# Clustering Algorithm Comparisons

Michael Carlstrom
Case Western Reserve University
United States
rmc170@case.edu

## 1 INTRODUCTION

This paper discusses various clustering algorithm for the cho and iyer datasets and evaluates their pros and cons. In this paper we will discuss K-Means clustering, Density Based clustering and Spectral clustering by calculating the Rand Index for each as well as comparing their Principal Component Analysis (PCA).

## 2 METHODOLOGY

### 2.1 Data Preprocessing

First we started by applying normalization on the features for improved performance and accuracy. This is done by preventing features with large value ranges from overpowering features with small value ranges by giving them uniform value ranges from 0 to 1.

*2.1.1 Normalization.*

For our data normalization we used Min-Max Scaling. We used scikit-learn's normalize() implementation to normalize via Min-Max scaling.

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

### 2.2 Evaluation Metrics

*2.2.1 Confusion Matrix.*

A confusion matrix is a table that summarize a classification models performance. Confusion matrices can be used to calculate all sorts of relevant metrics for classification including accuracy, precision, recall, F1 score, etc...

- TP (True Positives): Correct positive predictions
- TN (True Negatives): Correct negative predictions.
- FP (False Positives): Incorrect positive predictions.
- FP (False Negatives): Incorrect negative predictions.

| Confusion Matrix | Positive | Negative |
|---|---|---|
| Positive | TP | FN |
| Negative | FP | TN |

*2.2.2 Rand Index.*

The Rand Index can be calculated from a confusion matrix with the following formula. We use scikit-learn's rand_score() implementation for calculating.

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

*2.2.3 Principal Component Analysis.*

Principal Component Analysis (PCA) is a dimension reduction technique. PCA is typically used to counter the curse of dimensionality which is when there are lots of features for a data point which can lead to longer training times as well overfitting. PCA finds the n most meaningful dimension thus reducing the data size improving training times and reducing overfitting. Obviously if you remove to many features the accuracy and other relevant metrics will suffer. In this project it will be used for helping visualize higher dimension clusters into the second dimension rather than any data modification. This is done since it is impossible to visualize something above the third dimension in an image.

PCA works since it lowers the dimensionality but maximizes the variance of the data preserving the relationships. It works with the following process.

(1) Normalization:
We start by normalizing all features as described in the data preprocessing section.

(2) Relationships:
Next we find the relationship between the data using a covariance matrix.

$$cov(x1, x2) = \frac{\sum_{i=1}^{n}(x1_i - \overline{x1})(x2_i - \overline{x2})}{n-1}$$

(3) Principal Components:
We then use the covariance matrix to find the principal components. The principal components are the eigenvectors of the covariance matrix. Then we keep the n most information rich eigenvectors.
To calculate the eigenvectors use the following formulas

$$AX = \lambda X$$
$$AX - \lambda X = 0$$
$$A(-\lambda I)X = 0$$

Since A has to be invertible.

$$|A - \lambda I| = 0$$

Solving this equation gives the eigenvalues of the matrix.

$$AX = \lambda X$$

Plugging back into the original equation to find the eigenvectors.

## 2.3 K-Means

A partial clustering approach for k clusters. Each cluster has defined centroid which gets updated at each stage.

The algorithm for creating K-means clusters is defined below.

```
(1) Select K points as the initial centroids.
(2) repeat
(3)     Form K clusters by assigning all points to the closet centroid.
(4)     Recompute the centroid of each cluster.
(5) until The centroids don't change.
```

A common objective function for K-means is Sum of Square Error (SSE).

$$SSE = \sum_i \sum_{x \in C_i} (x - c_i)^2$$

## 2.4 Density Based

Clusters are defined as region's with high density that are separated by regions of low destiny. In Density Based classification there are three types of points. A point is a core point if has a minimum amount of points (MinPts) with radius $\epsilon$. This also includes points in the interior of the cluster. A border point is not a core point, but within the neighborhood of a core point. A noise point is any point remaining.

1. Label all points as core, border, or noise points.
2. Eliminate noise points.
3. Put an edge between all core points within a distance $\epsilon$ of each other.
4. Make each group of connected core points into a separate cluster.
5. Assign each border point to one of the clusters of its associated core points.

A common objective function for Density Based is Total Sum of Squares (TSS) which is the sum of previously mentioned SSE and new component SSB.

$$TSS = SSE + SSB$$
$$SSB = \sum_i |C_i|(c - c_i)^2$$

## 2.5 Spectral

Similar to K-means however focus on more complex geometries since K-means is designed for spherical clusters. This is done by generating a similarity graph then determining where to cut the graph into the respective clusters. The algorithm for this is defined below.

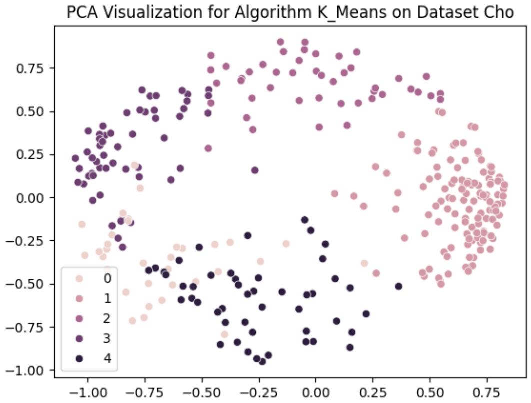
(1) Create a sparsified similarity graph $G$.
(2) Compute the graph Laplacian for $G$, $L$.
(3) Create a matrix $V$ from the first $k$ eigenvectors of $L$.
(4) Apply K-means clustering on $V$ to obtain $k$ clusters.

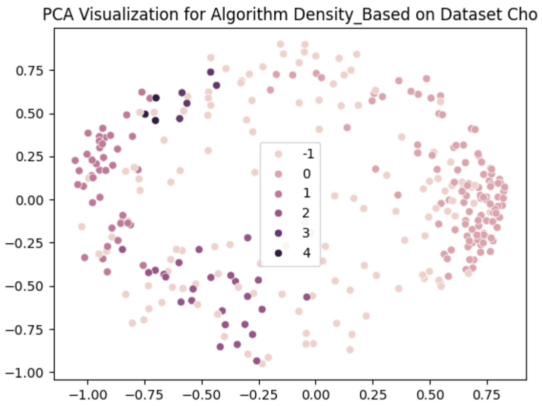
# 3 EVALUATION AND PERFORMANCE

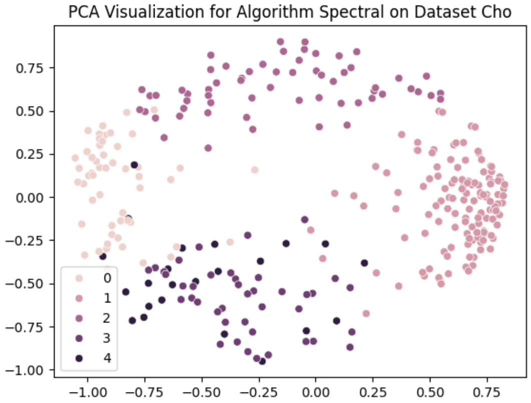
To assess the performance of the implemented cluster classification algorithms we computed the Rand Index for each cluster. We also created cluster visualization for comparisons of the algorithms.

Cho Dataset Rand Index Scores.

| Model | RI |
|---|---|
| K-means | 0.79387 |
| Density Based | 0.60000 |
| Spectral | 0.79453 |

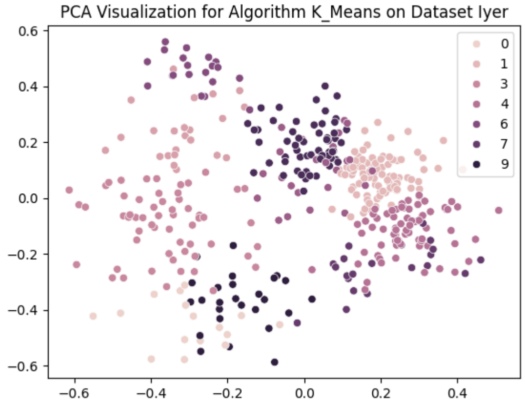


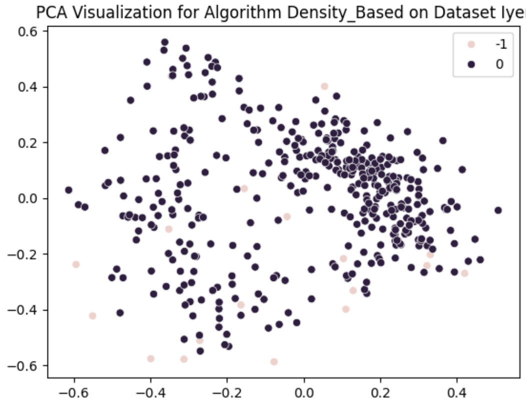


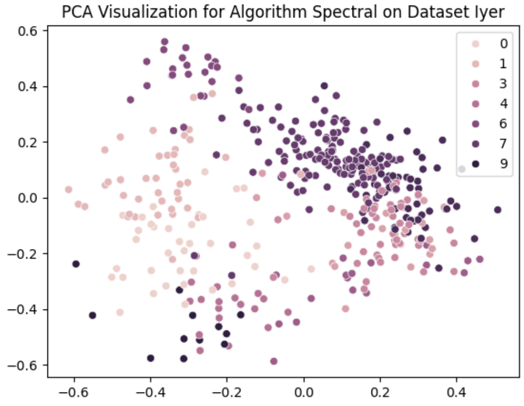


Iyer Dataset Rand Index Scores.

| Model | RI |
|---|---|
| K-means | 0.84372 |
| Density Based | 0.63666 |
| Spectral | 0.83738 |







# 4 CONCLUSION

For the Cho dataset the K-means and Spectral have comparable Rand Index scores. Their actual clusters are slightly different in relation to the bottom left clusters. Since Density Based clustering does not take some value for number of clusters it determined their was 6 clusters. The left and right most clusters mostly match up to the other two algorithms but, the rest are kind of nonsensical.

For the Iyer dataset the K-means and Spectral have comparable Rand Index scores. Their actual clusters are slightly different in relation to the upper right clusters. Something interesting to note is that some clusters don't appear at all in the PCA implying their features are more secondary when it comes to classifying them. Density Based really struggle on the Iyer dataset since it would very easily flip between far too many clusters and far too few. As seen by the plot using PCA only two clusters are identified.

## 4.1 Pros and Cons of Classification Algorithms

**K-means:** Extremely effective at finding spherical clusters, but heavily reliant on knowing how many clusters.

**Density Based:** Can handle noise and outliers but, as shown if cannot correctly determine number of clusters performance decreases.

**Spectral:** Similar to K-means but can find more advanced geometries than just high-dimensional spheres but, also struggles if the cluster count is wrong.

# 5 ACKNOWLEDGMENT

I would like to thank Dr. Jing Ma for her exemplary teaching this semester. I would also like to thank all the teacher assistants to data mining for helping answer questions.

# REFERENCES

(1) J. Ma, "CSDS 435 Data Mining Lectures" in https://canvas.case.edu/, 2025
(2) "scikit-learns Documentation" in https://scikit-learn.org, 2025