國立雲林科技大學

資料探勘專案作業三

利用分群演算法實作群集分析

成員：

M11021012 林承威

M11021028 劉軒瑋

M11021052 邱守燦

M11021059 李鴻庭

指導教授：

許中川 教授

日期：

2021 年 12 月 16 日

摘要

本研究使用了 Iris 與 Car Evaluation 資料集進行集群分析(Cluster Analysis)，集群分析被應用的領域十分廣泛，包括:機器學習、資料探勘、影像辨識等領域，集群分析將相似的物件透過不同變數考量(例如:距離、密度)的方法分成不同的組別或者更多的子集合，集群分析通常是非監督式學習的演算法建立而成；本研究使用到的演算法包括:K-means、階層式分群、DBSCAN，將 Iris 資料集根據「Label」欄位分為 3 群、Car Evaluation 根據「Class Values」欄位分為 4 群，且以 Purity 進行評估其績效，研究結果顯示 K-means 為其績效指標與其他兩者差距微小，但在花費時間上較快，故 K-means 為最佳的分群方法。

關鍵字：Clustering、K-means、Hierarchical、DBSCAN

一、 緒論

1.1 動機

1.1.1 Iris Dataset

花卉的種類繁多，因其觀賞性與經濟價值，除原生種之外，亦有透過人工培育與基因改造的品種，不同的花卉品種，在其外觀呈現上也會有所差異。台中花博曾經預估將為台中帶來大量商機，因此，在展場花卉的布置上，更應該具有主題性，透過分群演算法，可以將屬性(花瓣、萼片)各異，但相似度高的花卉分群，進而調整花博場地的佈置。

1.1.2 Car Evaluation Dataset

汽車是現代人常使用的代步工具，如若家庭經濟條件與環境許可，買輛汽車是值得考慮的選擇。人們購買車的考量大多包含：價格、舒適性、容量及安全性等。如何說服客戶買車也考驗行銷人員的能力，倘若能針對車輛的各種特性(價格、維護費用等)進行分群，便能針對最終的分群結果，制定合適的行銷計畫方案。

1.2 研究目的

1.2.1 Iris Dataset

分群演算法的方法很多，諸如 K-means、階層式分群、DBSCAN 等。本次研究除了替花卉分群外，將比較 K-means、階層式分群、DBSCAN 三個方法的分群所花費時間，並使用純度(Purity)指標衡量分群品質，最終決定出何者演算法適用於花卉分群。

1.2.2 Car Evaluation Dataset

根據最終的分群結果,制定良好的行銷計畫,也得仰賴良好的分群演算法，考量到行銷計劃對於公司收益的重大影響，分群演算法的比較也至關重要，本次研究除了替車輛進行分群外，將比較 K-means、階層式分群、DBSCAN 三個方法的分群所花費時間，並使用純度(Purity)指標衡量分群品質，最終決定出何者演算法適用於車輛分群。

二、 資料集

2.1 資料集

2.1.1 IRIS Dataset 說明

此資料集建立於 1936 年共有 150 筆，4 個欄位。

表 1 IRIS Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
sepal length in cm	萼片長度(cm)
sepal width in cm	萼片寬度(cm)
petal length in cm	花瓣長度(cm)
petal width in cm	花瓣寬度(cm)

2.1.2 Car Evaluation Dataset 說明

此資料集建立於 1997 年 6 月 1 日共有 1728 筆，6 個欄位。

表 2 Car Evaluation Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
buying	購買價格
maint	維護費用
doors	車門數
persons	可乘坐人數
lug_boot	行李箱大小
safety	安全性

三、 方法

3.1 程式架構

3.1.1 IRIS Dataset

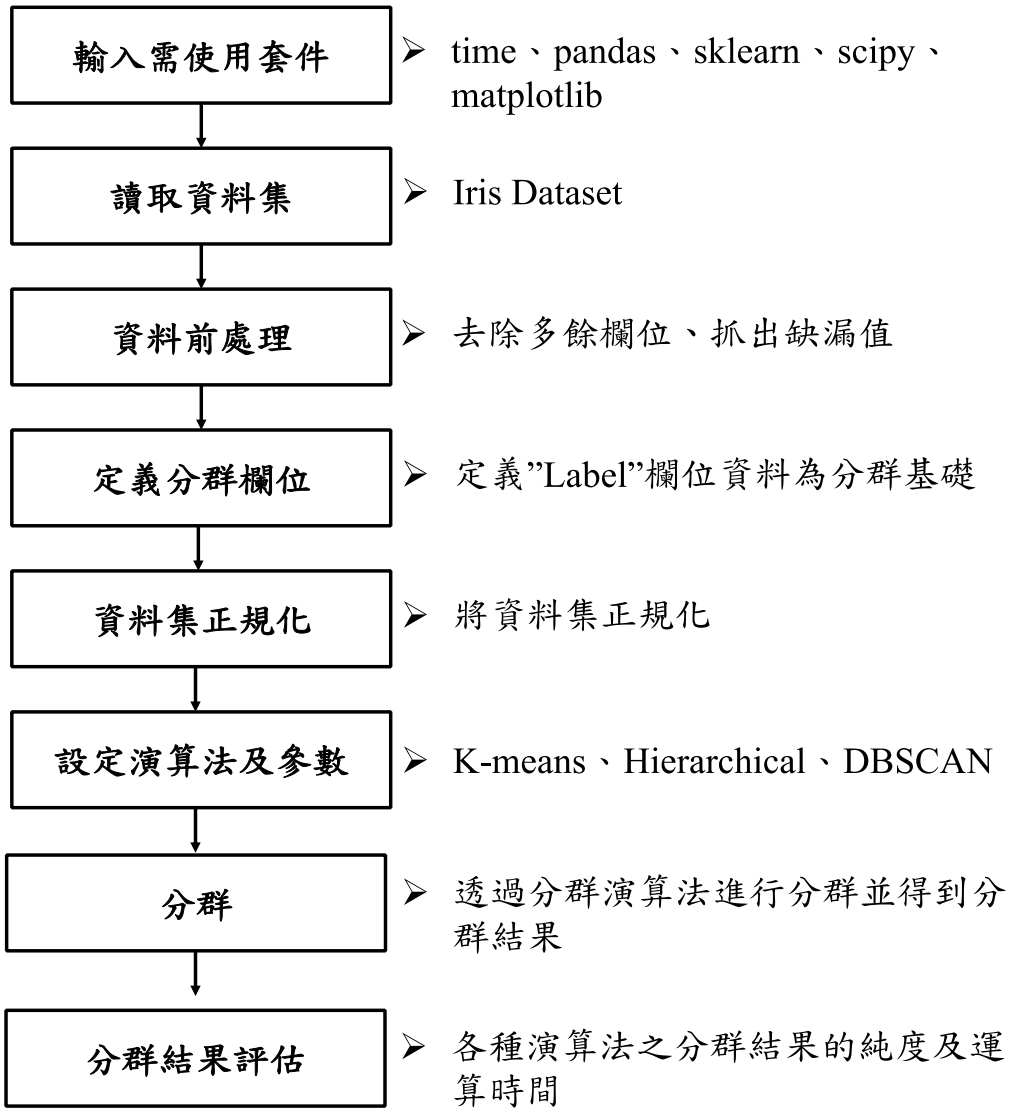

圖 1 IRIS Dataset 之程式架構流程圖及說明

3.1.2 Car Evaluation Dataset

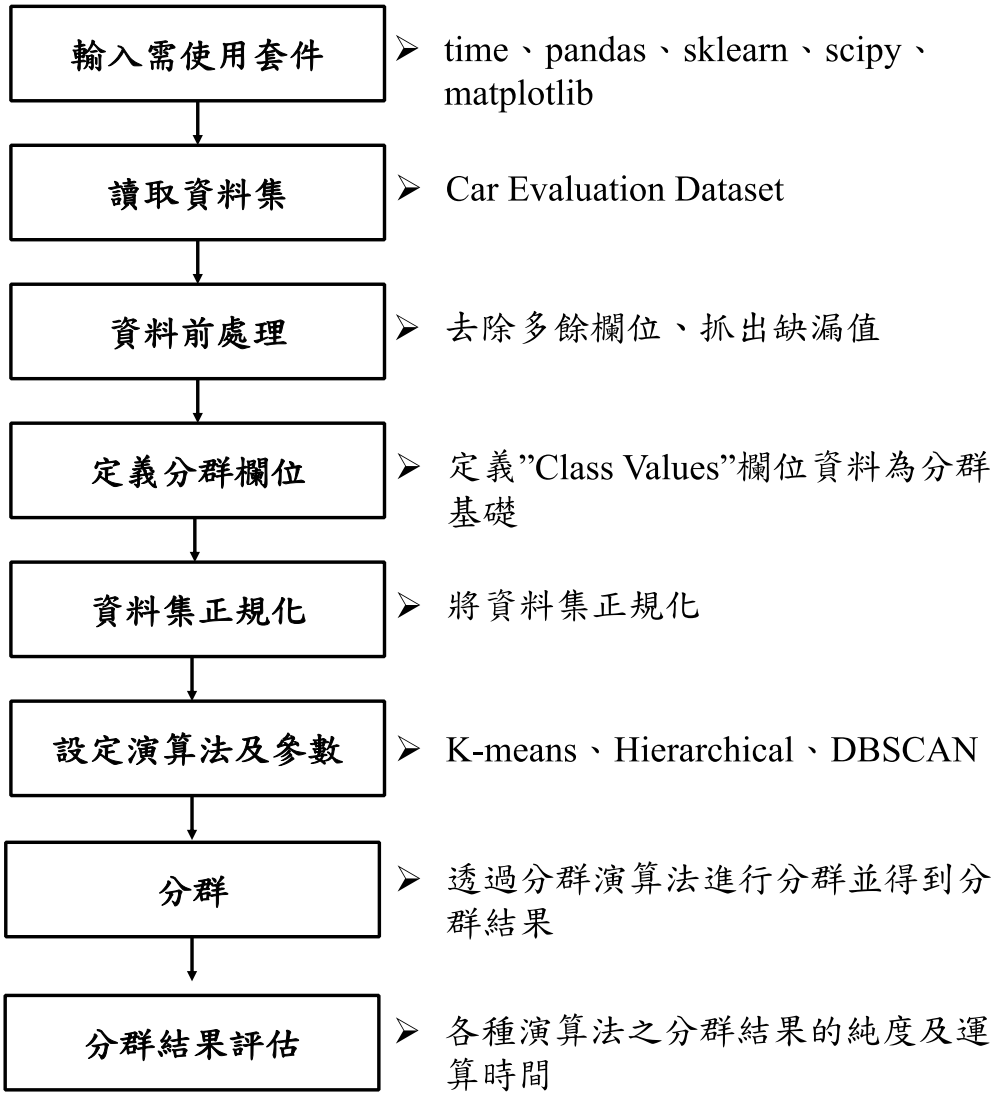

圖 2 Car Evaluation Dataset 之程式架構流程圖及說明

3.2 程式執行方法

3.2.1 K-means(k-means clustering)

K-means(k-means clustering) 是一種非監督式學習的演算法，將一群資料分成 k 群 (cluster)，演算法上是透過計算資料間的距離作為分群的依據，較相近的資料會成形成一群並透過加權計算或簡單平均可以找出中心點，透過多次反覆計算與更新各群中心點後，可以找出代表該群的中心點，之後便可以透過與中心點的距離來判定測試資料屬於哪一分群，或可進一步被用來資料壓縮，代表特定類別資料，以達到降低雜訊或填空值等議題。此為分割式分群法(partitional clustering)

中的一種，藉由反覆運算，逐次降低誤差目標值，直到目標函數值不再變化或更低，就達到分群的最後結果。

3.2.2 階層式分群法（hierarchical clustering）

階層式分群法（hierarchical clustering）是一種透過階層架構的方式，將資料層層反覆地進行分裂或聚合，以產生最後的樹狀結構，常見的方式有兩種： 如果採用聚合的方式，階層式分群法可由樹狀結構的底部開始，將資料或群聚逐次合併 如果採用分裂的方式，則由樹狀結構的頂端開始，將群聚逐次分裂。

3.2.3 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)是一個比較有代表性的基於密度的分群演算法。與劃分和層次分群方法不同，它將群集定義為密度相連的點的最大集合，能夠把具有足夠高密度的區域劃分為群集，並可在具有干擾的空間資料庫中發現任意形狀的群集。

四、 實驗

4.1 實驗數據

將原始資料集進行**正規化**處理。

（1）Iris Dataset 資料集

	Sepal length	Sepal width	Petal length	Petal width	Lable
0	0.222222	0.625000	0.067797	0.041667	0.0
1	0.166667	0.416667	0.067797	0.041667	0.0
2	0.111111	0.500000	0.050847	0.041667	0.0
3	0.083333	0.458333	0.084746	0.041667	0.0
4	0.194444	0.666667	0.067797	0.041667	0.0
...	...	...	...	...	...
145	0.666667	0.416667	0.711864	0.916667	1.0
146	0.555556	0.208333	0.677966	0.750000	1.0
147	0.611111	0.416667	0.711864	0.791667	1.0
148	0.527778	0.583333	0.745763	0.916667	1.0
149	0.444444	0.416667	0.694915	0.708333	1.0

150 rows × 5 columns

圖 3 Iris Dataset 資料集

（2）Car Evaluation Dataset 資料集

	buying	maint	doors	persons	lug_boot	safety	Class Values
0	1.000000	1.000000	0.0	0.0	1.0	0.5	0.666667
1	1.000000	1.000000	0.0	0.0	1.0	1.0	0.666667
2	1.000000	1.000000	0.0	0.0	1.0	0.0	0.666667
3	1.000000	1.000000	0.0	0.0	0.5	0.5	0.666667
4	1.000000	1.000000	0.0	0.0	0.5	1.0	0.666667
...	...	...	...	...	...	...	...
1723	0.333333	0.333333	1.0	1.0	0.5	1.0	0.333333
1724	0.333333	0.333333	1.0	1.0	0.5	0.0	1.000000
1725	0.333333	0.333333	1.0	1.0	0.0	0.5	0.666667
1726	0.333333	0.333333	1.0	1.0	0.0	1.0	0.333333
1727	0.333333	0.333333	1.0	1.0	0.0	0.0	1.000000

1728 rows × 7 columns

圖 4 Car Evaluation Dataset 資料集

4.2 前置處理

4.2.1 Iris Dataset

將 Iris 資料集由原來的 txt 檔案格式轉換為 csv 檔案格式，首先進行資料前處理，以利後續的實驗；本組利用 lambda 語法檢查是否具有缺漏值，完成資料清洗，接著將 Iris 資料集裡面文字屬性使用 Labelencoder 進行轉碼，最後讀取印出完成資料的前置處理。

4.2.2 Car Evaluation Dataset

將 Car Evaluation 資料集由原來的 txt 檔案格式轉換為 csv 檔案格式，首先進行資料前處理，以利後續的實驗；本組利用 lambda 語法檢查是否具有缺漏值，完成資料清洗，接著將 Car Evaluation 資料集裡面文字屬性使用 Labelencoder 進行轉碼，最後讀取印出完成資料的前置處理。

4.3 實驗設計

4.3.1 Iris Dataset

(1)套件載入：由 Scikit-Learn 開源軟體載入 K-means 分群（KMeans）套件、階層式分群（AgglomerativeClustering）套件，其中，階層式分群額外載入(Scipy.cluster.hierarchy)輔助繪圖、DBSCAN 分群（DBSCAN）套件，以上為主程式套件，接著載入計時器(time)、python 繪圖(matplotlib)，來完成後續的實驗。

(2)定義分群基礎:根據「Label」欄位的資料為分群基礎(Iris-setosa－0 , Iris-versicolour－1 , Iris-virginica－2)。

(3)K-means：將 KMeans 方法的參數設定為分 3 群(n_clusters=3)、初始化

方法為(init='k-means++')目的為加速群集分析的收斂、最大迭帶次數為300(max_iter=300)、而不同的質心進行運算次數最多為 10 次(n_init=10)、最後質心初始化的隨機性為 0(random_state=0)。

(4)階層式分群:將 AgglomerativeClustering 方法的參數設定為分 3 群(n_clusters=3)、使用 Ward's method 進行聯動(linkage='ward') 計算最小化被合併集群的方差，最後繪圖來獲得分群結果。

(5)DBSCAN:將 DBSCAN 方法的參數設定(eps=0.35,min_sample=10)。

(6)繪圖:k-means 分群使用散佈圖(plt.scatter)的方式呈現；階層式分群根據 Ward's method 透過 dendrogram 輔助繪圖參數的設定，最後再利用 matplotlib 套件完成繪圖。

(7)純度與運算時間:由計時器(time)可得知運算時間；純度則利用各分群結果的最大值在總和的占比進行評估。

4.3.2 Car Evaluation Dataset

(1)套件載入：由 Scikit-Learn 開源軟體載入 K-means 分群（KMeans）套件、階層式分群（AgglomerativeClustering）套件，其中，階層式分群額外載入(Scipy.cluster.hierarchy)輔助繪圖、DBSCAN 分群（DBSCAN）套件，以上為主程式套件，接著載入計時器(time)、python 繪圖(matplotlib)，來完成後續的實驗。

(2)定義分群基礎:根據「Class Values」欄位的資料為分群基礎(acc－0,good－1,unacc－2,vgood－3)。

(3)K-means：將 KMeans 方法的參數設定為分 4 群(n_clusters=4)、初始化方法為(init='k-means++')目的為加速群集分析的收斂、最大迭帶次數為300(max_iter=300)、而不同的質心進行運算次數最多為 10 次(n_init=10)、最後質心初始化的隨機性為 5(random_state=5)。

(4)階層式分群:將 AgglomerativeClustering 方法的參數設定為分 4 群(n_clusters=4)、使用 Ward's method 進行聯動(linkage='ward') 計算最小化被合併集群的方差，最後繪圖來獲得分群結果。

(5)DBSCAN:將 DBSCAN 方法的參數設定(eps=0.6,min_sample=15)。

(6)繪圖:階層式分群根據 Ward's method 透過 dendrogram 輔助繪圖參數的設定，最後再利用 matplotlib 套件完成繪圖。

(7)純度與運算時間:由計時器(time)可得知運算時間；純度則利用各分群結果的最大值在總和的占比進行評估。

4.4 實驗結果

4.4.1 Iris Dataset

本研究使用 K-means、DBSCAN、階層式分群方法來進行，由下表可得知 階層式分群效果最好，純度為 0.312，運算時間為 0.014 秒，其次為 K-means，純度為 0.304，但運算速度為 0.092 秒，狀結構圖如下圖所示。

表 3 Iris Dataset 績效評估彙總表

方法	Purity	運算時間（秒）
K-means	0.304	0.092
階層式分群	0.312	0.014
DBSCAN	0.302	0.022

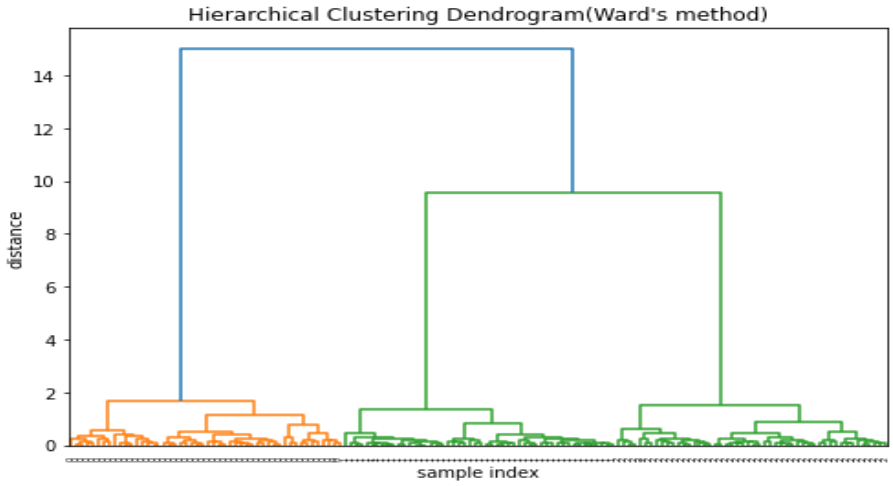

圖 5 Iris Dataset 階層式分群的階層樹

4.4.2 Car Evaluation Dataset

本研究使用 K-means、DBSCAN、階層式分群方法來進行，由下表可得知 階層式分群效果最好，純度為 0.212，運算時間為 0.312 秒，其次為 K-means，純度為 0.209，運算速度為 0.265 秒，狀結構圖如下圖所示。

表 4 Car Evaluation Dataset 績效評估彙總表

方法	Purity	運算時間（秒）
K-means	0.209	0.265
階層式分群	0.212	0.312
DBSCAN	0.184	0.168

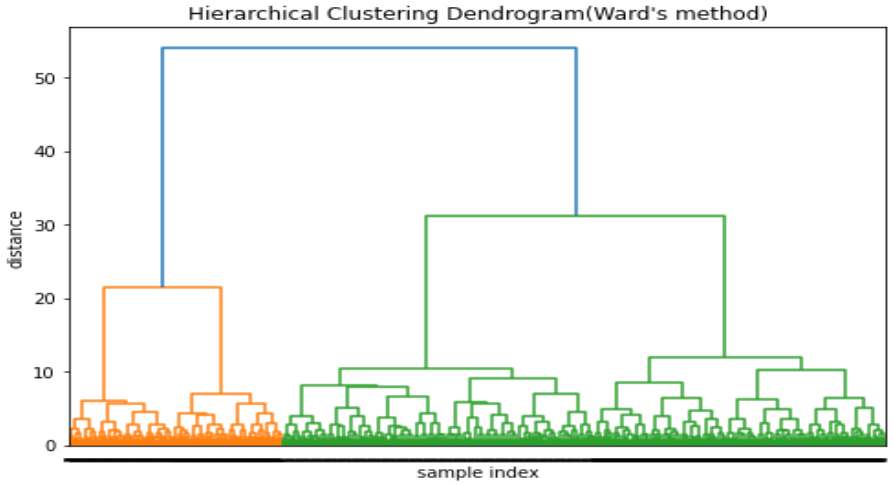

圖 6 Car Evaluation Dataset 階層式分群的階層樹

五、 結論

5.1 Iris Dataset

本研究利用 K-means、階層式分群、DBSCAN 三種方法，針對 Iris 資料集進行分群。研究結果顯示階層式分群法績效最好，其次為 K-means。

5.2 Car Evaluation Dataset

本研究利用 K-means、階層式分群、DBSCAN 三種方法，針對 Car Evaluation 資料集進行分群。研究結果顯示階層式分群法績效最好，其次為 K-means。

5.3 總結

由研究結果得知階層式分群方法雖然績效很好，但其花費時間較長，因此未來若有相關分群研究建議優先選擇 K-means 方法進行。

六、 參考文獻

[1]K-means 分群(K-means Clustering)
https://ithelp.ithome.com.tw/articles/10209058#:~:text=%5B%E6%BC%94%E7%AE%97%E6%B3%95%5D%20K%2Dmeans%20%E5%88%86%E7%BE%A4%20(K%2Dmeans%20Clustering)

[2]機器學習:集群分析 K-means Clustering
https://chih-sheng-huang821.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

[3]sklearn.cluster.KMeans — scikit-learn 1.0.1 documentation
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[4] [Python 實作]層次聚類 Hierarchical Clustering
https://pyecontech.com/2020/06/15/python_hierarchical_clustering/

[5]Python cluster.AgglomerativeClustering 方法代碼示例- 純淨天空
https://vimsky.com/zh-tw/examples/detail/python-method-sklearn.cluster.AgglomerativeClustering.html

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

[7]Ward's method – Wikipedia
https://en.wikipedia.org/wiki/Ward%27s_method

[8][Python 實作] 密度聚類 DBSCAN
https://pyecontech.com/2020/07/17/python_dbscan/

[9]Demo of DBSCAN clustering algorithm
https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html

[10]sklearn.cluster.DBSCAN — scikit-learn 1.0.1 documentation
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html