A-002

アンケートデータにおけるマイノリティの抽出手法およびその定量化に関する検討

A Study on Extraction Method and Quantification of Minority Groups in Questionnaire Data

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

1. はじめに

マーケティングにおいて，企業が新しい製品の開発をする際には，ターゲット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 [1]．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 [2] 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．そこで本稿では，Normalized cut[3] 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．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，少人数の回答者群とする．本稿では，上述の定義を定量的に表したマイノリティ指標を導入し，代表的なクラスタリング手法の 1 つである k-means 法との比較を行う．

2. Normalized cut

Normalized cut は，データを個体間の類似度に基づいてグラフ表現し，そのスペクトル（固有値）を用いてサブグラフに分割することでクラスタリングを行う手法である．

ある個体 i，j の間の類似度を $w(i,j)$ としたとき，グラフ全体をサブグラフ A と B に分割した際の 2 つの指標 cut と vol は，それぞれ以下のように定義される．

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

$$vol(A) = \sum_{i\in A} d_i, vol(B) = \sum_{i\in B} d_i \tag{2}$$

ただし，$d_i = \sum_j w(i,j)$ で定義され，分割によらず $vol(A\bigcup B) = vol(A) + vol(B) = const.$ である．cut はサブグラフ間の類似度，vol はサブグラフの大きさを表す値であるといえる．このとき，Normalized cut では，次の評価関数 $Ncut$ を最小化する分割を行う．

$$Ncut(A,B) = cut(A,B)\left(\frac{1}{vol(A)} + \frac{1}{vol(B)}\right) \tag{3}$$

ここで，(3) 式の最小化は，(1) 式の $cut(A,B)$ の最小化に対して，2 つのサブグラフの vol の偏りを小さくする制約を加えたものと解釈できる．つまり，他の全ての個体と低い類似度を持つ“外れ値”を切り取るような分割を防ぐことが可能となる．

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個体をクラスタ A に，それより小さい個体をクラスタ 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，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

前節で述べたように，Normalized cut はクラスタ間の大きさの偏りを小さくする働きがあるため，少数データであるマイノリティを抽出するには一見適していないと思われる．しかし一般にアンケートデータでは，多数の回答者が特定の質問に対し，高い／低い評点に偏って評点をつける傾向がある．そのため回答者間の類似関係としては，それらマジョリティが密に類似し，それらとの類似度は低いが，互いに類似したマイノリティグループがいくつか存在するデータ分布になると考えられる．よって Normalizde cut をアンケートデータに適用した場合，マジョリティグループを分割すると cut の値が大きくなるために，マジョリティ以外の比較的少数の類似グループであるマイノリティが切り出されることになる．以上の理由から，アンケートデータにおいて Normalized cut がマイノリティ抽出に適した手法であると考える．

次に提案手法における分割方法について説明する．2 節では 2 クラスタへの分割方法について述べたが，近年 Normalized cut を用いた複数クラスタへの分割法が報告されている [4]．しかし事前に適切なクラスタ数を決定することは困難なこと，また設定したクラスタ数によって結果が変化することが問題となると考えられる．そこで提案手法では，回答者数の最も多いクラスタに対する 2 分割を繰り返すことで，マイノリティを逐次的に抽出する方法を用いる．

† 名古屋大学大学院工学研究科

4. マイノリティの定量化

Normalized cut における個体 i,j 間の類似度 $w(i,j)$ は一般に，次式で表されるガウス関数によって定義される．

$$w(i,j) = \exp(-\frac{||\boldsymbol{x_i} - \boldsymbol{x_j}||^2}{\sigma^2}) \quad (4)$$

ここで，$\boldsymbol{x_i}$，$\boldsymbol{x_j}$はそれぞれ個体 i，j の特徴ベクトル，σ^2 はガウス分布の分散値を表すパラメータである．提案手法では，(3) 式の $Ncut$ を評価関数としているが，この値は σ^2 の値に依存するため，得られたマイノリティを評価するための指標として，$Ncut$ そのものを用いることはできない．そこで本稿では，以下の値をマイノリティ指標 MC と定義する．

$$MC = \frac{\sum_{i \in M} \sum_j (x_{ij} - \mu_{Mj})}{\sum_{i \in N} \sum_j (x_{ij} - \mu_{Mj})} \quad (5)$$

ここで，M はマイノリティ回答者の集合，μ_{Mj} は M における質問 j の平均評点，N はマイノリティ中心ベクトル$\boldsymbol{\mu_M}$とのユークリッド距離が小さい k 人（マイノリティ回答者の人数）のマジョリティ回答者の集合である．分子はマイノリティの密集度，分母はマイノリティとマジョリティの分離度を表しているため，MC の値が小さいほど M はマイノリティとみなすことができる．

5. 実験

5.1 概要

本実験では，あるアウトドア製品 α に関する実際のアンケートデータを用いて，代表的なクラスタリング手法である k-means 法との比較を行った．本調査では，製品 α を使用している 6 種類の動画を評価対象とし，10 項目の質問に対しそれぞれ {1,2,3,4,5} の 5 段階評点で評価してもらった．回答者数は 1,453 名である．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．提案手法の (4) 式における σ^2 の値は 5，k-means 法におけるクラスタ数は 100 とした．比較には提案手法で抽出した 4 クラスタと，k-means 法で得られたクラスタのうち回答者数が同数かつ (5) 式の指標値 MC が最も低いクラスタをそれぞれ比較した．

5.2 結果と考察

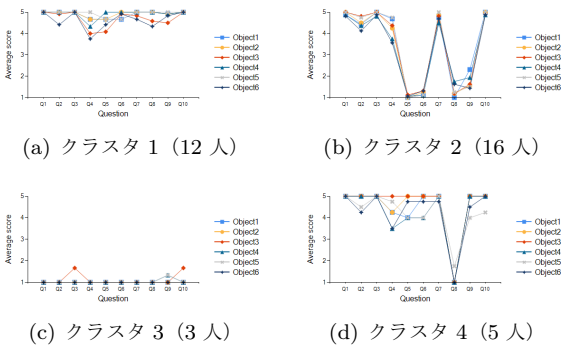
提案手法と k-means 法による各クラスタのマイノリティ指標値を表 1 に，平均評点をそれぞれ図 1，2 に示す．表 1 より，提案手法が k-means 法よりも指標の上で優れたマイノリティを抽出できたことがわかる．これは，k-means 法では他のクラスタとの分離度を考慮していないために，密に分布した回答者を細分化してしまい，指標値が高くなったと考えられる．

6. おわりに

本稿では，Normalized cut を用いた，アンケートデータにおけるマイノリティの抽出手法を提案した．またマイノリティの密集度とマジョリティとの分離度に基づくマイノリティ指標を導入した．実際のアンケートデータに適用し，指標の上で提案手法が k-means 法よりも優れたマイノリティを抽出できることを示した．今後の課題として，マイノリティ指標の妥当性に関する検証が挙げられる．

表 1: 各クラスタのマイノリティ指標値

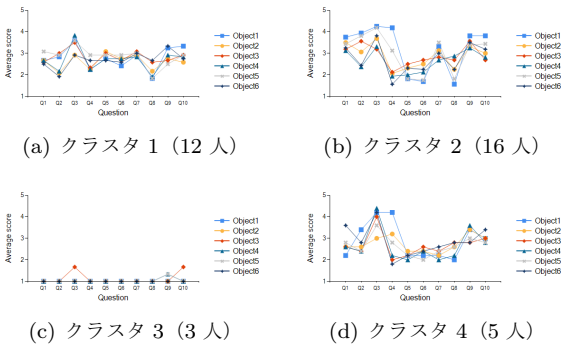
クラスタ	提案手法	k-means 法
クラスタ 1（12 人）	0.114	0.887
クラスタ 2（16 人）	0.355	0.838
クラスタ 3（3 人）	0.061	0.061
クラスタ 4（5 人）	0.389	0.820



(a) クラスタ 1（12 人） (b) クラスタ 2（16 人）

(c) クラスタ 3（3 人） (d) クラスタ 4（5 人）

図 1: 提案手法による各クラスタの平均評点



(a) クラスタ 1（12 人） (b) クラスタ 2（16 人）

(c) クラスタ 3（3 人） (d) クラスタ 4（5 人）

図 2: k-means 法による各クラスタの平均評点

参考文献

[1] 木下, 他. 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.

[2] 君山. データ分析入門 2 多変量解析法・MDS の応用, 第 2 巻. Data Analysis Instetute, Inc, 2008.

[3] Jianbo Shi, Malik, et al. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 888–905, 2000.

[4] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, Vol. 17, No. 4, pp. 395–416, 2007.