Keyword プライバシー保護　PPDP　データ拡張

プライバシー保護データ解析における匿名化処理およびデータ拡張

情報科学研究科 情報科学専攻 博士前期課程
2022年3月修了

趙文峰

主査 成凱　**副査** 安部恵介　石田健一

研究背景

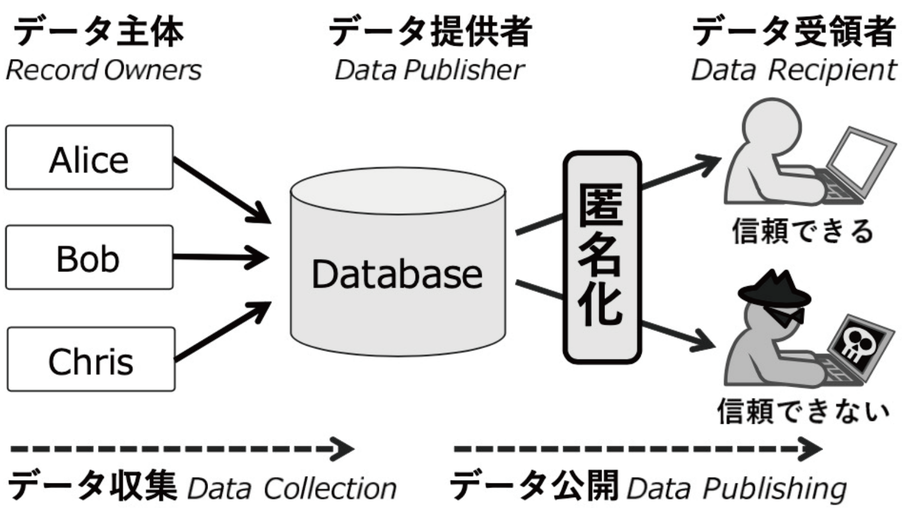
近年のITの進歩に伴い、多くの企業や公的機関ではパーソナルデータを蓄積し、機械学習やデータマイニングなどの手法でデータ利活用を行なっている。一つの機関のデータでは分析目的が達成できない場合があり、複数の機関間が保持するデータ流通させ、サービス開発や事業の最適化などに活用するニーズが高まっている。その中、個人に関するプライバシー保護の上、個人に関するデータの有用性を高める必要がある。

研究目的

本研究では匿名化技術によりデータの安全性を保つとともに、データ拡張に基づいて、データ拡張とプライバシー保護の匿名化技術を組み合わせることによってデータの有用性を高めることを目指すことを研究目的とする。データ拡張で作られたデータは偽データとして攻撃される心配がなく、たとえ攻撃されても、合成したデータには個人情報を含まれていないため、プライバシーなどの侵害が一切なくて、共有も容易になる。

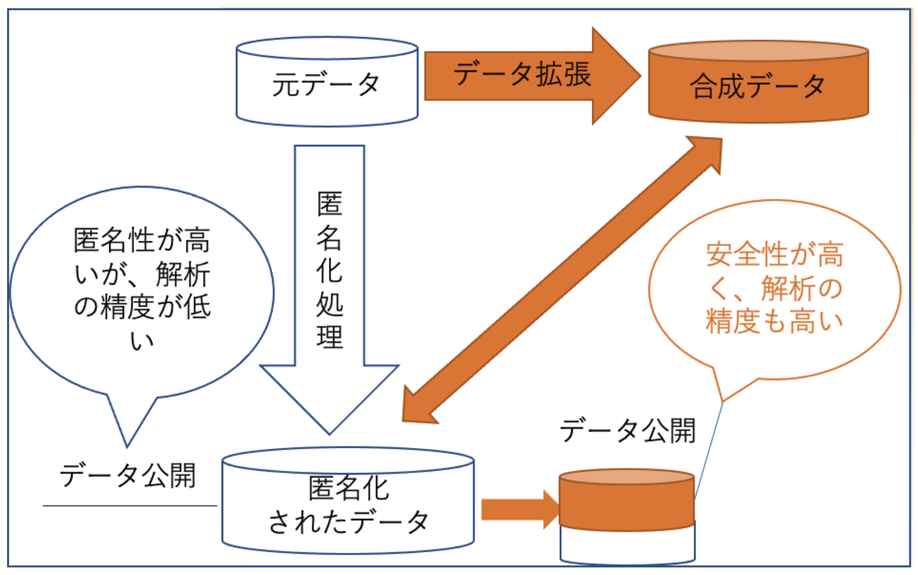
研究概要

PPDP：プライバシー保護データパブリッシング



参考文献：南 和宏, プライバシー保護データパブリッシング, 情報処理 Vol. 54 ,No. 9 ,Sep 2013 pp.938-946

本研究全体の流れ



データの匿名化加工

名前	年齢	性別	入院年月日	病名
鈴木	52	女	1/1/2020	肝炎
木村	52	男	2/1/2020	ねんざ
高橋	51	女	3/1/2020	エイズ
田中	46	男	9/9/2019	インフルエンザ
上田	48	女	8/9/2019	エイズ
岡本	61	男	7/9/2019	エイズ

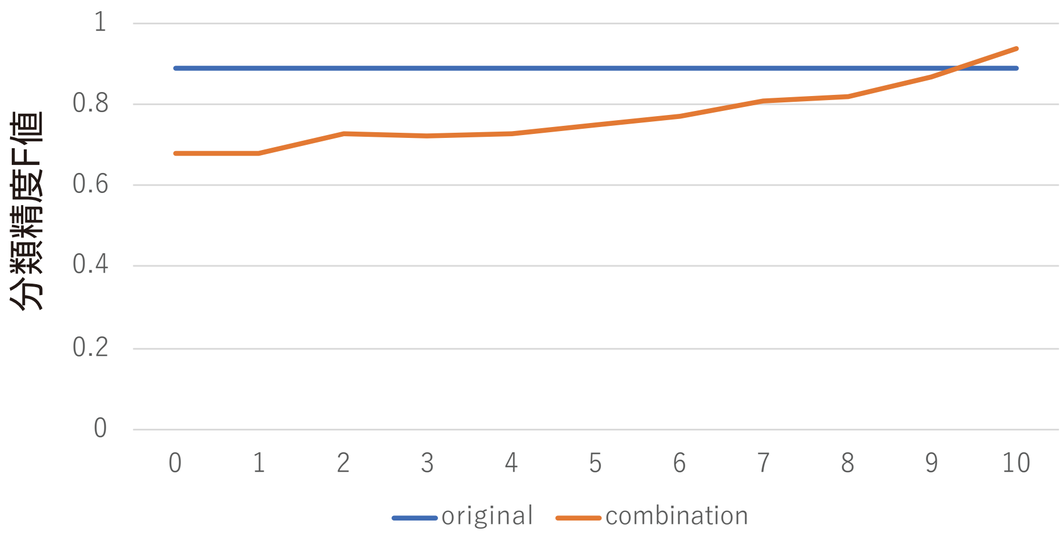
匿名化

削除　一般化　一般化

名前	年代	性別	入院年月	病名
鈴木	50	女	1/2020	肝炎
木村	50	男	1/2020	ねんざ
高橋	50	女	1/2020	エイズ
田中	40	男	9/2019	インフルエンザ
上田	40	女	9/2019	エイズ
岡本	60	男	9/2019	エイズ

評価実験

合成データと元データの割合を変えながら、組み合わせたデータセットでNavieBayes学習モデルの分類精度を評価した。



結果として、合成データを使うことで匿名性を高めながら、一定の分析精度が得られることがわかった。

成果・まとめ

本研究では、データ拡張とプライバシー保護の匿名化技術と組み合わせることによって、プライバシー保護と解析精度を両立させるため、評価実験を行った。合成データに1割ずつの元データを入れると、学習したモデルの精度が上がっていくことがわかった。元データと合成データの組み合わせデータの質が良いならば、元データのかわりに使用することができると考えられる。

指導教員コメント

本研究は、プライバシー保護データ解析における匿名化加工技術と、AI分野で高品質な教師データをアルゴリズムで合成するデータ拡張技術を融合させ、プライバシー保護とデータの有用性を両立させることを目的とし、学修データの匿名化加工とデータ拡張で得られた合成データを組み合わせて評価実験を行った。データ拡張の有用性を検証できた。

成凱