6W-04

重複データのある水平分割データベースのダミー値追加による匿名化

伊奈 優樹† 清 雄一‡ 田原 康之‡ 大須賀 昭彦‡

†電気通信大学情報理工学部総合情報学科 ‡電気通信大学大学院情報理工学研究科情報学専攻

1. はじめに

近年,情報化社会の進展に伴い，個人に関するデータ利用時のプライバシー保護が問題となっている．データ保有機関のもとの膨大なデータは，分析を通し有益な情報をもたらす一方で，個人に不利益な情報の暴露を発生させる危険がある．そのため,分析が可能な状態を保ちつつ，プライバシー侵害による被害が起こらないようデータを加工する技術として，Privacy Preserving Data Mining(PPDM)が提唱されている．

2. PPDM の手法

PPDM において，データを複数組み合わせることで個人を特定される可能性があるものを準識別子と呼ぶ．身長や出身国などといったデータは準識別子の例である．そして，個人が特定された場合に漏洩すると不利益となってしまうが，分析上重要なデータをセンシティブ属性値という．PPDM では，これらのデータを加工して，分析上の実益と安全性の担保の両立を実現していく．具体的な手法には，Sweeney による k-匿名化[1]などがある．これは,準識別子の一般化によって同一となったデータが,k 個以上データ集積中にあることを保証する匿名化手法である．

3. 水平分割データベースの匿名化と課題

3. 1 想定するデータ分析シナリオ

以降では，分析対象がデータベースであると考えて議論を進める．まず,データ分析の形態として,データ保有機関が複数あり，各々の持つデータベースを統合して分析を行う，という分散データベースにおける分析を想定する．ただし,各機関は自身の持つデータを他機関に開示せず秘匿する．データは行ごとの分割として各機関に集積されている,水平分割された状態にあるものとする．分析方法としては，各々の機関のデータベースを匿名化後,第三者機関が結合し，一つの分析用データベースを得るというモデルを考える．この第三者機関は信用でき,攻撃を行わないが，法律等の要請により各機関の匿名化前データを提供できないと想定する．このデータベースは分析者や元の機関に提供され，クロス集計による分析などの種々の用途に活用される．

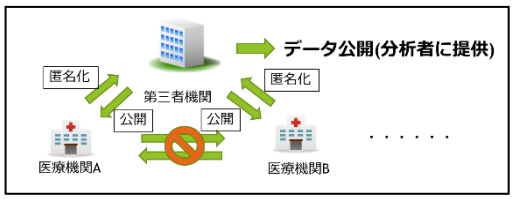

図 1 データ分析シナリオのモデル図

3. 2 重複データの定義

表 1 元データ

X			
氏名	年齢	身長	疾患
Alice	16	169	がん
Bob	13	158	脳卒中
Charlie	18	167	肺炎

Y			
氏名	年齢	身長	疾患
Alice	16	169	肺炎
David	25	178	かぜ
Ellen	19	168	心膜炎

表 2 匿名化データ

X*			
氏名	年齢	身長	疾患
削除	[11,20)	[150,170)	がん
削除	[11,20)	[150,170)	脳卒中
削除	[11,20)	[150,170)	はしか

Y*			
氏名	年齢	身長	疾患
削除	[15,30)	[165,180)	肺炎
削除	[15,30)	[165,180)	かぜ
削除	[15,30)	[165,180)	心膜炎

表 3 問題のある重複データ処理後のデータ

年齢	身長	疾患
[11,20)	[150,170)	がん, 肺炎
[11,20)	[150,170)	脳卒中
[11,20)	[150,170)	はしか
[15,30)	[165,180)	かぜ
[15,30)	[165,180)	心膜炎

第三者機関におけるデータ統合において，同一個人に帰することのできるデータを重複データと定義し，統合時そのセンシティブ属性値を第三者機関において集約することと規定する．これにより，例えば 20 歳代のある同一人物ががんと肺炎を同時に罹患しているという情報から，20 歳代にがんと肺炎の併発が多いかどうかというデータ分析が可能になる．この処理の例が表 3 である．これは医療機関が提供するデータを想定しており，準識別子は年齢・身長・体重，センシティブ属性値は疾患名である．X は医療機関 A が，Y は医療機関 B が保有する患者データである．この例では,両機関に通院している患者である Alice のレコードが重複データとなっている．処理の流れとしては，表 1 で示したデータを表 2 のように匿名化し，重複データである Alice のレコードをまとめ，センシティブ属性値を一つに集約する，というものになる[*1].

このようなデータでは次のような問題が発生する．医療機関 A は，X と公開された匿名化後のデータテーブルの両方を参照できる．表 3 では，重複データ処理により，センシティブ属性値を複数持つレコードが重複データであると分かってしまう．医療機関 A はこの重複データの値から Alice のレコードを特定し，さらに{がん，肺炎}

＊1 重複データの判定には，患者氏名と患者の電話番号などを組み合わせた文をハッシュ関数に入力し，その出力の一致をみることで実現する．

Anonymization for A Horizontal Distributed Database Including Duplicate Records by Adding Fake Sensitive Attribute Values

Yuki Ina†, Yuichi Sei‡, Yasuyuki Tahara‡, Akihiko Ohsuga‡

†Department of Informatics, Faculty of Informatics and Engineering, The University of Electro-Communications 182-8585, Tokyo, Japan

‡Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications 182-8585, Tokyo, Japan

という疾患データから，本来 A が知りえない「肺炎」という Alice の疾患名を特定してしまう．同様に，医療機関 B は A が保護したいデータである「がん」という疾患名を特定できる．このように，重複データがある場合，データ保有機関双方から情報漏洩が起こってしまう．

4. 問題解決のための提案手法

表 4 問題を解決した重複データ処理後のデータ

年齢	身長	疾患
[11,20)	[150,170)	がん, 肺炎
[11,20)	[150,170)	脳卒中
[11,20)	[150,170)	はしか, 心膜炎
[15,30)	[165,180)	がん, 肺炎
[15,30)	[165,180)	かぜ
[15,30)	[165,180)	はしか, 心膜炎

4. 1 ダミー値挿入による匿名化手法

この問題に対し，センシティブ属性値へのダミー値挿入という手法を提案する．例を表 4 に示す．このデータテーブルは表 3 から削除された重複データを復活させ，ダミー値として Charlie の疾患名にはしか，Ellen の疾患名に心膜炎を加えたものである．これにより，センシティブを複数有するレコードが真の重複データ以外にも出現している．そのため，表 3 とは違い，重複データは{がん，肺炎}を疾患名に持つレコードなのか，{はしか，心膜炎}を疾患名に持つものなのか分からなくなっている．

このダミー値挿入は，重複データの処理と同時に行われる．これらをアルゴリズムとしてまとめたのが図 2 である．図 2 の 1 から 9 行目までが重複データの処理，以降がダミー値挿入の処理になっている．QID_{Xi} はデータ保有機関 X_i の持つ匿名化後 QID 値グループ全体の順序付き集合であり，$QID_{Xi,k}$ はその k 番目の要素である．表 4 を例とすると，QID_{Xi} は {([11,20),[150,170)),([15,30),[165,180))}となる．レコード r の持つ QID 値グループを r．QID，センシティブ属性値を r．SA と表す．ここで，ダミー値を追加しても，QID 値範囲の共通部分をとると，他機関にダミーだと判別されてしまう場合がある．これを防ぐには，この共通部分に入るようなダミー値を挿入可能なレコードを特定することが必要であり，この処理は 10 行目から 12 行目の処理に該当する．この処理について詳述すると，これはダミー値を挿入する条件を備えたレコードの一部を確率 α によって選択する処理となっている．これにより，目的に応じたレベルでの匿名化が可能となる．

重複データ処理及びダミー値挿入アルゴリズム
1 n := 参加するデータ保有機関の数
2 r := データレコード
3 *for* $i = 1$ *to* n
4 *for* $k = 1$ *to* $|QID_{X_i}|$
5 *for* $m = 1$ *to* $|QID_{X_i,k}|$
6 r_m := $QID_{X_i,k}$ をQIDとしてもつレコード
7 *if* r_m がヨr'と重複データ AND r_mはアルゴリズム未処理 *then*
8 $QID_{r'}$:= r'が属しているQIDグループ
9 $r_m.SA \leftarrow \{r_m.SA, r'.SA\}$
8 $r'.SA \leftarrow \{r_m.SA, r'.SA\}$
10 D_{X_i} := $QID_{X_i} \wedge QID_{r'}$ を満たすQID値を持つQID_{X_i}のレコード集合
11 $D_{r'}$:= $QID_{X_i} \wedge QID_{r'}$ を満たすQID値を持つ$QID_{r'}$のレコード集合
12 $\min(\alpha \cdot |D_{X_i}|, \alpha \cdot |D_{r'}|)$ 個のレコード，$r_{D_{X_i}}$ 及び $r_{D_{r'}}$ を選択
13 $r_{D_{X_i}}.SA \leftarrow \{r_{D_{X_i}}.SA, r_{D_{r'}}.SA\}$
14 $r_{D_{r'}}.SA \leftarrow \{r_{D_{X_i}}.SA, r_{D_{r'}}.SA\}$
15 *Next*
16 *Next*
17 *Next*

図 2 提案する問題解決アルゴリズム

4. 2 真値推測の手法

提案手法によるダミー値の多数追加により，データ分析時の精度は減少する．より有用な結果を得るためには，提案手法のみならず，真値推測のためのアルゴリズムが必要である．Sei らの研究[2]によれば，ダミー値が含まれたセンシティブ属性値を持つデータベースに対しても，そこから統計的な情報を高精度で導くことができる．このデータベース中におけるセンシティブ属性値 s_i の出現回数を w_i，とすると，これは真の出現回数 t_i とダミーによる出現回数 d_i を用いて次のように表すことができる．

$$w_i = t_i + d_i \qquad (1)$$

このとき，w_i は既知であるから，d_i の期待値を求めれば，t_i を得ることができる．ただし，Sei 論文で扱われているのは単一のデータベースであるため，水平分割データベースでは d_i の算出方法が変化する．ここで，ある重複データのセンシティブ属性値(s_i, s_j)に対し，これがダミー値であるのは，s_i というセンシティブ属性値に s_j が，s_j というセンシティブ属性値に s_i が，それぞれ選ばれてダミー値として挿入された場合である．この選ばれる確率をそれぞれ p_i, p_j とする．また，ダミー値が追加されるのは確率 α による．以上から，s_i, s_j をそれぞれセンシティブ属性に持つレコードの数を r_i, r_j とすると，ダミー値の個数 d_{ij} は以下のように推測できる．

$$d_{ij} = \alpha(p_i \cdot r_i + p_j \cdot r_j) \qquad (2)$$

5. 関連研究

水平分割されたデータベースに対しての匿名化に関しては，いくつかの先行研究が確認できる．特に，竹之内らによる論文[3]では，本研究と同様に水平分割データベースの匿名化について，複数機関からの突合に対する観点から論じている．しかし，この論文では重複データが存在しないものとして扱われており，更に匿名化も第三者機関が受け取った後で行われる方式であり，根本的に想定するシナリオが異なる．

6. おわりに

水平分割データベースで生じる問題について，その匿名化手法と分析時におけるアルゴリズムについて述べた．今後は IPMUS-USA から入手したレコード数 50000 件のテストデータセットを用い，実験を行う．実験により，(2)における確率値 p_i, p_j を詳細に決定し，またアルゴリズムが有効であることの確認を行う予定である．

7. 参考文献

[1]Sweeney L, k-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Volume 10, Issue 05, p557-570, 2002
[2] Yuichi Sei, Takao Takenouchi, Akihiko Ohsuga, (l1, . . . , lq)-diversity for Anonymizing Sensitive Quasi-Identifiers, Proc. IEEE TrustCom, pp. 596-603, 2015
[3] 竹之内 隆夫, 側高 幸治, 豊田 由起, 高橋 翼, 森 拓也, 部分データセットとの突合に対する耐性を有するレセプト匿名化方式, 医療情報学Vol. 33, No. 3, p127-138, 2013