H-004

確率密度推定を用いた RDSP 法によるクラスタの階層構造の調査

Examination of the Hierarchical Structure of Clusters by Using the RDSP Method with Probability Density Estimation

斧城 悠大 † Onoshiro Yuta　**岩田 一貴** † Iwata Kazunori　**末松 伸朗** † Suematsu Nobuo　**林 朗** † Hayashi Akira

1 はじめに

RDSP 法は階層的クラスタリング手法の一つであり，標本の母集団の確率密度関数が混合分布であるときに有効であることが，最近の研究で示されている [1]．本論文では，確率密度推定を用いた RDSP 法を提案し，母集団の確率密度関数が混合分布に比較的近いと思われる気象統計データに対して階層的クラスタリングを行う．実験には，どのクラスタから発生しているかが既知である気象統計データを用い，各クラスタの階層構造を調査する．実験結果より，既存の手法と比較した場合の RDSP 法の優位性を確認し，確率密度推定における多変量カーネル関数のパラメータ調整についての考察を行う．

2 確率密度推定を用いた RDSP 法

標本の母集団が図 1 のように混合分布に従う場合を考える．混合分布では，各離散時間ステップにおいて，ある確率分布 ω により一つの部分母集団が選ばれ，選ばれた部分母集団の確率密度関数に従って標本が生成される．各部分母集団に番号をつけ，その番号をラベル番号と呼ぶ．d 次元ユークリッド空間 $\mathbb{R}^d$ 上の時間ステップ $i \in \mathbb{N}$ における確率変数を X_i，ラベル番号の全体集合を $\mathcal{L} \triangleq \{1,\ldots,M\}$ とする．標本空間 $\mathbb{R}^d$ 上の部分母集団 $m \in \mathcal{L}$ の確率密度関数を P_m，確率密度関数の集合を $\mathcal{P}(\mathcal{L}) \triangleq \{P_m | m \in \mathcal{L}\}$ とする．また，標本 $x \in \mathbb{R}^d$ が部分母集団 $m \in \mathcal{L}$ の確率分布に従うことを $x \sim P_m$ と表すと，部分母集団の選択確率 ω は $\omega(m) \triangleq \Pr(X_i \sim P_m)$ と表記できる．以上のような母集団において，同じ部分母集団から生成された標本の集合（クラスタ）の階層的クラスタリングを考える．このとき，クラスタ間の階層構造を適当に測るための非類似度として，次の RDSS と RDSP が提案されている [1].

定義 1 (RDSS[1]) 任意の部分集合 $\underline{\mathcal{L}} \subseteq \mathcal{L}$ に対して，各標本 $(x_1,\ldots,x_n) \in \mathbb{R}^{dn}$ が $\mathcal{P}(\underline{\mathcal{L}})$ の確率分布のいずれかに従うものとすると，複数の標本間の RDSS は以下のように定義される．

$$\left\{rds_{\mathcal{P}(\underline{\mathcal{L}})}(x_1,\ldots,x_n)\right\}^2 \triangleq \sum_{m\in\underline{\mathcal{L}}} \left| \sum_{i=1}^{n} I_{x_i\sim P_m} \log \frac{P_m(x_i)}{Q_{\underline{\mathcal{L}}}(x_i)} \right|. \quad (1)$$

ただし，I_C は条件 C が真のとき 1，それ以外のときは 0 となる指示関数であり，$Q_{\underline{\mathcal{L}}}$ は次のように定義される．

$$Q_{\underline{\mathcal{L}}}(x) \triangleq \sum_{m\in\underline{\mathcal{L}}} \lambda_{\underline{\mathcal{L}}}(m) P_m(x)\,,\ \lambda_{\underline{\mathcal{L}}}(m) \triangleq \frac{\omega(m)}{\sum_{m\in\underline{\mathcal{L}}}\omega(m)}. \quad (2)$$

定義 2 (RDSP[1]) 任意の部分集合 $\underline{\mathcal{L}} \subseteq \mathcal{L}$ に対して，複数の

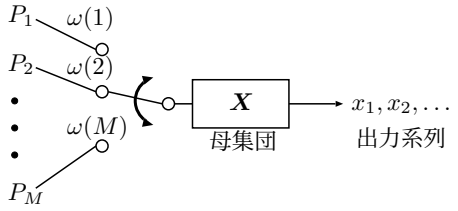


図 1　混合分布に従う母集団

確率密度関数 $\mathcal{P}(\underline{\mathcal{L}})$ 間の RDSP は以下のように定義される．

$$\{RDS(\mathcal{P}(\underline{\mathcal{L}}))\}^2 \triangleq \sum_{m\in\underline{\mathcal{L}}} \lambda_{\underline{\mathcal{L}}}(m) E_{P_m}\left[\log \frac{P_m(x)}{Q_{\underline{\mathcal{L}}}(x)}\right]. \quad (3)$$

任意の部分集合 $\underline{\mathcal{L}} \subseteq \mathcal{L}$ に対して，RDSS と RDSP の間には次のような漸近的関係が成立する [1].

$$\frac{1}{n}\left\{rds_{\mathcal{P}(\underline{\mathcal{L}})}(x_1,\ldots,x_n)\right\}^2 \to \{RDS(\mathcal{P}(\underline{\mathcal{L}}))\}^2 \text{ as } n\to\infty. \quad (4)$$

階層的クラスタリングにおいて，RDSP を用いてクラスタ間の距離を測る手法を RDSP 法と呼ぶ．RDSP の値を実際に計算する際には，期待値に関する積分計算を避けるために，RDSS を使って式 (4) の左辺で近似する．本論文では，RDSS を計算するときに必要な確率密度関数 P_m を確率密度推定により計算する．すなわち，標本 $(x_1,...,x_n) \in \mathbb{R}^{dn}$ を使って P_m を

$$\hat{P}_m(x;\mathbf{H}_m) = \frac{1}{n_m}\sum_{i=1}^{n} I_{x_i\sim P_m} |\mathbf{H}_m|^{-\frac{1}{2}} K\left(\mathbf{H}_m^{-\frac{1}{2}}(x-x_i)\right), \quad (5)$$

により推定する．ただし，$n_m \triangleq \left|\left\{x_i\in\mathbb{R}^d | x_i\sim P_m, i=1,...,n\right\}\right|$，$K$ は多変量カーネル関数，$\mathbf{H}_m$ はカーネル関数のパラメータで $d\times d$ 正定値実対称行列である．確率密度推定の精度は主にそのパラメータの選択によって決まることが知られている．そこで，Zhang ら [2] が提案したマルコフ連鎖モンテカルロ法に基づくパラメータの選択方法を使う．具体的には，$\mathbf{H}_m$ を $\mathbf{H}_m = \left(\mathbf{B}_m^{-1}\right)\left(\mathbf{B}_m^{-1}\right)^T$ と下三角行列 $\mathbf{B}_m^{-1}$ により Cholesky 分解し，式 (5) を $\mathbf{B}_m^{-1}$ について書き直した一つ抜き推定量 $\hat{P}_{m,i}(x_i;\mathbf{B}_m^{-1})$ の対数尤度 L_m，

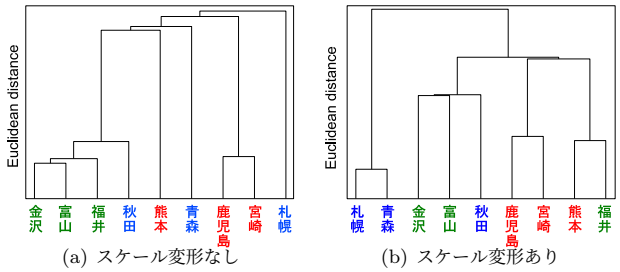
$$L_m(x_1,...,x_n;\mathbf{B}_m^{-1}) \triangleq \sum_{i=1}^{n} \log \hat{P}_{m,i}(x_i;\mathbf{B}_m^{-1}), \quad (6)$$

を最大化する $\left(\mathbf{B}_m^{-1}\right)^*$ を求めることで，最適な $\mathbf{H}_m^*$ を決定する．ここで，$\hat{P}_{m,i}$ は，$i=1,...,n$ に対して

$$\hat{P}_{m,i}(x_i;\mathbf{B}_m^{-1}) \triangleq \frac{1}{n_m-1}\sum_{j=1}^{n} I_{j\neq i} I_{x_j\sim P_m} |\mathbf{B}_m| K(\mathbf{B}_m(x_i-x_j)), \quad (7)$$

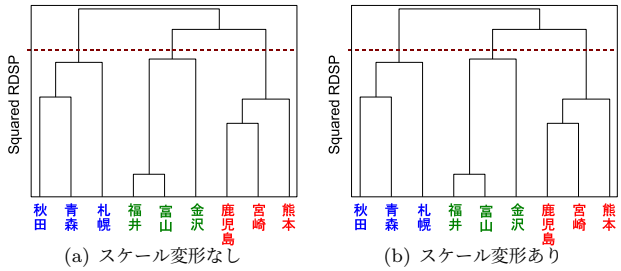
で与えられる．$\left(\mathbf{B}_m^{-1}\right)^*$ は次の事後分布 β から Metropolis-Hastings（M-H）アルゴリズムによってサンプリングされた $\mathbf{B}_m^{-1}$ のエルゴード平均値で近似される．

† 広島市立大学大学院情報科学研究科　〒731-3194 広島市安佐南区大塚東 3–4–1
Email: onoshiro@robotics.im.hiroshima-cu.ac.jp



(a) スケール変形なし　(b) スケール変形あり

図 2　最短距離法の結果（デンドログラム）



(a) スケール変形なし　(b) スケール変形あり

図 4　RDSP 法の結果（デンドログラム）

$$\beta(\mathbf{B}_m^{-1}|x_1,...,x_n)=\prod_{k=1}^{d}\prod_{l=1}^{k}\alpha(b_{kl};\lambda)\times\prod_{i=1}^{n}\hat{P}_{m,i}(x_i;\mathbf{B}_m^{-1}). \quad (8)$$

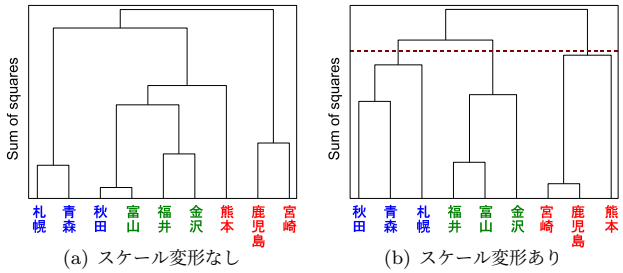
ただし，α は事前分布を表し，$\alpha(b_{kl};\lambda)=1/(1+\lambda b_{kl}^2)$ とした．ここで，b_{kl} は $\mathbf{B}_m^{-1}$ の (k,l) 要素，λ は事前分布のハイパーパラメータである．

3　実験

実データを用いた階層的クラスタリングにおける RDSP 法の優位性を検証するために，気象統計データ [4] に対して階層的クラスタリングを行った．使用するデータの特徴は，各都市における 1961 年–2000 年の間の 1 月–12 月の平均気圧，平均気温，降水量とした（つまり，$d=3$）．また，クラスタリングを行う都市は，気温が低く降水量があまり多くない地域である “札幌，青森，秋田”，気温が高く年間を通じて降水量が多い地域である “鹿児島，熊本，宮崎”，気温が低く冬の降水量が多い地域である “福井，金沢，富山” の 3 地域 9 都市とした．各都市における 40 年分のデータを同じ月ごとに分類した標本集合をクラスタとし，階層的クラスタリングによって各月ごとに 9 都市のクラスタ間の階層構造を調べた．クラスタリングに用いた手法は最短距離法，Ward 法 [3]，RDSP 法の 3 つで，各手法の成功率を比較した．ここでの成功とは，階層的クラスタリングの結果が各地域ごとに分類されたということである（例えば図 4 参照）．RDSP 法では，前節で述べたように RDSS における各クラスタの確率密度関数を M-H アルゴリズムにより求めた．ただし，多変量カーネル関数は 3 次元標準正規分布，M-H アルゴリズムにおける反復回数を 10000 回，burn-in 期間を 2500 回，$\mathbf{B}$ の初期値を単位行列，$\lambda=1$，proposal 分布を 6 次元正規分布とした．proposal 分布の共分散行列は，サンプリングの受理確率を 0.2–0.3 になるように調節した．Ward 法では，各変量の分散が

表 1　階層的クラスタリングの成功率

クラスタリング手法	スケール変形なし	スケール変形あり
最短距離法	33.3 %	33.3 %
Ward 法	16.7 %	58.3 %
RDSP 法	83.3 %	83.3 %



(a) スケール変形なし　(b) スケール変形あり

図 3　Ward 法の結果（デンドログラム）

大きく異なると階層構造を正しく認識できない場合があるため，スケール変形した標本 $\tilde{x}_i=\mathbf{S}_d^{-\frac{1}{2}}x_i$ に対しても階層的クラスタリングを行った．ここで，$\mathbf{S}_d$ は各変量の標本分散を対角要素に持つ対角行列である．確率密度推定において，スケール変形後の標本における最適な $\tilde{\mathbf{H}}_m^*$ は，スケール変形前の標本から推定した $\mathbf{H}_m^*$ を用いて $\tilde{\mathbf{H}}_m^*=\mathbf{S}_d^{-\frac{1}{2}}\mathbf{H}_m^*\mathbf{S}_d^{-\frac{1}{2}}$ と計算した．

階層的クラスタリングの成功率を表 1，各手法の特徴が現れている 4 月のデータのデンドログラムを図 2–4 に示す．初めに，最短距離法との比較から Ward 法と RDSP 法の有効性を確認する．最短距離法は図 2(a) のような鎖効果を引き起こす場合が多かったが，Ward 法と RDSP 法では鎖効果がほとんど起こらなかった．これは 2 つの手法が鎖効果を引き起こしにくいことを表す．次に，Ward 法と RDSP 法の成功率の比較から 2 つの手法の特徴を確認する．表 1 から標本のスケール変形の有無にかかわらず RDSP 法は Ward 法より高い成功率を示したことがわかる．これは，Ward 法はクラスタ間の非類似度にクラスタの平均と平方和のみしか使わないのに対し [3]，RDSP 法はクラスタ間の非類似度にクラスタの確率密度関数そのものを用いるからである（式 (3) 参照）．また，Ward 法は標本のスケール変形の有無によってクラスタリング結果が異なった（図 3 参照）．これは，各クラスタの平均と分散が変わることによって，クラスタ間の非類似度の大小関係が変化したからである．一方，RDSP 法は標本のスケール変形の有無によらず同じ結果となった（図 4 参照）．これは，RDSP がスケール変形の影響を受けないことを示している．最後に，確率密度推定における $\mathbf{H}$ の選択が RDSP 法に与える影響を考察する．$\mathbf{H}$ を単位行列とした場合，スケール変形した標本に対する RDSP 法の成功率は 58.3 %であった．よって，Zhang らの手法で選択した $\mathbf{H}$ を用いたほうが良い結果となった．これは，多変量カーネル関数に関する $\mathbf{H}$ の選択が比較的有効であったことを示している．

4　まとめ

気象統計データに対する階層的クラスタリングにおいて，確率密度推定を用いた RDSP 法の優位性を示した．また，多変量カーネル関数のパラメータ調整について考察した．

参考文献

[1] K. Iwata and A. Hayashi, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in press.

[2] X. Zhang, M. L. King and R. J. Hyndman, *Computational Statistics and Data Analysis*, vol. 50, pp. 3009–3031, 2006.

[3] J. H. Ward, *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[4] 気象庁, http://www.jma.go.jp/jma/index.html.