PWS CUP 2017 匿名加工・再識別コンテスト競技ルール

PWS CUP 実行委員会

Ver. 1.3
2017 年 10 月 17 日

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，中川，野島，波多野，濱田，村上，門田，山岡，山田，渡辺，「PWSCUP2017: 長期間の履歴データの再識別リスクを競う」，プライバシーワークショップ 2017, 情報処理学会, 2017.
[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．

コンテストルール

1. プレイヤーとして，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，購買履歴データから個人が特定されないための匿名化アルゴリズム A を作成する．A では，顧客 ID の仮名 ID への振替，日付や商品のランダムな変更，レコード (行) の消去などが行われる．そして，A によって加工された**匿名加工履歴データ**$A(T)$ を審判員に提出する．
3. (審判員) 審判員は，オリジナルの購買履歴データ T と，匿名加工者によって提出された匿名加工データ $A(T)$ から，顧客 ID と仮名 ID の対応関係を示す**仮名表**f を作成する．さらに，$A(T)$ の行を日付などでソートし縮約した**匿名化データ**S を作成する．審判員は S を再識別者に配布する．
4. (再識別者) 再識別者は，審判員から受け取った S と購買履歴データ T の一部である T_{α_1}, T_{α_2}, T_{α_3}, T_{α_4} (後述) を参照し, それぞれに対応する**予測仮名表**$\hat{f}$ を作成し，審判員に提出する．
5. (データセット)
   (a) 購買履歴データ T, 顧客マスターデータ M は, [2] からサンプリングして与える．M には, $n = 500$ 名の顧客が存在する．データセットの詳細な加工方法については [1] に示す．
   (b) 購買履歴データ T，及び顧客マスターデータ M は，PWSCUP2017 のホームページ (`https://pwscup.personal-data.biz/web/pws2017/index.php`) から配布する．
   (c) データセットの再配布は禁じないが，それを用いた論文を発表する時は，[1], [2] を参考文献として引用する．
   (d) 購買履歴データ T は，購買月によって，$T^{(1)}, \ldots, T^{(12)}$ に分割される．
   (e) 再識別者には，購買履歴データの行をサンプリング率 $\alpha_1 = 0.25, \alpha_2 = 0.5, \alpha_3 = 0.75, \alpha_4 = 1.0$ で

ランダムサンプリングした履歴データ $T_{\alpha_1}, \ldots, T_{\alpha_4}$ を部分知識として提供する．また，商品番号は上 2 桁だけを抽出した形に加工されている．

6. (規則第 19 条）「個人情報の保護に関する法律施行規則」(以下，規則) 第 19 条は，扱うデータセットや採用するアルゴリズムによって，様々な解釈が考えられる．本コンテストでは，規則第 19 条の多様な解釈や，それに対する対処方法について，多くの意見を得ることを目的としているので，次のように定める．
   (a) 3 項（「連結する符号」の削除）匿名化された S には，M のいかなる顧客 ID$c_{\cdot,1}$ とも一致する仮名 $s_{\cdot,1}$ が存在しない様にする
   (b) 1 項（個人情報の削除），4 項（特異な記述等の削除），5 項（適切な措置）の解釈と対処については各チームに委ねる．どのような対処を行なったのかは，最終プレゼンテーションにて発表する．このプレゼンテーションの内容は，総合順位には影響しない．
7. (加工フォーマット)
   (a) 購買履歴データ T のうち，削除する行は以下の例の様に書き換える．

$$\texttt{[17551,0,2010/12/15,14:12,22693,1.25,24]} \rightarrow \texttt{[DEL,,,,,,]}$$

   すなわち，列数は変更しない．顧客 ID に，“DEL” が指定されている行の他の値は無視する．
   (b) 期間データ $T^{(\ell)}$ に対する匿名化されたデータを $A(T^{(\ell)})$ とする．全期間データの匿名加工履歴データを，

$$A(T) = \begin{pmatrix} A(T^{(1)}) \\ \vdots \\ A(T^{(12)}) \end{pmatrix}$$

   とする．
   (c) 匿名加工履歴データ $A(T)$ の行の順序は，購買履歴データ T と同一とする．
   (d) 匿名加工履歴データ $A(T)$ には，購買履歴データ T に存在しない架空の行を新たに加えてはならない．
8. (仮名割り当て)
   (a) 加工は期間ごとに行われるため，期間を超えた変更は認めない．
   (b) 仮名は期間内では矛盾のない様に割当てる．（期間内では同一顧客の仮名は 1 つ）
   (c) 期間によって仮名は必ずしも変更しなくてよい．（期間を超えて同一人物に同じ仮名をつけても良い）
   (d) “DEL” という仮名は禁じる．
9. (匿名加工者の禁止事項) 匿名加工者の次の行為を禁じる．
   (a) チームで上限を超える数の匿名加工データを提出すること．上限は予備戦では 3 個，本戦では 1 個とする．ただし，匿名加工データの再提出（差し替え）は何回でも認められている．
   (b) データセットの書式「PWS CUP 2016 トランザクションデータの書式」及び, WEB サイトでの指定条件に従わない匿名加工データ等を提出すること．
   (c) 次の範囲を超えた匿名加工データ $A(T)$ を提出すること．

$$|T| = |A(T)|$$

$$|T|/2 < |S|$$

ここで，$|T|$, $|A(T)|$，$|S|$, は購買履歴データ，匿名加工履歴データ，及び削除行を除いた後の匿名加工履歴データの大きさ（行数）である．

（d）購買履歴データ T に含まれない商品 ID $(t_{\cdot,5})$ を含む匿名加工履歴データ $A(T)$ を加工すること．（単価 $t_{\cdot,6}$ ，数量 $t_{\cdot,7}$ はこの制約はない）

10. (匿名化データの生成）審判員は，仮名化，その他の項目のランダム化などを行なった履歴データ $A(T)$ に対して，次のようにして匿名化データ S を生成し，再識別者に提供する．

（a）$[DEL, , , , , ,]$ と指定されている行を消去し，さらに行をランダムに並び替える．各履歴データの長さは $|T| = |A(T)| \geq |S|$ となる．

（b）T と $A(T)$ の関係から，仮名表

$$F = \begin{pmatrix} c_{1,1} & f^{(1)}(c_{1,1}) & \cdots & f^{(12)}(c_{1,1}) \\ \vdots & & \ddots & \vdots \\ c_{n,1} & f^{(1)}(c_{n,1}) & \cdots & f^{(12)}(c_{n,1}) \end{pmatrix}$$

を求めて，秘密に管理する．ここで，$f^{(\ell)}(c_{i,1})$ は，期間 ℓ に，i 番目の顧客 ID $c_{i,1}$ に割り当てた仮名である．ただし, 仮名が存在しない場合は, DEL と記載する．

11. (有用性評価) 匿名化データ S の有用性は，

$$\mathrm{util}(S) = \max_{i=1,\ldots,6} E_i(S)$$

とする．ここで，E_1, E_2, E_3 は，[1] の 3.6.1 節で定めるアイテムベース類似度である．E_4, E_5 は T, $A(T)$ との間の購入日の差の平均と単価の (比率) 差の平均である．E_6 は，$|T|$ の行数における $A(T)$ で消去された行数の割合である．

12. (再識別）再識別者は，S と部分知識 T_α について評価した，推定仮名表 $\hat{F}$ を作成する

$$\hat{F} = \begin{pmatrix} c_{1,1} & \hat{f}^{(1)}(c_{1,1}) & \cdots & \hat{f}^{(12)}(c_{1,1}) \\ \vdots & & \ddots & \vdots \\ c_{n,1} & \hat{f}^{(1)}(c_{n,1}) & \cdots & \hat{f}^{(12)}(c_{n,1}) \end{pmatrix}$$

13. (仮名の再識別率)

（a）本指標は F の $12n$ 組の対応のうち，DEL を含めた 12 ヶ月分の仮名が全て正しく推定できたユーザの割合である．すなわち，指標値は，

$$\mathrm{reid}(F, \hat{F}) = \frac{|\{\forall \ell \in \{1, \ldots, 12\}\ f^{(\ell)}(c_{i,1}) = \hat{f}^{(\ell)}(c_{i,1})\}|}{n}$$

と定める．

（b）部分知識 T_α についての再識別率を $\mathrm{reid}(F, \hat{F_\alpha})$ とする．全部分知識の総合の再識別スコアは，それらの総和，すなわち，

$$\mathrm{reid}(F, \hat{F_*}) = \sum_{\alpha \in \{\alpha_1, \alpha_2, \alpha_3\, \alpha_4\}} \mathrm{reid}(F, \hat{F_\alpha})$$

とする．

14. (安全性評価) 匿名化データ S の安全性は，

$$\mathrm{reid}(S) = \max_i S_i(S)$$

とする．ただし，S_i は次の再識別アルゴリズムとする．$S_1 \sim S_6$ の各アルゴリズムは，それぞれ以下に示す複数の属性の組み合わせが，匿名加工前後で等しいレコード同士を同じ顧客とみなすものである．

S_1-datenum	購入日，数量
S_2-itemprice	商品 ID(2 桁)，単価
S_3-itemnum	商品 ID(2 桁)，数量
S_4-itemdate	商品 ID(2 桁)，購入日
S_5-item2pricenum	商品 ID(2 桁)，単価，数量
S_6-item2datenum	商品 ID(2 桁)，購入日，数量
その他	他のチームによる任意のアルゴリズム

15. (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．
16. (匿名加工の勝者) 最も有用性が高く，最も安全な $A(T)$ を提出した匿名加工者を勝者とする．
17. (審判員の禁止事項) 審判員の次の行為を禁じる．
    (a) 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)．
    (b) 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でコンテストで有利になるような情報を非公開にすること
    (c) 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，PWSCUP 実行委員会委員のコンテストへの参加を認める．
18. (再識別者の禁止事項) 再識別者は次を行ってはならない．
    (a) 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)．
    (b) 不正な形式の予測仮名表 $\hat{f}$ を提出すること．(1 列目を顧客 ID とする n 行 13 列の CSV ファイルとする．異なる顧客に同じ仮名を推定しても良い．仮名はその月に存在するものまたは，DEL のみとする)
    (c) 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．(10 回まではよい)
    (d) 部分知識 T_{α_i} を利用した再識別を行う際に，T_{α_i} 以外の背景知識（例えば，Online Retail の元データ，購買履歴データ T，他の部分知識 T_{α_j}（$j \neq i$)）を使って再識別すること．
19. (ソフトウェア，ネットワークなど) 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．
20. (本戦のルール)
    (a) 予備戦とは異なる 500 名の購買履歴データ T, 顧客マスターデータ M を事前に配布する．
    (b) 再識別に用いる部分知識 $T_{\alpha_1}, \ldots, T_{\alpha_4}$ は，本戦の会場で配布する．
    (c) 匿名加工データは 1 チームに付き 1 つのみ提出する．
    (d) ルール 11 で定められた有用性とルール 13 で定められた安全性の和，$U + E$ で順位付けをする．
    (e) ルール 15 で定められた基準で総合順位を定める．そして，上位 3 チームを**総合順位 1,2,3 位**として表彰する．また，総合順位 1 位チームの匿名加工データに対する再識別スコアが最も高いチームを**再識別賞**として表彰する．同点の場合は，その全チームを表彰する．
    (f) 総合順位と再識別スコアに基づき，2F2: PWS Cup FINAL での発表者を選定する．PWS CUP

実行委員会が若干名の推薦枠を持つ.

(g) ルール 6 の加工規則を十分に検討し，最も適切な措置を行ったチームを**匿名加工基準賞**として表彰する．2F2: PWS Cup FINAL でのプレゼンテーションに基づき，参加者による投票により選出する.

変更履歴

- Ver 1.0, 2017 年 9 月 11 日
- Ver 1.1, 2017 年 9 月 29 日，5 (e) データセット T_α，13 (再識別率)，14 （安全性評価）更新.
- Ver 1.2, 2017 年 10 月 6 日，10, 12 F の形式，13 (b) 追加，18 （再識別者禁止行為）追加
- Ver 1.3, 2017 年 10 月 16 日，5 条（データセット）(e) 「どの期間データ T_α にも必ず n 人」という制約を削除，$\alpha_4 = 1.0$ を追加．9 条 (匿名加工者の禁止事項)「システムに拒絶」の記述を削除. 20 条 (本戦ルール）の追加.