「新しい社会システムデザインに向けた情報基盤技術の創出」
2017 年度採択研究者

2018 年度
実績報告書

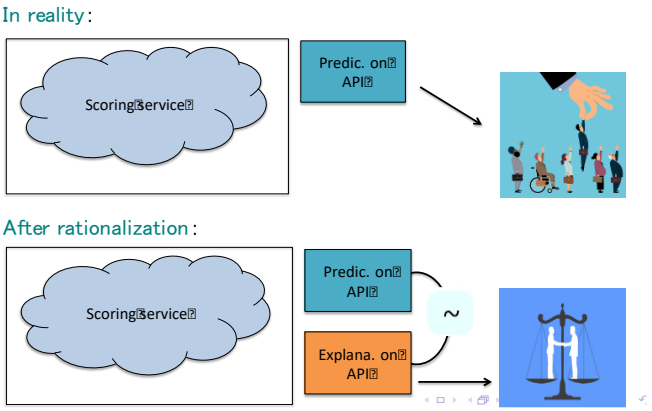
荒井　ひろみ

理化学研究所革新知能統合研究センター
研究員

安全かつ透明な個別化のためのプライバシ保護データマイニング

§1. 研究成果の概要

パーソナルデータを用いた情報提供サービスにおけるバイアスを説明する方法およびその説明方法の脆弱性について開発，分析を行った．大勢のパーソナルデータを集めたデータベースから学習をする際，ときに学習モデルは人間の理解が及ばないレベルの複雑さになる．それを人間に理解可能な形で説明する主要な方法の一つに，モデルの予測プロセスについて単純化や近似を行う方法がある．一方，近年機械学習モデルに公正性などのバイアスがあることが問題視されてきている．例えば人種や性別の違いが学習モデルの判断に影響し差別的な扱いをするなどである．申請者は，モデルを説明のために情報を落とし単純化する際に，上辺だけの公正性に配慮していると装う(Fairwashing)ことが可能であると考え，そのリスクについて評価方法を設計し検証を行った．具体的には学習した複雑な学習モデルを単純なルールリストで近似して説明を行う場合に，説明において人種や性別の情報を用いずに公正なプロセスで予測を行っていると装うこと(rationalization)が可能であることを示した．



図：Fairwashing の仕組み

また，パーソナルデータ利用のユーザー受容性の検証のためプライバシー意識についてのアンケート調査を実施し，パーソナルデータ提供における抵抗感や提供先機関のサービスの有用性とのトレードオフがある傾向を確認した．

§2. 研究実施体制

①研究者:荒井ひろみ （理化学研究所革新知能統合研究センター 研究員）
②研究項目

- Fairwashing についての評価方法設計，実験
- パーソナルデータ提供におけるプライバシー意識のアンケート調査の設計，ウェブ上のアンケートシステムの構築，クラウドソーシングサービスを用いた調査実施