國立雲林科技大學資訊管理系

資料探勘

National Yunlin University of Science and Technology

Department of Information Management

專案作業三

以三種演算法將鳶尾花資料集與基於飲食習慣和身體狀況估計肥胖水平資料集分群

M11123048 陳韋彤

M11123060 毛俐蘋

M11123019 張芸婷

指導老師：許中川

Advisor：Chung-Chain Hsu，Ph.D.

中華民國 111 年 12 月

December 2022

摘要

本研究透過三種不同演算法進行資料探勘，對 Iris 資料集以及 Estimation of obesity levels 資料集進行群聚分析，依據資料樣本間的共同相似屬性進行分類、分群，使資料同質性最大化還有各組之間的異質性能達到最大化。本研究計算了不同演算法下所花費的時間並利用 Purity 指標以評估分群品質。Iris 資料集中利用三種演算法，在 purity 指標的衡量下，其中階層式分群是最好的，但三種演算法分群品質皆在 0.9 以上，可知分群品質優良；Estimation of obesity levels 資料集中，三種演算法在 purity 指標的衡量下，階層式分群是最好的，三種演算法分群品質在 purity 皆在 0.4 至 0.5 左右，可見在 prurity 衡量下品質三者無太大差異

關鍵字：資料分群、資料探勘、鳶尾花、肥胖水平

一、 緒論

1.1 動機

群聚分析能夠將大量的原始資料，依據資料樣本間的共同相似屬性進行分類、分群。本研究選用 Iris 資料集以及 Estimation of obesity levels 資料集進行群聚分析。Iris 資料集提供了鳶尾花的花瓣以及花萼的長度與寬度的數據資料，利用其相似的特徵進行分群；Estimation of obesity levels 資料集包括墨西哥、祕魯和哥倫比亞國家的個人肥胖水平數據，將根據個人不同之飲食習慣和身體狀況等因素用來對肥胖水平進行分群分析。

1.2 目的

本次研究使用 K-means、階層式分群及 DBSCAN 三種演算法對資料集進行分群，並比較三者所花費的時間差異。於 Iris 資料集中，利用花瓣和花萼的長度及寬度作為特徵，將鳶尾花分群出 Setosa、Versicolour 和 Virginica 三種不同品種的鳶尾花，並以 Purity 指標衡量分群品質；Estimation of obesity levels 資料集中，本研究利用其中的身體基本資料、飲食習慣等 16 個屬性，將肥胖水平分群成體重不足、正常體重、超重 I 級、超重 II 級、肥胖 I 型、肥胖 II 型和肥胖 III 型七群，了解墨西哥、祕魯和哥倫比亞等國家國人的健康狀況，並以 Purity 及 Silhouette Coefficient 指標衡量分群品質。

二、 方法

本研究實驗流程如圖 1 所示。

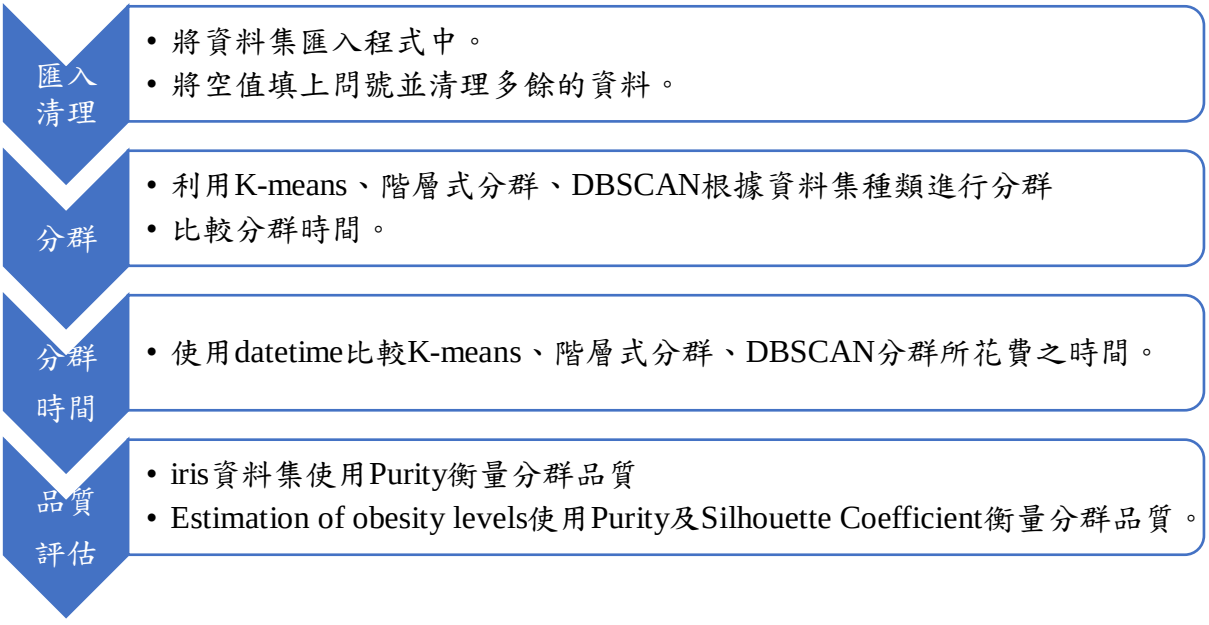

圖 1 實驗方法流程圖

三、 實驗

3.1 資料集

Iris 資料集包含了鳶尾花之花萼及花瓣的長寬數據以及三種鳶尾屬植物的類別。共有 5 種屬性，詳細屬性資訊如表 1 所示。

表 1

Iris 資料集屬性資料

屬性名稱	詳細資料描述
sepal length	sepal length in cm
sepal width	sepal width in cm
petal length	petal length in cm
petal width	petal width in cm
class	Iris Setosa, Iris Versicolour, Iris Virginica

Estimation of obesity levels 資料集包括根據飲食習慣和身體狀況估計墨西哥、秘魯和哥倫比亞國家的個人肥胖水平的數據。共有 17 種屬性，詳細屬性資料見表 2 所示。

表 2

Estimation of obesity levels 資料集屬性資料

屬性名稱	詳細資料描述
Gender	Female, Male
Age	Numeric value
Height	Numeric value in meters
Weight	Numeric value in kilograms
family_history_with_overweight	Yes, No
FAVC	Yes, No
FCVC	Never, Sometimes, Always
NCP	Between 1 y 2, Three, More than three
CAEC	No, Sometimes, Frequently, Always
SMOKE	Yes, No
CH2O	Less than a liter, Between 1 and 2 L, More than 2 L
SCC	Yes, No
FAF	I do not have, 1 or 2 days, 2 or 4 days, 4 or 5 days

(續下表)

(呈上表)

TUE	0–2 hours, 3–5 hours, More than 5 hours
CALC	I do not drink, Sometimes, Frequently, Always
MTRANS	Motorbike, Bike, Public Transportation, Walking
NObeyesda	result

3.2 前置處理

Iris 資料集以及 Estimation of obesity levels 資料集都用「?」代替缺失的資料，並且刪除空值。Iris 資料集將非數值屬性的 class 特徵值轉為數字類別資料，其中 Iris Setosa 轉為 0、Iris Versicolour 轉為 1、Iris Virginica 轉為 2，以利後續分群分析；Estimation of obesity levels 將 Gender(性別)、Age(年紀)、Height(身高)、Weight（體重）、family_history_with_overweight(家庭成員有超重史)、FAVC(是否經常食用高熱量食物)、FCVC（食用蔬菜的頻率）、NCP（主餐次數）、CAEC(兩餐之間的食物消耗量)、SMOKE(是否抽菸) 、CH2O（每天的水消耗量）、SCC(卡路里消耗監測)、FAF(身體活動頻率) 、TUE（使用技術設備的時間）、CALC(飲酒量)、MTRANS(使用的交通工具)、NObeyesda（結果）等非數值屬性的資料進行資料轉換，轉換為有順序大小之分的數字類別資料，方便進行分析。

3.3 實驗設計

在進行 Iris 資料集的分析時，本研究以「鳶尾花的種類」作為分群結果進行群聚分析，Estimation of obesity levels 資料集以「肥胖水平」作為分群結果進行分群。K-means 透過決定分成 k 群，並隨機選擇 k 點做為群集中心，將每個點分類至距離自己最近的群集中心，在群薪計算各組的群集中心，直到群集不變；階層式分群透過一種階層架構的方式，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；DBSCAN 透過更改畫出的圓形區域的半徑，以及高密度區域中的最少點數控制分群的數量。Iris 資料集是以 purity 作為衡量指標，Estimation of obesity levels 資料集則是以 Purity 和 Silhouette Coefficient 作為衡量指標比較分群結果。

3.4 實驗結果

3.4.1 Iris 資料集

在 Iris 資料集中，以花萼的長度(sepal length)、花萼的寬度(sepal width)、花瓣的長度(petal length)和花瓣的寬度(petal width)作為特徵進行分群，K-means 中，根據資料集 k = 3 分成三群；階層式分群中，使用 linkage = ward，affinity = euclidean 將資料分成三群；DBSCAN 中，Eps=0.5，min_samples=9 時分成三群。

0、1、2 分別代表 3 種不同的鳶尾花，其中 0 為「Iris-setosa」、1 為「Iris-versicolor」、2 為「Iris-virginica」，-1 則代表無法分群之雜訊。本研究以花瓣的長度和花瓣的寬度進行的分群結果，K-means 演算法如圖 2 所示，階層式分群演算法如圖 3 所示、DBSCAN 演算法如圖 4 所示。K-means、階層式分群、DBSCAN 分別所花費的分群時間如表 3 所示，三種演算法的 Purity 指標如表 4 所示。

表 3

Iris 資料集下三種演算法分群時間

演算法	分群時間
K-means	00.064822 秒
階層式分群	00.006981 秒
DBSCAN	00.003992 秒

表 4

Iris 資料集下三種演算法的 Purity 指標

演算法	Purity 指標
K-means	0.99
階層式分群	1.00
DBSCAN	0.94

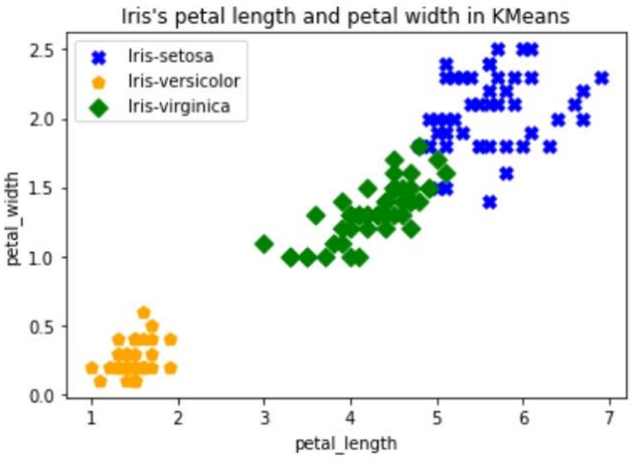

圖 2 K-means 下 Iris 資料集的分佈圖

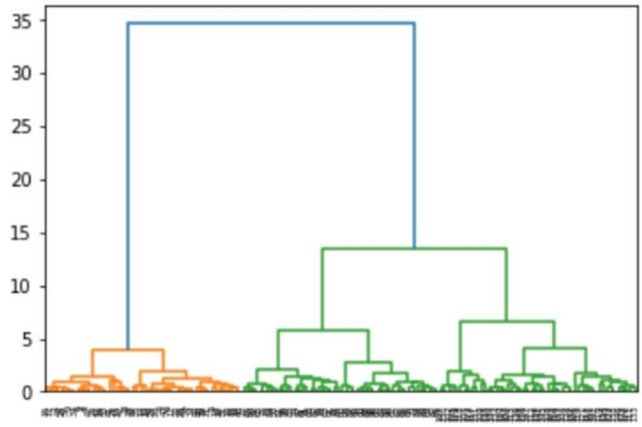

圖 3 階層式分群下 Iris 資料集的階層樹

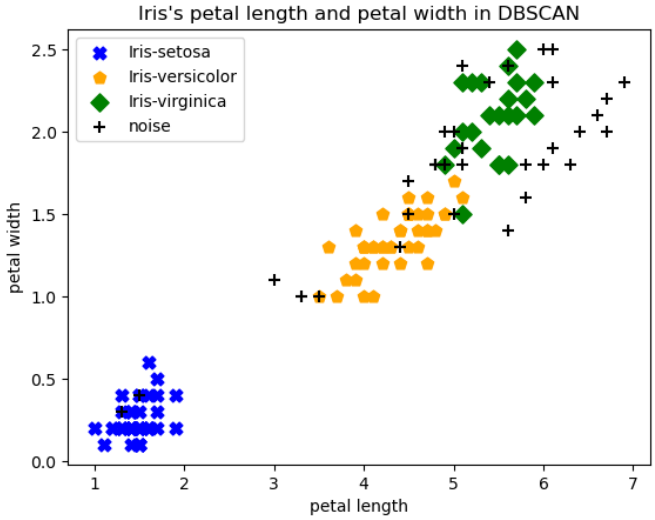

圖 4 DBSCAN 下 Iris 資料集的分佈圖

3.4.2 Estimation of obesity levels 資料集

在 Estimation of obesity levels 資料集中，該數據集包括根據飲食習慣和身體狀況估計墨西哥、秘魯和哥倫比亞國家的個人肥胖水平的數據，根據資料集所述，本研究將肥胖水平分為「體重不足」、「正常體重」、「超重 I 級」、「超重 II 級」、「肥胖 I 型」、「肥胖 II 型」和「肥胖 III 型」七種分群。

本研究以身高和體重進行的分群結果作為展示，K-means 演算法如圖 5 所示，階層式分群演算法如圖 6 所示、DBSCAN 演算法如圖 7 所示。在演算法 K-means 中，根據資料集使 k = 7 分成七群；階層式分群中，使用 linkage = ward，affinity = euclidean 將資料分成七群；DBSCAN 中，Eps=0.5，min_samples=25 時可剛好分成七群。K-means、階層式分群、DBSCAN 分別所花費的分群時間如表 5 所示。三種演算法的 Purity 和輪廓分析指標如表 6 所示。

表 5

Estimation of obesity levels 資料集下三種演算法分群時間

演算法	分群時間
K-means	00.187495 秒
階層式分群	00.223402 秒
DBSCAN	00.060837 秒

表 6

Estimation of obesity levels 資料集下三種演算法的 Purity 指標

演算法	Purity 指標	Silhouette Coefficient
K-means	0.52	0.43
階層式分群	0.53	0.37
DBSCAN	0.44	-0.33

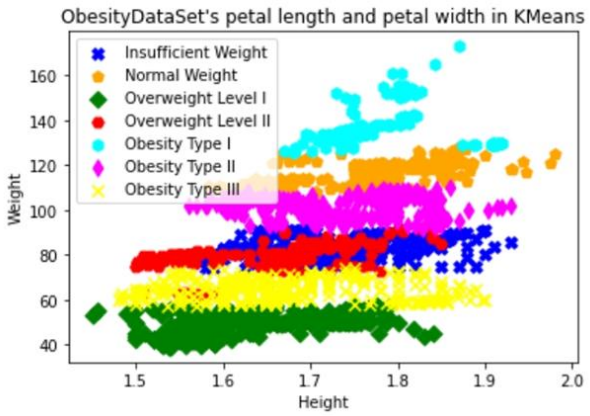

圖 5 K-means 下 Estimation of obesity levels 資料集的分佈圖

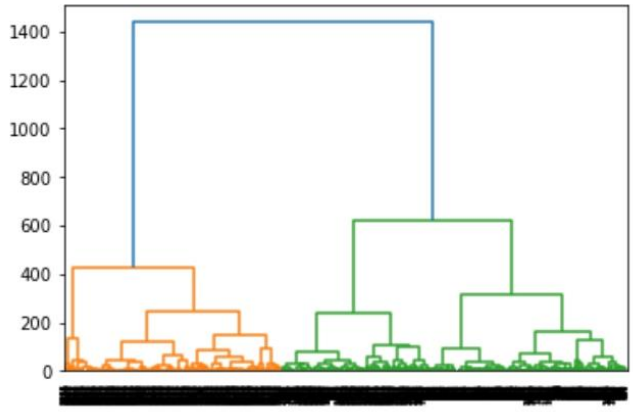

圖 6 階層式分群下 Estimation of obesity levels 資料集的階層樹

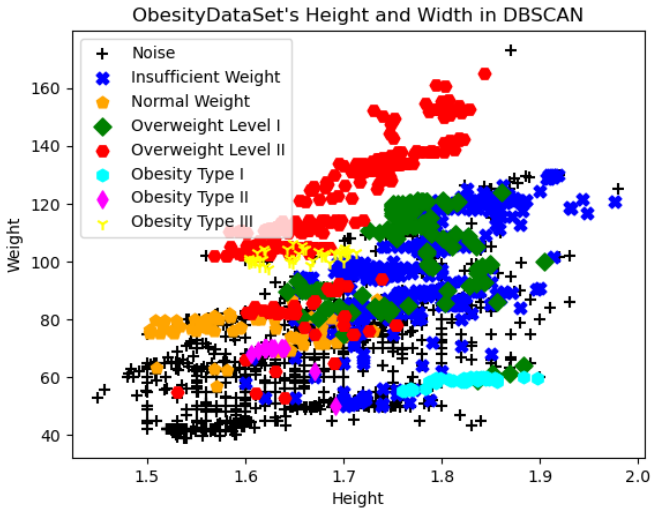

圖 7 DBSCAN 下 Estimation of obesity levels 資料集的分佈圖

四、 結論

根據結果所示，Iris 資料集的群聚分析中，DBSCAN 所花費的時間最少，再來是階層式分群，最後的 K-means 所花費的時間是最多的，在 purity 指標的衡量下，階層式分群是最好的，其次是 K-means，最後則是 DBSCAN，但三種演算法分群品質皆在 0.9 以上，可知分群品質優良。

Estimation of obesity levels 資料集的群聚分析中，DBSCAN 所花費的時間最少，再來是 K-means，最後的階層式分群所花費的時間是最多的，可見在 Estimation of obesity levels 資料集中，DBSCAN 的分群時間相較於其他兩種演算法，明顯所用時間較少。在 purity 指標的衡量下，階層式分群是最好的，其次是 K-means，最後則是 DBSCAN，三種演算法分群品質在 purity 皆在 0.4 至 0.5 左右，可見在 prurity 衡量下品質三者無太大差異；輪廓分析下只有 DBSCAN 的指標為負數，與其他兩種演算法相比較差。

參考文獻

(2019). *Iris Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/iris

(2019). *Estimation of obesity levels based on eating habits and physical condition Data Set*. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+#