國立雲林科技大學

資訊管理系碩士班

資料探勘

分群比較 Iris 與 Wine Quality 資料集

M11123028 林永沂

M11123017 洪國書

指 導 教 師 ：許 中 川

中 華 民 國 111 年 12 月

摘要

本研究蒐集分析了兩個資料集，首先，葡萄牙是世界前段的葡萄酒出口國，本次研究資料從 UCIDataSet 中蒐集葡萄酒品質資料集(Wine Quality Dataset)，其樣本來自葡萄牙西北部的紅色和白色「Vinho Verde」葡萄酒，資料集樣本屬性包括固定酸度、揮發性酸度、氯化物、二氧化硫以及 ph 值等。其次，蒐集了鳶尾花資料集(Iris Dataset)，此資料集包含來自三種鳶尾(Iris setosa、Iris virginica 及 Iris versicolor)各 50 個樣本，並從每個樣本測量四個特徵，分別是萼片長度、萼片寬度、花瓣長度以及花瓣寬度，以厘米為單位。本研究選用此兩個資料集，並採用 K-Means、DBSCAN 以及 Hierachiccal Clustering 三種群聚分析方法來比較分群結果。

Keyword：葡萄酒、葡萄酒品質、Vinho Verde、鳶尾

壹、緒論

1.1 研究背景

Wine Quality Dataset

葡萄酒是由新鮮葡萄果實或葡萄汁，經過發酵釀製而成的酒精飲料。而葡萄酒又有許多分類方式，常見的顏色分類，可分為紅葡萄酒、白葡萄酒及粉紅葡萄酒，以釀造方式來分類大致可分為平靜葡萄酒、氣泡葡萄酒、加烈葡萄酒及加味葡萄酒等。

葡萄酒的風味取決於釀製葡萄的品種，不同品種的果實所釀製出來的香味、喝的方式、收藏的方式都不同。而不同的地理位置、氣候、土質以及水質都有其適合栽培的葡萄品種，其中全球最大的三個產地為義大利、法國及西班牙。而葡萄酒也廣泛滲透至人們日常生活各種文化及行業領域，在歐洲地區尤其受歡迎，其選擇多樣、價格親民，成為家家戶戶每日必喝飲品，而觀光產業也因葡萄酒而衍伸了一系列的旅遊形式。

葡萄牙是世界前段的葡萄酒出口國，在世界葡萄酒產量排名在第十一位(糧農組織，2022)，本次研究資料從 UCIDataSet 中蒐集葡萄酒品質資料集(Wine Quality Dataset)，其樣本來自葡萄牙西北部的紅色和白色「Vinho Verde」葡萄酒，資料集樣本屬性包括固定酸度、揮發性酸度、氯化物、二氧化硫以及 ph 值等。

Iris Dataset

鳶尾屬(Iris)，是一類開花植物，屬於鳶尾科，其下包含 260-300 個種，因鳶尾花極其多樣的色彩，使得它在園藝中是十分受歡迎的花卉種類。

本次研究資料從 UCIDataSet 中蒐集鳶尾花資料集(Iris Dataset)，此資料集包含來自三種鳶尾(Iris setosa、Iris virginica 及 Iris versicolor)各 50 個樣本，並從每個樣本測量四個特徵，分別是萼片長度、萼片寬度、花瓣長度以及花瓣寬度，以厘米為單位。

1.2 研究目的

Wine Quality Dataset

針對葡萄牙葡萄酒的品質分析，從 UCIDataSets 的葡萄酒品質資料集蒐集資料，並利用 python 來分析該數據，而數據樣本與葡萄牙的西北部生產的「Vinho Verde」葡萄酒之紅色與白色變體有關，資料內容主要為物理化學(輸入)與感官(輸出)，而由於隱私與物流問題，此資料集沒有納入葡萄品種、品牌、售價等數據。

Iris Dataset

使用著名的鳶尾花資料集進行分析，此資料集數據樣本包含了三種鳶尾亞屬，分別是山鳶尾(Iris setosa)、維吉尼亞鳶尾(Iris virginica)、及變色鳶尾(Iris versicolor)，每個樣本皆包含四種特徵，分別是萼片長度、萼片寬度、花瓣長度以及花瓣寬度，並以厘米為單位。

最後，透過對以上兩個數據集進行群聚分析，來檢測分類優質與劣質的葡萄酒特性，與不同鳶尾品種的特徵，並比較不同分群方法之差異。

貳、資料集

2.1 真實資料集

本實驗所選擇的資料分別是 Iris Dataset、Wine Quality Dataset。

Wine Quality Dataset

Wine Quality Dataset 有 4898 筆資料、12 個欄位，分別是 fixed acidity、volatile acidity、citric acid、residual sugar、 chlorides、free sulfur dioxide、total sulfur dioxide、density、pH、sulphates、alcohol、quality ，詳見圖一。

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

圖 1 Wine Quality DataSet 資料表

Iris Dataset

Iris 有 150 筆資料、6 個欄位，分別是 ID、SepalLengthcm、SepalWidthCm、PetalLengthCm、PetalWidthCm、Species。

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

圖 2 Iris DataSet 資料表

參、方法

2.1 研究架構

本研究的研究架構如圖 3，將數據集的資料依序步驟進行統整、分析、決策樹建置、驗證、結論。並試著將葡萄酒品質數據集與鳶尾資料集內的資料利用 Python 將資料進行分群，並得出階層式分群之階層樹，得到的結論將有助於日後研究葡萄酒品質、鳶尾花種類變異的數據參考。

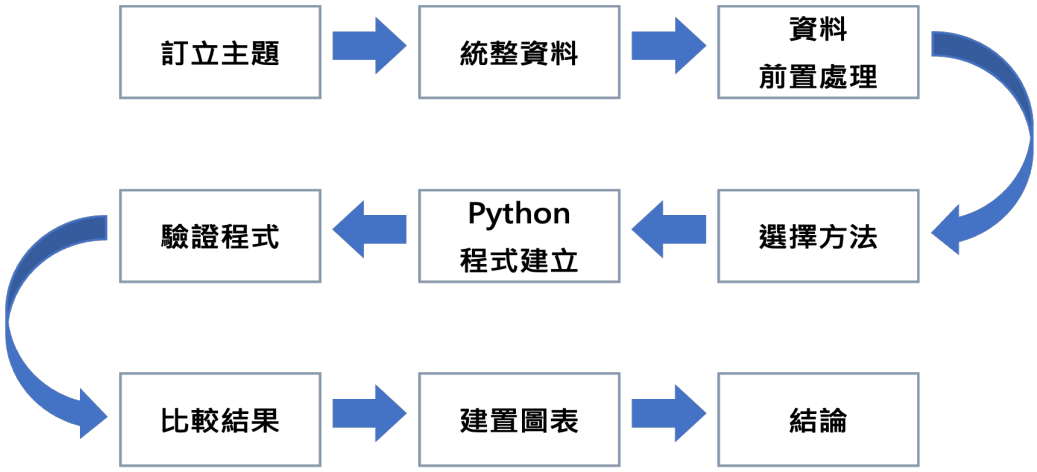

圖 3 研究架構

2.2 執行程式的方法

(1) 選擇資料集

從 UCI Dataset 中挑選資料集，並將挑選出來的變數簡化為目標資料集。

(2) 前置處理

針對資料的完整性去除不符合規則或不一致的資料，例如"空值"。

(3) 變換

將資料的輸入格式轉換成符合後續資料探勘步驟的格式。

(4) 群聚分析

資料經過轉換後，藉由 K-Means、階層式分群以及 DBSCAN 方法，來分析出資料集中的資料分布，並將其圖像化。

(5) 方法比較

透過以上方法分析後，經由統計及其他技術來確認其結果，設立衡量指標以利於比較不同群聚分析之品質。

肆、實驗

3.1 實驗設計

Wine Quality Dataset

在這項實驗中，本研究將嘗試找到最佳的群聚分析方法，以最有效和穩定的方式對不同的葡萄酒進行分組。為此，本實驗將利用 K-Means、Hierachiccal Clustering、DBSCAN 進行分析，並且列出 Hierachiccal Clustering 中的階層數，最後本實驗使用 Purity 與 Silhouette 來比較分群的品質與結果。

Iris Dataset

本實驗需要數據可視化和數據建模，所以使用 seaborn 和 python 的 sklearn 庫 matplotlib 來做題。首先 Import 本實驗會使用到的 Python 函式，把 Excel 裡面的欄位名稱輸入進去並且列出表格，接著從資料集刪除 Species 等欄位，本實驗將利用 K-Means、Hierachiccal Clustering、DBSCAN 來進行群聚分析，然後再列出 Hierachiccal Clustering 中的階層數，最後本實驗利用 Purity 來比較各個分群法的分群結果。

3.3 實驗結果

Wine Quality Dataset

透過上述實驗方式，使用 K-Means 分群法透過 Purity 得到的結果是 8.88%，而 Silhouette 結果是 0.355。使用 DBSCAN 分群法透過 Purity 得到的結果是 14.32%，Silhouette 則是 0.367，最後是 Hierachiccal Clustering 分群法，透過 Purity 得到 42.59%，Silhousette 則是 0.355，綜合評估下 Hierachiccal Clustering 的結果與花費時間是三個分群法中最佳的，如表 1 所示。

	Purity	Silhousette
K-Means	8.88%	0.355
DBSCAN	14.32%	0.367
Hierachiccal Clustering	42.59%	0.355

表 1 Wine Quality Dataset 三種分群法比對表格

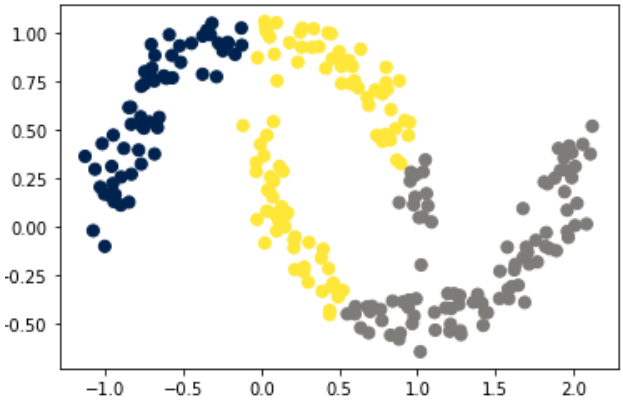

圖 4 K-Means

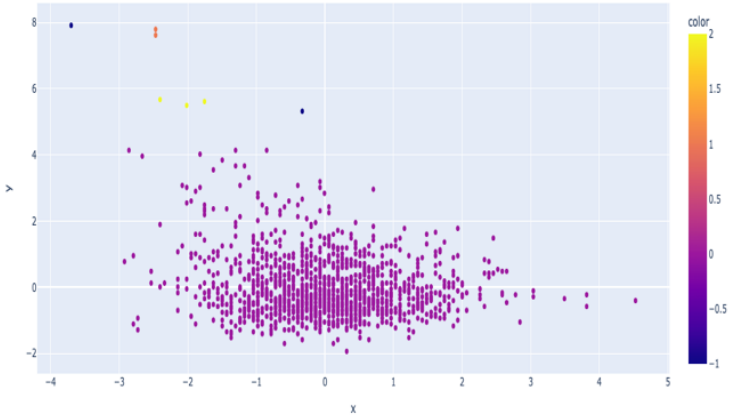

圖 5 DBSCAN

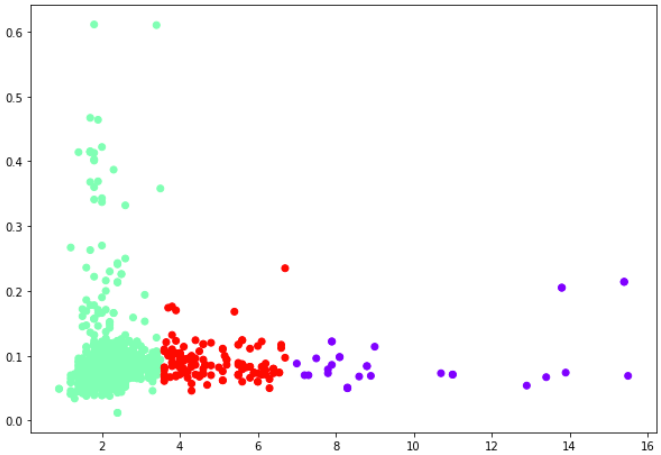

圖 6 Hierachiccal Clustering

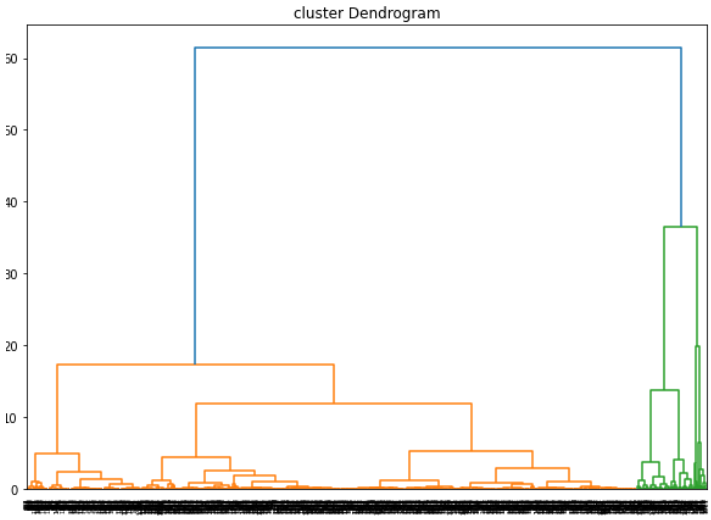

圖 7 Wine Quality Dataset Dendrogram

Iris Dataset

透過以上實驗方式，使用 K-Means 所得出的 Purity 指標是 89.33%，使用 Hierachiccal Clustering 所得出的 Purity 指標是 100%，使用 DBSCAN 所得出的 Purity 指標是 100%，綜合以上數據，Hierachiccal Clustering 與 DBSCAN 的結果是最佳的。

	Purity
K-Means	89.33%
Hierachiccal Clustering	100%
DBSCAN	100%

表 2 Iris Dataset 三種分群結果

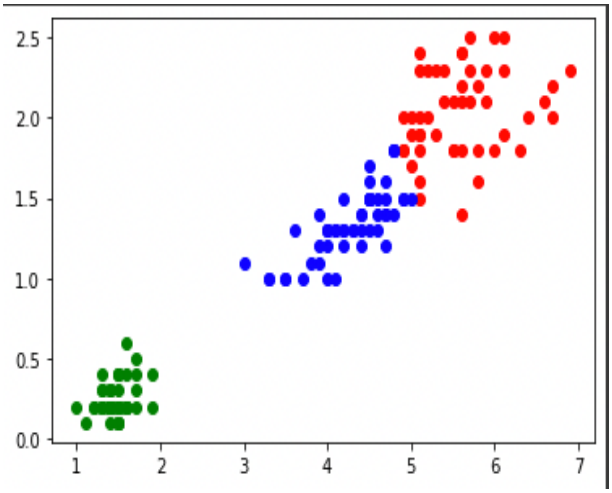

圖 8 K-Means

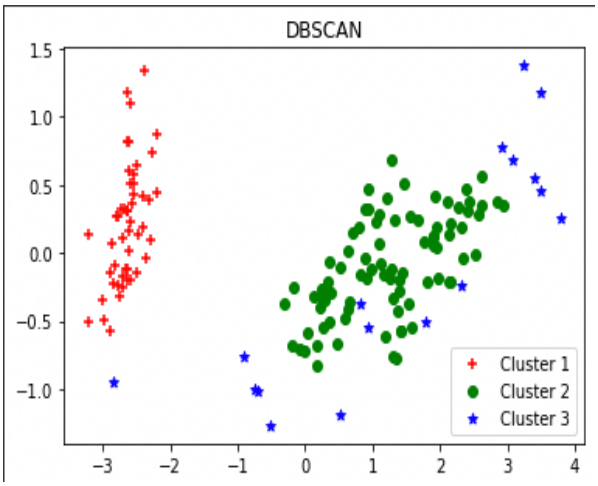

圖 9 DBSCAN

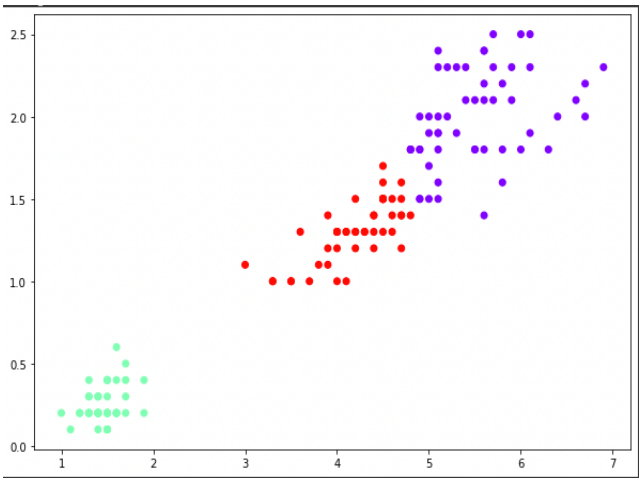

圖 10 Hierachiccal Clustering

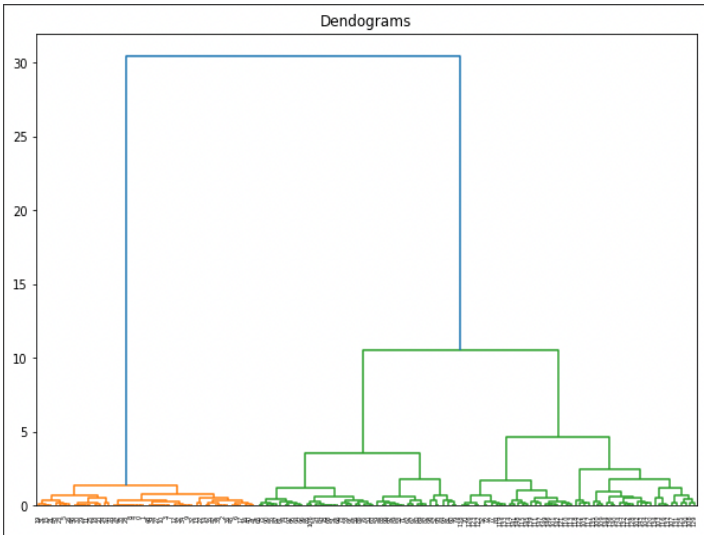

圖 11 Iris Dataset Dendrogram

肆、結論

本研究對 Iris Dataset、Wine Quality Dataset 進行不同的分群法進行了檢查展示了它們的細節，並且比較了三種分群法的群聚分析，且利用 Purity、Silhousette 得出各分群結果的指標，Wine Quality Dataset 在實驗上可能發生了一些錯誤導致各個衡量指標分數都在水準之下，未來本研究將會在針對程式碼進行修改以達到水準之上。

參考文獻

聯合國糧食及農業組織(糧農組織)2022 年 12 月
https://www.fao.org/faostat/en/#data/QCL