利用者名：コンピュータサイエンス学部　教授　中西 崇文

Scopus

Title: Bayesian-AIME: Quantifying Uncertainty and Enhancing Stability in Approximate Inverse Model Explanations （Bayesian-AIME：近似逆モデル説明における不確実性の定量化と安定性の向上）

Authors: Takafumi Nakanishi
（中西崇文(東京工科大　コンピュータサイエンス学部　教授)

Journal: IEEE Access, Vol. 13, 2025
掲載年月：2025 年 10 月

研究概要：説明可能 AI（XAI）において、説明の信頼性（不確実性）の定量化と、学習データの微小な変化に対する説明の安定性は重要な課題です。本研究では、既存の手法である AIME（Approximate Inverse Model Explanations）をベイズ推定の枠組みで拡張した「Bayesian-AIME」を提案しました 。ブラックボックスモデルの逆演算子を確率変数としてモデル化することで、特徴量重要度を事後分布として推定することに成功しました 。これにより、特徴量重要度に 95%信用区間（CI）を付与して信頼性を定量化できるとともに、事後平均を用いることで説明の安定性を向上させました 。

研究背景：医療や金融などの高リスク領域で AI が利用される中、その判断根拠を説明する XAI が不可欠となっています 。しかし、LIME や SHAP などの主流な手法は、信頼性を示さない単一値（点推定）の説明しか提供せず、データのわずかな変化で説明が大きく変わる「不安定性」の問題がありました 。ユーザーが安心して AI を利用するためには、説明がどの程度確実かを示す「不確実性の定量化（UQ）」と、再現性のある「安定した説明」が求められていました 。

研究成果：3 つの実験を通じて、以下の成果を確認しました。

1. 説明の安定性の向上：Iris データセットを用いたブートストラップ検証において、Bayesian-AIME は SHAP や LIME と比較して、順位相関係数が高く（0.969 vs 0.901）、標準偏差が低い（0.0110 vs 0.0261/0.0271）結果となり、最も安定した説明を提供できることが実証されました 。
2. 不確実性の適切な定量化：正解が既知の合成データを用いた検証により、提案手法による 95%信用区間が、真に重要な特徴量を正しく捉えつつ、無関係なノイズ特徴量に対してはゼロを含む区間を示すことで、「重要とは断定できない」という不確実性を適切に表現できることが確認されました 。
3. 実データでの有用性：タイタニック号の生存予測や乳がん診断のケーススタディにおいて、点推定だけでは見落とされがちな「判断の不確実さ（例：年齢は生存の決定的な要因ではない）」を可視化し、より信頼性の高い意思決定支援が可能であることを示しました 。

社会への影響：本手法により、AI の説明に対する「自信の度合い」を提示することが可能となります 。これにより、医師や専門家が AI の判断根拠を鵜呑みにせず、説明の信頼性が低い（信用区間が広い）場合には慎重な判断を行うなど、人間と AI のより適切な協調が可能となり、高リスク領域での AI 導入を促進することが期待されます 。

専門用語：

Bayesian-AIME ：本研究で提案する手法。AIME の逆演算子推定にベイズ線形回帰を導入し、特徴量重要度の事後分布を算出することで、説明の不確実性評価と安定化を実現したフレームワーク 。

信用区間 (Credible Interval, CI) ：ベイズ統計において、パラメータが特定の確率（本研究では 95%）で含まれる範囲。特徴量が予測にどの程度確実に寄与しているかの指標として利用される 。

近似逆モデル説明 (AIME) ：ブラックボックスモデルの入出力関係全体を線形な逆演算子として近似し、大域的および局所的な特徴量重要度を統一的に導出する XAI 手法 。

説明可能 AI (XAI) ：ブラックボックスになりがちな AI（特に深層学習など）の判断プロセスや根拠を、人間が理解できる形で提示する技術の総称 。

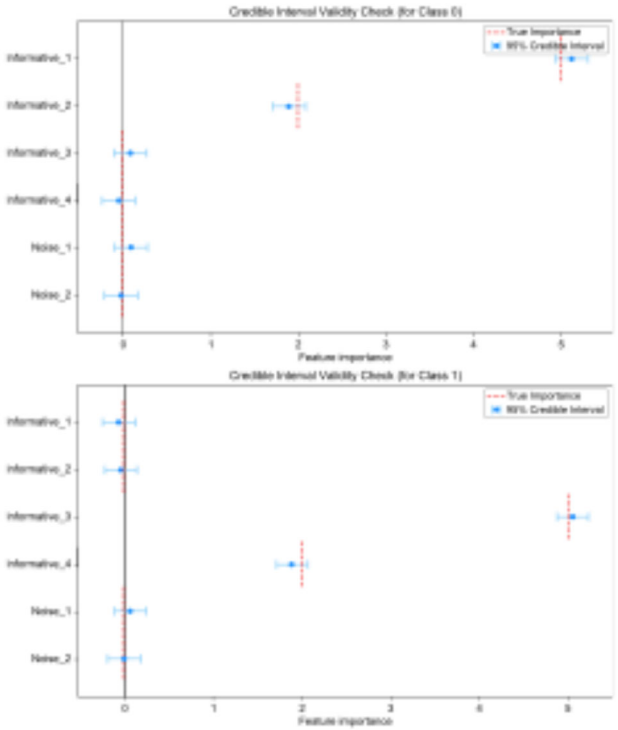

図 1　Bayesian-AIME による重要度推定の比較（合成データ）。青い点は推定された重要度の平均値、水色の線は 95%信用区間、赤い破線は真の重要度を示す。ノイズ特徴量（Noise_1，Noise_2）の信用区間がゼロを跨いでおり、不確実性が適切に表現されている。

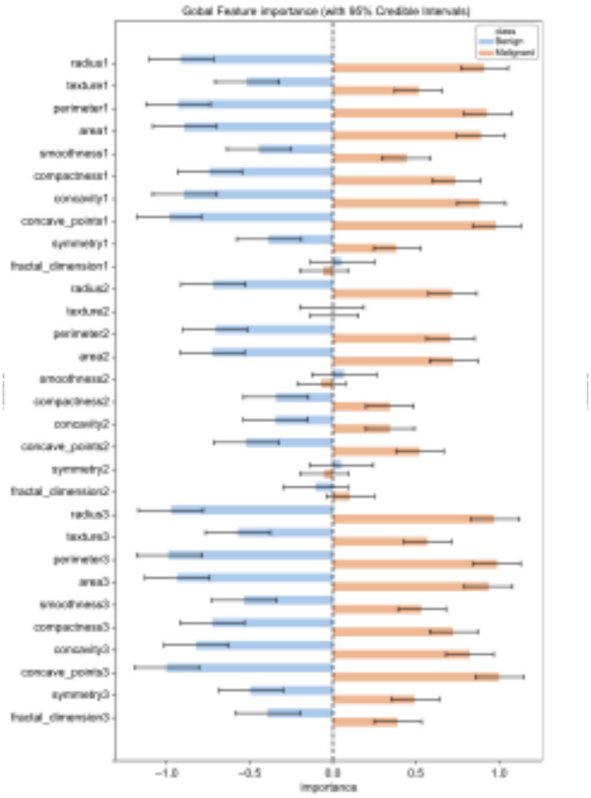

図 2　乳がんデータセットにおける大域的特徴量重要度と 95%信用区間。悪性（Malignant）のクラスに対して、凹みや面積などの特徴量が強い正の寄与を示し、その信用区間がゼロを含まないことから、これらが信頼できる診断根拠であることを示している 。