変数間の関係性を考慮してクラスター数を決定する k-means 法の改良

豊田 秀樹 池原 一哉 早稲田大学

An improved method using k-means to determine the optimal number of clusters, considering the relations between several variables

Hideki Toyoda and Kazuya Ikehara (Waseda University)

In this article, we propose a non-hierarchical clustering method that can consider the relations between several variables and determine the optimal number of clusters. By utilizing the Mahalanobis distance instead of the Euclidean distance, which is calculated in k-means, we could consider the relations between several variables and obtain better groupings. Assuming that the data are samples from a mixture normal distribution, we could also calculate Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) to determine the number of clusters. We used simulation and real data examples to confirm the usefulness of the proposed method. This method allows determination of the optimal number of clusters, considering the relations between several variables.

Key words: clustering, number of clusters, Mahalanobis distance, mixture distribution.

The Japanese Journal of Psychology
2011, Vol. 82, No. 1, pp. 32-40

心理学の調査・実験研究においては、因子分析や分散分析など、変数の性質を調べる手法が一般的に利用されてきた。近年では変数の性質だけでなく、個体（個人）の分類を行うことも重要となり、特に多変量を考慮した個体の分類は有効な分析手法である。荒木（2008）は防衛的悲観主義尺度の四つの下位尺度の因子得点を用いて、また、大内・長尾・櫻井（2008）は自己制御機能尺度の四つの下位尺度の標準得点を利用して k-means 法によるクラスター分析を行っている。このように、個体の分類は、心理学の分野において頻繁に行われている。

代表的な分類手法である k-means 法は、標本サイズが大きい場合の解析において非常に有効であり、Lloyd（1982）、Forgy（1965）、MacQueen（1967）、Hartigan & Wong（1979）などによって、これまで様々なアルゴリズムが提案されてきた。k-means 法の基本的な考え方は、まず初期条件としてクラスターを形成する核となる個体（シード点）を与え、次にその

シード点と残りの個体との距離を計算し、それを基準にして分類を行う。そして、設定したある収束条件を満たすまでシード点の変更を繰り返す。一般的に、距離の計算にはユークリッド距離が用いられる。

k-means 法のように個々の個体があるクラスターに属するか否かのみを考慮するのではなく、どの程度所属するかを考慮するファジィ c-means 法（Bezdek, 1980; Dunn, 1974）も提案されている。さらに、混合正規分布モデルによって分類を行う手法も提案されている。この手法では、データを混合分布に従う母集団からの標本とみなして、混合比率と各クラスターの平均ベクトル、分散共分散行列を推定する。その際、EM アルゴリズム（Dempster, Laird, & Rubin, 1977）を用いて不完全データに基づく最尤推定値をある繰り返し計算を行うことによって得る。具体的には、観測された n 個の p 変量データを、 K 個の多変量混合正規分布

$$f(x) = \sum_{k=1}^K \pi_k \varphi(x; \mu_k, \Sigma_k) \quad (1)$$

からの標本と見なす。 π_k は k 番目のクラスターの構成比率であり、 $\sum_{k=1}^K \pi_k = 1$ 、 $\pi_k > 0$ である。また、 $\varphi(x; \mu_k, \Sigma_k)$ は、平均 μ_k 、分散共分散行列 Σ_k に従う正規分布の確率密度関数を表す。混合正規分布モデ

Correspondence concerning this article should be sent to: Kazuya Ikehara, Graduate School of Letters, Arts and Sciences, Waseda University, Toyama, Shinjuku-ku, Tokyo 162-8644, Japan (e-mail: k.ikehara@ruri.waseda.jp)

ルでは、データ x_i の k 番目のクラスター C_k への帰属を示す変数 z_{ik} を導入する。変数 z_{ik} は未知であり、これを欠損値と考え、EM アルゴリズムによって、対数尤度関数 $l(\theta; x)$ を最大にするパラメータ θ を推定する。また、最適なクラスター数 K を求めるためのモデル選択基準として、AIC や BIC を算出する。しかし、EM アルゴリズムは標本サイズが大きいたく時には、尤度の局所最大値に収束してしまう可能性もあるので、注意が必要である。

ユークリッド距離を用いる k -means 法と比べて、混合正規分布モデルは、各クラスターの分散共分散行列を推定するため、クラスター内の変数間の関係性を考慮して分類を行うことができる。実際の調査・実験場面において、変数間に関係性がないデータが得られることのほうが稀である。したがって、変数間の関係性を考慮できるマハラノビス距離を用いるほうがより適切である。

最近では、 k -means 法のアルゴリズムを拡張させた x -means 法という手法が提案されている (石岡, 2000; Pelleg & Moore, 2000)。 x -means 法は、はじめに十分に小さなクラスター分割から始めて、各サブクラスターについて 2 分割が適切であると判断される限り分割を繰り返すというものである。具体的には、まず、 n 個の p 変量データを k_0 個のクラスターに分割する。次に、分割されたクラスター $C_j (j = 1, \dots, k_0)$ において、さらに k -means 法により、 C_j^1 と C_j^2 の二つのクラスターに分割する。ここで、 C_j に含まれるデータに多変量正規分布を仮定し、 C_j を分割しないモデルと C_j^1 と C_j^2 に 2 分割したモデルのそれぞれにおいて BIC を算出する。2 分割モデルの BIC が小さければ、さらに分割を続け、そうでなければ分割を終了する。この操作を、最初に k_0 個に分割したすべてのクラスターに対して行う。なお、最終的に算出される BIC を用いて、適切なクラスター数の決定を行う。

また、 x -means 法に改良を加えた方法に、石岡 (2006) がある。石岡 (2006) は、 x -means 法の 2 分割手続きによって生成されたクラスターを再評価し、必要ならばそれらのクラスターを併合するという手続きを加えることで、 x -means 法の性能が格段に向上することを、シミュレーションによって示している。

これまでの研究において、 x -means 法は非常に有効な手法であることが示されてきた。ただし、各サブクラスターで 2 分割しか許されないという本質的な制約がある。したがって、各サブクラスター内でどのように 2 分割されるかということに注目するため、全体的な視点に立って個体を分類することが薄れてしまう。石岡 (2006) の手法のように、再併合することも一つの方法であるが、2 分割を繰り返さずに、直接 K 個のクラスターに分類できれば、計算量を減らすことができ、また、全体的な視点で個体の分類が可能となる。

従来の非階層的クラスター分析におけるクラスター数の決定は、完全に解析者の主観に基づいて行われてきた。そのため、まったく同じ分析を行っても解析者によって結果が異なってしまう状況が生じる。一方、混合正規分布モデルや x -means 法は、統計的指標である情報量規準によってクラスター数の決定を行うため、この点に関しては、従来のクラスター分析よりも秀でている。しかしながら、前者は標本サイズが増加すると、EM アルゴリズムで推定を行う際に、推定に時間がかかったり、局所最大値に収束してしまったりする可能性があり、後者は局所的には 2 分割でなければならないという理論的な制約がある。

したがって、主観的なクラスター数の決定を回避し、かつ適切な分類を行える手法を提案することは非常に有用である。ゆえに、本研究では、変数間の関係性を考慮しながら、適切な統計的指標によって、より客観的にクラスター数の決定を行う分類手法を提案することを目的とする。

方 法

変数間の関係性を考慮した個体の分類手法と、適切なクラスター数を決定するための情報量規準の算出方法を示す。ここでは、個体数 n の p 変量データを、 K 個のクラスターに分類する場合を想定する。以下にアルゴリズムを示す。

アルゴリズム

1. MacQueen の方法により、個体を分類する¹。
2. 1 で求められたクラスター毎に、平均ベクトル μ_k と分散共分散行列 Σ_k を計算し、個体 $x_i (i = 1, \dots, n)$ と各クラスターの平均ベクトル μ_k とのマハラノビス距離を算出し、同時に尤度も求める。クラスター k における個体 i の尤度 L_{ik} は、次式で計算される。

$$L_{ik} = (2\pi)^{-n/2} |\Sigma_k|^{-1/2} \exp \left[-\frac{1}{2} (x_i - \mu_k)' \Sigma_k^{-1} (x_i - \mu_k) \right] \quad (2)$$

3. 各個体 x_i を、尤度の最も高い値のクラスターに割り当てる。この操作をクラスターが変更されなくなるまで繰り返す。

4. 最終決定したクラスター毎に、個体数および個体の所属割合 w_k を計算する。

5. 個体 x_i を、混合正規分布に従う母集団からの標本と考え、以下の式によって各個体 x_i の確率密度 (尤度) を算出する。

$$f(x) = \sum_{k=1}^K w_k f_k(x) = \sum_{k=1}^K w_k \varphi(x; \mu_k, \Sigma_k) \quad (3)$$

6. 混合正規分布からの標本と仮定することで算出

¹ クラスター内の個体数が 1 になると分散共分散行列が計算できないため、必ず二つ以上になるようにする。

Table 1
シミュレーション結果

クラスター数		2	3	4	5	6	7	8～
3 クラスター	AIC	1	879	78	11	8	4	19
	BIC	131	869	0	0	0	0	0
5 クラスター	AIC	0	0	0	880	93	15	12
	BIC	2	30	62	903	3	0	0

された個体ごとの尤度を用いて、以下の式で AIC および BIC を計算する。ただし、 l は自由母数の数である。

$$\text{AIC} = -2\log L(\mu_k, \Sigma_k; \mathbf{x}) + 2l \quad (4)$$

$$\text{BIC} = -2\log L(\mu_k, \Sigma_k; \mathbf{x}) + l \log n \quad (5)$$

7. 任意のクラスター数で分析を行い、AIC および BIC の値を参照して、適切なクラスター数を決定する。

1 は、MacQueen の方法により求められるシード点を初期値とするために行う。2 では、1 で求めたクラスター毎に平均ベクトルと分散共分散行列を算出する。ここで添え字 k はクラスターを表す ($k=1, \dots, K$)。そして、個体 i の変数ベクトルを $\mathbf{x}_i$ とし、各個体と各クラスターの平均ベクトル μ_k とのマハラノビス距離と尤度 L_{ik} を算出する。3 では、尤度の最も高い値のクラスターに各個体を割り当て、変更されなくなるまで繰り返し、4 において、個体数と所属割合 w_k を求める。5 では、各個体が混合正規分布に従う母集団からの標本であると仮定して、所属割合 w_k を混合比率とし、平均ベクトルと分散共分散行列を用いて、各個体の確率密度 (尤度) を計算する。ただし、 $\sum_{k=1}^K w_k = 1$ である。6 において、5 で求めた個体ごとの尤度をまとめて、全体の対数尤度を算出する。それにより AIC および BIC を求め、7 でこれらの指標を参考に適切なクラスター数を決定する。また、自由母数の数は、平均ベクトルと分散共分散行列の要素の数であるから、クラスター毎に $p + p(p+1)/2 = p(p+3)/2$ と計算される。したがって、モデル全体ではこれを K 倍して、自由母数の総数は $l = p(p+3)K/2$ となる。

本手法と混合正規分布モデルは、データが混合正規分布に従うというアイデアに基づいて分類するという点において共通している。しかし、所属クラスターの算出アルゴリズムにおいて、両者は大きく異なる。混合正規分布モデルでは、個体の所属クラスターを表す架空の変数 z_{ik} を導入し、それを欠損値と扱って EM アルゴリズムによって推定を行う。一方、本手法では、そのような欠損値を導入せず、 k -means 法によってある程度分類されたクラスターに対して、変数間の関係性、つまりマハラノビス距離を考慮して、分類の修正を行う。言い換えると、はじめに大まかに分類

し、次いでマハラノビス距離から計算される尤度によって、より適切な分類を行うということである。

シミュレーション

シミュレーションデータを用いて、提案手法の性能を調べる。具体的には、正しいクラスター数が決定できるかを検討することを目的とする。

3 クラスター

3 クラスターを想定して、2 変量正規乱数を各クラスター 50 個、計 150 個作成する。クラスター平均 μ_k はそれぞれ、 $\mu_1 = [2, 0]$ 、 $\mu_2 = [0, 2]$ 、 $\mu_3 = [-2, 0]$ であり、分散 σ_k^2 はそれぞれ、 $\sigma_1^2 = [0.5, 0.5]$ 、 $\sigma_2^2 = [0.4, 0.4]$ 、 $\sigma_3^2 = [0.4, 0.4]$ である。

シミュレーションデータを発生させて、本手法による分類を 1000 回実施した。2 変量正規乱数は、その都度、つまり 1000 組発生させた。本手法を適用して得られた最適なクラスター数をまとめたものが Table 1 の上段である。どちらの指標においても、3 クラスターを適切なモデルと選択する回数が 800 回以上であることがわかる。また、BIC を用いたほうが、より真のクラスター数に近い決定を行えることが示唆された。

5 クラスター

5 クラスターを想定して、2 変量正規乱数を各クラスター 50 個、計 250 個作成する。クラスター平均 μ_k はそれぞれ、 $\mu_1 = [2, 0]$ 、 $\mu_2 = [0, 2]$ 、 $\mu_3 = [-2, 0]$ 、 $\mu_4 = [0, -2]$ 、 $\mu_5 = [0, 0]$ であり、分散 σ_k^2 はそれぞれ、 $\sigma_1^2 = [0.5, 0.5]$ 、 $\sigma_2^2 = [0.4, 0.4]$ 、 $\sigma_3^2 = [0.4, 0.4]$ 、 $\sigma_4^2 = [0.4, 0.4]$ 、 $\sigma_5^2 = [0.2, 0.2]$ である。

先程と同様に、本手法による分類を 1000 回実施した。また、2 変量正規乱数はその都度発生させた。本手法を適用して得られた最適なクラスター数をまとめたものが Table 1 の下段である。どちらの指標を用いても、5 クラスターを適切なモデルと選択する回数が多いことが確認できる。また、AIC は真のクラスター数より多いモデルを、BIC は少ないモデルを採用しやすいことがわかる。

Table 2
アイリスデータの情報量規準

クラスター数	AIC	BIC
2	40734.28	40818.58
3	40698.57	40825.01
4	40712.27	40880.87
5	40712.92	40923.66
6	40716.65	40969.54
7	40718.84	41013.88

Table 4
お札データの情報量規準

クラスター数	AIC	BIC
1	74028.67	74045.16
2	73836.80	73869.79
3	73808.52	73857.99
4	73784.68	73850.64
5	73789.42	73871.88
6	73800.98	73899.93
7	73798.40	73913.84

数値例

続いて、実データに提案手法を適用させる。実データは、所属クラスターが既知であるアイリスのデータ (Fisher, 1936) とお札のデータ (Flury & Riedwyl, 1988) である。ここでは、本手法を両データに適用し、適切なクラスター数の決定が可能か、また、正分類が可能かについて検討することを目的とする。

アイリスデータ

アイリスデータは、3種類のアヤメのがく片の長さ、と幅、花卉の長さ、と幅に関する計測データである。各アヤメは50個ずつ測定され、合計150個の個体が4変数によって観測された。本手法をアイリスデータに適用し、情報量規準を算出した結果をTable 2に示す。AICでは3クラスターを、BICでは2クラスターを選択することが示された。続いて、3クラスターで分類した場合に、どのように分類されたかをTable 3に示す。セトサは完全に分類できており、パーシカラーとバージニカにおいては、数個の誤分類が見られる結果となった。全体の正答率は、 $0.98 = (50 + 48 + 49) / 150$ であり、上手く分類できていることが確認された。

お札データ

お札データは、真札と偽札の-margin幅と対角線の長さに関するデータである。真札と偽札が100枚ずつ測定され、合計200枚のお札が2変数によって観測さ

Table 3
3クラスターの分類結果 (アイリスデータ)

分析による分類	セトサ	パーシカラー	バージニカ	合計
真の分類 セトサ	50	0	0	50
パーシカラー	0	48	2	50
バージニカ	0	1	49	50
合計	50	49	51	150

Table 5
4クラスターの分類結果 (お札データ)

分析による分類	偽札(○)	偽札(×)	偽札(△)	真札(●)	合計
真の分類 偽札	45	14	40	1	100
真札	0	1	0	99	100
合計	45	15	40	100	200

れた。本手法をお札データに適用し、情報量規準を算出した結果をTable 4に示す。

AIC, BICともに4クラスターを選択することが分かる。実際には2クラスターであるので、最適なクラスター数を4とした原因を探るために、具体的にどのように分類されるのかを検討する。4クラスターの場合の分類結果をTable 5にまとめる。4クラスターの場合の誤分類は2枚で、正答率は非常に高く、適切に分類されている。具体的には、偽札がより詳細な三つの下位クラスターへ分類される。ここで、4クラスターにおけるお札の散布図をFigure 1に示す。Figure

Figure 1. 4クラスターにおけるお札データの散布図

Table 6
BtoB および BtoC データにおける BIC の変遷

クラスター数	BtoB	BtoC
2	214477.6	944985.6
3	214463.2	944941.9
4	214515.6	944841.8
5	214618.5	944966.9
6	214607.3	945030.9
7	214790.1	945001.5
8	214821.6	945033.6

Table 7
BtoB および BtoC のクラスター平均

調査の種類		04 年度	05 年度	06 年度	07 年度	08 年度	09 年度
BtoB	クラスター 1	1.054	0.828	0.149	0.047	-0.793	-1.286
	クラスター 2	0.372	0.214	0.122	0.596	-0.548	-0.756
	クラスター 3	0.121	0.693	1.125	0.108	-0.886	-1.161
BtoC	クラスター 1	0.302	0.087	0.876	0.519	-0.781	-1.004
	クラスター 2	0.513	1.230	-0.221	0.289	-0.593	-1.218
	クラスター 3	1.327	0.547	0.256	-0.166	-0.903	-1.061
	クラスター 4	-0.364	0.123	0.016	0.184	0.021	0.021

1 より、偽札は、中央 (○)、左下 (×)、右 (△) の三つの下位クラスターに分類され、マージン幅の違いによって詳細に分割される。

適 用 例

最後に、本手法を利用した適用例を示す。ここでは、消費者心理学の分野においてよく研究されるブランド価値のデータを利用する。心理学との関連で、ブランド・イメージや価値を扱った論文に、Aaker, Veronica, & Jordi (2001) や豊田・川端・片平・齋藤 (2006) が挙げられる。利用するブランドデータは、日経 BP コンサルティング (2009) によるブランドジャパンの調査データである。この調査では、消費者に対して消費行動上のブランド価値を示すコンシューマ指標 (以下 BtoC とする) とビジネスパーソンに対してビジネス上のブランド価値を示すビジネスパーソン指標 (以下 BtoB とする) が計算され、2001 年以降毎年 4 月にブランド指標が公開されている²。

今回は、調査法が統一された 2004 年以降の 6 カ年間の Web 調査データを用いて分析を行う³。BtoC で

は、四つの下位指標とそれをまとめた総合力指標が、BtoB では五つの下位指標とそれをまとめた総合力指標が算出されるが、ここでは総合力指標のみを利用し、年度を変数として提案手法によるブランドの分類を行う。また、ブランドの総合力指標は、平均値の高低によって左右されるため、個体内で標準化を行ったデータ (標準化データ) を用いる。6 カ年の間毎年評価されたブランドは、BtoB で 342 ブランド、BtoC で 717 ブランドであった。

BtoB 分析結果

BtoB における 342 ブランドに関して、本手法を適用した。クラスター数を変化させたときの BIC の変遷を Table 6 の BtoB の列に示す。Table 6 より、3 クラスターが適切であると判断する。そこで、各クラスターの平均を Table 7 の上段に示す。また、各クラスターにおいて、クラスター平均の変遷と最もよく似た傾向を示す 10 ブランドを調べ⁴、Table 8 にまとめる。Figure 2, Figure 3, Figure 4 には各クラスター平均と Table 8 で示した 10 ブランドの標準化された総合力指標の時系列を示す。

クラスターに関係なく、ここ数年において、ブランド力は低下傾向にある。近年の日本社会における不景

² BtoC と BtoB の調査データの解析には、構造方程式モデリング (狩野・三浦, 2002) を利用した。なお、豊田 (2003, pp. 742-743, 図 1・図 2 および表 1・表 2) において、それぞれのモデルにおけるパス図および観測変数が示されている。

³ Web 調査が有意抽出であるという問題は、傾向スコアウェイト法を用いることで対処されている (豊田・川端・中村・片平, 2007)。

⁴ 各個体に関して、所属するクラスターの平均と分散共分散行列によって尤度を計算し、尤度の値が高い上位 10 ブランドを抽出する。

Table 8
クラスター内で平均と類似した傾向を示す上位 10 ブランド (BtoB)

ランキング	クラスター 1	クラスター 2	クラスター 3
1	松井証券	積水ハウス	ライオン(家庭用品)
2	ヤクルト	住友化学	オンワード樫山
3	新潮社	森永製菓	ダスキン
4	ソニー銀行	グンゼ	キヤノン
5	ソニー・ミュージック	INAX	富士フィルム
6	島津製作所	全日本空輸	コジマ(家電量販)
7	マツダ	TBS	ツムラ
8	豊田自動織機	オムロン	第一生命保険
9	佐川急便	横浜ゴム	日興コーディアル証券
10	日本旅行	象印マホービン	ミサワホーム

Figure 2. クラスター 1 (BtoB)

注) 実線がクラスター平均を表し、点線が同様の变化を示すブランドを表す。

Figure 3. クラスター 2 (BtoB)

注) 実線がクラスター平均を表し、点線が同様の变化を示すブランドを表す。

Figure 4. クラスター 3 (BtoB)

注) 実線がクラスター平均を表し、点線が同様の变化を示すブランドを表す。

Table 9
クラスター内で平均と類似した傾向を示す上位 10 ブランド (BtoC)

ランキング	クラスター 1	クラスター 2	クラスター 3	クラスター 4
1	ケンウッド	毎日放送	日本ビクター	タニタ
2	コクヨ	アルク	ジョージア	ダイワハウス
3	オリコ	ミキモト	バーバリー	雪印
4	ソニー損保	シマンテック	コジマ(家電量販)	ジュピターテレコム
5	コスモ石油	エールフランス	高島屋	グーグル
6	エーザイ	ダスキン	ゲンゼ	CoCo 壱番屋
7	松竹	OCN	ベネトン	ミツカン
8	ホンダ	フレッシュネスバーガー	ボシュロム	アクオス
9	アリコ	カラリオ	全労済	セブン銀行
10	フォーシーズンズホテル	新潮社	コンパース	パッファロー

Figure 5. クラスター 1 (BtoC)

注) 実線がクラスター平均を表し、点線が同様の变化を示すブランドを表す。

Figure 6. クラスター 2 (BtoC)

注) 実線がクラスター平均を表し、点線が同様の变化を示すブランドを表す。

Figure 7. クラスター 3 (BtoC)

注) 実線がクラスター平均を表し、点線が同様の变化を示すブランドを表す。

Figure 8. クラスター 4 (BtoC)

注) 実線がクラスター平均を表し、点線が同様の变化を示すブランドを表す。

気や世界的な恐慌がその原因と考えられるが、その傾向の中でも、クラスターによって低下傾向にいくつかの種類が存在する。

まず、クラスター 1 は一貫して、総合力が低下しているクラスターである。自動車関連の会社（“マツダ”や“豊田自動織機”）や大手メーカーの関連会社（“ソニー・ミュージック”や“ソニー銀行”）などがこのクラスターに属する。大手の企業および関連会社は、直接不況の影響を受け、それがブランド力にも波及したと考えられる。次に、クラスター 2 は 2007 年度から 2008 年度にかけての低下が著しい。これは 2007 年度に起きたオイルショックが原因と考えられる。燃料が必要な“全日本空輸”や石油化学製品を製造する“住友化学”、また自動車に関連する“横浜ゴム”などは、オイルショックが起因となって、急激にブランド力が低下していったと考えられる。最後に、クラスター 3 は、2006 年度をピークに低下するクラスターである。クラスター内を見ると、家電製品を扱う“キヤノン”や“富士フィルム”、“コジマ”などが含まれている。これらは、2006 年度までは好調だった売り上げが、ここ数年の消費の低迷によりブランド力が影響を受けたと考えられる。

BtoC 分析結果

BtoC における 717 ブランドに関して、本手法を適用した。クラスター数を変化させた時の BIC の変遷を Table 6 の BtoC の列に示す。Table 6 より、4 クラスターが適切と判断する。そこで、各クラスターの平均を Table 7 の下段に示す。また、各クラスターにおいて、クラスター平均の変遷と最もよく似た傾向を示す 10 ブランドを調べ、Table 9 にまとめる。BtoB と同様に、Figure 5 から Figure 8 に各クラスターの平均と Table 9 で示した 10 ブランドの標準化された総合力指標の時系列を示す。

まず、クラスター 1 は、2006 年度をピークに低下するクラスターで、保険会社の“ソニー損保”や“アリコ”、自動車に関連する“ホンダ”や“コスモ石油”などが含まれる。次に、クラスター 2 には、毎年のようにブランド力が変動する変化の著しいブランドが所属する。マスメディア関連の“毎日放送”や“新潮社”などは、些細な事件や不祥事などによりブランド力が左右されやすいブランドと考えられる。続いて、クラスター 3 は、一貫してブランド力が低下するクラスターである。服飾や靴などのファッション関連のブランド（“バーバリー”や“グンゼ”、“ベネトン”や“コンパース”）は、消費者にとってより身近な存在であるため、近年の不況と連動して、一貫して低下傾向にあることが示唆される。最後に、クラスター 4 には、2005 年度以降大きな変動のないブランドが属する。通信・情報関連ブランドの“グーグル”や“ジュ

ピターテレコム”、なじみのある食品関連ブランドの“雪印”や“ミツカン”は、近年の不況の影響を受けずに、ブランド力を維持していると考えられる。

考 察

本手法の目的は、変数間の関係性を考慮しながら、適切な統計的基準によって客観的にクラスター数の決定が行える新しい分類手法を提案することであった。そして、その有用性を確認するために、シミュレーション・数値例・適用例において、様々なデータに本手法を適用した。

まず、シミュレーション結果より、2 変量 3 クラスターと 2 変量 5 クラスターのどちらの場合においても、本手法によって、真のクラスター数をより適切に決定できることが確認された。次に、数値例においては、3 クラスターであるアイリスデータを、上手く三つに分類し、かつ、高い正答率で分類することが分かった。また、お札データでは、4 クラスターが適切であると判断されたが、これは偽札をより詳細に分類しているためであった。最後に、適用例では、ブランド価値データを用いてブランドの分類を行った。BtoB では 3 クラスターが、BtoC では 4 クラスターが最適と判断され、それに関する適切な解釈も行うことができた。

以上より、はじめに大まかな分類をし、次いでマハラノビス距離を用いて分類の修正を行う本手法は、有用な方法であることが示唆された。*k*-means 法で頻繁に用いられてきたユークリッド距離ではなくマハラノビス距離を用いることで、変数間の関係性を考慮し、また、*x*-means 法とは異なり全体的な視点に立って分類の修正を行うことで、より適切な分類を行うことができると考えられる。また、主観的なクラスター数の決定を回避し、データに基づく客観的なクラスター数の決定を行うことができた。ただし本手法では、クラスター内の個体数が一つの場合には、マハラノビス距離が計算できないという問題が生じる。今後の課題として、極端に性質の異なる個体が存在し、クラスター内の個体数が一つの場合の対処法などについても検討する必要がある。

引用文献

- Aaker, J. L., Veronica B., & Jordi, G. (2001). Consumption symbols as carriers of culture: A study of Japanese and Spanish brand personality constructs. *Journal of Personality and Social Psychology*, **81**, 492-508.
- 荒木 友希子 (2008). 日本人大学生を対象とした学業達成場面における防衛的悲観主義の検討 心理学研究, **79**, 9-17.
- (Araki, Y. (2008). Japanese college student's pessi-

- mism, coping strategies and anxiety: Validation of the Japanese Defensive Pessimism Inventory (JDPI). *Japanese Journal of Psychology*, **79**, 9-17.)
- Bezdek, J. C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**, 1-8.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1-38.
- Dunn, J. C. (1974). A graph theoretic analysis of pattern classification via Tamura's fuzzy relation. *IEEE Transactions on Systems, Man, and Cybernetics*, **4**, 310-313.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.
- Flury, B., & Riedwyl, H. (1988). *Multivariate statistics: A practical approach*. London: Chapman & Hall.
- Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, **21**, 768-769.
- Hartigan, J. A., & Wong, M. A. (1979). A k -means clustering algorithm. *Applied Statistics*, **28**, 100-108.
- 石岡 恒憲 (2000). クラスター数を自動決定する k -means アルゴリズムの拡張について 応用統計学, **29**, 141-149.
(Ishioka, T. (2000). Extended k -means with an efficient estimation of the number of clusters. *Japanese Journal of Applied Statistics*, **29**, 141-149.)
- 石岡 恒憲 (2006). x -means 法改良の一提案—— k -means 法の逐次繰り返しとクラスターの再併合—— 計算機統計学, **18**, 3-13.
(Ishioka, T. (2006). An expansion of x -means: Progressive iteration of k -means and merging of the clusters. *Bulletin of the Computational Statistics of Japan*, **18**, 3-13.)
- 狩野 裕・三浦 麻子 (2002). グラフィカル多変量解析——AMOS, EQS, CALIS による目で見る共分散構造分析—— 現代数学社
(Kano, Y., & Miura, A.)
- Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28**, 29-137.
- MacQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.
- 日経 BP コンサルティング (2009). ブランドジャパン 2009 総合報告書“解説書” 日経 BP コンサルティング
(Nikkei BP Consulting)
- 大内 晶子・長尾 仁美・櫻井 茂男 (2008). 幼児の自己制御機能尺度の検討——社会的スキル・問題行動との関係を中心に—— 教育心理学研究, **56**, 414-425.
(Oh-uchi, A., Nagao, H., & Sakurai, S. (2008). Development of an Early Childhood Self-Regulation Scale: Social skills and problem behavior. *Japanese Journal of Educational Psychology*, **56**, 414-425.)
- Pelleg, D., & Moore, A. (2000). X -means: Extending K -means with efficient estimation of the number of clusters. *Proceedings of the Seventeenth International Conference on Machine Learning*, 727-734.
- 豊田 秀樹 (2003). ブランド指標の信頼性と信頼区間オペレーションズリサーチ, **48**, 741-746.
(Toyoda, H.)
- 豊田 秀樹・川端 一光・片平 秀貴・齋藤 朗宏 (2006). 残差変数の構成概念スコアを推定したランキングモデル 心理学研究, **77**, 115-123.
(Toyoda, H., Kawahashi, I., Katahira, H., & Saito, A. (2006). The ranking model which estimates the factor scores of residual variables. *Japanese Journal of Psychology*, **77**, 115-123.)
- 豊田 秀樹・川端 一光・中村 健太郎・片平 秀貴 (2007). 傾向スコア重み付け法による調査データの調整——ニューラルネットワークによる傾向スコアの推定—— 行動計量学, **34**, 101-110.
(Toyoda, H., Kawahashi, I., Nakamura, K., & Katahira, H. (2007). An adjustment of survey data using propensity score weighting: An estimation of the propensity score by a neural network. *Japanese Journal of Behaviormetrics*, **34**, 101-110.)

—— 2010. 1. 14 受稿, 2010. 7. 3 受理 ——