数理解析研究所講究録
第2078巻 2018年 127-133

ベクトルのスカラー化を用いたクラスター分析の新たな手法について

(A New Method of Cluster Analysis by Using Scalarization in Vector Optimization)

秋田県立大学　システム科学技術学部　経営システム工学科
Faculty of Systems Science and Technology, Akita Prefectural University
荒谷　洋輔 (ARAYA, Yousuke) *　斎藤　裕 (SAITO, Yutaka) †
木村　寛 (KIMURA, Yutaka) ‡

1　はじめに

クラスター分析とは、異なった性質のものが混ざり合っている対象の中で、互いに似た者同士を集めてクラスター（集落）を作り、それらを分類する方法のことである。生物学 (生活型の分類)、医学 (患者の病気を分類)、地質学（岩石や鉱石の分類）など多方面への応用がある。クラスター分析には次の２つの問題がある。

（１）点と点の類似度をどう決めるか？

（２）点と集合、集合と集合の類似度をどう決めるか？

上記の問について、（１）は**ユークリッド距離、重み付きユークリッド距離、ミンコフスキー距離、マハラノビス距離**などが、（２）は**最短距離法、最長距離法、群平均法、ウォード法**などが過去に提案された（[3, 5] やその参考文献を参照のこと)。

ところで、対象となる（ベクトル）データの中には、単に１つの目的だけではなく、あれもこれも良くしたいと思うことがしばしばあり、その考え方を取り入れた分類法も考えられる。本稿では、多目的最適化の考え方を取り入れた新しいクラスター分析手法を提案する。

尚、クラスター分析には「階層的な方法」（あらかじめクラスター数は決めない）と「非階層的な方法」（あらかじめクラスター数を決める）があるが、本稿では次の２次元のデータを**階層的な方法で分析する**。

番号	数学（x_1）	英語（x_2）
①	3	5
②	4	4
③	4	2
④	1	1
⑤	2	1

（上の表は、５人の数学と英語の成績を５段階評価で表したものである。）

*(*E-mail:* **y-araya@akita-pu.ac.jp**)
†(*E-mail:* **yutakasai@akita-pu.ac.jp**)
‡(*E-mail:* **yutaka@akita-pu.ac.jp**)

2　従来のクラスター分析手法

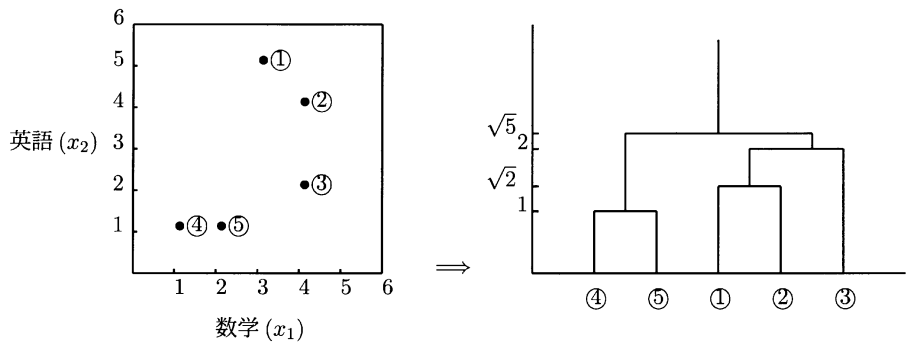
従来のクラスター分析手法の一例を紹介する。

(a) 2点 $a=(a_1,a_2)$、$b=(b_1,b_2)$ のユークリッド距離を $d_2(a,b)$ で表すことにすると

$$d_2(a,b) := \sqrt{(a_1-b_1)^2+(a_2-b_2)^2}$$

で定義される。

(b) 樹形図（デンドログラム）を作成する。クラスター間の距離測定は、**最短距離法**を用いる。



3　ベクトルのスカラー化手法を用いたクラスター分析手法

先ほどの例において、個人データは基本的に比較できない。これからは、データを比較する方法を考える。

定義 3.1 (ベクトル順序・平面ベクトル $\mathbb{R}^2$ の場合 [4]). ベクトル $a,b\in\mathbb{R}^2$ と凸錐

$$\mathbb{R}^2_+ := \{(x,y)|x\geqq 0, y\geqq 0\}$$

に対して、ベクトル順序を以下で定義する。

$$a\leqq_{\mathbb{R}^2_+} b \iff b-a\in\mathbb{R}^2_+$$

一般には、凸錐 $C\subset\mathbb{R}^2$ に対して、以下のように定義する。

$$a\leq_C b \iff b-a\in C$$

例 1. ベクトル $a=(2,5)$、$b=(3,8)$ とする。$b-a=(3,8)-(2,5)=(1,3)\in\mathbb{R}^2_+$ なので、$a\leqq_{\mathbb{R}^2_+} b$ である。

注意 1. 2つの実数は必ず比較できる（全順序）。しかし、ベクトルは順序関係を導入しても比較できないことがある。

例 2. ベクトル $a=(2,5)$、$b=(3,4)$ とする。$b-a=(3,4)-(2,5)=(1,-1)\notin\mathbb{R}^2_+$ なので、$a\not\leqq_{\mathbb{R}^2_+} b$ である。

定義 3.2 (Pareto 最適解・平面ベクトル $\mathbb{R}^2$ の場合 [4]). $f:\mathbb{R}\to\mathbb{R}^2$ をベクトル値関数とする。$f(a)\leqq_{\mathbb{R}^2_+} f(x)$ となるような $a\in\mathbb{R}$ が存在しないとき、$x\in\mathbb{R}$ を **Pareto 最適解**と言う。

Pareto 最適解を求めるために古くから最も使われている手法がスカラー化である。

定理 3.3 (Pareto 最適解の線形スカラー化関数による特徴づけ). $f:\mathbb{R}\to\mathbb{R}^2$ をベクトル値関数、L を線形スカラー化関数とする。$\cup_{x\in\mathbb{R}}\{f(x)\}$ **が凸集合** ならば、次が成り立つ。

$$\bar{x}\in\mathbb{R} \text{ が Pareto 最適解} \iff L(f(\bar{x}))\le L(f(x)) \quad \forall x\in\mathbb{R}$$

Proof. 凸集合同士の分離定理から示される（[2] とその参考文献を参照)。 □

上記の定理から「ベクトル順序で最適であること」と「スカラー化」は密接に関連していることが分かる。スカラーは取り扱いやすいという利点があるので、次の新しい分析手法の着想に至った。

3.1 データの各項目の平均（線形スカラー化）を用いる手法

手順

(a) データの各項目の平均をとることにより、各データをスカラー化する。

（注意）「平均」は一種の線形スカラー化である。

(b) スカラー化した各データ同士の距離を計算する。

(c) 樹形図（デンドログラム）を作成する。クラスター間の距離測定は、**最短距離法**を用いる。

実際の計算

(a) $A(①)=\dfrac{3+5}{2}=\boxed{4}$ $A(②)=\dfrac{4+4}{2}=\boxed{4}$ $A(③)=\dfrac{4+2}{2}=\boxed{3}$

$A(④)=\dfrac{1+1}{2}=\boxed{1}$ $A(⑤)=\dfrac{2+1}{2}=\boxed{1.5}$

(b) $|A(①)-A(②)|=0,\quad |A(①)-A(③)|=1,\quad |A(①)-A(④)|=3\quad |A(①)-A(⑤)|=2.5,$

$|A(②)-A(③)|=1,\quad |A(②)-A(④)|=3,\quad |A(②)-A(⑤)|=2.5,$

$|A(③)-A(④)|=2,\quad |A(③)-A(⑤)|=1.5,\quad |A(④)-A(⑤)|=0.5$

(c)

	①	②	③	④	⑤
①	0				
②	0	0			
③	1	1	0		
④	3	3	2	0	
⑤	2.5	2.5	1.5	0.5	0

$$|C(①,②)-A(③)|=\min\{d(①,③),d(②,③)\}=\min\{1,1\}=\boxed{1}$$

	C(①,②)	③	④	⑤
C(①,②)	0			
③	1	0		
④	3	2	0	
⑤	2.5	1.5	0.5	0

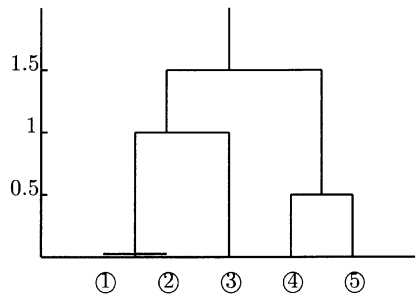
$$|C(④,⑤) - A(③)| = \min\{d(③,④), d(③,⑤)\} = \min\{2, 1.5\} = \boxed{1.5}$$

$$|C(①,②), C(④,⑤)| = \min\{d(①,④), d(①,⑤), d(②,④), d(②,⑤)\}$$
$$= \min\{3, 2.5, 3, 2.5\} = \boxed{2.5}$$

	C(①,②)	③	C(④,⑤)
C(①,②)	0		
③	1	0	
C(④,⑤)	2.5	1.5	0

$$|C(①,②,③), C(④,⑤)|$$
$$= \min\{d(①,④), d(①,⑤), d(②,④), d(②,⑤), d(③,④), d(③,⑤)\}$$
$$= \min\{3, 2.5, 3, 2.5, 2, 1.5\} = \boxed{1.5}$$

	C(①,②,③)	C(④, ⑤)
C(①,②,③)	0	
C(④,⑤)	1.5	0



3.2 劣線形ベクトルスカラー化関数を用いる手法

定理 3.4 (Pareto 最適解の**劣線形スカラー化関数による**特徴づけ). $f : \mathbb{R} \to \mathbb{R}^2$ をベクトル値関数、$k^0 \in C$、

$$h_{C,k^0}(y) = \inf\{t \in \mathbb{R} \mid y \in tk^0 - C\}$$

を**劣線形スカラー化関数**とすると、次が成り立つ。

$$\bar{x} \in \mathbb{R} \text{ が Pareto 最適解} \iff h_{C,k^0}(f(\bar{x})) \leq h_{C,k^0}(f(x)) \quad \forall x \in \mathbb{R}$$

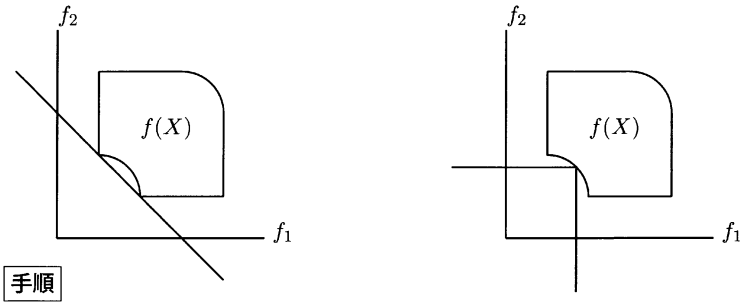
Proof. [2] を参照のこと。 □

< Tchebyshev スカラー化関数について [4] >

（重み付き）線形和スカラー化関数では、非凸な Pareto フロンティア上の一部の解を抽出できない。したがって、どのような Pareto 解でもスカラー化関数最小化の解として得ることができることが望ましい。

この性質をもつスカラー化関数は、図からの解釈で容易に分かるように、**等高線が逆 L 字型の形に折れ曲がった直線になるものである。**このような性質を持つ関数としては、上記で定義した「Tchebyshev スカラー化関数」がある。

（線形スカラー化関数）　　（Tchebyshev スカラー化関数）



手順

(a) ベクトルの劣線形スカラー化関数 h_{C,k^0}（C：閉凸錐、$k^0 \in C$）を用いて、各データをスカラー化する。

(b) スカラー化した各データ同士の距離を計算する。

(c) 樹形図（デンドログラム）を作成する。クラスター間の距離測定は、**最短距離法**を用いる。

(1) $C = \mathbb{R}^2_+ := \{(x,y) \mid x \geq 0, y \geq 0\}$、$k^0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ の場合

(a) $h_{C,k^0}(①) = \inf\{t \in \mathbb{R} \mid \begin{pmatrix} 3 \\ 5 \end{pmatrix} \in t \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \mathbb{R}^2\} = \boxed{5}$

$h_{C,k^0}(②) = \inf\{t \in \mathbb{R} \mid \begin{pmatrix} 4 \\ 4 \end{pmatrix} \in t \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \mathbb{R}^2\} = \boxed{4}$

$h_{C,k^0}(③) = \inf\{t \in \mathbb{R} \mid \begin{pmatrix} 4 \\ 2 \end{pmatrix} \in t \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \mathbb{R}^2\} = \boxed{4}$

$h_{C,k^0}(④) = \inf\{t \in \mathbb{R} \mid \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in t \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \mathbb{R}^2\} = \boxed{1}$

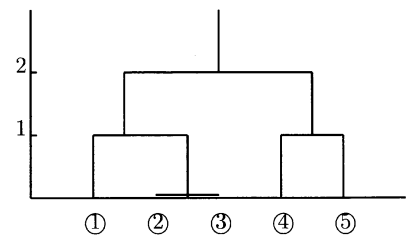
$h_{C,k^0}(⑤) = \inf\{t \in \mathbb{R} \mid \begin{pmatrix} 2 \\ 1 \end{pmatrix} \in t \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \mathbb{R}^2\} = \boxed{2}$

(b) $|h(①) - h(②)| = 1$, $|h(①) - h(③)| = 1$, $|h(①) - h(④)| = 4$, $|h(①) - h(⑤)| = 3$,

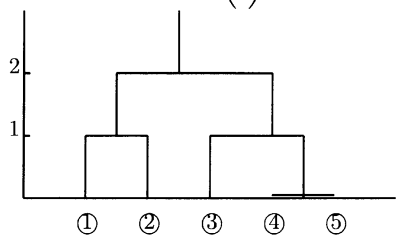
$|h(②) - h(③)| = 0$,　　$|h(②) - h(④)| = 3$,　　$|h(②) - h(⑤)| = 2$,

$|h(③) - h(④)| = 3$,　　$|h(③) - h(⑤)| = 2$,　　$|h(④) - h(⑤)| = 1$

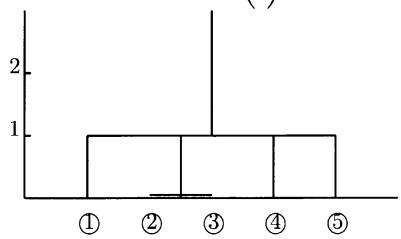
(c) 樹形図（デンドログラム）を作成する。



(2) $C=\mathbb{R}^2_+$、$k^0=\begin{pmatrix}2\\1\end{pmatrix}$ の場合



(3) $C=\mathbb{R}^2_+$、$k^0=\begin{pmatrix}1\\2\end{pmatrix}$ の場合



4 まとめ

ベクトルのスカラー化手法を取り入れた前章の２つのクラスター分析手法について、共通点と相違点は次のようになる。

(1) 共通点：**(A)** 各項目の重みづけの仕方によって、クラスタリングの構成が変わる。

(B) 通常のユークリッド距離からの観点で、遠い点同士を結びつけることもありうる。場合によっては、点と点の類似度の観点から疑問符がつく可能性がある。

(2) 相違点（２つの手法の比較）

	データの各項目の平均値を取る （線形スカラー化）	スカラー化関数 h_{C,k^0} を用いる （非線形スカラー化）
長所	(1) 多目的最適化と相性が良い。 （Pareto 最適解を **ある程度** 抽出できる） (2) 意味が分かりやすいので応用範囲が広い。	(1) 多目的最適化と相性が**とても良い。** （Pareto 最適解を **全て** 抽出できる） (2) データの情報が一部残る。
短所	共通点 (B) 参照	共通点 (B) 参照
例	＜学校の入学試験＞ **各科目の平均で評価** $\Longrightarrow$「各科目の（同時）最適化」という観点で最適解を見落とす可能性がある。	＜学校の入学試験＞ **得意科目（一番得点の高い科目）で評価** $\Longrightarrow$ 評価方法が極端である。

今回の新手法では、多目的最適化におけるスカラー化の定理が大切な役割を果たしている。この定理を実際に社会の問題に置き換えると、次のようなことを主張しているかも知れない。

- 平均的にすべての分野を伸ばすより、何か得意な分野を生かす方がより大切である。
- いろいろな分野で平均的に優れている人だけでは、何かに特化した人（ある意味で偏った人？）のようなタイプを見逃す可能性がある。人材には多様性が大切である。
- 集合の凸性は、制約・規律・想定内のこと … などに相当する。

5 今後の課題・展望

(1) ベクトル順序について：今回は話を簡単にするため、ベクトル順序を $\mathbb{R}^2_+$（第一象限）に限定して話を進めた。しかし、**問題によっては、一般のベクトル順序 $\leq_C$ で考えた方が良い場合もありうる。「問題によってどのように順序錐を決定すればよいのか」**という課題がある。

(2) ベクトルのスカラー化手法：新たなアイディアについていくつか紹介する。

- 上記のベクトルスカラー化の例について：「平均」と「最大値」を取り扱ったが**「中央値」「最頻値」でスカラー化する方法もあるのではないか。**
- 集合と点のスカラー化について：「最短距離法」を用いたが「群平均法」の方が良いのか？[3] によると、「群平均法」が一番よく使われているようなので、こちらを採用した方が良いのかも知れない。
- ２変数のベクトルスカラー化関数を用いる手法について

  (a) 次のベクトルの劣線形スカラー化関数を用いて、各データをスカラー化する。

$$h_{C,k^0}(y;a) = \inf\{t \in \mathbb{R} \mid y \in tk_0 + a - C\}$$

  （注意） 「$h_{C,k^0}(y;a) = h_{C,k^0}(a;y)$」ではない。

  (b) 樹形図を作成する。クラスター間の距離測定は、**最短距離法**を用いる。

(3) 集合のスカラー化関数を用いる手法について：クラスター間の距離測定の方法は、「集合と集合のスカラー化関数」も考えられる。

定義 5.1 (集合順序 [1])**.** 集合 $A, B \subset \mathbb{R}^n$ と凸錐 $\mathbb{R}^n_+$ に対して、集合順序を以下で定義する。

$$(\mathbf{L}\text{型}) \quad A \leqq^l_{\mathbb{R}^n_+} B \iff B \subset A + \mathbb{R}^n_+ \qquad (\mathbf{U}\text{型}) \quad A \leqq^u_{\mathbb{R}^n_+} B \iff A \subset B - \mathbb{R}^n_+$$

定義 5.2 (集合の劣線形スカラー化関数 [1])**.** 集合 $A, B \subset \mathbb{R}^n$、凸錐 $\mathbb{R}^n_+$、$k^0 \in \mathbb{R}^n_+$ に対して、集合スカラー化関数を以下で定義する。

$$(\mathbf{L}\text{型}) \quad F^l(A;B) = \inf\{t \in \mathbb{R} \mid tk^0 + B \subset A + \mathbb{R}^n_+\}$$

$$(\mathbf{U}\text{型}) \quad F^u(A;B) = \inf\{t \in \mathbb{R} \mid A \subset tk^0 + B - \mathbb{R}^n_+\}$$

謝辞 今回の講演では、たくさんの先生方から助言・アドバイスを頂きました。本稿の内容をさらに深めることにつながった、大切な役割を果たしたものもたくさんありました。ここに感謝の意を表します。

参考文献

[1] Y. Araya, *Four types of nonlinear scalarizations and some applications in set optimization*, Nonlinear Anal. 75, (2012) 3821–3835.

[2] A. Göpfert, H. Riahi, C. Tammer and C. Zălinescu *Variational methods in partially ordered spaces*, Springer-Verlag, New York 2003.

[3] H.Charles Romesburg (著), 西田 英郎 (翻訳), 佐藤 嗣二 (翻訳) 「実例　クラスター分析」内田老鶴圃, 1992 年.

[4] 中山 弘隆 (著), 岡部 達哉 (著), 荒川 雅生 (著), 尹 禮分 (著) 「多目的最適化と工学設計—しなやかシステム工学アプローチ」現代図書, 2008 年.

[5] 永田 靖 (著), 棟近 雅彦 (著) 「多変量解析法入門」サイエンス社, 2001 年.