

2 値行列データに対する 次元拡張を用いた κ^m -匿名化手法

藤岡研究室 201903692 小林雅弥

1 研究背景

購買履歴データや位置データなどの個人情報を含んだビッグデータの需要は高まっている。ビッグデータ活用には、プライバシー保護のために匿名化が必要不可欠である。 κ -匿名性は『データベースにおいて疑似 ID の組み合わせが同一の人は少なくとも κ 人存在する』ということを保証するプライバシー保護指標であり、 κ 人未満に個人を絞れないという直感的に分かりやすい指標のため、広く用いられている。

本研究では、購買履歴データなどのトランザクションデータにおける κ -匿名化を考える。購買履歴データの場合、誰が何を買ったかが示されており、購入対象商品そのものを疑似 ID として考えられる。そのため、商品種類数分の次元数のデータが必要であり、非常に高次元なデータとなる。

しかし Charu の研究によって高次元データでは有用性の高い匿名化データを作ることが困難であることが示されており、これは次元の呪いと呼ばれている。

次元の呪いを回避するため、Manolis らはトランザクションデータに対し、 κ -匿名性の仮定を緩めた κ^m -匿名性を提案した。これは、攻撃者がもつ背景知識を高々 m 個のアイテムまでと制限した場合に、少なくとも κ 個以上同じレコードが存在することを保証するプライバシー保護指標である。その後、Giorgos らは κ^m -匿名性を使用して位置情報軌跡データの匿名化手法を提案した。しかし匿名化処理で一般化を用いているため、Manolis らの手法では限定的なデータベースでしか、Giorgos らの手法では多値データでしかそれぞれ匿名化を行なうことができなかった。

長谷川らは五十嵐らが提案した $P\kappa$ -匿名性を使用して、多値データ・2 値行列データ共に一部有用性を満たす匿名化データの生成手法を提案したが、 $P\kappa$ -匿名性はデータの持ち主を $1/\kappa$ 以上の確信度に絞り込めない指標であるため、少なくとも κ 個以上同じレコードが存在することを保証する κ^m -匿名性を満たすことができなかった。

2 研究計画

本研究では、高次元な 2 値行列データに対する κ^m -匿名化手法を提案し、提案手法の有用性を検証する。

Manolis らや Giorgos らの提案した κ^m -匿名化手法では、一般化を用いているため 2 値行列データにおける κ^m -匿名性が満たせない。そのため、2 値行列データにおいて κ^m -匿名性を満たすための匿名化手法を確立させる事が本研究の目標である。はじめに 2 値行列データでは一般化を用いた匿名化手法が適用できないことを示し、一般化以外の匿名化処理で 2 値行列データに使用できるか確認する。次に 2 値行列データに使用できる κ^m -匿名化手法を提案する。これは 2 値行列データを購入者が多い順に列の置換を行った後、ある基準で κ^m -匿名性を満たすよう検査・ノイズ付与を繰り返し行うことで匿名化データを生成する手法を想定している。

その後、提案手法が実際の 2 値行列データに使用できるか確認する。低次元データから確認していくことで提案手法がどの次元まで使用することができると判断することが必須である。これは、Charu の研究により高次元データでは有用性の高い κ -匿名化データを作ることが困難であることが示されており、25 次元から 35 次元において 2-匿名性を達成する確率が急激に低下するため、提案手法で最低でも 35 列の 2 値行列データで 2^m -匿名性を満たす手法であれば、提案手法に有用性があることが分かる。

最後に、実際に使用されるようなビッグデータを使用して提案手法が有用性を満たすか確認する。人工的に作成した 1000×1000 の 2 値行列データを使用して数値実験を行い、情報損失の値を検証する。その結果を踏まえて有用性に関する考察を行う。

3 現時点の成果

2 値行列データでは一般化階層を作成できないことを検証した。また、一般化以外の匿名化処理を使用して 2 値行列データが匿名化できるか検証した。今後は 2 値行列データに対する κ^m -匿名化手法を提案したのち、有用性を満たすか検証していく。

参考文献

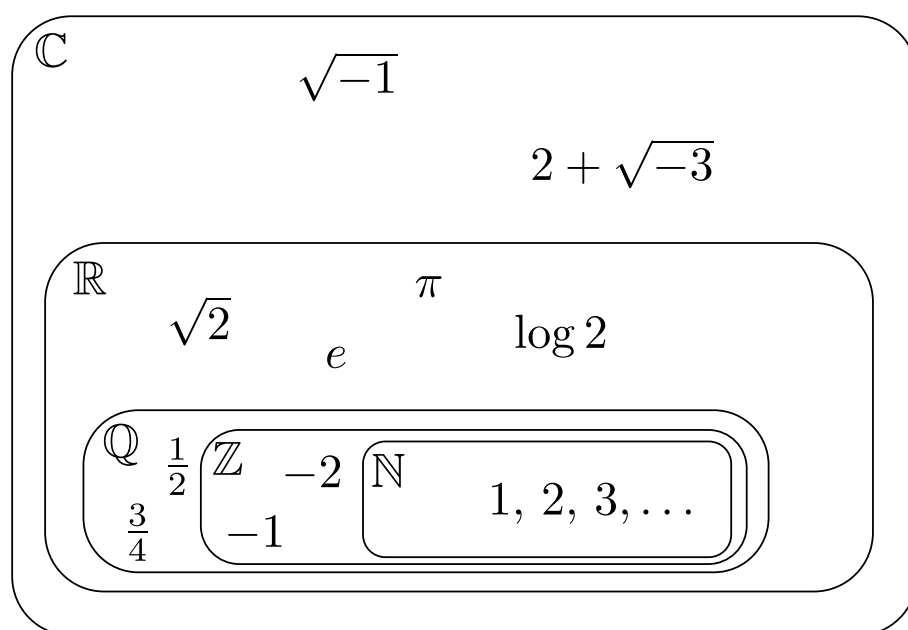
- [1] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. *Proc. VLDB Endow.*, 1(1):115–125, 2008.

- $X \cup Y := \{x \mid x \in X \text{ または } x \in Y\}$ を集合 X と集合 Y の**和集合**または**合併集合**という.
- $X \cap Y := \{x \mid x \in X \text{ かつ } x \in Y\}$ を集合 X と集合 Y の**積集合** または**共通部分**という.
- $X \setminus Y = X - Y := \{x \mid x \in X \text{ かつ } x \notin Y\}$ を集合 X と集合 Y の**差集合**という.

有名な集合の記号

- \mathbb{N} : 自然数全体の集合 ($0 \in \mathbb{N}$ とする流儀もあり)
- \mathbb{Z} : 整数全体の集合
- \mathbb{Q} : 有理数全体の集合
- \mathbb{R} : 実数全体の集合
- \mathbb{C} : 複素数全体の集合

ただし, 書籍等では $\mathbf{N}, \mathbf{Z}, \mathbf{Q}, \mathbf{R}, \mathbf{C}$ などのボールド体が使われることが多い.



使用例

- $1 \in \mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$
- $-1 \in \mathbb{Z} \setminus \mathbb{N}$, $\sqrt{2} \in \mathbb{R} \setminus \mathbb{Q}$, $\sqrt{-1} \in \mathbb{C} \setminus \mathbb{R}$
- $\mathbb{R} \setminus \mathbb{Q}$: 無理数全体の集合