國立雲林科技大學資訊管理系

機器學習

專案作業三

以 MDS、T-SNE 對高鐵經緯度資料與飲品資料作降維技術的評估

M10723047 黃靖媛

M10923012 陳羿欣

M10923021 吳青芬

M10923023 張凱淇

指導教授：許中川

2021 年 06 月

摘要

本研究目的利用 MDS 和 T-SNE 分別運用於兩種不同資料集，並評估降維後之結果，第一個資料集為高鐵經緯度資料，本研究想透過經緯度的位置，繪製高鐵站各站於二維平面上，查看各站之間的距離是否過於密集、是否能呈現出相對的距離關係。第二個資料為飲品資料，飲品的品項繁多，類型也不同，因此本研究想透過飲品資料集來了解各個不同種類的品項間，經由降維後分群的狀況，最終將資料視覺化於二維平面上，本研究實驗結果發現 MDS 在高鐵站經緯度資料中依然能呈現出每一高鐵站的距離關係；在飲品資料經由 T-SNE 降維處理後，僅有 Cappuccino 類別資料被分割為兩群，其他單一類別資料大致上都能有不錯的分群結果，另外本研究在該資料集的前處理中，分別使用了 1-of-K 與 Word2Vector 不同的文字編碼技術，觀察降維結果之差異，發現若將飲品類別分為汽水（Coke、Pepsi、Sprite、7up）以及咖啡（Latte、Espresso、Cappuccino），由 word2vec 二維平面圖可以發現較能呈現出兩種飲品的不同之處。

關鍵字： MDS、T-SNE、1-of-K、Word2Vector、降維。

一、緒論

1.1 動機

近年來，搭乘高鐵往返南北縣市越來越普遍，從原本板橋、台北、桃園、新竹、台中、嘉義、台南和左營站，後來又新增了南港、苗栗、彰化和雲林站，本研究想透過經緯度了解原本的高鐵站各站間的距離，查看各站間是否過於密集，因此在台中與台南之間的高鐵站，取中位站，將嘉義站替換成雲林站，在本研究中取得了高鐵站之經緯度資料，並計算各站距離矩陣，利用降維技術降至二維空間並呈現之。

飲品的品項越來越多，光是碳酸飲品就有 3,568 支（台灣區飲品工業同業工會，2020），因此本研究想透過飲品資料集來了解各個不同種類的品項間的分群效果，並利用不同的名目資料處理方法做比較不同飲品類型的相似度，最終呈現於二維空間上。

1.2 目的

本研究利用欲利用 MDS 技術以台北、桃園、新竹、台中、雲林、台南和高雄高鐵站的經緯度，在 2D 平面上畫出從台北到高雄間的高鐵站，藉以了解高鐵站各站間的距離。

第二個資料集飲品中，利用 1-of-k 和 word2vec 的方式將名目資料進行轉換，接著使用 T-SNE 將資料降維，並將降維的資料視覺化後，藉此了解兩種方法之間的差異。

二、方法

2.1 實作方法說明

本研究實驗資料包括高鐵站各經緯度、飲品資料集，首先將經緯度做地表距離換算轉換成距離矩陣放入 MDS 演算法進行轉換後視覺化、飲品資料對每個 class 的 Amount 的常態分佈隨機產生數量（count）的資料，對名目資料使用套件 gensim 的 word2vec 與 1-of-K 進行文字編碼，最終分別進行視覺化於二維空間。

2.2 程式執行方法說明

本研究利用 Anaconda3 的 Jupyter notebook 環境來進行開發，高鐵經緯度資料之實驗使用 numpy 進行地表距離進行轉換後，使用 MDS 進行降維至二維空間。在飲品資料之實驗使用 pnadas、numpy 對飲品資料進行隨機產生後整理資料，另外對於飲品資料集的名目資料分別使用 gensim 的 word2vec 以及 1-of-K 進行數值化，將資料轉換成數值後再放入 T-SNE 裡面進行降維轉換，最終兩實驗資料集用 matplotlib 與 plotly dash 套件來進行螢幕互動式的圈選資料群集且能取得該群組的資料點。

三、實驗

3.1 資料集

本研究使用高鐵各站經緯度資料集以及飲品資料集兩種資料集作為實驗資料，以下為兩種資料集的部分資料內容及說明。

3.1.1 台灣高鐵經緯度資料集

台灣高鐵經緯度資料集是由台北高鐵站、桃園高鐵站、新竹高鐵站、台中高鐵站、雲林高鐵站、台南高鐵站以及高雄高鐵站這些站的經度和緯度所組成，經度和緯度欄位為數值型欄位，如表 1。

表 1

台灣高鐵站經緯度資料集

City	經度	緯度
taipei	25.047919631252775	121.51624105197017
taoyuan	25.013079603615427	121.21478270089906
hsinchu	24.80833160925325	121.04023159113365
taichung	24.111865844301832	120.61572958016143
yunlin	23.73641758174236	120.41654313039452
tainan	22.924679495261397	120.28570491599645
kaohsiung	22.688067241088056	120.30908218780064

3.1.2 飲品資料集（飲品 Dataset）

飲品資料集中有一個類別欄位（Class）和三個特徵欄位，分別為 Drink、Rank、Amount，Drink 為名目型欄位，Rand 及 Amount 為數值型欄位，而類別欄位有 7 種類別，分別為 A、B、C、D、E、F、G，如表 2，針對每一個類別，依照 Amount 的常態分配，隨機產生 Count 數量的資料筆數。

表 2

飲品資料集

Class	Drink	Rank	Amount($N(\mu,\sigma)$)	Count
A	Coke	7	(100, 200)	200
B	Pepsi	6	(200, 10)	100
C	7Up	5	(200, 10)	100
D	Sprite	4	(400, 100)	200
E	Latte	3	(800, 10)	100
F	Espresso	2	(800, 10)	100
G	Cappuccino	1	(900, 400)	200

$N(\mu,\sigma)$：Normal Distribution

3.2 前置處理

3.2.1 台灣高鐵經緯度資料集

圖 1 台灣高鐵經緯度資料集 前置處理流程圖

- 數值轉換：地球赤道半徑約為 6378.140 千米，根據地球表面任意兩點的經緯度就可以計算出這兩點間的地表距離，運用此關係對各高鐵經緯度進行地表距離轉換。
- 產生地表距離矩陣：計算各站對其它高鐵站的地表距離矩陣

表 3

高鐵經緯度換算成距離矩陣之資料

	台北	桃園	新竹	台中	雲林	台南	高雄
台北	0	30.652835	54.957883	138.44771	183.69544	267.44247	290.01383
桃園	30.652835	0	28.810964	117.23114	163.54828	250.95123	274.75187
新竹	54.957883	28.810964	0	88.662928	135.07047	223.31108	247.50305
台中	138.44771	117.23114	88.662928	0	46.449953	136.38248	161.56279
雲林	183.69544	163.54828	135.07047	46.449953	0	91.346581	117.21852
台南	267.44247	250.95123	223.31108	136.38248	91.346581	0	26.44857
高雄	290.01383	274.75187	247.50305	161.56279	117.21852	26.44857	0

3.2.2 飲品資料集（飲品 Dataset）

圖 2 飲品資料集（飲品 Dataset）-前置處理流程圖

- 資料生成：根據原資料表給定每個 class 的平均值以及標準差生成常態分佈隨機產生 count 資料增加資料。
- 名目資料轉換：將名目資料分別用 1-of-k 和 word2vec 的方式進行轉換。

表 4

名目資料用 1-of-k 後的資料表

	Rank	Amount	7Up	Cappuccino	Coke	Espresso	Latte	Pepsi	Sprite
0	7	43.509716	0	0	1	0	0	0	0
1	7	7.156773	0	0	1	0	0	0	0
2	7	128.79138	0	0	1	0	0	0	0
3	7	247.3906	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...
997	1	1041.3875	0	1	0	0	0	0	0
998	1	2095.7814	0	1	0	0	0	0	0
999	1	2044.748	0	1	0	0	0	0	0

表 5

名目資料利用 word2vec 轉換後的資料表

	Rank	Amount	Vector-1	Vector-2	…	Vector-301
0	7	43.509716	-0.02356	0.045201	…	0.046427
1	7	7.156773	-0.02356	0.045201	…	0.046427
2	7	128.79138	-0.02356	0.045201	…	0.046427
…	…	…	…	…	…	…
997	1	1041.3875	-0.05172	0.009637	…	0.017269
998	1	2095.7814	-0.05172	0.009637	…	0.017269
999	1	2044.7478	-0.05172	0.009637	…	0.017269

3.3 實驗設計

3.3.1 台灣高鐵經緯度資料集

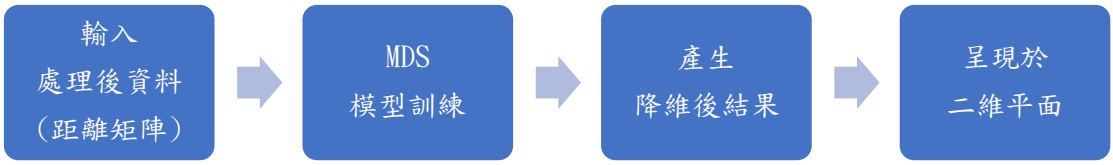

圖 3 台灣高鐵經緯度資料集實驗流程圖

在此實驗資料集中，主要任務為計算台灣高鐵各站之距離，並呈現於二維的圖表中，由於原始資料為各個高鐵站之經緯度資料，因此本實驗利用弧度轉換以及地表距離將原資料轉換成距離矩陣，如表 3，該距離矩陣擁有七個維度，無法映射到二維平面上，因此需要運用降維技術將其降維成兩個維度的空間，在本實驗方法中利用 MDS(Multi-Dimensional Scaling)來執行降維任務，MDS 保留了資料在原始空間的相對關係，且視覺化效果的呈現比較好。

MDS 主要是透過保持輸入資料的歐式距離（Euclidean distance），目的是希望降維後的低維資料中與輸入高維度資料的距離盡可能的保持最小化，在本研究中使用 scikit-learn 套件，進行 MDS 降維與結果展示，由於在本實驗中要呈現於二維圖表，因此在 MDS 中的維度參數設置為 2，最終將降維後的結果於 3.4 實驗結果展示。

3.3.2 飲品資料集（飲品 Dataset）

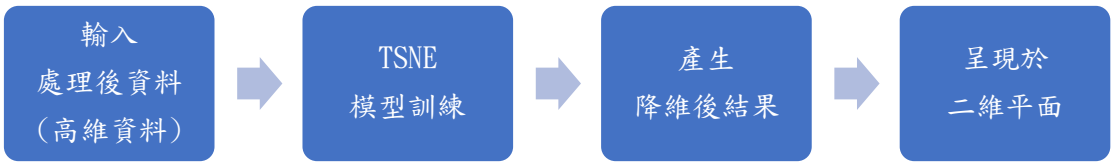

圖 4 飲品資料集實驗流程圖

在飲品資料集中，主要任務為利用名目資料以及其他資訊來將七種飲品資料呈現於 2 維平面空間上，原始資料集有三個特徵欄位，分別為(飲品, Rank, Amount)及一個類別欄位(Class)，飲品為名目型資料，然後再依照給定的 Amount 以常態分佈生成 Count，名目型資料分別 Word2Vector 與 One of K 來將名目型資料數值化，Word2Vector 的部分利用 fasttext-wiki-news-subwords-300 來詞向量轉換的 pre-model，另一為利用利用 One of K 來對名目資料進行編碼，完成以上前處理後將產生兩個高維度資料表，一為用 Word2Vectory 處理之資料表，二為 One of K 處理之資料表，再來利用 T-SNE 模型進行資料降維的任務。

T-SNE 主要是將高維度的資料用高斯分布的機率密度函數估計近似值，而在低維度的部分用 t 分布來求近似值，再使用 KL 計算兩分布的相似度，最終以梯度下降求得最佳解，在本研究中目的為將高維資料呈現於二維平面中，因此在 T-SNE 維度參數設置維 2，最終將降維結果呈現於二維空間，呈現於 3.4 章節中。

3.4 實驗結果

3.4.1 台灣高鐵經緯度資料集

在此實驗資料集，將原高維度距離矩陣經由 MDS 降維成二維資料，並將結果呈現於二維空間中，由下圖可以看出，本實驗資料集中的高鐵站與我們所認知的高鐵站距離是相對應的，由此可見將高維資料透過 MDS 降維後依然能呈現出原資料的距離關係。下圖為使用 ploty 套件所呈現的結果。

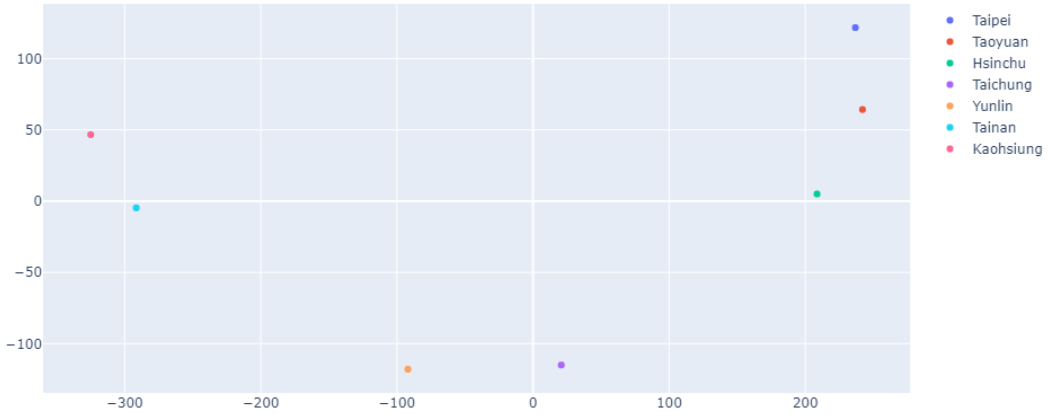

圖 5 台灣高鐵經緯度資料集

3.4.2 飲品資料集（飲品 Dataset）

在此實驗資料集中，將原高維度的資料經由 T-SNE 降維成二維資料，並呈現於二為平面空間上，由下面兩張圖可見，分別以 1-of-K、word2vec 處理的資料，在單一類別資料大致上都能有不錯的分群結果，僅有 Cappuccino 類別資料被分割為兩群，另外本研究發現若將飲品類別分為汽水（Coke、Pepsi、Sprite、7up）以及咖啡（Latte、Espresso、Cappuccino），由兩張圖得比較結果可以發現，word2vec 較能呈現出兩種飲品的不同之處，如圖中紅色虛線所示。

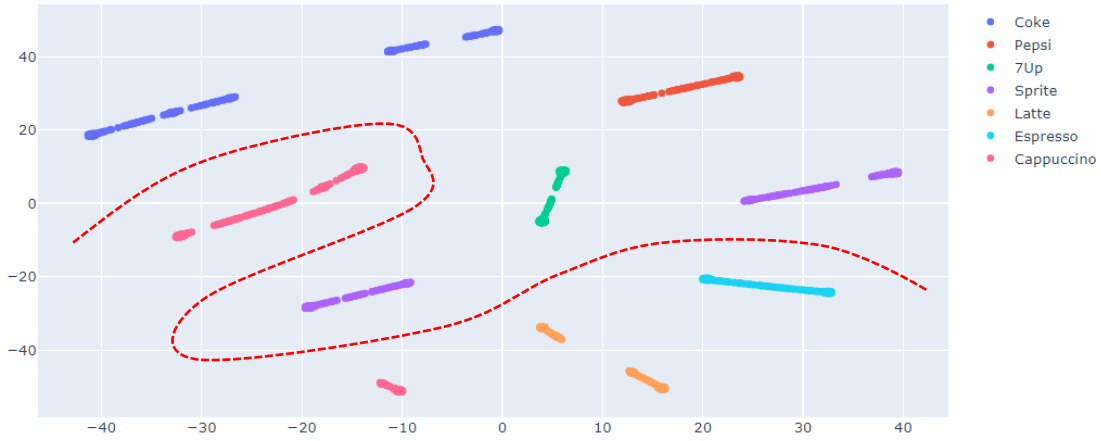

圖 6 以 1-of-K 處理之資料降維結果圖，紅線為人為繪製種類分割線

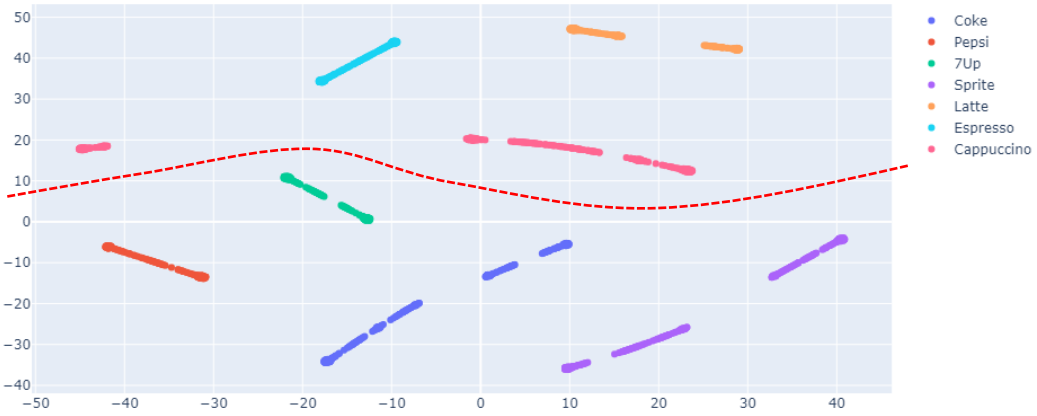

圖 7 以 word2vec 處理之資料降維結果圖，紅線為人為繪製種類分割線

四、 結論

在本研究主要探討將高維度資料經由降維之後，是否能在二維空間上表現出資料間的關係，本研究實驗資料包括高鐵站經緯度、飲品資料集，高鐵站經緯度首先將經緯度做地表距離換算轉換成距離矩陣放入 MDS 演算法進行降維後，視覺化於二維空間上，如 3.4 實驗結果所示；飲品資料集則是對每個 class 的 Amount 的用常態分佈隨機產生數量（count）的資料，並對名目資料使用 word2vec 與 1-of-k 方法將其轉成數值資料，分別進行 T-SNE 演算法進行降維後，視覺化後於二維空間上，並比較兩種文字轉數值方法之差異，結果如 3.4 實驗結果所示

在高鐵站經緯度結果中，可以發現利用 MDS 技術降維後，依然能呈現出每一高鐵站的距離關係；而在飲品資料實驗中，利用 T-SNE 技術降維後，僅有 Cappuccino 類別資料被分割為兩群，其他單一類別資料大致上都能有不錯的分群結果，另外本研究發現若將飲品類別分為汽水（Coke、Pepsi、Sprite、7up）以及咖啡（Latte、Espresso、Cappuccino），由 word2vec 二維平面圖可以發現較能呈現出兩種飲品的不同之處，本研究認為由於 1-of-K 僅僅將 7 種飲品類別以編碼方式進行數值化，而在 word2vec 不僅將名目資料轉換為向量資料，還將名目資料原本的文字語意也涵蓋進去，因此 word2vec 經由降維後呈現於二維平面中應當比 1-of-K 來得好。

參考文獻

台灣區飲品工業同業工會。(2020 年)。**臺灣飲品新品發展動向**，取自：
http://www.bia.org.tw/zh-tw/news-44033/%E8%87%BA%E7%81%A3%E9%A3%B2%E6%96%99%E6%96%B0%E5%93%81%E7%99%BC%E5%B1%95%E5%8B%95%E5%90%91.html