研究課題別事後評価結果

１．研究課題名：　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

２．個人研究者名

原　聡（大阪大学産業科学研究所　准教授）

３．事後評価結果

AI の基盤技術である機械学習の信頼性向上を目指し、その主要な問題点の一つであるモデルのブラックボックス性の解決に向けて、「機械学習モデルとユーザ間のコミュニケーション」を可能にする革新的な研究の枠組みを提唱し、機械学習モデルの説明・修正技術の開発と、その実現可能性について研究した極めて独創的な研究である。これは、従来の機械学習技術の在り方を根本から変える革新的な提案であり、モデルの説明と修正という双方向コミュニケーションを、機械と人間の間で実現することで、ユーザにとって望ましい AI システムの性質（安定性、頑健性、説明可能性、透明性、公平性など）を備えた機械学習モデルの実現を目指している。これは、機械学習の信頼性の向上に対して、統一的な視点からの解決策を与える極めて重要な研究といえる。現在、アドホックに研究されている機械学習運用における各種の問題点の解決に新展開をもたらすものであり、将来の学術的・産業的なインパクトは極めて大きい。

具体的な成果として、次の 3 つのテーマについて研究した。(1)モデルの説明に関して、説明のための関連性指標を研究開発し、モデル損失関数のパラメータ勾配のコサイン類似度が優れた指標であることを示した。一方、モデルの修正に関しては、(2)再学習による修正のための学習安定化問題に取り組み、アルゴリズム安定性理論を援用した決定木モデル向けの新しい手法を開発した。(3)さらに、完全分散化の環境で効率よく計算可能なハイパーパラメータ最適化法を初めて開発した。これらの研究成果は、機械学習分野における世界最高峰国際会議に採択されており、複数の国内学会賞を受賞し、報道掲載されるなど、国内外で高く評価されている。上記の研究成果は、「機械学習モデルとユーザ間でのコミュニケーション」という新しい概念を提唱し、独創的なアイディアと理論展開により「説明性」と「修正」という双方向コミュニケーションを実現する技術である。今後の AI の信頼性の研究開発の上で、研究進行のプロセス自体も重要な成果である。上記の成果は、従来の技術課題や分野の機械学習問題に狭く囚われることなく、本質的な解決策の議論を行なった上で、深層学習、画像類似性、差分プライバシー、安定アルゴリズム、分散計算などの複数の分野・技術を融合したアプローチの研究開発により生み出されている。このことは、本領域が目指す分野横断的研究を具現し、若きさきがけ研究者の良き手本となっている。また、本さきがけ領域の研究者とともに、当該分野の国内研究会において「信頼される AI」をテーマに企画セッションを企画・開催し、産業界との連携も進めるなど、提案する枠組みと研究成果の発信も積極的に行なっている。

このように、信頼される AI の中核となる機械学習モデルの信頼性向上のための研究に関して、本さきがけ研究を通して目覚ましい飛躍を成し遂げており、高く評価できる。今後は、機械学習の信頼性に関する第一人者として、またこれからの機械学習・AI 技術をリードする世界的研究者として、さらなる活躍を期待する。