PWS Cup 2018 安全性評価の詳細

2018 年 9 月 5 日

1 r の定義

関数 r は別ファイル「R.csv」(以下では R と呼ぶ) により定義する．R は各行に 2 つの整数がカンマ区切りで書かれたファイルで，

```
a,b
```

と書かれた行は

$$r(a) = b$$

であることを表す．$r(n') > n'$ である場合は，n' 人の仮顧客に対する再識別では n' 人全員を正しく再識別しても安全性の基準を満たさないとは判定されないことを表す．

次節では r の決め方の説明を行うが，R による定義と食い違っていた場合には，コンテストでは R による定義を優先する．

2 参考: r の設定の背景

今回のコンテストの安全性の評価では，公開加工トランザクション A' に対する再識別の推定対応表 F' が，F' に含まれる仮顧客の数を n'，F' による再識別の成功数を s とするとき，

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

であるときに A' が安全でないと判断する．本節では，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．

2.1 概要

今回のコンテストでは，まず，安全である公開加工トランザクションが満たすべき**安全性の条件** H_0 を設定する．H_0 を満たしていないと判断された公開加工トランザクションは安全でないと判断される．

公開加工トランザクション A' が H_0 を満たしていないかどうかは，A' に対する再識別の推定対応表 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．

2.2 安全性の条件

今回のコンテストでは，公開加工トランザクション 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す．この方針に則り，どの仮顧客集合 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安全性の条件とする．

より正確には，p を定数とし，公開加工トランザクション A' が以下の条件 H_0 を満たすとき，A' は安全であるとする．

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' の仮顧客の部分集合 S について，L によって S のすべての仮顧客が対応するトランザクション T の顧客へと正しく推定される確率が $p^{|S|}$ 以下である．

2.2.1 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.

最悪のケースは大きさ 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，$p^7 = 1/2187$ 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 に含まれる仮顧客の集合を S とすると，ランダムに推定しても S 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.

2.3 十分条件の設定

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

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

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

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

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

$$u(p, n', s) < \alpha \tag{3}$$

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.

2.4 パラメータの設定

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 に 20 回再識別される想定での Bonferroni 補正)