

背景知識を用いた推測を困難にしデータ歪曲度を極小化する プライバシー保護手法

村本 俊祐[†] 上土井陽子^{††} 若林 真一^{††}

^{† †} 広島市立大学大学院情報科学研究科

〒 731-3194 広島市安佐南区大塚東三丁目 4-1

E-mail: [†]shun@icl.ce.hiroshima-cu.ac.jp, ^{††}{yoko,wakaba}@ce.hiroshima-cu.ac.jp

あらまし 本稿ではデータベース上の入力データテーブルにおいてデータの一般化を行うことにより、 l -多様性という性質を保持させるプライバシー保護技法について考察する。我々は以前にデータテーブルに k -匿名性を保持させることによりデータ組合せによるデータ推測を防止し、かつデータ歪曲度（元のデータテーブルとのデータ値の変化の度合い）が極小な結果データテーブルを出力するアルゴリズムを提案した。しかし近年、 k -匿名性をもってしても防ぐことの出来ない新種の攻撃の存在が指摘されている。よって、 l -多様性という新しい性質を取り込むことにより新種の攻撃を防止し、プライバシー保護がより確実となるアルゴリズムへの改良を行い、データ歪曲度の低い結果データテーブルの出力を可能とするアルゴリズムを提案する。

キーワード データテーブル, プライバシー保護, 一般化, l -多様性, k -匿名性

Minimization of Data Distortion on a Privacy Protection Technique against Attacks Using Background Knowledge

Shunsuke MURAMOTO[†], Yoko KAMIDOI^{††}, and Shin'ichi WAKABAYASHI^{††}

^{† †}Graduate School of Information Sciences, Hiroshima City University

3-4-1, Ozuka-higashi, Asaminammi-ku, Hiroshima, 731-3194 Japan

E-mail: [†]shun@icl.ce.hiroshima-cu.ac.jp, ^{††}{yoko,wakaba}@ce.hiroshima-cu.ac.jp

Abstract In this paper, we consider a privacy protection technique to convert an input data table in a database into one maintaining l -diversity by generalizing data. In the previous work, we proposed an algorithm which prevents data guess by the data combination by letting data table maintain k -anonymity, and outputs result data tables with small data distortion degree (a degree of difference between original data and result data). Recently, however, there exists a new kind of attacks that we cannot prevent even with k -anonymity. Therefore, *Machanavajjhala* et al. introduced a new property as the l -diversity to prevent such new kinds of attacks, and we improve our algorithm for privacy protection against the new attacks so as to output result tables by minimal distortion.

Key words data table, privacy protection, generalization, l -diversity, k -anonymity

1. はじめに

統計調査や医療によって得られたデータで、かつ集計されるまえの個票データ（マイクロデータ）は分析者がそれぞれ独自の視点で再分析可能であることから一般に高い価値を持つ。マイクロデータに対するプライバシー保護の簡単な方法に重要な識別情報（名前など）を非公開にする方法がある。しかし、ただ単に識別情報を非公開にただけではデータテーブル数個を組み合わせることで非公開のデータ項目が推測できる可能性がある。データ項目の推測を防ぐために、データテーブルに

k -匿名性を持たせることが考えられていた [7]。従来手法 [1] [6] では k -匿名性保持のためのデータ操作で結果データを過度に歪曲したり、確実な推測防止が保証できないという欠点があった。我々はそれらの欠点の克服を目的としてデータ歪曲度という評価指標を提案し新しいプライバシー保護アルゴリズムに導入し、評価した。しかし、近年 k -匿名性を保持していても防ぐことの不可能な深刻な攻撃の存在が指摘されている [2] [3]。そこで、文献 [2], [3] で新たに定義された l -多様性という性質に注目し、この性質をアルゴリズムに取り入れ、更にデータ歪曲度算出関数 *DIS* と組み合わせることで、データ歪曲度が低く、か

表 1 2-匿名性保持を目的とした一般化

(a) 初期テーブル PT				(b) 一般化テーブル RT					
	Race	Birth	Gender	ZIP		Race	Birth	Gender	ZIP
t1	Black	1964	female	02138	t1'	Black	1964	female	02138
t2	Black	1964	female	02138	t2'	Black	1964	female	02138
t3	Black	1967	male	02141	t3'	Person	196*	male	02141
t4	White	1971	female	02139	t4'	White	1971	human	02139
t5	White	1967	male	02141	t5'	Person	196*	male	02141
t6	White	1971	male	02139	t6'	White	1971	human	02139
t7	White	1965	male	02141	t7'	Person	196*	male	02141

つ l -多様性を保持するデータテーブルを出力できるアルゴリズムを提案する。

2. k -匿名性

本研究ではデータテーブルとして表 1 のような有限個のタブル (行に対応) と属性 (列に対応) からなるものを考慮する。ここで各タブルは各属性に属するデータ値の属性数 n 個の組とする。また、データ推測から秘密にしたい情報 (ここでは個人) を特定する単独の識別子ではないが組み合わせることで同じ働きをする恐れのある属性の集合を準識別子 QI と呼ぶ。

従来手法 [1], [6] ではテーブルに k -匿名性を持たせるために一般化や、抑制というデータ操作を使用していた。まず抑制とはデータ値がすべて隠された状態を指す。データ値の状態を大きく分けると初期状態と抑制状態に分けられる。一般化とは、その二つの状態の中間の状態を示すために、データ値の一部分を隠す、またはより広い値域を指す値に変換するデータ操作である。

ここで k -匿名性を以下のように定義する。

データテーブル中の各タブルにおいて、そのタブルのもつデータ値情報 (各属性値の組合せ) と同じデータ値情報を持つタブルが自分自身を含め k 個以上存在する状態

k -匿名性を保持するテーブルの例を挙げる。表 1(a) のテーブル PT が与えられたとき、表 1(b) のテーブル RT に変換したとする。テーブル PT のタブルに注目すると、この状態では $t1, t2$ のタブルは同一データ値組合せであるが、その他のタブルは異なっている。一方、テーブル RT では、 $t1, t2$ のタブルが同一データ値組合せを持っており、同様に $t3, t5, t7$ の 3 つのタブル、 $t4$ と $t6$ の 2 つのタブルがそれぞれ同一データ値組合せを持っている。よって、テーブル RT では全てのタブルにおいて同一データ値組合せをもっているタブルが自分を含め 2 個以上存在する。このとき、テーブル RT は 2-匿名性を保持していると言う。 k -匿名性 ($k \geq 2$) を保持しているテーブルではどのタブルも公開前データのタブルに一意に対応していないので複数データ項目の組合せによる、データ推測が防止されているといえる。

3. 新種の攻撃

上記において、データ組合せによるデータ推測を防止することを考え、データテーブルに k -匿名性を保持させることでプライバシー保護を確実にしようとしてきた。しかし、近年ではこ

表 2 新種攻撃の例

(a) 初期テーブル				(b) 4-匿名テーブル			
ZIP	Nationality	Condition	Name	ZIP	Nationality	Condition	Name
13053	Russian	Heart Disease		1305*	*	Heart Disease	
13053	American	Heart Disease		1305*	*	Heart Disease	
13052	Japanese	Viral Infection	Umeko	1305*	*	Viral Infection	Umeko
13052	American	Viral Infection		1305*	*	Viral Infection	
13065	American	Cancer	Bob	1306*	*	Cancer	Bob
13068	India	Cancer		1306*	*	Cancer	
13067	Japanese	Cancer		1306*	*	Cancer	
13068	American	Cancer		1306*	*	Cancer	

の k -匿名性を持ってしても防ぎきれない攻撃が存在することが指摘されている。その攻撃とは

- ・同種攻撃 (Homogeneity Attack)
- ・背景知識攻撃 (Background Knowledge Attack)

の 2 種類の攻撃が挙げられる。

3.1 同種攻撃

今まで、準識別子としてきた属性の中で実際に一般化を行ってもかまわない属性 (Race, ZIP など) を総称して非注目属性 (Non-Sensitive-Attribute) と呼ぶ。また、データ解析者にとって重要な項目であるため一般化を行って欲しくない準識別子 (例として医療データであるならば病名など) を総称して注目属性 (Sensitive-Attribute) と呼ぶことにする。

例として、まずはじめに表 2(a) のテーブルが与えられ、Bob という人のデータは実際に下から 4 番目のタブルに当たるとする。このデータテーブルに 4-匿名性を保持させた結果を表 2(b) とする。

たしかに表 2(b) の結果が k -匿名性を保持していることから、Bob のデータがどのタブルであるかという確証は得られない。しかし、仮に攻撃者が Bob の非注目属性の値を知っていたとするならば、Bob がどのタブルグループに含まれているかわかってしまう可能性がある。なぜならば、一般化というデータ操作には弱点があるからである。一般化というデータ操作は、一般化後の値から一般化前の値を特定することは不可能だが、一般化前の値から一般化後の値を推測することが容易であるという弱点がある。さらに表 2(b) の結果において Bob のデータが含まれるであろうと推測されるタブルグループの注目属性が全て Cancer である。よって、Bob の非注目属性の値を知っている攻撃者は Bob のデータがどのタブルに当たるかは確証を得られないが、タブルグループ中の全ての注目属性の値が同一な為、Bob の病名が Cancer であると確信されてしまう。

3.2 背景知識攻撃

同種攻撃の例と同じく、まずはじめに表 2(a) のテーブルが与えられ、表 2(b) のテーブルに変換されたとする。また、日本人の患者である Umeko のデータは上から 3 番目のタブルに当たるとする。

仮に“日本人は Heart Disease の発病率が非常に低い”という事実があるとする。同種攻撃の時と同じように攻撃者は Umeko の非注目属性の値を知っているならば、Umeko のデータが含まれるタブルグループがどれか特定することは可能である。更にこの場合、“日本人である Umeko”のデータは注目属性の値が Heart Disease であるタブルには含まれないことから表 2(b) 中

の上から 1,2 番目のタプルが Umeko の候補から外れる．従って，一意的に Umeko の病名は Viral Infection であると確信されてしまう．

4. l -多様性

k -匿名性だけでは新種の攻撃を防ぐことが不可能であることがわかった．そこで l -多様性 [2] [3] という性質に注目する．

ここでは，非注目属性の属性値において，一般化されたデータが同一データ組合せを持ち，その組合せが仮に q^* となる場合のタプルグループをまとめて q^* -ブロックと呼ぶことにする． k -匿名性を保持しただけでは新種の攻撃を防ぐことが出来なかった理由として次の 2 点が挙げられる．

- ・ q^* -ブロック中の注目属性の多様性の欠如
- ・ 強力な背景知識

これらの原因において対処を可能とするための l -多様性の原則を簡単に示すと次のように表すことが出来る．

テーブル中の全ての q^* -ブロックにおいて出現する注目属性の値が少なくとも l 個の多様性を持つ状態を保つ

また l -多様性における l 個の多様性を保持している状態を定義するのに次の手法が挙げられる．

- ・ エントロピー l -多様性
- ・ 再帰的 (c, l) -多様性

4.1 エントロピー l -多様性

エントロピー l -多様性とはエントロピーを用いて l -多様性の l の値を導出する手法である．テーブル中の q^* -ブロックに注目し，以下の式が満たされる l の値がその q^* -ブロックの保持している l -多様性の l の値となる．式中の関数 $P(q^*, s)$ は q^* -ブロック中において非注目属性の値の組合せが q^* かつ注目属性の値が s であるタプルの確率を算出するものとする．また，注目属性の母集合を S とする．

$$-\sum_{s \in S} P(q^*, s) \log(P(q^*, s)) \quad \log(l)$$

テーブル中の全ての q^* -ブロックについて上記の式が満たされるとき，テーブルは l -多様性を保持していると言う．

4.2 再帰的 (c, l) -多様性

データテーブル中の各 q^* -ブロックにおいて出現する注目属性の値それぞれの出現確率を算出する．算出した確率値の中で最も高い注目属性の値の確率を r_1 と置き，次に高い確率値を $r_2 \dots$ と順に置く．仮にその q^* -ブロック中には n 種類の注目属性の値が存在するとし，定数 c ($c > 0$) を任意に定めたとき，次の式が成り立てば，この q^* -ブロックは再帰的 (c, l) -多様性を満たしているといえる．

$$r_1 \leq c(r_l + r_{l+1} + \dots + r_n), \quad 1 < l < n$$

同様に他の q^* -ブロックについても式を当てはめ，テーブル

中の全ての q^* -ブロックが再帰的 (c, l) -多様性を満たしているならば，このテーブルは再帰的 (c, l) -多様性を満たしているという．

4.2.1 Postive Disclosure-再帰的 (c, l) -多様性

いくつかの注目属性の値と個人との関係が公開されてもよい場合について考える．例として，一般的な病院における風邪と個人との関係や，あらかじめその病院には心臓病を患っている人が多く通院している，などの場合が挙げられる．この場合，風邪，心臓病それぞれと個人との関係が攻撃者にわかったとしても問題は発生しないといえる．このようにあるデータテーブルにおいて非常に高い確率で出現する注目属性の値のことを Positive Disclosure と呼ぶ．

Positive Disclosure でない値で最も出現頻度の高い値を y 番目に出現頻度が高い値とすると

$$\begin{aligned} \cdot y &= l-1 & r_y &= c \sum_{j=l}^m r_j \\ \cdot y &> l-1 & r_y &= c \sum_{j=l-1}^{y-1} r_j + c \sum_{j=y+1}^m r_j \end{aligned}$$

以上全ての y に対して 2 式のどちらかを満たしていれば，この q^* -ブロックは Postive Disclosure-再帰的 (c, l) -多様性を満たしているという．また同様にテーブル中の全ての q^* -ブロックが Postive Disclosure-再帰的 (c, l) -多様性を満たしているならば，このテーブルは Postive Disclosure-再帰的 (c, l) -多様性を満たしているという．

4.2.2 Negative/Postive Disclosure-再帰的 (c_1, c_2, l) -多様性

テーブル中には非注目属性の値と注目属性の値の組合せにおいて極端に出現確率の低いものがある．例を挙げると，日本人という非注目属性を含むタプル集合において心臓病という注目属性を含むタプルがテーブル中にほとんど現れない．これは日本人が心臓病にかかりにくいことを表しているといえる．このような非注目属性と注目属性の組合せを Negative Disclosure と呼ぶ．

Postive Disclosure-再帰的 (c_1, l) -多様性において Negative Disclosure を考慮する．Negative Disclosure になることが許されない注目属性の集合を W としたとき， $s \in W$ である全ての注目属性において， s を含むタプルが各 q^* -ブロックに少なくとも c_2 % (c_2 は任意に定めた 100 より小さい定数) 出現し，かつどの q^* -ブロックも Postive Disclosure-再帰的 (c_1, l) -多様性の条件を満たしている時，このテーブルは Negative/Postive Disclosure-再帰的 (c_1, c_2, l) -多様性を満たしているという．

4.3 複数の注目属性

これまでの例では，注目属性が単一の場合において考えてきた．実際のデータテーブルにおいて注目属性が必ずしも単一である保証はない．しかし，注目属性が複数設定されている場合でも基本的な l -多様性の定義に変更はない．

注目属性が仮に n 個設定されている場合を想定する． Q_1, Q_2, \dots, Q_m の非注目属性と S_1, S_2, \dots, S_n の注目属性からなるテーブルにおいてまず S_1 に注目したとすると， S_1 以外の注

Input: テーブル PT ; 準識別子 $QI = (A_1, \dots, A_n)$, 整数 $k(k \geq 2 \wedge |PT| \geq k)$, DGH_{A_i}, VGH_{A_i} , ここで $i = 1, \dots, n$

Output: k -匿名性を保持したテーブル MGT

step1. PT が k -匿名性を満たしているならば step3 へ .
 step2. ランダムに選んだ頻度 k 以下のタプルと一般化を行ったとき
 DIS の値が最も低いタプルを探し出し一般化を行い, step1 へ .

step3. $MGT \leftarrow PT$
 step4. **Return** MGT

図 1 k -匿名性を保持を目的としたアルゴリズム $MinDIS$

目属性である S_2, \dots, S_n を一時的に非注目属性とみなして, l -多様性の条件を満たしているか確認を行う. 確認が終了したら次は S_2 について注目し, S_1, S_3, \dots, S_n を非注目属性とみなして, 同様に l -多様性の条件について確認を行う. 以下同様に S_i ($1 \leq i \leq n$) に注目している間は $S_1, \dots, S_{i-1}, S_{i+1}, \dots, S_n$ を非注目属性とみなして l -多様性の条件について確認を行う. すべての注目属性において l -多様性の条件を満たしていれば, このテーブルは l -多様性を保持しているといえる.

5. データ歪曲度算出関数 DIS

文献 [4], [5] におけるプライバシー保護のための我々の従来のアルゴリズム $MinDIS$ では, k -匿名性保持を目的とした一般化をデータテーブルに適用することにより, 複数データ項目の組合せによるデータ推測を防ぐことを考えた. k -匿名性保持を目的としたアルゴリズム $MinDIS$ を図 1 に示す.

前述のとおり, データテーブルに k -匿名性を保持させるためには, 一般化等のデータ操作を必要とする. しかし, 一般化等のデータ操作は元のデータを歪曲してしまう. 従来のアルゴリズム $MinDIS$ では, データを利用する際に解析などが行いやすいように元のデータになるべく近い形で k -匿名性を保持したデータに変換することを考えた. したがって, より歪曲の少ない結果を出す必要があった.

また同様に, 本稿において提案する l -多様性を取り入れたアルゴリズム $DiverDIS$ でも, 同種攻撃, 背景知識攻撃を防ぐためにデータ操作に一般化を使用して, l -多様性を保持させることを考えている.

よって, 本稿でも一般化を行うことで得られたデータテーブルが元のデータテーブルに対して, どの程度変化したか (データ歪曲度) を評価するため, データ歪曲度算出関数 DIS [4], [5] をアルゴリズムに使用する.

提案アルゴリズムは属性を初期値の集合から最大一般化値までに一般化された回数で階層的に分ける属性一般化階層 DGH (Domain Generalization Hierarchies) と, 一般化前の値と後の値の関係を木 (最大一般化値を根とする) で表現した値一般化階層 VGH (Value Generalization Hierarchies) という一般化表現を使用している. 表 1 のデータテーブル中の属性 ZIP に関する属性一般化階層 DGH および値一般化階層 VGH の例を図 2 に示す. 属性一般化階層 DGH , 値一般化階層 VGH

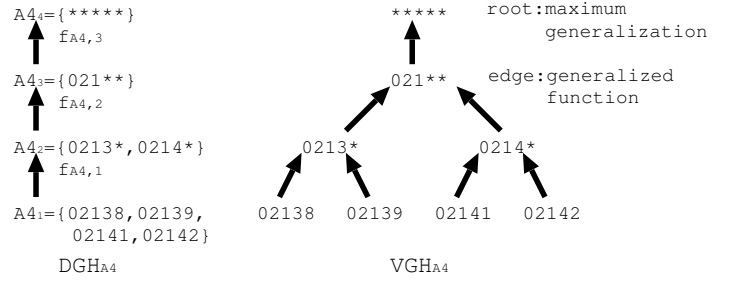


図 2 属性 A_4 (ZIP) の属性一般化階層 DGH , 値一般化階層 VGH

はデータテーブル上の各属性それぞれに対して定義され, 複数階層から成る. 基本的に属性一般化階層 DGH と値一般化階層 VGH はデータテーブルの管理者が任意に作成可能である. また DGH_{A_i}, VGH_{A_i} は属性 A_i の属性一般化階層 DGH と値一般化階層 VGH という意味をもつとする.

テーブル PT が一般化テーブル RT に変換されたときのデータ歪曲度算出関数 DIS の定義式を以下に示す.

$$DIS(RT) = \frac{\sum_{A_i \in QI} \sum_{t_j \in PT} \frac{h(VGH_{A_i}, t_j(A_i)) - h(VGH_{A_i}, t_j'(A_i))}{|DGH_{A_i}|}}{|PT| \cdot |QI|}$$

式中の t_j はタプルを指し, $t_j(A_i)$ でタプル t_j 中の属性 A_i に対応する値を示し, 関数 $h(tree, v)$ は木 $tree$ 中の値 v の高さを返す関数である. また DGH の絶対値は属性一般化階層関数 DGH の階層数を表わすとする. 一般化テーブル RT が一般化される前のテーブル PT とデータ値がまったく同じであれば $DIS(RT)$ は 0 となる. また, 一般化が行われるにつれて数値は大きくなり, 全てのデータ値が完全に抑制された状態 (すべてが * 等の情報が得られない状態) だと $DIS(RT)$ は 1 となる. したがって, データ歪曲度算出関数 DIS は 0 から 1 の値を取る.

6. 提案アルゴリズム $DiverDIS$

我々が以前提案したアルゴリズム $MinDIS$ は k -匿名性を取り入れたアルゴリズムだった. k -匿名性において同一データ組合せのタプルグループを作ることは l -多様性においての q^* -ブロックにそのまま当てはめることが可能であることから, l -多様性の条件を満たすことにより, 自動的に k -匿名性の条件を満たしたことになる.

また, 従来のアルゴリズム $MinDIS$ を改良するに当たり, 重要な注意すべきアルゴリズムの特性が存在する. その特性とは単調性質 (Monotonicity Property) である. 単調性質とは,

一度ある性質を保持したら, 同様の操作を行ったとしても性質が崩れることがない特性

と定義される. 従来のアルゴリズム $MinDIS$ は k -匿名性が単調性質を保持している故にボトムアップに q^* -ブロックを結合することで k -匿名性を満たす結果テーブルを見つけ出すことができる. また l -多様性も単調性質を保持していることは文

Input: テーブル PT ; 非注目属性 $Q = (Q_1, \dots, Q_m)$, 注目属性 $S = (S_1, \dots, S_n)$, 整数 $l (l \geq 2 \wedge |PT| \geq l)$, $c (c > 0)$, DGH_{Q_i}, VGH_{Q_i} , ここで $i = 1, \dots, n$
Output: l -多様性を保持したテーブル MGT

step1. PT から多様性リスト $d-list$ を作成する .
 step2. If (PT が l -多様性を満たしている) then do
 step2.1. $MGT \leftarrow PT$, step4 へ .
 step3. else do
 step3.1. ランダムに l -多様性を満たしていないタプルを選ぶ .
 step3.2. 仮に一般化を行った際に多様性が向上する
 もしくは, すでに l -多様性を満たしている一般化を行っても l -多様性が崩れないタプルの中で
 最も DIS が小さいタプルを探す .
 step3.3. 選ばれた 2 つのタプルを一般化して q^* -ブロックとし, $d-list$ を更新 .
 step3.4. q^* -ブロックが l -多様性を満たしているならば step1 へ
 そうでないならば step3.2 へ戻る .
 step4. Return MGT

図 3 提案アルゴリズム *DiverDIS*

献 [2], [3] で証明されている . よって我々の従来のアルゴリズム *MinDIS* において k -匿名性の判定条件を l -多様性の判定条件に変更したとしても提案アルゴリズムは成り立つことが保証される .

しかし, k -匿名性の条件判定を l -多様性の条件に変更することでデータ推測を防止し, 新種の攻撃も防止可能なアルゴリズムに改善出来たと言えるが, *MinDIS* で行っていた一般化部分のアルゴリズムは, あくまで k -匿名性保持を目指したアルゴリズムとなっている . したがって必ずしも l -多様性保持を目指したデータ一般化を行っているとは言い難い . つまり, プライバシー保護は保証されているが, 無駄なデータ一般化が多く行われる可能性があり, データ歪曲度が高い結果テーブルを出力する可能性がある .

したがって, データ一般化部分のアルゴリズムに関しても l -多様性を目指した一般化を行うように条件を設定しなければならない . 以上の点を考慮し, 新しいアルゴリズム *DiverDIS* を図 3 に提案する . アルゴリズム中の l -多様性の判定条件を実現する手法は 4 節で記述した手法のどれでも可能である . また多様性リスト $d-list$ とは, 各タプルを含む q^* -ブロックの保持している l -多様性の l の値をまとめたリストである .

6.1 実行例

提案アルゴリズム *DiverDIS* に図 3 のデータテーブルが入力テーブルとして与えられた場合の例を以下に示す . l -多様性の条件判定には再帰的 (c, l)-多様性を用い, l の値は 3, c の値は 1 と設定するとする . また, 属性 ZIP の VGH には図 2 と同じように最下位桁から一般化を行う階層を使用し, 属性 $Nationality$ は図 4 の階層を使用する .

まず step1 で多様性リスト $d-list$ を作成する . 入力された直後のデータテーブルは, どのタプルも独立しているため多様性は 1 となる . よって step2 の l -多様性条件では偽となりそのまま step3 に進む .

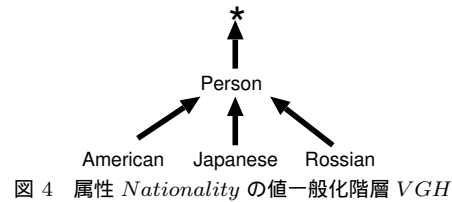


図 4 属性 $Nationality$ の値一般化階層 VGH

表 3 入力テーブル

	ZIP	Nationality	Condition
t1	13050	American	Heart Disease
t2	13050	Japanese	Cancer
t3	13051	American	Heart Disease
t4	13051	Rossian	Viral Infection
t5	13051	Japanese	Wheezing
t6	13052	Rossian	Cancer
t7	13063	American	Wheezing
t8	13063	Japanese	Viral Infection

step3.1 で $t1$ が選ばれたとする . step3.2 で一般化を行う対象のタプルを探し出す . step3.2 では, 多様性の数値の求め方は, l -多様性判定に使用している手法によって若干違いがある . エントロピー l -多様性の場合, 4.1 節で示した式を使用して l の値を算出し, その値をそのまま比較に使うことができる . 再帰的 (c, l)-多様性 (Positive Disclosure-再帰的 (c, l)-多様性及び Negative/Positive Disclosure-再帰的 ($c1, c2, l$)-多様性も含む) を使用した場合は 4.2 節で示した不等式の左辺と右辺の差を取ることで判定する . ここでは $t7$ が, 仮に $t1$ とともに一般化したとき構成される q^* -ブロックの多様性が向上し, かつデータ歪曲度 DIS の値が低いので $t1$ との一般化条件にあてはまる . よって $t7$ を対象とし, step3.3 で実際に一般化を行い, $d-list$ を更新する . $t1, t7$ の多様性が 2 となる . それ以外のタプルはデータ操作を行っていないので変更はされない .

step3.4 の条件では, 今作成した q^* -ブロックは 2-多様性なので, まだ l の条件を満たしていない . よって再度, step3.2 に戻り, また対象となるタプルをテーブルから探し出す . 次の対象となるのは $t2$ となる . このタプルと実際に一般化を行うと q^* -ブロックは 3-多様性となり条件を満たしたこととなり, step1 に戻る . この q^* -ブロックに関しては l -多様性を満たしているが, テーブル中にはまだ, l -多様性を満たしていないタプルが存在するので再度 step3 を行うことになる .

同様の手順により, 次に l -多様性を満たし完成する q^* -ブロックは $t3, t4, t5$ で構成される結果となる .

再度 step1 に戻ってくる . この時のテーブル状態は表 4(a) である . $t6, t8$ がまだ l -多様性を満たしていないことがわかる . 同じように step3 で $t6$ は $t3, t4, t5$ のタプルグループと, $t8$ は $t1, t2, t7$ のタプルグループと一般化を行うことになる . このように, 条件次第では, 一度 l -多様性を満たした q^* -ブロックと一般化を行う場合もありうる .

よって以上ですべてのタプルが l -多様性を保持している (つまり, そのタプルを含む q^* -ブロックも l -多様性を満たしているといえる) ので, 完成した結果テーブルを出力する . 結果テーブルは表 4(b) のようになる .

表 4 *DiverDIS* 実行例
(a) 途中結果

	ZIP	Nationality	Condition
t1	130**	Person	Heart Disease
t2	130**	Person	Canser
t3	1305*	Person	Heart Disease
t4	1305*	Person	Viral Infection
t5	1305*	Person	Wheezing
t6	13052	Rossian	Cancer
t7	130**	Person	Wheezing
t8	13063	Japanese	Viral Infection

(b) 出力結果

	ZIP	Nationality	Condition
t1	130**	Person	Heart Disease
t2	130**	Person	Canser
t3	1305*	Person	Heart Disease
t4	1305*	Person	Viral Infection
t5	1305*	Person	Wheezing
t6	1305*	Person	Cancer
t7	130**	Person	Wheezing
t8	130**	Person	Viral Infection

表 5 *MinDIS* が出力した結果テーブル

	ZIP	Nationality	Condition
t1	130**	American	Heart Disease
t2	130**	Person	Canser
t3	130**	American	Heart Disease
t4	130**	Person	Viral Infection
t5	130**	Person	Wheezing
t6	1305*	Person	Cancer
t7	130**	American	Wheezing
t8	130**	Person	Viral Infection

この提案アルゴリズム *DiverDIS* では k -匿名性では防ぎきれなかった新種の攻撃を l -多様性を取り入れることにより防ぐことが可能な結果テーブルを出力することがわかった。

また、従来手法 *MinDIS* のもう一つの特徴であったデータ歪曲算出関数 *DIS* を提案アルゴリズムに取り入れることにより、一般化を行った際に発生するデータ歪曲度を数値化し比較に使用することが可能となったので、出力結果テーブルのデータ歪曲度が小さくなるような一般化の選び方を行える。

一方、 l -多様性は k -匿名性と比べ新たな制限を含んでいるので、どうしても結果テーブルのデータ歪曲度に関しては劣るであろうと考えられる。実際に *MinDIS* に表 3 を入力したときの結果テーブルを表 5 に示す。このとき、*MinDIS* の結果テーブルにおけるデータ歪曲度は 0.356 に対して *DiverDIS* によって得られた表 4(b) の結果テーブルのデータ歪曲度は 0.400 となった。どちらもアルゴリズム中にランダムにタプルを選ぶステップを含んでいるので、一つの入力テーブルにおける結果テーブルは一通りではない。しかし、*DiverDIS* の出力しうる全ての結果テーブルは *MinDIS* の結果テーブルよりデータ歪曲度に対して劣るであろうと予想される。

7. 実験

提案アルゴリズム *DiverDIS* と従来アルゴリズム *MinDIS* を計算機上 (Pentium 4 CPU 3.80GHz, メモリーサイズ: 2 G byte) に C++ 言語で実装し、シミュレーション実験により出力される結果テーブルのデータ歪曲度 *DIS* について比較を行った。また、提案アルゴリズム *DiverDIS* と従来アルゴリズム *MinDIS* において取り入れる値一般化階層、属性一般化階層は同一の階層を使用する。

提案アルゴリズム *DiverDIS* に取り入れた l -多様性の判定にはエントロピー l -多様性と再帰的 l -多様性を使用した。また、再帰的 l -多様性における定数 c の値は本来、データベース所持者が自由に設定する物である。よって、今回は定数 c の値を 1.0 に固定して実装した。Positive Disclosure-再帰的 (c_1, l)-多様性、Negative/Positive Disclosure-再帰的 (c_1, c_2, l)-多様性についても同様に、Positive Disclosure と Negative Disclosure についての設定は基本的にデータベース所有者が決定する物であり、各種定数を定めることで条件の度合いを調整するものである。したがって、今回の実験においてはパラメータ設定に依存して客観的な考察が困難になるので、実装の候補から外した。

7.1 人工データによる実験

入力データとして独自に作成したランダムプロフィールデータ 5 種類 (*data1, data2, data3, data4, data5*) を使用した場合

の各種入力データテーブルにおける結果テーブルのデータ歪曲度 *DIS* を結果を表 6 にまとめた。

data1, data2, data3, data4 のデータテーブルの各種属性集合の領域は等しい設定にしているが、*data5* のデータテーブルはより広い領域設定としている。また、提案アルゴリズム *DiverDIS* と従来アルゴリズム *MinDIS* は共にアルゴリズム中にランダムにタプル選択を行うステップが含まれているので、毎回結果テーブルが異なる。したがって *data1, data2, data3* においては実行回数を 10 回とし、その結果の最小データ歪曲度と最大歪曲度を表 6 に載せた。*data4, data5* においては実行時間の都合上、1 回試行した結果のみを記す。また、表中の *min-tuplegroup* は、結果テーブル中のタプルグループの最小サイズを示している。すなわち、この値は k -匿名性の k の値と同じ意味を持つ。

7.2 ベンチマークデータによる実験

実データ (ベンチマークデータ) を入力したときに提案アルゴリズム *DiverDIS* と従来アルゴリズム *MinDIS* を用いて l -多様性及び k -匿名性を保持させる一般化を行い、それらの結果のデータ歪曲度を算出した。シミュレーション実験結果を表 7 に示す。データは *University of California, Irvine* の *KDD(Knowledge Discovery in Databases)* アーカイブ (<http://kdd.ics.uci.edu>) からの *coil2000(ticdata2000.txt)* の保険会社のデータ (*data6, data7*) と *Japanese Vowels(ae.test)* データ (*data8*) と *IPMUS* 国勢調査 (*ipmus9.la.97*) データからタプル数 10000、属性数 10 を抜き出したデータ (*data9*) を使用した。また、各入力テーブルにおける注目属性として、*data6* は属性 *MOSTYPE* (顧客サブタイプ)、*data7* は属性 *MINKGEM* (平均所得)、*data8* は第 12 番目の属性 (LPC 系列)、*data9* は属性 *nmothers* と設定した。

また人工データと同様に *data6, data7, data8* においては実行回数を 10 回とし、その最小データ歪曲度と最大データ歪曲度を載せた。*data9* においても同様に実行時間の都合上、1 回試行した結果を載せた。

7.3 考察

表 6 と表 7 の結果から、どのデータにおいても結果テーブルの最小データ歪曲度と最大データ歪曲度の差が比較的少ないといえる。これは、従来アルゴリズム *MinDIS* と同様にデータ歪曲算出関数 *DIS* を導入し、タプルに注目した一般化を行っていることから、結果テーブルは極小一般化である結果を導出しているからだといえる。

表 6 の人工データにおいて使用したデータは基本的に、今回提案したアルゴリズムをはじめとする、マイクロデータにおけ

表 6 人工データにおけるデータ歪曲度の比較

data name	tuple	NS-att	l	DiverDIS(Entropy)			DiverDIS(Recursive)			k	MinDIS	
				DIS		min-tuple group	DIS		min-tuple group		DIS	
				min	max		min	max			min	max
data1	100	6	2	0.256	0.287	2	0.289	0.295	2	2	0.193	0.198
			4	0.519	0.551	4~9	0.576	0.659	4~6	4	0.411	0.427
			8							8	0.559	0.604
data2	200	6	2	0.180	0.205	2	0.243	0.285	2	2	0.045	0.053
			4	0.487	0.554	4~9	0.685	0.720	4~7	4	0.214	0.236
			8							8	0.222	0.239
data3	1000	6	2	0.039	0.052	2	0.060	0.070	2	2	0.029	0.031
			4	0.125	0.130	4~9	0.290	0.300	4~9	4	0.047	0.049
			8							8	0.136	0.151
data4	10000	6	4	0.120		4	0.224		4	8	0.089	
			8	0.289		9	0.320		9	10	0.172	
			10	0.304		10	0.361		10	20	0.390	
data5	10000	6	4	0.413		4	0.466		4	8	0.416	
			8	0.468		12	0.483		10	10	0.452	
			10	0.484		12	0.472		12	20	0.469	

る知的情報の高度解析とプライバシー保護の両立を目指したアルゴリズムにおいて使用されると考えられるデータテーブルの構造を組み込んで人工的に作った。また注目属性の設定も適したと思われる属性を選んだ。よってデータテーブル中の値もある程度予測可能な領域内で分布している。したがって、結果テーブルのデータ歪曲度は比較的低い事がわかる。比較として data5 は他の人工データより、データテーブル中の属性値の分布領域を拡大させたデータとした。この data5 と同じデータテーブルサイズの data4 の結果を見ると、やはり分布領域の狭い data4 の結果テーブルのデータ歪曲度の方が低い値を示していることがわかる。また表 7 の実データにおける結果は人工データに比べてデータ歪曲度が高い傾向にある。原因としてあげられるのは、必ずしも理想的な注目属性の設定を行っているわけではないという点と、今回実装するにあたって、アルゴリズム中に導入した値一般化階層、属性一般化階層を、最下位の値（数字データなら一桁目から）からデータ抑制を行うという最もシンプルな階層として設定した点が挙げられる。この問題点は、入力データテーブルによって適切な注目属性、階層に指定することで解消されると考えられる。

またエントロピー l -多様性を取り入れた *DiverDIS* の結果テーブルのデータ歪曲度の方が再帰的 l -多様性を取り入れた *DiverDIS* より優れていることがわかる。これは再帰的 l -多様性を取り入れた *DiverDIS* において定数 c の値を 1.0 に固定したことが原因の一つであろうと考えている。再帰的 l -多様性及び Positive Disclosure-再帰的 (c_1, l) -多様性、Negative/Positive Disclosure-再帰的 (c_1, c_2, l) -多様性はエントロピー l -多様性と違い、各種定数を定めることで l -多様性の条件の高低をコントロールできる。故に入力データテーブルにおいて許容できる範囲で条件を緩和させる定数を定めることである程度データ歪曲度が下げられるのではないかと考えている。

実験結果を総合すると、必ずしも l -多様性の l の値と k -匿名性の k の値は同値ではないことがわかる。表 6,7 中の *min-tuplegroup* は結果テーブル中の最も出現頻度の低いタプルグ

ループのサイズを示している。つまり k -匿名性の k の値にあたる。これを見ると $k > l$ なる k -匿名性を保持していることがわかる。基本的に l -多様性を保持しているテーブルは最低 l -匿名性を保持しているといえる。故に、 l -多様性を保持させるということは一般的に l -匿名性を保持させることよりもデータを歪曲させてしまう傾向にある。特に data6, data7, data8 においてはその傾向が顕著に現れている。data6 で注目属性に選んだ属性は、テーブル中で最も多様性が豊富であると思われる属性を選んだ。しかし、データ歪曲度が *MinDIS* に比べて高いのは、data6 のデータテーブルの非注目属性と注目属性の間の相関関係が強かったり、テーブルサイズに対する注目属性全体の多様性が低いということが原因ではないかと考えている。

しかし、表 7 の data8 の結果においては l -多様性を保持させた結果テーブルのほうがデータ歪曲度が低い結果が得られた。提案アルゴリズム *DiverDIS* と従来アルゴリズム *MinDIS* 共に言えることなのだが、結果テーブルは必ずしも最小一般化ではなくあくまで極小一般化である結果の導出が保証されているだけである。よって、この場合では k -匿名性を目指した従来アルゴリズム *MinDIS* では到達出来なかった最小一般化により近い結果テーブルを提案アルゴリズム *DiverDIS* が導出出来たと言うことになる。ほとんどの場合、提案アルゴリズム *DiverDIS* は従来アルゴリズム *MinDIS* より高いデータ歪曲度を持つ結果テーブルを導出するが、稀に従来アルゴリズム *MinDIS* より低いデータ歪曲度を持つ結果テーブルを導出する可能性があるといえる。

また *MinDIS* において匿名性を強めていった場合、どの程度の l -多様性を満たしているかを表 8 に示した。表 8 からわかるように、従来アルゴリズム *MinDIS* においてはかなり高い匿名性を持たずことによっても 2-多様性を持つか持たないか程度の多様性しか保証されないことがわかる。この結果から、データテーブルに l -多様性を保持させることは難しい課題だということがわかる。

表 7 実データにおけるデータ歪曲度の比較

data name	tuple	NS-att	l	DiverDIS(Entropy)			DiverDIS(Recursive)			k	MinDIS	
				DIS min	DIS max	min-tuple group	DIS min	DIS max	min-tuple group		DIS min	DIS max
data6	5822	85	2	0.563	0.568	2	0.494	0.502	2	2	0.079	0.083
			4	0.819	0.831	4~24	0.769	0.791	4~14	4	0.135	0.136
			8							8	0.136	0.138
data7	5822	85	2	0.631	0.647	2	0.620	0.637	2	2	0.081	0.085
			4	0.829	0.866	4~18	0.799	0.823	4~12	4	0.136	0.138
			8							8	0.137	0.140
data8	5687	11	2	0.589	0.597	2	0.592	0.609	2	2	0.599	0.604
			4	0.718	0.724	4~6	0.718	0.725	4~6	4	0.764	0.775
			8	0.773	0.780	8~12	0.772	0.779	8~12	8	0.782	0.789
data9	10000	10	2	0.405		2	0.454		2	2	0.003	
			4	0.460		4	0.492		4	4	0.024	
			8	0.590		8	0.611		8	8	0.044	

表 8 *MinDIS* による結果テーブルの多様性

Inputted k	diversity (Entropy)		
	data1	data3	data6
2	1.0	1.0	1.0
10	1.0	1.0	1.0
20	1.74	1.0	1.0
50	2.82	1.89	1.42

一般化によるプライバシー保護”, DEWS2007, 2007.

- [5] 村本俊祐, 上土井陽子, 若林真一, “データを極小歪曲し k-匿名性を保持したデータに変換するプライバシー保護アルゴリズム”, DBSJ Letters Vol.6, No1, pp.97-100, 2007.
- [6] L. Sweeney, “Guaranteeing anonymity when sharing medical data, the Datafly system,” Journal of the American Medical Informatics Association, pp.1-5, 1997.
- [7] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp.571-588, 2002.

8. おわりに

本稿では、データ組み合わせによるデータ推測を防ぐために取り入れた k -匿名性の弱点を、同種攻撃及び背景知識攻撃という新種の攻撃を通して考察した。そして、弱点を補うために l -多様性という性質を取り入れることが我々の従来のアルゴリズム *MinDIS* の改善に繋がるということが判明した。また提案アルゴリズムは従来の k -匿名性を取り入れたアルゴリズムに比べ、データ歪曲度という評価指標に基づいてデータ歪曲度が極小な結果を出力するアルゴリズムであり、その長所は l -多様性を取り入れた提案アルゴリズム *DiverDIS* においても受け継がれる。

また実験結果より、提案アルゴリズム *DiverDIS* の結果は一般化ステップについては更なる改善の余地があると考えられるので、 l -多様性を保持させるのにより適切かつデータ歪曲度が最小に近づくような手法についてさらに研究を進め、考察する必要がある。

文 献

- [1] A. Hundepool, L. Willenborg, “ARGUS for protecting microdata and tables,” Seminar on New Techniques & Technologies for Statistics, 1998.
- [2] A. Machanavajjhala, J. Gehrke, D. Kifer. “l-Diversity: Privacy beyond k-anonymity,” Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006), 2006.
- [3] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkatasubramanian, “l-Diversity: Privacy beyond k-anonymity.” ACM Transaction on Knowledge Discovery from Data, Vol.1, No.1, Article 3, Publication data, pp.1-52, 2007.
- [4] 村本俊祐, 上土井陽子, 若林真一, “k-匿名性を利用したデータ