Nurdin Khoirurizka (諾丁)
312540025

Report Clustering Methods

Homework 1

NYCU

This report is present a comprehensive analysis of clustering algorithms applied to a variety of open-source datasets. The objective is to evaluate the performance and applicability of different clustering techniques on datasets that exhibit a range of complexities and structural characteristics. This evaluation is conducted using nine distinct datasets: Complex9, Curves2, Diamond9, Disk-5000n, Gausians1, Hypercube, Impossible, Disk-4000n, and Mopsi-Joensuu, and Sizes5 sourced from the Clustering Benchmark database.

The clustering methods that evaluate selected for this study include Mean Shift, K-means, Agglomerative, DBSCAN, HDBSCAN and OPTICS. These methods were chosen based on methods that we have studied, distinct methodological approaches and their potential for revealing intricate patterns in complex datasets. Each method will be thoroughly tested across all datasets, with a focus on tuning hyper-parameters to optimize performance on clustering pattern.

In the experiment, the dataset was displayed on a 2D graph as shown in Figure 1, arranged in the following order: compex9, curves2, diamond9, disk5000n, gaussians1, hypercube, impossible, disk-4000n, mopsi-joensuu, and sized5. Each dataset exhibits its own unique pattern and presents specific extraction challenges. Consequently, a method that is suitable for one dataset might not necessarily be effective for another. Additionally, the performance of clustering depends on the precise tuning of each hyper-parameter.

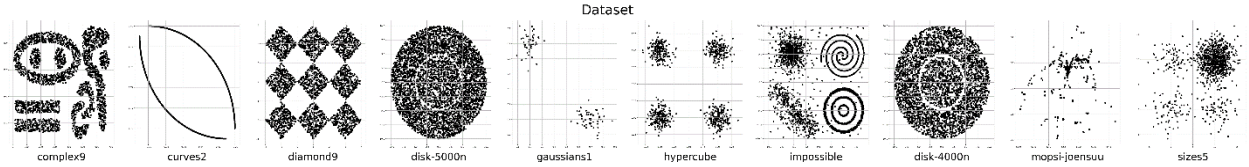

Figure 1. Dataset Collections

The reason of choosing MeanShift Clustering is because this method can discover blop in a smooth density of sample. Because it is centroid based algorithm, which work by updating candidate for centroid to be mean of points within a given region. Then this method will filter the candidates in post-processing stage to estimate near-duplicates from final set of centroid. The hyper parameter of this algorithm is based on bandwith that can automatically base on sized region.

Then, the K-means clustering is chosen, because it has simplicity, speed, and fixed number of clusters. This algorithm that work by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia. This algorithm has hyper-parameter of the number of clusters that should be specified.

Any other clustering algorithm that interested to discuss is Agglomerative Clustering. It is hierarchical clustering using a bottom to up approach. Each observation starts in its own clusters, and clusters are successively merge together. The hyper-parameter that will mainly be discussed on this report is n_clusters and linkages. The n_clusters is total number of clusters that will find. The linkage determines which distance to use between sets of features.

Different with the previous algorithm, the DBSCAN Clustering views clusters as areas of high density separated by areas of low density. Due to this rather generic view, BDSCAN can find

any type of shape of cluster. This algorithm has two parameters there are min_samples and epsilon. Higher main_samples or lower epsilon is mean higher density necessary to form a cluster.

The OPTICS Clustering has similarities with DBSCAN. The key of different of them are that OPTICS build reachability graph, which assigns each sample both reachability and cluster ordering. This algorithm has hyper parameter that was used on this report is min_samples, xi and min_clusters. The min_samples is the number of samples in a neighborhood for a point to be considered as a core point. Then the xi will determine the minimum steepness on the reachability plot that constitutes a cluster boundary. And the samples are minimum number of samples in each cluster.

The last algorithm that will discuss on this report is HDBSCAN. This algorithm is interesting to discuss because it is an extension of DBSCAN and OPTICS. This algorithm alleviates this assumption and explores all possible density scales by building an alternative representation of the clustering problem. This hyper parameter of this algorithm that will discuss is min_cluster_size, min_sample, max_cluster_size and leaf_sized. The min_cluster_size is minimum number of samples in a group. The min_sample is number of samples in a neighborhood for a point to be considered as a core point. The max_cluster_size is the maximum limit to the size of clusters returned. Then, leaf_sized is leaf size for trees responsible for fast nearest neighbour queries.

The hyper parameter that was used on the experiment described on Table 1-6. These hyper parameter is best value that was found on the clustering algorithm based on dataset.

Table 1. Mean Shift Hyper Parameter

No	Dataset	hyper parameter
		bandwidth
1.	Complex9	152.857
2.	Curves2	0.03242
3.	Diamond9	1.7941
4.	Disk-5000n	5.2169
5.	Gausians1	0.0788
6.	Hypercube	1.0193
7.	Impossible	4.3579
8.	Disk-4000n	5.2083
9.	Mopsi-Joensuu	0.2527
10.	Sizes5	4.0281

Table 2. K-mean Hyper Parameter

No	Dataset	hyper parameter
		k
1.	Complex9	10
2.	Curves2	10
3.	Diamond9	10
4.	Disk-5000n	10
5.	Gausians1	10
6.	Hypercube	10
7.	Impossible	10
8.	Disk-4000n	10
9.	Mopsi-Joensuu	10
10.	Sizes5	10

Nurdin Khoirurizka (諾丁)
312540025

Table 3. Agglomerative Hyper Parameter

No	Dataset	hyper parameter	
		n_clusters	linkages
1.	Complex9	10	ward
2.	Curves2	10	ward
3.	Diamond9	10	ward
4.	Disk-5000n	10	ward
5.	Gausians1	10	ward
6.	Hypercube	10	ward
7.	Impossible	10	ward
8.	Disk-4000n	10	ward
9.	Mopsi-Joensuu	10	ward
10.	Sizes5	10	ward

Table 4. DBSCAN Hyper Parameter

No	Dataset	hyper parameter	
		epsilon	min_samples
1.	Complex9	16	10
2.	Curves2	0.0016	10
3.	Diamond9	0.16	10
4.	Disk-5000n	1.6	2
5.	Gausians1	1.6	2
6.	Hypercube	0.16	10
7.	Impossible	1	8
8.	Disk-4000n	1.6	2
9.	Mopsi-Joensuu	0.2	10
10.	Sizes5	0.8	8

Table 5. OPTICS Hyper Parameter

No	Dataset	hyper parameter		
		min_samples	x_i	min_clusters
1.	Complex9	10	0.2	0.1
2.	Curves2	10	0.05	0.015
3.	Diamond9	10	0.015	0.1
4.	Disk-5000n	10	0.02	0.1
5.	Gausians1	10	0.15	0.1
6.	Hypercube	10	0.2	0.1
7.	Impossible	10	0.2	0.1
8.	Disk-4000n	10	0.025	0.1
9.	Mopsi-Joensuu	10	0.05	0.01
10.	Sizes5	10	0.05	0.015

Table 6. HDBSCAN Hyper Parameter

No	Dataset	hyper parameter				
		min_cluster_size	min_samples	algorithm	cluster_selection_method	
					(eom) max_cluster_size	(leaf) leaf_size
1.	Complex9	8	40	balltree	2272	-
2.	Curves2	2	40	kdtree	2846	-
3.	Diamond9	8	40	balltree	2846	-
4.	Disk-5000n	7	200	balltree	None	-

Nurdin Khoirurizka (諾丁)
312540025

No	Dataset	hyper parameter				
		min_cluster_size	min_samples	algorithm	cluster_selection_method	
					(eom) max_cluster_size	(leaf) leaf_size
5.	Gausians1	7	100	brute	None	-
6.	Hypercube	4	200	balltree	-	40
7.	Impossible	7	80	brute	-	100
8.	Disk-4000n	7	140	brute	None	-
9.	Mopsi-Joensuu	4	200	brute	-	80
10.	Sizes5	4	200	balltree	None	-

After the testing all methods and tuning the hyper-parameter, It's got the results of the clustering that is shown on Figure 2. On this results, it shows that each result has their own special pattern and different from each other. So this shows that each method will be useful according to the clustering problem to be solved.

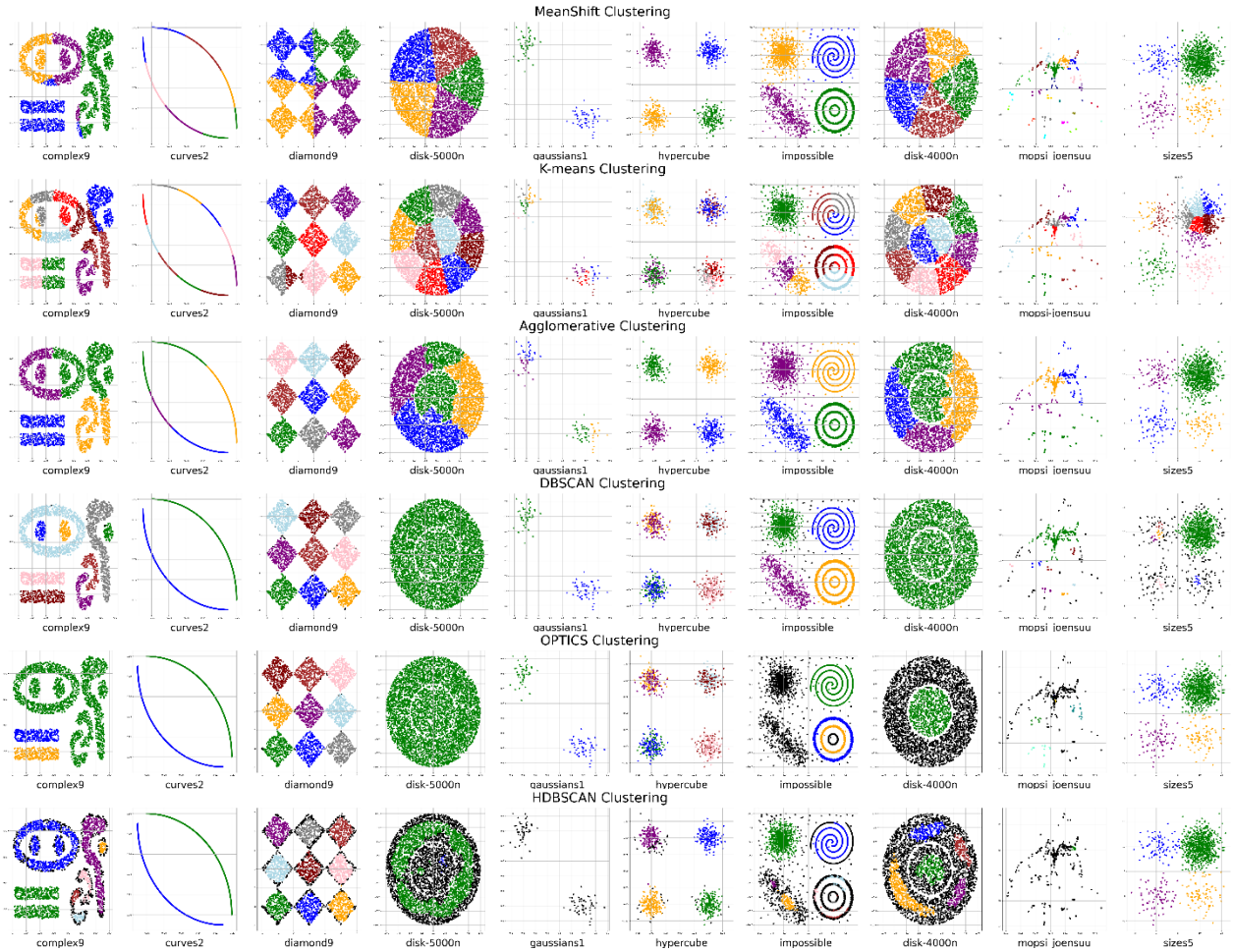

Figure 2 Results of Clustering Methods

Nurdin Khoirurizka (諾丁)
312540025

The result of internal evaluation for this experiment is using to Dunn's Index, Davis Bouldin Index, Xi-Beni Indexm and Silhouette Index. Larger Dunn's Index is mean better performance. This variable has range more than zero. On Davis Bouldin Index, it's also range more than zero, but less value will be better. The Xi-Beni Indexm is also range more than zero, but less value will be better. Than the Silhouette Index has range value from -1 to 1 with higher value is mean better performance.

The comparison result of each methods and dataset is shown on Table 7 until 12. This result has notes that * is mean larger is better and ** is mean smaller is better. The empty variable is mean that the result has not parameter to evaluate using that index variable.

Table 7. Mean Shift External Evaluation

No	Dataset	Dunn's Index (*)	Davis Bouldin Index (**)	Xi-Beni Index (**)	Silhouette Index (*)
1.	Complex9	0.5694	0.8617	0.3836	0.398
2.	Curves2	0.4537	0.636	0.2249	0.4756
3.	Diamond9	0.6879	0.808	0.5358	0.3756
4.	Disk-5000n	0.5942	0.8408	0.3861	0.3526
5.	Gausians1	0.095	0.1911	0.095	0.8613
6.	Hypercube	0.2326	0.4646	0.1459	0.7093
7.	Impossible	0.2573	0.5424	0.2121	0.6102
8.	Disk-4000n	0.6338	0.8502	0.4018	0.3461
9.	Mopsi-Joensuu	0.1607	0.3481	0.025	0.8430
10.	Sizes5	0.27	0.543	0.2303	0.5747

Table 8. K-mean External Evaluation

No	Dataset	Dunn's Index (*)	Davis Bouldin Index (**)	Xi-Beni Index (**)	Silhouette Index (*)
1.	Complex9	0.5361	0.8521	0.1956	0.4114
2.	Curves2	0.256	0.5045	0.1126	0.5387
3.	Diamond9	0.6303	0.6641	0.1649	0.5102
4.	Disk-5000n	0.4707	0.8093	0.2232	0.3642
5.	Gausians1	0.4644	0.8308	0.0777	0.3529
6.	Hypercube	0.7439	0.6333	0.041	0.7
7.	Impossible	0.4987	0.8458	0.163	0.4401
8.	Disk-4000n	0.5173	0.8167	0.222	0.3668
9.	Mopsi-Joensuu	0.5694	0.6599	0.0412	0.6457
10.	Sizes5	0.610	0.91	0.1755	0.3377

Table 9. Agglomerative External Evaluation

No	Dataset	Davis Bouldin Index (**)	Silhouette Index (*)
1.	Complex9	0.9087	0.399
2.	Curves2	0.7327	0.4421
3.	Diamond9	0.5543	0.5479
4.	Disk-5000n	1.1016	0.3104
5.	Gausians1	0.9766	0.3767
6.	Hypercube	0.4646	0.7093
7.	Impossible	0.5427	0.6097
8.	Disk-4000n	0.9251	0.3125

No	Dataset	Davis Bouldin Index (**)	Silhouette Index (*)
9.	Mopsi-Joensuu	0.5643	0.8225
10.	Sizes5	0.4966	0.5787

Table 10. DBSCAN External Evaluation

No	Dataset	Davis Bouldin Index (**)	Silhouette Index (*)
1.	Complex9	2.4651	-0.024
2.	Curves2	1.475	0.3475
3.	Diamond9	2.3378	0.5396
4.	Disk-5000n	-	-
5.	Gausians1	0.1911	0.8613
6.	Hypercube	0.2573	0.7932
7.	Impossible	1.6993	0.5914
8.	Disk-4000n	-	-
9.	Mopsi-Joensuu	2.0153	0.6088
10.	Sizes5	1.2599	0.3009

Table 11. OPTICS External Evaluation

No	Dataset	Davis Bouldin Index (**)	Silhouette Index (*)
1.	Complex9	1.8631	0.158
2.	Curves2	1.475	0.3475
3.	Diamond9	1.7874	0.5433
4.	Disk-5000n	-	-
5.	Gausians1	0.1911	0.8613
6.	Hypercube	0.2474	0.8276
7.	Impossible	46.7511	0.2019
8.	Disk-4000n	126.0775	-0.064
9.	Mopsi-Joensuu	2.1035	-0.2025
10.	Sizes5	3.0334	-0.2806

Table 12. HDBSCAN External Evaluation

No	Dataset	Davis Bouldin Index	Silhouette Index
1.	Complex9	1.6914	0.0513
2.	Curves2	1.475	0.3475
3.	Diamond9	3.9944	0.3156
4.	Disk-5000n	-	-
5.	Gausians1	-	-
6.	Hypercube	1.8869	0.5835
7.	Impossible	1.8109	0.0327
8.	Disk-4000n	53.0505	-0.1926
9.	Mopsi-Joensuu	3.4	-0.2869
10.	Sizes5	-	-

Than the external evaluation, the best algorithm with objective to extract pattern based on shape or pattern is BDSCAN. This algorithm also, have capabilities to filter the noise on shape pattern. Whereas the algorithm that can extract pattern completely and symmetrically is K-mean and mean shift.

Nurdin Khoirurizka (諾丁)
312540025

Based on the result of the experiment and my focus on control system development, I choose the implementation of the clustering algorithm to make a clustering of electric motor temperature. The propose is to identification of cluster that future more can implement on cooling electric motor based on the data. On this case, the algorithm that will be used on this implementation is Mean Shift and k-mean because the performance on previous experiment.

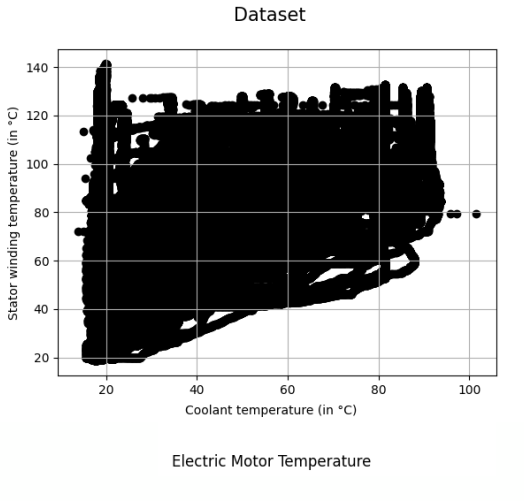

Figure 3 Dataset Electric Motor Temperature

After implementation and tuning hyper parameter the result is shown on Figure 4 for Mean Shift and Figure 5. For K-mean. Meanwhile the external evaluation of the both method is on Table 13.

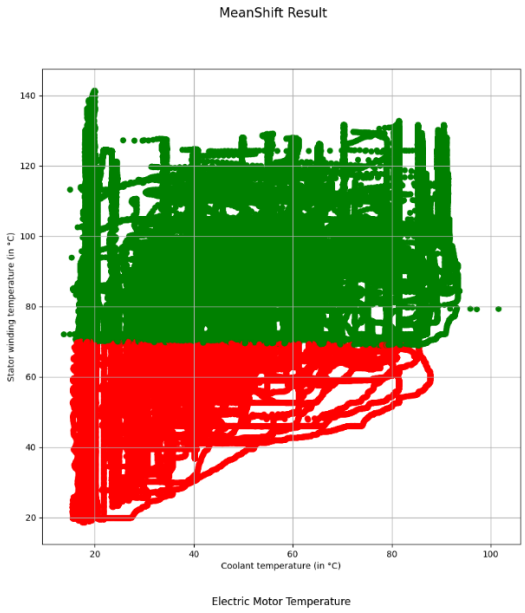

Figure 4 Results Mean Shift

Nurdin Khoirurizka (諾丁)
312540025

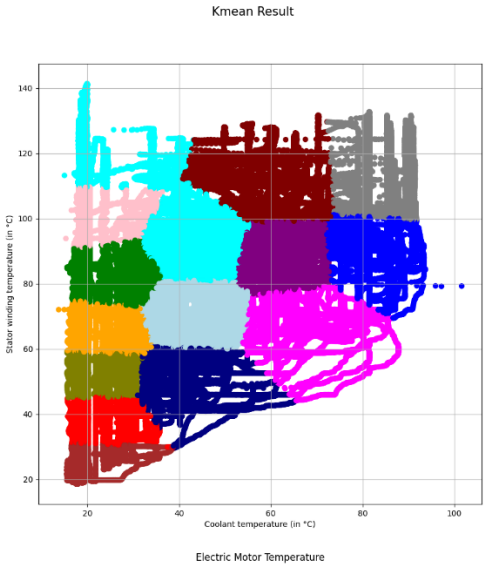

Figure 5 Results K-mean

Table 13. External Evaluation Comparison

No	Dataset	Davis Bouldin Index (**)
1.	Mean Shift	0.802
2.	K mean	0.7373

From the result is shown that k mean can make better cluster base on the image and also from Davis bounding index. This implementation just evaluates by Davis Bouldin Index because the other evaluation required larger computation.