センシティブ属性値の距離を考慮したダミー追加による l-多様性アルゴリズムの提案

大石 慶一朗　清 雄一　田原 康之　大須賀 昭彦

個人に関する情報を含んだデータベースを他事業者と共有する場合，プライバシへの配慮が必要不可欠であり，プライバシ指標である k-匿名性や l-多様性に従ってデータベースに加工を行うのが基本となっている．データベースが既存のプライバシ指標の一つである l-多様性を満たしていた場合，個人のセンシティブ属性値として l 種類以上の候補が存在することが保証される．しかしながら，例えばある人物のセンシティブ属性値として HIV-1(M)，HIV-1(N) のように l 種類以上の値の候補があったとしても，どのような型かわからないが，HIV であるということは特定されてしまう問題が存在する．このように l-多様性には事実上の多様性が考慮できていると言えない可能性が存在する．本論文では，既存の指標では事実上の多様性を満たすことができない可能性があるという問題を解決するために，より詳細な多様性を考慮できる (l, d)-意味的多様性を提案する．また，提案指標に適した匿名化アルゴリズム及び分析アルゴリズムも同様に提案し，検証実験を行う．

Consideration of privacy is crucial when sharing a database that contains personal information with other organizations. Many organizations have utilized personal information while realizing the importance of personal privacy protection by anonymizing personal information according to existing indicators, such as l-diversity. However, if a database containing certain personal information holds similar sensitive attribute values, there is a possibility that de facto diversity is not satisfied, even if anonymization is performed to satisfy l-diversity. In this research, we propose (l, d)-semantic diversity that is able to consider more actual diversity to solve the problem of not being able to satisfy de facto diversity with the existing indicator. We also propose an anonymization algorithm and analysis algorithm suitable for the proposal indicator, and we conduct evaluation experiments.

1　はじめに

近年，個人に関する情報を含んだデータベースは様々なことに活用されている．2015 年に個人情報の保護に関する法律の改正[17] により，個人に関する情報を含んだデータベースに匿名化を行い匿名加工情報にすることで円滑に活用，共有することが可能となった．そこで個人に関する情報を含んだデータベースに対して，k-匿名性[2] などのプライバシ指標に従って加工を行うことによって個人のプライバシ保護を実現しつつ，データベースを活用することが基本である．k-匿名性の拡張である l-多様性[3] はより良いプライバシ保護を実現するための指標である．l-多様性とは，ある攻撃者が，匿名加工後のデータベースを参照したとしても，ある人物の知られたくない情報であるセンシティブ属性値について，l 種類以上の候補が存在することを保証する指標である．例えば 3-多様性を満たしているデータベースの場合，攻撃者はある人物のセンシティブ属性値として，HIV-2, A 型インフル, 盲腸のどれであるか識別できないという状況となる．しかしながら，データベース内に類似したセンシティブ属性値が複数存在する場合，データベースが l-多様性を満たしていたとしても，事実上の多様性を満たしているとはいえない場合が存在する．ある攻撃者が 3-多様性を満たしている匿名化後のデータベースを見ることで，ある人物のセンシティブ属性値として，HIV-2, HIV-1(M), HIV-1(N) という候補を得たという状況を想定する．これは 3-多様性を確かに満たしているが，センシティブ属性値間の類似性を考慮しておらず,（どの種類の HIV かは不明だが）ある人物

Proposal of l-Diversity Algorithm Considering Distance between Sensitive Attribute Values

Keiichiro Oishi, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga, 電気通信大学情報理工学研究科, The University of Electro-Communications.

は HIV であるということが識別できてしまう．

そこで本論文では，事実上の多様性を考慮することができるような指標である (l,d)-意味的多様性及び (l,d)-意味的多様性に従った匿名化アルゴリズムを提案する．また，本論文では多くの既存研究[6]，[11] と同様に，データベース利用者がデータベースの統計的な情報を欲している状況を想定するので，匿名化アルゴリズムと共に統計値を得るための分析アルゴリズムも同様に提案する．

本論文の構成を示す．2 章で想定環境を記述する．3 章で既存指標の提示及び課題設定を示す．4 章で課題を解決するための新たな指標の提案を行う．5 章で新たな指標に従った匿名化アルゴリズム及び分析アルゴリズムの提案を行う．6 章で分析アルゴリズムの検証実験，7 章で実験結果の提示及び考察を示す．最後に 8 章で本論文をまとめる．

2　想定環境

個人に関する情報を含むデータベースを保持している人物をデータ保有者とする．個人に関する情報を含むデータベースとは識別子，準識別子 (QID)，センシティブ属性，その他の属性を含んでいるデータベースのことである．以下，本論文で述べるデータベースは全て個人に関する情報を含んだデータベースである．識別子は名前や電話番号など個人を一意に識別できる属性，QID は性別，年齢，郵便番号などそれ一つだけでは個人を識別できないが複数組み合わせることで個人を識別できる属性である．センシティブ属性は年収や病名など個人の知られたくない情報であり，匿名加工はセンシティブ属性を保護するために行われるものである．

データ保有者はデータベースをデータ利用者と共有したいと考えているが，データ利用者は必ずしも信頼できないものとする．データ利用者はデータベースを受け取った後，ある個人のセンシティブ属性値を推測しようとする可能性があるので，共有の際には個人のセンシティブ属性値を識別されないように匿名加工を行わなくてはならない．匿名加工は既存指標の k-匿名性や l-多様性などに従って行われ，特定のデータベースに対して一度のみ匿名加工がなされる．データ利用者は共有して得た匿名加工済みデータベースから統計値を利用するものとし，どのような属性を持つ人がどのようなセンシティブ属性値を保持しているのかを分析したいものとする．例えば，QID として性別，年齢を保持し，センシティブ属性として病状を保持しているデータベースの場合，ある病状がどの年代に多いのか，男性に多いのか，女性に多いのかなどを知りたいものとする．つまりデータ利用者はデータベースを得た際に，クロス集計表を作成し，ヒストグラムを用いた分析を行うものと想定する．

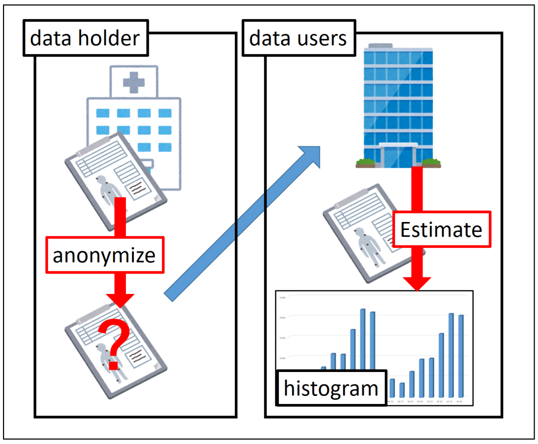

図 1　想定環境

想定される状況は図 1 のようになる．図 1 はデータ保有者がデータベースを匿名加工し，匿名加工後のデータベースをデータ利用者と共有，データ利用者はヒストグラムを作成するという一連の流れを表している．

3　既存指標

3.1　k-匿名性

k-匿名性[2] とは，データベースを匿名化する際に広く利用されているプライバシ指標の一つである．あるデータベースにおいて，全ての QID 値の組み合わせが同一であるレコードが少なくとも k 個存在する場合，そのデータベースは k-匿名性を満たしていると定義される．これにより，攻撃者がある人物の保持する QID 値を全て知っていたとしても同一の QID 値

表 1　病状データベース

名前	性別	年齢	郵便番号	病状
桃子	女性	24	999-4565	声帯ポリープ
夏実	女性	16	999-5636	マイコプラズマ肺炎
花音	女性	21	999-4557	インフルエンザ肺炎
愛理	女性	22	999-4531	インフルエンザ肺炎
五郎	男性	34	999-1332	虫歯
啓太	男性	34	999-1335	盲腸
光男	男性	48	999-1337	胃潰瘍

表 2　2-匿名性を満たす病状データベース

仮名	性別	年齢	郵便番号	病状
A	女性	10〜29	999-****	声帯ポリープ
B	女性	10〜29	999-****	マイコプラズマ肺炎
C	女性	2*	999-45**	インフルエンザ肺炎
D	女性	2*	999-45**	インフルエンザ肺炎
E	男性	30〜49	999-133*	虫歯
F	男性	30〜49	999-133*	盲腸
G	男性	30〜49	999-133*	胃潰瘍

の組み合わせとなっているレコードが少なくとも k 個存在するため，識別することはできない．

k-匿名性を実現するための匿名化手法として一般化がよくあげられる．一般化とは，QID 値の正確性を下げ，抽象化する手法であり，この一般化により QID 値の組み合わせが同一となるようにレコードをグループ分けすることができる．一般化は Mondrian [2] や Incognito [8] など様々な手法が存在しており，最も基本的な手法といえる．表 2 のデータベースは表 1 の病状データベースを一般化により匿名加工したデータベースである． 表 2 は仮名 A と B，仮名 C と D，仮名 E と F と G がそれぞれ同じ QID 値の組み合わせになっていることから，このデータベースは 2-匿名性を満たしているデータベースといえる．例えば桃子さんの QID 値を全て知っていたとしても匿名化データベースを見た際に A の人物なのか，B の人物なのか特定することができず，センシティブ属性値を特定できない．

しかしながら，k-匿名性にはある問題が存在する．表 2 の仮名 C，D のグループに注目すると，センシティブ属性値として “インフルエンザ肺炎” のみを保持していることになる．確かに 2-匿名性は満たしているが仮名 C と D がセンシティブ属性値として “インフルエンザ肺炎”を保持していることが識別できてしまう．k-匿名性では QID にしか着目しておらず，センシティブ属性は考慮していないので，QID 値の組み合わせとセンシティブ属性に相関がある場合，この問題は特に深刻となる場合がある．

3.2　l-多様性

l-多様性 [3] とは，k-匿名性を拡張した指標の一つである．あるデータベースにおいて，ある人物のセンシティブ属性値について，l 種類以上の候補が存在する場合，そのデータベースは l-多様性を満たしていると定義される．

表 3 は表 1 のデータベースを 2-多様性を満たすように匿名加工したデータベースである．仮名 A と D，仮名 B と C，仮名 E と F と G がそれぞれ同じ QID 値の組み合わせになっている．各グループの保持しているセンシティブ属性値を参照すると，少なくとも 2 種類のセンシティブ属性値を保持していることがわかる．これにより個人のセンシティブ属性値の識別を防ぐことが可能となっている．

表 3　2-多様性を満たす病状データベース

仮名	性別	年齢	郵便番号	病状
A	女性	2*	999-45**	声帯ポリープ
B	女性	10〜29	999-****	マイコプラズマ肺炎
C	女性	10〜29	999-****	インフルエンザ肺炎
D	女性	2*	999-45**	インフルエンザ肺炎
E	男性	30〜49	999-133*	虫歯
F	男性	30〜49	999-133*	盲腸
G	男性	30〜49	999-133*	胃潰瘍

3.3　l-多様性の課題点

l-多様性にも課題が存在する．あるデータベースが類似したセンシティブ属性値を保持していた場合，l-多様性を満たすように匿名加工したとしても，事実上の多様性が満たされているとはいえない可能性が存

在する．

表 3 は 2-多様性を満たしているデータベースである．しかしながら，仮名 B と C のグループの保持しているセンシティブ属性値は “マイコプラズマ肺炎” と “インフルエンザ肺炎”である．確かに 2 種類のセンシティブ属性値を保持しているので既存指標の l-多様性は満たしているが，“肺炎”であるという情報は漏洩してしまう．つまり指標は満たしているが，事実上の多様性が満たされているとは言えない状態になっているのである．

3.4 t-近接性

また，l-多様性の拡張として t-近接性[4] が存在する．t-近接性は分布と分布の距離を考慮した指標であり，あるデータベースが t-近接性を満たすとは，データベース全体の属性値の分布と，QID 値が同一のレコードグループ内における属性値の分布が距離 t 以下となっていることを保証するものである．この指標を利用することでプライバシ保護を実現できるが，上記に示した l-多様性の課題は解決できていない．例えば，比較的 HIV の患者が多いデータベースであれば，上記の問題が発生しやすくなることがわかる．さらに，出現回数が少ない属性値の情報のほとんどを失うとともに，QID 値グループに属したレコードの情報も大きく失われてしまうという問題点も存在する．

3.5 差分プライバシ

上記のような k-匿名性や l-多様性，t-近接性とは異なるプライバシ指標として，差分プライバシ[5]，[6] が存在する．差分プライバシは識別不能性に基づいたプライバシ指標であり，ある個人の情報を含むデータベースに問い合わせをした結果と，その個人の情報を含まないをデータベースに問い合わせした結果が，区別できない場合にその問い合わせは安全であることが保障される．さらに近年では，データベースにノイズを追加して公開することで，差分プライバシを満たす手法[7]，[18] がいくつも提案されている．差分プライバシを用いることで攻撃者の背景知識によらず数学的な安全性を保証できるというメリットがあるが，l-多様性とはプライバシに対する考え方が異なっており，状況に応じて使い分ける必要がある．本論文では，匿名化対象のデータ提供元である一般の人々がそのメリットを理解しやすいという点等を考慮して l-多様性を対象とする．

4 提案指標

本章では既存指標の l-多様性では扱うことができない事実上の多様性を考慮した指標として，(l, d)-意味的多様性の提案を行う．

4.1 記号定義

個人に関する情報を含んだデータベースを T とし，T の i 番目のレコードを r_i と表す．ただし $i = 1, \cdots, N$ である．データベース T 内のセンシティブ属性値 $a_1, \cdots, a_F$ を要素として持つセンシティブ属性の集合を S とする．また，レコード r の保持するセンシティブ属性値を $E(r)$ と表す．

4.2 (l, d)-意味的多様性

課題解消のための提案指標である (l, d)-意味的多様性の定義を示す．

表 4 ダミー追加によって 2-多様性を満たしている病状データベース

名前	性別	年齢	郵便番号	病状
A	女性	24	999-4565	声帯ポリープ，マイコプラズマ肺炎
B	女性	16	999-5636	マイコプラズマ肺炎，虫歯
C	女性	21	999-4557	インフルエンザ肺炎，盲腸
D	女性	22	999-4531	インフルエンザ肺炎，声帯ポリープ
E	男性	34	999-1332	虫歯，インフルエンザ肺炎
F	男性	34	999-1335	盲腸，胃潰瘍
G	男性	48	999-1337	胃潰瘍，インフルエンザ肺炎

定義 ((l, d)-意味的多様性)：自然数 l 及び d に対して，T が以下を満たすとき，T は (l, d)-意味的多様性を満たすという．

T 内の任意のレコード r_i が l 個以上のセンシティブ属性値の候補を持ち，かつ r_i のセンシティブ属性値間の最小距離 d_i が $d \leq d_i$ となる．ただし，$d \leq \frac{F}{2}$ である．

また，センシティブ属性値のカテゴリ分け及び距離の定義は匿名化するデータベース毎に定義する．例えばセンシティブ属性が数値で表されている場合，カテゴリ分け及び距離はその値の差によって定義される．例としてセンシティブ属性が年収の場合を考えると，カテゴリ分けは 100 万円台，200 万円台，300 万円台 $\cdots$ と分けられる．そして 100 万円台と 200 万円台のように差が 100 万円ならば距離 $d = 1$，100 万円台と 300 万円台のように差が 200 万円ならば距離 $d = 2$ となっていく．類似度をもっと細かく考慮したいならばカテゴリ分けを 10 万円台で行い，差が 10 万円を $d = 1$ とすることで詳細な考慮が可能となる．

さらに病状などカテゴリ分けが木構造のように成された場合は，深さによって定義することもできる．距離 $d = 1$ は親ノードの下にあるノード，つまり兄弟ノードの事を指す．距離 $d = 2$ ならば上位ノードへ二つ移動し，移動後のノードから見た子孫ノード全てを指す．同様にして距離 $d = 3$，距離 $d = 4$ と上位ノードに移っていきそこからの子孫ノード全てを指す．

5 提案アルゴリズム

5.1 匿名化アルゴリズム

上記で定義した (l, d)-意味的多様性を満たすように匿名加工を行う．本論文では，匿名化の手法としてダミー追加手法を用いる．

ダミー追加手法は Sei ら [1] の提案した手法であり，QID を一般化せず各属性にダミーを追加することでプライバシを保護する手法である．表 4 は表 1 の各属性にダミー追加を行い，2-多様性を満たしているデータベースである．

本論文では各レコードのセンシティブ属性のみにダミーを追加する．センシティブ属性のみにダミーを追加するのは，QID 値が同一のダミーレコードを複数生成し，生成されたダミーレコードにそれぞれ異なるセンシティブ属性値を設定することと同一である．よって，その人物の QID 値を全て知っていたとしても，センシティブ属性値を知らない場合，どのレコードが当該人物のレコードであるか特定することはできない．

Algorithm 1 Anonymization protocol for record r

```
Input: Domain of a sensitive attribute S, Privacy level l and d
Output: Set of anonymized sensitive values for record r
 Create set r
  /*Adds original sensitive attribute*/
 R ⇐ {E(r)}
  /*Adds dummy sensitive attributes*/
for i = 1, ··· , (l − 1) do
    R ⇐ R ∪ rand(S \ {dist(d, R)})
end for
return R
```

匿名化アルゴリズムでは保持してあるセンシティブ属性値を参照し，提案指標を満たすことのできるセンシティブ属性値をダミーとして選択，追加する．この匿名化アルゴリズムを Algorithm 1 として示す．

ここで関数 $dist(d, R)$ はレコード r の持つセンシティブ属性値 $E(R)$ から距離 d 以下にあるセンシティブ属性値を全て抽出している．関数 $rand(B)$ は集合 B からランダムに異なる要素を抽出する関数である．

5.2 分析アルゴリズム

データ利用者はデータベースの統計値を欲していると想定しているので，より良い統計値を与えることができるよう，提案した匿名化アルゴリズムに適した分析アルゴリズムを示す．

データ利用者は上記の匿名化アルゴリズムによりダミーが追加された匿名化データベースを得てヒストグラムを作成する．その際に提案する分析アルゴリズムを利用することにより，元のデータから作成したヒストグラムに類似したヒストグラムを作成することができる．ダミー追加後のデータベース内におけるセンシティブ属性値 a_i の総数を ω_i とし，その内のダミーではない真のセンシティブ属性値の総数を x_i とする．以下に示す式 1 の x_i を求めることで匿名化前

Algorithm 2 Analysis protocol

Input: Reported sets, category size F, database size N, parameters l and d
Output: Distribution of true categories
for $i = 1, \cdots, F$ **do**
  $F_i \Leftarrow$ size of $dist(d, \{a_i\})$
  if $a_j \notin dist(d, a_i)$ **then**
    $q_{(i,j)} \Leftarrow (l-1)/(F - F_i)$
  else
    $q_{(i,j)} \Leftarrow 0$
  end if
end for
  $\hat{x}_i \leftarrow gauss(\omega_i = x_i + \sum_{k \neq j} q_{(k,j)} * x_k)$ for all i
  /*Calculate Simultaneous Equation*/
return $\hat{x}_i (i = 1, \cdots, F)$

と類似した結果が得られる．

$$\omega_i = x_i + \sum_{k \neq i} q_{(i,k)} * x_k \tag{1}$$

式 1 は，ダミー追加後のデータベース内におけるあるセンシティブ属性値の総数が元から保持していた真値と追加されたダミー値の二種からなっており，ダミー値の個数を期待値として求めることで，真値を推測することができるというものである．ここで $q_{(i,j)}$ はあるレコードがセンシティブ属性値 a_i を保持しているときに a_j をダミーとして選択する確率であり次の式 2 で表される．

$$q_{(i,j)} = \begin{cases} \frac{l-1}{F-F_i} & (a_j \notin dist(d, a_i)) \\ 0 & (otherwise) \end{cases} \tag{2}$$

F_i は a_i を保持している際の選択することのできないセンシティブ属性値数であり，距離 d が与えられた時の a_i の属するカテゴリの要素数である．以上の式を利用した分析アルゴリズムを Algorithm 2 に示す．ここで，関数 $gauss(f(x))$ は連立方程式を解くための関数であり Algorithm 2 では式 1 を与えて x_i を求め $\hat{x}_i$ に代入している．

6 検証実験

提案した分析アルゴリズムの有用性を確認するために検証実験を行う．個人に関する情報を含んだデータベースを提案した匿名化アルゴリズムにより匿名化し，その後匿名化データベースに対して分析アルゴリズムを適用する．分析アルゴリズムにより生成されたヒストグラムと元のデータベースのヒストグラムの差異を検証することで有用性を評価する．

6.1 評価方法

推測アルゴリズムの有用性を評価する指標として，式 3 の平均二乗誤差 (MSE:Mean Squared Error) を利用する．MSE は正解データと求めたデータの誤差を求めるものであり，値が小さいほど元のデータと類似していることがわかる．

$$MSE = \frac{1}{|F|} \sum_{i=0}^{|F|-1} \left(\frac{x_i}{N} - \frac{\hat{x}_i}{N} \right)^2 \tag{3}$$

式 3 の x_i は元のデータベース内におけるセンシティブ属性値 a_i の総数であり，$\hat{x}_i$ は分析アルゴリズムにより推測されたセンシティブ属性値 a_i の総数である．真値 x_i と推測値 $\hat{x}_i$ の差を検証している．

また，提案した推測アルゴリズムの有用性を示すために，単純分析によって推測されたヒストグラム及び既存研究[1] 内の分析アルゴリズムによって推測されたヒストグラムと比較する．単純分析は次の式 4 を利用することでそれぞれヒストグラムの推測を行うことができる．

$$\hat{x}_i = \frac{\omega_i}{l} \tag{4}$$

既存研究における分析アルゴリズムは以下の式 5 及びダミーを選択する確率ための式 6 で表される．

$$\hat{x}_i = \frac{-qN + \omega_i}{1 - q} \tag{5}$$

$$q = \frac{l-1}{F-1} \tag{6}$$

式 4 及び式 5 によって求められた推測値と分析アルゴリズムによって求められた推測値を MSE で比較し評価する．

さらに，LeFevre ら [2] によって提案された一般化を用いた既存手法である，Mondrian を利用した匿名化も行い比較する．ただし，Mondrian は提案指標に従うことができないので，l-多様性に従うように匿名化を行う．

6.2 評価実験に用いたデータ

実験に用いたデータは独自に作成した二種類のデータセットと，一種類の公表されているデータセット，合計三種類のデータセットを利用した．

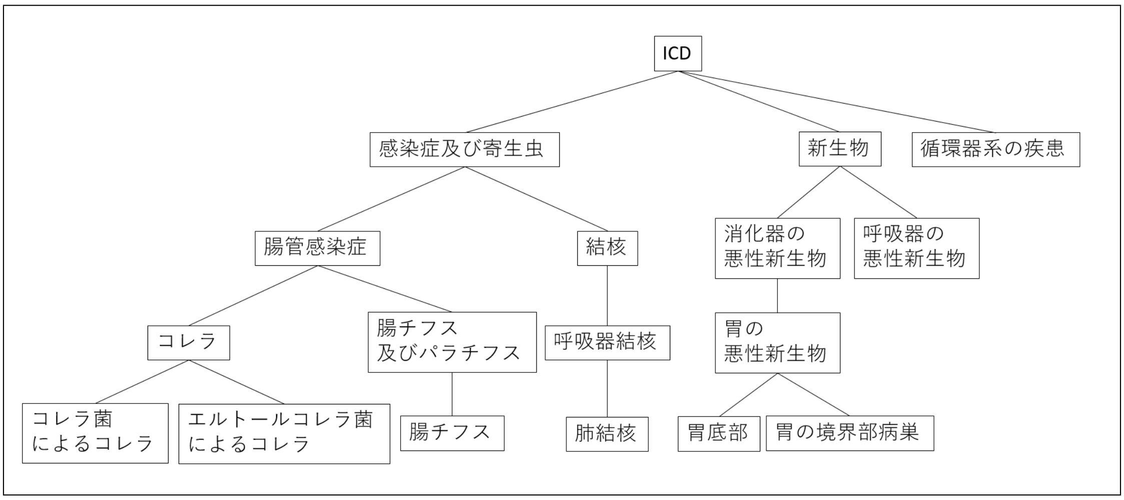

図 2　ICD の一部を木構造で表したもの

作成したデータセットはセンシティブ属性が年収である年収データセットと，センシティブ属性が病名である病状データセットの二種類で，両データセットともにレコード数は 10 万，QID として年齢と性別を含んである．病状は，世界保健機構 (WHO) によって公表されている，疾病及び関連保険問題の国際統計分類 (ICD:International Classification of Diseases)[15] を利用してセンシティブ属性値を設定しており，センシティブ属性の総数 $F = 8277$ である．年収は 100 万円以下, 200 万円以下,$\cdots$,1000 万円以下, 1500 万円以下, 2000 万円以下, 2500 万円以下, 2500 万円以上として，$F = 14$ である．

また，生成したデータセット内のセンシティブ属性値はランダムではなく，病状は厚生労働省から公開されている患者調査[16] を，年収は国税庁の公開している民間給与実態統計調査を参考にしている．よって，性別及び年齢と各センシティブ属性に相関があり，より現実に近いデータベースとなっている．

公開されているデータセットとして，UCI Machine Learning Repository より公開されている adult data set を利用した．adult data set は性別や年齢，学歴など 14 種類の QID を含んでおりレコード数は約三万と，多くの既存研究[1]，[2]，[11] でも活用されているデータセットである．本論文における検証実験では，既存研究[9]，[10] と同様に 14 種類の QID から性別，年齢，学歴など 8 種類の属性を QID として利用している．また，センシティブ属性としては独自に生成したデータセットと同様にして年収を付与した．表 5 は adult data set において利用した属性及び Mondrian を利用して匿名化を行う際の階層の深さを示している．各 QID の階層は既存研究を参考にしている．

表 5　adult data set の階層

	Attribute	Distinct Values	Generalizations
1	Age	74	4
2	Gender	2	1
3	Race	5	1
4	Marital Status	7	2
5	Education	16	3
6	Native country	41	2
7	Work class	7	2
8	Occupation	14	2

6.3　距離定義

匿名化アルゴリズム内における，センシティブ属性値のカテゴリ分け及び，センシティブ属性値間の距離定義は病状は ICD に準拠しておこなった．ICD は深さ 4 の木構造の統計分類となっており，センシティブ属性値としては深さ 4 の末端ノードのみを利用して

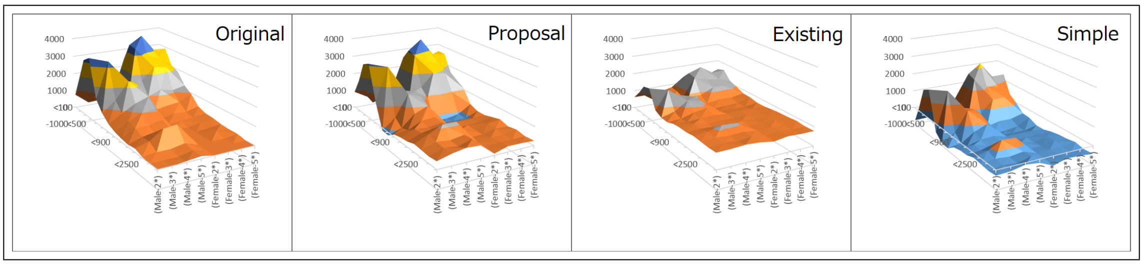

図 3　年収データセットに対する各ヒストグラム

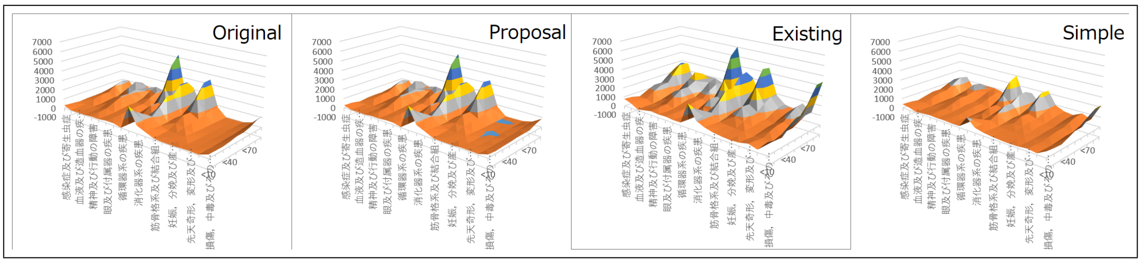

図 4　病状データセットに対する各ヒストグラム

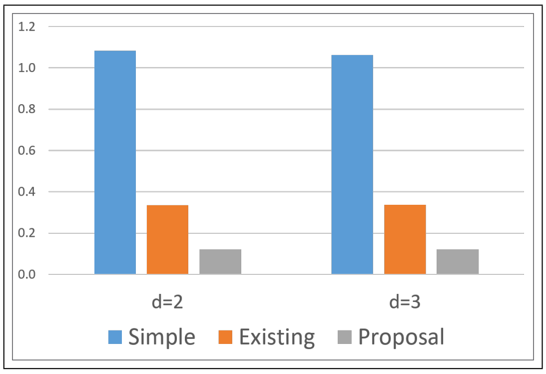

図 5　病状データセットの各 MSE

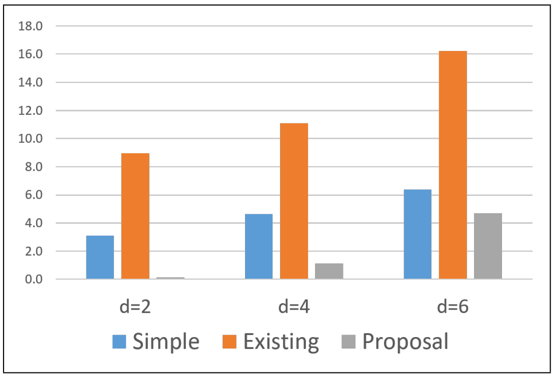

図 6　年収データセットの各 MSE

ある．カテゴリ分類は ICD に従い，センシティブ属性値間の距離定義は，共通の先祖ノードまでの距離で定義する．図 2 は ICD の一部を木構造で表現したものである．

図 2 より例えばあるレコードが“コレラ菌によるコレラ”というセンシティブ属性値を保持していたとする．距離 1 ならば“コレラ”の子孫ノード，つまり“コレラ菌によるコレラ”及び“エルトールコレラ菌によるコレラ”を指す．距離 2 ならば“腸管感染症”の子孫ノード，つまり“コレラ菌によるコレラ”，“エルトールコレラ菌によるコレラ”及び“腸チフス”を指す．なので (3, 2)-意味的多様性を満たすように匿名化した場合，“肺結核”，“胃底部”，“胃の境界部病巣”のいずれかから二つダミーとして選択され追加される．

また，年収は 100 万円を $d = 1$ として距離の定義をしている．例えばあるレコードが 500 万というセンシティブ属性値を保持していた場合，d=3 ならば，300 万以上 700 万以下の値を除いたいづれかのセンシティブ属性値よりダミーが選択される．

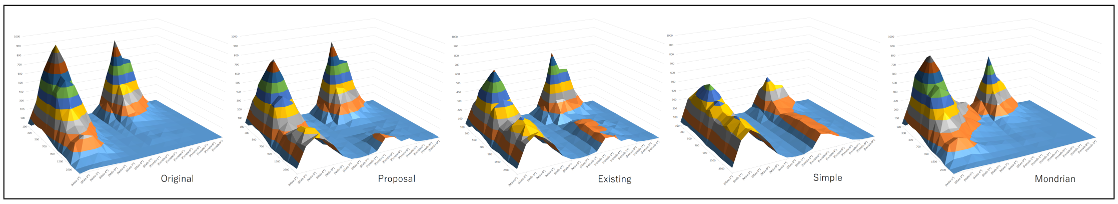

図 7　adult data set に対する各ヒストグラム

7　実験結果

7.1　検証結果

まず，独自に生成したデータセットに対しておこなった検証実験の結果を示す．図 3 は年収データセットをオリジナルデータとした，オリジナルデータのヒストグラム，分析アルゴリズムにより推測されたヒストグラム，単純分析により推測されたヒストグラムである．同様にして，図 4 は病状データセットをオリジナルデータとした，各ヒストグラムである．ヒストグラムにおける x 軸と y 軸はセンシティブ属性値の分類と年齢をそれぞれ表しており，z 軸が度数を表している．年齢は 10 歳ごと分類しており，図 3 のセンシティブ属性は ICD における最上位の分類を利用して 19 種に分類，図 4 のセンシティブ属性は 100 万円以下, 200 万円以下,···,1000 万円以下, 1500 万円以下, 2000 万円以下, 2500 万円以下, 2500 万円以上としている．

また，分析アルゴリズムによる推測の MSE と，単純分析，既存手法にによる推測の MSE をそれぞれ求めた結果は次の図 5，図 6 のようになった．

次に adult data set に対しておこなった検証実験に対する結果を示す．提案アルゴリズムを用いて匿名化した adulta data set に対して，上記と同様に検証実験をおこなった．さらに adult data set に対して Mondrina を利用した匿名化も行い，分析をおこなった．以上の 4 通りの結果に対して MSE を求めた結果を次に示す．ここで，提案手法の分析，既存手法の分析，単純分析の際の匿名化は $(2,4)$-意味的多様性を満たすようにおこなわれているが，Mondrian は (l,d)-意味的多様性ではなく，l-多様性に従っており，2-多様性を満たすように匿名化をおこなっている．また，図 7 はそれぞれの分析により作成されたヒストグラムである．

$$
\begin{aligned}
\text{MSE (Proposal)}: &\quad 1.0899 \times 10^{-4} \\
\text{MSE (Existing)}: &\quad 1.8748 \times 10^{-4} \\
\text{MSE (Simple)}: &\quad 3.0421 \times 10^{-4} \\
\text{MSE (Mondrian: 2-diversity)}: &\quad 0.7499 \times 10^{-4}
\end{aligned}
$$

7.2　考察

図 3，図 4 より提案した分析アルゴリズムによって推測されたヒストグラムが元のヒストグラムに最も類似していることがわかる．例えば単純分析の結果は全体的に平均化されており，特徴が薄れてしまっているが提案手法によるヒストグラムは特徴的な凹凸までしっかりと残っている．

また，図 5，図 6 のグラフより提案アルゴリズムが最も誤差が小さいことが確認できる．既存手法の MSE 値が単純分析の MSE 値よりも大きいものが存在する理由として，既存研究の分析アルゴリズムは既存研究の匿名化アルゴリズムに適しており，今回の匿名化アルゴリズムに必ずしも適していると言えず良い結果になると保証できないからである．以上の事から匿名化アルゴリズムと共に，匿名化アルゴリズムに適した分析アルゴリズムも提案すべきという事実も確認できる．

adulta data set に対する検証実験の結果において，提案手法の MSE が Mondrian の MSE よりも大きい原因として，Mondrian は (l,d)-多様性には従っていないからである．プライバシ保護度合いとデータの有用性はトレードオフの関係であり，プライバシ保護を優先している提案手法の方が誤差が大きい．しかしながら，提案手法と Mondrian の MSE の差は小さく，提案手

法が有用であると言える．

8 本論文のまとめ

本論文では，個人に関する情報を含んだデータベースを匿名加工する際，センシティブ属性値の類似から発生する恐れのある，事実上の多様性を考慮することができないという既存指標の課題を解消するために新たな指標 (l,d)-意味的多様性を提案した．(l,d)-意味的多様性はカテゴリ分けから定義されたセンシティブ属性値間の距離 d を，考慮することにより詳細な多様性まで扱っている．

また，提案指標に従ったダミー追加手法を用いた匿名化アルゴリズムの提案及び提案アルゴリズムに適した分析アルゴリズムの提案を行い，検証実験の結果から提案された手法が優れていることを確認した．本論文における提案を改善するために次のような項目があげられる．(1) 複数センシティブ属性値への適応．(2) よりよい推測アルゴリズムの提案．(3) 差分プライバシへの適用．以上のような点を検討していく．

謝 辞 本研究は JSPS 科研費 JP16K12411, JP17H04705 の助成を受けたものです．

参考文献

[1] Y. Sei, H. Okumura, T. Takenouchi, A. Ohsuga, *Anonymization of Sensitive Quasi-Identifiers for l-diversity and t-closeness*, IEEE Transactions on Dependable and Secure Computing, 2017, DOI: 10.1109/TDSC.2017.2698472.

[2] K. LeFevre, D. DeWitt, and R. Ramakrishnan, *Mondrian Multidimensional K-Anonymity*, in Proc. IEEE ICDE, 2006, p. 25.

[3] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, D. Kifer, and M. Venkitasubramaniam, *l-diversity: Privacy beyond K-Anonymity*, ACM TKDD, Vol. 1, no. 1, p. 3, 2007.

[4] N. Li, T. Li, and S. Venkatasubramanian, *t-closeness: Privacy beyond k-anonymity and l-diversity*, in Proc. IEEE ICDE, 2007, p. 106–115.

[5] C. Dwork, *Differential Privacy*, in Proc. ICALP, 2006, p. 1–12.

[6] G. Acs, C. Castelluccia, and R. Chen, *Differentially Private Histogram Publishing through Lossy Compression*, in Proc. IEEE ICDM, 2012, p. 1–10.

[7] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, *Enhancing data utility in differential privacy via microaggregation-based k-anonymity*, The VLDB Journal, vol. 23, no. 5, p. 771–794, 2014.

[8] K. LeFevre, D. DeWitt, and R. Ramakrishnan, *Incognito: Efficient full-domain k-anonymity*, in Proc. ACM SIGMOD, 2005, p. 49-60.

[9] V. Iyengar, *Transforming data to satisfy privacy constraints*, in Proc. ACM SIGKDD, 2002, p. 279-288.

[10] N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, *Centralized and Distributed Anonymization for High-Dimensional Healthcare Data*, ACM TKDD, Vol. 4, no. 4, p. 18, 2010.

[11] K. Mancuhan, and C. Clifton, *Statistical Learning Theory Approach for Data Classification with l-diversity*, in Proc. SIAM International Conference on Data Mining, 2017, p. 651-659.

[12] Z. Huang, and W. Du, *OptRR: Optimizing Randomized Response Schemes for Privacy-Preserving Data Mining*, in Proc. IEEE ICDE, 2008, p. 705–714.

[13] P. Kairouz, K. Bonawitz, and D. Ramage, *Discrete Distribution Estimation under Local Privacy*, in Proc. IEEE ICML, 2016, p. 2436-2444.

[14] R. Agrawal, and R. Srikant, *Privacy-Preserving Data Mining*, in Proc. ACM SIGMOD, 2000, p. 439-450.

[15] 厚生労働省, 疾病、傷害及び死因の統計分類, http://www.mhlw.go.jp/toukei/sippei/

[16] 厚生労働省, 平成 *26* 年 *(2014)* 患者調査, http://www.mhlw.go.jp/toukei/list/10-20.html

[17] 総務省法令データ提供システム, 個人情報の保護に関する法律, http://law.e-gov.go.jp/htmldata/H15/H15HO057.html

[18] 寺田雅之, 鈴木亮平, 山口高康, 本郷節之., 大規模集計データへの差分プライバシの適用, 情報処理学会論文誌, 56(9), p. 1801-1816, 2015.