2 値変量に基づく教師無し分類における類似係数の選択

石 田 実*, 西 尾 チヅル**, 椿 広 計***

Choosing a Similarity Coefficient for Classification by Binary Variables

Minoru ISHIDA*, Chizuru NISHIO** and Hiroe TSUBAKI***

Pairwise similarity coefficients are popular measure for binary variables. Many different measures of similarity have been proposed in the literature. Then we are interested in which one is the most effective for classifications. We focus on the fact that almost all measures of similarity are composed of interactions and main effects, and conjecture that the most useful similarity is an interaction because main effect don't play a role of classifications but totally order. All combinations of sixteen similarities coefficients and five clustering method were tested with music CD POS data. The cluster validation were assessed by interpretable, uniform, reproducible, external and internal criteria. As a result, the similarity coefficient which is more correlative with an interaction turns out more useful for classifications. That is, the best similarity is an interaction.

Key words: similarity coefficients, cluster validation, joint segmentation

キーワード: 類似係数, クラスターバリデーション, ジョイントセグメンテーション

1. 序 論

連続変量の記述多変量解析には変量の変動を特異値分析で近似要約する主成分分析および交互作用を近似する Mandel (1971) の交互作用解析と共に, 類似係数に基づくクラスター分析の2つの流れがある. 一方, 2 値変量の記述多変量解析においてはコレスポンデンス分析に代表される特異値分析等の近似要約はよく用いられるものの, クラスター分析については, どのような類似係数がどのような分析結果を導くかについて

殆ど研究されていない (Janson & Vegelius, 1981; Baulieu, 1989; Duarte, Santos, & Melo, 1999). そこで本研究では, 2 値変量の解析で提唱されている様々な類似係数について, クラスター分析の観点から好ましい結果を導く類似係数の特徴を実験的に示す. また, 類似係数の交互作用統計量による偏導関数を求め, クラスター分析に適していない類似係数と交互作用統計量との関係の特徴を示す.

本研究の結果として示す通り, いくつかの類似係数は空間収縮を引き起こすなど必ずしも好ましくないクラスター分析を導く. 特にコレスポンデンス分析における距離の役割を果たすカイ二乗距離は, 検定統計量に含まれる行列の交互作用の影響に比べて主効果情報の影響を強く受け, 主効果が大きくなると距離が小さくなる特徴を持つため, 結果として著しい空間収縮を見せる. この事は, カイ二乗距離行列の低次元近似を原理とするコレスポンデンス分析においても, 応用上の問題を引き起こす要因になると考えられ, 大規模疎行列に対するコレスポンデンス分析は上位2~3の固有値を用いても, その累積寄与率は極めて低くなるのが通例である. 従って, そのような行列に対する記述的な分析の第一段階として, 類似係数に基づくクラス

*株式会社アークエンジン
(ArkEngine Co., Ltd)

連絡先: 〒103-0022 東京都中央区日本橋室町1-10-10 LXS 室町1005

E-mail: ishida@arkengine.co.jp

**筑波大学ビジネス科学研究科

(Graduate School of Business Sciences, University of Tsukuba)

連絡先: 〒112-0012 東京都文京区大塚3丁目29-1

E-mail: nishio@gssm.otsuka.tsukuba.ac.jp

***統計数理研究所

(The Institute of Statistical Mathematics)

連絡先: 〒190-8562 東京都立川市緑町10-3

E-mail: tsubaki@ism.ac.jp

ター分析を行う事は有用と考える。例えば、膨大な2値データを記述的に解析するゲノム解析 (Duarte et al., 1999) やマーケティングのPOS データ解析 (芳賀, 2008) の分野において、類似係数の選択が課題とされている。

本研究では、類似係数が交互作用と主効果の関数で構成されている事に着目し、この2軸に沿って評価の優劣を検証する。この2要因を個別に分類に用いるとすると、まず観測対象に固有の値である主効果は観測対象に全順序を与え、分類というより主効果のランキングを与える。一方、交互作用は2項関係なので分類に適切な指標と期待できる。そこで主効果の影響が少なく、交互作用との相関が高い類似係数ほど分類に適切な指標になると予想する。

この予想を検証するために、主効果と交互作用が無相関となる事を踏まえて、類似係数を交互作用から主効果へと相関の高い順に並べる。そして、この順序と分類の妥当性の評価が一致する事を検証する。この検証プロセスの課題は、分類の妥当性の評価を客観的に示す事である。教師あり分類の場合であれば、真の分類との一致を評価すれば良い。しかし本論では教師無し分類を評価するため、クラスターバリデーション (Milligan, 1996; Gordon, 1999, 7章) の手順に従って、多面的な評価基準を用いる。具体的には再現性・均等性・外的基準・内的基準の4つの定量的基準と、定性的な解釈可能性の基準で評価する。

ところで、この検証には考察すべき課題がある。第1の課題は、交互作用の指標の選択である。本研究では、対応表の独立性の検定に用いるピアソンのカイ二乗統計量のルートに符号を付け総度数の平方値で割って $[-1, 1]$ の値域を持つように基準化した値を交互作用統計量とする (表2のM1)。二乗する前の符号を戻すのは、交互作用がプラスかマイナスかで分類結果が異なる事に配慮している。類似係数のPsi (表2のM7) は交互作用統計量に絶対値を付けた指標であるが、プラスの交互作用が存在するためにはマイナスの交互作用も必ず存在しており、符号が持つ情報を活かすかどうかで分類結果も異なる。また、ピアソンのカイ二乗統計量以外にも Goodman (1996) の連関モデルのように乗法的に交互作用を捉える考えもあろう。より良い交互作用の指標の選択については、当初の予想が肯定的に検証できた後の課題とし、本研究では主効果による偶然一致の影響を除いた指標で、主効果の影響を受ける標準偏差で基準化した指標が、他の類似

係数より妥当な分類を与える事のみを検証する。

第2の課題は、教師なし分類の検証に強力な基準が無い事である (Milligan, 1996; Gordon, 1999, 7章)。どのような分類結果に対しても、分析者は「これが私の意図した分類です。」と主張できる。このため、教師なし分類の検証では、複数の評価基準を併用する多面的評価が望ましい。また、個々の基準の評価についても、データに依存して偶然良い結果を得る可能性を無視できない。そこで本研究では、2-フォールド交差確認法を用いて分類を作成したり、検証指標の有意水準を評価する確率分布が分かっている場合は乱数を用いたシミュレーションで p 値に代わる経験的パーセンタイル点を求めたりして、検証の信頼性を高める事とする。

本論では、次の2章で検証の手順と分類の妥当性を評価するクラスターバリデーションについて述べる。続く3章で評価基準毎に実証分析を行い、結果を整理して考察する。実証データには音楽CDの個人ID付き販売POSデータを使用する。そして最後に4章で結果を考察し、今後の課題を述べる。

2. 検証の手法

2.1 類似係数

0と1を要素とする m 行 n 列の2値行列 $M = \{m_{ij}\}$ を、行方向が m 個の観測対象で、列方向が n 個の2値変量の属性に対応するデータセットとする。このとき観測対象 i_1 と i_2 の類似係数を、 i_1 と i_2 の n 個の属性値の 2×2 クロス集計表 (表1) と、各類似係数の算式 (表2) より計算する。

表1. 観測対象 i_1 と i_2 のクロス集計表
観測対象 i_2

観測対象 i_1		1	0	計
	1	a	b	r_1
	0	c	d	r_2
	計	c_1	c_2	n

ただし、集計値 a, b, c, d は、

$$a = \sum_j m_{i_1 j} \cdot m_{i_2 j},$$

$$b = \sum_j m_{i_1 j} \cdot (1 - m_{i_2 j}),$$

$$c = \sum_j (1 - m_{i_1 j}) \cdot m_{i_2 j},$$

$$d = \sum_j (1 - m_{i_1 j}) \cdot (1 - m_{i_2 j})$$

また行和と列和と総和を $r_1 = a + b$, $r_2 = c + d$, $c_1 = a + c$, $c_2 = b + d$, $n = a + b + c + d$.

表 2. (非) 類似係数の一覧

名称	類似係数 s_{ij}	非類似係数 d_{ij}
M1 交互作用統計量	$\frac{ad-bc}{\sqrt{r_1 r_2 c_1 c_2}}$ (*1)	$1 - \frac{ad-bc}{\sqrt{r_1 r_2 c_1 c_2}}$
M2 Cohen's Kappa	$\frac{2(ad-bc)}{2(ad-bc)+n(b+c)}$	$\frac{n(b+c)}{2(ad-bc)+n(b+c)}$
M3 Ochiai	$a/\sqrt{(a+b)(a+c)}$	$1 - a/\sqrt{(a+b)(a+c)}$
M4 Jaccard	$a/(a+b+c)$	$(b+c)/(a+b+c)$
M5 Sokal & Sneath 2	$a/(a+2b+2c)$	$(2b+2c)/(a+2b+2c)$
M6 Czesanowski	$2a/(2a+b+c)$	$(b+c)/(2a+b+c)$
M7 Psi	$\frac{ ad-bc }{\sqrt{r_1 r_2 c_1 c_2}}$	$1 - \frac{ ad-bc }{\sqrt{r_1 r_2 c_1 c_2}}$
M8 Kulczynski	$a/(b+c)$ (*2)	$\max_{i,j}\{s_{ij}\} - s_{ij}$
M9 Yule's Q	$(ad-bc)/(ad+bc)$	$2bc/(ad+bc)$
M10 Russel & Rao	a/n	$(b+c+d)/n$
M11 Rogers & Tanimoto	$(a+d)/(a+2b+2c+d)$	$2(b+c)/(a+2b+2c+d)$
M12 χ^2 metric	undefined	$\sum_{k=1}^n \frac{n}{m_{k\cdot}} \left(\frac{m_{ki}}{m_{\cdot i}} - \frac{m_{kj}}{m_{\cdot j}} \right)^2$ (*3)
M13 Simple matching	$(a+d)/n$	$(b+c)/n$
M14 Sokal & Sneath 1	$2(a+d)/(2a+2d+b+c)$	$(b+c)/(2a+2d+b+c)$
M15 Hamann	$(a+d-b-c)/n$	$(2b+2c)/n$
M16 主効果の相乗平均	$\sqrt{(a+b)(a+c)}/n$	$1 - \sqrt{(a+b)(a+c)}/n$

*1. a, b, c または d のいずれかの値が 5 以下の場合は、イエーツの修正を用いる。

*2. $b+c=0$ の場合は、 $a=(b+c)$ を a に置換える。

*3. $m_{i\cdot}$ は行列 $M=\{m_{ij}\}$ の i 行の行和、 $m_{\cdot j}$ は j 列の列和とする。

表 2 は Wadel & Kamakula (2000, p. 46, 表 5.1) が最も重要とする類似係数のリストに、M1 交互作用統計量と M2 Cohen's Kappa (Cohen, 1960) と M12 カイ二乗距離と M16 主効果の相乗平均の 4 指標を追加して作成している。Albatineh, Niewiadomska-Bugaj, & Mihalko (2006) は、このほとんどの類似係数の出典をサーベイしている。しかしながら、これまで交互作用統計量は類似係数として扱われていない。また、主効果の相乗平均は類似係数では無いけれど、比較の基準とするために検証する類似係数のリストに加える。この中で良く知られた類似係数は M4 Jaccard と M13 Simple matching であろう。これらの類似係数行列は半正定値となるので、非類似係数には自然な距離が対応する (Gower & Legendre, 1986; Gordon, 1999) こ

とが知られている。また M2 Cohens's Kappa は見かけの一致率から偶然一致率を差し引いているので、交互作用に近い特徴を持っている。ただし標準偏差で基準化していない。M12 カイ二乗距離は表 1 の 2×2 クロス集計表から計算する指標では無いが、同じ 2 値行列 M を用いて計算する指標であり、分布の同等性という優れた特性 (大隈他, 1994) を持っている事と、またコレスポンデンス分析に用いる距離として知られている事からリストに加えている¹⁾。

¹⁾ オッズ比については、M9 Yule's Q = $\frac{\text{オッズ比}-1}{\text{オッズ比}+1}$ の関係により、Yule's Q と単調に一对一対応する。先行研究ではオッズ比より Yule's Q を採り上げる事が多い慣例にならない (Wadel & Kamakula,

ところで、表 2 の類似係数の記載順は、後述する実証データにおいて交互作用統計量と相関が高い順としている。また表 2 の非類似係数は、分類のための距離行列に相当する非類似行列を求めるための算式で、非負となるように (定数 - 類似係数) の式で定義している。

多くの類似係数の比較あるいは評価をしている先行研究として、Batagelj & Bren (1995) と Warrens (2008) は類似係数が分数の形式を持つために、分母が 0 に近い指標を定義できない問題を指摘し、修正式を提案している。Janson & Vegelius (1981) と Baulieu (1989) は類似係数が持つべき好ましい複数の性質、例えば対称性などを示し、各類似係数がそれらの性質を満たすか否かを確認している。Fleiss (1975) は類似係数の計算において 2×2 クロス集計表の偶然一致を除く事の重要性を指摘し、偶然一致の確率分布について論じている。Duarte, Santos, & Melo (1999) は 8 個の類似係数で 27 個のサンプルをクラスター分析し、類似係数の順位相関とデンドログラムの類似について論じている。このように多数の類似係数を比較する研究はいくつかあるものの、類似係数毎の分類結果の妥当性を評価する研究は見当たらない。

次に、本論の予想に関わる議論として、類似係数の構成における交互作用と主効果の関係について考察する。表 1 の 2×2 クロス集計表は a, b, c, d の 4 つの観測度数情報を持っており、これは以下の 4 つのクロス集計表の統計量から一対一変換で表現できる同じ情報量を持つ。

- 1 総度数 $n = a + b + c + d$
- 2 行の主効果 $x_1 \stackrel{\text{def}}{=} (a + b) / n$
- 3 列の主効果 $x_2 \stackrel{\text{def}}{=} (a + c) / n$
- 4 交互作用統計量 $y \stackrel{\text{def}}{=} \frac{ad - bc}{\sqrt{r_1 r_2 c_1 c_2}}$

すなわち、度数 a, b, c, d と統計量 n, x_1, x_2, y の間に次式の関係式が成り立つ。

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = n \begin{pmatrix} x_1 x_2 \\ x_1 (1 - x_2) \\ (1 - x_1) x_2 \\ (1 - x_1) (1 - x_2) \end{pmatrix} + n y \sqrt{x_1 (1 - x_1) x_2 (1 - x_2)} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$

上式を表 2 の類似係数の算式に代入すれば、表 3 の

2000; Warrens, 2008), オッズ比を表 2 のリストから割愛している。

通り類似係数を主効果と交互作用統計量の関数として記述する事ができる。この変換により、M12 カイ二乗距離の行和 m_k を除いた全ての類似係数の分母分子で総度数 n がキャンセルされ、類似係数を 4 変数から 3 変数の関数に置き換えることができる。そして、大規模疎行列に類似係数を適用した場合の類似係数と主効果および交互作用統計量との関係を偏導関数を用いて評価できる利点がある。すなわち大規模疎行列では行と列の主効果 x_1, x_2 が小さくなる傾向を踏まえ、まず表 3 の関数式に $x_1 = x_2 \stackrel{\text{def}}{=} x$ を仮定して関数式を整理し (表 4 の第 2 列), これを交互作用統計量 y で偏微分して $x = 0$ を代入することで主効果が小さい場合の類似係数と交互作用統計量との関係が評価できる。

結果は表 4 の通り、行と列の主効果が等しい ($x_1 = x_2$) という条件の下で、類似係数を 3 つのタイプに分類できた。最初のタイプ I と分類した類似係数は、M1 交互作用統計量と M2 Cohen's Kappa と M7 Psi で、いずれも主効果の影響を受けない。次のタイプ II と分類した類似係数は、M3 Ochiai, M4 Jaccard, M5 Sokal & Sneath 2, M6 Czesanowski, M8 Kulczynski の 5 個で、興味深いことに交互作用統計量と主効果による 2 つの影響の程度が等しく対称な (1 - 主効果) (1 - 交互作用統計量) の関数として記述できる。その他のタイプ III は M9 Yule's Q 以下の 8 個の類似係数で、このうち 7 個は主効果 $x = 0$ の偏微分係数がゼロとなり、主効果がゼロに近づくとき交互作用統計量に対して無相関となる。またカイ二乗距離については、分母に主効果 x が単独で残るため、主効果 x が小さくなると交互作用統計量による影響より主効果の影響を強く受けて距離が長くなる。このため、タイプ III の類似係数は大規模疎行列をクラスタリングするために適さないと推測できる。

2.2 クラスタバリデーション

教師あり分類の妥当性は、真の分類との一致度で評価できる。しかし、本論では教師無し分類を扱うため、Milligan (1996) や Gordon (1999) の示すクラスタバリデーションの手順に従って表 5 の 5 つの基準を用いる。再現性・均等性・外的基準・内的基準の 4 評価は定量的な基準で、最後の解釈可能性は定性的な基準である。特に定量的な基準を用いる際には、各評価値の絶対水準の評価に加えて、類似係数間の相対的な評価

表 3. 主効果と交互作用統計量と類似係数の関係式 - その 1

名称	類似係数の算式 (*1)
M1 交互作用統計量	y
M2 Cohen's Kappa	$2y \cdot \frac{\sqrt{x_1(1-x_1)x_2(1-x_2)}}{x_1+x_2-2x_1x_2}$
M3 Ochiai	$\sqrt{x_1x_2+y\sqrt{(1-x_1)(1-x_2)}}$
M4 Jaccard	$\frac{x_1+x_2}{x_1+x_2-x_1x_2-y\sqrt{x_1(1-x_1)x_2(1-x_2)}}-1$
M5 Sokal & Sneath 2	$\frac{2}{3} \cdot \frac{x_1+x_2}{2x_1+2x_2-3x_1x_2-3y\sqrt{x_1(1-x_1)x_2(1-x_2)}}-\frac{1}{3}$
M6 Czesanowski	$2 \cdot \frac{x_1x_2+y\sqrt{x_1(1-x_1)x_2(1-x_2)}}{x_1+x_2}$
M7 Psi	$y(0 \leq y \text{ 場合}), -y(y < 0 \text{ の場合})$
M8 Kulczynski	$\frac{1}{2} \cdot \frac{x_1+x_2}{x_1+x_2-2x_1x_2-2y\sqrt{x_1(1-x_1)x_2(1-x_2)}}-\frac{1}{2}$
M9 Yule's Q	$\frac{y}{2(1+y^2)\sqrt{x_1(1-x_1)x_2(1-x_2)}+(1-2x_1)(1-2x_2)y}$
M10 Russel & Rao	$x_1+x_2+y\sqrt{x_1(1-x_1)x_2(1-x_2)}$
M11 Rogers & Tanimoto	$\frac{2}{1+x_1+x_2-2x_1x_2-2y\sqrt{x_1(1-x_1)x_2(1-x_2)}}-1$
M12 χ^2 metric (*2)	$\frac{1}{m_k} \cdot \left(\frac{1}{x_1} + \frac{1}{x_2} - 2 - 2y\sqrt{\frac{(1-x_1)(1-x_2)}{x_1x_2}} \right)$
M13 Simple matching	$1-x_1-x_2+2x_1x_2+2y\sqrt{x_1(1-x_1)x_2(1-x_2)}$
M14 Sokal & Sneath 1	$2 - \frac{2}{2-x_1-x_2+2x_1x_2+2y\sqrt{x_1(1-x_1)x_2(1-x_2)}}$
M15 Hamann	$(1-2x_1)(1-2x_2)+4y\sqrt{x_1(1-x_1)x_2(1-x_2)}$
M16 主効果の相乗平均	$\sqrt{x_1x_2}$

*1. x_1, x_2 はクロス集計表の行と列の主効果で, y はクロス集計表の交互作用統計量.

*2. χ^2 metric については行和 m_k が k に依存せず全て等しいと仮定した距離 (非類似係数) を記述.

順を確認する.

2.2.1 再現性

階層的クラスター法では最初に作成されるグループが次に作成するグループの選択に影響を与える. このため, クラスタリングの初期にグループを形成する少数の観測対象の振る舞いに分類結果が影響を受ける懸念がある. そこで, 階層的クラスター法について分類の再現性を評価する. 評価手法は Milligan (1996) と Gordon (1999, p 184) の手順に従って 2-フォールド交差確認法で評価する.

2-フォールド交差確認法の手順は以下の通り.

1. 観測対象を乱数を使って 2 等分し, それぞれ集合 A, B とする.
2. 集合 A を n 個のクラスター $C_1, \dots, C_n$ に分類する. その際に C_i と集合 B の各点との距離も計算する.
3. 集合 B の各点を最も距離に近い C_i に分類する.
4. 集合 A と B を入れ替えて 2 と 3 の手順を繰り返し, もう一つの分類 $C'_1, \dots, C'_n$ を得る.

表 4. 主効果と交互作用統計量と類似係数の関係式－その 2

名称	$x_1=x_2(=x)$ の場合の類似係数の算式 $f(x, y)$ (*1)	$\frac{\partial f(x, y)}{\partial y} \Big _{x=0}$	タイプ
M1 交互作用統計量	y	1	I
M2 Cohen's Kappa	y	1	I
M3 Ochiai	$1-(1-x)(1-y)$	1	II
M4 Jaccard	$\frac{2}{1+(1-x)(1-y)} - 1$	$\frac{2}{(2-y)^2}$	II
M5 Sokal & Sneath 2	$\frac{4}{3} \frac{1}{1+3(1-x)(1-y)} - \frac{1}{3}$	$\frac{4}{(4-3y)^2}$	II
M6 Czesanowski	$1-(1-x)(1-y)$	1	II
M7 Psi	$y(0 \leq y \text{ 場合})$ $-y(y < 0 \text{ の場合})$	$1(0 < y \text{ 場合})$ $-1(y < 0 \text{ の場合})$	I
M8 Kulczynski	$\frac{1}{2} \frac{1}{(1-x)(1-y)} - \frac{1}{2}$	$\frac{1}{2(1-y)^2}$	II
M9 Yule's Q	$\frac{y}{2x(1-x)(1+y^2)+(1-2x)^2y}$	0	III
M10 Russel & Rao	$x^2+x(1-x)y$	0	III
M11 Rogers & Tanimoto	$\frac{2}{1+2x-2x^2-2x(1-x)y} - 1$	0	III
M12 χ^2 metric (*2)	$\frac{2}{m_k} \frac{(1-x)(1-y)}{x}$	$-\infty$ に発散	III
M13 Simple matching	$1-2x+2x^2+2x(1-x)y$	0	III
M14 Sokal & Sneath 1	$2 - \frac{1}{1-x+x^2+x(1-x)y}$	0	III
M15 Hamann	$(1-2x)^2+4x(1-x)y$	0	III
M16 主効果の相乗平均	x	0	III

*1. x_1, x_2 はクロス集計表の行と列の主効果で, y はクロス集計表の交互作用統計量.

*2. χ^2 metric については行和 m_k が k に依存せず全て等しいと仮定した距離 (非類似係数) を記述.

表 5. クラスタバリデーションの 5 つの基準

1. 再現性	分類を再現できる.
2. 均等性	各分類の大きさが均等.
3. 外的基準	一般に社会で受け入れられている分類に整合する.
4. 内的基準	消費者分類×製品分類の購買度数表に強い交互作用がある.
5. 解釈可能性	消費者と製品の分類, およびその対応を自然に解釈ができる.

5. 2つの分類 $\{C_i\}$ と $\{C_j\}$ の一致度をランドの修正指標で評価する.

6. 以上の手続きを N 回繰り返す.

ランドの修正指標はラベルに自明な一対一対応が無い分類の一致の評価に用いる指標で, ランドの指標 (Rand 1971) に Hubert & Arabie (1985) が修正を加えている. この指標について Milligan & Cooper (1986) は同じ目的で使う他の指標に比べて優れている事を示している. しかし, この指標値の確率密度関数の形状は, 2つの分類の対応表の周辺度数に影響を受けるため複雑で, 有意水準を評価するのが難しい. そこで

50 回の交差確認法で得た 100 通りの分類からクラスターの大きさの分布の平均を求め、これを周辺度数として 2 つの独立な分類を乱数で 1 万通り作成し、この対応表のランドの修正指標の分布を求める。そして、実際の分析で得た 50 回の指標値の平均から 1 標準偏差を引いた値の水準値が先のシミュレーションで得た分布のどこに位置するかの経験的パーセンタイル点を求めて、分類の独立性の検定の p 値の代用とする。1 標準偏差を引いて保守的に評価した理由は、2-フォールド交差確認法ではクラスター連結による新しい距離の更新のステップが半数となるため、空間の歪みの効果も半分となる事を考慮している。

2.2.2 均等性

実務的な観点から、分類の大きさの均等さを評価する。100 個の観測対象を 98 個と 1 個と 1 個に 3 分類する事を分析者は望まないであろう。その結果が正しいか否かに関係無く、単に実務的な観点から均等な大きさの分類が望ましいと考える。均等さの指標は、観測対象を n 分類したグループを要素数の降順に並べ、上位半分のグループに含まれる要素数が全体に占める比率 (%) で評価する。ただし、先の 2-フォールド交差確認法で作成した 100 通りの分類について比率を計算し、その平均値を最終的な均等性の評価値とする。

ところで、階層的クラスタリング法を用いた場合の均等性は、デンドログラムの形状と密接な関係がある。そして、類似係数の主効果のウェイトがデンドログラムの形状に影響を与えると予想する。そこで、均等性の評価ではデンドログラムの形状も併せて判断する。図 1 のデンドログラムは典型的な 3 類型で、それぞれ A 空間保存的、B 空間膨張的、C 空間収縮的と呼ばれる。これらの類型の要因には、クラスタリング法の選択 (Lance & Williams, 1967; DuBien & Warde, 1979) と、データの性質 (Yadohisa, Takeuchi, & Inada 1999) の 2 つがあると知られている。

類似係数の主効果のウェイトが大きくなると、類似

係数が観測対象に固有の値となり、分類の階層構造も直線的なランキングに近くなる。このためデンドログラムは空間収縮的な形状を示し、分類の大きさは均等にならないと予想する。デンドログラムの類型の判断は、2-フォールド交差確認法では無く通常のクラスタリングの結果に対して行う。また類型の判断は筆者が定性的に行う。

2.2.3 外的基準

観測対象の分類として広く受け入れられている分類を外的基準として、この分類に似ている程度を評価する。外的な分類は唯一の手本では無いので分析結果と一致する必要は無いし、分類数も異なるが、分析結果と無関係である事は望ましくない。そこで外的基準の分類と分析結果の分類との対応表を作成し、この対応表の独立性の検定指標であるカイ二乗統計量を外的基準の評価とする。

2.2.4 内的基準

内的基準は、意図した分類が作成できたかどうかを外的基準に頼らずに評価する基準である。類似係数を用いた分類を行う意図は、属性値のパターンが似ている観測対象をグループ化する事である。よって、観測対象のグループ数が属性の数に対して十分小さい時、観測対象のグループに対応した属性のグループも存在すると考えられる。そこで内的基準を、図 2 の通り、観測対象と属性の各グループの対応によるジョイントセグメンテーションの集計表の独立性の検定指標とする。属性の分類は、2 値行列 M の行と列の役割を逆にして、観測対象を分類する場合と同様に作成する。そして観測対象と属性の分類を対応させて、ジョイントセグメンテーションの集計表を作成する。

2.2.5 解釈可能性

観測対象や属性に意味のあるラベルを与え、それらのラベルが同じ分類または違った分類に配置される事を、分析結果とは別の知見に照らして定性的に妥当な解釈が可能かを評価する。本実証分析の応用先である

図 1. デンドログラムの類型

図2. 分類の手順

消費者行動分析や企業のマーケティング戦略評価においては、分析結果の利用者が市場構造についての仮説を持っている。したがって、記述的なクラスタリング手法により導出された分類軸やセグメントは、その仮説に沿って解釈できることが極めて重要な要件となる。そこで本論では観測対象について多くの知見を持つデータ提供者に、解釈可能性の評価を求める。

2.3 Clustering

類似係数と相性の良いクラスタリング手法の組合せがあると、クラスタリング手法の選択によって検証結果が影響を受ける懸念がある。そこで非階層的クラスタリング法を含む5種類のクラスタリング法を検証に用いる。

PAM (Partitioning Around Medoids) 法は Kaufman & Rousseeuw (1990) によって提案された非階層的クラスタリング手法で、最適な分類を求めるアルゴリズムが k-means に似ているが、入力データが三角不等式を満たさない非類似係数でも使えるという利点がある。

ウォード法、群平均法、最長距離法、最短距離法の4つの手法は、最も近い2つの分類を逐次併合して新しい分類を作成する階層的なクラスタリング法である。表6の通り、新しく併合した分類と他の分類との距離を決めるパラメーター α_i , β , γ に違いがある。Lance & Williams (1967) と DuBien & Warde (1979) は、表6のパラメーターの差異がクラスタリングの階層構造における空間の歪みの原因となり、図1の通りデンドログラムの3類型が生じる事を示している。

表6. クラスタリングの手法

	クラスタリングの手法	α_i	β	γ
非階層的	1 PAM 法			
階層的	2 ウォード法	$\frac{n_i + n_k}{n_+}$	$-\frac{n_k}{n_+}$	0
	3 群平均法	$\frac{n_i}{n_i + n_j}$	0	0
	4 最長距離法	$\frac{1}{2}$	0	$\frac{1}{2}$
	5 最短距離法	$\frac{1}{2}$	0	$-\frac{1}{2}$

表中の n_i は分類 C_i の要素の個数、また $n_+ = n_i + n_j + n_k$ とする。

分類間の非類似度距離 $d(C_i, C_j)$ が所与のとき、 C_i と C_j を併合した新しい分類 $C_i \cup C_j$ と他の任意の分類 C_k との非類似度は、表中 α_i , β , γ を用いて次の通り。

$$d(C_i \cup C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|$$

3. 実証分析と結果

3.1 データ

大手セル・レンタル CD チェーン店から、個人が特定できないように加工した上で提供頂いた音楽 CD 販売の ID 付 POS データを用いる。データは 2002 年 11 月 1 日から 2003 年 12 月 21 日までの約 13 ヶ月間の、日本の大都市にある 5 店の計 707,085 件の購買履歴となっている。前処理として個別タイトルをアーティスト

表 7. 交互作用統計量と主効果との相関係数

非類似度係数	製品		消費者	
	M1 交互作用統計量	M16 主効果の相乗平均	M1 交互作用統計量	M16 主効果の相乗平均
M1 交互作用統計量	1.00	0.01	1.00	0.09
M2 Cohen's Kappa	0.96	0.01	0.94	0.04
M3 Ochiai	0.90	0.38	0.94	0.15
M4 Jaccard	0.90	0.34	0.95	0.11
M5 Sokal & Sneath 2	0.90	0.32	0.96	0.10
M6 Czesanowski	0.90	0.35	0.94	0.11
M7 Psi	0.88	0.27	1.00	0.13
M8 Kulczynski	0.80	0.27	0.95	0.10
M9 Yule's Q	0.58	0.19	0.64	0.15
M10 Russel & Rao	0.50	0.70	0.88	0.41
M11 Rogers & Tanimoto	0.07	-0.90	0.12	-0.94
M12 χ^2 metric	0.07	0.76	0.34	0.45
M13 Simple matching	0.07	-0.90	0.12	-0.94
M14 Sokal & Sneath 1	0.07	-0.89	0.11	-0.94
M15 Hamann	-0.07	0.90	-0.12	0.94
M16 主効果の相乗平均	0.01	1.00	0.09	1.00

表中の各非類似係数の M1 から M15 の番号は、製品についての交互作用統計量との相関係数の降順に一致させている。その相関の水準は、1.0 に近い M1~M8 と、0.5 付近の M9~M10 と、0.0 に近い M11~M16 の 3 グループに分かれる。

トに名寄せし、20 人以上の消費者に購入されたアーティストと 10 枚以上の CD を購入した消費者だけからなるデータセットに絞り込む。この結果、655 アーティスト $\times$ 3,577 消費者の 54,517 件の購買履歴を分析対象とし、これ以降、アーティストで名寄せした CD を便宜的に製品と呼称する。購買履歴と 2 値行列 $M = \{m_{ij}\}$ の関係は、製品 i を消費者 j が購入した履歴が POS データにある場合のみ $m_{ij}=1$ 、それ以外は $m_{ij}=0$ とする。したがって、消費者 j が同一製品を複数購入したり、繰返し購買した場合も $m_{ij}=1$ とする。

各類似係数と交互作用統計量、また主効果の相乗平均との相関係数を表 7 にまとめた。2 種類の類似係数行列 $S^1 = \{s_{i,j}^1\}$ と $S^2 = \{s_{i,j}^2\}$ の相関を計算する際は、それぞれ $s_{i,j}^k$ ($k=1, 2$) を第 $i+(j-1) \times m$ 要素に持つ列ベクトルを作成し、ベクトル間の相関係数として求め

る。

各類似係数の記載順の M1 から M16 は、製品間の類似係数について交互作用と相関の高い順にラベル付けしている。交互作用と主効果の相乗平均との相関係数は 0.01 (製品間) と 0.09 (消費者間) となり、2 つの類似係数はほぼ独立となっている。16 個の類似係数は、交互作用統計量と相関が高い M1~M8 までのグループと、交互作用統計量とはほぼ無相関だが主効果の相乗平均と相関が高い 11~M16 のグループ、およびどちらともいえない M9 と M10 の 3 個のグループに分かれる。分類に適切な類似係数は交互作用統計量と相関が高いという予測を、分類結果の妥当性の評価順位が M1 から M16 のラベルの順となる事で検証する。

クラスターバリデーションに関する検証は、クラス

タリング法の選択に結果が依存しないように、16種類の類似係数と5種類のクラスタリング法の組合せで行う。また分類数については敢えて分類数を決める理由と基準に乏しいので、敢えて最適な分類数一つ決める事はせず、観測対象と属性ともに10, 15, 20, 25, 30分類の5通りについて検証を行い、分類数の変化による分類結果の相違を確認する。

3.2 再現性の検証

階層的クラスタリング法について、50回の2-フォールド交差確認法により分類の再現性を評価する。分類の一致度はランドの修正指標で評価し、試行の平均値を分類数(10, 15, 20, 25, 30分類)毎に求める。分析の結果、各分類数の評価とも概ね同じ傾向を示したので、表8に15分類の場合の指標の有意水準をまとめ

ている。最長距離法とM1交互作用統計量またはM2Cohens's Kappaの組み合わせを除き、最長距離法と最短距離法は再現性に乏しいと分かる。またウォード法ではM15 Hamannを除き再現性があるが、群平均法ではM1交互作用統計量との相関が低いM10~M11およびM14~M16で再現性が乏しかった。

考察として、再現性が乏しくなるのは、分類の全体が少数の観測対象の振る舞いに左右されるからで、その要因はデンドログラムの形状に関係すると考える。図1の空間膨張的な場合は、併合するグループの決定がわずかな差で決まるため、分類結果が不安定になり易い。また空間収縮的な場合は、最初に形成されるグループが次に併合されるグループの選択に影響するため、最初のグループ形成が全体の分類に影響する。最長距離法と最短距離法の再現性が全般に乏しくなった

表8. 再現性の検証：ランドの修正指標と経験的パーセンタイル点

非類似係数	ウォード法		群平均法		最長距離法		最短距離法	
	ランドの修正指標	パーセンタイル点	ランドの修正指標	パーセンタイル点	ランドの修正指標	パーセンタイル点	ランドの修正指標	パーセンタイル点
M1 交互作用統計量	0.35	0.0%***	0.52	0.0%***	0.15	0.0%***	0.07	39%
M2 Cohen's Kappa	0.32	0.0%***	0.49	0.0%***	0.17	0.0%***	0.04	39%
M3 Psi	0.36	0.0%***	0.51	0.0%***	-0.15	100%	0.08	40%
M4 Sokal & Sneath 2	0.31	0.0%***	0.46	0.0%***	-0.15	100%	0.06	40%
M5 Jaccard	0.31	0.0%***	0.45	0.0%***	-0.15	100%	0.05	43%
M6 Czesanowski	0.32	0.0%***	0.45	0.0%***	-0.15	100%	0.07	42%
M7 Kulczynski	0.32	0.0%***	0.46	0.0%***	-0.15	100%	0.07	44%
M8 Ochiai	0.33	0.0%***	0.49	0.0%***	-0.14	100%	0.05	41%
M9 Yule's Q	0.38	0.0%***	0.40	0.0%***	-0.12	100%	-0.01	100%
M10 Russel & Rao	0.39	0.0%***	0.02	85%	-0.17	100%	-0.02	26%
M11 chi2 metric	0.39	0.0%***	-0.02	27%	-0.07	100%	-0.02	26%
M12 Rogers & Tanimoto	0.62	0.0%***	0.10	4%*	0.25	4%*	-0.02	100%
M13 Sokal & Sneath 1	0.58	0.0%***	0.10	4%*	0.21	19%	-0.01	100%
M14 Simple matching	0.59	0.0%***	0.08	30%	0.22	12%	-0.02	100%
M15 Hamann	-0.02	26%	-0.02	26%	0.04	7%	-0.02	26%
M16 主効果の相乗平均	0.57	0.0%***	-0.02	25%	-0.02	27%	-0.02	26%

優位水準 ***: 0.1% **: 1% *: 5%

ランドの修正指標は最大値が1で、偶然一致と同じ水準で0の値をとる。

経験的パーセンタイル点は、モンテカルロシミュレーションによって得たランドの修正指標の分布から求める。

のは、クラスタリング法の選択の要因であり、また群平均法において主効果と相関の高い類似係数の再現性が乏しくなるのは、類似係数の選択に要因があると考ええる。具体的なデンドログラムの類型は、次の均等性の検証で確認する。

3.3 均等性の検証

分類を構成する要素数の偏りを示す指標として、各分類を構成する要素数の多い上位半数の分類が全体の要素数に占める比率を計算する。例えば、100 個のメンバーを要素数が 80, 5, 5, 3, 2, 1, 1, 1, 1, 1 のクラスに 10 分類した場合には、 $(80+5+5+3+2)/100=95$ (%) と計算する。100% に近い値は、約半分 of 分類が空に近い事を示している。また、完全に一樣ならば 50% となる。階層的クラスタ法を用いる場合は、値の信

頼性を高めるために 50 回の交差確認法で作成した 100 通りの分類の平均値を求める。

表 9 に 15 分類の検証結果をまとめた。また同じく表 9 に、デンドログラムの 3 類型への分類を筆者が判断して記載した。図 3 の事例の通り、ほとんど全てのデンドログラムが明確な特徴を示し、判断に迷う事はなかった。図 4 の箱ひげ図は、交差確認法による各分類に含まれる要素数の変動を、横軸に $i (= 1, \dots, 15)$ 番目の大きさの分類を置いて示したものである。このグラフから、要素数の分布はどのデンドログラムの類型でも比較的安定しているので、上記の均等性の指標も同様に安定していると考えられる。また表 9 に記載した以外の条件である、10, 20, 25, 30 分類と消費者の分類についても評価したところ同様の傾向を示した。

表 9 より、空間収縮的なデンドログラムと均等性の

表 9. 均等性の検証：分類の類型と要素数の多い上位半数のグループに含まれる要素数の構成比

dissimilarity	PAM 法	ウォード法		群平均法		最長距離法		最短距離法	
	比率 %	類型	比率 %	類型	比率 %	類型	比率 %	類型	比率 %
M1 交互作用統計量	66	A	75	B	85	B	64	C	99
M2 Cohen's Kappa	60	A	75	B	81	B	62	C	99
M3 Ochiai	69	A	73	B	86	B	82	C	99
M4 Jaccard	62	A	76	B	85	B	82	C	99
M5 Sokal & Sneath 2	62	A	78	B	85	B	83	C	99
M6 Czesanowski	63	A	74	B	85	B	82	C	99
M7 Psi	67	A	77	B	86	B	82	C	99
M8 Kulczynski	67	A	80	B	85	B	82	C	99
M9 Yule's Q	61	A	66	C	76	B	81	C	99
M10 Russel & Rao	77	A	83	B/C	89	B	84	C	99
M11 Rogers & Tanimoto	99	C	98	C	99	C	99	C	99
M12 χ^2 metric	99	C	96	C	99	C	95	C	99
M13 Simple matching	99	C	98	C	99	C	99	C	99
M14 Sokal & Sneath 1	99	C	98	C	99	C	99	C	99
M15 Hamann	99	C	99	C	99	C	93	C	99
M16 主効果の相乗平均	99	C	98	C	99	C	99	C	99

分類の類型 A,B,C は図 1 の通り A. 空間保存的, B. 空間膨張的, C. 空間収縮的を示す。

表中の比率 (%) は、製品を 15 グループに分類して要素数の降順にグループを並べ、最初の半分 of グループに全体の要素数の何パーセントが含まれるか構成比を計算した値。ただし、第 8 グループについては要素数の半分以上を上位グループに含めた。

図 3. デンドログラムの空間の歪みの類型

2-フォールド交差確認法で15分類を50回作成し、各回において分類に含まれる要素数（縦軸）を降順に並べて（横軸）、値の分布を箱ひげ図に示している。

図 4. グループに含まれる要素数の分布の箱ひげ図（15分類の場合）

欠如は共通した現象であると分かる。また、主効果の相乗平均と相関が高い M11～M16 類似係数は、全順序の性質を持つ類似係数であるために、空間収縮的なデンドログラムを発生させていると考える。ところで、非階層的手法の PAM 法は非階層的な手法より均等な分類を作成している。M1～M10 の非類似係数について一つの例外を除き上位半分の占有率が階層的手法より小さく、PAM 法は均等な分類を作り易い事が分かる。M1 交互作用統計量と M2 Cohen's Kappa は他の類似係数に比べて最長距離法との組合せにおいて上位半分の構成比が20%ほど低い。この2つの類似係数は、偶然一致を差し引くためプラスの交互作用と同様にマ

イナスの交互作用も適正に評価できるので、最長距離法と相性が良いと考察する。

3.4 外的基準の検証

POS データの提供者から受領した音楽 CD の流通や管理目的に使用する音楽ジャンル区分を外的な基準として、購買履歴からクラスタリング法で作成した分類と、外的基準との一致の程度を検証する。対応の例として、表 10 に外的な音楽ジャンル区分を行方向に、交互作用統計量とウォード法の組み合わせで作成した分類を列方向として、各分類に該当する製品数を表に示している。

表 10. 外的基準の分類と分析の結果得られた分類との一致数の例

ジャンル	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	計
1 邦ポップ	19	104	35	22	40	15	36	42	38	0	0	0	0	0	0	351
2 ロック & ポップ	0	0	0	0	0	0	0	66	2	26	28	1	0	19	34	176
3 クラブ/ダンス	0	0	0	0	0	0	0	3	13	0	0	0	0	0	0	16
4 ハードロック/メタル	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	6
5 邦ジャズ	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
6 洋ジャズ	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
7 キッズ	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
8 アニメ/ゲーム	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3
9 ブラック/ソウル	0	0	0	0	1	0	0	8	1	0	2	33	0	0	0	45
10 ワールド	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	3
11 ヒップホップ/ラップ	0	0	0	0	0	0	0	3	0	0	0	4	32	0	0	39
12 イージーリスニング	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
13 クラシック Classic	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	3
14 レゲエ	0	0	0	0	2	0	0	3	0	0	0	1	1	0	0	7
計	19	107	35	22	43	15	36	141	54	26	32	39	33	19	34	655

行方向：外的基準は POS データと共にデータ提供者から提示された製品属性。

列方法：B1～B15 のラベルは交互作用統計量とウォード法の組合せで作成した 15 分類。

外的基準の音楽ジャンル区分の特徴として、表 10 の 1 行目と 2 行目に相当する「1 邦ポップ」と「2 ロック & ポップ」に該当する製品が多く、分類対象 655 製品の約 8 割の 527 製品がこの 2 ジャンルに区分されている。このように外的基準の分類は大きさが偏り、実務的に使いづらい分類となっている。これに対して交互作用統計量とウォード法の組み合わせで作成した分類は、外的基準で製品数の多い「1 邦ポップ」と「2 ロック & ポップ」を細分し、より均等で使い易い分類を与えている。同時に、この 2 分類以外は多数の分類に細分する事なく、14 行目の「14 レゲエ」を除き 1 分類か 2 分類への対応を与えている。このため 2 種類の分類に強い関係があると確認できる。このように交互作用統計量とウォード法の組み合わせで作成した分類は、外的基準と整合的で、結果として表 10 には 0 の値が多くなっている。

表 10 の行方向と列方向の分類の独立性をカイ二乗統計量で検証するのは、セルに 0 の値が多いという問題があるが、参考として 16 種類の非類似係数と 5 種類のクラスタリング法の組合せについて対応表のカイ

二乗統計量を求めて相対的な比較を行ったところ、M1～M10 の類似係数を選択した際のカイ二乗統計量は、M11 から M16 の類似係数を選択した場合に比べて総じて高く良い値を示した。またクラスター法の比較については、PAM 法とウォード法と群平均法が他の 2 つの方法に比べてカイ二乗統計量が大きく良い結果を示した。

3.5 内的基準の検証

表 11 に、交互作用の非類似係数とウォード法の組合せで作成したジョイントセグメンテーションの一例を示す。表中の各セルの値は、該当する分類の 2 値行列 $M = \{m_{ij}\}$ の小計に相当している。また分類名は、分類メンバーのアーティスト名を基に筆者が命名している。

これらのジョイントセグメンテーションのカイ二乗統計量を、外的基準の場合と同様に分析し、結果を表 12 にまとめた。消費者と製品それぞれ 5 通りの計 25 通りの分類数について計算し、表にはその集計値を記している。交互作用統計量と相関が高い類似係数に

表 11. ジョイントセグメンテーションの例

表中の値は購買履歴の該当数。ただし、同じ消費者による同一製品の繰り返し購買は重複カウントしない、製品と消費者の分類名は筆者による命名。

PAM 法かワード法を用いた組合せでカイ二乗統計量が 200 後半の高い値となり、好ましい結果を示している。特に M1 交互作用統計量, M8 Ochiai, M3 Psi, M2 Cohen's Kappa が上位を占め、カイ二乗統計量の順位の標準偏差も小さいため、順序関係も安定している事が分かる。

3.6 解釈可能性の検証

ラベルが M1~M8 の類似係数に PAM 法かワード法を適用した分類は、表 11 のジョイントセグメンテーションと概ね同じ結果になった。それ以外の類似係数とクラスタリング法の組み合わせを用いた分類は、著しく均等性が欠如していたり、あるいは同じ音楽性を持つアーティストが別の分類に属する不自然さがあり、第三者の意見を伺うまでも無く解釈可能でない分類と評価した。

そこで表 11 のアーティストのクラスタリング構造

についての解釈可能性のコメントを、職業的専門家であるデータ提供者に求めたところ、音楽 CD の店舗内の棚割に用いているジャンル区分と似ていて、邦楽と洋楽に分かれ、さらに細分している構造は違和感の無い自然な分類であるとの回答を得た。

3.7 結果

5 つのクラスターバリデーションの検証分析から 16 種類の類似係数の適切さを相対的に評価すると、まず主効果の相乗平均と高い相関を持つ M11~M16 の指標グループは再現性と均等性で劣後し、解釈に適したデンドログラムの形状を示さない事が分かる。また外的基準と内的基準で、交互作用統計量と相関が高い M1~M8 の類似係数のグループの下位に評価される。次に M9 と M10 の 2 指標は、再現性や均等性あるいは外的基準による評価で良好な結果もあるが、内的基準による評価が M1~M9 の指標の下位に位置してい

表 12. 内的基準の検証：ジョイントセグメンテーションのカイ二乗統計量のランキング

非類似係数	χ^2/df 統計量の順位の平均 (標準偏差)				χ^2/df の平均					
	PAM 法	ウォード法	群平均法	最長距離法	最短距離法	PAM 法	ウォード法	群平均法	最長距離法	最短距離法
M1 交互作用統計量	1.3 (0.5)	1.8 (1.0)	4.6 (1.9)	2.0 (0.0)	5.0 (1.8)	266	294	149	113	3
M2 Cohen's Kappa	4.1 (1.0)	3.2 (1.2)	1.9 (0.3)	1.0 (0.0)	10.0 (0.0)	249	294	163	139	2
M3 Psi	3.1 (1.9)	3.2 (0.8)	9.0 (0.0)	10.9 (1.4)	4.8 (1.6)	252	293	139	22	3
M4 Sokal and Sneath 2	5.1 (1.2)	6.0 (0.8)	6.0 (1.2)	7.2 (1.0)	7.5 (1.4)	246	282	145	30	2
M5 Jaccard	6.0 (1.2)	5.9 (0.7)	7.0 (0.9)	5.8 (1.0)	7.2 (1.2)	243	283	145	31	2
M6 Czesanowski	6.1 (1.4)	6.2 (1.1)	5.3 (1.3)	6.0 (1.3)	7.1 (1.8)	239	284	146	31	2
M7 Kulczynski	8.0 (0.7)	7.9 (0.3)	6.4 (1.4)	6.8 (1.5)	6.5 (1.8)	231	278	146	30	2
M8 Ochiai	2.6 (1.2)	1.8 (0.5)	3.8 (1.0)	4.4 (0.7)	3.8 (1.1)	258	296	149	31	3
M9 Yule's Q	9.6 (0.5)	9.6 (0.5)	1.1 (0.3)	3.0 (0.0)	11.4 (0.5)	187	257	168	34	1
M10 Russel and Rao	9.0 (1.1)	9.4 (0.5)	10.0 (0.0)	8.8 (0.7)	2.2 (0.8)	221	252	34	28	3
M11 カイ二乗距離	13.7 (0.8)	11.0 (0.0)	11.0 (0.0)	10.7 (0.7)	1.0 (0.0)	80	183	9	22	8
M12 Rogers and Tanimoto	11.0 (0.0)	12.0 (0.0)	13.9 (1.1)	12.9 (0.7)	14.1 (0.9)	139	118	1	16	1
M13 Sokal and Sneath 1	12.7 (0.4)	14.0 (0.0)	13.6 (1.1)	12.4 (1.0)	13.7 (1.1)	117	106	1	16	1
M14 Simple matching	12.7 (0.4)	13.0 (0.0)	13.4 (1.0)	13.0 (1.2)	13.7 (0.8)	117	113	1	16	1
M15 Hamann	15.0 (0.0)	16.0 (0.0)	15.0 (1.7)	15.0 (0.0)	15.8 (0.6)	3	1	1	2	0
M16 主効果の相乗平均	16.0 (0.0)	15.0 (0.0)	14.0 (1.6)	16.0 (0.0)	12.2 (1.6)	0	3	1	1	1

表の値は、消費者と製品の分類数をそれぞれ 10, 15, 20, 25 および 30 個の 5 通りで作成した計 5=25 パターンの対応表の $\chi^2=df$ 統計量の降順の順位と、その順位の標準偏差である。ただし、各分類パターンの χ^2/df は 2-フォールド交差確認法による 50 回のクラスタリングの平均値を採用している。

る。整理すると指標グループとして評価の良い順に M1～M8, M9 と M10, M11～M16 となる。

M1～M8 の指標は概ね同じ有効性を示しているが、特に M1 交互作用統計量と M3 Ochiai が外的基準と内的基準の評価で安定して上位にある。また M1 交互作用統計量と M2 Cohen's Kappa は最長距離法を用いた場合にも高い有効性を示し、交互作用統計量がマイナスの場合も適切に評価していると推測する。この結果と、16 指標の有効性が全体として交互作用と相関の高い順に並んでいることを勘案すると、相対的に優れた指標は M1 交互作用統計量と結論できる。

M9 以下の類似係数がクラスタリングに適さなかった事は、類似係数を主効果と交互作用統計量の関数として分類した表 4 の結果と整合する。大規模疎行列では主効果が小さくなり、このとき表 4 でタイプ III と分類した M9 以下の類似係数は交互作用統計量と無相関に近づく関数上の特徴を持っている。このため、これらの類似係数では、図 3 のカイ二乗距離の空間収縮

的なデンドログラムのように、主効果の大小の順にクラスターの凝縮を導く。

4. 考察と今後の課題

大規模疎行列データのクラスタリングに適した類似係数を実験的に評価し、また類似係数を交互作用統計量の導関数として評価した結果、クラスタリングに適した類似係数は主効果がゼロに近づいても交互作用統計量との相関を失わない類似係数である事を確認できた。そのような特徴を持つ類似係数は、主効果の影響を受けない類似係数と、主効果と交互作用の影響を等しく受ける類似係数の 2 つのタイプに分類された。本研究の応用分野として、大規模疎行列データを扱う事が多いゲノム解析やマーケティングの POS データ解析があろう。記述的なデータ解析の第一段階として、交互作用統計量との相関を失わない類似係数を用いて観測対象の類似性を評価する事は有益と考える。

特に amazon.com の購買推奨に利用されている仕組み (Linden, Smith, & York, 2003) として知られる協調フィルタリング (Goldberg, Nichols, Oki, & Terry, 1992; 芳賀, 2008) の計算では類似係数が中心的な役割を果たす。このため推奨の効果の違いで類似係数の選択基準を評価する事ができ、クラスターバリデーションで評価する場合に比べて検証の作業が容易となる。石田・西尾・佐藤 (2009) は、実証分析として協調フィルタリングの手法で過去の POS データから将来の購買を予測し、交互作用統計量が最も優位な予測力を持つ事を示している。別の目的と検証法でも交互作用統計量の優位が示された事は、分類に限らず類似係数に期待される共通の性質が存在する可能性を示唆していて興味深い。

今後の課題としては、序論で言及したようにピアソンのカイ二乗統計量以外の交互作用を検討する事がある。また、教師なし分類の妥当性を評価するクラスターバリデーションについて、まだ標準的な手続きや評価基準が定まっているとは言い難く、さらに議論を進めて評価の客観性を高める余地があると考ええる。

References

- Albatineh, A.N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On Similarity Indices and Correction for Chance Agreement. *Journal of Classification*, **23**, 2, 301-313.
- Batagelj, V., & Bren, M. (1995). Comparing resemblance measures. *Journal of Classification*, **12**, 73-90.
- Baulieu, F.B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, **6**, 233-246.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.
- Duarte, J.M., Santos, J.B., & Melo, L.C. (1999). Comparison of similarity coefficients based on RAPD markers in the common bean. *Genetics and Molecular Biology*, **22**, 427-432.
- Dubien, J.L., & Warde, W.D. (1979). A Mathematical Comparison of the Members of an Infinite Family of Agglomerative Clustering Algorithms. *Canadian Journal of Statistics*, **7**, 29-38.
- Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, **31**, 651-659.
- Goldberg, D., Nichols, D., Oki, B.M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, **35**, 21, 61-70.
- Goodman, L.A. (1996). A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *Journal of the American Statistical Association*, **91**, 408-428.
- Gordon, A.D. (1999). *Classification* (2nd ed.). Monographs on Statistics and Applied Probability, 82, Chapman & Hall/CRC.
- Gower, J.C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5-48.
- 芳賀麻誉美 (2008). レコメンデーション技術の理論と実際. 日本行動計量学会第36回大会チュートリアルセミナー.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, **2**, 193-218.
- 石田 実・西尾チヅル・佐藤忠彦 (2009). 同一製品の繰返し購買が無い市場の新製品の属性に対して協調フィルタリングを用いる競争市場構造分析手法の提案. 筑波大学企業科学専攻リサーチレポート, 09-01.
- Janson, S., & Vegelius, J. (1981). Measures of ecological association. *Oecologia*, **49**, 371-376.
- Kaufman, L., & Rousseeuw, P.J. (1990). *Finding Groups in Data*. New York, USA: John Wiley & Sons, Inc.
- Lance, G.N., & Williams, W.T. (1967). A general theory of classificatory sorting strategies, 1. Hierarchical systems. *The Computer Journal*, **9**, 373-380.
- Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, **7**, 1, 76-80.
- Mandel, J. (1971). A New Analysis of Variance Model for Non-additive Data. *Technometrics*, **13**, 1, 1-18.
- Milligan, G.W. (1996). Clustering validation: Results & implications for applied analyses. In *Clustering and Classification* (eds P. Arabie, L.J. Hubert, & G. De Soete), World Scientific, Singapore, 341-375.
- Milligan, G.W., & Cooper, M.C. (1986). A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behavioral Research*, **21**, 4, 441-458.
- 大隈 昇・L. ルバール・A. モリナウ・K.M. ワーウィック・馬場康維 (1994). 記述的多変量解析法, 日科技連出版社.
- Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 336, 846-850.
- Warrens, M.J. (2008). On the indeterminacy of resemblance measures for (presence/absence) data. *Journal of Classification*, **25**, 125-136.
- Wedel, M., & Kamakula, A.W. (2000). *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.). Kluwer Academic Publisher.

Yadohisa, H., Takeuchi, A., & Inada, K. (1999). Developing criteria for measuring space distortion in combinatorial cluster analysis and methods for controlling the distortion. *Journal of Classification*,

16, 1, 45-62.

(2009 年 9 月 18 日受付, 2010 年 11 月 19 日最終修正)