数値属性における, k-匿名性を満たすランダム化手法

五十嵐　大 †　　千田　浩司 †　　高橋　克巳 †

† 日本電信電話株式会社 NTT 情報流通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

{ ikarashi.dai, chida.koji, takahashi.katsumi }@lab.ntt.co.jp

あらまし 近年, プライバシーを保護しつつデータを活用する研究が活発に行われており, その中でも k-匿名性を実現する k-匿名法は特に注目されている. 一方, 確定的アルゴリズムを用いる k-匿名化ではなく, データのランダム化による匿名化手法も研究されている. CSS2009 において筆者らはランダム化したデータベース上の k-匿名性, Pk-匿名性を定義しその実現方法を示した. しかしその対象は例えば性別のように値の種類の少ない, カテゴリ属性のみであった. 本稿ではこれを身長や年収などの数値属性に拡張する.

まず, 非有界な数値属性がノイズの加算では Pk-匿名性を満たし得ないことを示し, 次に有界であっても Agrawal らによって提案された一様分布および正規分布の加算は Pk-匿名性を満たし得ないことを示す. そして最後に Laplase 分布と呼ばれるノイズの加算であれば有界な数値属性において Pk-匿名性を満たすことを示す.

Randomized k-Anonymization for Numeric Attributes

Dai IKARASHI†　　Koji CHIDA†　　Katsumi TAKAHASHI†

†NTT Information Sharing Platform Laboratories, NTT Corporation
3-9-11, Midoricho, Musashino-shi, Tokyo 180-8585, Japan

{ ikarashi.dai, chida.koji, takahashi.katsumi }@lab.ntt.co.jp

Abstract Recently, studies on the utilizing data privately, are widely conducted. Among them, k-anonymity is especially paid attention to. On the other hand, in community of database and statistics, not only k-anonymity, which adopt non-probabilistic algorithm, but also anonymization using data randomization is expected and has been studied. In CSS2009, we defined Pk-anonymity, which is k-anonymity on a randomized database, and showed a method for satisfying Pk-anonymity. However, it was applicable only to categorical attributes. In this paper, we extend it to numerical attributes.

First, we show that the addition of a noise on a unbounded numerical attribute can not satisfy Pk-anonymity, and that even on a bounded attribute, the addition of the uniform distribution and the normal distribution proposed by Aggrawal et al. can not satisfy it. Finally, we show that a noise called Laplase distribution satisfies Pk-anonymity on a bounded numerical attribute.

1　はじめに

k-匿名性 [1][2] は “データベース上である人のデータを k 個未満に絞り込むことができない” という, プライバシー保護データ分析の分野で最もポピュラーなプライバシーの概念である. k-匿名性は直感的で分かりやすいため安心感を得られやすいと考えられる一方, データのランダム化を用いた匿名化手法には適用することができなかった. ランダム化は, 統計学及びデータベースの両分野で有望視されている重要な匿名化手法である. しかしランダム化には k-匿名性のような直感的なプライバシーの概念が存在せず, 安全性の評価が困難であることが課題であった.

Pk-匿名性 [3] はこの課題に対して k-匿名性を

ランダム化に適用したプライバシーの概念であり, “データベース上である人のデータを $1/k$ 以上の確率で当てることができない” ことを保証する. [3] では維持-置換撹乱 [4] と呼ばれるランダム化が *Pk*-匿名性を満たすことが示された. しかし維持-置換撹乱は値の間に関係がない, 性別などといった, カテゴリ属性に対するランダム化手法であるため, 身長, 年収など数値属性に対する適用は限定的であった.

これに対し本稿では, 数値属性に対応するランダム化, 特に数値の加算による手法について考察し, *Pk*-匿名性を満たすノイズとして Laplase 分布と呼ばれるノイズの加算を提案する.

そのためにまず本稿では, [5] で我々が示した定理を用いて以下の 2 つの不可能性を証明する.

- 非有界な数値属性における, ノイズ加算による *Pk*-匿名性の実現不可能性
- 一様分布の加算及び正規分布の加算による *Pk*-匿名性の実現不可能性

前者は, 非有界, すなわち定義域が無限に拡がるような数値属性においては, そもそもノイズの加算というアプローチでは *Pk*-匿名性を満たすことができないことを示している. 後者は実数全域で定義される分布であっても, [6] で Agrawal らが提案した, 最もポピュラーな 2 つの分布の加算が *Pk*-匿名性を満たせないことを言っている.

最後に, 上記のように幾つかの分布が *Pk*-匿名性を満たせない中, Differential Privacy[7] の実現にも用いられる, Laplase 分布の加算であれば *Pk*-匿名性を満たすことができることを示す.

2 関連研究

ノイズの加算を用いたプライバシー保護手法に関しては, 統計の分野では 30 年ほど以前より研究されており, [8] などがある. またデータベースの分野では Agrawal らがプライバシー保護データマイニングとして, 各個人のデータにノイズを加算して, データマイニング結果のみを精度良く再構築するという, 再構築法を提案した [6]. 個人のデータにノイズを加えるのではないが, Dwork は対話型の統計データベースにおいて, Laplase 分布と呼ばれる分布をクエリに対する応答に加算してプライバシー保護を実現する手法及び, その場合のプライバシーである Differential Privacy を提案した [7].

ランダム化に対する *k*-匿名性の適用に関しては, [3] で我々が *Pk*-匿名性を提案したものの, 離散値を定義域とする属性を前提としたため, ノイズ加算に対しての適用可能性は十分ではなかった. そこで [5] では定義域を連続値も含む一般の確率空間に拡張し, ノイズ加算にも適用可能となった. 本稿ではこの結果を用いてノイズ加算の *Pk*-匿名性を評価する.

3 提案 0 : 維持-置換撹乱による連続属性の *Pk*-匿名化

連続属性に対して, 維持置換撹乱を用いて *Pk*-匿名化を行うことも, 限定的ながら可能である.

自明な適用方法

最も簡単な方法は, 属性値を離散化して維持-置換撹乱を施す方法である. 例えば, 年齢であれば元より離散的であるから各年齢を 1 属性値として維持-置換撹乱を行うなどである. この自明な方法の欠点は, 数値同士の距離関係を考慮していないため, 非常に高い確率で無関係な数値に置換されてしまうことである.

提案 0 : ビット分解による方法

次に, 数値をビット表現して, 各ビットを 1 属性と見て維持-置換撹乱を施す方法である. この方法では例えば最上位ビットを維持すればある程度数の大きさが維持されることから, 前記の自明な方法よりよさそうなことが分かる.

しかし, ビット分解による方法でも, 上位の方のビットが一つでも置換されれば大きな変化を生むため, 属性値は非常に高い確率でほぼ無関係な値に置換されてしまう.

このような問題を解決するため, 本稿では上記よりもさらに連続属性にマッチしたランダム化を探るため, ノイズの加算によるランダム化を解析する.

4 ノイズの加算による匿名化

[6] などでは, テーブル型データベースの各レコード, すなわち各個人のデータの, 各属性に対して, (確率の意味で) 独立に加法によってノイズを加算してプライバシー保護を行うことを試みている.

本稿では以降この方法と同様, 各レコードの各属性に対して, 独立にノイズを加算する種類のプライバシー保護アルゴリズムを考える.

まず, [5] で示された以下の定理を記しておく. 本稿では全面的に本定理を用いる.

定理 4.1 匿名化システム $(\Delta, \mathcal{R}, \mathcal{R}', \mathcal{V}, \mathcal{V}')$ をとる. Δ をレコード毎に独立にランダム化を行う匿名化とし, 値 $v \in \mathcal{V}$ のレコードが値 $v' \in \mathcal{V}'$ にランダム化される確率を $A_{v,v'}$ と書く. このとき Δ は k を以下の値として Pk-匿名化である.

$$k = 1 + (|\mathcal{R}| - 1)\operatorname*{ess\,inf}_{\substack{u,v\in\mathcal{V}\\ u',v'\in\mathcal{V}'}} \frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}$$

ess sup, ess inf は本質的上限 (下限) と呼ばれる概念で, 測度空間 (または確率空間) を $\mathcal{X}$, 可測関数 (または確率密度関数) を f とすれば以下のように定義される. 積分値を左右しない部分を除いた部分の上下限と思えばよいであろう.

$$\operatorname*{ess\,sup}_{x\in\mathcal{X}} f(x) = \inf\{a \in \mathbb{R} \mid \mu(\{f > a\}) = 0\}$$
$$\operatorname*{ess\,inf}_{x\in\mathcal{X}} f(x) = \sup\{a \in \mathbb{R} \mid \mu(\{f < a\}) = 0\}$$

$\mu(\{f > a\})$ は, 関数 f の関数値が a を超えるような定義域の部分の測度 ($\mathcal{X}$ が 1 次元なら長さ, 2 次元なら面積など) である.

5 幾つかのノイズに関する Pk-匿名性の実現不可能性

本節では 1 節で述べた, ノイズの加算に関する Pk-匿名性の実現の困難性を示す.

はじめに, 非有界な数値属性に対してはノイズ加算では Pk-匿名性を満たすことができないことを示す.

定理 5.1 n を正整数とし, $\mathcal{V}$ を $\mathbb{R}^n$ 上 *Lebesgue* 可測集合とする. Δ は加法ノイズ, すなわち以下を満たすとする.

- Δ はレコードごとにランダム化を行う匿名化である. レコードごとのランダム化を, 値 $v \in \mathbb{R}^n$ が値 $v' \in \mathbb{R}^n$ にランダム化される確率で表して, $A_{v,v'}$ と書く.
- ある確率密度関数 $f : \mathbb{R}^n \to \mathbb{R}$ が存在して任意の $v \in \mathcal{V}$, $x \in \mathbb{R}^n$ に対して $A_{v,v+x} = f(x)$

このとき, $\mathcal{V}$ が *Lebesgue* 零集合を除いても非有界となっていれば, Δ は $k = 1$ 以外では Pk-匿名化とならない.

Lebesgue 可測集合は, 通常想起するような $\mathbb{R}^n$ 上の集合であればほぼ当てはまると思ってよいであろう. Lebesgue 零集合は $\mathbb{R}^n$ 上では, $n-1$ 次元以下の部分空間のことと思ってほぼ差し支えない.

証明 まず, 任意の $\mathcal{V}$ の中で境界は零集合であり, ess sup には関与しない. よって境界を除いてしまい, $\mathcal{V}$ を開集合と仮定して一般性を失わない.

任意の点 $x \in \mathbb{R}^n$ と実数 l に対して, $C_{x,l}$ を, 重心 x, 一辺の長さ l の n 次元立方体とする. そして任意の点 $x \in \mathbb{R}^n$ と可測関数 f について, $\overline{f}$ を,

$$\overline{f}(x) = \lim_{l\to 0}\int_{C_{x,l}} \frac{f(y)}{l^n}dy$$

とする. すると任意の $\mathbb{R}^n$ 上開集合 X で

$$\inf_{x\in X}\overline{f}(x) = \operatorname*{ess\,inf}_{x\in X}\overline{f}(x) = \operatorname*{ess\,inf}_{x\in X} f(x)$$

となる. すなわち, ess inf を inf に置き換えることが可能となる.

次に $u, v \in \mathcal{V}$ に対して, $x \in \mathbb{R}^n$ をとり,

$$u' = u + \overline{f}(x),\ \ v' = v + \overline{f}(x),\ \ y = v - u$$

とする. すると

$$\operatorname*{ess\,inf}_{u,v\in\mathcal{V},u',v'\in\mathcal{V}'} \frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}} \le \inf_{x,y\in\mathbb{R}^n} \frac{\overline{f}(y)\overline{f}(-y)}{\overline{f}(x)^2}$$

である. ここで, $\sup_{x\in\mathbb{R}^n}\overline{f}(x) = 0$ とすると $\overline{f}$ の積分は 0 になってしまうはずで, $\overline{f}$ が確率密度関数であることに反するから, $\overline{f}(\hat{x}) \neq 0$ となる点 $\hat{x}$ は存在する.

次に, $\mathbb{R}^{n(+)}$ を, $\mathbb{R}^n$ の部分集合で全成分が非負である部分とする. そして関数 $\overline{g} : \mathbb{R}^{n(+)} \to \mathbb{R}$ を,

$$\overline{g}(y) = \sum_{S\in\{-1,1\}^n} \overline{f}(S_0y_0, S_1y_1, \ldots, S_{n-1}y_{n-1})$$

とする. すなわち, $\overline{g}$ は $\overline{f}$ の関数値を $\mathbb{R}^{n(+)}$ に折りたたんで足し込んだような関数である. すると

$$\int_{y\in\mathbb{R}^{n(+)}} \overline{g}(y)dy = \int_{y\in\mathbb{R}^n} \overline{f}(y)dy$$

となる. さらに, 任意の $y \in \mathbb{R}^n$ に対して y^+ を, y の全成分の符号を正に置き換えたものとする. すると $\overline{g}$ の定義より

$$\overline{f}(y) + \overline{f}(-y) \le \overline{g}(y^+)$$

が成り立つ. $\overline{f}$ は非負関数であるから, 相加相乗平均の関係より,

$$\overline{f}(y)\overline{f}(-y) \le \frac{(\overline{f}(y) + \overline{f}(-y))^2}{4} \le \frac{\overline{g}(y^+)^2}{4}$$

が成り立つ.

ここで $\mathbb{R}^{n(+)}$ が非有界で $\overline{g}$ の積分が有限であることより, $\inf_{y\in\mathbb{R}^{n(+)}}\overline{g}(y) = 0$ でなければならない. よって

$$\inf_{x,y\in\mathbb{R}^n} \frac{\overline{f}(y)\overline{f}(-y)}{\overline{f}(x)^2} \le \inf_{y\in\mathbb{R}^n} \frac{\overline{g}(y^+)^2}{4\overline{f}(\hat{x})} = 0$$

となり,

$$1 + (|\mathcal{R} - 1|) \operatorname*{ess\,inf}_{u,v\in\mathcal{V},u',v'\in\mathcal{V}'} \frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}} = 1$$

が分かる.

□ (定理 5.1)

定理 5.1 では任意のユークリッド空間 $\mathbb{R}^n$ の非有界な部分集合の元を属性値にもつ連続属性に加法ノイズを加えても, 決して Pk-匿名を ($k = 1$ を除けば) 満たすことができないことを見た.

以降では $\mathbb{R}^n$ の有界な連続属性の場合を見る.

[6] では加法ノイズとして 2 つの分布が提案されている. 連続一様分布と正規分布である. この 2 つのノイズでは有界属性であっても Pk-匿名を満たすことはできないことを示す. [6] では 1 次元の連続一様分布であったが, 連続一様分布は多次元でも定義される. 本稿では一般の次元で考える.

定理 5.2 $\mathcal{V}$ を任意の $\mathbb{R}^n$ 上 *Lebesgue* 可測集合, $\mathcal{N}$ を任意の $\mathbb{R}^n$ 上の, 零集合を除いて有界な *Lebesgue* 可測集合, Δ を $\mathcal{N}$ 上では任意の点で確率密度が 0 でなく, それ以外の領域では確率密度 0 であるような分布の加算とする. このとき Δ は $P1$-匿名化にしかならない.

証明 $\mathcal{V}$ が非有界であるときは定理 5.1 から分かる. $\mathcal{V}$ が有界であるときを考える. $\mathcal{V}, \mathcal{N}$ は開集合と仮定して一般性を失わない.

$$\mathcal{N} = [a, b] = [(a_0, a_1, \ldots, a_{n-1}), (b_0, b_1, \ldots, b_{n-1})]$$

とする. ただし各 i で $a_i \lneq b_i$ とする. このとき $\sup_{v\in\mathcal{V}} v_k \neq \inf_{v\in\mathcal{V}} v_k$ なる k が存在する. でなければ一点集合となってしまうからである. さらに $\mathcal{V}, \mathcal{N}$ が開集合であることより, ある正実数 ε が存在して, それより小さな任意の正実数 δ に対して点 $x, y, w, z \in \mathbb{R}^n$ が存在して

$$\sup_{v\in\mathcal{V}} v_k - x_k < \delta, y_k - \inf_{v\in\mathcal{V}} v_k < \delta$$

$$\sup_{r\in\mathcal{N}} r_k - w_k < \delta, z_k - \inf_{r\in\mathcal{N}} r_k < \delta$$

である. さらに δ を十分小さくすればこの x, y は $y_k \lneq x_k$ となるようにできる.

w, z を $w_k = b_k - \delta$, $z_k = a_k + \delta$, また $i \neq k$ では $w_i = z_i = (b_i - a_i)/2$ とおいて,

$$u = x,\ \ v = y,\ \ u' = x + w,\ \ v' = y + z$$

とおく. すると

$$u' - u = w,\ \ v' - v = z,$$

$$u' - v = w + (x - y),\ \ v' - u = z - (x - y)$$

となり, $u' - u, v - v' \in \mathcal{N}$, すなわち $A_{u,u'}, A_{v,v'} \gneq 0$ がすぐに分かる. また, $\delta \lneq 3(\sup_{v\in\mathcal{V}} v_k - \inf_{v\in\mathcal{V}} v_k)$ とすれば

$$\begin{aligned}(u' - v)_k &= w_k + (x_k - y_k)\\ &= \sup_{r\in\mathcal{N}} r_k + (\sup_{v\in\mathcal{V}} v_k - \inf_{v\in\mathcal{V}} v_k) - 3\delta \gneq \sup_{r\in\mathcal{N}} r_k,\end{aligned}$$

$$\begin{aligned}(v' - u)_k &= z_k + (x_k - y_k)\\ &= \inf_{r\in\mathcal{N}} r_k - ((\sup_{v\in\mathcal{V}} v_k - \inf_{v\in\mathcal{V}} v_k) - 3\delta) \lneq \inf_{r\in\mathcal{N}} r_k\end{aligned}$$

より, 逆に $u' - v, v' - u$ は $\mathcal{N}$ に属さない. よって $A_{v,u'}A_{u,v'} = 0$ であり,

$\operatorname{ess\,inf}_{\substack{u,v\in\mathcal{V}\\ u',v'\in\mathcal{V}'}} \dfrac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}} = 0$ となり, $P1$-匿名しか満たさないことが分かる.

□ (定理 5.2)

定理 5.2 は一様分布よりも一般の分布で Pk-匿名性の実現不可能性を証明した. 以下は一様分布に対する直接的な Pk-匿名性の実現不可能性である.

系 5.1 $\mathcal{V}$ を任意の $\mathbb{R}^n$ 上 *Lebesgue* 可測集合, $\mathcal{N}$ を任意の $\mathbb{R}^n$ 上有界 *Lebesgue* 可測集合, Δ を $\mathcal{N}$ 上一様分布の加算とする. このとき Δ は $P1$-匿名化にしかならない.

証明 $\mathcal{N}$ 上一様分布は $\mathcal{N}$ 上で確率密度 0 でなく, $\mathcal{N}$ の外では密度 0 であるから, 定理 5.2 より示された.

□ (系 5.1)

次に正規分布の加算が Pk-匿名を満たさないことを見る.

定理 5.3 $\mathcal{V}$ を任意の $\mathbb{R}$ 上 *Lebesgue* 可測集合, Δ を平均 0, 分散 σ^2 の正規分布の加算とする. このとき Δ は $P1$-匿名化にしかならない.

証明 定理 5.2 と同様, $\mathcal{V}$ が非有界であるときは定理 5.1 から分かる. $\mathcal{V}$ が有界であるときを考える.

正規分布は平均 μ と分散 σ により, x を変数として以下の確率密度関数で表される.

$$f_N(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$u \lneq v$ なる異なる 2 つの実数 $u, v \in \mathcal{V}$ をとる. また正実数 $r \in \mathbb{R}$ をとり, $u' = u - r$, $v' = v + r$, $l = v - u$ とおく. すると

$$\begin{aligned}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}} &= \frac{\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{(u'-v)^2 + (v'-u)^2}{\sigma^2}\right)}{\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{(u'-u)^2 + (v'-v)^2}{\sigma^2}\right)}\\ &= \exp\left(-\frac{(l+r)^2 - r^2)}{\sigma^2}\right) = \exp\left(-\frac{l^2 + 2rl}{\sigma^2}\right)\end{aligned}$$

極限をとると

$$\lim_{r\to\infty} \exp\left(-\frac{l^2 + 2rl}{\sigma^2}\right) = 0$$

より

$$\operatorname{ess\,inf}_{\substack{u,v\in\mathcal{V}\\ u',v'\in\mathcal{V}'}} \frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}} = 0$$

となる.

□ (定理 5.3)

6 提案手法 : Laplase 分布加算

本節では, 前節で様々な分布の加算に関する *Pk*-匿名性の実現不可能性を示したのに対して, Laplase 分布の加算を *Pk*-匿名化アルゴリズムとして提案し, その k の値を示す.

Laplase 分布は両側指数分布とも呼ばれ, これも一様分布や正規分布と並び, 一般的な分布である. [7] では統計値に Laplase 分布を加算することで, Differential Privacy と呼ばれるプライバシーを実現している. Differential Privacy はプライバシー保護の分野で非常に注目しており, ここで用いられている Laplase 分布が前節で紹介した様々な分布と異なり *Pk*-匿名性を実現するという事実は, 非常に興味深い事実である.

定理 6.1 $\mathcal{V}$ を任意の $\mathbb{R}^n$ 上の *Lebesgue* 可測集合, Δ を平均 0, 分散 σ^2 の *Laplace* 分布の加算とする. このとき Δ は

$$k = 1 + (|\mathcal{R}| - 1)\exp\left(-2\frac{\sup\limits_{u,v\in\mathcal{V}}(\|u-v\|_1)}{\sigma}\right)$$

として *Pk*-匿名化である.

ただし, $\|\cdot\|$ は $L1$-ノルムであり, 各成分の絶対値の総和である.

証明 $\mathcal{V}$ が非有界であるとき, 定理 5.1 より $k = 1$ である. このとき, 非有界より

$$\sup_{u,v\in\mathcal{V}}(\|u-v\|_1) = \infty$$

であり,

$$1 + (|\mathcal{R}| - 1)\exp\left(-2\frac{\sup\limits_{u,v\in\mathcal{V}}(\|u-v\|_1)}{\sigma}\right) = 1$$

であるから, 定理の主張は成り立つ.

次に $\mathcal{V}$ が有界であるときを考える. *Laplase* 分布は $x \in \mathbb{R}^n$ を変数として以下の確率密度関数で表される.

$$f_L(x) = \frac{1}{2\sigma}\exp\left(-\frac{\|x-\mu\|_1}{\sigma}\right)$$

ここで $u, v \in \mathcal{V} \subseteq \mathbb{R}^n$, $r, s \in \mathbb{R}^n$ をとり,

$$u' = u - r,\ \ v' = v + s,\ \ l = v - u$$

とおく. すると三角不等式を用いて, 以下の不等式が分かる.

$$\begin{aligned}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}} &= \frac{\frac{1}{2\sigma}\exp\left(-\frac{\|u'-v\|_1 + \|v'-u\|_1}{\sigma}\right)}{\frac{1}{2\sigma}\exp\left(-\frac{\|u'-u\|_1 + \|v'+v\|_1}{\sigma}\right)} \\ &= \exp\left(-\frac{\|l+r\|_1 + \|l+s\|_1 - \|r\|_1 - \|s\|_1}{\sigma}\right) \\ &\geq \exp\left(-\frac{2\|l\|_1}{\sigma}\right)\end{aligned}$$

(三角不等式より.)

等号成立は, 全ての $i < n$ で r, s の各成分 r_i, s_i が, 0 であるか, もしくは l の i 番目の成分 l_i と符号が等しいときである. よって

$$\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}} = \exp\left(-\frac{2\sup\limits_{u,v\in\mathcal{V}}\|u-v\|_1}{\sigma}\right)$$

であり,

$$k = 1 + (|\mathcal{R}| - 1)\exp\left(-2\frac{\sup\limits_{u,v\in\mathcal{V}}(\|u-v\|_1)}{\sigma}\right)$$

となる.

□ (定理 6.1)

6.1 Laplase 分布加算の適用法

ここでは Laplase 分布加算の適用法を述べる. すなわち, 所望の k の値で *Pk*-匿名性を実現することを考える. Laplase 分布を決定するためには σ を決定すればよい. 定理 6.1 の式は単調であることから, 二分法を用いて容易に σ を決定することができそうである. ただし σ の定義域が非有界であることから, その前に探索範囲を絞る必要がある. これも単調性から容易に絞り込むことができ, 結局以下のアルゴリズムで σ を決定できる.

ノイズ決定アルゴリズム (1 属性の場合)

1. σ の初期値を 1 とする.
2. 定理 6.1 により k を計算し, 所望の k 以下であれば, $\sigma_{\max} = \sigma$ とおく. そうでなければ σ を倍にして同じ操作を繰り返す.
3. 区間 $[0, \sigma_{\max}]$ で, 二分法により定理 6.1 の式の k が所望の値になるまで反復計算して最適な σ を求める.

上記は属性がただ一つの場合である. 次に属性が複数ある場合である.

各属性に独立にノイズを加える場合, 定理 4.1 の k の値はさらに以下のように変形できる.

$$1 + (|\mathcal{R}| - 1)\prod_{a:\text{属性}}\operatorname*{ess\,inf}_{\substack{u,v\in\mathcal{V}_a \\ u',v'\in\mathcal{V}'_a}}\frac{(A_a)_{v,u'}(A_a)_{u,v'}}{(A_a)_{u,u'}(A_a)_{v,v'}} \qquad (1)$$

ただし $\mathcal{V}_a, \mathcal{V}'_a$ は各属性に対するランダム化前後の定義域, A_a は各属性に対する属性値の変化の確率密度関数である. すなわち, “1 以上の部分” は, 各属性の $\operatorname*{ess\,inf}\limits_{\substack{u,v\in\mathcal{V}_a \\ u',v'\in\mathcal{V}'_a}}\frac{(A_a)_{v,u'}(A_a)_{u,v'}}{(A_a)_{u,u'}(A_a)_{v,v'}}$ の積になっている.

すると σ は属性間で一律とすれば, k の値は以下のように変形できる.

$$1 + (|\mathcal{R}| - 1) \prod_{a:\text{属性}} \exp\left(-2\frac{\sup_{u,v\in\mathcal{V}_a}(\|u-v\|_1)}{\sigma}\right) \quad (2)$$

典型的な場合である, 属性値が実数区間である場合について考える. 属性間でノイズの強度が意図せず異なるのを防ぐため, それぞれの属性の区間を正規化してから σ を決定する.

ノイズ決定アルゴリズム (n 属性の場合)

1. 各属性の区間を線形変換により [0, 1] に正規化する.
2. 定理 6.1 の式を式 (2) に変更して, 1 属性の場合のアルゴリズムを実行する.
3. ステップ 1 の逆変換により値のスケールを元に戻す.

ノイズ決定アルゴリズム (カテゴリ属性との混在の場合)

さらに, データベースには数値属性とカテゴリ属性は混在するのが普通である. この場合には, 維持-置換撹乱のパラメータである ρ と, Laplase 分布のパラメータである σ に, $\sigma = \tan\left(\frac{\pi}{4}(1-\rho)\right)$ などと重みをつける. そして二分法の際の評価式を, 式 (1) の $\frac{(A_a)_{v,u'}(A_a)_{u,v'}}{(A_a)_{u,u'}(A_a)_{v,v'}}$ に, 属性ごとに [3] で示された維持置換撹乱の場合の式,

$$\frac{1-\rho}{1+(|\mathcal{V}_a|-1)\rho)^2}$$

もしくは, 本稿で示した Laplase 分布加算の場合の式

$$\exp\left(-2\frac{\sup_{u,v\in\mathcal{V}_a}(\|u-v\|_1)}{\sigma}\right)$$

のいずれかを代入して用いればよい.

7 おわりに

本稿では [3] で提案されたランダム化を用いる匿名化のための匿名性である, Pk-匿名性について, 数値属性における加法ノイズに関する 4 つの定理を示した. すなわち, 不可能性の定理として

- 非有界な連続属性に対して, いかなる分布によるノイズの加算も Pk-匿名性を持ち得ない.
- 有界な連続属性であっても, 一様分布によるノイズの加算は Pk-匿名性を持ち得ない.
- 有界な連続属性であっても, 正規分布によるノイズの加算は Pk-匿名性を持ち得ない.

の 3 定理, そして最後の定理は Pk-匿名性を満たすポジティブな結果であり,

- 有界な連続属性において, Laplase 分布によるノイズの加算は Pk-匿名性を持つ.

であり, この Laplase 分布の加算及び, カテゴリ属性と混在するデータベースにおける, 維持-置換撹乱との混成を Pk-匿名化アルゴリズムとして提案した. Laplase 分布は近年非常に注目される Differential Privacy の実現にも用いられる分布であり, 他の分布が Pk-匿名性を実現しない中, Laplase 分布が Pk-匿名性に対しても有効性を発揮することは興味深い事実である.

今後の課題として, 提案したアルゴリズムの評価実験や, ノイズ加算だけでなく乗法的ノイズによる Pk-匿名性の評価などがある. また加法ノイズに関して, k-匿名性を強化した ℓ-多様性の拡張である $P\ell$-多様性 [9] の適用などが考えられる.

参考文献

[1] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). *Proc. of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, p. 188, Seattle, WA, 1998.*

[2] L. Sweeney. k-anonymity: a model for protecting privacy. *Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.10, Issue 5, pp.557-570, 2002.

[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応用例. CSS2009, 2009.

[4] R. Agrawal, R. Srikant and D. Thomas. Privacy Preserving OLAP. *SIGMOD Conference ACM*, pp. 251-262, 2005.

[5] 五十嵐 大, 千田 浩司, 高橋 克巳. ランダム化データベース上の k-匿名性の一般的算出法. CSS2011, 2011.

[6] R. Agrawal and R. Srikant. Privacy-preserving data mining. *Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data*, 2000.

[7] C. Dwork. Differential Privacy. *ICALP (2) 2006*, 2006.

[8] J. J. Kim. A method for limiting discloure in microdata based on random noise and transformation. In *Proceedings of the Section on Survey Research Methods*, 1986.

[9] 五十嵐 大, 千田 浩司, 高橋 克巳. $P\ell$-多様性:属性推定に対する再構築法のプライバシーの定量化. CSS2010, 2010.