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

資工三 B 411147144 李東晉 s1114714@o365st.pu.edu.tw

資工三 B 411147623 周宏澤 s1114762@o365st.pu.edu.tw

資工三 B 411147089 李侑叡 s1114708@o365st.pu.edu.tw

繳交日期：114 年 4 月 1 日

第一章 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。

第二章 進行方法及步驟

1. 實驗設計

1.1 數據獲取與前處理

1.2 降維方法

1.2.1 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 $I \times J$ 的矩陣 X，其

中通用元素為x_{ij}。矩陣 X 的秩為 L，滿足$L \leq min\{I,J\}$。

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$X^T 1 = 0$，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$\sqrt{I}$ 或$\sqrt{I-1}$，則該分析稱為共變異數 PCA（covariance PCA），此時矩陣$X^T X$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$X^T X$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：Δ^2 等於 A，其中 A 是$X^T X$和XX^T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：

$$\gamma_i^2 = \sum_i^I x_{i,j}^2$$

所有欄的γ^i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$1 \times J$的零向量0^T。

第 i 筆觀察的歐氏距離定義為：

$$d_{is}^2 = \sum_j^J \left(x_{ij} - s_j\right)^2$$

若資料已中心化（每欄平均值為 0），則上述公式簡化為：

$$d_{i,g}^2 = \sum_I^J x_{i,j}^2$$

注意，所有d_{is}^2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）：

$$X = P\Delta Q^T$$

F，即I×L的因子得分矩陣，定義為：

$$F = P\Delta$$

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$X = P\Delta Q^T$和$F = P\Delta$得出：

$$F = P\Delta = PAQ^TQ = XQ$$

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：

$$F = P\Delta = XQ$$

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：

$$X = FQ^T \text{其中} F^TF = \Delta^2，Q^TQ = 1$$

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$ctr_{i,t}$，其計算公式如下：

$$ctr_{i,t} = \frac{f_{i,t}^2}{\sum_i f_{i,t}^2} = \frac{f_{i,t}^2}{\lambda_t}$$

其中λ_t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$1/I$），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X_j相對於

資料點X_i的相似性，被定義為條件機率$P_{j|i}$，此機率代表若我們以資料點X_j為中心的高斯分布來選擇鄰近資料點，則選擇到資料點X_j的機率。對於接近的資料點而言，$P_{j|i}$的值會相對較高；而對於距離較遠的資料點而言，$P_{j|i}$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$P_{j|i}$表示為：

$$P_{j|i} = \frac{exp\left(-\|X_i - X_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i} exp(-\|X_i - X_k\|^2/2\sigma_i^2)}$$

對於高維資料點X_i與X_j在低維空間中的對應點Y_i和Y_j，我們同樣能計算一個條件機率$P_{j|i}$。

我們在低維空間中設定高斯分布的變異數為$\frac{1}{\sqrt{2}}$，因此資料點Y_j對資料點Y_i的相似性定義為：

$$q_{j|i} = \frac{exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq i} exp(-\|y_i - y_k\|^2)}$$

如果低維空間中的點Y_i和Y_j正確地表達了高維空間中的資料點X_i與X_j之間的相似性，那麼條件機率$P_{j|i}$與$q_{j|i}$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$p_{j|i}$與$q_{j|i}$之間的差異盡可能最小化。用來衡量$q_{j|i}$對於$p_{j|i}$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：

$$C = \sum_i KL\,(P_i\|Q_i) = \sum_i \sum_j P_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

Pi表示給定資料點X_i之後，對其他所有資料點X_j的條件機率分布，而Q_i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$q_{j|i}$去建模很大的$p_{j|i}$），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P_i。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P_i具備使用者事先設定的「困惑度」(perplexity)。定義為：

$$Perp(P_i) = 2^{H(P_i)}$$

這裡 H(P_i)代表P_i以 bits 為單位的 Shannon 熵：

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

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：

$$\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

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）：

$$\Upsilon^{(t)} = \Upsilon^{(t-1)} + \eta\frac{\delta C}{\delta \Upsilon} + \alpha(t)\left(\Upsilon^{(t-1)} - \Upsilon^{(t-2)}\right)$$

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：

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq 1}(1 + \|y_k - y_l\|^2)^{-1}}$$

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：

$$\frac{\delta C}{\delta y_i} = 4\sum_j (p_{ij} - q_{ij} + y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$$

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$p_{j|i}$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）

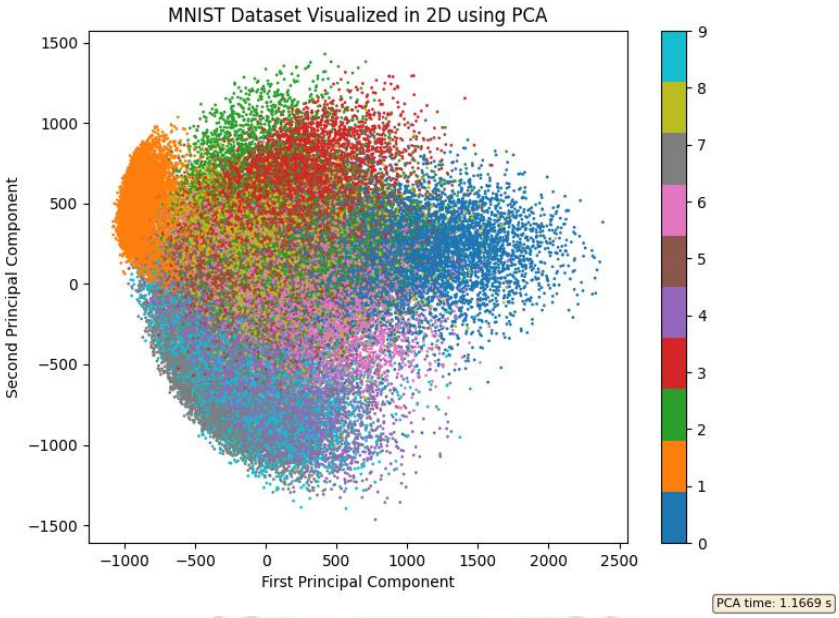
SFP 是一種改進版的頻繁模式挖掘演算法，專為大規模數據處理設計。它的核心概念是改進 Database Projection（DP）方法，透過「關聯分組（Association-based Grouping）」來減少 I/O 操作、提高計算效率，並且適應不同記憶體條件。他的主要特徵有(1)支援動態記憶體管理，意思是根據內存狀況自動調整算法。(2) 減少硬碟讀寫次數，根據文獻此方法相較 DP 方法更加有效率，因為減少了高昂的 I/O 成本。(3) 增強效能，根據文獻表明 SFP 算法在處理複雜數據時的執行時間顯著少於 DP 方法。

2 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**

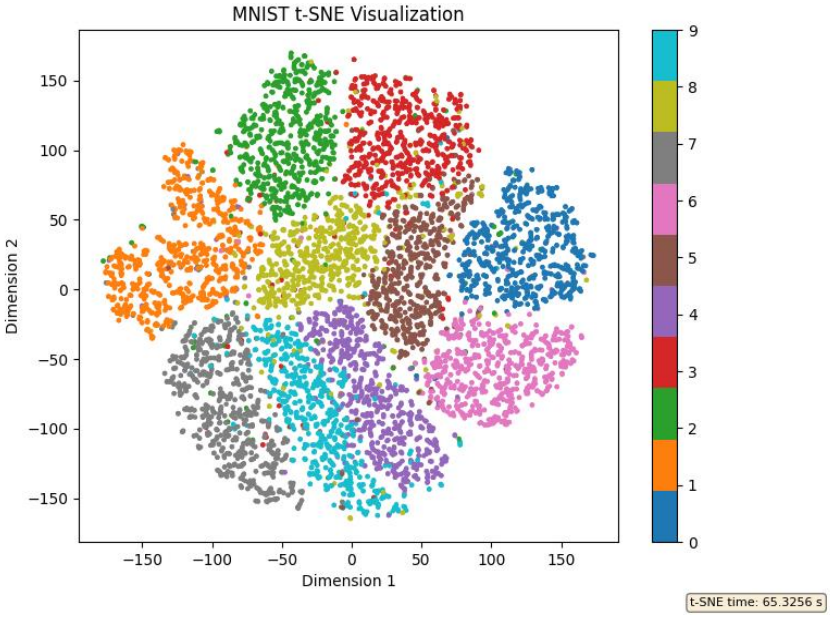
使用 MNIST 資料集完整 70000 筆資料降維至 2D 視覺化，並且記錄執行 PCA 所需時間 ：



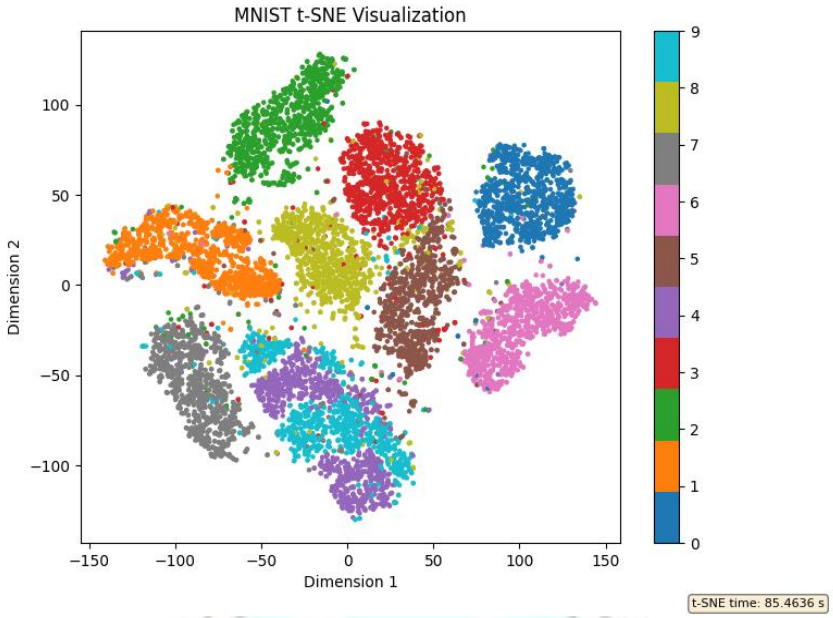
- **只使用 t-SNE**

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：

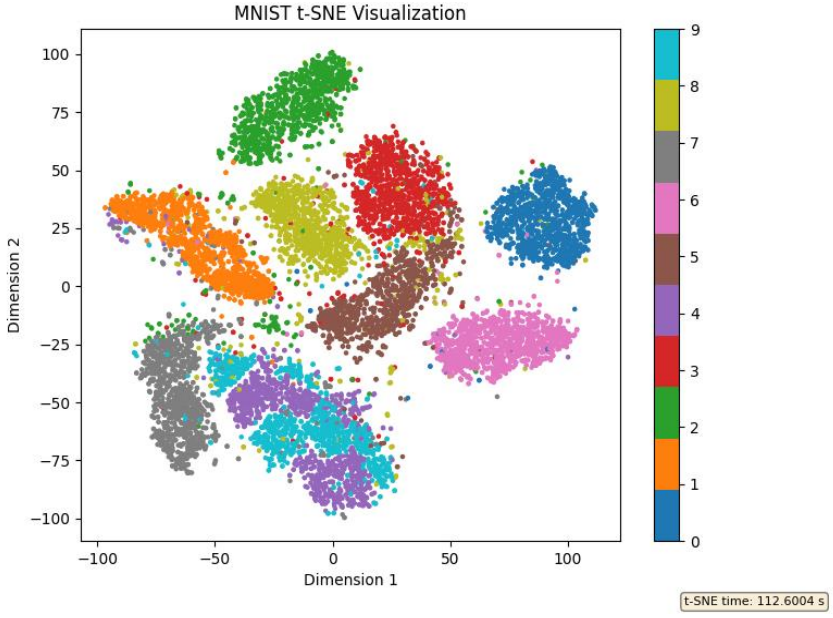
Perplexity 設定 5：



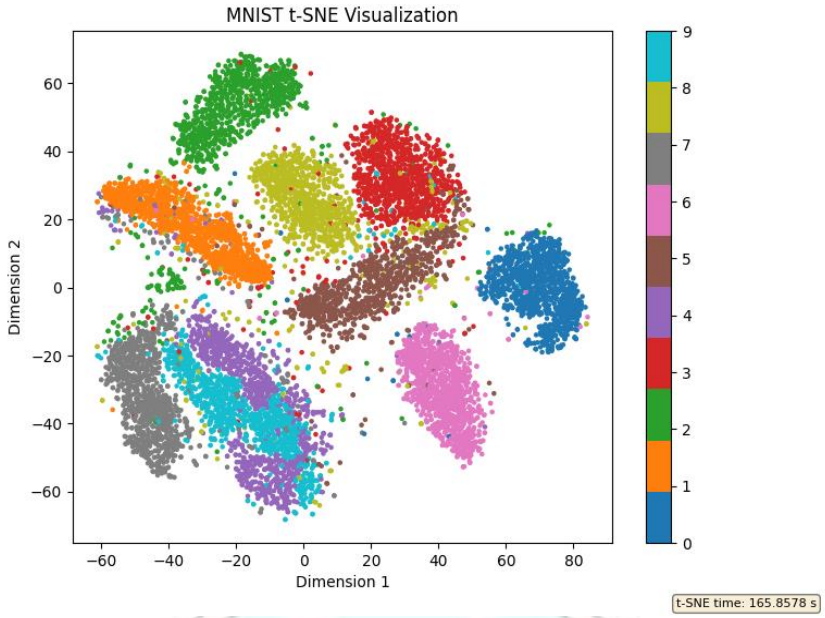
Perplexity 設定 25：



Perplexity 設定 50：



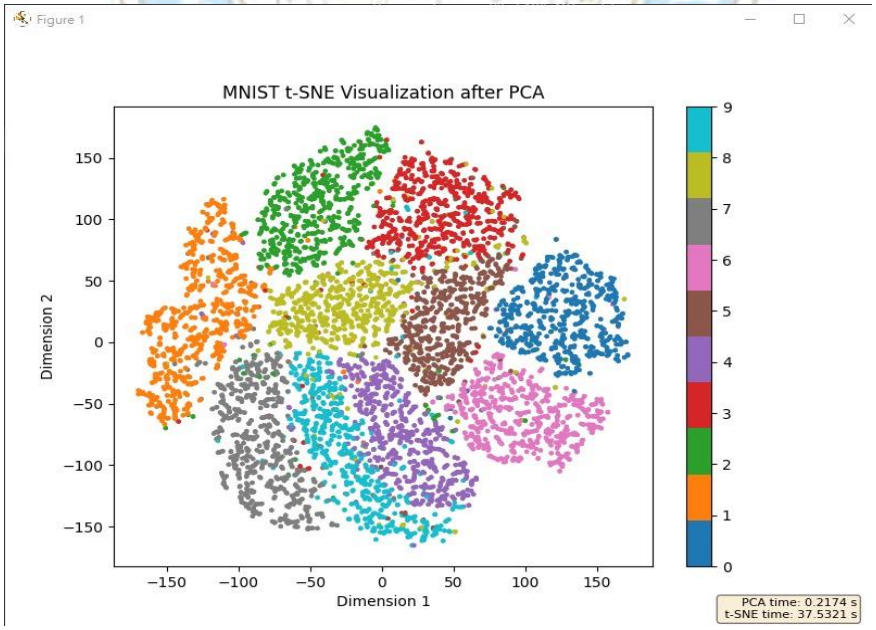
Perplexity 設定 100：



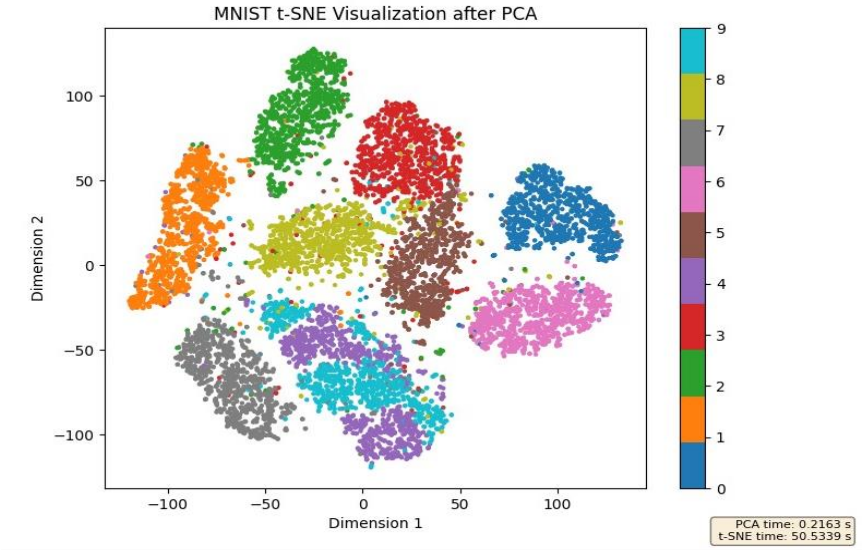
- **先使用 PCA，再使用 t-SNE 視覺化**

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 ：

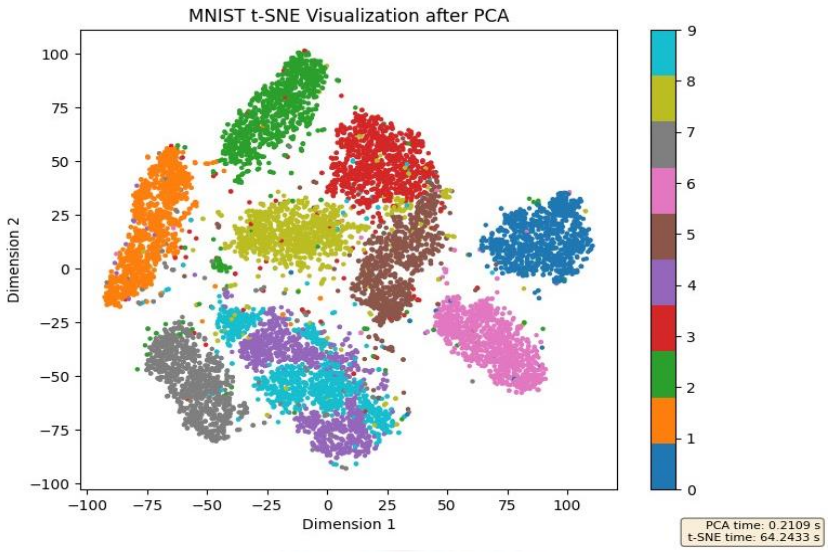
Perplexity 設定 5:



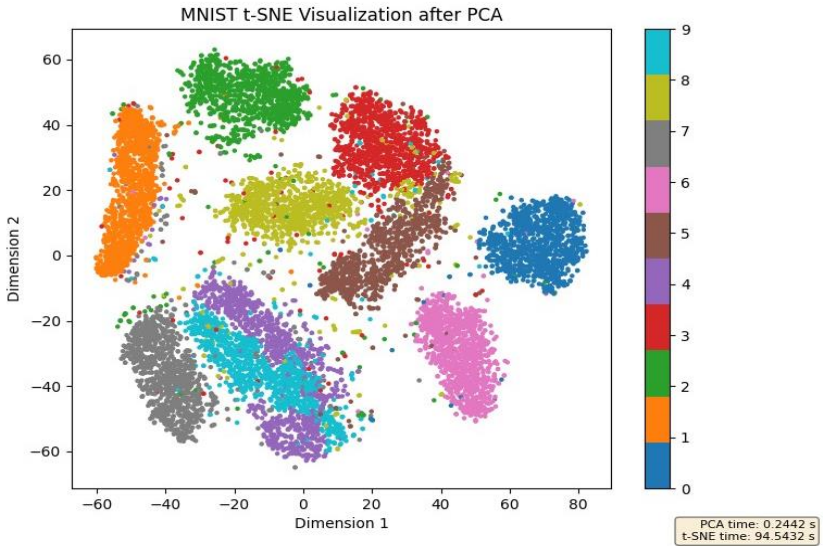
Perplexity 設定 25:



Perplexity 設定 50:



Perplexity 設定 100:



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。

3 效能與準確性對比

4 結果分析與討論

第三章 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	

第四章 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

第五章 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.