PWSCUP: 履歴データを安全に匿名加工せよ

菊池 浩明[1]　小栗 秀暢[2]　野島 良[3]　濱田 浩気[4]　村上 隆夫[5]　山岡 裕司[7]　山口 高康[6]
渡辺 知恵美[8]

概要：個人情報保護法の改正により，顧客に関る履歴データを匿名加工することで第三者提供できる法的な枠組みが整った．しかしながら，その安全性については，特定の個人を識別できないように個人情報保護委員会で定める基準に従って行うことが決まっているが，具体的な加工方法や再識別に対するリスク評価の基準については各分野の認定個人情報保護団体に委ねられており，信頼できる安全性を担保できるのかどうかが懸念されている．そこで，我々は共通のデータセットを用いて匿名加工を行い，加工されたデータの再識別の性能を競うコンテストを企画し，有用性が高く安全な匿名加工情報の技術開発を試みる．対象とするデータは，顧客情報を管理するマスターデータと個々の購買履歴を管理するトランザクションデータの 2 種類からなり，それらを仮名 ID により結びつける技術を競う．本稿では，このコンテストの基本定義，有用性の評価方法，安全性を定量的に定めるためのサンプルとなる再識別アルゴリズムなどについて述べる．

キーワード：個人情報保護，匿名加工，匿名性

PWSCUP Competition: De-identify Transaction Data Securely

Hiroaki Kikuchi[1]　Hidenobu Oguri[2]　Ryo Nojima[3]　Koki Hamada[4]　Takao Murakami[5]
Yuji Yamaoka[7]　Takayasu Yamaguchi[6]　Chiemi Watababe[8]

Abstract: Data anonymization is ready to go before the big-data business runs successfully while preserving privacy of personal information. While, it is not trivial to choose the best algorithm to make the given data anonymized to be secure for a given particular purpose. To access the risk to be compromised accurately, the data needs to balance the utility and the security. Hence, with a public online retail dataset, , we propose a new competition for best anonymization and re-identification algorithm. Our dataset consists of a customer dataset and a transaction dataset and these datasets are linked with pseudonyms, a random number assinged for each customer identities. The paper addresses the aim of the competition, the target dataset, sample algorithms, utility and security metrics.

1　明治大学, Meiji University
2　ニフティ (株), NIFTY Corporation
3　国立研究開発法人 情報通信研究機構, NICT
4　NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories
5　国立研究開発法人 産業技術総合研究所, AIST
6　(株)NTT ドコモ先進技術研究所
NTT DOCOMO, Inc.
7　株式会社富士通研究所
FUJITSU LABORATORIES LTD.
8　筑波大学
University of Tsukuba

1. はじめに

2016 年個人情報保護法が成立し，2017 年 1 月から全面施行する．この改正で，「匿名加工情報」が新設された．匿名加工情報とは，「(定められた措置を講じて）特定の個人を識別することができないように個人情報を加工して得られる個人に関する情報であって，当該個人情報を復元することができないようにしたもの（第二条 9 項)」と定められている．2016 年 8 月 2 日には，個人情報保護委員会が個人情報保護規定案を発表し，パブリックコメントを募集して

いる．その中で，匿名加工情報の作成の方法に関する基準（第十九条）は，

(1) 個人を識別する記述等の全部又は一部を削除する．
(2) 個人識別符号の全部を削除する．
(3) .. 個人情報を連結する符号を削除する．
(4) 特異な記述等を削除する．
(5) ... 他の個人情報に含まれる記述等の差異その他の当該個人情報データベース等の性質を勘案して，... 適切な措置を講じること．

と提案されているのみである．今後，業種ごとに定められた認定個人情報保護団体が，消費者や関係者の意見を聴いて，その対象事業に適した個人情報保護指針を整備していく予定である．しかしながら，再識別のリスクの評価や加工の方法には，高度な専門性やその分野に特化した背景知識の両方が必要であり，誰に対しても納得できる信頼できる基準には至っていない．

そこで我々は，2015 年に，匿名加工技術の開発と再識別に対する公平な安全性評価手法の確立を目的として，教育機関などの演習用として独立行政法人統計センターが作成した疑似ミクロデータを用いて，匿名加工と再識別のコンテストを実施した．総計 80 名を超える 17 のチームが集まり，熱心に加工技術と再識別の技巧が競われる結果となった．

ただし，ここで用いたミクロデータは，一個人がその年における年間の消費額が記述されたものであり，個人に対する複数の活動履歴の記録された形式ではなかった．一般的なサービス事業者が用いるデータベースの利用形態は，ほとんどが静的な顧客データベースとイベントの発生時刻を含む履歴データベースの組み合わせである．例えば，経済産業省の匿名加工情報作成マニュアル [4] では，代表的なユースケースとして

- 家庭における電力利用データ
- クレジットカードの購買データ
- 鉄道の乗降履歴データ

の 3 種類の履歴データを取り上げている．

本稿は，本コンテストで用いる技術の基本定義と有用性と安全性の評価指標を提案する．コンテストの設計においては，考慮した不正行為と対象としたリスクについて述べ，サンプルとするいくつかの再識別アルゴリズムを定義する．

2. 匿名加工コンテスト

2.1 目的

本コンテストは次を目的として実施する．

- 安全で有用性の高い匿名加工技術の開発を促進すること
- 再識別のリスクを正しく評価すること

匿名加工データに関わるリスクには，1) 匿名加工データからそのデータに対応する個人が識別されること（レコード再識別），2) 匿名加工データから，加工された属性の値が推定される（属性推定），3) 匿名加工データを他のデータと照合される，などがあるが，本コンテストでは，1) のリスクを低減することに焦点を定める．

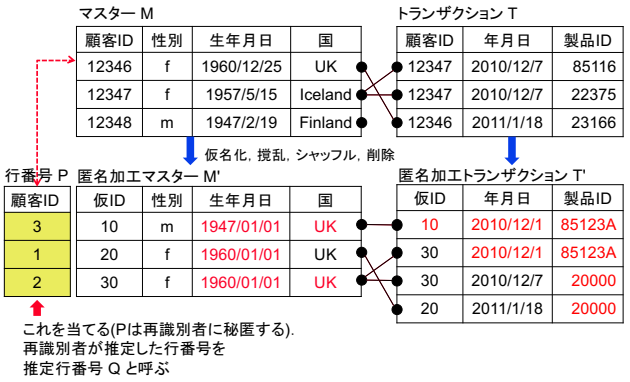

図 1　コンテスト概要

2.2 概要

図 1 に本コンテストの概要を示す．本コンテストでは，登録された個人情報である顧客マスターデータ M と顧客が行った購買取引（トランザクション）の履歴を表す購買履歴データ T を対象とする．M と T の間は，顧客 ID（または仮 ID）により結び付けられている．例えば，図の顧客 12347 は，12 月 7 日に 2 つの商品を購入している．この関係を維持したまま，M と T を匿名加工したデータをそれぞれ M', T' とする．

M' は M に対して顧客 ID が仮 ID に振り替えられ，データの値の撹乱，レコード（行）のシャッフル，特異な顧客などの削除などの処理が施されて加工される．図の例では，国名が全て UK に変更され，生年月日は月日が 1 月 1 日に変更され，更に，3 行目の男性が 1 行目に置換されている．この関係を行番号 P で表す．同様に，T' も購買年月日や購買した製品が変更され，時には図の仮 ID10 の様に，元の T に存在しない架空の購買履歴が追加される．

再識別者は，元の M と T の関係をヒントにして，加工された M' と T' を解析し，行番号 P を推測した Q を提出し，その精度を評価する．匿名加工者は，大きく変更して強く加工すれば，推定される再識別リスクを下げることができるが，匿名加工データの特徴は元のデータとは異なってしまい，その有用性が下がる．この安全性と有用性のトレードオフの中で，最適な加工方法を競う．

2.3 購買履歴データセット

Online Retail Data Set は，英国に現存する無店舗型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る．UCI Machine Learning Repository[*1]

[*1] https://archive.ics.uci.edu/ml/datasets/Online+Retail

表 1　取引データセット T の統計量

項目	値域, 値数
レコード数	$m = 397,625$
顧客数	$n = 4,333$
伝票数	18,513
商品数	3,663
国数	36
購入日時	2010/12/1 8:26
	– 2011/12/9 12:50

表 2　Online Retail Dataset, 購買属性

属性名	記述	T の利用順序
InvoiceNo	伝票 ID	2
StockCode	商品 ID	5
Description	商品名	-
Quantity	購入数	7
InvoiceDate	購入日時	3,4
UnitPrice	単価	6
CustomerID	顧客 ID	1
Country	国名	-

表 3　顧客マスターデータベース M

項目名	記述	生成方法
CustomerID	顧客 ID	取引データから抽出
Sex	性別 (f, m)	男女比を 1:5 となる確率で合成
Birthday	誕生日	年は 1930 年から 1989 年までの範囲で正規乱数（平均 1960，標準偏差 10）で，月日は一様乱数で合成
Country	国名	取引データの最後の取引履歴から抽出

から公開されている．Chen らは，このデータを顧客の市場分析のユースケースとして用い，顧客のクラスタリングとその購買頻度や購入金額などの分析を行った [3]．表 2 に，属性の記述と本コンテストでの利用した属性を示す．

不備のあるレコード（取消伝票，不良債権調整，単価が 0.01 未満，顧客 ID が空，国名が “Unspecified” ，商品 ID が “BANK CHARGES”) をクレンジングして，表 1 に示す約 40 万レコードの取引（トランザクション）データとした．また，顧客マスターデータベースは，表 3 に示す方法により我々が合成した．

3. 提案コンテスト

3.1 基本定義

マスターデータベース $\boldsymbol{M}$ は，n 人の顧客の情報を格納したレコードの集合 $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$ である．ここで，i 番目の顧客レコードは，4 つの属性を持つタプル $\boldsymbol{c}_i = (c_i^1, c_i^2, c_i^3, c_i^4)$ であり，それぞれ，表 3 に示される (顧客 ID, 性別, 誕生日, 国名) を表す．

取引データベース $\boldsymbol{T}$ は，m 個の購買履歴から成る集合 $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_m\}$ である．ここで，j 番目の購買履歴は，表 2 の順番に対応する 7 つの属性を表すタプル $\boldsymbol{t}_j = (t_j^1, \ldots, t_j^7)$ であり，(顧客 ID, 伝票 ID, 購買日，購買時，商品 ID, 単価，数量) を表す．

(例 3.1)　顧客マスター $\boldsymbol{M}$ と購買履歴 $\boldsymbol{T}$ の例を表 4,5 にそれぞれ示す．

表 4　顧客マスターデータ $\boldsymbol{M}$ の例

c^1 顧客 ID	c^2 性別	c^3 生年月日	c^4 国籍
12360	m	1976/2/24	Austria
12361	f	1954/2/14	Belgium
12362	f	1963/12/26	Belgium
12364	f	1960/9/16	Belgium

表 5　購買履歴データ $\boldsymbol{T}$ の例

t^1 顧 ID	t^2 伝票	t^3 購買日	t^4 時刻	t^5 商品	t^6 単価	t^7 数
12362	544203	2011/2/17	10:30	21913	3.75	4
12362	544203	2011/2/17	10:30	22431	1.95	6
12361	545017	2011/2/25	13:51	22630	1.95	12
12361	545017	2011/2/25	13:51	22555	1.65	12
12362	551346	2011/4/28	9:12	21866	1.25	12
12362	551346	2011/4/28	9:12	20750	7.95	2
12362	551346	2011/4/28	9:12	22908	0.85	12
12360	554132	2011/5/23	9:43	21094	0.85	12
12360	554132	2011/5/23	9:43	23007	14.95	6

3.2 匿名加工の定義

$\boldsymbol{M}$ には顧客が対応しており，顧客 ID が分かれば，（内部のデータベースを容易に照合することで）その特定の顧客は識別されたと考える．そこで，$\boldsymbol{M}$ と $\boldsymbol{T}$ のレコードの直接識別子や特異な値を削除したり，値を変更したりして，$\boldsymbol{M}'$ と $\boldsymbol{T}'$ に加工し，$\boldsymbol{M}'$ と元の $\boldsymbol{M}$ の関係が推測できないようにすることを（このコンテストにおける）**匿名加工**と呼ぶ．ただし，$\boldsymbol{T}$ と $\boldsymbol{T}'$ の関係は問わない．

ここで，匿名加工された $\boldsymbol{M}'$ のレコードは，写像

$$p : \{1, \ldots, n'\} \to \{1, \ldots, n\}$$

により任意に置換されている．すなわち，匿名加工マスターは $\boldsymbol{M}' = \{\boldsymbol{c}_{p(1)}, \ldots, \boldsymbol{c}_{p(n')}\}$ となる．なお，トップコーディングなどにより特異なレコード（例えば，100 歳以上の顧客）を削除することがあるので，$n \geq n'$ となることがあるが，架空の顧客を追加することは（本コンテストでは）考えないので，$n < n'$ はない．

この置換を表すために，レコードの行番号 $I = (1, \ldots, n)$ を導入する．置換は，匿名加工マスターの行番号 $P = (p(1), \ldots, p(n'))$ で表される [*2]．

[*2]　関数 p は加工で用いた置換の逆関数で定義する

3.3 仮名化と履歴の匿名加工

属性 c^1 は単体で特定の顧客を識別するので，匿名加工にするためには削除するか，仮名化する必要がある．**仮名化**とは，「復元することのできる規則性を有しない方法により直接識別子を他の記述等で置き換える操作」であり，置き換えられたものを仮 ID と呼ぶ．

顧客 ID(識別子)$c^1_{p^{-1}(i)}$ から，仮 IDc'^1_i を生成するには，例えば，一方向性ハッシュ関数 H を用いて，

$$c'^1_i = H(key + c^1_{p^{-1}(i)}) = h(c^1_{p^{-1}(i)})$$

と定める方法（鍵付きハッシュ）などが挙げられる．ここで，key は十分な定義域から選んだ秘密の乱数であり，$+$ は文字列の連結，$h()$ は key を暗黙に含んだ簡易表記とする．ただし，本コンテストのデータセットの様に長期間に渡って同一顧客に同一の仮 ID を割り当てると再識別されるリスクがあがる．従って，適切な期間や回数で仮 ID を更新したり，他の顧客の仮 ID と交換したりするなどの**仮名制御**を行う．

履歴データにおいても，元のレコードと匿名加工されたレコードの関係を表すために，写像

$$f : \{1, \ldots, m'\} \to \{1, \ldots, m\}$$

を導入する．ただし，この置換が分かっても，顧客 ID が直接判明するわけではないので，本コンテストではこれを推定することは問わないこととする．顧客マスターと同様に，履歴についても行番号データ $J = (1, \ldots, m)$ の記法を用いる．

(例 3.2)　例 1 のマスターを匿名加工した例を表 6 に示す．対応のため，1,2 列は元のマスター $\boldsymbol{M}$, 3 列目 c'^1 が匿名加工されたマスターの仮 ID であり，鍵付きハッシュにより生成している．4 列目 P が匿名加工マスターの置換を表す行番号である．

表 6　マスターの匿名加工と行番号データの例

I	c^1	c'^1	P
1	12360	$h(12361)$	$2 = p(1)$
2	12361	$h(12364)$	$4 = p(2)$
3	12362	$h(12360)$	$1 = p(3)$
4	12364	–	–
$\boldsymbol{M}$		$\boldsymbol{M'}$	

同様に，匿名加工した履歴 $\boldsymbol{T'}$ の例を表 7 に示す．3 列目が仮 IDt'^1, 4 列目の伝票 ID が 1(存在しない伝票 ID）により削除され，5 列目の購買日 t'^3 の 2 行目が書き換えられている．これにより，$\boldsymbol{T'}$ の順序が変わるため，匿名加工された履歴の行番号データ J' の 2 行目と 3 行目が入れ替わっている．

表 7　購買履歴データの匿名加工 $\boldsymbol{T'}$ の例

J	J'	t'^1	t'^2	t'^3	t'^4
1	1	$h(12362)$	1	2011/2/17	21913
2	3	$h(12362)$	1	2011/2/**25**	22431
3	2	$h(12361)$	1	2011/2/25	22630
4	4	$h(12361)$	1	2011/2/25	22555

3.4 安全性

匿名加工データの再識別リスクは，匿名加工情報の提供先の参照可能な情報や能力，および，再識別によって得られる価値によって決まる．しかし，これらを包括的に列挙するのは困難である．そこで，Domingo-Ferrer らによって提案されている最大知識攻撃者モデル [7] の下で，本コンテストでは P 以外のすべて，すなわち，$\boldsymbol{M'}, \boldsymbol{T'}$ に加えて，それらの元のデータ $\boldsymbol{M}, \boldsymbol{T}$ を用いて，置換 p を推定することとする．

匿名加工の安全性は，これらの再識別アルゴリズムによって，正しく識別されたマスターのレコード数によって評価する．匿名加工アルゴリズム E が出力したマスターの置換行番号 $P = (p(1), \ldots, p(n'))$ とある再識別アルゴリズムが出力した推定行番号 $Q = (q(1), \ldots, q(n'))$ が与えられたとき，**再識別率**を，

$$\text{re-id}^E(P, Q) = \frac{|\{i \in \{1, \ldots, n'\} | p(i) = q(i)\}|}{n'}$$

と定める．ここで，分母が n でなくて n' であることに注意しよう．

複数の再識別アルゴリズム $E_1, E_2, \ldots$ が与えられたとき，総合した再識別率はそれらの最大値で定義する．

3.5 有用性

匿名加工データの有用性は，そのユースケースに依存するところが大きい．しかし，本コンテストでは次の典型的なユースケースを想定し，それらを総合して加工データの有用性を評価する．

(1) RFM 分析による優良顧客の抽出．
購買履歴から，最終購買日 (Recency), 購買頻度 (Frequency), 累計購買金額 (Monetary) の 3 つの指標を計算し，優良顧客を選別する．Chen らは，Online Retail データを用いて RFM 分析を実施し， 顧客を 5 つのクラスタに分類して，それぞれに最適な広報戦略を選ぶことを提案している [3].

(2) バスケット分析と相関ルール抽出．
購買履歴には，伝票 ID(Invoice) が記録されており，顧客が同時刻に同時にバスケットに入れて購入した複数の商品が分かる．Apriori アルゴリズムをここに適用すると，「粉ミルクとオムツを購入する人は，ビールも購入する」の様な頻度の高いアイテム間の相関規則を抽出することができる．

(3) クロス集計.

特定の商品を購入している顧客の年齢分布や性別の分布が分かれば，それらを考慮して満足度の高い商品推薦を実現できる．国別の特徴を配慮してウェブページの対応言語を選定することにも活用できる．

そこで，これらを配慮して，次のような有用性指標を定義する．特に，匿名加工と置換行番号を全く不整合にする山岡匿名化 [2] に対して，有用性を損なうような指標が必要である．

3.5.1 クロス集計 cmae

マスターデータの属性値により履歴データを分割し，それぞれの部分集合についての集計値を求めて，オリジナルと匿名加工の平均絶対誤差 Mean Absolute Error (MAE) を求める．例えば，表 4,5 において，集計値として履歴データの単価 (t^6) の平均値を求める場合を考える．まず，顧客 $\boldsymbol{c}$ が条件 s を満たすことを述語 $s(\boldsymbol{c}_i)$ で表し，条件 s を満たす顧客で制約される履歴データの部分集合（制限）を

$$\boldsymbol{T}|_s = \{\boldsymbol{t}_j \in \boldsymbol{T} \mid \boldsymbol{c}_i \in \boldsymbol{M}, t^1_j = c^1_i, s(\boldsymbol{c}_i)\},$$

$$\boldsymbol{T}'|_s = \{\boldsymbol{t}'_j \in \boldsymbol{T}' \mid \boldsymbol{c}'_i \in \boldsymbol{M}', t'^1_j = c'^1_i, s(\boldsymbol{c}'_i)\}$$

と定める．例えば，顧客 $\boldsymbol{c}$ を性別により分割する条件の集合

$$C = \{c^2 = \mathrm{f}, c^2 = \mathrm{m}\}$$

の各条件により，表 5 の履歴データは

$$\begin{aligned}\boldsymbol{T}|_{c^2=\mathrm{f}} &= \{\boldsymbol{t}_j \in \boldsymbol{T} \mid \boldsymbol{c}_i \in \boldsymbol{M}, t^1_j = c^1_i, c^2_i = \mathrm{f}\}\\ &= \{\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3, \boldsymbol{t}_4, \boldsymbol{t}_5, \boldsymbol{t}_6, \boldsymbol{t}_7\},\\ \boldsymbol{T}|_{c^2=\mathrm{m}} &= \{\boldsymbol{t}_8, \boldsymbol{t}_9\}\end{aligned}$$

の二つに分割される．この各々について $\boldsymbol{T}'$ で同様に計算した値との差の平均を計算して

$$\mathrm{cmae}(\boldsymbol{M}, \boldsymbol{M}', \boldsymbol{T}, \boldsymbol{T}') = \sum_{s \in C} |\mu_{\mathrm{up}}(\boldsymbol{T}|_s) - \mu_{\mathrm{up}}(\boldsymbol{T}'|_s)|/|C|$$

と定める．ここで，$\mu_{\mathrm{up}}(\boldsymbol{T}|_s)$ は，$\boldsymbol{T}|_s$ における平均単価

$$\mu_{\mathrm{up}}(\boldsymbol{T}|_s) = \frac{\sum_{\boldsymbol{t}_j \in \boldsymbol{T}|_s} t^6_j t^7_j}{\sum_{\boldsymbol{t}_j \in \boldsymbol{T}|_s} t^7_j}$$

であり，

$$\mu_{\mathrm{up}}(\emptyset) = 0$$

とする．

(例 3.3) 表 4, 5, 6, 7 を，性別で制限した時は，$\mathrm{cmae}(\boldsymbol{M}, \boldsymbol{M}', \boldsymbol{T}, \boldsymbol{T}') = (|\mu_{\mathrm{up}}(\boldsymbol{T}|_{c^2=\mathrm{m}}) - \mu_{\mathrm{up}}(\boldsymbol{T}'|_{c^2=\mathrm{m}})| + |\mu_{\mathrm{up}}(\boldsymbol{T}|_{c^2=\mathrm{f}}) - \mu_{\mathrm{up}}(\boldsymbol{T}'|_{c^2=\mathrm{f}})|)/2 = (|(0.85 \times 12 + 14.95 \times 6)/(12+6) - 0| + |(3.75 \times 4 + 1.95 \times 6 + 1.95 \times 12 + 1.65 \times 12 + 1.25 \times 12 + 7.95 \times 2 + 0.85 \times 12)/(4+6+12+12+12+2+12) - (3.75 \times 4 + 1.95 \times 6 + 1.95 \times 12 + 1.65 \times 12)/(4+6+12+12)|)/2 = (|5.55 - 0| + |1.85 - 2.0558\cdots|)/2 = 2.877\cdots$.

一方，山岡匿名化をした時は，$\boldsymbol{M}$ と $\boldsymbol{T}'$ とが整合しないので，これを検出する次のクロス集計 2 を導入する．

$$\mathrm{cmae2}(\boldsymbol{M}, \boldsymbol{M}', p, \boldsymbol{T}, \boldsymbol{T}') = \sum_{s \in C} |\mu_{\mathrm{up}}(\boldsymbol{T}|_s) - \mu_{\mathrm{up}}(\boldsymbol{T}'p|_s)|/|C|,$$

ただし，ここで，

$$\boldsymbol{T}'p|_s = \{\boldsymbol{t}'_j \in \boldsymbol{T}' \mid \boldsymbol{c}'_i \in \boldsymbol{M}', t'^1_j = c'^1_i, s(\boldsymbol{c}_{p(i)})\}$$

である (性別での制限を $\boldsymbol{M}'$ ではなくて，元の $\boldsymbol{M}$ の上で行う)．

なお，ここでは性別についての条件の集合 C による分割の例でクロス集計を説明したが，コンテストでは $\boldsymbol{M}$ に含まれる性別と国別の組み合わせによる条件の集合

$$C' = \{c^2 = x \wedge c^4 = y \mid \boldsymbol{c}_i \in \boldsymbol{M}, c^2_i = x, c^4_i = y\}$$

による分割を用いる．

3.5.2 指標 subset

匿名加工された顧客マスター $\boldsymbol{M}'$ の大きさ 10 の部分集合 $\boldsymbol{X}' \subset \boldsymbol{M}'$ と，$\boldsymbol{X} = \{\boldsymbol{c}_{p(i)} \mid \boldsymbol{c}'_i \in \boldsymbol{X}'\}$ として匿名加工マスターの行番号 P により $\boldsymbol{X}'$ に対応付けられる $\boldsymbol{X} \subset \boldsymbol{M}$，$\boldsymbol{T}$ 上の連続した 30 日間 D に対し，期間 D における総購入額の $\boldsymbol{X}$ についての平均値 $\mu_{\mathrm{tp}}(\boldsymbol{X}, D, \boldsymbol{T})$ と期間 D における総購入額の $\boldsymbol{X}'$ についての平均値 $\mu_{\mathrm{tp}}(\boldsymbol{X}', D, \boldsymbol{T}')$ の差の最大値

$$Y(\boldsymbol{M}, \boldsymbol{M}', \boldsymbol{T}, \boldsymbol{T}', p) = \max_{X' \subset \boldsymbol{M}', D} (|\mu_{\mathrm{tp}}(\boldsymbol{X}', D, \boldsymbol{T}') - \mu_{\mathrm{tp}}(\boldsymbol{X}, D, \boldsymbol{T})|),$$

ただし，

$$\mu_{\mathrm{tp}}(\boldsymbol{X}', D, \boldsymbol{T}') = \sum_{i \text{ s.t. } t'^1_i = c'^1_j, t'^3_i \in D, \boldsymbol{c}'_j \in \boldsymbol{X}', \boldsymbol{t}'_i \in \boldsymbol{T}'} t'^6_i t'^7_i / |\boldsymbol{X}'|,$$

$$\mu_{\mathrm{tp}}(\boldsymbol{X}, D, \boldsymbol{T}) = \sum_{i \text{ s.t. } t^1_i = c^1_j, t^3_i \in D, \boldsymbol{c}_j \in \boldsymbol{X}, \boldsymbol{t}_i \in \boldsymbol{T}} t^6_i t^7_i / |\boldsymbol{X}|$$

である．

本指標値は最大値であるため，以下の手順により効率よく計算できる．

(1) $y := 0$.

(2) 各 $D \in \{\boldsymbol{T}$ 上の連続した 30 日間 $\}$ について，以下を実行する．

  (a) 各 $\boldsymbol{c}'_j \in \boldsymbol{M}'$ について，以下を実行する．

    (i) $k_j := \sum_{i \text{ s.t. } t'^1_i = c'^1_j, t'^3_i \in D, \boldsymbol{t}'_i \in \boldsymbol{T}'} t'^6_i t'^7_i - \sum_{i \text{ s.t. } t^1_i = c^1_{p(j)}, t^3_i \in D, \boldsymbol{t}_i \in \boldsymbol{T}} t^6_i t^7_i$.

  (b) 各 k_j のうち，最小のもの 10 個の値の総和を y_0 とする．

  (c) 各 k_j のうち，最大のもの 10 個の値の総和を y_1 とする．

  (d) $y := \max(y, y_1, -y_0)$.

(3) $y/10$ を出力する．

3.5.3 指標 ut-jaccard

顧客 i が $\boldsymbol{T}$ において購入した商品の多重集合を,

$$S(\boldsymbol{T}, i) = \{\underbrace{t_j^5, \ldots, t_j^5}_{t_j^6} \mid t_j \in \boldsymbol{T}|_{t^1=i}\}$$

と記載する. i の加工前, 加工後に購入した商品の多重集合の距離を, 多重集合版の Jaccard 係数を用いて

$$d(S(\boldsymbol{T}, p(i)), S(\boldsymbol{T}', i)) = 1 - \frac{|S(\boldsymbol{T}, p(i)) \cap S(\boldsymbol{T}', i)|}{|S(\boldsymbol{T}, p(i)) \cup S(\boldsymbol{T}', i)|}$$

と定義する. $\boldsymbol{M}'$ における全ユーザの距離の和,

$$\text{ut-jaccard} = \sum_{1 \le i \le n'} d(S(\boldsymbol{T}, p(i)), S(\boldsymbol{T}', i))$$

を加工の大きさを表す指標とする.

本コンテストにおいては,

$$\text{ut-jccard} > 0.7 \cdot n'$$

であるとき, $\boldsymbol{T}'$ を過加工とみなして排除している.

3.5.4 RFM 分析 ut-rfm_t_c.jar

$RFM(\boldsymbol{M}, \boldsymbol{M}', \boldsymbol{T}, \boldsymbol{T}')$ は，RFM 分析 [3] の観点で R(最後の購買日）, F （購買頻度）, M （購買額）の 3 つの条件で，表 8 の閾値でそれぞれ 10 分位値を目安に 10 ランクに分け, 計 1000 ランクに顧客を分類し，その二乗平均平方根誤差 RMSE とする．ただし，RMSE の最大値 $\sqrt{2n^2/1000}$ で割り，$[0, 1]$ に正規化している．

表 8 ランク閾値（閾値未満か以上かで分類）

項目	閾値
R	2011/03/27, 2011/06/19, 2011/08/28, 2011/10/02, 2011/10/23, 2011/11/13, 2011/11/20, 2011/11/27, 2011/12/04.
F	2, 3, 4, 5 ,6, 7, 8, 9, 10
M	200, 300, 400, 500, 700, 1000, 1400, 2100, 3700.

3.5.5 ハミング距離 ham.rb

顧客 ID を除くマスターの各セルの値のうち，匿名加工前後で異なるセルの割合を出力する.

$$e(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

とするとき，出力は

$$\frac{\sum_{i=1}^{n'} \sum_{j=2}^{4} e(c_i'^j, c_{p(i)}^j)}{3n'}$$

である [*3].

[*3] 本指標は，2016 年度コンテスト予備戦では用いていない.

3.5.6 頻出アイテム topitem

ある伝票 ID x において, 一度に購入された商品の組み合わせの集合, すなわち

$$item(x, \boldsymbol{T}) = \{y \mid y \subseteq \{t_j^5 \mid \boldsymbol{t}_j \in \boldsymbol{T}|_{t^2=x}\}\}$$

を考える. 例えば表 5 では,

$$item(12326, \text{表 5}) = \{\{21913, 22431\}, \{21913\}, \{22431\}\}$$

である. 商品の組み合わせ y が購入された回数を

$$count(y, \boldsymbol{T}) = |\{x \mid y \in item(x, \boldsymbol{T}), x = t_j^2, \boldsymbol{t}_j \in \boldsymbol{T}\}|$$

と記述する. あるしきい値 θ に対して, $Top(\boldsymbol{T})$ を

$$Top(\boldsymbol{T}) = \{y \mid count(y, \boldsymbol{T}) > \theta \wedge \forall y' \supset y \; count(y', \boldsymbol{T}) \not> \theta\}$$

と定義し, 有用性として

$$topitem(\boldsymbol{T}, \boldsymbol{T}') = \frac{|Top(\boldsymbol{T}) \cap Top(\boldsymbol{T}')|}{|Top(\boldsymbol{T})|}$$

を定める. 尚, $Top(\boldsymbol{T})$ については apdata.rb に記録されている.

3.6 匿名加工例

匿名加工アルゴリズムは，入力としてマスター $\boldsymbol{M}$ と履歴データ $\boldsymbol{T}$ を取り，加工した $\boldsymbol{M}'$ と $\boldsymbol{T}'$ と，$\boldsymbol{M}'$ の置換を表す行番号データ P を出力する.

以下に，いくつかの加工アルゴリズムの例を示す（これらのソースコードは本コンテストの参加者に提供する）.

3.6.1 山岡匿名加工 ano-ya.rb, ano-tya.rb

山岡匿名加工 [2] を行う．顧客 ID を除いて，$\boldsymbol{M}' = \boldsymbol{M}$, $\boldsymbol{T}' = \boldsymbol{T}$ とするが，匿名加工の行番号データはランダムに $P = (r_1, \ldots, r_n)$ に振り，不整合にする.

$(r_1, \ldots, r_n)$ として ano-ya.rb では $(1, \ldots, n)$ を一様ランダムに並べ変えた列を，ano-tya.rb では $(2, \ldots, n, 1)$ を，それぞれ用いる.

3.6.2 ano-sampling.rb

$\boldsymbol{M}$ の各行を独立に確率 0.9 で削除する．もしもすべての行が削除されてしまった場合は，最初からやり直す.

3.6.3 ano-shuffle.rb

置換 p をランダムに作り，それについて（整合させて）$\boldsymbol{M}'$, $\boldsymbol{T}'$ を作る.

3.6.4 ano-divt.rb

$\boldsymbol{M}'$ を，日付を均一に $c_1^3 = \cdots = c_n^3$ とする．$\boldsymbol{T}'$ の年間購買額と購買数を維持したままで，新規のレコードを追加する $(m < m')$.

3.7 再識別アルゴリズム

再識別アルゴリズムは，入力としてオリジナルのマスター $\boldsymbol{M}$ と履歴 $\boldsymbol{T}$, 加工されたマスター $\boldsymbol{M}'$ と履歴 $\boldsymbol{T}'$ を

取り，$\boldsymbol{M}'$ の置換を推定した推定行番号データ Q を出力する．

以下に，いくつかの再識別アルゴリズムの例を示す（これらのソースコードは本コンテストの参加者に提供する）．

3.7.1 マスター属性値によるソート re-sort.rb

$\boldsymbol{M}$ および $\boldsymbol{M}'$ を (性別，誕生日，国名) をキーとして昇順に安定ソートし， ソート後の順序で対応付けを行い行番号を推測する．表 $\boldsymbol{M}$ の i 行目を $\boldsymbol{M}[i]$ で参照することとする．ソート後の $\boldsymbol{M}, \boldsymbol{M}'$ をそれぞれ $\boldsymbol{N}, \boldsymbol{N}'$ とすると，ある全単射 x, y について $\boldsymbol{N}[i] = \boldsymbol{M}[x(i)]$，$\boldsymbol{N}'[i] = \boldsymbol{M}'[y(i)]$ が成り立つ．再識別アルゴリズム re-sort.rb は $\boldsymbol{N}'[i]$ と $\boldsymbol{N}[\lfloor (i-1) \times \frac{n}{n'} \rfloor + 1]$ が一致すると推測する．出力は，$i = 1, \ldots, n'$ について

$$q(i) = x(\lfloor (y^{-1}(i) - 1) \times \frac{n}{n'} \rfloor + 1)$$

を満たす推定行番号 $Q = (q(1), \ldots, q(n'))$ である．

3.7.2 トランザクション数によるソート re-tnum.rb

$\boldsymbol{M}$ および $\boldsymbol{M}'$ を対応する履歴の行数をキーとして昇順に安定ソートし， その後は 3.7.1 の re-sort.rb と同様にソート後の順序で対応付けを行い，行番号を推測する．$\boldsymbol{M}$ の i 行目の顧客に対応する履歴の行数は $|\{j \mid c_i^1 = t_j^1\}|$ により計算される．

3.7.3 総購入金額の距離最小化 re-totprice.py

総購入金額が最も近い顧客 ID に再識別を行う．$\boldsymbol{M}$ の i 行目の顧客の総購入金額は

$$a(\boldsymbol{M}, \boldsymbol{T}, i) = \sum_{j \text{ s.t. } c_i^1 = t_j^1} t_j^6 t_j^7$$

により計算される．re-totprice.py の出力は，$i = 1, \ldots, n'$ について，$q(i) = \min\{j \mid \forall k \in \{1, \ldots, n'\}, |a(\boldsymbol{M}', \boldsymbol{T}', j) - a(\boldsymbol{M}, \boldsymbol{T}, i)| \leq |a(\boldsymbol{M}', \boldsymbol{T}', k) - a(\boldsymbol{M}, \boldsymbol{T}, i)|\}$ を満たす推定行番号 $Q = (q(1), \ldots, q(n'))$ である．

3.8 匿名加工提出・評価システム

図 2 に，コンテストの参加者と与えられるデータ，求められるデータの関係を整理する．匿名加工者と再識別者とは，匿名加工提出・評価システムによりデータ交換が行われる．

4. おわりに

オンラインショッピングサイトの実際の購買データを対象とした 2016 年度の匿名加工・再識別コンテストとその評価指標の定義を行った．再識別のリスクを，顧客データベースと匿名加工データベースの置換（対応）を推定されることと定め，そのリスクを下げるための匿名加工の例や再識別アルゴリズムの例を示した．再識別行為を攪乱させるため，元のデータと整合しない行番号（置換）を提出する加工方式（山岡匿名加工）を防止するため，置換関数に

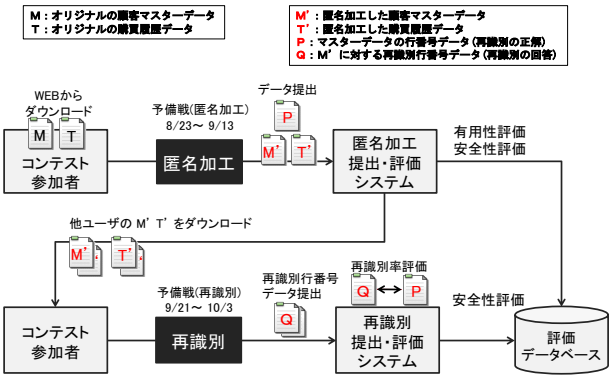

図 2 評価システムのインタフェース概要

基づいたクロス集計の有用性 ($cmae2$) を導入した．

本コンテストを通じて，各種のユースケースに幅広く適用可能な加工技術が開発され，その有用性と安全性が定量的に，かつ信頼できる品質で確立することを期待している．

謝辞

本コンテストの購買データを提供し，その加工に同意して頂いた London South Bank University の David Dqing Chen 博士に感謝する．

参考文献

[1] H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri and J. Sakuma, “Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization,” 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), Crans-Montana, pp. 1035-1042, 2016.

[2] 菊池，山口，濱田，山岡，小栗，佐久間，“匿名加工・再識別コンテスト Ice & Fire の設計”，コンピュータセキュリティシンポジウム (CSS 2015)，プライバシーワークショップ，2B2-1, pp. 1-8, 2015.

[3] Daqing Chen, Sai Liang Sain, and Kun Guo, “Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining,” Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197–208, 2012.

[4] 経済産業省，「事業者が匿名加工情報の具体的な作成方法を検討するにあたっての参考資料（「匿名加工情報作成マニュアル」）Ver1.0，2016. (`http://www.meti.go.jp/press/2016/08/20160808002/20160808002.html`)

[5] Information Commissioner’s Office (ICO), Anonymisation: managing data protection risk code of practice, 2012.

[6] Khaled El Emam, Luk Arbuckle, “Anonymizing Health Data Case Studies and Methods to Get You Started”, *O’Reilly*, 2013 (木村による和訳あり).

[7] Josep Domingo-Ferrer, Sara Ricci and Jordi Soria-Comas, “Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker”, 2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST), *IEEE*, 2015.