A Scalable Deterministic Global Optimization Algorithm for Clustering Problems

Kaixun Hua, Mingfei Shi, Yankai Cao

Cao's Research Group

Department of Chemical and Biological Enigneering

University of British Columbia

Problem Setup

We consider the following optimization problem:

$$\min_{\mu,d,b} \sum_{s\in\mathcal{S}} d_{s,*} \tag{1a}$$

$$\text{s.t.} \quad -N(1-b_{s,k}) \leq d_{s,*} - d_{s,k} \leq N(1-b_{s,k}) \tag{1b}$$

$$\boxed{d_{s,k} \geq ||x_s - \mu_k||^2} \tag{1c}$$

$$\sum_{k\in\mathcal{K}} b_{s,k} = 1 \tag{1d}$$

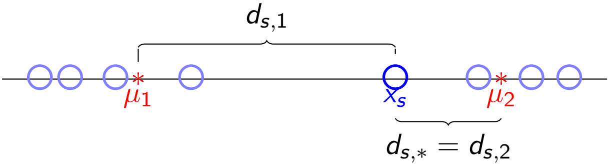
$$b_{s,k} \in \{0,1\} \tag{1e}$$

$$s \in \mathcal{S}, k \in \mathcal{K} \tag{1f}$$

$$b_{s,k} = \begin{cases} 1 & \text{If } x_s \text{ is in cluster } k \\ 0 & \text{otherwise} \end{cases}$$

$d_{s,k}$ is the distance between x_s and the cluster center μ_k

$d_{s,*}$ is $\min_k d_{s,k}$, $k \in \mathcal{K}$



Branch and Bound (BB) on Clustering Problem

Given a dataset D with m features and n samples, to cluster it into K clusters:

- The scale of such problem has $m(2n+K)$ variables
- A three-cluster, two dimensional dataset with 1000 samples consists of 4006 variables.
- BB in off-the-shelf solvers (CPLEX or Gurobi): branching on all (integer) variables.
- Our approach: branching only on space of centers (μ) is enough to guarantee the convergence. (# of branching variables: 6.)

LB Strategy: Scenario Decomposition

Problem 1 can be reformulated as following optimization problem:

$$z(M) = \min_{\mu \in M} \sum_{s \in \mathcal{S}} Q_s(\mu) \tag{2}$$

$$\begin{aligned} Q_s(\mu) = \min_{d_s, b_s} \; & d_{s,*} \\ \text{s.t.} \; & -N(1 - b_{s,k}) \leq d_{s,*} - d_{s,k} \leq N(1 - b_{s,k}) \\ & d_{s,k} \geq ||x_s - \mu_k||^2 \\ & \sum_{k \in \mathcal{K}} b_{s,k} = 1 \\ & b_{s,k} \in \{0, 1\}, k \in \mathcal{K} \end{aligned} \tag{3}$$

which is equivalent to:

$$\min_{\mu_s \in M} \quad \sum_{s \in \mathcal{S}} Q_s(\mu_s) \tag{4a}$$

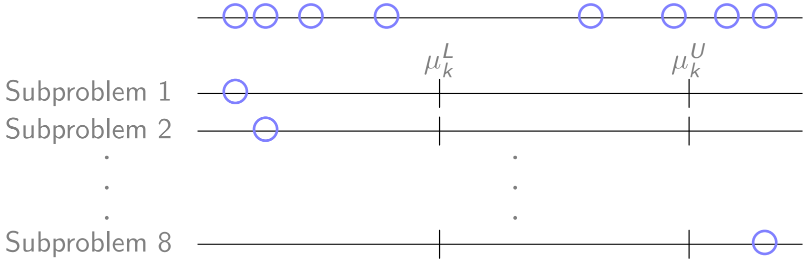
$$\text{s.t.} \quad \mu_s = \mu_{s+1}, s \in \{1, \cdots, \mathcal{S} - 1\} \tag{4b}$$

LB Strategy 1: Closed Form Solution

By relaxing the non-antipativity constraints 4b, we can obtain the lower bounding problem as follow:

$$\beta(M) = \sum_{s \in \mathcal{S}} \beta_s(M) = \min_{\mu_s \in M} \sum_{s \in \mathcal{S}} Q_s(\mu_s) \tag{LB1}$$

It is easy to decompose the Problem LB1 into n subproblems, where n is the number of samples of a dataset.



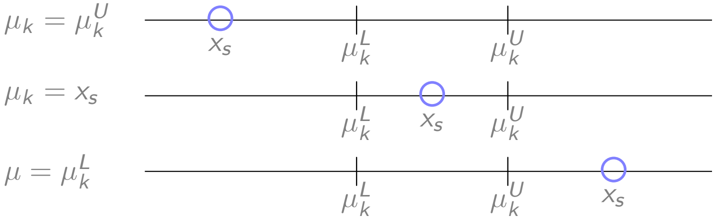
LB Strategy 1: Closed Form Solution

For each scenario s, the lower bound function β_s can be solved as follow:

$$\beta_s(M) = \min_k \beta_{s,k}(M_k) = \min_k \min_{\mu_k \in M_k} ||x_s - \mu_k||^2, \tag{5}$$

where $M_k := \{\mu_k \mid \mu_k^L \leq \mu_k \leq \mu_k^U\}$
Advantage: $\beta_{s,k}$ has a closed form solution:

$$\mu_{k,i} = \text{mid}\{\mu_{k,i}^L, x_{s,i}, \mu_{k,i}^U\}, \forall i \in \{1 \cdots m\} \tag{6}$$



LB Strategy 2: Lagrangean Decomposition

We dualized the non-anticipativity constraints and added to the objective functions with Lagrange multipliers λ:

$$\beta^{LD}(M,\lambda) := \min_{\mu \in M} \{ \sum_{s \in \mathcal{S}} Q_s(\mu_s) + \sum_{s=1}^{S-1} \lambda_s(\mu_s - \mu_{s+1}) \} \tag{7}$$

Thus we solve the lagrangean dual problem:

$$\beta^{LD}(M) = \max_{\lambda} \beta^{LD}(M,\lambda). \tag{LB2}$$

Lemma

$$\beta(M) = \beta^{LD}(M,0) \leq \beta^{LD}(M) \leq z(M)$$

LB Strategy 3: Adaptive Sample Grouping

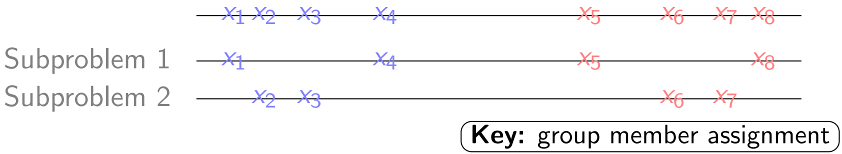
We assign group of samples into one subproblem

$$\min_{\mu_g \in M} \sum_{g \in \mathcal{G}} Q_g(\mu_g) \tag{8a}$$

$$\text{s.t.} \quad \mu_g = \mu_{g+1}, g \in \{1, \cdots, \mathcal{G} - 1\}. \tag{8b}$$

Thus, we have lower bounding problem by relaxing 8b:

$$\beta^{SG}(M) = \sum_{g \in \mathcal{G}} \beta_g^{SG}(M) = \min_{\mu_g \in M} \sum_{g \in \mathcal{G}} Q_g(\mu_g) \tag{LB3}$$



Lemma

$$\beta(M) \leq \beta^{SG}(M) \leq z(M)$$

Upper Bounding Problem

Two strategies to construct upper bound:

- Fix μ at a candidate solution $\hat{\mu} \in M$, we get an upper bound:

$$\alpha(M) = \sum_{s \in S} Q_s(\hat{\mu}) \tag{UB1}$$

- Solve the following NLP problem to local minimum:

$$\begin{aligned} \alpha(M) = \min_{\mu \in M, b} & \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}} b_{s,k} ||x_s - \mu_k||^2 \\ & \sum_{k \in \mathcal{K}} b_{s,k} = 1 \\ & 0 \leq b_{s,k} \leq 1, s \in \mathcal{S}, k \in \mathcal{K} \end{aligned} \tag{UB2}$$

Theorem

Construct $LB \in \{LB1, LB2, LB3\}$, $UB \in \{UB1, UB2\}$, the BB procedure converges by branching only on μ (center of clusters).

Experiment Analysis

Table 1: Comparison on datasets with state of the art. ($k = 2$)

Methods	UB	Nodes	Gap(%)
Padberg and Rinald's Dataset ($n = 2,392, d = 2$)			
Aloise et al., 2012	2.967×10^{10}	1	i[1] (50h)
serial	2.967×10^{10}	7	1.32 (4h)
serial	2.967×10^{10}	253	0.1 (11h)
20 cores	2.967×10^{10}	247	0.1 (1h)
Glass Identification ($n = 214, d = 9$)			
Aloise et al., 2012	cannot be solved		
serial	819.63	85	28.65 (4h)
serial	819.63	339	0.1 (9h)
20 cores	819.63	415	0.1 (1h)

Table 2: Performance of large dataset in parallel. (200 cores, $k = 3$)

Dataset	UB	Nodes	Gap(%)
Syn-210000	2.43×10^{6}	6	2.55

[1] Solved at the root node.

Contributions

Our work contributes for the following benefits:

- We provided a guaranteed global optimal solution for the minimum sum-of-squares clustering problem.
- By reformulating the clustering problem as a two-stage stochastic program problem, we proposed a tailed reduced space BB clustering algorithm that enables insensitivity to the scale of samples.
- By constructing proper upper and lower bounding problem, we are able to deal with datasets over 200,000 samples in a relatively short time ($\leqslant 4h$).