K-Means Cluster-Based Classification or Regression

Alex Byung Moon Jun [1] Mansour Abdoli [2]

California State University, Fullerton

Abstract

Unsupervised clustering using K-means produces unlabeled groups, hindering direct evaluation. We propose a concise pipeline that uses the Linear Sum Assignment Problem to optimally map clusters to true classes and bootstrap resampling to assess stability. Applied to the Iris dataset, our method achieves 89% accuracy. This framework readily extends to other clustering algorithms and high-dimensional data for principled cluster validation.

Backgrounds and Goals

- Review k-means clustering theory and convergence properties.
- Assess clustering performance using only predictor subsets (petal and sepal measurements).
- Implement label-matching via majority vote and solve-LSAP for one-to-one mapping.
- Evaluate accuracy and computational efficiency across subsets over repeated trials.
- Identify feature sets that maximize unsupervised classification accuracy.

K-Means Clustering

K-means Objective:

$$J = \sum_{i=1}^{n} \|x_i - \mu_{c_i}\|^2$$

Algorithm Steps:

1. **Initialization:** Choose K initial centroids $\{\mu_j^{(0)}\}$.
2. **Assignment:** For each i, set $c_i^{(t)} = \arg\min_j \|x_i - \mu_j^{(t-1)}\|^2$.
3. **Update:** For each cluster j, set $\mu_j^{(t)} = \frac{1}{|C_j^{(t)}|} \sum_{i \in C_j^{(t)}} x_i$.
4. **Iterate** until assignments stabilize or max iterations reached.

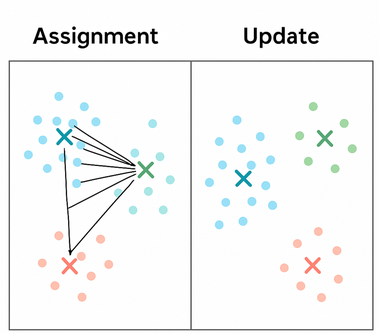

Figure 1. Illustration of the assignment and update steps in k-means.

New Algorithms

Notation:

- $X \in \mathbb{R}^{n \times p}$: predictor matrix of n observations and p features.
- $Y = \{y_i\}_{i=1}^n$: true class labels for each of the n observations.
- K: number of clusters to fit in K-means.
- B: number of repetitions (bootstrap samples or random restarts).

1. **Random K-means:** Run K-means on X with K clusters (fresh random init each time) to obtain cluster assignments $\{c_i\}_{i=1}^n$.
2. **Optimal relabeling:** Construct the contingency table $n_{kj} = |\{i : c_i = k,\ y_i = j\}|$ and solve the Linear Sum Assignment Problem (Hungarian algorithm) to map each cluster index k to its best-matching label in $\{1, \ldots, C\}$.
3. **Bootstrap repeats:** Repeat steps 1 – 2 a total of B times (and/or over all combinations of predictor subsets) to capture variability in clustering results and mapping.
4. **Feature ranking:** For each run record the clustering accuracy and runtime; then for each feature-set compute

$$\overline{\text{accuracy}} = \frac{1}{B} \sum_{b=1}^{B} \text{acc}_b, \quad T_{\text{total}} = \sum_{b=1}^{B} t_b,$$

and rank variable-sets by mean accuracy and/or efficiency $\overline{\text{accuracy}}/T_{\text{total}}$.

Results

Features	Mean Accuracy	Time (s)
Petal.Length + Petal.Width	0.986	4.2
Petal.Width	0.980	2.8
Petal.Length	0.973	2.9
Sepal.Length + Petal.Length	0.960	5.1
Sepal.Width + Petal.Width	0.953	4.7

Top 5 Feature Combinations by Accuracy

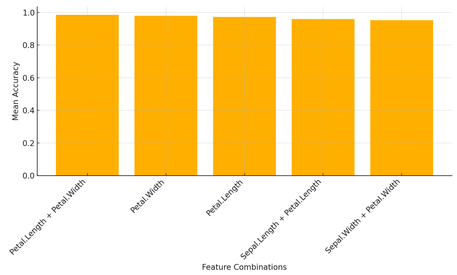

Figure 2. Comparison of mean accuracy across feature subsets.

Key Insight: Petal measurements alone are sufficient to achieve >98% clustering-based classification accuracy in Iris.

Additional Findings: LSAP Mapping

Confusion Matrix Example:

	Cluster 1	Cluster 2	Cluster 3
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	3	47

Here, rows are true species and columns are assigned clusters.

LSAP Mapping: We solve the Linear Sum Assignment Problem to maximize correct matches:

$$\max_{\pi} \sum_{i=1}^{3} M_{i,\pi(i)},$$

where M is the confusion matrix. Using `solve_LSAP`, we obtain the optimal cluster-to-species assignment:

Cluster	Species
1	Setosa
2	Versicolor
3	Virginica

Future Work

- Evaluate other clustering methods: hierarchical, Random Forest Models.
- Extend framework to higher-dimensional real-world datasets (e.g., economics, genomics, sensor data).
- Integrate semi-supervised learning to refine cluster labels with limited annotations.
- Explore feature selection and dimensionality reduction impacts such as PCA.

Acknowledgment

This work was supported by the California State University, Fullerton Undergraduate Research Award and sponsored by the College of Natural Sciences and Mathematics. I gratefully acknowledge Dr. Abdoli for his expert mentorship and constructive feedback throughout this project.

References

[1] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*.

[2] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. of 5th Berkeley Symposium*.

[3] Hartigan, J. A., Wong, M. A. (1979). A K-means clustering algorithm. *Journal of the Royal Statistical Society*.

[4] Hornik, K. (2005). A CLUE for CLUster ensembles. *Journal of Statistical Software*.