密度差推定に基づく分布クラスタリング手法の比較†

西原 詳*・中島 智晴*

本論文では、分布を対象とするクラスタリング手法の性能を調査する。2つの分布間の距離として L^2 距離を考え、 L^2 距離を2つの分布の密度差推定により求める。密度差推定の方法として、カーネル密度推定によりそれぞれの分布の密度関数を推定した後にその差を計算する2段階推定法と、最小二乗密度差推定法による直接推定法の2つを考え、それぞれの場合のクラスタリング性能を数値実験により調査する。

キーワード：クラスタリング、分布間距離、カーネル密度推定、最小二乗密度差推定

1. はじめに

クラスタリングは、データ集合を分割する教師なし学習手法の1つである。一般的なクラスタリングの対象は多次元空間内におけるデータ点の集合である。しかし、例えば、ある製品に対する多数のユーザによる評価のように、1つの対象を表すものが集合もしくは分布で表現されることがある。このような集合が多数ある場合にクラスタリングを適用しようとする、処理の対象は点ではなく集合もしくは分布になる。分布の集合をクラスタリングするためには、分布間の非類似度を定義する必要がある。Ngai ら [1] は分布間の非類似度として、クラスタ中心を用いて距離の期待値を求める手法を提案した。この手法では、距離を求める各分布の中心が同じであるとき、距離の期待値だと分布を区別できない。Kriegel ら [2] は密度ベースのクラスタリングである DBSCAN を拡張した手法を提案した。この手法では、分布の重なりが大きい場合の計算結果に大きなずれが生じるという問題がある。本研究では、このような分布が多数存在する状況下において分布の集合を適切にグループ分割する処理を考える。これを分布クラスタリングと呼ぶ。分布クラスタリングの概要を図1に示す。図1において、5つの分布が2つのクラスタに分割されている。クラスタリングを実行するためには、対象間の距離が定義されている必要がある。つまり、分布クラスタリングを実行するためには、異なる2つの分布間の距離が定義されなければならない。分布間の距離を調べる方法としては、KL-Divergence がよく知られているが、KL-Divergence で得られる値は対称性が成立しないなど、厳密な意味における距離ではない。数学的な距離の性質を満たすものとして L^2 距離がある。

本研究では、分布間の L^2 距離を推定する方法として2段階推定法と直接推定法の2つを考え、それぞれの場合におけるクラスタリングの性能を調査する。クラスタリング法として、

図1 分布クラスタリング

k -medoids 法 [4] を採用する。数値実験では、様々な正規分布から得られた分布の集合を用いてクラスタリングを行い、クラスタリングの結果を評価する。

2. 分布クラスタリング

分布クラスタリングは処理の対象が分布（もしくは何らかの集合）である以外は、点情報を処理対象とする一般的なクラスタリングと同じである。本研究では、クラスタリング法として k -medoids 法 [4] を用い、分布の集合をグループ分割する。また、生成されたクラスタの代表として medoid を用いることにする。medoid とは、クラスタ内の他の分布までの距離の総和が最小になる分布である。 k -medoids 法の手順を以下に示す。

手順1： k 個の初期 medoid を選択

手順2： 各分布を最近傍 medoid が所属するクラスタに割当て

手順3： 手順2の結果に基づいて各クラスタの medoid を更新

手順4： medoid に変更があれば手順2へ。そうでなければ終了

手順1における初期 medoid の選択方法は任意である。本研究では、KKZ 法 [5] を用いて初期の medoid を選択する。KKZ 法では、まず、最も離れている分布のペアを2つの medoid とする。次に、クラスタ数 k と選択した medoid の数が等しくなるまで、現在の medoid から最も離れた分布を新規に medoid として追加する。

クラスタリングを実行するためには、任意の2分布間に対する距離尺度が定義されていなければならない。本研究では、分布間の距離尺度として、密度差推定に基づく L^2 距離を使う。密度差推定の方法として、2段階推定法と直接推定法の2つを考える。これらについては3章で説明する。

† Comparative Study of Distribution Clustering Methods Based on Density-Difference Estimation

Syo NISHIHARA and Tomoharu NAKASHIMA

* 大阪府立大学大学院 人間社会システム科学研究科
Graduate School of Humanities and Sustainable System, Osaka Prefecture University

最適クラスタ数を決定するための評価基準として、式 (1) のような medoid 用に改良した Calinski-Harabasz 基準 [6] を用いる。この基準が最大となるクラスタ数を最適クラスタ数とする。

$$CH_k = \frac{\sum_{i=1}^k n_i L^2(p_{m_i}, p_m)}{\sum_{i=1}^k \sum_{p_j \in C_i} L^2(p_j, p_{m_i})} \times \frac{(N-k)}{(k-1)} \quad (1)$$

ここで、 N はデータ数、 k はクラスタ数、 $C_i (i = 1, \dots, k)$ はクラスタ、 p_{m_i} はクラスタ i 内の medoid、 p_m はデータ全体の medoid、 n_i はクラスタ i 内のデータ数を表す。

3. 密度差推定に基づく L^2 距離の算出

分布間距離として、分布間の密度差推定を用いた L^2 距離を用いる。いま、 d 次元空間において式 (2)、式 (3) のような 2 つの集合 (これ以降分布と呼ぶ) があるとする。

$$X = \{\mathbf{x}_i\}_{i=1}^n \stackrel{i.i.d.}{\sim} p(\mathbf{x}) \quad (2)$$

$$X' = \{\mathbf{x}'_j\}_{j=1}^m \stackrel{i.i.d.}{\sim} p'(\mathbf{x}) \quad (3)$$

ここで、確率密度関数 $p(\mathbf{x})$ 、 $p'(\mathbf{x})$ は知ることができないとする。確率密度関数 $p(\mathbf{x})$ と $p'(\mathbf{x})$ の距離を式 (4) で定義される L^2 距離により表す。

$$L^2(p(\mathbf{x}), p'(\mathbf{x})) = \int (p(\mathbf{x}) - p'(\mathbf{x}))^2 d\mathbf{x} \quad (4)$$

L^2 距離は、非負性、非退化性、対称性、三角不等式といった距離尺度の条件を満たすことが知られている。厳密に正確な L^2 距離を求めることは困難であるが、2 つの分布 X と X' を用いて推定することができる。本研究では、 L^2 距離を推定する方法として、2 段階推定法と直接推定法の 2 つを考える。

3.1 2 段階推定法

2 段階推定法では、まずカーネル密度推定により $p(\mathbf{x})$ 、 $p'(\mathbf{x})$ を推定し、次に、それらを使って L^2 距離を推定する。カーネル密度推定 [7] は、有限の標本から母集団の確率密度関数を推定するノンパラメトリック手法である。カーネル密度推定による確率密度関数はそれぞれ式 (5)、式 (6) のように定義される。

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K(\mathbf{x} - \mathbf{x}_i) \quad (5)$$

$$\hat{p}'(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^m K(\mathbf{x} - \mathbf{x}'_j) \quad (6)$$

ここで、 K はカーネル関数を表す。本研究では、カーネル関数として d 次元のガウシアンカーネルを用いる。

分布 X 、 X' の密度推定 $\hat{p}(\mathbf{x})$ 、 $\hat{p}'(\mathbf{x})$ を用いて、分布間の L^2 距離の推定を式 (7) により求める。

$$\hat{L}^2(p(\mathbf{x}), p'(\mathbf{x})) = \int (\hat{p}(\mathbf{x}) - \hat{p}'(\mathbf{x}))^2 d\mathbf{x} \quad (7)$$

3.2 直接推定法

2 段階推定法では、 $p(\mathbf{x})$ 、 $p'(\mathbf{x})$ それぞれの推定誤差が小さくても、密度差を推定する際に大きな誤差が生じる可能性がある。本節では、個々の確率密度関数を推定することなく密度差を直接推定する最小二乗密度差推定法 [3] により密度差を求める直接推定法を説明する。最小二乗密度差推定法では、密度差モデル $g(\mathbf{x})$ をできるだけ密度差に近づけようとする。具体的には、式 (8) のように二乗誤差が最小になることを目的とする。

$$\text{Minimize} \int \{g(\mathbf{x}) - (p(\mathbf{x}) - p'(\mathbf{x}))\}^2 d\mathbf{x} \quad (8)$$

密度差モデル $g(\mathbf{x})$ として式 (9) のようなガウシアンカーネルモデルを用いる。

$$g(\mathbf{x}) = \sum_{l=1}^{n+m} \theta_l \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_l\|^2}{2\sigma^2}\right) \quad (9)$$

ここで、 $(\mathbf{c}_1, \dots, \mathbf{c}_n, \mathbf{c}_{n+1}, \dots, \mathbf{c}_{n+m}) = (\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{x}'_1, \dots, \mathbf{x}'_m)$ はガウシアンカーネルの中心、 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{n+m})$ はモデルパラメータ、 σ はカーネル幅を表す。この時、最適なパラメータ $\boldsymbol{\theta}^*$ は式 (10) により求められる。

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg \min_{\boldsymbol{\theta}} \int \{g(\mathbf{x}) - (p(\mathbf{x}) - p'(\mathbf{x}))\}^2 d\mathbf{x} \\ &= \arg \min_{\boldsymbol{\theta}} [\boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta} - 2\mathbf{h}^T \boldsymbol{\theta}] \\ &= \mathbf{H}^{-1} \mathbf{h} \end{aligned} \quad (10)$$

ここで、 $\mathbf{H}$ は $(n+m) \times (n+m)$ 行列、 $\mathbf{h}$ は $(n+m)$ 次元ベクトルであり、それぞれ式 (11)、式 (12) のように定義される。

$$\begin{aligned} H_{l,l'} &= \int \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_l\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_{l'}\|^2}{2\sigma^2}\right) d\mathbf{x} \\ &= (\pi\sigma^2)^{\frac{d}{2}} \exp\left(-\frac{\|\mathbf{c}_l - \mathbf{c}_{l'}\|^2}{4\sigma^2}\right) \end{aligned} \quad (11)$$

$$\begin{aligned} h_l &= \int \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_l\|^2}{2\sigma^2}\right) p(\mathbf{x}) d\mathbf{x} \\ &\quad - \int \exp\left(-\frac{\|\mathbf{x}' - \mathbf{c}_l\|^2}{2\sigma^2}\right) p'(\mathbf{x}') d\mathbf{x}' \end{aligned} \quad (12)$$

目的関数の $\mathbf{h}$ を $\hat{\mathbf{h}}$ に置き換え、 l_2 正則化項を加えることにより、ガウシアンカーネルモデルのパラメータ決定は式 (13) の最適化問題に帰着される。

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} [\boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta} - 2\hat{\mathbf{h}}^T \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}] \quad (13)$$

ここで、 $\lambda (\geq 0)$ は正則化パラメータ、 $\hat{\mathbf{h}}$ は $(n+m)$ 次元ベクトルであり式 (14) のように定義される。

$$\begin{aligned} \hat{h}_l &= \frac{1}{n} \sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{c}_l\|^2}{2\sigma^2}\right) \\ &\quad - \frac{1}{m} \sum_{j=1}^m \exp\left(-\frac{\|\mathbf{x}'_j - \mathbf{c}_l\|^2}{2\sigma^2}\right) \end{aligned} \quad (14)$$

式 (13) より、最適化問題の解は式 (15) となる。

$$\hat{\boldsymbol{\theta}} = (\mathbf{H} + \lambda \mathbf{I})^{-1} \hat{\mathbf{h}} \quad (15)$$

この時、密度差の推定 $\hat{g}(\mathbf{x})$ は式 (16) で求められる。

表 1 実験に用いる正規分布

	平均	分散共分散行列
(1)	$N((0, 0), I)$	$10I$
(2)	$N((100, 100), I)$	$10I$
(3)	$N((-100, -100), I)$	$10I$
(4)	$N((0, 0), I)$	$100I$

$$\hat{g}(\mathbf{x}) = \sum_{l=1}^{n+m} \hat{\theta}_l \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_l\|^2}{2\sigma^2}\right) \quad (16)$$

最小二乗密度差推定法の精度は、カーネル幅 σ と正則化パラメータ λ に依存する。そこで、分布 X と X' を用いた交差検証によりカーネル幅と正則化パラメータを最適化する。詳細な交差検証の方法は本論文では省略する。

直接法による分布間の L^2 距離はバイアスを減らすために式 (17) のように推定する。

$$\hat{L}^2(p(\mathbf{x}), p'(\mathbf{x})) = 2\hat{\mathbf{h}}^T \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^T \mathbf{H} \hat{\boldsymbol{\theta}} \quad (17)$$

4. 数値実験

4.1 実験データ

3 章で述べた分布クラスタリング法の性能を数値実験により調べる。数値実験では、2 次元分布の集合に対するクラスタリング結果から各手法の性能を調べる。表 1 に示す 4 種類の正規分布を決められた実験設定にしたがって生成することで、クラスタリングの対象となる分布の集合を作成する。ここで、 $N(\mu, \Sigma)$ は平均 μ 、分散共分散行列 Σ の正規分布、 I は単位行列を表す。表 1 に示す正規分布を用いて、(a)~(h) に示す分布の集合からなる実験用データセットを乱数を用いて生成する。

- (a) (1) から生成した 1000 個のデータ点からなる分布 10 個と (2) から生成した 1000 個のデータ点からなる分布 10 個を要素とする分布の集合
- (b) (1) から生成した 1000 個のデータ点からなる分布 10 個と (2) から生成した 1000 個のデータ点からなる分布 10 個と (3) から生成した 1000 個のデータ点からなる分布 10 個を要素とする分布の集合
- (c) (1) から生成した 1000 個のデータ点からなる分布 10 個と (4) から生成した 1000 個のデータ点からなる分布 10 個を要素とする分布の集合
- (d) (1) から生成した 1000 個のデータ点からなる分布 10 個と (1) から生成した 500 個のデータ点、(2) から生成した 500 個のデータ点からなる分布 10 個を要素とする分布の集合
- (e) (2) から生成した 1000 個のデータ点からなる分布 10 個と (1) から生成した 500 個のデータ点、(2) から生成した 500 個のデータ点からなる分布 10 個を要素とする分布の集合
- (f) (1) から生成した 1000 個のデータ点からなる分布 10 個と (1) から生成した 100 個のデータ点、(2) から生成した 900 個のデータ点からなる分布 10 個を要素とする分布の集合
- (g) (1) から生成した 1000 個のデータ点からなる分布 10 個と (1) から生成した 500 個のデータ点、(4) から生成した 500 個のデータ点からなる分布 10 個を要素とする分布の集合

表 2 2 段階推定法を用いた分布クラスタリング結果

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
正解クラスタ数	2	3	2	2	2	2	2	2
最適クラスタ数	2	3	2	2	2	2	2	2
純度	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

表 3 直接推定法を用いた分布クラスタリング結果

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
正解クラスタ数	2	3	2	2	2	2	2	2
最適クラスタ数	2	3	2	3	2	2	2	2
純度	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00

- (h) (2) から生成した 1000 個のデータ点からなる分布 10 個と (1) から生成した 500 個のデータ点、(4) から生成した 500 個のデータ点からなる分布 10 個を要素とする分布の集合

4.2 評価手法

クラスタリングの評価を、推定した最適クラスタ数と正解クラスタ数との一致回数と純度により行う。純度は結果クラスタと正解クラスタの重なり具合を表し、値が大きいほどクラスタリング結果が良いと考えられる。純度は式 (18) で求められる。

$$\text{purity} = \frac{\sum_i \max_j n_{ij}}{\sum_{ij} n_{ij}} \quad (18)$$

ここで、 i は割り当てクラスタのインデックス、 j は正解クラスタのインデックス、 n_{ij} は正解クラスタ j からクラスタ i に割り当てられたデータ数である。

2 段階推定法による分布クラスタリング結果を表 2 に、直接推定法による分布クラスタリング結果を表 3 に示す。表 2 より、2 段階推定法の場合、(a)~(h) のすべてにおいて正解クラスタ数と最適クラスタ数が一致しており、純度も 1.00 であった。表 3 より直接推定法の場合、(d) 以外において正解クラスタ数と最適クラスタ数が一致しており、純度も 0.95 以上であった。このことから、2 段階推定法および直接推定法を用いた分布クラスタリングにより正解クラスタがほぼ求められていることがわかる。

5. おわりに

本論文では、分布クラスタリング手法の性能を調査した。2 つの分布間の距離を L^2 距離と定義し、 L^2 距離を 2 段階推定法、直接推定法により求める方法を示した。クラスタリング手法として k -medoids 法を適用した。数値実験では、それぞれの推定法において、集合から最適クラスタ数を推定し、正解クラスタ数と比較した。実験結果より、分布クラスタリングが有効に働いていることが分かった。

カーネル密度推定には、一般にデータ点の次元数が高くなるにつれて推定精度が悪くなるという性質がある。本論文では 2 次元データのみについて数値実験を行ったが、今後の課題として、多次元分布の集合に対する分布クラスタリング手法の性能を調査する必要がある。

参 考 文 献

- [1] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip, "Efficient Clustering of Uncertain Data," *Proceedings of the Sixth International Conference on Data Mining (ICDM)*, pp. 436–445, 2006.
- [2] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD)*, pp. 672–677, 2005.
- [3] M. Sugiyama, T. Kanamori, T. Suzuki, M. C. du Plessis, S. Liu, and I. Takeuchi, "Density-Difference Estimation," *Neural Computation*, Vol. 25, No. 10, pp. 2734–2775, 2013.
- [4] L. Kaufman and P. J. Rousseeuw, "Clustering by Means of Medoids," in *Statistical Data Analysis Based on the L_1 -Norm and Related Methods*, Y. Dodge (ed.), North-Holland, pp. 405–416, 1987.
- [5] I. Katsavounidis, C. C. J. Kuo, and Z. Zhang, "A New Initialization Technique for Generalized Lloyd Iteration," *IEEE Signal Processing Letters*, Vol. 1, Issue 10, pp. 144–146, 1994.
- [6] T. Calinski and J. Harabasz, "A Denrite Method for Cluster Analysis," *Communications in Statistics*, Vol. 3, No. 1, pp. 1–27, 1974.
- [7] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

(2017 年 4 月 4 日 受付)
(2017 年 7 月 27 日 採録)

[問い合わせ先]

〒599-8531 大阪府堺市中区学園町 1 番 1 号
大阪府立大学大学院人間社会システム科学研究科
中島 智晴
TEL: 072-254-9351
E-mail: tomoharu.nakashima@kis.osakafu-u.ac.jp

— 著 者 紹 介 —

にしはら しょう [非会員]

2016 年大阪府立大学大学 現代システム科学域卒業。現在、同大学大学院人間社会システム科学研究科在学中。センサを利用した高齢者見守りシステムの研究に従事。

なかしま ともはる [正会員]

1995 年大阪府立大学工学部卒業。2000 年同大学院博士後期課程修了。博士 (工学)。2000 年 4 月大阪府立大学大学院 助手。現在、同大学院 教授。2004 年度英国ノッティンガムトレント大学、2011 年度英国ノッティンガム大学 客員研究員。機械学習とファジィの融合、RoboCup、ケアサイエンス等の研究に従事。2017 年より日本知能情報ファジィ学会理事。

Comparative Study of Distribution Clustering Methods Based on Density-Difference Estimation

by

Syo NISHIHARA and Tomoharu NAKASHIMA

Abstract:

This paper investigates the performance of distribution clustering techniques. We use L^2 distance as a distance metric between two distributions. This distance is obtained by estimating the difference between the estimators of the two distributions. As the estimation methods of the density difference, we consider two methods. One is two-step method where first each of the two density functions is estimated by kernel method, and then calculate the difference between them. The other is direct method where the least-squares density-difference estimation method is used. The clustering performance with each method is investigated through a series of computational experiments.

Keywords: clustering, distance between distributions, kernel density estimation, least-square density-difference

Contact Address: Tomoharu NAKASHIMA

Department of Sustainable System Science, Graduate School of Humanities and Sustainable System, Osaka Prefecture University

1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan

TEL: +81-72-254-9351

E-mail: tomoharu.nakashima@kis.osakafu-u.ac.jp