國立雲林科技大學資訊管理系

資料探勘-作業三

Department of Information Management

National Yunlin University of Science & Technology

Assignment

使用聚類分析Banana資料集與sizes3資料集之績效

M11223018黃裕鳴

M11223020游棨翔

M11223043葉桔良

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

中華民國112年12月

December 2023

**摘要**

本研究旨在探討在當今數據科學環境中，不同聚類分析模型在處理真實資料集時的表現及其受超參數影響的情況。我們選用了包括K-means、階層式分群、以及DBSCAN等多種常見聚類模型，並在真實資料集中進行深入實驗與評估。

研究動機主要體現在面對當前大數據時代，聚類分析作為資料探勘的重要手段，其應用需求與挑戰持續擴大。我們關注不同聚類模型在應對真實資料的特性時的優勢和限制，並著重於超參數對模型性能的影響。透過實驗結果的比較，我們提供了針對不同模型的超參數設定建議，以優化聚類效果。

研究結果顯示，在處理banana和sizes3等真實資料集時，階層式分群模型表現出色，其優越的Accuracy和低entropy值使其成為這些資料集的優選模型。此外，我們強調了超參數對聚類模型性能的關鍵影響，提供了精細調整超參數的實用建議。

這項研究不僅有助於深入理解不同聚類模型在實際應用中的優勢，也為資料科學家提供了更具體的指導，以更有效地應對多樣性和複雜性的真實資料分析挑戰。

**關鍵字**：聚類分析、K-means、階層式分群、DBSCAN

1. **緒論**

在現代資料科學領域中，聚類分析是一種重要的技術，可用於發現資料中的隱藏模式和結構，進而提供對資料集的深入理解。隨著科技的進步，各種聚類演算法應運而生，以處理不同類型的資料，解決各種應用場景的挑戰。

1. **動機**

研究動機背後的主要考量源自於當前資訊時代中，資料科學與機器學習的蓬勃發展。這種快速發展不僅使得數據的產生與累積呈現爆炸性的增長，同時也凸顯了在處理這些大規模複雜資料時的挑戰。在此背景下，我們選擇聚類分析作為一種有力的資料探勘工具，旨在深入了解其在真實資料集上的應用效果。

此外，現實世界中的資料往往呈現出多樣性且複雜的特性，因此，不同聚類模型在面對這些真實挑戰時可能呈現出截然不同的表現。透過本研究，我們追求探索各種聚類模型在處理真實資料時的適應性，以期提供資料科學家更全面的選擇視角。

1. **目的**

近年來，隨著資料科學與機器學習的發展，聚類分析作為一種無監督學習的重要技術，扮演了發現資料潛在結構和模式的關鍵角色。本研究旨在深入研究三種主要的聚類模型（K-means、階層式分群、DBSCAN）在真實資料集上的應用情境，透過評估它們的性能指標，包括SSE、Accuracy和Entropy，以提供對不同聚類模型適用性的全面了解。

1. **資料集**
2. **真實資料集**
3. banana資料集介紹

一個人工資料集，資料在平面圖上的型態類似兩根香蕉，常用於做聚類分析。有兩個屬性x和y以及一個類別標籤（1和2），共有4812筆資料。

**表1**

banana資料集欄位介紹

|  |  |  |
| --- | --- | --- |
| **x** | **y** | **class** |
| 0.228 | 0.559 | 1 |
| 0.216 | 0.528 | 1 |
| 0.221 | 0.552 | 1 |
| 0.215 | 0.538 | 1 |
| 0.224 | 0.548 | 1 |
| 0.232 | 0.553 | 1 |

1. sizes3資料集介紹

一個人工資料集。有兩個屬性x和y以及一個類別標籤（1、2、3和4），共有1001筆資料。

**表2**

sizes3資料集欄位介紹

|  |  |  |
| --- | --- | --- |
| **x** | **y** | **class** |
| 9.28531 | 14.5879 | 0 |
| 12.3977 | 8.4489 | 0 |
| 8.73624 | 11.6803 | 0 |
| 9.10197 | 12.8327 | 0 |
| 12.2948 | 8.67373 | 0 |
| 11.027 | 10.6794 | 0 |

1. **方法**
2. **實作說明**

在分群分析模型的評估實驗中，本研究首先將「banana資料集」、「size3資料集」做數據的前置處理，其包括非結構化資料的轉換，並使用 K-means、階層式分群、 DBSCAN，三種不同的分群模型，以SSE指標衡量K-means模型，以Entropy與Accuracy指標衡量所有模型的分群品質，並計算在不同資料集中在每種分群演算法所花費時間。

1. **操作說明**

本研究執行環境皆為Python3.11.5，並且使用 Visual Studio Code與 Jupyter Notebook作為開發工具，利用 pandas、 sklearn來讀取資料以及做資料的前處理，預測模型利用 sklearn、 scipy套件來建構，顯示圖型的部分使用seaborn，最後再將測試資料導入並預測數值。

1. **實驗**
2. **前置處理**

**圖1**

banana前置處理流程圖

1. 檢測異常值：使用四分位距的方式，檢查是否有離群值。此資料集無異常值，故不做處理。
2. min-max Normalization：也稱為離差標準化，是對原始數據的線性變換，使得結果映射到0-1之間。因資料集圖形呈現方式不是常態分佈，故使用min-max Normalization。

**圖2**

sizes3前置處理流程圖

1. 檢測異常值：使用四分位距的方式，檢查是否有離群值。將資料集內異常值刪除。
2. Z-score Normalization：也稱為標準化值，各個樣本在經過轉換後，通常在正、負五到六之間不等。因資料集圖形呈現方式類似常態分佈，故使用Z-score Normalization。
3. **實驗設計**

實驗設計的K-means、階層式分群、DBSCAN三種方式來衡量banana資料集與size3資料集的SSE、Accuracy、Entropy指標。以下顯示本研究的實驗設計。

**4.2.1. 衡量指標流程步驟**

下圖(圖3)展示banana資料集與sizes3資料集在多種不同演算法應用上的流程步驟，需衡量不同演算法的指標。以下將解釋不同演算法之超參數數值設定說明。

**圖3**

流程圖

載入資料

訓練模型

衡量指標

比較差異

K-means

階層式分群

DBSCAN

**4.2.2. Banana資料集設定超參數之說明**

1. K-means超參數設定方式

將n\_clusters設定為2，數據將分為2個簇。同時，將n\_init設定為15，這表示演算法將運行15次，每次使用不同的初始中心點，以獲得最優的結果。使用init='k-means++'，這表示本研究使用智能的初始化方法，有助於加速算法的收斂。最後，設定max\_iter為50，這表示算法運行的最大迭代次數為50。

1. 階層式分群超參數設定方式

將n\_clusters設定為2，將數據分為2個簇。同時，將linkage設定為single，該方法將兩個群體中最近的點之間的距離作為群體之間的距離。

1. DBSCAN超參數設定方式

將eps設定為0.01953999321339954，表示將兩個鄰居間的最大距離。將min\_samples設定為10，表示指定一個樣本被視為核心點所需的最小鄰居數為10。由於無法得知何為 min\_samples的最佳值，且無法控制分群數量，因此將 min\_samples參數由4調整至10，並挑選出分群數量符合的最低值為10。

**4.2.3. Size3資料集設定超參數之說明**

1. K-means超參數設定方式

將n\_clusters設定為4，數據將分為4個簇。同時，將n\_init設定為15，這表示演算法將運行15次，每次使用不同的初始中心點，以獲得最優的結果。使用init='k-means++'，這表示本研究使用智能的初始化方法，有助於加速算法的收斂。最後，設定max\_iter為50，這表示算法運行的最大迭代次數為50。

1. 階層式分群超參數設定方式

將n\_clusters設定為4，將數據將分為4個簇。同時，將linkage設定為average，該方法計算兩個群體中所有點之間的平均距離作為群體之間的距離。

1. DBSCAN超參數設定方式

將eps設定為0.34832935391621983，表示將兩個鄰居間的最大距離。將min\_samples設定為8，表示指定一個樣本被視為核心點所需的最小鄰居數為8。由於無法得知何為 min\_samples的最佳值，且無法控制分群數量，因此將 min\_samples參數由4調整至10，並挑選出分群數量符合的最低值為8。

1. **實驗結果**

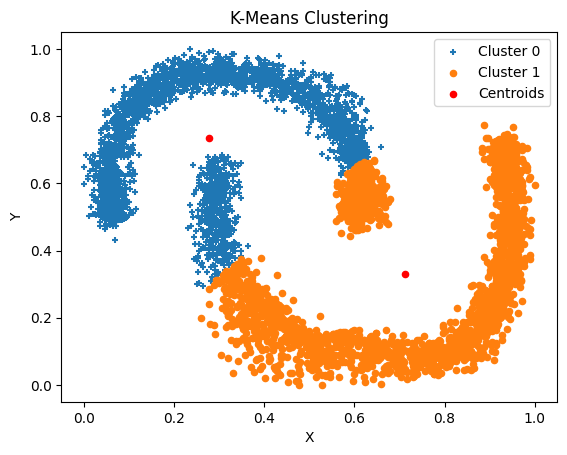
**4.3.1. banana資料集實驗結果**

1. K-means實驗結果

K-means分群模型的計算時間為0.034秒。相應的性能指標包括SSE為351.569、Accuracy為0.814，以及entropy為0.69。如圖4所示，分成兩個聚類，紅色的點為中心點。

**圖4**

banana資料集使用K-means做聚類分析結果圖

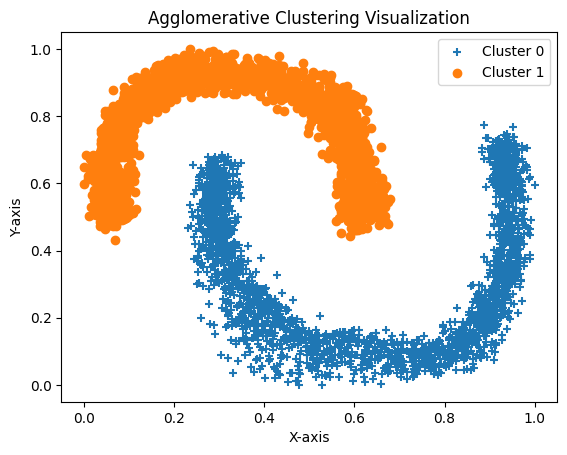


1. 階層式分群實驗結果

階層式分群模型的計算時間為0.111秒。相應的性能指標包括Accuracy為1.0，以及entropy為0。如圖5所示，分成兩個聚類，可見此模型為Banana資料集的最佳解。圖6為分群後階層圖，若階層分群的層數太高，則無法清楚顯示，故選擇適當的4層當作最後的視覺化結果。

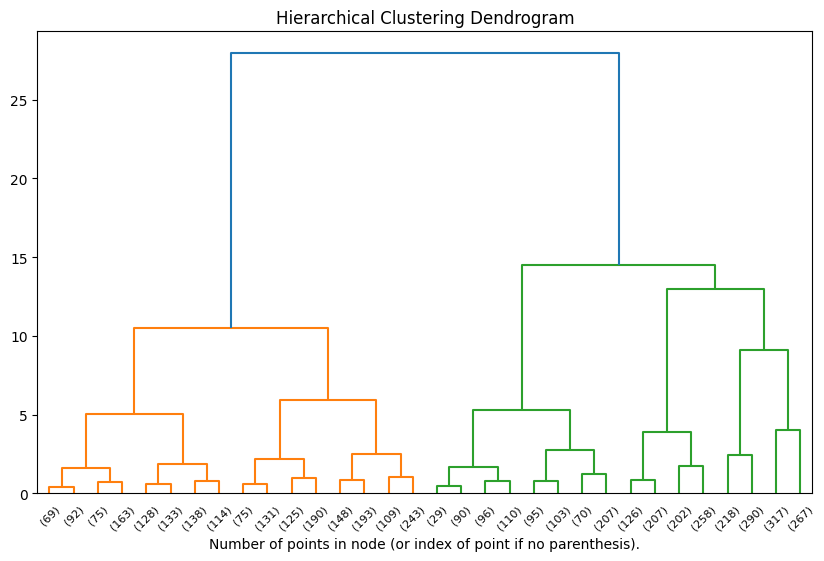
**圖5**

banana資料集使用階層式分群做聚類分析結果圖

****

**圖6**

banana資料集分群後階層圖

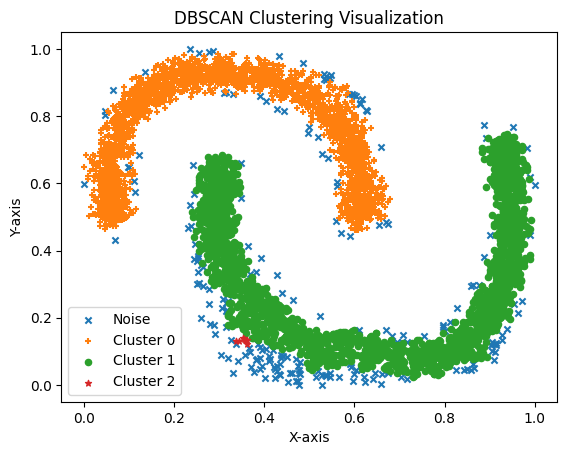
****

1. DBSCAN實驗結果

DBSCAN模型的計算時間為0.044秒。相應的性能指標包括Accuracy為0.954，以及entropy為0.035。如圖7所示，分成四個聚類，其中有兩個聚類為雜訊。

**圖7**

banana資料集使用DBSCAN做聚類分析

****

1. banana資料集不同模型衡量指標比較

如表1，可觀察到運行時間上無太大區別，但有發現階層式分群的準確率高達1，為此資料集最佳解。

**表1**

banana資料集不同模型衡量指標比較表

|  |  |  |  |
| --- | --- | --- | --- |
|  | **K-means** | **階層式分群** | **DBSCAN** |
| SSE | 351.569 |  |  |
| Accuracy | 0.814 | **1** | 0.954 |
| entropy | 0.69 | **0** | 0.035 |
| time(s) | **0.034** | 0.111 | 0.044 |

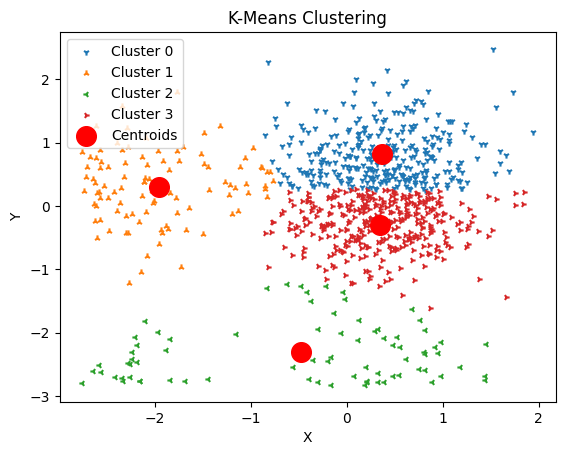
**4.3.2. sizes3資料集實驗結果**

1. K-means實驗結果

K-means分群模型的計算時間為0.066秒。相應的性能指標包括SSE為488.252、Accuracy為0.542，以及entropy為0.21。如圖8所示，分成四個聚類，紅色的點為中心點。

**圖8**

sizes3資料集使用K-means做聚類分析結果圖

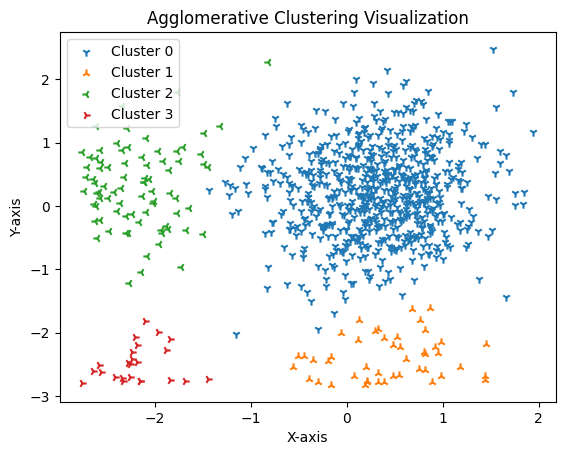


1. 階層式分群實驗結果

階層式分群模型的計算時間為0.023秒。相應的性能指標包括Accuracy為0.987，以及entropy為0.109。如圖9所示，分成四個聚類。圖10為分群後階層圖，若階層分群的層數太高，則無法清楚顯示，故選擇適當的4層當作最後的視覺化結果。

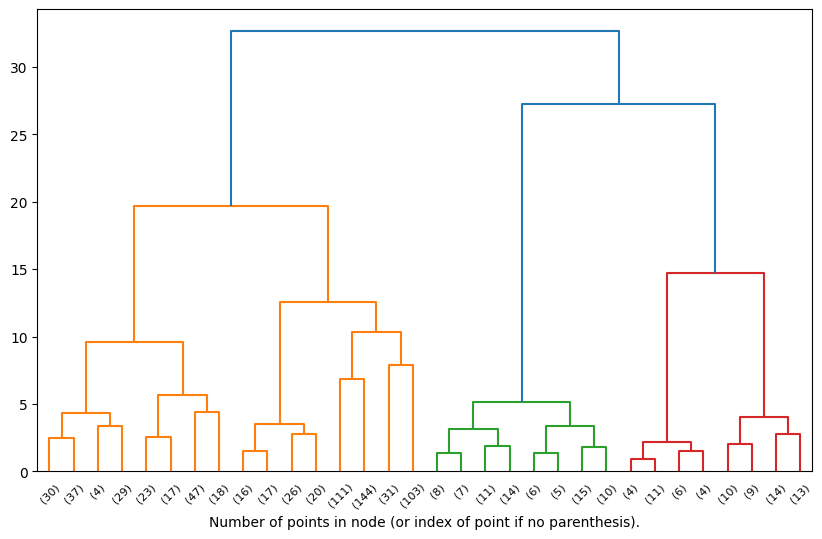
**圖9**

Sizes3資料集使用階層式分群做聚類分析結果圖

****

**圖10**

Sizes3資料集分群後階層圖

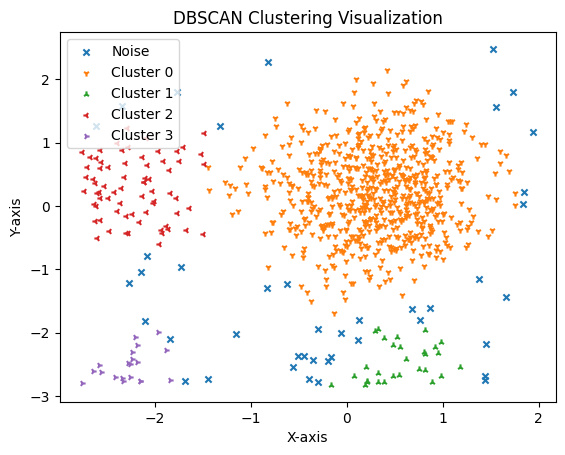
****

1. DBSCAN實驗結果

DBSCAN模型的計算時間為0.018秒。相應的性能指標包括Accuracy為0.942，以及entropy為0.147。如圖11所示，分成四種聚類，與一種雜訊。

**圖7**

sizes3資料集使用DBSCAN做聚類分析

****

1. sizes3資料集不同模型衡量指標比較

如表1，可觀察到運行時間上無太大區別，但有發現階層式分群的準確率高達0.987，為此資料集最佳解。

**表1**

banana資料集不同模型衡量指標比較表

|  |  |  |  |
| --- | --- | --- | --- |
|  | **K-means** | **階層式分群** | **DBSCAN** |
| SSE | 488.252 |  |  |
| Accuracy | 0.542 | **0.987** | 0.942 |
| entropy | 0.21 | **0.109** | 0.147 |
| time(s) | **0.066** | 0.023 | 0.018 |

1. **結論**

本研究針對不同的聚類分析模型在處理真實資料集時的性能進行了深入研究，包括K-means、階層式分群和DBSCAN等。透過豐富的實驗與評估，我們得出以下結論：

首先，在banana資料集中，階層式分群模型展現出優越的性能，其高Accuracy和低entropy值使其成為最佳選擇。這凸顯了在特定資料分布情境下，階層式分群模型能夠更準確地捕捉資料的結構，相對於其他模型更具優勢。

其次，在sizes3資料集中，同樣發現階層式分群模型表現卓越，其高Accuracy和低entropy值進一步印證其在處理複雜資料集時的強大能力。K-means和DBSCAN在這兩個資料集上也呈現出合理的表現，但相對而言未能超越階層式分群。

進一步觀察超參數的調整對模型性能的影響，我們發現對於不同資料集，最佳的超參數設定存在差異。這凸顯了在實際應用中，選擇適當的超參數對於模型性能至關重要。

總體而言，本研究不僅提供了對多種聚類模型在真實資料集上性能的深刻理解，也為從業者提供了針對超參數調整的實用建議，以提升聚類分析的實際應用效果。未來的研究可以進一步擴展到更多的資料集和模型，以提供更全面的指導。

1. **參考文獻**

andy6804tw (2021年04月02日)。[機器學習] 離群值處理。https://andy6804tw.github.io/2021/04/02/python-outliers-clean/

Renesh Bedre (2023, Feburary 6). DBSCAN clustering algorithm in Python (with example dataset). https://www.reneshbedre.com/blog/dbscan-python.html