Vol.2016-CSEC-73 No.25
Vol.2016-IOT-33 No.25
2016/5/27

ユークリッド距離を用いた再識別手法と PWSCup2015 の匿名加工データを用いた評価

伊藤聡志　菊池浩明
明治大学 総合数理学部 先端メディアサイエンス学科

概要：匿名加工データを再識別する手法として，元データの SA(機密属性)と匿名加工データの SA 間のユークリッド距離を用いる identify.euc を提案し，既存の手法と比較する．また，2015 年 10 月に開催された匿名加工・再識別コンテスト PWSCup の本戦に提出された匿名加工データを用い，この手法を評価する．

キーワード：匿名加工，再識別，PWSCup

Proposal on Re-identification method by using Euclidean distance and evaluation by using Anonymized Data of PWSCup2015

Satoshi Ito　Hiroaki Kikuchi
Department of Frontier Media Science
School of Interdisciplinary Mathematical Sciences, Meiji University

Abstract: We propose a new method to re-identify anonymized data by using Euclidean distance between the original record and the anonymized record and evaluate the accuracy of the proposed method. In order to clarify performance of several anonymization methods used in the competition of the PWSCup_2015, we examine each of single methods and attempt to estimate the accuracy of the combination of some methods.

Keyword: anonymization, re-identification, PWSCup

1　はじめに

企業間での顧客データの売買の際に，企業はデータを匿名加工し，個人を特定されないようにする必要がある．匿名加工データから個人を特定しようとする行為を再識別という．しかしながら，再識別手法は非常に多種多様であり，それらに対して頑強な匿名加工手法を開発するのは困難である．おりしも，2015 年 9 月に改正された個人情報保護法により匿名加工情報が定義され，様々な企業で活用が行われようとしている．そこで，匿名加工技術の開発と安全性の評価手法の確立を目的として，2015 年 10 月に匿名加工・再識別コンテスト(PWSCup)[1]が開催された．

本研究では，新たな再識別手法の提案とそのリスクを評価することを目的とする．既存の再識別手法としては，データの準識別子(Quasi Identifier, 以後 QI)や機密属性(Sensitive Attribute, 以後 SA)を用いるものなど様々な手法がある[2]．QI,SA の厳密な定義は[1]を参考されたい．本研究では PWSCup2015 で安全性の評価に用いられた 4 つの既存手法と新たに提案する手法を比較し，評価を行う．特に，再識別の際に用いる属性の数と，必要な計算時間に注目をする．なぜならば，既存手法の identify.sa は再識別に用いる属性の数が少なく，加工に弱いという弱点があったことに対し，提案手法では用いる属性の数を大幅に増やすことで小さな加工にも強いことが期待できるためである．しかし，そのコストとしてユークリッド距離の計算時間がかかるため，そのオーバヘッドも合わせて測定する．手法の実装・分析は R 言語で行う．

また，本研究では PWSCup2015 の本戦に提出された匿名加工データの解析も行う．コンテストに提出されたデータの統計的な性質や総合的な評価は[4]にて行われているが，これらのデータは複数の匿名加工手法を組み合わせて作成されているものが多いため，どの手法にどれだけの安全性向上の効果があったのか不明確であった．そこで，コンテストでも用いられていた擬似ミクロデータ[3]からサンプリングした実験データに対して，単一の匿名加工手法を適用してその有用性と安全性の評価を行った．本稿では，実験結果を基にして，どの匿名加工データがどの手法を組み合わせて加工され

ているのかを分析，予測する．新たに提案する再識別手法を評価する際にも，これらのデータを用いる．

2 有用性と安全性

2.1 擬似ミクロデータ

PWSCup では加工の対象として，独立行政法人統計センター[3]が作成した擬似ミクロデータが用いられた．このデータは 8333 レコード，25 属性のデータであり，平成 16 年世帯別年間支出額を表している．1～13 列目は世帯の人数，年齢などの離散的なデータを与えており，本研究ではこれを準識別子(QI)に分類する．一方，14～25 列目は食費や医療費などの支出額であり，連続値を持つ．本研究ではこれを属性値(SA)とする．

2.2 有用性と安全性

コンテストでは匿名加工データの有用性と安全性を評価するために，複数の指標[1]を用いていた．表 1 にそれらの内容と，求める際に用いる対象を示す．

表 1．有用性指標と安全性指標[1]

		指標名	指標の内容	対象
有用性	U1	meanMAE	SA 平均絶対誤差	SA
	U2	crossMean	クロス集計値の平均絶対誤差	QI SA
	U3	crossCnt	クロス集計数の平均絶対誤差	QI
	U4	corMAE	SA の相関係数の平均絶対誤差	SA
	U5	IL	匿名加工データの各値の平均絶対誤差	SA
	U6	nrow	匿名加工データのレコード数	行数
安全性	S1	k-anony	k-匿名性指標の最小値	QI
	S2	k-anonyMean	k-匿名性指標の平均値	QI
	E1	identify.rand	QI からランダムな再識別率	QI
	E2	identify.sa	QI から SA15 列目による再識別率	QI SA
	E3	identify.sort	SA の総和でソートによる再識別率	SA
	E4	Identify.sa21	SA21 列について再識別率	SA

2.3 既存再識別手法

本研究では提案手法との比較に，匿名加工・再識別コンテスト PWSCup2015 で匿名加工データの安全性の評価に用いられた以下の 4 つの手法を用いる．ここで，元データを X，X を匿名加工(SA へのノイズ付加)したデータを B とする．説明のために表 2a，2b にこれらの例を示す．3 つの QI 属性，2 つの SA 属性についての 4 つのレコード(行)から成っている．本編では QI 属性の値の組み合わせを，QI のベクトルと呼ぶ(SA も同様)．例えば，X の第 1 レコードの QI のベクトルは(2, 1, 1)であり，SA のベクトルは(100,100)である．

再識別率を再識別手法によって求めた行番号と匿名加工データの行番号との一致率，すなわち，一致したレコード数／元データのレコード数と定義する．

表 2a．サンプルオリジナルデータ X

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	2	300	200
1	1	2	400	500

表 2b．匿名加工データ(ノイズ付加)B

QI1	QI2	QI3	SA1	SA2
2	1	1	110	90
2	1	1	220	390
1	1	2	280	210
1	1	2	390	520

2.3.1 identify.rand(E1)

この手法では，B の再識別したいレコードと同じ QI のベクトルを持つレコードを X から探し，その中からランダムに再識別を行う．例えば，B の第 1 レコードと同じ QI のベクトル(2,1,1)を持つレコードは X の第 1，第 2 レコードであるため，それらの 2 レコードからランダムに 1 つを選ぶ．

2.3.2 identify.sa(E2)

この手法では，B の再識別したいレコードと同じ QI のベクトルを持つレコードを X から探し，その中から特定の SA が最も近いレコードを再識別する．例えば，B の第 1 レコードと同じ QI のベクトル(2,1,1)を持つ 2 レコードの中で SA1 の値が 110 に最も近いのは，100 と 200 の内第 1 レコードである．そこで，X の第 1 レコードを加工したと推定する．

2.3.3 identify.sort(E3)

この手法では，SA の和で X と B のレコードを昇順にソートし，その順位で対応するレコードを再識別する．表 1 の例では，SA の和(SA1+SA2)でソートした．X を昇順でソートすると第 1(200)，第 3(500)，第 2(600)，第 4 レコード(900)の順になり，B を昇順でソートすると第 1(200)，第 3(490)，第 2(610)，第 4 レコード(910)の順になるため，この順で推定レコードとする．

2.3.4 identify.sa21(E4)

この手法ではレコードの QI は考慮せず，特定の SA の値だけで再識別を行う．例えば，B の第 2 レコードの SA1 の値(220)と最も近い値を持つのは X の第 2 レコードの 200 であるため，これを推定レコードとする．

3 提案する再識別手法 identify.euc

3.1 identify.euc

本提案 identify.euc では，B の再識別したいレコードと同じ QI のベクトルを持つレコードを X から探し，それらの SA のユーク

リッド距離 $D(\boldsymbol{a},\boldsymbol{b})=\sqrt{\sum_{i\in s}^{n}(b_i-a_i)^2}$ を用いて再識別を行う．例えば，B の第 1 レコード(SA のベクトルは$\mathbf{b_1}$ =(110, 90))と同じ QI のベクトル(2, 1, 1)を持つ X のレコードは第 1 レコード(SA のベクトルは$\mathbf{a_1}$ =(100, 100))と第 2 レコード(SA のベクトルは$\mathbf{a_2}$ =(200, 400))であり，ユークリッド距離は，

$$D(\mathbf{a_1},\mathbf{b_1})=14.142<322.8=D(\mathbf{a_2},\mathbf{b_1})$$

より，$\mathbf{b_1}$を加工前のレコード$\mathbf{a_1}=\mathbf{b_1}$と推定する．

3.2 QI のベクトルが適合しない場合の動作

表 2b のサンプルデータは SA のみ加工されたデータであるため，元データと匿名加工データの QI 属性はすべて適合する．しかし QI 属性を加工されている場合，QI のベクトルが適合しないことがある．例えば，表 3 の匿名加工サンプルデータ D を考えよ．D は X の 3 つの QI 属性の内，QI3 の値をすべて 1 に統一した匿名加工データである．

表 3. 匿名加工データ(QI 統一)D(E)

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	1	300	200
1	1	1	400	500

D の 3,4 レコードを identify.euc で再識別する際，問題が生じる．(1, 1, 1)という QI のベクトルを持つレコードは X には存在しないためである．このように QI のベクトルが適合しない場合の動作を 2 種類用意し，それらを実装した identify.euc をそれぞれ EUC1, EUC2 とした．動作の詳細は表 4 に示す．EUC1, EUC2 のアルゴリズムを以下に示す．

表 4. EUC1 と EUC2

EUC1	QI のベクトルが適合しない場合，D の探索中のレコード$\mathbf{d_i}$の i を返す
EUC2	QI のベクトルが適合しない場合，そのレコードと，元データの全レコードの SA についてユークリッド距離を求め，最も近いものを探し，その ID を返す．

Algorithm: identify.euc (EUC1)

1. 入力: 元データ X，匿名加工データ B，B の長さ n，再識別に用いる QI のベクトル q，SA のベクトル s
2. key = q の QI，value = 対応するレコードでインデックス F を作成する．
3. B の第 i 番目のレコード$\mathbf{b_i}$と同じ QI を持つ全レコードを X から探し，それらの全てについてレコード間のユークリッド距離$D(\mathbf{a_j},\mathbf{b_i})$を s の SA で求める．最もユークリッド距離の近い j レコードを i の同一レコードと推定する．
4. QI のベクトルが適合しない場合，B の探索中のレコード$\mathbf{b_i}$の i を返す．
5. 3 または 4 を B のすべてのレコード i=1,…,n について行い，推定行番号 ID を返す．

例 1) X，B: 表 2a,2b のサンプルデータ

q = {1, 2, 3}　s = {4, 5}

q = {1, 2, 3}であるため，B の 1~3 列目を用いてインデックス F を作成する．この場合の F を表 5.1 に示す．

表 5.1. 例 1 における F

key	value
(2, 1, 1)	1, 2
(1, 1, 2)	3, 4

B の第 1 レコード$\mathbf{b_1}$を再識別する場合，X で QI のベクトルが$\mathbf{b_1}$と同じ(2, 1, 1)であるのは$\mathbf{a_1}$と$\mathbf{a_2}$である．
そのため，$\mathbf{a_1}$と$\mathbf{b_1}$，$\mathbf{a_2}$と$\mathbf{b_1}$間のユークリッド距離を求め，$\mathbf{b_1}$との距離が最小の X のレコードを$\mathbf{b_1}$の推定レコードとする．この工程を$\mathbf{b_2}$，$\mathbf{b_3}$，$\mathbf{b_4}$についても行い，推定行番号 ID を作成する．

例 2) X，D: 表 2a,4 のサンプルデータ

q = {1, 2, 3}　s = {4, 5}

q = {1, 2, 3}であるため，D の 1~3 列目を用いてインデックス F を作成する．この場合の F を表 5.2 に示す．

表 5.2. 例 2 における F

key	value
(2, 1, 1)	1, 2
(1, 1, 1)	3, 4

D の第 1 レコード$\mathbf{d_1}$を再識別する場合，X で QI のベクトルが$\mathbf{d_1}$と同じ(2, 1, 1)であるのは$\mathbf{a_1}$と$\mathbf{a_2}$である．
そのため，$\mathbf{a_1}$と$\mathbf{d_1}$，$\mathbf{a_2}$と$\mathbf{d_1}$間のユークリッド距離を求め，$\mathbf{d_1}$との距離が最小の X のレコードを$\mathbf{d_1}$の推定レコードとする．この工程を$\mathbf{d_2}$，$\mathbf{d_3}$，$\mathbf{d_4}$についても行い，推定行番号 ID を作成する．しかし，$\mathbf{d_3}$，$\mathbf{d_4}$の QI のベクトルは(1, 1, 1)であるが，X にはこの QI のベクトルを持つレコードは存在しない．
そのため，$\mathbf{d_3}$は 3，$\mathbf{d_4}$は 4 を推定行番号として返す．

Algorithm: identify.euc (EUC2)

1~3. EUC1 と同一である．

4. QI のベクトルが適合しない場合，そのレコードと，元データの全レコードの SA についてユークリッド距離を求め，最も近いものを探し，その ID を返す．
5. 3 または 4 を D のすべてのレコード i=1,…,n について行い，推定行番号 ID を返す．

例 3) X，D: 表 2a,4 のサンプルデータ

q = {1, 2, 3}　s = {4, 5}

q = {1, 2, 3}であるため，D の 1~3 列目を用いてインデックス F を作成する．この場合の F を表 5.3 に示す．

表 5.3. 例 3 における F

key	value
(2, 1, 1)	1, 2
(1, 1, 1)	3, 4

D の第 1 レコード$\mathbf{d_1}$を再識別する場合，X で QI のベクトルが$\mathbf{d_1}$と同じ(2, 1, 1)であるのは$\mathbf{a_1}$と$\mathbf{a_2}$である．

そのため，$\mathbf{a_1}$と$\mathbf{d_1}$，$\mathbf{a_2}$と$\mathbf{d_1}$間のユークリッド距離を求め，$\mathbf{d_1}$との距離が最小のXのレコードを$\mathbf{d_1}$の推定レコードとする．この工程を$\mathbf{d_2}$，$\mathbf{d_3}$，$\mathbf{d_4}$についても行い，推定行番号IDを作成する．しかし，$\mathbf{d_3}$，$\mathbf{d_4}$のQIのベクトルは$(1,1,1)$であるが，Xにはこの QI のベクトルを持つレコードは存在しない．
そのため，$\mathbf{d_3}$とXの全レコード間，$\mathbf{d_4}$とXの全レコード間のユークリッド距離を求め，距離が最も近いレコードの行番号IDを返す．

4 評価

本章の目的は，PWSCup2015 の匿名加工データの解析である．本章では，次の3つについて 評価する．

・単独匿名加工手法の効果(4.1,4.2,4.3 節)
・PWSCup2015 を用いた評価(4.4,4.5 節)
・提案手法(identify.euc)の性能評価(4.6 節)

4.1 PWSCup2015 の匿名加工データ

identify.euc を評価する際に用いるデータを，D_1,…,D_{12}とする．これらのデータは PWSCup2015 の本戦に参加した上位 3 チームを含む 5 チームから提出された，擬似ミクロデータを匿名加工したデータである．表 6 に示す．

表 6. PWSCup2016 のD_1,…,D_{12}の作成チーム

データ名	作成チーム	成績
D_1, D_2	T_A	
D_3, D_4	T_B	2 位
D_5, D_6	T_C	
D_7, D_8, D_9	T_D	1 位
D_{10}, D_{11}, D_{12}	T_E	3 位

4.2 単独手法による匿名加工データ

D_1,…,D_{12}は複数の匿名加工手法を組み合わせて作成されたデータである．そのため，どの加工手法が安全性と有用性にどれだけ効果していたかが不明であった．そこで，表 6 の加工に用いられていた主な手法を，単独に適用したデータを用いて，各々の効果を調査する．

疑似ミクロデータからランダムにサンプリングした小規模データ(100 レコード，25 属性)をもとに，8 つの匿名加工データを単独手法によって作成した．これらのデータをD_A,…,D_Hとし，詳細を表 7 に示す．D_A,…,D_Hの有用性と安全性を調べることにより，D_1,…,D_{12}がどの手法を組み合わせて加工されたデータであるかを予測する．また，D_A,…,D_Hを作成する際に用いた 8 つの匿名加工手法について，表 2a の X を用いて次節から説明する．ただし，k-匿名化については[2]を，山岡匿名化については[1]を参考されたい．

表 7. 単独加工データD_A,…,D_H

データ名	匿名加工手法	加工対象
DA	k-匿名化	QI
DB	SA ノイズ付加	SA
DC	山岡匿名化	ID
DD	QI 統一（対象外）	QI
DE	QI 統一（対象内）	QI
DF	SA 平均化	SA
DG	QI 内スワップ	SA
DH	レコード削除	レコード

4.2.1 SA ノイズ付加

元データの SA にノイズを付加する手法(ランダマイズ，摂動化)である．表 2b の B は X をこの手法で加工した匿名加工データである．この手法で加工を行うと，SA を対象とした有用性指標である U1,U2,U4,U5 が下がり，同様に SA を対象とした安全性指標である E3 と E4 が上がると考えられる．

4.2.2 QI 統一

元データ X の QI のいくつかの属性をある値に統一する手法である．表 4 の D は X の QI の内，QI3 を 1 に統一した匿名加工データである．この手法で加工を行うと，有用性を下げずに安全性を上げることが可能である．しかし QI の内，クロス集計の評価値 U2,U3 の対象である属性を統一してしまうと有用性が下がってしまう．例えば PWSCup2015 本戦の場合，U2,U3 で用いる QI の属性は 1 列目～6 列目であったため，1 列目～6 列目を統一してしまうと有用性が下がった．この手法の対象となる属性を加工に含めるか否かで，D と E の 2 種類を区別する．

4.2.3 SA 平均化(ミクロアグリゲーション)

元データのレコードの内，QI のベクトルが同じ値のレコードの SA を属性ごとに平均値で置き換える手法である．X を SA 平均化によって匿名加工した結果 F を表 8 に示す．この手法で加工を行うと，U4,U5 が下がり，E3,E4 が上がる．表 8 では，QI のベクトルによって，4 つのレコードが$\{1,2\},\{3,4\}$の 2 グループに分類されている．各グループの平均値を用いるため，全体の平均値 U1 には影響を与えない．

表 8. 匿名加工データ(SA 平均化)F

グループ	QI1	QI2	QI3	SA1	SA2
(1)	2	1	1	150	250
	2	1	1	150	250
(2)	1	1	2	350	350
	1	1	2	350	350

4.2.4 QI 内スワップ

元データのレコードの内，QI のベクトルが同じ値のレコードの SA を属性ごとにランダムにスワップする手法である．A を QI 内スワップによって匿名加工した結果 G を表 9 に示す．グループ(1)では SA1 を，(2)では SA2 を入れ替えている．スワップなので平均値は変わらず，QI 内なのでクロス集計値 U2,U3 も変化しない．この手法で加工を行うと，相関係数等の U4,U5 が下がり，安全性

E2,E3,E4 が上がる.

表 9. 匿名加工データ(QI 内スワップ)G

グループ	QI1	QI2	QI3	SA1	SA2
(1)	2	1	1	200	100
	2	1	1	100	400
(2)	1	1	2	300	500
	1	1	2	400	200

4.2.5 レコード削除

元データのレコードを削除する手法である. この手法で加工を行うと U1,U2,U3,U5,U6 が下がり, E3,E4 が上がる. (PWSCup2015 では該当データは提出されなかった)

4.2.6 k-匿名化, 山岡匿名化

これらの手法については参考文献[2],[1]を参考されたい. k-匿名化によって加工を行うと U2,U3 が下がり, S1,S2,E1,E2 が上がり, 山岡匿名化で加工を行うと U5 が下がり, E1~E4,が上がる.

4.3 期待される効果

データを匿名加工すると, 一般的に有用性が低くなり, 安全性が高くなる. 前節をもとに, D_A,…,D_H の作成に用いた 8 つの匿名加工手法が U1～U6,S1,S2,E1~E4,EUC1 の値をどのように変化させるのか定性的な効果の予測を表 10 に示す. 有用性 U1～U6 の欄における「×」は「大きく損なう」,「△」は「少し損なう」,「−」は「変化しない」を意味し, 安全性 S1,S2 における「○」は「高まる」,「×」は「変化しない」を意味し, E1~E4,EUC における「○」は「この手法には強い」,「△」は「この手法には少し強い」,「×」は「この手法には弱い」を意味する.

表 10. 期待される効果

		匿名加工手法							
		k-匿名化	ノイズ付加	YA	QI統一(対象外)	QI統一(対象内)	ミクロ	QI内スワップ	レコード削除
有用性	U1	−	△	−	−	−	−	−	×
	U2	×	△	−	−	×	−	−	×
	U3	×	△	−	−	×	−	−	×
	U4	−	△	−	−	−	×	×	×
	U5	−	△	×	−	−	×	×	×
	U6	−	−	−	−	−	−	−	×
安全性	S1	○	×	×	×	×	×	×	×
	S2	○	×	×	○	○	×	×	×
	E1	△	×	○	△	△	×	×	×
	E2	△	×	○	△	△	×	△	×
	E3	×	△	○	×	×	○	○	×
	E4	×	△	○	×	×	○	△	×
	EUC	△	×	○	△	△	×	○	×

4.4 評価結果

4.4.1 D_A,…,D_H の評価とD_1,…,D_{12}の手法予測

D_A,…,D_H の有用性・安全性を表 11 に示す. 表 11 の Original の列には D_A ,…, D_H の元データの指標値を示す. 表 11 の E4(identify.sa21)の値が全体的に低いように思えるが, これは擬似ミクロデータの 21 列目に 0 が多く, それらのレコードは正しく再識別することができないためである. このデータは 100 レコード中 76 レコードの 21 列目が 0 であるため, E4 の最大値は 0.24 となっている. 図 1 には, U4 と E3 についての散布図を示す.

また, 表 12 にはD_1,…,D_{12}の有用性と安全性を示し, 表 13 には D_1,…,D_{12}の評価結果と, それによる加工手法の予測を示す. 匿名加工手法は組み合わせると, それらの特徴を併せ持った加工になる. 例えば, D_{10}は k-匿名化と SA 平均化を組み合わせて匿名加工されたデータであるため, D_AとD_Fの特徴を併せ持っている(表 14 に示す). 表 2a の X 等のサンプルデータ X, B, C, D, F, G とD_A,…,D_Hは異なる. しかし, 記号が同じデータは同じ匿名加工手法で加工されている. 例えば, 表 2b の B とD_Bは両方 SA ノイズ付加で加工されたデータである.

表 11. 元データとD_A,…,D_Hの有用性と安全性

	Original	D_A	D_B	D_C	D_D	D_E	D_F	D_G	D_H
	元データ	k-匿名化	ノイズ付加	YA	QI統一(対象外)	QI統一(対象内)	ミクロ	QI内スワップ	レコード削除
U1	0	0	46.225	0	0	0	0	0	295.731
U2	0	38837.9	7808.7	0	0	15135.7	104.9	209.7	1094.4
U3	0	5.833	0	0	0	2	0	0	0.097
U4	0	0	0.020	0	0	0	0.000	0.000	0.049
U5	0	0	0.016	0.120	0	0	0.000	0.000	0
U6	0	0	0	0	0	0	0	0	10
S1	1	3	1	1	1	1	1	1	1
S2	1.031	7.692	1.031	1.031	1.053	1.053	1.031	1.031	1.034
E1	1	0.13	0.99	0	0.07	0.11	0.94	1	1
E2	1	0.17	1	0	1	1	1	1	1
E3	1	1	0.54	0	1	1	1	0.91	0.067
E4	0.24	0.24	0.22	0	0.24	0.24	0.24	0.24	0.089
EUC1	1	0.13	1	0	0.07	0.11	1	1	1
EUC2	1	0.17	1	0	1	1	1	1	1

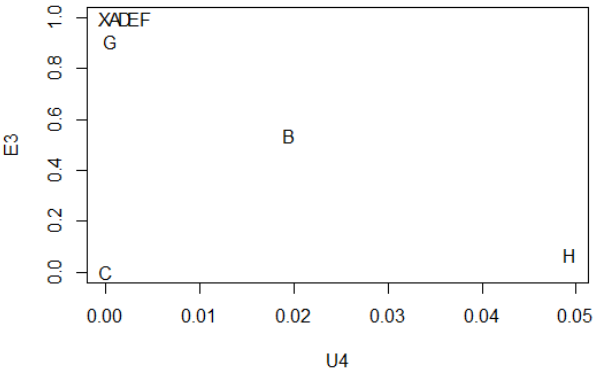

図 1. U4 と E3 についての散布図

表 14. D_A, D_FとD_{10}の特徴

	D_A	D_F	D_{10}
U1	−	−	−
U2	×	−	×
U3	×	−	×
U4	−	×	×
U5	−	×	△
U6	−	−	−
S1	○	×	○
S2	○	×	○
E1	△	×	△
E2	△	×	△
E3	×	○	○
E4	×	○	○
EUC1	△	×	○

Vol.2016-CSEC-73 No.25
Vol.2016-IOT-33 No.25
2016/5/27

4.4.2 EUC1 と既存手法との比較

EUC1 と既存の 4 つの手法との比較を行う．比較に用いるデータは，元データに擬似ミクロデータ(8333 レコード，25 属性)，匿名加工データにD_1,...,D_{12}を用いる．

既存手法の再識別成功率を表 15 に示す．赤い数値(*が付いている数値)は，その匿名加工データに対して最も再識別成功率が高かった再識別手法を示している．EUC1 は 12 個中 5 個が最高値であり，既存手法よりも多い．

表 15. 既存手法と EUC1 の再識別成功率

	既存方式				提案方式
匿名加工データ	Id-rand	Id-sa	Id-sort	Id-sa21	EUC1
D_1	0.0326	0.8238	*1.0000	0.1858	0.3010
D_2	0.6485	*0.6507	0.0012	0.0022	0.4780
D_3	0.1990	0.2412	*0.2482	0.0511	0.2070
D_4	0.1894	0.2401	*0.2526	0.0455	0.2110
D_5	0.0000	0.0223	0.0004	0.0002	*0.0743
D_6	0.0000	0.0223	0.0004	0.0002	*0.0743
D_7	0.0023	0.0223	0.0091	0.0014	*0.8762
D_8	0.0000	0.0000	0.0004	0.0002	*0.0011
D_9	0.0001	0.0002	0.0004	0.0000	*0.0024
D_{10}	0.0060	*0.0066	0.0001	0.0005	0.0043
D_{11}	*0.0180	0.0164	0.0001	0.0001	0.0080
D_{12}	*0.0214	*0.0214	0.0004	0.0001	0.0080
平均	0.0931	0.1723	0.1261	0.0240	*0.1871
標準偏差	0.1741	0.2578	0.2681	0.0499	0.2426
最適数	2	3	3	0	5

4.5 考察

EUC1 の最高値が従来手法より多かった理由は，提案手法で再識別に用いる属性の数が既存手法より多いためである． 例えば，既存手法の identify.sa は特定の SA からレコードを再識別する手法であるが，その特定の SA に大きいノイズが加えられると，正しく再識別することができなくなる．対して EUC1 は再識別の際に全ての SA を用いるため，それらの内 1 つが大きく加工されても，他の SA から再識別をすることができる．

しかし，D_1,...,D_4，D_{10},...,D_{12}においては identify.sa に再識別率で劣っている．これは identify.sa が EUC2 と同様に，「QI が適合しない場合，元データの全レコードの SA と計算を行い，ID を返す」という仕組みが実装されているためと考えられる．本来ならば EUC2 と identify.sa を比較するべきであるが，EUC2 は計算時間が identify.sa と比べてはるかに多く，再識別率を出すのが困難であるため断念した．また表 11 より，k-匿名化をされているデータでは EUC 1 よりも identify.sort の方が再識別率が高い．

また，もう一つの理由として，比較に用いた匿名加工データはコンテストに提出されたものであるため，安全性指標である 4 つの既存手法に対抗できるように作られたものが多い．そのため，提案手法が有利になった可能性も考えられる．

4.6 提案手法 identify.euc の処理性能評価

擬似ミクロデータには QI 属性が 13 あり，そのうちどれを identify.euc による再識別に用いるかによって計算時間と再識別率が変化する．

用いる QI 属性の数を|q|，SA 属性の数を|s|とおくと，|q|を増やせば計算量は少なくなるが，それに応じて QI の加工に弱くなり，再識別率が下がりやすい．図 2,...,図 5 に，100 レコード，25 属性のデータを用いた時の，|q|と|s|の変化に伴う計算時間と再識別成功率の変化を示す．図 2 より，計算時間は|q|について単調に減少しており，図 3 より，再識別率は|q|について単調に増加している(ただし，|q|=5 で飽和している)．図 4 より，計算時間は|s|について単調に増加しているが，小規模データでテストしているため誤差が大きい．また図 5 より，再識別率は|s|に依存しなかった．図 6 に|q|=1 のときのレコード数の増加に伴う計算時間の変化を示す．

図 6 より，計算時間はレコード数に対して増加している．なお，擬似ミクロデータの SA 属性にノイズを付加した匿名加工データでテストした結果，|q|=13 のとき計算時間は約 1 分，再識別率は約 17%であり，|q|=6 のとき計算時間は約 31 分，再識別率は約 20%であった．ただし，図 2,...,図 6 は EUC1 についてのものである．EUC2 は EUC1 に比べて計算量がはるかに多いため，EUC2 では計算時間と再識別成功率の両方が EUC1 より増加すると考えられる．

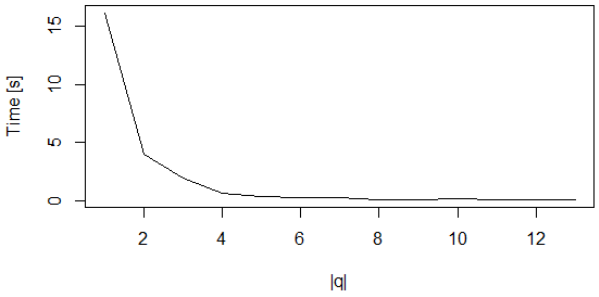

図 2 ． |q|についての計算時間

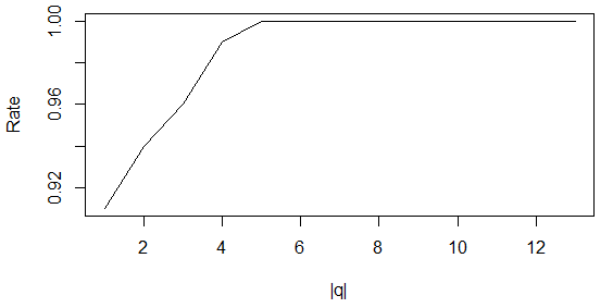

図 3. |q|についての再識別成功率

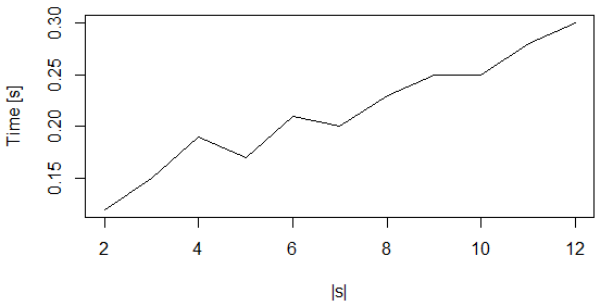

図 4. |s|についての計算時間

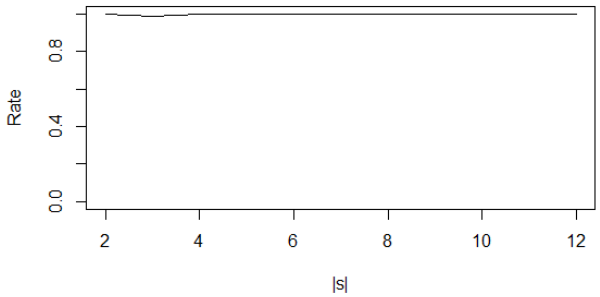

図 5. |s|についての再識別成功率

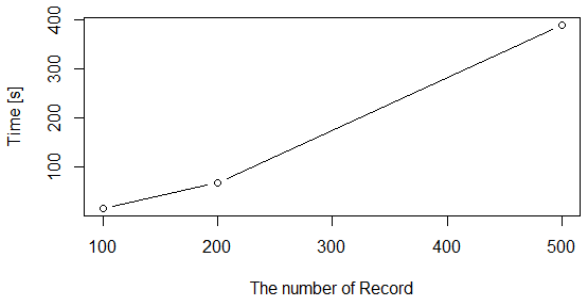

図 6. レコード数についての計算時間

5 おわりに

PWSCup2015 の匿名加工データを用いて提案再識別手法 identify.euc と既存手法の比較を行った．また，単独匿名加工手法で加工した小規模データを用いて PWSCup2015 の匿名加工データの解析を行った．

その結果，匿名加工手法を組み合わせてデータを加工すると，用いた複数の手法の有用性・安全性指標への影響を併せ持つ匿名加工データとなり，PWSCup2015 上位チームの匿名加工データはそれらをうまく組み合わせて作成されていること，また，identify.euc (EUC1)は既存手法(特に identify.sa)と比べて計算時間が大幅に多い割には再識別率にはあまり差はなく，差をつけるためには更に計算時間を増やす必要があることが判明した．しかし，EUC2 のように再識別にかなりの時間を必要とする手法のパフォーマンス性能は悪い．

identify.euc のさらなる改善，新たな再識別手法の開発，それらを考慮した上での新たな匿名加工手法の開発を今後の課題とする．

謝辞

identify.euc の評価を行うにあたり，匿名加工データとその行番号データを提供していただいた「匿名加工・再識別コンテスト PWSCup2015」の参加チームの山口高康氏，長谷川聡氏，濱田浩気氏，正木彰伍氏，田中哲士氏，藤田真浩氏に感謝いたします．

参考文献

[1] 菊池浩明，山口高康，濱田浩気，山岡裕司，小栗秀暢，佐久間淳，“匿名加工・再識別コンテスト Ice & Fire の設計”，コンピュータセキュリティシンポジウム 2015, pp. 363-370, 2015.

[2] 南和宏，“プライバシー保護データパブリッシング”，情報処理 Vol. 54，No. 9，pp. 938-946，2013.

[3] 秋山他，“教育用擬似ミクロデータの開発とその利用～平成 16 年全国消費実態調査を例として～”，統計センター製表技術参考資料, 16, pp.1-43, 2012.

[4] 菊池浩明，山口高康，濱田浩気，山岡裕司，小栗秀暢，佐久間淳，“匿名加工コンテスト PWSCUP2015 の報告と匿名加工方法の評価”，SCIS，2016.

表 12. $D_1,\ldots,D_{12}$の有用性と安全性

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
U1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
U2	58340.87	0.00	31572.91	31400.95	0.00	0.00	4321.75	0.00	0.00	65093.42	52975.02	46100.64
U3	18.60	0.00	1.01	0.99	0.00	0.00	1.54	0.00	0.00	7.28	2.97	1.85
U4	0.00	0.01	0.00	0.00	0.07	0.07	0.03	0.09	0.09	0.15	0.11	0.11
U5	0.00	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02
U6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S1	1.00	1.00	3.00	3.00	1.00	1.00	3.00	1.00	1.00	41.00	8.00	4.00
S2	2.66	1.88	4.91	4.86	36.07	36.07	36.07	13.71	13.68	106.83	42.30	31.09
E1	0.03	0.65	0.20	0.19	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02
E2	0.82	0.65	0.24	0.24	0.02	0.02	0.02	0.00	0.00	0.01	0.02	0.02
E3	1.00	0.00	0.25	0.25	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
E4	0.19	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EUC1	0.30	0.48	0.21	0.21	0.07	0.07	0.88	0.00	0.00	0.00	0.01	0.01

表 13. $D_1,\ldots,D_{12}$の評価結果とそれによる加工手法の予測

		匿名加工データ											
		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
有用性	U1	−	−	−	−	−	−	−	−	−	−	−	−
	U2	×	−	×	×	−	−	×	−	−	×	×	×
	U3	×	−	△	△	−	−	△	−	−	×	×	△
	U4	−	△	−	△	△	△	△	△	△	×	×	×
	U5	−	△	△	△	△	△	△	△	△	△	△	△
	U6	−	−	−	−	−	−	−	−	−	−	−	−
安全性	S1	×	×	△	△	×	×	△	×	×	○	○	△
	S2	×	×	△	△	△	○	○	○	○	○	○	○
	E1	△	×	×	×	◎	◎	○	◎	○	△	△	△
	E2	×	×	×	×	△	△	△	◎	○	△	△	△
	E3	×	○	×	×	○	○	○	○	○	○	○	○
	E4	×	○	△	△	○	○	○	○	◎	○	○	○
	EUC1	×	×	×	×	△	△	×	○	○	○	○	○
匿名加工手法	DA	−	−	○	○	−	−	○	−	−	○	○	○
	DB	−	−	−	−	−	−	−	−	−	−	−	−
	DC	−	−	−	−	○	○	−	○	○	−	−	−
	DD	−	−	−	−	○	○	○	−	−	−	−	−
	DE	○	−	−	−	−	−	−	○	○	−	−	−
	DF	−	○	−	−	−	−	−	−	−	○	○	○
	DG	−	−	○	○	−	−	○	○	○	−	−	−
	DH	−	−	−	−	−	−	−	−	−	−	−	−