F-003

精度と透明性を両立する AI を生成する技術
AI Simplification Technology for Accurate and Transparent AI

難波 博之†
Hiroyuki Namba

濱本 真生†
Masaki Hamamoto

1. はじめに

近年、インフラ制御や製品品質検査など失敗が許されない業務への AI 適用に向け、判断ロジックが単純で信頼できる AI の生成技術が研究されている。しかし、既存技術で得られる単純な AI は、I 精度が不十分である、あるいは II 判断根拠が単純なルールであっても現場の知識と整合しておらず納得されない場合がある、という課題があった。そこで、我々は高精度かつ複雑な従来 AI を最も良く近似する単純な関数を探索することで、I 高精度かつ単純な AI を生成する AI 単純化技術を開発した。さらに、探索する関数形を限定することで、II 現場の知識と整合させることもできる。また、オープンデータを用いた評価により、従来 AI を精度劣化をおさえながら大幅に単純化できることを確認し、提案技術の有効性を示す。

2. 前提と目的

2.1 問題設定

機械学習における分類問題及び回帰問題を考える。すなわち、計測できるいくつかの数値（特徴量）をもとに、ある指標（目的変数）を予測する問題を考える。例えば、製品の性質を表す複数の計測値から、その製品が良品か不良品かを予測する問題を考える。計測値から良品と不良品を正確に予測できると、従来人が手作業で実施していた不良品の判別作業を自動化できるため、生産現場における検査速度を向上できる。このように、予測を高い精度で実現することで、様々な業務を改善し、大きな価値を提供できる。

2.2 ブラックボックス AI：高精度だが透明性は低い

2.1 で述べた予測問題に対し、過去の実績データから規則性を見出し、予測規則を生成する既存技術がある[1]。深層学習などの技術の進展により、多くの事例で予測精度が高い AI が生成できる。一方で、このような AI は、数千から数十億ものパラメタをもとに予測を行うなどとても複雑であり、まさにブラックボックスと化している。ブラックボックス AI は、中身を人が理解・検証できないため、未知のデータに対して想定外の出力をする可能性がある。すなわち、ブラックボックス AI は、未知のデータに対して判定間違いをしてしまう可能性があるため、失敗が許されない業務に安心して導入することができない。

2.3 人が理解できる AI：透明性は高いが低精度

一方、中身を人が理解・検証できる AI なら、どのような入力データに対しても想定外の出力をする可能性を排除できる。これは、パラメタ数が数十個程度の単純な AI なら、あらゆる入力に対する出力の妥当性を事前に検証できるためである。このように、透明性の高い AI であれば、想定外の誤判別の可能性を無くすことができ、失敗が許されない業務にも安心して導入できる。透明性の高い AI を生成する技術としては、線形回帰・決定木をはじめとして様々な手法がある。特に、高精度なブラックボックス AI を近似する defragTrees[2]や、区分線形モデルを生成する手法[3,4]が挙げられる。しかし、これらの手法により得られる AI は、基本的にいくつかの定数関数および線形関数からなるため、以下の 2 つの課題がある。

- 課題 1：精度と透明性の両立が不十分である。すなわち、非線形の現象をうまく捉えることができず、精度が足りない場合がある。
- 課題 2：仮に精度と透明性を両立できていても、専門家から見ると予測式が単純すぎて納得されない。

本研究の目標は、この 2 つの課題を解決する AI の生成である。すなわち、精度と透明性を両立し、さらに専門家の知識と整合した AI の生成である。これが実現できれば、未知のデータに対しても原因不明の判定間違いの可能性がなくなるため、安心して AI を導入することができる。

3. 提案手法

2.3 節で述べた 2 つの課題を解決するために、提案技術は 3.1、3.2 節で述べる以下の 2 つの方針を採った：

3.1 方針 1：精度と透明性の両立に向けて

[3]では一定の精度と透明性を実現しているため、この手法をもとにした。すなわち、高精度なブラックボックス AI を単純化する方針や、ブラックボックス AI がとらえた重要特徴量の情報[5]をもとに領域を分割する方法は[3]を踏襲した。 [3]では予測式として線形関数のみを用いているが、精度のさらなる改善に向けて予測式の形式を拡張した。具体的には、少ないパラメタ数で高精度な予測式を生成できる Symbolic Pursuit[6]を参考に、各領域の予測式を $\sum\{af(L(x))+b\}$という形に拡張した。ただしa, b は定数で L は線形関数である。f としては 3.2 で述べる単純な非線形関数を用いた。最適な予測式を探索する方法は、Symbolic Pursuit と同様に勾配法と射影追跡回帰を用いた。

図 1 提案手法の方針 1

† (株) 日立製作所　Hitachi, Ltd.

3.2 方針 2（専門家の知識整合性の改善に向けて）

AI と整合させたい専門家の知識には多様な形式が考えられるが、本稿では、2 章で述べたような予測式の形式が原因で納得してもらえないという課題の解決が目標であるため、専門家の知識として予測式に用いる関数形のみを扱う。まず、方針 1 で述べたように各領域の予測式を非線形に拡張することで、単純すぎて納得してもらえないという課題は解決できる。しかし、Symbolic Pursuit で採用している G 関数をそのままfとして用いるとベッセル関数などの複雑な予測式も出てきてしまうため、G 関数の中でも産業界でも良く現れる物理法則の記述に登場する初等関数のみを用いた。具体的には 3 次以下の多項式・三角関数・指数関数・対数関数・双曲線関数・逆三角関数を用いた。さらに、この初等関数リストの中から、事前にドメイン知識と整合する関数形リストを指定して探索に制限をかけられるようにした。例えば、線形関数では単純すぎて現場に納得されないことが分かっている場合は f の候補から恒等関数を除いて探索する。また、対象が振動に関する現象で三角関数により予測式を記述したい場合、 f の候補として三角関数のみを用いる。

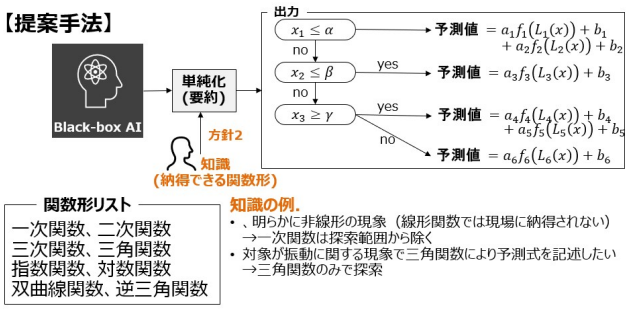

図 2 提案手法の方針 2

4. 数値実験

本章では、提案手法の評価方法および結果を述べる。UCI Machine Learning Repository[7]から Boston を用いて評価した。住宅価格を予測する回帰問題である。データの行数は 506、特徴量数は 13 である。複数の既存手法および提案手法を適用し、結果を比較した。既存手法としては、人が理解できる AI の代表である線形回帰および決定木、ブラックボックス AI の代表例である LightGBM、および 3.1 節で述べた手法[3]と Symbolic Pursuit[6]を用いた。また、単純化対象のブラックボックス AI としては、LightGBM を用いた。評価の観点は、精度および透明性とした。精度は、回帰問題の誤差関数の代表例である二乗誤差を用いて評価した。また、透明性については、モデルのパラメタ数で評価した。パラメタ数が数十個程度の単純な AI なら、あらゆる入力に対する出力の妥当性を事前に検証できるためである。

データを 5 分割し、5-fold 交差検証を実施した。各手法における二乗誤差 RMSE(Root Mean Squared Error）とモデルのパラメタ数を表 1 に示す。

表 1 各手法の精度と透明性

	RMSE	パラメタ数
LightGBM	3.56	9000
線形回帰	4.83	14
決定木	4.98	90
Symbolic Pursuit[6]	4.10	30
既存手法[3]	3.63	56
提案手法①	3.69	26
提案手法②	3.90	16

ただし、提案手法については、パラメタ数を変えて 2 通り実験した。①において、領域数は 1、和を取る関数の数 2、線形関数の入力として用いる特徴量の数は 9 とした。②では、領域数は 2、和を取る関数の数 2、線形関数の入力として用いる特徴量の数は 3 とした。

精度についてはブラックボックス AI と既存手法[3]、提案手法が優れている。また、パラメタ数については線形回帰、Symbolic Pursuit[6]、および提案手法が優れている。すなわち、提案手法は精度とパラメタ数の少なさを両立できており、本データにおける提案技術の有効性が確認された。具体的には、複雑な従来 AI である LightGBM を精度劣化 4%未満でパラメタ数 1/300 以下にまで単純化できたといえる。

5. おわりに

精度と透明性を両立し、さらに専門家の知識と整合した AI を生成する手法を提案した。また、本技術の有効性をオープンデータに適用し、精度と透明性を両立できることを検証した。本技術により生成した AI では、あらゆる入力に対する出力の妥当性を事前に検証できるため、想定外の誤判別のリスクが無く、失敗が許されない業務にも安心して導入できる。すなわち、本技術を用いることで、人が検証できる AI が必要なさまざまな業種における業務改善に貢献できるものと考える。ただし、専門家の知識との整合性に関する有効性評価は今後の課題である。

参考文献

[1] G. Ke, et al., Lightgbm: A highly efficient gradient boosting decision tree, In *Advances in Neural Information Processing Systems* 30, pp.3146--3154, 2017.
[2] S. Hara and K. Hayashi, Making tree ensembles interpretable: A bayesian model selection approach, In *International Conference on Artificial Intelligence and Statistics*. pp.77--85, 2018.
[3] H. Namba and M. Egi, Piecewise Simplification Approach for Accurate and Understandable Model, In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1--7, 2021.
[4] L. Yang, et al., Mathematical programming for piecewise linear regression analysis. *Expert Systems with Applications,* pp.156--167, 2016.
[5] S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, In *Advances in Neural Information Processing Systems,* pp. 4765—4774, 2017
[6] J. Crabbe, et al., Learning outside the black-box: The pursuit of interpretable models, In A*dvances in Neural Information Processing Systems,* 33, pp.17838--17849, 2020.
[7] A. Asuncion and D. Newman, UCI Machine Learning Repository, 2007.