Enhancing Clustering Results with KM-SML Approach

Colonescu Anna-Diana
Cloud Computing and Internet of Things
Politehnica University of Timisoara
Timisoara, Romania
Email: anna.colonescu@student.upt.ro

***Abstract*—This paper introduces a hybrid approach combining K-Means clustering with supervised learning to improve clustering quality. Misclassified points are identified using Split Criterion (SC) and silhouette scores and reclassified using a Random Forest model. Applied to the Iris dataset and another dataset, the approach significantly enhances metrics, improving precision, recall, and accuracy from 83% to 87% (SC) and up to 94%-95% (silhouette). The results demonstrate the effectiveness of post-processing in refining clustering results and optimizing misclassification detection thresholds.**

***Index Terms*—Clustering, K-Means, Supervised Learning, Post-Processing.**

I. Introduction

Clustering is a key unsupervised learning technique for grouping similar data points. K-Means is widely used for its simplicity and efficiency, but it struggles with misclassifications in overlapping clusters. This paper proposes a hybrid approach, KM-SML, that combines K-Means with supervised learning for post-processing. Misclassified points are detected using Split Criterion (SC) and silhouette scores, then reclassified using Random Forest.

The KM-SML approach is evaluated on the Iris dataset and a second dataset, demonstrating significant improvements in precision, recall, and accuracy. Threshold optimization ensures a balance between true misclassification detection and minimizing false positives. This study highlights the potential of integrating unsupervised and supervised techniques for clustering refinement.

The rest of the paper covers the theoretical background (Section II), datasets (Section III), methodology (Section IV), experimental results (Section V), and conclusions (Section VI).

II. Theory and Background

A. Clustering and K-Means

Clustering is an unsupervised machine learning method that groups similar data points into clusters, maximizing intra-cluster similarity and minimizing inter-cluster similarity. It is widely used in areas like market segmentation, image recognition, and anomaly detection.

K-Means is a popular partition-based clustering algorithm that divides data into k clusters by iteratively optimizing cluster centroids. The steps of the K-Means algorithm are as follows:

1) **Initialization**: Randomly select k initial centroids.
2) **Assignment**: Assign each data point to the cluster with the closest centroid using a distance metric (e.g., Euclidean distance).
3) **Update**: Compute new centroids as the mean of all data points assigned to each cluster.
4) **Stopping Criteria**: Repeat the assignment and update steps until centroids stabilize or a predefined number of iterations is reached.

K-Means is computationally efficient and straightforward but has limitations such as sensitivity to the initial centroid positions and difficulty handling non-spherical clusters.

B. Elbow Method

The Elbow Method is a heuristic used to determine the optimal number of clusters (k) for K-Means. It involves plotting the within-cluster sum of squares (WCSS) against various values of k. The WCSS is calculated as:

$$WCSS = \sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - \mu_j||^2,$$

where x_i is a data point, μ_j is the centroid of cluster C_j, and k is the number of clusters.

The "elbow point" on the curve, where the rate of decrease in WCSS slows significantly, indicates the optimal k. This ensures a balance between model simplicity and clustering quality.

C. Post-Processing

Post-processing refines clustering results by identifying and reassigning misclassified points using methods such as:

- **Split Criterion (SC)**: SC identifies potential misclassifications by evaluating a point's stability within its cluster. It compares distances to the assigned centroid and other centroids, assigning a value between (0, 1]. Higher values indicate a greater likelihood of misclassification. Thresholds are used to flag points, with lower thresholds detecting more misclassifications but risking false positives, while higher thresholds reduce false positives but may miss some misclassified points.

- **Silhouette Scores**: Silhouette scores assess clustering quality by measuring a point's cohesion within its cluster and separation from others. Scores range from -1 to 1:
  - **Close to 1**: Indicate well-clustered points.
  - **Near 0**: Suggest ambiguity near cluster boundaries.
  - **Close to -1**: Highlight potential misclassification.

  Average silhouette scores offer a global measure of clustering quality and are useful for optimizing cluster assignments.

D. Random Forest

Random Forest (RF) is a robust supervised learning algorithm that operates as an ensemble of decision trees. It is used for classification and regression tasks due to its ability to handle high-dimensional data and reduce overfitting through averaging.

The Random Forest algorithm works as follows:

1) **Bootstrapped Samples**: Random subsets of the training data are created with replacement.
2) **Decision Trees**: Each subset is used to train a decision tree. At each split, a random subset of features is considered.
3) **Majority Voting**: For classification, predictions from all trees are aggregated, and the majority vote determines the final output.

Random Forest is effective in reclassifying misclassified points identified during the post-processing stage, further enhancing clustering results.

III. Datasets

The experiments were conducted using two datasets: the widely recognized Iris dataset and a second dataset of unspecified origin.

A. Iris Dataset

The Iris dataset contains 150 records, each representing a plant. It includes five features: Sepal Length, Sepal Width, Petal Length, Petal Width (all measured in centimeters), and the class label, which identifies the plant's species.

B. Second Dataset

The second dataset comprises records with 13 features, most of which are fractional numbers, except for two attributes.

IV. Methodology

This chapter outlines the process of applying and evaluating the clustering and post-processing techniques used in this study. It is divided into four components: clustering with K-Means, post-processing using supervised learning, threshold optimization, and evaluation of results.

A. Clustering with K-Means

The K-Means algorithm is employed to partition datasets into clusters. The number of clusters (k) is determined using the elbow method. The steps include:

- **Initialization**: Randomly select k initial centroids.
- **Assignment**: Assign each data point to the nearest centroid based on Euclidean distance.
- **Update**: Recompute centroids as the mean of all points in their respective clusters.
- **Iteration**: Repeat the assignment and update steps until centroids stabilize or a predefined number of iterations is reached.

For the labeled Iris dataset, the Hungarian algorithm is used to align K-Means cluster labels with the true class labels, enabling effective comparison and evaluation.

B. Post-Processing Using Supervised Learning

Post-processing refines clustering results by addressing potential misclassifications:

- **Identifying Misclassified Points**: Two methods are utilized:
  - **Split Criterion (SC)**: Points with low stability within their clusters are flagged as potential misclassifications based on SC values.
  - **Silhouette Scores**: Points with low silhouette scores, indicating poor cohesion or separation, are identified as misclassified.
- **Reclassification with Random Forest**: Correctly classified points are used to train a Random Forest model, which reclassifies the flagged points, improving clustering quality.

C. Threshold Optimization

To determine the optimal thresholds for SC and silhouette scores, an iterative approach is employed:

- Define an array of threshold values for each method.
- Execute the code for each threshold and evaluate clustering results using precision, recall, and accuracy metrics.
- Select the threshold that provides the best values for precision, recall, and accuracy metrics.

D. Evaluation

The effectiveness of the clustering and post-processing methods is assessed using:

- **Metrics**: Precision, recall, and accuracy are calculated before and after post-processing to quantify improvements.
- **Visualizations**: Scatter plots and heatmaps are generated to compare clustering quality and misclassification rates.

This methodology combines the simplicity of K-Means, the robustness of supervised learning, and iterative threshold optimization to achieve enhanced clustering results.

V. Experimental Results

The machine used to process the Iris dataset has the following characteristics: Windows 10, Intel Core i5-6200U CPU @ 2.30GHz (dual-core, 4 logical processors), 8 GB of DDR3 RAM memory, and a standard hard drive.

A. Determining Optimal Clusters Using the Elbow Method

The elbow method was used to determine the optimal number of clusters (k):

- For the Iris dataset, the elbow point indicated $k = 3$.
- For Dataset 2, the elbow point suggested $k = 3$.

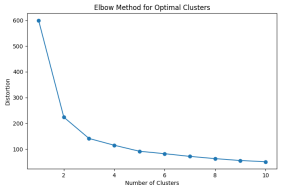

Fig. 1. Elbow Method for the Iris dataset.

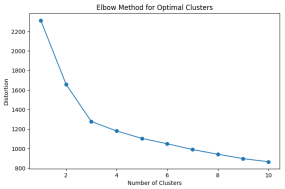

Fig. 2. Elbow Method for Dataset 2.

B. Initial K-Means Clustering

Scatter plots of the initial K-Means clustering results are presented. Misclassified points are evident, particularly in boundary regions:

- Iris dataset results show overlaps in cluster boundaries.
- Dataset 2 results highlight areas requiring further refinement.

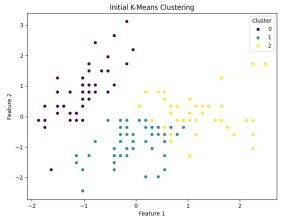

Fig. 3. Initial K-Means clustering for the Iris dataset.

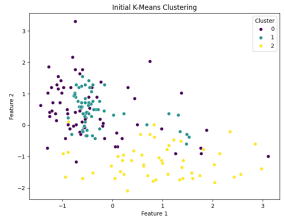

Fig. 4. Initial K-Means clustering for Dataset 2.

C. Identifying Stable and Misclassified Points

Two methods were applied to identify misclassified points:

- **Split Criterion (SC)**: Figures 5 and 7 show stable and misclassified points for the Iris dataset and Dataset 2, respectively.
- **Silhouette Scores**: For the Iris dataset, Figure 6 visualizes points flagged as misclassified using silhouette scores.

In the visualization, misclassified points are labeled as -1, while points labeled as 0, 1, and 2 represent correctly classified points corresponding to their respective clusters.

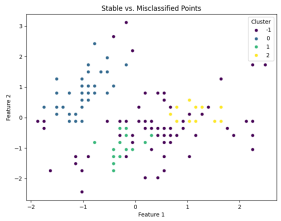

Fig. 5. Stable vs. misclassified points using SC for the Iris dataset.

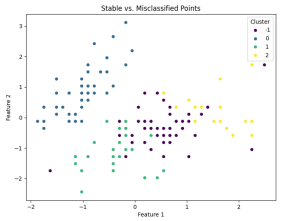

Fig. 6. Stable vs. misclassified points using silhouette scores for the Iris dataset.

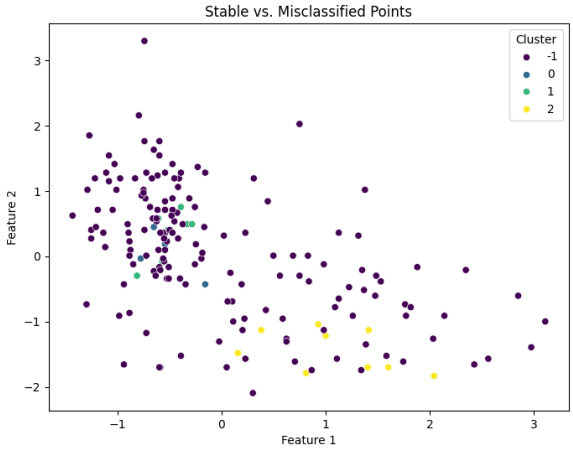

Fig. 7. Stable vs. misclassified points using SC for Dataset 2.

D. Post-Processing and Reclassification

Post-processing improved clustering quality by reclassifying misclassified points:

- **Using SC**: Figures 8 and 10 show reclassified points for the Iris dataset and Dataset 2.
- **Using Silhouette Scores**: Figure 9 presents results for the Iris dataset, demonstrating improved cluster cohesion and separation.

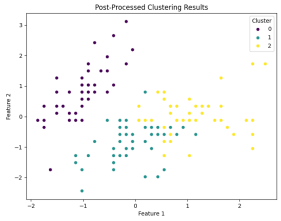

Fig. 8. Post-processed clustering using SC for the Iris dataset.

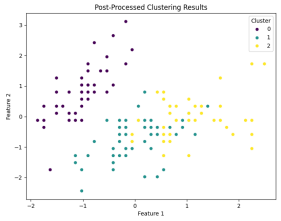

Fig. 9. Post-processed clustering using silhouette scores for the Iris dataset.

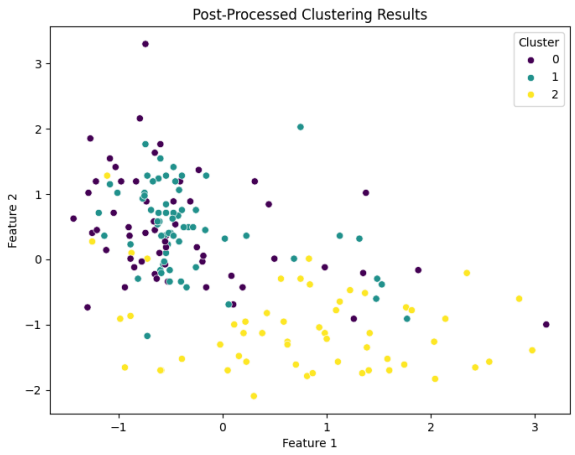

Fig. 10. Post-processed clustering using SC for Dataset 2.

E. Threshold Comparison for SC and Silhouette Scores

For the Iris dataset, thresholds for SC and silhouette scores were iteratively tested. Figure 11 shows the comparison of precision, recall, and accuracy across thresholds.

It can be observed that for both methods the optimal thresold is 0.4.

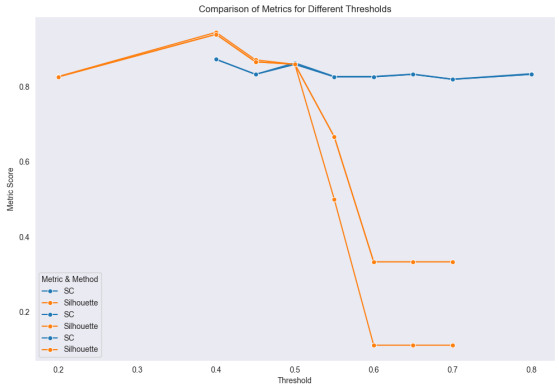

Fig. 11. Threshold comparison for SC and silhouette scores on the Iris dataset.

F. Final Metrics

The final clustering metrics highlight the improvements achieved:

- **SC-based metrics (Iris dataset)**: Table I shows precision, recall, and accuracy improvements.
- **Silhouette-based metrics (Iris dataset)**: Table II demonstrates similar enhancements.

TABLE I
FINAL METRICS USING SC FOR THE IRIS DATASET.

Step	Precision	Recall	Accuracy
K-Means	0.83	0.83	0.83
K-Means + Random Forest	0.87	0.87	0.87

TABLE II
FINAL METRICS USING SILHOUETTE SCORES FOR THE IRIS DATASET.

Step	Precision	Recall	Accuracy
K-Means	0.83	0.83	0.83
K-Means + Random Forest	0.95	0.94	0.94

VI. CONCLUSION

The KM-SML hybrid approach significantly improves clustering quality by addressing the limitations of K-Means through supervised learning. Misclassified points, identified using Split Criterion (SC) and silhouette scores, are effectively reclassified, resulting in substantial performance gains.

For the **Iris dataset**, SC increased precision, recall, and accuracy from 0.83 to **0.87**, while silhouette scores further boosted these metrics to **0.95**, **0.94**, and **0.94**, respectively. The **second dataset** demonstrated similar improvements, highlighting the robustness of the approach across diverse datasets. Optimal threshold selection ensures a balance between detecting true misclassifications and minimizing false positives, making the method adaptable to various scenarios.

These results demonstrate that combining clustering with supervised learning effectively refines clustering outcomes, particularly for complex or overlapping clusters. Future work can explore applying this approach to larger datasets, alternative clustering methods, and other supervised models like gradient boosting or neural networks.

REFERENCES

[1] I.-D. Borlea, R.-E. Precup, and A.-B. Borlea, "Improvement of K-means Cluster Quality by Post Processing Resulted Clusters," in *The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)*, Procedia Computer Science, Elsevier, 2022, pp. 63–70. Available: https://www.sciencedirect.com/science/article/pii/S1877050922000188

[2] F. Wang, H.-H. Franco-Penya, J. Kelleher, J. Pugh, and R. Ross, "An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity," in *International Conference on Artificial Intelligence and Cognitive Science*, Springer, Cham, 2017, pp. 257–268. doi: 10.1007/978-3-319-62416-7_21