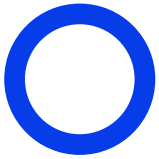
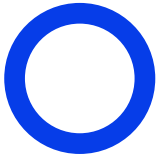
2016年度PWSCUP コンテストの計画について

菊池浩明 （明治大学）
PWS実行委員会 CUP WG

(復習)2015年度のコンテストの方針

- 基本方針
  - ❑多くのチームが実力を発揮出来る様にする
    - » 評価の技巧や難易度は妥協し，極力シンプルで分りやすいルールと環境とする
  - ❑競技者だけではなく，聴衆に対しても楽しいコンテストとする.
    - » 実行委員の参加を認めてもよい公平なルール.
  - ❑競技の後に，その結果を有益に活用する
    - » 再識別アルゴリズムの公開など





(復習) 2015年度の課題

- 匿名加工の方法
  - どの方式がどの程度の安全性を保ったか後から検証が困難
  - 本来は情報保護指針で公開する情報
  - 山岡攻撃一辺倒，多様化
- データやアルゴリズムの共有
  - 疑似ミクロデータは公開不可．参加者のアルゴリズムも提供少ない．
- 現実との乖離
  - 攻撃者の仮定が強すぎる．疑似ミクロの様な静的なデータばかりではない

2016年度の方針(案)

- コンテストの完成度を高める
  - 加工に用いた手法，再識別手法の再利用.
  - 匿名加工データの再利用.
- 匿名加工情報の有用性を高める
  - (A)静的な個人データ，(B)仮名付き履歴データの組合せ.
- 国際化…

マスター系，トランザクション系

マスター系 X

契約ID	氏名	性別	年齢	勤務先	住所	口座番号
A0123	山口太郎	男	41	A社	神奈川県川崎市	89101
A0456	渡辺花子	女	32	B社	東京都中野区	75555
B0789	山岡一郎	男	33	C庁	東京都荒川区	64444

トランザクション系 W

契約ID	利用日	金額	JANコード
A0123	2015年10月10日	12,000	49700001
A0123	2015年11月5日	5,000	49700020
A0456	2015年10月10日	12,000	49700001
A0456	2015年11月5日	250,011	49700030

匿名加工例

マスター系 X

契約ID	氏名	性別	年齢	勤務先	住所	口座番号
A0123	山口太郎	男	41	A社	神奈川県川崎市	89101
key変換	削除		カテゴリー	一般化	コード変換	削除

匿名加工 Y

5133		男	40	会社員	神奈川県川崎市	

トランザクション系 W

契約ID	利用日	金額	JANコード
A0123	2015年10月10日	12,000	49700001
key変換	加工なし		

匿名加工 Z

5133	2015年10月10日	12,000	49700001

トランザクション系Wの例と特徴

- 例
  - HEMSデータ
  - 乗降履歴データ
  - ブラウザ閲覧履歴データ
- 特徴
  - 動的（時刻情報が付随）
  - 個人識別性は低い（脈拍データ）
  - 長期間組み合わせると一意

業者側の主な要請

- 「匿名化されているのだから，履歴Wは無加工でいいでしょう」
- 「個人情報の非可逆なハッシュ化」
- 「履歴は26カ月分欲しい」
- 「（共有の）本人同意を取ってしまえ」

2016年度コンテスト案

- 「トランザクション匿名化（仮）」
- 匿名加工フェーズ
  - 対象データに対して，「情報保護指針」とアルゴリズムを提出
  - kなどの安全性レベルを可変にする
- 再識別フェーズ
  - AUCカーブで判定する
- 有用性フェーズ
  - 有用性関数を提出する.

課題

- 1. 対象データをどこから入手するか
- 2. トランザクション系の匿名加工はどうすればよいか.
- 3. (サンプルの)安全性評価関数定義
- 4. (サンプルの)有用性評価関数定義
- 5. 事前に提出する匿名加工方法の定義
- 6. 勝者を決めるルール

データ案1「人間特性データ」

生成した擬似データ

1000人分の身体計測データを元に，
10万人分のデータを擬似生成

測定地域	年代	性別	身長	体重	握力	・・・
中日本	30	男	175cm	67kg	720N	・・・
東日本	60	女	152cm	42kg	523N	・・・
西日本	20	男	190cm	62kg	230N	・・・

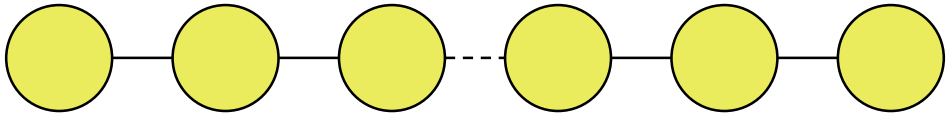
http://musashi.osdn.jp/artificialData/

データ案2「顧客購買履歴の人工データ」

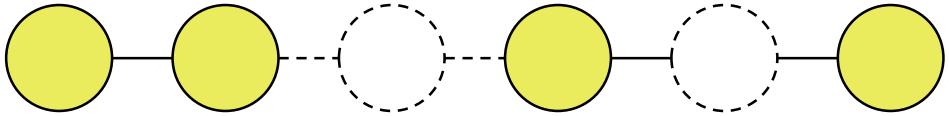
項目	デフォルト
商品数	平均4、標準偏差2の正規分布に基づく乱数
対象とする商品分類	中分類が11,12,13,14 のみ対象とする。
仕入単価	nrand(298円,100円)
店別顧客人数	A～Gの7店舗、合計2350顧客
顧客の来店時間	11時と16時が来店のピークとなるような来店時間
顧客の一来店あたりの購入数量	nrand(nrand(5個,2個),3) 但し、0以下は省く
顧客の来店間隔	nrand(30日,10日)
顧客の来店開始日	2001/07/01＋nrand(0,200日)
顧客の来店終了日	2003/07/01＋nrand(0,200日)
購入商品	nrand(商品番号の中点, 商品数/3)
性別	「男:女＝1:5」となるような一様乱数
顧客の生年月日	"1960年+nrand(0,10)"/01/01 + nrand(0,365日)
販売価格	最初の日付における単価は仕入単価 ×(1+nrand(0.3,0.1))で計算される。

トランザクション系の匿名加工

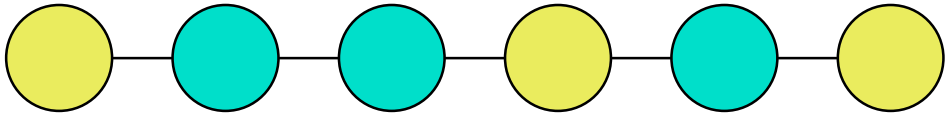
- 切断



- サブサンプリング

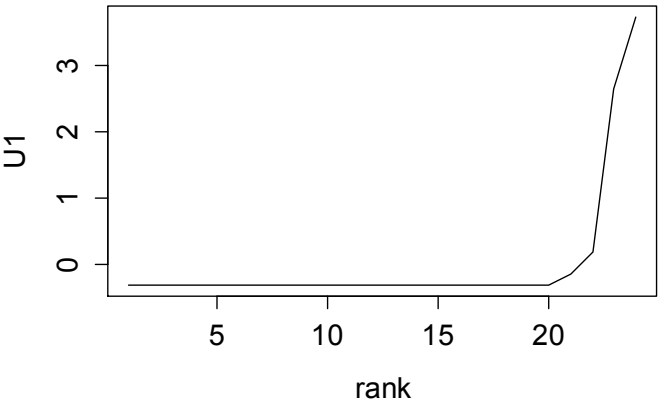


- ミックス

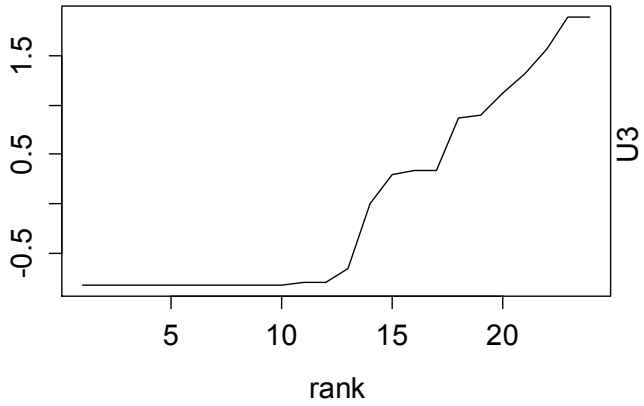


有用性 $U_1,..,U_5$ の分布

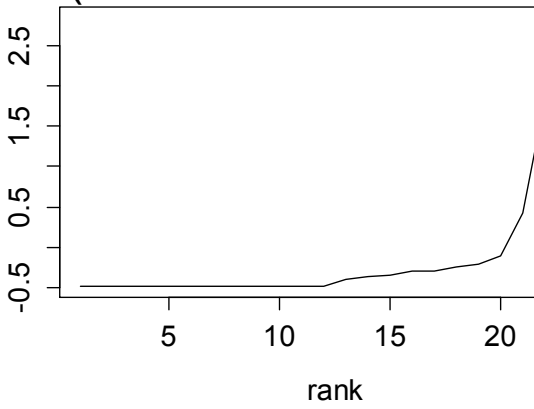
MeanMAE
(SA平均値のMAE)



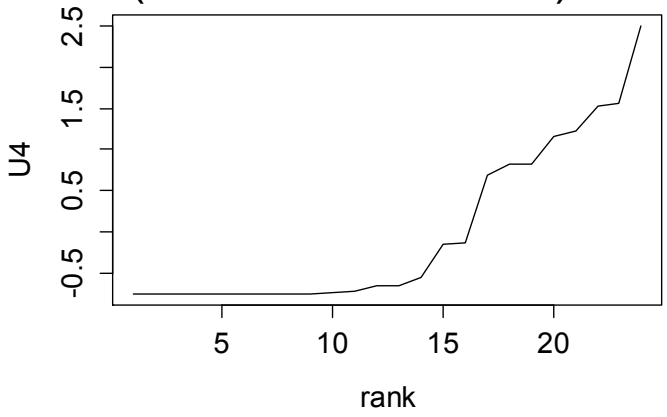
CrossMAE
(QIクロス集計値のMAE)



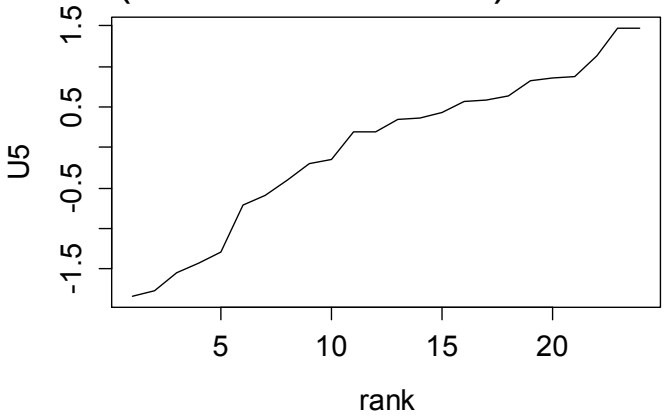
CrossCnt
(QIクロス集計数のMAE



CorMAE
(SA相関係数のMAE)



IL (Information Loss)
(SA加工値のMAE)

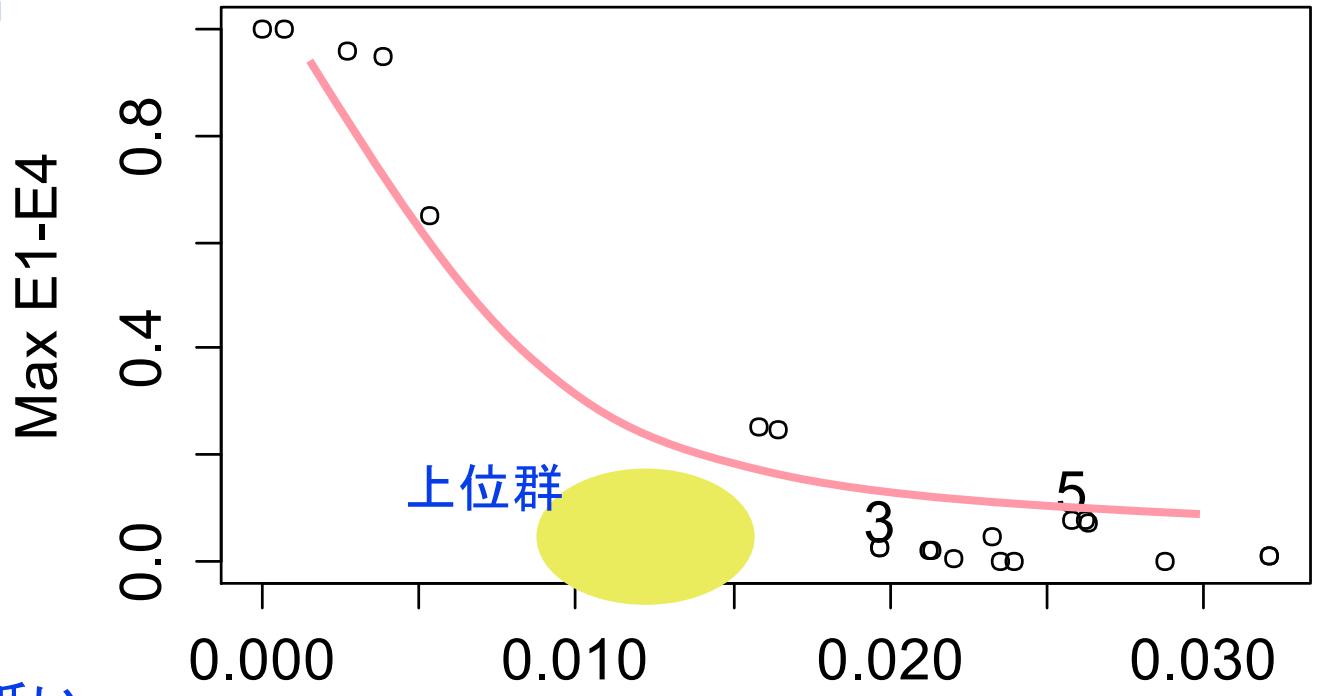


有用性Uと安全性Eのトレードオフ

危険
再識別率高

安全性
（再識別率）

安全
再識別率低い



小さな加工
有用性高い

U5
有用性

大きな加工
有用性低い

期待すること

- 加工方法の公開･提供
- サンプルデータの共有
- 安全性評価手法の普及