ユークリッド距離を用いた再識別手法とその評価

伊藤聡志
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．

キーワード：匿名加工，再識別

Proposal on Re-identification method By using Euclidean distance

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

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 proposed method.

Keyword: anonymization, re-identification

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，それらに対抗する匿名加工手法を開発するのは困難である．そこで，本研究では，再識別手法を評価することにより，更に優れた匿名加工手法を提案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(QI)や機密属性(SA)を用いるものなど様々な手法があるが[1]，本研究では 4 つの既存手法と新たに提案する手法を比較，評価を行う．手法の実装・分析は R 言語で行う．

2 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[2]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 A，A を匿名加工(SA へのノイズ付加)したデータを B とする．表 1 にこれらの例を示す．

QI1	QI2	QI3	SA1	SA2	QI1	QI2	QI3	SA1	SA2
2	1	1	336765	67413.87	2	1	1	340500.1	66196.79
2	1	1	151936.9	33783.49	2	1	1	155083.1	33942.66
1	1	1	217858	55406.89	1	1	1	214283.7	53411.23
1	1	1	188105.8	66204.9	1	1	1	185829.1	78337.58

A：元データ　　B：匿名加工データ

表 1．サンプルデータ

2.1 identify.rand

この手法では，B の再識別したいレコードと同じ QI を持つレコードを A 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(2,1,1)を持つレコードは A 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．

2.2 identify.sa

この手法では，B の再識別したいレコードと同じ QI を持つレコードを A 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(2,1,1)を持つ 2 レコードの中で SA1 の値が最も近い第 1 レコードを加工前のレコードと推定する．

2.3　identify.sort

この手法では，SA の和で A 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)でソートする．

2.4　identify.sa21

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 の値と最も近い値を持つのは A の第 2 レコードであり，これを第 2 レコードの推定行番号とする．

3　提案する再識別手法 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 を持つレコードを A 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(340500.1, 66196.79)と同じ QI(2, 1, 1)を持つ A のレコードは第 1 レコード(336765, 67413.87)と第 2 レコード(151936.9, 33783.49)であり，ユークリッド距離は，

$$D(\mathbf{a_1},\mathbf{b_1}) = 3928.39 < 191328.78 = D(\mathbf{a_2},\mathbf{b_1})$$

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．本アルゴリズムを以下に示す．

Algorithm: identify.euc

1. 入力: 元データ A，匿名加工データ B，
   再識別に用いる QI の列の部分集合 q，SA の列の部分集合 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. F をもとに，B の再識別したいレコードと同じ QI を持つレコードを A から探し，それらのレコード間のユークリッド距離を s の SA で求める．最もユークリッド距離の近い 2 レコードを同一レコードと推定する．
4. 2 を B のすべてのレコードについて行い，推定行番号を作成する．

例) A，B: 表 1 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 2 に示す．

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

表 2．例における F

B の第 1 レコード$\mathbf{b_1}$を再識別する場合，A で QI 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 A 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，推定行番号を作成する．

4　評価

4.1　既存手法との比較

Identify.euc 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，A に擬似ミクロデータ(8333 レコード，25 属性)，B に匿名加工・再識別コンテスト PWSCup2015 に参加した数チームが提出した 12 の匿名加工データを用いる．

再識別成功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．既存手法の再識別成功率を表 3 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．identify.euc は 12 個中 5 個が最高値である．

表 3．既存手法の再識別成功率

	既存方式				提案方式
匿名加工データ	id.rand	id.sa	id.sort	id.sa21	id.euc
1	0.0326	0.8238	*1.0000	0.1858	0.3010
2	0.6485	*0.6507	0.0012	0.0022	0.4780
3	0.1990	0.2412	*0.2482	0.0511	0.2070
4	0.1894	0.2401	*0.2526	0.0455	0.2110
5	0.0000	0.0223	0.0004	0.0002	*0.0743
6	0.0000	0.0223	0.0004	0.0002	*0.0743
7	0.0023	0.0223	0.0091	0.0014	*0.8762
8	0.0000	0.0000	0.0004	0.0002	*0.0011
9	0.0001	0.0002	0.0004	0.0000	*0.0024
10	0.0060	*0.0066	0.0001	0.0005	0.0043
11	*0.0180	0.0164	0.0001	0.0001	0.0080
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

4.2　提案手法 identify.euc の評価

擬似ミクロデータには QI にあたる属性が 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成功率が変化する．

用いる QI の数を|q|，SA 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，再識別成功率が下がりやすい．図 1~図 4 に，100 レコード，25 属性のデータを用いて，|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 5 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．

計算時間はレコード数に対して増加している．なお，レコード数が 8333 の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，再識別成功率は約 17%であり，|q|=6 のとき計算時間は約 31 分，再識別成功率は約 20%であった．

4.3　考察

identify.euc の最適数が既存手法より多かった理由は，提案手法で再識別に用いるデータや属性の数が既存手法より多いためと考えられる．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 identify.euc は再識別の際に複数の SA を用いるため，それらのうち 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．

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，4 つの既存手法に対抗できるように作られたものが多く，その分提案手法が有利になった可能性も考えられる．

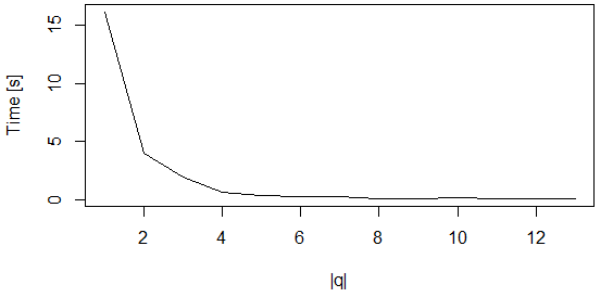

図 1：|q|についての計算時間

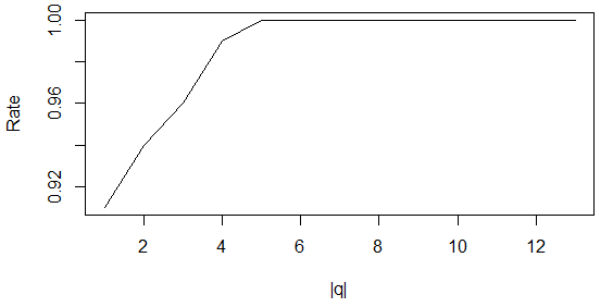

図 2：|q|についての再識別成功率

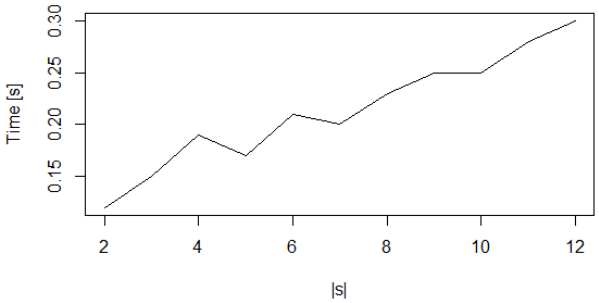

図 3：|s|についての計算時間

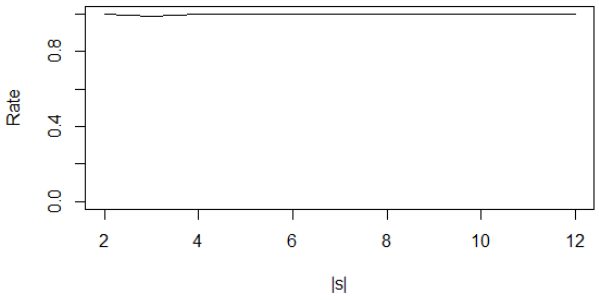

図 4：|s|についての再識別成功率

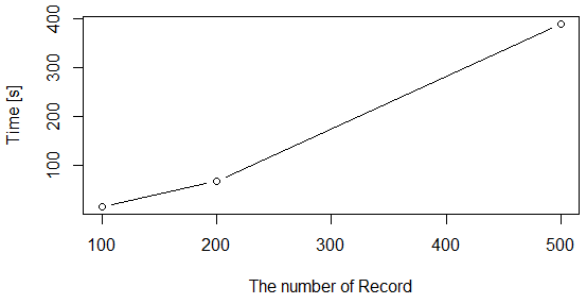

図 5：レコード数についての計算時間

4.4 改善すべき点

Identify.euc の欠点は QI への加工に弱い点である．例えば，再識別に用いる QI を 1~6 列目にしていた場合，1~6 列目のうちいずれかを加工されたとき正しく再識別できない．よって，加工に用いられていない QI を選ぶ仕組みを考える必要がある．

5 おわりに

identify.euc のさらなる改善，新たな手法の開発，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氏，長谷川氏，濱田氏，正木氏，田中氏，藤田氏に感謝いたします．

参考文献

[1] 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.

[2] 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.