［課程－２］

審査の結果の要旨

氏名　市川太祐

本研究は機械学習の手法を用いた健診結果予測モデルを構築・評価するため、複数の機械学習の手法を用いて現在の高尿酸血症の罹患有無を、特定健診の必須項目から予測するモデルを構築し、性能評価を行った上で最適な予測モデルの選択を目指したものであり、下記の結果を得ている。

1. 高尿酸血症の定義を 7mg/dl 、9mg/dl として、各基準において機械学習の手法として Gradient Boosting Decision Tree（GBDT)、ランダムフォレスト（RF)、L1 正則化ロジスティック回帰（LR)、Stacking 法（STACK）を用いて予測モデルを構築し、予測性能について比較評価を行った。予測モデルの予測性能を示す AUC（Area Under the Curve）が最高になった手法は前者が GBDT 及び STACK（ともに AUC 0.7）で後者が LR（AUC 0.8）であった。医学分野における予測モデルの AUC 基準に照合すれば、LR による高尿酸血症（9mg/dl）に関する予測モデルは妥当な予測性能を有していると判断された。
2. 今回用いた機械学習の手法間において構築した予測モデルの予測性能に差は無く、また少ない説明変数で予測が実行でき、過学習を起こしにくいことから、予測モデルの構築に用いる手法としては LR が最適であることが示唆された。
3. 構築した健診結果予測モデルのうち、GBDT、RF、LR を用いてモデル構築における必要データサイズの削減を試みた。LR による変数選択結果は全 39 の説明変数のうち、5 つのみが選択されるという結果であった。また、GBDT、RF、LR のいずれにおいても多数派データを元の 1.4%まで削減しても予測性能を示す AUC、F 値に大きな低下は認められなかった。したがって、予測モデル構築に必要なデータは、説明変数の観点からは前年の健診結果は不要であり、データ数の観点からは総データ数の 2.8 %まで削減できることを示した。

以上、本論文は機械学習の手法を用いることで高尿酸血症罹患を他の健診項目からある程度予測可能であり、その予測に必ずしも大量のデータが必要ではないことを示した。本研究は健診データに対する機械学習の適用について重要な貢献をなすと考えられ、学位の授与に値するものと考えられる。