機械学習を利用した腸内細菌データの関連解析法

九州大学病院メディカルインフォメーションセンター　奥井佑

近年, 次世代シークエンサーを用いて細菌のゲノム情報を網羅的にシークエンス解析するメタゲノム解析によって細菌叢の多様性や機能情報が得られるようになった. メタゲノム解析のうち, 細菌の 16S rRNA 遺伝子をシークエンスして得られる細菌組成データをもとに, 対象者の呼吸器や腸内において各菌種がどの程度の割合で存在するかの情報が得られる. 現在, 細菌データを利用して疾患有無と細菌種との関連解析が広く行われるようになり, 海外においては細菌データの統計解析手法に関する研究も急速に進展している.

細菌データは, 各対象者について, 各菌種がシークエンサーをもとにいくつ検出されたかを示すデータである. 統計学的はゼロが過剰な多変量のカウントデータといえるが, 各カウント値は対象者内での相対的な数を示すものであり, 一般的に前処理として割合データに変換するなどのデータの正規化を施す. 正規化法のうち rarefying 法と呼ばれる手法では, 各対象者のカウントデータから一定数のカウント数をサンプリングし, 各対象者のカウント値を同一とする. rarefying 法の欠点は各対象者のデータの一部を解析に用いないことである. ただ, 集団学習の方法論を応用して, データに rarefying 法を施し機械学習モデルを作成する工程を複数回繰り返すことで rarefying 法の欠点を克服できる可能性が考えられたため, 複数の機械学習法を用いて検証を行った. シミュレーション実験の結果, 提案法を用いることで既存手法よりも正確にアウトカムと関連する菌種を特定できると同時に, アウトカムを正確に予測するモデルを作成できる可能性が示唆された.

本発表では, 近年体系化されてきた細菌データの解析手法と解析を行う上での統計学的課題について概説するとともに, 正規化法の一つである rarefying 法と集団学習を組み合わせた関連解析手法の性能を評価した研究の内容について述べる.

参考文献

・Xia Y, Sun J, Chen DG (2018). ICSA Book Series in Statistics Statistical analysis of microbiome data with R. *Springer: Nature Singapore Pre Ltd*.
・Okui T, Matsuyama Y, Nakaji S (2019). A new association analysis method for gut microbial compositional data using ensemble learning. *Japanese journal of biometrics*,**39**,55-84.