Vol.2019-DPS-178 No.30
Vol.2019-CSEC-84 No.30
2019/3/5

匿名加工・再識別コンテストPWSCUP 2018の報告 ～購買履歴データの一般化加工の安全性と有用性評価～

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：2018 年 10 月に開催された匿名加工・再識別コンテスト (PWSCUP) において，購買履歴のトランザクションデータを一般化したデータを対象に，安全性と有用性の定量指標によって評価する試みがなされた．本稿ではコンテストのルールや評価指標について述べ，参加者から提出された匿名化データに関する安全性と有用性の評価を行う．またコンテストの開催によって得られた知見を報告する．

Koki Hamada[1,2]　Hiromi Arai[2]　Hidenobu Oguri[3]　Hiroaki Kikuchi[4,2]　Atsushi Kuromasa[5]
Hiroshi Nakagawa[2]　Kenshiro Nishiyama[6]　Takuma Hatano[7]　Takao Murakami[8]
Yuji Yamaoka[3]　Akira Yamada[9]　Chiemi Watanabe[10]

1. はじめに

近年のデータ分析技術の進歩に伴い，機械学習やデータマイニングなどのデータ分析の精度を向上させるため，より多くのデータが必要とされてきている．一方で一つの機関で集めることのできるデータは限られており，他の機関が所有するデータをも使用したいというニーズが高まっている．

2017 年 5 月 30 日には改正個人情報保護法が施行され，個人の識別性を低減した匿名加工情報と呼ばれる加工データについては，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．識別性の低減の度合いには統一基準がなく，基準の確立のため，加工データの安全性と有用性を定量的に評価する技術が求められている．

このような状況において，あらかじめ定められた安全性と有用性の評価指標の下で共通のデータセットをより安全かつ有用に加工する技術を競う匿名加工・再識別コンテスト (PWSCUP) [5][7][6] が 2015 年から開催されてきた．

本稿では，4 回目の開催となった PWSCUP 2018 [4] の結果について報告する．そして，PWSCUP 2018 で設定した安全性指標と有用性指標について，それらが設計の意図通りの指標となっていたかどうかを PWSCUP 2018 参加者からの提出データを使って評価する．

1.1 構成

本稿の構成は以下の通りである．2 節では，PWSCUP 2018 のルールや指標を概観する．3 節で PWSCUP 2018 の結果を報告する．4 節で PWSCUP 2018 で用いた安全性指標の設計の背景を説明する．5 節と 6 節では，それぞれ安全性指標と有用性指標を評価した結果を報告する．

2. PWSCUP 2018の概要

2018 年 9 月 6 日から 10 月 22 日にかけて，匿名加工・再識別コンテスト PWSCUP 2018 を実施した．コンテストは表 1 のスケジュールで，約 2 ヶ月間に渡って行われ，10 月 22 日の発表はホテルメトロポリタン長野で，その他はすべてオンラインで実施した．

PWSCUP 2018 には，日本国内外の大学や企業から 14 チームが参加した．

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

表 1 スケジュール

イベント	期間
参加登録	8 月 24 日 – 9 月 3 日
予備戦 (匿名加工)	9 月 6 日 – 9 月 11 日
予備戦 (再識別)	9 月 13 日 – 9 月 18 日
予備戦結果通知	9 月 20 日
本戦 (匿名加工)	9 月 27 日 – 10 月 9 日
本戦 (再識別)	10 月 11 日 – 10 月 16 日
プレゼン・ポスター発表，結果発表	10 月 22 日

表 2 PWSCUP 2018 で扱うデータの例

顧客 ID	購入日	商品 ID	単価	数量
12345	2010/12/1	22728	3.75	24
12345	2010/12/1	22727	5.20	3
12345	2010/12/8	22726	3.75	12
12603	2010/12/9	21724	0.85	10
12635	2010/12/9	21883	0.65	1

2.1 ルールの概要

コンテストのルールの詳細は [4] に記載されているので，ここでは概要を記す．PWSCUP 2018 では，各行が 1 つの商品の購入履歴を表す表形式のデータを扱う．各行は，購入を行った**顧客**ごとに固有の番号 (**顧客** *ID* と呼ぶ) と，購入内容からなる．購入内容は，**購入日**，購入した商品を表す**商品** *ID*，購入した商品の**単価**，購入した商品の**個数**，の 4 つの値の組からなる．PWSCUP 2018 で扱うデータの例を表 2 に示す．

コンテストに参加する各チームは，共通のデータセット (**元データ**と呼ぶ) を加工し，安全性の基準を満たし，かつ，有用性が高いデータ (**加工データ**と呼ぶ) を作成することが求められる．

2.1.1 安全性

PWSCUP 2018 では，安全性の基準として，加工データ上のどの顧客 (**仮顧客**と呼ぶ) についても攻撃者が元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要求する．攻撃者は，加工対象となった元データを知っているものとする (最大知識攻撃者を仮定する)．

2.1.2 有用性

PWSCUP 2018 では，加工データと元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 (加工データと元データが近い) ほど有用性が高いとみなす．

2.2 コンテストの流れ

コンテストは，審判と，複数の参加チームにより行われ，**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る．

(1) 加工フェイズ: 各チームは審判から受け取った加工対象の元データを加工し，審判に提出する．
(2) 再識別フェイズ: 各チームは他のチームが作成した加工データを審判から受け取り，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審判に提出する．
(3) 評価フェイズ: 審判は各チームから受け取った加工データと推定結果から各加工データの安全性と有用性を評価する．

2.3 安全性指標

2.3.1 狙い

PWSCUP 2018 の安全性の評価指標を設計するにあたっては，以下の 2 点の実現を目指した．

(1) あえて無加工のままにされる顧客が生じないこと．
(2) 安全性に一定の基準を定めた評価を可能にすること．

(1) は，前回までのコンテストで使用していた指標の課題であった，一部の顧客をあえて無加工にするという選択が合理的となっていた問題の解決を目指すものである．加工データでは，本来，最も保護の度合いが弱い顧客についても，個人識別性が十分に困難であることが保証されていることが望ましい．そのため，PWSCUP 2018 では，再識別は攻撃者が考える最も識別しやすい顧客たちを対象に実施することとし，それらの顧客に対して試行された再識別の結果から安全性の指標値を算出することとした．

(2) もまた，前回までのコンテストでの指標の課題の解決を目指すものである．前回までのコンテストでは，安全性と有用性のバランスを取った加工がなされるように安全性と有用性の評価値の和または最大値を全体の指標値としていた．しかし，実用上は安全とみなすことにする一定の基準を満たしているかどうかが重要であり，より安全なデータをより良いデータとみなす前回までの指標は実用上の要求と乖離していた．そのため，PWSCUP 2018 では，安全性に一定の基準を定め，その基準を有意に下回っていると考えられる加工データは安全でないとみなすこととした．

2.3.2 指標

上述の狙いに基づき，具体的に設定した安全性の指標 [4] の概要を説明する．各加工データは，他の参加者から一度ずつの再識別の試行を受け，そのうち一度でも以下で述べる基準を満たす再識別 (**効果的な再識別**と呼ぶ) を受けた場合には，その加工データは安全でないとみなすことにした．再識別の際には，再識別者は加工データの中から任意の人数の顧客を選択し，それらの顧客に対応した元のデータ上での顧客の ID を推定する．選択された人数ごとに顧客数の基準値があらかじめ定められており，基準値以上の人数が当てられた場合はその再識別が効果的な再識別であったと判断する．

2.4 有用性指標

2.4.1 狙い

PWSCUP 2018 の有用性の評価指標を設計するにあたっ

ては，
(1) 一般化による加工を評価できること，
(2) 単一で汎用的な評価指標であること
の 2 点を目指した．

前回までのコンテストでは，加工の方法は，元の値の維持，削除，摂動に限られており，有用性指標で評価可能な加工データはこれらの加工により作られたものに限定されていた．しかしながら，匿名化の主要な加工方法には他に一般化と呼ばれる方法があり，PWSCUP 2018 での有用性指標の設定にあたっては (1) をポイントの一つとした．

(2) は，前回までのコンテストにおける課題であった，加工データが特定のユースケースに特化して作られてしまう問題の解消を図るものである．匿名化によるデータ提供は，分析結果の提供とは異なり，加工データを受け取った者が加工データに対して分析を行うことを想定して行われる．従って，匿名化による加工データは任意の分析に対して有用であることが望ましい．しかしながら，前回までのコンテストでは特定のユースケースでの誤差の最小化に特化した加工データがよい評価値を得られてしまうような有用性指標となってしまっているという問題があった．そこで，PWSCUP 2018 では，(2) により任意のユースケースに対しての誤差の多寡を表現できる有用性指標の設定を目指すこととした．

2.4.2 指標

設計した有用性の評価指標は，次式の通りである．

$$U(A) := \frac{\sum_{i\in[1,m],j\in[2,5]} \mathsf{Err}(T[i,j],A[i,j])}{4m}$$

ここで，T は元データ，A は加工データ，m は T の行数，$\mathsf{Err}(x,y)$ は加工前のセルの値 x と加工後のセルの値 y に対し，x と y の誤差を表す．$\mathsf{Err}(x,y)$ の定義の概要は以下の通りである．

- y が削除されていることを表す値の場合は，1 とする．
- x,y が数値の場合は，x,y を標準化 [2] した値同士の差の絶対値とする．
- x,y が数値でない値の場合は，ハミング距離 ($x=y$ の場合は 0，$x\neq y$ の場合は 1) とする．
- y が (一般化などの結果) 集合の場合は，y から一様ランダムに選択された要素と x との誤差の期待値とする．

定義の詳細は [4] を参照されたい．

狙い (1) に対しては，加工前の値と加工後の集合との間の誤差を定義することにより対応を目指した．(2) に対しては，加工の最小単位であるセルごとの誤差の総和を評価指標とすることで，用途によらない元のデータとの乖離度合いを測ることを目指した．さらに，属性ごとの値のスケールによる違いを吸収するため，標準化を行った．

3. コンテストの結果

PWSCUP 2018 の本戦では，参加登録を行った 14 チームのうち 13 チームが加工データの提出と再識別を行った．提出された各加工データの評価結果を表 3 に示す．13 個の加工データのうち安全性の基準を満たしていると判断された加工データは，効果的な再識別を受けなかった 5 個であった．また，安全性の基準を満たしていると判断された加工データの中で最も良い有用性評価値であった加工データはチーム ID が 14 のチームによる加工データで，有用性評価値は 0.206 であった．

表 3 コンテスト (本戦) の結果

チーム ID	有用性評価値	効果的な再識別
01	0.206	受けた
02	0.365	受けなかった
03	0.294	受けた
05	0.262	受けなかった
06	0.209	受けた
07	0.457	受けなかった
08	0.277	受けなかった
09	0.430	受けた
10	0.193	受けた
11	0.206	受けた
12	0.476	受けた
13	0.255	受けた
14	0.206	受けなかった

4. 安全性指標の設計

加工データ A' と A' に対する再識別 F' について，F' に含まれる顧客の数を n'，F' による再識別の成功数を s とする．このとき，PWSCUP 2018 の安全性の評価では，

$$r(n') \leq s \tag{1}$$

が成り立つときに F' が A' に対する有効な再識別であったと判断していた．以下では，r がどのような根拠に基づいて設定されたのかを説明する．

4.1 安全性の条件 H_0 の設定

PWSCUP 2018 の安全性指標を設計するにあたり，まず，安全である加工データが満たすべき**安全性の条件** H_0 を設定した．具体的には，p を定数とし，加工データ A' が以下の条件 H_0 を満たすとき，A' は安全であるとした．

定義 1 (安全性の条件 H_0) 任意の再識別アルゴリズム L，任意の A' の顧客の部分集合 S について，L によって S のすべての顧客が元データ T 上の対応する顧客へと正しく推定される確率が $p^{|S|}$ 以下である．

H_0 の設定にあたっては，加工データ A' が

- どの仮顧客についても，正しく再識別される確率が十分に低いこと
- 複数の仮顧客が同時に正しく再識別される確率が十分に低いこと

を満たすときに安全となるような設計を目指した．すなわ

ち，どの加工データ上の顧客集合 S についても，S の全員が正しく再識別される確率が十分に低いことを要求するように H_0 を定めた．

4.1.1 条件 H_0 と k-匿名化の関係

例 1 $p = 1/3$ のとき，適切に (equivalent class 内では区別がつかないように) 7-匿名化された A' は H_0 を満たす．加工データにとって最悪のケースは攻撃者に大きさ 7 の eqivalent class を特定されてしまったときだが，そのときでも全員を当てられる確率は $1/(7!) = 1/5040$ であり，この確率は $p^7 = 1/2187$ よりも小さい．

例 2 $p = 1/3$ のとき，6-匿名化された A' は，大きさ 6 の equivalent class を特定されてしまうと H_0 を満たさない．

特定された equivalent class に含まれる顧客の集合を S とすると，ランダムに推定しても S の顧客全員を当てられる確率は $1/(6!) = 1/720$ である．一方，$p^{|S|} = p^6 = 1/729 < 1/720$ であるので，このとき A' は H_0 を満たさない．

4.2 十分条件の設定

加工データ A' が H_0 を満たしていないかどうかは，A' に対する再識別 F' を使って判断する．具体的には，F' により A' が H_0 を満たしていないと結論付けるための**十分条件**である式 (1) を用い，A' と F' が式 (1) を満たすかどうかを判定して式 (1) が満たされた場合に A' が H_0 を満たしていないと判断する．

A' と F' が与えられたときに H_0 が成り立っていないかどうかは，統計的仮説検定によって判断することができる．すなわち，H_0 を帰無仮説とみなし，α を有意水準として，n' 人に対して行った再識別により s 人当たったときに

$$\Pr(n' \text{ 人中 } s \text{ 人以上が当たる}) < \alpha \qquad (2)$$

が成り立つ場合に有意水準 α で H_0 を棄却できる．これは，実際には H_0 が成り立っているにもかかわらず H_0 が成り立っていないと判断する確率が α 未満ということである．

コンテストでは，式 (2) を直接計算することが難しいため，代わりに式 (2) の十分条件を利用する．H_0 が成り立つとき，式 (2) の左辺は

$$\begin{aligned}
&\Pr(n' \text{ 人中 } s \text{ 人以上が当たる}) \\
&= \sum_{k=s}^{n'} \Pr(n' \text{ 人中ちょうど } k \text{ 人が当たる}) \\
&= \sum_{k=s}^{n'} \binom{n'}{k} \Pr\left(\begin{array}{c} n' \text{ 人中 } k \text{ 人だけが当たる} \\ (\text{かつ，残り } n'-k \text{ 人がはずれる}) \end{array} \right) \\
&\le \sum_{k=s}^{n'} \binom{n'}{k} \Pr(k \text{ 人全員が当たる}) \\
&\le \sum_{k=s}^{n'} \binom{n'}{k} p^k
\end{aligned}$$

を満たすので，

$$u(p, n', s) := \sum_{k=s}^{n'} \binom{n'}{k} p^k$$

として，式 (2) の十分条件である

$$u(p, n', s) < \alpha \qquad (3)$$

を判定することにすると，式 (3) が成り立つとき，有意水準 α で H_0 が成り立たないと判断できる．ここで

$$r(n') := \begin{cases} \min\{s \mid u(p, n', s) < \alpha\} & \text{if } \exists s, u(p, n', s) < \alpha, \\ n' + 1 & \text{otherwise} \end{cases}$$

とすると，式 (1) と式 (3) は同値であるので，式 (1) も式 (2) の十分条件になっており，従って式 (1) が成り立つとき，有意水準 α で H_0 が成り立たないと判断できる．

4.3 PWSCUP 2018 で用いたパラメータ

PWSCUP 2018 では，安全性指標の各パラメータは以下のように設定した．

- $p = 1/3$
- $\alpha = 0.01/20$ (有意水準 0.01 相当で，20 回再識別される想定での Bonferroni 補正)

r の具体的な値の一部を表 4 に示す．

5. 安全性指標の評価

PWSCUP 2018 の安全性指標が 2 節で述べた狙いを達成するものであったかどうかを評価する．

5.1 コンテストの結果

本戦では 13 チームが加工データを提出した．これら 13 個の加工データの加工方法による分類と，安全であると判定された (再識別フェーズで有効な再識別を一度も受けなかった) データの個数を表 5 に示す．

5.2 考察

表 5 で安全であると判定された加工データには，一意に特定可能かつ無加工の顧客は一人も含まれていなかった．従って，PWSCUP 2018 では，安全性指標設計の狙い (1) が満たされる結果が得られた．

また，安全であると判定されたデータはすべて “2-匿名化”，“3-匿名化”，“3-匿名化に基づいた加工” のいずれかであった．一方で有効な再識別を一度以上受けて安全でないと判定されたデータは “一意顧客過多”，“2 人無加工+2-匿名化”，“2-匿名化” のいずれかであった．すなわち，“2-匿名化” を境に有効な再識別を一度でも受けたかどうかが分かれる結果が得られた．従って，PWSCUP 2018 では，安全性指標設計の狙い (2) の一部である，一定の基準を下回る加工データは安全でないとみなすことが “2-匿名化” を除いて実現できたことになる．

表 4 $p = 1/3$, $\alpha = 0.01/20$ のときの r の具体的な値の例．表中の*は該当する x については，正解数によらず有効な再識別とならないことを示す．

x	$r(x)$	x	$r(x)$	x	$r(x)$	x	$r(x)$	x	$r(x)$	x	$r(x)$	x	$r(x)$
0	*	10	10	20	16	30	22	40	28	90	59	990	606
1	*	11	10	21	17	31	23	41	29	91	59	991	607
2	*	12	11	22	17	32	23	42	29	92	60	992	607
3	*	13	11	23	18	33	24	43	30	93	60	993	608
4	*	14	12	24	18	34	25	44	31	94	61	994	609
5	*	15	13	25	19	35	25	45	31	95	62	995	609
6	*	16	13	26	20	36	26	46	32	96	62	996	610
7	7	17	14	27	20	37	26	47	32	97	63	997	610
8	8	18	15	28	21	38	27	48	33	98	63	998	611
9	9	19	15	29	21	39	28	49	34	99	64	999	612

表 5 13 個の加工データの内訳

加工方法	該当データ数	生存データ数
一意顧客過多	3	0
2 人無加工+2-匿名化	2	0
2-匿名化	6	3
3-匿名化	1	1
3-匿名化に基づいた加工	1	1

表中の “生存データ数” は再識別フェーズで有効な再識別を一度も受けなかったデータの数を，“一意顧客過多” は，一意に特定可能な顧客が 100 人以上残存している加工を，“2 人無加工+2-匿名化” は 2 人の顧客を無加工で残して残りを 2-匿名化した加工を，それぞれ表す．

5.3 課題

“2-匿名化” を一定の基準と見ると，一定の基準を上回る加工をしたデータと下回る加工をしたデータはそれぞれ安全，安全でないと判定される結果となり，狙いに近い状況が実現できた．しかしながら，“2-匿名化” を行ったデータに対しては判定が分かれることになった．これは，適切に 2-匿名化された加工データに対しては，残りの 13 チームが都合よく再識別を実施できた場合でも，一度も有効な再識別をされない確率が約 0.37 [*1] となり，安全と判定される確率も安全でないと判定される確率も十分に大きくなっていたことに依る．これに対しては，パラメータを変えることで任意の加工に対して常にいずれかの確率が 0 に非常に近づくように調整することができると望ましいが，それは難しい．というものも，加工者が a 人を無加工にして残りを k-匿名化するとき，a と k を変えることによって確率を調整できるためである．

また，“2-匿名化” は安全性指標の設計時の見積もりで得られた境界である 6-匿名化とは大きく離れている．これは十分条件の設定の際の p 値の上界が実際の上界とかけ離れていることに起因すると考えられる．よって，十分条件のよりタイトな解析が必要である．

[*1] 2-匿名化データですべての顧客の候補を 2 人にまで絞り込めた場合，24 人の顧客に再識別を試みて，有効な再識別となる確率は約 0.073 程度であることから計算される．

6. 有用性指標の評価

匿名化データ活用の典型的なユースケースを設定し，PWSCUP 2018 の有用性指標の評価値が各ユースケースでの評価値や各種統計量と相関があると言えるかどうかを見る．

6.1 ユースケース

典型的なユースケースとして，以下の 3 つを設定する．

- 頻出アイテム集合 [7]
- RFM 分析 [1] のための度数表計算 [7]
- アイテムベース協調フィルタリングのための Item-Item 類似度行列を計算の計算 [6]

これらはいずれも過去の PWSCUP での有用性指標として用いたものである．

6.1.1 頻出アイテム集合

このユースケースでは，バスケット分析を行うことを想定し，同一の伝票に含まれる商品の集合に関しての頻出アイテム集合を求める．指標値は，PWSCUP 2016 と同様に，元データに対する頻出アイテム集合を T_0，加工データに対する頻出アイテム集合を T_1 とするとき，

$$|T_0 \cap T_1|/|T_0|$$

とする [7].

6.1.2 RFM 分析の度数表

RFM 分析 [1] では，顧客を R(最後の購買日)，F(勾配頻度)，M(購買額) の 3 つの観点で顧客を分類する．本ユースケースでは，PWSCUP 2016 [7] と同様に，それぞれ 10 ランクに分けた 1000 ランクの度数表を作成することとし，度数表の二乗平均誤差を正規化した値を指標値とする．

6.1.3 アイテムベース協調フィルタリング

ある商品を購入した人に対して別の商品を推薦する際に，アイテムベース協調フィルタリング [3] を使用するユースケースを考える．アイテムベース協調フィルタリングでは，購買履歴から商品と商品の類似度を計算して格納した

Item-Item 類似度行列を作成し，この行列を使って推薦する商品を決定する．

加工データの評価指標として，元のデータと加工データそれぞれから作成される Item-Item 類似度行列の正規化 L_1 距離を使用する．また，類似度行列は，想定するデータセットの利活用状況を想定して，(1) 問屋，(2) 小売，(3) 上位の 3 種類を作成し，それぞれを指標値とする．(1) 問屋の類似度行列は，データ中の顧客，商品ごとに数量の合計 x を $\lfloor x/12 \rfloor$ に置き換えて作成される．(2) 小売の類似度行列は，データ中の顧客ごとに数量の合計が 11 以下の商品のみを使用して作成される．(3) 上位の類似度行列は，データ中で 1 度以上購入した顧客の数が多かった上位 100 の商品のみを使用して作成される．詳細は [6] を参照されたい．

6.2 統計量

以下の各統計量について，元データの統計量と加工データの統計量の誤差を測る．

- 行数
- (購入日，単価，個数の) 総和，分散，標準偏差，平均，最大値，最小値，中央値
- 購入者数が上位 100 の商品集合 [6]

購入者数が上位 100 の商品集合の誤差は，元データと加工データの購入者数が上位 100 の商品集合をそれぞれ U_0, U_1 とするとき，

$$|U_0 \setminus U_1|/100$$

とする．その他の統計量の誤差は，元データの統計量と加工データの統計量の差の絶対値とする．

6.3 評価実験

今回の加工は一般化であり，加工後のデータは集合や範囲になっている．そのため，加工前後の誤差を直接評価することは難しい．この実験では，加工データに矛盾しないデータの集合から一様ランダムにサンプリングしたデータを作成し，サンプリング後のデータと加工前のデータの間で各シナリオでの評価値を計算し，有用性評価値との相関を見る．

6.3.1 データ生成

一般化された加工データから元データと同じ形式のデータを以下のようにサンプリングと補完により作成する．

6.3.1.1 サンプリング

集合に一般化されたセルや削除されたセルの値は，ランダムサンプリングにより，以下のように設定する．

- 範囲に対しては，範囲内から一様ランダムサンプリングを行う．
- アイテム集合に対しては，は集合から一様ランダムサンプリングを行う．
- 削除された行は，そのまま削除されたものとする．
- 購入日・単価・数量の削除は，元のデータ T での最小値と最大値を x_0，x_1 とするとき，$[x_0, x_1]$ から一様ランダムサンプリングを行う．
- 商品の削除は，T に出現するすべての商品の集合から一様ランダムサンプリングを行う．

6.3.1.2 補完

PWSCUP2018 では，時刻や伝票 ID の属性を削除したデータを使用していた．そのため，時刻や伝票 ID を使用するシナリオでは評価値を計算することができない．年度間の差異の吸収のため，削除された属性を以下のように補完する．

- 伝票 ID は PWSCUP 2018 では存在していない．そのため，顧客 ID と購入日が一致するトランザクションを単一の伝票とみなし，顧客 ID と購入日の対ごとに異なる値を割り振ることにする．
- 時刻は PWSCUP 2018 では存在していないため，すべて 0:00 に設定する．

6.3.2 結果の評価方法

各シナリオや統計量ごとに PWSCUP 2018 の有用性指標との間の Pearson の相関係数を計算し，有用性指標との相関があるかどうかを見る．

6.4 結果

各チームごとに加工データから 100 回ずつランダムサンプリングしたデータを作成し，シナリオの評価値と統計量の元データとの誤差の計算を行った．結果を表 6，表 7，表 8，表 9，表 10 に示す．各表の末尾の行は，該当する列の評価値または統計量と PWSCUP 2018 の有用性指標の評価値との Pearson の相関係数を表す．

Pearson の相関係数に関する無相関検定の標本数 13 の場合の有意水準 0.05 での棄却限界となる相関係数の値は $0.552\cdots$ である．従って，ユースケースごとの評価値の誤差はいずれも有意水準 0.05 で有用性指標と相関があると言える．一方，統計量については，いずれも相関係数が計算できるものについては正の値となったものの，有意水準 0.05 で有用性指標と相関があると言えるものは商品集合の類似度，購入日の平均値，単価の最小値だけであった．

6.5 考察

ユースケースの数が限定されてはいるものの，今回あげたユースケースでは有意水準 0.05 で相関があった．しかしながら，統計量の多くとは相関があるとは言えなかった．

今回の評価では，3 つのユースケースしか試せておらず，相関についての理論的な保証も与えてない．そのため，この結果だけでは多くのユースケースでの指標値と相関のある指標だったとは言うことはできない．

表 6 有用性指標の評価値とユースケースごとの指標値の誤差

チーム ID	有用性指標	頻出アイテム集合 [7]	RFM 分析 [7]	類似度行列 (問屋)[6]	類似度行列 (小売)[6]	類似度行列 (上位)[6]
10	0.193	0.208	0.086	0.655	0.759	0.212
01	0.206	0.216	0.085	0.726	0.785	0.221
11	0.206	0.214	0.085	0.726	0.785	0.220
14	0.206	0.217	0.085	0.725	0.785	0.220
06	0.209	0.219	0.089	0.773	0.796	0.225
13	0.255	0.202	0.116	0.723	0.782	0.223
05	0.262	0.209	0.119	0.763	0.802	0.239
08	0.277	0.206	0.100	0.794	0.820	0.238
03	0.294	0.243	0.087	0.805	0.818	0.248
02	0.365	0.367	0.104	0.896	0.885	0.261
09	0.430	0.255	0.097	0.865	0.882	0.271
07	0.457	0.231	0.108	0.864	0.879	0.295
12	0.476	0.600	0.253	0.864	0.885	0.207
相関係数	-	0.654	0.612	0.877	0.949	0.594

表 7 商品 ID の統計量の誤差．商品集合は，購入者数が上位 100 の商品集合の誤差を表す．

チーム ID	行数	商品集合
10	1912	0.114
01	2622	0.108
11	2622	0.109
14	2622	0.109
06	2609	0.120
13	1074	0.123
05	1588	0.110
08	1284	0.117
03	4250	0.123
02	12423	0.176
09	5400	0.151
07	2622	0.145
12	0	0.199
相関係数	0.203	0.856

7. おわりに

本稿では，購買履歴のトランザクションデータに対する匿名加工・再識別コンテストである PWSCUP 2018 の結果を報告し，PWSCUP 2018 で使用した安全性と有用性の指標の評価を行った．評価には参加者から提出された加工データを用いた．評価の結果，有用性指標は今回設定したいずれのユースケースの評価値とも有意水準 0.05 で相関があると言えた．今後の課題は，より多くのユースケースでの検証である．

参考文献

[1] Chen, D., Sain, S. L. and Guo, K.: Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, *Journal of Database Marketing & Customer Strategy Management*, Vol. 19, No. 3, pp. 197–208 (2012).
[2] 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).
[3] Linden, G., Smith, B. and York, J.: Amazon. com recommendations: Item-to-item collaborative filtering, *IEEE Internet computing*, No. 1, pp. 76–80 (2003).
[4] 濱田浩気，荒井ひろみ，小栗秀暢，菊池浩明，黒政敦史，中川裕志，西山賢志郎，波多野卓磨，村上隆夫，山岡裕司，山田　明，渡辺知恵美：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〜，コンピュータセキュリティシンポジウム 2018 論文集 (2018).
[5] 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).
[6] 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，コンピュータセキュリティシンポジウム 2017 論文集 (2017).
[7] 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，コンピュータセキュリティシンポジウム 2016 論文集 (2016).

表 8 購入日の統計量の誤差

チーム ID	総和	平均値	分散	標準偏差	最大値	最小値	中央値
10	1012524.510	0.272	605.936	2.759	0	0	9.705
01	1343419.130	0.896	680.608	3.104	0	0	42.170
11	1343399.980	0.896	680.991	3.106	0	0	41.425
14	1343523.970	0.895	682.613	3.113	0	0	40.860
06	1341776.540	0.833	666.992	3.041	0	0	26.800
13	712427.120	1.478	1622.232	7.551	0	0	2
05	845338.140	0.192	2197.830	10.366	0	0	0.770
08	730917.230	0.398	1602.291	7.455	0	0	13.510
03	2207838.130	1.124	1697.699	7.916	0	0	25.645
02	6407586.200	4.233	1591.520	7.403	0	0	10.360
09	3527877.990	7.168	2956.987	14.204	0	0	40.655
07	1432050.780	0.160	3321.413	16.102	0	0	51.205
12	322002.510	3.607	172.068	0.777	0	0	60.060
相関係数	0.283	0.599	0.512	0.522	-	-	0.442

表 9 数量の統計量の誤差

チーム ID	総和	平均値	分散	標準偏差	最大値	最小値	中央値
10	21504567.050	252.435	14876936.250	3362.200	145.270	0	125.935
01	21044240.330	251.547	14818988.610	3354.846	159.860	0	582.695
11	21065201.260	251.788	14861317.520	3360.155	137.620	0	586.370
14	21095376.590	252.137	14868026.450	3360.991	148.090	0	573.050
06	28764438.240	340.543	19892290.490	3960.336	102.350	0	736.430
13	65563135.010	746.813	42525391.180	6011.711	49.890	1.170	505
05	57837663.360	664.776	37950519.640	5652.333	55.810	0.570	578.380
08	7403998.220	88.337	5319812.400	1835.185	386.180	0	457.290
03	1502531.800	3.302	86581.403	91.098	24727.530	1.020	685.585
02	56553391.610	782.315	44935363.060	6193.330	61.210	2.150	744.400
09	1219318.910	3.971	61114.859	62.456	20839.670	1.750	206.405
07	1499414.970	26.002	40122.198	39.223	19829.820	2.030	551.680
12	1351469273	15137.764	628951946.300	24553.983	2.310	0.910	40837.830
相関係数	0.515	0.516	0.508	0.365	0.502	0.794	0.528

表 10 個数の統計量の誤差

チーム ID	総和	平均値	分散	標準偏差	最大値	最小値	中央値
10	2050787.800	23.742	66213.665	216.073	3.340	0	5.845
01	2022186.970	23.707	66921.960	217.424	3.200	0	13.035
11	2014761.980	23.622	66641.357	216.886	2.480	0	14.125
14	2015408.170	23.629	66686.529	216.974	3.100	0	4.990
06	2023404.810	23.716	67196.689	217.947	2.680	0	5
13	4422427.840	50.288	134496.037	324.331	1.230	0	7.500
05	3804329	43.606	118416.545	301.894	1.410	0	5.220
08	4903499.120	55.905	149734.975	344.407	1.110	0	3.280
03	82847.060	0.361	683.784	8.367	1063.160	0	8
02	4343540.450	58.501	155247.526	351.419	1.100	0	15.540
09	127153.770	2.306	110.968	1.244	916.310	0	2.345
07	207315.260	2.764	116.936	1.293	860.760	0	2.710
12	61802525.510	692.248	1321850.718	1105.441	0	0	1904.295
相関係数	0.500	0.502	0.481	0.281	0.507	-	0.527