

小規模免疫実験データを学習に用いた機械学習による抗体価の評価		
黒田 研究室	学籍番号: 23261505	氏名 長谷川大樹

[背景・目的] バイオ医薬品開発では動物実験による効率的な条件探索が求められる。効率的な条件探索は開発コストの削減や実験期間の短縮、使用する実験動物数の削減につながる。特に免疫系に作用する場合、多くのマウスが使用される。本研究室では、小さなタンパク質でも凝集によって免疫原性が向上することを明らかにしている。現在、ペプチドタグの付与やアミノ酸変異、金属添加を行い、マウスで免疫応答を調査している。本研究では実験期間の短縮や実験動物使用数の削減を目指し、小規模免疫実験データの解析手法の開発を目的とした。

[材料・手法] 日本脳炎ウイルスのエンベロープタンパク質第三ドメイン(JEV ED3)を用いた 84 件の免疫応答実験データを解析した。このデータは同じ実験条件でも抗体価が大きく異なるデータが多く、ノイズの大きいデータセットである。本研究では、ランダムフォレストという単純な条件分岐を繰り返すことで予測を行うモデルを使用した。入力データは実験条件で、カテゴリ変数に対しワンホットエンコーディングを行った。目的変数は抗体価で、35 パーセントタイルを閾値とし、閾値以上を High、未満を Low の二種類にクラス分けし二値分類を行うモデルを作成した。入力には実験条件が同じデータが含まれるため、訓練データとテストデータに同じ実験条件が入らない調整した。3 分割交差検証を行い、各分割で Accuracy、ROC-AUC、F1Score を計算しその平均を実験データ全体のスコアとした。また、学習曲線から訓練データ数とスコアの関係性を解析した。

[結果・考察] High と Low をランダムに出力するダミーモデルとハイパーパラメータという事前に設定できるパラメータをほとんど設定せずに学習させたモデルとパラメータを調節して学習させたモデルの Accuracy, ROC-AUC, F1Score を比較すると大きな差が見られた(図 1)。このことから学習させたモデルは何らかのパターンを学習し、分類性能を発揮していることがわかる。特徴量重要度(図 2)を見ると、インジェクションするタンパク質の量が増加すると抗体価が上昇することや水和半径が大きくなると抗体価が上昇すること、コバルトが添加されると抗体価が上昇するという経験則に合致した特徴量が高い重要度を示しており、背景知識と一致するような学習をしていることが示唆される。

[結論・展望] 一般的な機械学習では大量のデータを学習に使用する。対して本研究では小規模かつノイズの大きなデータセットを用いて学習を行ったが、高いスコアを出すことができた。今後は JEV ED3 以外のタンパク質の免疫実験データも含めて学習を行い、より汎用性の高いモデルの作成を目指す予定である。

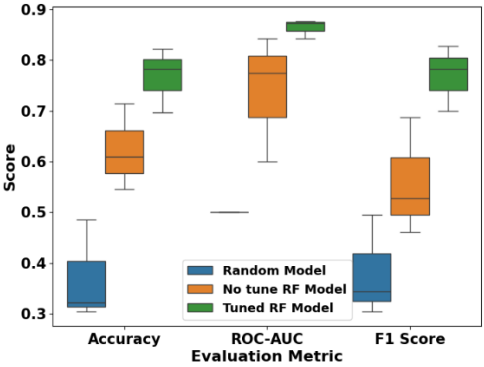


図 1. モデルのパフォーマンス比較

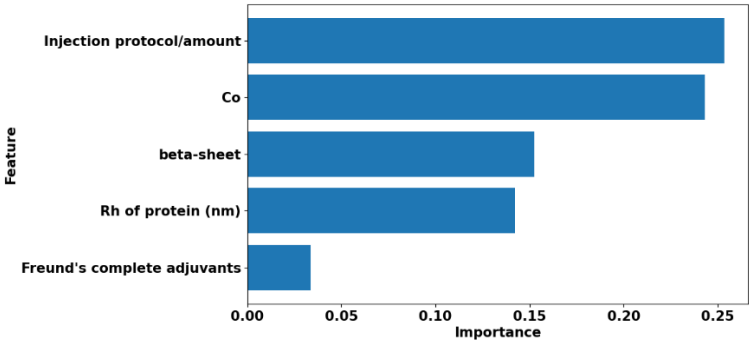


図 2. 特徴量重要度上位 5 件