Development of Coffee Classification by Feature Selection Algorithm Based on An Electronic Nose

特徵選擇演算法應用於電子鼻的咖啡氣體分類

專題領域：系統領域

組別：A404

指導教授：鄭桂忠

組員：沈亭、羅子翔、謝宇承

摘要

咖啡生豆的品種對於咖啡後製至關重要。咖啡莊園現行僅能以職人針對當下環境的狀況，透過經驗及試誤的方式，對咖啡豆進行分類與後製條件的調整，嚴重影響效率及成本。本專題試圖以一套電子鼻系統的特徵選擇演算法，針對咖啡豆氣味中的氣體分子，分類不同品種的咖啡豆，並自動化與標準化整個過程，降低試誤與人力的成本，提升咖啡莊園的生產量能。先以氣體感測器擷取不同品牌的咖啡豆氣味為數據，並對氣體分子進行初步特徵萃取，萃取出許多特徵後。計算每種特徵的分類指標 SI 值。最後進行 SVM，依特徵的 SI 值由大至小排序，提取不同數量的特徵做多次 SVM，並判斷選擇多少特徵數量時，能使得 SVM 的分類邊界越大，亦即分類準確度越高，以驗證演算法的結果。透過此演算法所找出的 SI 值最大的特徵組合與特徵數量，使得 SVM 的分類邊界產生最大值，分類準確度達100%。本專題使用的演算法對於咖啡生豆的氣味辨識有很高的精確度，得以更精確地對咖啡豆進行分類。

1. 研究背景與目的

咖啡生豆的品種對於咖啡後製至關重要。咖啡莊園現行僅能以職人針對當下環境的狀況，透過經驗及試誤的方式，對咖啡豆進行分類與後製條件的調整，嚴重影響效率及成本。本專題參考 NBME 實驗室學長的論文，和已畢業的學姊的論文，試圖復刻一套電子鼻系統的演算法，針對咖啡豆氣味中的氣體分子，分類不同品種的咖啡豆，並自動化與標準化整個過程，降低試誤與人力的成本，提升咖啡莊園的生產量能。

電子鼻系統是一連串的氣體感測器陣列，能夠根據空氣中不同的氣體分子，經過內部電路的運算系統，使感測器電路的電阻值發生變化，故對於不同的氣體分子，其電阻值有獨特性，也就是「特徵」。然而，電子鼻系統捕捉到的某些特徵，因為有空氣中的雜訊干擾，反而不利於辨別不同氣體。因此，為了更好的氣體辨識度，必須先進行「特徵萃取」。其目的在於保留最有辨識度的特徵。簡化分類問題，並提升模型的分類表現。

經過特徵萃取後，本專題試圖發展一套特徵選擇演算法。先計算不同特徵的分類指標(SI)，再依序納入不同數量的特徵，並降低資料維度，最後應用在 SVM，試圖找出最有辨識度的特徵組合以最佳化分類邊界。為了評估此演算法的效果，我們應用在三種不同品牌的咖啡豆，總共擷取四天的資料集。在本專題中，將以此資料集為樣本，執行多種不同的特徵選擇演算法，並比較最後的準確度與結果分析。

2. 研究方法

2-1. 數據採樣

圖一展示電子鼻系統的行進流程，主要有四大部分：感測器陣列、micro-pump with solenoid valves、氣體分子吸附劑、一組溫度與濕度感測器。感測器陣列由十四種 MOS 氣體感測器組成：TGS2600（氫氣、一氧化碳等）、TGS2602（氫氣、氨氣 ）、TGS2603（三甲基胺、甲硫醇等 ）、TGS2610（丙烷、丁烷 ）、TGS2611（甲烷 ）、TGS2612（甲烷、丙烷、丁烷 ）、TGS2620（乙醇、VOCs）、SB5100（硫化氫 ）、SB5300（氨氣）、SB-AQ1-06（VOCs）、SB-30-04（酒精）、SP-3S-AQ2-01（VOCs）、SP53B00（氨氣）和 SP3-6100（臭氧）。不同感測器負責專門偵測目標氣體。資料採集3/10、3/13、3/15和3/17共四天，A, B,C 三種不同品牌的咖啡豆氣體數據。

2-2. 特徵萃取

在電子鼻系統中，為了有效辨識與區分氣體。我們需要先對感測器中的原始資料進行基本的特徵萃取以降低複雜性，並去除數據中的雜訊，以增加後續演算法的效率與準確度。電子鼻常用的特徵有穩態特徵和瞬態特徵，穩態特徵是氣體和感測器表面的材質反應後達到穩定的狀態，像是大小量值、差值；瞬態特徵是相同時間內氣體流量的變化，像是微分值。因此，我們萃取了六種特徵。最後，因為使用了十四個氣體感測器，而每個氣體感測器取出六種特徵，故每筆資料共有八十四個特徵。氣體感測器的響應曲線如圖一與圖二。

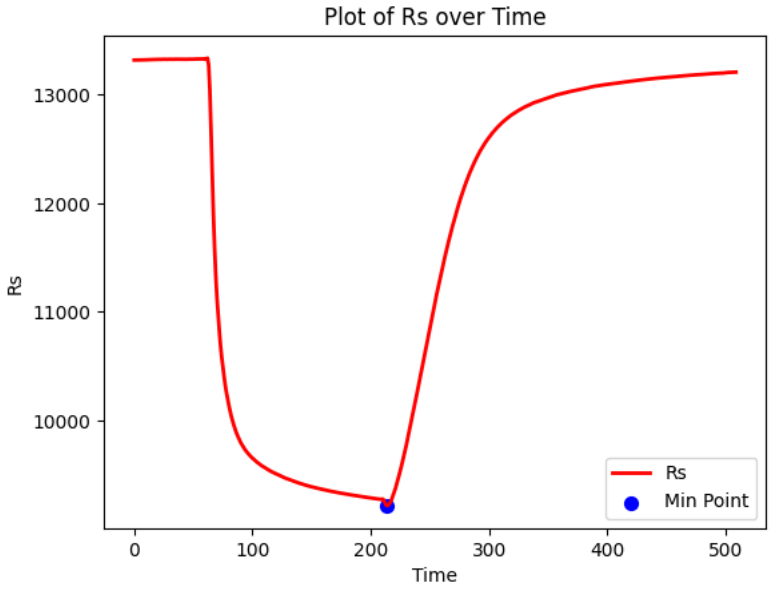


圖 一 氣體感測器的響應曲線

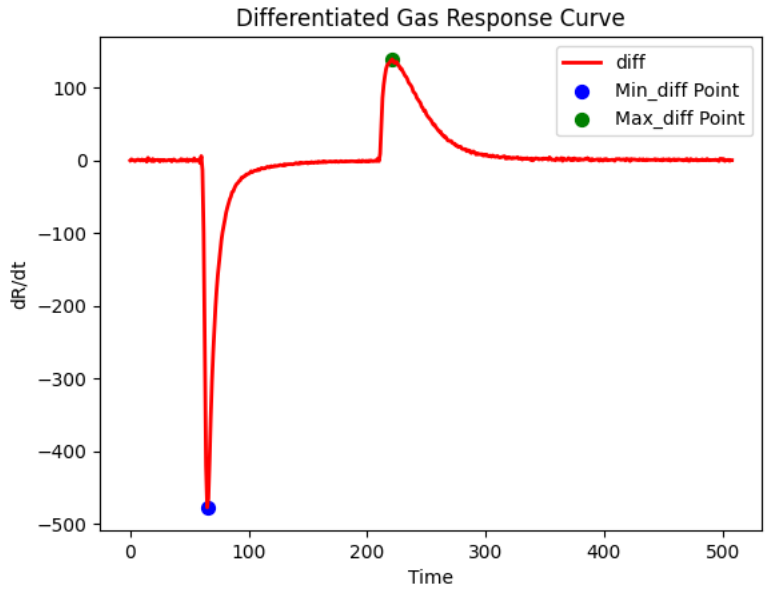


圖 二 氣體感測器響應微分曲線

2-3. 特徵參數

1. 分類指標 Separability Indicator (SI)

為了找出最佳的分類特徵，並評估不同特徵的分類表現。我們試圖計算出各個類別的特徵相關參數：各類別的平均值、變異數、類別間變異數、類別內變異數，以計算分類指標(SI)：

$$\mathrm{SI} = \frac{\sigma_{bc}^2}{\sigma_{wc}^2}$$

σ_{bc}^2：類別間變異數
σ_{wc}^2：類別內變異數

越大的σ_{bc}^2值代表類別間變異越大，亦即分類效果越好。相反的，越小的σ_{bc}^2代表類別間越無辨識度。越小的σ_{wc}^2值代表類別內變異越小，亦即在一個類別內，特徵收斂程度越好，特徵越具有代表性。相對的，越大的σ_{wc}^2代表一個類別內，特徵收斂程度越差，特徵較不具有代表性。因此，結合σ_{bc}^2以及σ_{wc}^2，得出分離指標(SI)。由以上解釋，SI 值越大代表分類的表現越好。

2. SVM 分類邊界

支持向量機(Support Vector Machine)是傳統的機器學習分類器，能有效的進行類別區分和樣本辨識。SVM 的主要功能為找出一個最佳的超平面，可在高維度下，最大程度區分不同類別之間的資料點。

SVM 的兩個主要觀念為支持向量與分類邊界。支持向量(support vector)為最靠近決策邊界（超平面）的資料點，代表最具挑戰的分類樣本，所以是決定決策邊界最關鍵的資料點。邊界(margin)是決策邊界和最近支持向量之間的距離。SVM 的目標為最大化邊界以完成分類任務和確保模型穩定度。

2-4. 特徵選擇演算法

我們同時使用分類指標(SI)與支持向量機(SVM)。有兩個主要的目的，其一為降低數據集的維度（也就是複雜度），以方便進行資料管理與處理。其二為去除多餘雜訊，留下更能精確分類的特徵，並最大化分類邊界以提升分類準確

度。詳細過程如以下特徵選擇演算法所示。另外，在進行 SVM 之前，先進行二維 PCA (2-D Principal Component Analysis) 將高維度的資料集降低至二維。此步驟能確保每個特徵都能擁有相同的維度，並能視覺化結果，使 SVM 在二維平面上計算分類邊界。

表 1 特徵選擇演算法

輸入	Data $X \in \mathbb{R}^{N*M}$, N samples with M features
輸出	最佳特徵集 S
步驟1.	計算每個特徵的 SI 值並依大到小排序
步驟2.	**執行迴圈** $k = 2, 3, \ldots M$ 執行以下步驟
步驟3.	納入k個特徵值為S_k集合，特徵值的選擇依 SI 值由大到小
步驟4.	對S_k進行 PCA 降維至二維
步驟5.	將步驟4.的結果應用至 SVM，並計算分類邊界m_k
步驟6.	**如果** SVM 無分類錯誤，則回傳(S_k, m_k)
步驟7.	**結束迴圈**
步驟8.	選擇最大的$m_{k,max}$及$S_{k,max}$
步驟9.	**回傳** S = $S_{k,max}$

3. 研究結果

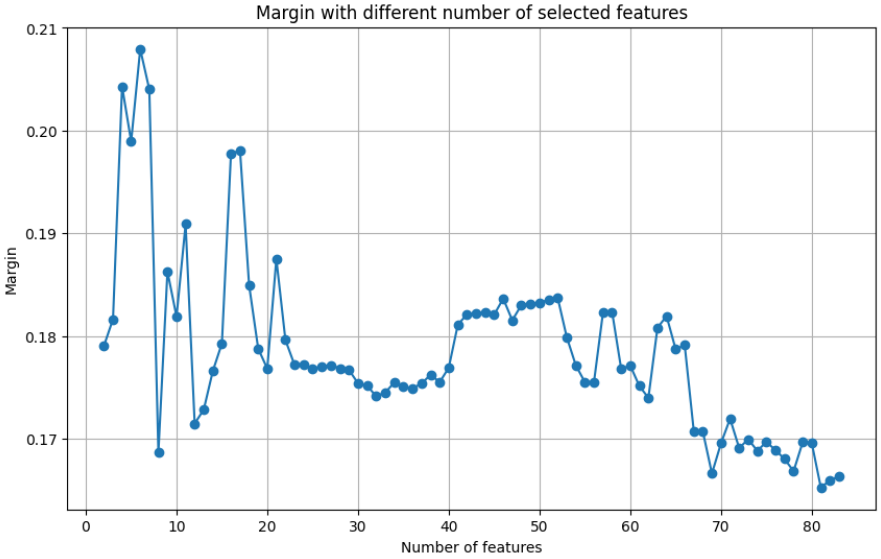
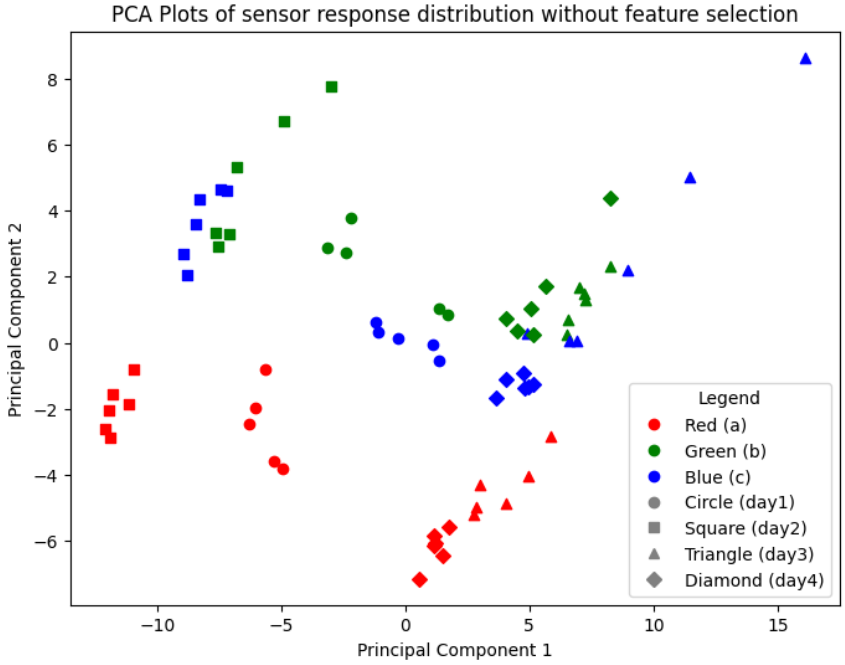
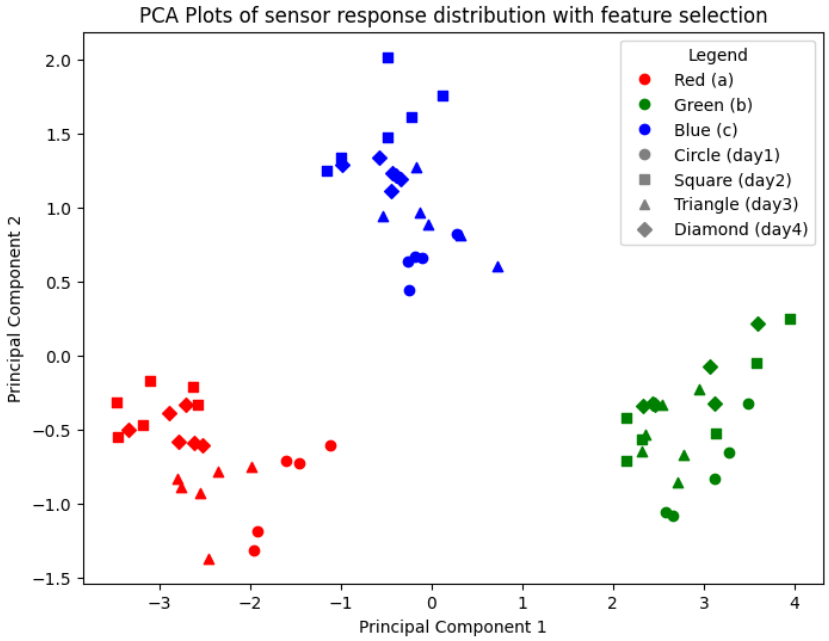


圖 三 選擇不同特徵數的邊界(margin)



(a)



(b)

圖 四　PCA 散佈圖(a)沒有經過特徵選擇(b)有經過特徵選擇

圖三的結果顯示，在特徵選擇演算法執行的過程中，在選擇不同特徵數量的情形下，各經過 SVM 後得出的邊界值(margin)。可以發現選擇六個特徵時會有最大的邊界，能最大化決策邊界與不同類別之間支持向量的距離，也就是 SI 最高的六個特徵最具有分類咖啡的代表性，能最大掌握不同類別間的辨識度。

圖四分別展示了沒有經過特徵選擇和有經過特徵選擇的 PCA 結果。圖四(a)是以所有84個特徵進行主成分分析，圖中顯示在不同類別的分布之間，有很大一部分重疊在一起，代表感測器數據的分類效果很差。對比之下，圖四(b)是以 SI 最高的六個特徵進行主成分分析，展現了在經過特徵選擇演算法後，不同類別之間有很明顯的被區分開來，並且分類邊界也都被最大化。圖四的結果展示了特徵萃取確實能大幅度的提升不同類別的辨識度。

為了更進一步的評估先前演算法的表現，我們使用 SVM 分類器，但分別套用兩種也是非常著名的氣體感測器的特徵選擇演算法：mRMR 和 SVM-RFE 。我們依照順序使用了不同天數的氣體資料作為訓練集，而剩下三天作為驗證結果的測試集。結果如表三所示，展現了不同演算法的精確度，並且隨著不同的演算法，測試集的平均精確度在0.83到1之間。我們的特徵選擇演算法則是達到了最高的分類精確度，驗證其分類的表現。

表 2 不同特徵選擇方法下的分類精確度

	訓練集	測試集	精確度
無特徵選擇	03/10	03/13, 03/15, 03/17	0.72
	03/13	03/10, 03/15, 03/17	0.65
	03/15	03/10, 03/13, 03/17	1.0
	03/17	03/10, 03/13, 03/15	0.94
平均值			0.83
mRMR	03/10	03/13, 03/15, 03/17	0.94
	03/13	03/10, 03/15, 03/17	0.94
	03/15	03/10, 03/13, 03/17	1.0
	03/17	03/10, 03/13, 03/15	0.98
平均值			0.97
SVM-RFE	03/10	03/13, 03/15, 03/17	0.85
	03/13	03/10, 03/15, 03/17	0.80
	03/15	03/10, 03/13, 03/17	1.0
	03/17	03/10, 03/13, 03/15	0.96
平均精確度			0.90

特徵選擇演算法	03/10	03/13, 03/15, 03/17	1
	03/13	03/10, 03/15, 03/17	1
	03/15	03/10, 03/13, 03/17	1
	03/17	03/10, 03/13, 03/15	1
平均值			1

4. 總結

本專題參考實驗室學姊[1]和學長[2]的論文，以一套特徵選擇演算法降低資料維度並最大化分類邊界，應用於電子鼻系統的咖啡豆氣體分類。此演算法結合了分離指標(SI)以及支持向量機(SVM)的分類邊界，試圖找出最具辨識度的特徵，並應用於資料集：三種不同品牌的咖啡豆，共採樣四天。結果展示了無論是在分類精確度，還是在分類邊界上，這套特徵選擇演算法的準確度都優於 mRMR[6]和 SVM-RFE[7]這兩套常用的氣體辨識與分類演算法。因此，本專題的特徵選擇演算法提供了一套更有效率與可靠性更高的方式以辨別咖啡豆的氣體。期許未來能夠實際應用於咖啡莊園，提升咖啡莊園的生產量能，也期許未來的研究能夠更進一步優化這套演算法，甚至應用在其他的領域。

心得感想

鄭桂忠老師於上學期初安排兩位具有豐富實作經驗的碩班學長指導我們。上學期的前半段，學長們指定我們閱讀實驗室相關題材的論文，並讓我們於每兩個禮拜一次的會議中報告，確保我們對於論文的理解是正確的，使我們具備機器學習及其演算法的基本觀念。上學期的後半段，學長們讓我們實作基本的 Python 演算法，包含但不限於本專題使用的 PCA 和 SVM-RFE，並在會議中不吝指正我們的錯誤，與提供明確的指引。寒假時讓我們繼續練習其他的演算法，並推薦閱讀其他論文，以加深我們的底子。下學期我們便全力投入於本專題的研究。因為有了先前的基礎知識與實作經驗，再加上學長們提供的論文也給了我們明確的指引，並改成每個禮拜一次的會議頻率，掌控我們的狀況，最後產出本專題。在繳件前，學長們與鄭桂忠老師也不厭其煩的細心審閱報告，提供我們疏忽的細節與更精確的陳述方式。

一路上非常感謝老師與學長們的指導。除了到學習電子鼻系統的原理、感測器原理、系統實作、程式設計，以及系統於實際場域的應用，使我們得到了難能可貴的實作經驗，並產出此份專題。期許未來在研究所能在這門領域繼續努力。