様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１４　日現在

機関番号：　３２６４３
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７８０２４０

研究課題名（和文）大規模マーケティングデータへのクラスタリング手法の適用

研究課題名（英文）An Application of Clustering Analysis for lage scale marketing data

研究代表者
横山　暁（Yokoyama, Satoru）

帝京大学・経済学部・講師

研究者番号：９０５８２８６７

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：本研究では，1つの対象が複数のクラスターに所属することを許容した重複クラスター分析法に大規模なデータを適用して新た知見を得ようとする取り組み，またそのための分析法の理論的研究について行った．
その結果，従来の分析では得ることのできない結果の解釈を得られる可能性があることを確認することができた．また理論的研究については，大規模なデータを分析するにあたり，分析のアルゴリズムにいくつかの問題点があることが分かり，その改善案を提案するとともに，研究を継続している．

研究成果の概要（英文）：In this research, I focused on the overlapping cluster analysis which is one of the cluster analysis and is allowed the object can belong to more than one cluster. I tried to analyze the large scale data using overlapping cluster analysis, and to get new results which could not get by traditional analysis. In addition, I performed a theoretical research of overlapping cluster analysis which which is applied to the large scale data.
As the result, I confirmed the analysis can obtained the new interpretation. However, I discovered several problems in algorithm of overlapping cluster analysis, so I suggested some improvement of these problems. Now I write articles based on the result of research and continue a study for further improvement of the analysis.

研究分野：　データマイニング

キーワード：　クラスター分析　親近度データ　マーケティングデータ

１．研究開始当初の背景

本研究開始時点において大規模データ（ビッグデータ）という言葉に関心が寄せられ，大規模データが取得・蓄積され，分析する必要性に着目が集まっていた．それ以前よりデータから何らかの知見を得ようとする「データマイニング」は行われていたが，大規模なデータを従来のデータマイニング手法にどう適用するかが一つの課題であった．そこで本研究において，研究代表者が主として研究している「重複クラスター分析法」に大規模データを適用することで，新たな知見を得ようとすることはもちろんのこと，分析法そのものの理論的研究を行う必要性があった．

２．研究の目的

上記「1. 研究開始当初の背景」でも触れたように，大規模データの取得・蓄積・分析の必要性が高まっている状況において，本研究では，大規模データの中でも特にマーケティングデータに着目し，マーケット・セグメンテーションを行うことを目的とするクラスタリング手法を用いて大規模データを分析し結果を解釈することを目的の一つとした．

しかしながら，従来から用いられているクラスタリング手法は，大規模なデータを分析することを目的とはしていない．例えば，クラスタリング手法で一般的な階層的クラスター分析法 (Johnson, 1967) は，対象間の類似度を逐次比較するアルゴリズムであるため，大規模なデータを扱うことには不向きである上，その結果を示すデンドログラムは対象数が多くなることで視認性が著しく低下するという問題点がある．また，階層的クラスター分析法に加え，非階層クラスター分析法である K-means 法 (MacQueen, 1967) を含めた従来のクラスター分析法の多くは，1 つの対象が 1 つのクラスターに分けられるという性質は，現実的ではないと考えられた．

そこで本研究では，第一に，大規模マーケティングデータに対し，クラスタリング手法，特に 1 つの対象を複数のクラスターに所属することを許容している「重複クラスター分析法」を用い，有益な情報を得ようとすること，またそのための方法論の研究を行うこととした．さらに，実際のマーケティングデータに適用することで，従来の分析では得ることのできなかったような新たな知見を得ることを第二の目的とした．

本研究で扱う重複クラスター分析法は，Shepard & Arabie (1979) および Arabie & Carroll (1980) によって提案されたクラスター分析法の 1 つであり，1 つの対象が複数（より厳密には 0 以上）のクラスターに所属することができる分析法である．従来のクラスター分析法の多くと比較して，現実の状況に近い結果を得ることができると考えられ，大規模データへの適用によって新たな知見が得られると考えられる．

３．研究の方法

研究は以下の要領で行った．

(1) 大規模マーケティングデータ分析の基礎研究

大規模マーケティングデータに対して重複クラスター分析法を適用するにあたり，大規模データ分析の基礎的な研究について調査･検討を行った．主に Mayer-Schoenberger & Cukier (2014) や O’Neil, & Schutt (2013) などの文献や，各種学会発表・論文から情報を収集するとともに，最新のクラスター分析法の情勢を Everitt, et al. (2011) などにより調査した．

⑵ 重複クラスター分析に関する理論的研究

Shepard & Arabie (1979) および Arabie & Carroll (1980) によって提案された重複クラスター分析法では，分析には対象間の親近性を表すデータを用いるが，多くの対象間の親近性を表すデータを分析する場合，計算時間の増大や結果の不安定性，つまりアルゴリズム上の問題点があることが想定される．

そこでこの問題を詳細に把握し，解決するために，多くの対象数をもつ人工データ等を用いてシミュレーションを行う．具体的には，人工的に作成したクラスター構造から，重複クラスター分析法のモデル式を用いて親近度データを算出し，そのデータを分析することで，結果の再現性・安定性を評価するとともに，計算時間についても算出する．この手続きをいくつかの特徴的なクラスター構造で繰り返し行うことにより，アルゴリズム上の問題点を検出する．

この研究より，アルゴリズムの改善・改良案を提案し，分析用プログラムに実装していく．

⑶ 大規模マーケティングデータの取得・基礎集計と分析，評価

上記⑵の研究と並行して，大規模マーケティングデータを取得し，基礎集計を行った上で，重複クラスター分析法やその他のクラスター分析法を用いて分析する．

具体的なデータとしては，スーパーやコンビニエンスストアの購買記録データやアンケート調査のデータを用い，商品・商品カテゴリーや被験者間の親近度データを集計によって作成する．

作成されたデータを，重複クラスター分析法に加え，いくつかのクラスター分析法等により分析し比較することで，新たな知見を得られるかどうかについて評価を行っていく．

４．研究成果

上記３の研究の方法に基づき研究を実施た．特に学会発表等においては，⑵の理論的研究について実施し研究発表を行った．

まず，従来の重複クラスター分析法のアルゴリズムについて検討を行った．Arabie & Carroll（1980）によって提案された MAPCLUS（MAthematical Programming CLUSterng）と呼ばれるアルゴリズムは，2 つの段階から構成されており，第 1 段階は主として交互最小二乗法を用いて，クラスター構造と各クラスターの重みを算出する段階，第 2 段階は組み合わせ最適化として，クラスター構造を少しずつ入れ替えながら，より良いクラスター構造を探索していく段階となっている．

まず，各段階における処理の問題点について検討するために，対象数が 15 程度と比較的少ない購買記録データと，対象数が 60 程度と比較的多いアンケート調査のデータを用い，通常の分析（第 1 段階＋第 2 段階）と比較して，第 1 段階のみ，第 2 段階のみの分析で，計算時間はどの程度かかるかを調べた．

その結果，クラスター数等の条件にもよるが，対象数が少ない場合においては，通常の分析より第 1 段階のみ，第 2 段階のみともに 50%～70%程度の計算時間で許容できる結果が得られたが，対象数が多い場合においては，第 1 段階のみの場合では通常の分析の 3%～5%の計算時間であったが，第 2 段階のみ場合では 1.2 倍～1.5 倍程度の計算時間がかかることが判明した．但し得られた結果の妥当性についての検討が不十分であり，この点についてを踏まえたさらなる研究が必要であることも判明した．なお，この研究は日本行動計量学会第 42 回大会において「重複クラスター分析法を用いたマーケティングデータの分析とその課題」として発表を行った．

さらに，国際学会である IFCS2015 において，「A study on the overlapping cluster analysis for the large data」というタイトルで発表を行った．重複クラスター分析法では，分析の初頭に用いる初期クラスター構造によっても結果が変わってしまう，つまり局所解に陥ってしまうという問題点がある．通常，この問題点については，初期クラスターを様々に変え分析を行い，分散比と呼ばれる指標を用いて評価をすることで解決を図る．この問題点について本研究で検討を行った．さらに，MAPCLUS のアルゴリズムにおける 2 つの段階の組み合わせに関して，計算時間及び分散比が許容できる結果であるかどうか（最良の結果に対して一定程度の分散比であるかどうか）についても検討を行った．

分析では，まず実データを用いた分析を行った．その結果，対象数が比較的少ない場合，通常の分析（第 1 段階＋第 2 段階）では許容できる結果が 9 割程度と十分な結果であったが，第 1 段階のみの場合，計算時間は約半分程度であったが許容できる結果が極端に少ない結果となった．また，第 2 段階のみの場合における計算時間は 6 割程度で許容できる結果は 2 割程度という結果であった．一方，対象数が比較的多い場合では，計算時間においては，第 1 段階のみにおいては 3%程度，第 2 段階のみでは 140%程度と大きく差がつく結果であったが，許容できる結果の傾向は対象数が比較的少ない場合とほぼ同一であった．

これらの結果を踏まえ，さらに人工データによるシミュレーションを行った．シミュレーションでは対象数を 10, 25, 50, 100 の 4 種類とし，クラスター数と各クラスターの重みは固定した．また，正しいクラスター構造はパターン化したものとランダムなものを設定した上で重複クラスター分析法のモデル式より親近度データを逆算し，分析を行った．その結果，クラスター構造にかかわらず第 1 段階のみの分析では通常の分析と比較して対象数が増加するとともに計算時間は減少するものの，許容できる結果が著しく減少するという結果が得られ，第 2 段階のみの分析では通常の分析と比較して許容できる結果は 4 割～5 割程度であり，計算時間は対象数が増えるにつれ 150%程度に増加するという結果が得られた．これにより，対象数において第 1 段階と第 2 段階を組み合わせる従来のアルゴリズムが最適であることが確認されると同時に，第 1 段階におけるクラスター構造と重みの算出における繰り返し数を変化させた際に計算時間や結果に変動がみられるため，第 1 段階のアルゴリズムの詳細な検討が必要であることも確かめられた．また，本発表における質疑等において，アルゴリズムに関する新たなアイディアを得ることができ，本報告書作成時点において継続的に研究を行っている．

次に，日本計算機統計学会第 30 回大会において「重複クラスター分析法における非小規模データの分析」というタイトルで発表を行った．この研究では，これまでの研究を踏まえ，アルゴリズムの内部に踏み込んだ研究を行った．具体的には，MAPCLUS のアルゴリズムの第 1 段階の繰り返し数に着目し，繰り返し数が結果の妥当性や安定性に影響があるかを検証した．MAPCLUS の第 1 段階の交互最小二乗法を用いてクラスターを推定する際の繰り返し数の問題は，類似するアルゴリズムを用いている多次元尺度構成法のアルゴリズム（Carroll and Chang, 1970; INDSCAL）においても課題となっており，繰り返し数を多くしてしまうと，多次元尺度構成法で出力される布置が退化してしまうことがある．退化とは対象を表現する点が重なり合い，少数の点の近くに集中してしまうことであり，結果として分析から有益な情報

が得られないことになる．重複クラスター分析法においても，退化が起こることで少数の対象のみがクラスターに所属するという現象が起きてしまうことが考えられる．また，逆に十分な繰り返し数を設定しないと，収束が不十分となっている可能性も考えられる．そこで本研究では，繰り返し数を変更して分析を行うことで，得られた結果の退化性および収束性を検討することとした．なお，分析には主として前述の人工データおよび購買記録データとアンケート調査のデータを用いた．

その結果，対象数が少ない場合には，数回の繰り返し数で交互最小二乗法が完了しており，収束性および退化に関しては問題ない結果となっていた．しかし，対象数が多い場合，初期クラスターによっては 10 回程度では収束しないケースも散見され，いくつかの解をループしているケースも確認された．また，結果によっては退化に近い状況と判断される構造もあった．

この研究より，対象数によっては現状のアルゴリズムでは十分な収束がなされないということが判明したことになる．

また，これらの本研究課題の研究を通し，実データの分析において，従来のクラスター分析との比較も行った（研究の方法の(3)に対応）．従来の分析では，1 つの対象は 1 つのクラスターにしか所属しないため，簡潔な結果が得られる一方，実際に即したクラスター構造とは言えない状況であった．例えば，比較的小規模な大砲数である購買記録データの場合においても，階層的クラスター分析法では図 1 のようなデンドログラムが出力される．一方で重複クラスター分析法を行うと，例えばクラスター数 5 の結果の場合，表 1 となる．

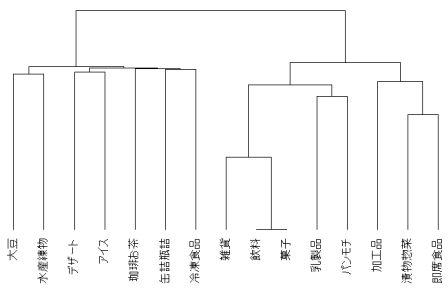

図 1 階層的クラスター分析法の結果

表 1 重複クラスター分析法の結果

	クラスター				
	1	2	3	4	5
大豆	0	0	0	0	0
漬物惣菜	1	1	0	0	0
水産練物	1	0	0	0	0
加工品	1	0	0	0	0
乳製品	1	0	0	1	0
デザート	0	0	1	0	0
飲料	1	1	1	1	1
即席食品	1	1	0	0	0
缶詰瓶詰	0	0	0	0	0
パンモチ	0	1	0	1	0
珈琲お茶	0	0	0	0	0
菓子	1	1	1	1	1
雑貨	1	1	0	0	1
冷凍食品	0	0	0	0	0
アイス	0	0	1	0	0
重み	0.57	1.14	0.67	1.24	2.56

これより，階層的クラスター分析法では，冷凍食品と雑貨を境目として左右で大きく 2 つのクラスターに分かれることが分かる一方，重複クラスター分析法では，第 1 クラスターが図 1 の右側のクラスターにほぼ対応するのに加え，飲料やお菓子が全てのクラスターに所属する一夫，大豆や缶詰瓶詰，冷凍食品は度のクラスターにも所属しないという結果を得ることが分かる．つまり，従来の分析では決して得られることがなかった結果を得ることができることが確認された．大規模なデータの結果は紙面の都合上割愛するが，同様の傾向を見て取れる．

以上より，本研究において，重複クラスター分析法に大規模なデータを適用する際の理論面，特に MAPCLUS のアルゴリズム面においては，

(1) 第 2 段階の所要時間増大する問題点があることが判明した
(2) 第 1 段階の繰り返し数の設定，また収束性に課題があることが判明した

本研究では，初期クラスター構造をいくつも与えることで，一定程度満足いく結果を得ることができていると考えている．

しかし，この課題は今後改善していく必要があることも事実であり，課題に関して，既存研究よりいくつかのアルゴリズムの改良案を提示可能であると考えているが，研究当初想定していた以上に抜本的な改良が必要と考えられる．この点に関しては今後も継続して研究を行っていく予定である．

また，重複クラスター分析法を適用することよって新たな知見を得るという目的に関しては，一定の成果があったと考えられる．適用するデータや目的によって，必ずしも従

来の分析より良い結果が得られるとは限らない．しかし，上記の例にもあるように，新たな知見を得るという観点においては，効果的な分析であると考えられ，実際のマーケティングデータ等に積極的に適用していく必要があると考えている．

しかし，重複クラスター分析法の結果は，クラスター数だけ，クラスター構造が出力されるとともに，各クラスターに重みが付与される．対象数が増えるにつれ，クラスター構造が複雑になる上，重みの解釈も容易ではなくなるという点が問題である．この結果の表現方法については，今後の 1 つの研究課題としていきたい．

＜ここまでの引用文献＞

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 45*, 211-235.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an *N*-way generalization of ``Eckart-Young'' decomposition. *Psychometrika, 35*, 283-319.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis (5th Edition)*. UK., Wiley.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241-254.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1* (pp. 281-297). Berkeley, CA: University of California Press.

Mayer-Schoenberger, V., & Cukier, K. (2014). *Big data: A revolution that will transform how we live, work, and think*. Boston, Mariner.

O'Neil, C., & Schutt, R. (2013). *Doing data science*. O'Reilly Media, Inc..

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*, 87-123.

５．主な発表論文等

〔学会発表〕（計　3　件）

[1] **横山暁** (2014). **重複クラスター分析法を用いたマーケティングデータの分析とその課題［要旨］. 日本行動計量学会第 42 回大会抄録集**, pp. 66-67.

[2] Yokoyama, S. (2015). A study on the overlapping cluster analysis for the large data [summary]. Proceedings of Conference of the International Federation of Classification Societies (IFCS 2015), pp. 173-174.

[3] **横山暁** (2016). **重複クラスター分析法における非小規模データの分析［要旨］. 日本計算機統計学会第 30 回大会講演論文集**, pp. 55-56.

６．研究組織

(1)研究代表者

横山　暁　（YOKOYAMA, Satoru）

帝京大学・経済学部・講師

研究者番号：90582867