28X-pm13

回帰分析における精度向上のための化学構造データ選択手法

○菅野 泰弘[1], 金子 弘昌[1]（[1]明大理工）

莫大な医薬品開発コストの削減およびヒット率の上昇を目的として、近年コンピュータを利用して創薬をサポートすることに高い関心が寄せられている。化合物の活性を目的変数 **y**、化学構造の情報を数値化した構造記述子を説明変数 **X** とし、2 つの間の関係を数値モデル化することで、仮想的な化学構造の活性値を実験することなく推定することが可能となる。**X** の値はソフトウェア等により計算できることが多いのに対して **y** の値を得るには実際に実験しなければならないため、**y** の値が測定されている化合物(教師ありデータ)が少なく、**y** の値が測定されていない化合物もしくは化学構造(教師なしデータ)が多く存在する。教師なしデータを有効活用して学習を行う方法は半教師あり学習と呼ばれ近年盛んに研究されている。教師ありデータと教師なしデータを合わせて主成分分析によりデータを低次元化し、得られた主成分と **y** との間で教師ありデータを用いて部分的最小二乗法により回帰モデルを構築する手法 PCAPLS がよく知られている。しかし、半教師あり学習を行う際に使用する教師なしデータに関する議論は進んでいない。回帰モデルの精度に悪影響を与える教師なしデータを含めてしまう可能性がある。

本研究は、回帰モデルの性能が十分に発揮されるデータ範囲である、モデルの適用範囲に着目した教師なしデータ選別手法を提案する。モデル構築用データのデータ密度の高い領域から教師なしデータを選択する。PCAPLS および提案手法を用いて、1213 個の化合物に関する 50%成長阻害濃度(IGC50)データと医薬品候補になりうる約 50 万化合物のデータを解析したところ、提案手法を用いることで PCAPLS よりも回帰モデルの推定性能が向上したことを確認した。