公開

研 究 終 了 報 告 書

「誤りがないことを保証する検証器つき機械学習の研究」

研究期間:2020 年 11 月～2024 年 3 月
研 究 者: 西野 正彬

1. 研究のねらい

背景:検証可能な誤りと信頼される AI

現代の AI 技術の中核をなす機械学習技術は確率統計の枠組の上に構築されている。近年の深層学習の発展により、機械学習モデルの予測精度は飛躍的に向上しつつあるが、確率統計を用いているために機械学習モデルの予測誤りをゼロにすることは本質的に困難である。そのため、機械学習モデルを利用するシステム（AI システム）は、モデルの出力が誤りに対処するために複雑になりがちであるほか、システムのデバッグ・テストにかかるコスト画像代するといった問題が発生する。このように、予測誤りを避けることが本質的に困難であるという統計的機械学習の性質が、AI 技術が広く利用されることを妨げる障壁となっている。

本研究では上記の**機械学習モデルの出力に誤りがないことを保証できないこと**こそが、AI 技術を信頼できない主要因であると考え、そのような誤りをなくすことで「信頼される AI」を実現する方法を検討した。機械学習モデルの予測誤りを完全になくすことは難しい。一方で、誤りの種類によってはその有無を外部から判定することが可能である。外部からその有無を判定可能な誤りのことを**検証可能な誤り**とよぶ。

検証可能な誤りの例をいくつか示す。近年注目を集めている大規模言語モデルなどは、人が作成したものと遜色ない、自然な文書を生成することが可能である。一方で深層学習モデルによって生成された文書は事実と異なる内容を含む可能性がある。このような種類の誤りも外部から検証可能である。すなわち、事実を集めたデータベースを用意して、言語モデルが生成した文書とデータベースとを比較することで、生成された文書とデータベースのエントリとの間に矛盾がないかを判別することができる。そのほか、医療や自動操縦等、安全性への配慮が求められる場面では、予測を誤ることのリスクが大きい。こうした場面では、特に危険が大きいと考えられる誤りをあらかじめデータベースとして準備することで、機械学習モデルがそのような誤りを起こさないことを保証できる。

アプローチ:検証器つき機械学習

本研究では外部から検証可能な誤りを検出できる検証器を、機械学習モデルに接続して利用する**検証器つき機械学習モデル**を検討する。背景知識や制約条件を扱うことができる機械学習モデルは検証器つき機械学習のほかにも存在するが、既存方式と比較して以下の利点がある。

(1) **汎用性**: 検証器は多様な機械学習モデルと組み合わせることができるため、様々なタスクに特化した、高性能な機械学習モデルの出力に関する保証を与えることができる。

(2) **予測時の制約追加**: 検証器を用いることで、モデルの学習時には明らかでなかった仕様を、予測時に考慮できる。モデルの学習と、学習したモデルを用いた予測とでは、一般的に前者がより計算機リソースを消費する。そのため、近年では大規模な計算リソースを用いて学習した

モデル（学習済みモデル）を様々な場面で利用することが一般的である。検証器はこのような利用形態にも適用できる。

(3) **保証可能**：制約を用いる既存のアプローチは、機械学習モデルの予測結果が常に制約を満たすことを保証できるわけではなかった。検証器は機械学習モデルの入出力が与えられた仕様を満たすことを保証可能である。

利点(1), (2) は、現代的な機械学習モデルの利用形態に合致している。また、利点(3)は機械学習モデルを利用した AI システムを構築・運用するうえでコストを低減することに貢献する。以上より、検証器つき機械学習モデルは、現代的な機械学習モデルを用いた AI システム構築のコスト低減への貢献ができる。

検証器つき機械学習の課題

検証器の利用によって既存の機械学習モデルの予測結果に対する保証つけることができる一方で、検証器を接続して利用することによる影響は明らかではなかった。すなわち、機械学習モデルの予測精度（予測誤差）が検証器の利用によってどのように変化するか、検証器を用いたときに学習・推論アルゴリズムの実行時間にどの程度のオーバーヘッドが発生するかといった点が知られていなかった。仮に検証器を接続することによって、学習が失敗したり計算効率が大幅に悪化したりといった副作用があるならば、検証器が有用な場面が制限される。また、検証器によってどのような仕様を扱うことができるかも明らかにする必要がある。そこで本研究では検証器つき機械学習のコンセプトの確立と、機械学習モデルにおいて重要な性質である、理論的な基礎を明らかにすることを目的として検討をすすめた。

2. 研究成果

(1) 概要

本研究プロジェクトでは統計的機械学習モデルの検証に関する研究を遂行して、いくつかの成果を創出した。代表的な研究として、機械学習モデルの出力が仕様を満たすかどうかを検証する検証器を機械学習モデルに接続した、**検証器つき機械学習モデル**のコンセプトを世界で初めて提唱するとともに、検証器を機械学習に接続することの影響を理論的に解析した。具体的には、検証器で扱う仕様をいくつかの典型的なカテゴリに分類するとともに、それらのカテゴリの仕様を扱う検証器を用いたときに機械学習モデルの予測誤差（汎化誤差）の理論的な上限がどのように変化するかを明らかにした。また、検証器を利用したときに機械学習の効率がどのように変動するのかも明らかにした。

代表的な理論的な成果は以下のとおり。機械学習モデルの汎化誤差に関しては、学習時と推論時の両方で検証器を用いたならば、元の機械学習モデルの Rademacher 複雑に基づく汎化誤差の上限が求まることを示した。この結果は、任意の仕様のもとで検証器を用いても、元のモデルのもつ汎化誤差の上限が悪化しないことを示している。すなわち、学習がうまくいくモデルに検証器を接続しても、汎化誤差の意味では悪影響を与えないことを示唆している。また、学習時には検証器を用いずに、推論時にのみ検証器を用いる場合、元のモデルが PAC 学習可能であれば汎化誤差が小さくなることを示した。一方で PAC 学習可能ではない

場合、最適でないモデルが得られる可能性があることも示した。この結果は、PAC 学習可能かどうかが推論時にのみ検証器を用いて誤差の小さなモデルを得られるかどうかの判定に利用できることを示唆している。

(2) 詳細

主要な成果 1: 検証器つき機械学習モデルの汎化性能解析 (Neural Information Processing Systems (NeurIPS) 2022 採録, 成果リスト 1)

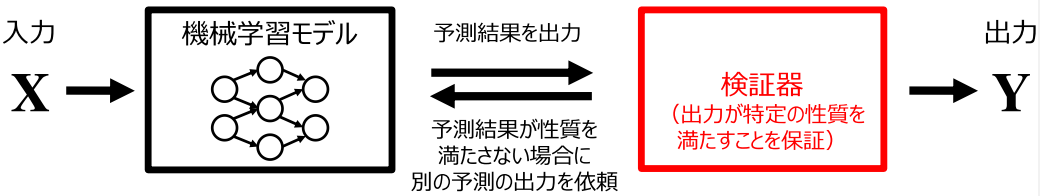

図 1 検証器つき機械学習モデルの構成図。従来の機械学習モデルに、出力を検証するための検証器を追加することで、モデルの出力が特定の性質を満たすことを保証する。

多クラス分類を行う機械学習モデルに、単一の入出力ペアに関する仕様を扱う検証器を接続した場合に、機械学習モデルの汎化誤差の上限がどのように変化するかを解析した。K クラスへの多クラス分類を行う機械学習モデルは、入力 $x \in \mathcal{X}$を受け取り、出力$y \in \mathcal{Y} = \{1, \dots, K\}$を出力する関数$h: \mathcal{X} \to \mathcal{Y}$として表現できる。単一ペアに関する仕様とは機械学習モデルの入力x と出力 y のペア (x, y)を受け取り、0 または 1 を返す関数$c(x, y)$として表現できるような仕様のことを指す。関数cは入出力ペアが仕様を満たすときに 1, そうでないときに 0 を返す。例えば入力xが特定の値aをとったときには出力yは値bをとってはいけない、という仕様は、入力が$x = a, y = b$ の場合に 0 を、それ以外の場合に 1 を出力する関数cとして表現できる。ある仕様cを満たすかどうかを判定する検証器を機械学習モデルに接続することは、機械学習モデル$h: \mathcal{X} \to \mathcal{Y}$を関数$h_c: \mathcal{X} \to \mathcal{Y}$に変換することに相当する。ここで$h_c$は仕様$c$を満たすように$h$ の出力を補正した関数であり、

$$h_c(x) = \begin{cases} h(x) & if \quad c(x, h(x)) = 1 \\ y' & if\ c(x, h(x)) = 0 \end{cases}$$

として定義される。すなわち、機械学習モデルhの入出力ペア$\big(x, h(x)\big)$が仕様を満たすのであれば$h_c(x)$ は $h(x)$ と等しく、そうでないのであれば$c(x, y') = 1$を満たす$y' \in \mathcal{Y}$を出力する。検証器つき機械学習モデルの概要を図1に示す。

汎化性能解析は機械学習モデルが未知の入力に対してどの程度正確な予測を行えるかを理論的に明らかにすることを目的とする解析方法である。機械学習モデルは訓練例を用いてモデルを学習する。そのため、あるモデルが既知の訓練データに対してどの程度正しく予想可能かを知ることは容易であるが、未知の入力に対してどの程度正しく予測可能かは明らかではない。汎化誤差解析では、未知の入力に対する予測誤差(汎化誤差)が、訓練例に対する予測誤差(訓練誤差)よりどの程度大きくなりうるかを理論的に明らかにすることで、ある

機械学習モデルが未知の入力に対しても誤差を小さくできることを示す。

機械学習モデルhの汎化誤差に関する理論的な上限が知られていたとしても、検証器を接続することで得られるモデルh_cの汎化誤差の上限は明らかではない。そこで本成果ではh_cの汎化誤差の上限を汎化誤差解析によって理論的に導出した。解析を行うにあたり、検証器の利用形態を(1)学習済みのモデルに検証器を接続して推論時に用いる設定 (Inference Time Verification, ITV) と(2) 学習時と推論時のいずれにも検証器を接続して利用する設定 (Learning Time Verification, LTV) の 2 種類に分類した。ITV 設定はモデルの学習を行う際に仕様が未知であってもよいことから、LTV 設定よりも広いユースケースに当てはまる。一方でモデルの学習時に仕様に関する情報が利用できないため、LTV 設定よりも予測誤差が大きくなる可能性がある。

ITV 設定における主要な結果として、機械学習モデルが PAC (Probably approximately correct)-学習可能であるならば、学習されたモデルに検証器を用いた場合の汎化誤差が、高い確率で仕様を満たす他のモデルよりも大きくならないことを示した。一方で PAC 学習可能でないならば、ITV 設定では予測誤差が最小となるモデルが得られないことを示した。この結果は PAC 学習可能性によって ITV 設定で予測誤差が小さなモデルが得られるかを判定できることを示しており、検証器をどのように利用すべきかについて示唆を与えることができる。例えば、学習済みの大規模言語モデルを用いた言語生成において、その出力を仕様に合わせて制御する制約付き言語生成手法など、制約を用いて学習済みモデルの出力を制御する手法は近年活発に研究されている。こうした手法は ITV 設定の一形態とみなすことができる。そのため、もしこれらの手法が PAC 学習可能性を満たしていないのであれば、学習時に仕様の情報を利用することで予測精度の改良の余地があることが示唆される。

学習時と推論時の両方で検証器を扱う LTV 設定に関する主要な結果として、学習されたモデルh_cの汎化誤差が、h_cの訓練誤差と**元の機械学習モデルの** Rademacher 複雑性に基づく項、および訓練例のサンプル数に依存する項の和によって求まる上限によって制限されることを明らかにした。ここで Rademacher 複雑性とは機械学習モデルの表現能力を表す尺度であり、モデルの Rademacher 複雑性が小さいほど、訓練誤差と汎化誤差の上限との差が小さくなる。本結果の重要な点は仕様cに非依存なh_cの汎化誤差の上限が得られたことである。すなわち、学習時にも検証器を用いるのであれば、Rademacher 複雑性に基づく汎化誤差のタイトな上限が得られている機械学習モデルに検証器を接続しても、仕様の種類によらず汎化誤差のタイト上限が得られることを表している。この結果は、機械学習モデルの出力が仕様を満たすことを保証するために検証器を用いても、汎化誤差の上限に与える影響がないこと、すなわち検証器を接続することによって元々学習がうまくいっていたモデルの学習が困難にならないことを示唆しており、検証器の活用を肯定する結果である。

主要な成果 1 の論文では、ほかにも検証器の利用が汎化誤差に与える影響について解析している。具体的には、LTV の問題設定において、Rademacher 複雑性よりもよりタイトな上限を

与えることができる、Local Rademacher 複雑性に基づく汎化誤差の上限も、仕様cに依存しないことを示した。また、多クラス分類の一種である構造化予測問題において、Structured Rademacher 複雑性に基づく汎化誤差の上限も示した。これらの結果はいずれも LTV 設定において検証器を利用しても汎化誤差の上限が悪化しないことを示しており、検証器の LTV 設定での利用を肯定する結果である。

主要な成果 2: ITV 設定における予測誤差の変化と損失関数の関係の理論解析（国際会議 International Conference on Machine Learning (ICML 2024) 採録決定、成果リスト 3）
主要な成果 1 の論文では、検証器つき機械学習のコンセプトを提案するとともに、ITV、LTV の 2 種類の問題設定において検証器の利用が汎化誤差に与える影響を理論的に解析し、ITV 設定においては PAC-学習可能であるならば、推論時に検証器を利用しても汎化誤差の上限が与えられることを示した。一方で PAC-学習可能でない場合には汎化誤差の上限は与えられていなかった。PAC 学習可能性は強い条件であり、実際の機械学習タスクにおいては PAC 学習可能性が満たされない場合も多い。また、ITV は高コストな学習を行うことなく仕様を満たす機械学習モデルを得ることができる方法であることから、その潜在的な需要は大きい。そこで本研究では ITV の設定において、検証器の追加におよって予測誤差がどのように変動するのかを理論的に明らかにした。

本成果では、特定の誤差関数を利用したならば、多クラス分類問題で推論時に使用を追加しても、相対誤差が保存されることを示した。相対誤差とは、理想的なモデルを利用したときの汎化誤差と、実際のモデルの汎化誤差との差を表す。相対誤差が保存されるとは、仕様を追加しないときのあるモデルの相対誤差を a であったとき、使用を追加してモデルの出力を ITV 設定で修正したあとの、モデルの相対誤差（仕様を満たす最適なモデルの汎化誤差との差分）が a よりも小さくなることをいう。相対誤差が保存されるということは、もともと相対誤差が小さなモデルに、推論時に検証器をとりつけても予測誤差は大きくならないということである。すなわち、汎化誤差が保存されているならば、「誤差が小さな良いモデルは、検証器を利用しても良いモデルである」ことがわかるため、検証器を利用する際に学習が必要なのか（LTV）、それとも推論時に検証器を追加する（ITV）だけで十分なのかの判断を的確に行うことが可能となる。

3. 今後の展開
一連の研究によって、検証器つき機械学習モデルのコンセプトの提案および、理論的な基盤の構築を成し遂げることができたと考えている。今後の展望として、効率的な検証器の実装に取り組む予定である。仕様を満たす出力 y を見つける問題は NP 困難な制約充足問題となることもあるため、効率的な学習・推論を実現する検証器の実装は実用上重要である。二分決定グラフや確率回路などの、制約や論理式を扱う効率的なデータ構造やアルゴリズムを用いることで効率的な実装が可能であろうと考え、検討を進めている。

検証器は既存の機械学習モデルと組み合わせて利用することになるため、検証器の実課題への

活用を進めるためには、既存の機械学習フレームワークと接続して利用可能な検証器パッケージを実装し公開することが重要であると考える。さきがけの研究終了後2年以内を目処に検証器を実装したパッケージを公開し、機能を充実させることを通じて、検証器つき機械学習モデルの社会実装をすすめる予定である。

本研究プロジェクトを開始してからの三年間で AI 技術を取り巻く環境は大きく変化した。代表的な変化として、画像等の各種生成モデルや ChatGPT に代表される大規模言語モデルの台頭が挙げられるだろう。これら最新の AI 技術は各種タスクにおいて高い予測性能を発揮する一方で、既存のモデルと比較してパラメータ数が増加し、ブラックボックス化が進んでいる。また、生成モデルの可能な入出力をすべて調べ上げることは実質的に不可能であり、大規模言語モデルのハルシネーションなどの予測できない振る舞いが問題視されている。このような現状を鑑みると、ブラックボックスである機械学習モデルに検証器を接続することで、誤りがないことを保証しようとする本研究のコンセプトの重要性は、研究開始時以上に増していると実感している。このさきがけの研究で確立した検証器つき機械学習モデルを一つの分野として育てていきたい。

4. 自己評価

本研究を進めるにあたっては、「信頼される AI」というゴールを実現するためにはどのような技術が必要かをまず考え、現代の AI 技術が順当に進化し、予測精度がさらに向上したとしても解決しない課題こそが、信頼される AI を実現するための課題であると考えた。そして、機械学習モデルの出力が仕様を満たす保証がないことが、AI が信頼されない大きな要因であると考え、その課題を解決する手段として検証器つき機械学習モデルのコンセプトを立案した。検証器つき機械学習モデルが目的に合致すると考える理由は 1 ページに記載のとおりである。また、個人でできる範囲でインパクトを最大化するために、汎用的な理論の整備に注力した。

検証器つき機械学習という新規の概念に基づく研究であること、トップダウンな研究アプローチのため自身が必ずしも明るくない技術を使う必要があり、研究の遂行には苦労した面もあった。しかし検証器つき機械学習モデルという枠組みの提案と、その汎用的な理論的解析の研究成果を、機械学習分野の最重要会議の一つである NeurIPS および ICML で発表できたことで、コンセプトを打ち出すという目的は達成できたと考えている。研究の到達目標としていた検証器つき機械学習モデルの基礎の確立についても、理論的な部分について満足のいく成果を得られたと考えている。一方で、新しいコンセプトの研究がコミュニティに認識され、受け入れられるようになるためには、研究成果の厚みがまだ足りないと考えている。3 節の今後の展開で述べたように、研究開始から3年が経ち、大規模言語モデルなどの革新的な技術が出てきても検証器つきモデルの重要性は変わらないことから、研究の方向の正しさは確信している。今後は研究の方向性に賛同する仲間作りなども含めて、引き続き検証器つき機械学習の技術を育てていきたい。

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:3 件 (うち採録決定済み論文1件)

1. Masaaki Nishino, Kengo Nakamura, Norihito Yasuda, “Generalization Analysis on Learning with a Concurrent Verifier”, Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022), 2022.
検証器つき機械学習モデルを提案するとともに、検証器を機械学習モデルに接続したときにその汎化性能がどのように変化するかを汎化性能解析によって確認した。検証器を推論時のみに利用する場合に、汎化誤差が悪化する条件を示したほか、検証器を学習時に利用する場合には Rademacher 複雑性に基づく汎化誤差の上限が悪化しないことを示した。
2. Hikaru Tomonari, Masaaki Nishino, Akihiro Yamamoto, “Robustness Evaluation of Text Classification Models Using Mathematical Optimization and Its Application to Adversarial Training”, The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL/IJCNLP 2022 (Findings)), 2022.
自然言語処理のための機械学習モデルの、入力の微小な変化に対する頑健性を、数理最適化ソルバを用いて厳密に検証する方法を提案した。同時に、提案した検証手法を、微小変化に対して頑健なモデルを学習する Adversarial Training に適用し、性能の向上を確認した。
3.（採録決定済）Masaaki Nishino, Kengo Nakamura, Norihito Yasuda, “Understanding the Impact of Introducing Constraints at Inference Time on Generalization Error”, 41st International Conference on Machine Learning (ICML 2024), 2024.
検証器を利用した機械学習において、学習時には不明であった仕様を推論時のみに追加した際に、汎化誤差がどのように変化するのかを理論的に解析し、多クラス分類問題において、特定の性質を満たす誤差関数を利用すれば、仕様の追加の前後で相対的な予測誤差が保存されることを明らかにした。

(2) 特許出願

研究期間全出願件数： 0 件（特許公開前のものも含む）

(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）

解説記事 1 件：西野、「Sentential Decision Diagrams と機械学習」、人工知能学会誌 2021 年 7 月号、2021 年

学会招待講演 1 件、「検証器つき機械学習モデル」、人工知能学会 第 125 回人工知能基本問題研究会、2023 年 8 月 30 日