学　位　論　文　内　容　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

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, 公平性や説明責任, 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需要が高まっている. 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可能な形式で提示できる説明可能性 (explainability) の実現が特に重要視されており, 説明可能な機械学習 (explainable machine learning) に関する研究が注目を集めている.

説明可能性を実現するための研究には, 大きく分けて, (1) モデルの予測結果から局所的な説明を抽出するアプローチと, (2) 大域的に解釈可能なモデルを学習するアプローチの二つが存在する. いずれのアプローチにおいても, 多くの手法がそのタスクを最適化問題として定式化することで説明可能性の実現を数理モデル化することを試みており, これまでに様々な説明手法とその最適化問題が提案されてきた. 本研究では, 各アプローチにおける主要手法である反実仮想説明法 (counterfactual explanation, CE) と決定木 (decision tree) に着目する. CE は, モデルから所望の予測結果を得るための摂動ベクトルを「アクション」として提示する説明手法であり, 決定木は, if-then-else 形式の予測ルールを二分木で表現した解釈可能なモデルである. これらの手法は, 人間がモデルの予測結果を理解し, そこからより良い知見を得るための助けとなることが期待されるため, 理論と応用の両面で近年活発に研究が行われている.

しかし, 現状の説明手法には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. これらの課題を解決するためには, 説明が満たすべき要求を実用上の課題から明示的に特定し, その要求を達成するための最適化問題を適切に定式化することが不可欠だが, 具体的に, (i) それぞれの説明手法で提示される説明がどのような要求を満たすべきかと, (ii) その要求を適切な目的関数や制約式として数学的に表現する方法, および (iii) 定式化された最適化問題に対する柔軟かつ効率良い解法の設計方法は, まだ十分に解明されておらず, 説明可能な機械学習における重要な課題となっている.

本研究の目的は, 実用的な説明を提示する説明可能な機械学習の新しいフレームワークの設計方

法を明らかにすることである. そのために, CE や決定木の課題について実用性の観点から議論を行い, 説明が満たすべきいくつかの要求を特定する. 特定した要求を数学的に表現するための適切な目的関数と制約式を導入することで, CE と決定木の実用性を向上させるための新たな最適化問題を定式化する. さらに, 定式化した最適化問題に対して, 混合整数線形計画法 (mixed-integer linear optimization, MILO) に基づく柔軟かつ効率良い解法を提案する.

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する. 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する.

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する. 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) を提案する. 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する.

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する. 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する. 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する.

本研究を通して, 機械学習の説明可能性を実現するための新しい最適化問題について考察し, 混合整数線形計画法に基づく解法を開発してきた. 説明可能性は, 機械学習技術の実社会意思決定タスクへの応用において必要不可欠な要素であり, その実現には, 最適化問題としての適切な数理モデル化とそれらに対する柔軟かつ効率良い解法の設計が鍵となる. 本研究の成果は, 説明可能性の実現するための新たな方法を提案するものであり, 産業・学術の分野を問わず, 機械学習技術の実社会応用における多様な活動の推進に貢献するものである.