1E5-4

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の回答者分類に関する検討

A Study on Clustering of Respondents in Questionnaire Data based on Spectral Clustering

稲垣　和人　　吉川　大弘　　古橋　武
Kazuto Inagaki　Tomohiro Yoshikawa　Takeshi Furuhashi

名古屋大学大学院工学研究科
Graduated School of Engineering Nagoya University

In the field of marketing, questionnaire is often carried out in order to design a marketing strategy by analyzing collected data. Recently, people have a multiple of individuality, so respondents have various impressions. It is important to focus on minority groups which have strong impression but are different from general groups. It is, however, difficult to extract minority groups by conventional cluster analysis methods. This paper aims to extract minority groups in questionnaire. We focus on the spectral clustering method that considers local similarity. This paper applies the proposed method to actual questionnaire data and shows the effectiveness.

1. はじめに

マーケティングにおいて，企業が市場調査を通して，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 [Kinoshita 08]．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 [Kimiyama 08] を用いて解析される．しかしこれらのアプローチでは，解析結果に影響を与える可能性があるノイズとなる回答や，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．そこで本稿では，スペクトラルクラスタリング [Shi 00] 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．

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，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．このとき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が対応する．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，ここでは代表的な $Ncut$[Shi 00] 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 u，v の間でのエッジの重みを $w(u,v)$ 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．

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

連絡先: 稲垣和人，名古屋大学大学院工学研究科，名古屋市千種区不老町，052-789-2793，052-789-3166，inagaki@cmplx.cse.nagoya-u.ac.jp

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示されている．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．本稿では，各カット位置に対する $Ncut\,(A,B)$ の値を算出し，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. マイノリティ抽出手法

ここでは，2. 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．

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，ある程度以上の距離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．この特徴は σ の値が小さいほど極端になる．これにより，他の回答者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が強調される．

3.2 スペクトラルクラスタリングの適用

2. 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 3.1 のように定義することで，強調された回答者群がマイノリティとして抽出されることが期待できる．

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 [Luxburg 07]．しかしクラスタ数を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. で述べた 2 分割を，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 つずつ抽出する本手法が，実用上は有用であると思われる．

4. 実験

4.1 実験内容

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，各 3 つのサービス（アフターサービス，ユビキタス，リサイクル）についての曖昧な説明と具体的な説明の合わせて 6 つの説明を評価対象とし，評定尺度法により，10 個の質問項目に対してそれぞれ 5 段階の評点 {1,2,3,4,5} で評価してもらった．

得られたアンケートデータに対して，3. で述べた方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 質問の合計 60 個の評点を並べたものを用いた．ただし，(3) 式における $\sigma^2 = 5$ とし，回答者群の抽出を 3 回行った．

4.2 結果と考察

多次元尺度構成法による回答者の可視化結果を図 1 に示す．ただし，多次元尺度構成法における回答者 a，b 間の距離基準は $||\boldsymbol{x_a} - \boldsymbol{x_b}||$（ユークリッド距離）とした．全回答者の平均評点と，各クラスタにおける平均評点をそれぞれ図 2，図 3 に示す．

図 3(a) のクラスタ 1 は，図 2 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．図 3(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 付近であり，極端な評点を付けた回答者群であることがわかる．またクラスタ 3（図 3(c)）については，一般的なアンケートに多くみられる，全ての質問に対して 3 付近の評点を付けた回答者群であり，250 人と多人数であるものの，強い特徴を持った回答者群であるといえる．このように，提案手法を用いることで，本稿で対象とするマイノリティである特徴的なクラスタが抽出されたと考えられる．

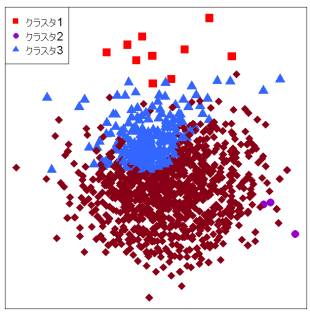


図 1: 回答者の可視化結果

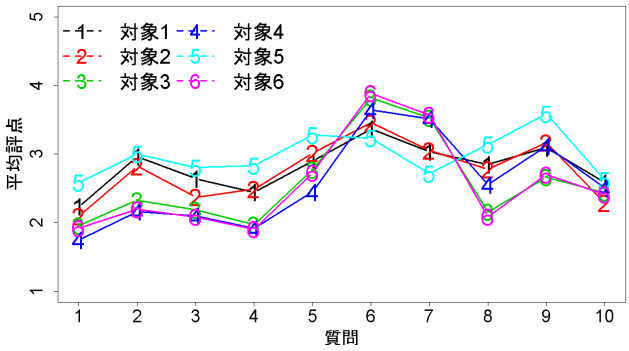
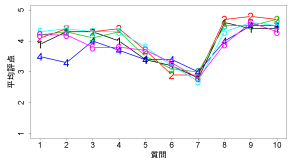
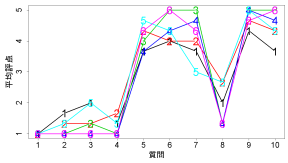


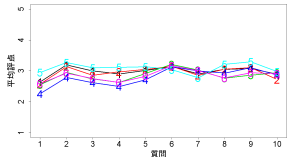
図 2: 全回答者の平均評点



(a) クラスタ 1（10 人）



(b) クラスタ 2（3 人）



(c) クラスタ 3（250 人）

図 3: 各クラスタの平均評点

5. 終わりに

本稿では，スペクトラルクラスタリングを用いて，アンケートデータにおける回答者のクラスタリングを行った．実際の Web アンケートデータに適用し，特徴的な回答傾向を持つ少数の回答者群が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

[Kinoshita 08] 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，感性工学会研究論文集，Vol.7，No.3，pp.449-460，2008.

[Kimiyama 08] 君山由良：『データ分析入門 2 多変量解析法・MDS の応用』，データ分析研究所，2008.

[Shi 00] J.Shi and J.Malik：Normalized cuts and image segmentation，IEEE Trans，Pattern Analysis and Machine Intelligence，Vol.22，N0.8，pp.888-905，2000.

[Luxburg 07] Ulrike von Luxburg：A Tutorial on Spectral Clustering，Statistics and Computing，17(4)，2007.