4ZD-01

混合型データセットに対する k^m–匿名化手法

金澤 理乃 † 小林 雅弥 ‡ 藤岡 淳 § 千田 浩司 ¶
神奈川大学 † 神奈川大学大学院 ‡ 神奈川大学 § 群馬大学 ¶
永井 彰 ‖ 安田 幹 **
NTT 社会情報研究所 ‖ NTT 社会情報研究所 **

1 はじめに

人工知能やデータマイニングの研究・開発が進むにつれ，個人情報を含んだビッグデータの需要は高まってきている．しかしこのようなデータはプライバシを侵害する可能性を含み，ビッグデータの活用には k–匿名化のような個人の特定を防ぐための処理が必要不可欠である．本研究では，ヘルスケアデータなどの高次元トランザクションデータに対する匿名化を考える．Aggarwal らの研究により，高次元データに対して k–匿名化手法の使用は困難であると示されたため，Terrovitis らは k–匿名性の制約を強めた k^m–匿名性を提案し，一般化のみを用いた手法を構築した．Poulis らは軌跡データの k^m–匿名化手法を提案した [2] が，提案された手法は 2 値行列データにおいて使用できなかった．また，小林らは，2 値行列データにおいて一般化を用いない手法を提案した [1] が，多値データが混在した場合には使用できないことが問題点として挙げられる．そのため本稿では，2 値と多値の両方が混在する行列データ (混合型データ) に対する k^m–匿名化手法を提案し，その有用性を検証する．

k^m-anonymization method for mixed-type data sets

† Ayano Kanazawa, Kanagawa University
‡ Masaya Kobayashi, Kanagawa University
§ Atsushi Fujioka, Kanagawa University
¶ Koji Chida, Gunma University
‖ Akira Nagai, NTT Social Informatics Laboratories
** Kan Yasuda, NTT Social Informatics Laboratories

2 準備

2.1 k–匿名性

k–匿名性とは Sweeney が考案したプライバシ保護指標である．これは，個人情報からなるデータベースにおいて，同じ準識別子の組を持つデータ主体が k 人以上いる状態を意味する．

2.2 k^m–匿名性

k^m–匿名性とは Terrovitis らが提案した匿名性指標であり，これは攻撃者が持つ背景知識が高々 m 個のアイテムまでと制限した場合に，少なくともそれら m 個のアイテムが同一となるレコードが k 個以上存在することを保証するプライバシ保護指標である．

3 提案手法

本節では，混合型データに対する k^m–匿名化手法を提案する．以下は提案手法となる匿名化のブロック図を示す (図 1)．

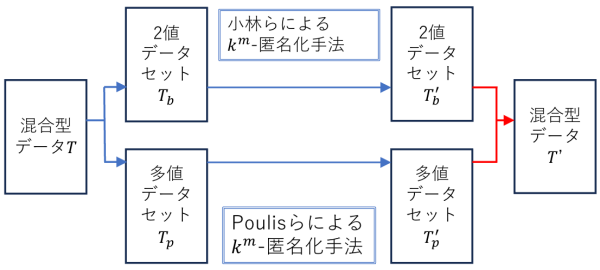

図 1 提案手法のブロック図

まず，混合型データを 2 値データと多値データに分割し，2 値データは小林らによる 2 値行列に対する k^m–匿名化手法を用い，多値データは Poulis らによる一般化を用いた k^m–匿名化

Algorithm 1 結合処理

Input: 2 値データセット $\mathcal{T}_b$，多値データセット $\mathcal{T}_p$, 匿名化パラメータ k，アイテム数 m

Output: k^m–匿名性を満たした混合型データセット $\mathcal{T}'$

1: $A = \mathcal{T}_b$の列数，$V = \mathcal{T}_b$の列番号の組からなる部分集合 ($|V| = {}_AC_{m-1}$)
2: $\mathcal{T}' = \mathcal{T}_b$
3: **for** i **in** $\mathcal{T}_p$ の列数 **do**
4: 　$\mathcal{T}'$ の最終列に $\mathcal{T}_p$ の i 列目を追加
5: 　**for** v **in** V **do**
6: 　　$\mathcal{T}'$ の v 列目と $A+i$ 列目で構成されるデータ τ を作成
7: 　　**if** τ が k^m–匿名性を満たさない **then**
8: 　　　τ が k^m–匿名性を満たすよう $\mathcal{T}'$ の $A+i$ 列目を一般化
9: **return** $\mathcal{T}'$

手法を用いて，匿名化を行う．その後，今回，新たに提案する k^m–匿名性を満たすための結合処理を行う (Algorithm 1)．

4 実験

人工的に作成した 1000×30 の混合型データで数値実験を行い，情報損失の量を検証する．

4.1 実験設定

混合型データを 20 列が 2 値データ，10 列が多値 (7 値) データで構成し匿名化パラメータ k は 10，アイテム数 m は 2 とする．また，評価指標としては匿名化前後においてどの程度値が変更されたかを測るために情報損失として以下の式を用いる．このとき，N をデータ行数，M をデータ列数，ij をそれぞれ行番号，列番号，n を多値数，$\#X$ を値の要素数とする．

$$loss := \frac{1}{NM} \sum_{ij} \begin{cases} |X_{ij} - X'_{ij}| & \text{2 値データ} \\ \dfrac{\#X'_{ij} - \#X_{ij}}{n} & \text{多値データ} \end{cases}$$

実験は 25 回ずつ行ない，その平均値 ($loss\_ave$)，最大値 ($loss\_max$)，最小値 ($loss\_min$) を求める．

4.2 実験結果

実験結果を表 1 に示す．ここで，$\mathcal{T}'_b$，$\mathcal{T}'_p$，$\mathcal{T}'$ における値はそれぞれ M が，20，10，30 のときのものであることに注意されたい．

表 1 **情報損失量の評価**

	$\mathcal{T}'_b$	$\mathcal{T}'_p$	$\mathcal{T}'$
$loss\_ave$	0.00117	0.04713	0.01649
$loss\_max$	0.00125	0.04713	0.01654
$loss\_min$	0.00105	0.04713	0.01641

5 考察

混合型データを分割した後，それぞれに匿名化処理を行なう手法であるため，情報損失量が大きくなると思われていた．しかし，図 1 を見ると，全体の 1.7% 程度しか損失がなく，その最大，最小の幅が小さいこともわかる．また，結合処理では，匿名化データの多値部分に対してさらに一般化を用いているが，一般化は階層構造であるため，元データに対する匿名化処理として問題がないと言える．

6 まとめ

本稿では，混合型データに対する k^m–匿名化手法を提案した．加えて，数値実験により，少ない情報損失量で，匿名化が行えることを確かめた．今後の課題として，匿名化パラメータやデータ列数などの値を変更し, 本手法で今回使用したデータセットと異なる内容や形式のデータセットを用いても汎用的に使用出来るかどうか確認することが挙げられる．

参考文献

[1] M. Kobayashi, et al. Extended k^m–anonymity for randomization applied to binary data. In *PST2023*, pp. 221–227. IEEE, 2023.

[2] G. Poulis, et al. Apriori-based algorithms for k^m-anonymizing trajectory data. *TDP*, 7(2):165–194, 2014.