信頼される AI の基盤技術
2020 年度採択研究者

2021 年度
年次報告書

西野 正彬

日本電信電話(株) NTT コミュニケーション科学基礎研究所
特別研究員

誤りがないことを保証する検証器つき機械学習の研究

§1. 研究成果の概要

本研究課題では、統計的機械学習において機械学習モデルが誤った推測結果を出力する問題に対処する方法として、検証器つき機械学習モデルという新たな数理モデルの実現を目指す。機械学習モデルは確率統計に基づいているため、機械学習モデルの予測誤りを完全になくすことは本質的に困難である。そのため、機械学習モデルを用いたシステム(AI システム)は誤りに対処するためのコストがかかり、AI 技術が広く社会で活用される妨げとなっている。

機械学習モデルの誤りを完全になくすことは本質的に困難である一方、ある種類の誤りについては、機械学習モデルの出力が誤りであるかどうかを外部から検証することが可能である。外部から検証可能な誤りが存在しないことを保証するための技術の実現を目指す方法として、検証器つき機械学習モデルを考える。このモデルでは機械学習モデルの入出力が仕様に沿ったものであるかを、外部に接続されている検証器が都度検証することで、モデルの出力にある種の誤りがないことを保証でき、AI システム構築のコスト削減に寄与できると考える。

2021 度は、検証器つき機械学習モデルの基本的な性質を明らかにすることを目的として、検証器を用いることによって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について解析を行った。機械学習モデルの推論時にのみ検証器を用いる設定(推論時の検証)と、学習時と推論時の両方で検証器を用いる設定(学習・推論時の検証)の 2 通りの設定を検討し、推論時の検証において汎化誤差が小さくなる条件を明らかにした。また、学習・推論時の検証においては検証器と機械学習モデルの組合せによらず汎化誤差の条件が変化しないことを明らかにした。

また、機械学習モデルの性質検証の一つの具体例として、自然言語処理分野で用いられる文書分類モデルが外乱に対して頑健であるかどうかを、数理最適化ソルバを用いて検証する手法を考案し、その有用性を示した。

【代表的な原著論文情報】

1) 友成光、西野正彬、山本章博、“数理最適ソルバを用いたテキスト判別モデルの検証”、2021 年度 人工知能学会全国大会
2) 友成光、西野正彬、山本章博、“数理最適ソルバを用いたテキスト判別モデルに対する敵対的サンプルの生成”、NLP 若手の会 第 16 回シンポジウム
3) 西野正彬、 “Sentential Decision Diagram と機械学習”、人工知能学会誌 Vol. 36, No. 4, 2021