國立雲林科技大學

資訊管理所

資料探勘專案作業三

比較機器學習模型分類聚集：
家族基因影響肥胖與鳶尾花種類之研究

指導教授：許中川教授

學生：工業工程與管理學系 碩一

M11121019 劉驊維

M11121027 洪源廷

M11121036 杜佳容

M11121047 黃惠妤

摘要

台灣地狹人稠，人口稠密，空氣及環境的污染相對提高，近年政府致力推動都市綠化，因而栽種許多植物，而非行家的民眾無法辨識植物品種，為了讓民眾對植物有更多的認識並探討模型間分群能力，本文利用不同分群方式並進行分析比較，分群方式為 K-means、階層式分群與 DBSCAN 群聚分析，找出判斷可能品種準確率最高的模型，且比較所花費時間差異，最後採用 Purity 指標衡量分群的品質。研究結果顯示使用 DBSCAN 的群聚分析為最佳模型，結果的準確率最高為 89%，且花費時間也是三者中最短的。

「減重」已成為全民運動，許多連喝水都會胖的人對吃不胖的人總是羨慕不已，然而科學家發現，人與人之間先天上確實存在差異，因此本文利用 Obesity Data Set 進行資料分析，藉由比較 K-means、階層式分群與 DBSCAN，探討家族基因是否導致易胖體質，並比較不同分群法所花費的時間差異，且採用 Purity 和 Accuracy 指標進行分群品質的衡量。研究結果為使用 DBSCAN 在 Purity 指標中高達 100%，但在 Accuracy 為最低，且需花費時間也最長；如針對 Accuracy 與時間進行考量，K-means 與階層式分析在 Purity 和 Accuracy 的績效差異並不大，但 K-means 在花費時間上比階層式分析短，因此兩者相比 K-means 為較佳模型。

一、緒論

1.1 動機

Iris Data Set

台灣的土地寸土寸金，為了可以善加利用，都市興建許多高樓大廈，但卻忽略都市綠化的必要性，而都市綠化可以帶來的好處為一、可以使都市氣象緩和；二、使都市的生活品質提升；三、節省能源，且抑制二氧化碳排放；四、多層效益，包括防止都市水災、噪音、防風、空氣污染、光合成……。近年政府意識到都市綠化的重要性，也在都市中規劃土地進行植物栽種，但民眾往往無法直接辨識出植物品種，因此為了增加民眾對植物的認識與知識，本文致力找出分群準確率最高的模型，讓民眾可以將圖片丟入並告訴民眾其品種，分群方式採用 K-means、階層式分群與 DBSCAN，就三種分群方式進行比較花費的時間及採用 Purity 指標衡量分群的品質。

Obesity Data Set

根據世界衛生組織的統計，2016 年 18 歲以上的成年人中，過重者超過 19 億人口，其中肥胖者超過 6.5 億人，表示肥胖已成為全球重要的公衛議題。而肥胖不僅容易影響生理健康，更會連帶影響心理健康，包括三高糖尿病、心腦疾病、退化性關節炎、癌症、憂鬱症……。而造成肥胖的原因不僅有人為因素，家庭基因遺傳也是主要原因之一，因此本文針對 Obesity Data Set 探討肥胖是否為家族遺傳，利用 K-means、階層式分群與 DBSCAN 三種不同的分群方式，找出最佳分群模型，並比較不同分群方式所花費的時間，最後採用 Purity 和 Accuracy 指標衡量分群的品質。

1.2 目的

Iris Data Set

民眾對於身邊的植物通常都無法清楚辨識，甚至搞混或認錯品種，因此栽種植物不僅可以讓城市綠化，同時也希望民眾可以多認識植物品種，本研究在 IRIS 資料集中，找出不同鳶尾花品種的特定特徵，並利用機器學習的方式來判定這些植物的可能品種，而使用四個特定特徵分別為花萼和花瓣的長度與寬度，目的為在不同分群的方式下，找出準確率最高的分群分類器。

Obesity Data Set

21 世紀最重要的公共衛生問題之一「肥胖」，是全世界主要的可預防死因，其困擾現今許多人，於是本文深入探討其肥胖成因，發現家族遺傳也是導致肥胖的因素之一，而出生家庭是不變的事實，因此本文探究家族遺傳是否真為影響肥胖起因，透過 UC Irvine Machine Learning Repository 中的 Obesity Data Set 進行資料分析，並比較 K-means、階層式分群與 DBSCAN 三種不同的分群方式，就其找出分群品質最佳的分類器。

二、資料集

2.1 Iris Data Set

Iris Data Set 資料集來自 UCI 機器學習儲存庫，包含 4 欄位總共 150 筆，紀錄了鳶尾花花朵部位的尺寸，資料屬性如表 2.2.1；資料部分內容表 2.2.2。

表 2.1.1 Irsi Data Set 資料屬性

屬性名稱	屬性說明	屬性類別
sepal lenth (cm)	花萼長	Ratio
sepal width (cm)	花萼寬	Ratio
petal lenth (cm)	花瓣長	Ratio
petal width (cm)	花瓣寬	Ratio
class	品種	Nominal

表 2.1.2 Irsi Data Set 資料集部分內容

sepal lenth (cm)	sepal width (cm)	petal lenth (cm)	petal width (cm)	class
5.1	3.5	1.4	0.2	Iris-sentosa
4.9	3	1.4	0.2	Iris-sentosa
4.7	3.2	1.3	0.2	Iris-sentosa
4.6	3.1	1.5	0.2	Iris-sentosa

2.2 Obesity Data Set

此研究資料集來自 UCI 機器學習儲存庫，名稱為 Obesity Data Set，資料集內有 17 個欄位，共有 2111 筆資料，紀錄了墨西哥、祕魯和歌倫比亞人民的飲食習慣和身體狀況，資料屬性如下表 2.2.1；資料部分內容如表 2.2.2、表 2.2.3、表 2.2.4。

表 2.1.1 Obesity Data Set 資料屬性

屬性名稱	屬性說明	屬性類別
Gender	性別	Nominal
Age	年齡	Ratio
Height	身高	Ratio
Weight	體重	Ratio
family_history_with_overweight	家族肥胖史	Nominal
FAVC	高熱量食物頻率	Interval

接續下頁

FCVC	食用蔬菜的頻率	Interval
NCP	主餐次數	Interval
CAEC	兩餐之間的食物消耗	Ordinal
SMOKE	吸菸	Nominal
CH2O	每日飲水消耗	Ratio
SCC	卡路里消耗監測	Nominal
FAF	身體活動頻率	Interval
TUE	技術設備使用時間	Interval
CALC	飲酒	Ordinal
MTRANS	使用的交通工具	Nominal
NObeyesdad	肥胖等級	Ordinal

表 2.1.2 Obesity Data Set 資料集部分內容

Gender	Age	Height	Weight	family_history with_overweight	FAVC	FCVC	NCP
Female	21	1.62	64	yes	no	2	3
Female	21	1.52	56	yes	no	3	3
Male	23	1.8	77	yes	no	2	3
Male	27	1.8	87	no	no	3	3

表 2.1.3 Obesity Data Set 資料集部分內容

CAEC	SMOKE	CH2O	SCC	FAF	TUE
Sometimes	no	2	no	0	1
Sometimes	yes	3	yes	3	0
Sometimes	no	2	no	2	1
Sometimes	no	2	no	2	0

表 2.1.4 Obesity Data Set 資料集部分內容

CALC	MTRANS	NObeyesdad
no	Public_Transportation	Normal_Weight
Sometimes	Public_Transportation	Normal_Weight
Frequently	Public_Transportation	Normal_Weight
Frequently	Walking	Overweight_Level_I

三、方法

3.1 實作說明

Iris Data

本研究使用 K-mean 分類、階層式分群以及 DBSCAN 進行分類，程式架構如圖 3.1.1 所示。把 iris.data 資料集投入多種模型後進行績效評估。

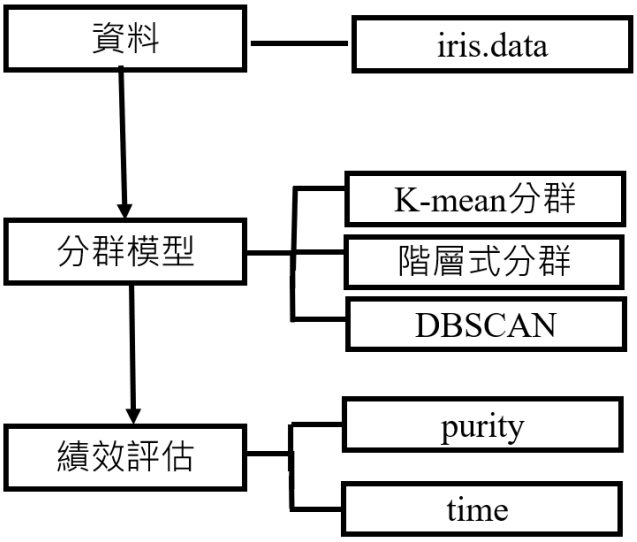

圖 3.1.1

Obesity DataSet

本研究使用 K-mean 分類、階層式分群、 DBSCAN 進行分類，程式架構如圖 3.1.2 所示。首先把 ObesityDataSet_raw_and_data_sinthetic 資料集進行資料轉換，接著做類別化，最後透過 PCA 降維降至 2 個特徵屬性，接著再將資料投入多種模型並進行績效評估。

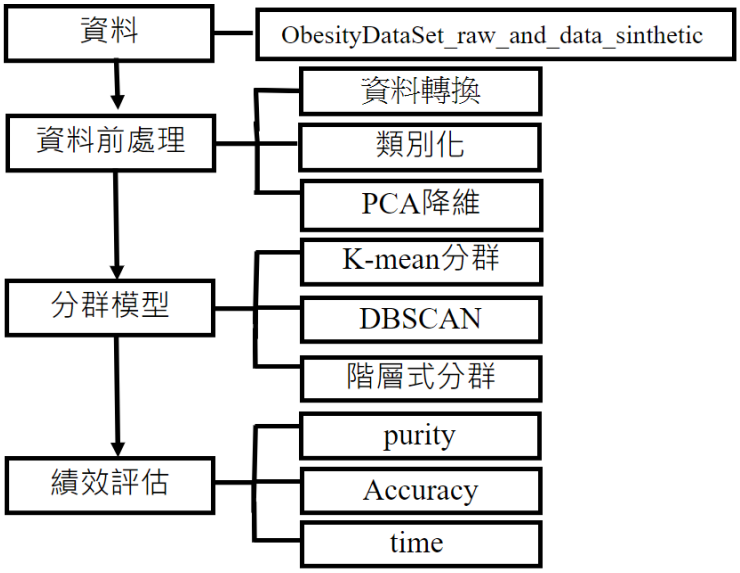

圖 3.1.2

3.2 操作說明

環境裡安裝套件

- Sklearn
- Pandas
- Numpy
- matplotlib
- time
- scipy.cluster.hierarchy

Iris Data

兩者模型訓練集和測試集皆必須與程式在同一個資料夾。程式結束後比較 3 個分類的 purity 和 time，如圖 3.2.1 所示，分別為 K-mean 分類、階層式分群、DBSCAN 分類的 purity 和 time 績效指標。

```
   kmeans_purity  kmeans_time  hierach_purity  hierach_time  dbscan_purity  \
0       0.906667     0.097376             1.0      0.037068       0.886667

   dbscan_time
0     0.007024
```

圖 3.2.1

Obesity DataSet

兩者模型訓練集和測試集皆必須與程式在同一個資料夾。程式結束後比較 3 個分類的 purity、acc 和 time，如圖 3.2.2 所示，分別為 K-mean 分類、階層式分群、 DBSCAN 分類的 purity、 accuracy 和 time 績效指標。

```
   kmeans_purity  kmean_acc  kmeans_time  hierach_purity  hierach_acc  \
0       0.609664   0.431549     0.111862        0.615822      0.43676

   hierach_time  dbscan_purity  dbscan_acc  dbscan_time
0      0.185743       0.998579    0.182032     3.878412
```

圖 3.2.2

四、實驗

4.1 前置處理

Iris Data

- 未做任何前置處理

Obesity DataSet

- 類別化：資料集內包含名目資料，因此先將這些資料利用 labelencoder 類別化。

- 資料轉換：資料集含有身高與體重的資料，但因為身高體重並無法直接的得出是否有肥胖標準，所以將透過兩個特徵屬性，統計出 BMI 的特徵屬性。
- 降維度：因特徵屬性眾多，為了避免高維度災難，將使用主成分要因模型（PCA）降維度，從 18 個維度降至 2 個維度。

4.2 實驗設計

資料前處理過後，本研究提出三種分群聚集模型，分別為 K-means、階層式分群與 Dbscan，並透過 Purity、Accuracy 以及訓練模型時間的績效指標，評估三種各個在資料集的分群能力，進而選取該資料集較適用的分群模型。

4.3 實驗結果

Iris Data

由表 3.4.1 結果得知使用 single link 的階層式分群在 Purity 指標中是三者最佳，高達 100%並且比在 Purity 指標第二高的模型所花費的時間，還更低，綜合評比此資料集較適用於階層式分群的模型，而圖 3.4.1 為階層式分群的階層樹狀圖，能更了解分群的狀況。

表 4.3.1 各模型績效評比

	K-means	Hierach	Dbscan
Purity	91%	100%	89%
Time(s)	0.0409	0.0070	0.0020

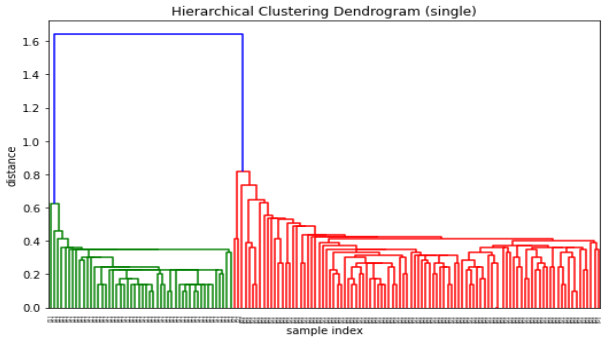

圖 4.3.1

Obesity DataSet

由表 3.4.2 結果得知使用 Dbscan 在 Purity 指標中是三者最佳，高達 100%，但在 Accuaracy 僅有 18%並且花費時間是最高，高達 93 秒整體績效除了 Purity 皆無比其餘兩者模型績效佳，若只針對 Accuracy 與時間的考量選擇 K-means 為較佳的模型，而圖 3.4.2 為階層式分群的階層樹狀圖，能更了解分群的狀況。

表 4.3.2 各模型績效評比

	K-means	Hierach	Dbscan
Purity	61%	62%	100%
Accuracy	57%	56%	18%
Time(s)	0.0279	0.2254	93.6903

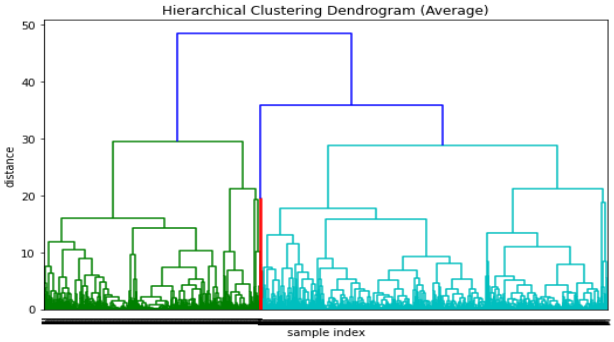

圖 4.3.2

五、結論

本研究分別利用 K-means、階層式分群與 DBSCAN 3 種分群聚集模型，針對 Iris Data 資料集進行分析，並採用 Purity 以及時間作為模型績效評比。研究結果顯示使用階層式分群的績效為最佳，準確率高達 89%，且花費時間也是三者中最短，因此若要針對 Iris 進行分類用階層式分群最為適合。

另一筆資料集使用相同實驗架構，探討家族基因是否導致易胖體質，比對 Purity、Accuracy 以及訓練模型時間，結果顯示 Purity 指標中 Dbscan 最佳，但 Accuaracy 僅有 18%並且花費時間是最高，若只考慮 Accuracy 與時間的考量選擇 K-means 為較佳的模型。

參考文獻

1. https://wwwwec.ntut.edu.tw/var/file/95/1095/img/3012/239539684.pdf
2. *https://health.udn.com/health/story/7426/3666021*
3. https://health.udn.com/health/story/6032/6165658
4. *https://www.cxyzjd.com/article/weixin_45529837/106313295*
5. *https://medium.com/ai-academy-taiwan/clustering-method-4-ed927a5b4377*
6. *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%BA-k-%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*
7. *https://axk51013.medium.com/%E4%B8%8D%E8%A6%81%E5%86%8D%E7%94%A8k-means-%E8%B6%85%E5%AF%A6%E7%94%A8%E5%88%86%E7%BE%A4%E6%B3%95dbscan%E8%A9%B3%E8%A7%A3-a33fa287c0e*