

消費者 DATA から情報を得る方法

Methods to mine the knowledge from the purchased consumer data

流通科学大学・商学部 野口 博司¹ 神戸大学・海事科学部 磯貝 恭史²

Hiroshi Noguchi, Faculty of Commerce, University of Marketing and Distribution Sciences¹

Takafumi Isogai, Faculty of Maritime Sciences, Kobe University²

要旨

流通業界では、POS 等の消費者購買 data から、売れ筋商品の傾向を発見し、また優良顧客を発掘する方法について研究している。この時代の要請に応じ、我々は、大量の database から business に役立つ情報を得るための方法について研究を行った。本報告は、購買 data の分析に有効とされている数量化の方法、correspondence analysis, market basket analysis を取り上げて、実用面からどのような目的の時にどの手法が有効であるかを比較検討する。そして、今後の活用の留意点を提言する。特に correspondence analysis においては、反応数の少ない項目を含む行と列の同時布置を行う際に、その重みを調整する距離尺度のあり方について提案する。

Keywords: 数量化 3 類, correspondence analysis, market basket analysis, decision tree, data mining, positioning methods.

1. はじめに

business 現場では、顧客記録の database が普及して、売れ筋商品の傾向や優良顧客を発見するための手法について研究がなされている。しかし、必ずしも、data 解析の専門家がその研究に取り組んでいるとは限らず、結果が興味を引くものであれば、各手法のもつ解法特性を理解せずに、その結果を信じている。各手法には、それぞれ特性があり、その特性をよく理解した上で、その解析の結果を考察していく必要があると考える。そこで、我々は、これら顧客購買 data を扱う代表的な手法を取り上げて、実用面の立場から、同じ data からでも、どのような結論が導けるかを示す。また、各手法間の特性について比較を行い、今後の活用上での留意点を提言する。

2. 研究方法について

2. 1 購入 data の収集とその内容について

神戸の大学生 2, 3 年生 42 人を対象にして、「午後から数人の談話会を開催することを前提として、a convenience store でどのような snack 菓子と飲み物を購入するかを模擬実験してもらった。予算金額は 1500 円である。購入された延べ品目は、snack 菓子は、ポテトチップス、クッキー、チョコレート、ケーキ、ビスケット、パン、和菓子、7 種の 9 品目であった。飲み物は、紅茶、コーラ、ジュース、コーヒー、日本茶、ウーロン茶、ミルク、スポーツ飲料の 8 品目である。得られた data を表 1 に示す。表 1 で、1 となっているのは、被験者が購入した品目であり、0 は購入しなかった品目である。1 番多かった品目数は、8 品目であり、1 番少なかったのは、2 品目であった。

次に今回検討の対象とした解析法を示す。

表1 今回の実験で得たデータ

	コーヒー	紅茶	コーラ	ジュース	日本茶	ウーロン茶	スポーツ飲料	ミルク	チョコレート	ポテトチップス	アられ	ハン	ビスケット	ケーキ	クッキー	和菓子	ポッキー	性別
藤田	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0
伊藤	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
赤木	1.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
江尻	0.0	0.0	1.0	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	1.0	1.0
谷村	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	2.0
森	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
新井	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0
小澤	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0
清下	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
濱岡	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
野田	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0
植本	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
近藤	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
八木	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
大塚布佐	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	2.0
穴吹	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
鈴木	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
国宝	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0
長濱	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
天野	0.0	1.0	1.0	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	1.0	1.0
明尾	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
岩本	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
中垣	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
井村	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	2.0
松崎	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
水田	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0
佐藤	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
岡本	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
平田	0.0	1.0	1.0	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	1.0	1.0
村中	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0
大濱	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	2.0
長畑	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0
森下	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0
岡内	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	2.0
石黒	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0
内田	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
有吉	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
西口	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
畑岡	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
中尾	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0
竹元	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
浅田	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0

2. 2 検討した解析法について

今回、検討対象にした解析方法は、消費者購買 data で、1 番よく活用されている下記の三つである。即ち、

- (1) 数量化法による購入品目の pattern 分析
- (2) correspondence analysis による snack 菓子と飲み物の関係
- (3) market basket analysis による 購入品目間の関連である pattern の抽出、即ち data mining である。

3. 数量化法と correspondence analysis について

(1) 数量化法と (2) correspondence analysis についてであるが、分割表や林の数量化法との数値理論的な関係は、既に、磯貝・野口の「特異値分解とその応用」[1]で報告してあるので、数量化法についての説明は省略する。本報告では、correspondence analysis における反応の少ない項目を含む行と列との同時布置を行う際の調整方法について提言したいので、correspondence analysis の距離尺度[2]については、説明することにする。

3. 1 correspondence analysis の距離尺度について

correspondence analysis の距離尺度を述べる前に、まず特異値分解について説明する。

$n \times p$ 行列 X が与えられた時、 X の列 vector についての内積を考える時に、 $n \times n$ の正定符号行列 M をおく。そして、 X の行 vector について内積を考える時に $p \times p$ の正定符号行列 N とおいて、重みとして与えられている重み付きの Euclid 内積を考えることにする。

この時、一般化特異値分解とは、rank r ($r \leq p$) の data 行列 $n \times p$ の X が与えられた時、 $X = U D_p V^T$ と分解されることを言う。ここで、 $D_p = \text{diag}(d_1, d_2, \dots, d_r)$ ($d_1 \geq d_2 \geq \dots \geq d_r$) であり、 $U^T M U = I_r$ 、 $V^T N V = I_r$ を満たす。即ち、 U は $n \times r$ 、 D_p は $r \times r$ 、 V^T は $r \times p$ の行列である。

そこで、 $m \times t$ の確率行列 $P = (p_{ij})$ が与えられた時、次の式を計算して、この Q の特異値分解を求めることにする。

$$Q = R^{-1/2} P C^{-1/2} \quad \{R = \text{diag}(p_{1.}, p_{2.}, \dots, p_{m.}), C = \text{diag}(p_{.1}, p_{.2}, \dots, p_{.t})\}. \quad (1)$$

Q の持つ rank を $r+1$ と考えて、 $S = Q^T Q$ の spectral 分解を求める。

$$Q^T Q = \tilde{V} \Lambda \tilde{V}^T, \quad \Lambda = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_r), \quad \lambda_0 \geq \lambda_1 \geq \dots \geq \lambda_r > 0$$

$$\tilde{V} = (\tilde{v}_0, \dots, \tilde{v}_r), \quad \tilde{V}^T \tilde{V} = I_{r+1} \quad (2)$$

ここで、 $\sqrt{\lambda_i} = \rho_i$ ($\tilde{D} = \text{diag}(\rho_0, \rho_1, \dots, \rho_r)$) とおき、

$$\tilde{U} = Q \tilde{V} \tilde{D}^{-1} \quad (3)$$

とおけば、 Q の特異値分解は次の (4) 式となる。

$$Q = \tilde{U} \tilde{D} \tilde{V}^T, \quad \tilde{U} = (\tilde{u}_0, \tilde{u}_1, \dots, \tilde{u}_r), \quad \tilde{U}^T \tilde{U} = I_{r+1} \quad (4)$$

そこで、 $U = R^{-1/2} \tilde{U}$ 、 $V = C^{-1/2} \tilde{V}$ と変換すると、

$$R^{-1} P C^{-1} = R^{-1/2} Q C^{-1/2} = R^{-1/2} \tilde{U} \tilde{D} \tilde{V}^T C^{-1/2} = U \tilde{D} V^T \quad (5)$$

ここで $\tilde{D} = \text{diag}(\rho_0, \rho_1, \dots, \rho_r)$ 、 $\rho_0 = 1 \geq \rho_1 \geq \dots \geq \rho_r > 0$ であり、 R は基準化条件 $U^T R U = I_{r+1}$ を満たし、 C も $V^T C V = I_{r+1}$ の基準化条件を満たす。このとき、 $P = (p_{ij})$ の確率分布において、 ρ は $P = (p_{ij})$ の確率分布の I と J についての 2 つの実数値確率変数 $f(I)$ と $g(J)$ の相関係数に相当する。そして、 ρ が最大になるような score x 、 y の score vector が特異値 ρ_K ($K=1, 2, \dots, r$) に対応する固有 vector u_K と v_K の (u_K, v_K) ($K=1, 2, \dots, r$) となる。即ち、行列の固有値問題にして表現すると $R^{-1} P y = \rho x$ 、 $C^{-1} P x = \rho y$ であり、これら K 次元での図示表現は、自明な解 $\rho_0 = 1$ 、 $x_0 = 1_m$ 、 $y_0 = 1_t$ を取り除き、上から K 番目までの解を用いることになる。

$$U_K \tilde{D}_K = F, \quad \{F^T = (f_1 : f_2 : \dots : f_t)\} \quad (6), \quad V_K \tilde{D}_K = G, \quad \{G^T = (g_1 : g_2 : \dots : g_m)\} \quad (7)$$

となる。(6) 式の F は主成分分析の $U_0 D_p (a)$ の主成分得点に対応する。 F の行 vector f_i の布置における f_i の Euclid 相当の距離関係 $\|f_i - f_j\|$ は、 m 個の行 vector 中に対応する第 i 行 vector と第 j 行 vector の Euclid 距離関係

$$\left\{ \sum_{s=1}^t \left(\frac{p_{is}}{p_{i.} \sqrt{p_{.s}}} - \frac{p_{js}}{p_{j.} \sqrt{p_{.s}}} \right)^2 \right\}^{1/2} \quad (8)$$

を近似する。

G は共分散型 bi-plot $D_p (2) V^T (2)$ に対応する。そして、 G の行 vector g_j の布置における g_j の Euclid 距離相当の関係 $\|g_i - g_j\|$ は、 t 個の列 vector 中に対応する第 i 列 vector と第 j 列 vector の Euclid 距離関係

$$\left\{ \sum_{s=1}^m \left(\frac{p_{si}}{p_{i.} \sqrt{p_{.s}}} - \frac{p_{sj}}{p_{j.} \sqrt{p_{.s}}} \right)^2 \right\}^{1/2} \quad (9)$$

を近似する。

(c) (d) を、共に Benzecri の χ^2 (chi-square) 距離という。

F, G間の関連を調べることは、FとGとの行 vector 間の内積情報を取ることであり、即ち、行・列間の分布間距離を比べていることになる。この考えは Pearson の χ^2 統計量の考え方と一致する。即ち、correspondence analysis は、主成分分析や bi-plot のような行列の近似を目的としたものとは異なる。行・列間の分布間距離を Euclid 距離と同じ扱いで、correspondence analysis を用いるならば、用いる data 行列 X の入出力項目の単位（例えば金額や得点で全て示せるとか）を揃えておく必要がある。

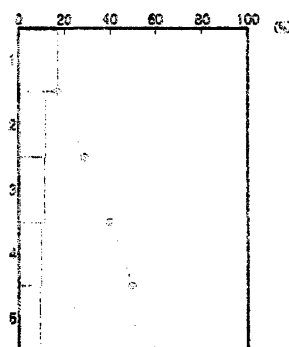
4. 解析結果

4. 1 数量化法の結果について

表 1 の data から購入品目間の関係が pattern として現れないか数量化法で解析した。

(1) 固有値と固有 vector

変数の数: 17 成分数: 5



No	固有値	寄与率	累積寄与率
1	0.383	0.171	0.171
2	0.252	0.118	0.289
3	0.236	0.111	0.400
4	0.209	0.098	0.498
5	0.199	0.094	0.592
6	0.181	0.078	0.667
7	0.148	0.070	0.737
8	0.133	0.062	0.799
9	0.109	0.051	0.851
10	0.079	0.037	0.888
11	0.073	0.034	0.922
12	0.059	0.028	0.950
13	0.042	0.020	0.969
14	0.027	0.013	0.982
15	0.024	0.011	0.993
16	0.014	0.007	1.000

図 1 固有値の出方

求めた固有値は図 1 である。通常、数量化法の固有値の出方はなだらかであるが、今回の解析結果もなだらかであり、pattern がまとまる傾向はなさそうである。

表 2 固有 vector の表

変数名	成分1	成分2	成分3	成分4	成分5
コーヒー	0.078	-0.146	-0.275	-0.402	0.101
紅茶	-0.241	-0.243	0.145	0.211	-0.181
コーラ	-0.206	-0.146	0.307	-0.266	-0.237
ジュース	0.112	0.253	0.007	0.086	-0.335
日本茶	0.498	-0.207	0.123	0.226	0.209
ウーロン茶	-0.258	0.151	-0.503	0.208	0.495
スポーツ飲料	0.191	0.555	0.435	-0.006	0.347
ミルク	0.221	0.135	-0.422	-0.246	-0.241
チョコレート	0.049	0.288	-0.016	0.283	-0.054
ポテトチップス	-0.073	0.059	0.040	0.005	-0.003
アラレ	0.374	0.095	-0.175	0.268	-0.257
ハン	0.422	-0.031	0.004	-0.409	0.201
ビスケット	-0.113	-0.228	-0.104	0.025	0.386
ケーキ	0.105	-0.271	-0.192	0.129	-0.179
クッキー	-0.273	0.205	-0.034	0.158	-0.054
和菓子	0.168	-0.423	0.229	0.294	0.165
ホッキー	-0.168	0.010	0.199	-0.335	0.083

表 2 は、各成分の固有値に対する固有 vector の表である。これより、成分 1 は+に日本茶と

お腹が膨れるようなパン、アレルが関係し、成分2は+にスポーツ飲料、-に和菓子となる。しかし、これらの購入 pattern の学生数は少なく特異な傾向である。その他の成分の意味については判明できなかった。

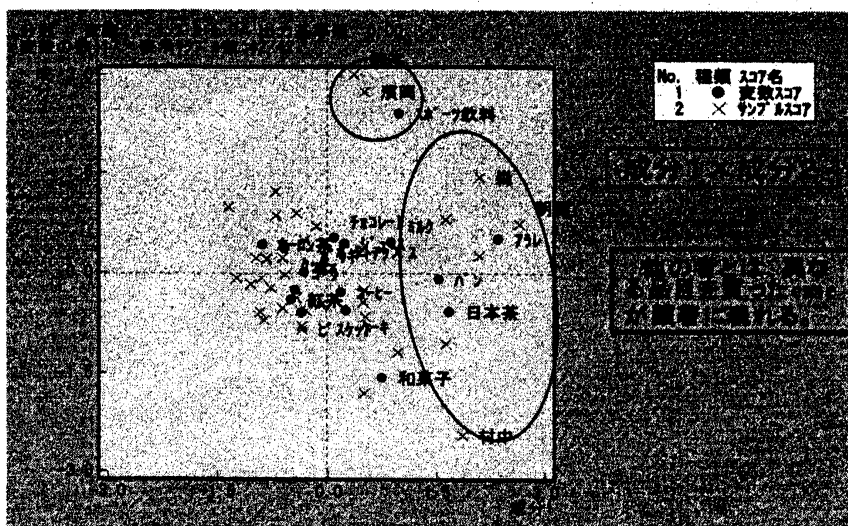


図2 品目(変数 score)と購入者(氏名 score)のMap

図2は成分1と成分2における購入者と購入された品目の score を同時布置にした図であるが、前述のように、他の者と異なる品目を購入した学生とその品目が外側に現れている。

(2) 数量化法の特徴について

以上より、数量化法の特徴は

- ①従来から言われているように、特異な反応をする data に対して、非常に感度が高い。従って、特異な品目を購入する消費者の抽出には向くといえる。
- ②逆に、平均的に多い pattern 傾向を抽出するのには不適といえる。
- ③一般的に各成分の意味づけは難しいことが多く、各成分は数理理論上異なる成分を抽出しているにも拘わらず、固有 vector は重複して各成分に出てくる傾向にある。

4. 2 品目間の独立性の検定結果(χ^2 検定)について

表3 品目間の関係表

製品名	コーヒー	紅茶	コーラ	ジュース	日本茶	ウーロン茶	スポーツ飲料	ミルク	チョコレート	ポテトチップス	アレル	パン	ビスケット	ケーキ	クッキー	和菓子	ポッキー
コーヒー	-	3.055	0.000	0.000	0.088	0.088	1.815	1.816	1.235	1.815	1.050	0.875	1.235	0.223	0.438	0.138	0.000
紅茶	3.055	-	2.299	0.982	1.222	0.076	2.937	1.736	0.004	0.284	1.909	6.364*	0.877	1.462	0.795	1.329	0.127
コーラ	0.000	2.299	-	1.167	2.800	2.800	2.423	1.211	0.867	2.423	1.575	0.259	1.287	3.008	0.656	0.092	5.477*
ジュース	0.000	0.982	1.167	-	0.000	2.800	0.748	0.881	1.188	0.120	2.800	0.259	1.287	0.928	0.032	0.576	0.856
日本茶	0.088	1.222	2.800	0.000	-	1.680	0.846	0.045	0.020	0.646	10.500**	5.600*	0.124	0.557	6.300*	10.832**	2.800
ウーロン茶	0.088	0.076	2.800	2.800	1.680	-	0.646	0.045	0.020	0.646	0.420	1.400	0.494	0.022	2.800	0.884	0.000
スポーツ飲料	1.815	2.937	2.423	0.748	0.846	0.646	-	0.437	0.920	0.249	0.162	0.957	0.760	1.448	0.120	0.340	0.120
ミルク	1.816	1.736	1.211	0.881	0.045	0.045	0.437	-	0.898	1.415	2.906	3.065	1.335	0.217	0.019	0.597	0.881
チョコレート	1.235	0.004	0.667	1.186	0.020	0.020	0.920	0.898	-	0.920	0.368	0.148	0.036	2.366	2.973	0.166	5.567*
ポテトチップス	1.815	0.284	2.423	0.120	0.646	0.646	0.249	1.415	0.920	-	0.162	7.239**	4.751*	0.009	0.120	2.126	0.748
アレル	1.050	1.909	1.575	2.800	10.500**	0.420	0.162	2.906	3.068	0.162	-	2.188	0.494	0.358	1.575	0.221	2.800
パン	0.875	6.364*	0.259	0.259	5.600*	1.400	0.957	3.065	0.148	7.239**	2.188	-	0.026	0.019	5.250*	0.414	0.146
ビスケット	1.235	0.877	1.287	1.287	0.124	0.494	0.760	1.335	0.036	4.751*	0.494	0.026	-	0.198	0.116	0.102	1.287
ケーキ	0.223	1.462	3.008	0.928	0.557	0.022	1.448	0.217	2.366	0.009	0.356	0.019	0.198	-	3.008	0.751	2.883
クッキー	0.438	0.795	0.656	0.032	6.300*	2.800	0.120	0.019	2.973	0.120	1.575	5.250*	0.116	3.008	-	0.576	0.032
和菓子	0.138	1.329	0.092	0.576	10.832**	0.884	0.340	0.597	0.166	2.126	0.221	0.414	0.102	0.751	0.576	-	1.865
ポッキー	0.000	0.127	5.477*	0.856	2.800	0.000	0.120	0.881	5.567*	0.748	2.800	0.148	1.287	2.883	0.032	1.865	-
性別	1.371	1.909	6.300*	0.700	0.034	9.909**	0.846	0.045	0.968	0.646	0.420	1.400	7.906**	0.557	2.800	0.221	0.000

さて、数量化の 0-1 反応 data を cross 集計して pattern の傾向を探るのが correspondence analysis である。そこで、個々の品目間の関係を表 3 のように求めた。即ち、correspondence analysis の特性を検討するに当たって、飲み物と食べ物 (snack 菓子) の関連を探ることとした。

飲み物と食べ物 (snack 菓子) の関係を cross 集計すると表 4 のようになった。

表 4 飲み物と食べ物の cross 集計表

飲み物/スナック菓子	チョコレート	ポテトチップス	アヲレ	パン	ビスケット	ケーキ	クッキー	和菓子	ポッキー
コーヒー	4	12	0	3	4	5	5	1	8
紅茶	8	19	0	0	5	8	10	3	12
コーラ	6	18	0	2	2	3	9	2	14
ジュース	9	17	2	2	2	7	8	1	9
日本茶	3	6	2	3	1	3	0	3	2
ウーロン茶	3	7	0	0	2	2	5	0	4
スポーツ飲料	2	3	0	1	0	0	1	0	2
ミルク	3	4	1	2	0	2	2	0	2

表 4 より、ポテトチップスはよく購入され、その時の飲み物は紅茶、コーラ、ジュース、コーヒーとなる。次いでポッキーがよく購入され、飲み物はコーラ、紅茶となる。逆にあまり購入されないのはアヲレ、パン、和菓子であり、アヲレの時の飲み物はジュース、日本茶となる。パンのときはコーヒー、日本茶となり、和菓子は日本茶、紅茶となっているが、これらは極めて例数が少ない。

表 5 は、表 4 を元に飲み物の品目と食べ物の品目とにおいて関係があるのかを χ^2 検定した結果を示している。

表 5 飲み物と食べ物において個々品目間の独立性の検定結果 (χ^2 検定)

飲み物/スナック菓子	チョコレート	ポテトチップス	アヲレ	パン	ビスケット	ケーキ	クッキー	和菓子	ポッキー
コーヒー	1.235	1.615	1.050	0.875	1.235	0.223	0.438	0.138	0.000
紅茶	0.004	0.264	1.909	6.384*	0.877	1.462	0.795	1.329	0.127
コーラ	0.667	2.423	1.575	0.259	1.287	3.008	0.656	0.092	5.477*
ジュース	1.186	0.120	2.800	0.259	1.287	0.928	0.032	0.576	0.656
日本茶	0.020	0.646	10.500**	5.600*	0.124	0.557	6.300*	10.832**	2.800
ウーロン茶	0.020	0.646	0.420	1.400	0.494	0.022	2.800	0.884	0.000
スポーツ飲料	0.920	0.249	0.162	0.957	0.760	1.448	0.120	0.340	0.120
ミルク	0.898	1.415	2.906	3.065	1.335	0.217	0.019	0.597	0.681

各項目間では、アヲレの時の飲み物は日本茶であり、和菓子の時も日本茶という関係は存在する結果となっている。しかし、例数は少ない。

ここで、全体の行の食べ物項目と列の飲み物項目との項目間が独立であるかどうかの検定を行った。行の数は 8、列の数は 9 であり、自由度 56 となり、 χ^2_0 統計量 = 51.185 となる。P 値上側 0.657 で、全体の飲み物項目と食べ物項目間の頻度の出方においては独立であるという仮説は棄却されない。即ち、食べ物と飲み物においては関連がないとなる。

correspondence analysis の実施は、本来 χ^2 検定が有意になった場合に用いるべきであると考えが、一般的には、このことは考慮されなく解析されている場合が多い。

4. 3 correspondence analysisの結果について

χ^2 値では有意にならなかったが, correspondence analysis を実施した. その結果が表 6 である. 表 7 は, 食べ物品目と飲み物品目間の分布間距離を示す 2 次元までの score を求めたものである. この score は, 3. 1 節の (8) および (9) 式の距離尺度に相当する.

(1) 固有値と次元 score について

表 6 Correspondence Analysis の解析結果

次元	特異値	要約パーセント	カ2乗	有意確率	パーセントの寄与率		信頼特異値	
					説明	累積	標準偏差	相関
1	.323	.104			.593	.593	.067	-.054
2	.180	.032			.183	.776	.051	
3	.145	.021			.119	.895		
4	.116	.013			.077	.972		
5	.050	.003			.014	.986		
6	.043	.002			.011	.997		
7	.023	.001			.003	1.000		
要約合計		.176	51.185	.657 ^a	1.000	1.000		

a. 自由度56

表 7 元の Correspondence Analysis における各次元の score

購入品	1次元	2次元	層別
4コーヒー	0.087	-0.148	飲み物
1紅茶	0.348	-0.438	飲み物
2コーラ	0.307	0.148	飲み物
3ジュース	-0.121	0.229	飲み物
5日本茶	-1.616	-0.587	飲み物
6ウーロン茶	0.623	0.052	飲み物
8スポーツ飲	-0.082	1.182	飲み物
7ミルク	-0.815	0.938	飲み物
dチョコレート	-0.138	0.356	食べ物
aポテトチップ	0.083	0.039	食べ物
iアヲレ	-2.656	0.246	食べ物
gパン	-1.411	0.688	食べ物
fバスケット	0.405	-0.874	食べ物
eケーキ	-0.200	-0.367	食べ物
cクッキー	0.551	0.190	食べ物
h和菓子	-0.998	-1.503	食べ物
bホッケー	0.325	0.102	食べ物

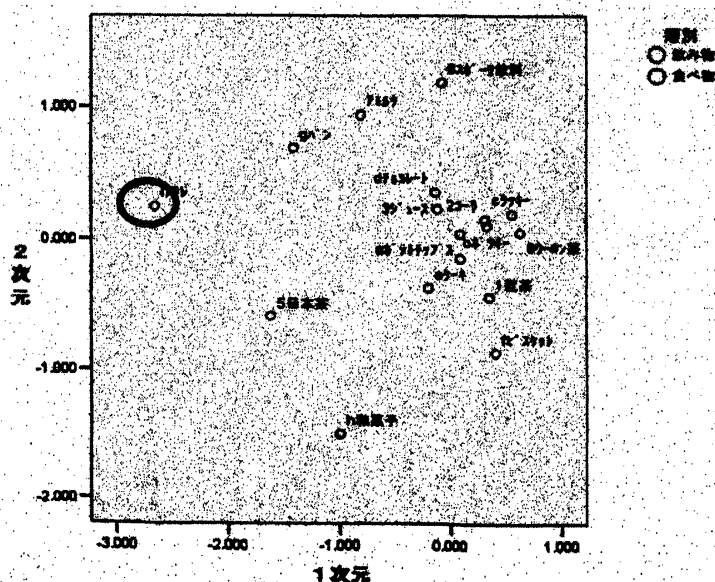


図 3 元の Correspondence Analysis における飲み物と食べ物の同時布置

表 7 の各品目の score について, 同時布置を描いたのが図 3 である. ポテトチップスと紅茶やコーラ

は近く、ホッケーもその近くにある。独立性の検定では有意にならなかったが、表3で、品目間の関係があるとみなしたものは、やはり布置においても近い位置関係にある。しかし、各品目間においては、有意となったアールと日本茶との位置や和菓子の位置等は離れている。また、アールの購入の際には日本茶とジュースとが同時購入されているにも拘わらず外側に現れて、反応数の少ない case の特徴として、数量化法と似た結果が、correspondence analysis にも表れている。

そこで、我々は、反応数を調整して、行と列とを同時布置で比較できるための correspondence analysis 距離尺度を提言する。

即ち、correspondence analysis の図示表現を、主成分分析または bi-plot 表現のような行列の近似を目的とするものに近づけるためには、座標表現を行うための行列として

$$F = R^{1/2} U_K D_K, \quad G = C^{1/2} V_K D_K \text{ を採用した.}$$

即ち F の行 vector f_i^T の布置における f_i の Euclid 相当の距離関係 $\|f_i - f_j\|$ は

$$\left\{ \sum_{s=1}^k \left(\frac{p_{is}}{\sqrt{p_{i.}} \sqrt{p_{s.}}} - \frac{p_{js}}{\sqrt{p_{j.}} \sqrt{p_{s.}}} \right)^2 \right\}^{1/2} \quad (10)$$

とおき、G の行 vector g_j^T の布置における g_j の Euclid 相当の距離関係 $\|g_j - g_l\|$ として

$$\left\{ \sum_{s=1}^m \left(\frac{p_{si}}{\sqrt{p_{j.}} \sqrt{p_{s.}}} - \frac{p_{sl}}{\sqrt{p_{l.}} \sqrt{p_{s.}}} \right)^2 \right\}^{1/2} \quad (11)$$

とおく。(10)、(11) 式の分母を見て判るように、これは行と列とを反応数に応じて同じ重みで調整した距離尺度となっており、行と列は同じ距離にて同時に比較可能となる。

表8 我々が提言した距離尺度
における各次元の score

購入品	1次元	2次元	層別
4コーヒ	0.033	-0.056	飲み物
1紅茶	0.164	-0.207	飲み物
2コーラ	0.135	0.065	飲み物
3ジュース	-0.054	0.101	飲み物
5日本茶	-0.454	-0.165	飲み物
6ウーロン茶	0.175	0.015	飲み物
8スポーツ飲料	-0.014	0.208	飲み物
7ミルク	-0.191	0.220	飲み物
dチョコレート	-0.050	0.129	食べ物
aホテチップス	0.045	0.021	食べ物
iアール	-0.346	0.032	食べ物
gパン	-0.299	0.146	食べ物
fスケート	0.095	-0.205	食べ物
eケーキ	-0.642	-0.118	食べ物
cクッキー	0.204	0.070	食べ物
h和菓子	-0.184	-0.277	食べ物
bホッケー	0.139	0.044	食べ物

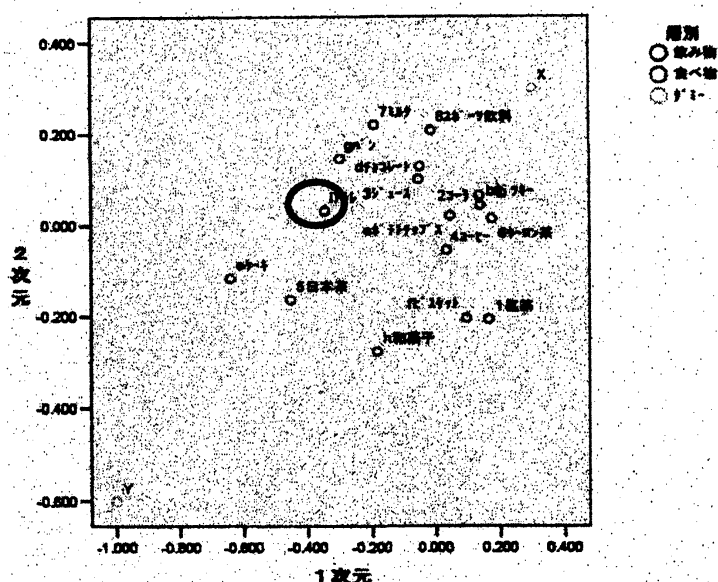


図4 我々が提言する correspondence analysis の距離尺度での同時布置

表8は、我々が提言した距離尺度の式(10)、(11)で求めた correspondence analysis の各品目

の2次元までの score を求めたものである。図4は、表8の同時布置を図示したものである。
 ○73の位置を見て判るように、73は反応数が少なかったで、元の方法では外側にあったが、その反応数により同じ weight で調整した我々の提言式を用いると、図3に比べて中の方へ移動している。そして、同時購入された日本茶とジュースと近くなり、その距離の balance が取れていることが判る。反応数が多いお茶とジュースとの関係も図3に比べてより接近している。

今回の事例では、食べ物と飲み物の項目間の関係が有意でなかったが、 χ^2 検定で従属関係がいえる事例においては、我々が提言した距離尺度により、関係のある品目はより近く、ないものはより遠くになると考えている。今後は、 χ^2 検定で有意となる多品目間において我々の距離尺度の有効性を検証したい。

(2) correspondence analysis の特性について

以上より、correspondence analysis の特性をまとめると、

- ①固有値の出方は、数量化の方法よりも差がついて出る。今回の data 例では χ^2 検定は有意ではなかったで、飲み物と snack 菓子の関係は独立となった。
- ②しかし χ^2 検定で有意とならなくても、数量化法よりも全体的な項目間の傾向が出易い。 χ^2 統計量に対応する固有値が導かれるので、行と列の関係が統計的に有意かが明確にできる。そして、独立であっても項目間の関係は同時布置に再現され易いので有用といえる。
- ③また、我々が提言した距離尺度にて同時布置を導けば、元の correspondence analysis の距離に比べて、反応数に応じた重みが均等化されるので、行も列も同じ距離にて比較することができる。より品目間の位置関係を視覚的に捉えるのに有効となる。
- ④平均的な傾向と特異な場合の関係も同時に表れるので、購入 data から何らかの情報を得るのには向いていると考えられる。
- ⑤逆に、 χ^2 統計量が独立でも、あたかも関係があるかのように結果を見てしまう危険もある。

以上は、消費者購買 data から全体的な傾向を抽出する解析法であった。次に部分的な傾向を抽出方法として、今話題の market basket analysis を実施する。

4. 4 market basket analysis の結果について

(1) market basket analysis について

market basket analysis は特定の顧客が買う品物間の関連を測ることを言う。market basket という呼び名は、顧客が食料品店で shopping cart の「market basket」に一緒に入れた商品群はどのようなものになるかを探ることから生まれた data mining の一つの技法である。

(2) 相関 rule について

market basket analysis[3]では、全 data を取り扱う model の概念はなく、data set の部分集合、例えば、変数の部分集合、観測値の部分集合に注目していくことから始まる。

まず pattern から、商品(項目)間の関連を捉える相関 rule について説明する。

pattern[4]とは、database 項目に関する2値の data 行列において、行として transaction、列として項目を置き、pattern が

$$\alpha = (\text{Age} < 30 \wedge \text{Income} > 100)$$

$\wedge$: AND

$$\beta = (\text{Gender} = \text{male} \vee \text{Education} = \text{High})$$

$\vee$: OR

なら、相関 rule として、 $\alpha \rightarrow \beta$ なら、 α と β が一緒に発生すると定義する。即ち、 α が起れば、

β も起ると考える. "if α occurs, then β also occurs, if condition, then result." である. ここで, pattern を定式化する.

(a) 2 値変数 $A_1, A_2, \dots, A_p$ としたとき, 項目 A_i が発生 $\Leftrightarrow A_i = 1$ とおくと, pattern の定義は,

$$\text{pattern } A = (A_{j1} = 1 \wedge \dots \wedge A_{jk} = 1) \quad (12)$$

$$\text{相関 rule } A \rightarrow B \text{ は, } (A_{j1} = 1 \wedge \dots \wedge A_{jk} = 1) \rightarrow A_{jk+1} \quad (13)$$

となる. この時の $k+1$ を order (位数、順位) と呼ぶ.

例: 位数 3 の相関 rule (Milk $\wedge$ Tea) $\rightarrow$ Biscuits

(b) 評価尺度としては, (14) 式の支持度と (15) 式の信頼度を定める.

$$\text{支持度 } \text{support}\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N} \quad (14)$$

$N_{\cdot}$: pattern $\cdot$ を持つ transaction の個数

N : transaction の全個数

$$\text{信頼度 } \text{confidence}\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N_{A \rightarrow A}} = \frac{\text{support}\{A \rightarrow B\}}{\text{support}\{A\}} \quad (15)$$

$$\text{Lift}\{A \rightarrow B\} = \frac{\text{confidence}\{A \rightarrow B\}}{\text{support}\{B\}} = \frac{\text{support}\{A \rightarrow B\}}{\text{support}\{A\} \text{support}\{B\}} \quad (16)$$

(c) software としての Apriori の計算原理は,

- ・ 頻出 item 集合 (最小支持度を越える item 集合) を見つける.
- ・ 頻出 item 集合を用いて相関 rule を求める.
- ・ 計算原理は, 頻出 item 集合 (最小支持度を越える item 集合) の任意の部分集合は再び頻出 item 集合である.

(3) Apriori アルゴリズムについて

(a) ここで, C_k : 大きさ k の候補 item 集合, L_k : 大きさ k の頻出 item 集合, とする.

Joint Step: L_{k-1} を自分自身と join して, C_k を生成する.

Prune Step: 頻出でない任意の $(k-1)$ -item 集合は k -item 集合の部分集合にはなれない.

(b) 相関 rule を求める.

Step 1: 各頻出 item 集合 m に対して, それをすべての方法で 2 つの空でない部分集合 $s, m-s$ に分割する.

Step 2: rule 候補 $s \Rightarrow (m-s)$ に対する信頼度を求める. もしこの値が最小信頼度よりも大きいと等しければ, それを相関 rule とする.

(c) 上記の (14), (15) 式から, 支持度及び信頼度を計算する.

(d) 相関 rule において, 支持度の高い pattern を発見する.

(4) 表 1 データの Market basket analysis の結果について

第 1 段階: 表 1 の購入品を記号化して, 表 9 のようにする.

表9 表1のdataの購入品を記号化して品目の多い順に並べたdata

	a	b	1	2	3c	d	4e	f	5	6g	7h	8i					
	ホテトチップ	ポッキー	紅茶	コーラ	ジュース	クッキー	チョコレート	コーヒー	ケーキ	ビスケット	日本茶	ウーロン茶	パン	ミルク	和菓子	スポーツ飲料	アラレ
藤田	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
伊藤	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
赤木	1.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
谷村	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0
明尾	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0
松崎	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
森下	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0
水田	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0
新井	1.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
横本	1.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
大塚	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
国宝	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
岩本	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
岡本	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
村中	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0
大滝	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
内田	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
西口	1.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
森	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
小澤	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
	a	b	1	2	3c	d	4e	f	5	6g	7h	8i					
濱田	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
野田	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
近藤	1.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
八木	1.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
鈴木	1.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
中塚	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
井村	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
佐藤	1.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
長畑	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
岡内	0.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
有吉	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
柳田	1.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
中野	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
浅田	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
穴吹	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
天野	1.0	1.0	1.0	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
平田	1.0	1.0	1.0	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
石川	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
江尻	1.0	1.0	0.0	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
滝下	1.0	1.0	0.0	0.0	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
竹元	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
長濱	1.0	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

第2段階：表10より各購入品の単独の支持度を求める。

表10 各購入品別の全体に対する支持度

TID	TID	記号	Itemset	C ₁ , L ₁	支持度
ab1234ef	a3c4e	a	ホテトチップス	39	0.929
ab123c4	a12cd	b	ポッキー	24	0.571
ab23c47	ab24g	1	紅茶	20	0.476
a1cde67	a3cd8	2	コーラ	18	0.429
a3d5g7i	a123e	3	ジュース	18	0.429
ab123cd	ab3cf	c	クッキー	18	0.429
a13de5h	ab12c	d	チョコレート	17	0.405
ab125h	ab2eg	4	コーヒー	14	0.333
a12cdh	b1cdf	e	ケーキ	13	0.310
ab3cd6	a1d4f	f	ビスケット	8	0.190
ab1ef6	a23cd	5	日本茶	7	0.167
ab134e	acd46	6	ウーロン茶	7	0.167
a3de5i	3d4g7	g	パン	6	0.143
ab2c46	a13e	7	ミルク	5	0.119
4ef5gh	ab12	h	和菓子	4	0.095
ab1of6	ab12	8	スポーツ飲料	3	0.071
ab2d4f	a1ce	i	アラレ	2	0.048
ab12cd	ab2			42	
ab5g8	ab3				
ab4e7	a4				
ab3d8	a6				

表10より、学生の談話会では、aホテトチップスの支持度は0.929であり、aホテトチップスは殆ど購入

されることが解る.

第2段階： 相関 rule, 一つの品目 $A \rightarrow B$ の支持度を求める.

表10 相関 rule 支持度

記号	内容	N	Itemset	C_2, L_2	支持度	Itemset	C_2, L_2	支持度
a	ポテトチップス	39	ab	23	0.590	1d	8	0.400
b	クッキー	24	a1	19	0.487	14	3	0.150
1	紅茶	20	a2	19	0.487	1e	8	0.400
2	コーラ	18	a3	17	0.436	23	6	0.333
3	ジュース	18	ac	16	0.410	2c	9	0.500
c	クッキー	18	ad	13	0.333	2d	6	0.333
d	チョコレート	17	a4	12	0.308	24	5	0.278
4	コーヒー	14	ae	12	0.308	2e	4	0.222
e	ケーキ	13	b1	11	0.458	3c	9	0.500
			b2	14	0.583	3d	9	0.500
			b3	9	0.375	34	6	0.333
			bc	9	0.375	3e	6	0.333
			bd	6	0.250	cd	9	0.500
			b4	8	0.333	c4	5	0.278
			be	4	0.167	ce	3	0.167
			12	11	0.550	d4	3	0.176
			13	7	0.350	de	3	0.176
			1c	10	0.500	4e	3	0.214

aポテトチップスが購入されると, bクッキーの支持度0.590, 同様に, 1紅茶0.487, 2コーラ0.487, 3ジュース0.436, cクッキー0.410である. 他に, bクッキーが購入されると, 2コーラ0.583, 1紅茶0.458の支持度となる. これらの pattern は, 表4の cross 集計の結果と同じようになっている.

第3段階: 相関 rule, 二つの品目 A, B が購入されると C が購入されるという支持度を求める.

表11 相関 rule $A, B \rightarrow C$ の支持度

記号	内容	N	Itemset	C_2, L_2	Itemset	C_2, L_2
a	ポテトチップス	39	ab	23	1d	8
b	クッキー	24	a1	19	14	3
1	紅茶	20	a2	19	1e	8
2	コーラ	18	a3	17	23	6
3	ジュース	18	ac	16	2c	9
c	クッキー	18	ad	13	2d	6
d	チョコレート	17	a4	12	24	5
e	ケーキ	13	ae	12	2e	4

支持率の MAX 導出

Itemset	C_2, L_2	Itemset	C_2, L_2
ab	23	1d	8
a1	19	14	3
a2	19	1e	8
a3	17	23	6
ac	16	2c	9
ad	13	2d	6
a4	12	24	5
ae	12	2e	4
b1	11	3c	9
b2	14	3d	9
b3	9	34	6
bc	9	3e	6
bd	6	cd	9
b4	8	c4	5
be	4	ce	3
12	11	d4	3
13	7	de	3
1c	10	4e	3

Itemset	C_2, L_2	前提	支持度	Itemset	C_2, L_2	前提	支持度
ab1	11	b1 $\Rightarrow$ a	1.000	b2d	1	—	—
ab2	14	b2 $\Rightarrow$ a	1.000	b2e	1	—	—
ab3	9	b3 $\Rightarrow$ a	1.000	b3c	5	bc $\Rightarrow$ 3	0.556
abc	9	bc $\Rightarrow$ a	1.000	b3d	3	bd $\Rightarrow$ 3	0.500
abd	5	bd $\Rightarrow$ a	0.833	b3e	1	—	—
abe	3	be $\Rightarrow$ a	0.750	bcd	3	bd $\Rightarrow$ e	0.500
a12	11	12 $\Rightarrow$ a	1.000	bce	0	—	—
a13	7	13 $\Rightarrow$ a	1.000	bde	0	—	—
a1c	8	1c $\Rightarrow$ a	0.727	123	4	23 $\Rightarrow$ 1	0.667
a1d	7	1d $\Rightarrow$ a	0.875	12c	6	2c $\Rightarrow$ 1	0.667
a1e	7	1e $\Rightarrow$ a	0.875	12d	4	2d $\Rightarrow$ 1	0.667
a23	6	23 $\Rightarrow$ a	1.000	12e	2	2e $\Rightarrow$ 1	0.500
a2c	9	2c $\Rightarrow$ a	1.000	13c	1	—	—
a2d	4	2d $\Rightarrow$ a	0.667	13d	2	13 $\Rightarrow$ d	0.286
a2e	3	2e $\Rightarrow$ a	0.750	13e	5	3e $\Rightarrow$ 1	0.833
a3c	7	3c $\Rightarrow$ a	0.778	1cd	6	cd $\Rightarrow$ 1	0.667
a3d	8	3d $\Rightarrow$ a	0.889	1ce	2	ce $\Rightarrow$ 1	0.667
a3e	6	3e $\Rightarrow$ a	1.000	1de	2	de $\Rightarrow$ 1	0.667
acd	8	cd $\Rightarrow$ a	0.889	23c	4	23 $\Rightarrow$ c	0.667
ace	3	ce $\Rightarrow$ a	1.000	23d	2	2d $\Rightarrow$ 3	0.333
ade	2	de $\Rightarrow$ a	0.667	23e	2	2e $\Rightarrow$ 3	0.500
b12	8	b1 $\Rightarrow$ 2	0.727	2cd	5	2d $\Rightarrow$ c	0.833
b13	4	13 $\Rightarrow$ b	0.571	2ce	0	—	—
b1c	6	bc $\Rightarrow$ 1	0.667	2de	0	—	—
b1d	3	bd $\Rightarrow$ 1	0.500	3cd	4	cd $\Rightarrow$ 3	0.444
b1e	2	be $\Rightarrow$ 1	0.500	3ce	1	—	—
b23	3	23 $\Rightarrow$ b	0.500	3de	2	de $\Rightarrow$ 3	0.667
b2c	5	2c $\Rightarrow$ b	0.556	cde	1	—	—

表 1 1 より, b ホッキーと 1 紅茶, b ホッキーと 2 コーラ, b ホッキーと 3 ジュース, b ホッキーと c クッキーが購入された時は, a ホテトチップスの支持度が 1.000 で必ず購入される。また, 1 紅茶と 2 コーラ, 1 紅茶と 3 ジュース, 2 コーラと 3 ジュース, 2 コーラと c クッキー, 3 ジュースと e ケーキ, c クッキーと e ケーキが購入された時も, a ホテトチップスの支持度が 1.000 で必ず購入されるという pattern が見つけれられる。これらの pattern は, 表 4 からでは得られない。

以下同様にして, market basket analysis では, 組合せを多くして, 新たな pattern を探索することができる。

別の段階: 例数は少ないが, 特異な購入品である h 和菓子に着目して pattern を考える。

表 1 2 和菓子に着目した pattern の表

記号	Item	N	TID	TID	Itemset	前提	支持度
a	ホテトチップス	39	ab1234ef	a3c4e	5h	3 h⇒5	0.750
b	ホッキー	24	ab123c4	a12cd	ah	3 h⇒a	0.750
1	紅茶	20	ab23c47	ab24g	1h	3 h⇒1	0.750
2	コーラ	18	a1cde67	a3cd8	ah	3 1h⇒a	1.000
3	ジュース	18	a3d5g7i	a123e			
c	クッキー	18	ab123cd	ab3cf			
d	チョコレート	17	a13de5h	ab12c			
4	コーヒー	14	ab125h	ab2eg			
e	ケーキ	13	a12cdh	b1cdf			
f	ビスケット	8	ab3cd6	a1d4f			
5	日本茶	7	ab1ef8	a23cd			
6	ウーロン茶	7	ab134e	acd46			
g	ハン	6	a3de5i	3d4g7			
7	ミルク	5	ab2c46	a13e			
h	和菓子	4	4ef5gh	ab12			
8	スポーツ飲料	3	ab1cf8	ab12			
i	アヲレ	2	ab2d4f	a1ce			
			ab12cd	ab2			
			ab5g8	ab3			
			ab4e7	a4			
			ab3d8	a6			

1 紅茶と h 和菓子が購入されると a ホテトチップスの支持度は 1.000 となり必ず同時に購入される。また, h 和菓子が購入されると, 5 日本茶, a ホテトチップス, 1 紅茶の支持度は 0.750 の pattern となる。表 4 の結果に, 加えて他の品目が加わり新たな情報を得ることができる。

このように market basket analysis を用いれば, 組合せを幾つも考えて, 支持度の高い pattern を探索していくことができる。そして, ある品目間の相関 rule には, その関連性が理解できる場合と, 理解できない場合とが出てくる。data mining では, 後者の発見が大切なわけであるが, 偽相関のような場合も考えられ, 更に深く考察することを忘れてはならない。

(5) market basket analysis の特性について

以上より, market basket analysis の特性をまとめると,

- ①二つ以上の同時購買に対する他の品目の購買関連を探るのに適している。
- ②部分最適解をいろんな切口から求められ, 新しい組合せによる pattern の発見に役立つ。
- ③そして, 既成の関係や概念を捨てなければならないような結果が多く出てくる。従って, 結果については仮説実験を行い, 確認する必要がある。
- ④部分の最適解であるので, 既成の概念を逸脱することが多い。従って, 短期的な傾向であるとも考えられるので, 常に傾向の変化が生じることに注意を払う必要がある。

4. まとめ

消費者の購買 data から、どのような情報が得られるか、各手法の特性を活かして行う解析のあり方について検討した。その結果、

- ①特異な購買 pattern を抽出するのには数量化3類が有効である。
- ②行列間の関連(snack 菓子と飲み物、商品とその購入者の特徴等)を mapping して考察するのには correspondence analysis がよい。しかし、correspondence Analysis は、結局、行と列間の内積の情報を取っていることであり、 χ^2 統計量の考え方と同じである。従って、行と列間の関係が有意となる case に適用するのが好ましい。よく correspondence analysis の適用例として紹介されている Greenace(1984)の喫煙習慣の例[5]は χ^2 統計量が有意とならなく不適切な例である。適切な適用例[6]は他にもあるので、今後はそのような例を紹介すべきである。また、行と列の項目で、項目に反応する例数が極端に少ない場合などには、今回提言した我々の行と列とを同じ重みで調整した距離尺度を用いるのがよい。提言した距離尺度を用いれば、行と列とを同じ距離尺度で同時に比較できる。
- ③品目の組合せが多い場合については、その相関有無の pattern 情報を抽出するのには、market basket analysis が適する。しかし、その関係の因果については、仮説を置いて確認を行うことも忘れてはならない。安易に結果を利用すると短期的な現象にしかすぎない場合もあるので注意が必要である。market basket analysis は部分解であり、その部分解から関連する次の解を得る姿勢が大切である。

今後は、行と列間の関連がある事例において、我々が提言している correspondence analysis の距離尺度が有効性であることを確認したい。また、ある商品を購入する顧客層を特定化していく decision tree[7]による分類手法も加えて、market basket analysis と decision tree との組合せによる行と列間の関連付け法と correspondence analysis による行と列間の捉え方との違いについての研究を進めていきたい。

参考文献

- [1]Isogai.T and Noguchi.H., “ Singular value decomposition and the application”, Journal of Social Science Research, Osaka University, No.40, (1992),pp.63-101.
- [2]Benzecri,J.P., “Statistical analysis as a tool to make patterns emerge from data”, In Methodologies of Pattern Recognition (Watanabe.S, ed, New York: Academic (1969)), pp.35-60.
- [3]Pralo Giudici., “Applied Data Mining”, Principles of Data Mining MIT Press, Cambridge MA, Wiley, (2003).
- [4]Hand, Mannila and Smyth, ”Principles of Data Mining, MIT Press, Cambridge MA, (2001).
- [5]Greenace,M.J., ”Theory and Application of correspondence analysis”, London, Academic,(1984).
- [6]Hair, Anderson, Tatham, Black., “ Multivariate Data Analysis”, Fifth Edition, Prentice Hall, (1998).
- [7]Rissanen.J., “Stochastic complexity in learning”, Jr. Computer System and Sciences,55(1), (1997), pp.89-95.