4ZD-02

ランダム化を用いたk^m–匿名化手法

岩城 早汰 † 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

深層学習やデータマイニングの研究・開発が進むにつれ，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きている．

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個人の特定を防ぐためには，k-匿名化のような処理が必要不可欠である．しかし，Aggarwal らの研究により，高次元データを k–匿名化することは困難であることが示されている (次元の呪い)．

次元の呪いを回避するため，Terrovitis らにより，k–匿名性の制約を強めたk^m–匿名性および匿名化手法が提案された．k^m–匿名性は，攻撃者の背景知識が最大 m 属性に限定されるとき，少なくともそれら m 属性の値が同一となるレコードが k 個存在することを保証するが，彼らの手法では 2 値行列データは扱えなかった．

一方，小林らは 2 値行列データに対する k^m–匿名化手法を提案したが，彼らの手法には一部データに対し匿名性が満たせないという問題点が存在している．

本稿では，小林らの手法に確率的処理を導入し，その有用性を検証する．

k^m–anonymization method using randomization.
† Sota Iwaki, 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，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，攻撃者の背景知識が最大 m 属性に限定される時，それら m 属性の値が同一となる k 個のレコードが存在することを保証するプライバシ保護指標である．

2.3 小林らの手法

ここでは，提案手法の元となる小林らの手法 [1] について紹介する．小林らの手法は，m 列の組からなるデータを抽出し，最終列 (1 桁目) のみ異なるレコードに対して，条件 (表 1) に応じてアイテムの書き換えを行っていた．この条件は，同一のレコードが k 個以上である (“+”)，k 個である (“=”)，k 個未満である (“–”)，0 個である (“0”) の 4 つで構成されており，アイテムの書き換えにより，どちらのレコードとも “+” もしくは “=” になるため，最終的に k^m–匿名性を満たすことができる．ただ，一部データセットに対し，書き換え条件として 12 型や 15 型が選ばれる，このような状態が起こってしまうと，k^m–匿名性を満たすことができない．

3 提案手法

今回提案する手法は小林らの手法で問題となる 12 型ないし 15 型が起こった際に，該当する

表1　アイテム書き換え条件

1桁目	0	1	1桁目	0	1
1型	+	+	9型	−	+
2型	+	=	10型	−	=
3型	+	−	11型	−	−
4型	+	0	12型	−	0
5型	=	+	13型	0	+
6型	=	=	14型	0	=
7型	=	−	15型	0	−
8型	=	0	16型	0	0

データに向けてランダムなノイズ付与を行うものとなる．ランダムなノイズ付与としては，維持-置換撹乱を用いる．

維持-置換撹乱とは，維持確率 ρ で属性値を維持し，$1-\rho$ の確率で属性値をランダムに変更することでデータを秘匿化する処理である．あるカテゴリ属性 A_j の属性値 $v \in A_j$ が $v' \in A_j$ に変わる確率 $P^{A_j}_{y|x}(v'|v)$ は維持確率 ρ_j により，以下のように表せる．

$$P^{A_j}_{y|x}(v'|v) = \begin{cases} \rho_j + \frac{1-\rho_j}{|A_j|} & (v' = v) \\ \frac{1-\rho_j}{|A_j|} & (v' \neq v) \end{cases}$$

4　実験

小林らの手法と提案手法の比較実験を行なう．実験としては，「匿名化手法の成功確率」と「成功時の有用性評価」の二つを比較していく．

4.1　実験設定

ランダムに作成した 50×30 の二値データを用い，行列の要素は 5.0% の確率で 1 を持つものとする．また，評価指標として匿名化前後においてどの程度誤差が生じるかを測るために，以下の式を情報損失量として用いる．ここで，N をデータ行数，M をデータ列数とし，i, j はそれぞれ行番号，列番号とする．

$$loss := \frac{1}{NM} \sum_{ij} |X_{ij} - X'_{ij}|$$

4.2　実験結果

実験を 50 回ずつ行ない，成功確率と情報損失量の平均 ($loss\_ave$) をそれぞれ求めた (表 2)．

表2　**比較実験**

	成功確率	$loss\_ave$
小林らの手法	0.78000	0.00249
提案手法	1.00000	0.00252

5　考察

提案手法において，12 型，15 型によるエラーがなくなったため，成功確率は 100.0 % となったが，加工量が増えたため，情報損失量の平均は小林らの手法に比べ僅かに高くなった．

小林らの手法でも匿名性を満たさない列を削除することで処理を継続できるが，提案手法では全ての場合において加工を行なうことができるため，次元数を維持したままデータの加工が可能となっているし，列の削除を行った場合，情報損失量はかなり増加すると予想される．

また，今回のようなノイズ処理による手法では，k^m–匿名性を満たすことは保証できないが，k^m–匿名性の確率的な拡張である Pk^m–匿名性 [1] であれば満足できる可能性がある．

6　まとめ

本稿では，小林らが提案した 2 値行列データに対する k^m–匿名化手法の問題点を確率的処理による改良した手法を提案した．数値実験の結果により，小林らの手法とほぼ同等の情報損失量で匿名化データを作成することができた．

今後の課題として，提案手法が Pk^m–匿名性を満たせるかを示すことが挙げられる．

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.