Clustering Analysis Report: K-Means vs DBSCAN

2023101136

1 Introduction

This report presents a clustering analysis on a dataset containing 21 2D points representing three letter groups: **N**, **A**, and **V**. The objective is to evaluate the clustering performance using two different algorithms: **K-Means** and **DBSCAN**. We examine how well each method identifies natural groupings based on spatial proximity.

2 Clustering Setup

- **Number of data points:** 21 (7 from each group: N, A, V)
- **Features:** 2D Coordinates (X, Y)
- **Methods used:**
  - K-Means with good and bad initial centroids
  - DBSCAN with $\varepsilon = 10$ and `MinPts` $= 3$

3 K-Means Clustering Results

Good Initial Centroids

- Initial centroids placed near centers of letter groups:
  - Cluster 0 (N): [5, 7.5]
  - Cluster 1 (A): [25, 7.5]
  - Cluster 2 (V): [45, 7.5]
- K-Means converged in 2 iterations.

Final Clusters

Cluster	Points
0	N1, N2, N3, N4, N5, N6, N7
1	A1, A2, A3, A4, A5, A6, A7
2	V1, V2, V3, V4, V5, V6, V7

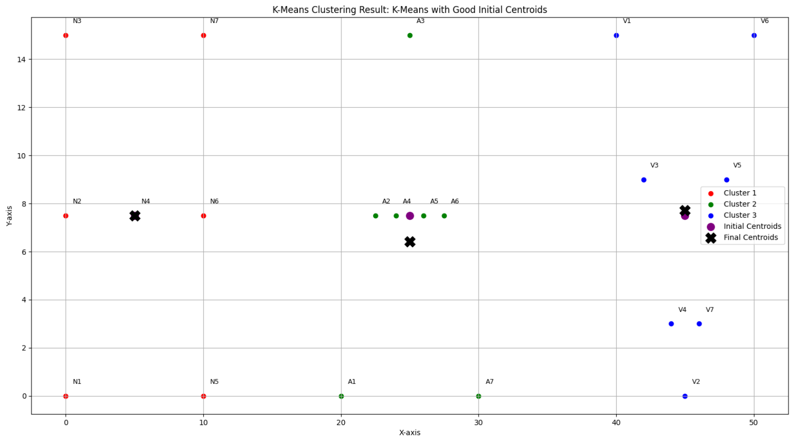

Figure 1: K-Means Clustering Output (Good Initial Centroids)

Bad Initial Centroids

- Initial centroids were intentionally placed far from optimal cluster centers:
  - Cluster 0: [0, 0]
  - Cluster 1: [5, 5]
  - Cluster 2: [45, 5]
- This setup caused overlapping and confusion during early clustering iterations.

Even with poorly chosen initial centroids, the algorithm eventually converged to the correct clusters after 5 iterations. However, early iterations had misclassified points, showing that K-Means is sensitive to centroid initialization.

Final Clusters (After Convergence)

Cluster	Points
0	N1, N2, N3, N4, N5, N6, N7
1	A1, A2, A3, A4, A5, A6, A7
2	V1, V2, V3, V4, V5, V6, V7

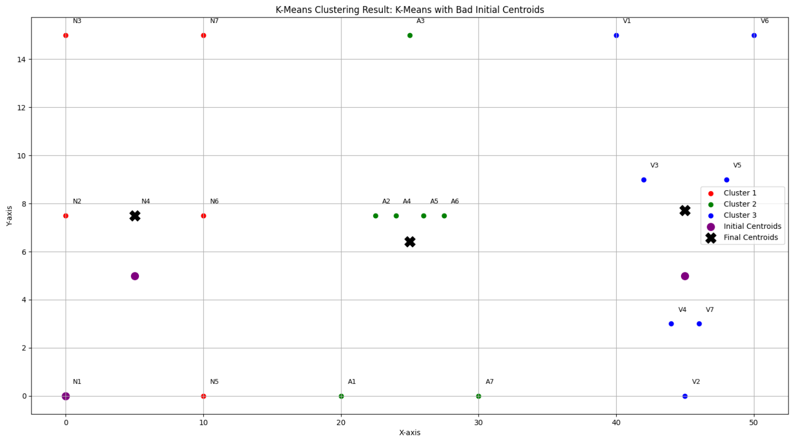

Figure 2: K-Means Clustering Output (Bad Initial Centroids)

4 DBSCAN Clustering Results

Parameters

- $\varepsilon = 10$
- MinPts $= 3$

Core Point Detection

All points were identified as core points due to high density and connectivity within each letter group.

Final Clusters

Cluster	Points
0	N1, N2, N3, N4, N5, N6, N7, A1, A2, A3, A4, A5, A6, A7
1	V1, V2, V3, V4, V5, V6, V7

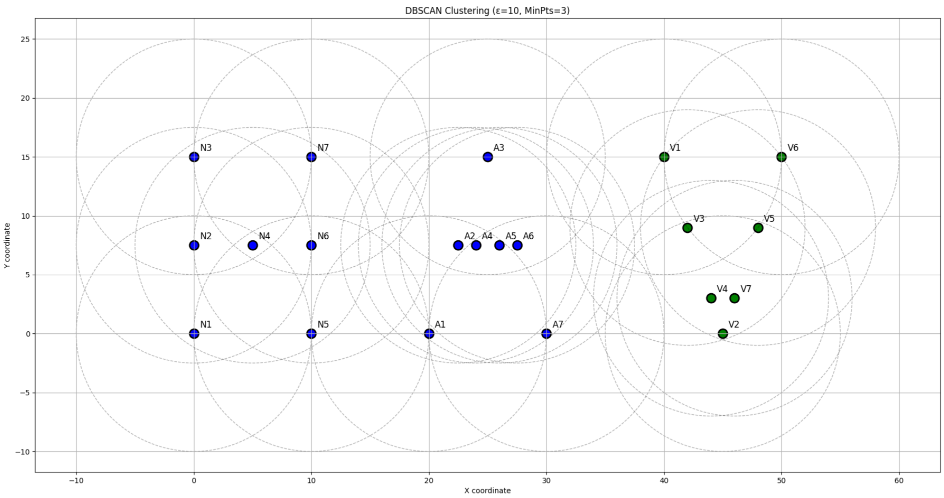

Figure 3: DBSCAN Clustering Output

Observations

- Letters N and A are considered a single cluster due to spatial proximity.
- DBSCAN correctly identified V as a separate cluster.
- No noise points were detected.

5 Analysis and Comparison

Distinct Clusters

- **K-Means:** Successfully formed distinct clusters for N, A, and V (3 clusters total).
- **DBSCAN:** Formed only 2 clusters—grouping N and A together due to their closeness.

Comparison Table

Aspect	K-Means	DBSCAN
Number of Clusters	3	2
Sensitivity to Initialization	High	Low
Can detect noise?	No	Yes
Handles arbitrary shape?	No	Yes
Performance on this dataset	Excellent	Good (minor merging)

Table 1: Comparison of K-Means and DBSCAN

6 Conclusion

Both K-Means and DBSCAN performed well on the dataset, but with distinct behaviors:

- **K-Means** produced clearly defined clusters for each letter group, provided good initial centroids were used. It is effective when the number of clusters is known and clusters are roughly spherical.
- **DBSCAN** grouped letters N and A into one cluster due to density overlap, demonstrating its strength in discovering clusters of arbitrary shape but sensitivity to parameter tuning (ε and MinPts).

Recommendations

- Use **K-Means** when the number of clusters is known and well-separated.
- Use **DBSCAN** for discovering clusters in noisy datasets or when the number of clusters is unknown.
- Potential improvements: try fine-tuning ε or using hierarchical clustering for better separation.