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」～

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

異なる調査年次データのクラスター分析結果を、相互に比較することはよく行われることである。異なる調査年次の結果の相違点を比較するためには、類似したクラスターを正しく対応づけることが必要である。しかし、R 言語の *hclust*( ) 関数などの計算パッケージソフトでは、当然ながら他のクラスター分析結果との類似性を考慮したクラスター番号の付与を行うことはできない。このため研究者がクラスター分析結果を目視比較することによって、主観的にクラスター番号を振りなおす必要があった。

クラスター分析はいわゆる「集合の分割」である。今回の問題は、等しい数に分割された異なる集合分割結果のアラインメント問題と考えることができる。アラインメント判定手法次第では「組合せ最適化問題」になり、その多くは NP 困難となってしまう。「組合せ最適化問題」は現実的手法として数理最適化手法が用いられることが多い。現実的な時間で近似的に最適解を求める手法がいろいろと提案されている [1]。

本稿では、ward 法階層的クラスター分析の原理 [2] に着目し、異なる調査年次データに対する ward 法クラスター分析結果間における類似クラスターを多項式時間でアラインメントする手法を提案する。比較したいクラスター分析結果間の「クラスター重心間距離」が最小となるクラスター対を同定し、クラスター番号の振りなおしを行う手法「クラスター重心間距離最小法」である。本稿で定義した「クラスター重心間距離」は、2 つのクラスター分析結果間におけるクラスター重心間の距離である。今回提案する手法が適用できるための条件としては、

⑴ 参照用のクラスター分析に用いた元データと、同定の対象となるクラスター分析に用いた元データがともに利用可能であること。
⑵ 参照用の元データと同定対象の元データの変数（次元）が等しい（もしくは、合わせられる）こと。

である。これは、多くの研究で可能な条件であると考える。ward 法クラスター分析において用いられるユークリッド距離に基づく「非類似度」をクラスター類似同定にも活用する方式であり、納得性が高いと言えよう。これにより「組

合せ最適化問題」になることも回避している。

上記⑴と⑵の条件を満たさない場合においては、クラスター分析結果として出力される「クラスターベクトル」だけを用いたクラスター同定法を別途提案する。簡単に言えば、クラスター分析結果を「集合の類似性」で比較し、全体としての集合的な差異が最も少なくなるクラスター対を同定する方法である。これらの手法については別途、別稿「その2」として、多項式時間では求解できない厳密解法と多項式時間で求解できる近似解法について詳細を述べたい。

本稿では、今回提示する「クラスター重心間距離最小法」を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も示す。また、「クラスター重心間距離最小法」のアルゴリズムをR言語で実装したスクリプトもエストレーラWebで公開する[注1]。

2. 用語の定義

同一の観測対象・変数に対して異なる時点で測定したデータが存在する場合で考える。複数の調査時点における多変量データの1つを「参照用データ」と定め、そのほかの調査時点における多変量データを「対象データ」と呼ぶことにする。観測数N、変数Pとすると、この場合すべての多変量データは要素値の異なるN行P列の行列で表現できる。通常、クラスター分析を行う際には、これらの多変量データを標準化したデータを用いて距離行列を作成する。この距離を算定するために用いる標準化したデータを「標準化データ」と呼ぶことにする。「参照用データ」の「標準化データ」行列を*RefND*、「対象データ」の「標準化データ」行列を*ND*とする。いずれもN行P列の行列である。

N個の観測値をK個のクラスターに分ける場合は、N個の観測値から成る集合をK個の部分集合に分割することと等しい。クラスター分析の結果はN次元のベクトルで表すことができる。このN次元のベクトルは、1からK（$K \leqq N$）までの連続した自然数（＝クラスター識別番号）のすべての数字をベクトルの要素の値として保持するベクトル$\boldsymbol{A} = (A_1, A_2, \cdots, A_N)$である。このベクトル$\boldsymbol{A}$を「**クラスターベクトル*N-K***」、このベクトルの集合を「**クラスターベクトル*N-K*集合**」と呼ぶことにする。以後本稿では、ベクトルを太文字で表記する。

「クラスターベクトル*N-K*」$\boldsymbol{A} = (A_1, A_2, \cdots, A_N)$の各要素の添え数字を「**要素次数**」と定義する（要素A_1の「要素次数」は1、要素A_mの「要素次数」はmである）。「要素次数」はN個の観測値を区別する識別番号と考えることができる。上記$\boldsymbol{A} = (A_1, A_2, \cdots, A_N)$の要素の中で、同じクラスター識別番号を保持するすべての要素の「要素次数」の集合を**クラスター**と呼ぶ。「クラスターベクトル*N-K*」にはK個のクラスターが存在する。$\boldsymbol{A}$のすべての「要素次数」からなる集合を$\{\boldsymbol{A}\}$と記載するとき

$$\{\boldsymbol{A}\} = \{1,2,\cdots,N\}$$

である。クラスターは$\{\boldsymbol{A}\}$の部分集合であり、識別番号がiであるクラスターを$\{_i\boldsymbol{A}\}$と記載する。例えば、識別番号2のクラスター$\{_2\boldsymbol{A}\}$が観測値の6番目と9番目と20番目のみを含むとすれば、

$$\{_2\boldsymbol{A}\} = \{6, 9, 20\}$$

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。

である。また、集合 $\{\boldsymbol{A}\}$ の要素数を $\left|\{\boldsymbol{A}\}\right|$ で表す。直前の例で言えば、

$$\left|\{\boldsymbol{A}\}\right| = N$$
$$\left|\{_2\boldsymbol{A}\}\right| = 3$$

である。

集合の分割であるから、

$$\{\boldsymbol{A}\} = \{_1\boldsymbol{A}\} \cup \cdots \cup \{_i\boldsymbol{A}\} \cup \cdots \cup \{_K\boldsymbol{A}\}$$

また、$i \neq j$ の場合、

$$\{_i\boldsymbol{A}\} \cap \{_j\boldsymbol{A}\} = \Phi$$

である。また、定義から明らかなように、$\{_i\boldsymbol{A}\}$ に含まれる $\boldsymbol{A}$ の要素が保持する値（クラスター識別番号）は i である。これを、

$$[\{_i\boldsymbol{A}\}] = i$$

と記載することにする。このとき、$i \neq j$ ならば、当然ながら

$$[\{_i\boldsymbol{A}\}] \neq [\{_j\boldsymbol{A}\}]$$

である。

3. 重心間距離最小法の基本的考え方

行列 ND の n 行目の要素からなる P 次元ベクトルを $\boldsymbol{NDn}$ とする。ND のクラスター分析結果における i 番目のクラスターに関する「クラスター i 重心ベクトル」$_i\boldsymbol{CCV}$ を以下のように定義する。ここで、$\boldsymbol{CN}$ は行列 ND の「クラスターベクトル N-K」であるとする。

$$_i\boldsymbol{CCV} = \left(\sum_{m \in \{_i\boldsymbol{CN}\}} \boldsymbol{NDm} \right) \div \left|\{_i\boldsymbol{CN}\}\right| \quad \text{（式 1）}$$

$_1\boldsymbol{CCV}$ から $_K\boldsymbol{CCV}$ を縦に並べた K 行 P 列の行列を「クラスター重心行列」$CCVM$ と定義する。$CCVM$ の i 行 j 列の要素を $CCVM_{ij}$ とし、$_i\boldsymbol{CCV}$ の j 番目の要素を $_i\boldsymbol{CCV}_j$ とすると、

$$CCVM_{ij} = {}_i\boldsymbol{CCV}_j \quad \text{（式 2）}$$

である。$RefND$ についても同様に $RefCCVM$ を定義する。ここで、$\boldsymbol{RefCN}$ は行列 $RefND$ の「クラスターベクトル N-K」であるとする。

$$_i\boldsymbol{RefCCV} = \left(\sum_{m \in \{_i\boldsymbol{RefCN}\}} (\boldsymbol{RefND}m) \right) \div \left|\{_i\boldsymbol{RefCN}\}\right| \quad \text{（式 3）}$$

$_1\boldsymbol{RefCCV}$ から $_K\boldsymbol{RefCCV}$ を縦に並べた K 行 P 列の行列を「クラスター重心行列」$RefCCVM$ と定義する。$RefCCVM$ の i 行 j 列の要素を $RefCCVM_{ij}$ とし、$_i\boldsymbol{RefCCV}$ の j 番目の要素を $_i\boldsymbol{RefCCV}_j$ とすると、

$$RefCCVM_{ij} = {}_i\boldsymbol{RefCCV}_j \quad \text{（式 4）}$$

である。ここで「**クラスター重心距離行列**」CCD を以下のように定義する。CCD は K 行 K 列の行列である。ここで、$sqrt(x)$ は x の平方根である。

$$CCD_{ij} = sqrt\left(\sum_{m=1}^{P} (RefCCVM_{im} - CCVM_{jm})^2 \right) \quad \text{（式 5）}$$

CCD_{ij} は、*RefND* をクラスター分析した結果における i 番目のクラスター重心と *ND* をクラスター分析した j 番目のクラスター重心との間の N 次元ユークリッド距離である。

「クラスター重心間距離最小法」のアルゴリズムは以下のとおりである。

⑴ *CCD* の中の最小値の要素の 1 つ（最小値が複数存在した場合はどれか 1 つを選ぶ）が CCD_{xy} であるとき、*RefND* のクラスター分析結果 ***RefCN*** のクラスター x と *ND* のクラスター分析結果 ***CN*** のクラスター y を「類似クラスター」であると同定する。

⑵ 引き続き、*CCD* の x 行および y 列のすべての要素を除いた残りの *CCD* の要素の中の最小値が CCD_{uv} であるとき、*RefND* のクラスター分析結果 ***RefCN*** のクラスター u と *ND* のクラスター分析結果 ***CN*** のクラスター v を「類似クラスター」であると同定する。

⑶ 以後同じ操作⑵を繰り返す。K 対のクラスター同定が完了すれば終わりである。

ここでは、行列 *ND* と行列 *RefND* の列数がともに P である場合で説明したが、列数が異なる場合は、*CCD* の算出に先立って列数を合わせておく必要がある。家計調査の場合は、調査年次によって調査品目の分割・統合が行われることがある。このため、異なる調査年次では列数が異なる場合がありうる。しかし、調査品目の分割・統合の推移が明らかにされているため、調査品目数を合わせこむことが可能である。次項の例示においても *CCD* を算定する前に *RefND* と *ND* の列数の合わせこみを実施している。

4. 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（参考文献 [3],[4]）

総務省統計局の家計調査の異なる年次のデータを用いたクラスター分析結果の類似クラスター同定に適用してみよう。家計調査の食料品目を抜き出して分析を行うこととする。食料品目数は調査年次によって変動がある。変数の次元をそろえるために、品目が少ない年次のデータに次元を合わせて元データを加工しておく。

日本の 47 都道府県を 12 クラスターに分割する場合で考える。これは「クラスターベクトル *47-12* 集合」の 1 つの元として表すことができる。都道府県のクラスター分析の結果を表現する場合には、「クラスターベクトル *47-12*」の各要素の値に 1 から 12 のどれかを割り当てることになる。同じクラスターに含まれる都道府県に対応する要素の値は同じ値に割り当てる。

R 言語の *hclust*( ) 関数の出力に対して、クラスター数を 12 に指定して *cutree*( ) 関数にかけることによって、「クラスターベクトル *47-12*」が得られる。各要素の値をパターン番号に対応させて日本地図に表現すれば、都道府県がクラスター番号毎に 12 パターンに塗り分けされた日本分県地図ができあがる。冒頭述べたとおり、*hclust*( ) 関数の複数の出力ベクトル相互間では「クラスター重心間距離最小」は必ずしも成り立っていない。

⑴ 2007 年～ 2009 年分析結果との比較

図 1 は、2017 年～ 2019 年の総務省統計局の家計調査データのうち、食料品目（213 品目）の世帯別年間支出額（3 年間の平均値）を用いて ward 法クラスター分析を行った結果である [3]。

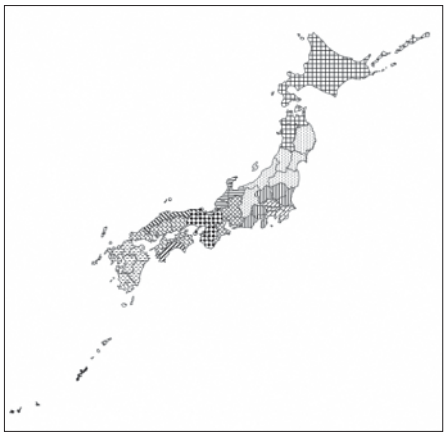

図１　木村・高部 [3] の図３

一方、図２は、家計調査の中から 2007 年～ 2009 年の３年間の調査結果（食料 207 品目）を平均したデータを用いて ward 法クラスター分析を行った結果である。図１と比べてみると、三重県が東海３県地域クラスターから瀬戸内地域クラスターに移ったために、瀬戸内地域クラスターの採番にずれが生じた。この影響で、近畿地域クラスターや山陰地域クラスターにも採番にずれが波及した。これによって日本区分地図における塗り分けパターンの割り当てにも影響が出てしまっている。近畿地域クラスターはクラスター自体には変更がないにもかかわらず、図１と図２では全く異なるパターンで表示されてしまっている。

図２の「元データ」に対して「クラスター重心間距離最小法」を用いて描画しなおしたものが図３である。「クラスター重心間距離最小法」を適用するにあたって「参照用元データ」として図１の元データを 207 品目に集計しなおしたものを用いている。図２と図３は全く同じクラスター分析結果なのであるが、「クラスター重心間距離最小法」を適用して図１に対する類似クラスターに同じ採番を行った図３で比較するほうが、図２と図１を比較するよりも分析結果の比較がしやすいことが分かる。

図１と図３を比べると、岩手県、長野県、静岡県、三重県だけがそれぞれ異なるクラスターに移動したことがよく分かる。

図２

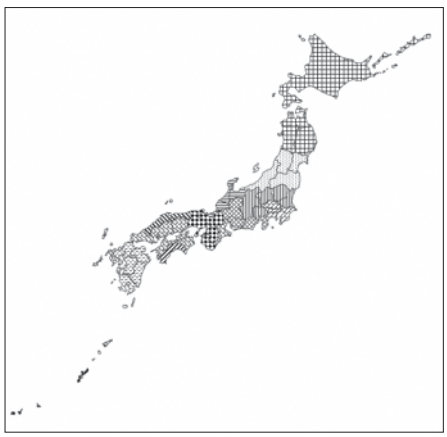

図３

⑵　2012 年～ 2014 年分析結果との比較

図４は 2012 年～ 2014 年の家計調査データのうち食料品目（208 品目）の世帯別年間支出額（3 年間の平均値）を用いて同様に ward 法クラスター分析を行った結果である。

図１と見比べると西日本エリアの塗り分けパターンが大きく異なっているが、これも類似ク

ラスターに異なるクラスター番号が付与されてしまっているために発生する事象である。

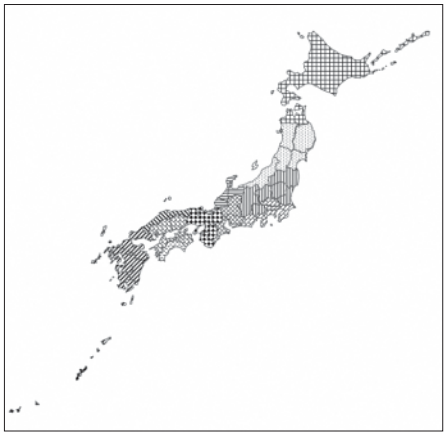

図４

これを改善するために、図４の「元データ」を用いて「クラスター重心間距離最小法」を適用してクラスター番号を採番しなおしたものが図５である。

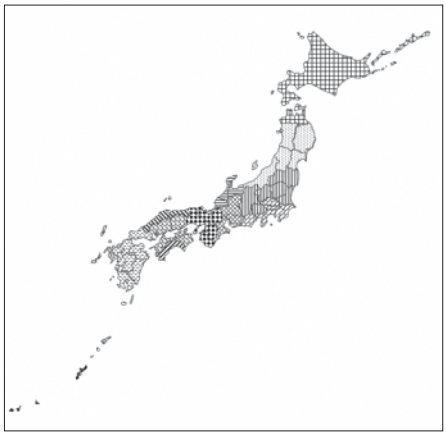

図５

図４と図５も全く同じクラスター分析結果なのであるが、「クラスター重心間距離最小法」を適用して図１に対する類似クラスターに同じ採番を行った図５で比較するほうが、図４と図１を比較するよりも分析結果の比較がしやすいことが分かる。秋田県、福島県、長野県、静岡県、山口県だけがそれぞれ隣接するクラスターに移動していることが一目で分かる。「対象元データ」と「参照元データ」の変数次元が不一致であるため、213 品目の「参照元データ」を「対象元データ」の 208 品目と合わせるために、図1の元データを208品目に編集しなおして「参照用元データ」として使用している。

以上のように「クラスター重心間距離最小法」を用いて相互の分析結果間の類似したクラスターのクラスター番号を合わせることによって、比較したいクラスター分析結果間の違いを把握することが容易になる。クラスター間を移動した都道府県だけを彩色して表示させるなどの処理も適切に行うことが可能となる。

5. おわりに

複数の異なる条件で作成したクラスター分析結果の相互比較を行う際に、類似したクラスターを多項式時間で同定して同じクラスター番号を採番するアルゴリズム「クラスター重心間距離最小法」を提案した。さらに、本手法を木村・高部 [3] の結果に適用し、良好な結果が得られることを検証した。

「クラスター重心間距離最小法」を R 言語で実装した例を関連データとともにエストレーラ Web[注1] に掲載しておいたので、参考にしていただければ幸いである。

＊参考文献

[1] 梅谷俊治（2020）『しっかり学ぶ数理最適化　モデルからアルゴリズムまで』講談社 .

[2] Joe H. Ward, Jr.(1963)“Hierarchical Grouping to Optimize an Objective Function”, Journal of the American Statistical Association, Vol. 58, No. 301 (Mar., 1963), pp.236-244.

[3] 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 .

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