属性間の相関を考慮した撹乱再構築法の提案

齋藤恆和 †　　五十嵐大 †　　菊池亮 †　　廣田啓一 †　　正木彰伍 †

†NTT セキュアプラットフォーム研究所
180-8585 東京都武蔵野市緑町 3-9-11
saito.tsunekazu@lab.nt.co.jp

あらまし　パーソナルデータの安全な二次利用に向けて, プライバシーを保護しつつ情報を有効活用できるようなデータの加工技術が必要となる. このような技術の一つに撹乱再構築法を用いた Pk-匿名化がある. Pk-匿名化はマーケティング分野等の多少の誤差が許容される分野では十分に利用可能であるが, 更に分析精度を向上させれば適用範囲の拡大が見込める. 分析精度を下げている原因の一つに, 元データに存在しない属性値の組み合わせを扱い Pk-匿名化を行っていることが挙げられる. 本稿では, 属性間の相関を考慮して属性を組合せ, 存在しない属性の組み合わせを排除する Pk-匿名化によって分析精度の向上を目指す.

Proposal of Perturbation-Reconstruction Method with Correlations Between Attributes

Tsunekazu SAITO †　　Dai IKARASHI †　　Ryo KIKUCHI †
Kei-ichi HIROTA †　　Shogo MASAKI †

†NTT Secure Platform Laboratories.
3-9-11, Midori-cho, Musashino-shi, Tokyo 180-8585, JAPAN
saito.tsunekazu@lab.ntt.co.jp

Abstract　From the point of view of the privacy risk for secondary use of the personal data, a method that converts the personal data so that the privacy of each individual in the data is protected is necessary. The method is also required to keep the personal data effectively utilizable. The perturbation-reconstruction method is one of the possible solutions. However, in some cases, this method makes a significant difference between the original and the converted data. One of the reasons is that the converted data could have the combinations of attribute values that do not exit in the original data. In this paper we propose a novel method that considers correlations between attributes to perturb the attribute values. The method prevents the converted data from including the unnatural combinations of attributes values and improves the usability of the converted data.

1　はじめに

近年, 購買データや移動データなどのパーソナルデータを二次利用するニーズが高まっている. これに伴い, データマイニングにおいてプライバシーを保護しながらも分析可能な形にデータを加工する匿名化の研究が注目されている. 具体的な匿名化の手法は k-匿名化や Pk-匿名化等が知られている. k-匿名化とはデータの属性

値の抽象化やレコードの削除によって, ある個人のレコードを k 個以下に絞り込めないようにする手法である. この k-匿名化を確率的に拡張し, ある個人のレコードを $1/k$ 以上の確信度に絞り込めないようにする手法が Pk-匿名化である [1], [2], [3]. この Pk-匿名化は属性の抽象化を伴う k-匿名化と比較して, 等価な匿名性を保証しつつも細かい粒度での分析が行える. Pk-匿名化は属性値の置換やノイズ付与などにより確率的にレコードを変換し匿名化する撹乱の処理と, 撹乱されたデータから元のデータを推定する再構築の処理によって実現される.

Pk-匿名化の撹乱処理は確率的な処理であるために再構築後の推定したデータと元のデータには誤差が生じることが避けられない. これに対して, 確保する匿名性と分析精度の関係性について廣田らによる実験があり, マーケティング分野等の厳密な分析精度が必要ない場合には Pk-匿名化は実用的であることが示されている [4]. 更なる分析精度の向上によって, 高い分析精度を求められる分野でも Pk-匿名化を適用することができるようになる.

そもそも, 撹乱処理による分析精度の低下の原因の 1 つに, 属性を組み合わせる際に単純に組み合わせることで属性値の組み合わせの取りうる範囲が大きくなってしまい, 再構築処理における推定が困難になってしまうことが挙げられる. ゆえに, 属性値の組み合わせの範囲を小さくするような属性の組み合わせ方をした Pk-匿名化を行えば, 分析精度の向上が望める.

本稿では, 属性の相関に注目し属性値の組み合わせの範囲を小さくさせ撹乱する手法を提案する. 例えば, 年齢という属性と自動車の運転免許の有無という属性に対して (10 代前半, 有) といった属性値の組合せはありえない. この相関を考慮した属性の組み合わせを行い実際にはありえない属性値の組み合わせを外すことでその取りうる範囲を制限し, 分析精度の向上を図る.

2 従来の Pk-匿名化

プライバシー保護データ公開技術を実現させる技術の一つである, 撹乱再構築法を用いた Pk-匿名化の概要を [2] と [3] に従って述べる.

2.1 テーブルとクロス集計

保護の対象となるデータは複数の情報提供者から集約されたデータの集合であり, テーブルとして表現される. 各情報提供者からのデータはテーブル上では 1 行に表現され各行をレコードと呼ぶ. そして各レコードはあらかじめ定められた項目に対する値から成り立っており, この項目を属性, 属性の値を属性値といい, 属性値の範囲を属性の値域という. 表 1 にテーブルの例を示す. “性別” および “年代” が属性であり, 各行 “ 女性, 20 代 ” 等がレコードである.

表 1: テーブルの例

ID	性別	年代
1	女性	20 代
2	男性	20 代
3	女性	10 代
⋮	⋮	⋮

クロス集計とはテーブルに対して着目したすべての属性に関して値が等しいようなレコードをカウントしたものである. このクロス集計はアンケート集計で使われるほかに数多くのデータマイニング手法の中で使用される基本的かつ重要な演算である. クロス集計の大きさは各属性の組み合わせの総数つまり各属性の値域の積であり, 要素数の和はテーブルのレコード数と一致している. 下の式が表 1 におけるテーブルでのクロス集計である. ベクトルの第 2 要素である “{ 男性, 20 代 }” はテーブルの中にある男性かつ 20 代である人数の総和であり, 右辺のベクトルの第二要素はそれが 45 人であることを示している.

$$
\begin{aligned}
&(\{\text{男性}, 10\text{代}\}, \{\text{男性}, 20\text{代}\},\\
&\qquad \{\text{女性}, 10\text{代}\}, \{\text{女性}, 20\text{代}\})\\
&\quad = (61, 45, 73, 50).
\end{aligned}
$$

一般的な各属性の値域が $M_1, \ldots, M_n$ (n は自然数) のテーブルについて, 各属性の取りうる範囲を $M_i = \{1, \ldots, m_i\}$ までの値に数値化をしたテーブルからクロス集計を求めるためには次のような操作を行う. まず初期のクロス集計として長さが $m = \prod_{i=1}^{n} m_i$ の 0 ベクトル $(0, \ldots, 0)$ を定める. 次に j を 1 からレコード数までの整数を亙らせて j 番目のレコード

$$(a_{(j-1)n+1}, a_{(j-1)n+2}, \ldots, a_{jn})$$

に対して初期に定めた 0 ベクトルの

$$1 + \sum_{i=1}^{n} \left((a_{(j-1)n+i} - 1) \frac{m}{\prod_{s=1}^{i} m_s} \right)$$

番目の要素について順次カウントすればよい.

2.2 撹乱と再構築

Pk-匿名化は, テーブルに対して行うプライバシー保護のための撹乱と, 撹乱されたテーブルからクロス集計のみの統計量を取り出す再構築からなる.

撹乱では各レコードを確率的に変化させる. 確率的な変化の方法は維持置換撹乱を用いる, すなわち各属性値を一定の確率のもとで維持させ, それ以外であれば属性値から一様ランダムに遷移させる. この確率は維持確率と呼ばれ ρ_i と表記する.

撹乱の操作は属性の値域 M_i に応じて次のようにサイズが $m_i \times m_i$ であるような遷移確率行列 A_i で表現できる.

$$A_i = ((1 - \rho_i)/m_i) U(m_i) + \rho_i E(m_i),$$

ここで $U(m_i)$ と $E(m_i)$ はそれぞれ, サイズが $m_i \times m_i$ の要素が全て 1 の行列と単位行列である. s 行 t 列目の要素 $(A_i)_{st}$ は値が s から t に遷移する確率を意味し具体的に以下のようになる.

$$(A_i)_{st} = \begin{cases} \rho_i + (1 - \rho_i)/m_i, & \text{if } \ s = t, \\ (1 - \rho_i)/m_i, & \text{otherwise.} \end{cases}$$

複数の属性に対して遷移確率行列 $\{A_i\}_{i=1,\ldots,n}$ が与えられている際に, レコードが撹乱後にあるレコードに遷移する確率を遷移確率行列のクロネッカー積 $\bigotimes A_i = A_1 \otimes \cdots \otimes A_n$ で表現できる.

再構築では, 撹乱で用いた遷移行列のクロネッカー積 $A = \bigotimes A_i$ と撹乱後のテーブルから得られるクロス集計 y からベイズ推定法に基づいたもとのテーブルのクロス集計の推定を行う. 具体的なアルゴリズムは [2] 等に述べられている.

2.3 Pk-匿名性と維持確率の関係

従来の k-匿名性はテーブル形式のデータベースの保護処理において, “保護処理後のテーブル中に, どのレコードに関しても同じレコードが自身を含めて k 個以上存在する” ことであり直観的に個人のレコードを k 個以下に絞れ込めないことを保証する指標である. この指標の確率的な拡張として [3] において, Pk-匿名性が定義されている, すなわち Pk-匿名性を “ある個人のレコードを $1/k$ 以上の確信度に絞り込めない” として定義している. 任意のテーブル T と維持置換撹乱に限らず任意の遷移確率行列 A を用いて撹乱する際の k と A の関係式は以下である [3].

$$k \geq 1 + (\#T - 1) \min_{\substack{u,v \in \mathcal{R} \\ u',v' \in \mathcal{R}}} \frac{A_{uv'} A_{vu'}}{A_{uu'} A_{vv'}}.$$

ここで, $\mathcal{R}$ はレコード空間であり, レコードの取るべき値の集合である. この結果を維持置換撹乱に用いれば以下のように維持確率 ρ_i と k の関係式が示される. 具体的には, 属性の値域が $M_i = \{1, \ldots, m_i\}$ でありレコード数が $\#T$ のテーブルにおいて, 維持確率 ρ_i での維持置換撹乱は

$$k \geq 1 + (\#T - 1) \prod_{i=1}^{n} \left(\frac{1 - \rho_i}{1 + (m_i - 1)\rho_i} \right)^2$$

として, Pk-匿名性を満たす.

3 属性間の相関を考慮した Pk-匿名化

3.1 属性間の相関

予め定められている属性には相関がある場合がある. まずこの相関の具体例と一般の定義について述べる. 例えば, 年齢という属性と運転免許の種類の属性には年齢の要素を 1 つ定めた時に運転免許の種類のとるべき要素の部分集合が定まる. 具体的には, 値域を

$$M_{\text{年齢}} = \{\ldots, 16\text{ 歳}, 17\text{ 歳}, 18\text{ 歳}, \ldots\},$$

$$M_{\text{免許}} = \{\text{ 無し, 普通自動車, 原付, }\ldots\}$$

とする. この時に $2^{M_{\text{免許}}}$ を $M_{\text{免許}}$ の部分集合族として以下のような写像が定まる;

$$\begin{aligned}
\pi : M_{\text{年齢}} &\longrightarrow 2^{M_{\text{免許}}} \\
&\cdots\ , \\
17\text{ 歳} &\longmapsto \{\text{ 無し, 原付, 小型特殊,}\ldots\}, \\
18\text{ 歳} &\longmapsto \{\text{ 無し, 普通 1 種, 原付,}\ldots\}, \\
&\cdots\ .
\end{aligned}$$

この例に従って, 属性の相関の一般化を行う.

定義 3.1 属性の値域 $M_1, M_2, \ldots, M_n$ に対してそれらの相関とは以下のように定義される写像の列 $\{\pi_2, \ldots, \pi_n\}$ のことである.

$$\pi_j : \prod_{s=1}^{j-1} M_s \longrightarrow 2^{M_j}$$

ここで, 集合 M に対して, 2^M はその部分集合族を表す.

従来の相関を考慮しない場合については属性の値域 $M_1, M_2, \ldots, M_n$ に対して,

$$\begin{aligned}
\pi_j : \prod_{s=1}^{j-1} M_s &\longrightarrow 2^{M_j}, \\
(a_1, \ldots, a_{j-1}) &\longmapsto M_j
\end{aligned}$$

という定写像を与えていたことになり, この相関の定義が従来の場合の拡張であることがわかる. また $2 \leq j \leq n$ とベクトル

$$(i^{(1)}, i^{(2)}, \ldots, i^{(j-1)}) \in M_1 \times \cdots \times M_{j-1}$$

に対して集合 $\pi_j(i^{(1)}, i^{(2)}, \ldots, i^{(j-1)}) \subseteq M_j$ を形式的に

$$\{1^{(j)}_{i^{(1)},\ldots,i^{(j-1)}}, 2^{(j)}_{i^{(1)},\ldots,i^{(j-1)}} \ldots, p^{(j)}_{i^{(1)},\ldots,i^{(j-1)}}\}$$

と表記できる. 逆に $p^{(j)}_{i^{(1)},\ldots,i^{(j-1)}}$ としたときには集合 $\pi_j(i^{(1)}, i^{(2)}, \ldots, i^{(j-1)})$ の要素の個数を表すものとする.

この相関に伴い, レコードに条件が加わる. 従来のレコードでは属性の値域 $M_1, \ldots, M_n$ の元を単純に並べていたが, 一方で相関を考慮した場合はレコード $(i^{(1)}, i^{(2)}, \ldots, i^{(n)})$ について $i^{(j)} \in \pi_j(i^{(1)}, \ldots, i^{(j-1)})$ を満すようにする.

また, この相関に伴いクロス集計の定め方も変わる. 従来のクロス集計のでは 2 つの属性の場合では

$$((i, M_2\text{の元のブロック})|i \in M_1)$$

という形をしている. 相関を考慮した場合には

$$((i, \pi_2(i)\text{ の元のブロック})|i \in M_1)$$

となる. なお, この際のクロス集計を表すベクトルの長さは値域 M_2 の部分集合 $\pi_2(i^{(1)})$ の位数の総和 $L = \sum_{i^{(1)} \in M_1} p^{(2)}_{i^{(1)}}$ である. また, レコード $(i^{(1)}, i^{(2)})$ に対して, クロス集計を表す長さ L のベクトルに対して $\sum_{t^{(1)} < i^{(1)}} p^{(2)}_{t^{(1)}} + i^{(2)}$ 番目をカウントさせればよい.

任意の属性の個数の場合, クロス集計を表すベクトルの長さは n 番目の値域 M_n の部分集合 $\pi_n(i^{(1)}, \ldots, i^{(n-1)})$ の位数の総和

$$L = \sum_{(i^{(1)}, i^{(2)}, \ldots, i^{(n-1)})} p^{(n)}_{i^{(1)}, i^{(2)}, \ldots, i^{(n-1)}}$$

となる. また, レコード $(i^{(1)}, i^{(2)}, \ldots, i^{(n)})$ に対して, クロス集計を表す長さ L のベクトルの

$$\sum_{(t^{(1)}, \ldots, t^{(n-1)}) < (i^{(1)}, \ldots, i^{(n-1)})} p^{(n)}_{t^{(1)}, \ldots, t^{(n-1)}} + i^{(n)}$$

番目をカウントさせればよい. ここで, ベクトルの大小関係 $(t^{(1)}, \ldots, t^{(n-1)}) < (s^{(1)}, \ldots, s^{(n-1)})$ は辞書式順序であり, すなわち適当な $1 \leq a \leq n-1$ が存在して, $j < a$ に対して $t^{(j)} = s^{(j)}$ と $t^{(a)} < s^{(a)}$ が成立することである.

3.2 相関を考慮した撹乱

属性間に相関 $\{\pi_2,\ldots,\pi_n\}$ がある場合に撹乱の方法が以下のように変化させる.

Input. $(a^{(1)},\ldots,a^{(n)}) \in M_1\times\cdots\times M_n$.
ここで, $a^{(j)} \in \pi_j(a^{(1)},\ldots,a^{(j-1)})$ を満足する.

Output. $(a'^{(1)},\ldots,a'^{(n)}) \in M_1\times\cdots\times M_n$.

Step 1. 属性 A_1 に対しては従来と同じくパラメータ ρ_1 で維持置換撹乱を行い, 元 $a^{(1)} \in M_1$ に対して $a'^{(1)} \in M_1$ を定め撹乱する.

Step 2. 元 $a^{(2)} \in \pi_2(a^{(1)})$ を撹乱する場合, $a^{(1)} = a'^{(1)}$ であれば $\pi_2(a'^{(1)})$ の中でパラメータ ρ_2 で維持置換撹乱を行い $a'^{(2)}$ を定め, そうでなければ $\pi_2(a'^{(1)})$ の中の元を一様ランダムに取り $a'^{(2)}$ を定める.

⋮

Step n. 元 $a^{(n)} \in \pi_3(a^{(1)},a^{(2)},\ldots,a^{(n-1)})$ を撹乱する場合, $j=1,\ldots,n-1$ に対して $a^{(j)} = a'^{(j)}$ であれば $\pi_n(a'^{(1)},a'^{(2)},\ldots,a'^{(n-1)})$ の中でパラメータ ρ_n で維持置換撹乱を行い $a'^{(n)}$ を定め, そうでなければ

$$\pi_n(a'^{(1)},a'^{(2)},\ldots,a'^{(n-1)})$$

の中の元を一様ランダムに取り $a'^{(n)}$ を定める.

従来の相関がない場合の撹乱と同様にこの相関付の撹乱は遷移確率行列として表現できる.

属性 A_1 については従来と同じ遷移確率行列が定まる;

$$A_1 = \rho_1 E(m_1) + ((1-\rho_1)/m_1)U(m_1).$$

ここで, $E(m_1)$ と $U(m_1)$ ははそれぞれサイズが $m_1\times m_1$ の単位行列と要素が凡て 1 の行列である. 2 つ目の属性に対する遷移確率行列 A_2 は各 $\pi_2(i^{(1)})\times\pi_2(j^{(1)})$ ごとのブロック行列

$$A_2 = (V_{\pi_2(i^{(1)})\times\pi_2(j^{(1)})})$$

として表現でき, $i^{(1)}=j^{(1)}$ の時には,

$$V_{\pi_2(i^{(1)})\times\pi_2(i^{(1)})} = \rho_2 E(p^{(2)}_{i^{(1)}}) + \frac{(1-\rho_2)}{p^{(2)}_{i^{(1)}}}U(p^{(2)}_{i^{(1)}})$$

であり, $i^{(1)}\neq j^{(1)}$ の時には, $U(n_1,n_2)$ をサイズが $n_1\times n_2$ の要素が凡て 1 の行列として

$$V_{\pi_2(i^{(1)})\times\pi_2(j^{(1)})} = (1/p^{(2)}_{j^{(1)}})U(p^{(2)}_{i^{(1)}},p^{(2)}_{j^{(1)}})$$

となる. 以降同様に $A_3,A_4,\ldots,A_n$ が定まる. 具体的には $s=2,\ldots,n$ に対して遷移確率行列 A_s は $\pi_s(i^{(1)},\ldots,i^{(s-1)})\times\pi_s(j^{(1)},\ldots,j^{(s-1)})$ 毎のブロック行列 $(V_{\pi_s(i^{(1)},\ldots,i^{(s-1)})\times\pi_s(j^{(1)},\ldots,j^{(s-1)})})$ として表現できる. それぞれのブロック行列は $(i^{(1)},\ldots,i^{(s-1)})=(j^{(1)},\ldots,j^{(s-1)})$ の時には,

$$V_{\pi_s(i^{(1)},\ldots,i^{(s-1)})\times\pi_s(i^{(1)},\ldots,i^{(s-1)})} = \rho_s E(p^{(s)}_{i^{(1)},\ldots,i^{(s-1)}}) + \frac{(1-\rho_s)}{p^{(s)}_{i^{(1)},\ldots,i^{(s-1)}}}U(p^{(s)}_{i^{(1)},\ldots,i^{(s-1)}})$$

であり, $(i^{(1)},\ldots,i^{(s-1)})\neq(j^{(1)},\ldots,j^{(s-1)})$ の時には

$$V_{\pi_s(i^{(1)},\ldots,i^{(s-1)})\times\pi_s(j^{(1)},\ldots,j^{(s-1)})} = (1/p^{(s)}_{j^{(1)},\ldots,j^{(s-1)}})U(p^{(s)}_{i^{(1)},\ldots,i^{(s-1)}},p^{(s)}_{j^{(1)},\ldots,j^{(s-1)}})$$

となる. 次に個々に定められた遷移確率行列 A_s に関して, 複数の属性に関する遷移確率行列にするために以下の行列の演算を定義する.

定義 3.2 サイズが $p_A\times p_A$ の行列 $A=(a_{ij})$ と $p_A\times p_A$ 個のブロック行列で生成される行列

$$B=(C_{ij}),\quad (\text{ここで } C_{ij}\text{は行列})$$

に対して, 演算 $\otimes_\pi$ を以下のように定める;

$$A\otimes_\pi B = (a_{ij}C_{ij}).$$

ここで, $a_{ij}C_{ij}$ は C_{ij} の a_{ij} スカラー倍である.

一つの属性ごとに定めた行列に対して A_i に対して複数の属性を含めた行列は上で定義される行列の演算を使えば表現でき A がそれである;

$$A = A_1\otimes_\pi A_2\otimes_\pi\cdots\otimes_\pi A_n.$$

3.3 相関を考慮した撹乱の Pk-匿名性

上記で定めた相関を考慮した場合の撹乱方法と安全性の指標 k の関係式を導く.

任意のテーブル T に対して, 維持置換撹乱に限らず任意の遷移確率行列 A を用いて撹乱する際の A と k の関係式は以下であった,

$$k \geq 1 + (\#T - 1) \min_{\substack{u,v \in \mathcal{R} \\ u',v' \in \mathcal{R}}} \frac{A_{uv'}A_{vu'}}{A_{uu'}A_{vv'}}.$$

この関係式より $A = \bigotimes_\pi A_i$ の場合に, $z = \min \frac{A_{uv'}A_{vu'}}{A_{uu'}A_{vv'}}$ を計算することで ρ_i と k の関係式を導く.

まず簡単の場合に $n = 2$ の場合について述べる. 以下の条件を仮定しても一般性を失わない

$$p^{(2)}_{1^{(1)}} \geq p^{(2)}_{2^{(1)}} \geq \cdots \geq p^{(2)}_{m_1^{(1)}}.$$

また記号として $\delta_0 = 1 - \rho_1, \delta_1 = 1 + (\#M_1 - 1)\rho_1$ とする.

レコード u, u', v, v' をそれぞれ

$$\pi_2(i^{(1)}_1), \pi_2(i^{(1)}_2), \pi_2(j^{(1)}_1), \pi_2(j^{(1)}_2)$$

の元から選ぶのだが, 4 つのペア

$$(i^{(1)}_1, j^{(1)}_1), (i^{(1)}_2, j^{(1)}_2), (i^{(1)}_1, j^{(1)}_2), (i^{(1)}_2, j^{(1)}_1)$$

が一致しているかどうかで場合を分け, z を計算させる. 場合分けでの計算より, 最小値を与えるのは以下の 3 つの場合である;

(1) $i^{(1)}_1 \neq j^{(1)}_1$ かつ $i^{(1)}_2 \neq j^{(1)}_2$ かつ $i^{(1)}_1 = j^{(1)}_2$ かつ $i^{(1)}_2 = j^{(1)}_1$ の場合で

$$z = \frac{\delta_0^2}{\delta_1^2}(1 - \rho_2)^2.$$

(2) $i^{(1)}_1 = j^{(1)}_1$ かつ $i^{(1)}_2 = j^{(1)}_2$ かつ $i^{(1)}_1 \neq j^{(1)}_2$ かつ $i^{(1)}_2 \neq j^{(1)}_1$ の場合で

$$z = \frac{\delta_0^2}{\delta_1^2} \frac{1}{(1 + (p^{(2)}_{j^{(1)}_1} - 1)\rho_2)(1 + (p^{(2)}_{j^{(1)}_2} - 1)\rho_2)}.$$

(3) $i^{(1)}_1 = j^{(1)}_1$ かつ $i^{(1)}_2 = j^{(1)}_2$ かつ $i^{(1)}_1 = j^{(1)}_2$ かつ $i^{(1)}_2 = j^{(1)}_1$ の場合で,

$$z = \left(\frac{(1 - \rho_2)}{(1 + (p^{(2)}_{j^{(1)}_1} - 1)\rho_2)} \right)^2.$$

まとめると, 最小値 $\min \frac{A_{uv'}A_{vu'}}{A_{uu'}A_{vv'}}$ は

$$\min \left\{ \frac{\delta_0^2}{\delta_1^2}(1 - \rho_2)^2, \right.$$
$$\frac{\delta_0^2}{\delta_1^2} \frac{1}{(1 + (p^{(2)}_{1^{(1)}} - 1)\rho_2)(1 + (p^{(2)}_{2^{(1)}} - 1)\rho_2)},$$
$$\left. \left(\frac{(1 - \rho_2)}{(1 + (p^{(2)}_{1^{(1)}} - 1)\rho_2)} \right)^2 \right\}.$$

一般の属性の個数の場合について以下に記す. $1 \leq s \leq n - 1$ に対して $\{p^{(s+1)}_{i^{(1)},\ldots,i^{(s)}}\}$ の最大値を与えるベクトルを $\alpha_s = (\alpha^{(1)}, \ldots, \alpha^{(s)})$ とし, 2 番目に大きい値を与えるベクトルを $\beta_s = (\beta^{(1)}, \ldots, \beta^{(s)})$ とする.

レコード u, u', v, v' を $t = 1, 2$ としてそれぞれ

$$\pi_n(i^{(1)}_t, \ldots, i^{(n-1)}_t), \pi_n(j^{(1)}_t, \ldots, j^{(n-1)}_t)$$

の元から選ぶのだが, まずステップ 1 として第 1 要素の $i^{(1)}_1, j^{(1)}_1, i^{(1)}_2, j^{(1)}_2$ の一致で場合分けを行う. 先の $n = 2$ の場合と同じく, 以下の三つの場合が考えられる,

(1) $i^{(1)}_1 \neq j^{(1)}_1$ かつ $i^{(1)}_2 \neq j^{(1)}_2$ かつ $i^{(1)}_1 = j^{(1)}_2$ かつ $i^{(1)}_2 = j^{(1)}_1$ の場合.

(2) $i^{(1)}_1 = j^{(1)}_1$ かつ $i^{(1)}_2 = j^{(1)}_2$ かつ $i^{(1)}_1 \neq j^{(1)}_2$ かつ $i^{(1)}_2 \neq j^{(1)}_1$ の場合.

(3) $i^{(1)}_1 = j^{(1)}_1$ かつ $i^{(1)}_2 = j^{(1)}_2$ かつ $i^{(1)}_1 = j^{(1)}_2$ かつ $i^{(1)}_2 = j^{(1)}_1$ の場合.

(1) の場合の次のステップとして, 第 2 要素以降での比較を行う. $i^{(1)}_1 \neq j^{(1)}_1$ かつ $i^{(1)}_2 \neq j^{(1)}_2$ であるので, 凡ての $s \geq 2$ において $i^{(s)}_1 \neq j^{(s)}_1$ かつ $i^{(s)}_2 \neq j^{(s)}_2$ が従う. この時に, 最小値 z の候補を与えるのは凡ての $s \geq 2$ において $i^{(s)}_1 = j^{(s)}_2$ かつ $i^{(s)}_2 = j^{(s)}_1$ のときであり, その値は

$$\frac{\delta_0^2}{\delta_1^2} \prod_{s=2}^{n-1} (1 - \rho_s)^2 \tag{1}$$

である.

(2) の場合も (1) の場合と同様に第 2 要素以降の比較を行うことで, 最小値の候補を与えることができる. 次のステップとして, 第 2 要素以降での比較を行う. その最小値 z の候補は

$$\frac{\delta_0^2}{\delta_1^2} \prod_{s=2}^{n-1} \frac{1}{(1 + (p^{(s)}_{\alpha_{s-1}} - 1)\rho_s)(1 + (p^{(s)}_{\beta_{s-1}} - 1)\rho_s)} \tag{2}$$

である.

(3) の場合次のステップとして, 第 2 要素での比較を行う. $i_1^{(2)}, j_1^{(2)}, i_2^{(2)}, j_2^{(2)}$, の場合について第 1 要素と同じく 3 つの場合に分かれる ;

(3-1) $i_1^{(2)} \neq j_1^{(1)}$ かつ $i_2^{(1)} \neq j_2^{(2)}$ かつ $i_1^{(2)} = j_2^{(2)}$ かつ $i_2^{(2)} = j_1^{(1)}$ の場合.

(3-2) $i_1^{(2)} = j_1^{(2)}$ かつ $i_2^{(2)} = j_2^{(2)}$ かつ $i_1^{(2)} \neq j_2^{(2)}$ かつ $i_2^{(2)} \neq j_1^{(1)}$ の場合.

(3-3) $i_1^{(2)} = j_1^{(2)}$ かつ $i_2^{(2)} = j_2^{(2)}$ かつ $i_1^{(2)} = j_2^{(2)}$ かつ $i_2^{(2)} = j_1^{(2)}$ の場合.

(3-1) や (3-2) の場合には (1) や (2) の場合と同じ場合に帰着し, 最小値の候補の計算が可能である. (3-3) の場合には, 再度第 3 要素の比較で最少値 z の候補を求めればよい. 以下帰納的求めた最小値の候補を示す. $s' = 2, \ldots, n-1$ に対して

$$\prod_{s=2}^{s'} \left(\frac{1-\rho_s}{1+(p_{\alpha_{s-1}}^{(s)}-1)\rho_s} \right)^2 \prod_{s=s'+1}^{n-1} (1-\rho_s)^2 \tag{3}$$

及び

$$\prod_{s=2}^{s'} \left(\frac{1-\rho_s}{1+(p_{\alpha_{s-1}}^{(s)}-1)\rho_s} \right)^2 \times \prod_{s=s'+1}^{n-1} \frac{1}{(1+(p_{\alpha_{s-1}}^{(s)}-1)\rho_s)(1+(p_{\beta_{s-1}}^{(s)}-1)\rho_s)} \tag{4}$$

である. 以上より, 最小値 $z = \min \frac{A_{uv'}A_{vu'}}{A_{uu'}A_{vv'}}$ は式 (1) から (4) の最少値となる.

3.4 相関を考慮した再構築

属性間の相関を考慮した撹乱の方法を提案した. この撹乱に対する再構築の方法は, 相関がない場合と同じように, 行列 $A = \bigotimes_\pi A_i$ を用いて撹乱後のテーブルのクロス集計についてベイズ推定法を行えばよい [2].

4 実装結果

本章では提案手法を実装し, 従来法との比較実験を行った結果を示す. 計算環境は CPU : Core i7-3770K, メモリ : 16GB, OS : Ubuntu 13.10, 使用ソフト: Magma, R の環境で行った.

使用したデータは米国の 1990 年の国勢調査データ USCensus 1990 (2,458,285 レコード) である [6]. その中の年齢, 学歴と性別の 3 つの属性を抜きだした. それぞれの属性の値域は年齢は 5 歳刻みの 19 種類, 学歴は { 3 歳未満 or 無回答, 無し, 保育園, 幼稚園, 4 年生以内, 8 年生以内, 9 年生, 10 年生, 11 年生, 12 年生, 高卒, 大学入学, 職業準学士, 学問準学生, 学士, 修士, 専門学位, 博士 } の 18 種類と性別は男女の 2 種類である. 単純にこれらの属性を組み合わせた場合にレコードの取りうる値の種類は 684 である. 単純に属性を組み合わせるのに対して, 3 つの属性のうち年齢と学歴の 2 つの属性に対して相関を定義し, 実際にはありえない属性値の組み合わせを削除する. 具体的には, 0〜4 歳等の若年齢層に対して大学入学等の高い学歴の組み合わせや 20 代後半以上の成人年齢層の対して義務教育のため 10 年生以下の学歴の組み合わせはあり得ない. 削除した結果, レコードの取りうる値の種類は 306 個になった. なお, 実際のテーブルには定義した相関による属性の組み合わせに属さないレコードが存在するが, そのようなレコードは削除して 2,002,755 レコードのテーブルを実験に用いた.

撹乱再構築後の分析精度はデータマイニング手法の中で使用される一般的なクロス集計と関連ルール抽出の 2 つによって測定する.

4.1 L_1 精度による分析精度

L_1 距離はクロス集計の各要素の差の絶対値の総和をレコード数で割ったものであり, 直感的にはクロス集計の結果が総レコード数の外れ具合を表す. L_1 距離が小さいほど, 元のデータのクロス集計と近い結果が得られていることになり, データとしての有用性が高いと考えられる. 撹乱前のテーブルと従来の Pk-匿名化後のテーブルの L_1 距離及び撹乱前のテーブルと相関を考慮した Pk-匿名化後のテーブルの L_1 距離を比較する. また, 撹乱処理の際に匿名性は $k = 2, 3, 10$ を満足するように行った.

L_1 距離は従来法と比較して, 提案法では平均して 1.22 倍低下していることが確認できた.

表 2: L_1 距離

	$k=2$	$k=3$	$k=10$
従来法	0.210	0.247	0.325
提案法	0.174	0.198	0.267
精度向上の比率	1.21	1.25	1.22

4.2 関連ルールによる分析精度

関連ルールとは, テーブル内に頻出する各属性値の組み合わせすなわちトランザクションを支持度と確信度の 2 つの評価指標によって判定し, この 2 つの指標が予め定められた閾値よりも高いトランザクションをいう [5]. 今回の実装では, 支持度の閾値を 0.05 とし確信度の閾値を 0.05 に設定し関連ルールを抽出した. 元データから抽出した関連ルールを支持度でソートし, 上位 10 個のトランザクションについて撹乱後に同じトランザクションが抽出できている割合 (抽出率), 抽出したトランザクションの順序の正答率, 支持度の差分の平均及び確信度の差分の平均について調べた. また, 撹乱処理の際に匿名性は $k=2$ を満足するように行った.

表 3: 関連ルールにおける誤差

	抽出率	順序	支持度	確信度
従来法	90%	60%	2.2%	2.1%
提案法	90%	80%	2.1%	2.5%

実験の結果, 順序の正答率に 20%の向上が確認ができたが, 従来法と新手法について抽出した関連ルールの抽出率, 支持度と確信度の差分はほぼ影響がなかった. 従来法では大きな支持度を持つトランザクションに対してはその支持度は保存されやすいが小さい支持度を持つトランザクションに関してはその支持度が保存されにくい傾向があった. しかし, 提案法では相関を考慮したことで小さい支持度を持つトランザクションについても, 支持度の分布が保存されやすくなったので順序の正答率が向上したと考察できる.

5 まとめ

Pk-匿名化の分析精度の向上による適用範囲の拡張に向けて, 属性間の相関を考慮し属性値の組み合わせの取りうる範囲を小さくするような Pk-匿名化の手法を提案し, 実験を行った. 実験では, L_1 距離の評価と関連ルール抽出において分析精度の向上が確認できた.

課題として, 更なる向上を目指すためには, 元のテーブルに依存するレコードの値の分布や関連ルール等の適切な情報を用いて撹乱の属性値の組み合わせの範囲の制限を行えばよいと思われる. ただし, その際には元のテーブルから適切な情報を公開し撹乱をしているので, その情報が匿名性に対してどのように影響するかを考察する必要がある.

参考文献

[1] R. Agrawal, R. Srikant and D. Thomas. Privacy-preserving Data Mining. Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data, 2000.

[2] 五十嵐大, 千田浩司, 高橋克巳. 多値属性に適用可能な効率的プライバシー保護クロス集計, コンピューターセキュリティシンポジウム 2008 (CSS2008).

[3] 五十嵐大, 千田浩司, 高橋克巳. k-匿名性の確率的指標への拡張とその適用例, コンピュータセキュリティシンポジウム 2008(CSS2009), 2009.

[4] 廣田啓一, 正木彰伍, 齋藤恆和, et al. 確率的な安全性の指標に基づくパーソナルデータ匿名化システムの構築と評価, セキュリティ心理学とトラスト研究会 2014.

[5] X. Wu , V Kumar, J. R. Quinlan, et al. Top 10 Algorithms in Data Mining. Knowledge and Information Systems 2008, Volume 14.

[6] USCensus(1990) Data Set. 1990, http://archive.ics.uci.edu/ml/datasets.html.