多変量解析のための処理効率の良い再構築法

長谷川 聡 †　　濱田 浩気 †　　菊池 亮 †

†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
{ hasegawa.satoshi, hamada.koki, kikuchi.ryo}@lab.ntt.co.jp

あらまし 個人情報の保護とビッグデータの活用を両立させるための技術として, 個人情報の匿名化技術が研究されている. 匿名化技術として, 既存の k-匿名化と同等のプライバシ保護強度を担保した, 確率的な処理による k-匿名化 (Pk-匿名化) が提案されている. Pk-匿名化は, ランダムに値を書き換えるだけで匿名化処理が行えるため処理効率が良い. しかしながら Pk-匿名化データから所望の分析を行う再構築処理は, これまで多次元ヒストグラムを推定していたため, 多変量になるにつれ計算量が指数的に増える問題があった. そこで本研究では, 確率密度比を再構築するアプローチによりこれらの問題を克服する, 多変量解析のための新たな再構築法を提案する.

Efficient Reconstruction Algorithm for Multivariate Statistics

Satoshi Hasegawa†　　Koki Hamada†　　Ryo Kikuchi†

†NTT Secure Platform Laboratories.
3-9-11 Midori-cho, Musashino-shi, Tokyo 180-8585, JAPAN
{ hasegawa.satoshi, hamada.koki, kikuchi.ryo}@lab.ntt.co.jp

Abstract Anonymization technology has been researched as a technique of both protected for personal information and utilized big data. One of the anonymization technology is *Pk*-anonymity. *Pk*-anonymity consists of two methods as "perturbation" for protecting personal information and "reconstruction" for obtaining statistics. Existing reconstruction algorithm is exponentially increasing the computational complexity by the increase of attributes. We propose an efficient reconstruction algorhithm for multivariate statistics.

1　はじめに

近年のデータ分析基盤の高性能化やデータ分析技術の発達により, 大量のデータを分析し, 有益な知見を得ることが可能となった. それに伴い, 購買データや位置データといった個人に紐づく情報 (パーソナル情報) の利活用が注目を浴びている. しかしながらこのようなパーソナル情報は, 個人を特定可能な情報が含まれており, 安易に他者に提供 (第三者提供) などを行うと個人のプライバシを侵害するリスクが生じる恐れがある. こうしたプライバシ侵害のリスクを低減し, データの第三者提供を可能にする方法として匿名化技術が研究開発されている.

1.1　匿名化

匿名化とは, データベース中に含まれる個人のデータを加工し, 個人の特定を困難にすることでプライバシを保護する技術のことをいう. 主にデータベース保持者が第三者にデータを提供する際に個人のプライバシを保護する技術である.

匿名化したデータがどの程度プライバシを保護しているかを示す代表的な指標として, k-匿名性 [9] が提案されている. k-匿名性は, 「データベース中に同じ間接識別子 (QI)[1] を持つレコードが少なくとも k 個以上存在する」ことで個人のプライバシを保護する指標である.

1.1.1 k-匿名化とその問題点

k-匿名性を満たす匿名化の方法として, 属性の一般化による手法 [7, 6], クラスタリングによる手法 [2], レコード削除による手法が存在する. これらの手法はいずれにも問題がある.

一般化による手法は, 属性値がどのように一般化されるかを示す属性の階層構造を自ら定義せねばならず, 属性数が非常に多い場合は設定が困難となる. また性別といった属性値のとりうる値が少ないようなものは, 一般化により属性の値に意味を持たなくなってしまう. クラスタリングによる手法は, クラスタリングを行うため距離を適切に定義する必要がある. 数値属性の場合はユークリッド距離などを用いることで問題がないが, カテゴリ属性に対しては距離をどのように定義するか問題が生じる. レコード削除による手法は, 匿名性を満たすために, 極端に偏ったデータ削除が行われる可能性があり, データの傾向が大きく変わってしまう恐れがある.

1.1.2 Pk-匿名化 (撹乱)

これらの問題を解決できる方法として Pk-匿名化 [11] がある. Pk-匿名化は, Pk-匿名性 [11] と呼ばれる k-匿名性と同等の匿名性の評価が可能なランダム化手法による匿名化方法である. Pk-匿名性を満たすランダム化手法として, カテゴリ属性はランダムに値を入れ替える維持置換撹乱処理 [11, 1] を, 数値属性はランダムなノイズを付与する有界ラプラスノイズを付与する方法 [12] が存在する.

1.1.3 再構築とその問題点

Pk-匿名化は, 値が撹乱されていることから, そのままでは分析の正確性が損なわれてしまう. そこで, ランダム化されたデータから元データの統計値を推定する再構築というアプローチをとることで, できるだけ正確な統計分析を行う[2]. 再構築する統計値として, 元データの確率密度関数を推定する方法が主に研究されており [10, 12, 1, 13], 確率密度関数を多次元ヒストグラムで表現し推定する方法 [10, 12] および, 無相関化により属性ごとに 1 次元ヒストグラムを推定する手法 [13] が提案されている. 前者は属性数が増えるごとに, 多次元ヒストグラムのとりうる値が指数的に増えるため, たちまち計算が困難になる. 後者は, あらかじめ主成分分析でデータを無相関化する. それにより各変数は独立であると捉えることができ, 各属性ごとに独立に撹乱および再構築を行うことで, 前者の計算量の問題を回避できるアプローチである. しかし「無相関ならば変数間は独立」という仮定は, データが正規分布に従っている場合は成立するが, それ以外は一般に成り立つとは限らない. また両アプローチとも, 数値属性をヒストグラムで再構築するため, 必ず離散化を伴い情報を損なってしまう. それゆえ離散化を伴わず高次元データの確率密度関数を再構築する必要があるといえる.

1.2 本論文の貢献

元データの生成分布である確率密度関数を推定することは, データの生成規則を知ることゆえ, あらゆる統計分析に適用可能である. それゆえこれまでの再構築法では, 確率密度関数の推定を行ってきた. しかしながら, 元データの確率密度関数を推定すること, 特に高次元データの推定は一般に難しい問題として知られている [8]. そこで我々は, 元データの確率密度関数を推定せず, 元データと匿名化データの確率密度の比を推定するアプローチ [8] をとることを考える. この確率密度比により, これまでとほぼ同様の統計分析が可能である.

[1] 直接個人を識別する属性ではないが, 組み合わせることで個人を識別することができる属性のこと

[2] ランダム化処理 (撹乱処理) と再構築処理を合わせて撹乱再構築法という

我々の貢献は大きく 2 つある.

- 密度比推定を介した再構築により, 指数的に計算量が増える問題を改善. 多変量データの統計分析が可能
- 離散化を伴わずにカテゴリ属性と数値属性を同時に再構築することで精度向上

密度比の推定を行う形にすることで, 属性数に依存して計算量が指数的に増えない再構築アルゴリズムの構成が可能である. 本アルゴリズムにより, 既存の再構築法では困難であった属性数を多く使用する統計分析が可能となる. また密度比の推定を行う形にすることで, 離散化を伴わない数値属性[3] とカテゴリ属性の両方を同時に再構築することが可能となる. それにより, 再構築の精度向上が見込める.

本論文の構成を示す. 2 章では, 記法の定義および既存の Pk-匿名化法および再構築法の課題について示す. 3 章では, 再構築法の課題を解決する提案手法を示す. 4 章では, 提案手法を適用した多変量解析法を示す. 5 章では実験を行い, 6 章にまとめを示す.

2 準備

2.1 記法

記法の定義を行う. bold 体は列ベクトルを表し, 大文字の bold 体は行列を表す. 元データのデータ数を N, 属性数を M とする. 属性はカテゴリ属性と数値属性が混在して含まれており, j 番目の属性の取りうる値の集合を A_j と表す. そして全属性の集合を $A = A_1 \times \cdots \times A_M$ とする.

元データは確率密度 $P_X(\mathbf{x})$ を持つ確率分布に, i.i.d[4] 標本 $\{\mathbf{x}_i\}_{i=1}^N$ で与えられるとする ($\mathbf{x}_i \in A$). また匿名化データは, 条件付き確率 $P_{Y|X}$ に従ってデータがランダムに撹乱されており, 撹乱後の匿名化データを $\{\mathbf{y}_i\}_{i=1}^N$ とする. また $x_i^{(j)}$ は, i 番目のデータで, j 番目の属性値の値を表す.

[3] 著者らによって数値属性のみの場合は離散化を伴わずに再構築が可能であったが, カテゴリ属性が含まれる場合再構築ができなかった

[4] i.i.d = 独立同一分布

2.2 撹乱再構築法

撹乱再構築法とは, 元データを条件付き確率 $P_{Y|X}$ に従いランダムに変更を加えることでデータを秘匿化し (撹乱と呼ぶ), 秘匿化されたデータから統計値を得る (再構築と呼ぶ) ことにより, プライバシを保護したまま統計分析を行う手法のことである. 既存の撹乱手法として, カテゴリ属性を撹乱する維持置換撹乱 [11], 数値属性を撹乱する方法として有界ラプラスノイズ [12] があり, 撹乱されたデータから元データの確率密度関数を推定する方法として逐次ベイズ法 [10, 12] が提案されている.

2.2.1 撹乱

維持置換撹乱

カテゴリ属性を撹乱する方法として, 維持置換撹乱が提案されている [1]. 維持置換撹乱とは, 維持確率 ρ で属性値を維持し, $1-\rho$ の確率で属性値をランダムに変更することで, データを秘匿化する処理である. あるカテゴリ属性 A_j の属性値 $v \in A_j$ が $v' \in A_j$ に変わる確率 $P_{Y|X}^{A_j}(v'|v)$ は, 維持確率 ρ_j により,

$$P_{Y|X}^{A_j}(v'|v) = \begin{cases} \rho_j + \frac{1-\rho_j}{|A_j|} & (v' = v) \\ \frac{1-\rho_j}{|A_j|} & (v \neq v') \end{cases} \tag{1}$$

と表すことができる.

有界ラプラスノイズ付与

数値属性を撹乱する方法として, 有界ラプラスノイズ付与が提案されている [12]. 有界ラプラス分布とは, ラプラス分布の上限と下限が定まっている分布 (有界ラプラス分布) のことであり, この有界ラプラス分布に従う乱数を付与することで, データを秘匿化する. ある数値属性 A_j (値域が $[a_j, b_j], a_j \in \mathbb{R}, b_j \in \mathbb{R}$) の属性値 v が v' に変わる確率密度は, 有界ラプラス分布のパラメータ ϕ_j により,

$$P_{Y|X}^{A_j}(v'|v) = \frac{1}{\gamma_j(v)} \frac{1}{2\phi_j} \exp(-\frac{|v-v'|}{\phi_j}) \tag{2}$$

となる. ここで $\gamma_j(v)$ は以下である.

$$\gamma_j(v) = \int_{a_j-v}^{b_j-v} \frac{1}{2\phi_j} \exp(-\frac{|z|}{\phi_j}) dz \tag{3}$$

詳細ついては, 文献 [12] を参照されたい.

Pk-匿名性を満たすパラメータ決定法

ρ_j および ϕ_j は,「撹乱後のテーブルのある人のレコードを $1/k$ 以上に確信することができない」(Pk-匿名性) を満たすようにする. カテゴリ属性を $A_1, ..., A_l$, 数値属性を $A_{l+1}, ..., A_M$ とすると,

$$k = 1 + (N-1)\left(\prod_{1\le j\le l}\left(\frac{1-\rho_j}{1+(|A_j|-1)\rho_j}\right)^2\right) \prod_{l+1\le j\le M}\exp\left(-2\frac{|b_j-a_j|}{\phi_j}\right) \quad (4)$$

が成立するよう ρ_j, ϕ_j を決めることにより, 維持置換撹乱および有界ラプラスノイズ付与による撹乱データは, Pk-匿名性を満たす.

2.2.2 再構築

逐次ベイズ法

撹乱されたから元データの統計量を推定する方法として, 撹乱方法である $P_{Y|X}$ および撹乱データ $\{\mathbf{y}_i\}_{i=1}^N$ から, 元データの確率密度関数 P_X を推定する方法が提案されている [10]. 以下の対数尤度関数を最大化する (最尤推定法) ことにより, P_X を推定する.

$$\arg\max_{P_X} \int_{\mathbf{y}} P_Y(\mathbf{y}) \log\left(\int_{\mathbf{x}} P_{Y|X}(\mathbf{y}|\mathbf{x}) P_X(\mathbf{x})\right) d\mathbf{x} \quad (5)$$

Agrawal らや五十嵐らは, $P_{Y|X}$ に維持置換撹乱, P_Y として撹乱されたデータの度数分布を, P_X として多次元ヒストグラムを仮定したアルゴリズムを提案している. 式 (6) を解くことにより, 元データの確率密度関数の推定を行う [1].

$$\arg\max_{P_X} \sum_i P_Y(\mathbf{y}_i) \log\left(\sum_{\mathbf{x}} P_{Y|X}(\mathbf{y}_i|\mathbf{x}) P_X(\mathbf{x})\right) \quad (6)$$

式 (6) の解法として, EM アルゴリズムを適用した逐次ベイズ法を提案している.

五十嵐らは, 数値属性の撹乱データでも再構築ができるよう, $P_{Y|X}$ として有界ラプラスノイズ付与を用いて逐次ベイズ法を適用する手法を提案している [12]. 多次元ヒストグラムの推定となるため, 予め数値属性を離散化して適用する.

これらの具体的なアルゴリズムについては, 文献 [1, 12] を参照されたい.

既存の再構築法の問題点

これら多次元ヒストグラムを推定する手法は, 属性数が少ない場合には計算可能であるが, 属性数が増えるにつれ途端に計算が困難となる. なぜなら, 全属性をカテゴリ属性と仮定した際, 多次元ヒストグラムの空間計算量は $O(|A_1|\times\cdots\times|A_M|)$ となり, たちまち計算困難となるからである [5].

元データの生成分布である P_X の正確な推定は, 様々な分析に適用が可能であるゆえ有益である. しかしながら, 確率密度関数の正確な推定は, 一般的に難しい問題として知られており, 特に高次元になるほど推定が困難となる. そこで, 本稿では, 計算量, 推定精度の両観点からも困難である確率密度関数の推定を行わず, 様々な多変量解析に適用可能な確率密度比を推定するアプローチをとることで, 多変量解析を行うこととする.

3 提案手法

本稿では P_X を推定せず, P_X/P_Y という元データと撹乱データの確率密度の比を推定するアプローチをとる. この確率密度比の推定問題に置き換えることにより, 既存の再構築法での空間計算量の問題を解決することに加え, 数値属性の離散化を伴わないアルゴリズムを構築できる.

3.1 確率密度比の再構築

密度比を $w(\mathbf{z}) = P_X(\mathbf{z})/P_Y(\mathbf{z})$ と表す. 提案手法では, $P_X(\mathbf{z})$ および $P_Y(\mathbf{z})$ を求めずに, $w(\mathbf{z})$ を直接求めることを目標とする.

密度比を推定する際も, 同様に式 (5) を解く.

$$\arg\max_{\hat{P}_Y} \int_{\mathbf{y}} P_Y(\mathbf{y}) \log \hat{P}_Y(\mathbf{y}) d\mathbf{y} \quad (7)$$

ここで, $\hat{P}_Y(\mathbf{y}) = \int_{\mathbf{x}} P_{Y|X}(\mathbf{y}|\mathbf{x}) P_X(\mathbf{x}) d\mathbf{x}$ とした.

$\hat{P}_Y(\mathbf{y})$ は密度比モデル $w(\mathbf{z}) = P_X(\mathbf{z})/P_Y(\mathbf{z})$ を

[5] 数値属性が含まれる場合, 精度よく再構築するためには, 細かく離散化する必要があり, $|A_j|$ が大きくなることは想像できる

用いることにより, 以下のようになる.

$$\hat{P}_Y(\mathbf{y}) = \int_{\mathbf{x}} P_{Y|X}(\mathbf{y}|\mathbf{x})P_X(\mathbf{x})d\mathbf{x} \quad (8)$$

$$= \int_{\mathbf{x}} P_{Y|X}(\mathbf{y}|\mathbf{x})w(\mathbf{x})P_Y(\mathbf{x})d\mathbf{x} \quad (9)$$

$$\approx \frac{1}{N}\sum_i P_{Y|X}(\mathbf{y}|\mathbf{y}_i)w(\mathbf{y}_i) \quad (10)$$

ここで $P_X(\mathbf{z}) = w(\mathbf{z})P_Y(\mathbf{z})$ という関係, および式 (9) から式 (10) の近似は撹乱データの標本で近似していることを意味する.

式 (7) に, 式 (10) を適用し, 撹乱データの標本で近似することで以下になる [6].

$$\arg\max_{\hat{w}} \int_{\mathbf{y}} P_Y(\mathbf{y}) \log \sum_j P_{Y|X}(\mathbf{y}|\mathbf{y}_j)w(\mathbf{y}_j)d\mathbf{y}$$
$$\rightarrow \arg\max_{\hat{w}} \sum_i \log \sum_j P_{Y|X}(\mathbf{y}_i|\mathbf{y}_j)w(\mathbf{y}_j)$$

また, 最適化問題を解く際に, P_X に以下の制約を設ける.

$$\int_{\mathbf{x}} w(\mathbf{x})P_Y(\mathbf{x})d\mathbf{x} = 1$$
$$\rightarrow \frac{1}{N}\sum_i w(\mathbf{y}_i) = 1 \quad (11)$$

これらを用いて, 最終的に求めるべき最適化問題は以下である.

$$\arg\max_{\hat{w}} \sum_i \log \sum_j P_{Y|X}(\mathbf{y}_i|\mathbf{y}_j)w(\mathbf{y}_j)$$
$$\text{s.t.} \quad \frac{1}{N}\sum_i w(\mathbf{y}_i) = 1 \quad (12)$$

ここで, $P_{Y|X}$ は各属性ごとの条件付き確率の積であり, カテゴリ属性なら式 (1) を, 数値属性なら式 (2) を適用する.

$$P_{Y|X}(\mathbf{z}'|\mathbf{z}) = \prod_j P^{A_j}_{Y|X}(z'^{(j)}|z^{(j)}) \quad (13)$$

数値属性の場合, 式 (2) は確率密度であることから, 一般的には確率に変換するため積分計算を行う必要が生じる. しかし以下の定理により, 確率密度の値をそのまま用いることができる.

定理 1 条件付き確率密度 $P^{A_j}_{Y|X}(v'|v)$ が式 *(2)* で与えられているとき, ある区間 $[v'-\xi, v'+\xi]$ $(\xi > 0)$ で積分した結果は, 式 *(2)* の定数倍である.

定理の証明は付録に示す. 定数倍は, 式 (12) の結果に影響しない. 定理 1 より, 式 (2) をある区間で積分した結果は, 式 (2) の定数倍ゆえ式 (2) をそのまま用いて良い.

密度比モデルについては, 具体的にどのようなものかを議論してこなかった. そのため, 密度比モデル w をどのような形にするかは自由度がある. 例えばカーネル密度比モデル $w(\mathbf{z}_j) = \sum_i^N \boldsymbol{\alpha}^{(i)} k(\mathbf{z}_j, \mathbf{z}_i)$ や線形密度比モデル $w(\mathbf{z}_j) = \boldsymbol{\alpha}^T \mathbf{z}_j$ などが提案されている [8]. ここで $\boldsymbol{\alpha}$ はモデルパラメータと呼ばれるもので, 密度比モデルを求めることは, このパラメータを求めることと等価である. カーネル密度比モデルにある $k(\mathbf{z}_i, \mathbf{z}_j)$ は, [8] よりガウスカーネル $k(\mathbf{z}_i, \mathbf{z}_j; \sigma^2) = \exp\left(-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|_2^2}{\sigma^2}\right)$ を用いることとする [7].

カーネル密度比モデルは, 線形密度比モデルと比べ表現能力が高いことが利点である. しかしながら表現能力を高くするため, すべてのデータ同士の関係を用いることから計算コストが高い. それに比べ線形密度比モデルはモデルの表現能力が低いが, 計算コストが低いことが利点である.

3.2 アルゴリズム

本節では, まずカーネル密度比モデルを用いたアルゴリズムを示し, 後に線形密度比モデルを用いた手法を示す. 式 (12) に対し, カーネル密度比モデルを適用した形のものが以下となる.

$$\arg\max_{\boldsymbol{\alpha}} \sum_i^N \log \sum_j^N P_{Y|X}(\mathbf{y}_i|\mathbf{y}_j) \sum_r^N \boldsymbol{\alpha}^{(r)} k(\mathbf{y}_j, \mathbf{y}_r)$$
$$\text{s.t.} \quad \frac{1}{N}\sum_i^N \sum_r^N \boldsymbol{\alpha}^{(r)} k(\mathbf{y}_i, \mathbf{y}_r) = 1 \quad (14)$$

式 (14) は, 勾配法によるパラメータ更新および制約条件の充足により, 解くことが可能である. パラメータ α を求めるアルゴリズムをアル

[6] 定数倍は, arg max に影響しないため, 予め式から除いている

[7] ガウスカーネルの計算の際は, 予めカテゴリ変数はダミー変数化しておく必要がある

表 1: 再構築法の計算量の比較

	初期計算の時間計算量	ループの時間計算量	空間計算量
逐次ベイズ法 [10]	-	$O((\lvert A_1\rvert \times \cdots \times \lvert A_M\rvert)\log(\lvert A_1\rvert \times \cdots \times \lvert A_M\rvert))$	$O(\lvert A_1\rvert \times \cdots \times \lvert A_M\rvert)$
カーネル密度比再構築法	$O(N^2M)$	$O(N^2)$	$O(N^2)$
線形密度比再構築法	$O(N^2M)$	$O(M'N)$	$O(M'N)$

ゴリズム 1 に示す. ここで $\mathbf{P}$ は, i 行 j 列の要素を $P_{Y|X}(\mathbf{y}_i|\mathbf{y}_j)$ として持つ $N \times N$ 行列, $\mathbf{K}$ は i 行 j 列の要素を $k(\mathbf{y}_i, \mathbf{y}_j)$ として持つ $N \times N$ 行列である. また β は学習率と呼ばれる値で, rowSum() は行単位の総和を表す. 本アルゴズムは, $\mathbf{P}$ および $\mathbf{K}$ の計算を除けば, 空間計算量が $O(N^2)$ であり, 時間計算量は行列計算の順番を工夫することにより $O(N^2)$ で計算が可能である.

Algorithm 1 カーネル密度比モデルを用いた密度比再構築法 (カーネル密度比再構築法)

Require: $\mathbf{P}, \mathbf{K}, \beta, e$
Ensure: $\boldsymbol{\alpha}$
  Initialize $\boldsymbol{\alpha}^0$ の要素を 0 以上で初期化
  $\mathbf{b} \leftarrow \frac{1}{n}\mathrm{rowSum}(\mathbf{K})$
  for $t = 0, 1, 2...$ **do**
    $\boldsymbol{\alpha}^{t+1} \leftarrow \boldsymbol{\alpha}^{t+1} + \beta\mathbf{K}^T\mathbf{P}^T(\mathbf{1}./\mathbf{PK}\boldsymbol{\alpha}^t)$
    $\boldsymbol{\alpha}^{t+1} \leftarrow \boldsymbol{\alpha}^{t+1} + (1 - (\mathbf{b}^T\boldsymbol{\alpha}^{t+1})\mathbf{b}/(\mathbf{b}^T\mathbf{b}))$
    $\boldsymbol{\alpha}^{t+1} \leftarrow \max(\mathbf{0}, \boldsymbol{\alpha}^{t+1})$
    $\boldsymbol{\alpha}^{t+1} \leftarrow \boldsymbol{\alpha}^{t+1}/(\mathbf{b}^T\boldsymbol{\alpha}^{t+1})$
    if $\|\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^t\|^2 < e$ **then**
      $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}^{t+1}$
      break
    end if
  end for

線形密度比モデルによる高速な計算

カーネル密度比モデルは, $O(N^2)$ 必要となるため, N が増えると計算が困難となる. そこで, 表現能力を犠牲にしつつも, 計算が可能となる線形密度比モデルを用いた予測法を考える.

撹乱データ $\{\mathbf{y}\}_{i=1}^N$ のカテゴリ変数を予めダミー変数で置き換える. ダミー変数化した場合の属性数を M' とする. そして撹乱データを $N \times M'$ 行列 $\mathbf{Y}$ で表現する. アルゴリズム中の $\mathbf{K}$ を $\mathbf{Y}$ に置き換えることで, 線形密度比モデルを用いた推定が可能となる. 本アルゴリズムの場合, $\mathbf{P}$ および $\mathbf{K}$ の計算を除けば, 時間計算量および空間計算量ともに $O(NM')$ となる.

4 確率密度比の適用

確率密度比は, 統計で用いられる重点サンプリングで重要な役割を果たす. この重点サンプリングを行うことで, 尤度最大化に基づく統計分析手法や, 損失最小化に基づく機械学習手法での分析が可能となる.

4.1 重点サンプリング

重点サンプリングとは, ある関数 Q の P_X に関する期待値を, 別の確率密度 P_Y を用いて計算する統計的な手法のことをいう. 匿名化の場合は, 元データの確率密度 $P_X(z)$ を推定することが容易ではないため, 代わりに得られている $P_Y(z)$ のサンプルを用いて期待値を取る.

$$\begin{aligned} \mathrm{E}[Q(\mathbf{z})] &= \int Q(\mathbf{z})P_X(\mathbf{z})d\mathbf{z} \\ &= \int Q(\mathbf{z})\frac{P_X(\mathbf{z})}{P_Y(\mathbf{z})}P_Y(\mathbf{z})d\mathbf{z} \\ &\approx \sum_{i=1}^{N} Q(\mathbf{z}_i)\frac{P_X(\mathbf{z}_i)}{P_Y(\mathbf{z}_i)} \end{aligned} \tag{15}$$

$\frac{P_X(\mathbf{z}_i)}{P_Y(\mathbf{z}_i)}$ が密度比 $w(\mathbf{z}_i)$ である. Q を例えば,

$$Q(\mathbf{z}_i) = -(z_{y_i}\log\sigma(\boldsymbol{\gamma}^T\mathbf{z}_{x_i}) + (1-z_{y_i})\log(1-\sigma(\boldsymbol{\gamma}^T\mathbf{z}_{x_i}))) \tag{16}$$

とすると ($\mathbf{z}_i$ のうち, 説明変数を $\mathbf{z}_{x_i}$, 目的変数を z_{y_i} とする), ロジスティック回帰分析が可能となる. ほかにも重回帰分析や SVM といったアルゴリズムを適用可能である.

5 実験

数値実験により, 提案手法の有用性について示す. 既存の再構築アルゴリズムおよび k-匿名化アルゴリズムとの比較を行う. 実験に用いるデータセットとして, UCI Machine Learning Repository

表 2: 実験で用いるデータについて

	age	workclass	education num	martial status	occupation	relationship	sex	capital gain	capital loss	hours week	income
属性	カテゴリ	カテゴリ	数値	カテゴリ	カテゴリ	カテゴリ	カテゴリ	数値	数値	数値	カテゴリ
識別子	QI	QI	QI	QI	QI	QI	QI	QI	QI	QI	SA
変数	説明	説明	説明	説明	説明	説明	説明	説明	説明	説明	目的

の Adult Dataset[8] を用いる. Adult Dataset はレコード数 48842, 属性数 15 のデータセットで, 14 属性を用いて, 15 属性目の収入 (50K 以上か以下か) を予測する. 今回は 14 属性のうち, 予測にあまり寄与しない変数を予め削除し, 10 属性のデータに変換した (表 2 を参照). また, データ数はデータセット中の 1/3 を訓練データ (匿名化および多変量解析の対象) とし, 残り 2/3 をテストデータとし, 評価を行った. 匿名性は, $k = 3, 5, 10, 50$ と変化させて匿名化処理を行った.

評価方法

再構築結果の評価は, 多変量解析手法の 1 種であるロジスティック回帰分析の分析の正確性で測定した. 評価指標として, 2 値判別分析の評価で良く用いられる AUC[4] と呼ばれる指標を用いた. 評価指標として, 真陽性のみで評価する「正答率」もあるが, AUC では真陽性の多さに加え偽陽性ができるだけ少ないということも評価される. AUC の値が 0.5 のときランダムな予測で, 1.0 の場合予測の正確性が最も良いということを示す.

比較手法

比較手法として, k-匿名化手法との比較を行う. k-匿名化の方法として, 一般化による手法と削除による手法の 2 手法との比較を行う. k-匿名化を実現するツールとして, ARX Powerful Data Anonymization[3] を用い, また一般化階層の設計は [5] に従った [9].

結果

元データでの判別分析結果, 撹乱データでの判別分析結果, 再構築データでの判別分析結果, k-匿名化による結果 (一般化および削除) を図 1 に示す.

再構築の分析結果は, カーネル密度比再構築法を用い, カーネルにはガウスカーネルを用いている. カーネルの値 σ^2 は 1000 とした ([8] を参考に, 撹乱データのみを用いた交差検定を行いパラメータを決定した).

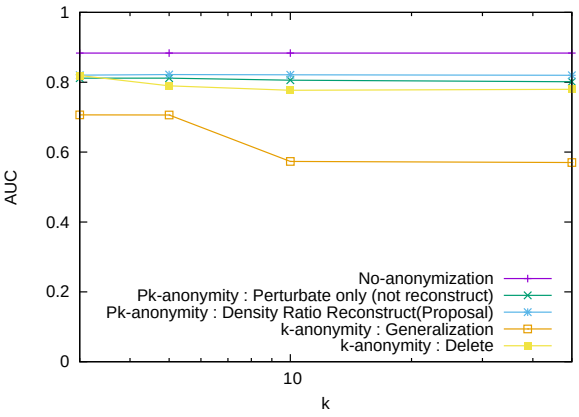

図 1: ロジスティック回帰分析による, 各匿名化手法との分析の比較実験

結果より, 提案法による分析結果が今回実験したすべての k において最も AUC が高いとわかる. $k = 3$ の結果の場合, Pk-匿名化による撹乱データおよび再構築データでの分析結果, 削除の k-匿名化による分析結果がほぼ同等で, 一般化による k-匿名化の結果が悪かった. 削除の k-匿名化の AUC が良かった理由は, $k = 50$ でもデータ数 16281 のうち 2 割のデータ削除ですんだことが大きいと考えられる. また Pk-匿名化では, 撹乱のみでの分析結果は k を上げることで, 徐々に AUC が下がっているが, 再構築での分析結果はほぼ一定であった. これは, k を上げたとしても, ランダム化量が極端に増えず, また再構築により正確性が上がっているからといえる.

6 まとめ

Pk-匿名化で多変量解析を行う際に問題となっていた空間計算量を改善した再構築アルゴリズムを提案した. 本手法により, 多変量解析に Pk-匿名化を適用可能となる. また実験により, 本提

[8] https://archive.ics.uci.edu/ml/datasets/Adult
[9] http://ddm.cs.sfu.ca/software.html

案手法を適用した分析結果が最も良くなるとわかった.

今後の課題として, 大規模データで適用可能な線形密度比モデルを用いたアルゴリズムでの実験および, 再構築での分析結果の理論的な評価などが挙げられる.

参考文献

[1] Rakesh Agrawal, Ramakrishnan Srikant, and Dilys Thomas. Privacy preserving olap. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 251–262, 2005.

[2] Ji-Won Byun, Ashish Kamra, Elisa Bertino, and Ninghui Li. Efficient k-anonymization using clustering techniques. In *Proceedings of the 12th International Conference on Database Systems for Advanced Applications*, pp. 188–200, 2007.

[3] Prasser Fabian, Kohlmayer Florian, Lautenschlaeger Ronald, and A. Kuhn Klaus. Arx a comprehensive tool for anonymizing biomedical data. In *AMIA Annual Symposium Proceedings*, pp. 984–993, 2014.

[4] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, Vol. 27, No. 8, pp. 861–874, 2006.

[5] Benjamin Fung, Ke Wang, and Philip S Yu. Top-down specialization for information and privacy preservation. In *Proceedings. 21st International Conference on Data Engineering*, pp. 205–216, 2005.

[6] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60, 2005.

[7] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*, pp. 25–25, 2006.

[8] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pp. 1433–1440, 2008.

[9] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570, 2002.

[10] 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. コンピュータセキュリティシンポジウム 2008, pp. 497–502, 2008.

[11] 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 2009, pp. 1–6, 2009.

[12] 五十嵐大, 長谷川聡, 納竜也, 菊池亮, 千田浩司. 数値属性に適用可能な, ランダム化により k-匿名性を保証するプライバシー保護クロス集計. コンピュータセキュリティシンポジウム 2012, pp. 639–646, 2012.

[13] 坂野鋭. 相関保存 pk 匿名化法. 暗号と情報セキュリティシンポジウム 2015, 2015.

A 定理 1 の証明

$P^{A_j}_{Y|X}(v'|v) = \frac{1}{\gamma_j(v)}\frac{1}{2\phi_j}\exp(-\frac{|v-v'|}{\phi_j})$ の条件付き確率を求めるため, v が与えられているとし, $[c-\xi, c+\xi]$ 区間で積分を行う.

$v < c-\xi$ のとき

$$\begin{aligned}&\int_{c-\xi}^{c+\xi}\frac{1}{2\phi_j\gamma_j(v)}\exp\left(-\frac{v'-v}{\phi_j}\right)dv'\\ =\ &\frac{1}{2\gamma_j(v)}\left(-\exp(-\frac{c+\xi-v}{\phi_j})+\exp(-\frac{c-\xi-v}{\phi_j})\right)\end{aligned}$$

これを条件付き確率密度 $\frac{1}{\gamma_j(v)}\frac{1}{2\phi_j}\exp(-\frac{v-c|}{\phi_j})$ で除算した結果が倍数となる. 除算した結果は以下となる.

$$-\phi\exp(-\xi/\phi)+\phi\exp(\xi/\phi) \tag{17}$$

$v > c+\xi$ のとき

紙面の都合上省略する. 符号がかわるのみで, 結果は式 (17) と同じとなる.

$c-\xi < v < c+\xi$ のとき

$$\int_{c-\xi}^{c+\xi}\frac{1}{2\phi_j\gamma_j(v)}\exp\left(-\frac{-v'+v}{\phi_j}\right)dv' \tag{18}$$

$$=\ \frac{1}{\gamma_j(v)}-\frac{1}{2\gamma_j(v)}\exp(\frac{c-v-\xi}{\phi})$$

$$-\ \frac{1}{2\gamma_j(v)}\exp(\frac{c-v+\xi}{\phi}) \tag{19}$$

同様に条件付き確率 $\frac{1}{\gamma_j(v)}\frac{1}{2\phi_j}\exp(0/\phi_j)$ で除算すると以下となる.

$$2\phi-2\phi(\exp(-\xi/\phi)) \tag{20}$$

式 (17) と式 (20) は一見すると等価であるとはわからない. しかし, 式 (20) を式 (17) の引き算により,

$$\begin{aligned}&2\phi-2\phi\exp(-\xi/\phi)-(-\phi\exp(-\xi/\phi)+\phi\exp(\xi/\phi))\\ =\ &2\phi-\phi(\exp(-\xi/\phi)+\exp(\xi/\phi))\end{aligned} \tag{21}$$

ξ/ϕ が小さければ (積分区間を狭ければ), $2\phi-2\phi$ はほぼ 0 となり, 等価とみなすことができる. また, 式 (17) は, v と c に依存しないため, 定数倍であるといえる.