28X-pm12

少数サンプルにおける活性予測モデルの性能評価および精度向上

○清水 直斗[1], 金子 弘昌[1]（[1]明大理工）

近年、医薬品の研究開発費が増加している中で、コンピュータを活用した効率的な創薬研究が注目されている。本研究では、化合物の活性を目的変数 y、その化合物の構造記述子を説明変数 X とし、コンピュータを使った相関モデルの構築を行う。一般的に、サンプルが少ない場合、テスト用のデータでの検証ができず、モデル構築用データにのみ適合した予測性能の低いモデルが構築される危険性が高い。本研究では、クロスバリデーションを入れ子構造にして 2 回行うことで、クロスバリデーションよりも相関モデルの予測性能を適切に評価できるダブルクロスバリデーションを使い、サンプル数が少ない中で高性能な相関モデル構築を目指す。

今回は、729 個の化合物の Human NAD-dependent protein deacetylase sirtuin 1（Sirtuin 1）のアセチル化に対する阻害度 IC50 の対数を y、RDKit を用いて計算された構造記述子を X としたデータを使用した。

化合物データから少数のサンプルをランダムに選び、線形回帰手法と非線形回帰手法で相関モデルを構築した。それぞれの回帰手法におけるモデル構築用データの IC50 予測結果を、ダブルクロスバリデーション推定値による決定係数 R_{dcv}^2、Root Mean Squared Error （$RMSE_{dcv}$）を指標として評価した。モデル構築に使用したサンプル以外のデータの IC50 を各モデルで予測して結果を比較することで、ダブルクロスバリデーションによる評価の検証を行い、その有効性を確認した。