國立雲林科技大學  
資訊管理研究所

資料探勘專案作業3

指導教授：許中川教授  
學生：M11123039 廖顯億  
 M11123046 周子懷  
 M11123052 賴俊佑  
 M11123062 陳靖穎

Data Mining Homework: The Iris Dataset Mining

摘要

植物的花辨識相當重要，是分辨該植物的重要觀察之外，透過訓練可以協助生物判別分析發現資料集裡的花種分布。本研究以鳶尾花的資料集進行K-means、階層式分群、DBSCAN群聚分析。發現階層式分群跟DBSCAN的速度最快，K-means的純度最高。

**關鍵字**：生物判別分析、鳶尾花、K-means、階層式分群、DBSCAN

1. **緒論**

植物的辨識相當重要，尤其如何辨識該植物屬於哪一個生物分類，是藉由觀察植物的生殖器官，也就是植物的花，觀察植物的花來辨識該植物的品種，也可以分辨出該植物是否為新種，或是暫時歸類在哪一個生物分類底下。

* 1. **動機**

鳶尾花的花瓣明顯而且花朵分類簡單，尤其又是雌雄同株的特性，只需要一朵花就可以辨識該植物。辨識上可以透過花萼長度、寬度，花瓣長度、寬度等四個特徵，利用Fisher（1936）所建立的資料集進行花種分群。

* 1. **目的**

透過K-means、階層式分群、DBSCAN等三種演算法進行分群。可以藉由訓練出來的模型，套用在日後採集分類上，希望可以快速且準確地進行生物判別分析，辨識分類出資料集裡的花種類別。

1. **方法**

匯入資料後進行前處理，將資料進行正規化。接著定義純度，這邊使用相依矩陣表示。其資料屬性分為花萼長度、寬度，花瓣長度、寬度及對應到的鳶尾花學名。本研究預計使用K-means、階層式分群、DBSCAN，其中階層式分群包含Ward’s method、Median-linkage、Complete-linkage並對三種演算法進行比較分群所花費的時間並用純度指標衡量分群品質。

1. 實驗
2. 資料集

* 資料集名稱：Iris
* 資料筆數：150
* 屬性數量：5

表1 Iris資料集屬性簡介

|  |  |  |
| --- | --- | --- |
| 屬性名稱 | 型態 | 尺度 |
| sepal length | Int | 比例 |
| sepal width | Int | 比例 |
| petal length | Int | 比例 |
| petal width | Int | 比例 |
| class | String | 名目 |

1. 前置處理

本研究前處理步驟為將資料進行標準化，統一度量。

1. 實驗設計

一開始先進行預處理，統一度量，並且利用K-means、階層式分群、DBSCAN，其中階層式分群包含Ward’s method、Median-linkage、Complete-linkage並對三種演算法進行比較分群所花費的時間並用純度指標衡量分群品質。

1. 實驗結果

表2 各分群分所花費時間

|  |  |
| --- | --- |
| 演算法 | 花費時間（秒） |
| K-means | .141 |
| Ward’s method  Median-linkage  Complete-linkage | .001  .001  .001 |
| DBSCAN | .001 |

表3 各分群純度

|  |  |
| --- | --- |
| 演算法 | 純度 |
| K-means | .967 |
| Ward’s method  Median-linkage  Complete-linkage | .960  .953  .840 |
| DBSCAN | .740 |

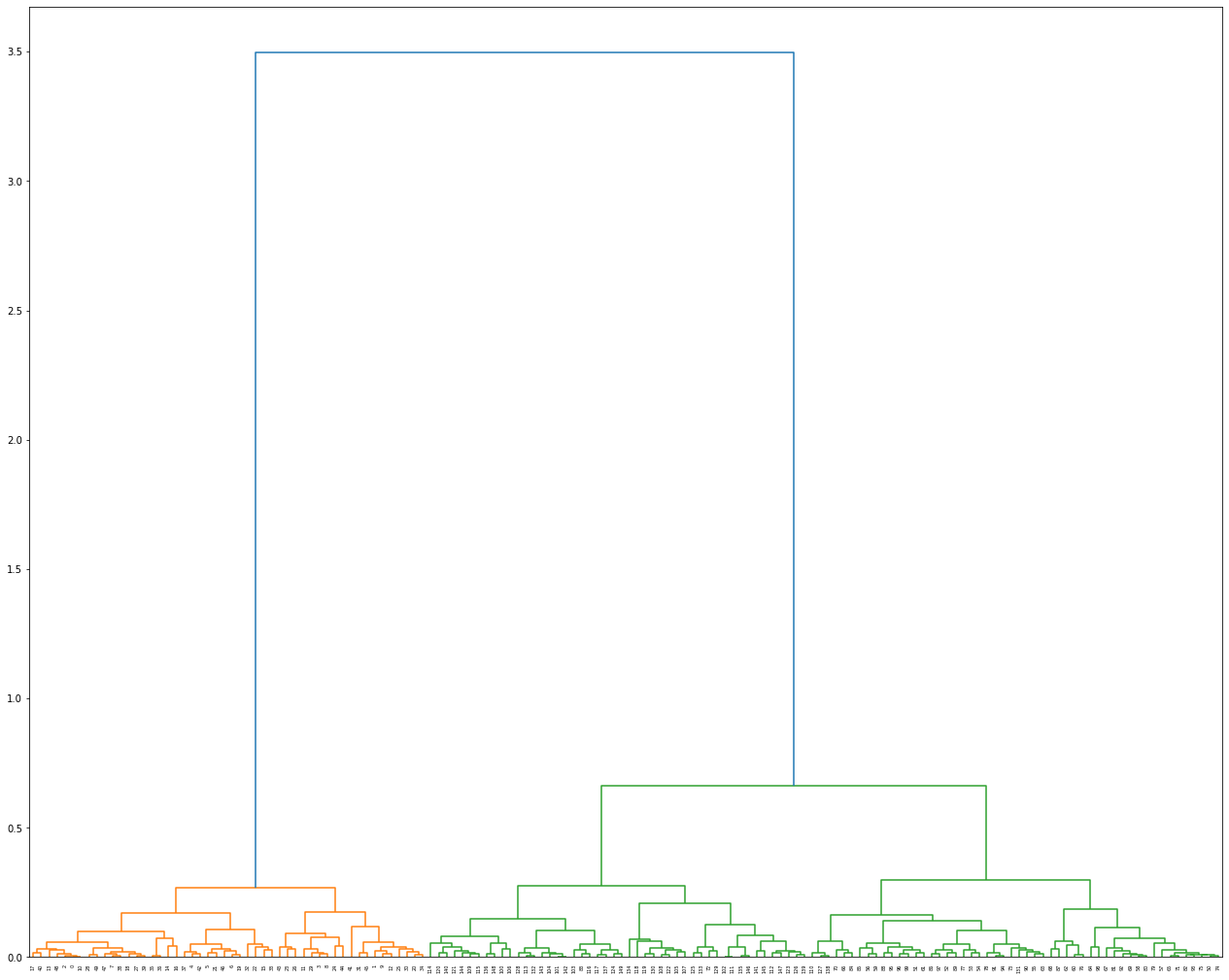


圖1 Dendrogram Using Ward’s method

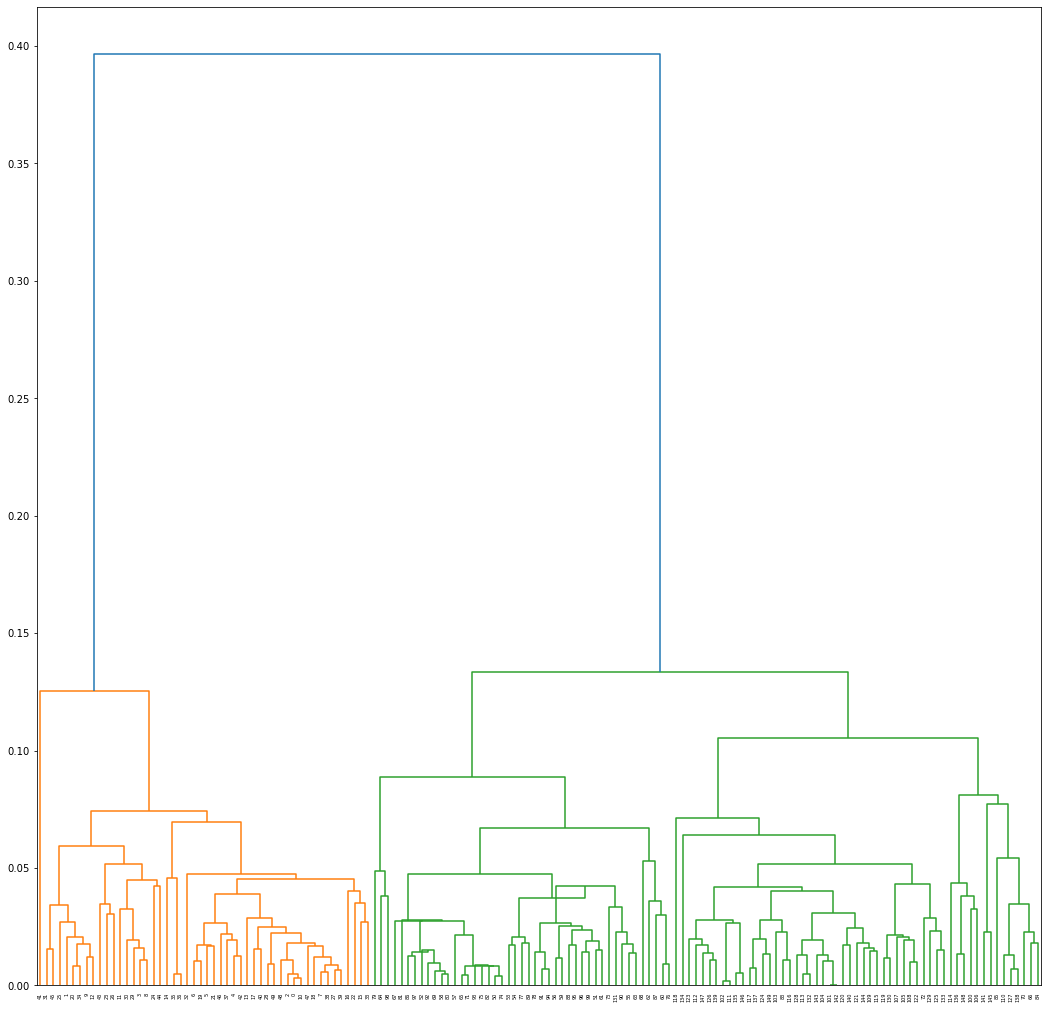


圖2 Dendrogram Using Median-linkage

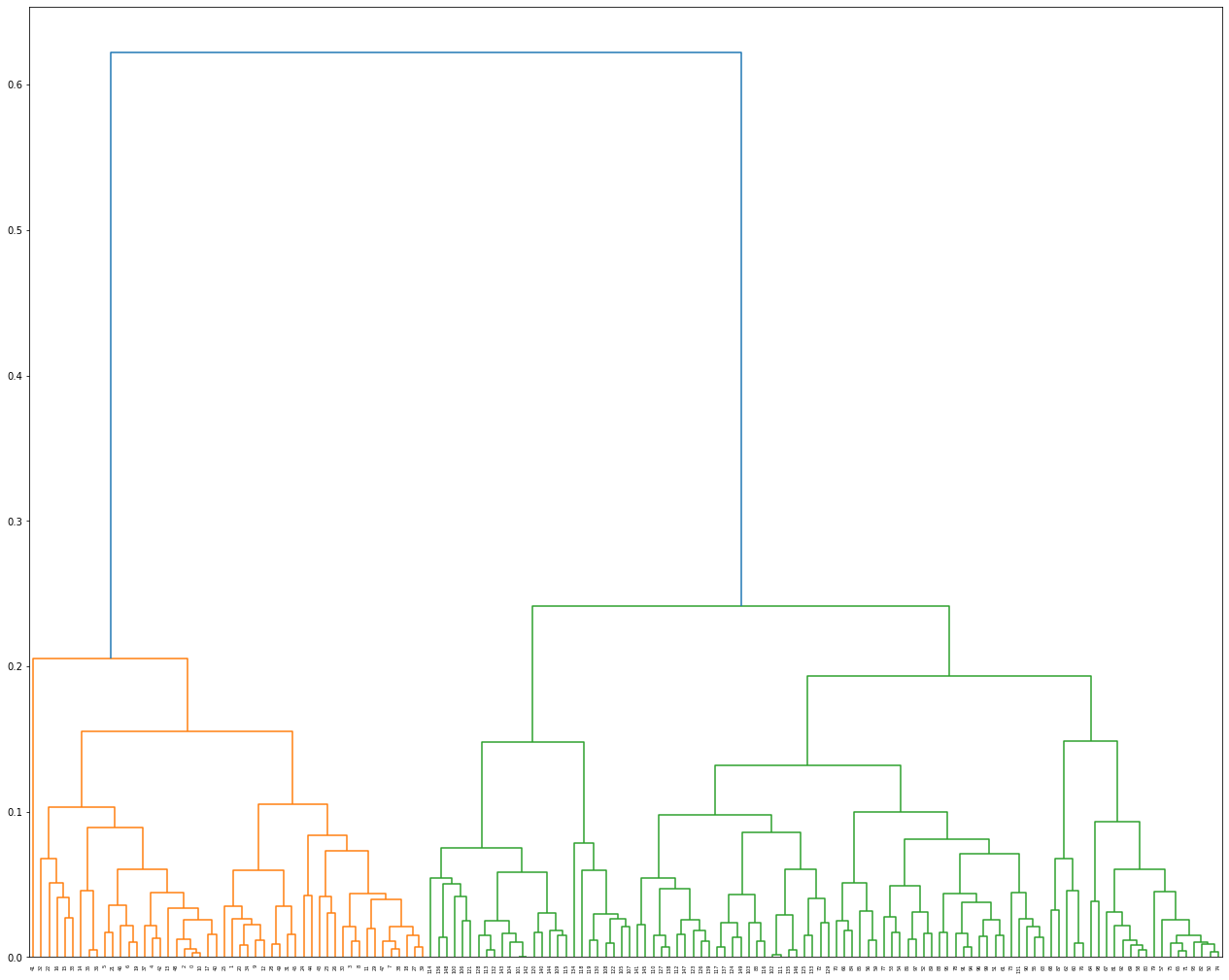


圖3 Dendrogram Using Complete-linkage

1. 結論

透過演算法結果可以發現K-means演算法所花費的時間最多，階層式分群跟DBSCAN所花費的時間最少，但是透過純度可以看到K-means的純度最高。可以依據現況選擇不同演算法，來達到分析後目的。

參考文獻

Agnes（2019年12月26日）。K-means Clustering 實戰篇：Iris data。WordPress.com。<https://artsdatascience.wordpress.com/2019/12/26/k-means-clustering-%E5%AF%A6%E6%88%B0%E7%AF%87%EF%BC%9Airis-data/>。

Fisher,R.A.. (1988). Iris. UCI Machine Learning Repository.。<https://archive-beta.ics.uci.edu/dataset/53/iris>。

imehrdadmahdavi（2019年08月01號）。iris-hierarchical-clustering。GitHub。<https://github.com/imehrdadmahdavi/iris-hierarchical-clustering/blob/master/iris-hierarchical-clustering.ipynb>。

PyInvest（2020年07月17日）。[Python實作] 密度聚類。PyInvest。<https://pyecontech.com/2020/07/17/python_dbscan/>。

Data Mining Homework: Shill Bidding Dataset Mining

摘要

網路購物詐騙事件頻傳，如何有效避免網路詐騙在現在這個社會顯得越來越重要。本研究以eBay上流行產品拍賣的資料集進行K-means、階層式分群與DBSCAN做分群分析。結果發現K-means的分群速度最快，且利用CH分數衡量分群品質時品質最高，但若是使用輪廓係數衡量品質時則為DBSCAN品質最好。

**關鍵字**：線上網路購物、網路購物詐騙、K-means、階層式分群、DBSCAN

1. **緒論**

現在這個資訊科技發達的時代，越來越多人從線下實體店面購物轉往線上網路購物，網路購物方便也不會受到時間限制，從早到晚甚至是凌晨都能在網路上下訂單買商品，網路購物的商機可以說是越來越大。

* 1. **動機**

隨著網際網路的蓬勃發展，連帶帶動的線上網路購物的商機，但當商機越來越大，伴隨而來的就是不法人士的覬覦，近年來消費者在網路購物時遇到詐騙的事件頻傳，因此本研究利用資料集分析詐欺的數據。

* 1. **目的**

透過K-means、階層式分群與DBSCAN三種演算法進行分群。可以透過訓練出來的模型，套用在日後網路詐騙的事件上，希望能夠簡單明瞭的查看拍賣數據並幫助消費者辨識詐騙行為避免受騙。

1. **方法**

匯入資料後進行預處理，將資料進行正規化、降維。接著利用輪廓係數與CH分數指標衡量分群的品質。本研究預計使用K-means、階層式分群與DBSCAN三種演算法，其中階層式分群採用single link，比較三種演算法分群所花費的時間並用輪廓係數以及CH分數指標衡量分群品質。

1. 實驗
2. 資料集

* 資料集名稱：Shill Bidding
* 資料筆數：6321
* 屬性數量：13

表 1 Shill Bidding資料集屬性簡介

|  |  |  |
| --- | --- | --- |
| 屬性名稱 | 型態 | 尺度 |
| Record ID | Int | 名目 |
| Auction ID | Int | 名目 |
| Bidder ID | Int | 名目 |
| Bidder Tendency | Int | 名目 |
| Bidding Ratio | Int | 名目 |
| Successive Outbidding | Int | 名目 |
| Last Bidding | Int | 名目 |
| Auction Bids | Int | 名目 |
| Auction Starting Price | Int | 名目 |
| Early Bidding | Int | 名目 |
| Winning Ratio | Int | 名目 |
| Auction Duration | Int | 名目 |
| Class | Boolean | 名目 |

1. 前置處理

本研究前處理為刪除ID特徵與類別，接著進行正規化，並將資料降維。

1. 實驗設計

一開始先進行預處理，將資料降維，利用K-means、階層式分群與DBSCAN將資料分群，對三種演算法分群所花費的時間進行比較並用輪廓係數以及CH分數衡量分群品質。

1. 實驗結果

表2 各分群分所花費時間

|  |  |
| --- | --- |
| 演算法 | 花費時間（秒） |
| K-means | 0.096 |
| 階層式分群 | 1.102 |
| DBSCAN | 0.112 |

表 3各分群輪廓係數

|  |  |
| --- | --- |
| 演算法 | 輪廓係數 |
| K-means | .454 |
| 階層式分群 | .322 |
| DBSCAN | .490 |

表 4各分群CH分數

|  |  |
| --- | --- |
| 演算法 | CH分數 |
| K-means | 5968.606 |
| 階層式分群 | 3583.354 |
| DBSCAN | 139.287 |

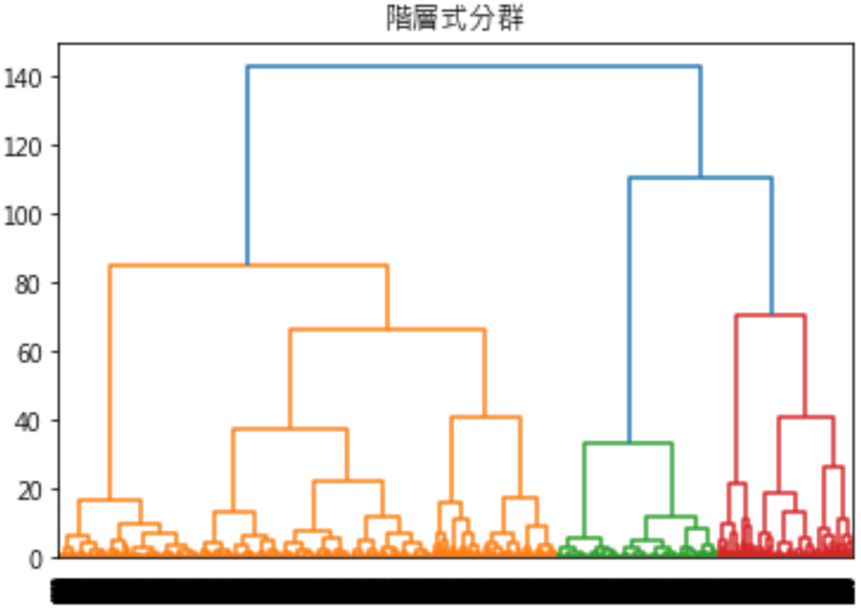


圖1 階層式分群之階層樹

1. 結論

透過訓練結果可以發現K-means演算法分群花費的時間最少，階層式分群則花費最多的時間。利用輪廓係數衡量分群品質為DBSCAN品質最高，但若利用CH分數衡量分群品質則是K-means品質最好。

參考文獻

Ahmad Alzahrani and Samira Sadaoui (2020)。Shill Bidding Dataset Data Set。UCI Machine Learning Repository。

<https://archive.ics.uci.edu/ml/datasets/Shill+Bidding+Dataset>

LoveMIss-Y (2019, July 3). 基於sklearn的聚類算法的聚類效果指標. 台部落.

<https://www.twblogs.net/a/5d1bbc22bd9eee1ede05c0d2>

PyInvest (2020, May 27). [Python實作] 聚類分析 K-Means / K-Medoids. PyInvest.

<https://pyecontech.com/2020/05/27/python_k-means_k-medoids/>

PyInvest (2020, June 15). [Python實作] 層次聚類 Hierarchical Clustering. PyInvest.

<https://pyecontech.com/2020/06/15/python_hierarchical_clustering/>

PyInvest (2020, July 15). [機器學習首部曲] 密度聚類 DBSCAN. PyInvest.

<https://pyecontech.com/2020/07/15/dbscan/>

陳元裕（2018年7月6日）。機器學習-特徵工程-降維。TAROBALLZ。

<http://www.taroballz.com/2018/07/06/ML_DecreaseFeature/>