3J3-2

Normalized cutを用いたマイノリティの抽出手法に関する検討

A Study on Extraction Method of Minority Groups based on Normalized cut

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 normalized cut 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 [木下 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 [君山 08] 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[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．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．

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，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，分割のための評価関数 $Ncut$ は以下で表される．ただし，V は全個体の集合である．

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

この $Ncut\,(A,B)$ の値を最小化する分割を行う．これは，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

ここでは，前節で示した Normalized cut を用いて，マイノリティを抽出する提案手法について説明する．

3.1 逐次抽出

一般にアンケートデータでは，多数の回答者が，中心評点付近，あるいは特定の質問に対し，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グループが密に類似し，それらとの類似度は低いが，互いに類似したマジョリティグループがいくつか存在すると考えられる．そこで提案手法では，2. で示した Normalized cut 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 クラスタずつ逐次的に抽出する方法を用いる．

3.2 類似度関数におけるパラメータの決定法

個体間の類似度関数には以下のガウス関数を用いる．

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

$\boldsymbol{x_a}$，$\boldsymbol{x_b}$は各個体を表すベクトル，σ^2 は分散値を表すパラメータである．σ^2 はクラスタリング実行前に決定する値であるが，この値はクラスタリング結果に大きな影響を与え，予め適切な

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

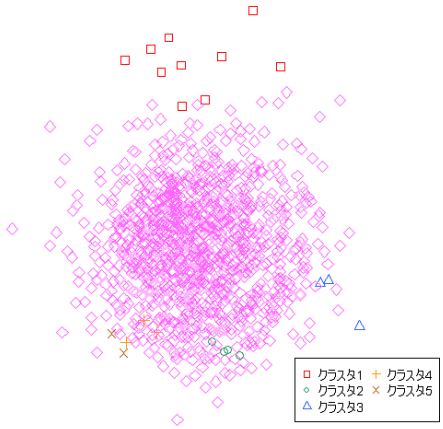


図 1: 提案手法によるクラスタリング結果

値を設定することは難しい．そこで本稿では，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 σ^2 を自動で決定する方法を提案する．文献 [Pelleg 00] 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）[Schwarz 78] 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^2 の値を一定の範囲内で変化させ，各 σ^2 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，その値が最小となるときの σ^2 の値を用いる．

4. 実験

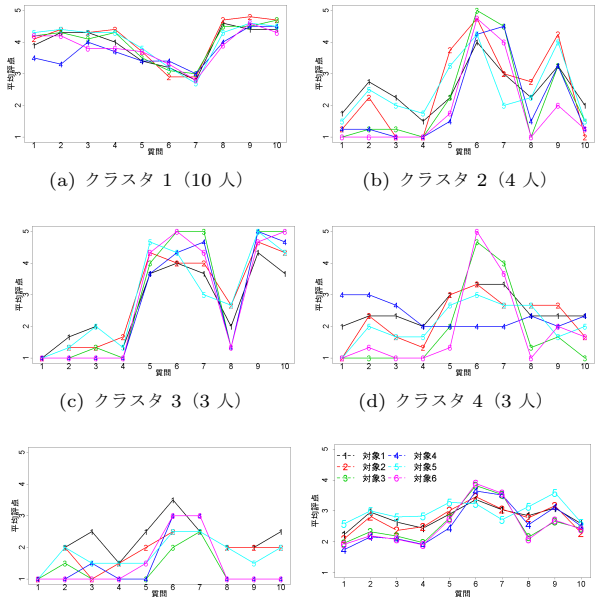
4.1 概要

実際の Web アンケートを用いて実験を行った．本調査では，1014 名の回答者に対して，6 つの次世代サービスについての説明文を評価対象とし，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．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．

4.2 結果と考察

提案手法により抽出されたクラスタ 1〜5 および全回答者の平均評点を図 2 に示す．各抽出において得られた σ^2 の値はそれぞれ 2.5,1.0,1.5,1.0,1.0 であった．

図 2(a) のクラスタ 1 は，図 2(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 2(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，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 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．



(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 2: 各クラスタの人数および平均評点

5. 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グループの抽出手法を提案した．実際の Web アンケートデータに適用し，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

[Pelleg 00] Pelleg, D., Moore, A. W., et al.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters., in *ICML*, pp. 727–734 (2000)

[Schwarz 78] Schwarz, G.: Estimating the dimension of a model, *The annals of statistics*, Vol. 6, No. 2, pp. 461–464 (1978)

[Shi 00] 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)

[君山 08] 君山 由良：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)

[木下 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, 感性工学研究論文集, Vol. 7, No. 3, pp. 449–460 (2008)