Khulna University of Engineering & Technology (KUET), Khulna

Department of Computer Science and Engineering (CSE)

Course No: CSE4239

Course Title: Data Mining

Assignment on Cluster Analysis (Assignment 1)

Colab Link:
https://colab.research.google.com/drive/1iiufzP44ssfgukrRyhcg8DT-Ej71jTn6?usp=sharing

Submitted To

Animesh Kumar Paul

Assistant Professor

Department of CSE, KUET

Submitted By

Md. Noyan Ali (1607021)

Department of CSE, KUET

Date of Submission: 14/02/2021

Cluster Analysis

Introduction

Clustering is the most important unsupervised learning. It finds structures in a collection of unlabeled data. Actually, it is a process of organizing objects into groups whose elements are similar in some way. As a result, a cluster is similar to the objects belonging to it and dissimilar to the objects belonging to other clusters. There are several techniques for clustering. A comparison on different clustering algorithms using different datasets with performance measurements is shown here.

Clustering Algorithms

Partitional clustering, hierarchical clustering and density-based clustering techniques have been used to cluster analysis. The algorithms that are used are listed below:

- Partitional Clustering Approach
    - KMeans
    - KMedoids

- Hierarchical Clustering Approach
    - Agglomerative Nesting (**AGNES**)
    - Balanced Iterative Reducing & Clustering Using Hierarchies (**BIRCH**)

- Density-based Clustering Approach
    - Density-Based Spatial Clustering of Applications with Noise (**DBSCAN**)

For implementation purpose, these algorithms are used from *sklearn* library in python.[1]

Datasets

The datasets that are used to experiment different clustering algorithms are generated from *sklearn datasets* library [2]. The name of the datasets is listed below:

- *Sklearn dataset blobs*
- *Sklearn dataset moons*
- *Sklearn dataset circles*

Each dataset contains 1500 samples with 2 features. Artificial noises have been added. Each dataset has been preprocessed using feature scaling before using to the clustering algorithms.[3]

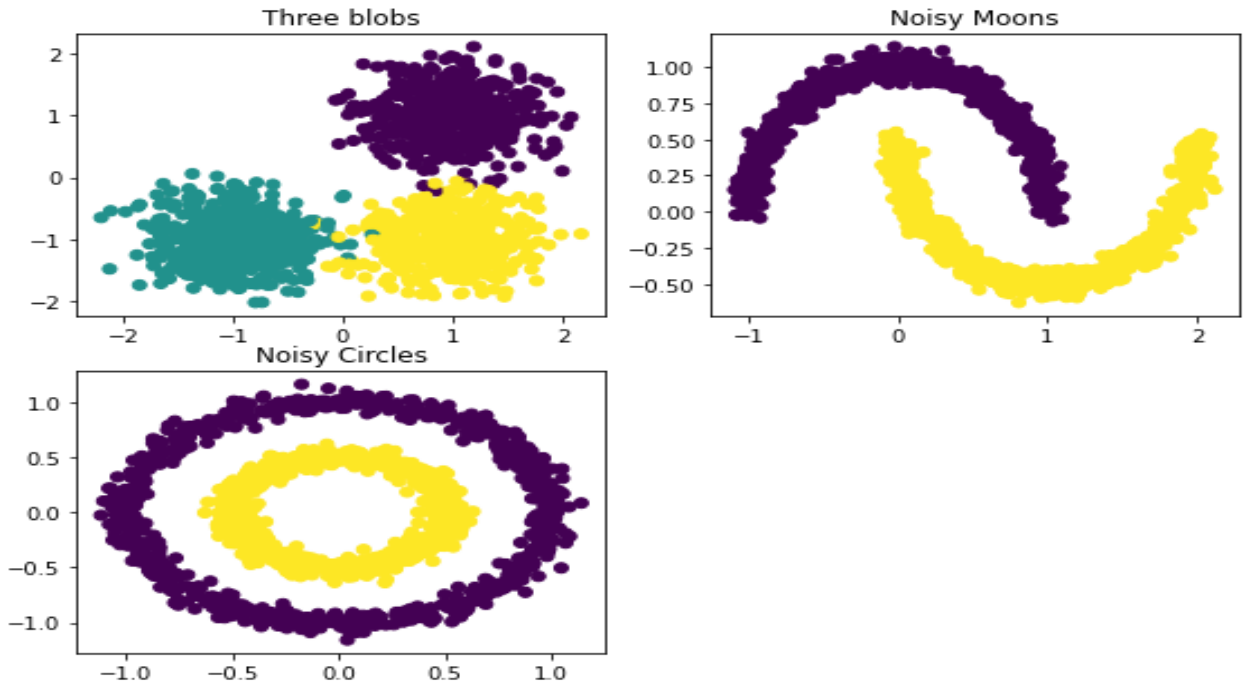

Figure 1: Datasets used in cluster analysis

Evaluating Metrics

Different evaluating metrics are used to show the performance of different algorithms on different datasets. Silhouette score, Adjusted Rand Index (ARI) score and Normalized Mutual Information (NMI) score have been used. Short description about these metrics is given below:

- Silhouette score's value ranges from -1 to 1. 1 means clusters are well apart from each other and clearly distinguished. 0 Means clusters are indifferent or the distance between clusters is not significant and -1 Means clusters are assigned in the wrong way.[4]

- ARI score's value ranges from 0 to 1. 0 means random labelling and 1 means perfect labelling.[5]

- NMI score's value also ranges from 0 to 1 and have same meaning like ARI.[6]

Experiment on *Sklearn Blobs* Dataset

KMeans, KMedoids, AGNES, BIRCH and DBSCAN algorithms have been used on *sklearn blobs* dataset. The results are shown in Figure 2. It shows that all algorithms perform very good clustering on it. Partitional clustering approaches such as KMeans and KMedoids perform well because cluster shapes are very simple with same density and spherical.

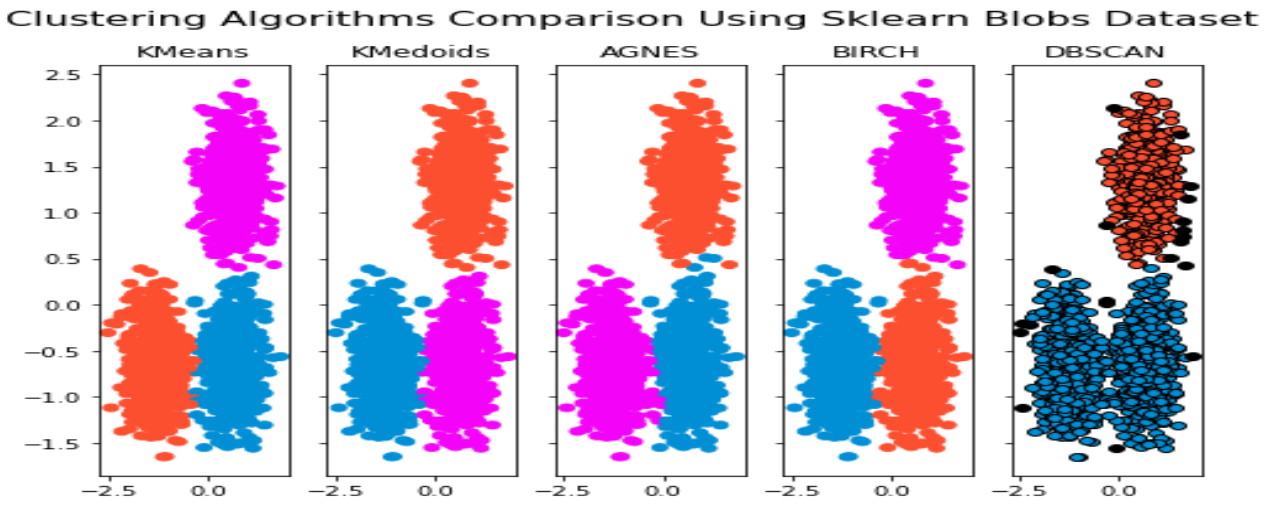

Figure 2: Clustering Algorithms Comparison using *sklearn blobs* dataset

Evaluating metrics of KMeans, KMedoids, AGNES, BIRCH and DBSCAN algorithms on *sklearn blobs* dataset is shown in Table 1. ARI score and NMI score show that all algorithms perform well on this dataset but silhouette score fails to show it. However, the table shows that all algorithms perform well on this dataset.

Table 1: Performance of clustering algorithms on *sklearn blobs* dataset using different evaluating metrics

Evaluating Metrics	KMeans	KMedoids	AGNES	BIRCH	DBSCAN
Silhouette	0.64	0.64	0.64	0.64	0.45
ARI	0.97	0.97	0.95	0.97	0.54
NMI	0.95	0.95	0.92	0.95	0.67

Experiment on *Sklearn Moons* Dataset

KMeans, KMedoids, AGNES, BIRCH and DBSCAN algorithms have been used on *sklearn moons* dataset. The results are shown in Figure 3. It shows that DBSCAN performs very well on this dataset because it is capable of finding clusters of arbitrary shapes and sizes. But the KMeans, KMedoids, AGNES and BIRCH fail to cluster this dataset because it has complex cluster shapes and sizes.

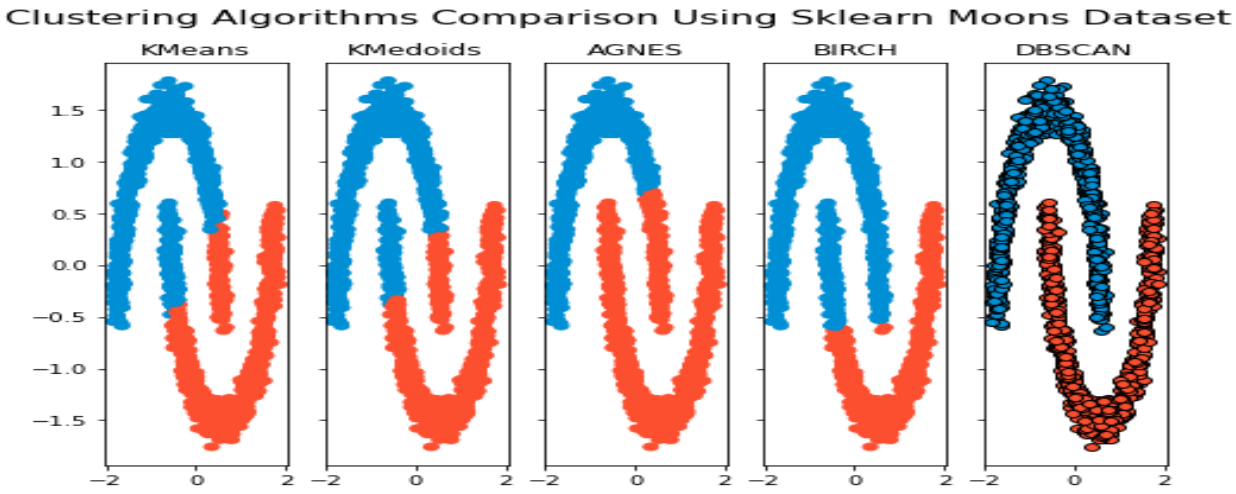

Figure 3: Clustering Algorithms Comparison using *sklearn moons* dataset

Evaluating metrics of KMeans, KMedoids, AGNES, BIRCH and DBSCAN algorithms on *sklearn moons* dataset is shown in Table 2. All the metrics show that DBSCAN performs very well on this complex clustering shapes and other algorithms fail.

Table 2: Performance of clustering algorithms on *sklearn moons* dataset using different evaluating metrics

Evaluating Metrics	KMeans	KMedoids	AGNES	BIRCH	DBSCAN
Silhouette	0.5	0.49	0.46	0.46	0.39
ARI	0.49	0.51	0.66	0.59	1.0
NMI	0.39	0.42	0.64	0.58	1.0

Experiment on *Sklearn Circles* Dataset

KMeans, KMedoids, AGNES, BIRCH and DBSCAN algorithms have been used on *sklearn circles* dataset. The results are shown in Figure 4. It shows that DBSCAN performs very well on this dataset because it can perform well on arbitrary shape and size. But the KMeans, KMedoids, AGNES and BIRCH fail to cluster this dataset because it has a circle shaped cluster which is very complex and these algorithms can't deal with complex cluster shapes.

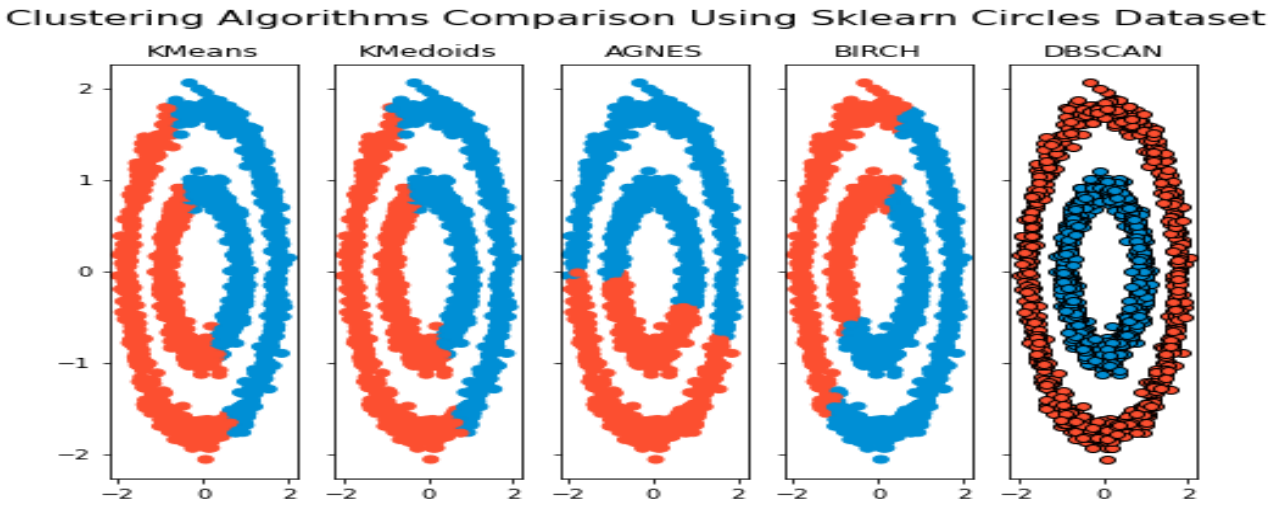

Figure 4: Clustering algorithms comparison using *sklearn circles* dataset

Evaluating metrics of KMeans, KMedoids, AGNES, BIRCH and DBSCAN algorithms on *sklearn circles* dataset is shown in Table 3. It shows DBSCAN performs very well and other algorithms fail to cluster this dataset.

Table 3: Performance of clustering algorithms on *sklearn circles* dataset using different evaluating metrics

Evaluating Metrics	KMeans	KMedoids	AGNES	BIRCH	DBSCAN
Silhouette	0.35	0.35	0.35	0.31	0.11
ARI	0.0	0.0	0.01	0.01	1.0
NMI	0.0	0.0	0.0	0.01	1.0

References

[1] https://scikit-learn.org/stable/modules/clustering.html

[2] https://scikit-learn.org/stable/datasets/sample_generators.html

[3] https://realpython.com/k-means-clustering-python/

[4] https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c?fbclid=IwAR19fWzXXF_8t8oy1Oe7o1eCjqIpRsEul3Mu9h3_fKPGDG4_674sB0PecF4

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html