5U-08

健康診断データによる動脈硬化指数の予測

松原 享佑[†1]　矢野 裕一朗[†2]　長尾 智晴[†3]
横浜国立大学 大学院環境情報学府[†1]　横浜国立大学 大学院環境情報研究院[†3]
滋賀医科大学 NCD 疫学研究センター 最先端疫学部門[†2]

1 はじめに

心筋梗塞や脳卒中などの疾患を未然に防ぐため，それらの疾患に繋がる動脈硬化の進行度を測定することに需要がある．動脈硬化の指数として用いられるのは心臓足首血管指数(Cardio Ankle Vascular Index; CAVI)である．この値は検査機器を用い，仰臥位で両腕・両足首の血圧と脈波を 5 分程度測定することで導出できる．しかし CAVI には検査義務がないため，個人で検査を受けない限り動脈硬化の進行度を利用できない現状がある．また検査機器が特殊機器のため一部の医院などでのみ取り扱いがあり，地方によっては検査を受けにくいなどの欠点がある．

本研究では，多くの人が日常的に取得可能な生体データである健康診断データから，機械学習の手法を用いて CAVI の予測を行う．これにより健康診断を受けた人であれば CAVI 検査を受けずに動脈硬化の進行度を利用できるようになり，健康診断の結果に伴った動脈硬化のサービスや社会実装が可能になる．予測結果は追加検査や治療などの意思決定に用いられるため，CAVI の実測値を予測するだけでなく，範囲と確率を多段的に示すことで利用者の次の行動に繋げられる手法の構築を目指す．

2 提案手法

ニューラルネットワーク(Neural Network; NN)による予測の最終的な目標は出力をもとに意思決定を行うことであり，モデル予測の確信度を知ることは重要であるため，多くの問題において不確実性の推論が効果的である[1]．そのような推論手法のうち，分位点回帰はピンボール損失を用いることにより確率的な推論を複数の分位点の予測に帰着させることができる[2]．ピンボール損失とは，図 1 に示すような損失関数で，範囲内の任意の分位τを指定することができる．

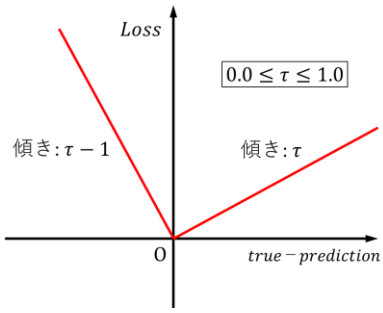

図 1: ピンボール損失

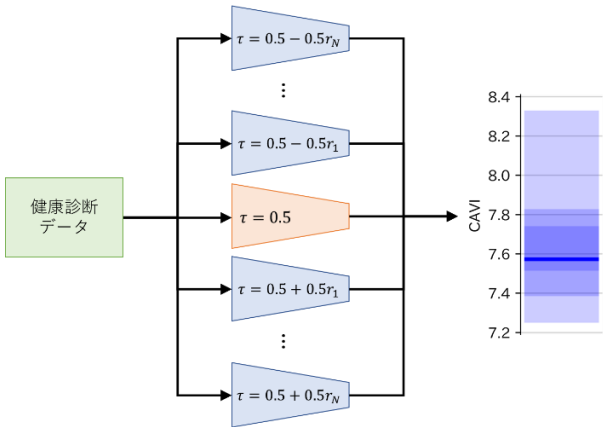

図 2: 手法概要（τ:分位数，$r_1, \ldots, r_N$:範囲）

提案手法を図 2 に示す．予め分位点の範囲を複数決めておき，そこから導かれる分位数を用いて互いに分位数の異なる NN を学習させ，CAVI の実測値と分位点の値を予測させる．なお実測値の予測は$\tau = 0.5$とすることで得られる．

予測結果は浮動棒グラフで可視化する．確信度は色の濃淡で示し，色が濃いほど確信度が高く，薄いほど網羅率が高い．このように示すことで，例えば「CAVI が 7.4～7.8 の間である確率が 80%である」のような範囲と確率による説明が可能になる．

3 実験

3.1 使用するデータセット

Prediction of Arteriosclerosis Index from Medical Examination Data
†1 Kyosuke Matsubara　†2 Yuichiro Yano
†3 Tomoharu Nagao
†1 Graduate School of Environment and Information Sciences, Yokohama National University
†2 NCD Epidemiology Research Center, Shiga University of Medical Science
†3 Faculty of Environment and Information Sciences, Yokohama National University

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

117709 人による 1〜16 件（計 305891 件）の，健康診断により得られる項目と CAVI からなる 134 項目のデータを用いた．具体的には，ID，検査回数と経過年数，年齢，性別，居住地域，体重や血圧などの基本計測値，心電図，血液検査，尿，便潜血，眼底，問診の各質問に対する回答，CAVI 実測値からなる．データは欠損を含んでおり，全体の欠損率は 23.4%である．また 70%以上が欠損している項目が 20 項目存在する．これらのデータから，ID など変数に使用できない値を除外し，項目ごとに適切な前処理を施したのち，CAVI 実測値を目的変数，それ以外を説明変数としてデータセットを作成した．なお，本研究では個人の経時的変化は考慮せず，305891 件を個別のデータとして扱った．

3.2 項目の選別と欠損処理

医師の方から，居住地域，心電図，便潜血，問診の回答は CAVI と直接関係はないという知見をいただいたため，全ての項目を用いた場合とこれらの項目を除いた場合で予測誤差を比較した．また，NN は欠損を含むデータを学習することができないため欠損処理が必要となる．今回は欠損を一つでも含むケースを除外するリストワイズと，値を 0.1〜0.9 に正規化したのち欠損を 0.0 で埋める単一代入法を行い，予測誤差を比較した．得られたデータセットは訓練データとテストデータに 8:2 で分割し，NN に訓練データを学習させた後，テストデータで誤差の評価を行った．

結果を表 1 に示す．特定項目を除外し，かつリストワイズを行った場合で最も誤差が小さくなった．医師の方から CAVI の予測誤差は 0.3 程度が望ましいという意見をいただいていたため，目標値を概ね達成していると言える．

全ての項目を使用した場合ではリストワイズが効果的ではなかったが，特定項目を除外した場合は効果があり，用いた項目に良質なデータが集まっていると考えられる．

3.3 提案手法を用いた説明の検証

分位点の範囲を，正規分布の標準偏差の 2 倍，4 倍，6 倍に入るデータの割合から，68%，95%，99.7%と 3 段階に定めた．データセットは，3.2 で得られた特定項目を除外しリストワイズを行ったデータを，同じく訓練データとテストデータに 8:2 で分割し，NN に訓練データを学習させた後，テストデータで誤差の評価を行った．

可視化結果を図 3 に示す．医師の方に紹介したところ好意的な評価をいただいた．ただし分位点の範囲について新たに知見をいただいたため，

表 1: 各処理法ごとのテスト誤差(MAE)

	全ての項目を使用	特定項目を除外
リストワイズ	0.6031	**0.2966**
0 代入	0.4119	0.4098

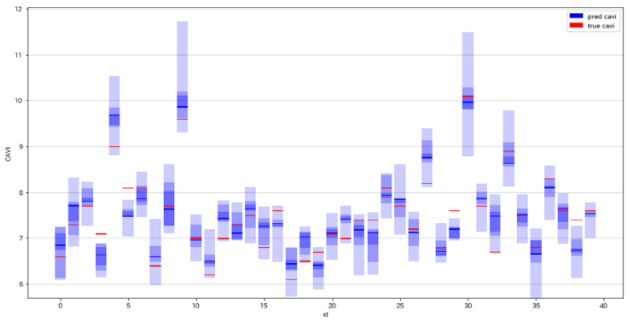

図 3: 分位点回帰の結果（40 人抜粋）
縦軸: CAVI（青=予測　赤=正解）

引き続き検証を行うことを考える．

4 まとめ

本稿では，範囲と確率を多段的に示すことで患者の意思決定に貢献する CAVI 予測モデルを提案し，実際の健康診断データから効果的な項目と前処理を選択したのち，提案手法に適用し医師の知見を基に有効性を検証した．問題点として，現在は各 NN を個別に学習しているため，分位点の上下が逆転してしまうサンプルが存在する．今後はこの問題への対策を検討する．

謝辞

本研究は滋賀医科大学様との共同研究です．大規模な健康診断データをお譲り下さいました矢野裕一朗先生に深く感謝いたします．

参考文献

[1] F. Rodrigues and F. C. Pereira, "Beyond Expectation: Deep Joint Mean and Quantile Regression for Spatiotemporal Problems," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, pp. 5377-5389, Dec. 2020.

[2] Koenker, Roger, and Kevin F. Hallock. 2001. "Quantile Regression." Journal of Economic Perspectives, 15 (4): 143-156.

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.