信頼される AI の基盤技術　　2022 年度
2020 年度採択研究代表者　　年次報告書

原 聡

大阪大学 産業科学研究所
准教授

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

研究成果の概要

研究項目 2「“修正”のための関連性指標および効果的なモデル修正方法の開発」に注力し研究を進めた。特に (i) モデルの再学習の際にモデルの変化度合いを抑える方法、および (ii) モデルを所望の方向に修正する方法、の二点についての検討を進めた。

成果 1. モデルの学習の安定化

「モデルの再学習」は学習の設定を少し変えてモデルを再度学習し直すことでより良いモデルを作成する方法である。しかし、モデルの学習プロセス自体が不安定な場合には、再学習によって全く異なったモデルが得られてしまうことも珍しくない。この場合、“修正前のモデルの性質”が保持されることは望めず好ましくない“修正”となってしまう。そこで、機械学習の基本的なモデルの一つである決定木および階層クラスタリングについてその学習プロセスの安定化を行った。従来の決定木学習および階層クラスタリングでは貪欲法によるデータの分割を再帰的に繰り返すことで木を構成していた。それに対し、提案法では分割の選択を適当な方法で確率化する。これにより、学習に用いるデータが微小に変化しても構築される木構造がほぼ変化しないこと、つまり学習プロセスの安定化を保証できることを理論的および実験的に示した。

成果 2. 分散学習におけるハイパーパラメータ最適化

モデルをユーザの所望する方向へ修正するためには、ユーザがモデルに介入する手段が必要である。このような介入手段の一例として、ユーザ所望の方向性を損失関数、そして“修正”作業を損失関数に対するハイパーパラメータ最適化として定式化することが考えられる。本研究では、特に複数のユーザが協同して行う分散学習問題における“修正”作業を考え、そのためのハイパーパラメータ最適化手法、特に分散環境下でハイパーパラメータの勾配を効率的に推定する手法を提案し、その理論的な性質を示した。

【代表的な原著論文情報】

1) Naoyuki Terashita, Satoshi Hara. Personalized Decentralized Bilevel Optimization over Stochastic and Directed Networks. arXiv:2210.02129, 2022.
2) Gabriel Laberge, Ulrich Aïvodji, Satoshi Hara, Mario Marchand, Foutse Khomh. Fooling SHAP with Stealthily Biased Sampling. arXiv:2205.15419, 2022.