利用機器學習與集成學習快速偵測多重抗藥性與各年份質譜訊號差異分析

Wei- Tso Chen（陳維佐）

Department of Applied Mathematics, National Sun Yat-sen University

摘要

本研究旨在利用 Python 取代過去繁瑣醫療檢測過程，實現一個能更直接且完整，預測患者體內是否具有超級細菌的研究。本研究利用 Python 套件 Pytemics 替代過去廣泛使用的 mMass 系統，將質譜數據 mzml 檔轉換成表格形式的 csv 檔，並利用聚類演算法 forward alignment 校正峰值。最後用 stacking 的方式，集成了九種機器學習模型，對患者體內是否具有超級細菌做預測，並在初步實驗中達到了約 91%的準確度，證明了利用 Python 取代過去繁瑣醫療檢測過程的可能性。

研究中，我們使用了不同的峰值聚類演算法、增加樣本數量等方法，試圖提高模型的準確度。我們發現 Forward Alignment 聚類演算法在絕大多數的情況，能達到更高的準確度；但增加樣本數，準確度則不一定會上升，甚至可能導致準確度下降，為此我們特別對各年病患質譜資料質赫比（m/z)之分佈做比較，發現某些不同年份患者的質譜質赫比分佈不盡然相同，且用年份跨度過大的資料中取出的代表性峰值做矯正的情況下，有無抗藥性之菌株的峰值訊號並沒有明顯差異，這可能是導致廣增資料，反而使模型準確度降低的原因。這意味著，在模型部署上，需要定期去檢視模型、更新模型，以保持預測結果的品質。

最後，我們用了 2021 年 1 月到 2021 年 12 月之資料訓練了一個混合 SVM、ADABoosting、Logistic Regression、Logistic Regression with LASSO 等四種機器學習模型的 stacking 集成學習模型，在對 2021 年 1 月到 2021 年 12 月之資料做預測的實驗中得到了 83%的準確度；對 2022 年 1 月到 2022 年 12 月之資料做預測的實驗中得到了 76%的準確度，並成功部署並上線，供高雄榮民總醫院做相關使用與測試。

關鍵詞：Pyteomics 套件、質譜數據分析、峰值校正、聚類演算法、監督式機器學習、集成學習