購買履歴データを用いた匿名加工データの最頻アイテムに注目した再識別手法 freqItem の提案とその評価

岡本　健太郎†
明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室†

1　はじめに

近年，ビッグデータをデータマイニングする取り組みが盛んになってきているが，データの一部を削除しても個人が特定されてしまうような事態が不安視されている．そこで,2015 年に匿名加工に関する個人情報保護法が改正されこのようなプライバシー保護の観点から匿名加工という概念が生まれた．

本稿では，情報処理学会の CSS(Computer Security Symposium)にて開催された匿名加工コンテスト，通称 PWSCUP2016 において各チームによってさまざまな加工をされた購買データを用いて本提案手法である最頻アイテム再識別にて再識別を行い，提案した再識別率を評価し，他の手法との比較を示す．

2　匿名加工・再識別

本研究において，再識別の手法を述べる際にまず匿名加工の基礎知識を述べることとする．匿名加工とは、ビッグデータに含まれる個人のプライバシーを守りつつデータの有用性を保つようなデータ加工である．また，ビッグデータは名前やマイナンバーのような個人を直接的に特定しうる属性を ID，誕生日や住所，性別など組み合わせることで間接的に個人を特定しうる属性を QI,その他の保護すべき対象の属性を SA と定義し，各属性をこの 3 要素に分類する．匿名加工において ID は当然削除対象なのだが，QI を組み合わせることによって個人を特定できる場面がしばしば生じる．

そこで k-匿名化やトップコーディングなどといった加工を施すことにより QI からも個人を特定できないようにするのが主流である．また，その加工データは有用性と安全性という二つの観点から評価をされ，その価値を問われる．加工データの有用性を示す指標には MAE，安全性を示す指標には l-多様性や m-不変性などがある．これらの指標は同じ QI を持つグループに分類したとき，どのグループの平均値も元データと近似するかどうかや，どのグループの SA が l 個，または m 個持つように加工を施されているかを示す指標である．

再識別とは匿名加工されたデータから個人を特定するための一連の手順のことを指す．特に，K-匿名化された加工データにおいては個人を再識別できる確率は 1/K になる．

一般的に匿名加工されたデータにおいて L-多様性や M-不変性を満たすためにノイズを加えることがあるが，これらの*加工を行うことで安全性が増す代わりに有用性が下がることが知られており，今日の匿名加工の分野では有用性を下げないように安全性を高める手法の提案が議論の中心である．

3　コンテストで用いたデータについて

匿名加工には静的データと動的データの 2 種類が存在する．含有されるタプルが時間によって変化しないものを静的データ，変化するものを動的データとする．例えば，本大会 PWSCUP2016 で扱ったオンラインショッピングの履歴データは動的データである．

3.1　元データ

今回 PWSCUP2016 で用いたデータは,UCI データセットのオンラインショッピングサイトの購買データにおいて顧客数を 10%にサンプリングしたデータを用いている．元データを構成する要素を表 1 に示す．また，顧客の個人情報において性別や生年月日は乱数を用いて合成されている．

このデータセットは顧客個人の情報が入ったマスターデータ，購入履歴が入ったトランザクションデータで構成され，元マスターデータ 4333 顧客中 400 人を無作為に抽出したものをコンテストマスターデータ，その 400 人が含まれる行だけを残したものをコンテストトランザクションデータと呼ぶ．また，PWSCUP2016 で使用したコンテストデータを構成する要素は表 1 の下部のようになっている．

表 1　データの詳細

	マスターデータ	トランザクションデータ
含有情報	顧客の情報	各顧客の購買履歴
属性	顧客 ID，性別，生年月日，国籍の 4 属性	顧客 ID，伝票 ID，購買日時，購買時間，製品 ID，単価，購買数の 7 属性
行数	4333 行	397625 行
コンテストデータ		
行数	400 行	38087 行

†Kentaro Okamoto, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory.

3.2 PWSCUP について

匿名加工されたデータは有用性と安全性に二つの観点からその価値を評価され,PWSCUP2016 では有用性指標を平均絶対誤差 CMAE，ハミング距離 ham，購買アイテム集合 topitem の 5 つ，安全性指標を再識別されたユーザー数の割合と定義した[5]．コンテストは予備戦と本戦に分かれていてそれぞれでデータを加工する匿名加工フェイズ，チーム同士でそれぞれ提出されたデータを再識別する再識別フェイズを行い，予備戦本戦の結果が 1：9 の割合で順位を決定する．匿名加工フェイズでは顧客マスターデータ M と履歴トランザクションデータ T が配布される．そして各々の手法で加工し，加工マスターデータ M′ と加工トランザクションデータ T′，そしてマスターデータの行番号データ P を提出し，各有用性指標の最大値をそのデータの有用性の値とした．その後再識別フェイズではコンテストサイトにおいて各チームの M′ と T′ が公開されているので各自ダウンロードし再識別を行ったのち，行番号データ Q のみを提出する．そこで P と Q を照合し，正解率を再識別率の値とした．

4 提案手法 freqItem

本研究では,有用性指標の中でもとりわけ Y2-jaccard を下げないような匿名加工データを対象にした再識別手法 freqItem を提案する．Y2-jaccard とは購入データの中でユーザーごとに購入アイテムで多重集合を形成し，この集合の差が加工前と加工後で大きいと過加工とみなされ排除されてしまう．この指標は YA-匿名化という，PWSCUP において有用性を全く下げることなく安全性を高める手法を防ぐために用意された指標である，したがって，各チームが最も気にした指標がこの Ý2-jaccard であり，かつ商品集合は加工をしづらいファクターであるため最も高い効果が得られると私は考えた．

4.1 コンテストでの加工手法について

本コンテストでは，事前に定められた有用性指標スクリプトに加工データを入力として値が出力されることによってそのデータの価値を定めている．そのため，闇雲にノイズを加えるのではなく有用性指標を下げないような加工法を提案することが重要である．そこで，本節ではどの有用性指標を対策しているかに注目しながら用いられた主な手法を紹介する．

4.1.1 YA-匿名化

YA-匿名化とは M＝M′，T＝T′ となるようなデータを提出するのだが，行番号データ P をランダムに入れ替える加工手法である．元データとまったく同じデータのため，どの有用性指標も変化せず，その有用性指標は最小値 0 を示し，すべてのユーザーを特定できたとしても P がランダムに入れ替わっているため再識別率は大幅に下がってしまうよう．

4.1.2 Jaccard 最適化

Y2-Jaccard は 4.1.1 節で記した YA-匿名化を防ぐための足切り指標であるが，YA-匿名化をしながら Y2-Jaccard が足切りにかからないようなギリギリの YA-匿名化データを探索するのが Jaccard 最適化である．

4.1.3 Jaccard ランダム化

4.1.2 節では Y2-Jaccard が足切りにかからないギリギリの探索を行う手法であったが，最適化を行うとそれが手掛かりとなってしまい，YA-匿名をしたにも関わらず再識別されてしまう恐れがある．それを防ぐのが Jaccard ランダム化であり,最適化はしないが最適化された Y2-Jaccard の値に近いデータを選択することで Jaccard 最適化されたデータよりも再識別されにくい YA-匿名化をする．

4.1.4 最適化

CMAE1，CMAE2，ut-rfm などの有用性指標を最小にするようなトランザクションデータの加工を探索する加工手法であり，YA-匿名化だけでなく偽造タプルの挿入やノイズを加えた後に CMAE1 を回復するようなタプルの挿入などが最適化に当てはまる．

4.1.5 仮名化

トランザクションデータには ID を示す列があるため，ID の列を見ればどのユーザーをどの程度加工したのかが分かってしまう．それを防ぐのが仮名化であり，マスターデータの顧客 ID に対してハッシュ関数を用いて仮 ID を割り振り，トランザクションデータの ID の列を仮 ID で表現することで元データの顧客と結び付けられないようにする加工法である．表 2，表 3 は仮名化する前と後のマスターデータを表したものである．この例を元にすると Abe→b，Baba→d，Chiaki→c，Doi→a のように ID が変換されているがタプルの内容は変化していないことが分かる．

表 2　仮名化前のマスターデータ

ID	Country	Sex	Birthday
Abe	America	M	Feb
Baba	Bulgaria	M	Dec
Chiaki	Canada	F	May
Doi	Denmark	M	Aug

表 3　仮名化後のマスターデータ

ID	Country	Sex	Birthday
B	America	M	Feb
D	Bulgaria	M	Dec
C	Canada	F	May
A	Denmark	M	Aug

4.1.6 列統一

time の列に関する有用性指標が定められてないためどれほどの加工を施してもどの有用性指標も下がらない．したがって time の列の値をランダムな値に統一する加工で

ある．ランダム化でも結果はほとんど変わらない．

4.1.7 属性値追加

元データの集合に含まれていなかった属性値を追加する手法である．CMAE や RFM の有用性は元データの集合に対して平均絶対誤差などをとるため，新たな値に関しては平均絶対誤差は 0 となる．

4.1.8 グループ内スワップ

グループ内スワップは有用性指標を下げないようにグループ内で値だけをユーザー同士で入れ替える手法である．例えば,表 4 は表 2 のユーザーを性別グループ内で country の列をスワップした例である．例えば，元データでは Abe の country は America であるが表 4 でスワップした後は Bulgaria になっている．なお，Birthday の列の値は変化していない．

表 4　性別グループスワップの例

ID	Country	Sex	Birthday
Abe	Bulgaria	M	Feb
Baba	Denmark	M	Dec
Chiaki	Canada	F	May
Doi	America	M	Aug

4.2 再識別手法 freqItem について

freqItem とは最頻アイテムに注目した再識別手法である．Y2-jaccard は足切指標であるので加工しづらい要素である．しかし，上位チームは Y2-jaccard に関して当然有効な加工をしてくるであろうと考え，Y2-jaccard と同様の freqitem という概念を提唱する．freqitem とはユーザーごとに最も購入したアイテムを単一に定めた値であり，全ユーザーの人気商品の多重集合 topitem とは異なり，最頻アイテムが複数存在した場合はその中でランダムに決定する．例えば,表 5 において,topitem が{doughnut,eraser}だとすると各ユーザーの topitem と freqitem は表 8 のようになる．どのユーザーにおいても，topitem に freqItem に含まれている．topitem とは item 集合の中で頻度の高いアイテムであるのでこのような結果になることが多いと考えられる．例えば，匿名性を高めるために，2 行目を削除すると topitem は eraser が削除され doughnut のみとなるが freqItem は変化しない．

表 5　履歴データの例

	Name	Item
1	Abe	Doughnut
2	Abe	Eraser
3	Baba	Doughnut
4	Abe	Doughnut
5	Chiaki	Fork
6	Baba	Eraser
7	Chiaki	Eraser
8	Baba	Eraser

表 8　ユーザーごとの topitem と freqItem の違い

Name	Topitem	freqItem
Abe	doughnut, eraser	doughnut
Baba	doughnut,eraser	Eraser
Chiaki	Eraser	eraser または fork

4.3 jaccard 再識別について

Jaccrad 再識別[6]は，多重集合の類似度 jaccard 係数[5]を用いて jaccard 係数をユーザーごとに求め，最も近いユーザーを同一ユーザーと再識別している．なお，この再識別手法は PWSCUP2016 において再識別賞を受賞した．

5 再識別結果

表 9 は,PWSCUP2016 の各チームの加工手法と各加工データに対する freqItem 再識別と jaccard 再識別の識別率を示している．ここで，チーム名は，上から順にランキングの上位にいることを表している．表 9 に示された通り，jaccard 再識別は freqItem 再識別の再識別率の平均 2.90 倍の精度があることがわかる．とりわけ，チーム T と K においては 10 倍以上の精度があり，チーム I に対しては 88.50%とほとんどのユーザーの再識別に成功している．すべてのチームで仮名化，列統一が，チーム T 以外のすべてのチームで YA-匿名化を採用している．また，上位 3 チームでは Jaccard 最適化およびランダム化を行っている．それと同様に上位 3 チームにおいて freqItem 再識別の識別率が低く，特にチーム T と K に関してはほとんどのユーザーを識別できていない．

表 9　各チームの加工データに対する freqItem 再識別と Jaccard 再識別の再識別率

チーム名	freqItem 再識別率 (%)	Jaccard 再識別率(%)
T	1.25	22.25
K	0.75	25.50
J	9.50	27.50
B	14.75	30.25
N	15.25	27.50
M	13.00	38.50
I	44.75	88.50
平均	14.18	41.07

表 10　各チームの用いた加工手法

チーム名	YA-匿名化	JCCARD最適化	JCCARDランダム化	最適化	CMAE2最適化	RFM最適化	仮名化	列統一	属性値追加	国グループ内スワップ	国&性別グループ内スワップ	購入傾向グループ内スワップ
T			○	○	○		○	○				
K	○	○		○	○		○	○				○
J	○		○	○	○	○	○	○				○
B	○			○	○		○	○			○	○
N	○	○			○		○	○	○	○	○	○
M	○	○					○	○		○	○	
I	○						○	○				

6　考察

表 9 の加工方法により，freqItem 再識別は仮名化，YA-匿名化などといった手法には識別率を左右されず，スワップに関しては購入グループ内でのスワップに最も強い．それに対し，jaccard 係数を変えずに行われたアイテムの加工手法 Jaccard 最適化・Jaccard ランダム化に対して極端に弱く，ほとんど識別できないことが分かった．Jaccard 係数を変えずに行われたアイテムの加工手法は topitem 以外のアイテムについて，表 11 のような加工が施されていたために識別率が低い結果となった．表 11 は上位チームの購入アイテム集合を変えないように購入したアイテムをユーザーごとに一つだけ残し，他のアイテムはすべて「POST」に変換した加工である．左表が加工前の履歴データ，右表が加工後の履歴データとなっている．つまり，T の Item の列のほとんどが POST になったことにより，ほとんどすべてのユーザーの freqItem が POST になったので正しく再識別できなかった．そこで，freqItem 再識別に加えて Jaccard 係数も計算し，識別要素とすることでさらに再識別率を高めることができるのではないかと期待している．

表 11　上位チームの加工例

	ID	Item
1	Abe	Eraser
2	Abe	Doughnuts
3	Abe	Doughnuts
4	Baba	Eraser
5	Baba	Folk
6	Abe	Eraser
7	Baba	Eraser
8	Baba	Folk
9	Abe	Folk
10	Baba	Doughnuts
11	Abe	Folk

	ID	Item
1	Abe	Eraser
2	Abe	Doughnuts
3	Abe	POST
4	Baba	Eraser
5	Baba	Folk
6	Abe	POST
7	Baba	POST
8	Baba	POST
9	Abe	Folk
10	Baba	Doughnuts
11	Abe	POST

7　おわりに

本研究では，freqItem 再識別と jaccard 再識別の比較を行うことで手法の評価，ならびに改善策を講じることができた．この結果をもとに，現手法の改善や新たな再識別手法の提案をすることで PPDP の研究の発展に貢献していきたい所存である．

参考文献

[1] 南和宏，“プライバシー保護データパブリッシング”，2013 年 6 月

[2] 菊池亮，五十嵐大，濱田浩気，千田浩司，“データを逐次公開する際のプライバシー保護”2015 年 4 月

[3] 上土井陽子，堀内敦史，沖田梨絵子，若林真一，“動的データのプライバシ保護再公開における精確な安全性の評価について”SCIS，2016 年 1 月

[4] 伊藤聡志，菊池浩明“ユークリッド距離を用いた再識別手法と PWSCup2015 の匿名加工データを使用した評価”，情報処理学会 CSEC，2016 年 5 月

[5] 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”，2016 年 9 月

[6] 原田玲央，伊藤聡志，菊池浩明，“商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案”，電子情報通信学会 SCIS，2017 年 1 月