國立雲林科技大學資訊管理系

資料探勘

專案作業三

組員

M11123037 許禾諭

M11123043 謝永盛

M11123050 李京樺

M11123053 葉家豪

**摘要**

Iris是典型的花朵資料集，將山鳶尾、變色鳶尾和維吉尼亞鳶尾三類進行標示，利用花瓣與花萼的長寬資料，協助喜愛鳶尾花的賞花人士辨別種類，本研究採用K-means、Hierarchical clustering和DBSCAN進行分群預測，計算分群花費時間與純度(Purity)並比較三種分群方法，結果顯示DBSCAN執行時間最快，而K-means與Hierarchical clustering純度相同且優於DBSCAN。

最新數據顯示，2013-16年國內成人平均肥胖率為45.4%，逐年攀升，2016-19年國內成人肥胖率來到47.9%，創下新高(衛福部國健署，2021)。由此可知肥胖的問題已成為多數台灣人民都會遇到的問題。因此本研究選用Estimation of obesity levels based on eating habits and physical condition Data Set收集相關數據，以分析何種行為可能會導致肥胖，以減少國人肥胖的情況。本研究採用K-means、Hierarchical clustering和DBSCAN三種不同的分群演算法來訓練模型，並透過Purity以及V-measure來評估績效，結果顯示DBSCAN執行時間最快，而Hierarchical clustering的純度和V-measure分群指標是最高的。

關鍵字: K-means、Hierarchical clustering、DBSCAN、鳶尾花、肥胖程度

1. 緒論
2. 動機

由於鳶尾花的多樣性，使剛接觸鳶尾花的賞花人士不易分辨鳶尾花，這種情況令人感到挫折，故本研究使用Iris資料集，利用花瓣與花萼的資料進行分群，期望能協助賞花人士辨別鳶尾花。

近年來由於生活繁忙、飲食不均衡以及缺乏定期的運動等可能的原因導致臺灣的肥胖比例逐年上升，由此可知肥胖已成為國人健康的一大隱患。故本研究希望藉由Estimation of obesity levels based on eating habits and physical condition 資料集來分析是何種生活因素是導致肥胖的因素。

1. 目的

本研究計畫利用Iris資料集將不同的鳶尾花特徵進行分群，例如:花辮長度、寬度等等來對同種不同屬的鳶尾花進行分群，並將分群結果提供給對於鳶尾花有興趣的賞花人士，以方便他們辨識。

本研究計畫利用Estimation of obesity levels based on eating habits and physical condition 資料集來對不同的生活因素進行分群，例如: 每天的水消耗量。來對可能導致的生活因素進行分群。並將結果提供給有肥胖隱憂的民眾，以期減少國人的肥胖比例，提升國人的健康。

1. 資料集
2. Iris Data Set

表 1

Iris資料集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set Characteristics: | Multivariate | Number of Instances: | 150 | Area: | Life |
| Attribute Characteristics: | Real | Number of Attributes: | 4 | Date Donated | 1988-07-01 |
| Associated Tasks: | Classification | Missing Values? | No | Number of Web Hits: | 5044495 |

表 2

Iris資料集-屬性資料表

|  |  |
| --- | --- |
| 屬性名稱 | 屬性資訊 |
| sepal length(萼片長度) | 單位:公分(cm) |
| sepal width(萼片寬度) | 單位:公分(cm) |
| petal length(花瓣長度) | 單位:公分(cm) |
| petal width(花瓣寬度) | 單位:公分(cm) |
| class | Iris Setosa、Iris Versicolour、Iris Virginica |

1. Estimation of obesity levels based on eating habits and physical condition Data Set

表 3

基於飲食習慣和身體狀況的肥胖水平估計資料集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set Characteristics: | Multivariate | Number of Instances: | 2111 | Area: | Life |
| Attribute Characteristics: | Integer | Number of Attributes: | 2111 | Date Donated | 2019-08-27 |
| Associated Tasks: | Classification, Regression, Clustering | Missing Values? | N/A | Number of Web Hits: | 104120 |

表 4

基於飲食習慣和身體狀況的肥胖水平估計資料集-屬性資料表

|  |  |
| --- | --- |
| 屬性名稱 | 屬性資訊 |
| Gender | 性別 |
| Age | 年齡 |
| Height | 身高 |
| Weight | 體重 |
| family\_history\_with\_overweight | 有超重的家族史 |
| FAVC | 經常食用高熱量食物 |
| FCVC | 食用蔬菜的頻率 |
| NCP | 正餐次數 |
| CAEC | 兩餐之間的食物消耗量 |
| SMOKE | 吸菸 |
| 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 |
| NObeyesdad | 體重不足、正常體重、超重 I 級、超重 II 級、I 型肥胖、II 型肥胖和 III 型肥胖 |

1. 方法
2. 實作說明

首先載入所需的函示庫，再進行資料的前處理，處理需正規化的欄位，分別是有順序性和無順序性的欄位，做不同的正規化，並且將數值控制在小數點第二位，接下來分別以KMeans、階層式分群、DBSCAN，三種演算法訓練模型，並且計算出模型執行時間和分群指標等方式評估模型。

1. 操作說明

下載GitHub上的資料集以及程式，使用Jupyter開發環境讀取檔案的路徑，確定檔案可成功讀取後，即可執行程式。

1. 實驗
2. 前置處理

在自選集的部分，本研究在Jupyter開發環境下匯入自選集的csv檔，將 family\_history\_with\_overweigh、FAVC、CAEC、SMOKE、SCC、CALC、MTRANS和 NObeyesdad，透過 LabelEncoder將資料集內的文字轉換成數值（如表5）），此外注意到性別沒有順序性，所以做另一種的正規化，透過dummies將原本Gender的欄位變成Gender\_Female和Gender\_Male兩個欄位，有出現數值1的欄位對應到的屬性名稱代表這筆資料的性別是男或女，因為資料集原先的數值有的到小數點第二位，有的到小數點第五位，為了統一數值，根據Height到小數點第二位為原則，將所有到小數點第五位的數值，全部四捨五入到小數點第二位。

表 5

基於飲食習慣和身體狀況的肥胖水平估計資料集-LabelEncoder的前處理

|  |  |
| --- | --- |
| 屬性名稱 | 數值 |
| family\_history\_with\_overweigh | 0、1 |
| FAVC | 0、1 |
| CAEC | 0、1、2、3 |
| SMOKE | 0、1 |
| SCC | 0、1 |
| CALC | 0、1、2、3 |
| MTRANS | 0、1、2、3、4 |
| NObeyesdad | 0、1、2、3、4、5、6 |

1. 實驗設計
2. Iris Data Set

使用Scikit-learn套件將資料集載入程式中，本研究採用K-MEAN、Hierar-chical clustering和DBSCAN三種分群演算法，計算分群所花費的時間，以及使用Purity指標衡量分群品質。

1. Estimation of obesity levels based on eating habits and physical condition Data Set

資料前置處理後，本研究將使用K-MEAN、Hierarchical clustering和DBS-CAN三種不同的分群演算法來訓練模型，並且透過purity和V-measure兩個績效指標來計算出不同演算法訓練的模型績效，並且比較了不同分群算法的計算時間。

1. 實驗結果
2. Iris Data Set
3. K-means

Iris資料集使用K-means分成三群，執行時間約0.05100秒。

分群結果與真實資料計算Purity，純度為0.89333。

1. Hierarchical clustering

Iris資料集使用階層式分群分成三群，執行時間約0.01049秒。

分群結果與真實資料計算Purity，純度為0.89333。

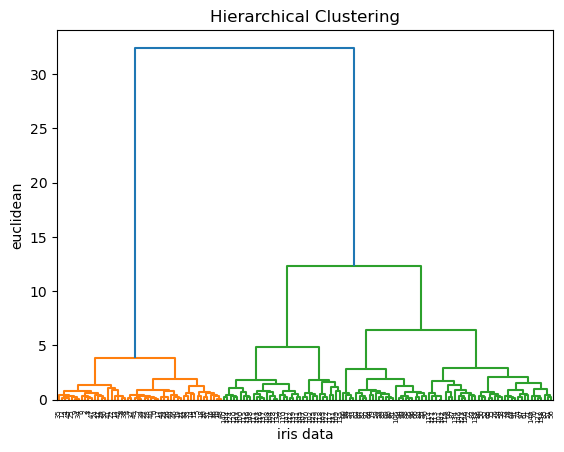


圖 1階層式分群的階層樹(Dendrogram)

1. DBSCAN

Iris資料集使用DBSCAN分成三群，執行時間約0.00212秒。

分群結果與真實資料計算Purity，純度為0.53333。

1. Estimation of obesity levels based on eating habits and physical condition Data Set
2. K-means

Estimation of obesity levels based on eating habits and physical condition資料集使用K-means分成七群，執行時間約0.06157秒。

分群結果計算Purity，純度為0.52534。

分群結果計算V-measure為0.48099。

1. Hierarchical clustering

Estimation of obesity levels based on eating habits and physical condition資料集使用階層式分群分成七群，執行時間約0.09124秒。

分群結果計算Purity，純度為0.53103。

分群結果計算V-measure為0.51460。

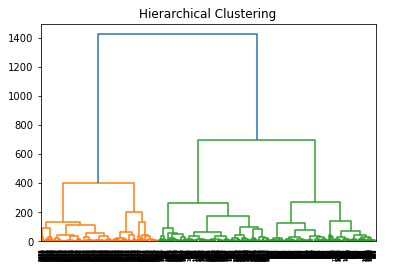


圖 2階層式分群的階層樹(Dendrogram)

1. DBSCAN

Estimation of obesity levels based on eating habits and physical condition資料集使用 DBSCAN分成七群，執行時間約0.02237秒。

分群結果計算Purity，純度為0.2198。

分群結果計算V-measure為0.09719。

1. 結論

本實驗可以得知在K-means、階層式分群、DBSCAN這些分群演算法中，以鳶尾花資料集（Iris Data Set）來說，根據Purity分群指標的結果得知，使用K-means分群演算法優於另外兩種分群演算法，但以肥胖程度資料集（physical condition Data Set）來說，使用 Hierarchical clustering 分群演算法，相較於另外兩種分群演算法的分群指標 Purity 、V-measure，結果顯示都是最好的，而 DBSCAN的執行時間不管在哪個資料集都是三者分群演算法中運算最快的。使用不同的資料集也會影響測試結果，像使用鳶尾花資料集（Iris Data Set）所測試出的結果，不管是執行時間或是績效指標都會優於肥胖程度資料集（physical condition Data Set）的測試結果，我們推測是因為鳶尾花資料集本身的資料整理得較為完整，才能導致執行時間和績效指標較優。

1. 參考文獻
2. 邱芷柔。2021年7月30日。又胖了！ 台灣壯年男10人有6人過重。自由健康網。<https://health.ltn.com.tw/article/paper/1463808>
3. Tonykuoyj。2016年12月24日。[第 24 天] 機器學習（4）分群演算法。iT邦幫忙。<https://ithelp.ithome.com.tw/articles/10187314>
4. Yeh James。2017年9月22日。[資料分析&機器學習] 第2.1講： 如何獲取資料？ Sklearn內建資料集。medium。<https://reurl.cc/bG3vDX>
5. Muhammad Maisam Abbas。2021年3月21日。在 Python 中計算所經過的時間。DelftStack。<https://www.delftstack.com/zh-tw/howto/python/python-elapsed-time/>
6. 阿新。2020年12月31日。聚類效果的外部評價指標——純度(Purity)及其Python和matlab實現。程式人生。<https://www.796t.com/article.php?id=210746>
7. PyInvest。2020年6月15日。[Python實作] 層次聚類 Hierarchical Clustering。PyInvest。<https://pyecontech.com/2020/06/15/python_hierarchical_clustering/>
8. PyInvest。2020年7月17日。[Python實作] 密度聚類 DBSCAN。PyInvest。<https://pyecontech.com/2020/07/17/python_dbscan/>
9. Scikit-learn。 Homogeneity, completeness。<https://scikit-learn.org/stable/modules/clustering.html>
10. 2020年1月14日。Python學習筆記-Pandas更改欄位名稱。 Kuo's 3C 筆記。<http://icekuo.blogspot.com/2020/01/python-pandas_14.html?view=flipcard>
11. Yeh James。2017年10月10日。 [資料分析&機器學習] 第2.4講：資料前處理(Missing data, One-hot encoding, Feature Scaling)。Medium。<https://reurl.cc/eWM7a7>