A p-median heuristic for very large-scale data clustering

Igor Vasilyev

Institute of System Dynamics and Control Theory
Siberian Branch of Russian Academy of Sciences

SYM-OP-IS'09, Ivanjica, September 24

Outline

1. Introduction.
2. Lagrangean relaxation and subgradient optimization.
3. Upper bounding heuristics:
    1. Core heuristic
    2. Aggregation procedure
4. Conclusions

Problem statement

Cluster analysis

Cluster analysis consists of dividing the set of objects into subsets (clusters) basing on similarity.

Given:

Set of objects

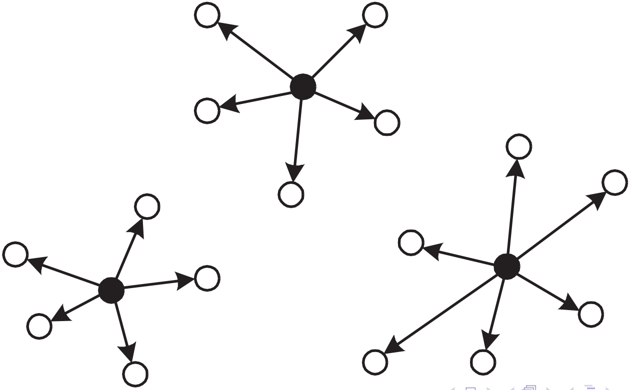
$$V = \{1, \ldots, m\}$$

$$a^u \in \mathbb{R}^n$$

$$d_{uv} = d(u, v) = \|a^u - a^v\|, \quad \forall u, v \in V$$

Problem statement

$G(V, A)$ – complete digraph.
V – set of nodes,
$A = \{uv : u \in V, v \in V, u \neq v\}$ – set of arcs,
d_{uv} – arc weight,
p – number of clusters.



$$Z^* = \min_{(x,y)} \sum_{u \in V} \sum_{v \in V} d_{uv} x_{uv} \tag{1}$$

$$\sum_{u \in V} x_{uv} + y_v = 1 \qquad \forall v \in V, \tag{2}$$

$$x_{uv} \leq y_u \qquad \forall u, v \in V, \tag{3}$$

$$\sum_{v \in V} y_v = p \tag{4}$$

$$y_u \in \{0, 1\} \qquad \forall v \in V, \tag{5}$$

$$x_{uv} \in \{0, 1\} \qquad \forall u, v \in V. \tag{6}$$

Large-scale heuristics

1. Hybrid Heuristic – GRASP (M. G. C. Resende, R. F. F. Werneck)
2. Primal–Dual Variable Neighborhood Search – PDVNS (P. Hansen, J. Brunberg, D. Urosevic, N. Mladenovic)

Motivation

Clustering quality

Klastorin T., The p-median problem for cluster analysis: A comparative test using the mixture model approach, Management Science, 31, 1985.

Complexity

- The problem is NP-hard.
- Metric problem – small duality gap.

Research progress

PDVNS can solve problems with more than 20,000 objects with 1% of quality.

Lagrangean relaxation

$$\theta(\lambda) = \min_{(x,y)} \sum_{v \in V} \Bigl(\sum_{u \in V} (d_{uv} - \lambda_v) x_{uv} - \lambda_v y_v + \lambda_v \Bigr) \tag{7}$$

$$x_{uv} \le y_u \qquad \forall u, v \in V, \tag{8}$$

$$\sum_{v \in V} y_v = p, \tag{9}$$

$$y_u \in \{0, 1\} \qquad \forall v \in V, \tag{10}$$

$$x_{uv} \in \{0, 1\} \qquad \forall u, v \in V. \tag{11}$$

Subgradient optimization

Lagrangean reduced cost

$$r_u(\lambda) = \sum_{v \in V} (d_{uv} - \lambda_v)^- - \lambda_u, \quad \forall u \in V \tag{12}$$

$$T(\lambda) = \{u_1, \ldots, u_p\}$$

$$\theta(\lambda) = \sum_{u \in T(\lambda)} r_u(\lambda) + \sum_{u \in V} \lambda_u \tag{13}$$

Subgradient optimization

$$\theta(\lambda) \to \max_{\lambda \in \mathbb{R}^m} \tag{14}$$

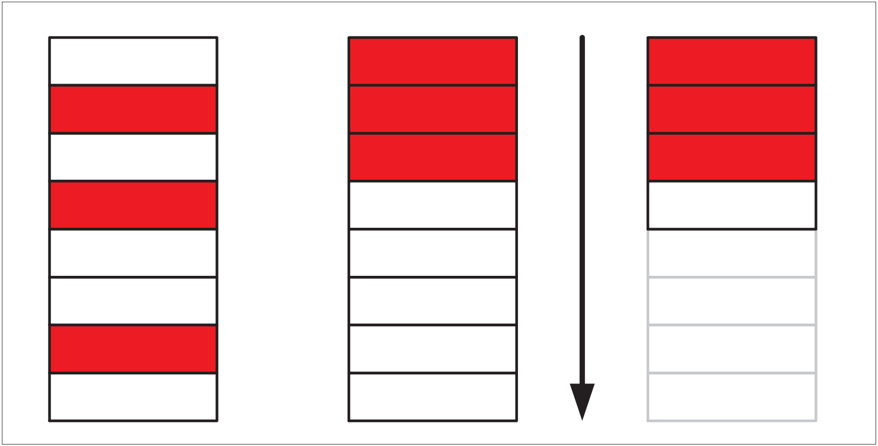
$$\lambda^{k+1} = \lambda^k + \alpha_k g(\lambda^k), \tag{15}$$

$$g_v(\lambda^k) = 1 - \sum_{u \in V} x_{uv}(\lambda^k) - y_v(\lambda^k), \tag{16}$$

$$\alpha_k = \frac{\phi_k(1.05 \cdot BUB - \theta(\lambda^k))}{\|g(\lambda^k)\|_2^2} \tag{17}$$

Lagrangean reduced cost

$$r_u(\lambda) = \sum_{v \in V} (d_{uv} - \lambda_v)^- - \lambda_u, \quad \forall u \in V$$



m	p	GAP(%) VNS	GAP(%) CH	Time VNS	Time CH	GAP(%) VNS	GAP(%) CH	Time VNS	Time CH
		BIRCH of type 1				BIRCH instances of type 3			
10000	100	0.021	0.011	565	28	0.096	0.109	1738	37
15000	100	0.213	0.024	2014	59	0.094	0.077	2238	71
20000	100	0.000	0.007	2497	108	0.181	0.095	2238	138
9600	64	0.023	0.013	969	27	0.123	0.101	939	34
12800	64	0.015	0.006	1981	49	0.117	0.175	1688	60
16000	64	0.000	0.009	2233	75	1.890	0.306	2231	105
19200	64	0.021	0.010	2478	137	0.907	0.728	2483	172
10000	25	0.065	0.005	989	41	0.834	0.323	889	63
12500	25	0.049	0.012	1734	60	0.788	0.509	1461	80
15000	25	0.028	0.011	1932	95	3.099	0.203	2160	139
17500	25	0.026	0.008	2234	127	1.141	0.944	2231	238
20000	25	0.001	0.008	2489	198	2.060	0.788	2479	281

Core heuristic

Core problem

Choose variables with the best lagrangean reduced cost

$$\begin{array}{lcll} y_u & - & \sum_{v \in V} (d_{uv} - \lambda_v)^- - \lambda_u & u \in V \\ x_{uv} & - & d_{uv} - \lambda_v & uv \in A \end{array}$$

Algorithm

Step 0. Generate initial feasible solution (upper bound).

Step 1. Subgradient algorithm (lower bound).

Step 2. Choose and solve core problem (upper bound).

Core heuristic

n	p	BUB	Err(%) GR	Err(%) VNS	Err(%) CH	GAP(%) VNS	GAP(%) CH	Time GR	Time VNS	Time CH
		BIRCH instances of type 1								
10000	100	12428.5	0.000	0.004	0.000	0.021	0.001	54	786	47
15000	100	18639.3	–	0.015	0.000	0.213	0.002	–	3386	101
20000	100	24840.3	–	0.000	0.000	0.000	0.001	–	3982	210
9600	64	11934.8	0.000	0.000	0.000	0.023	0.002	57	1205	56
12800	64	15863.8	0.000	0.000	0.000	0.015	0.001	99	2451	84
16000	64	20004.6	–	0.000	0.000	0.000	0.001	–	2739	129
19200	64	24018.3	–	0.000	0.000	0.021	0.002	–	3698	219
10000	25	12455.7	0.000	0.000	0.000	0.065	0.001	95	1091	82
12500	25	15597.1	0.000	0.005	8.792	0.049	8.794	151	2073	115
15000	25	18949.3	–	0.000	16.677	0.028	16.681	–	2353	175
17500	25	21937.4	–	0.000	8.434	0.026	8.437	–	2615	241
20000	25	25096.8	–	0.000	10.166	0.001	10.168	–	3055	365

Core heuristic

n	p	BUB	Err(%) GR	Err(%) VNS	Err(%) CH	GAP(%) VNS	GAP(%) CH	Time GR	Time VNS	Time CH
		BIRCH instances of type 3								
10000	100	9624.79	0.000	0.050	0.00	0.096	0.002	377	2609	60
15000	100	15904.12	–	0.000	21.72	0.094	21.767	–	3495	121
20000	100	19989.02	–	0.000	27.92	0.181	27.983	–	3429	222
9600	64	8225.58	0.000	0.055	21.89	0.123	21.912	377	1483	57
12800	64	10210.36	0.000	0.062	11.40	0.117	11.412	413	2503	98
16000	64	13340.47	–	0.000	23.10	1.890	23.142	–	3169	170
19200	64	15207.56	–	0.000	38.69	0.907	38.925	–	3243	229
10000	25	7203.39	0.000	0.000	11.18	0.834	11.349	316	1016	94
12500	25	8576.10	0.000	0.339	0.82	0.788	0.956	203	1606	144
15000	25	9513.64	–	0.000	51.96	3.099	52.041	–	2742	192
17500	25	12535.68	–	0.000	37.75	1.141	38.387	–	2803	250
20000	25	13052.81	–	0.000	54.33	2.060	54.700	–	3364	364

Aggregation procedure

Idea

1. Construct partition of the nodes by solving the p-median problem with big number of clusters (5000).
2. Aggregate the nodes of each cluster into one, constructing a new reduced graph.
3. Solve the p-median problem over the reduced graph, recover the solution for the original graph.
4. Find the lower bound by the subgradient algorithm.

Core heuristic with aggregation procedure

n	p	BUB	Err(%) GR	Err(%) VNS	Err(%) CH	GAP(%) VNS	GAP(%) CH	Time GR	Time VNS	Time CH
		BIRCH instances of type 3								
12500	25	15597.12	0.000	0.005	0.086	0.049	0.096	151	2073	162
15000	25	18949.26	–	0.000	0.117	0.028	0.125	–	2353	194
17500	25	21937.40	–	0.000	0.103	0.026	0.111	–	2615	228
20000	25	25096.82	–	0.000	0.130	0.001	0.136	–	3055	312
		BIRCH instances of type 1								
15000	100	15904.12	–	0.000	0.312	0.094	0.383	–	3495	186
20000	100	19989.02	–	0.000	0.359	0.181	0.462	–	3429	289
9600	64	8225.58	0.000	0.055	0.163	0.123	0.190	377	1483	123
12800	64	10210.36	0.000	0.062	0.228	0.117	0.331	413	2503	171
16000	64	13340.47	–	0.000	0.172	1.890	0.375	–	3169	240
19200	64	15207.56	–	0.000	0.272	0.907	0.999	–	3243	321
10000	25	7203.39	0.000	0.000	0.058	0.834	0.330	316	1016	160
12500	25	8576.10	0.000	0.339	0.097	0.788	0.286	203	1606	188
15000	25	9513.64	–	0.000	0.076	3.099	0.256	–	2742	261
17500	25	12535.68	–	0.000	1.044	1.141	2.105	–	2803	359
20000	25	13036.63	–	0.124	0.000	2.060	0.674	–	3364	480

		BUB	GAP(%)	Time	BUB	GAP(%)	Time
n	p	type 1			type 3		
25000	25	31282.6	0.181	447	17718.6	0.210	527
36000	36	45226.3	0.261	780	27476.1	0.682	913
49000	49	61569.7	0.319	1216	44282.5	0.663	1760
64000	64	80337.4	0.369	2258	58991.5	0.467	2624
30000	25	37617.1	0.161	559	21865.1	0.610	832
43200	36	54305.8	0.226	1003	32391.6	0.575	1873
58800	49	73854.7	0.324	1691	50985.1	0.394	2692
76800	64	96393.4	0.384	2834	66944.7	0.816	4393
35000	25	43972.1	0.180	768	24833.7	0.183	972
50400	36	63329.2	0.267	1472	38162.3	0.407	2297
68600	49	86082.0	0.300	2441	62007.4	0.981	3556
89600	64	112485.2	0.393	4501	79245.3	0.978	5779

Conclusions

Results

The approach is fast and gives quite good solutions.

Future prospects

1. Application to SFLP.
2. Combine with PDVNS.
3. Parallelization.
4. Deal with millions objects.

Contacts

email: vil@icc.ru
http://iv.icc.ru/