修士学位論文要約（令和４年３月）

不均衡データ学習とシャープレイ値の組み合わせによる２型糖尿病と生活習慣の関係解析

矢島 亮介
指導教員：張山 昌論

Analysis of the relation between type 2 diabetes and lifestyle by combining imbalanced data learning and Shapley value

Ryosuke YAJIMA
Supervisor: Masanori HARIYAMA

Type 2 diabetes is one of most familiar diseases in Japan, and is lifestyle-related. However, it has not been clarified how the elements of lifestyle are intertwined to cause type 2 diabetes. In this study, we propose a method to analyze the relationship between type 2 diabetes and lifestyle-related habits. We demonstrate that the proposed method can find the many factors of type 2 diabetes in a uniform manner.

1. はじめに

２型糖尿病は予備群を含めると日本人の成人６人に１人が罹患すると推計される病気であり，国民にとって非常に馴染みがある．生活習慣病として定義されているが，未だに詳しい原因解明はされていない．解明が進むことで，早期発見や生活習慣指導に応用することができ，病気の進行による合併症を未然に防ぐことが期待される．本研究では，木構造モデルの中でも複雑で高精度な手法として知られるLightGBM[1] を用いて２型糖尿病のデータを学習し，その推論結果に対し，近年研究が盛んな「説明可能なAI」として注目されているSHapley Additive exPlanations(SHAP)[2] を用いることにより，２型糖尿病と生活習慣の関係性について解析する．また，不均衡データ学習を導入し，不均衡の度合いと変数の関係性を調査することにより，網羅的な評価を行い，２型糖尿病に対する各変数の重要度・効果推定の高信頼化を実現する．

2. シャープレイ値に基づく解析手法に関する基礎的考察

既存の機械学習手法では２型糖尿病が各要素とどのように関係し合っているか理解しにくい．そこでLightGBM による推論結果に対して，SHAP を用いることにより，２型糖尿病のデータから重要な変数を解析する．また，その解析結果に対して考察を行う．

今回使用した２型糖尿病のデータは人間ドックの種々の測定結果と生活習慣に関するアンケートから構成され，男性5,971 名から成る．今回目的変数として空腹時血糖値 (FPG) を使用した．FPG は日本人間ドック学会の判定区分に基づき以下の表１のように分類し，「要経過観察」と「要医療」の二クラスを使用した．また，説明変数は215 変数あり，年齢，血圧などの生体情報のほか，食事の内容や食事の摂り方，睡眠，喫煙，ストレス，栄養素，他の病気のバイオマーカーなどを含む．ここで，より重要な説明変数に注目した解析を行うために，多重共線性の除去を行い，61 変数の除去を行った．

表１．人間ドックの判定区分による FPG の人数比

	異常なし・軽度異常	要経過観察	要医療
FPG[mg/mL]	5,293 名	454 名	182 名

LightGBM を使用した FPG の二値分類の推論結果を表２に示す．一般化のため，初期乱数を 10 回変化させて平均値を算出した．表２の結果から F 値とRecall が低いことが確認できる．理由として，負と正のクラス間で約 7:3 の偏りがあるためと考えられる．全員を負と診断することで 0.7 に近い Accuracy を達成することができるが，この場合，Recall は 0 に近づくことになる．医療分野においてRecall が重要であるため，クラス間の偏りを小さくし，Recall と F 値を増加させる必要がある．

また，LightGBM の結果に対し，SHAP を使用することで FPG に対する重要変数を算出した．これを，図１に示す．上から順に変数重要度が高い順に並べられている．赤色や青色は各変数の値の大きさ

表 2．Accuracy と F 値，Recall の 10 回の平均値

	10 回の平均値
Accuracy	0.669
F 値	0.331
Recall	0.301

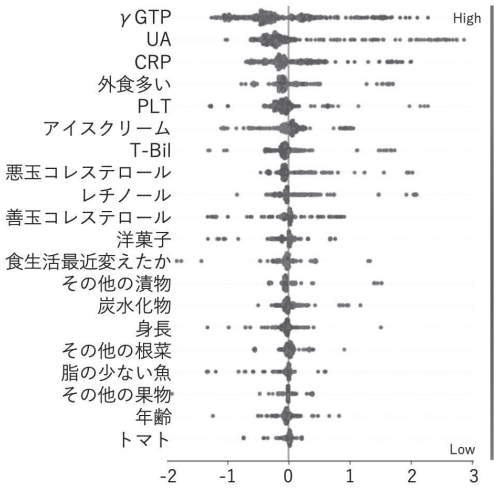

図 1．重要変数と SHAP 値の関係

を表し，x 軸は SHAP 値の大きさを表している．しかし，この重要変数と SHAP 値の関係図はある初期乱数における一時的な図であり，初期乱数を変化させると重要変数が変化してしまう問題がある．そのため，10 回初期乱数を変化させたときの平均を取るなど，一般的に評価する方法を考慮する必要がある．

3.　不均衡度の網羅的評価に基づく変数重要度・効果推定の高信頼化

本章ではクラス間の不均衡さに関する問題や乱数変化における問題の解決のために不均衡データ学習の導入や，変数重要度・効果推定のアプローチ法を考案し，結果の考察を行った．

LightGBM による評価結果から，不均衡データ学習の手法として Under Sampling を使用した．また，事前に以降の実験で使用しない他の初期乱数を用いた学習を行い，50 変数まで選択した．

多人数クラスに対する少人数クラスの比を「不均衡度」と定義し，SHAP を用いて不均衡度に対する変数重要度を算出した．ランキングの変化を図 2 に示す．図から不均衡度に対して変数重要度ランキングの変化が激しいことが確認できる．そのため，FPG に一定の効果を持つ変数を調べる必要がある． 次に，FPG に対して正と負の一定の効果を持つ変数を推定する方法を考案した．各変数と各変数の SHAP 値のグラフから最小二乗法の傾きを求め，その傾きが正か負かで効果を推定する．この値を不均衡度，初期乱数毎に推計することによってその重要変数が

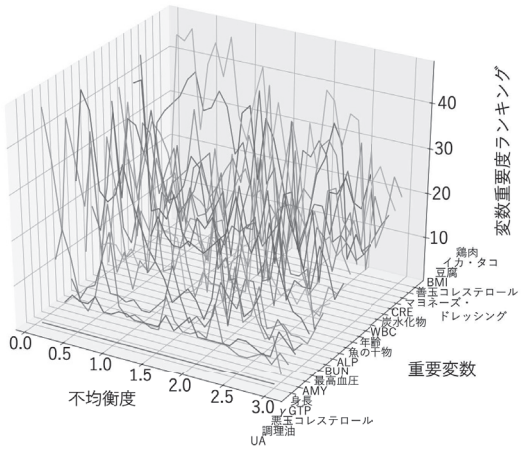

図 2．不均衡度に対する変数重要度ランキングの変化

表 3．FPG に対して正の効果を持つ変数

正の効果を持つ変数	負の効果を持つ変数
体脂肪率, 心拍数,ALP, γ GTP,CRP, 緑茶, マヨネーズ・ドレッシング, 味噌汁, 最低血圧, 卵など	UA, 善玉コレステロール, AMY, 炭水化物, 脂の少ない魚, パン, 豚肉・牛肉, 豆腐,PLT,LDH など

FPG に対して正の効果を持つか負の効果を持つかを推定することができる．推定の結果を表に示す．

この結果について専門家の方にフィードバックをいただいた．その結果医師の考えと一致する変数が多いことがわかった．また，医師の考えと異なる変数として「味噌汁」,「緑黄色野菜」,「炭水化物」などが挙げられた．

4.　まとめ

LightGBM で学習した推論結果に対し，SHAP を用いることによって，不均衡度毎の重要変数の調査を行った．また，FPG に対する重要変数の一定の効果を網羅的に推定する手法を考案し，推定を行った．専門家にフィードバックをいただき，専門家の考えと一致することを確認した．今後因果関係解析の手法などと組み合わせることで，2 型糖尿病のメカニズムをさらに解明し，医療現場で活用されることが期待される．

文献

1) Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang,Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu,“LightGBM: A Highly Efficient Gradient Boosting Decision Tree“.Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.
2) Lundberg, Scott M., and Su-In Lee. ”A unified approach to interpreting model predictions.” 31st conference on neural information processing systems (NIPS 2017), Long Beach, CA; 2017.