公開

研 究 終 了 報 告 書

「機械学習モデルとユーザのコミュニケーション:モデルの説明と修正」

研究期間:2020 年 12 月～2024 年 3 月
研 究 者:原 聡

1. 研究のねらい

AI の基盤技術として機械学習は欠かせない技術である。近年、大量のデータ(e.g. 画像と物体ラベル、英文と和訳文)を使って、高性能な機械学習モデル(e.g. 画像認識モデル、機械翻訳モデル)の学習が可能となっている。しかし、「信頼される AI」の実現には機械学習モデルの性能が高いだけでは不十分である。機械学習モデルの信頼性において主要な問題の一つが「モデルが“ブラックボックス”である」ことである。機械学習モデルは完璧ではなく、ときに間違いをおかす。しかし、モデルがブラックボックスであるために、その間違いの理由やその修正方法をユーザが窺い知ることは困難である。結果、「モデルの判断基準がおかしいのではないか」など、ユーザはモデルへの不信を募らせることとなる。

本研究の大目標は「モデルのブラックボックス性に対するユーザの不信を解消する」ための技術を確立することである。この目標の達成にはモデルとユーザとの間での適切な(言語的な手段に限らない)“コミュニケーション手段”の開発が必要である。この大目標に向けて、本研究ではコミュニケーション手段として特に研究代表者が有望だと考えているモデルの“説明”と“修正”のための技術の開発に取り組む。

本研究で考える“説明”と“修正”はそれぞれがモデル→ユーザ方向とユーザ→モデル方向のコミュニケーション手段に相当する。“説明”とはモデルが予測結果だけでなく、さらにその予測を補助する追加情報を提供する機能である。“修正”はユーザがモデルに干渉して、モデルをユーザの期待に沿ったものへと改変する機能である。これらのコミュニケーション手段の提供が機械学習モデルへのユーザの信頼醸成に効果的であることは、既にいくつかの研究で実証されている。

本研究ではコミュニケーション手段としての“説明”と“修正”の技術の開発を通じて、ユーザの機械学習モデルへの信頼醸成、ひいては「信頼される AI」の実現へと貢献することを目指す。

2. 研究成果

(1)概要

研究のねらいに対して、“説明”および“修正”技術の研究に取り組んだ。“説明”については主要な説明方法の一つである「事例ベースの説明」を対象に「(研究テーマ A) “説明”のための関連性指標の開発と性能検証」の研究に取り組んだ。“修正”についても当初は事例ベースの方法の研究開発を想定していたが、研究を通じて当初想定よりも問題が難しいことがわかり、研究期間中盤以降はより多様な方法案をベースに研究を行った。“修正”については特に有望なアプローチとして「(研究テーマ B) 再学習による“修正”のための学習の安定化」「(研究テーマ C) モデルの“修正”のためのハイパーパラメータ最適化」に注力して研究を行った。

“説明”の研究の成果

「事例ベースの説明」で用いられるデータ間の関連性・類似性の指標を対象に、「指標の性能

評価方法の検討」および「指標の設計」の研究に取り組んだ。前者については「同一クラステスト」「同一サブクラステスト」という評価方法を考案し、複数の指標の良し悪しの比較を可能とした。また、既存の指標の評価を通して、「モデルの損失関数のパラメータ勾配のコサイン類似度（Cosine of Gradient; GC）」が特に良い指標であることがわかった。また、その良さを分析した知見を「指標の設計」へと流用し、既存の指標を改善した指標を設計した。しかし、これら改善した指標はGCと同等の性能であったため、実用的にはGCが最も優れた指標であると結論づけた。

“修正”の研究の成果

“修正”の方法として「再学習」と「ハイパーパラメータ最適化」の二つの観点から研究に取り組んだ。「再学習」の目的は、“修正”のためにモデルを再学習する過程で当初のモデルとは全くの別物になってしまうことを回避することである。この目的に対して、代表的なモデルである決定木を対象に学習を安定化させるアルゴリズムを考案した。「ハイパーパラメータ最適化」の目的はモデルを所望の挙動に近づけるようにハイパーパラメータをチューニングすることである。最適化のためにはハイパーパラメータの勾配が必要となるため、その計算方法に着目して研究を行った。特に近年のビッグデータでの学習に対応するため、分散学習に着目し分散環境下での効率的なハイパーパラメータ勾配の計算方法を開発した。

上記の研究成果は機械学習の主要国際会議論文として採択（一部は投稿中）され、また一部は国内学会での受賞やメディア掲載にもつながった。

（2）詳細

研究テーマA「“説明”のための関連性指標の開発と性能検証」

本テーマでは機械学習モデルの代表的な説明方法の一つである「事例ベースの説明」の研究に取り組んだ。

【事例ベースの説明】

「事例ベースの説明」の目的はあるデータに対するモデルの予測への説明として、そのデータに何らかの尺度で関連・類似するデータを提示する方法である。例えば図1のように、ある画像をモデルが「タゲリ（鳥の一種）」と分類したときに、同種の「タゲリ」の画像をユーザに提示することで、ユーザが鳥の専門家でなくとも予測の正しさを確認し納得することができる。

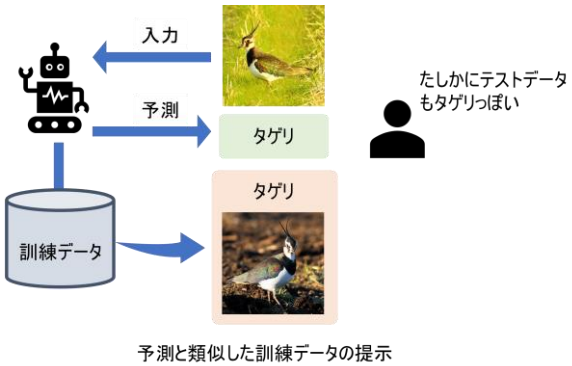


図1. 事例ベースの説明の例

【研究内容】

「事例ベースの説明」にはデータ間の関連性・類似性を測る指標が必要である。本研究では「指標の性能評価方法の検討」および「指標の設計」に取り組んだ。

「指標の性能評価方法の検討」 では、問題を分類問題に限定し、「同一クラステスト」と「同一サブクラステスト」という二つの性能評価方法を考案した。「同一クラステスト」では、指標によって関連性・類似性が高いとされたデータが、元々の説明対象のデータと同じクラスである場合にその「事例ベースの説明」を「成功」と評価する。「同一クラステスト」を多数のテストデータに適用することで指標の成功率を測定し、その成功率が高い指標を良い「事例ベースの説明」の指標と判定する。「同一サブクラステスト」は「同一クラステスト」をより一般化してクラス内構造までも考慮して評価する指標である。

「指標の設計」では既存の指標を含めた複数の指標およびデータ表現ベクトルを用いて「同一クラステスト」および「同一サブクラステスト」を行うことで、成功率の高い良い指標を特定した。結果、図 2 のように「モデルの損失関数のパラメータ勾配のコサイン類似度（Cosine of Gradient; GC）」が特に高い成功率を記録し、良い指標であることがわかった。興味深い点として、Neural Tangent Kernel に代表されるような勾配の内積は成功率が低く、内積に正規化を取り入れてコサイン類似度とすることで成功率が大きく改善することがわかった。この知見をもとに、既存の指標に正規化を取り入れた指標の改善にも取り組んだ。しかし、これら改善した指標も元々の GC と同等の性能であり、その単純さから GC が最も優れた指標であると結論づけた。図 3 は GC による類似データの提示の例である。単に鳥の画像を類似データとして提示しているだけでなく、画像のより詳細なコンテキスト（鳥の向きや種類）を反映した類似データの提示ができている。

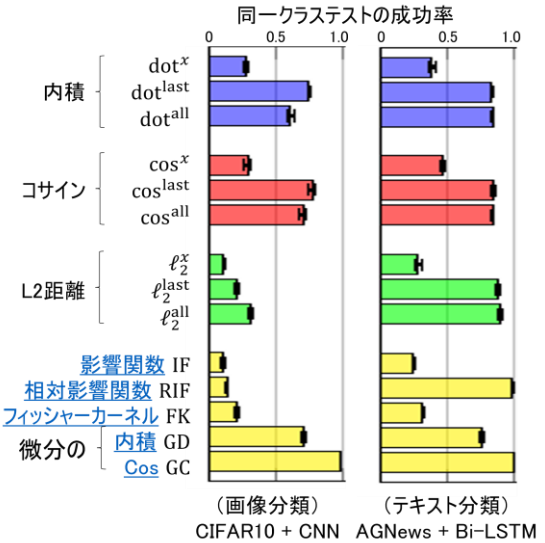


図 2. 同一クラステストの成功率

図 3. GC による類似データの例

【参考文献】

- Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, Kentaro Inui. Evaluation of Similarity-based Explanations. The 9th International Conference on Learning Representations (ICLR'21), 2021.

研究テーマ B「再学習による“修正”のための学習の安定化」

機械学習モデルにユーザが干渉する方法として「学習の安定化」の研究に取り組んだ。

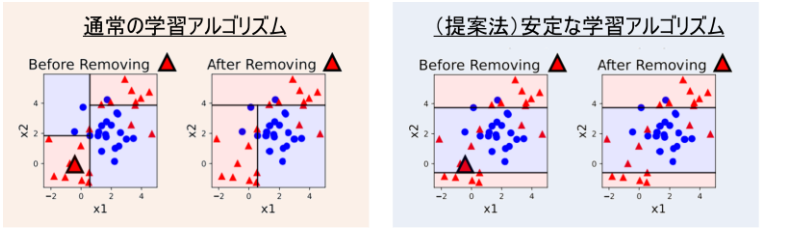
【モデルへのユーザの干渉と安定化の関係】

ユーザが機械学習モデルの挙動について不満を持ったときに、モデルを満足のいくものに“修正”することができれば、ユーザの満足度が高まりモデルへの信頼が醸成されると期待される。“修正”の最も単純な方法はモデルの再学習である。再学習の際に正則化などによりモデルをユーザの所望するものへと近づけることができる。しかし、再学習の過程でモデルが当初のモデルとは全くの別物になってしまうことは好ましくない。そのため、モデルの学習に多少の摂動を加えても依然として（多少の変化はあれど）当初のモデルに似たモデルが学習されること、つまりモデルの学習アルゴリズムが摂動に対して安定であることが望ましい。

【研究内容】

機械学習モデルの学習アルゴリズムを安定化させる方法について研究を行った。ここでは摂動として特に「ランダムな訓練データの削除」を対象とした。これは「ランダムな訓練データの削除」が最も単純かつ本質的な摂動だからである。実際、「ランダムな訓練データの削除」は訓練データの分布を（平均的には）変化させない。そのため、同じ分布のデータから学習したモデルは基本的にはほぼ同一のものとなることが期待される。しかし、多くの機械学習モデルの学習アルゴリズムは「ランダムな訓練データの削除」に対して安定ではない。

本研究では基本的なモデルの一つである決定木を対象にその学習アルゴリズムの安定化に取り組んだ。決定木はそれ単独で解釈性が高く説明に有効なモデルなだけでなく、ランダムフォレストや GBDT などのアンサンブルモデルの構成要素でもある。そのため、決定木の学習アルゴリズムを安定化させることで、決定木の解釈性の安定化、そしてランダムフォレストやGBDTの学習の安定化にもつながる。提案法では、従来の学習アルゴリズムの貪欲法を指数メカニズムにより乱択化する。そして、この乱択化により、「ランダムな訓練データの削除」に対して学習アルゴリズムが安定化されることを理論的・実験的に示した。図 4 は従来の学習アルゴリズムと提案法による安定化の例である。



データ（▲）の削除で決定木が変わる。
→“不安定”な学習アルゴリズム

データ（▲）の削除で決定木が変わらない。
→“安定”な学習アルゴリズム

図 4.決定木の学習の安定化の例

【参考文献】

- Satoshi Hara, Yuichi Yoshida. Average Sensitivity of Decision Tree Learning. The 11th International Conference on Learning Representations (ICLR'23), 2023.
- 原聡, 竹内孝, 吉田悠一. 階層クラスタリングの安定化. 人工知能学会全国大会, 2023.

研究テーマ C「モデルの“修正”のためのハイパーパラメータ最適化」

本テーマでは機械学習モデルにユーザが干渉する方法として「ハイパーパラメータ最適化」の研究に取り組んだ。

【モデルへのユーザの干渉とハイパーパラメータ最適化の関係】

ユーザがモデルの挙動に不満を持った時に、モデルの学習の際に正則化などによりモデルをユーザの所望するものへと近づけることができる。このとき正則化の強さなど、外部的に決定する必要のあるハイパーパラメータが存在する。ユーザの所望の“修正”を実現するためにはハイパーパラメータを適切に決定する必要がある。

【研究内容】

ハイパーパラメータ最適化のために、ハイパーパラメータ勾配の計算方法について研究を行った。ハイパーパラメータ勾配は一般にヘッセ行列の逆行列を含む重い計算が必要となる。本研究では特に分散学習のような複数クライアントが共同してモデルの学習を行う設定下での効率的なハイパーパラメータ勾配の計算方法を開発した。従来の方法ではクライアント間でヘッセ行列の通信を必要としたり、また中央集権的なサーバの存在を前提としていた。これに対して、提案法ではヘッセ行列の通信を不要とし、また完全分散型かつ動的な有向ネットワーク上でのハイパーパラメータ勾配の計算を可能とした。これにより、初めて完全分散型の環境下でハイパーパラメータ勾配が効率的に計算可能となった。

【参考文献】

- Naoyuki Terashita, Satoshi Hara. Decentralized Hyper-Gradient Computation over Time-Varying Directed Networks. arXiv2210.02129, 2022.
- Naoyuki Terashita, Satoshi Hara. Cooperative Personalized Bilevel Optimization over Random Directed Networks. The 14th Workshop on Optimization and Learning in Multiagent Systems, 2023.

当さきがけ領域内外の研究者や産業界との連携

研究テーマ B は吉田悠一氏（NII/さきがけ数理構造活用）との協業から立ち上がったテーマである。機械学習の不安定性に関心があった研究代表者とアルゴリズムの安定性を研究していた吉田氏との興味とが合致して共同研究が始まった。また当該テーマに竹内孝氏（京大/さきがけ信頼される AI）が興味を持ちさらなる共同研究へと発展した。

また、研究代表者が委員を務める IBISML 研究会で「信頼される AI」をテーマに企画セッションを組んだ際にさきがけ・CREST 領域内でのつながりで何名かの研究者に講演を依頼した。

産業界との連携の一つとしては、日立研究所との「ハイパーパラメータ最適化」の共同研究がある。これは本さきがけ研究の元になった研究代表者の研究の一つを発展させる形で始まった共同研究である。

3. 今後の展開

本研究の“説明”の成果のうち、「事例ベースの説明」の評価方法である「同一クラステスト」「同一

サブクラステスト」は特に難しい計算や実装は不要なため、誰でも必要に応じて利用できる状況である。「指標の設計」において特に有効であると判明した GC についても、必要な計算リソースさえあれば現段階でも実用可能である。ただし、近年の巨大な深層ニューラルネットワークについては必要な GPU メモリが膨大なため、GC を手軽に使えるようにするためには近似計算などの導入による計算の簡略化などを検討する必要がある。これら計算効率化の方法はおそらく数年あれば実現可能だろうと思われる。

“修正”については、まだ多くの研究が必要だと考えられる。本研究の成果である「学習の安定化」や「ハイパーパラメータ最適化」はどれもそれ単体では利用可能であるが、特定のモデルや問題設定に紐づいているため、汎用的な“修正”方法とは言い難い。また、“修正”の研究の過程において、ユーザの所望の挙動へとモデルを改変する過程で、モデルの性能が劣化したり、またその挙動が元のモデルから著しく変化するなどの好ましくない変化が多く観察された。これら好ましくない変化を避けつつ、所望のモデルへと“修正”することが、そもそも原理的に可能なのか、可能であるならばどういった方法が良いのか、といった点についての研究は今後も継続していく必要がある。これらの研究から実用的な方法が確立されるまでには数年から 10 年程度の地道な研究が必要ではないかと思われる。ただし、近年モデルから特定のデータを忘却させる Unlearning の技術への社会的・技術的な関心が高まっている。Unlearning もモデルへとユーザが干渉する方法の一つであるため、Unlearning の研究から汎用的な“修正”方法への糸口が見つかるかもしれない。

4. 自己評価

【研究目的の達成状況】 全体としては概ね目的に沿った研究が遂行できたと考えている。ただし、その内訳は当初の予定からは多少変更があった。“説明”についてはある程度は予定の通りに遂行できたと考えている。“修正”については当初計画の事例ベースの方法では上手くいかないことがわかったために、研究期間中盤以降に選択肢を広げて複数のアプローチを検討した。その結果、当初計画とは異なる方法へと研究を発展させることができた。

【研究の進め方】 研究実施体制および研究費執行は当初計画から何点か変更があった。体制については当初は学生を RA として雇用することを計画していたが、RA を希望する学生がいなかったために断念した。研究費については、国際会議参加や海外大学への研究訪問を想定していたが、コロナ禍のためにこれらの計画の多くは断念することとなった。

【研究成果の科学技術および社会・経済への波及効果】 “説明”の研究成果については、すでにある程度は実用可能な状況となっている。特に本研究で明らかになった類似度指標 GC は「事例ベースの説明」の指標として極めて有望であると考えている。今後、GC の改良（特に計算効率性）およびその有効性の実証を積み重ねていくことで、機械学習モデルの説明方法の一つとして確立できると考えている。今後、社会や科学の分野で機械学習技術の活用が広がっていくにあたって、GC は「事例ベースの説明」の有効な方法として重要な役割を果たすと考えている。

“修正”の研究成果については、汎用的な“修正”方法を確立するには至らなかった。その要因は“修正”がそもそも難しい問題であるためである。しかし、“修正”が可能になればユーザがより積極的に機械学習モデルに干渉できるようになり、「信頼される AI」の確立に大きく貢献すると考え

られる。そのため、“修正”の研究は当さきがけ研究終了後も継続して取り組むべき重要な課題であると考えている。

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数: 4件

1. Satoshi Hara, Yuichi Yoshida. Average Sensitivity of Decision Tree Learning. The 11th International Conference on Learning Representations (ICLR'23), 2023.
決定木は解釈性の高いモデルの一つである。しかし、従来の決定木の学習方法では、訓練データの摂動(例えばデータ 1 点の削除)に対して決定木の構造が大きく変化することがある。これは決定木を通じて我々が読み取るデータの解釈や意思決定のプロセスが大きく変化しえる不安定なものであることを示唆している。本研究ではこの問題を解決するために、決定木の学習を安定化させ、構造変化の少ない安定した決定木を学習する方法を提案した。
2. Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, Satoshi Hara. Characterizing the risk of fairwashing. Advances in Neural Information Processing Systems 34 (NeurIPS'21), 2021.
Fairwashing とは、差別的なモデルに対して嘘の説明を生成することでモデルを公平だと誤認させる方法である。本研究ではこの Fairwashing の性質を包括的に調査し、その潜在的なリスクを明らかにした。具体的には、適当な方法で生成された嘘の説明が「異なる説明対象および異なるモデルに対しても一定程度有効であること」を実験により明らかにした。この結果は嘘の説明が汎用的であること、そしてその検知が極めて難しいことを示唆している。
3. Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, Kentaro Inui. Evaluation of Similarity-based Explanations. The 9th International Conference on Learning Representations (ICLR'21), 2021
類似データの提示は、機械学習における説明法の一つである。これまの研究でデータ間の類似性の様々な指標が提案されてきた。本研究ではこれら類似性指標の評価方法として、同一のクラスのデータを類似データとして取得できることを良い指標の要件とする「同一クラステスト」を提案した。既存の指標を同一クラステストで評価した結果、中には類似性指標として好ましくない指標が存在すること、また損失勾配のコサイン類似度が特に良い類似性指標であることが明らかとなった。

(2)特許出願

研究期間全出願件数: 0 件(特許公開前のものは件数にのみ含む)

1	発明者	
	発明の名称	
	出願人	
	出願日	
	出願番号	
	概要	
2	発明者	
	発明の名称	
	出願人	

	出　願　日	
	出 願 番 号	
	概　　要	

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

【受賞】

- 2023 年度人工知能学会全国大会優秀賞
  - （対象発表）原聡, 竹内孝, 吉田悠一. 階層クラスタリングの安定化.
- 最優秀プレゼンテーション賞, 第 25 回情報論的学習理論ワークショップ(IBIS2022), 2022.
  - （対象発表）原聡, 吉田悠一. 決定木学習の安定化

【報道・プレスリリース】

- 阪大など、AI の階層型クラスタ分析を安定化　分岐構造、同型作成で高性能実現, 日刊工業新聞, 2023/6/21.
- 説明可能な AI は本当に適切な根拠を示せているのか　AI の説明能力を客観的に評価するための方法論の構築, 阪大産研, 2021/4/27