データ構造に対して頑健なクラスタリング手法の開発

代表研究者　　濵　砂　幸　裕　　近畿大学　理工学部情報学科　講師

1 はじめに

情報通信技術の飛躍的発展により、自然科学や社会科学など様々な分野において、多種多様な大量のデータが蓄積されている。また、蓄積されたビッグデータを活用したデータマイニングの重要性は極めて強く認識されている[1]。例えば、EC サイトにおける購買履歴の分析による商品の推薦、GPS を利用した行動履歴の分析による人流の予測、SNS 上のコミュニティ抽出によるトレンド予測など、その応用範囲は多岐にわたる。情報通信機器の性能向上や低コスト化が著しい現代において、データマイニングの重要性は今後益々進展すると予想され、その中核をなすデータ解析手法の重要性は高まるばかりである。

データ解析手法の中でも、クラスタリングはデータの構造や規則性を明らかにする手法として知られており、データの前処理から分析まで幅広く用いられる手法として知られている[2]。クラスタリングとは、対象とするデータをクラスタと呼ばれる複数のデータの集まりに分割する手法である。その際に、同じクラスタに含まれるデータは似た特徴を持ち、特徴の似ていないデータは異なるクラスタに含まれるように分割される。クラスタリングとは、先に述べたように特徴の似たデータで 1 つのクラスタを構成するようにデータを分割する手法の総称であり、代表的な手法として k-means、ファジィ c-平均法[3]、階層的クラスタリング[4]が知られている。また、クラスタリングが対象とするデータはベクトルデータに限らず、近年ではグラフ構造で表されるネットワークデータも重要な解析対象の一つとなっている[5]。

クラスタリング手法は階層的手法と非階層的手法の 2 つに大別される。階層的クラスタリングは一つ一つのデータをクラスタとみなし、逐次的に結合することでクラスタ分割を生成する手法である。階層的手法は非階層的手法に比べて計算量は多いものの、クラスタ分割の生成過程を樹形図として可視化することも可能であることから、比較的小規模のデータを対象とする場合に有効な手法と考えられる[4]。一方、k-means をはじめとする非階層的クラスタリングの多くの手法は、帰属度などの制約条件の下で、目的関数を最小化する交互最適化を行うことで、クラスタ分割を生成する手法である。非階層的手法の代表例である k-means はクラスタ中心と呼ばれるクラスタの代表点と帰属度と呼ばれるデータがクラスタへ所属するか否かを示す変数の交互最適化によりクラスタ分割を生成する手法である。非階層的手法は、階層的クラスタリングに比べて計算量が少ないことから比較的規模の大きなデータに対しても適用することが可能であり、大量のデータを少数のデータへ要約する前処理やデータの構造や規則性の分析にも用いられる[2,3,4]。また、k-means の帰属度を 0 か 1 の二値ではなく、0 から 1 の範囲を取るように拡張したファジィクラスタリングは、人間の感覚に沿った柔軟なクラスタ分割を生成するという点に加えて、確率モデルの分野における混合ガウス分布との関連性が議論されるなど、理論と応用の両面から重要な研究対象となっている[6]。

k-means やファジィクラスタリングを代表とする非階層的手法は、比較的規模の大きなデータを扱えるという特徴から様々な分野において利用されている。しかしながら、クラスタリング手法にはいくつかの問題点があるため、今後も増大かつ複雑化が進行すると考えられるビッグデータを扱うには、それらの問題点を克服した手法が望まれる。ここでは、クラスタリングにおける代表的な問題とそれらに対応するために提案されたクラスタリング手法について説明する。まず、k-means などの非階層的手法は、交互最適化を行うため、最終的に得られるクラスタ分割が初期値に依存するという初期値依存の問題がある。この初期値依存に対する手法として、初期値として与えるクラスタ中心をデータ個体の中から、ある確率に従って選択する k-means++が提案されている[7]。k-means++はすでに選択されたデータ個体から遠いデータ個体ほど、次のクラスタ中心として選択されやすくする初期値選択の方法を取ることで、目的関数が収まる範囲を理論的に解析した手法である。k-means++は通常の k-means よりも早く収束し、良好なクラスタ分割を得られることから、scikit-learn (https://scikit-learn.org/stable/index.html)などのデータ解析ライブラリにも収録されている。また、クラスタリング手法の多くは非類似度としてユークリッド距離の自乗を用いるため、結果として得られるクラスタ分割がボロノイ図を構成するという特徴がある。そのため、データの塊から離れた位置に存在する外れ値の影響を受けやすいという特徴がある。非類似度としてユークリッド距離の自乗を用いるもう一つの影響として、生成されるクラスタ分割が等方性を持つ超球状になるという性質もある。外れ値

公益財団法人電気通信普及財団
研究調査助成報告書　第35号　2020年度

に対する影響を低減した手法としては、可能性クラスタリング[8]、ノイズクラスタリング[9]、DBSCAN[10]、HDBSCAN[11]などが代表的である。また、多様なクラスタ分割を生成する手法としてカーネル法を用いたクラスタリング手法が提案されている[3]。このように、クラスタリングに関する諸問題を解決するには、非類似度や帰属度に関する数理モデルを拡張し、新たな目的関数に基づいたアルゴリズムを構築する方法が取られている。

そこで本研究では、外れ値などの影響を低減し、データの分布や構造に依存することなく良好なクラスタ分割を生成する手法として、Jensen-Shannon divergence（JS-ダイバージェンス）[12]に基づく k-medoids（JS-divergence based k-medoids, JSKMdd）を提案する。従来手法では、単一のデータが持つ座標などの情報に対して非類似度を与えるため、外れ値やデータの分布を考慮することが困難となっている。そのため提案手法では、データが持つ情報だけでなく、データの周辺に分布するデータも併せて考慮する非類似度を用いる。具体的には、カーネル密度推定(Kernel Density Estimation, KDE)[13]を用いて、近傍に存在するデータを用いて一つの個体をデータ分布に変換する。さらに、KDE を用いて推定されたデータ分布間の非類似度を Jensen-Shannon Divergence（JS-ダイバージェンス）[12]により算出する。JS-ダイバージェンスの算出は計算コストが大きいため、提案手法ではクラスタ中心をクラスタ内に存在する個体から選択する k-medoids 型のアルゴリズム[14]を用いる。

提案手法はクラスタリングの代表的手法である k-medoids を拡張したアルゴリズムである。数値実験では 5 種類の人工データを用いて、クラスタリングの代表的手法である k-means、k-medoids、スペクトラルクラスタリング（Spectral clustering, SC）[15]と比較を行った。SC はカーネル法との関連性も示されており、複雑な構造を持つデータを適切に分割する有用な手法として知られている。数値実験の結果から、提案手法 JSKMdd は線形な境界を持つデータおよび非線形な境界を持つデータの両者を適切に扱える手法であり、パラメータを適切に設定することで SC を上回る性能を示すことが示唆された。

2 準備

クラスタリングの対象となるデータ集合を$X = \{x_k \mid x_k \in R^p,\ k = 1{\sim}n\}$とする。クラスタを$G_i$とし、クラスタリングの結果として得られるクラスタ分割を$G = \{G_1, \ldots, G_c\}$とする。また、データx_kがクラスタG_iに所属する度合いを示す帰属度をu_{ki}とし、帰属度の集合を$U = \{u_{ki} \mid k = 1{\sim}n,\ i = 1{\sim}c\}$とする。

2-1 k-medoids

k-medoids はクラスタリングの代表的手法の一つとして知られている[14]。最も代表的な手法である k-means はクラスタの代表点であるクラスタ中心をクラスタ内の個体の平均とする手法である[2]。一方で、k-medoids はクラスタ内の個体から代表点を選択する手法となっている。そのため、k-medoids はベクトルデータのみならず、拡散カーネル[16]などを用いた重み付けを行うことでネットワークデータなどの関係データを扱うことも可能である[17]。k-medoids の目的関数を以下に示す。

$$J(U, W) = \sum_{i=1}^{c} \sum_{k=1}^{n} \sum_{l=1}^{n} u_{ki} w_{li} r_{kl} \tag{1}$$

ここで、r_{kl}はデータ間の関係性を示す尺度であり、ユークリッド距離、マンハッタン距離、ユークリッド距離の自乗などが用いられる。また、$W = \{w_{li} \mid l = 1{\sim}n,\ i = 1{\sim}c\}$はクラスタの代表点を決める重みである。$k$-medoids のアルゴリズムは$u_{ki}$と$w_{li}$の制約条件の下での交互最適化により構成される。$u_{ki}$と$w_{li}$の制約条件を以下に示す。

$$U = \left\{ (u_{ki}) :\ u_{ki} \in \{0, 1\}, \sum_{i=1}^{c} u_{ki} = 1,\ {}^{\forall}k \right\} \tag{2}$$

$$W = \left\{ (w_{li}) :\ w_{li} \in \{0, 1\},\ \sum_{l=1}^{n} w_{li} = 1,\ {}^{\forall}i \right\} \tag{3}$$

公益財団法人電気通信普及財団
研究調査助成報告書 第35号 2020年度

k-medoids では、$w_{li} = 1$となったl番目の個体がクラスタの代表点となる。目的関数と制約条件よりu_{ki}とw_{li}の更新式は以下で与えられる。

$$u_{ki} = \begin{cases} 1 & \left(i = \arg\min_{s} \sum_{l=1}^{n} w_{ls} r_{kl} \right) \\ 0 & (\text{otherwise}) \end{cases} \tag{4}$$

$$w_{li} = \begin{cases} 1 & \left(i = \arg\min_{t} \sum_{k=1}^{n} u_{ki} r_{kt} \right) \\ 0 & (\text{otherwise}) \end{cases} \tag{5}$$

k-medoids のアルゴリズムは以下で表される。

Algorithm 1 KMdd

KMdd1 Set cluster number c and initial medoids $w_{li} \in \boldsymbol{W}$ by choosing objects at random.

KMdd2 Calculate $u_{ki} \in \boldsymbol{U}$ using (4).

KMdd3 Calculate $w_{li} \in \boldsymbol{W}$ using (5).

KMdd4 If the convergence criterion is satisfied, stop. Otherwise, return to **KMdd2**.

KMdd4 における収束条件として、最大繰り返し回数、変数の収束、目的関数の収束が考えられる。

2-2 カーネル密度推定

カーネル密度推定は、特定の分布を仮定せずに与えられたデータセットに基づいて確率密度関数を推定する手法である[13]。x_kを独立同一分布に従う確率変数とすると、以下のカーネル密度推定により確率密度関数は以下で表される。

$$p(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h} \right) \tag{6}$$

ここで、$K(\cdot)$はカーネル関数であり、$h > 0$は滑らかさを調整するパラメータである。また、カーネル関数$K(\cdot)$は以下の条件を満たす関数が用いられる。

$$K(x) \geq 0, \qquad \int K(x)\, dx = 1, \qquad \int x\, K(x) dx = 0, \qquad \int x^2\, K(x) dx > 0$$

$K(\cdot)$の代表的な例として、以下で表されるガウシアンカーネルがある。

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

カーネル密度推定により求められる多次元の確率密度関数は以下で表される。

$$p(x) = \frac{1}{n}\sum_{k=1}^{n}\prod_{j=1}^{p}\frac{1}{h_j}K\left(\frac{x^j - x_k^j}{h_j}\right) \tag{7}$$

ここで、$h_j > 0$はj次元に対する滑らかさを調整するパラメータであり、x_k^jはx_kのj次元目の要素である。

2-3 Jensen-Shannon Divergence

JS-ダイバージェンスは 2 つの確率密度関数に対する非類似度と考えることできる[12]。JS-ダイバージェンスは Kullbuck-Leibler ダイバージェンス（KL-ダイバージェンス）[18]に基づいている。2 つの確率密度関数$p(x)$と$q(x)$間の KL-ダイバージェンスは以下で与えられる。

$$KL(p(x) \,||\, q(x)) = \int_{-\infty}^{+\infty} p(x)\log\frac{p(x)}{q(x)}dx$$

クラスタリングにおける非類似度は対称性を満たすことが一般的であるため、KL-ダイバージェンスをそのまま適用するには不都合が生じる場合がある。そこで、対称性を満たす尺度として、JS-ダイバージェンスが知られている[12]。JS-ダイバージェンスは以下で表される。

$$JS\big(p(x) \,\|\, q(x)\big) = \frac{1}{2}KL\big(p(x) \,\|\, m(x)\big) + \frac{1}{2}KL\big(q(x) \,\|\, m(x)\big) \tag{8}$$

ここで、$m(x) = \frac{1}{2}\big(p(x) + q(x)\big)$である。

3 提案手法

本章では、本研究で提案する JS-ダイバージェンスを用いた k-medoids（JSKMdd）について説明する。JSKMdd は k-medoids の拡張であり、データの構造や周辺の分布を考慮するために、KDE で推定したデータ分布間の非類似度を JS-ダイバージェンスで求める手法である。

はじめに KDE を用いてデータ分布を推定する方法について説明する。データx_kの近傍個体の集合を$N(x_k)$とし、以下で定める。

$$N(x_k) = \{x \in X \mid d(x_k, x) \leq D\} \tag{9}$$

ここで、$D > 0$はパラメータであり、$d(x_k, x)$はユークリッド距離などの距離尺度である。式(9)では、データx_kからD以下の非類似度となるデータが$N(x_k)$に含まれることになる。また、任意の個数のデータを$N(x_k)$に含む場合には、以下のように定義することも考えられる。

$$N(x_k) = \big\{x \in X \mid d(x_k, x) \leq d\big(x_k, x_{q(t)}\big)\big\} \tag{10}$$

ここで、$q_{(t)} \in \{1\sim n\}$は$d(x_k, x_1), \ldots, d(x_k, x_n)$を昇順に並び替えた場合の個体番号を示しており、$t$番目に小さい値を持つ個体番号を指す。式(9)と式(10)を比較した場合、式(10)で与える近傍では、必ずt個の個体を含むことになる。

データx_kと近傍に含まれるデータ集合$N(x_k)$から KDE により推定される確率密度関数$p(x_k)$は以下で表される。

$$p(x_k) = \frac{1}{|N(x_k)|} \sum_{x \in N(x_k)} \prod_{j=1}^{p} \frac{1}{h_j} K\left(\frac{x^j - x_k^j}{h_j}\right) \tag{11}$$

ここで、$|N(x_k)|$は集合$N(x_k)$に含まれる個体数を意味している。式(11)で求めたデータx_kの確率密度関数に対して、式(8)を用いて個体間の非類似度を算出する。

提案手法の目的関数は式(1)のr_{kl}を JS-ダイバージェンスに置き換えた以下の式で表される。また、制約条件は k-medoids と同じく、式(2), (3)で表される。

$$J(U,W) = \sum_{i=1}^{c} \sum_{k=1}^{n} \sum_{l=1}^{n} u_{ki} w_{li} r'_{kl}$$

$$U = \left\{ (u_{ki}) : \ u_{ki} \in \{0,1\}, \sum_{i=1}^{c} u_{ki} = 1, \ ^{\forall}k \right\}$$

$$W = \left\{ (w_{li}) : \ w_{li} \in \{0,1\}, \ \sum_{l=1}^{n} w_{li} = 1, \ ^{\forall}i \right\}$$

提案手法のアルゴリズムは以下で表される。

Algorithm 2 JSKMdd

JSKMdd1 Set parameter D or t. Calculate $p(x_k)$ by (12).

JSKMdd2 Calculate r'_{kl} by (9).

JSKMdd3 Set cluster number c and initial medoids $w_{li} \in \boldsymbol{W}$ by choosing objects at random..

JSKMdd4 Calculate $u_{ki} \in U$ using (4).

JSKMdd5 Calculate $w_{li} \in W$ using (5).

JSKMdd6 If convergence criterion is satisfied, stop. Otherwise, return to **JSKMdd4**.

k-medoids との大きな違いは JSKMdd1 においてパラメータを設定し、KDE を用いて確率密度関数を推定する点、JSKMdd2 において JS-ダイバージェンスを算出する点である。u_{ki}とw_{li}の更新式は式(4), (5)においてr_{kl}をr'_{kl}に置き換えることで求められる。JSKMdd6 における収束条件として、KMdd4 と同じく、最大繰り返し回数、変数の収束、目的関数の収束が考えられる。

4 数値実験

提案手法の有効性を検証するために 5 種類の人工データを用いて数値実験を行った。はじめに、計算条件について述べ、その後実験結果について説明する。最後に提案手法の特徴について示す。本研究では、k-means、k-medoids、スペクトラルクラスタリングと比較を行った。

4-1 計算条件

今回使用する人工データは分割結果が既知のデータである。提案手法と上記の比較手法について、それぞれの手法が生成するクラスタ分割と人工データの分割を Adjusted rand index (ARI)[19]を用いて評価した。ARI は 2 つの分割結果の一致度合いを示す指標であり、完璧に一致した場合に ARI の値は 1 となる。

はじめに、使用した人工データのデータ数、次元数、クラスタ数について表 1 に示す。

表 1: 数値実験で使用したデータの詳細

	n	p	c
Polaris	51	2	3
Artificial	72	2	3
Double circle	150	2	2
Double circle 2	150	2	2
Two moon	400	2	2

また、図 1 から 5 にそれぞれの人工データを適切に分割した際の結果を示す。図 3, 4 では外側のクラスタに 100 個のデータが含まれ、内側のクラスタに 50 個のデータが含まれている。図 4 は図 3 の内側のクラスタを外側のクラスタに近づけたものとなっている。

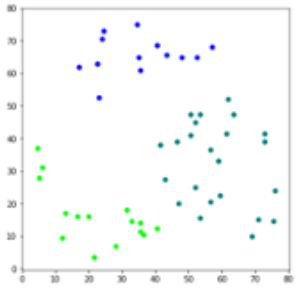

図1 Polaris dataset

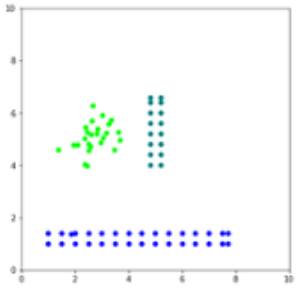

図2 Artificial dataset

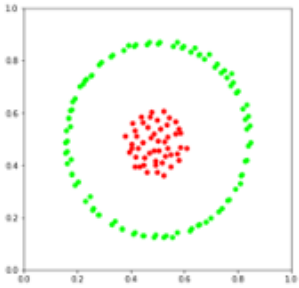

図3 Double circle dataset

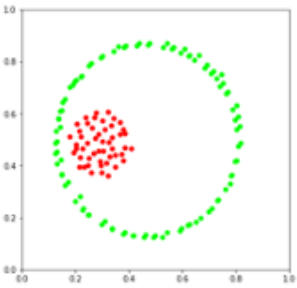

図4 Double circle 2 dataset

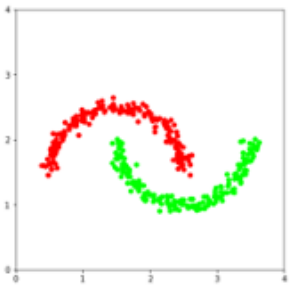

図5 Two moon dataset

提案手法では、データ分布を推定する際のパラメータとして、Dもしくはtを設定する必要があり、KDE で確率密度関数を推定する際に、バンド幅h_jを設定する必要がある。適切なパラメータはデータセットによって異なるため、表 2 に示されるパラメータを用いて数値実験を行った。実験では、全てのバンド幅h_jに対して同じ値を与えた。また、バンド幅は表 2 に示した値の範囲を 100 分割し、それぞれの値を与えて実験を行った。また、提案手法は初期値に依存するため、初期値を 100 回変えて実行し最も値が良かった場合の ARI で評価した。

表 2: 数値実験で使用したパラメータの詳細

Data	t	h_j
Polaris	$t \in \{4, 6, 8\}$	$h_j \in [0.001, 5.000]$
Artificial	$t \in \{4, 9, 14\}$	$h_j \in [0.001, 1.000]$
Double circle	$t \in \{4, 6, 8\}$	$h_j \in [0.001, 0.100]$
Double circle 2	$t \in \{4, 9, 12\}$	$h_j \in [0.001, 0.100]$
Two moon	$t \in \{4, 8, 12\}$	$h_j \in [0.001, 0.500]$

4-2 実験結果

図 6 から 11 にバンド幅h_jと近傍数を変えた場合の ARI の結果を示す。図の縦軸は ARI の値、横軸はバンド幅h_jの値を示している。図 6, 7 は線形に分離可能なデータセットであるため、ARI が 1 となる場合が多いことが確認できる。また、図 8 から 11 は非線形境界を持つデータセットであるため、図 6, 7 と比較すると ARI が低くなっている個所が多いことが確認できる。また、図 9 と 10 は、Double circle 2 dataset に対して近傍数を増加させた場合の結果である。図 10 の結果より、近傍数が大きすぎる場合には良好な分割結果が得られないことが確認できる。図 11 の結果から提案手法は Two moon dataset に対して良好なクラスタ分割を生成することが難しいことが確認できる。

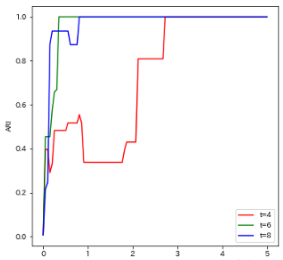

図6 Polaris datasetに対するJSKMddのARIの結果

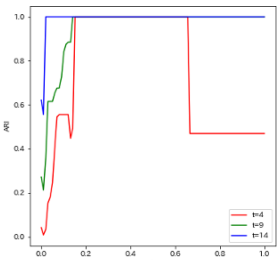

図7 Artificial datasetに対するJSKMddのARIの結果

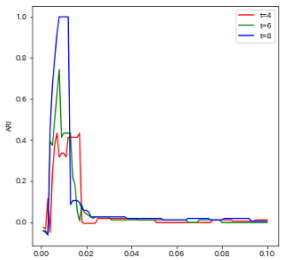

図8 Double circle datasetに対するJSKMddのARIの結果

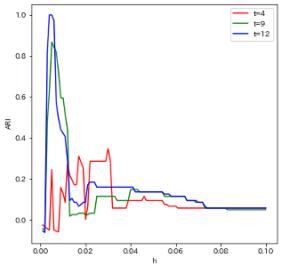

図9 Double circle 2 datasetに対するJSKMddのARIの結果1

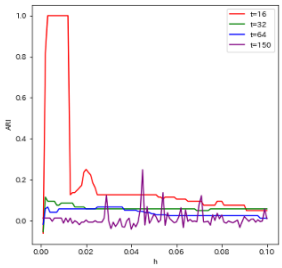

図10 Double circle 2 datasetに対するJSKMddのARIの結果2

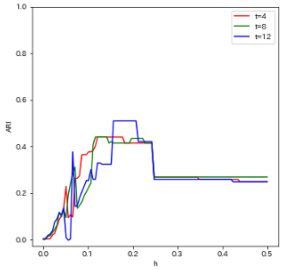

図11 Two moon datasetに対するJSKMddのARIの結果

次に表 3 にそれぞれの人工データに対して、k-means、k-medoids、スペクトラルクラスタリング、提案手法を適用した場合に最も良好な値となった ARI の値を示している。太字で示された値はそれぞれのデータセットに対して、最も良好な結果を得たことを意味している。Polaris データセットは線形分離可能であるため、全ての手法で適切なクラスタ分割を得ていることが確認できる。Artificial データセットは、線形分離可能に見えるが、k-means や k-medoids では扱いにくいクラスタの形状となっているため、ARI の値が小さくなっている。一方で、カーネル法を用いているスペクトラルクラスタリングや提案手法では適切な分割結果を得ている。非線形なクラスタ境界を持つ Double circle や Two moon データセットは、k-means や k-medoids で扱いにくいクラスタの構造であるため ARI の値が小さくなっている。Double circle に対しては SC と提案

手法がどちらとも良好な分割を得ている。Double circle 2 ではスペクトラルクラスタリングの結果が悪く、Two moon では提案手法の結果が悪くなっていることからが確認できる。

表 3: ARI の結果

Data	KM	KMdd	SC	JSKMdd
Polaris	**1.000**	**1.000**	**1.000**	**1.000**
Artificial	0.469	0.469	**1.000**	**1.000**
Double circle	0.008	0.005	**1.000**	**1.000**
Double circle 2	0.058	0.033	0.478	**1.000**
Two moon	0.233	0.274	**1.000**	0.510

図 6 から 11 および表 3 の結果より、提案手法は線形や非線形の構造を持つデータセットに対して良好なクラスタ分割を得ることができる手法であることが確認できる。また提案手法では 2 つのパラメータが必要となるが、複雑な構造を持つデータセットに対しては、近傍数の影響がより強いことが示唆される。

提案手法で用いる KDE による確率密度関数の推定および JS-ダイバージェンスの計算は大きな計算コストを必要とする。そのため、実データを対象とする場合には、それぞれの計算効率の向上および最適なパラメータの自動推定が必要となることが考えられる。

5 おわりに

本研究では、JS-ダイバージェンスに基づく k-medoids (JSKMdd)を提案した。提案手法は k-medoids の拡張であり、KDE と JS-ダイバージェンスを用いるころで、線形のクラスタ構造を持つデータセットおよび非線形なクラスタ構造を持つデータセットの両者を適切に扱える手法である。数値実験により提案手法を従来手法と比較し、その有効性を示した。数値実験の結果より、提案手法は適切なパラメータを設定することで、従来手法を上回る性能を示すことが示唆された。

今後の課題として、ベンチマークデータを用いた提案手法の評価が挙げられる。また、KDE と JS-ダイバージェンスにより求められる非類似度がどのような特徴を持つか、詳細に分析する必要がある。さらに、パラメータの自動推定やネットワークデータなどへの適用も重要な課題と考えられる。

【参考文献】

[1] Han J., Pei J., and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann, 2011.

[2] Jain, A.K., Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. Vol. 31, No. 8, pp. 651-666, 2010.

[3] Miyamoto, S., Ichihashi, H., Honda, K., Algorithms for Fuzzy Clustering. Springer, Heidelberg, 2008.

[4] 宮本定明, クラスター分析入門—ファジィクラスタリングの理論と応用, 森北出版, 1999.

[5] Newman M., Networks: An Introduction, Oxford University Press, New York, 2010.

[6] Ichihashi H., Honda K., and Tani N., Gaussian mixture PDF approximation and fuzzy c-means clustering with entropy regularization, Proc. 4th Asian Fuzzy System Symp., pp. 217-221, 2000.

[7] D. Arthur, Vassilvitskii S., k-means++: the advantages of careful seeding, Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035, 2007.

[8] Krishnapuram R., Keller J.M., A possibilistic approach to clustering, IEEE Trans. on Fuzzy Systems, Vol. 1, No. 2, pp. 98-110, 1993.

[9] Dave R.N., Krishnapuram R., Robust clustering methods: a unified view, IEEE Trans. on Fuzzy Systems, Vol. 5, No. 2, pp. 270-293, 1997.
[10] Ester, M., Kriegel H. P., Sander J., and Xu, X., A density-based algorithm for discovering clusters in large spatial databases with noise, Proc. of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), pp. 226-231, 1996.
[11] Campello R. J. G. B., Moulavi D., and Sander J., Density-based clustering based on hierarchical density estimates, Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2013), LNCS7819, Springer, Berlin, Heidelberg, pp. 160-172, 2013.
[12] Fuglede, B., Topsoe, F., Jensen-Shannon divergence and Hilbert space embedding, Proc. International Symposium on Information Theory (ISIT2004), 2004.
[13] Epanechnikov V. A., Non-parametric estimation of a multivariate probability density, Theory of Probability and its Application, Vol. 14, pp. 153-158, 1969.
[14] Kaufman L. and Rousseeuw P. J., Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.
[15] Luxburg U. von, A tutorial on spectral clustering, Statistics and Computing, Vol. 17, No. 4, pp. 395-416, 2007.
[16] Kondor R. I. and Lafferty J. D., Diffusion kernels on graphs and other discrete input spaces, Proc. of ICML, pp. 315–322, 2002.
[17] Nakano S., Hamasuna Y., Endo Y., A study on controlled node sized network clustering for unweighted network data, Joint 10th International Conference on Soft Computing and Intelligent Systems and 19th International Symposium on Advanced Intelligent Systems (SCIS & ISIS 2018), pp. 826–831, 2018.
[18] Kullback, S., and Leibler, R. A., On information and sufficiency, Annals of Mathematical Statistics, Vol. 22, pp. 79-86, 1951.
[19] Hubert L., and Arabie P., Comparing Partitions, Journal of Classification, Vol. 2, No. 1, pp. 193–218, 1985.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
k-medoids Clustering based on Kernel Density Estimation and Jensen-Shannon Divergence	The 16th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2019), Springer, LNCS 11676, pp. 272-282, 2019	2019 年 9 月