Vol.2013-MPS-96 No.18
Vol.2013-BIO-36 No.18
2013/12/12

スペクトラルクラスタリングを用いたマイノリティの抽出手法に関する検討

稲垣 和人[1]　吉川 大弘[1]　古橋 武[1]

概要：マーケティングにおいて，アンケート調査は市場調査を行う上での重要な手段の一つである．一方近年，個性の多様化により，回答者が評価対象に対して抱く印象は様々なものとなってきている．これにより，得られたアンケートの解析においては，全体の傾向を捉えるだけでなく，他の回答者とは大きく異なるが強い印象を持った少人数のグループ，いわゆるマイノリティグループの発見が有益である．しかし，アンケートデータ解析に用いられる従来のクラスタ解析手法は，基本的にマジョリティグループを抽出するものや，データを大きく分類することを目的としたものが多いため，これらマイノリティを適切に抽出することは困難である．そこで本稿では，局所的な類似性を考慮した上で，全体とのつながりの低いクラスタ抽出を特徴とするスペクトラルクラスタリングを用いた，アンケートデータにおけるマイノリティグループの抽出手法を提案する．

キーワード：アンケートデータ解析，スペクトラルクラスタリング，マイノリティ抽出

Abstract: In the field of marketing, a questionnaire is one of the most important approaches in order to research the market or to design a marketing strategy. On the other hand, people have a variety of individuality recently, then respondents have various impressions to evaluation objects. In the analysis of collected questionnaire data, it is important not only to analyze overall trends but also to discover minority groups which have strong impressions but are different from general groups. It is, however, difficult to extract minority groups by conventional cluster analyses applied to questionnaire data, because they generally aim at extracting majority groups or making a rough clustering. In this paper, we propose the extraction method of minority groups in questionnaire data using the spectral clustering method which considers local similarity and extracts the clusters having less connection to general groups.

Keywords: Questionnaire data analysis, Spectral clustering, Minority Extraction

1. はじめに

マーケティングにおいて，市場調査に基づき，企業が自社の製品やサービスに対する顧客の需要や評価を把握することは極めて重要である．例えば，企業が新しい製品の開発をする際には，対象となる顧客の需要を理解した上で企画をし，また既製品に対する顧客の評価なども考慮して販売戦略が立てられる [12].

このような市場調査の方法の 1 つがアンケート調査であり，評価対象に対する各質問項目に複数段階の評点を付けることで，回答者の対象に対する印象が数値化されたアンケートデータを得ることができる．得られたアンケートデータは一般的に，クラスター分析や，主成分分析，多次元尺度構成法などに代表される多変量解析手法 [11] を用いて解析される．しかしこれらのアプローチは基本的に，回答者全体の回答傾向や特徴抽出を行うことを目的としたものが多く，全体傾向とは大きく異なる回答は，解析結果に影響を与える可能性があるノイズとみなされてしまう．またそれにより，少数ではあるが解析の上で有益な特徴を持った，いわゆるマイノリティを抽出することは難しい．

本稿では，スペクトラルクラスタリング [7] を用いることで，少数の特徴的な回答者群を抽出することを試みる．また本稿では，ガウス関数に基づく回答者間の類似度を定義し，2 分割の繰り返しによりマイノリティ候補を 1 グループずつ抽出する手法を提案する．さらに，回答者間の類似度指標に用いられるパラメータを，ベイズ情報量規準を用いて自動で決定する方法を提案する．初めに，仮想アンケートデータにおける評価実験により，提案手法を用いることで，想定したマイノリティグループが，適切に抽出

[1] 名古屋大学大学院工学研究科
Graduated School of Engineering Nagoya University, Furocho, Nagoya 464–8601, Japan

できることを示す．次に，実際のアンケートデータに提案手法を適用し，特徴的な評点傾向を持つ少人数のグループが複数抽出されることを示す．

2. スペクトラルクラスタリング

スペクトラルクラスタリングは，グラフのノードにデータ，ノード間のエッジの重みにデータ間の類似度を対応させ，クラスタリングをグラフ分割の問題として解く手法である．このように表されたグラフについて，枝切りを行うことで全体のグラフをいくつかのサブグラフに分割する．その際に，サブグラフ内のエッジが密になり，サブグラフ間のエッジが疎になるような評価関数を設定する．このための評価関数はいくつか提案されているが，本稿では代表的な $Ncut$[7] を利用する．まず，グラフのノード集合 V を 2 つのサブグラフ A と B に分けることを考える．あるノード i, j の間でのエッジの重みを $w(i,j)$ としたとき，サブグラフ A と B の類似度 $cut(A,B)$ を以下のように定義する．

$$cut(A,B) = \sum_{i \in A, j \in B} w(i,j) \tag{1}$$

このとき，評価関数 $Ncut$ は以下で表される．

$$Ncut(A,B) = \frac{cut(A,B)}{cut(A,V)} + \frac{cut(A,B)}{cut(B,V)} \tag{2}$$

この式を最小化することは，サブグラフ内の類似度を大きく，かつサブグラフ間の類似度を小さくすることに等しい．またこの最小化問題は，一般化固有値問題に帰着することが知られている．W をデータ間の類似度行列，D を W の次数を対角成分に持つ行列とすると，$D^{-1}(D-W)$ の固有ベクトルがグラフの分割を与える．ただし最小固有値は 0 となるため，2 番目に小さな固有値に対する固有ベクトルを用い，ある値以上の要素値を持つノードをクラスタ A に，それより小さいノードをクラスタ B に対応させることでクラスタリングを行う．この要素値に対応するしきい値は，0 とするもの，中央値を用いるもの，以下のように $Ncut$ の最小となる値を用いるものが代表的である．本稿では，各カット位置，すなわちすべての要素値をしきい値としてそれぞれ $Ncut\,(A,B)$ の値を算出し，$Ncut\,(A,B)$ が最小となるカット位置でのクラスタリング結果を得る．

3. 提案手法

ここでは，前節で示したスペクトラルクラスタリングを用いて，マイノリティを抽出する手法について述べる．なお，本研究におけるマイノリティの定義は，他の回答者群との類似度は低い一方で，グループ内の類似度は高い回答者群とする．基本的には少人数のグループを想定しているが，この条件を満たしていれば人数は問わない．

3.1 回答者間の類似度の定義

アンケートデータにおける，回答者 a，b の各質問項目に対する評点を要素としたベクトルをそれぞれ $\boldsymbol{x_a}$，$\boldsymbol{x_b}$ としたとき，回答者 a，b 間の類似度 $w(a,b)$ を次式で定義する．

$$w(a,b) = \exp(-\frac{||\boldsymbol{x_a} - \boldsymbol{x_b}||^2}{\sigma^2}) \tag{3}$$

(3) 式はガウス関数であり，σ^2 は分散値を表すパラメータである．この関数は，$||\boldsymbol{x_a} - \boldsymbol{x_b}||$ が小さいときの類似度を強調し，ある程度以上評点が離れていることに対する類似度を極めて低い値とするという特徴を持つ．またこの特徴は，σ の値が小さいほど極端になる．

2 節で述べたように，スペクトラルクラスタリングでは，サブグラフ内の類似度が高く，サブグラフ間の類似度が低くなるようにグラフを分割する．そのため回答者間の類似度を (3) 式のように定義することで，グループ内の類似度，および他との非類似度が強調され，本研究で求めるマイノリティが適切に抽出されることが期待できる．

3.2 パラメータ σ の決定法

(3) 式の回答者間類似度関数における σ は，クラスタリング実行前に決定すべきパラメータとなる．しかし σ の値の違いはクラスタリング結果に大きな影響を与え，適切な値を設定することは難しい．通常，アンケートデータ解析は，試行錯誤を繰り返すことで，多角的にデータの特徴を捉える必要があるため，σ の値を変化させてそれぞれ得られたグループを解析する方法も有効であると考えられるが，本稿では，マイノリティグループが局所的に密な多変量正規分布に従うという仮定のもとに，最適な σ を決定する方法を提案する．文献 [5] では，代表的なクラスタリング手法の一つである K-means 法において，ベイズ情報量規準（Bayesian Information Criterion: BIC）[6] を用いて最適なクラスタ数を決定する手法として，X-means 法が提案されている．BIC は以下の式で表される．

$$BIC = -2\log L + k\log n \tag{4}$$

ここで，L は尤度関数，n は標本数，k は母数の数である．本手法では，σ の値を一定の範囲内で変化させ，各 σ 値で抽出されたマイノリティグループの多変量正規分布に対する BIC を算出し，最小となるときの σ の値を最適値とする．

3.3 2 分割の繰り返しによるマイノリティの抽出

スペクトラルクラスタリングにおいて，任意のクラスタ数への分割に拡張された方法が報告されている [9]．しかしこの方法では，クラスタ数を事前に決定する必要があるため，存在するマイノリティの数が不明であるアンケートデータ解析での適用は難しい．そこで本稿では，2 節で述べた 2 分割によるクラスタリングを，回答者数の多いグループに対して繰り返すことで，マイノリティ候補を 1 クラスタずつ抽出する方法を用いる．従来のクラスタリング手法において，クラスタ数を十分に大きくして分類を行う

Vol.2013-MPS-96 No.18
Vol.2013-BIO-36 No.18
2013/12/12

ことでも，本稿で対象とするマイノリティの抽出を行うことも可能であると考えられるが，得られた多数のクラスタから，特徴を持ったマイノリティを探索することが必要となるため，マイノリティ候補を 1 つずつ抽出し，特徴解析を行うアプローチが，実用上は有用であると思われる．

4. 関連研究

基本的に，アンケートデータ解析におけるマイノリティの抽出を目的とした研究報告は少ない．マイノリティ抽出に用いることが可能であると考えられる手法はいくつか存在するが，それらの多くは，どちらかといえばはずれ値や特異なデータの抽出を行うためのものであり，実際のデータに適用すると，ほとんどの場合，特異なデータが 1 つずつ抽出されてしまう [3], [10]．その中で，マイノリティグループのクラスタリングを目的とした研究として，安藤らは情報理論的クラスタリングを用いて，大域的に分布する典型事例（マジョリティ）と局所的に分布する特異事例の混在したデータのクラスタリングを行っている [2]．また Gonzalez らは，計算コストの低い Weak Clustering[8] を複数回試行し，それらの結果をもとに局所的に密集したデータの抽出を行っている [4]．しかしこれらの手法は，分布の中で最も密な集団を抽出するという特徴がある．一般にアンケートデータには，中心評点付近をつける回答者が多く存在するため，上述の手法を適用すると，これら特徴の少ない，いわばマジョリティグループが抽出されることになる．前述の通り，本研究では，グループ内の類似度が高い密な集団であると同時に，他の回答者との類似度が低いグループを抽出することを目的としており，上述の手法ではこれらの抽出は困難であると考えられる．さらに文献 [2] では，マジョリティ，マイノリティそれぞれが従う分布形状を事前に仮定する必要がある．

5. 評価実験

本節では，3 節で述べたマイノリティ抽出手法を仮想アンケートデータ，実際のアンケートデータそれぞれに適用し，その性能を評価する．

5.1 仮想アンケートデータへの適用

5.1.1 仮想アンケートデータ

仮想的なアンケートデータとして，評価対象数 1，質問数 10，回答者数 650 の 5 段階評定尺度法によるアンケートデータを用意した．また回答者は，**表 1** に示す 7 つの回答傾向を持つグループを設定した．ここで，グループ 5，6，7 が，全体傾向とは大きく異なる回答傾向を持つ，マイノリティと想定したグループである．各回答者において，10 個の質問に対する評点を要素とした 10 次元ベクトル間のユークリッド距離を用い，多次元尺度構成法（MDS）により回答者評点を可視化した結果を**図 1** に，各グループおよび全回答者の平均評点を**図 2** にそれぞれ示す．

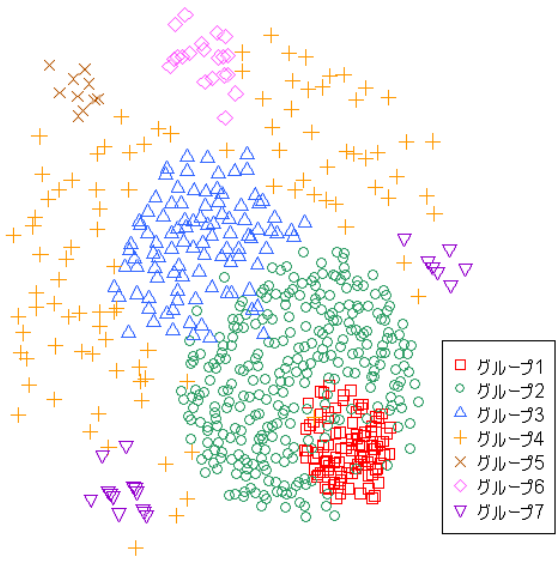


図 1 回答者評点の可視化結果
Fig. 1 Visualization of evaluation scores

5.1.2 実験方法

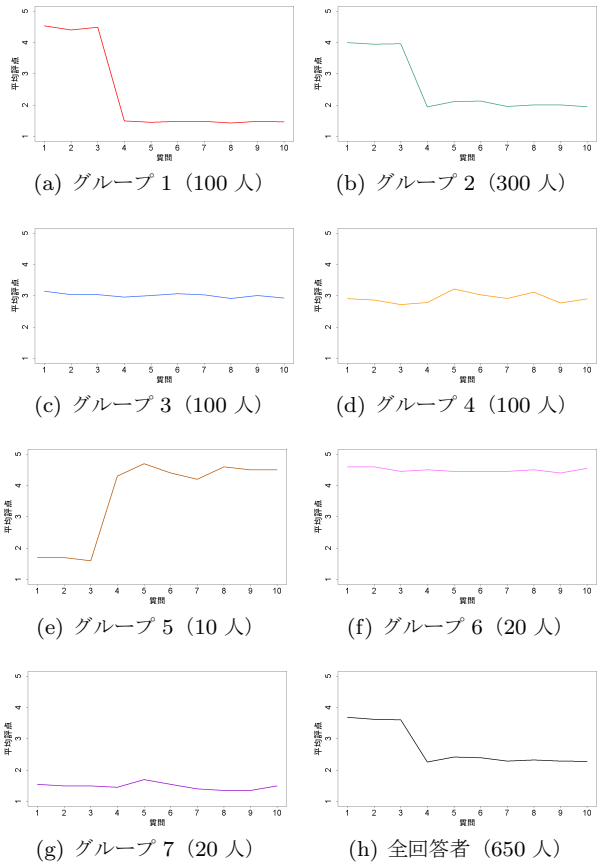
5.1.1 で示した仮想アンケートデータに対して，3 節で述べた提案手法を用いてマイノリティの抽出を 3 回行った．各抽出において，3.2 で示した方法により，σ^2 の値を 1 から 10 の範囲（刻み幅 0.2）で決定した．またここでは，代表的なクラスター分析手法であるデンドログラムとの比較を行った．

5.1.3 結果と考察

MDS による回答者評点の可視化結果に対して，提案手法により抽出された 3 つのクラスタを示したもの，抽出された各クラスタの平均評点をそれぞれ**図 3**，**図 4** に示す．

表 1 各グループの特徴
Table 1 Characteristics of each group

グループ	特徴
グループ 1	質問 1〜3 に 4 または 5 の評点を，質問 4〜10 に 1 または 2 の評点をそれぞれランダムに付けた回答者（100 人）
グループ 2	質問 1〜3 に 3〜5 のいずれかの評点を，質問 4〜10 に 1〜3 のいずれかの評点をそれぞれランダムに付けた回答者（300 人）
グループ 3	全ての質問に 2〜4 のいずれかの評点をランダムに付けた回答者（100 人）
グループ 4	全ての質問に 1〜5 のいずれかの評点をランダムに付けた回答者（100 人）
グループ 5	質問 1〜3 に 1 または 2 の評点を，質問 4〜10 に 4 または 5 の評点をそれぞれランダムに付けた回答者（10 人）
グループ 6	全ての質問に 4 または 5 の評点をランダムに付けた回答者（20 人）
グループ 7	全ての質問に 1 または 2 の評点をランダムに付けた回答者（20 人）



(a) グループ 1（100 人） (b) グループ 2（300 人）
(c) グループ 3（100 人） (d) グループ 4（100 人）
(e) グループ 5（10 人） (f) グループ 6（20 人）
(g) グループ 7（20 人） (h) 全回答者（650 人）

図 2 各グループの人数および平均評点
Fig. 2 Number of respondents and average scores in each group

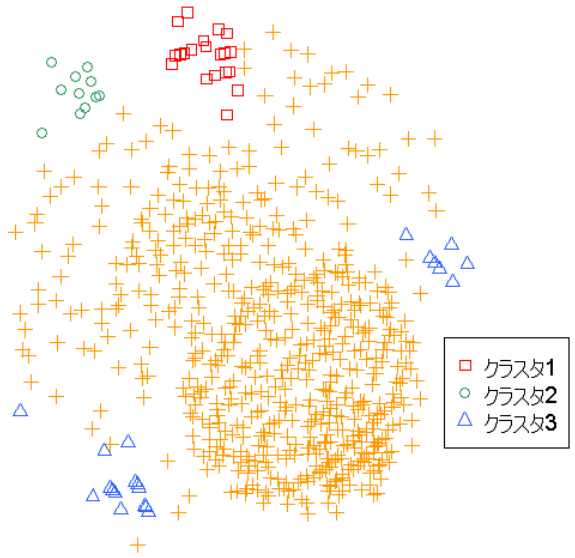
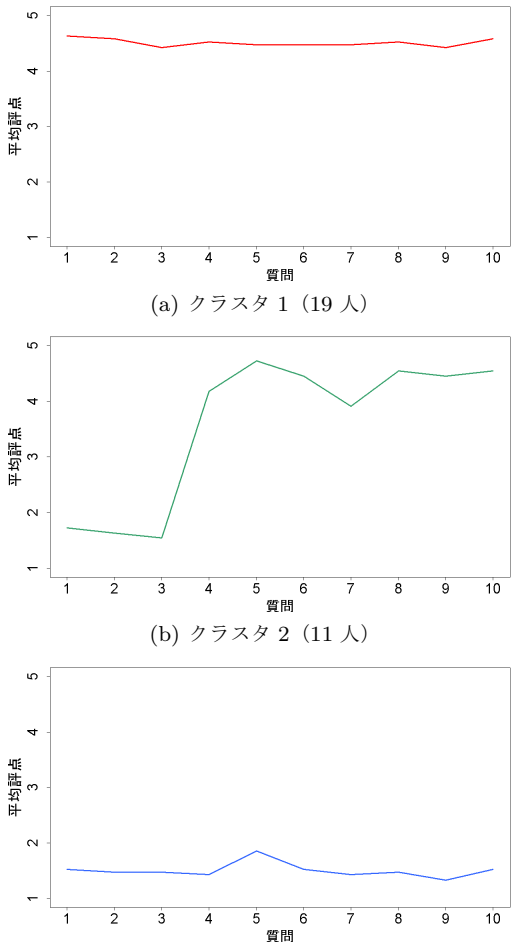


図 3 提案手法によるクラスタリング結果（仮想データ）
Fig. 3 Clustering result by proposed method (virtual data)

また各抽出において，(4) 式により算出された各 σ^2 値に対する BIC の値を**図 5** に示す．

想定したマイノリティであるグループ 5，6，7 が，それぞれグループ 6（クラスタ 1），グループ 5（クラスタ 2），グループ 7（クラスタ 3）の順で抽出された．クラスタ 2，



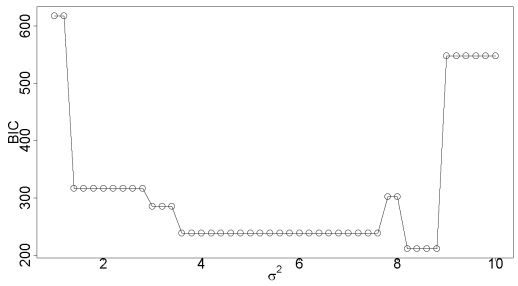
(a) クラスタ 1（19 人）
(b) クラスタ 2（11 人）
(c) クラスタ 3（21 人）

図 4 抽出された各クラスタの平均評点（仮想データ）
Fig. 4 Average scores of extracted clusters (virtual data)

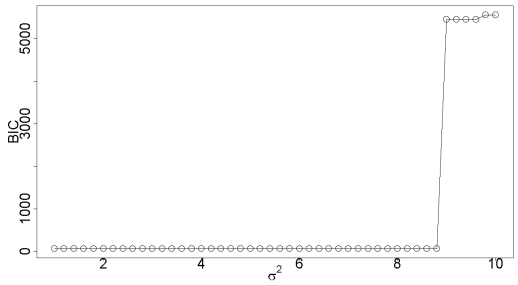
3 がそれぞれ図 2 で示した設定よりも 1 人ずつ多いのは，全ての質問にランダムな評点を与えたグループ 4 の中に，グループ 5 や 7 と近い評点となった回答者がいたためである．逆にマイノリティとして設定したグループ 5，6，7 の回答者は，グループ 6 の 1 人を除き，全てクラスタ 2，1，3 に網羅されていた．これらの結果から，提案手法によってマイノリティが適切に抽出可能であることが確認できた．また図 5 から，本実験では，σ^2 の値は比較的広い範囲で最適値となることが確認できた．

次に，デンドログラムによる結果を**図 6** に示す．クラスタの結合には，最も分類感度が高いといわれるウォード法 [1] を用いた．図 6 において，グループ 5，6，7 はそれぞれクラスタを形成し，他のグループとも比較的離れて結合されてはいるものの，階層的に分割していくことを考えた場合，3，4，6 番目の分割でそれぞれグループ 7，6，5 が抽出され，これらのグループのみを単体でマイノリティとして抽出することは難しいことがわかる．今回用いたウォード法では，グループ内の分散に対するグループ間の分散を最大化する基準でクラスタリングを行うため，各クラスタに属するデータ数が等しくなる傾向がある．各クラ

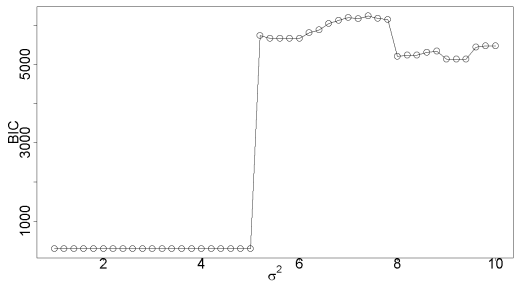
Vol.2013-MPS-96 No.18
Vol.2013-BIO-36 No.18
2013/12/12



(a) 1 回目（σ^2 最適値:8.2-8.8）



(b) 2 回目（σ^2 最適値:1.0-8.8）



(c) 3 回目（σ^2 最適値:1.0-5.0）

図 5 σ^2 の探索結果
Fig. 5 Results of σ^2

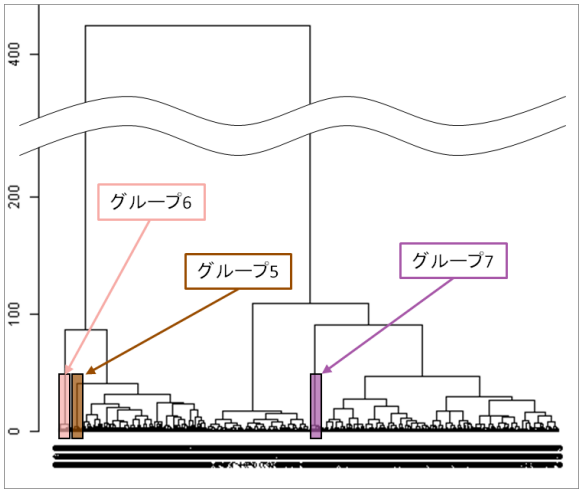


図 6 デンドログラムによる結果（仮想データ）
Fig. 6 Result of dendrogram (virtual data)

スタに属するデータ数の偏りを許容する方法として単連結法 [1] が存在するが，この方法では，連鎖と呼ばれる，ある 1 つのクラスタにデータが 1 つずつ結合していく現象が起きやすく，明確なクラスタリングを行ってマイノリティのグループを抽出するという方法に向いていない．

5.2 実際のアンケートデータへの適用

5.2.1 アンケートデータ

次に，実際のアンケートデータを用いて実験を行った．1014 名の回答者に対して，次世代型サービスに関する Web アンケートを行った．本アンケート調査では，**表 2** に示す次世代型サービスに対する 6 つの説明を評価対象とし，各対象に対し，**表 3** に示す 10 の質問項目（合計質問数:60）について，“まったく当てはまらない”から“当てはまる”までの 5 段階の数値で評価してもらった．

5.2.2 実験方法

5.1.2 と同様，提案手法を用いてマイノリティの抽出を行った．各回答者の評点ベクトルは，6 対象× 10 質問に対する評点，計 60 次元のベクトルで表したものを用いた．σ^2 の値は，1 から 10 の範囲（刻み幅 0.5）で決定した．

5.3 結果と考察

提案手法により抽出されたクラスタ 1〜5 を示した，MDS による回答者評点の可視化結果を**図 7** に示す．また，各クラスタおよび全回答者の平均評点を**図 8** に示す．

図 8(a) のクラスタ 1 は，**図 8(f)** に示す全回答者の平均評点に対し，ほぼ逆の回答傾向を持つ回答者群であることがわかる．また**図 8(b)** のクラスタ 2 については，全ての質問に対して平均評点が 1 または 5 付近となっており，比較的極端な評点を付けた回答者群であることがわかる．クラスタ 3，4 についても，クラスタ 2 とほぼ同様の傾向で，

表 2 評価対象
Table 2 Evaluation objects

対象	内容
対象 1	アフターサービスに関する曖昧な説明
対象 2	ユビキタスに関する曖昧な説明
対象 3	リサイクルに関する曖昧な説明
対象 4	アフターサービスに関する詳細な説明
対象 5	ユビキタスに関する詳細な説明
対象 6	リサイクルに関する詳細な説明

表 3 質問項目
Table 3 Questions

質問	内容
質問 1	どんなものか興味がある
質問 2	周りの人にも勧めたいと思う
質問 3	社会的需要が高く，普及しそうだ
質問 4	提供する企業のイメージが向上しそうだ
質問 5	提供企業の負担が大きすぎると思う
質問 6	狙う方向が間違っていると思う
質問 7	特定の人々にしか評価されないだろう
質問 8	社会の課題の本質を突いていると思う
質問 9	社会的には重要だが，提供企業の負担が大きいので公的機関が補助すべきだと思う
質問 10	近未来的なサービスだと思う

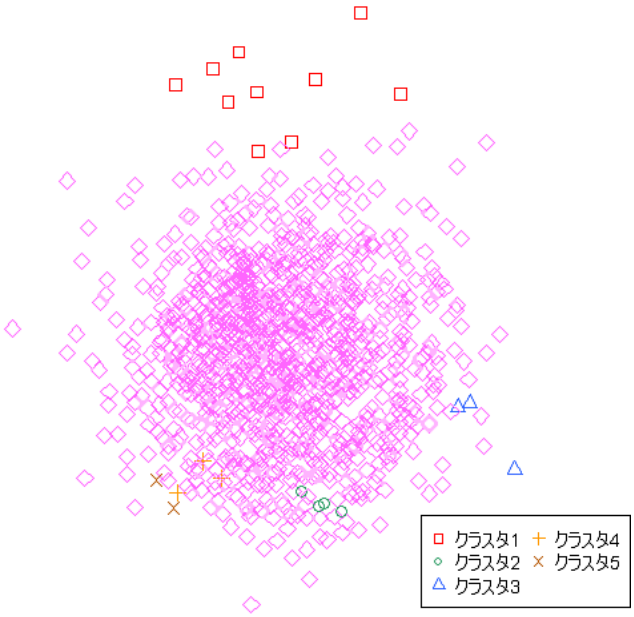
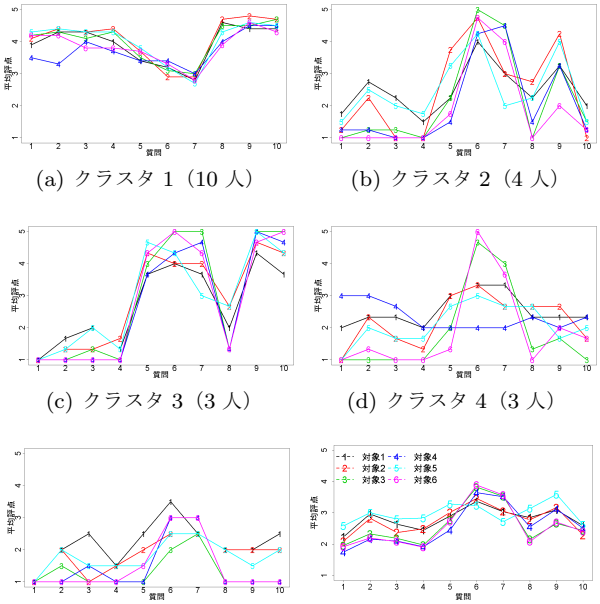


図 7 提案手法によるクラスタリング結果（実際のデータ）
Fig. 7 Clustering result by proposed method (real data)



(a) クラスタ 1（10 人） (b) クラスタ 2（4 人）
(c) クラスタ 3（3 人） (d) クラスタ 4（3 人）
(e) クラスタ 5（2 人） (f) 全回答者（1014 人）

図 8 各クラスタの人数および平均評点（実際のデータ）
Fig. 8 Number of respondents and average score in each cluster (real data)

主に質問 9，10 に対する評点の違いがクラスタを分けていると考えられる．さらにクラスタ 5 では，ほぼ全ての質問に低い評点を付けた回答者群であった．このように，提案手法を用いることで，特徴的な評点傾向を持つクラスタが抽出されたと考えられる．

6. おわりに

本稿では，スペクトラルクラスタリングを用いた，アンケートデータにおけるマイノリティグループの抽出手法を提案した．提案手法では，ガウス関数に基づいて回答者間類似度を定義し，2 分割の繰り返しによるマイノリティ候補を 1 グループずつ抽出する．さらに，回答者間の類似度指標に用いられるパラメータ σ を，ベイズ情報量規準を用いて自動で決定する．初めに，仮想アンケートデータにおける評価実験により，提案手法を用いることで，想定したマイノリティグループが，適切に抽出できることを示した．次に，実際の Web アンケートデータに提案手法を適用し，特徴的な評点傾向を持つ少人数のグループが複数抽出されることを示した．今後の課題として，抽出されたマイノリティの妥当性に関する検証や，回答者間の類似度関数と得られる結果との関係性の解析などが挙げられる．

参考文献

[1] Anderberg, M. R.，西田英郎：クラスター分析とその応用，内田老鶴圃 (1988).
[2] Ando, S. and Suzuki, E.: Detecting Clusters of Outliers with Information Theoretic Clustering, *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 23, pp. 344–354 (2008).
[3] Comaniciu, D. and Meer, P.: Mean shift: A robust approach toward feature space analysis, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 24, No. 5, pp. 603–619 (2002).
[4] Gonzàlez, E. and Turmo, J.: Unsupervised ensemble minority clustering, *Machine Learning*, pp. 1–52 (2012).
[5] Pelleg, D., Moore, A. W. et al.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters., *ICML*, pp. 727–734 (2000).
[6] Schwarz, G.: Estimating the dimension of a model, *The annals of statistics*, Vol. 6, No. 2, pp. 461–464 (1978).
[7] Shi, J. and Malik, J.: Normalized cuts and image segmentation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 22, No. 8, pp. 888–905 (2000).
[8] Topchy, A., Jain, A. K. and Punch, W.: Combining multiple weak clusterings, *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, IEEE, pp. 331–338 (2003).
[9] Von Luxburg, U.: A tutorial on spectral clustering, *Statistics and computing*, Vol. 17, No. 4, pp. 395–416 (2007).
[10] Wang, B., Xiao, G., Yu, H. and Yang, X.: Distance-based outlier detection on uncertain data, *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on*, Vol. 1, IEEE, pp. 293–298 (2009).
[11] 君山由良：データ分析入門 2 多変量解析法・MDS の応用，Vol. 2, Data Analysis Instetute, Inc (2008).
[12] 木下祐介，井上勝雄，酒井正幸：携帯電話機デザインの男女差の調査分析，感性工学研究論文集，Vol. 7, No. 3, pp. 449–460 (2008).