プライバシー保護手法における攻撃者知識と安全性に関する評価
A study on the relationship between attacker's background knowledge and privacy for anonymization

情報工学専攻　林 弘悦
Hiroyoshi Hayashi

1　はじめに

個人の情報を含むデータをプライバシー侵害することなく分析可能な形式に変換する匿名加工の安全性を評価するためには，適切な攻撃者モデルの設定は不可欠である．従来想定である最大知識攻撃者モデル [1] は最も強い攻撃者を想定しているため, 最悪の場合の安全性を評価するという点では適切と考えられる．しかし, 現実的にはオリジナルデータをすべて知る攻撃者という想定が当てはまることは少ない．攻撃者モデルを変えた場合は安全性への影響が生じるはずであり，これらの関係についても適切に評価しておく必要がある．本研究では, 新たに想定される匿名加工の状況に応じた複数の攻撃者モデルを提案する．データ内に親しい知り合いがいる場合や, 分散管理されているようなデータ内の一部を持つ場合, それらよりも, より限られた知識を持つ場合, 反対にデータのほとんどを知る場合の攻撃者を想定したモデルについて検討した．提案モデルについて, 実データセットを用い, 安全性と攻撃者知識との関係について実験評価を示す.

2　データ匿名加工の安全性と有用性

本研究では, 有限個のレコード (行に対応) と属性 (列に対応) からなるテーブル形式の DB を扱う．テーブルの各行をレコード, 各列を属性と呼び, 各セルに格納されている値を属性値と呼ぶ．本稿で扱うテーブルにおいて, レコードはある個人に関する各属性の属性値の組である．属性には有限種類の属性値分類を格納するカテゴリ属性や, 整数値や実数値などを格納する数値属性がある．テーブルにおける特定のレコードを識別することを Record Linkage と呼び, テーブルを開示した際に個人の識別が起こる危険性を個人識別リスクと呼ぶ.

個人識別リスクを低減する匿名加工手法として, 属性やレコードの削除, 乱数に基づき数値属性を変化させるランダムノイズ加算, 似た属性値をグルーピングし曖昧化する一般化, 属性値を入れ替えるスワッピングなどが使用される．匿名加工データの安全性の測定手法として, k-匿名性や l-多様性などの匿名性指標の他に実際に攻撃し, どれだけ正しくマッチングできたかによって識別リスクを評価する方法がある．再識別リスク測定に用いられるマッチング手法としては, 属性値の一致でマッチングする確定的マッチングや, 属性をソートし, その順位が同じもの同士をマッチングする順位ベースマッチング, 距離尺度の小さいものをマッチングする距離ベースマッチングがある．一般にオリジナルデータセットを変換し, 匿名加工を行う過程で必ず情報が失われる．既存の有用性指標としては, 数値属性であれば平均絶対誤差や平均変化率などがあり, 一般化に匿名加工に対しては Discernibility Metric などがある．匿名加工データの有用性と安全性はトレードオフの関係にあると言われ, 多くの研究がそのトレードオフの中で, 最適な解, すなわち評価基準を最もバランスよく満たす匿名加工の作成を目的としている.

3　攻撃者モデル

匿名加工データの安全性は, 想定する攻撃者モデルに応じて評価方法が異なる．本研究では, Record Linkage を攻撃者の目的と想定する．従来手法である最大知識攻撃者モデル [1] は以下で定義される．攻撃者は全オリジナルデータセット T と, 対応する匿名加工データセット $T^{'}$ を知っている．攻撃者の目的はオリジナルデータと匿名加工データの間の正しい Record Linkage を得ることである.

最大知識攻撃者モデルは攻撃者が匿名加工前後のテーブルについてすべての情報を有するという, 最悪の場合を想定している．しかしながら現実では攻撃者がすべての情報を有した状態で対応付けを図ることは考えにくい．よって, 匿名加工データの安全性を図る際には, 最大知識攻撃者モデルだけを想定するのではなく, 攻撃者の持つ知識を変化させ, 実際の攻撃に対する安全性を測定するのが望ましい.

攻撃者知識を制限している既存の研究として, [3] では公的統計などで用いられるリサンプリングによる識別リスクの低減を評価する際に攻撃者の背景知識を制限している．具体的には, 攻撃者はテーブル内の全 n レコードのうち, その一部である n_0 個のレコードを知っており, 知らない部分については知識レコードと同じ分布であると仮定し推定を行った場合のリスク証明を与えている．この知識制限は本研究における部分レコード知識モデル (4.1 節) と同様のモデルであるが, [3] ではサンプリングによるリスクの低減の性質のみを対象としている．また, [4] では単独でレコード識別可能となる属性を疑似 ID に置き換える仮名化という手法を履歴データに適用したときの安全性について, 背景知識を制限したものについても評価している．具体的には, 攻撃者が具体的な属性値を知っている (知識の質が高い) 場合, 具体的な情報を知らず 1 ビットの情報のみを持つ (知識の質が低い) 場合の 2 パターンについて, 情報の漏洩量を評価する手法を提案している．[4] は仮名化による漏洩量の算出が主な提案で, 攻撃者モデルの制限に関しては 2 パターンの想定にとどまっている．本研究は複数パターンかつ強さの調整が可能な攻撃者モデルを提案しているため, より攻撃者モデルの分類に踏み込んでいるといえる.

4　提案

本研究では, 想定される匿名加工の状況に応じた 4 パターンの攻撃者モデルを提案する.

4.1　部分レコード知識モデル

攻撃者はオリジナルテーブル T のすべてのレコード N のうち, ランダムな N_0 個のレコードのみを知識として持つ. これはテーブル内に攻撃者の親しい知人の情報が含まれている場合などに当てはまる. 図 1 は部分レコード知識モデルの例である. 攻撃者には, T の 1,2,3,4 レコードと T' 全体が与えられ, 攻撃者の目的は T の 1,2,3,4 レコードが T' のどのレコードに対応するか推測することである. 攻撃者は知識として T の一部のレコードのみを有すため, T' 内に特徴的なレコードがあっても, それに対応するレコードを T 内に見つけられない場合があり, 最大知識攻撃者モデルと比べて識別リスクが低くなることが期待できる.

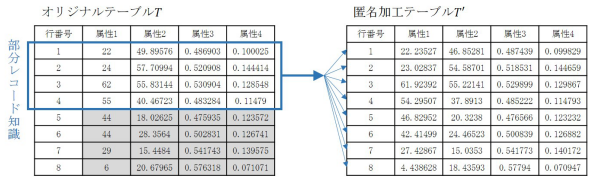

図 1　部分レコード知識モデルの例

4.2　部分属性知識モデル

攻撃者 T の M 個の属性のうち M_0 個の属性のみを知識として持つ. これはテーブル内の攻撃者にとって入手しやすい属性情報と入手しづらい属性情報が明確である場合に当てはまる. 図 2 は部分属性知識モデルの例である. 攻撃者には, T の全レコードの属性 3, 属性 4 と T' 全体が与えられる. 攻撃者の目的は T 内の各レコードが T' のどのレコードに対応するか推測することである. 攻撃者は知識として, T' の全属性を有しているものの, テーブル T の属性 1, 属性 2 については知らないため, 再識別を試みる際には属性値の比較ができない. 例えば, T のあるレコードと T' のあるレコードとの間で属性 1 や属性 2 を用いて距離を比較することができないため, 最大知識攻撃者モデルと比べて識別リスクが低くなることが期待できる.

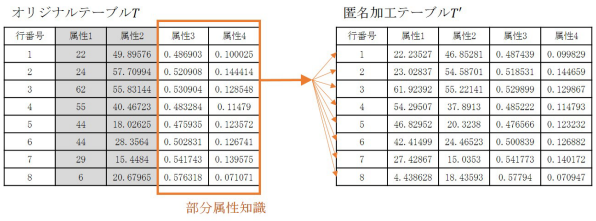

図 2　部分属性知識モデルの例

4.3　部分テーブル知識モデル

T 内の N レコード $\times M$ 属性のうち部分集合 $N_0 \times M_0$ 個のレコード, 属性のみを知識として持つ. これは部分レコード知識モデル, 部分属性知識モデル双方よりも弱い攻撃者想定である. あまり親しくない知人や web 上から収集した他人の情報を知識として持つ攻撃者に当てはまる. 図 3 は部分テーブル知識モデルの例である. 攻撃者には, T の 2, 3, 4, 5 レコードの属性 3, 属性 4 と T' 全体が与えられる. 攻撃者の目的は T 内の 2, 3, 4, 5 レコードがテーブル T' のどのレコードに対応するか推測することである. 4.1 節の部分レコード知識モデルと比較すると, N_0 の値が同じであれば, 知識属性数が M_0 個だけ少ない. 4.2 節の部分属性知識モデルと比較すると, M_0 の値が同じであれば, 知識レコード数が N_0 個だけ少ない. そのため, 最大知識攻撃者モデルと比べて識別リスクが低くなることが期待できる.

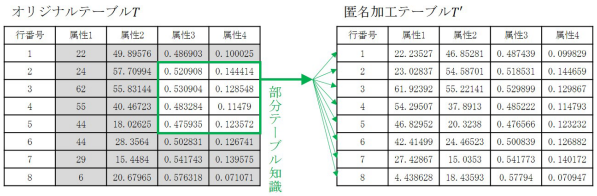

図 3　部分テーブル知識モデルの例

4.4　準完全知識モデル

T 内の N レコードのうち, $N \times R$ レコードについてはすべての属性 M 個を知っており, 残りの $N_1 = N \times (1 - R)$ $(0 \leq R \leq 1)$ レコードについては M_0 個の属性のみを知識として持つ想定. 言い換えると, 攻撃者は $(N \times (1 - R)) \times (M - M_0)$ 個の属性値のみを知らず, それ以外のオリジナルテーブルすべてを知識として持つ. 部分テーブル知識モデルは部分レコード知識モデルと部分属性知識モデルの知識範囲の共通部分を知識として持っていたのに対して, 準完全知識モデルは 2 モデルの知識範囲の和集合を知識として持つと考えられる. このモデルでは, 例えば 1 属性値のみを知らず, それ以外のすべてのオリジナルテーブルの属性値を知識とすることが可能である. そのため, 今回提案する知識制限の中で最も弱い制限をとることが可能で, 最も最大知識攻撃者モデルに近い知識を設定可能なモデルである.

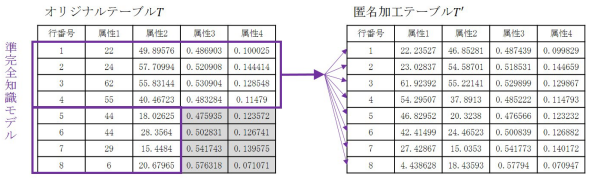

図 4　準完全知識モデルの例

5　実験と評価

実験の流れ　オリジナルテーブル T に匿名加工を行い, $T^{'}$ を得る. T の一部分の知識と $T^{''}$ を用いてマッチングを行う. マッチングの正解率に基づく指標を評価する. モデルによる差異を定量的に評価することが目的の一つであるため, 匿名化手法とマッチング手法として簡潔なものを採用する.

実験データ　定量評価の容易性のため, 無相関な数値属性を複数もつデータセットを用いる. UCI Machine Learning Repository で公開される Diabetic Retinopathy Data Set のそれぞれ無相関な数値属性である, 属性番号 3,9,17,18 を用いる.

匿名加工　T の i 番目の属性の標準偏差 σ_i に対し, 平均 0, 標準偏差 $p \times \sigma_i$ となる正規分布に従うランダムノイズ加算を行う. 係数パラメータ p について, 本稿では $p = 0.05$ を用いる.

マッチング方法　攻撃者が知識として有するレコード (T の一部) を匿名加工テーブル T'' のレコードへ対応付けを行う方法として以下 2 つを用いる.

[距離ベースマッチング] 正規化ユークリッド距離の値が最小なレコード同士を対応付ける

[順位ベースマッチング] T, T'' 内で各属性を降順でソートし, 同じ順位のレコードを対応付ける

準完全知識モデルに関してのみ, 距離ベースマッチングを用いた段階的なレコードマッチングを行う.

評価指標　本稿では [2] を参考にした以下 2 指標を用いる.

$$entire = \frac{\text{マッチング正解数}}{T' \text{全体のレコード数}} \tag{1}$$

$$restricted = \frac{\text{マッチング正解数}}{\text{知識レコード数}} \tag{2}$$

知識制限

[部分レコード知識モデル] $N_0 = N \times r$ として, r の値を 0.5 ～ 1.0 まで 0.1 刻み (つまり 50% ～ 100% まで 10% 刻み) で試行した.

[部分属性知識モデル] 4 つの数値属性 $3, 9, 17, 18$ から知識属性数 1 ～ 4 のすべての組み合わせを試行する.

[部分テーブル知識モデル] 知識レコード数が $N_0 = N \times r$, 知識属性数が 1 ～ 4 となるような知識を与える. すなわち, 上記 2 モデルを同時に満たすような知識制限を与える.

[準完全知識モデル] すべての属性を知っているレコードの数 $N \times R$ について, R の値を 90% ～ 10% まで 10% 刻みで, 知識属性の数については 1 ～ 3 のすべての組み合わせを試行する. このモデルに対してのみ下記指標 3, 4, 5 を用いる.

$$r1 = \frac{\text{属性完全レコードの再識別正解数}}{\text{属性完全レコード数}} \tag{3}$$

$$r2 = \frac{\text{属性欠損レコードの再識別正解数}}{\text{属性欠損レコード数}} \tag{4}$$

$$E = \frac{\text{全体での再識別正解数}}{\text{全体レコード数}} \tag{5}$$

$r1, r2$ はそれぞれ 1 段階目,2 段階単体での再識別率を表す. E は 1,2 段階目の再識別処理全体での再識別率を表す.

5.1　距離ベースマッチングによる実験結果と評価

距離ベースマッチングによる実験結果を図 5, 6, 7,8, 9 に示す. 図 5, 6 における高さは指標値の大きさを表す. 横軸は知識属性の組み合わせであり, 左の方ほど知識属性数が大きい. 奥行はオリジナルテーブルに対して知識レコード数が占める割合であり, 奥の方ほど知識レコード数が大きい.

5.1.1　評価指標 entire についての評価

距離ベースマッチングについて, 指標 entire は知識レコード数が大きければ大きく, 小さければ小さくなると予想していた. 知識属性数が 4 の場合は予想通りである. 知識属性数が小さいときは知識レコード数の影響を受けにくいことが今回の実験データの特徴である (図 5 の知識属性数が 1 の場合). 実験データに依存するが, 知識属性数が大きい場合はノイズ加算だけでなく, 置換やレコーディングなどの他の匿名加工手法を用いる必要がある.

5.1.2　評価指標 restricted についての評価

知識レコード数が変化しても, 指標値は変化しないと予想した. 今回の実験では予想通り, 同じ属性知識条件下において, 知識レコード数が変化してもほぼ等しい結果が得られた (図 6).

マッチングについて, 正規化ユークリッド距離最短なレコード同士を対応付けているため, 知識レコード数が変化しても再識別率が大きく変化することは無かった.

5.1.3　2 段階再識別についての評価

以下では, すべての属性について知っているレコードを属性完全レコードと呼び, M_0 個の属性のみを知るレコードを属性欠損レコードと呼ぶ.

[1 段階目の再識別に対する評価] 図 7 は T 内の全ての属性 (3, 9, 17, 18) を知っているレコードを用いた, 1 段階目の距離ベースマッチングについての $r1$(式 3) を示したグラフである. 1 段階目における restricted 指標である $r1$ は, すなわち restricted 指標の結果 (図 6) に対して サンプリングした場合と捉えることができるため, どの知識属性, 知識レコードについてもほぼ一定の指標値が得られた.

[2 段階目の再識別に対する評価] 図 8 は 2 段階目のマッチング, すなわち, T 内の全ての属性 (3, 9, 17, 18) に対して, 1 つ以上の属性を知らない状況で, 距離ベースマッチングを行った際の $r2$(式 4) を示したグラフである. $r2$ については欠損属性の数が増えるほど (グラフ右側ほど), 低い指標値を示している. これは部分属性知識モデルの結果である図 6 の横軸の推移と同様, 知識属性が少ないほどマッチングの精度が下がっているためである. また, 欠損レコード数が増えるほど低い指標値になっている. 部分レコード知識モデルの結果である図 6 では一定の値を示したが, こちらは, 属性知識にも欠損が生じているため, 欠損レコードの数が増えるほど低い指標値を示している.

[2 段階再識別全体に対する評価] 2 段階再識別処理全体での結果が図 9 である. また, 奥に行くほど, すなわち属性完全レコード割合が大きいほど大きい指標値を示しており, 識別リスクが高まっている. さらにグラフの左側に行くほど, すなわち攻撃者が持つ知識属性数が大きいほど, 大きい指標値を示しており識別リスクが高まっていることがわかる. この性質自体は 2 段階再識別ではない, 通常の距離ベースマッチングの結果 (5.1.1) と似ているが, 異なる部分としては, 2 段階目の再識別での指標 $r2$, すなわち再識別精度が低いためにこのような結果となっている.

5.2　順位ベースマッチングについての評価

順位ベースマッチングでの entire 指標の実験結果を表 1 に示す. 順位ベースマッチングはノイズの大きさのパラメータである p を 0.05 や 0.1 にしてもマッチングは成功しなかった. 更にノイズを小さく, 具体的には p=0.01 とした結果を表 1 に示す. 表 1 内の知識属性が

属性 3 のみの箇所を見ると, ノイズが付与されていないときに指標が 0.05 と低く, ノイズが p=0.01 のときには, それよりも指標が大きく 0.10 となっている. 今回実験で用いた順位ベースマッチングのアルゴリズムでは同順位の箇所はすべて最初に出現したレコードにマッチングしている. またランダムノイズによる匿名加工ではオリジナルデータで整数属性のものも, 実数のランダムノイズを与え, 実数属性に置き換えている. そのため, オリジナルデータにおける整数属性である属性 3 は, 数値の重複が多く存在し, 同じ順位のレコードについてマッチングが成功する確率が低くなっている. 属性 3 のみの箇所を除くと, ノイズが付与されていないときは指標が上限である 1.0 に近い, 大きい値となっており, ノイズが p=0.01 のときは, 指標が 0.09 ～ 0.17 と非常に低い値となっている. 特に単一属性 (属性 3 のみの場合を除く) での結果から, 今回実験で用いたデータは少しでもノイズを与えると順位が変動するデータであると言える.

6 結論

従来研究ではデータ利用の場面によっては過大にリスクが見積もられてしまう可能性があった. データを利用する環境に応じたモデル設定をできるようにするため, 本研究では, 攻撃者が有する知識について, 想定される知識量や性質に応じて複数の攻撃者モデルを提案した. 提案攻撃者モデルについて, シンプルな状況を想定し, 実データセットを対象として, 再識別リスクの実験評価を行った. 知識モデルに応じて, 再識別リスクが大きく異なるという結果を得た. 準完全知識モデルに対する実験では, すべての属性について知っているレコードがあると, そのほかのレコードの属性知識に欠損があっても, より大きい再識別リスクとなることがわかった. 今回提案した攻撃者モデルに関して, 再識別リスクと有用性との関係の検証や, 匿名加工手法と再識別アルゴリズムの複雑化が今後の課題である.

謝辞

本研究を進めるにあたり御指導を頂きました, 中央大学理工学部趙晋輝教授, ならびに情報セキュリティ大学院大学土井洋教授に感謝申し上げます.

関連発表

林弘悦, 西山賢志郎, 土井洋, 趙晋輝, “匿名加工における攻撃者知識と安全性に関する評価”, 情報処理学会全国大会, 2017.

参考文献

[1] J. Domingo-Ferrer, K. Muralidhar. *New directions in anonymization: permutation paradigm, veriability by subjects and intruders, transparency to users*, CoRR, abs/1501.04186, 2015.

[2] 菊池浩明, 小栗秀暢, 野島良, 濱田浩気, 村上隆夫, 山岡裕司, 渡辺知恵美. *PWSCUP:* **履歴データを安全に匿名加工せよ**, コンピュータセキュリティシンポジウム, pp.271-278, 2016.

[3] 菊池亮, **サンプリングを用いた際の個人識別リスクの評価**, コンピュータセキュリティシンポジウム, pp.173-179, 2016.

[4] 早稲田篤志, 野島良, 盛合志帆, 菊池浩明, **良い仮名化悪い仮名化**, コンピュータセキュリティシンポジウム, 2D1-6, 2017.

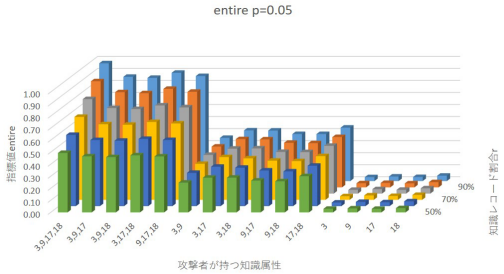

図 5　entire p=0.05

図 6　restricted p=0.05

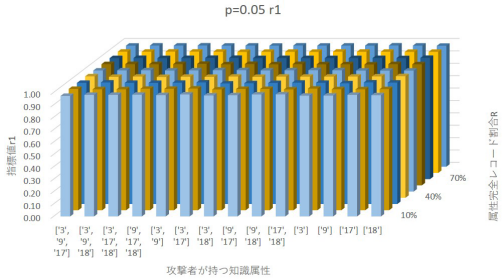

図 7　2 段階再識別 p=0.05 指標 r1

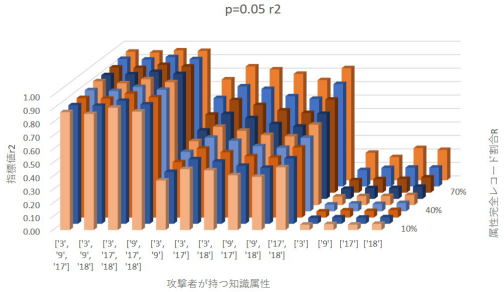

図 8　2 段階再識別 p=0.05 指標 r2

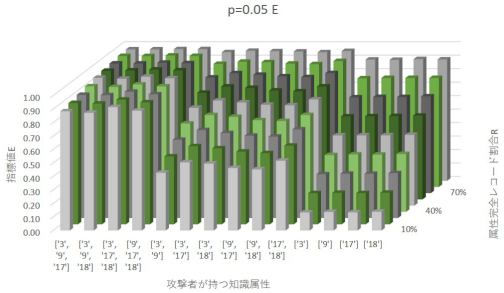

図 9　2 段階再識別 p=0.05 指標 E

表 1　entire 順位ベースマッチング

	知識属性														
	3,9,17,18	3,9,17	3,9,18	3,17,18	9,17,18	3,9	3,17	3,18	9,17	9,18	17,18	3	9	17	18
ノイズなし	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.98	0.98	0.05	0.98	0.97	0.93
ノイズ p=0.01	0.12	0.15	0.14	0.10	0.14	0.14	0.09	0.09	0.14	0.14	0.15	0.10	0.14	0.14	0.17
ノイズ p=0.05	0.04	0.04	0.04	0.05	0.05	0.03	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.05
ノイズ p=0.1	0.03	0.02	0.03	0.03	0.03	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.03