

PWS CUP 2016 匿名加工・再識別コンテスト

競技ルール Ver. 1.3

本コンテストは、匿名加工の技術と再識別のリスクの評価技術を競うものである。基本ルールの理解とサンプルコードなどの利用方法を習得するための予備戦と限られた時間で効率的に匿名加工と再識別を行う本戦から構成されている。

使用するソフトウェアや OS には制限を加えない。参加者は自分の実験環境を会場に持参する。ネットワークに繋いでもよい。

本ルールの記号やアルゴリズムの詳細、およびデータセットは、次の文献にて与えられている。

[1] 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」, プライバシーワークショップ 2016.

[2] UCI Machine Learning Repository, Online Retail Data Set (英国のオンラインショップにおける 1 年間の購買履歴) <https://archive.ics.uci.edu/ml/datasets/Online+Retail>

コンテストのルールは次の通り。なお、以下のルールは合理的で公平なコンテストの実施の為に、予告なく変更することがある。

1. (プレイヤー) 匿名加工者, 再識別者, 審判員の 3 者が係る。
2. (匿名加工者) 匿名加工者は, オリジナルの顧客マスターデータ M と購買履歴データ T を与えられ, 匿名加工したマスターデータ M' と購買履歴データ T' と, M と M' のレコードの対応を表した行番号データ P を生成する。再識別者に M', T' を, 審判員に $M', T',$ と P を提出する。
3. (再識別者) 再識別者は, 顧客マスターデータ M と購買履歴データ T を参照して, 匿名加工された M' と T' から推定した推定行番号データ Q を審判員に提出する。
4. (匿名加工の勝者) 最も有用性が高く, 最も安全な匿名加工データを提出した匿名加工者を勝者とする。有用性と安全性を総合して, 勝者を決定する。
5. (再識別者) M' の全レコード数に対する P と等しい Q のレコードの数の比をその匿名加工データの再識別アルゴリズム E による再識別率 Re-id^E と呼ぶ。再識別を行うことで他の参加者の匿名加工データの安全性を下げることができる(再識別の勝者は設けない。どのチームのデータを再識別してもよく, 全てのデータを再識別しなくてもよい)。
6. (過加工の除外) 匿名加工データ M', T' について次の条件を過加工(山岡匿名化)とみなし, 受け付けない。

(1) Y1(subset)

$$Y1(M, M', T, T', P) = \max_{X' \subset M', D} (|\mu_{X', D}(t'^6 t'^7) - \mu_{p(X'), D}(t^6 t^7)|)$$

と定める Y1 について、 $Y1 > 50,000$ となること．ここで、 X' は匿名加工顧客マスターデータ M' の大きさ 10 の部分集合、 D は購買履歴データ T における任意の連続する 30 日間であり、 $\mu_{X', D}(t^6 t^7)$ は D における X' の顧客ごとの購買総額（単価と個数の積の D の総和）の平均値である．また、 p は匿名加工の際に行ったレコードの置換を表す行番号であり、 $p(X')$ で p により X' の要素と対応付けられる M の要素の集合を表すこととする．すなわち、もしも山岡匿名加工を行っているとき X と $p(X')$ が異なる顧客集合になるので、購買総額の平均値が大きく変化する．(Ver. 1.3 にて、Y1 の閾値が 5,000 から 50,000 に引き上げられた)

(2) Y2(Jaccard)

$$Y2(M, M', T, T', P) = 1 - \frac{1}{n'} \sum_{x=1}^{n'} \frac{|S_{p(x)} \cap S'_x|}{|S_{p(x)} \cup S'_x|}$$

と定める Y2 について、 $Y2 > 0.7$ となること．ここで、 $S_{p(x)}$ 、 S'_x は、顧客 x が購買した T と T' について記録されている全ての商品の多重集合である．すなわち、

$$S_{p(x)} = \{t^5, \dots, (t^7 \text{ 回}) \dots, t^5 \mid (t^1, \dots, t^5, \dots, t^7) \in T, t^1 = p(x)\},$$

$$S'_x = \{t^5, \dots, (t^7 \text{ 回}) \dots, t^5 \mid (t^1, \dots, t^5, \dots, t^7) \in T', t^1 = x\},$$

とする．

7. 有用性総合評価

$$U = \text{Max} (U1 (Cmae1), U2 (Cmae2), U3 (RFM), U4(\text{TopItem}))$$

ここで、

$$U1 = Cmae1(M, M', T, T')$$

M の {c² 性別, c⁴ 国名} でクロス集計した T の平均単価（購買数を考慮）と M' の {c² 性別, c⁴ 国名} でクロス集計した T' の平均単価（購買数を考慮）の平均絶対誤差 MAE

$$U2 = Cmae2(M, M', p, T, T')$$

M の {c² 性別, c⁴ 国名} でクロス集計した T の平均単価（購買数を考慮）と M の {c² 性別, c⁴ 国名} と行番号 p でクロス集計した T' の平均単価（購買数を考慮）の平均絶対誤差 MAE

$$U3 = \text{RFM}(M, M', T, T')$$

M, M' の顧客を、それぞれ T, T' の Recency (最後の購買日), Frequency (購買頻

度), Monetary (購買額)の3つの条件でクラス分けした顧客数の二乗平均平方根誤差 RMSE を(最大値で割ることで0~1に)正規化した値

$U4 = \text{Topitem}(T, T')$

出現頻度の高い商品の頻出集合の T と T' の相対誤差 (1- 正規化した積集合の大きさ)

とする. 各式の算出については, [1]を参照されたい.

8. 安全性評価

$E = \text{Max}$ (以下のサンプルと再識別者による再識別の再識別率)

$E1\text{-birthday}$ = 生年月日同士の距離が最小となる顧客 ID に再識別

$E2\text{-eqi}$ = マスターの属性(仮 ID を除く)とトランザクションが完全一致するレコードを推測. なければランダム

$E3\text{-sort}$ = (性別, 生年月日, 国)でソート

$E4\text{-sort2}$ = 生年月日でソート

$E5\text{-recnum}$ = レコード数マッチング (トランザクションのレコード数同士の距離が最も近い顧客に再識別)

$E6\text{-eqtr}$ = トランザクションが完全一致するレコードを推測. なければランダム

$E7\text{-tnum}$ = トランザクション数でソート

$E8\text{-meantime}$ = 平均購入時刻同士の距離が最小となる顧客 ID に再識別

$E9\text{-re}$ = 常に 1, 2, 3, ..., $|M|$ と推測

$E10\text{-tnum-bi}$ = (トランザクション数, 生年月日)でソート

$E11\text{-totprice}$ = 総価格同士の距離が最小となる顧客 ID に再識別

$E12\text{-cid}$ = ID が一致する行番号を(複数あればランダムに選んで)出力. 存在しなければ[1, $|M|$]の乱数を出力

$E13\text{-random}$ = [1, $|M|$] の乱数を $|M|$ 個重複ありで出力

サンプル再識別アルゴリズムの一部は, [1]で定めている.

9. (総合評価)

予備戦の順位と本戦の順位を, 1:9 の割合で合計して総合評価とする.

同点の場合は同順位とする.

10. (匿名加工者の禁止事項) 匿名加工者の次の行為を禁じる. なお, 次の行為はいずれもシステムに拒絶されるため, その行為が理由で失格になるようなことは起きない.

- (1) チームで上限を超える匿名加工データを提出すること. 上限は予備戦では3個まで, 本戦では1個までとする. ただし, 何回でも匿名加工データの再提出(差し替え)が認められている.

- (2) 行番号データ P が一意でない（同じ行番号データを複数用いてはならない．ただし，全行番号を含める必要はなく，いわゆる行削除は認める）
- (3) データセットの書式「PWS CUP 2016 マスターデータの書式」，「PWS CUP 2016 トランザクションデータの書式」に従わない加工データ M' , T' を提出すること．
- (4) 次の範囲を超えた匿名加工データ T' や M' を提出すること．

$$|T|/2 \leq |T'| \leq 2|T|$$

$$10 \leq |M'| \leq |M|$$

ここで， $|T|$, $|M|$ は元の顧客データ，履歴データの大きさ（行数）， $|T'|$, $|M'|$ は匿名加工した対応するデータの大きさである．

- (5) 匿名加工された履歴データ T' と顧客マスター M' の仮名 ID が整合しないこと．（ T' のある仮名 ID t'_i が M' のいかなる仮名 ID c'_j にも一致しないこと）．
- (6) 同じ伝票 ID t^2 が異なる履歴データに割り当てられて矛盾した，すなわち，
伝票 ID は同じ $t^2_i = t^2_j$ かつ
(仮名 ID が異なる $t^1_i \neq t^1_j$ ，または，
購買日が異なる $t^3_i \neq t^3_j$)
となる t^1_i と t^1_j が存在する履歴データ T' を加工すること．
- (7) 履歴データ T に含まれない商品 ID t^5 を含む履歴データ T' を加工すること．（単価 t^6 ，数量 t^7 はこの制約はない）
- (8) 仮名 ID に重複がある顧客データ M' を加工すること．

11. （再識別者の禁止事項）再識別者の次の行為を禁じる．

- (1) 匿名加工者と結託すること（行番号データなどを教えてもらうこと）．
- (2) 不正な形式の推定行番号データ，あるいは M の行数と異なる推定行番号データ Q を提出すること．（ただし， Q は一意でなくてもよい．推定行番号データの形式は行番号データの形式と同じで，各行に行番号（1, 2, ...）1 つを記載したテキストファイルである．）
- (3) 一つの匿名加工データに対して，11 回以上推定行番号データを提出して，再識別を試みること．

12. （審判員の禁止事項）審判員の次の行為を禁じる．

- (1) 匿名加工者や再識別者と結託すること（審判員の特権により知った情報（行番号データなど）を教えること）．
- (2) PWS CUP 実行委員会委員として，匿名加工者や再識別者がそれを知ること

ンテストで有利になるような情報を非公開にすること

- (3) コンテスト参加者として匿名加工者や再識別者を兼ねる場合、データ提出受付期間中に審判員の特権を使うこと（他チームの行番号データなどを知ること）以上の禁止行為が守られている条件の下で、PWS CUP 実行委員会委員のコンテストへの参加を認める。

13. (本戦のルール)

- (1) 用いるデータセットは、予備戦と同一(Master-Customer400.csv, Transaction-Customer400.csv)とする。
- (2) 匿名加工データは1チームにつき1つのみ提出する。
- (3) 7で定められた有用性と8で定められた安全性の和、U+Eで順位付けをする。
- (4) 6の(1), (2)の条件で検出された過加工は除外する。
- (5) 再識別による各チームで得るポイントはない。(競合する他のいくつかのチームの安全性を下げる効果がある)
- (6) 9で定められた基準で総合順位を定める。本戦の出場をキャンセルしたチームの本戦の順位は最下位とする。

2016年8月2日 Ver 0.1

2016年8月24日 Ver 1.0

2016年9月6日 Ver 1.1

2016年9月27日 Ver. 1.2

2016年10月3日 Ver. 1.3

PWS CUP 実行委員会