

資料探勘

專案作業三

組員：黃凱祥、吳俊園、賀鈞嗣、吳鈞達

2022 年 12 月

摘要

此次作業使用三種分群法: K-means、階層式分群、DBSCAN 來對資料分群，針對 Iris 資料集和自選的身體胖瘦度資料集進行資料預處理，藉由各別資料集中的屬性之間的相關性進行分群，找出規律，並藉由調整參數來增進分群的 purity。

在 Iris 資料集中，主要藉由花萼和花瓣的長寬來分群和判斷，自選資料集中則有運動習慣、飲食習慣等欄位來分群判斷，最後使用 scikit-learn 的 purity_score 計算分群資料純度。

關鍵字:python、K-means、hierarchical clustering、DBSCAN

第一章、緒論

1.1 動機

透過對課堂上所教授的 K-means、階層式分群、DBSCAN 分群法，先透過較為簡單、範例較多的 Iris 資料集進行實作練習，以了解學習的成果和累積實作之經驗。

完成後，再行自選一資料集身體胖瘦度資料集，用同樣的方法運算一遍，觀察結果差異。

1.2 目的

在 Iris 資料集中進行分群，主要觀看 K 值(分為幾群)對分群資料的純度，以及資料雜訊、離群值多寡等對分群造成的效應影響。

在身體胖瘦度資料集中，此資料集主要使用問卷調查，提供選項來蒐集身體的 BMI、身材比例等，和其填答者的運動、飲食習慣，因此會使用最後的身材比例來分群，調整群數 purity 盡可能提高。

第二章、方法

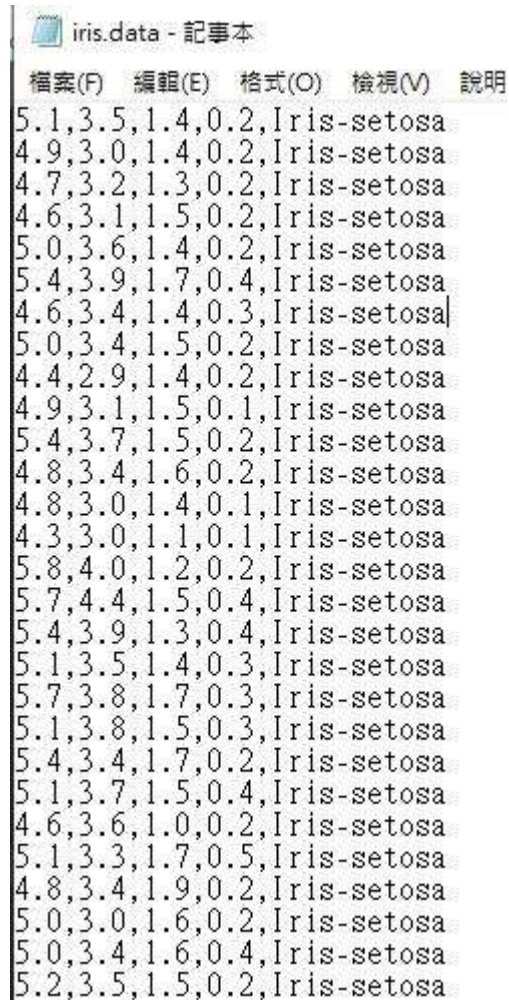
首先引入 `pandas`、`sklearn`、`numpy`、`cluster` 等函式庫，讀取資料，在自選資料集中進行資料預處理，將一些問卷填答選項數值化，對三種使用的分群法分別用不同的 python 程式撰寫和編譯執行(實作中在同一程式運行，後續的資料分群效益會受影響，可能由於實驗環境運行速度或其他未知原因)。

程式輸出結果如計算 `purity` 以及建立一個 PDF 用於儲存階層式分群的階層樹。

第三章、實驗

3.1 資料集

3.1.1 Iris 資料集



檔案(F)	編輯(E)	格式(O)	檢視(V)	說明
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3.0	1.4	0.1	Iris-setosa
4.3	3.0	1.1	0.1	Iris-setosa
5.8	4.0	1.2	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
5.4	3.9	1.3	0.4	Iris-setosa
5.1	3.5	1.4	0.3	Iris-setosa
5.7	3.8	1.7	0.3	Iris-setosa
5.1	3.8	1.5	0.3	Iris-setosa
5.4	3.4	1.7	0.2	Iris-setosa
5.1	3.7	1.5	0.4	Iris-setosa
4.6	3.6	1.0	0.2	Iris-setosa
5.1	3.3	1.7	0.5	Iris-setosa
4.8	3.4	1.9	0.2	Iris-setosa
5.0	3.0	1.6	0.2	Iris-setosa
5.0	3.4	1.6	0.4	Iris-setosa
5.2	3.5	1.5	0.2	Iris-setosa

Figure 1 adult.train 部分資料

Iris 資料如下：

- (a) 資料筆數：150。
- (b) 資料屬性欄位數：4。
- (c) 資料型態：小數資料、分類型別資料。
- (d) 缺失值：無。

其屬性的資訊為

- (a) 花萼長度：單位為 cm。
- (b) 花萼寬度：單位為 cm。
- (c) 花瓣長度：單位為 cm。
- (d) 花瓣長度：單位為 cm。
- (e) 花朵分類：山鳶尾、變色鳶尾和維吉尼亞鳶尾。

3.1.2 身體胖瘦資料集

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Gender	Age	Height	Weight	family_his	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObesidad		
2	Female	21	1.62	64 yes	no		2		3 Sometimes no			2 no		0	1 no	Public_Tra	Normal_Weight		
3	Female	21	1.52	56 yes	no		3		3 Sometimes yes			3 yes		3	0 Sometimes	Public_Tra	Normal_Weight		
4	Male	23	1.8	77 yes	no		2		3 Sometimes no			2 no		2	1 Frequently	Public_Tra	Normal_Weight		
5	Male	27	1.8	87 no	no		3		3 Sometimes no			2 no		2	0 Frequently Walking	Overweight_Level_I			
6	Male	22	1.78	89.8 no	no		2		1 Sometimes no			2 no		0	0 Sometimes	Public_Tra	Overweight_Level_II		
7	Male	29	1.62	53 no	yes		2		3 Sometimes no			2 no		0	0 Sometimes	Automobil	Normal_Weight		
8	Female	23	1.5	55 yes	yes		3		3 Sometimes no			2 no		1	0 Sometimes	Motorbike	Normal_Weight		
9	Male	22	1.64	53 no	no		2		3 Sometimes no			2 no		3	0 Sometimes	Public_Tra	Normal_Weight		
10	Male	24	1.78	64 yes	yes		3		3 Sometimes no			2 no		1	1 Frequently	Public_Tra	Normal_Weight		
11	Male	22	1.72	68 yes	yes		2		3 Sometimes no			2 no		1	1 no	Public_Tra	Normal_Weight		
12	Male	26	1.85	105 yes	yes		3		3 Frequently no			3 no		2	2 Sometimes	Public_Tra	Obesity_Type_I		
13	Female	21	1.72	80 yes	yes		2		3 Frequently no			2 yes		2	1 Sometimes	Public_Tra	Overweight_Level_II		
14	Male	22	1.65	56 no	no		3		3 Sometimes no			3 no		2	0 Sometimes	Public_Tra	Normal_Weight		
15	Male	41	1.8	99 no	yes		2		3 Sometimes no			2 no		2	1 Frequently	Automobil	Obesity_Type_I		
16	Male	23	1.77	60 yes	yes		3		1 Sometimes no			1 no		1	1 Sometimes	Public_Tra	Normal_Weight		
17	Female	22	1.7	66 yes	no		3		3 Always no			2 yes		2	1 Sometimes	Public_Tra	Normal_Weight		
18	Male	27	1.93	102 yes	yes		2		1 Sometimes no			1 no		1	0 Sometimes	Public_Tra	Overweight_Level_II		

Figure 2 身體胖瘦部分資料

身體胖瘦資料如下：

- (a) 資料筆數：2111。
- (b) 資料屬性欄位數：17。
- (c) 資料型態：問卷選項資料。
- (d) 缺失值：無。

其屬性的資訊為：

- (a) 性別：男或女。

- (b) 年齡：歲數。
- (c) 身高：單位(公尺)用於計算 BMI。
- (d) 體重：單位(公斤)。
- (e) 家族病史：是否家族內有人過重。
- (f) 飲食習慣一：是否常吃高熱量食物。
- (g) 飲食習慣二：吃蔬菜的頻率(從不、有時、總是)。
- (h) 飲食習慣三：每天吃幾餐(1、2、3 或 3 以上)。
- (i) 飲食習慣四：正餐之間是否會進食(從不、有時、經常、總是)。
- (j) 抽菸：是或否。
- (k) 喝水習慣：每日飲水量(未滿 1 公升、1 至 2 公升、2 公升以上)。
- (l) 控制熱量：是否每天檢查每日攝取熱量。
- (m) 運動習慣：每周運動天數(從不、1 至 2 天、2 至 4 天、4 至 5 天)。
- (n) 每日使用科技產品時間：每日小時數(0 至 2 小時、3 至 5 小時、5 小時以上)。
- (o) 飲酒習慣：飲酒頻率(從不、有時、時常、總是)。

- (p) 通勤習慣：使用的交通工具(汽車、摩托車、自行車、大眾運輸、步行)。

3.2 前置處理

在 Iris 資料集中，並未進行預處理，因為資料集十分明確。

在身體胖瘦資料集中，由於許多資料都是問卷資料，因此利用選項或區間對資料進行正規化處理，使其資料呈現數值化的狀態。

3.3 實驗設計

現在 scikit-learn 中已經包含 Iris 資料集，直接引入資料集，分別使用 K-means、階層式分群、DBSCAN 個別分群，調整 `n_clusters` 參數，也就是分群數量，並且撰寫 `purity` 函式個別運算純度，其中階層式分群另外使用 `matplotlib.pyplot` 的繪圖將其階層樹繪製並儲存成 PDF 檔案。

Iris 資料集中，群數設置為 2 或 3 的 `purity` 都非常好，而自選的身體胖瘦資料集，群數設置為 6 或更高才會有比較好的 `purity`，可能是由於資料欄位較多，又或者是資料離群值影響較大。

3.4 實驗結果

首先將兩種資料集的實驗結果製作成表格。如 Table 1、Table 2 所示。

iris↵	Kmeans↵	Hierarchical↵	DBSCAN↵
執行時間(ms)↵	89.9741↵	1.0287↵	1.0004↵
Purity↵	0.893↵	0.893↵	0.686↵

Table 1 Iris 資料集之三種分群法純度比較

身體胖瘦↵	Kmeans↵	Hierarchical↵	DBSCAN↵
執行時間(ms)↵	106.714487↵	49.893141↵	18.952847↵
Purity↵	0.4978↵	0.3912↵	0.2965↵

Table 2 身體胖瘦資料集之三種分群法純度比較

首先將兩種資料集的階層樹輸出如 Figure 3、Figure 4

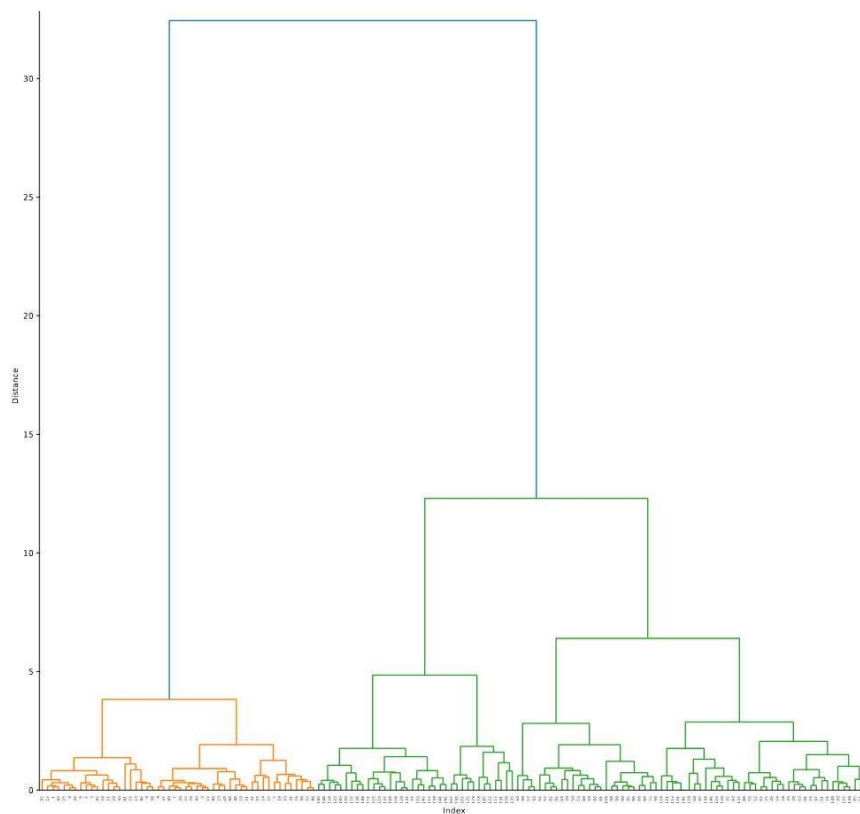


Figure 3 Iris 資料集階層樹

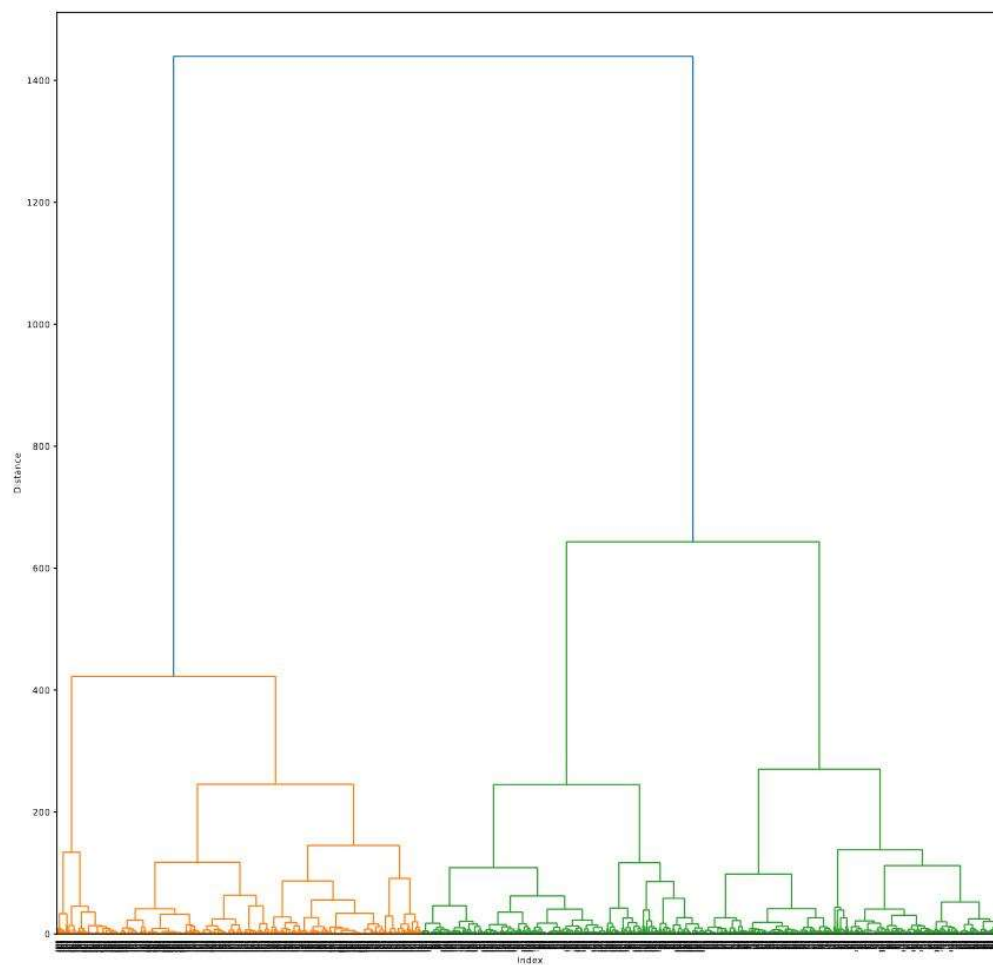


Figure 4 身體胖瘦資料集之階層樹

第四章、結論

在兩個資料集中，Iris 不需要預處理資料，純度都能達到 0.8 至 0.9 以上，由於資料集簡單，資料也不複雜，分群表現很好。

而自選的資料集資料比較複雜，經過預處理之後有把 purity 提高，但沒有到特別的好，大致上都落在 0.5 上下，可能資料的差異性不夠明顯，或者要去除一些較無相關的欄位等。

第五章、參考文獻

- [1] *Iris*. (n.d.). UCIdataset. <https://archive.ics.uci.edu/ml/datasets/iris>
- [2] *Estimation of Obesity Levels Based on Eating Habits and Physical Condition*. (n.d.). UCIdataset.
<https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+>
- [3] *K-Means*. (n.d.). Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
- [4] *Clustering-2.3.6. Hierarchical Clustering*. (n.d.). Scikit-Learn. <https://scikit-learn.org/stable/modules/clustering.html>
- [5] *DBSCAN*. (n.d.). Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>