國立雲林科技大學資訊管理系

機器學習-作業三

Department of Information Management

National Yunlin University of Science & Technology

Assignment

Banana 資料集和 Size3 資料集分群演算法分析

Banana Datasets and Size3 Dataset Cluster Analysis

楊欣蓓、陳怡君、鄭皓名、陳郁云

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

中華民國112年12月

December 2023

摘要

本研究對 Banana 資料集和 Size3資料集進行分群分析，由於這兩種資料集的資料分佈密度及資料群聚的狀態不同，因此透過觀察兩種資料集在 K-means、階層式分群 (Hierarchical Clustering)、DBSCAN 三種分群演算法的分群結果，來研究出在不同型態的資料集的情況下，較適合使用何者分群演算法。此外，本研究欲找出 DBSCAN 分群演算法的最佳超參數組合。在執行分群演算法前，對資料刪除離群值以及做正規化，使分群結果不受前兩因素影響，另外透過視覺化圖形，如：Dendrogram、Scatter 可以更清楚地檢視資料分佈狀況以及了解分群資訊，最後以 SSE (Sum of Squared Error)、Accuracy、Entropy 三種績效評估指標來衡量分群結果。實驗結果於 Banana 資料集和 Size3資料集中，DBSCAN 皆為表現最佳，其最佳超參數組合在 Banana 資料集中為 eps=0.018、n=16，Accuracy 為0.969；在 Size3資料集中為 eps=0.084、n=12， Accuracy 為0.957。

關鍵字：分群演算法、K-means、階層式分群 (Hierarchical Clustering)、DBSCAN

一、緒論

1.1 研究動機

本研究主要是在眾多的群聚分析中進行分析，因其資料集的資料雜亂，所以需要資料前處理才能進行分群分析，而選擇 Banana、Size3的主要原因是各個資料集的資料分佈不同，使得本研究可以使用不同的分群演算法（例如:K-means、Hierarchical Clustering、DBSCAN）來進行實驗，通過比較這些分群演算法表現，可以評估它們在處理不同形狀和密度的分群時的績效。

本研究發現 Banana、Size3的二維結構使得群集可視化相對容易，並且可以使用不同的分群演算法將數據點劃分到群集中，以及可視化這些群集的分佈，這有助於理解不同算法的群集結果，由於這些資料集包含不同形狀和密度的群集，所以可以嘗試調整不同群集算法的參數，以查看參數對結果的影響如何。如果這個資料集與某個實際應用相關，例如顧客購買行為、產品銷售模式等，使用 K-means、Hierarchical Clustering、DBSCAN 可以幫助本研究識別和理解樣本之間的相似性和差異性，並且為業務決策提供有價值的資訊。總上所述，本研究進行 K-means、Hierarchical Clustering、DBSCAN 分群演算法是為了從數據中挖掘有效資訊，理解潛在的結構和模式，並且比較不同算法在這個情境下的效能表現。

1.2 研究目的

本研究欲以 Banana 資料集和 Size3 資料集以各種分群來進行比較 K-means、Hierarchical Clustering、DBSCAN 三種分群所花費的時間，將此分群以 SSE、Accuracy、 Entropy 為此衡量指標，並劃分出這三種分群所呈現的結果，其中本研究想以 DBSCAN 分群來進行不同的參數設定，此研究透過各種參數設定來進行比較後， 藉由比較後的參數設定上分析出以 Banana 資料集和 Size3 資料集中的試驗以達成 Banana 資料集和 Size3 資料集的最佳設定值，也了解每種分群在不同情況下的優缺點及此限制，在碰到資料集中存在雜訊時，該如何使用各種分群 來解決資料集的問題，並透過 K-means、Hierarchical Clustering、DBSCAN 分群來分析資料 集的 SSE、Accuracy、Entropy 的結果。

二、實驗方法

2.1 實作說明

本研究使用 K-means、Hierarchical Clustering、DBSCAN 三種分群演算法對 Banana、Size3進行分群實驗。在資料前處理的部分，對資料刪除了離群值，使分群結果不被離群值影響、對資料作正規化，使不同屬性的數值差距不會甚多；調整了分群演算法的參數，利用 SSE、Accuracy、Entropy 的結果去調整參數，使分群結果更加完確。此外本次實驗為分群分析，並透過 Dendrogram 和 Scatter 的資料視覺化，去觀察不同演算法對資料的分群效果，進而去比較各個演算法的分群效果為何，最後記錄各個演算法的執行時長及分群結果並做出研究分析。

2.2 操作說明

本研究執行環境採用 Python3.10.10，以 Visual Studio Code 作為開發工具，利用 K-means、Hierarchical Clustering、DBSCAN 三種演算法進行分群實驗，並使用 Pandas、Numpy、Scikit-learn、Matplotlib 等函式庫來讀取資料、分析分群結果及資料視覺化呈現。於資料前處理，利用 skewness 和 kurtosis 去檢視資料離群值並刪除離群值、用 MinMaxScaler 將資料數值之間的差距縮至0至1之間，上述對資料的操作可以使資料在做分群時分得更完整、提高分群績效。

三、實驗設計

3.1 資料集

名稱：Banana 資料集

資料筆數：4811筆

表 1

Banana datasets 欄位介紹

欄位	屬性	內容
0	x	numeric
1	y	numeric
2	class	nominal

名稱：Size3 資料集

原始資料筆數：1000筆

刪除離群值後的資料筆數：820筆

表 2

Size3 datasets 欄位介紹

欄位	屬性	內容
0	x	numeric
1	y	numeric
2	class	nominal

3.2 資料前處理

3.2.1 Banana 資料集

- 分析資料集：
  - 透過將 Banana 資料集繪製成常態分布圖，如下圖1所示，觀察資料分布的偏度 (Skewness)與峰度 (Kurtosis) 是否呈現常態分佈，以檢測資料集是否有離群值。
  - 由於 Banana 資料集並非呈現常態分布，因此本研究利用四分位距法 (IQR[1]) 檢測資料集中是否有離群值。經四分位距法檢測後，發現並無未在區間外的資料，故 Banana 資料集不須再刪除離群值。
  - 由於 Banana 資料集轉成散佈圖後，可發現所以資料點的數值皆位在0到1之間，故無需做正規化。

3.2.2 Size3資料集

- 資料前處理：
  - 透過 IQR 方法檢查資料前後筆數及 IQR 箱型圖確認資料是否存在離群值，並一一將離群值做刪除，使資料分群結果不受離群值影響，刪除離群值後的結果如下表 3 所示。
  - 從資料散佈圖看出若干筆資料數值未落在 0 至 1 之間，為了不讓數值之間的差距影響分群結果，使用正規化技術 (MinMaxScaler) 將資料數值落在 0 到 1 之間。

表 3

部分經資料處理後的 Size3 資料集

資料 / 特徵	No.0	No.1	No.2	No.3	No.4	No.5
X	0.60260	0.78808	0.56988	0.59167	0.78194	0.70639
Y	0.87058	0.48759	0.68918	0.76108	0.50162	0.62674

[1] 四分位距法 (IQR)：將資料做排序後，取第三個四分位數 (Q3) 減去第一個四分位數 (Q1) 可求得 IQR。當有資料數值落在 Q1−1.5×IQR 至 Q3+1.5×IQR 此區間外，則是為該筆資料為離群值，應做刪除。

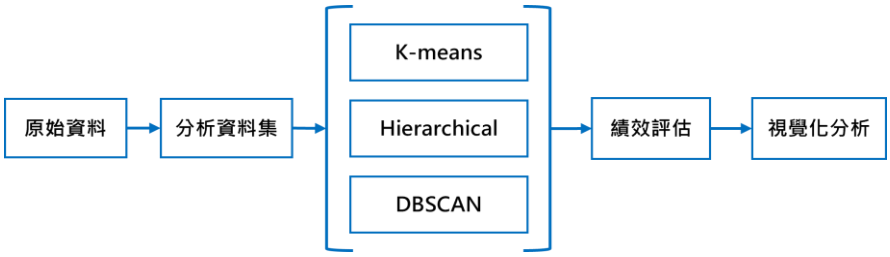
3.3 實驗設計

3.3.1 Banana 資料集

Banana 資料集的實驗設計如圖1所示。在資料進行分群前，首先觀察 Banana 資料集的資料分佈與檢測有無離群值，本研究透過 IQR 方法進行離群值檢測。接著，使用 K-means、Hierarchical Clustering、DBSCAN 分群演算法進行資料分群，以 SSE 、Accuracy、Entropy 三種評估指標來衡量分群績效。最後，透過將分群後的資料以視覺化圖形來觀察分群結果，其中，包含以 Dendrogram 和散佈圖呈現各個演算法分群結果。從該結果來決定是否再訓練及超參數的調整。

圖 1

Banana 資料集實驗設計流程圖

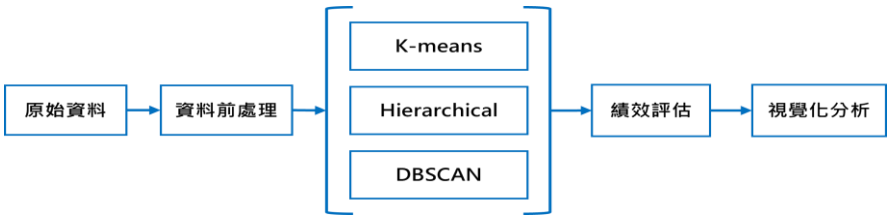


3.3.2 Size3資料集

Size3資料集實驗設計如圖2所示。於資料前處理，本研究利用 IQR 方法發現資料集有離群值，因此將其刪除後做正規化 (MinMaxScaler)，將數值範圍落在0至1之間。分析資料時，透過 K-means、Hierarchical、DBSCAN 三種演算法分別做分群，並依 SSE、Accuracy、Entropy 三種績效指標調整參數。本實驗將Size3資料集分成四群，利用散布圖將分群結果可視化，其四群標籤分別為「1」、「2」、「3」、「4」，此外，也利用樹狀圖比較不同度量距離的參數，用來評估各個分群結果。

圖 2

Size3實驗設計流程圖



3.4 實驗結果

本研究實驗結果如下列兩資料集實驗結果所述，分為兩個部分。第一部分主要討論三種分群演算法 (K-means、Hierarchical Clustering、DBSCAN) 分群所耗費的時間及透過三種衡量指標 (SSE、Accuracy、Entropy) 評估分群結果；第二部分主要比較 DBSCAN 分群演算法在不同超參數組合下的分群準確度，以找到準確度最高的超參數組合。

3.4.1 Banana 資料集實驗結果

首先對 Banana 資料集使用三種分群演算法 (K-means、Hierarchical Clustering、DBSCAN) 進行分群，分群結果如圖3，以三種衡量指標 (SSE、Accuracy、Entropy) 評估分群結果與分群演算法所耗費的時間，如表4。其中，分群資料散佈圖將兩群的標籤分別標為「+」及「O」以清楚地觀察分群的結果。

K-means 的超參數設定為 n_clusters 為2、n_init 為 auto、更新次數的上限值 (max_iter) 採演算法預設數值、每個 Cluster 的中心點收斂容忍度 (tol)，預設為 0.0001。Hierarchical Clustering 使用 AgglomerativeClustering 方式做分群，階層圖如圖4所示，超參數設定為 n_clusters 為2、metric 為 euclidean、計算群間資料點的距離 linkage 為 average，其中，階層圖以 Dendrogram 呈現，可觀察群內兩兩資料之間合併的結果和群與群之間合併的過程，藉由上述去分析在不同度量距離下，分群結果會產生多少變化，進而去挑選出最適當的分群方式。

DBSCAN 的超參數設定以第二部分比較 DBSCAN 分群演算法在不同超參數組合下的分群準確度之實驗，找到分群準確度最高的超參數組合做 DBSCAN 演算法的分群實驗，組合為 eps 為0.01769180601295415及 min_samples 為16。

比較三種分群演算法的結果，可以發現，以 DBSCAN 作為分群演算法在 Accuracy 與 Entropy 兩種評估指標下，表現皆為最優異，因此，可推論 DBSCAN 演算法在針對似月形的資料做分群時，可以做到較好的分群效果，準確度高達 0.97、Entropy 為0.02、執行時間為0.08秒

表4

各演算法績效表現

模型	SSE	Accuracy	Entropy	Run Time(s)
K-means	185.21	0.83	0.66	**0.01**
Hierarchical Clustering	—	0.82	0.58	0.30
DBSCAN	—	**0.97**	**0.02**	0.08

圖 3

Banana 資料集分群散佈圖-K-means 、Hierarchical Clustering 、DBSCAN

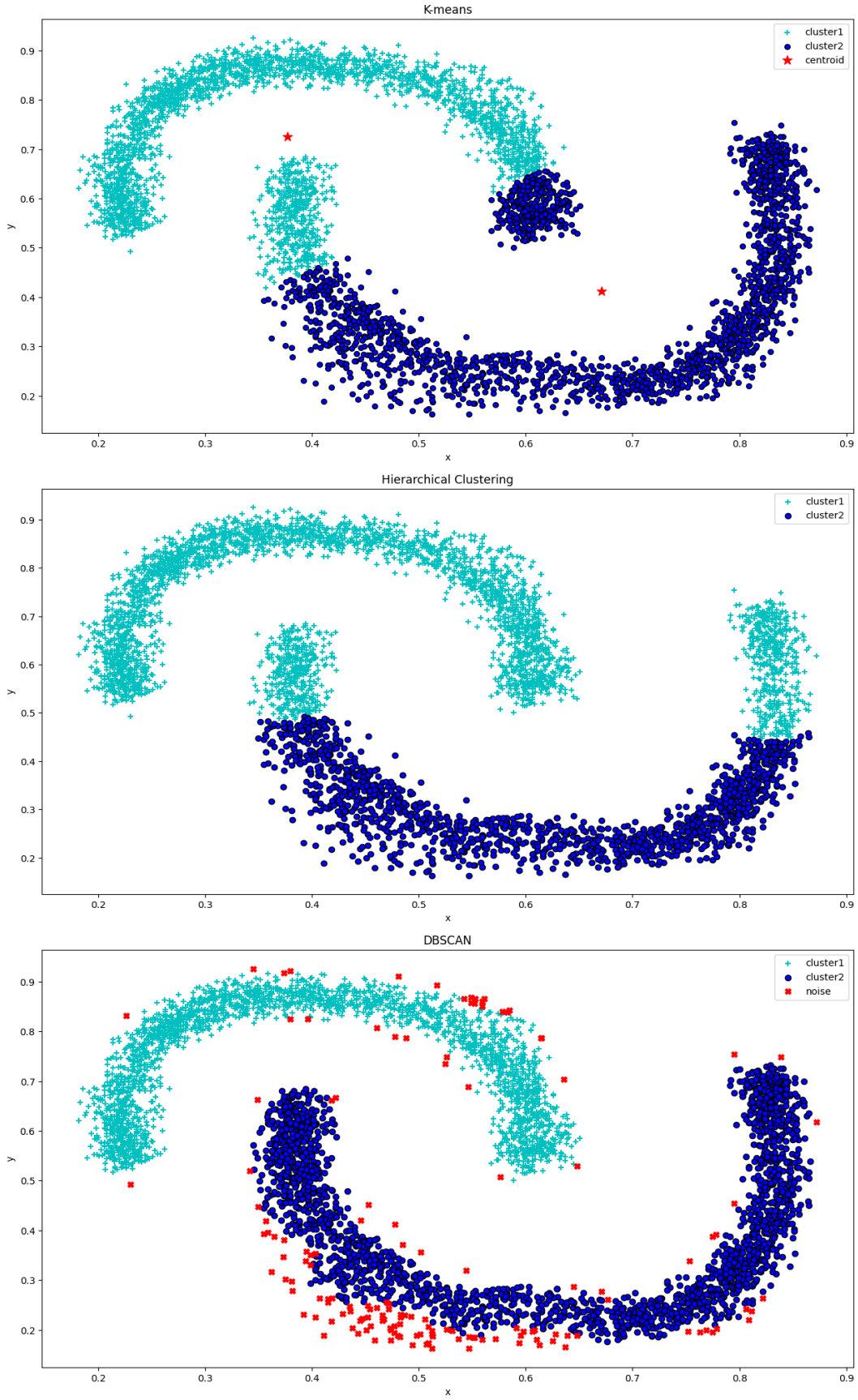
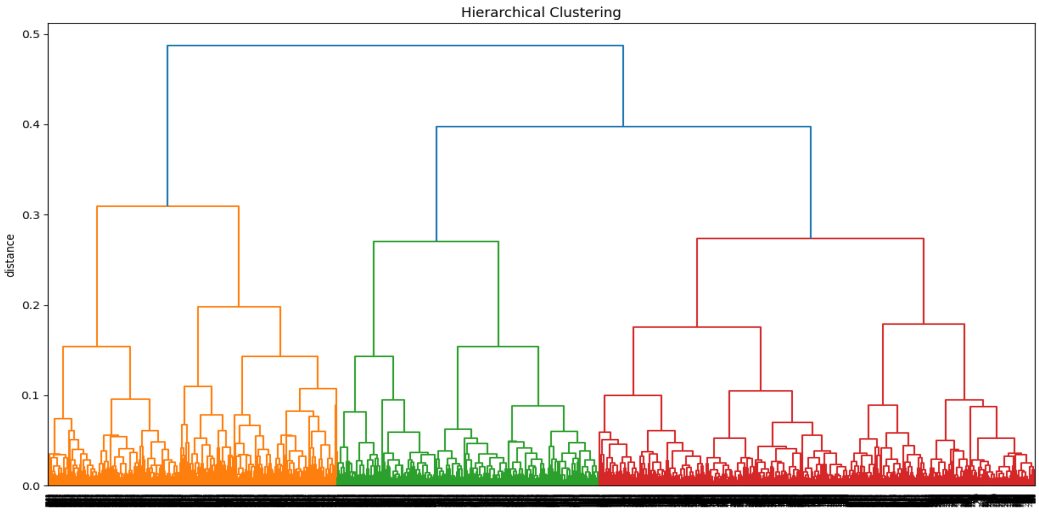


圖 4

Hierarchical Clustering 階層圖



接著，針對不同超參數組合的 DBSCAN 分群演算法對 Banana 資料集分群的結果作分析與觀察其準確度的變化，如下表5所示。圖5為第二部分實驗：找出 DBSCAN 中最佳超參數組合，以最近鄰居演算法找出最佳的 Eps。

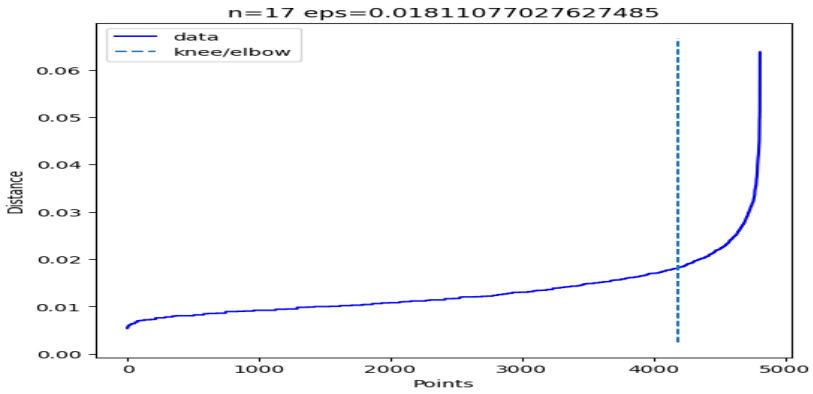
表 5

DBSCAN 演算法在 Banana 資料集下各參數組合之 Accuracy

組合	eps	min_samples	Accuracy
1	**0.018**	**16**	**0.969**
2	0.018	17	0.968
3	0.016	13	0.964
4	0.017	15	0.964
5	0.016	14	0.964

圖 5

搜尋 DBSCAN 的最佳 Eps



3.4.2 Size3資料集實驗結果

本實驗針對Size3資料集使用K-means、Hierarchical Clustering、DBSCAN三種演算法做分群，群數設定為四群，如圖6所示。其中K-means的超參數設定為n_clusters為4、random_state為1、n_init為k-means++、每個Cluster的中心點收斂容忍度(tol)，預設為0.0001。

Hierarchical Clustering使用AgglomerativeClustering方式做分群，階層圖如上圖5所示，超參數設定為n_clusters為4、metric為euclidean、計算群間資料點的距離linkage為average。

DBSCAN的超參數設定以第二部分比較DBSCAN分群演算法在不同超參數組合下的分群準確度之實驗，找到分群準確度最高的超參數組合做DBSCAN演算法的分群實驗，組合為eps為0.07741470944658732及min_samples為10。

比較三種分群演算法的結果，如下表6所示，可以發現，以DBSCAN作為分群演算法在Accuracy評估指標下，表現最優異，而Entropy評估指標則是Hierarchical Clustering與DBSCAN並列。在實驗過程中，發現未做資料前處理時，DBSCAN分群結果與預期分群數不一致績效也不盡理想，而在完成資料前處理後，DBSCAN才能正常分為四群，由上述結果，可得知DBSCAN演算法在針對Size3資料集做分群時，可以做到較好的分群效果，準確度高達0.96、Entropy為0.17、執行時間為0.01秒。

表 6

各演算法績效表現

模型	SSE	Accuracy	Entropy	Run Time(s)
K-means	20.63	0.35	0.22	0.08
Hierarchical Clustering	—	0.89	**0.17**	0.02
DBSCAN	—	**0.96**	**0.17**	**0.01**

圖 6

Size3 資料集分群散佈圖-K-means 、Hierarchical Clustering 、DBSCAN

接著，針對不同超參數組合的 DBSCAN 分群演算法對 Size3資料集的分群結果作分析與觀察其準確度的變化，如下表7所示。圖7中的 DBSCAN 為本研究找出的最佳超參數組合，以最近鄰居演算法找出最佳的 Eps。

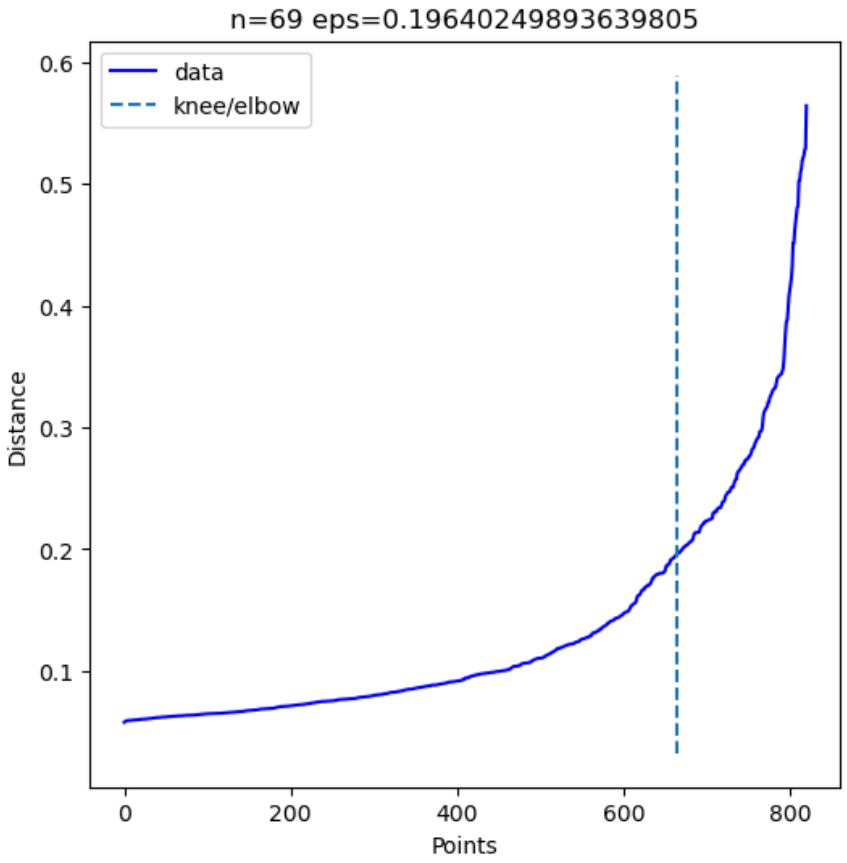
表 7

DBSCAN 演算法在 Size3 資料集下各參數組合之 Accuracy

組合	eps	min_samples	Accuracy
1	**0.084**	**12**	**0.957**
2	0.095	15	0.955
3	0.099	16	0.955
4	0.099	17	0.955
5	0.081	11	0.954

圖 7

搜尋 DBSCAN 的最佳 Eps



結論

本研究主要為兩個部分。第一項實驗中，針對 Banana 與 Size3兩資料集討論三種分群演算法 (K-means、Hierarchical Clustering、DBSCAN) 分群結果。實驗過程，先透過 IQR 找出離群值並刪除後，透過 SSE、Accuracy、Entropy 三項指標做評估並做為調整超參數之依據。在 Banana 資料集中，DBSCAN 在 Accuracy 與 Entropy 兩項指標皆最佳，其 Accuracy 達到0.97、Entropy 為0.02、執行總時長為0.08；另外，在 Size3資料集中，DBSCAN 各項指標也最為亮眼，其 Accuracy 達到0.96；Entropy 為0.17，與 Hierarchical Clustering 相同；執行總時長為0.01。

第二項實驗，主要比較 DBSCAN 分群演算法在不同超參數組合下的分群準確度。實驗結果發現，Banana 資料集在 eps=0.018且 n=16的超參數組合下，Accuracy 評估指標為0.969，表現最優異；Size3資料集在 eps=0.084且 n=12的操參數組合下，Accuracy 評估指標為0.957，表現最佳。

參考文獻

Jason (2023)。【學習筆記】K-means 實作篇。
https://medium.com/@jason8410271027/%E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98-k-means%E5%AF%A6%E4%BD%9C%E7%AF%87-5c3fb9faf17

Pylnvest (2020)。層次聚類 Hierarchical Clustering。
https://pyecontech.com/2020/06/15/python_hierarchical_clustering/

Pylnvest (2020)。密度聚類 DBSCAN。
https://pyecontech.com/2020/07/17/python_dbscan/