研究課題別事後評価結果

1. 研究課題名： 誤りがないことを保証する検証器つき機械学習の研究

2. 個人研究者名

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）

3. 事後評価結果

従来の機械学習アルゴリズムは、本質的に統計的な仕組みを内包するため、予測の出力仕様の性能保証が困難であった。本研究は、自然言語による文章生成システムや、医療や自動操縦システムなど、実際の多くの AI システム応用において、検証可能な誤りをデータベースとして準備可能である点に着目した。そして、検証器つき機械学習モデルという新しいアイデアに基づき、動作保証付き AI システムの実現方式を開発するという斬新なアイデアを提示し、統計学習理論に基づく解析により、この提案の有効性を示している。成果として、外付けの検証器付き機械学習システムにおいて、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や、元のモデルが一定の予測性能をもつ場合には、推論時にのみ検証器を付加しても性能が悪くならないことなどの理論的性能保証を与えている。本研究の提案の適用対象としては、最近、注目されている自然言語による文章生成システムにおける事実との整合性保証や、医療や自動操縦での危険行動回避による安全性保証などが該当する。このような一見すると複雑な AI タスクにおいて、基本的かつ理論的性能保証をもつ方式を提案し、その有用性を示したことは、高く評価できる。さらに、仕様がわかっている場合に積極的に検証器を利用することで、予測の汎化誤差の改善に繋がれるという興味深い結果も得られている。また自然言語処理システムにおける機械学習アルゴリズムの信頼性検証のための手法として、数理計画法ソルバーを用いて機械学習モデルの訓練データの微小変動に対する頑健性を検証する手法を研究開発している。

本さがけ研究において中心となる検証器つき機械学習アルゴリズムの性能解析に関する成果は、機械学習分野における最難関会議に採択・発表されている。また関連する機械学習の信頼性向上に関する研究について、国内招待講演と学術誌の解説記事を発表するなど、その研究活動は高く評価されている。今後、AI 技術の社会への受容が進むとともに、動作保証付きの AI システムへの要求はより大きくなると予想される。着実な基礎研究を継続するとともに、さまざまな社会応用分野での実証を進めることで、本さがけ研究で得られた成果が、将来の信頼される AI システム構築の基盤技術として大きく発展することを期待する。