PWS Cup 2018 競技ルール Ver 1.2

PWS CUP 実行委員会

2018 年 9 月 27 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．

- CSS 2018 に投稿された PWS Cup 2018 のルール論文 [1] の改変版である「PWS Cup 2018 コンテスト設計」 [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) [3]

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**トランザクション** T を与えられ，T から個人が特定されないように加工した**加工トランザクション** A を作成する．加工の際には，顧客 ID の仮名への振替，日付や商品の一般化，レコード (行) や値の消去などが行われる．そして，A を審判員に提出する．
3. (審判員) 審判員は，オリジナルのトランザクション T と匿名加工者によって提出された加工トランザクション A から，A の行を辞書順で整列し縮約した**公開加工トランザクション** A' を作成する．審判員は A' 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 A' と T を参照し, それぞれに対応する**推定対応表** F' を作成し，審判員に提出する．
5. (データセット)
   (a) トランザクション T は，[3] からサンプリングして与える．T には，$n = 1000$ 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 [2] に示す．
   (b) トランザクション T は PWS Cup 2018 のホームページ [4] 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，[2], [3] を参考文献として引用する．
6. (加工フォーマット)
   (a) トランザクション T のうち，削除する行は以下の例の様に書き換える．

   `17551,2010/12/15,22693,1.25,24` → `*,*,*,*,*`

   すなわち，行数は変更しない．

(b) 加工トランザクション A の行の順序は，トランザクション T と同一とする．

(c) 加工トランザクション A には，トランザクション T に存在しない架空の行を新たに加えてはならない．

7. (仮名割り当て)
   (a) 仮名は矛盾のない様に割当てる．(同一顧客の仮名は 1 つ)
   (b) 異なる顧客に対して同じ仮名を割り当てることは禁じる．
8. (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 1 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 2018 データの書式」及び「PWS Cup 2018 コンテスト設計」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) 製品 ID を要素数が 100 を超える集合に一般化すること．(すなわち，「PWS CUP 2018 データの書式」のパラメータ e は $e = 100$ とする．)
   (d) 次の条件を満たさない加工トランザクション A を提出すること．

$$|T| = |A|$$

   ここで，$|T|, |A|$ はトランザクション，加工トランザクションの大きさ (行数) である．
   (e) トランザクション T に含まれない商品 ID ($T[\cdot, 3]$) を含む加工トランザクション A を作成すること．(単価 $T[\cdot, 4]$，数量 $T[\cdot, 5]$ はこの制約はない)
   (f) 他の匿名加工者や再識別者と結託すること (A や A' などを教えてもらうこと)．ただし，A や A' が推定できない範囲で，T の分析情報や，プログラムモジュールを共有すること等は，結託に当たらないとする．
9. (匿名化データの生成) 審判員は，匿名加工者が仮名化，その他の項目の一般化などを行なった加工トランザクション A に対して，次のようにして公開加工トランザクション A' を生成し，再識別者に提供する．
   (a) `*,*,*,*,*` と指定されている行を消去し，さらに行を辞書順に並び替える．各履歴データの長さは $|T| = |A| \geq |A'|$ となる．
   (b) T と A の関係から，**対応表**

$$F = \begin{pmatrix} \text{仮名 } i_1 & \text{顧客 ID} j_1 \\ \vdots & \vdots \\ \text{仮名 } i_{n'} & \text{顧客 ID} j_{n'} \end{pmatrix}$$

   を求めて，秘密に管理する．ここで，n' は A に含まれる仮名の個数である．F の 1 列目は A に含まれる仮名をちょうど 1 度ずつ並べた列ベクトルである．F の 2 列目の各要素は 1 列目の仮名に対応した顧客 ID である．
10. (有用性評価) 公開加工トランザクション A' の有用性は，

$$U(A') = \frac{\sum_{i \in [1,m], j \in [2,5]} \mathsf{Err}(T[i,j], A[i,j])}{4m}$$

とする．ここで，A は A' に対応した加工トランザクション，$\mathsf{Err}(\cdot, \cdot)$ は [2] で定める要素間の誤差である．

11. (再識別) 再識別者は，A' と T について評価した，**推定対応表**

$$F' = \begin{pmatrix} \text{仮名 } i_1 & \text{顧客 ID} j_1 \\ \vdots & \vdots \\ \text{仮名 } i_{n'} & \text{顧客 ID} j_{n'} \end{pmatrix}$$

を作成する．ここで，n' は再識別者が再識別を試みる顧客の人数である．

12. (仮名の再識別成功人数) 推定対応表 F' による公開加工トランザクション A' に対する再識別成功人数は，

$$\mathsf{Suc}(F, F') = |\{i \mid F'[i,\cdot] = F[j,\cdot]\}|$$

である．ただし，A は A' に対応した加工トランザクション，F は A の対応表である．

13. (安全性評価) 公開加工トランザション A' の安全性評価値 $S(A')$ は，

$$S(A') = \begin{cases} 1 & \text{if } \exists F', r(n') \leq \mathsf{Suc}(F, F'), \\ 0 & \text{otherwise} \end{cases}$$

とする．ただし，F は A' に対する対応表，F' は A' に対する推定対応表，n' は F' の行数，$r(\cdot)$ は再識別を試みた人数 x ごとの安全性の基準を満たしていないと判断する再識別成功人数の下限 $r(x)$ を与える関数である．$S(A') \neq 0$ である公開加工トランザクション A' は安全性の基準を満たしていないとみなす．

14. (匿名加工の順位付け) $S(A') = 0$ である A' の中で，有用性が高い (U が小さい) 順に順位付けを行う．ただし，予備戦の順位付けでは，予備戦で A の差し替えを行ったチームは順位付けの際に有用性の評価値を 0.1 大きいとみなして順位付けを行う．$S(A') \neq 0$ である A' は最下位 (A を提出しなかった場合，A が指摘条件に従っていなかった場合と同等) とする．

15. (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 (加工トランザクションなど) 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 (他チームの加工トランザクションなどを知ること)

    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．

16. (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) 他の匿名加工者や再識別者と結託すること (F や F' などを教えてもらうこと)．ただし，F や F' が推定できない範囲で，プログラムモジュールを共有すること等は，結託に当たらないとする．
    (b) 推定対応表の書式「PWS CUP 2018 推定対応表の書式」及び「PWS Cup 2018 コンテスト設計」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) 一つの公開加工トランザクションに対して，2 回以上推定対応表を提出して，再識別を試みること．(1 回まではよい)

17. (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．

18. (本戦のルール)
    (a) 予備戦とは別に作成された 1000 名のトランザクション T を事前に配布する．

19. (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．発表を行ったチームを対象に総合評価で順位付けを行い，総合順位とする．同点の場合は同順位とする．発表を行わなかったチームの総合評価は参考記録とし，総合順位は与えられない．
20. (再識別スコア) 提出した推定対応表のうち，

$$r(n') \leq \mathsf{Suc}(F, F')$$

を満たした推定対応表 F' の数を，**再識別スコア**とする．ただし，n' は F' の行数，F は A' に対する対応表，A' は F' の推定対象である公開加工トランザクション，$r(\cdot)$ は再識別を試みた人数 x ごとの安全性の基準を満たしていないと判断する再識別成功人数の下限 $r(x)$ を与える関数である．
21. (総合再識別スコア) 予備戦の再識別スコアと本戦の再識別スコアを 1:9 の割合で合計して**総合再識別スコア**とする．発表を行ったチームを対象に総合再識別スコアで順位付けを行い，**総合再識別順位**とする．同点の場合は同順位とする．発表を行わなかったチームの総合再識別スコアは参考記録とし，総合再識別順位は与えられない．
22. (表彰) 発表を行ったチームのうち，いくつかのチームを表彰する．
    (a) ルール 19 で定められる総合順位が上位の 3 チームを**総合優勝，総合第 *2* 位，総合第 *3* 位**として表彰する．
    (b) ルール 21 で定められる総合再識別順位が最も上位のチームを**再識別賞**として表彰する．
    (c) 最もよいプレゼン・ポスター発表を行ったチームを**優秀発表賞**として表彰する．

変更履歴

- Ver 1.2, 2018 年 9 月 25 日．再識別者の禁止事項，表彰，匿名加工者の禁止事項を更新．本戦のルールを追加．
- Ver 1.1, 2018 年 9 月 13 日．再識別者の禁止事項，匿名加工の禁止事項，を更新．再識別スコア，総合再識別スコア，を追加
- Ver 1.0, 2018 年 9 月 5 日，公開

参考文献

[1] 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 ～. In *CSS*, 2018.

[2] PWS Cup コンテスト設計. `https://www.iwsec.org/pws/2018/cup18.html`.

[3] UCI Machine Learning Repository. `https://archive.ics.uci.edu/ml/datasets/Online+Retail`. Accessed: 2018-08-15.

[4] PWS Cup 2018 ホームページ. `https://www.iwsec.org/pws/2018/cup18.html`. Accessed: 2018-08-15.