様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ４ 月 １６ 日現在

機関番号：１４４０１
研究種目：若手研究
研究期間：2020～2023
課題番号：２０Ｋ１９８６０

研究課題名（和文）機械学習モデルの説明駆動開発のための基盤技術

研究課題名（英文）Explanation-guided Machine Learning Model Development

研究代表者
原　聡（Hara, Satoshi）

大阪大学・産業科学研究所・准教授

研究者番号：４０７８０７２１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：実用に供するレベルの精度の高い機械学習モデルを作ることは必ずしも容易ではなく、開発者の熟練度によって出来上がるモデルの精度には大きな開きがある。本研究では説明可能AI（Explainable AI; XAI）の技術を基盤に、開発者により良いモデルの構築方法について適切にアドバイスする仕組み「説明駆動モデル開発」の研究に取り組んだ。
研究の成果として、モデルの性能を改善するためのデータクレンジング法やそれを拡張した類似データ説明法、そしてモデルの部分的な可読化に基づくモデルの修正技術などが得られた。

研究成果の学術的意義や社会的意義
研究成果の学術的意義としては「説明駆動モデル開発という新たな機械学習モデルの開発の仕組みの提案」、そして「XAI技術のさらなる発展による機械学習モデルの解釈性の向上」があげられる。これらにより、モデルを効率的に改善する方法や、モデルがなぜ特定の予測や判断を下したのかを理解することが容易になり、モデルへの信頼性向上が期待される。
研究成果の社会的意義としては「高性能なモデルが効率的に開発可能になることで、機械学習モデルの社会的な活用がより一層進む」ことがあげられる。

研究成果の概要（英文）：In this reserach project, we focused on the development of "Explanation-guided model development," which provides appropriate advice to developers on how to construct better machine learning models, leveraging Explainable AI technology as its foundation.
Creating highly accurate machine learning models at a level suitable for practical use is not always straightforward, and there can be significant variations in the accuracy of models produced depending on the skill level of the developers.
In this project, we developed methods for data cleansing to improve model performance, its extesion for similarity-based explanation, as well as model correction techniques based on partial model explanation.

研究分野：機械学習

キーワード：機械学習　深層学習　説明可能AI

様　式　C－１９、F－１９－１、Z－１９（共通）

1. 研究開始当初の背景

械学習技術は今や情報分野を超えて様々な分野 (e.g. 材料、天文、広告) での利用が広がっている。特に、便利な開発ツールのおかげで、機械学習の非専門家であっても機械学習モデルを作れるようになってきている。機械学習の大目標は、データから予測・認識精度の高いモデルを学習により作ることである。これらモデルにより、所望の特性を持つ材料の推定、天文画像からの流星の検出、顧客の購買行動の予測、など様々な問題を高い精度で解くことができる。

しかし、現行の開発ツールをもってしても精度の高いモデルを作ることは必ずしも容易ではない。例えば学習データの質が悪い (e.g. ノイズが多い) 場合には高精度な予測は難しくなる。また、モデルがデータや解きたい問題に適していない場合にも精度が低下してしまうことがままある。機械学習の実用現場では、このようなデータの質やモデルの改善などについて開発者が様々な対処法を検討し、試行錯誤をしながら徐々にモデルの精度を高めているのが実情である。そのため、機械学習モデルの構築は高度な職人技と化しており、同じデータ・同じ計算資源を用いても個人の熟練度合いによって出来上がるモデルの精度には大きな開きがある。機械学習の適用範囲を広げるためにはこのような属人性を解消して、誰でも精度の高いモデルを作れるようにする必要がある。

このような開発者間の格差は各個人の積み重ねた経験と勘によるものであるため、経験の浅い開発者であっても熟練開発者のアドバイスがあればより精度の高いモデルが構築できるようになる。つまり、開発者に必要なのはより良いモデルの構築方法について適切にアドバイスする仕組みである。このように機械学習モデルから人間へと情報をフィードバックする方法の枠組みとして説明可能 AI (Explainable AI; XAI) がある。本研究の狙いは機械学習モデルの開発へと XAI の枠組みを拡張し「説明駆動モデル開発」の基盤技術を確立することである。

2. 研究の目的

本研究では、具体的に以下の 2 つの課題に取り組む。

研究課題1．良い学習方法の推定

開発者に良いモデルの学習方法をアドバイスする最も単純な方法は、大量の学習方法の候補について実際にモデルを学習させてみることである。これら無数の候補の中から得られた最も良いモデルおよびその学習方法を開発者にフィードバックすれば良い。しかし、当然ながらこの方法は大量の時間と計算資源を必要とするため現実的ではない。現実的な時間・計算資源の中で開発者へ良いアドバイスをするためには、大量の学習方法の候補について実際にモデルを学習させることなしに良い学習方法を見積もる必要がある。そこで、研究課題 1 では、学習なしに良い学習方法を見積もる推定法の研究に取り組む。

研究課題2．悪いモデルの修理

既に学習によりモデルが得られている場合には、より良いモデルを得るためにゼロから再度学習して新しいモデルを構築するのは無駄が多い。このような場合、開発者が欲するアドバイスはモデルを改善できる簡便な修理法である。研究課題1と同様に、この問題についても原理的には大量の修理法の候補について実際にモデルを修理してみることで、最も良い修理法を開発者にフィードバックできる。もちろん、研究課題1の場合と同様にこの方法は大量の時間と計算資源を必要とするため現実的ではない。

本研究課題特有の問題として、機械学習モデルの修理法には現状確立された技術が存在しないという点がある。この点に対して、本研究課題では継続学習 [1, 2] という方法に着目する。継続学習とは、データの追加・変更やモデルの拡張など、データやモデル周りに変化があった場合に、モデルに追加の学習を施すことでそれら変化に適応するようにモデルを変更する方法である。元のモデルが正しく予測できなかったデータを正しく予測できるように追加の学習をする、という意味でモデルの修理も継続学習の一種として捉えることができる。そこで、本研究課題では修理法として継続学習を対象とし、実際に修理することなしに良いモデルの修理法を見積もる方法の研究に取り組む。

3. 研究の方法

研究課題1．良い学習方法の推定

本研究課題では特に XAI の「影響の大きいデータを見つける問題 [1, 2]」と、「良い学習方法を見積もる問題」との以下のような類似点に着目する。

- 影響の大きいデータを見つける問題
  もしもモデルの学習時点であるデータがなかったとしたら得られるモデルはどう変わるか？

- 良い学習方法を見積もる問題

もしもモデルの学習にある方法を採用したら得られるモデルはどう変わるか？

これらの類似点をもとに、現行の XAI の技術 [1, 2] を「良い学習方法を見積もる問題」へと拡張することで、本研究課題の解決に取り組む。このために、本研究課題では以下の通り 3 つの小課題を設ける。

- 「良い学習方法を見積もる問題」の数理的な定式化と推定アルゴリズムの考案
  - 既存の XAI 研究をベースとして「良い学習方法を見積もる問題」を数理的な問題として定式化し、推定アルゴリズムを考案する。既存研究では「あるデータがなかったとしたら」という条件を、モデルを学習する際の最適化の目的関数を適切に書き換えることで、影響の大きいデータの推定問題を数理的に定式化している。これに対し、「ある学習方法を採用したら」という条件は、最適化の目的関数以外にも、最適化を解くためのアルゴリズムやデータの処理など様々な箇所への変更を要求する。このような広範な変更を記述して定式化する枠組みとしては、研究提案者が過去に考案した確率勾配降下法に基づく方法が有望だと考えている。そのため、研究提案者の方法を基本として問題の定式化および推定アルゴリズムの考案に取り組む。
- アルゴリズムの効率化
  - 既存研究で提案されている方法はどれも相応の計算コストが必要となる。XAI 研究を基盤とする本研究においても、同様にアルゴリズムの計算コストが課題となることが想定される。ここでは特に問題を近似的に解く計算効率の良い近似アルゴリズムの確立を目指す。
- 実データを用いた評価
  - これまでに開発したアルゴリズムの有効性を評価する実験を行う。ここでは特に、開発したアルゴリズムを使うことで、(i) 学習なしに良い学習方法が見積れるか、そして (ii) 見積もりをもとに精度の高いモデルが得られるか、の 2 点について評価を行う。加えて、アルゴリズムの汎用性を調べるために様々な機械学習の問題、例えば画像識別や音声認識、機械翻訳など複数の応用分野のデータ・問題において評価を行う。

研究課題2．悪いモデルの修理

本研究課題では以下の通り 3 つの小課題を設ける。

- モデル修理の継続学習による定式化と有効性の検証
  - まず継続学習により実際にモデルの修理が可能なことを検証する。そのために、まずはモデル修理の問題を継続学習の問題として再定式化する。そして、既存の継続学習のアルゴリズムを適用することで実際にモデル修理の有効性を実験的に検証する。
- 「良い修理法を見積もる問題」の数理的な定式化と推定アルゴリズムの考案
  - 継続学習の問題をベースに研究課題 1 の定式化を拡張する。具体的には、「ある修理法を採用したら」という条件を考え、継続学習の目的関数やモデルの拡張、修理アルゴリズムの変更、など広範な変更を記述して定式化する枠組みを考案する。
- 実データを用いた評価
  - 開発したアルゴリズムの有効性を評価する実験を行う。ここでは特に、開発したアルゴリズムを使うことで、(i) 学習なしに良い修理法が見積れるか、そして (ii) 見積もりをもとにモデルの精度が向上する修理ができるか、の 2 点について評価を行う。研究課題 1 と同様に、こちらについてもアルゴリズムの汎用性を調べるために様々な機械学習の問題、例えば画像識別や音声認識、機械翻訳など複数の応用分野のデータ・問題において評価を行う。

4. 研究成果

(1) 「良い学習方法の推定」の成果

「良い学習方法の推定」の代表的な方法はデータクレンジングである。機械学習モデルを学習する際に、学習データに不適切なデータが含まれている場合、学習されたモデルの精度は悪化してしまう。データクレンジングの目的はこのような不適切なデータを自動的に検知することである。本研究ではこのデータクレンジングの問題を強化学習まで拡張し、強化学習に対するデータクレンジング手法を開発した。

本研究ではまた従来のデータクレンジング手法を改変・拡張することで、類似データによる説明手法を構築できることも明らかになった。従来のデータクレンジングでは外れ値のような異常値を際立たせる手法が効果的であったが、類似データによる説明ではむしろ外れ値を無視して“通常の”データを際立たせる必要がある。本研究において、従来のデータクレンジング手法があるベクトルの長さに強く依存していることを発見し、この長さが異常を際立たせる要因になっていることを明らかにした。そこで、このベクトルの長さを正規化して長さの影響を排除したところ、改変後の手法は“通常の”データの中から類似データを適切に検索できることがわかった。

(2) 「悪いモデルの修理」の成果

「悪いモデルの修理」ではモデルを再学習することなく、モデルの誤分類を削減する方法の研究に取り組んだ。本研究ではブースティングなどのように弱学習器を用いて逐次的にモデルをアップデートす

る手法に着目した。そして、この弱学習器を適切に設計する問題として「悪いモデルの修理」の問題を定式化した。ここでは特に「修正においてモデルの挙動ができる限り変化しない」ことを要請として設け、このような制約を満たす弱学習器を学習する方法を開発した。具体的には、弱学習器として決定木を採用し、多くの葉ノードの値がゼロとなるように正則化を加えて学習する方法を考案した。これにより、値が非ゼロの葉ノードでのみモデルの予測が修正される「悪いモデルの修理」方法が実現できる。

本研究ではさらにこの弱学習器を用いた修正方法を発展させる研究にも取り組んだ。上記の決定木を用いた修正方法では、一通りの修正パターンしか得られなかったのに対し、発展させたパターンマイニング手法ではモデルの多様な修正パターンを発見することができる。これにより開発者は様々な修正パターンの中から適切なものを選択することが可能となった。

本研究ではまたモデルの判断の不確実性を修正する方法として、ユーザの意思決定を考慮した不確実性を推定手法を開発した。これは多くの機械学習モデルが不確実性を適切に推定できない点を改善するための方法であり、特にユーザがモデルの判断に基づいて意思決定を行う場合に効果的な方法である。

(3) 本研究のインパクト及び今後の展望

本研究により、「良い学習方法の推定」及び「悪いモデルの修理」という 2 通りの説明駆動型の機械学習モデル開発手法を考案した。これらの方法により、モデル開発者は自身の経験や勘に強く頼ることなく自動的にモデルを改良・修理することが可能となる。ひいては、機械学習モデル開発の属人性を解消して、誰でも精度の高いモデルを作れるようになる道が開かれた。

もちろん本研究の成果は説明駆動型の機械学習モデル開発手法への第一歩である。個別の手法の高度化には今後も引き続き取り組む必要がある。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Hirofumi Suzuki, Hiroaki Iwashita, Takuya Takagi, Keisuke Goto, Yuta Fujishige, Satoshi Hara
２．発表標題 Explainable and Local Correction of Classification Models Using Decision Trees
３．学会等名 The 36th AAAI Conference on Artificial Intelligence（国際学会）
４．発表年 2022年

１．発表者名 Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, Kentaro Inui
２．発表標題 Evaluation of Similarity-based Explanations
３．学会等名 The 9th International Conference on Learning Representations (ICLR'21)（国際学会）
４．発表年 2021年

１．発表者名 Ulrich Aivodji, Hiromi Arai, Sebastien Gambs, Satoshi Hara
２．発表標題 Characterizing the risk of fairwashing
３．学会等名 Neural Information Processing Systems 34 (NeurIPS'21)（国際学会）
４．発表年 2021年

１．発表者名 藩丹青
２．発表標題 Data Cleansing for Reinforcement Learning with Least Squares Temporal Difference
３．学会等名 第23回情報論的学習理論ワークショップ（IBIS2020）
４．発表年 2020年

１．発表者名 Hirofumi Suzuki, Hiroaki Iwashita, Takuya Takagi, Yuta Fujishige, Satoshi Hara
２．発表標題 Rule Mining for Correcting Classification Models
３．学会等名 23rd IEEE International Conference on Data Mining（国際学会）
４．発表年 2023年

１．発表者名 宗近康平，原聡
２．発表標題 決定損失の期待値と分散を用いた分類モデルの較正
３．学会等名 2023年度人工知能学会全国大会
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関