計測自動制御学会東北支部 第 346 回研究集会 (2023.12.20)
資料番号 346-14

説明可能 AI を用いた２型糖尿病と生活習慣の関係性解析

Analysis of the Relationship between Type 2 Diabetes and Lifestyle using Explainable AI

○奥村満 *, 張山昌論 *, 望月和樹 **

○ Mitsuru Okumura*, Masanori Hariyama*, Kazuki Mochizuki**

*東北大学, **山梨大学

*Tohoku University, **Yamanashi University

キーワード： 機械学習 (machine learning), ビッグデータ解析 (big data analysis), XAI

連絡先： 〒 980-8579 宮城県仙台市青葉区荒巻字青葉 6-3-09
東北大学 大学院情報科学研究科 張山研究室
奥村満, Tel.: 022-795-7155 E-mail: mitsuru.okumura.s4@dc.tohoku.ac.jp

1. はじめに

2 型糖尿病は遺伝的要因に加え, 過食や運動不足などの環境的要因が加わることで高血糖が慢性的に続く病気である. 特徴として初期症状が乏しいこと, 病気が進行することで網膜症, 腎症, 神経障害といった深刻な合併症を伴うといったことが挙げられる. これら合併症は患者自身の生活に大きな影響を及ぼすのみならず, 医療費の観点から社会にとっても大きな負担を強いることとなる. 本研究の目的は, 機械学習を用いて糖尿病と生活習慣の関係性を探索することで, 生活習慣の改善による発症予防, 早期発見, 血糖コントロールを基本とした合併症予防に寄与することである.

近年, 機械学習アルゴリズムの複雑化に伴い, モデルの中で説明変数がどのように影響しているか分からないというモデルのブラックボックス化が発生している. また, 従来の機械学習の解析構造では, 単一の目的変数と説明変数による一階層の解析が行われることが多い. そのため, 目的変数と説明変数間の関係性を理解することはできても, 説明変数間の関係性を把握することはできない. そこで, 本研究ではブラックボックスの解消のために「説明可能 AI」を利用することで, 説明変数がモデルや目的変数に及ぼす影響を定量的に評価する. また, 解析データに含まれるすべての変数を目的変数の対象とする多階層の解析を行い, 関係性のネットワークを構築することで変数間の関係性を網羅的に表現することを目指す.

2. LightGBM と SHAP に基づくネットワーク構築

機械学習モデルの一つである LightGBM[1)] と説明可能 AI の代表的な手法である SHAP[2)] を用いて, 変数間の関係性ネットワークを構築する. 関係性ネットワークの構築は一階層ネットワークの構築とその統合からなる.

2.1　SHAP

SHAP は説明可能 AI の代表的な手法である. 説明可能 AI(XAI) とは, モデルの推論プロセスを人間が理解し, 信頼できるようにするための一連の手法を指す. モデルの予測精度を損なうことなく, 予測に一定の解釈を与えることができる. XAI はモデルの出力結果に入力された変数がどのように影響しているかが評価でき, 推論過程の正当性・公平性を特徴づけることが可能である.

SHAP はゲーム理論の「シャープレイ値」を機械学習に応用した手法である.SHAP 値は目的変数の予測結果への各特徴量の寄与度を定量的に算出したものであり, これにより, 目的変数に対する重要な変数を探索することが可能となる. SHAP 値は式 (1) により計算する.

$$\text{貢献度}\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!\,(M - |S| - 1)!}{M!} \left(v\left(S \cup \{i\}\right) - v\left(S\right)\right) \tag{1}$$

但し, ここで M は変数の数, S は変数 i を含まない変数の全集合の一つの集合, v はモデルの出力結果を示す関数である. 式 (1) では, 変数 i を用いた場合と用いなかった場合に生じる推論結果の差に, 使用した集合の組み合わせの出現確率を掛け合わせたものが変数 i の SHAP 値 (貢献度) であることを示している.

2.2　一階層ネットワークの構築

一階層ネットワークは LightGBM で作成した機械学習モデルに対して SHAP を適用することで構築する. 構築の流れを図 1 に示す.

1) LightGBM でモデルを作成し, 各説明変数に対する SHAP 値を計算する. ネットワークのエッジの重みに SHAP 値を割り当てる.

2) エッジの重みに割り当てた SHAP 値の絶対値に対して正規化を施し, エッジの重みの最大値が 1 となるよう変換する.

3) 閾値 (0.5) 以下のエッジを削除する.

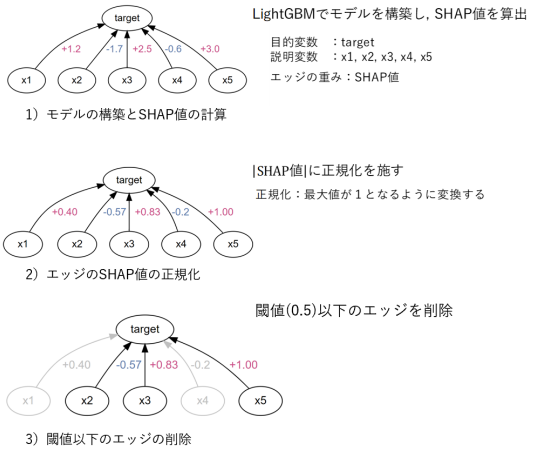

Fig. 1　一階層ネットワークの構築の流れ

2.3　多階層ネットワークの構築

前節で構築した一階層ネットワークを統合することで, 全変数の関係性を含んだ多階層ネットワークを構築する. 実際の構築の流れを図 2 に示す. 一つの変数に着目して得られた左側の一階層ネットワークの情報を, 右側の全体の関係性ネットワークに階層的に組み込んでいくことで多階層のネットワークを構築する.

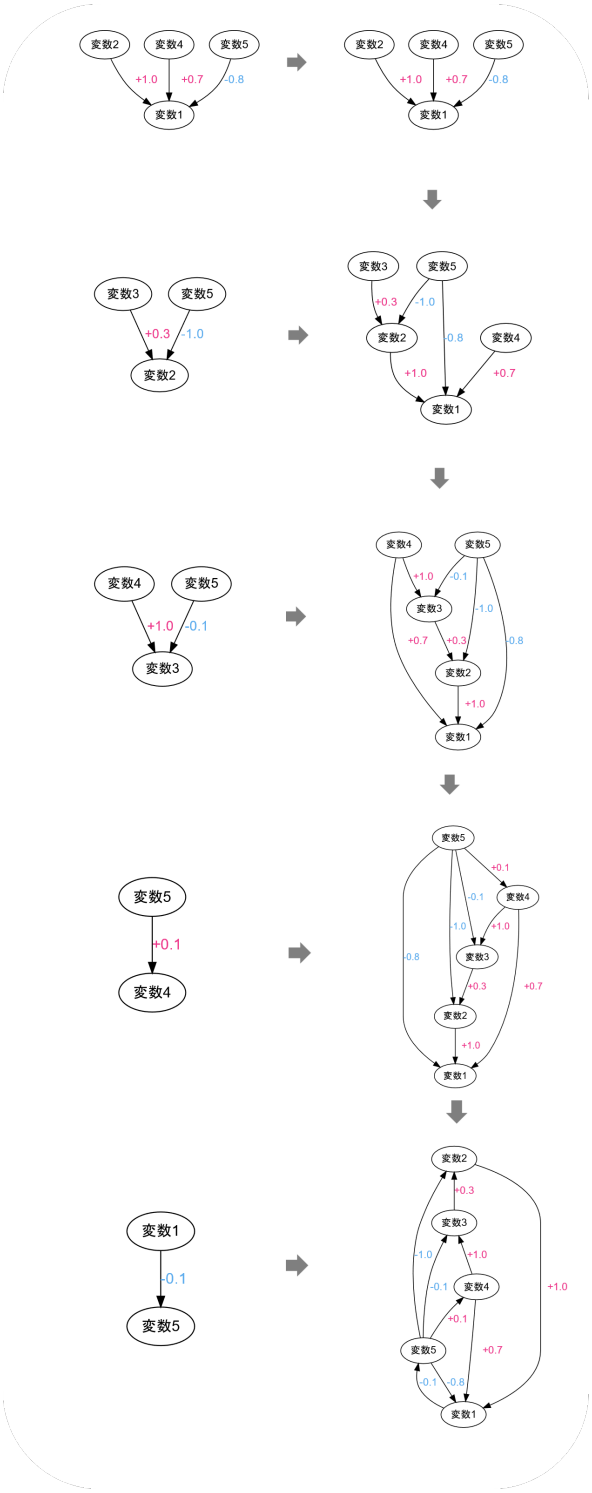

Fig. 2　多階層ネットワークの構築の流れ

3.　糖尿病データへの適用

3.1　糖尿病データ

前章で述べた多階層ネットワークの構築手法を糖尿病データに対して適用する．解析に使用した変数を表 1 に示す．

Table 1　本解析で使用する変数

生体情報	BMI
生活習慣	喫煙本数 (日)
	運動回数 (週)
	食事時間不規則か
	睡眠時間
	夕食 9 時以降か
	外食が多いか
	腹いっぱい食べるか
	脂肪分の多い食事か
	朝食抜くか
	間食をよくするか
	野菜果物たべないか
	ひとりで食事をするか
	糖分飲料よく飲むか
	ストレスを感じているか
	飲酒の量
	食事速度
血液成分	TG
	HDL
	LDL
	HbA1c
	UA
	CRE

表 1 に示された変数の中で, 特に重要な変数が HbA1c である．HbA1c は糖尿病診断に用いられる血液成分である．従って, HbA1c と生活習慣変数の関係性を探ることが本研究の目的となる．また, データの前処理として, 生活習慣に分類される変数である「食事時間不規則か」は不規則でないと答えた人を 0, どちらでもないと答えた人を 1, 不規則だと答えた人を 2, となるよう変換処理をした．他の疑問形で記述されている変数も同様な前処理を施している．

3.2　ドメイン知識を用いた関係性の方向指定

多階層ネットワークの構築手法を, 現実のデータに適用させた時に関係性の方向として成り立

たないものが結びつく可能性がある．例えば, 血液成分に分類される変数から生活習慣に分類される変数にエッジが伸びる場合である．この関係性の方向は通常逆であるべきである．つまり, 生活習慣に分類される変数により血液成分に影響が出ると考える方が自然である．本研究ではこの現象への対応として, ドメイン知識を用いた関係性の方向指定を行う．ここでのドメイン知識とは, 解析対象となる専門分野における学術的知見をはじめとした, 解析の前に既に明らかである知識を指す．本研究で指定した関係性の方向を図 3 に示す．図 3 で指定された関係性の方向性を保つよう解析に使用する変数を選択する．例えば, BMI の解析に使用する説明変数は喫煙, 睡眠時間, 運動回数, 飲酒の量といった生活習慣に分類される変数のみである．

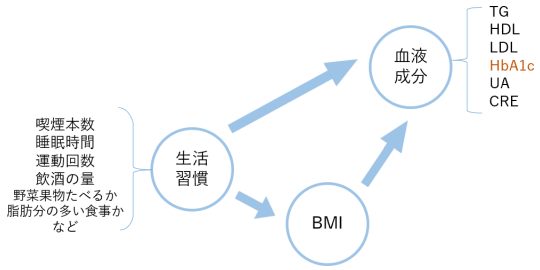

Fig. 3　ドメイン知識を用いた関係性の方向指定

3.3　解析結果

前章で述べた多階層ネットワークの構築手法と前節で述べたドメイン知識を用いた関係性の方向指定を組み合わせた解析結果を図 4 に示す．

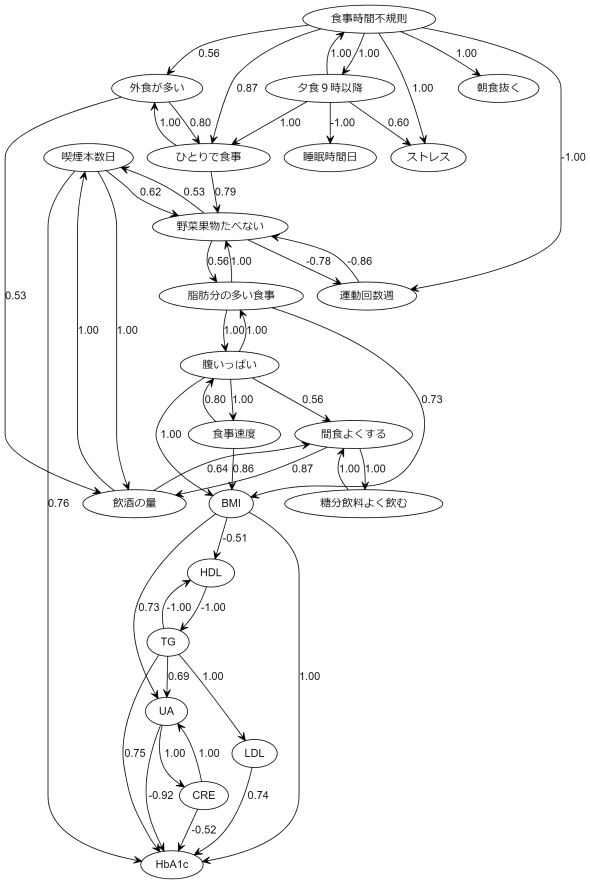

Fig. 4　糖尿病データにおける多階層関係性ネットワーク

エッジの重みは変数がもつ寄与度を数値化したものである．エッジの重みが 1 に近いほどエッジ先の変数にとって重要な変数であることを意味している．また, エッジの符号が正の場合は, エッジ元の変数の数値が増加するとエッジ先の変数の値も増加することを意味している．本図の BMI と HbA1c の関係性がこれに該当する．また, エッジの符号が負の場合は, エッジ元の変数の数値が増加するとエッジ先の変数の値が減少することを意味している．本図の BMI と HDL(善玉コレステロール) の関係性がこれに該当する．

3.4　評価

図 4 より与えられる関係性と従来の研究から得られている関係性との関連について述べる．BMI と HbA1c を含む他の血液成分との関係性

は, 低 HDL コレステロール血症, 高血糖の出現率と正の関連が認められるとした, 従来の研究[3] と合致する結果を示した. BMI と関連の深い変数として, 食事速度, 腹いっぱい食べるか, 脂肪分の多い食事をするか, が本解析によって抽出されている. 食事速度, お腹いっぱいまで食べるか, と BMI の関係性はこれら変数と肥満の関係性を調査した研究[4] と一致する正の関係性を示している. また, 喫煙と HbA1c の関係性に関しても, 能動喫煙が糖尿病リスクを高めるとする研究[5] と合致する結果を示した.

3.5 考察

解析結果から, 糖尿病予防において BMI の改善が重要であること, BMI 改善にはゆっくりとした食事や腹八分目の意識が重要であることが示唆されている. 食事速度や意識改善のアプローチは, 食事内容や運動パターンを急激に変えることに比べ, 個人の生活スタイルに合わせて段階的に実践できる利点がある。

また, 生活習慣同士の変数の関係性として興味深いものが存在する. ひとりで食事する人は野菜果物をたべない傾向にあり栄養が偏ること, 食事時間が不規則な人ほどストレスを感じていること, 飲酒と喫煙の相互関係などである. これらの関係性は, 個々の生活習慣や行動がお互いに影響を与え合う可能性を示している. 従って, これらの生活習慣の相互関係を考慮することで, より効果的かつ包括的な予防・改善策の提言につながると考えられる.

4. おわりに

本研究では機械学習を活用した全変数間の関係性ネットワークを構築する手法を用いて, 糖尿病と生活習慣の多階層関係性ネットワークを構築した. 生活習慣と糖尿病の関係性だけでなく, 生活習慣の変数同士の関係性についても解析を行った. 今後の方針として, 専門家のドメイン知識を加えた定性的なネットワークの評価に加え, 定量的な評価方法の検討を行っていく.

参考文献

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.

3) 吉池信男, 西信雄, 松島松翠, 伊藤千賀子, 池田義雄, 樫原英俊, 吉永英世, 小倉浩, 小峰慎吾, 佐藤祐造, 佐藤則之, 佐々木陽, 藤岡滋典, 奥淳治, 雨宮禎子, 坂田利家, 井上修二: Body Mass Index に基づく肥満の程度と糖尿病, 高血圧, 高脂血症の危険因子との関連－多施設共同研究による疫学的検討－, 肥満研究 6 (1) 4-17 (2000)

4) Maruyama K, Sato S, Ohira T, Maeda K, Noda H, Kubota Y et al. The joint impact on being overweight of self reported behaviours of eating quickly and eating until full : cross sectional survey BMJ 2008;

5) Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J: Active smoking and the risk of type 2 diabetes a systematic review and meta-analysis, JAMA, 2007 Dec 12;298(22):2654-64.