PWS Cup 2018: 匿名加工再識別コンテストの設計 ～履歴データの一般化・再識別～

濱田 浩気[1,2]　荒井 ひろみ[2]　小栗 秀暢[3]　菊池 浩明[4,2]　黒政 敦史[5]　中川 裕志[2]　西山 賢志郎[6]
波多野 卓磨[7]　村上 隆夫[8]　山岡 裕司[3]　山田 明[9]　渡辺 知恵美[10]

概要：2017 年 5 月に改正個人情報保護法が施行され，データを「匿名加工情報」に加工することにより本人同意なしに第三者提供ができるようになった．しかしながら，匿名加工情報の標準的な加工手法や評価手法は定まっておらず，匿名加工情報の活用の活発化のために優れた加工手法や評価手法を明らかにすることが求められている．そこで我々は匿名加工したデータの安全性と有用性を競うコンテストを企画し，コンテストの実施を通して適切な評価手法の確立と高度な匿名加工技術の開発を図る．コンテストでは参加者は履歴データを加工し，再識別リスクに関する一定の条件を満たしながら元のデータにより近いデータを作成することを目指す．本稿では，このコンテストの基本定義について述べる．

キーワード：個人情報保護，匿名加工，匿名性

1. はじめに

近年のデータ分析技術の進歩により，多くの企業でパーソナルデータを蓄積し，機械学習やデータマイニングなどのデータ利活用を行っている．しかし，それらの利活用技術は学習データが多量に必要であり，一つの機関のデータだけでは分析目的が達成できない場合がある．そこで，他の機関が保持するデータを使用し，データ分析の精度を向上させたいというニーズが高まっている．

2017 年 5 月 30 日に全面施行された改正個人情報保護法によって，匿名加工情報という新たな情報の類型が定義されたことにより，一定の条件の下で，本人の同意がなくても第三者提供や目的外利用が可能となった．そこで，保持するデータを匿名加工して第三者に提供するという，新しい情報流通方式が徐々に定着し始めている．

しかしながら，データの特徴に応じて最適な匿名加工を行う技術は，必ずしも自明ではない．匿名加工の手法や安全性指標は，数多く存在しており，個人情報保護委員会が作成したガイドライン [1] 等に示された情報だけでは不十分である．

欧州連合 (EU) におけるデータ保護のアドバイザリー機関である第 29 条作業部会は「匿名化技術に関する意見書 [2]」にて，匿名化アルゴリズムを選択する基準として以下を定めている．

(1) 個人を識別すること (single out) は可能か
(2) 個人に関する記録と紐付けることは可能か
(3) 個人を推定することは可能か

これらの基準に合わせて匿名化技術を選定することが必要とされるが「それぞれの手法に長所と短所があるため，目的に即した適切な手法を選択すること」を求めている．

より良い匿名加工データを作成するには，個々のデータ加工技術者によって，有用性を維持する加工と，安全性指標の選択・検証作業が重要となる．しかし，匿名加工情報の加工技術の公開は，既に流通している情報の安全性を損ねる可能性があるため，安全に管理することが定められており，匿名加工の技術を高める機会は限られていることが課題であった．

これらの課題により，個々の技術者，研究者による匿名加工技術の向上を促す機会と，特定の匿名加工情報に依存しない技術共有の場が求められていた．

このような状況において，情報処理学会コンピュータセキュリティ研究会 (CSEC) は，産学が共同してプライバシー保護技術の研究開発を活性化し，議論するため，2015

[1] NTT セキュアプラットフォーム研究所
[2] 国立研究開発法人 理化学研究所
[3] 株式会社富士通研究所
[4] 明治大学
[5] 富士通クラウドテクノロジーズ株式会社
[6] 株式会社ジーニー
[7] 新日鉄住金ソリューションズ株式会社
[8] 国立研究開発法人 産業技術総合研究所
[9] 株式会社 KDDI 総合研究所
[10] 筑波大学

年にプライバシーワークショップ (PWS) を発足した．その中でも匿名加工技術の発展のために，毎年行われているのが，匿名加工・再識別コンテスト (PWS Cup)[3][4][5] である．

過去 3 回の PWS Cup では，毎回異なる有用性と安全性の指標を提案し，参加者間での技術の向上と知識の共有を促してきた．第 4 回目となる PWS Cup 2018 では，加工方法としての頻繁に利用されてきた「一般化」の手法に着目し，ルールの設計を行った．

本稿は，コンテストの概要，データセットの定義，有用性と安全性の基準などについて詳説することで，コンテスト参加者の理解を深め，匿名加工技術に関する議論を活性化することを目的とする．

1.1 本年度のコンテストの特徴

まず，本年度のルールについて検討する前に，昨年度までの課題と本年度の特徴についてまとめる，

1.1.1 一般化による加工

過去に行われたコンテストでは，マスターデータ型，トランザクション型 (履歴型)，トランザクション型の仮名制御，というパーソナルデータの「型」に着目したルール設計を行ってきた．

しかし，その反面，データに含まれる値の加工方法については，摂動 (数値の変更，ノイズ付与，スワップ等の処理) が主に利用されており，広く利用されているデータの一般化 (抽象化，集合化，抑制等の処理) については，使用できる手法が制限されており，加工方法の効果を検証することができなかった．

一般化の処理を施した匿名加工データの有用性は，元データの分布に依存する部分が大きい．例えば，偏りの大きいロングテール型の性質を持つトランザクションデータに一般化処理を施すと，情報の有用性が著しく低下する危険性があり，現実的には困難であると言われている．しかし，常に元情報と矛盾しない情報を出力することから，分析データに対する正確性が求められる分野において，広く活用されている．例えば災害や医療などの分野で匿名加工情報を利用する場合，摂動法を一部でも用いた場合，元情報の正確性が失われてしまい，利活用の目標が達成できない場合がある．

そこで PWS Cup 2018 では，元データの正確性が必要な分析分野を想定し，元データの値と匿名加工された値が常に正しく対応するように，元データを必ず含む一般化の加工に限定するようルールを定めた．これらの加工の制限については 5 節にて述べる．

1.1.2 安全性の基準の設定

昨年までのルールでは，安全性の評価値と有用性の評価値は等価であり，その総和で総合評価を行ってきた．しかし，実際にデータ利活用する側の立場で考えると，安全性には一定の基準を定めて，その基準を満たす範囲で，できる限り元データに近い加工を行う方が自然である．

そこで今回は，安全性に一定の基準を定め，その基準を満たした上で最も有用性の高いデータを作成することを目的とする．

1.1.3 安全性の評価方法の変更

昨年までのルールでは，匿名加工されたデータに対して，他のプレイヤーが再識別攻撃を行い，その再識別された割合によって安全性を評価した．これは，安全性の定量的な測定方法として優れた手法であるが，その反面，ある個人の識別されやすさには制限が無いことが課題であった．例えば，有用性を高めるために，最も購入数が多い個人を，元データそのままで提出することが可能である．これにより安全性は 1 名分低くなるが，購入数が多いことから有用性は高くなるため，コンテストとしては有利に働くことがあった．これは求めるプライバシー保護基準としては不十分と判断した．

そこで今回は，全ての個人に対して，一定のしきい値以上に再識別されにくく加工されていることを，安全性の基準とする．

1.1.4 攻撃者知識の設定

PWS Cup 開催の目的は，匿名加工技術の進展であり，最も強い攻撃者を想定した最大知識攻撃者モデル [6] は有益である．しかし，現実には部分的な知識のみを保持する攻撃者 (部分知識攻撃者) も珍しくないと考えられ，部分知識攻撃者を想定した評価もできることが望ましい．

そのため，昨年度のコンテストでは部分知識攻撃者モデルを想定してルールの設定を行った．しかしながら昨年度のルールはすべてのコンテスト参加者がルールに従うことを前提としており，参加者が悪意なく間違ってルールを逸脱してしまった場合にもコンテスト運営を公正に行うことが難しくなってしまうという問題があった．

この問題は依然解決できておらず，また，今回は匿名加工データにどれだけの危険が含まれているかを安全に見積もることを重視することとした．そのため，今回のコンテストでは最大知識攻撃者モデルを採用する．

2. 準備

2.1 記法

行列 M の (i, j) 要素を $M[i, j]$ で参照する．M の第 i 行ベクトル，第 j 列ベクトルを，それぞれ $M[i, \cdot]$，$M[\cdot, j]$ で参照する．

集合 S の要素の数を $|S|$ と表記する．

3. コンテストの概要

本コンテストでは，表形式の購買履歴データを対象に，安全性の基準を満たしながら有用性の高いデータに加工する技術を競う．

表 1 クレンジング後の Online Retail Data Set [7]

属性名	値域，数
レコード数	397625
顧客数	4333
商品数	3663
購入日時	2010/12/1 8:26
	– 2011/12/9/12:50

3.1 加工対象のデータ

本コンテストで扱うデータは，各行が 1 つの商品の購入履歴を表す表形式のデータである．各行には，購入を行った**顧客**ごとに固有の番号 (**顧客** *ID*) と，購入内容が含まれる．

3.2 加工後のデータに求められる性質

加工後のデータは，安全性の基準を満たし，有用性が高いことが求められる．

3.2.1 安全性

本コンテストでは，加工後のデータに含まれるどの顧客 (**仮顧客**と呼ぶ) についても攻撃者が元の加工対象のデータ上のどの顧客に対応しているかを正しく言い当てること (**再識別**と呼ぶ) が十分に難しいことを安全性の基準として要求する．攻撃者は，加工対象のデータを知っているが (最大知識攻撃者)，加工に用いられたアルゴリズムは知らないものとする．

3.2.2 有用性

本コンテストでは，加工後のデータと元のデータの遠さを表す評価関数を定め，その評価値が小さい (加工後のデータと元のデータが近い) ほど有用性が高いものとする．

3.3 コンテストの流れ

コンテストは，**審判** J と，複数の**チーム** $P_1, P_2, \ldots$ により行われる．コンテストは**加工フェイズ**，**再識別フェイズ**，**評価フェイズ**の 3 つのフェイズにより構成され，順に実施される．

(1) 加工フェイズ: 各チーム P_ℓ は J から受け取った加工対象のデータを加工し，J に提出する．加工の際には，加工後のデータが安全性の基準を満たしながら元のデータに近い (有用性が高い) ことを目指す．

(2) 再識別フェイズ: 各チーム P_ℓ は他のチームが作成した加工後のデータを J から受け取り，再識別を行う．再識別により得た仮顧客と顧客の対応の推定結果を J に提出する．

(3) 評価フェイズ: J は各チームから受け取った加工後のデータと推定結果から各加工後のデータの安全性と有用性を評価する．

4. データセット

本コンテストでは, 加工対象のデータとして, 昨年度 [5] と同じく Online Retail Data Set [7] を用いる．Online Retail Data Set は英国のオンラインショッピングサイトにおける 2010 年からの 1 年間の購買履歴であり，541909 行 8 列のデータである．属性は InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country からなり，昨年度 [5] と同様にデータをクレンジングし，表 1 に示すような購買履歴データとした．さらに，表 2 に示す 5 属性を抜粋した．コンテストではこのデータセットから必要な量をサンプリングして課題データとして使用する．

表 2 本コンテストで使用する属性

[7] での属性名	本稿での呼称	元のデータとの違い
CustomerID	顧客 ID	
InvoiceDate	購入日	時刻を削除
StockCode	商品 ID	
UnitPrice	単価	
Quantity	数量	

表 3 トランザクション T の例

顧客 ID	購入日	商品 ID	単価	数量
14667	2011/11/14	21745	3.75	2
14974	2011/11/2	23392	2.08	1
17042	2011/4/6	22439	0.65	10
15039	2011/5/9	21974	1.45	3
14911	2011/9/30	22818	0.42	12

コンテストで課題データとして使用する上述の購買履歴データからサンプリングしたデータを**トランザクション** T と呼ぶことにする．T は，m 行 5 列の行列

$$T = \begin{pmatrix} T[1,1] & \ldots & T[1,5] \\ \vdots & \ddots & \vdots \\ T[m,1] & \ldots & T[m,5] \end{pmatrix}$$

である．T の各行は 1 種類の商品の購入履歴を，各列は表 2 の順番に対応する 5 つの属性 (顧客 ID，購入日，商品 ID，単価，数量) をそれぞれ表す．T の例を表 3 に示す．

T に含まれる顧客 ID の集合を $C = \{T[i,1] \mid i \in [1,m]\}$ とする．T に含まれる顧客の人数を $n = |C|$ とする．T の j 列目 $T[\cdot, j]$ の取りうる値の集合を D_j とする．D_1 は (9 桁以下の自然数，のような) 可能な顧客 ID の集合，D_2 は (2010 年 1 月 1 日から 2011 年 12 月 31 日，のような)1 日を単位とした年月日の (離散値の) 閉区間，D_3 は (大文字英字と数字の 1 文字以上の文字列，のような) 可能な商品 ID の集合，D_4 は (0.01 以上 99999.99 以下，のような)0.01 を単位とした実数の (離散値の) 閉区間，D_5 は (6 桁以下の自然数，のような)1 を単位とした自然数の (離散値の) 閉区間とする．

5. 加工

加工フェイズでは，各 P_ℓ は T を匿名加工したデータで

ある**公開加工トランザクション** A'_ℓ と，A'_ℓ を作成する際の中間データである**加工トランザクション** A_ℓ を作成する．

A_ℓ は T と同じ行数，列数 (m 行 5 列) の行列

$$A_\ell = \begin{pmatrix} A_\ell[1,1] & \dots & A_\ell[1,5] \\ \vdots & \ddots & \vdots \\ A_\ell[m,1] & \dots & A_\ell[m,5] \end{pmatrix}$$

であり，T の各要素 (または行) に**維持**，**削除**，**顧客** ID **の変更**，**一般化**のいずれかの加工を行って作成される．

A'_ℓ は A_ℓ から削除された行を除去し，残った行を辞書順で整列することで得られる m_ℓ 行 5 列の行列で，

$$A'_\ell = \begin{pmatrix} A'_\ell[1,1] & \dots & A'_\ell[1,5] \\ \vdots & \ddots & \vdots \\ A'_\ell[m_\ell,1] & \dots & A'_\ell[m_\ell,5] \end{pmatrix}$$

である．

A_ℓ は A'_ℓ の有用性および安全性を評価するために用いられる．A_ℓ や A'_ℓ に含まれる顧客 ID を特に**仮名**と呼び，A'_ℓ に含まれる仮名の集合を

$$C_\ell = \{A'_\ell[i,1] \mid i \in [1, m_\ell]\}$$

とする．

5.1 維持

T の要素の値を加工せずに対応する A_ℓ の要素の値としてもよい．

5.2 削除

T に対する加工では行単位の削除と顧客 ID 以外の要素単位の削除を行ってもよい．行単位の削除の場合，その行のすべての要素の値を，削除されたことを示す値 DEL に書き換える．要素単位の削除の場合，その要素の値を DEL に書き換える．特定の顧客に関する行をすべて削除しても構わない．

5.3 顧客 ID の変更

顧客 ID (T の 1 列目) は D_1 に含まれる任意の値 (仮名) へ変更してもよい．ただし，同一の顧客 ID は行単位の削除により削除される場合を除いてすべて同一の仮名に変更しなくてはならない ($\forall i \forall i', T[i,1] = T[i',1] \Rightarrow A_\ell[i,1] = \mathsf{DEL} \vee A_\ell[i',1] = \mathsf{DEL} \vee A_\ell[i,1] = A_\ell[i',1]$)．また，異なる顧客 ID を同一の仮名に変更することは禁止する ($\forall i \forall i', T[i,1] \neq T[i',1] \Rightarrow A_\ell[i,1] = \mathsf{DEL} \vee A_\ell[i',1] = \mathsf{DEL} \vee A_\ell[i,1] \neq A_\ell[i',1]$)．

5.4 一般化

顧客 ID 以外の属性の各要素は，一般化を行ってもよい．一般化は要素ごとに独立に行ってよい (いわゆる local recoding でよい)．一般化として認められる加工は，以下で規定するように，属性値の尺度水準により異なる．

5.4.1 名義尺度 (商品 ID) の一般化

商品 ID のように，属性値が名義尺度 (カテゴリ値) であるときは，その属性値の取り得る値の集合を D_j として，元の値を含む，D_j の部分集合への一般化をしてもよい．例えば，$a, b, e, g \in D_j$ とするとき，

$$a \to \{a, b, e, g\},$$

$$a \to \{a\}$$

のような一般化が認められる．一方，$c \notin D_j$ であるとき，

$$a \to \{b, e, g\} \quad (a \notin \{b, e, g\} \text{ である*}),$$

$$a \to \{a, c\} \quad (\{a, c\} \not\subset D_j \text{ である*})$$

のような一般化は禁止される．

5.4.2 順序尺度の一般化

属性値が順序尺度 (値の差には意味がないが，大小には意味がある) である属性は T には含まれないため，順序尺度の一般化は定義しない．

5.4.3 間隔尺度 (購入日)，比例尺度 (単価，数量) の一般化

属性値が購入日のように間隔尺度 (値の差には意味があるが比には意味がない) であるとき，または，単価や数量のように比例尺度 (値の比にも意味がある) であるときは，元の値を含む単一の閉区間への一般化をしてもよい．ただし，その属性値の取り得る値の集合を D_j とするとき，一般化後の閉区間は D_j の部分集合でなくてはならない．すなわち，購入日，単価，数量はそれぞれ D_2，D_4，D_5 上での (離散値の) 閉区間への一般化が認められる．例えば，

$$3 \to [3, 10],$$

$$3 \to [1, 10],$$

$$2.3 \to [1.2, 5.6],$$

$$2011/9/6 \to [2011/8/3, 2011/9/6],$$

$$2011/9/6 \to [2011/9/1, 2011/9/30]$$

のような一般化が認められる．例えば，購入日での

$$2011/9/6 \to 2011/9 \quad (2011/9 \text{ は閉区間はでない*})$$

数量での

$$3 \to [1.5, 10] \quad (1.5 \notin D_5 \text{ である*})$$

$$3 \to [1, 10], [13, 18] \quad (\text{単一の閉区間ではない*})$$

のような一般化は禁止される．

6. 再識別

再識別フェイズでは，各 P_ℓ は各 $A'_{\ell'}$ $(\ell' \neq \ell)$ に対して，$A'_{\ell'}$ に含まれる仮顧客の一部を選択して再識別を行う．再識別の結果，$A'_{\ell'}$ に含まれる仮名と T に含まれる顧客 ID の対応を推測した**推定写像** $f'_{\ell',\ell} : C'_{\ell',\ell} \to C$ が作成される．ここで，$C'_{\ell',\ell} \subset C_{\ell'}$ は $A'_{\ell'}$ に含まれる仮名の集合 $C_{\ell'}$ の一部であり，再識別を行う P_ℓ が自由に選択する．

昨年度までのコンテストとは異なり，すべての仮顧客に対して再識別を行う必要はない．また，$f'_{\ell',\ell}$ は単射でなくてよい．すなわち，複数の仮顧客に対して同一の顧客を推定してもよい．

7. 有用性評価

公開加工トランザクション A'_ℓ の**有用性評価値** $U(A'_\ell)$ は小さいほどよい (有用性が高い) 値であり，T と A_ℓ の対応する要素ごと (第 1 列の要素を除く) の**誤差** $\mathsf{Err}(T[i,j], A_\ell[i,j])$ の平均値として定義される．すなわち，

$$U(A'_\ell) := \frac{\sum_{i\in[1,m],j\in[2,5]} \mathsf{Err}(T[i,j], A_\ell[i,j])}{4m}$$

である．なお，削除された行は，その行のすべてのセルが削除されたものとして扱う．

7.1 $x \in D_j$ と $y \in D_j$ の誤差

7.1.1 $T[\cdot, j]$ が名義尺度のとき

$T[\cdot, j]$ が名義尺度のとき，$x \in D_j$ と $y \in D_j$ の誤差を以下のように定義する．

$$\mathsf{Err}(x, y) := \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{otherwise.} \end{cases}$$

7.1.2 $T[\cdot, j]$ が順序尺度のとき

本コンテストでは順序尺度は扱わないため，定義しない．

7.1.3 $T[\cdot, j]$ が間隔尺度のとき

$T[\cdot, j]$ が間隔尺度のとき，D_j から任意の値 d を選び，D_j の各値 $x \in D_j$ を $x - d$ で置き換え，$T[\cdot, j]$ を比例尺度とみなして $\mathsf{Err}(x, y)$ を計算する．

7.1.4 $T[\cdot, j]$ が比例尺度のとき

$T[\cdot, j]$ が比例尺度のとき，$T[\cdot, j]$ の標準偏差を σ_j として，$x \in D_j$ と $y \in D_j$ の誤差を以下のように定義する．

$$\mathsf{Err}(x, y) := \frac{|x - y|}{\sigma_j}.$$

これは標準偏差で割ることによる標準化 [6] を行った値同士の絶対誤差と等しい．

7.2 $x \in D_j$ と DEL の誤差

2 つの値のうち一方が削除されているときの誤差を定義する．7.1.4 節のように属性ごとに標準化を行なっているため，$x \in D_j$ と DEL の誤差は属性によらず一定の値とし，

$$\mathsf{Err}(x, \mathsf{DEL}) := 1$$

とする．

7.3 $x \in D_j$ と $Y \subset D_j$ の誤差

2 つの値のうち一方が一般化されているときの誤差を定義する．確率変数 Y' の分布を Y 上の一様分布として，$x \in D_j$ と $Y \subset D_j$ の誤差を $\mathsf{Err}(x, Y')$ の期待値と定義する．すなわち，

$$\mathsf{Err}(x, Y) := E[\mathsf{Err}(x, Y')].$$

7.3.1 Y が集合の場合の計算例

$j = 3$，すなわち属性が商品 ID の場合が該当する．

$$\Pr(Y' = y) = \begin{cases} \frac{1}{|Y|} & \text{if } y \in Y, \\ 0 & \text{otherwise} \end{cases}$$

であるので，加工のルールにより $x \in Y$ であることに注意すると，

$$\begin{aligned} \mathsf{Err}(x, Y) &= \sum_{y\in Y} \mathsf{Err}(x, y) \Pr(Y' = y) \\ &= \frac{1}{|Y|} \sum_{y\in Y} \mathsf{Err}(x, y) \\ &= \frac{|Y| - 1}{|Y|} \\ &= 1 - \frac{1}{|Y|}. \end{aligned}$$

7.3.2 Y が要素間の間隔が一定の閉区間の場合の計算例

$j = 2, 4, 5$，すなわち属性が購入日，単価，数量の場合が該当する．　$Y = [a, b]$，Y の k 番目に小さい要素を y_k，$\delta = y_2 - y_1$ とすると，$a = y_1, b = y_{|Y|}$ で，$y_k = a + (k-1)\delta = b - (n-k)\delta$ が成り立つ．

$$\Pr(Y' = y) = \begin{cases} \frac{1}{|Y|} & \text{if } y \in Y, \\ 0 & \text{otherwise} \end{cases}$$

であるので，加工のルールにより $x \in Y$ であることに注意すると，$x = y_l$ として，

$$
\begin{aligned}
\mathsf{Err}(x,Y) &= \sum_{y\in Y}\mathsf{Err}(x,y)\Pr(Y'=y)\\
&= \frac{1}{|Y|}\sum_{i=1}^{|Y|}\mathsf{Err}(x,y_i)\\
&= \frac{1}{|Y|}\sum_{i=1}^{|Y|}\frac{|x-y_i|}{\sigma_j}\\
&= \frac{1}{|Y|\sigma_j}(\sum_{i=1}^{l}|x-y_i|+\sum_{i=l}^{|Y|}|x-y_i|)\\
&= \frac{\delta((l-1)^2+(|Y|-l)^2+(|Y|-1))}{2|Y|\sigma_j}\\
&= \frac{(x-a)^2+(b-x)^2+(|Y|-1)\delta^2}{2|Y|\sigma_j\delta}\\
&= \frac{(a-x)^2+(b-x)^2+(b-a)\delta}{2(b-a+\delta)\sigma_j}.
\end{aligned}
$$

8. 安全性評価

公開加工トランザクション A'_ℓ の**安全性評価値** $S(A'_\ell)$ は小さいほどよい (安全性が高い) 値であり，A'_ℓ に対する推定写像の集合を**推定写像集合** $F'_\ell := \{f'_{\ell,\ell'} \mid (P_{\ell'}\text{ が }A'_\ell\text{ に対する再識別を実施})\}$ として，T，A_ℓ，F'_ℓ により以下のように計算される．

$$
S(A'_\ell) := \begin{cases} 1 & \text{if } \exists f'_{\ell,\ell'}\in F'_\ell, r(n'_{\ell,\ell'}) \le \mathsf{Suc}(f_\ell, f'_{\ell,\ell'}),\\ 0 & \text{otherwise.}\end{cases}
$$

ここで，$r:[0,n]\to[0,n+1]$ は J があらかじめ定める関数であり，再識別を試みられた人数 $n'\in[0,n]$ に対し，安全性の基準を満たしていないと判断する再識別成功人数の下限 $r(n')$ を与える．$\mathsf{Suc}(f_\ell, f'_{\ell,\ell'})$ は $f'_{\ell,\ell'}$ による再識別の成功数を表し，

$$
\mathsf{Suc}(f_\ell, f'_{\ell,\ell'}) := |\{q \mid f'_{\ell,\ell'}(q) = f_\ell(q)\}|
$$

により定義される．また，$f_\ell : C_\ell \to C$ は A'_ℓ に含まれる仮名と T に含まれる顧客 ID の対応を表す写像であり，

$$
\forall i\in[1,m], A_\ell[i,1] = \mathsf{DEL} \vee T[i,1] = f_\ell(A_\ell[i,1])
$$

を満たす．f_ℓ を**正解写像**と呼ぶ．$n'_{\ell,\ell'}$ は再識別を実施した人数であり，$n'_{\ell,\ell'} = |C'_{\ell',\ell}|$ である．

$S(A'_\ell)$ は少なくとも一つの $f'_{\ell,\ell'}\in F'_\ell$ について，$r(n'_{\ell,\ell'}) \le \mathsf{Suc}(f_\ell, f'_{\ell,\ell'})$ であった場合に 1，それ以外の場合に 0 となる．$S(A'_\ell)=1$ であった場合，A'_ℓ は安全性の基準を満たしていないと判断される．

9. コンテストの実装

本年のコンテストの実装について述べる．コンテストは 3 節から 8 節までで述べた設計の通り加工フェイズ，再識別フェイズ，評価フェイズの 3 つのフェイズにより構成され，審判 J と参加者 $P_1, P_2, \ldots$ により実行される．ただし，運用の効率化のために一部のデータの形式や作成者が設計と異なって実装されている．本節では設計と実装の差異を指摘しながらコンテストの実装を説明する．

9.1 加工フェイズ

加工フェイズの実装は，5 節で述べた設計から変更されている．5 節では P_ℓ が A_ℓ と A'_ℓ を作成するとしたが，実装では P_ℓ は A_ℓ のみを作成し，A'_ℓ は A_ℓ を受け取った後に J が作成することとした．これにより，設計通りならば P_ℓ から J には A'_ℓ と A_ℓ の両方を送る必要があったが，変更により P_ℓ から J に A'_ℓ を送る必要がなくなった．A'_ℓ は A_ℓ から決定的に計算可能であるので，この変更によりコンテストの結果が変わることはない．

加工フェイズの実装は以下の通りである．

(1) J は T を公開する
(2) 各 P_ℓ は公開された T を加工して A_ℓ を作成し，J に提出する
(3) J は $A_1, A_2\ldots$ からそれぞれ $A'_1, A'_2, \ldots$ を作成する

9.2 再識別フェイズ

再識別フェイズの実装では，各チーム P_ℓ の提出する推定写像 $f'_{\ell,\ell'}$ の提出方法を明確化する．

P_ℓ は $A'_{\ell'}$ に対する再識別結果である推定写像 $f'_{\ell',\ell}$ を J に提出する際に，$f'_{\ell',\ell}$ の代わりに $f'_{\ell',\ell}$ を表す $n_{\ell',\ell} = |C'_{\ell',\ell}|$ 行 2 列の行列である**推定対応表** $F'_{\ell',\ell}$ を提出する．$F'_{\ell',\ell}$ は 1 列目に仮名，2 列目に顧客 ID を含み，$C'_{\ell',\ell}$ に含まれる仮名を $c_1,\ldots,c_{n_{\ell',\ell}}$ として，

$$
F'_{\ell',\ell} = \begin{pmatrix} c_1 & f'_{\ell',\ell}(c_1)\\ \vdots & \vdots\\ c_{n_{\ell',\ell}} & f'_{\ell',\ell}(c_{n_{\ell',\ell}})\end{pmatrix}
$$

である．

再識別フェイズの実装は以下の通りである．

(1) J は各 $A'_1, A'_2, \ldots$ を公開する
(2) 各 P_ℓ は公開された $A'_{\ell'}$ に再識別を行い，$f'_{\ell,\ell'}$ を作る
(3) 各 P_ℓ は $f'_{\ell,\ell'}$ から $F'_{\ell,\ell'}$ を作成し，J に提出する

9.3 評価フェイズ

評価フェイズの実装は，以下の通り設計通りである．

(1) J は各 A'_ℓ に対し，T と A_ℓ から $U(A'_\ell)$ を計算し，公開する
(2) J は各 A'_ℓ に対し，T，A_ℓ，F'_ℓ から $S(A'_\ell)$ を計算し，公開する

10. まとめ

本稿では，購買履歴データを用いた匿名加工再識別コンテストの基本ルールと安全性，有用性の提案を行った．

参考文献

[1] 個人情報保護委員会：個人情報の保護に関する法律についてのガイドライン (匿名加工情報編) (2016).

[2] Emam, K. and Alvarez, C.: A critical appraisal of the article 29 working party opinion 05/2014 on data anonymisation techniques, *Int Data Priv Law*, Vol. 5, pp. 73–87 (2015).

[3] 菊池浩明，山口高康，濱田浩気，山岡裕司，小栗秀暢，佐久間淳：匿名加工・再識別コンテスト Ice & Fire の設計，Vol. 2015, No. 3, pp. 363–370 (2015).

[4] 菊池浩明，小栗秀暢，野島　良，濱田浩気，村上隆夫，山岡裕司，山口高康，渡辺知恵美：PWSCUP: 履歴データを安全に匿名加工せよ， Vol. 2016, No. 2, pp. 271–278 (2016).

[5] 菊池浩明，小栗秀暢，中川裕志，野島　良，波多野卓磨，濱田浩気，村上隆夫，門田将徳，山岡裕司，山田　明，渡辺知恵美：PWSCUP2017: 長期間の履歴データの再識別リスクを競う， Vol. 2017, No. 2 (2017).

[6] Domingo-Ferrer, J. and Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation, *Data Mining and Knowledge Discovery*, Vol. 11, No. 2, pp. 195–212 (2005).

[7] UCI Machine Learning Repository: Online Retail Data Set, , available from ⟨https://archive.ics.uci.edu/ml/datasets/Online+Retail⟩ (accessed 15 Aug 2018).