Global Optimal K-Medoids Clustering of One Million Samples

Jiayang Ren, Kaixun Hua, Yankai Cao

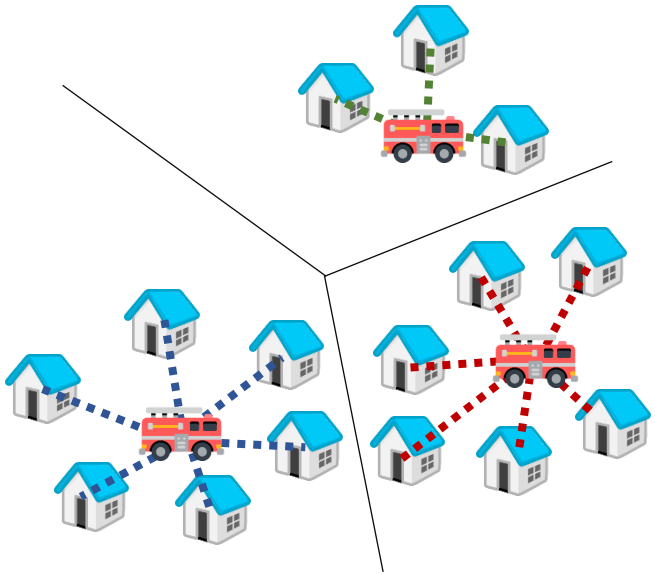
Department of Chemical and Biological Engineering

University of British Columbia

K-Medoids Clustering

Select *K* existing samples to minimize the sum of distance to any other samples

- Input:
  - Dataset of S samples and A dimensions, $X = \{x_1, \dots, x_S\} \in \mathbb{R}^{A \times S}$
  - Desired cluster number, K
- Output:
  - Center set $\mu \in X$ with $|\mu| = K$
  - $\mu \in X$: "medoids on samples" constraint
- Object:
  - $\min_{\mu \in X} \sum_{s \in \mathcal{S}} \min_{k \in \mathcal{K}} \|x_s - \mu^k\|_2^2$ (1)



$S = 18,\ K = 3$

Extensive Form

- $\boldsymbol{b_{s,j}} = \begin{cases} 1, x_s \ belongs \ to \ cluster \\ \quad with \ x_j \ as \ medoid \\ 0, \end{cases}$
- $\boldsymbol{y_j} = \begin{cases} 1, sample \ x_j \ is \ a \ medoid \\ \quad of \ a \ clsuter \\ 0, \end{cases}$
- $d_{s,j} = \|x_s - x_j\|_2^2$: the distance between x_s and x_j
- $s, j \in \mathcal{S}$

MIP

$$\min_{b,y} \sum_{s\in\mathcal{S}} \sum_{j\in\mathcal{S}} d_{s,j} b_{s,j} \quad (2a)$$

s.t.

$$\sum_{j\in\mathcal{S}} b_{s,j} = 1 \quad (2b)$$

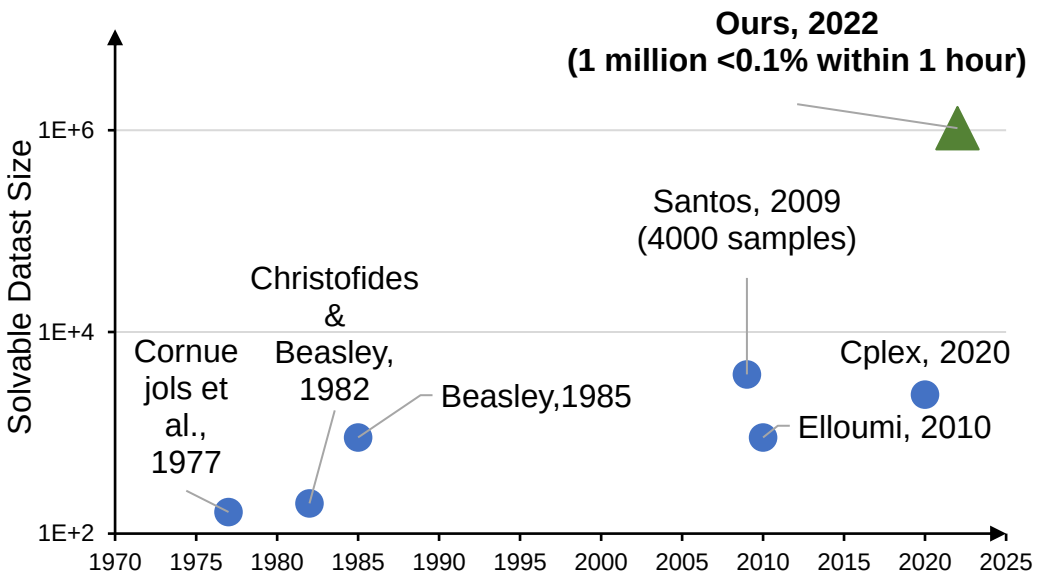
$$\sum_{j\in\mathcal{S}} y_j = K \quad (2c)$$

$$b_{s,j} \le y_j \quad (2d)$$

$$b_{s,j}, y_j \in \{0,1\} \quad (2e)$$

$$s, j \in \mathcal{S} \quad (2f)$$

Classical Branch and Bound

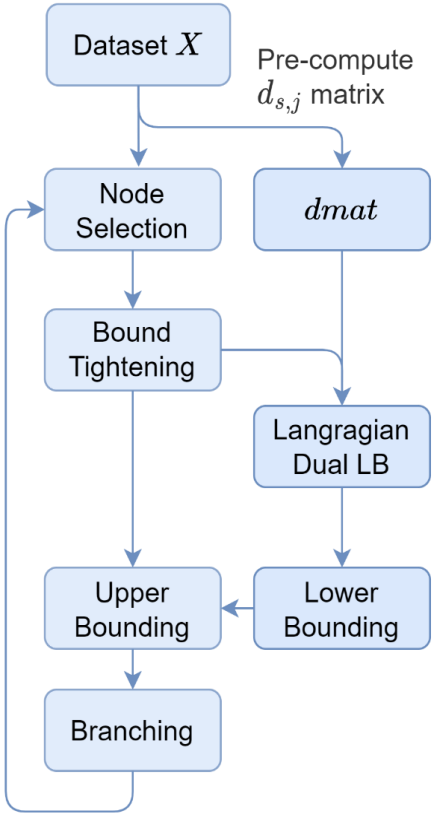


- Branch on integer variable, y_j
- Number of y_j equal to sample number
- E.g., a 1000-sample , 2-dimensional dataset with 3 clusters consists of 1000 integer variables, y_j.

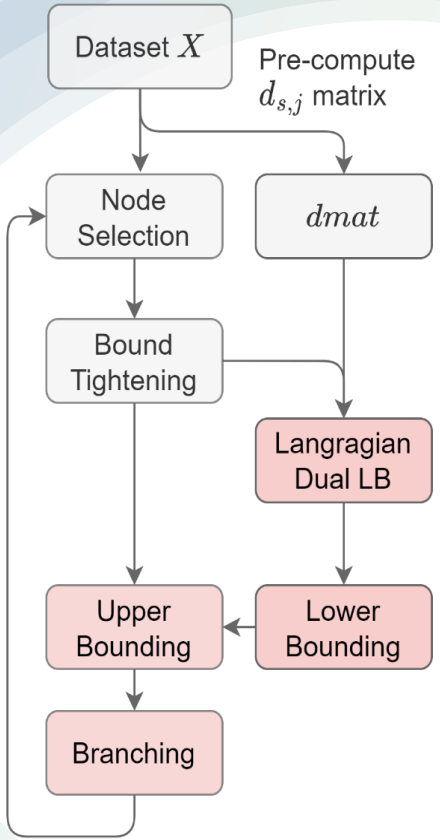
Hard to solve large-scale problems

Our Works

- Branching on regions of medoids can guarantee the convergence to global optimum.
  (e.g., branching variables: $1000 \longrightarrow 6$).
- Combination of basic and Lagrangian based lower bound methods, both with analytic solutions.
- Bound Tightening to reduce the search space and speed up the branch and bound.

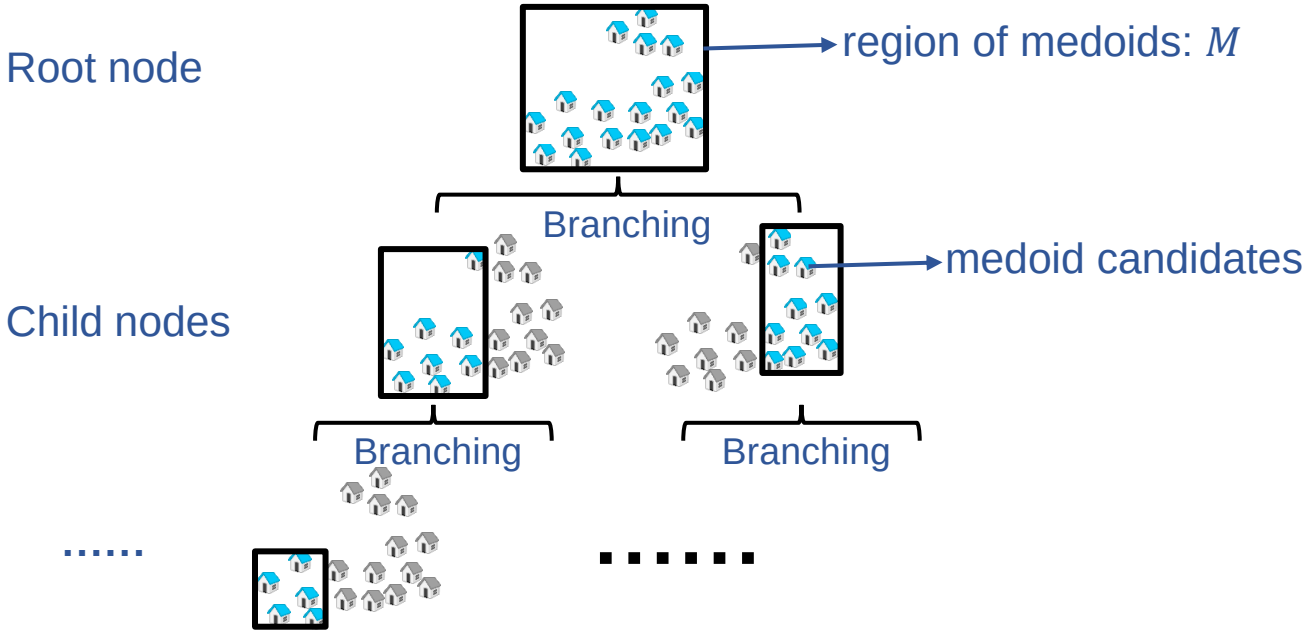


Tailored Branch and Bound Scheme



Branching

- Branch on regions of medoids ($M \coloneqq \{\mu \mid \underline{\mu} \leq \mu \leq \bar{\mu}\}$, or called medoid region)
- Branching variable number: $K * A$ (cluster $*$ dimension)
- Medoid candidates: $\mathrm{u} \in M \cap X,\ |M \cap X| > 1 \quad (4)$



Basic Lower Bound

- At each child node, the K-Medoids problem

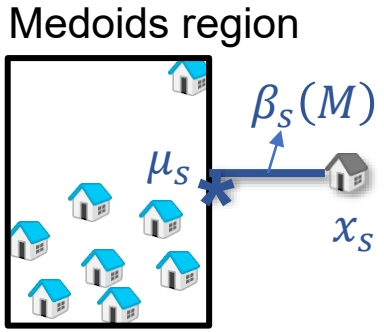
$$z(M) = \min_{\mu \in M \cap X} \sum_{s \in \mathcal{S}} \min_{k \in \mathcal{K}} \|x_s - \mu_s^k\|_2^2 \quad (5a)$$

$$s.t.\ \ \mu_s = \mu_{s+1}, s \in \{1, \dots, S-1\} \quad (5b)$$

- Relax non-anticipativity constraints (5b) and "medoids on samples" constraint

$$\beta_{basic}(M) = \min_{\mu_s \in M} \sum_{s \in \mathcal{S}} \min_{k \in \mathcal{K}} \|x_s - \mu_s^k\|_2^2 = \sum_{s \in \mathcal{S}} \underbrace{\min_{\mu_s \in M} \min_{k \in \mathcal{K}} \|x_s - \mu_s^k\|_2^2}_{\beta_s(M)} \quad (6)$$

- Closed-form solution of $\beta_s(M)$: distance from the sample to medoid region



- Characters: low computational cost but not tight bound

Theorem 1

The algorithm converges to the global optimal solution after a finite step by branching on the regions of medoids and the basic lower bound.

Lagrangian lower bound

- Extensive Form for child node

$$\min_{b,y} \sum_{s\in\mathcal{S}} \sum_{j\in\mathcal{S}(M)} d_{s,j} b_{s,j} \quad (8a)$$

$$\text{s.t.} \sum_{j\in\mathcal{S}(M)} b_{s,j} = 1 \quad (8b)$$

$$\sum_{j\in\mathcal{S}^k(M)} y_j^k = 1 \quad (8c)$$

$$\sum_{k\in\mathcal{K}} y_j^k \le 1 \quad (8d)$$

$$b_{s,j} \le \sum_{k\in\mathcal{K}} y_j^k \quad (8e)$$

$$y_j^k = 0, j \in \mathcal{S}\backslash\mathcal{S}^k(M) \quad (8f)$$

$$b_{s,j}, y_j^k \in \{0,1\} \quad (8g)$$

$$\mathcal{S}(M) = \mathcal{S}^1(M) \cup \cdots \cup \mathcal{S}^K(M) \quad (8h)$$

- Lagrangian Relaxation

$$\beta_{LD}(M,\lambda) = \min_{b,y} \sum_{s\in\mathcal{S}} \left[\sum_{j\in\mathcal{S}(M)} (d_{s,j} - \lambda_s) b_{s,j} + \lambda_s \right] \quad (9a)$$

- Closed-form solution (for a given λ):

Define contribution $\rho_j(\lambda) := \sum_{s\in\mathcal{S}} \min(0, d_{s,j} - \lambda_s)$,

$$\beta_{LD}(M,\lambda) = \min_{y} \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{S}^k(M)} \rho_j(\lambda) y_j^k + \sum_{s\in\mathcal{S}} \lambda_s \quad (9b)$$

Select samples with smallest ρ_j in each medoid region

Lagrangian lower bound

- Lagrangian Dual

$$\beta_{LD}(M) = \max_{\lambda} \beta_{LD}(M, \lambda). \quad (9c)$$

- Certain optimality gap*

→ Quality guarantee of LB

$$Gap = \frac{z - \beta_{LD}}{z_r - \beta_{LD}} < \frac{1}{e},$$

$$where\ \mathrm{z_r} = \sum_{s \in \mathcal{S}} \max_{j \in \mathcal{S}} d_{s,j}.$$

Method	Basic	Lagrangian
Symbol	$\beta_{Basic}(M)$	$\beta_{LD}(M)$
Cost	Low	High
Bound	Loose	Tight
Convergence	Yes	No

* Gerard Cornuejols, Marshall L. Fisher, and George L. Nemhauser. Location of Bank Accounts to Optimize Float: An Analytic Study of Exact and Approximate Algorithms. Management Science, 23(8):789–810, 1977

Combined lower bound

Algorithm 1 Combined Lower Bound

1: Compute basic LB, $\beta_{basic}(M)$.
2: **if** $|\beta_{basic}(M) -$ Upper Bound $| \geq \epsilon$ **then**
3: Compute Lagrange Dual LB, $\beta_{LD}(M)$.
4: $\beta(M) = max\{\beta_{basic}(M), \beta_{LD}(M)\}$.
5: **else**
6: $\beta(M) = \beta_{basic}(M)$.
7: **end if**

Method	Combined	Basic	Lagrangian
Symbol	$\beta(M)$	$\beta_{Basic}(M)$	$\beta_{LD}(M)$
Cost	Medium	Low	High
Bound	Tight	Loose	Tight
Convergence	Yes	Yes	No

Upper Bound

- Given a feasible solution $\hat{\mu}$, upper bound is

$$UB = \alpha(M) = \sum_{s \in \mathcal{S}} \min_{k \in \mathcal{K}} \|x_s - \hat{\mu}^k\|_2^2 \qquad (10)$$

Root node

- K-Means-like method

↓

Initial guesses of solutions

↓

- Evolutionary Centers Algorithm

Child node

- Candidate solutions obtained in the lower bounding process

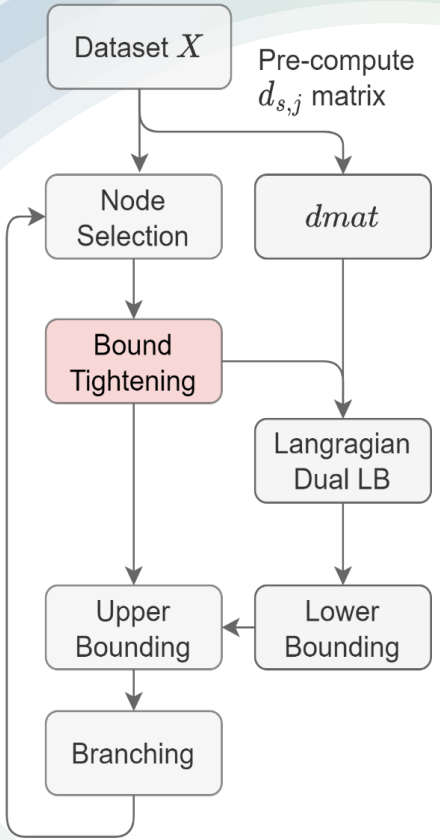
↓

Initial guesses of solutions

↓

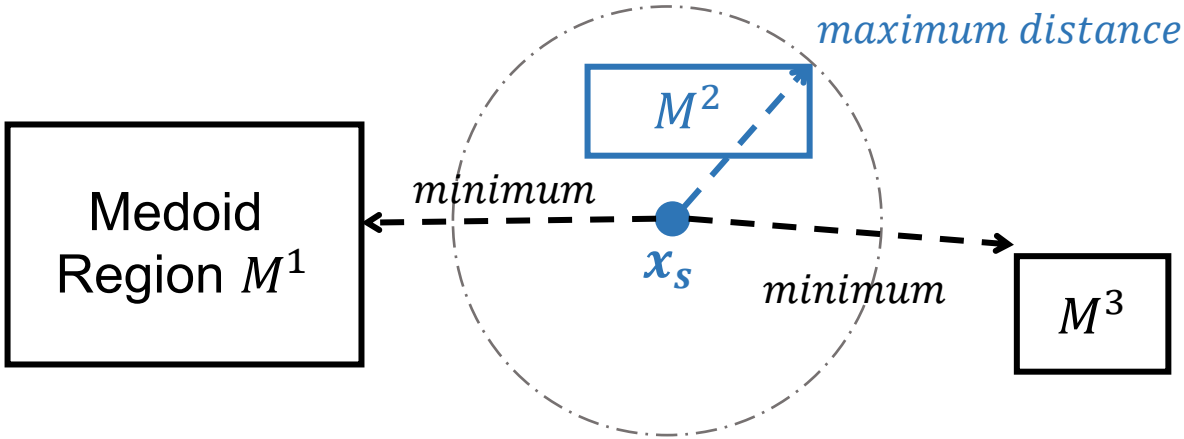
- K-Means-like method

Bound Tightening



Feasibility-Based Bound Tightening (FBBT)

- Cluster Assignment (for a child node with M)
  - If the maximum distance from sample s to a medoid region k is smaller than the minimum distances to other medoid regions
  - Then this sample s must belong to this cluster k.



Feasibility-Based Bound Tightening (FBBT)

- Denote the index sets of samples assigned to cluster k as $\mathcal{S}_A^k$.
- For cluster k, tighten the medoid region, M, by the feasibility inequation:

$$\beta(M)^- + \sum_{s \in \mathcal{S}_A^k} \left\| x_s - \mu^k \right\|_2^2 \leq \alpha_l, \qquad (11)$$

  - Where $\beta(M)^-$ represents the lower bound contributions of samples unassigned to cluster k, $\beta(M)^- = \sum_{s \in \mathcal{S} \setminus \mathcal{S}_A^k} \min_{k \in \mathcal{K}} d_{s,min}^k(M)$.

- Inequation (11) is a simple quadratic inequation, it can be solved analytically to obtain the tightened medoid region, $\widehat{M}$.

Bound Tightening

- Probing
  - Divide the center region, M, into several sub-regions, M_{sub}
  - Compute basic LB, $\beta_{basic}(M_{sub})$, for each sub-region
  - Delete the sub-region if $\beta_{basic}(M_{sub}) > \alpha_l$
  - Execute only at the root node
- Symmetric-breaking
  - Enforce a symmetric-breaking constraint on the first attribute
  - $\mu_1^k \leq \mu_1^{k+1}, \forall k = 1, \ldots, K-1$
- Benefits
  - Reduce the number of child nodes to need be explored.
  - Accelerate the calculation of lower bounds
    - Only need to calculate the subproblems in the tightened medoid region.

Numerical Results

Serial results – small datasets

- 21 datasets from 100 samples to 10K samples in the serial mode.
- Identical results with CPLEX when datasets are small.
- Solving all 21 datasets to 0.1% gap with less computational time.
- Obtain a slightly better objective value in a small part of problems – Good Heuristic!

Dataset	Sample	Dimension	Method	Objective value	Nodes	Gap(%)	Time(s)
TR	980	10	Heuristic	1136.93	-	-	-
			CPLEX	1134.45	3	≤0.10	855
			BB+LD	**1134.45**	**191**	**≤0.10**	**145**
PR2392	2392	2	Heuristic	2.13E+10	-	-	-
			CPLEX	2.13E+10	1	≤0.10	5339
			BB+LD	**2.13E+10**	**37**	**≤0.10**	**123**
HTRU2	17898	8	Heuristic	8.21E+07	-	-	-
			CPLEX	-	-	-	-
			BB+LD	**8.21E+07**	**465**	**≤0.10**	**1555**
URBAN GB_10	100000	2	Heuristic	1.26E+05	-	-	-
			CPLEX	-	-	-	-
			BB+LD	**1.15E+05**	**49**	**≤0.10**	**6834**

Parallel results – huge datasets

- 7 datasets from 10K samples to 2 million samples in the parallel mode.
- Capable to solve 1 million sample datasets within 1 hour.
- Solving 2 million sample datasets to 6.33% gap with 4 hours

Dataset	Sample	Dimension	Cores	Objective value	Nodes	Gap(%)	Time(s)
rng_agr	199,843	7	1600	8.23E+14	99	≤0.10	341.0
urbanGB	360,177	2	1600	4.14E+05	57	≤0.10	327.0
spnet3D	434,874	3	1600	2.28E+07	115	≤0.10	865.0
retail	541,909	2	1600	6.80E+09	1	≤0.10	80.0
synthetic	1,000,000	2	6000	9.44E+06	3	≤0.10	171.0
retail-II	1,046,910	2	6000	2.31E+10	214	≤0.10	2515.0
USC1990*	2,458,285	68	3000	6.91E+08	25	6.33	4h

Conclusion

- A tailored reduce-space branch and bound algorithm for K-Medoids.
- A Lagrangian-based lower bounding method.
  - No need to solve any optimization sub-problems.
- Bound tightening methods.
  - Reduce the search space and speed up the BB procedure.
- Enlarge the solvable K-Medoids problem for global optimum.
  - 1 million samples within 1 hours.

Reference

1. Gerard Cornuejols, Marshall L. Fisher, and George L. Nemhauser. Location of Bank Accounts to Optimize Float: An Analytic Study of Exact and Approximate Algorithms. Management Science, 23(8):789–810, 1977
2. N. Christofides and J.E. Beasley. A tree search algorithm for the p-median problem. European Journal of Operational Research, 10(2):196–204, 1982.
3. Beasley, John E. "A note on solving large p-median problems." European Journal of Operational Research 21.2 (1985): 270-273.
4. N. Christofides and J.E. Beasley. A tree search algorithm for the p-median problem. European Journal of Operational Research, 10(2):196–204, 1982.
5. Sourour Elloumi. A tighter formulation of the p-median problem. Journal of Combinatorial Optimization, 19(1):69–83, 2010.
6. IBM ILOG Cplex. V20.1.0: User's Manual for CPLEX. International Business Machines Corporation, 2020.
7. Hua, K., Shi, M., and Cao, Y. A Scalable Deterministic Global Optimization Algorithm for Clustering Problems. In International Conference on Machine Learning, pp. 4391–4401. PMLR, 2021.
8. Cao, Y. and Zavala, V. M. A scalable global optimization algorithm for stochastic nonlinear programs. Journal of Global Optimization, 75(2):393–416, 2019.

Thank you!