論文の内容の要旨

論文題目　機械学習を用いた健診結果予測モデルの構築および検証

氏名　市川太祐

目的

日本においては第二次世界大戦後の結核対策を契機に、疾病スクリーニングを中心とした疾病予防施策がとられてきた。当初、スクリーニングは結核等の感染症を対象としていたが、衛生状況の改善に伴い、生活習慣病を代表とする慢性疾患を対象とする健康診断・健康診査（以下、健診）へと変わっていった。2008 年 4 月から、高齢者の医療の確保に関する法律により、医療保険者に対して、内臓脂肪の蓄積等に着目した生活習慣病に関する健康診査（以下、特定健診）及び特定健診の結果により健康の保持に努める必要がある者に対する保健指導（以下、特定保健指導）の実施が義務づけられた。さらに保険者には、健診結果の電子化及び保存が義務付けられた。これにより日本においては 2008 年以降莫大な健診結果データが蓄積されている。

一方、健診項目は保険者によって異なるという現状がある。健診は法制度に基づき受診が必須となっている健診項目（以下、必須項目）と保険者の判断により提供している健診項目（以下、非必須項目）で構成されている。加入する保険者によって提供される健診項目が異なっている事実は、加入者が有用な疾病スクリーニングを受ける機会を失っているという見方もできる。

そこで本研究は非必須項目から得られる罹患の有無を必須項目から予測するモデルを構築することを考えた。予測モデルの構築には保険者から得られた健診データを利用する。構築したモデルを非必須項目の検査を実施していない保険者が利用することで、最低限の健診項目しか受診できない保険者の加入者に対してもコストを抑えた形で現状の疾病スクリーニングを補完できることが期待される。

以上より、本研究では、機械学習の手法を用いた健診結果予測モデルの構築とその性能評価を目的とする。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す。

方法

機械学習を用いた健診結果予測モデルの構築

本研究では機械学習を用いて特定健診の検査項目を説明変数として高尿酸血症か否かを目的変数として予測する予測モデルを構築した。

予測モデル構築・検証にあたって 1 健康保険組合の協力を得てデータを取得し、予測モデル構築用の訓練用データセットとモデルの検証を目的とした検証用データセットに分割した。最終的に訓練用データセットには 43,524 人、検証用データセットは 17,789 人のデータとなった。

健診で取得したデータ項目は年齢、BMI、高血圧関連項目として収縮期血圧と拡張期血圧、糖尿病関連項目として空腹時血糖と HbA1c、脂質障害関連項目として中性脂肪、HDL コレステロール、LDL コレステロール、肝機能障害関連項目としてγGTP、GOT、GPT、そして血清尿酸であった。血清尿酸以外の項目については 2 年分の値（前年、後年）及びその差を予測モデルの説明変数として利用した。また検査結果から計算できる脈圧（収縮期血圧と拡張期血圧の差)、GOT/GPT 比、LDL コレステロール/HDL コレステロール比も説明変数として利用した。

データセット内における高尿酸血症の判定は学会基準に従って、診断基準の 7.0 mg/dl と薬物療法が開始する基準の 9.0 mg/dl を用いた。

機械学習の手法については Gradient Boosting Decision Tree（以下、GBDT)、ランダムフォレスト（以下、RF)、L1 正則化ロジスティック回帰（以下、LR）及び Stacking 法（以下、STACK）を比較して評価した。評価指標には AUC（Area Under the Curve)、感度、特異度を用いた。AUC については 95%信頼区間（以下 95%CI）も算出した。

学習データサイズの削減

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。削減の方針としては 1) 説明変数の削減と 2) データ数の削減の 2 つを検討した。

1) については LR を用いて予測に必要な説明変数の選択を行った。2) についてランダムアンダーサンプリング法を採用した。ランダムアンダーサンプリングは少数派データ（本研究においては高尿酸血症患者のデータ）はそのままに多数派データ（本研究においては非高尿酸血症患者のデータ）からランダムにデータを抽出してデータを削減する手法である。

2)については GBDT、RF、LR を用いて訓練用データにランダムアンダーサンプリング法を適用し、高尿酸血症患者が占める割合を 10%から 50%まで 10%刻みで変化させ、予測性

能がどのように変化するかを検証することとした。予測性能評価には AUC、感度、陽性的中率、F 値を用いた。また高尿酸血症の閾値については前章と同様に 9mg/dl を用いることとした。

結果

機械学習を用いた健診結果予測モデルの構築

高尿酸血症の閾値を 7mg/dl とした場合、各予測モデルの AUC は GBDT で 0.70[95% CI: 0.69‐0.71]、RF で 0.66 [95% CI: 0.65‐0.67]、LR で 0.69 [95% CI: 0.69‐0.70]、STACK で 0.70 [95% CI: 0.69‐0.71]であった。特異度は RF が最も高く、次いで STACK、LR、GBDT の順であった。感度は GBDT、LR、STACK がほぼ同等で、RF がそれに次いだ。

高尿酸血症の閾値を 9mg/dl とした場合、各予測モデルの AUC は GBDT で 0.76 [95 % CI: 0.72‐0.79]、RF で 0.78 [95 % CI: 0.75‐0.81]、LR で 0.80 [95 % CI: 0.77‐0.83]、STACK で 0.78 [95 % CI: 0.74‐0.81]であった。特異度は LR が最も高く、ついで RF、STACK、GBDT の順であった。感度は GBDT と STACK がほぼ同等で、LR、RF の順に高い値となっていた。

今回用いた手法間において予測性能の差は少なく、また少ない説明変数で予測が実行できることから、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。

学習データサイズの削減

説明変数の削減については、全説明変数 39 のうち、標準偏回帰回帰係数の絶対値が 0 以上となった説明変数の数は 5 であった。標準偏回帰回帰係数の絶対値が最大だった説明変数は BMI（後年）であり、最小は拡張期血圧（後年）であった。また選択された説明変数はいずれも後年の健診項目であった。

データ数の削減については、ランダムアンダーサンプリングを用いて多数派データを削減し AUC の変化を確認した結果、多数派データのデータ数を元の 42910 からポジティブデータと 1 対 1 になる 614 まで削減しても AUC、F 値に大きな変化は認められなかった。

考察

機械学習を用いた健診結果予測モデルの構築

今回設定した高尿酸血症の定義に関わらず、機械学習を用いて構築した予測モデルは 4 つのモデルの間でその性能に大きな差は認められなかった。一般に集団学習を用いた予測モデルは仮説空間が多岐にわたる、つまり説明変数の数が膨大で、その説明変数間に複雑な交互作用が認められるような場合に他手法を凌駕する性能を示すことが報告されている。

今回、決定木を弱学習器として用いた RF と GBDT の 2 つの手法と LR との間に大きな差は無かったことから、健診データにはそのような特性は認められなかったものと考えられる。また、交互作用が認められる場合は、クロスバリデーションを用いて決定した木の深さのパラメータに反映される。最終的なモデルのパラメータをみると、AUC が高かった 9mg/dl の場合では GBDT の場合は 3、RF は 2 であり、個々の決定木は深くなく、複雑な交互作用を仮定しないモデルの予測性能が高くなかったと考える

なお、高尿酸血症の定義を 9mg/dl とした場合の予測モデルにおいて AUC が最高になった手法は LR でその AUC は 0.80 だった。予測モデルの基準に寄れば moderate accuracy であり予測モデルとしては妥当な結果といえる。

学習データサイズの削減

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。これらの選択された説明変数はいずれも後年の検査項目であり、前年の検査項目は選択されなかった。本研究においては健診結果の経年変化も考慮して予測モデルに説明変数として組み入れたが、高尿酸血症予測モデルにおいては単年度の健診結果のみで予測が可能という結果になった。

アンダーサンプリングを用いて多数派データを削減した結果の AUC、F 値の変化を確認した。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 AUC、F 値に大きな低下は認められなかった。これにより多数派データの多くが予測性能には影響を及ぼさないデータであることが示唆された。

結論

本研究では保険者より収集した健診データに機械学習の手法を適用し、高尿酸血症罹患を例に、特定健診の非必須項目から判定される疾患の罹患の有無を必須項目から予測するモデルを構築した。最良の手法は LR であり、その予測性能は AUC 0.8 で予測モデルとして妥当な結果が得られた。また、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。今後は本研究を通して得られた予測モデル構築のアプローチを用いて他の検査項目を対象とした予測モデルを構築する。