数理解析研究所講究録 1373 巻 2004 年 222-228

ランクつき投票モデルの多次元尺度法による類似度分析

大分大学 工学部 小畑 経史 (Tsuneshi OBATA)
Faculty of Engineering, Oita University
obata@csis.oita-u.ac.jp
大阪大学大学院 情報科学研究科 石井 博昭 (Hiroaki ISHII)
Graduate School of Information Science and Technology, Osaka University
ishii@ist.osaka-u.ac.jp

Abstract: 投票モデルにおいて単一投票にはいくつかの問題があることが知られており, そのため, それぞれの投票者が複数の票を持つような投票モデルが望ましい. ランクつき投票モデルは投票者が複数の候補者に順位をつけて投票するものである. このようなモデルから得られる得票データは通常おのおのの順位ごとに集計されたのち, 候補者の評価に使用される. これまでこのようなデータを分析し, 候補者を順位つける, あるいは, 候補者の中から最も望ましいものを決定する手法がいくつも提案されてきた. しかし, 票を集計する段階で, 特定の候補を支持した投票者が同時にどの候補を支持しているかという情報が失われてしまう. 同一の投票者によって高く評価された候補者どうしは, その投票者にとって「似ている」と見なされていると考えられる. したがって, 候補者のある組み合わせを高く支持する投票者が多ければ, 彼らが高い類似性を持っていると判断してもよいであろう. この仮説に基づき, 我々は集計前の得票データを多次元尺度法とクラスター分析を用いて分析することで, ランクつき投票モデルにおける候補者の類似性と配置を評価する手法を提案する. そして, シミュレーション実験を行い, この手法の妥当性を調べる.

Keywords: ランクつき投票モデル, 候補者の類似性, 多次元尺度法 (Multidimensional Scaling, MDS), クラスター分析, データ包絡分析法 (Data Envelopment Analysis, DEA)

1 はじめに

複数の人々の意見を集約し, 候補／選択肢から最も望ましいものを選択する, あるいはそれらを望ましい順に順位づけするために, しばしば投票という手段がとられる. その際に, おのおのの投票者が最も望ましい候補一人だけに票を投じる単一投票は, 必ずしも妥当とはいえないことが知られている [8]. そのため複数の候補に投票するモデルが望ましい. 特に, 投票者が複数の候補者を順位つきで投票するモデルをランクつき投票モデルと呼ぶ. このようなモデルにおいて, 投票結果から当選者を決定する, あるいは候補者の順位づけを行うには, ランクごとの得票数に何らかのウェイトをつけて集計したスコアにより各候補の選好度合いを数値化し, これを比較することが自然な方法である. しかし, ウェイトをあらかじめ決定することは何らかの恣意性が含まれる. そのため DEA (data envelopment analysis, データ包絡分析法) をベースにして, 各候補にとって有利なウェイトで評価することのできる手法が提案されてきた [2, 3, 7].

DEA においては, 類似したデータが存在するかどうかが, データの効率性評価に大きく影響する. ランクつき投票モデルにおいては, 観測データはランクごとに集計された得票数の組となる.

したがってこのデータが類似していることが必ずしも候補がその政策や特徴の面で似ていることを意味しない. 特に複数の候補が当選するケースでは当選者が政策や特徴の面で互いに似ているかどうかは, 投票者の意思を広く反映するかどうかに関わる重要なポイントといえる.

それでは, ランクつき投票モデルで得られるデータには, このような候補者間の類似性に関する情報は含まれないのであろうか. 先にあげた分析手法は投票データをランクごとに集計して使用する. 確かにこのように集計した後には類似性の情報は含まれない. しかし, 集計する前のデータはそうではない. ある候補者を支持する投票者が同時にどの候補を支持しているのか, という情報を含む. これはその投票者にとってはそれらの候補が何らかの意味で類似していることを示しているといえる. さらに, そのように考える投票者が多ければ, すなわち, ある候補のペアをともに支持する投票者が多ければ, そのペアが高い類似性を持つと見なすことができるであろう.

このような仮説に基づき, 我々は2節でランクつき投票データから候補者の空間的な配置を評価する手法を提案する. 各候補者は空間上の点として表現され, これらの点の近さがすなわち候補間の類似性を表す. 3節では提案手法の妥当性を確認するためのシミュレーション実験について述べる. この実験では候補者, 投票者がいくつかの政党の支持者に分かれている状況を想定し, 擬似的に生成した投票者によって得られた投票データにより, 元の候補者の配置がどの程度再現されるかを調査する.

2 候補者間の類似度

ここでは n 人の投票者 $\mathrm{V}_1, \ldots, \mathrm{V}_n$ が m 人の候補者 $\mathrm{C}_1, \ldots, \mathrm{C}_m$ の中から $k(\leq m)$ 人を選び, 順位つきで投票するモデルを考える. 投票者 V_l により j-位にランクづけられた候補のインデックスを i_{lj} と表す.

まず, $k = 2$ の場合, すなわち, 各投票者が上位二人を投票する場合を考える. このとき, 投票者 V_l は $\mathrm{C}_{i_{l1}}$ を1位, $\mathrm{C}_{i_{l2}}$ を2位として票を投じる. ここで, C_i を1位に C_j を2位にランクづけた投票者の人数を s_{ij} とおく, すなわち,

$$s_{ij} = \#\{\mathrm{V}_l | \mathrm{C}_i = \mathrm{C}_{i_{l1}} \text{ and } \mathrm{C}_j = \mathrm{C}_{i_{l2}}\}, \quad i, j = 1, \ldots, m,$$

ただし, 記号 # は集合の要素数を意味する. もし候補者 C_i と C_j が似ていれば多くの投票者が C_i と C_j をともに支持し s_{ij} が大きくなると考えられる. したがって s_{ij} が大きければ, C_i と C_j が何らかの類似性を持っていると判断してよいだろう.

また, $k > 2$ の場合には, s_{ij} を

$$s_{ij} = \sum_{q=1}^{k-1} s_{ij}^{(q)}, \quad i, j = 1, \ldots, m,$$

ただし

$$s_{ij}^{(q)} = \#\{\mathrm{V}_l | \mathrm{C}_i = \mathrm{C}_{i_{lq}} \text{ and } \mathrm{C}_j = \mathrm{C}_{i_{lq+1}}\}, \quad i, j = 1, \ldots, m; q = 1, \ldots, k-1,$$

と定める. これは C_i と C_j を隣り合った順位にランクづけた投票者の人数をカウントしたものである. これには投票者が好まないが似通った候補者の情報が含まれる反面, 似ていないが好まれないというだけでたまたま隣り合った順位にランクづけられた候補の情報も含まれる.

このような類似性の高さを表すデータから, 対象間の距離や空間的な配置を分析するための手法として, MDS (multidimensional scaling, 多次元尺度法) [5, 6, 9] が, また, 対象間の近さをもと

に対象をクラスタリングする手法としてクラスター分析 [4] がある. ランクつき投票データから候補者間の類似性を評価するにはこれらの手法が利用できる.

しかし, s_{ij} を並べてできる行列 $S = (s_{ij})$ をそのまま MDS の類似度行列として扱うことはできない. MDS を適用するために多少の手直しが必要である.

対称化: MDS は必ずしも対称な類似度データにしか適用できないわけではないが, 分析を簡単にするために次のような変換によりデータを対称化する.

$$\bar{s}_{ij} = s_{ij} + s_{ji}, \quad i, j = 1, \ldots, m.$$

これは, C_i と C_j をこの順でランクづけした投票者と, 逆順でランクづけした投票者とを区別せずに集計することに相当する.

基準化: もし C_i と C_j の支持者がそもそも少なければ, たとえ彼らが非常に似ていたとしても s_{ij} の値は大きくはならない. そのため支持者の数により次のような基準化を行う.

$$\bar{\bar{s}}_{ij} = \frac{\bar{s}_{ij}}{\bar{s}_{i+} + \bar{s}_{j+} - \bar{s}_{ij}}, \quad i, j = 1, \ldots, m,$$

ただし $\bar{s}_{i+} = \sum_k \bar{s}_{ik}$. この式の分母は, C_i と C_j のいずれか (あるいは両方) を支持する投票者の人数を意味する.

これらの変換を施した $\bar{\bar{s}}_{ij}$ をあらためて s_{ij} と書き表し, s_{ij} を成分に持つ行列 $S = (s_{ij})$ を類似度行列と呼ぶ.

こうして得られた類似度行列 S を MDS やクラスター分析で分析することにより, 候補者間の距離, 候補者の幾何的な配置の評価, クラスタリングを行うことができる.

[Proposal method]

Step 1 投票データを集計し類似度行列を作成する.

Step 2 類似度行列に対称化, 基準化を施す.

Step 3 MDS を用いて候補者の配置と距離を評価する.

Step 4 クラスター分析を用いて候補をクラスタリングする.

3 投票行動モデルと実験

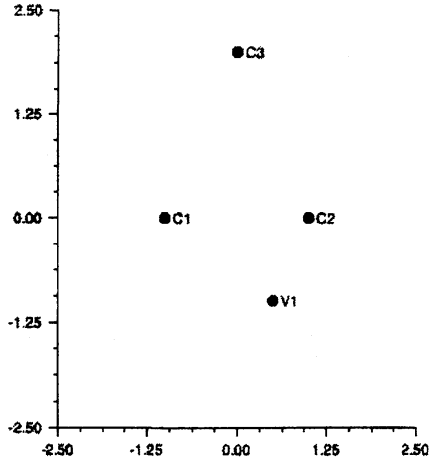
我々の提案手法の妥当性を確かめるためにシミュレーション実験を行う. 実験に先立ち, 投票者の行動をシミュレートするためのモデルを提案する. このモデルは Gill and Gainous [1] が提案した投票空間モデルに似たものである.

我々の提案するモデルでは投票者 (および候補者) は以下のように行動するものとする.

1. おのおのの候補者は r 次元ユークリッド空間上の点として配置される. このとき点の座標はその候補の性格や政策によって定まる.

2. おのおのの投票者は同じ空間上の点として配置される. このとき点の座標はその投票者の理想とする性格や政策によって定まる.

図 1: 候補者と投票者の配置の例



3. おのおのの投票者は自分が理想とする点に近い候補ほどより高い好感度を持つ.

4. おのおのの投票者は好感度の高い候補から順に候補者にランクをつけ, それにしたがってランクつき投票を行う.

たとえば, 3 人の候補者 C_1, C_2, C_3 と投票者 V_1 が図 1 のように 2 次元空間に配置されるとする. このとき V_1 は C_2, C_1, C_3 の順に候補をランクづけ, 投票する.

このモデルにしたがって以下のようなシミュレーション実験を行う. この実験では現実の状況を模すために, 次のように想定する.

1. 候補者と投票者が 2 次元空間に配置される (すなわち $r = 2$).

2. すべての候補者と投票者は 4 つのグループ (3 つの政党と無党派層) のいずれかに属する.

3. おのおののグループに属する候補者／投票者は 2 次元正規分布にしたがって分布する.

[Experiment]

Step 1 m 人の候補者を生成する.

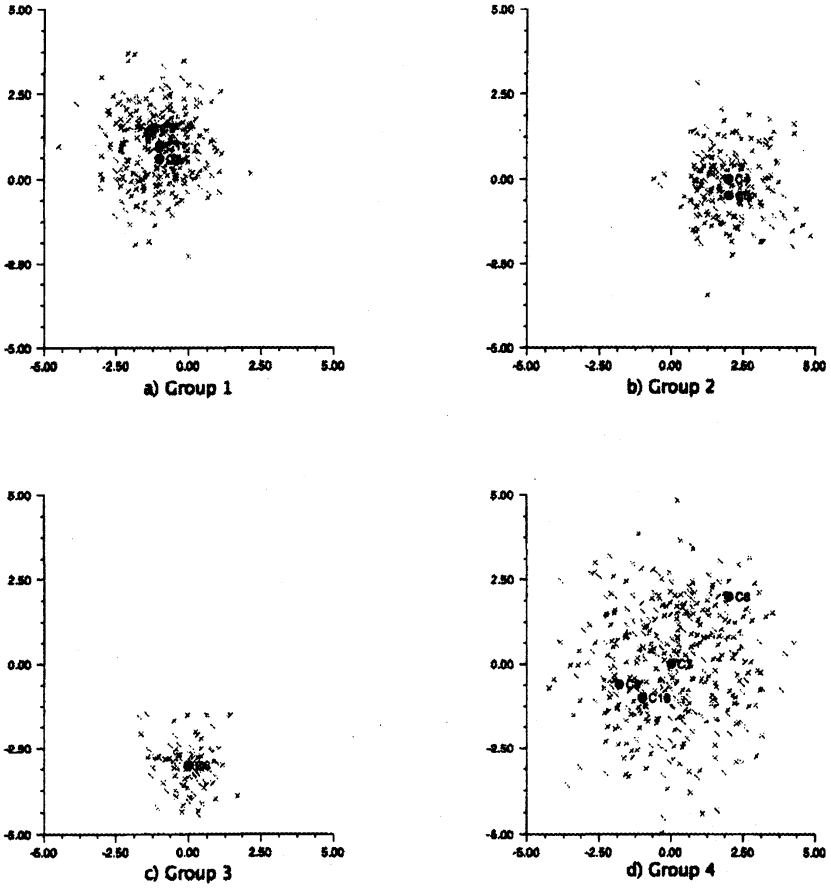
Step 2 $p = 1, 2, 3, 4$ のそれぞれについて n_p 人の投票者を 2 次元正規分布 $N(\mu^p, \Sigma^p)$ にしたがうランダムベクトルとして生成する, ただし n_p は $n = n_1 + n_2 + n_3 + n_4$ をみたす.

Step 3 おのおのの投票者から投票者までの距離を求める.

Step 4 おのおのの投票者が投票する上位 k 人の候補者のランクを決定し, 投票データを作成する.

Step 5 前節で述べた我々の提案手法を用いて投票データを分析し, 候補者の配置および候補者間の距離を評価する.

図 2: 候補者とランダムに生成された投票者



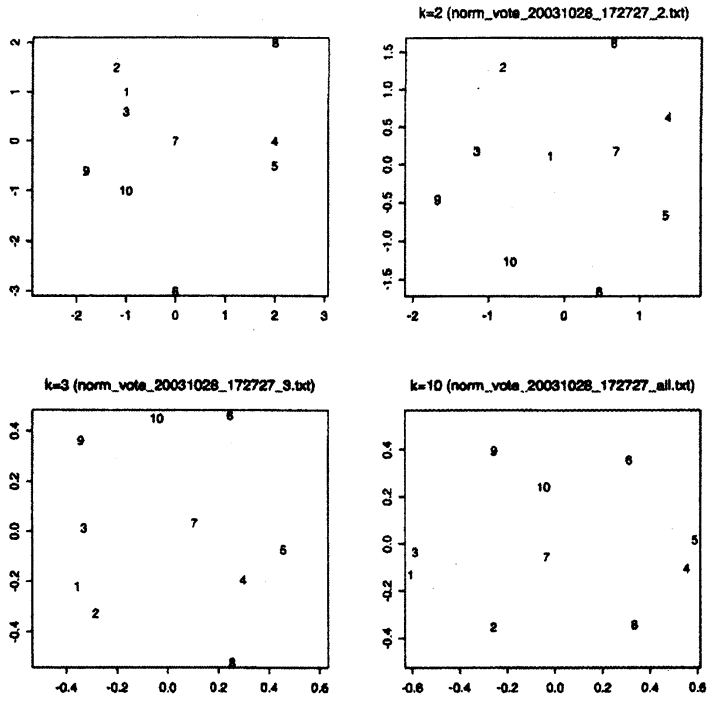
今回の実験では $m = 10$, $n = 1000$, $n_1 = 300$, $n_2 = 200$, $n_3 = 100$, $n_4 = 400$, $\mu^1 = (-1, 1)^{\mathrm{T}}$, $\Sigma^1 = \mathrm{diag}(1, 1)$, $\mu^2 = (2, 0)^{\mathrm{T}}$, $\Sigma^2 = \mathrm{diag}(1, 1)$, $\mu^3 = (0, -3)^{\mathrm{T}}$, $\Sigma^3 = \mathrm{diag}(0.5, 0.5)$, $\mu^4 = (0, 0)^{\mathrm{T}}$, $\Sigma^4 = \mathrm{diag}(3, 3)$ とした. グループ 1, 2, 3 を 3 つの政党, グループ 4 を全体に広く分布する無党派層と想定している. また, 10 人の候補者を各政党の中心付近に配置した. すなわち $\mathrm{C}_1 = (-1, 1)^{\mathrm{T}}$, $\mathrm{C}_2 = (-1.2, 1.5)^{\mathrm{T}}$, $\mathrm{C}_3 = (-1, 0.6)^{\mathrm{T}}$ がグループ 1 に; $\mathrm{C}_4 = (2, 0)^{\mathrm{T}}$, $\mathrm{C}_5 = (2, -0.5)^{\mathrm{T}}$ がグループ 2 に; $\mathrm{C}_6 = (0, -3)^{\mathrm{T}}$ がグループ 3 に属する候補; そして $\mathrm{C}_7 = (0, 0)^{\mathrm{T}}$, $\mathrm{C}_8 = (2, 2)^{\mathrm{T}}$, $\mathrm{C}_9 = (-1.8, -0.6)^{\mathrm{T}}$, $\mathrm{C}_{10} = (-1, -1)^{\mathrm{T}}$ が無所属の候補として 2 次元空間に配置した (図 2).

投票人数を $k = 2, 3, 10(= m)$ としたときの結果を図 3, 4 に示す.

図 3 が候補者の配置, 図 4 がそのときの候補者間の距離を用いてクラスター分析を行って得られたデンドログラムである. いずれの図も左上が候補者の真の配置／デンドログラム, 右上が $k = 2$ のときの分析結果, 左下が $k = 3$ のときの分析結果, そして右下が $k = 10$ のときの分析結果である.

候補者の配置に関してはいずれのケースでももともとの配置の特徴をおおまかにとらえることができている. 候補 C_7 が中心的な位置にあり, C_4 と C_5 が $\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3$ の反対側に, また C_8 が C_6,

図 3: 分析結果 (候補者の配置)



C_9, C_{10} の反対側に現れている.

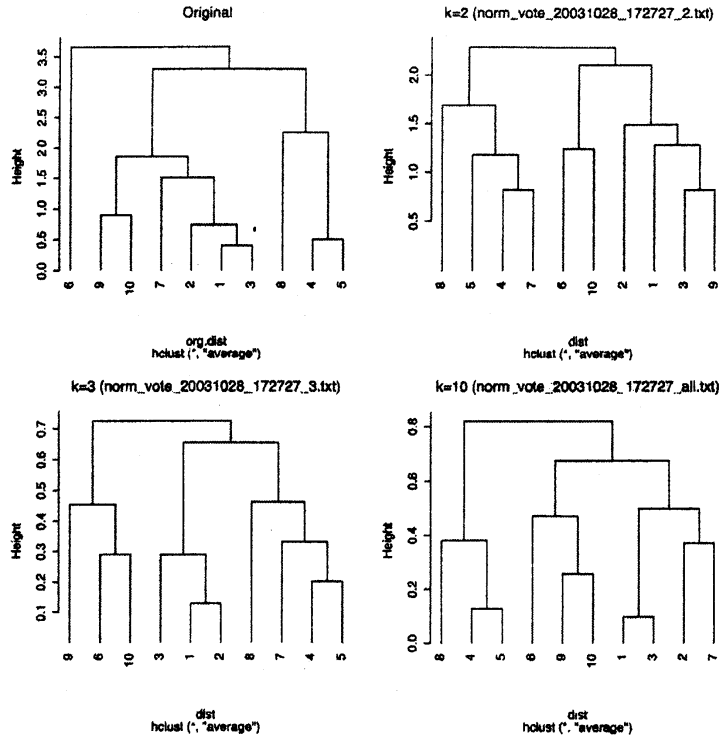
また, クラスター分析では同じ政党に属する候補同士が比較的早い段階で統合されており, まずまずうまくクラスタリングされているといえる.

4 おわりに

本稿で我々はランクつき投票モデルのもとで候補者の空間的な配置と類似性を評価する手法を提案した. 乱数を用いたシミュレーション実験では我々の手法により本来の候補者の特徴をおおまかにとらえることができることが示された. しかしながら, この実験はあくまでも擬似的な人工データに対するものであり, 実際の投票データに対しても有効かどうかを確認するためにより現実に近い状況での実験, すなわち何らかのテーマに関して実際にランクつき投票を行うことが必要であろう.

我々の最終的な目的は単に候補者間の類似性を評価することではなく, 適切な候補者の選択のために, 得られた類似性を用いることにある. 複数の候補が当選する選挙の場合, 一緒に当選する候補同士が類似性の高い候補かどうかは大きな意味を持つであろう. 一般に投票者の意思をより広く汲み上げるには類似性の低い候補が選ばれるほうがよいと考えられる. また逆に事業体の経営陣を選ぶようなケースではスムースな意思統一のために類似性の高いものが選ばれるほうがよいこともあるかもしれない. 今後は今回提案した手法で得られた類似性の評価結果を, ランクつき投票モデルに対する候補者順位決定手法と組み合わせ, 当選する候補の類似性の高低をコントロールする方法について検討したい.

図 4: 分析結果 (デンドログラム)



参考文献

[1] J. Gill and J. Gainous, Why does voting get so complicated? A review of theories for analyzing democratic participation, Statistical Science 17 (2002), 383–404.

[2] R. H. Green, J. R. Doyle and W. D. Cook, Preference voting and project ranking using DEA and cross-evaluation, European Journal of Operational Research 90 (1996), 461–472.

[3] A. Hashimoto, A ranked voting system using a DEA/AR exclusion model: A note, European Journal of Operational Research 97 (1997), 600–604.

[4] 河口至商, 多変量解析入門II, 森北出版, 1978.

[5] J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964), 1–27.

[6] J. B. Kruskal, Nonmetric multidimensional scaling: A numerical method, Psychometrika 29 (1964), 115–129.

[7] T. Obata and H. Ishii, A method for discriminating efficient candidates with ranked voting data, European Journal of Operational Research 151 (2003), 233–237.

[8] 佐伯胖, 「きめ方」の論理, 東京大学出版会, 1980.

[9] 齋藤堯幸, 多次元尺度構成法, 朝倉書店, 1980.