学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　金森 憲太朗

審査担当者　主　査　教　授　有村 博紀
副　査　教　授　堀山 貴史
副　査　教　授　吉岡 真治
副　査　教　授　中村 篤祥

学位論文題名

Studies on Explainable Machine Learning Based on Integer Linear Optimization
（整数計画法に基づく説明可能な機械学習）

深層学習に代表される機械学習技術の発展により, 機械学習アルゴリズムで学習された予測モデル (機械学習モデル) が, 医療や金融, 司法などといった実社会の意思決定に活用されつつある. このような社会的に重要な意思決定タスクでは, 機械学習モデルの予測に基づく意思決定結果が, 人間の生活に 大きな影響を与える可能性がある. したがって, 機械学習モデルの予測精度だけでなく, 公平性や, 説明可能性, 透明性などといった, 信頼性 (trustworthiness) を向上させるための技術の実現が, 社会的な課題となっている. とくに,2010 年代に入って, 信頼性を向上させるための鍵として, 機械学習モデルの予測根拠や判断基準を人間が理解できような説明可能な機械学習 (explainable machine learning) の研究が盛んになっている.

本論文では, このような機械学習モデルの予測の根拠や, 理由を人間が理解できような説明可能な機械学習を研究する. 従来から, 機械学習モデルによる予測の説明手法が提案されてきたが, これらには, 以下のようないくつかの実用上の課題が存在する. 一つ目の課題は, 説明の妥当性である. 例えば, CE によって提示される摂動ベクトルは, 所望の予測結果を得るためのアクションとして解釈されるが, 従来の CE 手法では, 特徴量間の相関関係や外れ値リスク, 因果効果などといった予測タスク特有の性質を十分に考慮できないため, ユーザが実際に実行できる実用的なアクションを提示できないことが多い. 二つ目の課題は, 説明の安定性である. 決定木に代表される解釈可能なモデルは, モデル自体が予測結果の説明を提示することができるが, データの変化や公平性の実用上の追加制約により, モデルが提示する説明が運用中に大きく変化するため, 信頼性の問題が生じることがある. そこで, 本論文では, 次の問題について研究し, 上記の問題を解決または低減できる説明手法を提案している.

まず第 4 章では, CE において, 元の予測タスクのデータ分布を考慮することで実現可能なアクションを提示する方法について議論している. 具体的には, データ分布の特性として, 特徴量間の相関関係と外れ値指標を考慮してアクションのコストを評価することで, ユーザにとって現実的なアクションを提示する新たな CE のフレームワーク (distribution-aware counterfactual explanation, DACE) を提案する. このために, 統計学分野で広く用いられている距離関数であるマハラノビス距離と外れ値指標である局所外れ値因子に基づく新たなコスト関数を導入する. さらに, 導入したコスト関数に対して, それを近似する代理関数を設計することで, MILO に基づく効率良い最適化方法を提案している.

次に第 5 章では, アクションの実現可能性を向上させるための補助情報として, アクションの適切な実行順序を提示する方法について議論している. 従来の CE 手法は, アクションとして特徴量の変更方法を示す摂動ベクトルのみを提示する. しかし, 特徴量間には因果効果などの相互作用が存在するため, アクションの実行コストは実際には各特徴量の変更順序にも依存する. そこで, アクションとその特徴量の適切な変更順序を同時に提示する新たな CE のフレームワーク (ordered counterfactual explanation, OrdCE) を提案している. このために, 事前に推定した因果 DAG に基づいて特徴量の変更順序を評価するコスト関数を新たに導入し, アクションとその変更順序を同時に最適化する問題を定式化する. 定式化した最適化問題に対して, MILO に基づく解法を提案している.

最後に第 6 章では, 決定木に対して, その予測と解釈を維持したまま所与の制約を満たすように修正する新たな学習問題について議論している. とくに, 機械学習分野において近年重要視されている性質である公平性に着目し, 学習済みの決定木の予測と解釈を維持しつつ公平性制約が満たされるように編集するフレームワーク (fairness-aware decision tree editing, FADE) を提案する. このために, 予測多重性と根付き順序木の編集距離に基づく決定木間の非類似度指標をそれぞれ導入し, それらの指標を公平性制約のもとで最小化する最適化問題として決定木編集問題を定式化し, さらに, 定式化した最適化問題に対して, MILO に基づく解法を提案している.

これを要するに, 著者は, 人工知能分野における機械学習の説明可能性の問題において, 各種の拡張された予測結果の説明問題について, 新しい定式化を与え, それらの効率良い最適化手法を開発し, その有効性を理論的・実験的に示したものであり, 機械学習の説明可能性における情報科学において貢献するところ大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.