Vol.2015-CSEC-69 No.27
Vol.2015-IOT-29 No.27
2015/5/22

有限混合モデルを用いた新たな再構築法

長谷川 聡[1,a)]　菊池 亮[1]　五十嵐 大[1]　濱田 浩気[1]　千田 浩司[1]

概要：近年, プライバシを保護しながら統計分析を行うことができる技術として, データを撹乱してプライバシを保護し, その後データの分布を推定して得る (再構築と呼ぶ), 撹乱再構築法が注目されている. 従来の撹乱再構築法では, 元データに対する一切の仮定をおかず, 元データの分布を推定することから, 精度良く再構築を行うためには大量のデータを必要としていた. しかしながら, 実際には再構築に十分なデータ数がない場合も多く, そのような際でも精度よく再構築したいニーズがある. そこで, 十分にデータがない場合でも精度よく再構築を行えるよう, 分布の推定によく用いられる有限混合モデルと呼ばれる確率分布を仮定した新たな再構築法を提案する.

1. はじめに

近年のデータ分析基盤およびデータ分析技術の発展により, 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情報といったパーソナル情報を含むデータの利活用が注目を浴びてきている. 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, 安易に利用してしまうと, プライバシを侵害するリスクが伴う. こうしたプライバシ侵害のリスクを低くすることでデータの利活用を可能にする方法として, 匿名化技術が盛んに研究されている.

1.1 匿名化

匿名化とは, データベース中のデータを加工し, 個人の特定を難しくすることでプライバシを保護する技術である. 主にデータを保有している人が, データを分析したい人に, プライバシを保護してデータを提供したい場面を想定した技術である. データ保有者が匿名化処理を施す際, どの程度プライバシが保護できているかを表す指標として, 様々なものが提案されており, それを満たすアルゴリズムも同様に提案されている.

代表的な指標として k-匿名性 [1] が提案されている. k-匿名性とは, データベース中に同じレコードが少なくとも k 個以上存在すればプライバシが保護されていると考える指標である. k-匿名性を満たす匿名化方法として [2] や [3] が提案されている. これらは一般化 (データの抽象化) により, k-匿名性を満たす方法であるため, 元データと加工後のデータの粒度が異なってしまう恐れがある. これに対し, データを抽象化せずに, k-匿名性と同等のプライバシ保護性能を満たす匿名化方法として, 撹乱再構築法を基とした Pk-匿名化 [4] が提案されている.

表 1 撹乱再構築のイメージ

元データ

身長	体重
170.5	80.4
164.3	60.2
158.4	74.3
⋮	⋮

撹乱データ

身長	体重
182.4	65.4
152.3	87.2
190.4	60.3
⋮	⋮

再構築データ

(クロス集計, 縦軸:身長, 横軸:体重)

	[41-50]	[51-60]	⋯
[141-150]	5	1	⋯
[151-160]	6	2	⋯
[161-170]	4	2	
⋮	⋮	⋮	⋱

Pk-匿名化の基となっている撹乱再構築法は, データを確率的に変更しプライバシを保護する撹乱処理, その後統計値のみ推定して得る再構築処理の 2 つからなる手法である. 撹乱再構築法のイメージを表 1 に示す. 撹乱処理では, 元の身長や体重データに対し, ランダムな数値を加算することで匿名化処理を行う. その後再構築処理では, 匿名化後のデータからノイズの影響をなるべく排除しながら, 所望の統計値の推定を行う (表 1 の場合, 再構築として 2 次元の度数表 (クロス集計表) を推定している). 撹乱再構築法では, 元データと同じ粒度での匿名化が可能となる.

撹乱処理の方法として, カテゴリ属性に対する維持置換撹乱 [5] や数値属性に対するラプラスノイズを付与する方

[1] 日本電信電話株式会社
Nippon Telegraph and Telephone Corporation
[a)] hasegawa.satoshi@lab.ntt.co.jp

法 [6] が提案されており, データの特徴に応じた様々な撹乱方法が提案されている [7], [8]. また再構築方法としてクロス集計表を得る方法 [4], [5], [9] が提案されている. 既存のこれらの再構築アルゴリズムは, 元データの分布を仮定しない手法であるゆえ, 精度よく再構築するためには, 多量のデータ数が必要となることが課題となっている.

1.2 貢献

本論文では, 再構築アルゴリズムに注目し, 少量のデータでも精度良く再構築を行う方法を提案する. 既存の再構築アルゴリズムは, 先ほども述べたとおり元データの分布を仮定しないため, 精度よく再構築するために多量のデータが必要となる. そこで, 従来より少ないデータ数でも精度よく再構築が行えるよう, データの生成分布を仮定する新たな再構築アルゴリズムを提案する. 特にデータとして数値属性 (撹乱方法としてラプラスノイズ付与 [6]) に特化したものを提案する.

提案する再構築アルゴリズムは, データの生成分布として, 有限混合モデルと呼ばれる有限個の確率分布の線形和で記述される分布を想定する. 有限混合モデルは, 加算する分布の数を増やすことでほぼ任意の確率分布を表現できることから表現能力が高く, また位置情報データの分布推定 [10] や分割最適化ベースのクラスタリング [11] などに用いられており, 応用でも良く用いられている確率分布である. 本提案手法では, 混合モデルの中でも最も用いられているガウス分布の線形和で記述される混合ガウス分布に注目する.

本論文で提案するアルゴリズムは以下の特徴を持つ.

- 既存手法より少数データでの再構築精度が良い
- 分割最適化ベースのクラスタリングを行える

本論文の構成を示す. まず 2 章では, 撹乱再構築法および混合モデルを導入し, 3 章では, 1 次元データおよび多次元データ (制約がある) の場合での提案手法を示す. その後 4 章では, 3 章で用いていた制約を取り除き一般的な場合に適用する手法を示す. 5 章では, 数値実験の結果を示し, 6 章で結論を述べる.

2. 準備

2.1 記法

基本的に, ベクトルを bold 体 $\mathbf{a}$ で記述し, スカラ値をイタリック体 a で記述し, ベクトルの j 番目の要素を $a^{(j)}$ と記述することにする.

撹乱前のデータを X, 撹乱後のデータを Y とおき, それぞれを確率変数として考える. 撹乱前データが確率分布 $p(X)$ から発生し, また撹乱後のデータが確率分布 $p(Y)$ から発生しているとする. またデータの撹乱を条件付き確率 $p(Y|X)$ で表す. 撹乱前および後のデータ数を N とする.

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 $p(Y|X)$ に従い確率的に変更を加える (撹乱と呼ぶ) ことでデータを秘匿し, 秘匿したデータから統計値を得る (再構築と呼ぶ) 処理によって, プライバシを保護したまま統計分析を行う手法である. まず, 本手法の撹乱方法として想定している五十嵐らの提案したラプラスノイズ付加 [6] を以下に示し, 後に再構築法について述べる.

2.2.1 ラプラスノイズ付加

数値属性に対する撹乱方法として, 五十嵐らはラプラスノイズを付与する方法を提案している [6]. ラプラスノイズ付加は, 平均 0, パラメータ ϕ の d 次元のラプラス分布

$$\mathcal{L}(\mathbf{x}; 0, \phi) = \frac{1}{(2\phi)^d} \exp\left(- \sum_{j=1}^{d} \frac{|x^{(j)} - 0|}{\phi} \right) \tag{1}$$

に従う乱数を付加したものである. これは撹乱後のデータ $\mathbf{y}$ の確率が, パラメータ ϕ, 平均 $\mathbf{x}$(元データ) に従うラプラス分布と解釈できることから, 撹乱処理である条件付き確率 $p(Y|X)$ は,

$$p(Y|X) = \mathcal{L}(\mathbf{y}; \mathbf{x}, \phi) = \frac{1}{(2\phi)^d} \exp\left(- \sum_{j=1}^{d} \frac{|y^{(j)} - x^{(j)}|}{\phi} \right) \tag{2}$$

と表される. パラメータ ϕ は五十嵐ら [6] の手法により決まるものとする.

2.2.2 再構築法

再構築法では, データの撹乱方法である条件付き確率 $p(Y|X)$ および, 撹乱後データの確率分布 $p(Y)$ が与えられているもとで, $p(X)$ を求めることが目標である.

Agrawal らは, $p(X)$ を求める方法として, 対数尤度を最大化する手法 (最尤推定法) を用いている [5]. 次式を解くことに相当する.

$$\begin{aligned} &\arg\max_{p(X)} \sum_{i=1}^{N} h(\mathbf{y}_i) \log\left(\sum_{\mathbf{x}} p(Y = \mathbf{y}_i | X = \mathbf{x}) p(X = \mathbf{x}) \right) \\ &\text{subject to} \sum_{\mathbf{x}} p(X = \mathbf{x}) = 1, 0 \le p(X = \mathbf{x}) \le 1 \end{aligned} \tag{3}$$

ここで, $\arg\max_{p(x)}$ は関数を最大にする $p(x)$ を求めること, また subject to は制約を表す. $h(\mathbf{y_i})$ は観測された撹乱済みデータ $\mathbf{y}_i$ の度数を表す. Agrawal らは, $p(Y|X)$ として遷移確率行列を用いた維持置換撹乱を, また $p(Y)$ の分布として多項分布を考え, その尤度が最大となる $p(X)$ を求めている. この尤度最大化問題を解く方法として, EM アルゴリズムを適用した Iterative Baysian Technique を提案している.

ラプラスノイズを付加する撹乱方法の再構築方法として, 五十嵐ら [6] は, ラプラスノイズを離散化し遷移確率行列を求め, Iterative Baysian Technique を適用する方法を提案している.

2.3 有限混合モデル

有限混合モデルとは, 有限個の確率分布の線形和で表さ

れるものである. k 番目の確率分布を $\mathcal{P}(x;\theta_k)$ で表すと,

$$\sum_{k=1}^{K}\rho_k\mathcal{P}(x;\theta_k)$$
$$\text{subject to}\forall k\ 0\le\rho_k\le 1,\sum_{k=1}^{K}\rho_k=1 \tag{4}$$

となる. ここで, ρ_k は混合比と呼ばれるものである.

$\mathcal{P}(x;\theta_k)$ として, ガウス分布やベルヌーイ分布などの指数分布族と呼ばれる部類の確率分布がよく用いられている [11].

2.3.1 混合ガウス分布

有限混合モデルとしてよく用いられるものとして, 混合ガウス分布モデルがある. 混合ガウス分布は, ガウス分布を複数加算したモデルである. ガウス分布を $\mathcal{N}(x_i;\mu,\sigma^2)$ とすると, 混合ガウス分布は以下のようになる.

$$p(x_i;\rho,\mu,\sigma^2)=\sum_{k=1}^{K}\rho_k\mathcal{N}(x_i;\mu_k,\sigma_k^2)$$
$$\text{subject to}\forall k\ 0\le\rho_k\le 1,\sum_{k=1}^{K}\rho_k=1 \tag{5}$$

混合ガウス分布モデルは, クラスタリングやデータの分布の推定によく用いられる手法である [11].

3. 有限混合モデルを仮定した再構築法

この章では, 本提案手法である, 有限混合モデルを仮定した再構築方法について示す. 特に, 混合ガウス分布モデルを用いた手法について示す. 本手法では, 撹乱方法である条件付き確率 $p(Y|X)$ として, ラプラスノイズ付加を, 元データの分布 $p(X)$ として, 混合ガウス分布を考え, 混合ガウス分布のパラメータ ρ,μ,σ を求める ($p(X)$ を求めることに相当) 方法を提案する.

まずは簡単化のため, 1 次元データに限った場合で議論し, 後に多次元データでの場合の手法について示す.

3.1 1 次元データの場合

まず, 1 次元データに限定した混合ガウスモデルを仮定した再構築処理法について示す. 元データが混合ガウスモデル $\sum_{k=1}^{K}\rho_k\mathcal{N}(x;\mu_k,\sigma_k^2)$ で生成されているものとし, 平均 0, パラメータ ϕ のラプラス分布 $\frac{1}{(2\phi)^d}\exp\left(-\frac{|x-0|}{\phi}\right)$ のノイズが加わっているものとする. 撹乱後のデータが従う確率分布の確率変数は, 元データの確率変数とラプラスノイズの確率変数の確率変数同士の和と考えられる. 確率変数同士の和は, 確率分布の畳み込みであることから, 混合ガウス分布とラプラス分布の畳み込み,

$$\begin{aligned}&g(y_i;\rho,\mu,\sigma^2)\\ =&\int_x\sum_{k=1}^{K}\rho_k\mathcal{N}(x;\mu_k,\sigma_k^2)\mathcal{L}(y_i;x,\phi)dx\\ =&\int_x\sum_{k=1}^{K}\rho_k\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x-\mu_k)^2}{\sigma_k^2}\right)\frac{1}{2\phi}\exp\left(-\frac{|y_i-x|}{\phi}\right)\\ =&\sum_{k=1}^{K}\rho_k\frac{1}{4\phi}\left(\exp\left(\frac{\sigma_k^2+2\phi\mu_k-2\phi y_i}{2\phi^2}\right)\mathrm{erfc}\left(\frac{\sigma_k^2+\phi\mu_k-\phi y_i}{\sqrt{2}\phi\sigma_k}\right)\right.\\ &+\left.\exp\left(\frac{\sigma_k^2-2\phi\mu_k+2\phi y_i}{2\phi^2}\right)\mathrm{erfc}\left(\frac{\sigma_k^2-\phi\mu_k+\phi y_i}{\sqrt{2}\phi\sigma_k}\right)\right)\\ =&\sum_{k=1}^{K}\rho_k f(y_i;\mu_k,\sigma_k^2,\phi).\end{aligned} \tag{6}$$

となる. ここで, $\mathrm{erfc}\,(x)$ 関数は,

$$\mathrm{erfc}\,(x)=\int_x^{\infty}\exp(-t^2)dt \tag{7}$$

である. また,

$$\begin{aligned}&f(y_i;\mu_k,\sigma_k^2,\phi)\\ =&\frac{1}{4\phi}\left(\exp\left(\frac{\sigma_k^2+2\phi\mu_k-2\phi y_i}{2\phi^2}\right)\mathrm{erfc}\left(\frac{\sigma_k^2+\phi\mu_k-\phi y_i}{\sqrt{2}\phi\sigma_k}\right)\right.\\ &+\left.\exp\left(\frac{\sigma_k^2-2\phi\mu_k+2\phi y_i}{2\phi^2}\right)\mathrm{erfc}\left(\frac{\sigma_k^2-\phi\mu_k+\phi y_i}{\sqrt{2}\phi\sigma_k}\right)\right)\end{aligned} \tag{8}$$

とする. この確率分布は, ϕ が小さいほど, ラプラスノイズの影響が少なく混合ガウス分布に近づき, ϕ が大きいほど, ラプラスノイズの影響が大きくなり混合ガウス分布からかけ離れたものとなる.

撹乱データ y_i はラプラスノイズが付与された混合ガウス分布 $g(y_i;\rho,\mu,\sigma^2)$ に従い生成されているものと考えられ, 観測済み撹乱データ y_i に尤もフィットする $g(y_i;\rho,\mu,\sigma^2)$ のパラメータ ρ,μ,σ^2 を求める問題を解くことで, 元データの確率分布を求めることができる. すなわち, 以下に示す対数尤度を最大とする問題を解くことで解を得ることができる.

$$\arg\max_{\rho,\mu,\sigma^2}\sum_{i}^{N}\log g(y_i;\rho,\mu,\sigma^2)$$
$$\text{subject to}\ \forall k\ \ 0\le\rho_k\le 1,\sum_{k=1}^{K}\rho_k=1 \tag{9}$$

この制約付き最大化問題は, 様々な方法を用いて解くことができる. 例えばニュートン法や最急勾配法, EM アルゴリズム [12] などを利用できる. ここでは, 大域解を得る保証はないが, 局所解を必ず得ることができる EM アルゴリズムを用いた手法を示す.

3.1.1 E ステップ, M ステップの導出

EM アルゴリズムは, 最尤推定を行うための汎用的な計算法である [11], [12], [13]. 混合ガウスモデルのパラメータ推定にも用いられている手法であり, 本手法も同様な形で適用する.

E ステップ E ステップでは, i 番目のデータが l 番目のラプラス分布を畳み込んだガウス分布 f にどれだけ属しているかを表す, 負担率 $\gamma(i,l)$ と呼ばれる値を計算する [11]. i 番目のデータが l 番目の f に属している割合を表すことから,

$$\gamma(i,l) = \frac{\rho_l f(y_i;\mu_l,\sigma_l^2,\phi)}{\sum_{k=1}^{K}\rho_k f(y_i;\mu_k,\sigma_k^2,\phi)} \tag{10}$$

となる.

M ステップ M ステップでは [11] と同様にパラメータ μ,σ,ρ を更新するステップとなる. しかし本手法では, ρ は解析的に求めることができるが, μ,σ については解析的に求めることができない. そこで, 一般化 EM アルゴリズムと呼ばれる, M ステップのパラメータ算出に勾配法 (パラメータの微分情報を用いて逐次的にパラメータを更新する手法) を用いた手法を適用する. それぞれのパラメータの算出方法は以下のようになる.

$$\begin{aligned}\mu_l^{\text{new}} &\leftarrow \mu_l^{\text{old}} + \alpha\frac{\partial L}{\partial \mu_l}\\ \sigma_l^{\text{new}} &\leftarrow \sigma_l^{\text{old}} + \alpha\frac{\partial L}{\partial \sigma_l}\\ \rho_l^{\text{new}} &\leftarrow \frac{N_l}{N}\end{aligned} \tag{11}$$

ここで, N_l は,

$$N_l = \sum_{i=1}^{N}\gamma(i,l) \tag{12}$$

とし, L を,

$$L = \sum_{i=1}^{N}\log g(y_i;\rho,\mu,\sigma^2) \tag{13}$$

とする. また, α は学習率と呼ばれるものであり, ユーザが適切な値を設定する. 一般的にはアルゴリズムが発散しない程度 (0.01 から 0.000001 程度) に設定することが多い. $\frac{\partial L}{\partial \mu_l}, \frac{\partial L}{\partial \sigma_l}$ はそれぞれ,

$$\begin{aligned}&\frac{\partial L}{\partial \mu_l}\\ &= \sum_{i=1}^{N}\frac{\rho_l}{4g_i\phi}\left(\frac{1}{\phi}\exp\left(\frac{\sigma_l^2+2\phi\mu_l-2\phi y_i}{2\phi^2}\right)\operatorname{erfc}\left(\frac{\sigma_l^2+\phi\mu_l-\phi y_i}{\sqrt{2}\phi\sigma_l}\right)\right.\\ &- \frac{2}{\sqrt{2\pi}\sigma_l}\exp\left(\frac{\sigma_l^2+2\phi\mu_l-2\phi y_i}{2\phi^2}\right)\exp\left(-\left(\frac{\sigma_l^2+\phi\mu_l-\phi y_i}{\sqrt{2}\phi\sigma_l}\right)^2\right)\\ &- \frac{1}{\phi}\exp\left(\frac{\sigma_l^2-2\phi\mu_l+2\phi y_i}{2\phi^2}\right)\operatorname{erfc}\left(\frac{\sigma_l^2-\phi\mu_l+\phi y_i}{\sqrt{2}\phi\sigma_l}\right)\\ &+ \left.\frac{2}{\sqrt{2\pi}\sigma_l}\exp\left(\frac{\sigma_l^2-2\phi\mu_l+2\phi y_i}{2\phi^2}\right)\exp\left(-\left(\frac{\sigma_l^2-\phi\mu_l+\phi y_i}{\sqrt{2}\phi\sigma_l}\right)^2\right)\right),\end{aligned}$$

$$\begin{aligned}&\frac{\partial L}{\partial \sigma_l}\\ &= \sum_{i=1}^{N}\frac{\rho_l}{4g_i\phi}\left(\frac{\sigma_l}{\phi^2}\exp\left(\frac{\sigma_l^2+2\phi\mu_l-2\phi y_i}{2\phi^2}\right)\operatorname{erfc}\left(\frac{\sigma_l^2+\phi\mu_l-\phi y_i}{\sqrt{2}\phi\sigma_l}\right)\right.\\ &- \frac{2}{\sqrt{2\pi}\phi}\exp\left(\frac{\sigma_l^2+2\phi\mu_l-2\phi y_i}{2\phi^2}\right)\exp\left(-\left(\frac{\sigma_l^2+\phi\mu_l-\phi y_i}{\sqrt{2}\phi\sigma_l}\right)^2\right)\\ &+ \frac{\sigma_l}{\phi^2}\exp\left(\frac{\sigma_l^2-2\phi\mu_l+2\phi y_i}{2\phi^2}\right)\operatorname{erfc}\left(\frac{\sigma_l^2-\phi\mu_l+\phi y_i}{\sqrt{2}\phi\sigma_l}\right)\\ &- \left.\frac{2}{\sqrt{2\pi}\phi}\exp\left(\frac{\sigma_l^2-2\phi\mu_l+2\phi y_i}{2\phi^2}\right)\exp\left(-\left(\frac{\sigma_l^2-\phi\mu_l+\phi y_i}{\sqrt{2}\phi\sigma_l}\right)^2\right)\right)\end{aligned}$$

ここで, $g_i = g(y_i;\rho,\mu,\sigma^2)$ とする.

3.1.2 アルゴリズムの全体

EM アルゴリズムを適用した解法をアルゴリズム 1 に示す. EM アルゴリズムは, さきほど述べたとおり局所解への収束は保証されているが, 大域解への収束は保証されていない. そのため, 初期値を適切に選ぶ必要がある.

Algorithm 1 1 次元データでのアルゴリズム

Input: $y_i\ (i=1,...,N), K, \alpha, \epsilon$
Output: $\rho_k,\mu_k,\sigma_k^2 (k=1,...,K)$

1: ρ_k,μ_k,σ_k^2 を適切に初期化する (ランダムに初期化など). ただし, ρ_k については, $\rho_k \geq 0, \sum_{k=1}^{K}\rho_k = 1$ を満たすように初期化すること.
2: 以下の E ステップと M ステップを, 収束条件を満たすまで交互に繰り返す.
3: [E ステップ] 負担率 $\gamma(i,l)$ を計算する.
4: [M ステップ] $l=1,...,K$ について以下を求め, パラメータ μ_k,σ_k,ρ_k を更新する.

$$\begin{aligned}\mu_l^{\text{new}} &\leftarrow \mu_l^{\text{old}} + \alpha\frac{\partial L}{\partial \mu_l}\\ \sigma_l^{\text{new}} &\leftarrow \sigma_l^{\text{old}} + \alpha\frac{\partial L}{\partial \sigma_l}\\ \rho_l^{\text{new}} &= N_l/N\end{aligned}$$

ここで N_k は,

$$N_l = \sum_{i=1}^{N}\gamma(i,l) \tag{14}$$

であり, L は,

$$L = \sum_{i=1}^{N}\log g(y_i,\rho,\mu,\sigma^2) \tag{15}$$

である.

5: [収束条件] $|L^{\text{old}} - L^{\text{new}}| < \epsilon$ なら収束とする.

3.1.3 特徴

従来, ラプラスノイズを撹乱として用いた場合の再構築アルゴリズムは, ラプラスノイズを一度離散化し遷移確率行列を求め, Iterative Baysian Technique の枠組みで再構築を行っていた. それに対し, 本手法は離散化を行わず連続値のまま取り扱うことが可能なため, 離散化による誤差が生じないメリットがある.

3.2 多次元の場合 (制約付き)

これまで, 1 次元データに限った混合ガウス分布を仮定した再構築方法について議論を行ってきた. しかしながら実際のデータは多次元であり, 1 次元の再構築方法だけでは不十分である. そこで, 本節では多次元への拡張を行う.

3.2.1 一般的な混合ガウス分布を考えた際の困難性

多次元データの一般的な場合を考える. すなわち, ラプラス分布および, 混合ガウス分布が式 16,17 のような場合を考える.

$$\frac{1}{2\phi^d}\exp\left(-\sum_{j=1}^{d}\frac{|y_i^{(j)}-x^{(j)}|}{\phi}\right) \tag{16}$$

$$\sum_{k=1}^{K}\rho_k \frac{1}{\sqrt{2\pi}^d\sqrt{|\mathbf{\Sigma}_k|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right) \quad (17)$$

ここで,$\mathbf{\Sigma}$ は共分散行列を表す.

一次元データの場合と同様に, g を求め EM アルゴリズムを適用しようとするが, 1 次元の場合と異なり混合ガウス分布に共分散行列が含まれているため, 解析的に積分計算を行うことが困難となる. そこで一旦, 共分散が 0 の多次元混合ガウス分布の場合を考えることとする.

3.2.2 共分散が 0 の混合ガウス分布の場合

共分散が 0 の場合の多次元混合ガウス分布は, 以下のようになる.

$$\sum_{k=1}^{K}\rho_k \frac{1}{\sqrt{2\pi}^d\prod_{j=1}^{d}(\sigma_k^{(j)})^2}\exp\left(-\sum_{j=1}^{d}\frac{(x^{(j)}-\mu_k^{(j)})^2}{2(\sigma_k^{(j)})^2}\right) \quad (18)$$

これより, 共分散が 0 と仮定した場合の $g(\mathbf{y}_i;\rho,\boldsymbol{\mu},\boldsymbol{\sigma}^2)$ は,

$$\begin{aligned}&\int_{x^{(1)}}\cdots\int_{x^{(d)}}\sum_{k=1}^{K}\rho_k\frac{1}{\sqrt{2\pi}^d\prod_{j=1}^{d}(\sigma_k^{(j)})^2}\exp\left(-\sum_{j=1}^{d}\frac{(x^{(j)}-\mu_k^{(j)})^2}{2(\sigma_k^{(j)})^2}\right)\\ &\cdot\ \frac{1}{2\phi^d}\exp\left(-\sum_{j=1}^{d}\frac{|y_i^{(j)}-x^{(j)}|}{\phi}\right)d_x^{(1)}\cdots d_x^{(d)}\\ &=\sum_{k=1}^{K}\rho_k\prod_{j=1}^{d}f(y_i^{(j)};\mu_k^{(j)},(\sigma_k^{(j)})^2,\phi) \quad (19)\end{aligned}$$

となる. これにより, 共分散が 0 という制約はあるが, 多次元への拡張が可能となる. 次の章では, これを用いて, 共分散が 0 という制約を取り除いたアルゴリズムを示す.

4. 制約のない多次元データへ適用できるアルゴリズム

3 章までは, 共分散が 0 という制約のある多次元混合ガウスモデルを用いた再構築法について取り扱ってきた. しかし実際のデータは相関があることから, このまま適用すると再構築精度が低い恐れがある. そこで, データを無相関 (共分散が 0) にすることで, 制約のない多次元データへの適用を可能とする方法を示す.

4.1 データの無相関化

データを無相関にする方法として, 主成分分析 [14] を用いた方法が代表的である. 本手法も同様に, 主成分分析を用いた手法を用いることとする.

4.1.1 主成分分析に基づく手法

d 次元ベクトルで表現されたデータ $\mathbf{z}_i$ を N 個考える. これを行列 $\mathbf{Z}=(\mathbf{z}_0,...,\mathbf{z}_{N-1})^T$ で表す. データを無相関化するような回転行列 $\mathbf{P}$ は, データの共分散行列 $\mathbf{S}=\mathrm{cov}(\mathbf{Z})$ を対角化する (すなわち, 共分散が 0) ことにより得ることができる.

$$\mathbf{P}^T\mathbf{S}\mathbf{P}=\mathbf{\Lambda} \quad (20)$$

このような $\mathbf{P}$ および $\mathbf{\Lambda}$ は, データの共分散行列 $\mathbf{S}$ を固有値分解することで得られる. これはまさに主成分分析そのものである. 無相関になるようデータを回転させるには, さきほど求めた $\mathbf{P}$ を用いて, $\mathbf{P}^T\mathbf{Z}$ とすれば良い.

4.2 多次元データでのアルゴリズム

データの無相関化を活用することで, 多次元データでも共分散が 0 である制約付きのアルゴリズムが適用可能となる. 多次元データにおける本手法のアルゴリズムをアルゴリズム 2 に示す.

Algorithm 2 多次元データでのアルゴリズム (データの無相関化を利用)

Input: $\mathbf{y}_i\ (i=1,...,N), K, \alpha, \epsilon$

Output: $\rho_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 (k=1,...,K)$

1: $\rho_k, \boldsymbol{\mu}_k, \sigma_k^2$ をランダムに初期化する. ただし, ρ_k については, $\rho_k \le 0, \sum_{k=1}^{K}\rho_k = 1$ を満たすようにランダムに初期化すること.

2: 以下の E ステップと M ステップを, 収束条件を満たすまで交互に繰り返す.

3: [E ステップ-0] 負担率 $\gamma(i,l)$ を計算する.

4: [E ステップ-1] K 個のデータ集合 $(\mathbb{D}_1,...,\mathbb{D}_K)$ を作成する. このデータ集合は以下のような規則で生成する.

- i 番目のデータ $\mathbf{y}_i$ を, $k=1,...,K$ の中で負担率 $\gamma(i,k)$ の中で最も高い k のデータ集合 $\mathbb{D}_k$ に含める.
- これをすべての i について行い, K 個のデータ集合を生成する.

5: [E ステップ-2] K 個のデータ集合にそれぞれ対し, 主成分分析を行う (データが無相関になるよう回転する) ことで, 無相関な K 個のデータ集合を生成する. $\boldsymbol{\sigma}_l$ の要素を対角要素として並べた対角行列 $\mathbf{\Sigma}$ および, $\boldsymbol{\mu}_l$ にも同様に主成分分析を行い, 無相関化する.

6: [M ステップ-0] $l=1,...,K$ について以下を求め, パラメータ $\boldsymbol{\mu},\boldsymbol{\sigma},\rho$ を更新する.

$$\begin{aligned}\boldsymbol{\mu}_l^{\mathrm{new}} &\leftarrow \boldsymbol{\mu}_l^{\mathrm{old}} + \alpha\frac{\partial L}{\partial\boldsymbol{\mu}_l}\\ \boldsymbol{\sigma}_l^{\mathrm{new}} &\leftarrow \boldsymbol{\sigma}_l^{\mathrm{old}} + \alpha\frac{\partial L}{\partial\boldsymbol{\sigma}_l}\\ \rho_l^{\mathrm{new}} &= N_l/N\end{aligned}$$

ここで N_k は,

$$N_l=\sum_{i=1}^{N}\gamma(i,l) \quad (21)$$

であり, L は,

$$L=\sum_{i=1}^{N}\log g(y_i,\rho,\boldsymbol{\mu},\boldsymbol{\sigma}^2) \quad (22)$$

である.

7: [M ステップ-1] $\boldsymbol{\sigma}_l^{new}$ の要素を対角要素として並べた対角行列 $\mathbf{\Sigma}^{new}$ および, $\boldsymbol{\mu}_l^{new}$ に対し, E ステップ-2 で行った回転とは逆の回転を施す.

8: [収束条件] $|L^{\mathrm{old}}-L^{\mathrm{new}}|<\epsilon$ なら収束.

4.3 クラスタリングとしての解釈

混合ガウス分布モデルは, クラスタリング手法としてもよく用いられている手法である [11]. 本提案手法は, 撹乱前データが混合ガウス分布にしたがって生成されていると仮

Vol.2015-CSEC-69 No.27
Vol.2015-IOT-29 No.27
2015/5/22

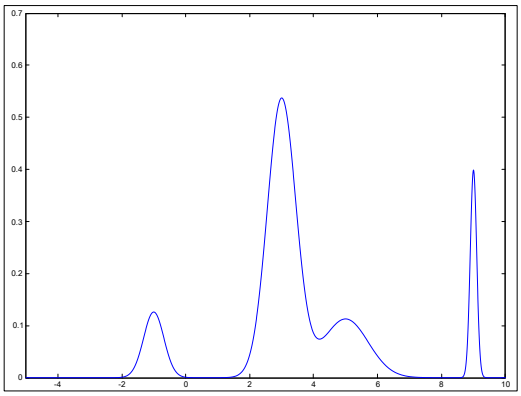

図 1　1 次元混合ガウス分布の確率密度関数, パラメータは, $\mu = (-1, 3, 5, 9)^T$,$\rho = (0.1, 0.6, 0.2, 0.1)^T$, $\sigma^2 = (0.1, 0.2, 0.5, 0.01)^T$ である

定したもとでの推定法であるゆえ, 推定した結果をクラスタリングの結果として用いることができる. すなわち, 本手法は元データの確率分布を求めると同時に, クラスタリングを行うことが可能な再構築方法であるともいえる.

5. 数値実験

この章では, 人工データおよびベンチマークデータを用いたデータの分布推定精度の比較実験を行う. 比較対象は, Agrawal らの再構築法を数値属性に拡張した五十嵐ら [6] の手法である. 提案手法では, 連続値の値が求まるが離散化し評価を行う. 評価方法は, 元データとの再構築データとの L1 距離が最大だと 0, 最小だと 100 になるような評価値 (L1 精度と以降は呼ぶ) を用いて評価を行う.

5.1 人工データによる実験

5.1.1 設定

ここでは, 人工的に混合ガウス分布から生成したデータに対しラプラスノイズを付与したデータに対する, 再構築結果の比較を行う. 人工データは, 以下の様な仕様の 1 次元混合ガウス分布から, 乱数を発生させた.

- $\mu = (-1, 3, 5, 9)^T$
- $\rho = (0.1, 0.6, 0.2, 0.1)^T$
- $\sigma^2 = (0.1, 0.2, 0.5, 0.01)^T$

この混合ガウス分布の密度関数を図 1 に示す. データ数は, $N = 1000, 10000, 100000, 1000000$ を用意し, それぞれに k-匿名性の $k = 2$ となるラプラスノイズ (ϕ =3.18, 2.45, 2.03, 1.75) を付与し, 再構築処理を行う実験を行った. また, 今回は K の値を変化 $2, ..., 6$ まで変化させ実験を行った. また $\alpha = 0.001$, $\epsilon = 0.001$ とした.

5.1.2 結果

実験結果を図 2 に示す. データ数が 1,000 や 10,000 の場合, 本手法のほうが L1 精度が良く, 100,000 や 1000,000 の場合, 既存手法の方が L1 精度が良かった. 提案手法は, EM アルゴリズムを用いているゆえ局所解に陥る可能性がある

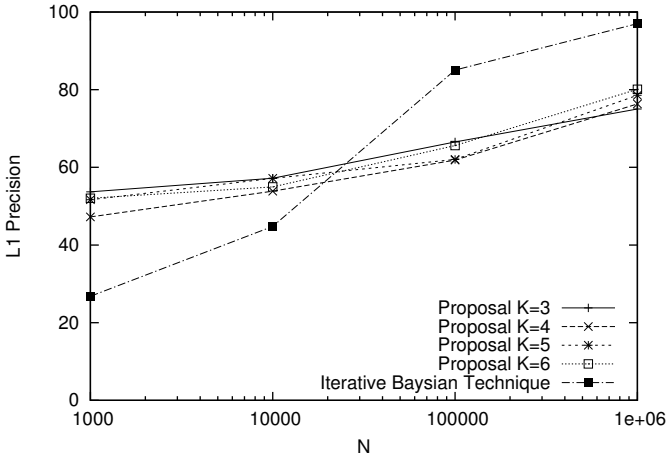

図 2　K を変化させた場合の 1 次元データに対する再構築結果. 横軸がデータ数, 縦軸が L1 精度を表している.

表 2　Adult Dataset での L1 精度の比較

	age	capital-gain	education-num
Iterative Baysian Technique	86.02	91.60	58.20
提案手法	88.03	91.67	62.76

ことから, 特にデータ数が多い場合での精度向上に結びつきにくかったと考えられる. データ数が少ない場合, 元データの分布が混合ガウス分布に従う乱数であったことから, 提案手法で精度良く再構築できたと考えられる.

5.2 ベンチマークデータ

5.2.1 設定

ベンチマークデータとして, UCI Machine Learning Repository 内にある, Adult Dataset[15] を用いて精度比較実験を行った. Adult Dataset はデータ数 32561, 属性数 15 のデータセットで, 今回は age, capital-gain, education-num の 3 つの属性に対し, それぞれ 1 属性ずつ再構築結果の精度比較を行った. ベンチマークでも同様に, k-匿名性の $k = 2$ となるラプラスノイズを (age は $\phi = 14.05$,capital-gain は $\phi = 19247.53$, education-num は $\phi = 2.88$) 付与し, パラメータは $\alpha = 0.0001$,$\epsilon = 0.001$ とした.　またこの実験では先ほどと異なり, $K = 1, ..., 5$ と変化させ, その中でもっとも L が高い K を採用した.

5.2.2 結果

実験結果を表 2 に示す. capital-gain については, ほぼ同様の結果であるが, age や education-num の両方とも,　提案手法の方が精度よく再構築できていることがわかる. 実際のデータの分布は, 混合ガウス分布の形で近似できる場合が多く, 今回のデータセットでも, そのような特徴があったことから, 再構築精度がよかったと考えられる.

6. おわりに

本論文では, 元データの生成分布として混合ガウス分布を, 撹乱方法としてラプラスノイズを仮定したもとでの再構築アルゴリズムを提案した. 1 次元データでの場合および, 多次元データの場合でのアルゴリズムを示し, また 1 次元データでの実験を行った. その結果, データの分布をある程度仮定できる場合, 本手法のほうが優位であることがわかった.

今後の課題として, 他のベンチマークデータでの実験, 特に多次元データでの実験が挙げられる. 多次元拡張の際に, 共分散を求める代わりにデータを無相関化することが, 実際にどれくらい精度に影響するかを評価する必要がある. また, EM アルゴリズムであるゆえ, 局所解への収束保証はあるが, 大域解への保証がない. 初期値の適切な設定方法や, 大域解を得ることができる MCMC[11] の適用などは今後の課題である.

参考文献

[1] Sweeney, L.: k-anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570 (2002).
[2] LeFevre, K., DeWitt, D. J. and Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity, *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, ACM, pp. 49–60 (2005).
[3] LeFevre, K., DeWitt, D. J. and Ramakrishnan, R.: Mondrian multidimensional k-anonymity, *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, IEEE, pp. 25–25 (2006).
[4] 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 論文集，Vol. 2009, pp. 1–6 (2009).
[5] Agrawal, R., Srikant, R. and Thomas, D.: Privacy preserving OLAP, *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, ACM, pp. 251–262 (2005).
[6] 五十嵐大，千田浩司，高橋克巳：数値属性における, k-匿名性を満たすランダム化手法，コンピュータセキュリティシンポジウム 2011 論文集，Vol. 2011, No. 3, pp. 450–455 (2011).
[7] 齋藤恆和，五十嵐大， 菊池亮，廣田啓一，正木彰伍：属性間の相関を考慮した撹乱再構築法の提案，コンピュータセキュリティシンポジウム 2014 論文集，Vol. 2014, No. 2, pp. 1042–1049 (2014).
[8] 柿澤美穂，渡辺知恵美， 古川諒， 高橋翼：属性値のグループ分類を用いた Pk-匿名化手法の検討，コンピュータセキュリティシンポジウム 2014 論文集，Vol. 2014, No. 2, pp. 1050–1056 (2014).
[9] 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 論文集，Vol. 2008, No. 8, pp. 497–502 (2008).
[10] Lichman, M. and Smyth, P.: Modeling human location data with mixtures of kernel densities, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 35–44 (2014).
[11] Bishop, C. M. et al.: *Pattern recognition and machine learning*, Vol. 4, No. 4, springer New York (2006).
[12] Dempster, A. P., Laird, N. M. and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38 (1977).
[13] Bilmes, J. A. et al.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, *International Computer Science Institute*, Vol. 4, No. 510, p. 126 (1998).
[14] Jolliffe, I.: *Principal component analysis*, Wiley Online Library (2002).
[15] Lichman, M.: UCI Machine Learning Repository (2013).