靜宜大學資訊工程學系畢業專題計畫書

**基於一可擴展性且高效率之大數據降維演算法分析**

**-應用於COVID-19患者資料**

指導教師：鄭婉淑

專題學生：

資工三B 411147322 姜駿楷 [s1114732@o365st.pu.edu.tw](mailto:s1114732@o365st.pu.edu.tw)

資工三B 411147144 李東晉 [s1114714@o365st.pu.edu.tw](mailto:s1114714@o365st.pu.edu.tw)

資工三B 411147623 周宏澤 [s1114762@o365st.pu.edu.tw](mailto:s1114762@o365st.pu.edu.tw)

資工三B 411147089 李侑叡 [s1114708@o365st.pu.edu.tw](mailto:s1114708@o365st.pu.edu.tw)

繳交日期：114年4月1日

1. **摘要**

在大數據時代，高維數據的有效處理與分析方法已成為重要的研究課題。隨著科技的迅速發展，生物醫學、金融科技及環境科學等領域的數據維度與複雜性不斷提升，高維數據中冗餘資訊的存在往往導致計算效率瓶頸。因此，探索高效的數據降維方法對於提升數據分析的準確性與效率至關重要。

在高維數據分析過程中，降維方法的選擇至關重要，需在保留關鍵資訊的同時兼顧計算效率。本研究整合主成分分析法(Principal Component Analysis, PCA)、ｔ分佈隨機鄰居嵌入法(t-distributed Stochastic Neighbor Embedding, t-SNE)及可擴展頻繁模式挖掘SFP演算法(Scalable Frequent Pattern Mining,SFP)，進行多層次的數據處理與分析。為驗證所提出方法的適用性與效能，首先使用 MNIST 資料集進行初步測試，隨後將方法應用於 COVID-19 患者數據集，以進一步分析其可行性。

預期研究成果將顯示，整合前述三種方法可在計算效率、結果準確性及特徵選取等方面展現顯著優勢。本研究不僅為高維數據分析提供了新的方法參考，也為各領域在降維策略的應用上提供了不同的解決方案。

1. **進行方法及步驟**
2. **實驗設計**
   1. **數據獲取與前處理**
   2. **降維方法**
      1. **PCA（Abdi & Williams,2010）**

PCA是一種多變量技術，用來分析一個資料表，其中觀察值是由多個彼此相關的定量依變數所描述。其目標是從該資料表中提取重要資訊，將其表示為一組新的正交變數，稱為主成分（principal components），並以此展示觀察值與變數之間的相似性模式，矩陣以大寫粗體字母表示，向量以小寫粗體字母表示，元素以小寫斜體字母表示。來自同一矩陣的元素都使用相同字母（例如：A,a,*a*）。轉置操作使用上標T表示。單位矩陣表示為 I。

PCA要分析的資料表包含*Ｉ*筆觀察，每筆觀察由*J*個變數描述，表示為的矩陣X，其中通用元素為。矩陣X的秩為*L*，滿足。

一般而言，資料會在分析前做前處理。幾乎總是會將X的各欄置中，使每欄平均值為零（即，其中0是一個長度為*J*的0向量，1是長度為*I*的全1向量）。若再進一步將X的每個元素除以 或，則該分析稱為共變異數PCA（covariance PCA），此時矩陣是共變異數矩陣。

除了置中外，當變數的單位不同時，通常還會將每個變數標準化為單位範數。這是藉由將每變數除以其範數（即該變數平方和的平方根）完成的。此時，分析稱為相關係數PCA （correlation PCA），因為矩陣是相關係數矩陣（大多數統計套件預設使用相關係數的 預處理方式）。

**奇異值與慣性（Inertia）**

Δ 是奇異值的對角矩陣。注意： 等於A，其中A是和的非零特徵值的對角矩 陣。欄的慣性（inertia of a column）定義為該欄元素平方和，表示如下：

所有欄的總和記為*I*，稱為資料表的總慣性（total inertia）或簡稱慣性。總慣性也等於資料表奇異值平方和。

列的重心（center of gravity of the rows），也叫做質心或重心，是每欄的平均值所組成的向量s。當矩陣X已中心化（每欄平均值為零），其重心為的零向量。

第i筆觀察的歐氏距離定義為：

若資料已中心化（每欄平均值為 0），則上述公式簡化為：

注意，所有加總等於總慣性*I*。

**PCA 的目標（GOALS OF PCA）:**

1. 從資料表中提取最重要的資訊；

2. 透過保留最重要的資訊來壓縮資料集；

3. 簡化資料的描述；

4. 分析觀察值與變數的結構關係。

為了達成以上目標，PCA會計算新變數，稱為主成分，這些主成分是原始變數的線性組合。

• 第一主成分會有最大的變異量（或稱慣性），也就是能夠「解釋」資料中最大部分的總慣性；

• 第二主成分在與第一主成分正交的前提下，擁有最大的剩餘變異量；

• 其他主成分依此類推。

觀察值在這些新變數上的得分稱為因子得分（factor scores），這些得分可以透過幾何方式解釋為投影到主成分上的位置。

**主成分（Finding the Components）**

在 PCA中，主成分來自對資料矩陣X的奇異值分解（SVD）：

F，即𝐼×𝐿的因子得分矩陣，定義為：

矩陣Q提供了線性組合的係數，這些係數用於計算主成分得分。此矩陣也可視為投影矩陣 （projection matrix），因為將資料X乘上Q可以得到觀察值在主成分上的投影。這可合併 公式和得出：

矩陣 Q 也表示用來計算因子得分的線性組合係數。此矩陣也可以被視為一個投影矩陣，因為將資料矩陣X乘上Q，就可以得到每個觀察值在主成分上的投影值。這可以由以下公式表示：

主成分也可以從幾何角度來看，表示為原始軸的旋轉。在這個脈絡下，矩陣 Q 可以視為方向 餘弦矩陣（direction cosines），因為Q是正交矩陣。同時，Q也稱為載荷矩陣（loading matrix）。

在這種解釋下，原始資料矩陣 X 可以被看作是因子得分矩陣與載荷矩陣的乘積：

其中，

主成分也可視為原始軸的旋轉。在幾何上，如果X是兩變數的資料，例如某字詞的長度(Y)和其詞義定義的數量(W)，PCA就是用兩個正交軸表示資料的方式。

**某筆觀察對主成分的貢獻（Contribution of an Observation to a Component）**

一個主成分的特徵值（eigenvalue）等於所有觀察在該主成分上的因子得分平方和。因此，可以透過某筆觀察的得分平方除以該主成分的特徵值來衡量該觀察對此主成分的貢獻。

這個比例稱為該觀察對主成分*t*的貢獻值，記為，其計算公式如下：

其中是第*t*個主成分的特徵值。

• 𝑐𝑡𝑟𝑖,𝑡的值介於0到1之間；

•對某一主成分而言，所有觀察的貢獻值總和為 1；

•一般而言，若某筆資料的貢獻值大於平均貢獻值（即大於），就表示它對該主成分的重要性較高。

這些高或低貢獻的觀察值可幫助解釋主成分的意義。

**1.2.2** **t-SNE(Laurens van der Maaten，2008):**

我們使用名為「t-SNE」的降維技術，可將高維資料視覺化，將每個資料點對應到二維或 三維空間中的位置。這種技術是隨機鄰域嵌入法（Stochastic Neighbor Embedding, SNE； Hinton&Roweis,2002）的變種，更易於優化，且能夠顯著改善視覺化效果，尤其在降低資料點於視覺化圖中央過度密集的傾向方面有明顯進步。t-SNE相較於現有技術，更能以單一圖表展現不同尺度的資料結構。這對於那些分布在不同但相關的低維流形（manifold）上的高維資料尤為重要。

高維資料的視覺化是許多不同領域的重要問題，其所處理的資料維度也各異。SNE的起始步驟，是將高維空間中資料點之間的歐式距離轉換成表示相似性的條件機率。資料點相對於資料點的相似性，被定義為條件機率，此機率代表若我們以資料點為中心的高斯分布來選擇鄰近資料點，則選擇到資料點的機率。對於接近的資料點而言，的值會相對較高；而對於距離較遠的資料點而言，將會趨近於零（假設高斯分布的變異數σ取適當值），數學上，條件機率表示為：

對於高維資料點與在低維空間中的對應點和，我們同樣能計算一個條件機率。

我們在低維空間中設定高斯分布的變異數為，因此資料點對資料點的相似性定義為：

如果低維空間中的點和正確地表達了高維空間中的資料點與之間的相似性，那麼條件機率與的值將會非常接近。受此啟發，SNE的目的就是找出一個低維空間的資料表示法，使得條件機率分布與之間的差異盡可能最小化。用來衡量對於建模效果的自然標準，是Kullback-Leibler（KL）divergence。因此SNE透過梯度下降法來最小化所有資料點之KL散度總和，其損失函數定義為：

𝑃𝑖表示給定資料點之後，對其他所有資料點的條件機率分布，而則是在低維空間中的對應條件機率分布。由於KL散度並非對稱的，因此在低維空間的成對距離中，不同型態的誤差具有不同的權重。特別是，若用距離遙遠的點去表示相近的資料點（亦即用很小的去建模很大的），將導致很大的損失。反之，若用相近的低維空間資料點去表示原本距離遙遠的資料點，則僅產生較小的損失，這表示SNE方法專注於保留資料的局部結構。

SNE還需要選擇的另一個參數，是高維空間中每個資料點中心的高斯分布變異數σ。由 於資料密度在各處可能不同，因此不可能只用單一的變異數值就能適用於整個資料集。對於較密集的區域來說，通常需要較小的變異數，而稀疏區域則需較大的變異數。特定的變異數值會對所有資料點產生一個機率分布。SNE透過二元搜尋找出一個合適的變異數σ，使得產生的分布具備使用者事先設定的「困惑度」（perplexity）。定義為：

這裡H()代表以bits為單位的Shannon熵：

困惑度可以直觀解釋為資料點的有效鄰近點數量的平滑量測。SNE的效能對於困惑度的選取相當穩健，通常困惑度的設定值在5到50之間。SNE透過梯度下降法來最小化上述損失函數，其梯度具有相當簡潔的數學形式：

梯度下降的初始值，是從以原點為中心且變異數極小的高斯分布隨機抽取低維空間的資料點。為了提高優化速度並避免局部最佳解(Local minimum)，通常會在梯度下降中加入一個較大的動量項（momentum）:

SNE 雖能產生相當好的視覺化結果，但存在兩個明顯問題：一是其損失函數難以優化，二是有所謂的「擁擠問題」（crowding problem）。考慮一組資料點位於嵌入在高維空間內的低維流形（manifold）上，當此流形的固有維度（intrinsic dimension）增加時，二維空間中用來表示適中距離資料點的區域將變得不足，這會造成大量的資料點被壓縮到視覺化圖表的中心區域。儘管每個資料點對此產生的吸引力很小，但大量資料點合在一起卻足以使中心區域點群被過度壓縮，以致無法清晰呈現出自然的群集。

由於對稱SNE實際上是在匹配資料點對之間的聯合機率而非距離，因此我們可以透過在 低維空間使用比高斯分布尾部更厚的分布，來緩解擁擠問題。在t-SNE中，我們使用自由度為1的Student-t分布作為低維空間中的厚尾分布：

這種分布有一個優點：即使兩點距離變大，它們之間的機率密度仍以平方倒數率（inverse square law）衰減，這使得點群可以更自由地展開，避免過度集中。此外，這也使得t-SNE更易於優化，避免陷入不好的局部最佳解。t-SNE損失函數的梯度公式為：

**t-SNE的優化方法（Optimization Methods for t-SNE），具有兩種優化方式:**

1.“early compression”: 在優化初期，施加一個額外的L2正則化項，使資料點靠近原點。

2.“early exaggeration”: 初期將所有高維聯合機率乘上一個誇大因子，讓資料在初期 更明確地聚集成群，幫助形成良好的全局結構。

**1.2.3** **SFP（Cheng et al.,2024）**

SFP 是一種改進版的頻繁模式挖掘演算法，專為大規模數據處理設計。它的核心概念是改進 Database Projection（DP）方法，透過「關聯分組（Association-based Grouping）」來減少 I/O 操作、提高計算效率，並且適應不同記憶體條件。他的主要特徵有(1)支援動態記憶體管理，意思是根據內存狀況自動調整算法。(2) 減少硬碟讀寫次數，根據文獻此方法相較DP方法更加有效率，因為減少了高昂的 I/O 成本。(3) 增強效能，根據文獻表明SFP 算法在處理複雜數據時的執行時間顯著少於 DP 方法。

* 1. **實驗流程**

**2.1.1 使用MNIST資料集進行以下降維測試：**

• **只使用PCA**

使用MNIST資料集完整70000筆資料降維至2D視覺化，並且記錄執行PCA所需時間 :

一張含有 文字, 螢幕擷取畫面, 鮮豔, 圖表 的圖片

AI 產生的內容可能不正確。

• **只使用t-SNE**

使用MNIST資料集抽樣10000筆資料(分別為不同perplexity)來進行2D視覺化比較，並 且記錄執行t-SNE所需時間：

Perplexity設定5：

一張含有 文字, 螢幕擷取畫面, 圖表, 地圖 的圖片

AI 產生的內容可能不正確。

Perplexity設定25：

一張含有 文字, 螢幕擷取畫面, 圖表, 地圖 的圖片

AI 產生的內容可能不正確。

Perplexity設定50：

一張含有 文字, 螢幕擷取畫面, 圖表, 鮮豔 的圖片

AI 產生的內容可能不正確。

Perplexity設定100：

一張含有 文字, 地圖, 螢幕擷取畫面, 圖表 的圖片

AI 產生的內容可能不正確。

• **先使用PCA，再使用t-SNE視覺化**

使用MNIST資料集抽樣10000筆資料(分別為不同perplexity)來進行2D視覺化比較，並且記錄先使用PCA降維至50維後再使用t-SNE所需時間 :

Perplexity設定5:

一張含有 文字, 螢幕擷取畫面, 地圖 的圖片

AI 產生的內容可能不正確。

Perplexity設定25:

一張含有 文字, 地圖, 螢幕擷取畫面 的圖片

AI 產生的內容可能不正確。

Perplexity設定50:

一張含有 文字, 地圖, 螢幕擷取畫面, 鮮豔 的圖片

AI 產生的內容可能不正確。

Perplexity設定100:

一張含有 文字, 地圖, 螢幕擷取畫面, 鮮豔 的圖片

AI 產生的內容可能不正確。

**我們將在實驗中使用SFP演算法對數據前處理，然後使用PCA降維，再使用t-SNE視覺化**

**2.1.2 未來在實驗中使用 COVID-19 的 PBMC 資料集進行以下降分析：**

• PCA（主成分分析）。

• t-SNE（t-Distributed Stochastic Neighbor Embedding）。

• 先使用PCA降維，再使用t-SNE視覺化。

• 先使用SFP演算法對數據前處理，然後使用PCA降維，再使用t-SNE視覺化。

1. **效能與準確性對比**
2. **結果分析與討論**
3. **經費預算需求表**

**編列預算範本**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 項 目 名 稱 | 說 明 | 單位 | 數量 | 單 價 | 小 計 | 備 註 |
| 臺幣(元) | 臺幣(元) |
| 個人電腦 | 專案之進行 | 部 | 2 | 26000 | 52000 | 由系上實驗室提供 |
| 共 計 | | | | | 52000 |  |

1. **工作分配**

姜駿楷：期刊探討、測試結果分析、文書處理

李東晉：測試MNIST資料集、期刊探討、測試結果分析

周宏澤：程式優化、期刊探討、測試結果分析

李侑叡：程式優化、測試MNIST資料集、期刊探討、測試結果分析

1. **預期完成之工作項目及具體成果**

目前我們使用MNIST手寫資料集來模擬患者PBMC的數據分析情形。透過以下多組測試資料。

首先，分別使用單純PCA降維分析和t-SNE降維分析，直接用原始784維資料做視覺化結果。接著，將PCA 降到50維，再使用t-SNE將50維資料進一步降到2 D做視覺化模型。最後，在開始PCA降維前，先進行SFP前置處理。PCA的運算成本與t-SNE的迭代計算量主要取決於原始資料維度。由於MNIST中很多背景像素數值接近0，但仍被納入運算，這可能會增加計算負擔。SFP前處理增加了二值化與頻繁特徵篩選，過濾出頻繁項目，再藉由動態調整分群數量，將資料集拆分成較小的子集進行局部挖掘，從而有效避免記憶體不足的問題。最後，將各子集的挖掘結果整合成全局頻繁模式。此外，使用SFP前置處理後預期t-SNE的視覺化效果更佳，因為數據中的關鍵結構會更突出，使低維投影能夠更清楚地展示數據的分佈。

本研究預期SFP搭配傳統方法將在數據降低維度、計算效率與細胞群分類準確性方面優於未使用SFP的傳統方法。透過SFP先行篩選PBMC數據中的頻繁模式與關鍵特徵，預計能保留更多與細胞類型相關的重要資訊，減少降低維度過程中的訊息損失，使後續的數據映射與可視化結果的結果更加清晰，提升不同細胞群之間的可分性。與傳統方法相比，SFP 預期能更有效地區分PBMC亞群，如CD8+ T細胞、CD4+ T細胞、B細胞、單核細胞與NK細胞，甚至可能發掘新的細胞亞型，從而提升細胞分類的準確性與解釋性。

整體而言，本研究預期SFP在降維前的特徵篩選將顯著優化數據結構，使後續的模式辨識與視覺化分析結果更加穩定，並為此研究中的PBMC數據的進一步解析與免疫研究提供更具效能與準確性的策略，也預期此研究將能支持SFP在大數據降維過程中扮演不可或缺的部分。

**參考文獻**

1. van der Maaten, L. & Hinton, G., “Visualizing Data using t-SNE”, Journal of Machine Learning Research, Vol. 9, pp. 2579–2605, 2008.
2. Cheng, W.-S., Lin, Y.-T., Huang, P.-Y., Chen, J.-C. & Lin, K.W., “A fast and highly scalable frequent pattern mining algorithm”, Future Generation Computer Systems, Vol. 160, pp. 854–868, 2024.
3. Abdi, H. & Williams, L.J., “Principal component analysis”, Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 2, No. 4, pp. 433–459, 2010.
4. Su, Y., Chen, D., Yuan, D., et al., “Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19”, Cell, Vol. 183, No. 5, pp. 1479–1495, 2020.
5. Wen, W., Su, W., Tang, H., et al., “Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing”, Cell Discovery, Vol. 6, No. 31, 2020.
6. Kobak, D. & Berens, P., “The art of using t-SNE for single-cell transcriptomics”, Nature Communications, Vol. 10, Article No. 5416, 2019.