissimilarity measure between parts
1) $C = \{s_1, \ldots, s_p\}$
2) While $C \neq \emptyset$, repeat :
   - 2a) Choose $s_i \in C$; Let R the sub-problem composed of the set of the r closest parts of s_i ($s_i \in R$)
   - 2b) Optimize R
   - 2c) If R improved, set $C \leftarrow C \cup R$ else set $C \leftarrow C \setminus \{s_i\}$.

For centroid clustering:

One part = one centre + entities allocated to the centre

Distance between parts: dissimilarity between centres

Optimization of sub-problems: CLS(s)

Complexity of LOPT for centroid clustering

Loop 2 is repeated $\hat{O}(p^{\gamma}\cdot n^{\lambda})$ times. $\gamma \cong 0.9$, $\lambda \cong 0.2$

Step 2a) $O(rp)$

Step 2b) $\hat{O}(s\cdot r^{\alpha}\cdot(r\cdot n/p)^{\beta})$

Step 2c) $O(p)$

Total complexity: $\hat{O}(r\cdot p^{\gamma+1}\cdot n^{\lambda} + s\cdot r^{\alpha+\beta}\cdot p^{\gamma-\beta}\cdot n^{\lambda+\beta})$

$p = O(n)$: $\hat{O}(n^{\lambda+\gamma+1}) \cong \hat{O}(n^{2.1})$

Most "difficult" part: finding the r closest centres

Parallel LOPT

0) Input : Solution s composed of parts s_1, …, s_p, dissimilarity measure between parts.
1) $C = \{s_1, \ldots, s_p\}$
2) While $C \neq \emptyset$, repeat ,for all $s_i \in C$ in parallel:
   2a) Let R_i the sub-problem composed of the set of the r closest parts of s_i ($s_i \in R_i$)
   2b) Optimize R_i
3) Extract a subset $\boldsymbol{R} \subset \{R_1, \ldots, R_p\}$ of independent sub-problems that have been improved.
4) Set $C \leftarrow C \setminus \{s_i\}$ for all R_i not improved.
5) Set $C \leftarrow C \cup \boldsymbol{R}$

Practical implementation of 3: ?

Step 2) Consider a priori independent sub-problems only (Taillard, 1993)

Good initial solutions generator : DEC

0) Input : Set of entities with dissimilarity measure.
1) Solve an auxiliary problem with t centres with CLS(u).
2) The subsets of entities allocated to the same centre form t independent subproblems. Set $f_{ij} = \infty$, $1 \leq i \leq t$, $1 \leq j \leq n_i$.
3) For each subproblem do : Solve subproblem i with CLS(v) with $\max(1, \lfloor p/t \rfloor - 1) \leq j \leq \min(n_i, \lfloor p/t \rfloor + 1)$ centres and update the f_{ij} values associated.
4) Find a collection $j_1^*, \ldots, j_t^*$ of optimum number of centres to attribute to each subproblem with dynamic programming.
5) While $K = \{(i, j_i) \mid f_{ij_i} = \infty, j_i = \min(n_i, j_i^* + 1) \text{ or } j_i = \max(1, j_i^* - 1), 1 \leq i \leq t\} \neq \emptyset$, repeat :
   - 5a) For all $(i, j_i) \in K$, solve subproblem i with j_i centres, using CLS(v) and update the f_{ij_i} associated.
   - 5b) Find a new collection $j_1^*, \ldots, j_t^*$ of optimum number of centres to attribute to each subproblem with dynamic programming.

Complexity of DEC

Step 1) $\hat{O}(u \cdot t^{\alpha} \cdot n^{\beta})$

Steps 3) + all repetitions of 5a): t sub-problems with p/t centres and n/t entities, solved with CLS(v): $\hat{O}(v \cdot t^{1-\alpha-\beta} \cdot p^{\alpha} \cdot n^{\beta})$

Steps 5) + all repetitions of 5b) $\hat{O}(t \cdot p)$

Choice of parameter $t = \lfloor \sqrt{p} \rfloor$

Total complexity: $\hat{O}(u \cdot p^{\alpha/2} \cdot n^{\beta} + v \cdot p^{(1+\alpha-\beta)/2} \cdot n^{\beta} + p^{3/2})$

For $p = O(n) \Rightarrow \hat{O}(n^{\alpha/2+\beta})$ (To be compared to ALT: $\hat{O}(n^{\alpha+\beta})$)

NUMERICAL RESULTS (MWP)

n	*p*	Quality [%] DEC (20, 50)	DEC (20, 200)	LOPT (7, 50)	Time [s. SG] DEC (20, 50)	DEC (20, 200)	LOPT (7, 50)	Proportion DEC (20, 50)	DEC (20, 200)	LOPT (7, 50)
14051	100	2.4	2.38	0.39	458	1931	3109	4.4	4.4	4.4
	200	1.9	2.03	0.30	336	1379	1632	4.3	4.6	3.3
	300	2.2	2.08	0.30	252	1073	1055	5.0	5.0	2.6
	400	2.2	2.04	0.26	214	915	885	4.9	5.2	2.4
	500	2.5	2.12	0.27	195	909	838	4.9	5.8	2.5
	600	2.4	1.99	0.26	184	799	707	5.1	5.8	2.3
	700	2.5	1.93	0.23	177	829	632	5.7	6.8	2.2
	800	2.5	1.97	0.24	149	760	555	4.8	6.8	2.1
	900	2.6	2.00	0.26	146	736	505	6.2	7.3	2.1
	1000	3.0	2.17	0.31	129	667	437	4.6	6.6	2.0
	1500	3.0	2.41	0.43	97	491	288	4.0	5.8	1.9
	2000	4.0	2.81	0.59	88	379	227	4.7	4.8	1.8
	2500	4.5	3.28	0.91	88	345	193	4.4	4.9	1.8
	3000	4.7	3.65	1.15	82	347	173	4.6	4.9	1.8
	5000	4.4	3.59	1.28	73	309	123	4.3	4.5	1.5
85900	1000	1.78	1.53	0.09	3557	9415	7634	4.2	5.2	2.9
	1500	1.97	1.70	0.17	3149	7885	5343	4.3	5.7	2.8
	2000	1.81	1.46	0.12	2819	6923	4750	4.1	5.1	2.5
	2500	1.84	1.48	0.05	3100	7031	4640	4.8	6.1	2.6
	3000	1.74	1.30	0.10	2405	5959	4532	4.1	5.2	2.6
	4000	1.72	1.30	0.04	2440	5503	4098	3.9	4.5	2.2
	5000	1.87	1.41	0.00	2597	5214	3423	4.1	4.3	2.0
	6000	1.67	1.37	0.05	2328	5392	2872	4.0	4.2	1.9
	7000	2.03	1.50	0.13	2276	4770	2526	4.3	4.3	1.9
	8000	2.07	1.53	0.03	2681	4685	2344	4.1	4.6	2.0
	9000	2.55	1.67	0.16	2796	4658	1992	4.1	4.3	1.8
	10000	2.78	1.80	0.16	2629	4863	1813	5.1	4.5	1.8
	15000	3.71	2.61	0.58	3144	5242	1552	5.3	6.0	1.8

Quality (% above best known), Time (seconds, Silicon Graphics 195MHz), Proportion (sub-problem solved /*t* (DEC), /*p* (LOPT)) for Multi-source Weber problems

Numerical Results
Problem instance Pcb3038

Sum-of-squares clustering

			Quality [%]		Time [s. Sparc 10]		
p	Best known	RVNS	DEC	LOPT	RVNS	DEC	LOPT
100	47721950.8	2.27	5.24	1.04	152.8	50	82
150	30524769.8	3.13	4.63	1.09	153.0	48	55
200	21885997.1	2.44	5.55	0.90	159.9	48	44
250	16621446.5	2.56	6.98	1.59	182.1	46	39
300	13290304.8	2.50	7.22	1.44	229.3	49	33
350	11027516.8	2.53	7.24	1.42	230.9	44	30
400	9362179.2	3.35	7.56	1.70	165.0	43	27
450	8101618.7	3.47	7.18	1.66	242.6	42	26
500	7102678.4	2.85	7.47	1.73	204.4	43	25

Comparison of DEC(20, 50), LOPT(6, 40) and RVNS

p-median problem

			Quality [%]		Time [s. Sparc 10]		
p	Best known	RVNS	DEC	LOPT	RVNS	DEC	LOPT
100	352704.86	1.12	4.04	0.65	132.4	197	485
150	281193.96	0.65	4.41	0.74	128.5	141	277
200	238432.02	1.23	4.12	0.74	107.6	106	187
250	209241.25	0.71	4.17	0.59	150.3	85	150
300	187723.46	0.53	4.08	0.64	130.6	84	125
350	170973.34	0.83	3.99	0.67	153.1	72	110
400	157030.46	1.13	4.04	0.83	158.7	64	97
450	145422.94	1.13	4.14	0.76	179.5	65	87
500	135467.85	0.88	4.01	0.71	209.7	59	81

Comparison of DEC(20, 50), LOPT(7, 50) and RVNS