PWSCUP2017: 長期間の履歴データの再識別リスクを競う

菊池 浩明[1]　小栗 秀暢[2]　中川裕志[10,11]　野島 良[3]　波多野 卓磨[4]　濱田 浩気[5]　村上 隆夫[6]
門田 将徳[10]　山岡 裕司[7]　山田 明[8]　渡辺 知恵美[9]

概要：2017 年 5 月の改正個人情報保護法施行により，本人同意がなくとも第三者提供ができる「匿名加工情報」の利用が始まった．個人情報保護委員会は「個人情報の保護に関する法律施行規則」を公布し，その第 19 条にて, 匿名加工情報の作成方法に関する 5 つの基準を定めている．基準を満たす匿名加工情報は一定の条件のもとで自由に利活用できるため，匿名加工情報を利用した新しいビジネスの創出が期待されている．しかしながら，匿名加工情報の再識別リスクはそれほど自明ではなく，標準的な評価手法は定まっていない．例えば，匿名加工時に想定していた以上の長期間に渡って履歴が流通すると，再識別リスクが高まることが懸念される．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，高度な匿名加工技術の開発を試みる．コンテストで扱うデータでは，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 種類が，仮名 ID により結び付けられている．コンテストの参加者は, 仮名 ID を適切に変更することによって，より高い有用性とより低い再識別リスクを持つ匿名加工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と, 有用性・安全性の評価方法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

PWSCUP2017: Evaluate re-identification risk on long time history data

Hiroaki Kikuchi[1]　Hidenobu Oguri[2]　Hiroshi Nakagawa[10,11]　Ryo Nojima[3]　Takuma Hatano[4]
Koki Hamada[5]　Takao Murakami[6]　Masanori Monda[10]　Yuji Yamaoka[7]　Akira Yamada[8]
Chiemi Watanabe[9]

Abstract: On May 30, 2017, the amended Act on the Protection of Personal Information has been enforced fully in Japan. Under the new regulation, big-data businesses are about to run efficiently with the newly introduced notion, a de-identified information (anonymously processed data). While, it is not trivial to choose the best algorithm to make the given data anonymized to be secure for a given particular purpose. Especially, the risk of re-identification could increase as the anonymized data are accumulated for a long term observation. Hence, business parties are required the long-term history to be divided into small datasets so that identification to individual is impossible.
To assess the risk to be compromised accurately, the data needs to be balanced in the tradeoff between the utility and the security. 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, assigned for each customer identities. The paper addresses the aim of the competition, the target dataset, sample algorithms, utility and security metrics.

Keywords: personal identifiable information, anonymization, anonymity

[1] 明治大学, Meiji University
[2] 富士通クラウドテクノロジーズ株式会社, Fujitsu Cloud Technologies
[3] 国立研究開発法人 情報通信研究機構, NICT
[4] 新日鉄住金ソリューションズ株式会社, NS Solutions Corporation
[5] NTT セキュアプラットフォーム研究所, NTT Secure Platform Laboratories
[6] 国立研究開発法人 産業技術総合研究所, AIST
[7] 株式会社富士通研究所, FUJITSU LABORATORIES LTD.

1. はじめに

2017 年 5 月，改正個人情報保護法 [7] の本施行を迎え，その第二条 9 号により，本人同意がなくとも第三者提供ができる匿名加工情報が定められた．2016 年には内閣府による法律施行令 [8] により，個人識別符号などが定義された．2017 年，個人情報保護委員会は「個人情報の保護に関する法律施行規則」[9]，及び，ガイドライン類 [10] を公表した．施行規則第 19 条にて, 匿名加工情報の作成方法に関する 5 つの基準が定められた．これを受けて，保護委員会により分野ごとに認定された認定個人情報保護団体は，個人からの苦情申し立ての対応や匿名加工の作成方法などを定めた個人情報保護指針を作成している．こうして，定められた法制度のもと，基準を満たす匿名加工情報は自由に利活用できるため, 匿名加工情報を利用した新しいビジネスの創出が期待されている．

しかしながら，匿名加工情報が再識別されるリスクを正しく評価することはそれほど自明ではない．施行規則による第 19 条の基準についても，「特異な記述等」を削除することは決められているが，何をもって特異とするかは定かではない．ガイドラインの中では，「116 歳」の様な例示はされているが，単純にデータセットの統計量から判断される外れ値とすることにも批判がある．例えば，個人情報保護委員会事務局レポート [11] では，一般人が知りえる社会通念上特異であるもの[*1] と条件付けされており，データセットの統計を取ることは一般人には困難であるとみなされる可能性がある．頼りになるはずの，認定個人情報保護団体の情報保護指針も，その分野固有の基準を記載するには至っていない．

そこで，我々は，共通のデータセットを指定して，多くの研究者により様々な匿名加工方法や再識別リスクを評価するコンテスト PWSCUP[1], [3], [4], [6] を 2015 年より開催している．2015 年には，全国消費実態調査を基に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 [18] を用いて，各世帯の 25 種類の属性を加工した．2016 年には，UCI Machine Learning Repository[*2] にて公開されている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 年間の購買履歴の Online Retail Data Set を用いた．このデータセットから，購買日，商品，単価，購買数などの 7 種類の属性についての，1 万 8 千レコードを匿名加工するコンテストを実施し，15 チームがそれぞれの匿名加工データを提出した．15 チームの平均で，47% の匿名加工されたレコードが正しく再識別され，優勝チームの場合でも 22%レコードが識別されることが分かった．サンプルとして設定した共通の再識別アルゴリズムによる再識別率が 18% だったのに対して，各チームが互いに試みた再識別率の平均は 47 % であり，再識別の正しい評価がいかに困難であるかを明らかにした．

これまでの結果から，コンテストには次のような課題があることを認識している．

- 匿名加工されたレコードの正しい行番号を識別するコンテストのルールに対して，匿名加工者が偽りの置換行番号を提出する，いわゆる「山岡匿名加工」[6] を行う参加者が生じる．
- 匿名加工する前のオリジナルの個人情報データベースを攻撃者の背景知識として与える最大知識攻撃者モデルの仮定は強すぎて，現実とは乖離している．事務局レポート [11] で主張される，いわゆる「一般人基準」による評価とはみなされない．
- 匿名加工に用いたアルゴリズムやソースコードが非公開の為，コンテスト以外のデータに適用することが出来ない．また，ROC などによる加工アルゴリズムの厳密な評価をすることも出来ない．
- 独自の有用性，安全性の基準を用いているため，コンテストの後に策定された施行規則の基準に照らしてみると，不適合とみなされる可能性がある．
- 1 年間もの長期間に，顧客に単一の仮名を割り当てるのは識別リスクが高い．識別される前に，別の新たな仮名を割り当てることが出来ない．

そこで，これらの課題に対して，本年度は次のような工夫を行う．

(1) 長期間履歴の分割.
1 年間の履歴を月ごとの 12 個の期間に分けて提供する．仮名の割り当ては期間ごとに行い，別の期間で変更することを認める．単一の顧客に全ての期間で同じ仮名を割り当てれば有用性が高いが，識別されるリスクが高まる．参加者のこの有用性と再識別リスクのトレードオフの中で，仮名制御を最適化する．

(2) 部分知識攻撃者モデル.
最大知識攻撃者モデルに代わり，攻撃者には確率的にサンプリングした部分的な履歴情報を基にして再識別を行わせる．複数の部分的な背景知識に基づいて安全性と有用性を評価する．

(3) 施行規則 19 条への対応.
匿名加工の規則に照らし合わせて，コンテストのルールと評価方法を設計する．ただし，「特異な記述」などを統一的な方法で基準を設けるにはまだ議論が足りな

8 株式会社 KDDI 総合研究所, KDDI Research, Inc.
9 筑波大学, University of Tsukuba
10 東京大学，the University of Tokyo
11 国立研究開発法人 理化学研究所, RIKEN

*1 「その情報の項目の性質や集団の大きさ、集団の分布の特徴等を考慮して判断されるべきものであるが、社会通念上特異であるものが対象になるため、特異であるものであっても、分布の調査結果が存在しないもの、存在したとしても一般人には知りえないものについては、本号の「特異」には該当しないものと考えられる。」同 24p.

*2 https://archive.ics.uci.edu/ml/datasets/Online+Retail

表 1　PWSCUP コンテストエディションの比較

	2015	2016	2017
開催日	10/21,22	10/11,12	10/23,24
開催地	長崎ブリックホール	秋田キャッスルホテル	山形国際ホテル
参加者	13 チーム (20 名)	15 チーム (42 名)	–
データセット	疑似ミクロデータ	UCI Dataset Online Retail	
属性数	25	11 (顧客 4 + 履歴 7)	
個人数 n	8,333	400	500
履歴数 m	N/A	18,524	44,917
履歴データ　T	N/A	年間 T	月単位 $T^1, \ldots, T^{(12)}$
攻撃者モデル	最大知識攻撃者		部分知識攻撃者

いと判断し，各チームに解釈を委ねるところを残す．上位チームによる最終プレゼンテーションにおいて，それぞれの解釈を報告する機会を設ける．

(4) 匿名加工データのレコード並べ替えシステム．
山岡匿名加工の課題に対して，加工方法を明示したデータの並べ替えをシステム側で行う．これにより，置換行番号の提出がなくなり，山岡匿名加工が困難となることが期待できる．

(5) アイテムベース協調フィルタリングのユースケース．
単一の顧客に複数の仮名を振ることを許すため，顧客当たりの平均購買数などの統計量を公平に評価することが難しい．そこで，顧客数に依らない匿名加工情報の応用として，アイテムとアイテムの関係を推測する，推薦システムなどに活用されているアイテムベース協調フィルタリングなどのユースケースを想定する．

ソースコードの公開などについても検討を行ったが，知的財産の観点で参加が困難になる事業者などを考慮して，本コンテストでは見送り，今後検討をする．以上の本年度の特色と，過去のコンテストの関係を表 1 に整理する．本稿では，このコンテストの基本定義と, 有用性・安全性の評価方法などについて述べる．

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
- 長期間に渡る履歴データの加工基準を算出すること

2.2 匿名加工の概要

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，顧客識別子 cid により結び付けられている．例えば，図 1 の顧客 Alice は，1 月 20 日と 22 日に商品 Chocolate と Candy を購入

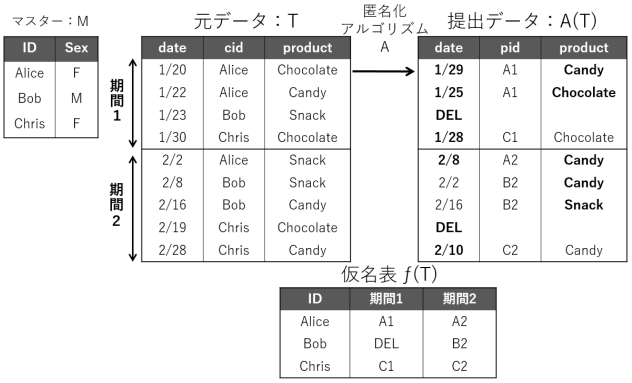

図 1　匿名加工の概要

している．

この購買履歴が Alice のものであることが識別できないように，匿名化アルゴリズム A により加工したデータを $A(T)$ とする．$A(T)$ は T に対して顧客 cid が仮名 pid に振り替えられ，日付や商品のランダムな変更（摂動化），レコード（行）削除などが行われる．例えば，Alice の 1 月 20 日の Chocolate は，1 月 29 日に Candy を購入したことに加工され，顧客 Bob の履歴は識別されるリスクが高いと判断されて，削除（図の “DEL” で指定）されている．この T と $A(T)$ の各行は，一対一に対応しており，この二つを参照すると匿名加工者が行った加工の内容が分かる．本コンテストでは，行の削除は認めるが，T に存在しない架空の行を追加することは認めない．従って，T と $A(T)$ の行数は一致し，$A(T)$ の実質的な履歴の数は T より少ない．

本コンテストで特徴的なのは，T の履歴が長期間（実際には 1 年間）に渡っている点である．加工は短期間（1 か月）で定期的に行われて，第三者に提供される．図では，履歴 T は期間 1(1 月，赤) $T^{(1)}$ と期間 2 (2 月，青) $T^{(2)}$ の二つの期間に分割されている．

この期間と加工について，次の制約がある．

- 加工はこの期間ごとに行われるため，期間を超えた変更は起きない．例えば，Alice の 1 月 22 日の履歴は，加工されて 1 月末に第三者提供されるので，2 月に変更することは出来ない．
- 仮名は期間内では矛盾のない様に割当てる．Alice の

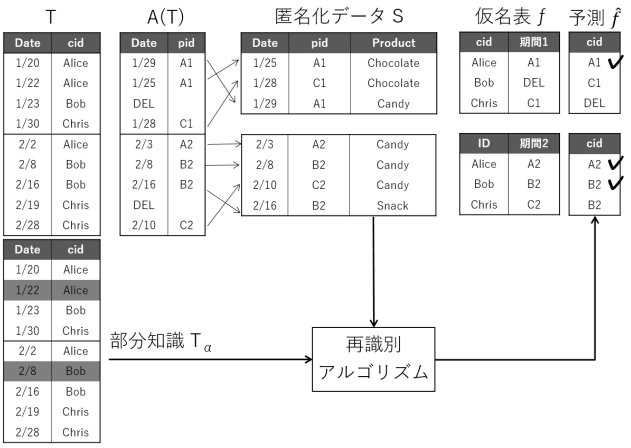

図 2 部分知識攻撃者モデル

1 月の仮名 $A1$ は一つである．

- 期間によって仮名は必ずしも変更しなくてよい．すなわち，$A1 = A2$ となることも許されている．

同一の仮名を長期間使い続けたり，多量の購買が行われたりすれば，その顧客の識別されるリスクが高まる．一方，毎回仮名を変更していると，データの持つ特性（例えば，顧客あたりの購買数など）が変わり，加工データの有用性が落ちる．従って，仮名の利用の長さ（寿命）について，識別の安全性とデータの有用性の間にトレードオフが生じる．匿名加工者は，適切な長さに仮名の寿命を制御しなくてはならない．

仮名は，上記のルールの元，期間ごとに顧客 ID について一意に割当てられる．従って，顧客 ID と期間ごとの仮名の関係は，図の仮名表 $f(A(T))$ で表される．

2.3 部分知識攻撃者モデル

本コンテストで想定する攻撃者のモデルを図 2 に示す．

購入履歴 T と匿名加工された履歴 $A(T)$ から，仮名表 f が出来る．それに基づいて，DEL で指定された行を削除し，日付などの順番でソートして縮約されたデータを匿名化データ S と呼ぶ．再識別者には，この $A(T), f$ は伏せられ，S だけが渡される．履歴 T の部分知識を基に，S から真の加工表 f を予測し，予測加工表 $\hat{f}$ を提出する．$f \approx \hat{f}$ の精度で，再識別率を定義する．図 2 では，正しく再識別されている行（顧客）にチェックをつけている．

再識別者に与えられる背景知識は，元の履歴 T の一部のみに制約される．図 2 の例では，1 月 22 日と 2 月 8 日の履歴（薄く表示）は再識別者には渡されない．この再識別者に秘匿される条件（列，行の割合など）を α で表す．すなわち，再識別者には，部分知識 T_α と匿名化データ S が提供される．α が大きくなると識別の精度があがり，1 に近づくと，最大知識攻撃者モデルになる．本コンテストでは，複数の部分知識についての再識別率で加工データのリスクを検討する．

表 2 Online Retail Dataset， 統計量

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

表 3 Online Retail Dataset, 購買属性

属性名	記述
CustomerID	顧客 ID
InvoiceDate	購入日時
StockCode	商品 ID
UnitPrice	単価
Quantity	購入数

表 4 顧客マスターデータベース M

項目名	記述	加工方針
CustomerID	顧客 ID	T と接続する顧客 ID
Sex	性別 (f, m)	そのまま利用
Generation	年代	誕生日属性を 1930-1980 までの年代に抽象化
Country	国名	United Kingdom, France，Germany, Others の 4 種類.

2.4 課題データセット

購買履歴データセットには，昨年度 [6] と同じく Online Retail Data Set を用いる．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[*3] から公開されている．これらのデータをクレンジングし，表 2，3 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，今年度の加工対象ではないため，引き続き，昨年度に合成されたデータを用いる．今年度は，年齢及び国名属性を抽象化し，性別，年齢，国名属性の組み合わせで 2-匿名化を達成できるよう属性の組み合わせ出現数が特殊となるデータを排除した．これにより，識別子以外からは個人特定リスクが低減されるよう加工されている．

購買履歴データセット，及び顧客マスターデータベースは，コンテストの進行に応じて必要な量がサンプリングされ，課題データとして参加者に提供される．

表 4 に，顧客マスターデータベースの属性と加工方針を示す．

3. 提案コンテスト

3.1 基本定義

マスターデータベース M は，n 人の顧客の情報を格納

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

した n 行 4 列の行列

$$M = \begin{pmatrix} c_{1,1} & \cdots & c_{1,4} \\ \vdots & \ddots & \vdots \\ c_{n,1} & \cdots & c_{n,4} \end{pmatrix}$$

である．列は，表 4 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，期間 $\ell = 1, \ldots, 12$ についての履歴 $T^{(1)}$, ..., $T^{(12)}$ から構成される．期間 ℓ の履歴 $T^{(\ell)}$ は，m 行 7 列の行列

$$T^{(\ell)} = \begin{pmatrix} t^{(\ell)}_{1,1} & \cdots & t^{(\ell)}_{1,7} \\ \vdots & \ddots & \vdots \\ t^{(\ell)}_{m,1} & \cdots & t^{(\ell)}_{m,7} \end{pmatrix}$$

である．行は，m 個のレコード（履歴），列は，ここで，j 番目の購買履歴は，表 3 の順番に対応する 7 つの属性 (顧客 ID, 伝票 ID, 購買日，購買時，商品 ID, 単価，数量) を表す．この内，伝票 ID 属性は今年度の課題で使用しないため，値が 0 に統一されている．

(例 3.1)　顧客マスター $\boldsymbol{M}$ と購買履歴 $\boldsymbol{T}$ の例を表 5，6 にそれぞれ示す．

表 5　顧客マスターデータ $\boldsymbol{M}$ の例

$c_{\cdot,1}$ 顧客 ID	$c_{\cdot,2}$ 性別	$c_{\cdot,3}$ 年代	$c_{\cdot,4}$ 国籍
12360	f	1950/1/1	Others
12361	m	1960/1/1	Germany
12362	m	1950/1/1	France
12363	f	1970/1/1	United Kingdom

表 6　購買履歴データ $\boldsymbol{T}$ の例

$t_{\cdot,1}$ 顧 ID	$t_{\cdot,2}$ 伝票	$t_{\cdot,3}$ 購買日	$t_{\cdot,4}$ 時刻	$t_{\cdot,5}$ 商品	$t_{\cdot,6}$ 単価	$t_{\cdot,7}$ 数
12362	0	2011/2/17	10:30	21913	3.75	4
12362	0	2011/2/17	10:30	22431	1.95	6
12361	0	2011/2/25	13:51	22630	1.95	12
12361	0	2011/2/25	13:51	22555	1.65	12
12362	0	2011/4/28	9:12	21866	1.25	12
12362	0	2011/4/28	9:12	20750	7.95	2
12362	0	2011/4/28	9:12	22908	0.85	12
12360	0	2011/5/23	9:43	21094	0.85	12
12360	0	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と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，M と T のレコードの直接識別子や特異な値を削除したり，値を変更するなどして，S に加工し，顧客 cid と仮名 ID の仮名表 f 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．

仮名表 f は，期間と顧客 ID から仮名 ID への写像

$$f : \{1, \ldots, 12\} \times Dom(M_1) \to Dom(T_1^{(\ell)})$$

である．ただし，$Dom(M_1)$ は M の 1 行目の全ての値と削除されたことを示す値 DEL からなる集合とする．

一方，再識別者が提出する推定仮名表は，期間と仮名 ID から推定した顧客 ID への写像

$$h : \{1, \ldots, 12\} \times Dom(T_1^{(\ell)}) \to Dom(M_1)$$

である．h は，顧客 ID の集合 $Dom(M_1)$ への全射である必要も（削除された顧客もあるので），単射である必要もない（複数の仮 ID を同一の顧客に推定することも認める）ことに注意しよう．

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．

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．

3.4　規則 19 条の基準適合性

法第 36 条第 1 項では，匿名加工情報を作成するときの基準は，個人情報保護委員会規則に従うこととされている．保護委員会の規則第 19 条は以下の基準を定めている．

(1) 個人情報に含まれる特定の個人を識別することができる記述等の全部又は一部を削除[*4] すること
(2) 個人情報に含まれる個人識別符号の全部を削除すること
(3) 個人情報と当該個人情報に措置を講じて得られる情報とを連結する符号（現に個人情報取扱事業者において取り扱う情報を相互に連結する符号に限る。）を削除すること
(4) 特異な記述等を削除すること
(5) 前各号に掲げる措置のほか、個人情報に含まれる記述等と当該個人情報を含む個人情報データベース等を構成する他の個人情報に含まれる記述等との差異その他

[*4] 「削除」には，（当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む。）という代替法が認められている．他の規準についても同様

表 7　規則第 19 条規準と本コンテストの解釈

号	基準	解釈
(1)	個人情報	各チームで検討
(2)	個人識別符号	該当なし
(3)	連結する符号	顧客 ID
(4)	特異な記述	各チームで検討
(5)	他の記述等との差異	各チームで検討

の当該個人情報データベース等の性質を勘案し、その結果を踏まえて適切な措置を講ずること。

この規則に対する本コンテストの解釈を表 7 に示す．

まず，(1) と (5) については，攻撃者の背景知識と識別アルゴリズムに依存するので，統一的な解釈は各チームに委ねる．(2) については，該当なし．(3) は，分散管理の為の情報を相互に連結する符号を意味しているので，本コンテストの M と T を結び付ける顧客 ID が該当する．本コンテストでは，顧客 ID を複数の仮名に置換えることを加工者に求めているが，その仮名が偶然に元の顧客 ID と重なることがあるかも知れない．しかし，それを考慮すると判定が困難なため，本コンテストでは仮名は顧客 ID と重複しないことを求める．

(4) については，ガイドラインの中では「症例数の極めて少ない病歴」,「116 歳」の 2 例が挙げられている．例えば，その候補には，

(1) 年間で 1 回しか購入されていない「特異な」商品

(2) 単価の平均値 ± 4 標準偏差を超える商品

などが考えられる．しかし,「特異」の考え方や対象とする属性などによって，様々な場合が考えられる．そこで，もしも特異な記述が残っていれば，それは再識別率に反映されると考え，その対処については各チームで解釈することとする．どのような対処を行なったのか，最終プレゼンテーションにて発表する機会を与える．

3.5　安全性

3.5.1　安全性指標：仮名の安全性

攻撃アルゴリズムは仮名表 f を推定するものとする．本指標は f の $12n$ 組の対応のうち正しく推定できた組の割合を指標値とする．すなわち，攻撃アルゴリズムの出力を h とするとき，指標値は，$reid_c(f,h) =$

$$\frac{|\{(i,j) \mid i \in \{1,\ldots,12\}, j \in Dom(T_1^{(i)}), f(i,h(i,j)) = j\}|}{12n}$$

と定める．

例えば，図 2 の時，$h(1,A1) = Alice$, $f(1,h(1,A1)) = A1$ となり，再識別率は，$reid_c(f,h) = |\{(1,A1),(2,A2),(2,B2)\}|/2 \times 3 = 3/6$ で与えられる．

3.5.2　安全性指標：各トランザクションの安全性

攻撃アルゴリズムは S の各トランザクションが M のどの顧客と対応しているかを推定するものとする．本指標は S の各トランザクションのうち正しい顧客を推定できたものの割合を指標値とする．すなわち，攻撃アルゴリズムの出力を $R : \{1,\ldots,|S|\} \to \{1,\ldots,n\}$ として，指標値は以下の式で与えられる．

$$\frac{|\{i \mid i \in \{1,\ldots,|S|\} \wedge c_{R(i),1} = t_{P(i),1}\}|}{|S|}.$$

3.6　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) アイテムベース協調フィルタリング

アイテムベース協調フィルタリング [17] は，ある商品を購入した人に対して，別の商品を推薦するためのアルゴリズムである．アイテムベース協調フィルタリングでは，購買履歴を基に商品 (Item) 同士の類似度を格納した Item-Item 行列を計算する．この Item-Item 行列から，ある商品を購入した人が，他にどのような商品を購入しやすいかの傾向を知ることができる．

(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 [5] に対して，有用性を損なうような指標が必要である．

3.6.1　アイテムベース協調フィルタリング ut-ItemCF

アイテムベース協調フィルタリングに基づく有用性指標の概要を図 3 に示す．まず元データ T から，User-Item 行列 $\mathbf{V}$ を計算する．$\mathbf{V}$ の (i,j) 要素 $v_{i,j}$ には,「i 番目の顧客 ID（cid）による j 番目の商品 ID の購入回数」が格納される．$\mathbf{V}$ の i 列目を取り出したベクトルを「アイテムベクトル」と呼ぶことにし，$\mathbf{v}_i$ と表記する．例えば図 3 では，$\mathbf{v}_1 = (56,12,28,0,0,0)^T$ である．次に，User-Item 行列 $\mathbf{V}$ を用いて，Item-Item 行列 $\mathbf{W}$ を計算する．$\mathbf{W}$ の (i,j) 要素 $w_{i,j}$ には,「アイテムベクトル $\mathbf{v}_i$ と $\mathbf{v}_j$ のコサイン類似度 $\mathbf{cos}(\mathbf{v}_i,\mathbf{v}_j)$」が格納される．コサイン類似度 $\mathbf{cos}(\mathbf{v}_i,\mathbf{v}_j)$ は,

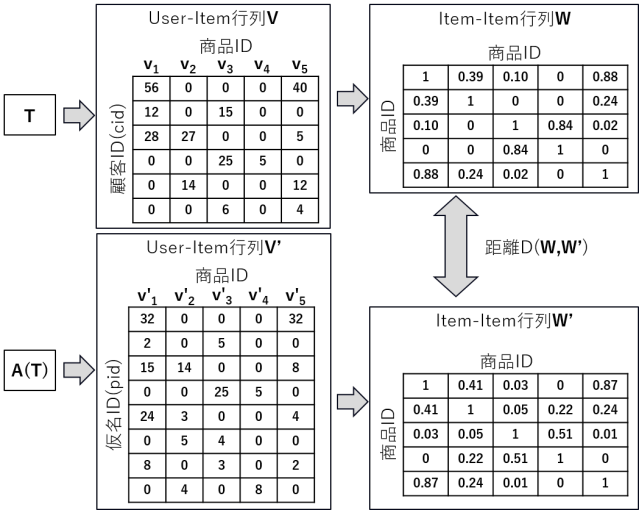

図 3 アイテムベース協調フィルタリングに基づく有用性指標

$$\mathbf{cos}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i^T \mathbf{v}_j}{|\mathbf{v}_i||\mathbf{v}_j|} \tag{1}$$

で定義される類似度である．

同様に，提出データ $A(T)$ からも User-Item 行列 $\mathbf{V}'$ を計算し，Item-Item 行列 $\mathbf{W}'$ を計算する．提出データ $A(T)$ では，計 12 個の各期間で異なる仮名 ID（pid）を付与している可能性があるため，$\mathbf{V}'$ の行数は $\mathbf{V}$ の行数より多くなる可能性がある．（複数の期間で同じ仮名 ID（pid）を用いる場合もあるため，最大で 12 倍）．ただし，商品数に変化はないため，$\mathbf{W}'$ の行数（=列数）は $\mathbf{W}$ と同じである．

Item-Item 行列 $\mathbf{W}$ と $\mathbf{W}'$ の距離を，有用性指標として用いる．具体的には，以下で定義される正規化 L_1 距離を使用する．

$$D(\mathbf{W}, \mathbf{W}') = \frac{\sum_{(i,j)\in W_D} |w_{i,j} - w'_{i,j}|}{\sum_{(i,j)\in W_D} |w_{i,j}|} \tag{2}$$

ここで，W_D は，行列 $\mathbf{W}$ のうち 0 でない（即ち，dense な）要素の集合である．例えば図 3 では，$W_D = \{(1,1),(1,2),(1,3),(1,5),\cdots(5,3),(5,5)\}$ である．$\mathbf{W}'$ が全て 0 のとき，$D(\mathbf{W}, \mathbf{W}') = 1$ となる．尚，$D(\mathbf{W}, \mathbf{W}')$ が 1 を超える可能性もあるが，その場合は有用性の値が 1 になるようにした．即ち，$\min\{D(\mathbf{W}, \mathbf{W}'), 1\}$ を最終的な有用性の値として求めた．

次に，上記のアイテムベース協調フィルタリング ut-ItemCF をベースとした有用性指標として，以下の 3 指標を紹介する．

(1) 指標：ut-ItemCF-supply と ut-ItemCF-retail

データセットの利活用者として，Supplier（問屋），Retailer（小売）の 2 者を想定し，匿名加工されたデータがそれぞれに対して価値のあるデータであるかを評価するために，2 種類の Item-Item 行列 $\mathbf{W}$ を定義する．$\mathbf{W}^S$ を Supplier 向け，$\mathbf{W}^R$ を Retailer 向けの Item-Item 行列とし，これらは User-Item 行列 $\mathbf{V}^S$ および，$\mathbf{V}^R$ によって作成されたものである．なお，

- $\mathbf{V}^S$ は，$\mathbf{V}$ の要素のうち，12 以上のものはそのままにして，11 以下のものを 0 と変換したもの
- $\mathbf{V}^R$ は $\mathbf{V}$ の要素のうち，11 以下のものはそのままにして，12 以上のものを 0 と変換したもの

とする．

(a) 指標：ut-ItemCF-supply

T，$A(T)$ から Supplier 向けの Item-Item 行列 $\mathbf{W}^S$，$\mathbf{W'}^S$ を作成し，$\min\{D(\mathbf{W}^S, \mathbf{W'}^S), 1\}$ を計算するアルゴリズムである．なお，ut-ItemCF-supply では，問屋向けのデータセットを想定し，User-Item 行列 $\mathbf{V}^S$ の要素 x は，$\lfloor x/12 \rfloor$ と変換する（即ち，$12 \le x \le 23$ のときは 1，$24 \le x \le 35$ のときは 2，というように変換する）．

(b) 指標：ut-ItemCF-retail

$T, A(T)$ から Retailer 向けの Item-Item 行列 $\mathbf{W}^R$，$\mathbf{W'}^R$ を作成し，$\min\{D(\mathbf{W}^R, \mathbf{W'}^R), 1\}$ を計算するアルゴリズムである．

(2) 指標：ut-topk

このアルゴリズムは，最も多くの顧客に購入された商品の上位 $\mathbf{k}$ 個に対して評価を行うものである．商品 j の購入顧客の集合 $\mathbf{C}_j$ は以下のように定義できる．

$$\mathbf{C}_j = \{i \in Dom(M_1) \mid v_{ij} > 0\}$$

$\mathbf{C}_j$ の要素数は商品 j を購入した顧客の数を表している．全ての商品に対して購入した顧客数 $|\mathbf{C}_j|$ を算出し，購入顧客数が上位 k 個の商品の商品 ID リストを $\mathbf{U}_{topk}$ と定義する．T，$A(T)$ から作成された $\mathbf{V}$，$\mathbf{V}'$ から $\mathbf{U}_{topk}$，$\mathbf{U'}_{topk}$ を作成し，これらの差集合の大きさを k で割ったものを，新たな有用性指標 $topk$ として用いる．すなわち，

$$topk(T, A(T)) = |\mathbf{U}_{topk} \setminus \mathbf{U'}_{topk}|/|k|$$

である．この値が小さいほど，匿名加工されたデータセットに対しても「最も多くの顧客に購入された k 個の商品」が正しく抽出されていることを示している．

3.7 サンプルプログラム

月毎に仮名を変更する仮名化プログラムの実装例は PWSCUP2017 組織委員会で準備しておく．また, 部分知識攻撃者に対する安全性を評価するため,

(1) 同じ日に購入した顧客は同じとみなす,

(2) 同じ商品を購入した顧客は同じとみなす,

(3) 同じ月に同じ商品を購入した顧客は同じとみなす

の少なくとも 3 パターンの再識別プログラムを準備する．

3.8 匿名加工提出・評価システム

図 4 にコンテストにおけるデータの入出力概要を示す．

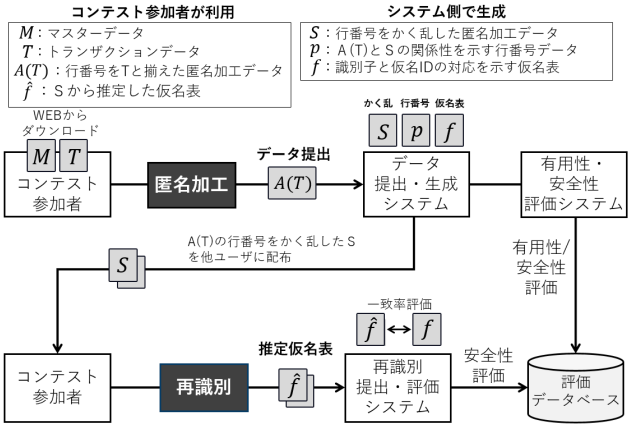

図 4 コンテスト概要

まず，匿名加工フェイズでは，公式 HP からマスターデータ M とトランザクションデータ T をダウンロードし，参加者は匿名加工データ $A(T)$ を作成する．そのデータを評価システムに投入する際，システム側では前述した 19 条に対応する条件をチェックする．その後，有用性・安全性を計測し，その結果をデータベースに記録する．

その後の再識別フェイズでは，$A(T)$ をかく乱した S を生成し，その関係性 p と仮名表 f を隠した状態で参加者に S を提供する．再識別攻撃者は予測仮名表 $\hat{f}$ を提出し，再識別成功率を競う．

コンテストでは，本稿にて示した指標群からいくつかを選択して参加者に対する課題として与えられる．また，本稿で述べた，評価手法やデータ生成に用いたサンプルプログラム等は，公式 HP にて公開されている．

4. おわりに

長期間にわたる購買履歴データを用いた匿名加工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．

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] 菊池 浩明, “匿名加工・再識別コンテスト Ice and Fire：匿名加工方式とその安全性を評価する試み”, 情報処理学会論文誌, 57(9), pp. 1900-1910, IPSJ, 2016.

[4] Hiroaki Kikuchi, Takayasu Yamaguchi, Koki Hamada, Yuji Yamaoka, Hidenobu Oguri, Jun Sakuma, “A Study from the Data Anonymization Competition Pwscup 2015”, Data Privacy Management and Security Assurance (DPM 2016), pp. 230-237, Springer, 2016.

[5] 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.

[6] 菊池, 小栗, 野島, 濱田, 村上, 山岡, 山口, 渡辺, “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 2016 論文集,2016(2), pp. 271-278, IPSJ, 2016.

[7] 個人情報の保護に関する法律（平成 15 年法律第 57 号．平成 27 年法律第 65 号及び平成 28 年法律第 51 号により改正）, 2016.

[8] 個人情報の保護に関する法律施行令（平成 15 年政令第 507 号）（平成 28 年政令第 324 号による改正）， 内閣府，2016.

[9] 個人情報の保護に関する法律施行規則（平成 28 年 10 月 5 日個人情報保護委員会規則第 3 号）, 個人情報保護委員会，2017.

[10] 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）（平成 28 年個人情報保護委員会告示第 6 号ないし第 9 号），個人情報保護委員会，2017.

[11] 個人情報保護委員会事務局, “個人情報保護委員会事務局レポート：匿名加工情報–パーソナルデータの利活用促進と消費者の信頼性確保の両立に向けて”, 2017.

[12] 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.

[13] 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`)

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

[15] 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).

[16] 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.

[17] Greg Linden, Brent Smith and Jeremy York, “Amazon.com Recommendations Item-to-Item Collaborative Filtering”, IEEE Internet Computing, vol. 7, no.1, pp.76-80, 2003.

[18] 秋山 裕美，山口 幸三，伊藤 伸介，星野 なおみ，後藤 武彦, “教育用擬似ミクロデータの開発とその利用～平成 16 年全国消費実態調査を例として～”, 統計センター製表技術参考資料, 16, pp. 1-43, 2012.