複数のデータ提供のための匿名化

竹之内 隆夫† 古川 諒† 宮川 伸也†

†日本電気株式会社 クラウドシステム研究所
211-8666 神奈川県川崎市中原区下沼部 1753
takenouchi@bu.jp.nec.com, r-furukawa@cb.jp.nec.com, s-miyakawa@ce.jp.nec.com

あらまし 属性数が多いパーソナルデータの分析では，一部の属性のみ使われることが多い．また，様々な分析において，それぞれ異なる属性が使われる．パーソナルデータを外部に提供して分析する際にプライバシを保護するためには，個人特定を防ぐようにデータを加工する k-匿名化や ℓ-多様化のような匿名化が有用である．しかし，分析毎に異なる属性を匿名化すると，それぞれの匿名化結果の突合によって個人特定の危険性がある．そこで本論文では，そのような個人特定を防ぎつつ，分析精度の悪化を抑制することができる匿名化手法を提案する．そして，提案手法を評価し，既存手法よりも分析精度が向上することを示す．

An Anonymization Method for Multiple Releases

Takao Takenouchi† Ryo Furukawa† Shinya Miyakawa†

†Cloud System Research Laboratories, NEC Corporation
1753 Shimonumabe Nakahara-ku, Kawasaki, Kanagawa 211-8666, JAPAN
takenouchi@bu.jp.nec.com, r-furukawa@cb.jp.nec.com, s-miyakawa@ce.jp.nec.com

Abstract To analyze personal information which contains a lot of attributes, analysts select subsets of the attributes according to the purposes of the analyses. When data holders release personal information to analysts, the data holders should anonymize the information to protect privacy. However, if subsets of the information are anonymized separately, there is a risk that the analysts can identify a person by combining the anonymized subsets. We therefore propose a new anonymization method which prevents the identification. Our evaluation results show that the data utility of our method is higher than that of existing methods.

1 はじめに

近年，いくつかのサービス事業者は年齢や購買情報などユーザに関する様々なデータ（パーソナルデータ）を収集している．これら収集された大量のデータは，ビッグデータと呼ばれている．そして，ビッグデータを分析することで有用な知見を得て，より良いサービスの提供などに役立てることが求められている．ビッグデータは多くの属性から構成される場合があるが，その場合の分析では一部の属性のみが使われることが多い．そこで本論文では，以下のような分析のユースケースを想定する．まず分析者は，いくつかの仮説を立てる．そして，その仮説を検証・分析するために必要な属性を選択し，選択した属性を用いて分析を行う．つまり，全属性を含んだパーソナルデータのテーブルのうち，一部の属性を含んだテーブルをビューとして出力して，そのビューを用いて分析するというユースケースである．

一方，ある事業者が収集したパーソナルデータを広く他の事業者へ提供し，提供先の事業者によって様々な分析が行われることが期待されている．パーソナルデータはプライバシにかかわるため，提供するパーソナルデータを匿名化し，k-匿名性 [1] や ℓ-多様性 [2] を満たすように加工する処理が注目されている．

k-匿名性とは，個人を k 人よりも少ない人数に特定できないことを意味するプライバシ指標であり，ℓ-多様性とは k-匿名性を拡張した指標である．k-匿名性を提案した Sweeney や Samarati の研究 [1, 3] では，複数の属性の値の組み合わせによって個人を識別できる恐れがある属性の集合を準識別子 (Quasi Identifier, QI)，個人の知られたくない属性をセンシティブ属性 (Sensitive Attribute, SA) と呼んでいる．ℓ-多様性を提案している Machanavajjhala らの研究 [2] では，1 名の個人のデータが 1 レコードで表現されているテーブルにおいて，QI によって識別できるレコード群の SA の値が ℓ 種類以上であるとき，そのテーブルは ℓ-多様性を満たすと定義している．ℓ-多様性を満たすようにテーブルを加工し，匿名化することを ℓ-多様化と呼ぶ．匿名化の既存手法としては，QI の値を曖昧な値に汎化する方法などが知られている [4]．テーブルを ℓ-多様化することで，たとえ分析者がある個人の QI の属性値を知っていたとしても，その個人の SA を特定することが出来なくなる．

単一のビューを提供する際には，そのビューを ℓ-多様化することでプライバシを保護できるが，本論文が想定するユースケースでは，ある分析者が複数の分析を行うために複数のビューを得ることや，複数の分析者が結託し，それぞれが得たビューを突合することが考えられる．このような場合，複数のビューが突合され，SA の値が推測されてしまう可能性がある．本論文では，ビューの突合によって推測される SA の値が ℓ 種類以上であることを「複数ビュー ℓ-多様性」を満たすと表現する．Ganta らは複数ビュー ℓ-多様性の指標を提案し，指標を確認するアルゴリズムを示した [5].

しかし，既存手法を用いて複数ビュー ℓ-多様性を満たす匿名化手法では，一部のビューの属性値が他のビューに比べて，より曖昧な値に汎化されてしまう問題がある．その結果，そのビューを用いた分析の精度が著しく悪化する恐れがある．この問題は，属性値の汎化の度合いがビュー間で大きく偏ることにより発生する．

そこで本論文では，複数ビュー ℓ-多様性を満たしつつ，属性値の汎化の度合いがビュー間で偏らないような匿名化手法を提案する．さらに，提案手法を評価し，既存手法よりも分析精度が向上することを示す．本論文は以下のような構成になっている．まず，2 章にて関連研究を示す．続いて 3 章にて，本論文が対象としている問題について説明する．そして，4 章でこの課題を解決する手法を提案し，5 章で評価を行う．最後に 6 章で本論文をまとめる．

2 関連研究

複数のテーブルやビューを提供する際の匿名化の研究は，いくつか存在する．Ganta らの研究 [5] では，複数の機関が同一のユーザのデータを保持する場合に，各機関が独立して ℓ-多様化し，データを公開することを想定している．そして，その際の複数ビュー ℓ-多様性の確認方法を提案している．Ganta らは，異なる機関の連携は困難であるため，この問題の解決は困難であるとし，データをランダム化する Differential Privacy を拡張した指標を提案している．

それに対し本論文では，Ganta らの想定と異なり，単一機関が保持するデータに対して複数のビューを公開する場合を想定している．そして，複数ビュー ℓ-多様性を満たすように値を汎化させる手法を提案している．

Yao らの研究 [6] でも複数ビュー ℓ-多様性の指標が提案されている．この研究では属性間の関数従属性も考慮した指標の確認アルゴリズムが提案されているが，匿名化手法は提案されていない．

Jin らの研究 [7] では，1 つのテーブルを複数のビューにわけて匿名化する際に，データの有用性が維持されるようなビューの形式を決定する方法が提案されている．それに対し本論文では，ビューの形式は分析者によって予め決定さ

れている前提とし，その形式においてビューを匿名化する手法を提案している．

以上に挙げた研究のような複数のビューを一度に匿名化し提供する手法と異なり，ビューを順番に提供する匿名化手法の研究も存在する [8]．また，複数のビューを匿名化するのではなく，複数の集計表 (marginal) を提供するような研究も存在する [9]．

3 課題定義

3.1 複数ビュー ℓ-多様性

本論文では，元テーブル T から一部の属性を抜き出したビュー $T_i(i = \{1, 2, \cdots, n\})$ を生成し，それらのビュー T_i を匿名化するという前提である．T は $\{ID, A_1, \cdots, A_m, SA\}$ という属性を持つとする．ここで，ID はユーザの識別子，$\{A_1, \cdots, A_m\}$ は準識別子であり，QI と表記する ($QI = \{A_1, \cdots, A_m\}$)．$QI$ は 1 つ以上の数値属性の集合とする．SA はセンシティブ属性であり，カテゴリ値とする．また，1 名の個人のデータは 1 レコードで表現されるとする．

そして T_i は，QI の部分集合となる QI_i ($QI_i \subset QI$) と SA が含まれる形式とする．つまり，T_i の属性は $\{QI_i, SA\}$ である．T_i を分析者に公開する際には，匿名化されることとし，匿名化した T_i を T_i^* と表現する．

表 1 に元テーブル (T)，表 2 と表 3 に匿名化されたビュー T_1^* と T_2^* の例を示す．{ 年齢, 身長 } が QI，病気が SA である．ここで，分析者が年齢と病気の相関と，身長と病気の相関を検証する場合を考える．この場合，年齢と病気を抜き出したビュー T_1 と身長と病気を抜き出したビュー T_2 を生成し，それぞれを匿名化した T_1^* と T_2^* を提供することになる．表 2 と表 3 は，それぞれ T_1 と T_2 を個別に 2-多様性を満たすように匿名化を行った T_1^* と T_2^* の例である．

ℓ-多様化を行うことで個人の SA の値が特定されることを防げるが，個人の SA の値が特定するために，各ビューを突合する Intersection Attack という攻撃が存在する [5]．この攻撃は，次のような手順である．まず，匿名化された複数のビューを受け取った分析者 (攻撃者) は，あるユーザ u の QI の値 q_u を知っているとする．そして攻撃者は，T_i^* について q_u に該当するレコードを複数抜き出す．これらのレコードの SA の値を $I(T_i^*, u)$ とする．つまり $I(T_i^*, u)$ は，攻撃者が T_i^* から推測できるユーザ u の SA の値の集合である．そして，$\{T_1^*, \cdots, T_n^*\}$ の各ビューから推測した SA の値の集合の積集合 ($I(T_1^*, u) \cap \cdots \cap I(T_n^*, u)$) を得る．この積集合が，Intersection Attack で推測されるユーザ u の SA の値の集合となる．

例えば，表 2 の T_1^* と表 3 の T_2^* を受け取った分析者 (攻撃者) が，user4 の年齢と身長を知っているとする．すると，この攻撃者は表 2 を参照すると，user4 の年齢 23 才に該当するレコードは「23-24」に汎化された 2 レコードであることから，user4 の病気は「HIV」か「肺炎」のいずれかであることがわかる．同様に，この攻撃者は表 3 を参照すると，身長「160」に該当する 2 レコードを見ることで user4 の病気は「HIV」か「かぜ」のいずれかであることがわかる．これら 2 つのことから，この攻撃者は user4 の病気は「HIV」であることが特定できてしまう．

本論文では，あるテーブルの各個人 (レコード) について，Intersection Attack で推測できる SA の値の集合の要素数が ℓ 個以上であるとき，そのテーブルは「複数ビュー ℓ-多様性」を満たすと表現する．

3.2 複数ビュー ℓ-多様化の課題

T_1 と T_2 を複数ビュー 2-多様性を満たすように匿名化する方法として，T_1 を 2-多様性を満たすように匿名化し，その後 T_2 を複数ビュー 2-多様性を満たすように匿名化する方法が考えられる．しかし，この方法では後に行う T_2 の属性値が汎化され過ぎてしまう場合がある．例えば，T が表 1 であるとき，T_1^* を表 2 のように 2-多様性を満たすように匿名化した後に，T_2^* を複数ビュー 2-多様性を満たすように匿名化する場合を考える．この場合，T_2^* の全レコードの「身長」を「160-189」と汎化することになってしまい，属性値が曖昧な値になり過ぎ，T_2^* を用いた分析結果の精度が悪くなってしまう．

表 1: 元データ (T)

ID	年齢	身長	病気
user1	20	180	かぜ
user2	21	180	肺炎
user3	22	175	かぜ
user4	23	160	HIV
user5	24	185	肺炎
user6	25	170	HIV
user7	26	165	かぜ

表 2: T_1^*

	年齢	病気
user1	20-22	かぜ
user2	20-22	肺炎
user3	20-22	かぜ
user4	23-24	HIV
user5	23-24	肺炎
user6	25-26	HIV
user7	25-26	かぜ

表 3: T_2^*

	身長	病気
user4	160-169	HIV
user7	160-169	かぜ
user6	170-179	HIV
user3	170-179	かぜ
user1	180-189	かぜ
user2	180-189	肺炎
user5	180-189	肺炎

表 4: 複数ビュー 2-多様性を満たす T_1^*

	年齢	病気
user1	20-21	かぜ
user2	20-21	肺炎
user3	22-23	かぜ
user4	22-23	HIV
user5	24-26	肺炎
user6	24-26	HIV
user7	24-26	かぜ

このように，T_1^* と比較して T_2^* の情報精度が悪化し，情報精度の偏りが大きくなってしまう．これは，T_1^* を匿名化する際に T_2^* を匿名化することが考慮されていないからである．もし，T_1^* と T_2^* の双方を考慮して匿名化することが出来れば，T_1^* と T_2^* を表 4 と表 3 のように，複数ビュー 2-多様性を満たしつつ情報精度の偏りを小さくできる．

以下に本論文が扱う問題について，問題の前提とプライバシ要件，さらに提案方式が目指すべき目標を整理する．

前提 各 T_i の形式は事前に与えられているとする

要件 $\{T_1^*, \cdots, T_n^*\}$ は複数ビュー ℓ-多様性を満たすこと

目標 各 T_i^* の情報精度の偏りを小さくしつつ，可能な限り詳細な情報を出力する

4 提案手法

複数ビューの ℓ-多様化を行うために，既存手法の Mondrian アルゴリズム [4] を拡張する．Mondrian は，Top-down アプローチと呼ばれる匿名化方式であり，比較的データ精度が良く処理効率が良い匿名化アルゴリズムとして知られている．Top-down アプローチとは，最初に各 QI の属性値をもっとも曖昧な値に汎化し，徐々に詳細化する手法である．ここで詳細化とは，QI の属性値で識別されるユーザ集合 (等価クラス, Equivalence class) を，ある境目で分割することである．この分割の境目となる属性値を分割点と呼ぶ．例えば,「20 才」という分割点で分割すると,「20 才以上」と「20 才未満」に分割することになる．分割点は，分割点決定関数によって選ばれる．Mondrian[4] では値域 (range) が大きい属性の中央値 (median) が選ばれるというアルゴリズムになっている．

提案手法では，Mondrian を 2 つの観点で拡張している．1 つ目は，複数ビューを同時に匿名化するという拡張である．既存の Mondrian は単一のテーブルを匿名化するためのアルゴリズムであるため，複数のビューを匿名化する場合は個々のテーブルを個別に匿名化する処理になる．その結果，複数ビューを匿名化する際にビュー間で情報精度の偏りが出てしまう．本拡張により，各ビューの情報精度が偏らない匿名化を実現できる．

2 つ目の拡張は，分割点決定関数を複数ビュー ℓ-多様性を満たしやすいような分割点が選ばれるようにする拡張である．この拡張により，既存の Mondrian が採用している分割点決定関数よりも多くの分割を行うことができ，より詳細な情報を出力することが可能となる．以降に，これらの拡張について詳しく説明する．

4.1 複数ビューの同時匿名化

図 1 に，複数ビューを同時に匿名化する提案手法のアルゴリズムを示す．このアルゴリズムでは，他のビューを考慮して分割点を決定し，各ビューが同時に徐々に分割を行っていく点が特徴である．これにより，各ビュー間で分割回数が偏ることを防止でき，各ビュー間での情報精度の偏りを抑えた匿名化が実現できる．

このアルゴリズムでは，まず，ビュー T_i ($i \in$

```
function anonymizeViews({T_1, ..., T_n})
1:  Views ⇐ T_i の QI_i を最も汎化し {T*_1, ..., T*_n} を作成
2:  Points ⇐ chooseDivisionPointsOfViews(Views)
3:  while (Points ≠ φ)
4:    for each point ∈ Points
5:      分割対象の view ⇐ point.view
6:      view を point で分割
7:      if Views が複数ビュー ℓ-多様性を満たさない
8:        view の point での分割をキャンセル
9:        point を次回以降の分割点候補から除く
10:   Points ⇐ chooseDivisionPointInViews(Views)
11: return Views
```

図 1: 提案手法のアルゴリズム

$\{1, \cdots, n\}$) の QI_i を最も汎化し，初期化された T_i^* を生成する (図 1 の 1 行目)．続いて，4.2 節で説明する分割点決定関数を用いて，各ビュー T_i^* の分割点を決定する (2 行目)．この分割点は，各ビューにおいて 1 つずつ選ばれる．そして，決定した分割点で各ビューを分割する (3〜6 行目)．ここで，ビューを分割するとは，ビューの中の 1 つの等価クラスを，ある分割点で分割することを意味する．ここで，分割後に匿名性を確認し，もし匿名性を満たさない場合は分割をキャンセルする (7〜8 行目)．分割がキャンセルされた場合は，分割点を記録し，次回以降の分割において，この分割点が選ばれないようにする (9 行目)．そして，分割可能な分割点が無くなるまで処理を行い，最後に分割済みの $\{T_1^*, \cdots, T_n^*\}$ を出力する (10〜11 行目)．

4.2 分割点決定関数

続いて，複数ビュー ℓ-多様性を満たしやすいように拡張した分割点決定関数について説明する．Intersection Attack は，$\{T_1^*, \cdots, T_n^*\}$ から推測される SA の値の集合の積が小さくなることで発生する．よって，SA の値の集合の積が小さくならない分割点を選べば，より多く分割することが出来ると考える．

そこで，本論文では SA の値の集合の類似度という概念を導入する．SA の値の集合の類似度は，分割後のビューにおいて，各ユーザについて推測される SA の値の集合が，各ビュー間でどのくらい似ているかを示す指標である．提案手法では，ある分割点候補 c で分割した際の SA の値の集合の類似度 S を以下のように定義した．

$$S(c) = -\sum_{u \in Users} Dist(c, u)^2 \quad (1)$$

ここで，$Users$ は全ユーザである．$Dist$ はユーザ u について，分割点候補 c で分割後のビュー T_i^* から推測できる SA の値を $I(T_i^*, u)$[1] としたとき，全ビュー T_i^* ($i \in \{1, \cdots, n\}$) の任意の組合せ ($\{T_s^*, T_t^*\}$) の $I(T_s^*, u)$ と $I(T_t^*, u)$ の編集距離の合計である．つまり，類似度 S は $Dist$ を 2 乗し，各ユーザ u について合計した値を負数にした値である．$Dist$ を 2 乗しているのは，距離が離れている場合の類似度を急激に低下させるためである．負数にしているのは，距離が大きいほど類似度を小さくさせるためである．なお，S は全ユーザについて計算するため，少なくとも初回の分割ではユーザ分 (レコード分) の S 計算が必要であり，既存の Mondrian に比べて計算量が多くなる．

そして，類似度 S を利用してさらに全分割点候補 C における，ある分割点候補 c のスコア $Score$ の計算式を以下のように定義した．

$$Score(c) = \frac{\alpha S(c)}{\max_{p \in C} S(p)} - \frac{(1-\alpha) M(c)}{\max_{p \in C} M(p)} \quad (2)$$

ここで M は，分割点候補と中央値からの距離である．負数にしているのは，中央値から遠いほど選ばれにくくするためである．$Score(c)$ は，S と M を全分割点候補の最大値で割ることで 0〜1 に補正し，補正した値を重み $\alpha (0 \leq \alpha \leq 1)$ 付きで合計した値となる．

そして，提案手法の分割点決定関数は，分割可能な分割点候補を列挙し，その中で一番 $Score$ が大きい分割点候補を分割点として選択するように設計した (図 2)．ただし，分割点候補が多くなると計算量が増えすぎてしまう恐れがあるので，各 QI のグループにおける候補数に上限値を設定し，各グループの中央値付近の分割点候補を中心に候補を列挙している．また，$Score$ は前回の分割によって影響がない個所については前回の値をそのまま利用し，再計算は行わない．このような設計により，2 回目以降の分割の計算量を削減することができる．

[1] $I(T_i^*, u)$ の詳細は 3.1 節にて説明している

表 5: 24 才 170cm で分割時の編集距離の計算例

ID	T_1^* の SA 集合	T_2^* の SA 集合	編集距離
user1	{A,A,B,C}	{A,C}	2
user2	{A,A,B,C}	{A,C}	2
user3	{A,A,B,C}	{A,A,B,B,C}	1
user4	{A,A,B,C}	{A,A,B,B,C}	1
user5	{A,B,C}	{A,A,B,B,C}	2
user6	{A,B,C}	{A,A,B,B,C}	2
user7	{A,B,C}	{A,A,B,B,C}	2

(凡例 A:かぜ, B:肺炎, C: HIV)

function chooseDivisionPointInViews($Views$)
1: $Candidates \Leftarrow Views$ の分割点候補を列挙
2: **for each** $c \in Candidates$
3: $c.score \Leftarrow c$ のスコア S を計算 (更新)
4: $winner \Leftarrow c.score$ が最大となる c を選択
5: $Points \Leftarrow winner$ の各ビューの分割点
6: **return** $Points$

図 2: 分割点決定関数のアルゴリズム

次に，$S(c)$ と $M(c)$ の計算例を示す．表 5 に，3 章で示した T_1 と T_2 のビューを，T_1^* を年齢 24 才，T_2 を身長 170cm で分割する分割点候補 c について，類似度を計算した例を示す．この表に示したように，分割後の T_1^* について推測できる user1 の SA の値の集合は { かぜ, かぜ, 肺炎, HIV} であり，分割後の T_2^* については { かぜ, HIV} である．この場合の編集距離は 2 となる．つまり，この分割点候補 c での user1 の編集距離 ($Dist(c, user1)$) は 2 である．同様に，この分割点候補 c の全ユーザの編集距離を計算し，$S(c)$ を求めると -22 となる．そして，年齢 24 才の中央値からの距離は 1 で，身長 175cm は 1 である．よって $M(c) = 1 + 1 = 2$ となる．このデータ例では，$\alpha = 0.8$ ではこの分割店候補のスコア値が最も大きくなり，初回の分割点はこの分割点が選ばれる．そして最終的に表 4 と表 3 のように 2 回分割される．

5 評価実験

提案手法と既存手法を Java で実装し，有効性の評価を行った．

5.1 評価データ

評価で用いたデータは，匿名化の研究で頻繁に用いられる UCI Repository[10] の Adult データである．Adult データは米国の国勢調査をもとに作られたデータであり，15 種類の属性を持つ約 3 万レコードのデータである．このデータから，既存の Mondrian の研究 [4] や ℓ-多様化の研究 [2] と同様に，age(年齢), sex(性別), work-class(職種), education(学歴), marital-status (結婚歴), race(人種), country(母国), occupation(職種) の 8 属性を抜き出し，このうち occupation を SA，それ以外を QI として T を作成した．ただし，提案手法では QI は数値である前提を置いているため，これらの属性のうち age と occupation 以外のカテゴリ値は，アルファベット順で数値で置き換えた．

そして，T の属性のうち {age, sex} を基本的な属性，{education, work-class} を学歴に関する情報，{native-country, race, marital-status} を身体に関する情報と考え，学歴と職業の関係や身体と職業の関係を分析するユースケースを想定し，T_1 と T_2 の 2 つのビューを用いて評価を行った．T_1 が持つ属性は，{age, sex, $work\text{-}class$, $education$, $occupation$} であり，T_2 が持つ属性は，{age, sex, $marital\text{-}status$, $race$, $country$, $occupation$} である．

なお，現状の提案手法は数値属性のみの対応であるが，汎化ツリーを用いることで，カテゴリ値にも対応することが可能である．また，各評価値は，Adult データから 200 行をランダムに選択し，30 回計測を行った平均値である．分割点決定関数の重み α は 0.8 と置き，中央値よりも類似度が高い分割点が選ばれるようにしている．また，計算量削減のために各 QI のグループ内の分割点候補数の上限を 6 個と設定した．

5.2 評価指標

評価指標は，Mondrian の研究 [4] と同じく Discernibility Metric（DM）[11] を用いた．DM は匿名化による情報精度の低下を表現した指標であり，小さい値であるほど情報精度が高いこ

とを意味する．DM を用いることで，他の手法と情報精度を比較することが出来る．

DM の値は，ある匿名化されたテーブル T^* について，QI の属性値が同一のレコードを等価クラス (Equivalence Class) と表現した際に，以下に示したように各等価クラスにおけるレコード数の 2 乗した値を合計することで求めることが出来る[2]．

$$DM = \sum_{E \in EquivClass} |E|^2 \qquad (3)$$

ここで $EquivClass$ は，T^* における全等価クラスの集合を意味し，$|E|$ はある等価クラス E のレコード数を意味する．

5.3 比較対象となる既存手法

比較対象とした既存手法は，既存の Mondrian を用いた手法である「全属性 Mondrian」と「順序 Mondrian」である．「全属性 Mondrian」は，複数ビューの ℓ-多様化を満たした T_1^* と T_2^* を生成するために，まず T を Mondrian を用いて ℓ-多様化し，匿名化した T^* を生成する．その後，匿名化した T^* から必要な属性を抜き出し，T_1^* と T_2^* を生成する．このように生成した T_1^* と T_2^* は，もともと同一の T^* から抜き出したビューであるため，T_1^* と T_2^* を突合しても T^* よりも詳しい情報はわからない．つまり，T_1^* と T_2^* は複数ビュー ℓ-多様性を満たすビューとなる．全属性 Mondrian は，最も単純に複数ビューの ℓ-多様化を実現する方法であるため，評価の基準となる手法と考える．

「順序 Mondrian」は T_1^* と T_2^* のうち，片方のビューを先に Mondrian で ℓ-多様性を満たすように匿名化した後，もう片方のビューを複数ビュー ℓ-多様性を満たすように匿名化するという手法である．この方法は，3 章で説明したとおり，後に匿名化したビューはあまり分割できくなり，ビュー間での情報精度の偏りが大きくなる傾向がある．この手法は，どちらのビューから匿名化するかに応じて，情報精度が大きく変わるので，匿名化の順序として「T_1^*，T_2^*」という順序と「T_2^*，T_1^*」という順序の両方で評価を行った．

[2] [11] ではレコード削除（suppression）をした際の DM 値が定義されているが，提案手法ではレコード削除は行わないので無視している．

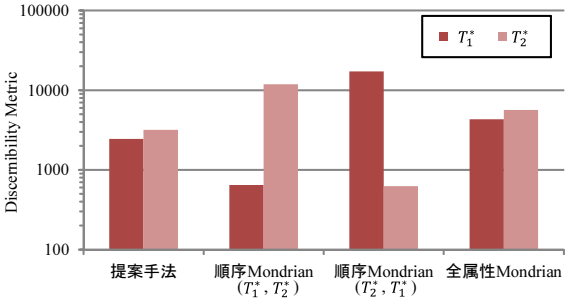

図 3: T_1^* と T_2^* における DM

5.4 評価結果

まず，提案手法と既存手法を用いて，複数ビュー 2-多様性を満たすように匿名化し，結果の T_1^* と T_2^* の DM 値を計測した．結果を図 3 に示す．このグラフからわかるとおり，順序 Mondrian を「T_1^*，T_2^*」という順序で匿名化した場合，T_1^* の DM 値は 700 程度であるが T_2^* の DM 値は約 10000 であり，T_2^* の情報精度が著しく悪くなっている．順序 Mondrian を「T_2^*，T_1^*」という逆の順序で匿名化した場合は，T_2^* の DM 値は 700 程度であるが T_1^* の DM 値は約 15000 となり，T_1^* の情報精度が著しく悪くなっている．このことから，順序 Mondrian は後に匿名化したビューは，先に匿名化したビューと比較して情報精度が悪化することがわかる．

それに対し提案手法は，T_1^* も T_2^* も DM 値が約 2500〜3000 程度であり，順序 Mondrian に比べて T_1^* と T_2^* の DM 値の差は小さい．ただし，提案手法の T_1^* の DM 値は，「T_1^*，T_2^*」という順序で匿名化した順序 Mondrian の T_1^* の DM 値よりも悪化している．同様に，提案手法の T_2^* の DM 値は「T_2^*，T_1^*」という順序で匿名化した順序 Mondrian の T_2^* よりも悪化している．つまり，提案手法は順序 Mondrian のように片方のビューの情報精度を高くするのではなく，T_1^* と T_2^* の両方の情報精度を同じ程度にした匿名化ができることがわかった．

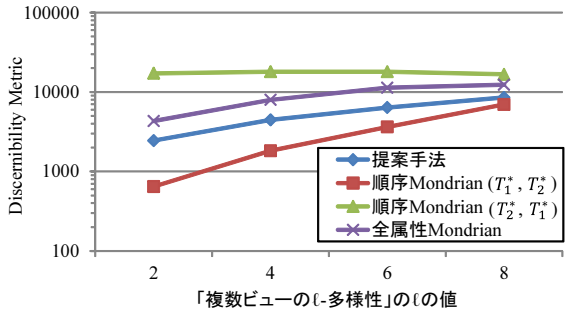

図 4: T_1^* の DM 値

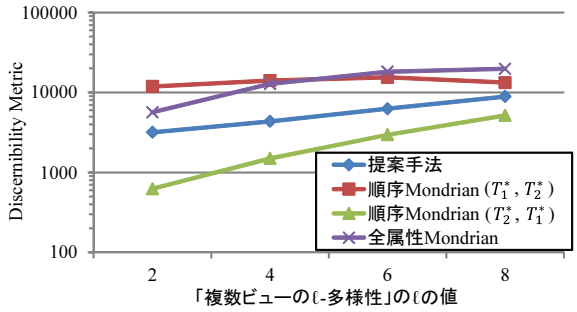

図 5: T_2^* の DM 値

また，全属性 Mondrian については T_1^* と T_2^* とも DM 値は 4000〜6000 であり，提案手法よりも情報精度が悪いことがわかる．これは，提案手法は全属性 Mondrian とは違い，T から T_1^* と T_2^* のビューを生成してから匿名化し，さらに，T_1^* と T_2^* の双方を考慮して分割点を選んでいるためである．これにより複数ビュー ℓ-多様性を満たしつつ多くの分割が可能となり，結果として T_1^* と T_2^* ともある程度の情報精度を保った匿名化が可能となっている．

続いて，複数ビュー ℓ-多様性の ℓ の値を 2〜8 まで変えて DM 値を計測した結果を図 4 と図 5 に示す．図 4 が T_1^* の DM 値，図 5 が T_2^* の DM 値である．これらのグラフから分かるとおり，ℓ が小さいほど順序 Mondrian の T_1^* と T_2^* の DM 値の差が大きくなる．それに対し，提案手法は ℓ がどの値であっても T_1^* と T_2^* の DM 値の差は小さいままである．また，提案手法は全属性 Mondrian よりも DM 値は小さい．

以上の 2 つの評価結果から，提案手法は，既存手法である順序 Mondrian や全属性 Mondrian と比較して，各ビューの情報精度の偏りを小さくしつつ，さらに，各ビューの情報精度を保った匿名化ができることが分かった．

6 まとめ

本論文では，各ビューの情報精度 (属性値の加工の度合い) がビュー間で偏らないように複数のビューを匿名化する手法を提案した．そして，提案手法を評価し，既存手法よりも分析精度が向上することを確認した．

今後は，提案手法の計算量オーダーの算出と実行速度の計測を行い，データ量が増えた場合のスケーラビリティ評価を行う予定である．また，様々なデータを用いて，より詳細な有効性評価を行い，提案手法の改良を行う予定である．

参考文献

[1] L. Sweeney, “k-anonymity: a model for protecting privacy,” Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol.10, pp.557–570, 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “ℓ-diversity: Privacy beyond k-anonymity,” Proc. ICDE’06, p.24.

[3] P. Samarati, “Protecting respondents’ identities in microdata release,” IEEE Trans. on Knowl. and Data Eng., vol.13, pp.1010–1027, 2001.

[4] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” Proc. ICDE’06, p.25.

[5] S.R. Ganta, S.P. Kasiviswanathan, and A. Smith, “Composition attacks and auxiliary information in data privacy,” Proc. KDD’08, pp.265–273.

[6] C. Yao, X.S. Wang, and S. Jajodia, “Checking for k-anonymity violation by views,” Proc. VLDB’05, pp.910–921.

[7] X. Jin, M. Zhang, N. Zhang, and G. Das, “Versatile publishing for privacy preservation,” Proc. KDD’10, pp.353–362.

[8] K. Wang and B.C.M. Fung, “Anonymizing sequential releases,” Proc. KDD’06, pp.414–423.

[9] D. Kifer and J. Gehrke, “Injecting utility into anonymized datasets,” Proc. SIGMOD’06, pp.217–228.

[10] C.L. Blake and C.J. Merz, “Uci repository of machine learning databases.,” 1998. http://archive.ics.uci.edu/ml/

[11] R.J. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” Proc. ICDE’05, pp.217–228.