國立雲林科技大學

資訊管理系

資料探勘

作業三

群聚分析實作

D10630002 陳冠臻

M11023006 謝媄衣

M11123044 沈俊良

M11123061 王成綱

指導教授:許中川

２０２２　年　１２　月　１５　日

摘要

本研究採用 Online Shoppers Purchasing Intention 資料集測試 Purity 及 time 後，由於 Online Shoppers Purchasing Intention 資料集之 Purity 較低，Purity 為 0.786 之純度，time 為 2.03 秒;接著透過 Iris 資料集進行驗證，使用 MS SQL Server 進行資料清洗、正規化資料，透過 Python 訓練演算法(K-means、DBSCAN、Hierarchical Clustering)、迴圈找出兩資料集之最佳純度及分群數，並畫出各分群圖及階層圖，並比較兩資料集在各演算法間哪項最佳，最後資料集純度(Purity)為 0.97。

關鍵字: MS SQL Server、K-means、DBSCAN、Hierarchical Clustering、Purity

一、 緒論

1.1 動機

受到 COVID-19 的影響，全球宅經濟快速發展，根據 eMarketer 預估，到 2024 年，全球電子商務銷售額將來到近 6.4 兆美元（約 177 兆台幣），佔整體零售額的 5 分之 1。以美國為例，在疫情開始後幾周，電商銷售在短短 3 個月內達到過去 10 年的電子商務量，並在全球零售總額占比 16.4%，也是歷史新高，而且年紀較長的消費者也投入網購的世界，更帶動驚人的成長率。同時也是由於疫情的關係，消費者的習慣發生改，更加速了原本就在進行的企業數位化，而民眾的辦公、生活、飲食及消費習慣也已逐漸改變，電子商務儼然成為疫情中最大的受益者。一場疫情而改變了我們習以為常的生活，同時也造就新一波電商革命與龐大的商機；當我們開始習慣透過網際網路來購物以及滿足生活中的大小事時，第一個接觸到的就是各個購物網站的頁面，而在當消費者瀏覽各電商平台網頁時，就可藉由數位技術將消費者在網頁上進行所有瀏覽行為以及購物動作記錄下來，並轉化為行為資料，而對於相關店家或平台業者就可以依據此數據資料，更近一步改善消費者對網頁的使用習慣以及可分析網頁頁面所影響消費者的購買能力，依據數據分析來做更客觀且細緻的評估與調整，使這種方式所獲得的資料更貼近於消費者真實的購物經驗與需求。本研究挑選一份線上消費者購買意向數據集，進行消費行為的資料分析，並且使用三種不同的演算法，來了解相同數據及對於不同演算法的過程進行比較分析，進而發現演算法的各自特色，以及可以協助我們發現資料的不同面向。

1.2 目的

傳統的行銷方式早已體認到網際網路的重要性，但仍只是將網站視為是一個提供給消費者認識的管道，但卻忽略了網站能帶來商機的潛力。但疫情的到來無形中創造了一個良好的契機。身處於科技時代即便早已設立網站的電商平台或是公司行號，似乎沒有善加利用從網站中所帶來的相關資料，經分析後可獲取商業上更好的報酬，而其中最重要的因素應該是未善加利用數位資料轉換成有用的資訊來協助相關單位進行決策判斷。因此分析線上購物平台的使用習慣就是一個改變過往對於網站單一面向的認識，而採用不同資料探勘的相關工具，可以提供更多我們挑選一個更合適的探勘方式，來協助提供更好的數據品質與數據內容。

二、真實資料集

2.1資料集

表 1

Online Shoppers Purchasing Intention Dataset Data Set 資料屬性一覽

欄位名稱	欄位意義	型態
Administrative	管理	Integer
AdministrativeDuration	管理持續時間	Integer
Informational	資訊	Integer
InformationalDuration	資訊持續時間	Integer
ProductRelated	產品相關	Integer
ProductRelatedDuration	產品相關持續時間	Float
BounceRates	跳出率	Float
ExitRates	退出率	Float
PageValues	頁面值	Float
SpecialDay	特殊日	Float
Month	月份	Integer
OperatingSystems	作業系統	Integer
Browser	瀏覽器	Integer
Region	地區	Integer
TrafficType	流量類型	Integer
VisitorType	訪客類型	String
Weekend	是否為周末	Boolean
Revenue	是否有購物	Boolean
樣本數		屬性個數
12330		18

表 2

Iris Data Set 資料屬性一覽

欄位名稱	欄位意義	型態
Sepal Length	萼片長度	float
Sepal Width	萼片寬度	float
Petal Length	花瓣長度	float
Petal Width	花瓣寬度	float
Name	鳶尾花名稱	String
樣本數		屬性個數
150		5

三、方法

3.1 實作說明

將 Online Shoppers Purchasing Intention Dataset Data Set 與 Iris Data Set 使用 Microsoft SQL Server Management Studio 匯入至 Microsoft SQL Server 資料表中，透過 T-SQL 語法進行正規化處理，並匯出成 csv 檔。

使用 Python 作為開發程式語言，透過 K-means、階層式分群及 DBSCAN 演算法套件，對兩資料集進行群聚分析，並計算純度 (Purity) 與執行時間分析各演算法之最佳參數值，再依個別演算法之品質衡量指標比較分群結果。

3.2 操作說明

對 K-means 演算法程式之幾何中心(centroids)數量進行 3 至 10 設置，並記錄各組幾何中心之執行時間與純度。

對階層式分群演算法程式之四種計算方式(ward、single、complete、average)之分群數量進行 3 至 10 之設置，並記錄各計算方式執行時間與純度。

對 DBSCAN 演算法程式分別對半徑(EPS)與最小樣本數(min_samples)進行 0.3 至 1.0 與 10 至 50 之設置，並記錄執行時間與純度。

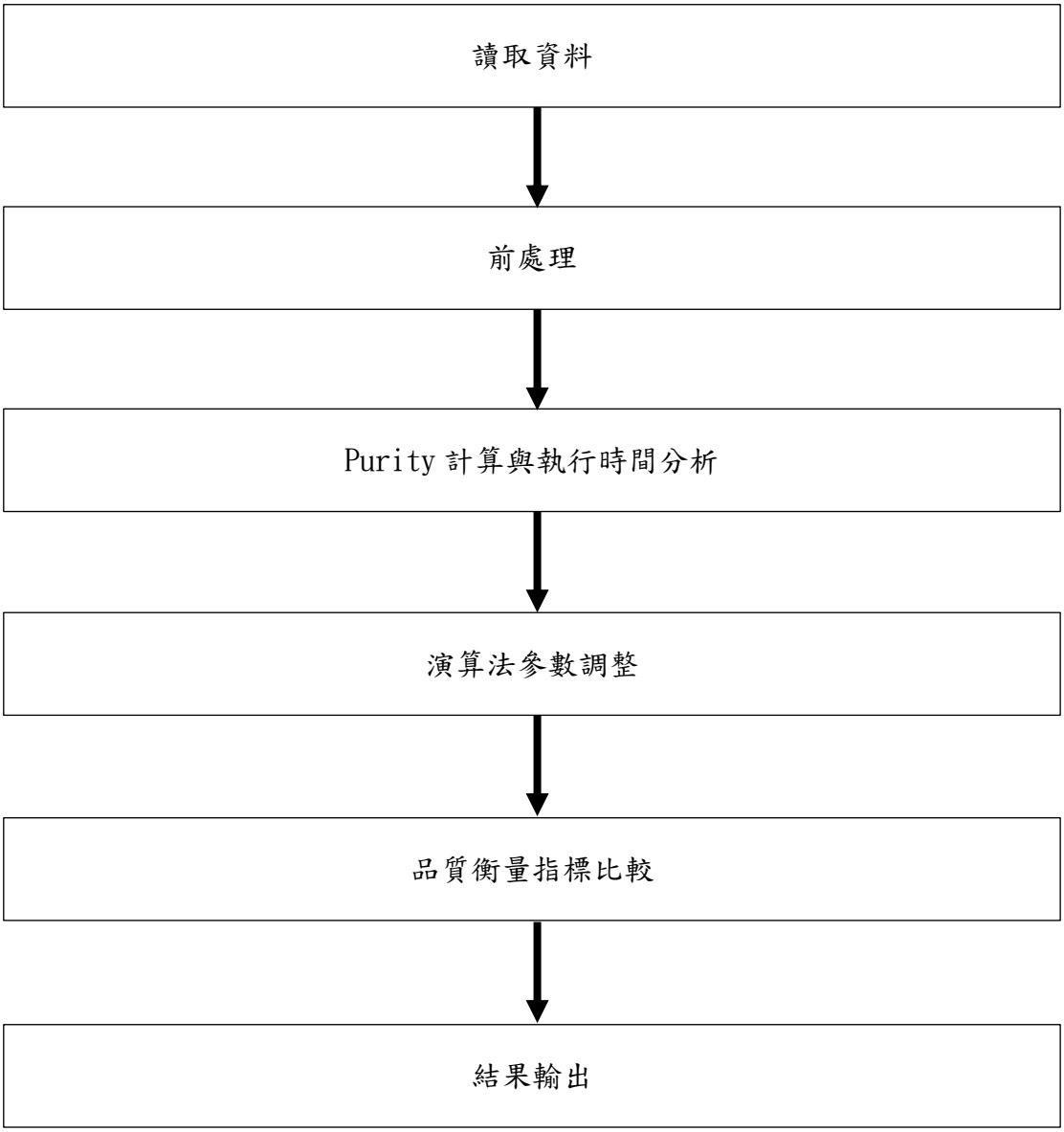

圖 1 資料集分群流程圖

四、實驗

4.1 前置處理

對於 Online Shoppers Purchasing Intention Dataset Data Set 之名目尺度特徵屬性進行 One Hot Encoding，以及將布林(Boolean)資料值之 FALSE 與 TRUE 轉換為 0 與 1。

Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	0	0	0	1	0	0.2	0.2	0	0	2	1	1	1	1	Returning_Visitor	FALSE	FALSE
0	0	0	0	2	64	0	0.1	0	0	2	2	2	1	2	Returning_Visitor	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0	2	4	1	9	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	2	2.666666667	0.05	0.14	0	0	2	3	2	2	4	Returning_Visitor	FALSE	FALSE
0	0	0	0	10	627.5	0.02	0.05	0	0	2	3	3	1	4	Returning_Visitor	TRUE	FALSE
0	0	0	0	19	154.2166667	0.015789474	0.024561404	0	0	2	2	2	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0.4	2	2	4	3	3	Returning_Visitor	FALSE	FALSE
1	0	0	0	0	0	0.2	0.2	0	0	2	1	2	1	5	Returning_Visitor	TRUE	FALSE
0	0	0	0	2	37	0	0.1	0	0.8	2	2	2	2	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	3	738	0	0.022222222	0	0.4	2	2	4	1	2	Returning_Visitor	FALSE	FALSE
0	0	0	0	3	395	0	0.066666667	0	0	2	1	1	3	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	16	407.75	0.01875	0.025833333	0	0.4	2	1	1	4	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	7	280.5	0	0.028571429	0	0	2	1	1	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	6	98	0	0.066666667	0	0	2	2	5	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	2	68	0	0.1	0	0	2	3	2	3	3	Returning_Visitor	FALSE	FALSE
2	53	0	0	23	1668.285119	0.008333333	0.016312696	0	0	2	1	1	9	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0	2	1	1	4	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	13	334.9666667	0	0.007692308	0	0	2	1	1	1	4	Returning_Visitor	TRUE	FALSE
0	0	0	0	2	32	0	0.1	0	0	2	2	2	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	20	2981.166667	0	0.01	0	0	2	2	4	4	4	Returning_Visitor	FALSE	FALSE
0	0	0	0	8	136.1666667	0	0.008333333	0	1	2	2	2	5	1	Returning_Visitor	TRUE	FALSE
0	0	0	0	2	0	0.2	0.2	0	0	2	3	3	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	3	105	0	0.033333333	0	0	2	3	2	1	5	Returning_Visitor	FALSE	FALSE
0	0	0	0	2	15	0	0.1	0	0.8	2	2	4	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	1	0	0.2	0.2	0	0	2	2	2	4	1	Returning_Visitor	TRUE	FALSE
0	0	0	0	5	156	0	0.04	0	0	2	1	1	9	3	Returning_Visitor	FALSE	FALSE
4	64.6	0	0	32	1135.444444	0.002857143	0.00952381	0	0	2	2	2	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	4	76	0.05	0.1	0	0	2	1	1	1	3	Returning_Visitor	FALSE	FALSE
0	0	0	0	4	63	0	0.05	0	0.2	2	2	6	1	3	Returning_Visitor	FALSE	FALSE
1	6	1	0	45	1582.75	0.043478261	0.050821256	54.17976...	0.4	2	3	2	1	1	Returning_Visitor	FALSE	FALSE
0	0	0	0	2	35	0	0.1	0	0	2	1	1	6	3	Returning_Visitor	FALSE	FALSE

圖 2 Online Shoppers Purchasing Intention Dataset Data Set 原始資料

Administrative	Admini...	Infor...	Informationa...	ProductRelated	Product...	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType_New_Visitor	VisitorType...	VisitorTyp...	Weekend	Revenue
0	0	0	0	0	0	0.18	0.181818182	0	0	6	2	2	7	6	0	1	0	0	0
0	0	0	0	0	0	0.2	0.2	0	0	11	1	1	8	8	0	1	0	0	0
0	0	0	0	0	0	0.2	0.2	0	0	11	1	1	9	2	0	1	0	0	0
0	0	0	0	0	0	0.2	0.2	0	0	11	1	1	9	3	0	1	0	0	0
0	0	0	0	0	0	0.2	0.2	0	0	5	2	2	8	3	0	1	0	0	0
0	0	0	0	0	0	0.2	0.2	0	0	5	3	2	3	18	0	1	0	1	0
0	0	0	0	1	0	0	0.1	0	0	11	2	6	1	3	0	1	0	0	0
0	0	0	0	1	0	0	0.1	0	0	12	8	13	9	20	0	0	1	0	0
0	0	0	0	1	0	0	0.1	0	0	3	2	2	2	6	0	1	0	0	0
0	0	0	0	1	0	0	0.1	0	0	5	2	2	3	18	0	1	0	1	0
0	0	0	0	1	0	0	0.2	0	0	12	3	2	4	1	0	1	0	0	0
0	0	0	0	1	0	0	0.2	0	0	3	2	2	1	1	0	1	0	0	0
0	0	0	0	1	0	0	0.2	0	0	3	2	2	1	2	0	1	0	0	0
0	0	0	0	1	0	0	0.2	0	0	3	2	2	5	1	0	1	0	0	0
0	0	0	0	1	0	0	0.2	0	0	7	1	1	4	4	0	1	0	0	0
0	0	0	0	1	0	0	0.2	0	0.2	5	2	2	6	1	0	1	0	0	0
0	0	0	0	1	0	0.1	0.2	0	0	7	2	2	2	5	1	0	0	1	0
0	0	0	0	1	0	0.2	0.2	0	0	10	1	1	1	1	1	0	0	1	0
0	0	0	0	1	0	0.2	0.2	0	0	10	1	1	8	1	0	1	0	1	0
0	0	0	0	1	0	0.2	0.2	0	0	10	2	2	6	2	0	1	0	0	0
0	0	0	0	1	0	0.2	0.2	0	0	10	2	5	1	20	0	1	0	0	1
0	0	0	0	1	0	0.2	0.2	0	0	10	3	2	2	3	1	0	0	1	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	1	1	0	0	1	0	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	1	1	1	0	0	0	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	1	20	0	1	0	0	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	1	3	0	1	0	0	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	2	2	0	1	0	0	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	3	2	0	1	0	0	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	3	2	0	1	0	0	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	3	3	0	1	0	0	0
0	0	0	0	1	0	0.2	0.2	0	0	11	1	1	3	3	0	1	0	0	0

圖 3 Online Shoppers Purchasing Intention Dataset Data Set 正規化後資料

4.2 實驗設計

利用範圍 3 至 10 的迴圈對 K-means 套件參數(n_clusters)進行操作，並記錄每一筆迴圈數值帶入參數後所執行的時間與純度。

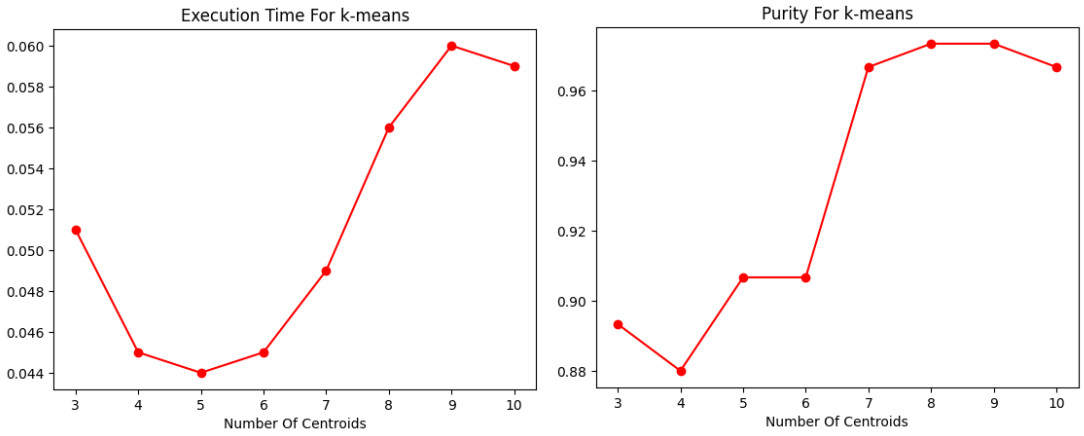

圖 4 Iris Data Set 之 K-means 各參數執行時間與純度

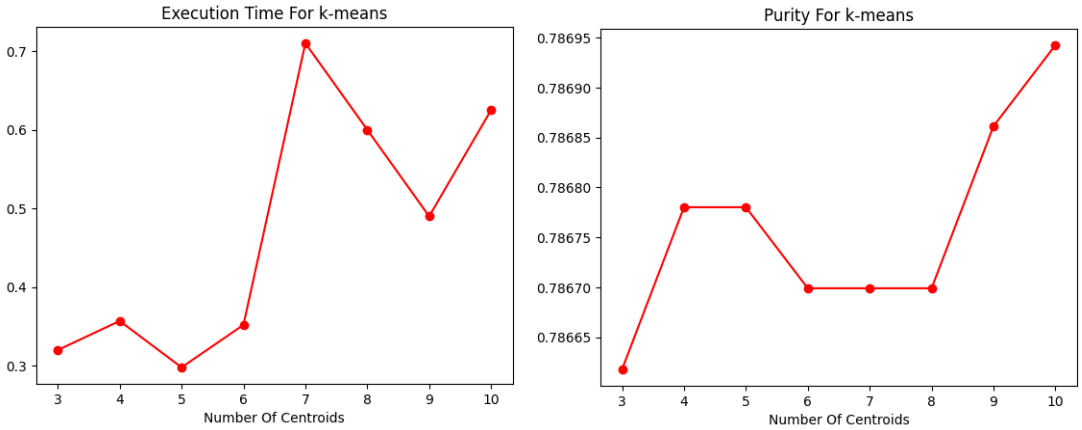

圖 5 Online Shoppers Purchasing Intention Dataset Data Set 之 K-means 各參數執行時間與純度

利用範圍 3 至 10 的迴圈對階層式分群套件之不同計算方式與參數(n_clusters)進行操作，並記錄每一筆迴圈數值帶入參數後所執行的時間與純度。

利用範圍 0.3 至 10 的迴圈對 DBSCAN 套件參數(eps)進行操作，並記錄每一筆迴圈數值帶入參數後所執行的時間與純度。

利用範圍 10 至 50 的迴圈對 DBSCAN 套件參數(min_samples)進行操作，並記錄每一筆迴圈數值帶入參數後所執行的時間與純度。

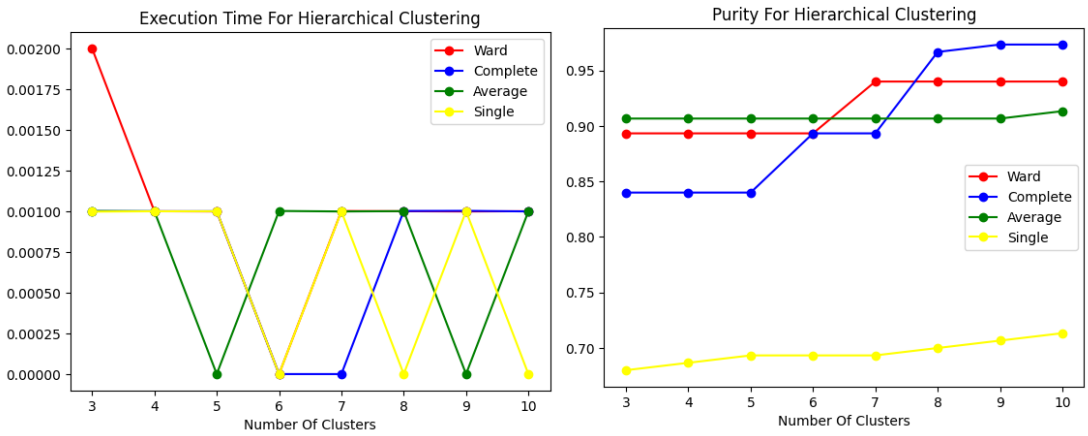

圖 6 Iris Data Set 之階層式分群各參數執行時間與純度

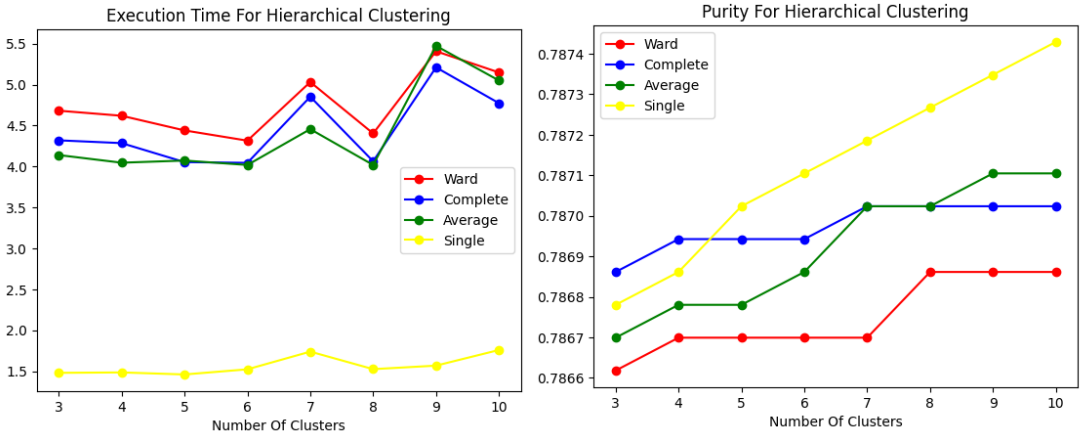

圖 7 Online Shoppers Purchasing Intention Dataset Data Set 之階層式分群各參數執行時間與純度

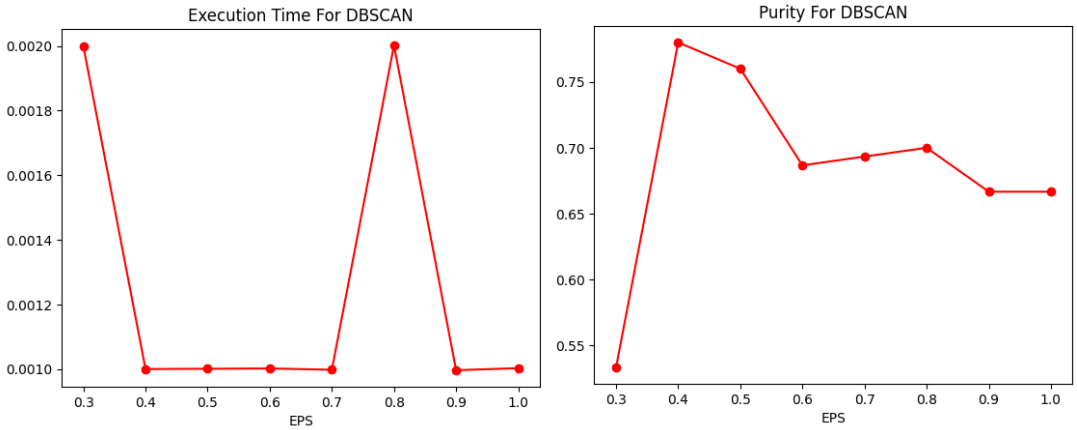

圖 8 Iris Data Set 之 DBSCAN EPS 執行時間與純度

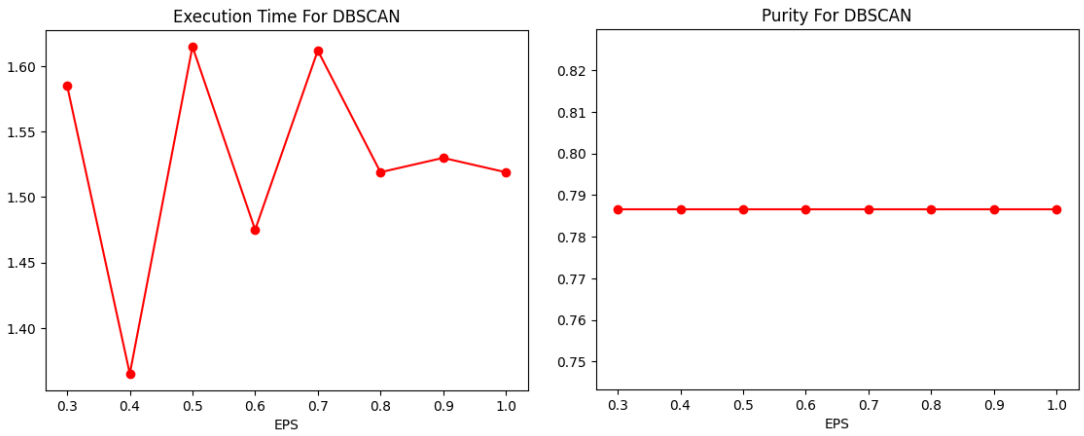

圖 9 Online Shoppers Purchasing Intention Dataset Data Set 之 DBSCAN EPS 執行時間與純度

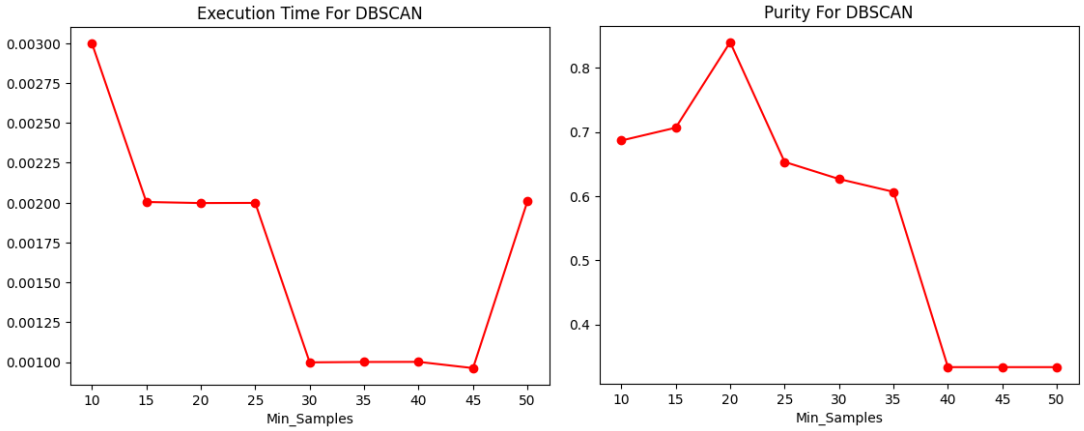

圖 10 Iris Data Set 之 DBSCAN Min_Samples 執行時間與純度

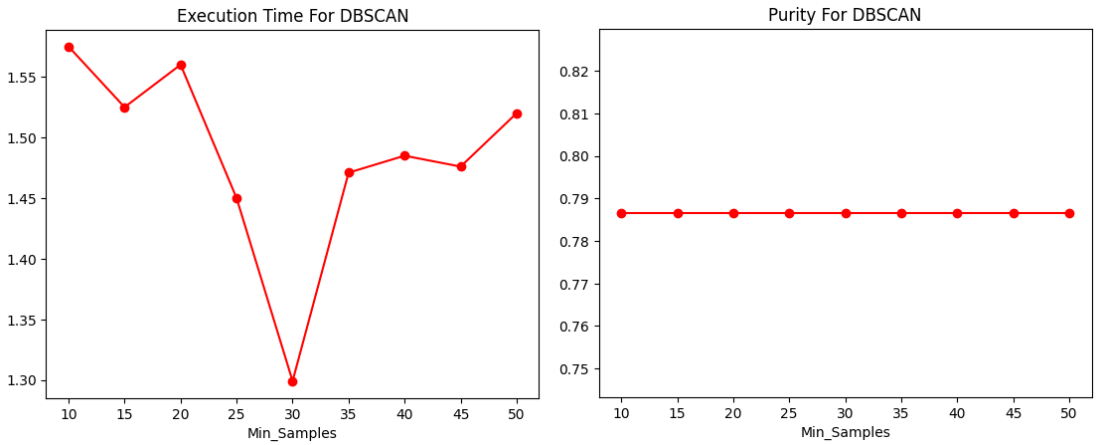

圖 11 Online Shoppers Purchasing Intention Dataset Data Set 之 DBSCAN Min_Samples 執行時間與純度

4.3 實驗結果

透過實驗設計獲得各演算法之較佳參數，再使用各演算法之品質衡量指標，得知對於不同資料集而言，較佳之演算法也不同。

表 3

Iris Data Set 品質衡量指標

演算法	品質衡量指標	數值
K-means	Silhouette	.553
階層式分群	Silhouette	.767
DBSCAN	Silhouette	.414

表 4

Online Shoppers Purchasing Intention Dataset Data Set 品質衡量指標

演算法	品質衡量指標	數值
K-means	Silhouette	.684
階層式分群	Silhouette	.835
DBSCAN	Silhouette	-.365

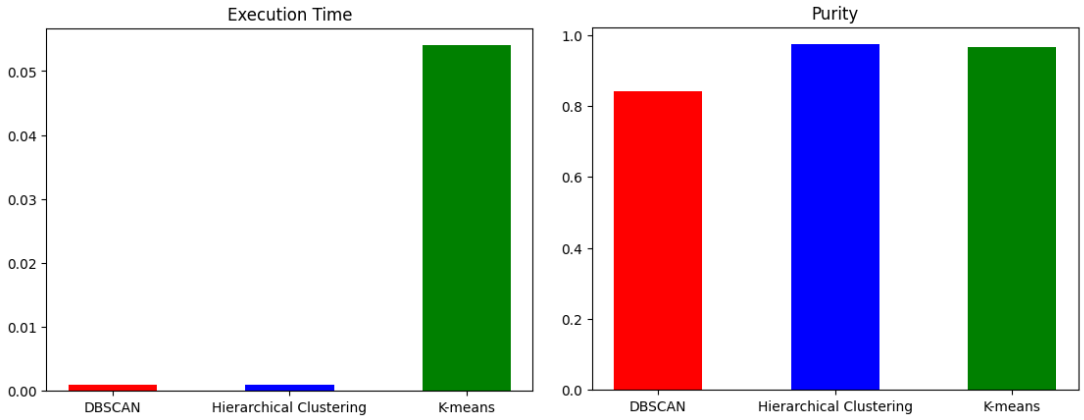

圖 12 Iris Data Set 演算法比較

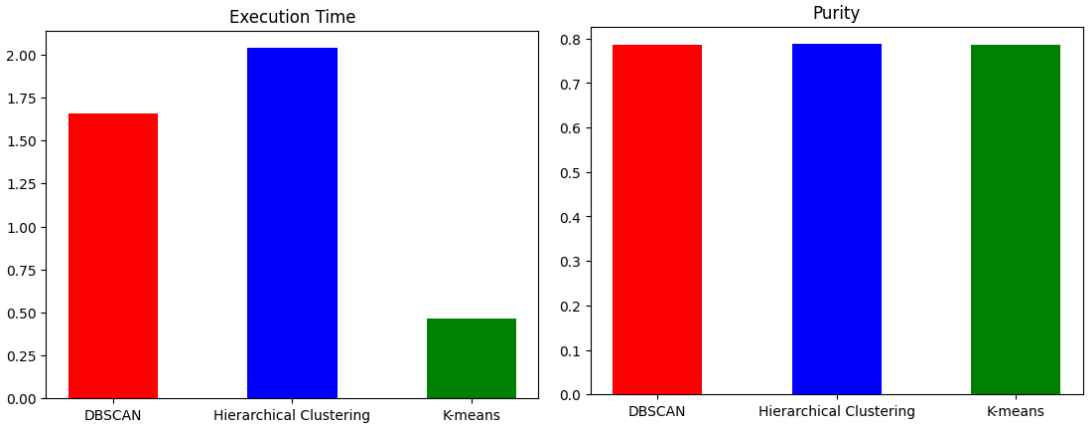

圖 13 Online Shoppers Purchasing Intention Dataset Data Set 演算法比較

五、結論

本研究使用 SQL Server 進行前置處理，後利用 python 進行各演算法(K-means、DBSCAN、Hierarchical Clustering)分析，計算出兩資料集在各演算法間哪項最佳，結果得知 Online Shoppers Purchasing Intention 在資料集之 Purity 較高，而 Iris 鳶尾花資料集 Purity 為 Hierarchical Clustering 階層式分群最高。

六、參考文獻

mokainzi (Dec 2020)。online_shoppers_purchasing_intention.ipynb。
https://github.com/mokainzi/machine_learning/blob/main/final_project/online_shoppers_purchasing_intention.ipynb

tonykuoyj (Dec 2016)。機器學習 (4) 分群演算法。
https://ithelp.ithome.com.tw/articles/10187314

Tommy Huang (Apr 2018) 。機器學習: 集群分析 K-means Clustering。
https://chih_shenghuang821.medium.com/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E9%9B%86%E7%BE%A4%E5%88%86%E6%9E%90-k-means-clustering-e608a7fe1b43

Yeh James (Oct 2017) 。[資料分析&機器學習] 第 2.4 講：資料前處理 (Missing data, One-hot encoding, Feature Scaling) 。
https://medium.com/jameslearningnote/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E7%AC%AC24%E8%AC%9B%E8%B3%87%E6%96%99%E5%89%8D%E8%99%95%E7%90%86-missing-data-one-hot-encoding-feature-scaling-3b70a7839b4a

Alan Wang (Aug 2020) 。 KMeans：能從資料中找出 K 個分類的非監督式機器學習演算法 。
https://alankrantas.medium.com/kmeans%E8%83%BD%E5%BE%9E%E8%B3%87%E6%96%99%E4%B8%AD%E6%89%BE%E5%87%BAk%E5%80%8B%E5%88%86%E9%A1%9E%E7%9A%84%E9%9D%9E%E7%9B%A3%E7%9D%A3%E5%BC%8F%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E6%BC%94%E7%AE%97%E6%B3%95%E6%89%80%E4%BB%A5%E5%AE%83%E5%88%B0%E5%BA%95%E6%9C%89%E5%95%A5%E7%94%A8%E4%BD%BF%E7%94%A8-scikit-learn-%E8%88%87-python-5dd8c0c8b167

Soner Yıldırım (Jan 2021)。 Evaluation Metrics for Clustering Models。
https://towardsdatascience.com/evaluation-metrics-for-clustering-models-5dde821dd6cd

Yomi Kastro(aug 2018)。Online Shoppers Purchasing Intention Dataset Data Set。
https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

Michael Marshall(JUL 2018)。Iris Data Set。
https://archive.ics.uci.edu/ml/datasets/Iris

Alan Wang(2020，8 月 17 日)。KMeans：能從資料中找出 K 個分類的非監督式機器學習演算法 — — 所以它到底有啥用？(使用 scikit-learn 與 Python)。Medium。https://reurl.cc/MX2qqL

PyInvest(2020，6 月 15 日)。[Python 實作] 層次聚類 Hierarchical

Clustering。Pyecontech。https://reurl.cc/lZp8LY
林倢愷(2021，4 月 7 日)。不要再用 K-means！ 超實用分群法 DBSCAN 詳解。Medium。https://reurl.cc/7jEO39
Soner Yıldırım. (2021,January 10). Evaluation Metrics for Clustering Models. Medium. https://reurl.cc/nZpkZv
scikit-learn. sklearn.cluster.AgglomerativeClustering. scikit-learn. https://reurl.cc/85Y39g
newaurora(2019，7 月 22 日)。python 使用 matplotlib 畫折線圖(Line chart)。痞客邦。https://reurl.cc/QWNLRM
STEAM 教育學習網。長條圖 Bar Chart。STEAM 教育學習網。https://steam.oxxostudio.tw/category/python/example/matplotlib-bar.html