研究ノート

# 異なるデータのクラスター分析結果を比較する
## ～その２：類似クラスターのアラインメント　２つの集合的同定法とＲ言語実装～


木村 敦 | Kimura Atsushi

㈱統計センター 理事・CIO

■ NTTにてICT関連開発に長年携わり、㈱NTTファシリティーズ総合研究所 取締役情報技術本部長を経て、2019年4月から現職。1988年3月名古屋大学大学院理学研究科博士課程（前期）修了、修士（理学）、専門統計調査士。


## 1. はじめに

前回「その1」[2]で紹介した「クラスター重心間距離最小法」は、計算量が少なく結果の納得性も高い手法であるが、クラスター分析に用いた元データに立ち返った計算が必要となることと、距離を測るための元データのベクトルの次元が等しい（もしくは合わせられる）ことが適用可能条件であった。

今回提案する手法は、異なるクラスター番号となる都道府県の数が最も少なくなるようにクラスター番号を振りなおす手法であり、クラスター分析結果のみあれば適用可能である。今回定義する「**クラスター差異度**」とは、2つのクラスター分析結果間（同一の集合に対する異なる分割結果間）における対称差集合要素数の総和であり、2つのクラスター分析結果の集合的な変化の大きさを示すものである。

クラスター数が$K$の場合「クラスター差異度最小化法」の厳密な求解には最悪の場合$K!$回の計算が必要となる。このため本稿では多項式時間で求解できる近似的な手法もあわせて提案し、各手法結果の比較評価も実施する。

今回提案する手法を実装したRスクリプトはエストレーラWeb[注1]で公開する（厳密解法の*cluster_identification*( )関数と、近似解法の*Jaccard_identification*( )関数）。

## 2. 用語の定義と基本的考え方

「**クラスターベクトル*N-K***」、「**クラスターベクトル*N-K*集合**」などの用語については前稿[2]での定義を参照願いたい。識別番号が$i$であるクラスターは、前稿同様$\{_i\boldsymbol{A}\}$と記述する。

ここからは、本稿で新たに定義する用語である。「クラスターベクトル$N$-$K$」$\boldsymbol{A}$の異なる2つのクラスター$\{_i\boldsymbol{A}\}$,$\{_j\boldsymbol{A}\}$において、お互いの元をすべて入れ替える操作を「**元交換操作**」と呼ぶことにする。$\{_i\boldsymbol{A}\}$と$\{_j\boldsymbol{A}\}$の間で「元交換操作」を行った後の「クラスターベクトル$N$-$K$」$\boldsymbol{A}'$を

$$\boldsymbol{A}' = \boldsymbol{EXCH}_{i,j}(\boldsymbol{A})$$

と表記する。このとき、

$$\{_i\boldsymbol{EXCH}_{i,j}(\boldsymbol{A})\} = \{_i\boldsymbol{A}'\} = \{_j\boldsymbol{A}\}$$
$$\{_j\boldsymbol{EXCH}_{i,j}(\boldsymbol{A})\} = \{_j\boldsymbol{A}'\} = \{_i\boldsymbol{A}\}$$

となる。

異なる「元交換操作」を任意の回数行った結果のクラスターベクトルを「**元交換ベクトル**」と呼

注1）（公財）統計情報研究開発センターHP（https://www.sinfonica.or.jp/）内の［刊行物］＞［エストレーラ］＞［参考］に掲載した。

ぶことにする。また、「元交換ベクトル」の集合を「**元交換ベクトル集合**」と呼ぶことにする。この場合「元交換ベクトル集合」$Y$ は $K!$ 個の「元交換ベクトル」を持つことになる。

$$|Y| = K!$$

$N$ 個の観測結果群を $K$ 個のクラスターに分けた結果は、「クラスターベクトル $N$-$K$ 集合」の中の 1 つの「クラスターベクトル $N$-$K$」の「元交換ベクトル集合」として表すことができる。同一の「元交換ベクトル集合」に含まれる任意の「クラスターベクトル $N$-$K$」はクラスター分析結果としては完全に同一である。

「クラスターベクトル $N$-$K$」$\boldsymbol{A} = (A_1, A_2, \cdots, A_N)$ の「**展開行列 $K$-$N$**」($K$ 行 $N$ 列) を以下のように定義する。$\boldsymbol{A}$ の「展開行列 $K$-$N$」の $i$ 行 $j$ 列の要素を $\boldsymbol{EXPM(A)_{ij}}$ と表記する。$m$ を 1 から $N$ の自然数としたとき

(1) $A_m = i$ の場合、 $EXPM(\boldsymbol{A})_{im} = 1$

(2) その他の場合、 $EXPM(\boldsymbol{A})_{ij} = 0$

とする。このとき、次の式が成り立つ。

$$\sum_{i=1}^{K} (EXPM(\boldsymbol{A})_{ij}) = 1$$

$$\sum_{i=1}^{K} \left( \sum_{j=1}^{N} (EXPM(\boldsymbol{A})_{ij}) \right) = N$$

次に、操作 $\boldsymbol{EXCH}_{i,j}(\boldsymbol{A})$ を $EXPM(\boldsymbol{A})$ で表現しよう。実例で考える。

$\boldsymbol{A} = (4,4,4,2,3,2,2,3,3,1,1,8,1,6,6,7,7,5,3)$ の場合。$\boldsymbol{A}$ は「クラスターベクトル *19-8*」である。このとき、$\{\boldsymbol{A}\} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19\}$、

$\{{}_1\boldsymbol{A}\} = \{10,11,13\}$、$\{{}_2\boldsymbol{A}\} = \{4,6,7\}$、

$\{{}_3\boldsymbol{A}\} = \{5,8,9,19\}$、$\{{}_4\boldsymbol{A}\} = \{1,2,3\}$、

$\{{}_5\boldsymbol{A}\} = \{18\}$、$\{{}_6\boldsymbol{A}\} = \{14,15\}$、

$\{{}_7\boldsymbol{A}\} = \{16,17\}$、$\{{}_8\boldsymbol{A}\} = \{12\}$ である。

ここで $\boldsymbol{A}$ に対して、$\{{}_2\boldsymbol{A}\}$ と $\{{}_6\boldsymbol{A}\}$ に対する「元交換操作」$\boldsymbol{EXCH}_{2,6}(\boldsymbol{A})$ を行う。この「元交換操作」は、$\{{}_2\boldsymbol{A}\} = \{4,6,7\}$、$\{{}_6\boldsymbol{A}\} = \{14,15\}$ であるから、

$\boldsymbol{EXCH}_{2,6}(\boldsymbol{A}) =$

$(4,4,4,\underline{6},3,\underline{6},\underline{6},3,3,1,1,8,1,\underline{2},\underline{2},7,7,5,3)$

である。

一方「元交換操作」は「展開行列 *8-19*」に対する以下の操作で代替できる。

① $\boldsymbol{A} = (4,4,4,2,3,2,2,3,3,1,1,8,1,6,6,7,7,5,3)$ を「展開行列 *8-19*」に変換すると、

$EXPM(\boldsymbol{A}) =$

0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0
0,0,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,1,0,0,1,1,0,0,0,0,0,0,0,0,0,1
1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0
0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0

となる。

② $EXPM(\boldsymbol{A})_{2j}$ と $EXPM(\boldsymbol{A})_{6j}$ を丸ごと入れ替えた行列をつくる。ここで、行列の $n$ 行と $m$ 行を入れ替える「**行交換操作**」を $\boldsymbol{EXCHROW_{n,m}}()$ と表記すると、

$EXCHROW_{2,6}(EXPM(\boldsymbol{A})) =$

0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0
$\underline{0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0}$
0,0,0,0,1,0,0,1,1,0,0,0,0,0,0,0,0,0,1
1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0
$\underline{0,0,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0}$
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0

となる。

③「展開行列」をベクトルに戻す操作を $\boldsymbol{CNT}(\ )$ と表記する。ここで、

$$D = EXCHROW_{2,6}(EXPM(\boldsymbol{A}))$$

と置き換える。行列 $D$ を「クラスターベクトル 19-8」に戻せば、

$$\boldsymbol{CNT}(D) = (4,4,4,\underline{6},3,\underline{6},\underline{6},3,3,1,1,8,1,\underline{2},\underline{2},7,7,5,3) = \boldsymbol{EXCH}_{2,6}(\boldsymbol{A})$$

つまり、

$$\boldsymbol{EXCH}_{i,j}(\boldsymbol{A}) = \boldsymbol{CNT}(EXCHROW_{i,j}(EXPM(\boldsymbol{A})))$$

と記述することができる。このように「クラスターベクトル」$\boldsymbol{A}$ の「元交換操作」は「展開行列」$EXPM(\boldsymbol{A})$ における「行交換操作」で置き換えることができる。R言語での実装においてはこれを利用している。

## 3.「クラスター差異度最小化法」について

「**クラスター差異度**」を $\boldsymbol{CD}$ と表記する。$CD$ はふたつの「クラスターベクトル $N$-$K$」$\boldsymbol{A}, \boldsymbol{B}$ 間の差異（違い）の大きさを表すもので、

$$CD = \frac{1}{2} \times \sum_{i=1}^{K}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{A})_{ij} - EXPM(\boldsymbol{B})_{ij})^2\right) \quad \text{(式 1)}$$

と定義する。1/2 倍しているのは、1 つの差異が $EXPM(\boldsymbol{A})$ 側と $EXPM(\boldsymbol{B})$ 側で二重にカウントされるためである。「**クラスター差異度最小化法**」とは、「クラスターベクトル $N$-$K$ 集合」に含まれる 2 つのベクトル $\boldsymbol{RefCN}$ と $\boldsymbol{CN}$ が与えられたとき、$\boldsymbol{CN}$ の「元交換ベクトル集合」に含まれる「元交換ベクトル」$\boldsymbol{CNm}$ の中から、「クラスター差異度」が最小となる $\boldsymbol{CNmin}$ を選び出す手法である。ここで、$m$ には 1 から $K$！までの自然数が入る。式で表せば、

$$\sum_{i=1}^{K}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CNmin})_{ij})^2\right) = min\left(\sum_{i=1}^{K}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CNm})_{ij})^2\right)\right) \quad \text{(式 2)}$$

である。言い換えれば、$\boldsymbol{RefCN}$ と $\boldsymbol{CNm}$ の値が異なる要素の数が最小になる $\boldsymbol{CNmin}$ を求めることに等しい。

$\boldsymbol{RefCN}$ と $\boldsymbol{CNm}$ が同一の「元交換ベクトル集合」に含まれる場合は、

$$min\left(\sum_{i=1}^{K}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CNm})_{ij})^2\right)\right) = 0$$

であり、$\boldsymbol{RefCN} = \boldsymbol{CNmin}$ となる。一方、$\boldsymbol{RefCN}$ と $\boldsymbol{CNm}$ が異なる「元交換ベクトル集合」に属する場合は、

$$min\left(\sum_{i=1}^{K}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CNm})_{ij})^2\right)\right) > 0$$

となり、$\boldsymbol{RefCN} \neq \boldsymbol{CNmin}$ である。

$\boldsymbol{CN}$ が含まれる「元交換ベクトル集合」には $K$！個の「元交換ベクトル」が存在するので、（式 2）を満足する $\boldsymbol{CNmin}$ を求めるには、$K$！個すべての「元交換ベクトル」$\boldsymbol{CNm}$ について（式 1）を計算して値を比較する必要がある。なお、（式 2）を満足する $\boldsymbol{CNmin}$ は複数存在する可能性があ

るが、今回のＲ言語の実装では解の中の１つの ***CNmin*** だけを返すようにしている。

## 4. 計算回数削減のための実装アルゴリズム

### ⑴　一致するクラスターの合わせこみ

計算回数 $K!$ を減らすために、前処理として完全一致クラスターの合わせこみを行う。

＜前処理＞

クラスター$\{_s\boldsymbol{RefCN}\}$とクラスター$\{_t\boldsymbol{CN}\}$に関して、

$$\{_s\boldsymbol{RefCN}\} = \{_t\boldsymbol{CN}\}$$

かつ

$$s \neq t$$

が成立するとき、$\boldsymbol{EXCH_{s,t}(CN)}$ を１回行ったクラスターベクトルを $\boldsymbol{CN_{(1)}}$ と表記することにしよう。これは、***RefCN*** のクラスターと一致する ***CN*** のクラスターのクラスター識別番号を合わせこむことに等しい。すべての ***RefCN*** のクラスター ($s = 1 \sim K$) に対して、一致する他の ***CN*** のクラスターについても「元交換操作」を繰り返し実施する。

最初の ***CN*** に対して、$u$ 回の「元交換操作」で＜前処理＞が完了したとき、終了時点での ***CN***（***CN*** の「元交換ベクトル」の１つ）を $\boldsymbol{CN_{(u)}}$ と呼ぶことにする。最大でＫ－１回で前処理は完了するが、$K-1$ 回前処理が実行できた場合は $u = K-1$ で、$\boldsymbol{RefCN} = \boldsymbol{CN_{(K\text{-}1)}} = \boldsymbol{CNmin}$ である。つまり、この場合は ***RefCN*** と $\boldsymbol{CN_{(K\text{-}1)}}$ は「クラスター差異度」が０である。これ以外の場合は、$u \leqq K-2$ で、$\boldsymbol{RefCN} \neq \boldsymbol{CN_{(u)}}$ である。＜前処理＞において、$u$ 回数行った「元交換操作」における $s$ の集合を $X$ とする。集合 $X$ の要素数は $u$ 個である。

### ⑵　差異度最小化計算回数削減の考え方

***RefCN*** と $\boldsymbol{CN_{(u)}}$ から各々の「展開行列 $K$-$N$」$EXPM(\boldsymbol{RefCN})$ と $EXPM(\boldsymbol{CN_{(u)}})$ を作成する。このとき、$i \in X$ において、

$$\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CN_{(u)}})_{ij})^2 = 0$$

つまり、

$$\sum_{i \in X}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CN_{(u)}})_{ij})^2\right) = 0$$

が成立する。

したがって (式 2) は、以下のように変形できる。

$$min\left(\sum_{i=1}^{K}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CN_{(u)}})_{ij})^2\right)\right)$$

$$= min\left(\sum_{i \in X}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CN_{(u)}})_{ij})^2\right)\right.$$

$$\left.+\sum_{i \notin X}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CN_{(u)}})_{ij})^2\right)\right)$$

$$= min\left(\sum_{i \notin X}\left(\sum_{j=1}^{N}(EXPM(\boldsymbol{RefCN})_{ij} - EXPM(\boldsymbol{CN_{(u)}})_{ij})^2\right)\right)$$

(式 3)

「元交換操作」は「展開行列」における「行交換操作」で置き換えることができることから、(式 2) の解 ***CNmin*** を求めるために $K!$ 回必要であった計算が、(式 3) によって $i \notin X$ なる行の中での「行交換操作」$(K-u)!$ 回に減らすことができる。$K-u$ は $X$ の補集合の要素の数である。比較を行いたいクラスター分析結果は比較的類似している場合が多いことが想定されるため、実効上は計算回数の削減がかなり期待できる。

$(K-u)!$ 個存在する $\boldsymbol{CN_{(u)}}$ すべてについて計算を行うことになる。Ｒ言語で計算を行う場合に

は $permutations(\,)$ 関数で $permutations(K-u, K-u)$ を求めて計算を行うことになる。R 言語のメモリサイズの制約により、$K-u>10$ の場合は $permutations(K-u, K-u)$ の戻り値が入る行列の領域を確保することができずに、エラーとなるため注意が必要である。計算回数を削減するためにもう少し削減のための工夫を考えることとしたい。次のような行列 $D$ を定義する。

$$D_{ij} = \sum_{h=1}^{N} (EXPM(\boldsymbol{RefCN})_{ih} - EXPM(\boldsymbol{CN})_{jh})^2$$

（式 4）

$D_{ij}$ は、$EXPM(\boldsymbol{RefCN})$ の $i$ 行目と $EXPM(\boldsymbol{CN})$ の $j$ 行目の各要素を引いたものの二乗和である。$D_{ij}$ の値は $\{_i\boldsymbol{RefCN}\}$ と $\{_j\boldsymbol{CN}\}$ の差集合と $\{_j\boldsymbol{CN}\}$ と $\{_i\boldsymbol{RefCN}\}$ の差集合の和、つまりクラスター間の非一致元の数である。以後、行列 $D$ を **RefCN** と **CN** の「**差異度行列**」と呼ぶことにする。行列 $D$ は、$K$ 行 $K$ 列の正方行列である。**RefCN** と **CN** の「クラスター差異度」$CD$ と $D_{ij}$ の関係は、

$$CD = \frac{1}{2} \times \sum_{i=1}^{K} D_{ii} \qquad \text{（式 5）}$$

が成り立つ（（式 4）において $D_{ii}$ として、（式 1）に代入すると導くことができる）。$D_{ij}$ の対角要素の和の 1/2 が「クラスター差異度」$CD$ と等しい。「クラスター差異度最小化法」は、$D_{ij}$ の列の交換操作によって対角要素の和が最小となる組み合わせを求めていることと同等である。

$D_{ij}=0$ の場合は、$D$ の $i$ 行には他に 0 となる要素は存在しない。これは $\{_i\boldsymbol{RefCN}\}$ と $\{_j\boldsymbol{CN}\}$ が一致クラスターであることを示しており、合わせこみ操作に該当する。前項ではこれが $u$ 個存在していたわけである。さらに、行列 $E$ を次のように定義する。

$$E = EXPM(\boldsymbol{RefCN}) \times (EXPM(\boldsymbol{CN}))^T$$

ここで $(EXPM(\boldsymbol{CN}))^T$ は $EXPM(\boldsymbol{CN})$ の転置行列である。$E_{ij}$ の値は $\{_i\boldsymbol{RefCN}\}$ と $\{_j\boldsymbol{CN}\}$ の共通元の数、つまりクラスター間の一致元（積集合）の数である。行列 $E$ を **RefCN** と **CN** の「一致行列」と呼ぶことにする。行列 $E$ も、$K$ 行 $K$ 列の正方行列である。

$D$ と $E$ の関係において次が成り立つ、

条件(1)：$D_{ij}+1 \leqq E_{ij}$ のとき、
$D_{ij}$ は $D$ の $i$ 行の要素の中で唯一最も小さい。

条件(1)を満足する $D_{ij}$ のうち $D_{ij} \neq 0$ であるものの数を $v$ とする。この場合 $\{_i\boldsymbol{RefCN}\}$ と $\{_j\boldsymbol{CN}\}$ は「最も類似したクラスター」である。このため、$D_{ij} \neq 0$ の中で $(i, j)$ が条件(1)を満足する場合は $\{_i\boldsymbol{RefCN}\}$ と $\{_j\boldsymbol{CN}\}$ を合わせこむことができる。つまり、$v$ がさらなる計算回数削減可能な数となる。つまり計算回数は $(K-u-v)!$ に削減可能である。クラスター差異度最小を厳密に求める手法として、$cluster\_identification(\,)$ 関数として実装した。

## 5.「Jaccard 係数 [4] を用いた近似的同定法」

集合 $A$ と $B$ の類似性を示す Jaccard 係数は、

$$|A \cap B| \,/\, |A \cup B|$$

で定義される。今回の場合であれば、

$$|\{_i\boldsymbol{RefCN}\} \cap \{_j\boldsymbol{CN}\}| \,/\, |\{_i\boldsymbol{RefCN}\} \cup \{_j\boldsymbol{CN}\}|$$

である。この Jaccard 係数の行列を作成して、Jaccard 係数の大きいクラスター対から順に同定する近似的な手法も評価し確認しておく。

今回の R 言語の実装では「差異度行列」$D$ と「一致行列」$E$ から Jaccard 係数を以下の式を用いて求めている。$\{_i\boldsymbol{RefCN}\}$ と $\{_j\boldsymbol{CN}\}$ の Jaccard 係数を $JAC_{ij}$ とすると、

$$JAC_{ij} = E_{ij} / (E_{ij} + D_{ij})$$

クラスター同定のアルゴリズムとしては、$JAC_{ij}$ の最大要素から順番に、***RefCN*** のクラスター $i$ と ***CN*** のクラスター $j$ とを同定してゆくものである。あくまでも近似的手法なので「クラスター差異度」の最小性は必ずしも保障されない。そのかわり、計算量としては $K$ の二乗のオーダーであり多項式時間で求解できる。この $JAC_{ij}$ を用いた「近似的同定法」は *Jaccard_identification*( ) 関数として実装しておいた。

## 6. 家計調査データのクラスター分析において適用した事例 [1],[3]

「その1」[2] 同様に、日本の47都道府県を12クラスターに分割する場合に適用してみよう。

### ⑴　2007年～2009年調査結果の場合

***RefCN*** と ***CN*** の「クラスター差異度」は18であるが、「クラスター差異度最小化法」で求めた ***CNmin*** では4に減少する。

### ⑵　2017年～2019年調査結果の場合

***RefCN*** と ***CN*** の「クラスター差異度」は13であるが、「クラスター差異度最小化法」で求めた ***CNmin*** では5に減少する。

今回検証に用いた家計調査データによるクラスター分析の事例では「クラスター差異度最小化法」で求めた解と、前稿 [2] の「クラスター重心間距離最小法」で求めた解、さらに「Jaccard 係数を用いた近似的同定法」で求めた解も全く同じとなった。異なる調査年次の家計調査データのクラスター分析結果間レベルの同定では、3つの手法で結果が異なることはあまりなさそうである。

しかし、求解の考え方が異なるため、これら3手法の結果が全て同じになる保証は無い。

## 7. おわりに

本稿で提案した「クラスター差異度最小化法」は、クラスター分析結果のみを用いてクラスター同定することが可能な点が特徴であるが多項式時間では求解できない点がデメリットとして挙げられる。一方、「Jaccard 係数を用いた近似的同定法」は多項式時間で求解可能な手法であり、「クラスター差異度最小化法」の計算量が膨大となり実用に耐えないような場合においても実行可能な手法である。クラスター分析結果のみを用いて多項式時間で求解できる「Jaccard 係数を用いた近似的同定法」は最も手軽で実用的と言えよう。

いずれの手法もR言語のスクリプトをエストレーラWeb注1に掲載しておくので、参考にしてもらいたい。表1に「その1」[2] と本稿で提案した3種のクラスター同定法の特徴をまとめておく。

表1　クラスター同定法の比較表

| | クラスター重心間距離最小法 | クラスター差異度最小化法 | Jaccard 係数を用いた近似的同定法 |
|---|---|---|---|
| 手法の概要 | ward 法の原理を同定にも活用 | 集合的同定の厳密解法 | 集合的同定の近似解法 |
| 計算量 | 多項式時間 | 多項式時間で解けない | 多項式時間 |
| 適用条件 | 元データ必要<br>次元合わせ必要 | クラスター分析結果があれば同定可能 | クラスター分析結果があれば同定可能 |

＊参考文献

[1] 木村敦・高部勲（2021）「家計消費データから見る日本の食料嗜好地域性～人文社会系知見との連携も見据えて～」,『ESTRELA』No. 324, pp. 36-42, 統計情報研究開発センター .

[2] 木村敦（2023）「異なるデータのクラスター分析結果を比較する～その1：類似クラスターのアラインメント「クラスター重心間距離最小法」～」,『ESTRELA』No. 350, pp. 26-31, 統計情報研究開発センター .

[3] ㈱統計センター「SSDSE（教育用費用準データセット）」https://www.nstac.go.jp/SSDSE/index.html

[4] Dr. Jaccard paul（1901）“Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines”, Bulletin de la Societe Vaudoise des Sciences Naturelles, January 1901.