國立雲林科技大學資訊管理系

Department of Information Management National Yunlin University of Science & Technology Assignment

高維資料降維與視覺化：基於MDS和t-SNE的比較分析與互動式探索

Comparative Study of Multiple Dimensionality Reduction Techniques with Interactive Visualization: Applications of MDS and t-SNE on Mixed-Type Data

M11323039、M11323021、M11323023、M11323041

蔡昆彥、蔣育承、毛淑娟、林承勳

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

中華民國114年6月

June 2025

**摘要**

在許多實際應用中，高維度資料往往難以理解和處理，且1-of-k編碼無法有效處理類別間的相似關係，限制了降維表現與視覺化解釋力。本研究透過比較不同降維方法的效果、改善類別資料的處理方式，以及透過採用實用的視覺化工具，解決高維資料處理的關鍵問題。研究採用模組化程式架構，以Python進行實驗，設計兩個主要資料集：地理空間資料集使用台灣七個火車站的距離資料進行MDS降維分析，Drink Dataset為模擬生成的混合型資料，使用t-SNE技術並比較1-of-k編碼與Word2Vec編碼兩種處理方法，同時採用Plotly Dash框架建構互動式視覺化介面。實驗結果顯示，MDS方法在處理地理空間距離資料時能準確保持火車站間的相對位置關係，與實際地圖分佈高度吻合；Word2Vec編碼相較於1-of-k編碼能更有效捕捉飲料間的語義關係，在t-SNE視覺化圖中呈現更明顯的聚類效果；互動式視覺化系統成功實現即時資料圈選功能，提升資料探索的直觀性。本研究驗證了不同降維技術的適用場景，發現Word2Vec等語義嵌入方法在處理類別資料上的顯著優勢，建議對於地理空間資料採用MDS方法，對於混合型資料則優先考慮結合語義嵌入技術的t-SNE方法，為後續相關研究提供有效的分析工具與技術選擇指南。

關鍵字：降維技術、t-SNE、1-of-k編碼、Word2Vec編碼、互動式視覺化

1. **緒論**
2. **動機**

在許多實際應用中，高維度資料往往難以理解和處理，因此降維與視覺化技術變得越來越重要。這些技術能幫助我們發現資料中隱藏的規律，讓研究者更容易掌握資料特性，提升分析品質(McInnes et al., 2018; Van der Maaten & Hinton, 2008)。此外，過去的研究往往忽略名目型屬性在相似度計算中對降維結果的潛在影響。1-of-k編碼無法有效處理類別間的相似關係，進而限制降維表現與視覺化解釋力(Guo & Berkhahn, 2016; Mikolov et al., 2013)。而目前處理混合型資料時，仍面臨前置處理複雜、評估標準不統一等問題。透過MDS和t-SNE等降維技術，可以將複雜的高維資料轉換為容易理解的低維視覺化圖形(KRUSKAL, 1964; Van der Maaten & Hinton, 2008)。

本研究針對地理空間資料與混合型類別資料，探討不同降維方法的效能表現與適用性，並特別關注名目型屬性的相似度計算對降維品質的影響，達成更精確的資料視覺化與互動式分析能力。

綜合上述背景，如何有效運用多種降維技術處理不同類型資料，並建立考慮屬性相似度的名目型資料處理方法，以及實現高品質的互動式視覺化分析，成為本研究的核心議題。因此，本研究聚焦於以下三點，並制定三項具體研究問題：

1. 探討MDS方法在地理空間距離資料上的降維效能，以台灣主要火車站距離為例，分析其二維投影的準確性與Google地圖實際位置的對應關係。
2. 分析t-SNE技術處理混合型資料的效果，比較1-of-k編碼與Word2Vec編碼兩種名目型資料處理方法，探討如何在飲料資料集上達到更佳的群集視覺化效果。
3. 開發互動式資料探索系統，運用Dash的技術實現動態資料圈選與群集分析功能，提升降維結果的可解釋性與實用性。

本研究透過系統性評估不同降維方法的特性與應用場景，探究多元化降維技術的有效性，歸納最佳實踐方法並可作為後續相關研究的精準分析工具。

1. **目的**

基於上述研究動機與背景分析，本研究旨在建立一套有效的降維分析方法，解決高維資料處理的關鍵問題。

具體而言，本研究的目的包含以下三個層面：

1. **比較不同降維方法的效果：**

透過實際操作MDS和t-SNE技術，分析它們在地理空間資料（火車站距離）和混合型資料（Drink Dataset）上的表現差異，找出各自的優缺點和適用情況。

1. **改善類別資料的處理方式：**

針對1-of-k編碼的限制，比較兩種處理名目型資料的方法：標準的1-of-k編碼和Word2Vec編碼，探討如何更好地捕捉類別間的關係，提升降維品質。

1. **建立實用的視覺化工具：**

利用Dash的技術，讓使用者能夠直接在視覺化圖形上圈選和分析資料群組，提供更直觀的資料分析體驗。

綜合上述三個層面，本研究希望透過系統性的實驗和比較，提供研究者一套清楚的降維技術選擇指南，同時分析出何者為較實用的分析工具，讓高維資料的理解變得更加容易和有效。

1. **方法**
2. **架構**

本研究採用模組化程式架構，主要分為四個核心模組：資料處理模組、降維分析模組、視覺化模組及互動式操作模組。整體系統以Python為主要開發語言，整合NumPy、Pandas、Scikit-learn、Matplotlib、Plotly等函式庫，確保各模組間的高效協作與資料流通。

1. **資料收集與資料前處理**
   * 1. 資料集一：地理空間資料集

* 採用Google Maps取得台北、新竹、台中、斗六、高雄、花蓮玉里、台東知本共七個火車站彼此之間的距離。
* 然後用MDS方法投影在2D平面上。
* 使用Google地圖標記其位置。
  + 1. 資料集二：Drink Dataset
* 透過資料敘述生成四個特徵欄位（Drink、Rank、Amount、Quantity）及一個類別欄位（Class）的模擬資料。
* Amount欄位按常態分配生成。
* Quantity欄位在指定範圍內隨機亂數生成。
* 各類別隨機產生Count數量的資料筆數。
* 最後以t-SNE將資料投影在2D平面。
  + 1. 資料正規化處理
* 針對Drink Dataset的數值型欄位採用Min-Max正規化方法，將所有數值縮放至0-1區間。
  + 1. 降維實作
* 地理空間資料集：使用MDS方法，透過特徵值將距離矩陣降維，將七個火車站的相對位置投影至2D平面。
* Drink Dataset：採用標準t-SNE演算法處理Drink Dataset，特別針對名目型屬性實作兩種處理方式：
  + 1-of-k編碼方法：將名目型屬性Drink直接轉換為one-hot向量，與數值型屬性合併後輸入t-SNE演算法。
  + 相似度考量方法：建立類別間相似度矩陣，採用Word2Vec編碼等嵌入技術學習類別間的語義距離。
    1. 互動式視覺化
* 採用Plotly Dash框架建構互動式視覺化介面，實現在降維結果圖上選取特定區域時，系統即時顯示該區域內資料點的詳細資訊。

1. **實驗設計**
2. **資料集一：地理空間資料集**

採用Google Maps取得台北火車站、新竹火車站、台中火車站、斗六火車站、高雄火車站、花蓮玉里及台東知本的經緯度座標。

**表 1**

*各火車站的經緯度座標*

|  |  |  |
| --- | --- | --- |
| **車站** | **緯度** | **經度** |
| 台北火車站 | 25.047637204053995 | 121.5171273798563 |
| 新竹火車站 | 24.801750331885113 | 120.9716203981461 |
| 台中火車站 | 24.137523918186687 | 120.68683583244295 |
| 斗六火車站 | 23.712201175089422 | 120.54104266971164 |
| 高雄火車站 | 22.6396706176077 | 120.30261518317573 |
| 花蓮玉里火車站 | 23.727506999999997 | 120.29999999999999 |
| 台東知本火車站 | 23.712201175089422 | 120.54104266971164 |

1. **資料集二：Drink Dataset**

Drink Dataset共有四個特徵欄位(Drink, Rank, Amount, Quantity)及一個類別欄位(Class)，其中Drink為名目型欄位，而Rank、Amount、Quantity為數值型欄位。

**表 2**

*Drink Dataset (生成處理前*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class** | **Drink** | **Rank** | **Amount**  **(N (μσ))** | **Quantity** | **Count** |
| A | Coke | 8 | (100, 200) | Random(500, 1000) | 100 |
| B | Pepsi | 7 | (200, 10) | Random(500, 1000) | 200 |
| C | Sprite | 6 | (200, 10) | Random(500, 1000) | 100 |
| D | 7Up | 5 | (400, 100) | Random(500, 1000) | 400 |
| E | Cappuccino | 4 | (700, 10) | Random(1, 500) | 400 |
| F | Espresso | 3 | (700, 10) | Random(1, 500) | 200 |
| G | Latte | 2 | (800, 300) | Random(1, 500) | 100 |
| H | Mocha | 1 | (900, 400) | Random(1, 500) | 100 |

本研究針對每一個類別(Class)，依照Amount的常態分配及亂數分配Quantity，隨機產生Count數量的資料筆數。

**表 3**

*Drink Dataset (生成處理後*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Class | Drink | Rank | Amount | Quantity |
| 0 | A | Coke | 8 | -140.028960 | 774.985814 |
| 1 | A | Coke | 8 | 318.755235 | 757.997651 |
| 2 | A | Coke | 8 | -101.136205 | 713.600572 |
| 3 | A | Coke | 8 | -67.477304 | 543.120963 |
| 4 | A | Coke | 8 | -104.616745 | 584.091803 |
| … | ... | ... | ... | ... | ... |
| 1595 | H | Mocha | 1 | 995.692659 | 117.574864 |
| 1596 | H | Mocha | 1 | 1042.645044 | 306.298079 |

1. **資料正規化**

針對Drink Dataset的數值型欄位採用Min-Max正規化方法，將所有數值縮放至0-1區間，確保不同欄位間的值域一致性，避免某些特徵因數值範圍較大而主導距離計算結果。

1. **實驗設計**

* 地理空間資料集：使用MDS方法，透過特徵值將距離矩陣降維，將各個火車站的相對位置投影至2D平面。
* Drink Dataset：採用標準t-SNE演算法處理Drink Dataset，特別針對名目型屬性實作兩種處理方式：
  + 1 of-k編碼方法：將名目型屬性Drink直接轉換為one-hot向量，與數值型屬性合併後輸入t-SNE演算法。
  + 相似度考量方法：建立類別間相似度矩陣，同類飲料設定較小的不相似度值（如2），不同類飲料間設定較大的不相似度值（如4），採用Word2Vec編碼等嵌入技術學習類別間的語義距離。

1. **互動式視覺化**

採用Plotly Dash框架建構互動式視覺化介面，實現在降維結果圖上選取特定區域時，系統即時顯示該區域內資料點的詳細資訊。

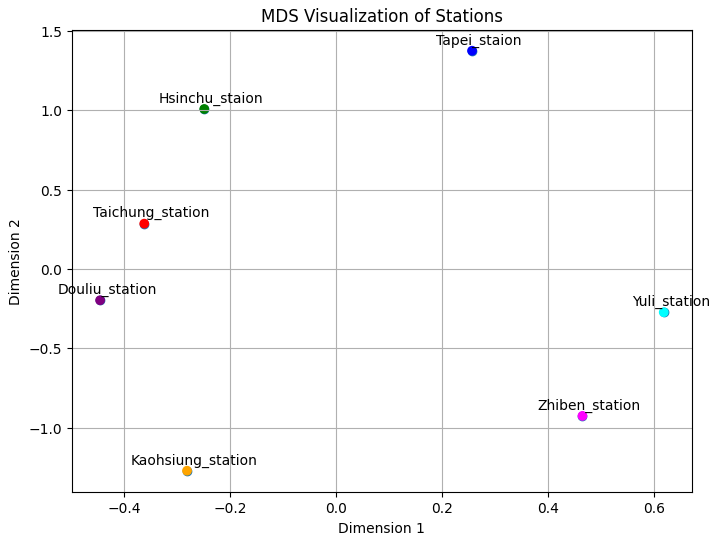
1. **實驗結果**

* **資料集一：地理空間資料集**

本實驗利用MDS方法，以視覺化方式展現台北、新竹、台中、斗六、高雄、花蓮玉里、台東知本共七個火車站的地理距離關係，透過散點圖（**圖 1**）的方式呈現降維後之各個火車站間的相對距離，本研究通過與實際地圖（**圖 2**）的對比發現與實際地理分佈高度吻合，進一步驗證了MDS方法在可視化地理距離上的有效性。

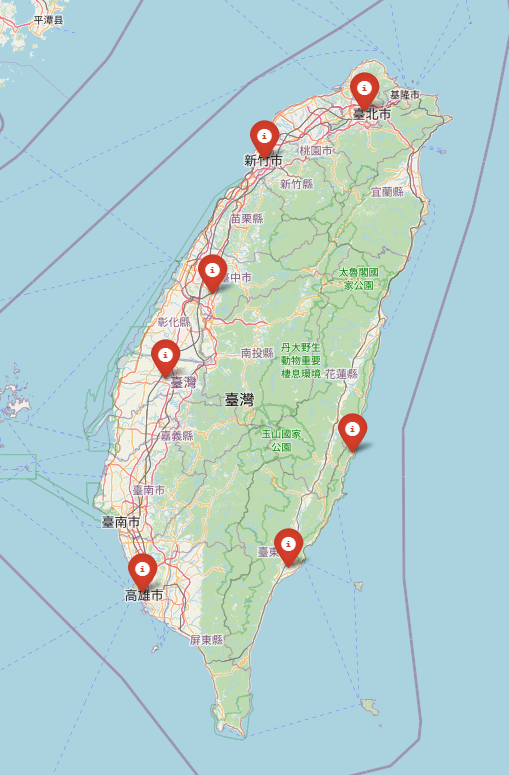
**圖 1**

*各火車站距離之MDS視覺化*



**圖 2**

*各個火車站真實座標圖*



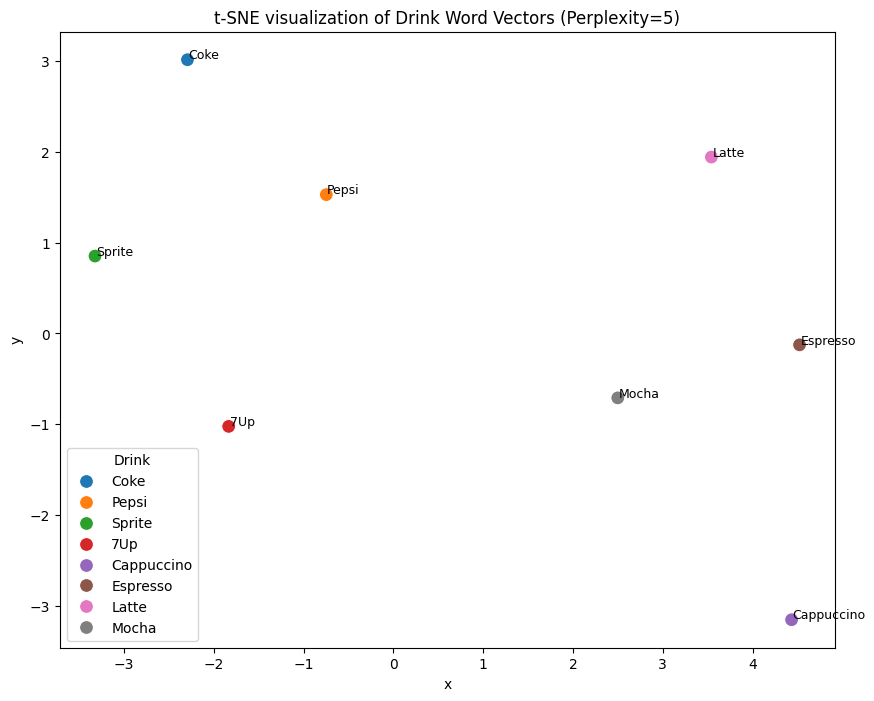
* **資料集二：Drink Dataset**

本研究針對1-of-k編碼的限制，分析t-SNE技術處理混合型資料的效果，比較1-of-k編碼與Word2Vec編碼這兩種處理名目型資料的方法。實驗結果表明，Word2Vec編碼在視覺化圖表中能更好的展現飲料之間的相似度（**圖 3**），其相似飲品在t-SNE視覺化圖中呈現明顯聚類，反觀1-of-k編碼僅將飲料單純轉化為獨立類別（**圖 4**），未能體現彼此的相似度，圖表也相對分散。

本研究根據Word2Vec編碼另外繪製了heatmap（**圖 5**），顏色越深則相似度越高，明確的體現了Word2Vec編碼在分析兩種飲料之間語意相似度的有效性。例如:Cappuccino與Expresso相似度較高，而1-of-k編碼在圖表呈現上卻分離的很遠，說明了其無法良好的呈現飲料間的語意關係。

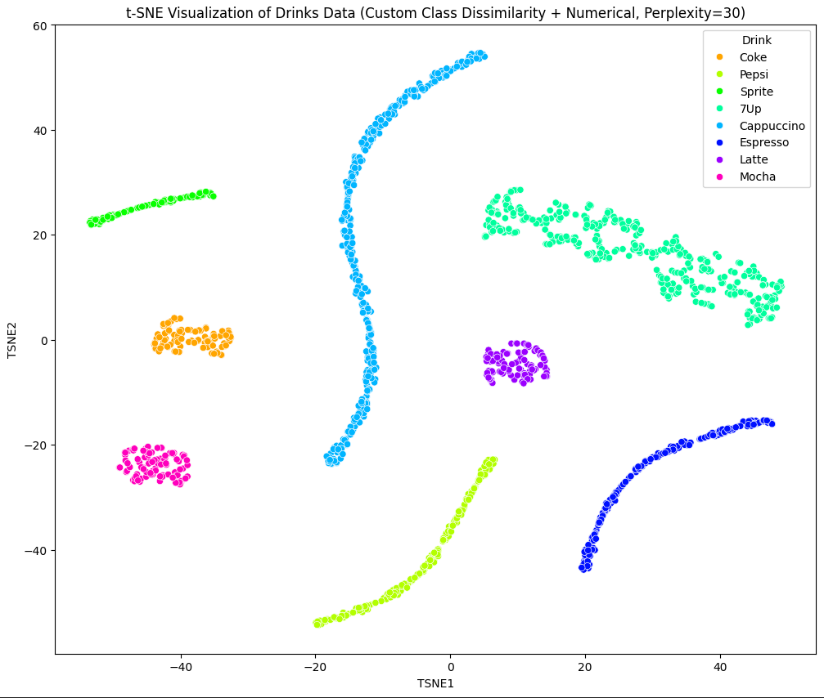
**圖 3**

*Drink Dataset使用Word2Vec編碼之t-SNE視覺化圖*

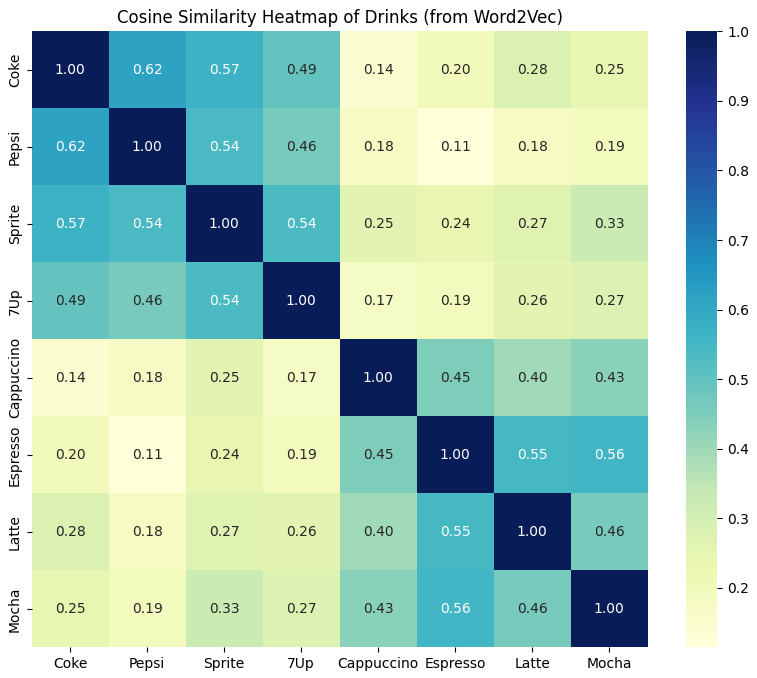


**圖 4**

*Drink Dataset 使用1-of-k編碼之t-SNE視覺化圖*



**圖 5**

*Drink* *Dataset heatmap視覺化*

* **互動式視覺化**

**圖 6**採用Plotly Dash框架建構互動式視覺化介面，實現在降維結果圖上選取特定區域時，系統能即時顯示該區域內資料點的詳細資訊。當使用者在降維結果圖上選取特定區域時，系統即時顯示該區域內資料點的詳細資訊，包括原始特徵值、類別標籤等。

|  |  |
| --- | --- |
| **圖 6**  *Drink Dataset Dash之視覺化* | |
|  |  |

1. **結論**

本研究基於三個主要目的進行了系統性的降維技術比較與應用研究，並獲得了重要的研究成果與發現。

**1. 比較不同降維方法的效果：**

透過地理空間資料的實驗結果顯示，MDS方法在處理距離矩陣時表現優異，能夠準確保持火車站間的相對地理位置關係，與實際地圖分佈呈現高度吻合。此發現證實了MDS在地理空間視覺化應用上的有效性，為後續相關研究提供了可靠的方法選擇依據。

**2. 改善類別資料的處理方式：**

Word2Vec編碼相較於1-of-k編碼，能夠更有效地捕捉飲料間的語義關係，在t-SNE視覺化圖中呈現更明顯的聚類效果。特別是相似飲品（如Cappuccino與Espresso）在Word2Vec編碼下展現了較高的相似度，而1-of-k編碼則無法體現此種語義關聯。此次實驗明確證實了Word2Vec編碼優於1-of-k編碼。

**3. 建立實用的視覺化工具：**

透過採用Plotly Dash框架的互動式視覺化系統，實現了即時資料圈選與群集分析功能，顯著提升了資料探索的直觀性與便利性。

綜合而言，本研究透過系統性的實驗設計與比較分析，認為關於地理空間等距離資料建議採用MDS方法，若是包含類別變數的混合型資料則應優先考慮結合語義嵌入技術的t-SNE方法，此次實驗不僅驗證了不同降維技術的特性與適用場景，更重要的是發現了Word2Vec編碼等語義嵌入方法在處理類別資料上的顯著優勢。同時，互動式視覺化系統的成功開發也為後續研究提供了有效的分析平台，使高維資料的理解與應用變得更加容易和直觀。

References

1. Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.
2. KRUSKAL, J. (1964). MULTIDIMENSIONAL SCALING BY OPTIMIZING GOODNESS OF FIT TO A NONMETRIC HYPOTHESIS. *Psychometrika*, *9*(1).
3. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.
5. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).