Optimal K-means and K-medoids Clustering

Zack Horton and Tanner Street

December 13th, 2023

MIT MANAGEMENT
SLOAN SCHOOL

Problem & Motivation

- **Background:** Clustering groups similar observations in distinct groups
  - K-Means uses centroids
  - K-Medoids uses medoids
- **Problem:** Clustering is performed with heuristic algorithms
  - No guarantee of optimality
  - Results vary based on random initializations
- **Goal:** Formulate MIO models for K-Means and K-Medoids clustering
- **Motivation:** Address the shortcomings of heuristic algorithms while remaining scalable and time efficient

MIT MANAGEMENT SLOAN SCHOOL

Data Collection

- Two datasets from UC Irvine Machine Learning Repository: **Abalone** and **Similarity Prediction**
- **Abalone:** 4,177 observations, 9 features (mostly continuous)
- **Similarity Prediction:** 100 observations, 17 features (mostly categorical)

MIT MANAGEMENT SLOAN SCHOOL

Methodology

MIT MANAGEMENT
SLOAN SCHOOL

Optimal K-Means Formulation

Variables for MIO:

Observation i assigned to cluster k: $z_{ik} \in \{0, 1\}$

Values for cluster $k's$ centroids: $c_{kj} \in \mathbb{R}$

Distance between i and cluster $k's$ centroid: $d_{ij} \geq 0$

Constraints for MIO:

Ensure Cluster Assignment:

$$\sum_{k=1}^{K} z_{ik} = 1\,, \quad \forall i \in [1, n]$$

Distance (if Manhattan): $d_{ij} = |x_{ij} - c_{kj}|$

$$d_{ij} \geq x_{ij} - c_{kj} - \mathbf{M} \times (1 - z_{ik})\,, \quad \forall i \in [1, n], \quad \forall j \in [1, p], \quad \forall k \in [1, K]$$

$$d_{ij} \geq c_{kj} - x_{ij} - \mathbf{M} \times (1 - z_{ik})\,, \quad \forall i \in [1, n], \quad \forall j \in [1, p], \quad \forall k \in [1, K]$$

Distance (if Euclidean): $d_{ij} = ||x_{ij} - c_{kj}||_2^2$

$$d_{ij} \geq (x_{ij} - c_{kj})^2 - \mathbf{M} \times (1 - z_{ik})\,, \quad \forall i \in [1, n], \quad \forall j \in [1, p], \quad \forall k \in [1, K]$$

Objective Function:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{p} d_{ij}$$

Optimal K-Medoids Formulation

Variables for MIO:

Observation i assigned to cluster k: $z_{ik} \in \{0,1\}$

Observation i is medoid of cluster k: $m_{ik} \in \{0,1\}$

Values for cluster $k's$ medoid: $c_{kj} \in \mathbb{R}$

Distance between i and cluster $k's$ medoid: $d_{ij} \geq 0$

Constraints for MIO:

Ensure Cluster Assignment:

$$\sum_{k=1}^{K} z_{ik} = 1 , \quad \forall i \in [1,n]$$

Ensure Medoid Assignment:

$$\sum_{i=1}^{n} m_{ik} = 1 , \quad \forall k \in [1,K]$$

$$\sum_{k=1}^{K} m_{ik} \leq 1 , \quad \forall i \in [1,n]$$

$$m_{ik} \leq z_{ik} , \quad \forall i \in [1,n], \quad \forall k \in [1,K]$$

Determine Medoid Values:

$$x'_{ij} - \mathbf{M} \times (1 - m_{ik}) \leq c_{kj} , \quad \forall i \in [1,n], \quad \forall j \in [1,p], \quad \forall k \in [1,K]$$

$$c_{kj} \leq x'_{ij} + \mathbf{M} \times (1 - m_{ik}) , \quad \forall i \in [1,n], \quad \forall j \in [1,p], \quad \forall k \in [1,K]$$

Distance (if Manhattan): $d_{ij} = |x_{ij} - c_{kj}|$

$$d_{ij} \geq x_{ij} - c_{kj} - \mathbf{M} \times (1 - z_{ik}) , \quad \forall i \in [1,n], \quad \forall j \in [1,p], \quad \forall k \in [1,K]$$

$$d_{ij} \geq c_{kj} - x_{ij} - \mathbf{M} \times (1 - z_{ik}) , \quad \forall i \in [1,n], \quad \forall j \in [1,p], \quad \forall k \in [1,K]$$

Distance (if Euclidean): $d_{ij} = ||x_{ij} - c_{kj}||_2^2$

$$d_{ij} \geq (x_{ij} - c_{kj})^2 - \mathbf{M} \times (1 - z_{ik}) , \quad \forall i \in [1,n], \quad \forall j \in [1,p], \quad \forall k \in [1,K]$$

Objective Function:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{p} d_{ij}$$

MIT MANAGEMENT SLOAN SCHOOL

Evaluation Procedure

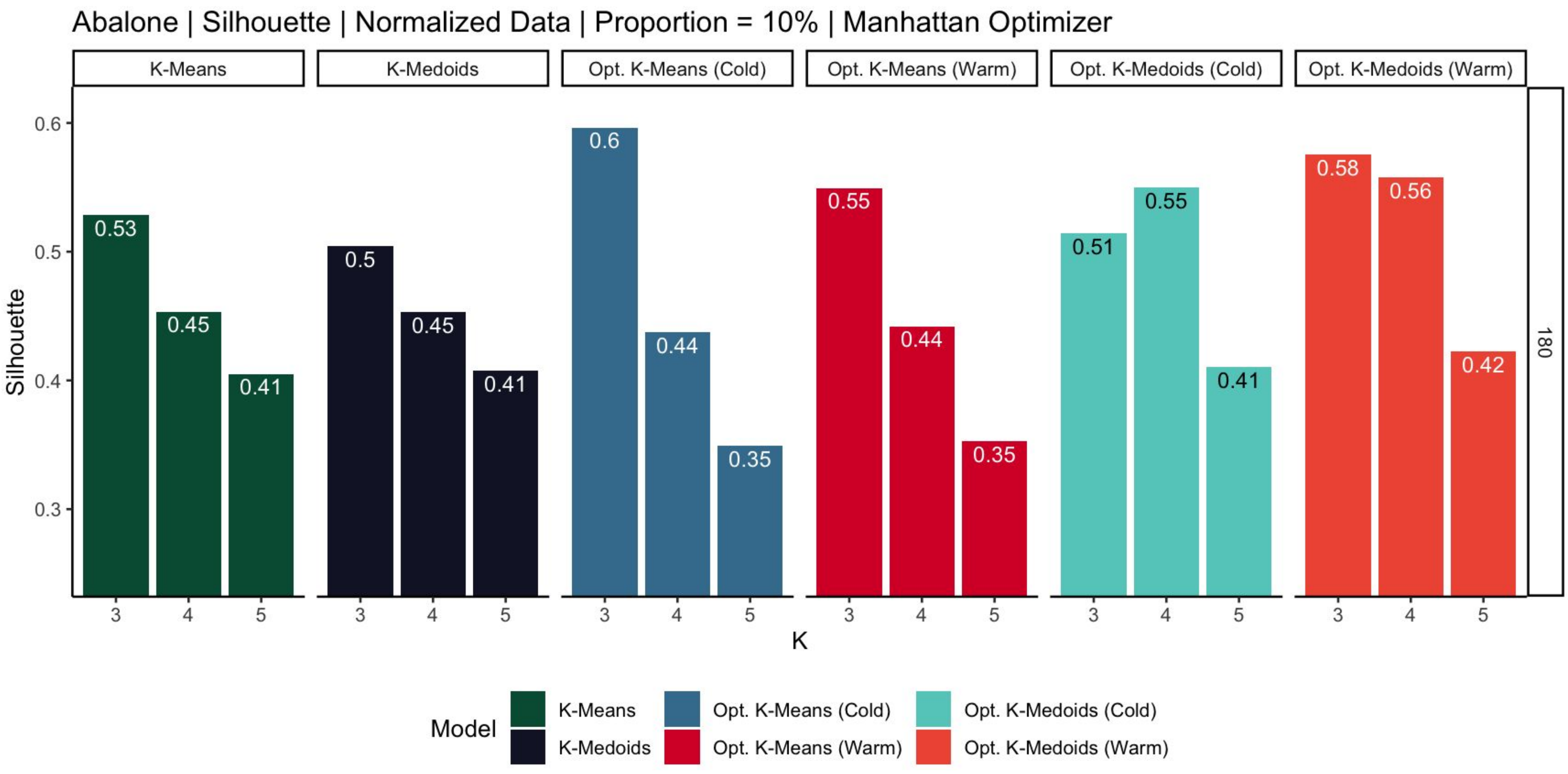
- Compare heuristic and MIO models on each dataset using large variety of parameters
- Reasonable ranges for number of clusters and time limit were determined
- Metrics: Within-cluster sum of squares and average silhouette scores

Preprocessing	Percentage of Observations	Warm Start	Distance Metric	Number of Clusters	Time Limit [sec]
Normalized	10%	Yes	Euclidean	3	30
Scaled	25%	No	Manhattan	4	90
	75%			5	180

MIT MANAGEMENT SLOAN SCHOOL

Key Findings

MIT MANAGEMENT
SLOAN SCHOOL


Abalone | Silhouette | Normalized Data | Proportion = 10% | Manhattan Optimizer
K-Means
K-Medoids
Opt. K-Means (Cold)
Opt. K-Means (Warm)
Opt. K-Medoids (Cold)
Opt. K-Medoids (Warm)
Silhouette
0.6
0.5
0.4
0.3
0.53
0.45
0.41
0.5
0.45
0.41
0.6
0.44
0.35
0.55
0.44
0.35
0.51
0.55
0.41
0.58
0.56
0.42
180
3
4
5
K
Model
K-Means
K-Medoids
Opt. K-Means (Cold)
Opt. K-Means (Warm)
Opt. K-Medoids (Cold)
Opt. K-Medoids (Warm)

Analyzing Results

	Similarity		Abalone	
	K-Means	K-Medoids	K-Means	K-Medoids
Avg % Improvement in WCSS: Heuristics to Opt. Warm Start	14%	15%	0.4%	-5.2%
Avg % Improvement in WCSS: Opt. Cold Start to Opt. Warm Start	4%	1%	50%	50%
Avg % Improvement in Silhouette: Heuristics to Opt. Warm Start	-1%	7%	-1%	14%
Avg % Improvement in Silhouette: Opt. Cold Start to Opt. Warm Start	4%	1%	80%	54%
Total % Euclidean Models w/o solution or no change to Heuristic	21%	11%	80%	14%

MIT MANAGEMENT SLOAN SCHOOL

THANK YOU
Questions?

Zack Horton

zhhorton@mit.edu

Tanner Street

street24@mit.edu

Similarity Dataset		Normalized	Scaled
Avg % Improvement in Silhouette: Heuristics to Opt. Warm Start	K-Means	-5%	2%
	K-Medoids	14%	-1%
Avg % Improvement in Silhouette: Heuristics to Opt. Cold Start	K-Means	-7%	-1%
	K-Medoids	14%	1%

Abalone Dataset		Normalized	Scaled
Avg % Improvement in Silhouette: Heuristics to Opt. Warm Start	K-Means	-1%	-1%
	K-Medoids	21%	6%
Avg % Improvement in Silhouette: Heuristics to Opt. Cold Start	K-Means	-29%	-35%
	K-Medoids	-20%	-32%

Similarity Dataset		10% Obs.	25% Obs.	75% Obs.
Avg % Improvement in Silhouette: Heuristics to Opt. Warm Start	K-Means	-6%	1%	1%
	K-Medoids	5%	11%	5%
Avg % Improvement in Silhouette: Heuristics to Opt. Cold Start	K-Means	-5%	-1%	-6%
	K-Medoids	6%	15%	2%

Abalone Dataset		10% Obs.	25% Obs.	75% Obs.
Avg % Improvement in Silhouette: Heuristics to Opt. Warm Start	K-Means	-1%	-1%	0%
	K-Medoids	15%	27%	-1%
Avg % Improvement in Silhouette: Heuristics to Opt. Cold Start	K-Means	-16%	-40%	-50%
	K-Medoids	-9%	-25%	-50%