USING CLUSTERING MODEL TO PREDICT ONLINE SHILL BIDDING BEHAVIOR AND CATEGORIZE NEWS ARTICLE TOPICS

以分群模型預測網路競標行為以及分類新聞文章主題

林郁凱,彭冠穎,彭嘉瑋,陳羿欣
Department of Information Management, National Yunlin University of Science and Technology

摘要

Knowledge Discovery in Database（KDD）已廣泛應用於各個領域之中，現今網路商機無窮，購物詐欺行為常出現於網路競標之中，因此在本研究中以網路競標行為判斷作為實驗主題，以 Shill Bidding Dataset 公開資料集作 k-means、階層式分群、DBSCAN 三種方法之分群演算，計算不同分群演算法所花費的時間，以及使用 Purity 指標以衡量其分群的品質，為了比較在不同資料集中使用三種不同分群方法之績效，因此在本研究中另外使用 Mini 20 Newsgroups 作新聞文章分群實驗。在本研究實驗結果顯示，在 Mini 20 Newsgroups 資料集實驗結果，DBSCAN 計算時間最短，K-mean 的分群品質最佳，而在 Shill Bidding 資料集實驗結果中，K-mean 計算時間最短，而分群品質皆為一致。

關鍵字：KDD、K-means、階層式分群、DBSCAN、Purity。

一、緒論

1.1 動機

在訊息爆炸的大數據時代，Knowledge Discovery in Database (KDD)已經廣泛應用於商業行銷、製造、媒體等各領域，透過探索這些龐大的數據，可以從中找尋有用的資訊，或是判斷異常的資訊。

在行銷領域中，企業常利用蒐集得來的客戶消費相關數據來得知該客戶之消費行為，針對消費者習慣作不同的行銷策略，或是異常的消費行為，如購物詐欺等，現今網路商機無窮，購物詐欺行為常出現於網路競標之中，因此本研究使用 Shill Bidding Dataset 公開資料集，利用三種不同分群演算法（Clustering）來探討競標者之拍賣行為，是否為正常的競標行為。

為了比較在不同資料集，三種分群方法之績效，在本研究中另外使用一套小型的新聞文章資料集 Mini 20 Newsgroups，對二十種不同的新聞主題作分群演算，並比較不同分群演算法之分群品質。

1.2 目的

本研究使用 Mini 20 Newsgroups、Shill Bidding Dataset 資料集，以三種不同的分群演算法（Clustering），分別為 K-means、階層式分群、DBSCAN，三種不同之分群演算法，其演算方式各有不同，其運算時間也有不同，因此在本研究中除了將資料分群之外，將計算在不同資料集中在每種分群演算法所花費時間，以及使用 Purity 指標衡量其分群的品質。

二、資料集

2.1 真實資料集

本研究中使用 Mini 20 Newsgroups 以及 Shill Bidding Dataset 兩種資料集作為本研究的實驗資料，以下為兩種資料集之規格與說明。

2.1.1 新聞文章資料集（Mini 20 Newsgroups）

Mini 20 Newsgroups 資料集包含了休閒、談話、汽車、科學、宗教等等 20 個新聞主題，每一個新聞主題有 100 篇的文章[1]，文章為非結構化資料，在本研究中將會為此作結構化處理。

- 資料集名稱：Mini 20 Newsgroups 資料集
- 總資料筆數：2,000

2.1.2 網路競標資料集（Shill Bidding Dataset）

Shill Bidding 資料集為在 eBay 拍賣上之相關數據，如競標者趨勢、競標比率、拍賣持續時間、獲勝率等等，皆為數值資料[2]。

- 資料集名稱：Shill Bidding 資料集
- 總資料筆數：6321

三、方法

3.1 實作說明

在分群分析模型的評估實驗中，本研究首先將「新聞文章資料集（Mini 20 Newsgroups）」、「網路競標資料集（Shill Bidding Dataset）」做數據的前置處理，其包括非結構化資料的轉換，以及資料清理，並使用 K-means、階層式分群、DBSCAN，三種不同的分群模型，以 Purity 指標衡量每一個模型的分群品質，並計算在不同資料集中在每種分群演算法所花費時間。

3.2 操作說明

本研究執行環境皆為 Python3.6，並且使用 Jupyter Notebook 作為分析工具，利用 os、re、pandas、sensim、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，最後再將測試資料導入並預測數值。

四、實驗

4.1 前置處理

4.1.1 Mini 20 Newsgroups 資料集

圖 1 Mini 20 Newsgroups 資料集-前置處理流程圖

- 資料前處理：將資料做轉小寫、移除數字、移除符號、移除停用詞等處理。
- 文字轉向量：將文字轉為向量以便後續分群使用。

表 1 Mini 20 Newsgroups-資料前處理後部分資料

欄位名稱 / 資料編號	bir	bird	birds	birth	birthday	bis	bishop	bit	…	bits
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.186551	…	0.000000
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	…	0.000000
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	…	0.000000
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	…	0.000000
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.104231	…	0.154641

4.1.2Shill Bidding 資料集

圖 2 Shill Bidding 資料集-前置處理流程圖

- 資料分割：將資料之類別與欲用來訓練之資料分割。
- 資料清理：將資料中與預測結果不相關資訊過濾。

表 2 Shill Bidding 資料集-資料前處理後部分資料

欄位名稱 / 資料編號	Bidder_ Tendency	Bidding_ Ratio	Successive_ Outbidding	Last_ Bidding	Auction_ Bids	…
1	0.200000	0.400000	0.0	0.000028	0.0	…
2	0.024390	0.200000	0.0	0.013123	0.0	…
3	0.142857	0.200000	0.0	0.003042	0.0	…
4	0.100000	0.200000	0.0	0.097477	0.0	…
5	0.051282	0.222222	0.0	0.001318	0.0	…

4.2 實驗設計

以下為 Mini 20 Newsgroups 資料集與 Shill Bidding 資料集依序的流程步驟：

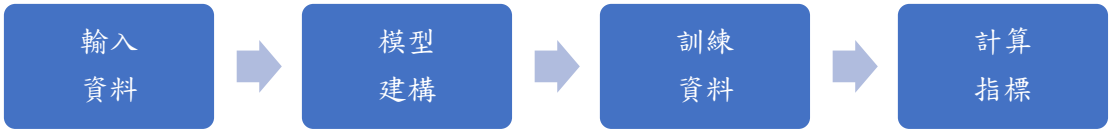

圖 3 實驗流程圖

4.2.1 Mini 20 Newsgroups 資料集

1. 設定模型初始參數
2. 將資料匯入模型中並進行訓練
3. 算出 Purity 分群品質衡量指標
4. 測試多次參數並採用最佳結果

4.2.2 Shill Bidding 資料集

1. 設定模型初始參數
2. 將資料匯入模型中並進行訓練
3. 算出 Purity 分群品質衡量指標
4. 測試多次參數並採用最佳結果

4.3 實驗結果

4.3.1 Mini 20 Newsgroups 資料集

4.3.1.1 K-means

將 K-means 分群模型參數 n_clusters 設置為題目所要求之群數 20，而 init 則設置為 k-means++，由於 K-means 當中隨機初始點會有不良的分類狀況，因此使用 k-means++來避免此狀況，而 k-means++的概念為初始分群中心之間的相互距離要盡可能的遠，績效指標結果 Purity 為 0.436 而 Times 為 15.407。

4.3.1.2 階層式分群

將階層式分群模型參數 n_clusters 設置為題目所要求之群數 20，在階層式分群當中的績效指標結果 Purity 為 0.328 而 Times 為 35.719。圖 4 為整體的階層樹，而圖 5 為分 群後之階層樹。

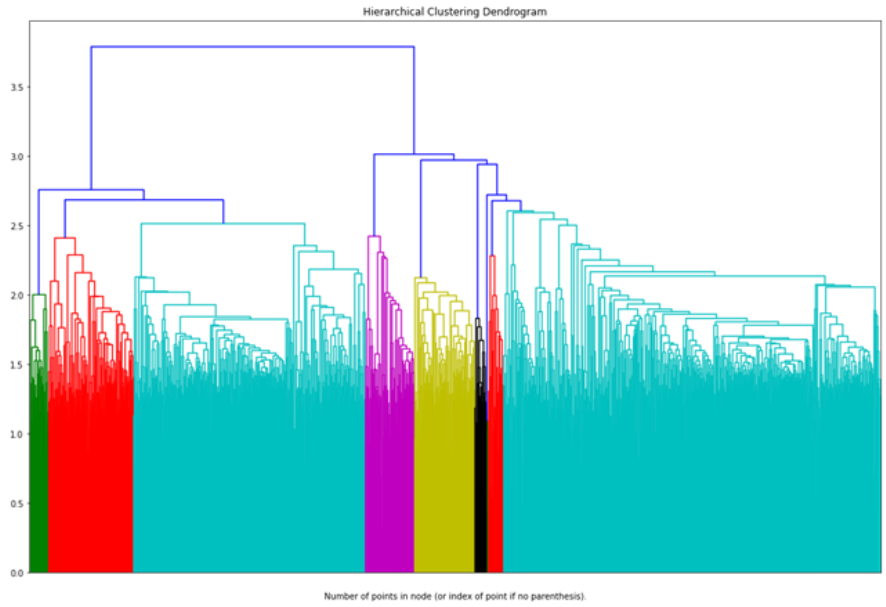

圖 4 Mini 20 Newsgroups 資料集-整體階層式分群之階層樹

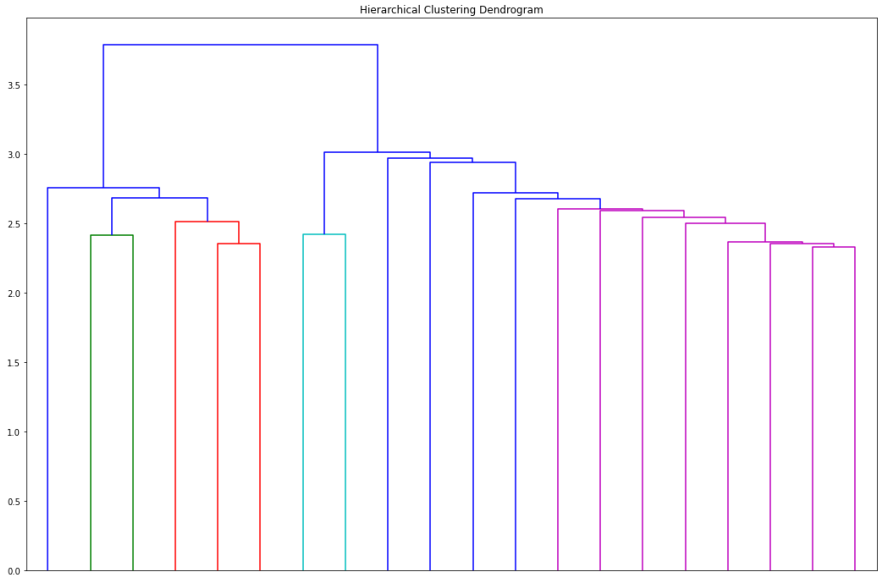

圖 5 Mini 20 Newsgroups 資料集-分群後階層式分群之階層樹

4.3.1.3 DBSCAN

將DBSCAN模型參數eps由0.05調整至1且每次調整0.005，而 min_samples 則由 2 調至 9 每次調整 1，將兩個參數交互配對後產出如表 3 所示，依據題目要求須將資料分為 20 群，分群後的分布為只有一群有三個資料點，其餘十九群則為兩個資料點，因此將此表當中為 20 群的結果印出如表 4，由表 4 之結果顯示績效指標結果 Purity 為 0.0655 而 Times 最低的為 0.1659。

表 3 Mini 20 Newsgroups 資料集-資料 0 到 4 筆之參數配對結果

欄位名稱 / 編號	eps	min_samples	n_clusters	outliners	purity	time
0	0.050	2	8	1984	0.054	0.223176
1	0.050	3	0	2000	0.050	0.188293
2	0.050	4	0	2000	0.050	0.169035
3	0.050	5	0	2000	0.050	0.171015
4	0.050	6	0	2000	0.050	0.169006

表 4 Mini 20 Newsgroups 資料集-n_clusters 為 20 之參數配對結果

欄位名稱 / 編號	eps	min_samples	n_clusters	outliners	purity	time
776	0.535	2	20	1959	0.0655	0.166945
784	0.540	2	20	1959	0.0655	0.177372
792	0.545	2	20	1959	0.0655	0.180813
800	0.550	2	20	1959	0.0655	0.165999
808	0.555	2	20	1959	0.0655	0.167972

4.3.2 Shill Bidding 資料集

4.3.2.1 K-means

將 K-means 分群模型參數 init 設置為 k-means++，由於 K-means 當中隨機初始點會有不良的分類狀況，因此使用 k-means++來避免此狀況，而 k-means++的概念為初始分群中心之間的相互距離要盡可能的遠，其績效指標結果 Purity 為 0.893 而 Times 為 0.0639。

4.3.2.2 階層式分群

將階層式分群模型參數 n_clusters 設置為該資料集之類別個數 2，其績效指標結果 Purity 為 0.893，Times 為 1.4944。圖 6 為整體的階層樹，而圖 7 為分群後之階層樹。

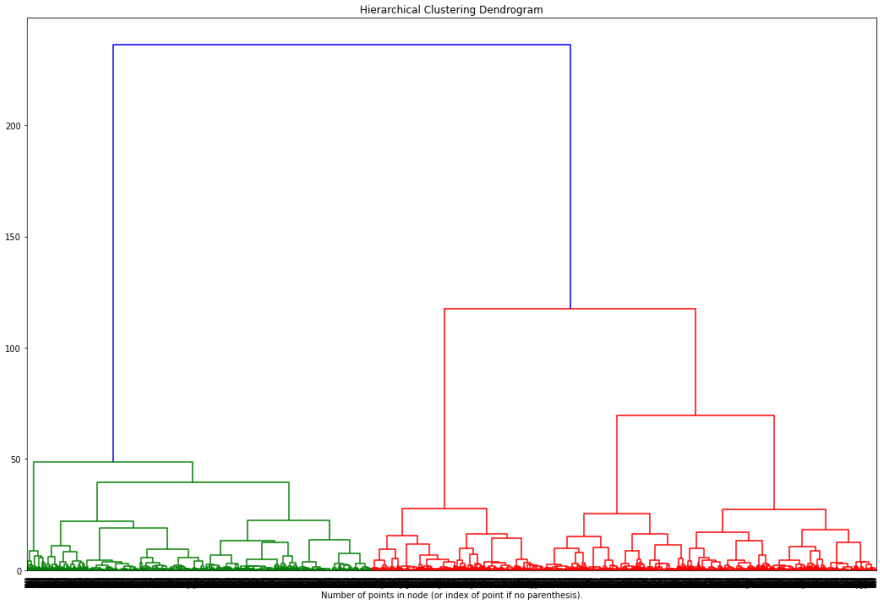

圖 6 Shill Bidding 資料集-整體階層式分群之階層樹

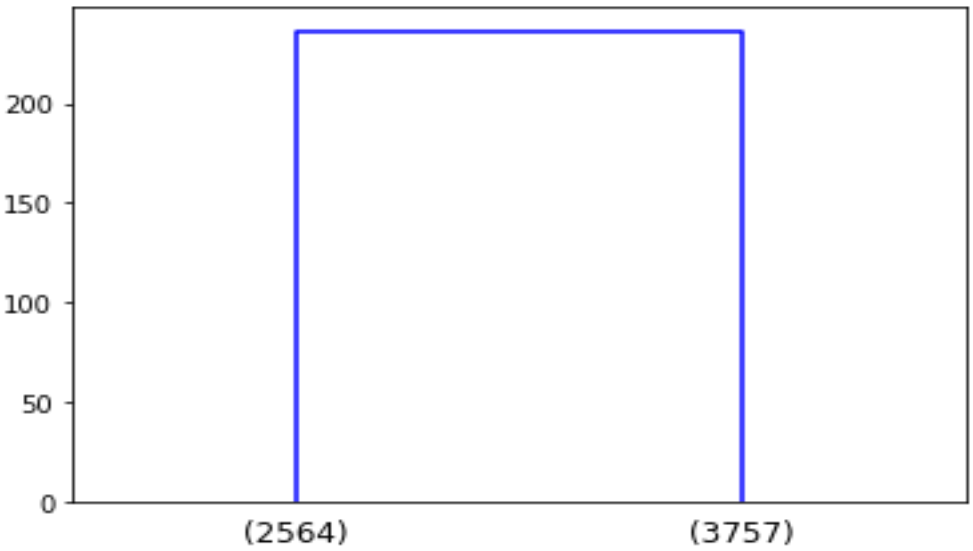

圖 7 Shill Bidding 資料集-分群後階層式分群之階層樹

4.3.2.3 DBSCAN

將 DBSCAN 模型參數 eps 由 1 調整至 2 且每次調整 0.05，而 min_samples 則由 5 調至 14 每次調整 1，將兩個參數交互配對後產出如表 5 所示，由表 5 之結果顯示最佳績效指標結果 Purity 為 0.8932，Times 為 0.5314。

表 5 Shill Bidding 資料集-資料 0 到 4 筆之參數配對結果

欄位名稱 / 編號	**eps**	**min_samples**	**n_clusters**	**outliners**	**purity**	**time**
0	1.00	5	5	0	0.893213	0.615398
1	1.00	6	5	0	0.893213	0.538424
2	1.00	7	5	0	0.893213	0.541450
3	1.00	8	5	2	0.893213	0.531460
4	1.00	9	5	2	0.893213	0.534454

五、結論

本研究使用 Mini 20 Newsgroups、Shill Bidding Dataset 兩種資料集，分別以 K-means、階層式分群、DBSCAN 三種不同演算進行分群分析，由實驗結果顯示於 Mini 20 Newsgroups 資料集中，績效指標 Purity 表現最好的為 K-means，其次為階層式分群，而在 Times 當中表現最好的為 DBSCAN，其次則為 K-means；於 Shill Bidding 資料集中，績效指標 Purity 在各個模型皆一致，但在時間方面表現最好的是 K-means，其次則為 DBSCAN。

在本研究實驗結果中，每個分群模型之演算時間 Times 會產生這樣的結果，是因為階層式分群法由樹狀結構的底部開始，將資料或群聚逐次合併，導致階層式分群運算時間較長。

在 Purity 方面，Shill Bidding 資料集中皆相同，本研究認為此結果是因為 Shill Bidding 資料集只有分為兩類，導致 Purity 指標一致；然而在 Mini 20 Newsgroups 資料集中資料分部較為分散，使得 DBSCAN 無法有較好的分群結果，多數都只有兩個資料點成群。

表 6 統整績效表

Mini 20 Newsgroups 資料集			
模型 / 指標	**K-means**	**階層式分群**	**DBSCAN**
Purity	0.436	0.328	0.0655
Times	15.407	35.719	0.1659
Shill Bidding 資料集			
模型 / 指標	**K-means**	**階層式分群**	**DBSCAN**
Purity	0.893	0.893	0.893
Times	0.0639	1.4944	0.5314

六、參考文獻

[1] Tom Mitchell. (1999) 20 Newsgroups [Dataset].

[2] Ahmad Alzahrani and Samira Sadaoui. (2020) Shill Bidding Dataset Data Set [Dataset].

[3] K-Means clustering. (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)

[4] Agglomerative Clustering. (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html)

[5] DBSCAN clustering. (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html)