Vol.2011-MPS-84 No.10
2011/7/18

確率分布間における類似関係の階層的可視化を用いたプロファイリング

伊藤　晃[†1]　吉川大弘[†1]　古橋　武[†1]

アンケートは，企業が市場の動向を調査するための重要な手段の一つであり，様々な解析がなされている．アンケートデータに対する最も重要な解析の一つに，ターゲット層についてのプロファイリングが挙げられる．プロファイリングに用いられる代表的な解析手法には，コレスポンデンス分析，アソシエーション分析などがあり，これらの手法は，データ中の要素間の関係を把握する上で極めて有用である．しかし，得られる結果はあくまでデータ内の偏り等を捉えたものであり，アンケートデータのような比較的小規模のデータに対する適用では，しばしば母集団の傾向とは異なる結果となり得る．そこで本稿では，確率モデルを用いることにより母集団の傾向を考慮可能とした，新たなプロファイリング手法を提案する．提案手法では，ある属性に当てはまる回答者から，任意の質問に対する各回答の得られる確率を確率変数として扱い，その確率分布をベイズ的アプローチにより得る．また，すべての属性，もしくはそれらの組み合わせについての確率分布とそれらの間の類似関係を算出し，その類似関係に基づいて各属性を可視空間上にプロットする．本稿では，スキャナ製品に関するアンケート調査データに提案手法を適用し，解析者が，様々な属性と特定の質問の関係において，ルールの信頼性を考慮に入れたプロファイリングが可能となることを示す．また，提案手法による可視化により，属性間における回答傾向の類似関係を確率分布の観点から把握できること，および特徴的な属性の抽出を支援することが可能となることを示す．

Hierarchical Visualization of Similarities between Probabilistic Distributions for Profiling

Akira Ito,[†1] Tomohiro Yoshikawa[†1]
and Takeshi Furuhashi[†1]

One of the most important purposes in the analysis of questionnaire data is to get profiles for the target group(s). As the result of profiling gives a great influence on the planning marketing strategy, the reliability of profiling is very important. Correspondence Analysis, Association Rule mining are typical methods for profiling and are useful to grasp the relationships between categorical variables in data. However, the results of these methods may "overfit" to the data and be different from the behavior in the general population, especially when the sample size is small. This paper proposes a new profiling method considering behavior in the general population by probabilities. It derives the probabilistic distributions of each choice on a question by Bayesian approach for each attribute, and visualizes the similarities between these distributions. This paper applies the proposed method to an actual questionnaire data on a scanner product. It shows that a user can profile the data considering the uncertainty of extracted rules for the relationship between each attribute and the answer to the essential question. It also shows that the visualization supports a user to grasp the similarities between the probabilistic distributions for each attribute and to extract the characteristics of attributes.

1. はじめに

アンケートは，企業が市場の動向を調査するための重要な手段の一つであり，アンケートデータを解析し，得られた知見に基いてマーケティング戦略の立案を行うことがしばしば行われている[1)2)]．例えば，新製品の販売を企画する際に，その製品に関する印象調査を行い，得られたデータを解析して購買層に関する知見を得ることで，購買規模の予測，コマーシャル戦略の立案，製品デザインへのフィードバックなどが行われる．

アンケートデータに対する解析には，様々な目的が存在するが，最も重要な解析の一つに，例えば製品の購買層における属性情報の抽出など，ターゲット層についてのプロファイリングが挙げられる．プロファイリングは，購買規模予測や製品デザインへのフィードバックなど，後のプロセスに大きな影響を与えるため，その解析手法や得られる結果には信頼性の高いものが求められる．

プロファイリングには様々な解析手法が用いられるが，それらの代表的な手法として，コレスポンデンス分析[3)]，アソシエーション分析[4)] などが挙げられる．コレスポンデンス分析は，クロス集計表の行の要素と列の要素の相関関係が最大になるように数量化し，その行の要素間や列の要素間の関係，行と列の要素間の関係を可視空間で表現する手法である．また，アソシエーション分析は，データに内在する要素間の相関関係を，ルール形式 {A ⇒ B} で抽出する手法である．ルールには複数の評価指標が存在し，それらの中でユーザが求める評

†1 名古屋大学
Nagoya University

Vol.2011-MPS-84 No.10
2011/7/18

価指標に基づいてルール抽出が行われる．これらの手法は，データ中の要素間の関係を把握する上で極めて有用であり，プロファイリングへの応用も多く見られる[5)6)]．しかし，得られる結果はあくまでデータ内の偏り等を捉えたものであり，アンケートデータのような比較的小規模のデータに対する適用では，しばしば母集団の傾向とは異なる結果となり得る[7)8)]．

そこで本稿では，確率モデルを用いることにより母集団の傾向を考慮可能とした，新たなプロファイリング手法を提案する．提案手法では，ある属性に当てはまる回答者から，任意の質問に対する各回答の得られる確率を確率変数として扱い，その確率分布をベイズ的アプローチにより得る．また，すべての属性，もしくはそれらの組み合わせについての確率分布とそれらの間の類似関係を算出し，その類似関係に基づいて各属性を可視空間上にプロットする．

本稿では，スキャナ製品に関するアンケート調査データに提案手法を適用し，解析者が，様々な属性と特定の質問の関係において，ルールの信頼性を考慮に入れたプロファイリングが可能となることを示す．また，提案手法による可視化により，属性間における回答傾向の類似関係を確率分布の観点から把握できること，および特徴的な属性の抽出を支援することが可能となることを示す．

2. 提 案 手 法

今，ある属性 X に当てはまる回答者から，任意の質問（例えば，解析者が注目した質問やアンケートの調査目的となる質問など）に対して，k 通りの異なる回答のいずれか 1 つが得られることを考え，回答 1 を得る確率を θ_1，回答 2 を得る確率を θ_2，...，回答 k を得る確率を θ_k とする (図 1)．$\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_k\}$ は確率変数として扱い，その確率分布をベイズ的アプローチにより求める．また，確率分布における期待値，Highest Posterior Density（HPD）区間[9)] を，属性 X の評価指標として扱う．提案手法では，すべての属性についての確率分布とそれらの間の類似関係を求め，その類似関係に基づいて各属性を可視空間上にプロットする．

2.1 確率分布の推定

2.1.1 確 率 分 布

属性 X に当てはまる回答者群において，回答 1〜k の各回答が得られる確率を $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_k\}$ としたとき，アンケートデータ上での各回答を選択した人数 $\boldsymbol{x} = \{x_1, x_2, ..., x_k\}$ が得られる確率は

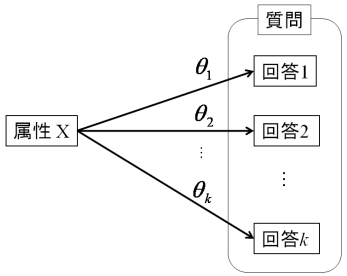


図 1 属性の各回答への選択確率

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{(\Sigma_i x_i)!}{\prod_i x_i!} \prod_{i=1}^{k} \theta_i^{x_i} \tag{1}$$

となり，これは尤度と呼ばれる．今，求めたいものは，$\boldsymbol{x}$ を得たときの $\boldsymbol{\theta}$ の確率分布 $p(\boldsymbol{\theta}|\boldsymbol{x})$ であり，これはベイズの定理を用いて以下のように展開できる．

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{2}$$

ここで，$p(\boldsymbol{\theta})$ は $\boldsymbol{x}$ を得る前の $\boldsymbol{\theta}$ の確率分布であり，事前分布と呼ばれる．また，これに対し，$p(\boldsymbol{\theta}|\boldsymbol{x})$ を事後分布と呼ぶ．本手法では，事前分布として以下のような共役事前分布を用いる．

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\Sigma_i \alpha_i)}{\prod_i \Gamma \alpha_i} \prod_{i=1}^{k} \theta_i^{\alpha_i} \tag{3}$$

ここで，$\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_k\}$ であり，ハイパーパラメータと呼ばれる．式 (1)，(2)，(3) から，事後分布は以下のように求められる．

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{\Gamma(\Sigma_i (x_i + \alpha_i))}{\prod_i \Gamma(x_i + \alpha_i)} \prod_{i=1}^{k} \theta_i^{x_i + \alpha_i - 1} \\ &= \frac{1}{B(\boldsymbol{x} + \boldsymbol{\alpha})} \prod_{i=1}^{k} \theta_i^{x_i + \alpha_i - 1} \end{aligned} \tag{4}$$

この事後分布の分布形は Dirichlet 分布と呼ばれる[10)]．

2.1.2 評価指標

期待値

θ_i に対する期待値は以下の式で与えられる．

$$E[\theta_i] = \int p(\boldsymbol{\theta}|\boldsymbol{x})\theta_i d\boldsymbol{\theta} \tag{5}$$

これは，Dirichlet 分布において解析的に求められることが知られており[10]，式 (4) の場合，

$$E[\theta_i] = \frac{x_i + \alpha_i}{\Sigma_i(x_i + \alpha_i)} \tag{6}$$

となる．

HPD 区間

確率分布における $100(1-\epsilon)\%$ HPD 区間は次の二つの性質を持つ．

(1) 区間内のあらゆる点における確率密度が区間外のあらゆる点における確率密度より大きい．

(2) 確率が全体の $100(1-\epsilon)\%$ となる区間の中で最も小さい区間である．

HPD 区間は確率変数の数 k の次元をもっており，数値計算により求めることが可能である[9]．

図 2 は，ある確率分布における評価指標を表している．属性 X に当てはまる回答者数が多くなるほど，期待値は最頻値（単純な条件付き確率）に近づき，HPD 区間は小さくなる．

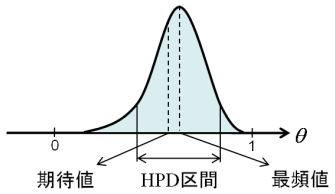


図 2 評価指標

2.2 確率分布間における類似関係の可視化

2.2.1 確率分布間の類似関係

本手法では，確率分布間の類似度として Bhattacharyya 距離[11] を用いる．二つの確率分布 p_a，p_b 間の Bhattacharyya 距離 $J_B(p_a, p_b)$ は以下の式で与えられる．

$$J_B(p_a, p_b) = -\ln \int_{\boldsymbol{\theta}} \sqrt{p_a p_b} d\boldsymbol{\theta} \tag{7}$$

p_a，p_b が以下のような Dirichlet 分布の場合を考える．

$$p_a = \frac{1}{B(\boldsymbol{\lambda}_a)} \prod_i^k \theta_i^{\lambda_{ai}-1} \qquad \boldsymbol{\lambda}_a = \{\lambda_{a1}, \lambda_{a2}, ...\lambda_{ak}\}$$
$$p_b = \frac{1}{B(\boldsymbol{\lambda}_b)} \prod_i^k \theta_i^{\lambda_{bi}-1} \qquad \boldsymbol{\lambda}_b = \{\lambda_{b1}, \lambda_{b2}, ...\lambda_{bk}\}$$

このとき，$J_B(p_a, p_b)$ は，以下のように解析的に求めることが可能である[11]．

$$J_B(p_a, p_b) = \ln \frac{\sqrt{B(\boldsymbol{\lambda}_a)B(\boldsymbol{\lambda}_b)}}{B(\frac{1}{2}\boldsymbol{\lambda}_a + \frac{1}{2}\boldsymbol{\lambda}_b)} \tag{8}$$

2.2.2 ばねモデルを用いた多次元尺度構成法

ばねモデルを用いた多次元尺度構成法[12] により，元空間上での要素間の類似関係（距離）を可視空間上で表すことが可能である．要素 i，j 間の，元空間における距離と可視空間における距離をそれぞれ d_{ij}，d^*_{ij} としたとき，可視空間上における各要素の座標は，以下のエネルギー関数 E を最小化することにより求められる．

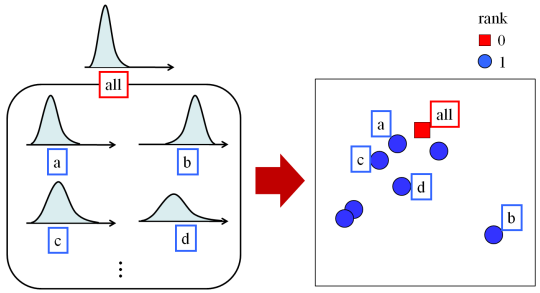
$$E = \Sigma_{i=1}^{n-1}\Sigma_{j=i+1}^{n} k_{ij}(d_{ij} - d^*_{ij})^2 \tag{9}$$

k_{ij} はばね係数であり，k_{ij} が大きくなるほど，相対的に，d_{ij} と d^*_{ij} の誤差が小さくなる．

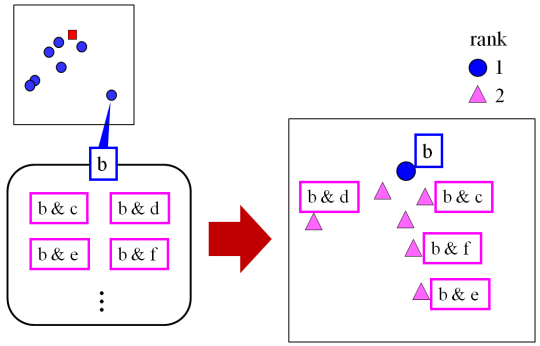
2.2.3 階層的可視化

各属性は，確率分布間の類似関係に基づき，ばねモデルを用いた多次元尺度構成法により可視空間上にプロットされる．図 3 は可視化のイメージを表している．rank は属性における条件の数を表しており，例えば { 男性 }, {20 代 } といった属性は rank1，{ 男性 & 30 代 }, { 女性 & パートタイマー } などは rank2 となる．また，rank0 は条件なし，すなわち全体を表している．

はじめに，rank0 と rank1 の可視化を行う（図 3(a)）．これらの属性は同一の可視空間上にプロットされるが，このとき，異なる rank の属性間に対して，元空間における距離が保持されるよう，これらの要素間のばね係数を十分に大きな値とする．ここから解析者は，注目した属性に関して掘り下げて解析を行うことが可能である．解析者が選択した属性を細分化し，rank 0-1 と同様に rank 1-2 の可視化を行う（図 3(b)）．以下の階層についても同様である．本可視化により，解析者が属性間における回答傾向の類似関係を確率分布の観点から把握でき，特徴的な属性の抽出を支援することが可能となる．



(a) rank 0-1



(b) rank 1-2

図 3 可視化のイメージ

3. 実験と考察

3.1 実 験 条 件

本実験では，スキャナ製品に関するアンケートデータを解析対象とする．用いる質問は，“スキャナ使用経験” (あなたは最近 1 年以内に，スキャナを使用しましたか．) と属性（年齢，性別，職業，世帯年収）に関するものである．回答者数は 1564 名であった．また，本実験では，事前分布において $\alpha_i = 1$（一様分布），HPD 区間を 95% と設定した．

3.2 アソシエーション分析との比較

本節では，提案手法における評価指標である期待値，HPD 区間と，アソシエーション分析[4)6)] における評価指標の一つである $confidence$ を，推定誤差の観点から比較する．

アソシエーション分析は，データに内在する要素間の相関関係を，ルール形式 {A ⇒ B} で抽出する手法であり，その抽出はルールの評価指標に基づいている．評価指標には，$support$；$p(A, B)$，$confidence$；$p(B|A)$ が多く用いられる．

実験は以下の手順で行った．

(1) アンケートデータの全サンプル D（1564 名）からランダムに 100 サンプル D_s を抽出する．

(2) 結論部を { スキャナ使用経験 = なし } とし，$support \geq a$，$rank \leq 2$ という条件下で，D_s からルール群 R を抽出する．R_{10} を R における $confidence$ 上位 10 個のルールとする．

(3) R_{10} について，D_s における $confidence$ の値と，D における $confidence$ の値との差をそれぞれ求め，その平均を $E1$ とする．

(4) R_{10} について，D_s における期待値の値と，D における $confidence$ の値との差をそれぞれ求め，その平均を $E2$ とする．

(5) R_{10} について，HPD 区間から，D における $confidence$ の値が外れる割合を $E3$ とする．

ここでは，$E1$ を $confidence$ 誤差，$E2$ を期待値誤差，$E3$ を HPD 区間誤差とする．10 試行行ったときの，$E1$，$E2$，$E3$ の平均値の比較結果を図 4 に示す．図 4 の横軸は $support$ の閾値 a を表している．図 4 から，a を低く設定するほど $confidence$ 誤差が大きくなっているのに対し，期待値誤差は，a が低い場合でも比較的小さいことがわかる．また，HPD 区間誤差は 5% 以下となっており，設定した値（5%）の範囲に収まっている．なお，本実験では rank を 2 以下と設定したが，この閾値をより大きな値に設定すれば，わずかな人数にしか当てはまらないルールが上位に増え，$confidence$ 誤差はさらに大きくなると考えられる．

以上より，期待値を評価指標とすることで，$confidence$ と比較して精度の高い推定を行うことができ，また，HPD 区間によりルールの信頼性を考慮することが可能となると考えられる．これらのことから，提案手法における評価指標は，アソシエーション分析における

評価指標と比較してプロファイリングに適しており，特にサンプル数が少ない場合に有効だと考えられる．

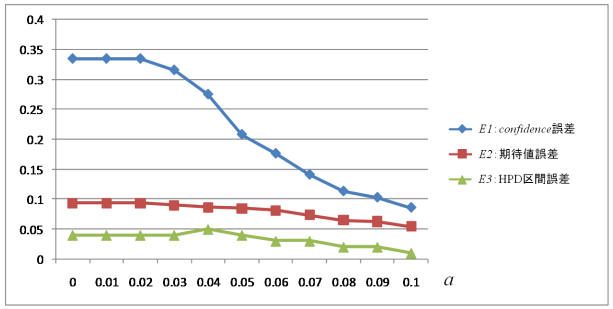


図 4 推定誤差

3.3 プロファイリング

提案手法による可視化結果を図 5 に示す．また，図 5 に現れているいくつかの属性の評価指標を表 1 に示す．表 1 における **A**-**F** は，図 5 の記号と対応している．

rank 0-1 に対する可視化結果を図 5(a) に示す．図において，**A**, **B**, **C** が rank0 の点から見て遠くに布置されていることがわかる．図 5(a) と表 1 から, 全体での傾向と比較して, { 職業 = 主婦 }, { 職業 = パートタイマー } はスキャナ使用経験が少ない属性であり，逆に，{ 世帯年収 = 2000 万円以上 }, { 職業 = 従業者 (管理職)} はスキャナ使用経験が多い属性であることがわかる．また，これらの属性は全体の場合と HPD 区間が重複しておらず，このことからも，スキャナ使用経験において全体とは大きく異なった傾向を持つ属性であるといえる．

ここから，{ 職業 = 従業者 (管理職)}, { 世帯年収 = 2000 万円以上 } に関して掘り下げて解析を行う．図 5(b) は { 職業 = 従業者 (管理職)} に関する rank 1-2 の結果を表している．上位属性 { 職業 = 従業者 (管理職)} に比べ，{ 職業 = 従業者 (管理職) & 世帯年収 = 800-1000 万円 } における { スキャナ使用経験 = あり } の期待値は低く，{ 職業 = 従業者 (管理職) & 世帯年収 = 1800-2000 万円 } の期待値は高くなっていることがわかる．このことから，管理職の従業者の中でも世帯年収の多寡がスキャナの使用経験に関係しているように思えるが，{ 職業 = 従業者 (管理職) & 世帯年収 = 1800-2000 万円 } における HPD 区間は大きく，その下限値は 0.23 となっており，それほど信頼性は高くない情報であるといえる．

表 1 評価指標

ID	属性	スキャナ使用経験 = なし	スキャナ使用経験 = あり
	全体	0.79 (0.77-0.81)	0.21 (0.19-0.23)
A	職業 = 主婦	0.87 (0.83-0.91)	0.13 (0.09-0.17)
	職業 = パートタイマー	0.90 (0.83-0.95)	0.10 (0.05-0.17)
B	世帯年収 = 2000 万円以上	0.57 (0.41-0.73)	0.43(0.27-0.59)
C	職業 = 従業者（管理職）	0.64 (0.57-0.71)	0.36 (0.29-0.43)
D	職業 = 従業者（管理職） & 世帯年収 = 800-1000 万円	0.75 (0.59-0.90)	0.25 (0.10-0.41)
E	職業 = 従業者（管理職） & 世帯年収 = 1800-2000 万円	0.40 (0.04-0.77)	0.60 (0.23-0.96)
	職業 = 従業者（管理職） & 世帯年収 = わからない	0.40 (0.17-0.64)	0.60 (0.36-0.83)
F	世帯年収 = 2000 万円以上 & 性別 = 男性	0.58 (0.36-0.79)	0.42 (0.21-0.64)
	世帯年収 = 2000 万円以上 & 年齢 = 20 代	0.57 (0.24-0.90)	0.43 (0.10-0.76)

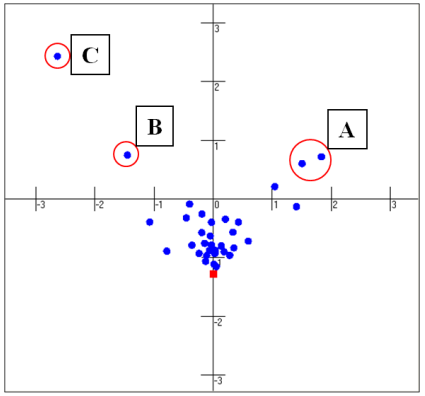
※期待値 (HPD 区間)

{ 世帯年収 = 2000 万円以上 } に関する rank 1-2 の結果を図 5(c) に示す．**F** は rank2 の属性の一部を表している．図 5(c) において，点のばらつきが小さいことから，{ 世帯年収 = 2000 万円以上 } という属性がスキャナの使用経験に関して支配的になっており，それを細分化したところで，確率分布に大きな変化は見られない，すなわち回答傾向に大きな違いはないといえる．

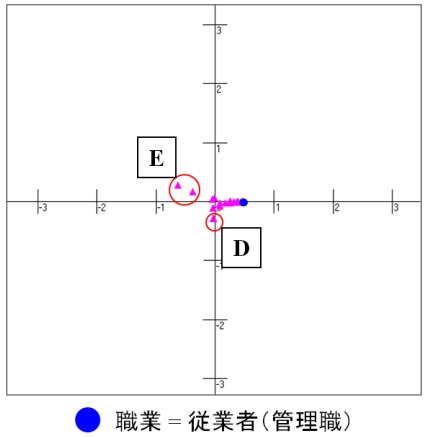
4. おわりに

本稿では，確率モデルを用いることにより母集団の傾向を考慮可能とした，新たなプロファイリング手法を提案した．提案手法では，ある属性に当てはまる回答者から，任意の質問に対する各回答の得られる確率を確率変数として扱い，その確率分布をベイズ的アプローチにより得る．また，すべての属性についての確率分布とそれらの間の類似関係を求め，その類似関係に基づいて各属性を可視空間上にプロットする．提案手法をスキャナ製品に関するアンケート調査データに適用し，解析者が，様々な属性と特定の質問の関係において，ルールの信頼性を考慮に入れたプロファイリングが可能となることを示した．また，提案手法により，属性間における回答傾向の類似関係を確率分布の観点から把握できること，および特徴的な属性の抽出を支援することが可能となることを示した．

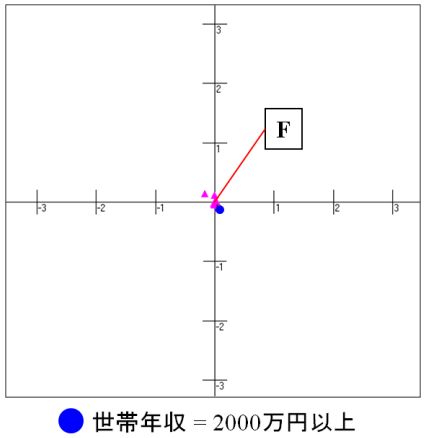
今後の課題としては，事前分布における適切なハイパーパラメータの設定方法に関する検



(a) rank 0-1



(b) rank 1-2 (rank1:**C**)



(c) rank 1-2 (rank1:**B**)

図 5 可視化結果

討や，確率分布間の類似度における他の距離測度の適用等が挙げられる．

謝辞 本研究の一部は，文部科学省科学研究費（基盤研究 (C)，No.22500088）の補助を得て遂行された．

参 考 文 献

1) Liao, S., Chen, C., Hsieh, C. and Hsiao, S.: Mining information users knowledge for one-to-one marketing on information appliance, *Expert Systems with Applications*, Vol.36, No.3, pp.4967–4979 (2009).
2) Kuroda, S., Yoshikawa, T. and Furuhashi, T.: A Proposal for Analysis of SD Evaluation Data by Using Clustering Method Focused on Data Distribution, *Proc. of the International Symposium on Frontiers of Computational Science 2005*, pp.317–320 (2007).
3) Hill, M.O.: Correspondence Analysis: A Neglected Multivariate Method, *Applied Statistics*, Vol.23, No.3, pp.340–354 (1974).
4) Agrawal, R. and Srikant, R.: Fast Algorithms for MiningAssociation Rules, *Proc. of the 20th VLDB Conf.*, pp.487–499 (1994).
5) Greenacre, M.J.: *Correspondence Analysis in Practice*, Chapman and Hall, London (2007).
6) Jiao, J. and Zhang, Y.: Product portfolio identification based on association rule mining, *Computer-Aided Design*, Vol.37, No.2, pp.149–172 (2005).
7) Yin, X. and Han, J.: CPAR: classification based on predictive association rules, *Proc. of the 3th SIAM International Conf.*, pp.331–335 (2003).
8) Scheffer, T.: Finding association rules that trade support optimally against confidence, *Intelligent Data Analysis*, Vol.9, No.4, pp.381–395 (2005).
9) Chen, M. and Shao, Q.: Monte Carlo estimation of bayesian credible and HPD intervals, *Computational and Graphical Statistics*, Vol.8, No.1, pp.69–92 (1999).
10) Minka, T.: Estimating a dirichlet distribution, Technical report, M.I.T (2000).
11) Rauber, T.W., Braun, T. and Berns, K.: Probabilistic distance measures of the Dirichlet and Beta distributions, *Pattern Recognition*, Vol. 41, No. 2, pp. 637–645 (2008).
12) Pham, M.T., Yoshikawa, T., Furuhashi, T. and Tachibana, K.: Pattern Recognition based on Two-dimensional Dendrogram Map using Spring Model, *Proc. of the 1th International Workshop on Aware Computing*, pp.614–619 (2009).