信頼される AI の基盤技術　　2022 年度
2021 年度採択研究代表者　　年次報告書

ホーランド マシュー ジェームズ

大阪大学 産業科学研究所
助教

学習過程における価値観の多様化と性能保証の両立

研究成果の概要

1 年半にわたる研究活動を経て、初年度の初期成果を足掛かりに、独自の汎化指標を基軸に据えた機械学習アルゴリズムの開発と解析は大きく前進した。まず、本提案の汎化指標を従来のいわゆるリスク関数（特に OCE 系や DRO 系）と比較しながら、広く「不確実性の下での意思決定」という歴史的な背景を整理するために、調査論文を新たに執筆した [1)]。先行研究のサーベイはもちろん、従来の学習問題の定式化を大幅に緩和し、機械学習システム設計者の「価値観」をより柔軟に表現する枠組みの可能性と技術的な課題についても考察した。

これまでの本提案のコア部分にあたる汎化指標族を threshold risk（T-risk）と呼んでおり、簡単にいうと既存の指標との最大の違いは損失のバラツキを双方向的に捉えている点である。リスク関数の公理でいうと「単調性」を満たさないため目的関数としての凸性は犠牲になるが、その代わりに単調性のある汎化指標と違って「モデル候補の過剰な自信」を容易に予防することが可能である。単調性を外すだけなら期待値と分散の和でも良いが、弱い凸性を有する形式や高いロバスト性を有する形式など、バラツキを測るために柔軟な関数族を導入することで学習アルゴリズムの安定性、汎化能力、外れ値への感度、テストデータにおける不確実性など、さまざまな要素を統一的に調整する仕組みの原型となっている。

先述の「双方向性」は端的にいえば、非対称な損失の分布の向きの逆転に対する不変性があり、この特性を可視化した実験、既存リスクとの公理的な相違点、また確率勾配法に基づく学習法の理論保証および実験的検証をまとめた論文を執筆した。この論文はすでに AISTATS 2023 で発表している [2)]。特に「最適解の多様性」の可視化法をめぐって、汎化指標の調節によって、学習アルゴリズムを柔軟に誘導できることを明確に示す補足的なデモンストレーションとして、平易な解説とともに「offgen」というリポジトリを GitHub で作成し、複数の Jupyter ノートブックで初版を公開している [3)]。先述の取り組みの初期成果とともにこの中身を中心に、JSAI2022 と NEURO2022 で発表し、後者では優秀発表賞を受賞した。

これまでの成果を軸として、学習法のロバスト化、差分プライバシーの保証、深層学習の汎化性能向上と学習の安定化、公平性の保証など、さまざまな問題領域との接点を探っていくとともに、ユーザーとのインタラクションを想定した半自動的な汎化指標選択の方法を探求していく予定である。

【代表的な原著論文情報】

1) Learning criteria going beyond the usual risk. Matthew J. Holland and Kazuki Tanabe (submitted).
2) Matthew J. Holland Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023), PMLR 206:1586-1623, 2023.
3) offgen: A visual "explainer" for off-sample generalization metrics https://github.com/feedbackward/offgen