ポスター
ポスター1
医療データ分析１（ NDB・レセプト）
2018年11月23日(金) 14:20 〜 15:10 J会場(ポスター) (2F 多目的ホール)

[2-J-1-2] 疾病罹患有無予測モデルにおけるデータバランスの検討
○市川 太祐, 斎藤 季, 香山 綾子, 小山 博史 （東京大学大学院医学系研究科臨床情報工学分野）

日本においては2008年の特定健康診査・特定保健指導制度（以下、特定健診）の開始以降保険者において、莫大な健診結果データが蓄積されている。健診データを予測モデルに利用する際、非患者データが多数派データ、患者データが少数派データとなる。したがって健診データを用いた予測モデルを構築する際、多数派データへの過学習を防ぐためにサンプルサイズの不均衡を調節する必要がある。本研究では、特定健診データを用いて高尿酸血症の罹患有無を予測するモデルを構築し、不均衡の調節を行った上で、予測性能を保持できるデータバランス（多数派データ-少数派データ比）について検討を行った。予測モデル構築には特定健診データ61,313件を用いて、訓練用データと検証用データがそれぞれ70%、30%になるよう分割した。手法としては 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）を用いた。評価として訓練用データにおいて多数派データが占める割合を10 %から、少数派データと同数の50 %まで10 %刻みで変化させ、検証用データに対する予測性能がどのように変化するかを検証した。不均衡の調節には多数派データに対するアンダーサンプリングを、予測性能の評価指標には Area Under the Curve（ AUC）を用いた。結果として少数派データとほぼ同数まで多数派データを削減しても予測モデルの予測性能に大きな変化は認められず、予測モデル構築におけるデータ収集については少数派データの収集が重要であることが示唆された。

疾病罹患有無予測モデルにおけるデータバランスの検討

市川太祐, 斎藤季, 香山綾子, 小山博史,
東京大学大学院医学系研究科臨床情報工学教室

Daisuke Ichikawa, Toki Saito, Ayako Kohyama, Hiroshi Oyama
Dept. of Clinical Information Engineering, University of Tokyo

Abstract
There is an imbalance classification problem between patient data and non-patient data in building prediction models from the health check-up data. It is required to correct the imbalance to prevent from the overfitting to the non-patient data. In the present study, we built the classification model for the hyperuricemia from data of the Specific Checkup in Japan and examined the appropriate balance. Machine learning approaches were used to build the classification model and random under-sampling was used for the correction for the imbalance. Area Under the Curve (AUC) was used for the metric of the model. The health check-up data was split to two datasets, the training dataset and the validation dataset. The number of the training dataset was 43,524 (the number of the hyperuricemia was 614) and the number of the validation dataset was 17,789 (the number of the hyperuricemia was 215). When non-hyperuricemia patient data was under-sampled from full size to 1:1 (patient : non-patient), the decrease of AUC were 0.01 in the classification models. The result indicates the collection of patient data was important to build classification models from health check-up data.

Keywords: Health check, Hyperuricemia, Prediction model, Machine learning, Imbalanced data

1. 緒論

日本においては 2008 年の特定健康診査・特定保健指導制度（特定健診）の開始以降、保険者において、莫大な健診結果データが蓄積されている。蓄積されているデータは特定健診で受診が必須となっている項目（必須項目）のデータが中心となるが、保険者によっては血清尿酸等の受診が必須ではない項目（非必須項目）を必須項目に加えて健診メニューとして提供し、そのデータを蓄積している。

我々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検討してきた[1][2]。構築したモデルを非必須項目の検査を実施していない協会けんぽ及び国民健康保険等の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。

一方で、健診データを用いた予測モデルを構築する際、2 つの課題がある。1 つは健診データにおける不均衡である。予測対象の疾病に罹患している患者データは健診データ全体に占める割合が非常に少ない不均衡データになっている。このような不均衡データにおいて、不均衡の調節を行わずに予測モデルを構築した場合、非患者データに過学習してしまうという問題が知られている[3]。2 つ目の問題としてデータ確保の問題が挙げられる。予測モデルを構築する際には保険者に健診データの提供を依頼するが、保険者側としては必要なデータのみ外部に提供したいという要請がある。したがって、予測モデルに必要なサンプルサイズをあらかじめ見積もった上でデータ提供を依頼する必要があるが、予測モデル構築におけるサンプルサイズの見積もりについては明確な指針は無い。

2. 目的

本研究では、不均衡の調節を行った上で、予測性能を保持できるサンプルサイズについて検討を行った。

3. 方法

予測モデルの構築には1健康保険組合から取得した健診データを用いた。被保険者で 40 歳から 60 歳の男性で服薬歴が無く、かつ 2011 年から 2013 年の3年間にわたって毎年健診を受診した集団を本研究の対象とした。女性は高尿酸血症の有病率が非常に少ないため、研究対象からは除外した。最終的に本集団の全データ件数は 61,313 人となった。

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。

本研究では予測モデルの予測対象を高尿酸血症の罹患有無とした。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、薬物療法が開始する基準の 9.0 mg/dl を用いた。

予測モデルを構築する手法については 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）を用いた。全ての予測モデルは R を用いて構築した。予測モデルのパラメータについてはグリッドサーチを用いて決定した。予測モデルを構築する際の健診データの欠測値補完には knnImpute 法を用いた[3]。

予測モデルの構築・評価に際して、健診データを訓練用データセットと検証用データセットの2つに分割した。それぞれ、訓練用データは予測モデルの構築に利用し、検証用データは構築した予測モデルの性能の検証に用いた。分割にはランダムサンプリングを用いて、70%を訓練用データ、30%を検証用データとした。最終的に訓練用データは 43,524 人、検証用データセットは 17,789 人のデータとなった。

評価に際しては、訓練用データにおいて高尿酸血症患者データ（少数派データ）が占める割合を 10 %から、非高尿酸血症患者データ（多数派データ）と同数の 50 %まで 10 %刻みで変化させ、検証用データにおける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 ROC（Receiver Operating Characteristic）曲線から求められる AUC（Area Under the Curve）を用いた。

訓練用データにおいて多数はデータが占める割合を変化させる際はアンダーサンプリング法を適用した。不均衡を調整する方法としては少数派データを繰り返し抽出する方法（オーバーサンプリング法）、多数派データから目的とする数までデータを繰り返し抽出する方法（アンダーサンプリング法）、オーバーサンプリング法とアンダーサンプリング法を組みわせたコンビネーション法の 3 つがある。医療データへの適用においてはアンダーサンプリング法はオーバーサンプリング法及びコンビネーション法より良い結果が得られるとされている[4][5]。本研究ではアンダーサンプリング法の中でも、ランダムに多数はデータからデータを抽出するランダムアンダーサンプリングを用いた。

4. 結果

少数派データの数は訓練用データでは 614 人（総数に対して 1.4 %）、検証用データでは 215 人（総数に対して 1.2 %）となった。

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 から少数派データと 1 対 1 になる 614 まで削減しても書く予測モデル共に AUC に大きな変化は認められなかった（表 1）。

表 1 予測モデルにおける AUC の変化（割合は総データ数に占める少数派データの割合）

割合 (%)	GBDT	LR	RF
サンプリング前	0.76	0.80	0.80
10	0.76	0.80	0.79
20	0.76	0.80	0.80
30	0.75	0.80	0.80
40	0.75	0.79	0.81
50	0.75	0.79	0.80

5. 考察

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 の変化を確認した結果、LR、GBDT、RF 共に多数派データのデータ数を元の 42,910 から少数派データとの比が 1 対 1 になる 614 まで削減しても AUC に大きな変化は認められなかった。多数派データを元のデータの 1.4 %(= 614 / 42,910)まで削減しても予測性能には影響は無かったことになる。予測モデル構築を目的としてデータ収集する際は少数派データを重点的に収集することが望ましい。

なお、LR の基礎手法であるロジスティック回帰が統計モデルとして用いられることの多い症例対照研究では、多数派データと少数派データの比が 1:1 であることが望ましいとされている。今回得られた LR の結果においても多数派データと少数派データの数が1対1となるまで削減しても予測性能には影響は無かったことから、これらの報告と一致する[6][7]。一方で、GBDT 及び RF においても LR と同様に予測性能に大きな変化は認められなかった。これは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る。なお、近年の報告ではアンダーサンプリングの適用は予測モデルの性能を改善するとされているが、本研究においては改善にまでは至らなかった[8][9]。

6. 結論

本研究では、必要なサンプルサイズの削減の観点から現状の性能を保持したデータ面でより効率的な予測モデルの構築を検討した。結果として予測モデル構築におけるデータ収集については少数派データすなわち患者データに着目したデータ収集が重要であることが示唆された。

7. 文献

[1] Ichikawa D, Saito T, Ujita W, Oyama H：How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. Journal of Biomedical Informatics 64: 20-24, 2016.

[2] Ichikawa D, Saito T, Oyama H：Impact of predicting health-guidance candidates using massive health check-up data: A data-driven analysis. International Journal of Medical Informatics 106: 32-36, 2017.

[3] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520–525

[3] Drammond C, Holte R： C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In Workshop on learning from imbalanced datasets II 11, 2003

[4] Blagus R, Lusa L： Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. 2012 11th International Conference on Machine Learning and Applications 2: 89‐94, 2012.

[5] Lewallen S, Courtright P：. Epidemiology in Practice: Case-Control Studies. Community Eye Health, 11: 57‐58.

[6] Hennessy S, Bilker WB, Berlin JA, Strom BL ： Factors influencing the optimal control-to-case ratio in matched case-control studies. American journal of epidemiology, 149: 195-197, 1999.

[7] Lusa L, Blagus R：The class-imbalance problem for high-dimensional class prediction. 2012 11th International Conference on Machine Learning and Applications 2: 123‐126, 2012.

[8] Menardi G, Torelli N: Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery 28: 92-122, 2012