


DS5230 Final: Phase 3

Dimensionality Reduction, Clustering, and Evaluation



Michael Massone and Nelson Farrell

Spring 2024



Data Overview:

Dry Bean Dataset

- Size: (13611 x 17) - 13611 rows, 17 columns
- This dataset is composed of data derived from 13,611 images of 7 species of beans. Bean images obtained by computer vision system were subjected to segmentation and feature extraction stages, and a total of 16 features; 12 dimensions and 4 shape forms, were generated.
- Koklu, M. and Ozkan, I.A., (2020), "Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques." Computers and Electronics in Agriculture, 174, 105507.
- DOI:
<https://doi.org/10.1016/j.compag.2020.105507>
- Link:
<https://archive.ics.uci.edu/dataset/602/dry+bean+dataset>

Feature Descriptions

- **Area (A)** - *float64 - Ratio* - The area of a bean zone and the number of pixels within its boundaries.
- **Perimeter (P)** - *float64 - Ratio* - Bean circumference is defined as the length of its border.
- **MajorAxisLength (L)** - *float64 - Ratio* - The distance between the ends of the longest line that can be drawn from a bean.
- **MinorAxisLength (l)** - *float64 - Ratio* - The longest line that can be drawn from the bean while standing perpendicular to the main axis.
- **AspectRatio (K)** - *float64 - Ratio* - Defines the relationship between L and l.
- **Eccentricity (Ec)** - *float64 - Ratio* - Eccentricity of the ellipse having the same moments as the region.
- **ConvexArea (C)** - *float64 - Ratio* - Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
- **EquivDiameter (Ed)** - *float64 - Ratio* - The diameter of a circle having the same area as a bean seed area.
- **Extent (Ex)** - *float64 - Ratio* - The ratio of the pixels in the bounding box to the bean area.
- **Solidity (S)** - *float64 - Ratio* - Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
- **Roundness (R)** - *float64 - Ratio* - Calculated with the following formula: $\frac{(4\pi A)}{(P^2)}$
- **Compactness (CO)** - *float64 - Ratio* - Measures the roundness of an object: $\frac{Ed}{L}$
- **ShapeFactor1 (SF1)** - *float64 - Ratio* - $\frac{L}{A}$
- **ShapeFactor2 (SF2)** - *float64 - Ratio* - $\frac{l}{A}$
- **ShapeFactor3 (SF3)** - *float64 - Ratio* - $\frac{A}{\frac{L}{2} \cdot \frac{l}{2} \cdot \pi}$
- **ShapeFactor4 (SF4)** - *float64 - Ratio* - $\frac{A}{\frac{L}{2} \cdot \frac{l}{2} \cdot \pi}$

Optimum UMAP Hyperparameters

UMAP Gridsearch:

```
min_dist_list = [0.0]
n_neighbors_list = [50, 100, 250, 500, 750, 1000]
metric_list = ['cosine', 'euclidean']
n_components_list = range(2, 9)
```

Gridsearch was tuned using a sample size of $n=1000$. The following parameter ranges were decided based on these preliminary results.

- Min_dist was set to 0 to encourage the formation of very tight, dense clusters.
- Large values of n_neighbors was used to capture global structure of the data over local. The values of n_neighbors used for tuning on sampled data were increased by an order of magnitude to better relate to the true size of the dataset.
- A euclidean and cosine distance metric were chosen to explore different relationships between the data objects. The cosine distance was included to capture linear relationships between features.
- N_components range was established from 2 up to half the number of features.

Optimum UMAP Parameters

KMEANS:

Hyperparameters:

```
n_neighbors: 500
min_dist: 0.0
metric: euclidean
n_components: 7
```

Results:

```
Hopkin's Statistic = 0.0028945957265941
Test2: Pass
Number of Clusters: 15
Silhouette Score: 0.51362556
```

DBSCAN:

Hyperparameters:

```
n_neighbors: 750
min_dist: 0.0
metric: cosine
n_components: 7
```

Results:

```
Hopkin's Statistic = 0.013332919591190
Number of Clusters: 4
Validity Index: 0.9992647931722908
Noise Ratio: 0.0
```

Note: Both clustering algorithms found their optimum solution at value of `n_components=7`. This points to the possible existence of a latent manifold in 7-dimensions. They also used relatively high values of `n_neighbors`, favoring a global representation of the data.

Optimum Clustering Algorithm

KMEANS:

Kmeans Hyperparameters:

k^1 : 15

Results:

Hopkin's Statistic = 0.0028945957265941

Test2: Pass

Number of Clusters: 15

Silhouette Score: 0.51362556

DBSCAN:

DBSCAN Hyperparameters:

Eps²: 0.265387

Min_samples³: 0.5

Metric⁴: cosine

Results:

Hopkin's Statistic = 0.013332919591190

Number of Clusters: 4

Validity Index: 0.9992647931722908

Noise Ratio: 0.0

1. The number of clusters the algorithm will find. We ranged over 2-16 for each UMAP embedding.
2. The length of the radius that defines the neighborhood of a point in clustering with center-based density.
3. The number of other points within a points eps radius necessary to define it as a core point.
4. The pairwise distance measure. For DBSCAN the embedding is transformed into distance matrix using the specified metric before computing.

Internal Indices

KMEANS:

Internal Indices:

Silhouette Score¹: 0.51362556

Hopkin's Statistic² = 0.0028945957265941

Results:

Number of Clusters: 15

DBSCAN:

Internal Indices:

Validity Index³: 0.9992647931722908

Noise Ratio⁴: 0.0

Hopkin's Statistic = 0.013332919591190

Results:

Number of Clusters: 4

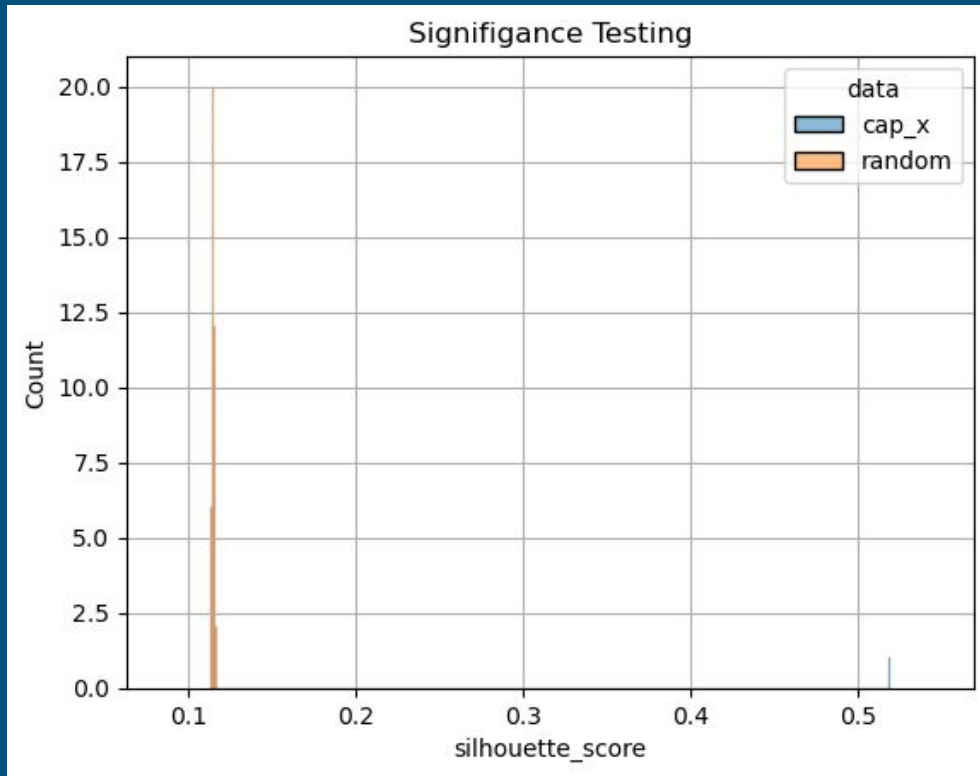
1. The mean silhouette coefficient over all data objects - measure cluster cohesion and separation on a scale of $[-1, 1]$. A good score (closer to 1) indicates well separated, globular clusters.
2. An indicator of how clustered the data is within the data space. Bound by $[0, 1]$, with values closer to 0 indicating well clustered data.
3. A density based clustering index. Bounded by $[-1, 1]$, with greater values indicating better clustering.
4. The percentage data objects classified as noise points. We used a threshold of 10% to eliminate clustering solutions that discounted too much of the data.

Results Table

index	70	69
algo	dbscan	k_means
n_clusters_found	4	15
n_clusters_db_score_is_min	NaN	15
n_clusters_ch_score_is_max	NaN	15
n_clusters_silhouette_score_is_max	NaN	15
silhouette_score	NaN	0.519526
hopkins_statistic	0.00941	0.002861
umap_n_neighbors	1000	750
umap_min_dist	0	0
umap_metric	cosine	euclidean
umap_n_components	7	7
trustworthiness	0.84725	0.999954
eps	0.265387	NaN
dbscan_min_samples	5	NaN
dbscan_metric	cosine	NaN
validity_index	0.999869	NaN
noise_ratio	0	NaN

Checking For False Discoveries: Kmeans

- This graphic displays the observed silhouette score versus the distribution of silhouette scores from a data set with the same dimensionality, uniformly distributed, and within the range (min, max) of the observed data set.
- This graphic helps demonstrate that our discovered clustering solution is not a false discovery, i.e., the observed silhouette score is unlikely to be achieved on random data.
- We base this conclusion on the fact that there appears to be a significant difference between the observed silhouette score and the random distribution.
- We conclude that we have discovered a valid clustering solution.



Validation with External Indices

- Using true labels cluster solution validation is possible using external indices.
- This is not the case in a true unsupervised setting.
- This analysis will validate the clustering solution using the following external indices:
 - Adjusted Rand Score
 - Fawlk's and Mallows Score
 - F-1 Score
 - Jaccard Score
 - Purity Score (Cluster Homogeneity)

Validation with External Indices

Algorithm: k_means

Number of Clusters Found: 15

Adjusted Rand Score: 0.42887

Fawlks and Mallows Score: 0.53950

F-1 Score: 0.07050

Jaccard Score: 0.04816

Normalized Mutual Info Score: 0.72172

Homogeneity Score (Purity): 0.89315

REMAPPED CONTINGENCY MATRIX

Best Mapping: {0.0: 4.0, 1.0: 14, 2.0: 6.0, 3.0: 0.0, 4.0: 2.0, 5.0: 5.0, 6.0: 9, 7.0: 10, 8.0: 8, 9.0: 3.0, 10.0: 11, 11.0: 12, 12.0: 15, 13.0: 13, 14.0: 16}

Contingency Matrix:

[[711	0	0	0	0	0	0	0	0	0	0	0	0	0	611]
[165	0	0	0	0	0	0	0	0	0	0	0	0	357	0]
[0	876	0	0	0	0	0	0	0	117	0	0	0	637	0]
[0	0	941	0	0	0	0	61	0	0	882	832	830	0	0]
[0	0	0	975	0	0	125	0	0	828	0	0	0	0	0]
[0	0	0	0	833	0	0	0	832	0	0	0	0	0	362]
[0	0	0	0	0	0	927	869	840	0	0	0	0	0	0]

Looking at the contingency matrix we can see that kmeans came very close to finding the true solution. Although 15 clusters were found, it seems that kmeans was subdividing the true clusters with most true labels being split fairly evenly between 2 or more predicted labels. The homogeneity score is quite high as a result. Due to the dimensionality of the embedding, we cannot actually visualize this data to confirm this is what is happening. With additional post processing we believe a valid clustering solution could be obtained from this result.

Validation with External Indices

Algorithm: dbscan

Number of Clusters Found: 4

Adjusted Rand Score: 0.14241

Fawlks and Mallows Score: 0.47102

F-1 Score: 0.00079

Jaccard Score: 0.00040

Normalized Mutual Info Score: 0.43696

Homogeneity Score (Purity): 0.29794

REMAPPED CONTINGENCY MATRIX

Best Mapping: {0.0: 5.0, 1.0: 3.0, 2.0: 8, 3.0: 1.0}

Contingency Matrix:

[0	1315	7	0]
[522	0	0	0]
[0	1630	0	0]
[0	3546	0	0]
[0	1928	0	0]
[0	63	1188	776]
[0	2636	0	0]]

- Looking at the contingency matrix we can see that DBscan did not perform particularly well. The only true label it successfully identified is **1**. DBscan broke true label **5** into predominately **2** clusters, and the rest of the true labels it predicted label **1**.

External Validation Results

true_num_clusters	7	7
umap_n_components	7	7
umap_min_dist	0	0
umap_n_neighbors	1000	750
umap_metric	cosine	euclidean
trustworthiness	0.84725	0.999954
algo	dbscan	k_means
n_clusters_found	4	15
silhouette_score	NaN	0.519526
validity_index	0.999869	NaN
adj_rand_score	0.142414	0.428874
fawlks_and_mallows	0.471018	0.539496
nmi	0.436957	0.721724
jaccard_score	0.000398	0.048161
f1_score	0.000795	0.070500

Discussion of Results

- An overall assessment of the results based primarily on Adjusted Rand Score, Cluster Homogeneity, and Contingency Matrices.
- **Adjusted Rand Score:**
 - Adjusted Rand Score is a similarity measure between two clustering solutions.
 - The Rand index computes the similarity between two clustering solutions by analyzing pairs of samples and whether those are in the same or different clusters in the predicted clusters and the true clusters.
 - The Adjusted Rand Score adjusts this metric for chance.
 - Adjusted Rand Score is a good choice for metric because it satisfies the following:
 - Bounded: $[-1, 1]$
 - Distinct score for random labeling: 0
 - Does not make assumptions about the cluster structure.
 - DBscan performed poorly (**Adj. Rand Score = 0.14241**).
 - Kmeans performed adequately (**Adj. Rand Score = 0.42887**).
 -
- **Cluster Homogeneity:**
 - Homogeneity (purity) is a measure of how “pure” the clusters are, i.e., to what extent the clusters contain a single true label.
 - Kmeans performed well (**Cluster Homogeneity = 0.89315**).
 - DBscan performed poorly (**Cluster Homogeneity = 0.29794**).

Acknowledgments

- KOKLU, M. and OZKAN, I.A., (2020), "Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques." Computers and Electronics in Agriculture, 174, 105507. DOI: <https://doi.org/10.1016/j.compag.2020.105507>
- Dr. Steve Morin – Class slides and labs.
- UC Irvine Machine Learning Repository
- Sci-kit Learn Documentation:
 - scikit-learn.org
- UMAP Documentation:
 - umap-learn.readthedocs.io/en/latest/
- HDBSCAN Documentation:
 - clusteringjl.readthedocs.io/en/latest/dbscan.html