# Evaluation of Clustering Algorithms

Panagiotis Michael
U204N2110
michael.p15@live.unic.ac.cy

## I. Introduction

### A. Goals Of The Project

The goal of this project is to successfully investigate the application of 3 clustering algorithms on 3 datasets. The clustering algorithms are k-Means, Agglomerative Clustering and DBSCAN. The data from the 3 different datasets are related to countries, fish and mall customers respectively.

### B. Summary of the results

The results from the clustering experiments were okay. All algorithms performed adequately.

## II. Datasets

### A. Country Dataset

The *country* dataset consists of 167 entries of 167 countries with 10 attributes. The attributes are *country*, *child_mort*, *exports*, *health*, *imports*, *income*, *inflation*, *life_expec*, *total_fer* and *gdpp*. The *country* attribute is nominal and it contains the names of the countries. The rest of the attributes are numerical. *child_mort* represents the death of children under 5 years of age per 1000 live births. *exports* are the exports of goods and services per capita. *health* is the total health spending per capita. *imports* are the imports of goods and services per capita. *income* is the net income per person. *inflation* is the measurement of the annual growth rate of the total GDP. *life_expec* is the average number of years a new born child would live if the current mortality patterns are to remain the same. *total_fer* is the number of children that would be born to each woman if the current age-fertility rates remain the same. *gdpp* is the GDP per capita and it is calculated as the total GDP divided by the total population [1]. You can see the first 5 entries of the *country* dataset in Figure 1.

| Row No. | country | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Afghanistan | 90.200 | 10 | 7.580 | 44.900 | 1610 | 9.440 | 56.200 | 5.820 | 553 |
| 2 | Albania | 16.600 | 28 | 6.550 | 48.600 | 9930 | 4.490 | 76.300 | 1.650 | 4090 |
| 3 | Algeria | 27.300 | 38.400 | 4.170 | 31.400 | 12900 | 16.100 | 76.500 | 2.890 | 4460 |
| 4 | Angola | 119 | 62.300 | 2.850 | 42.900 | 5900 | 22.400 | 60.100 | 6.160 | 3530 |
| 5 | Antigua and ... | 10.300 | 45.500 | 6.030 | 58.900 | 19100 | 1.440 | 76.800 | 2.130 | 12200 |

Fig. 1. The first 5 entries of the *country* dataset.

### B. Fish Dataset

The *fish* dataset consists of 159 entries of 7 fish species with 7 attributes. The attributes are *Species*, *Weight*, *Length1*, *Length2*, *Length3*, *Height* and *Width*. The *Species* attribute is nominal and it contains the name of the species for each fish entry. The rest of the attributes are numerical. *Weight* is the weight of the fish in grams. *Length1* is their vertical length in cm. *Length2* is the diagonal length in cm. *Length3* is the cross length in cm. *Height* is their height in cm and *Width* is the diagonal width in cm [2]. You can see the first 5 entries of the *fish* dataset in Figure 2.

| Row No. | Species | Weight | Length1 | Length2 | Length3 | Height | Width |
|---|---|---|---|---|---|---|---|
| 1 | Bream | 242 | 23.200 | 25.400 | 30 | 11.520 | 4.020 |
| 2 | Bream | 290 | 24 | 26.300 | 31.200 | 12.480 | 4.306 |
| 3 | Bream | 340 | 23.900 | 26.500 | 31.100 | 12.378 | 4.696 |
| 4 | Bream | 363 | 26.300 | 29 | 33.500 | 12.730 | 4.455 |
| 5 | Bream | 430 | 26.500 | 29 | 34 | 12.444 | 5.134 |

Fig. 2. The first 5 entries of the *fish* dataset.

### C. Mall Customers Dataset

The *Mall Customers* dataset consists of 200 entries of mall customers with 5 attributes. The attributes are *CustomerID*, *Gender*, *Age*, *Annual Income (k$)* and *Spending Score (1-100)*. The *CustomerID* and *Gender* are nominal with *CustomerID* being a unique ID assigned to each customer and *Gender* being the gender of the customer. The rest of the attributes are numerical. *Age* is the age of the customer, *Annual Income (k$)* is the annual income of the customer and *Spending Score (1-100)* is the score assigned by the mall based on customer behavior and spending nature [3]. You can see the first 5 entries of the *mall customers* dataset in Figure 3.

| Row No. | CustomerID | Gender | Age | Annual Inco... | Spending Score (1-100) |
|---|---|---|---|---|---|
| 1 | 1 | Male | 19 | 15 | 39 |
| 2 | 2 | Male | 21 | 15 | 81 |
| 3 | 3 | Female | 20 | 16 | 6 |
| 4 | 4 | Female | 23 | 16 | 77 |
| 5 | 5 | Female | 31 | 17 | 40 |

Fig. 3. The first 5 entries of the *mall customers* dataset.

## III. Clustering Algorithms

### A. k-Means Algorithm

k-Means is a clustering algorithm used to classify a set of n-dimensional points into k clusters. The algorithm works by first randomly initializing k ”centroids” (the center point of each cluster), and then iteratively assigning each point to the cluster whose centroid is closest to it, and then recalculating the position of the centroid based on the points assigned to it. This process continues until the centroids no longer move or a maximum number of iterations is reached. The result is k clusters, each represented by a centroid, and with the points closest to that centroid in that cluster [4]. You can see a visual representation of the algorithm steps in Figure 4.

### B. DBSCAN Algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups together data points that are closely packed together (points with many nearby neighbors), and marks as outliers the points that are in low-density regions (points with few or no nearby neighbors). The algorithm takes in two main parameters, epsilon (eps) and minimum number of points (minPts), which are used to define the density of a cluster. Eps is the maximum distance between two points for them to be considered as part of the same neighborhood, and minPts is the minimum number of points required to form a dense region.

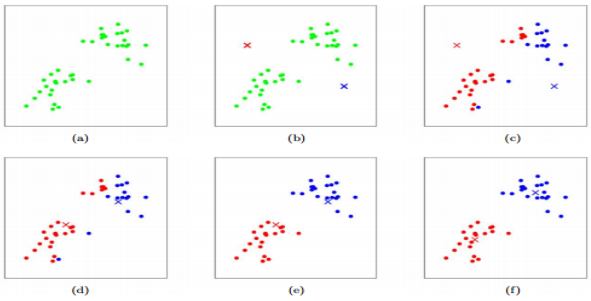

Fig. 4. Visual representation of the k-Means algorithm. Adapted from [4]

Points that are in the dense regions are considered as core points, and points that are not in the dense regions, but are close to them, are considered as border points. In order to apply the algorithm first a point is randomly selected, and if it has at least minPts within distance eps, it starts a new cluster with all of those points, then the cluster is expanded by adding any point within distance eps that is not already in the cluster and if a point has less than minPts within distance eps, it is considered as noise [5]. You can see a possible result of the DBSCAN algorithm in Figure 5.

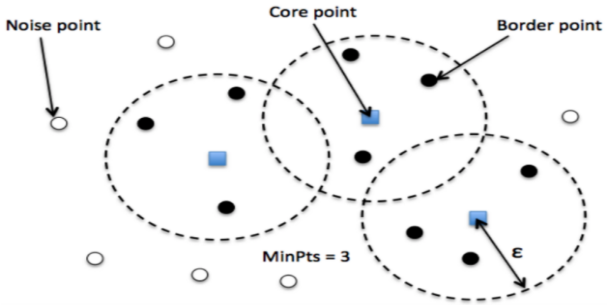


Fig. 5. Sample result of the DBSCAN algorithm. Adapted from [5]

### *C. Agglomerative Clustering Algorithm*

Agglomerative Clustering is a type of hierarchical clustering algorithm that builds a hierarchy of clusters by merging smaller clusters into larger ones. The process starts by treating each data point as an individual cluster and then iteratively merging the two closest clusters until a stopping criterion is met. The stopping criterion can be the number of clusters desired, or a threshold for the distance between clusters. The final result is a tree-like structure called a dendrogram, where each leaf node represents a single data point and each internal node represents a merged cluster. It can be decided at which level of the dendrogram to cut the tree to obtain the desired number of clusters. In order to measure the distance between the clusters linkage methods are used, which can be single linkage, complete linkage, average linkage, and others. Each linkage method uses a different distance metric to measure the distance between two clusters. The algorithm is a bottom-up method, so it starts with individual points and builds clusters by merging them. You can see a sample result of the agglomerative clustering algorithm in Figure 6.

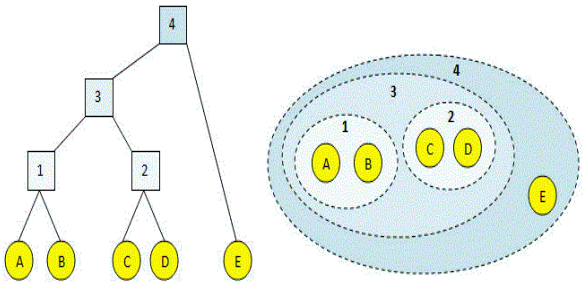


Fig. 6. Sample result of agglomerative clustering. With yellow are the points that are initially considered as individual clusters. The clusters formed are indicated with dotted lines. Adapted from [6]

## IV. Results

To get the necessary results, appropriate data pre-processing was applied with the inclusion of principal component analysis (PCA). Note: Plots can be viewed within RapidMiner whilst following the report, since the processing is not demanding at all and RapidMiner gave me a hard time exporting the plots. k plots can be manually created within the "Optimize Parameters" tab in the results section. In the *Country*, *Fish* datasets, you can set the labels as colors in order to get a different version of the results, which helps compare the results, in the case of the *Fish* dataset, and find the outliers by name, in the case of the *Country* dataset.

### A. k-Means Results

To help determine the value of k for each dataset, an average within centroid distance plot was used, which had in the other axis values of k up until 10. This wasn't the only method, as observation also played a big role.

*a) Country Dataset:* k = 3. (pc1,pc4), (pc1,pc2) and (pc1,pc3) showed clear segmentations.

*b) Fish Dataset:* k = 7. Tried to replicate the 7 class labels. Plot suggested 4. Clear segmentations with clusters close to original. Can be compared when setting color to Species.

*c) Mall Customers Dataset:* k = 2. Clear segmentations.

### B. DBSCAN Results

*a) Country Dataset:* eps = 0.4, minpts= 5. Came up with 1 cluster and some noise/outlier points, which represent countries that are outside of the pack.

*b) Fish Dataset:* eps = 0.3, minpts = 4. Came up with 9 clusters, which is not initially as optimal but could be in the future. The number of clusters exceed the number of available classes. Could aim for less clusters but the dataset is not fit for that.

*c) Mall Customers Dataset:* eps = 1.3, minpts = 4. Came up with 2 clusters and 2 noise points. The segmentations are clear.

### C. Agglomerative Clustering Results

Here the "Flatten" procedure was used, where the tree is cut and separate clusters are formed in order to get observable and plottable results.

*a) Country Dataset:* mode = complete link, number of clusters = 3. (pc1,pc4), (pc1,pc2) and (pc1,pc3) showed clear segmentations.

*b) Fish Dataset:* mode = complete link, number of clusters = 4. Clear Segmentations. Trying to replicate original clusters from labels did not work.

*c) Mall Customers Dataset:* mode = single link, number of clusters = 2. All component combinations gave good segmentations.

## V. Discussion

All datasets were equally fun to cluster and were equal in difficulty. Some tweaks were required for each dataset, regarding the algorithms, since they couldn't work with the same parameters. The experiments gave some very distinct and distinguishable clusters from all the algorithms. PCA helped to create uncorrelated features which were also capable of being clustered. From the application of DBSCAN on the *Country* dataset, some countries were classsified as noise/outliers which are Singapore, Malta, Luxemburg, Qatar, United States. These countries can be considered upper outliers or rich countries . Futhermore, it can be observed that agglomerative clustering gave similar results to k-Means applications. Lastly, it is of worth to note that in the k-Means clustering of the *Mall Customers* dataset, (pc2,pc3) gives a very intriguing segmentation, where maybe the gender attribute plays a role.

## VI. Conclusions

- Proper pre-processing is needed in order to execute clustering algorithms efficiently.
- Many tests are required to find the right parameters that are suited for the problem and the dataset.
- Multiple combinations of parameters can give "good" results.
- External means such as class labels, help to provide further inside to the segmentations from the algorithms.

Clustering is a very interesting topic that requires a lot of investigation and understanding. There are cases that there is not definitive correct answer which makes it even more a challenge.

## References


[1] R. Kokkula, "Unsupervised learning on country data," Kaggle, 17-Jun-2020. [Online]. Available: https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data. [Accessed: 20-Jan-2023].

[2] A. Pyae, "Fish market," Kaggle, 13-Jun-2019. [Online]. Available: https://www.kaggle.com/datasets/aungpyaeap/fish-market. [Accessed: 20-Jan-2023].

[3] V. Choudhary, "Mall Customer Segmentation Data," Kaggle, 11-Aug-2018. [Online]. Available: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python. [Accessed: 20-Jan-2023].

[4] C. Piech, "K Means ," CS221, 2012. [Online]. Available: http://stanford.edu/ cpiech/cs221/handouts/kmeans.html. [Accessed: 20-Jan-2023].

[5] N. S. Chauhan, "DBSCAN clustering algorithm in machine learning," KDnuggets, 2022. [Online]. Available: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html. [Accessed: 20-Jan-2023].

[6] R. Edgar, "Agglomerative clustering," drive5. [Online]. Available: https://drive5.com/usearch/manual/agg.html. [Accessed: 20-Jan-2023].