順序のクラスタリング

神嶌 敏弘（産業技術総合研究所）

k-o'means：順序をクラスタリングする新手法

※ 順序：何らか基準の順に対象を整列したもの

例：Aさんが好き[基準]な寿司[対象]を整列

とろ ≻ いか ≻ かっぱ巻

"いか"より"とろ"が好き

しかし，どれくらい好きかは不明

実験

- 人工データ: 本手法の従来手法に対する利点
- 嗜好調査データ: 探索的解析ツールとしての適用例

なぜ順序を使うのか？：SD法の問題点

官能検査(味覚など)や主観的判断の計量の方法

一般的な手法

- SD (Semantic Differential) 法 [Osgood 57]

  例：食べ物の嗜好に関する調査で，被験者に「好き」から「嫌い」までを，5段階に区切って評価してもらう

欠点

- 全体の長さが等しく，等間隔に分割された尺度を，被験者が共有しているという非現実的仮定
- 必ず対義語が必要

なぜ順序を使うのか？：順位法の利点

順位法

・被験者に「好き」な方から順にアイテムを並べてもらう

利点

・「好き」や「嫌い」に関する定量的尺度を被験者全員が共有するという非現実的仮定は不要

・「嫌い」といった対義語が不要

・採用した対義語で結果が違う：きれい ⇔ 醜い or 汚い?

・対義語を決めにくい：エレガント ⇔ ワイルド？ ポップ？

問題点：解析手法が少ない ➡ 研究課題

順序のクラスタリング

対象 $x^i \in X^*$ ：整列される個体 (X^*:対象の全集合)

サンプル順序 $O_i = x^1 \succ \cdots \succ x^{|X_i|}$:

対象の部分集合 $X_i \subseteq X^*$ 中の対象を何らかの基準 (例: 嗜好，価格) で整列したもの

サンプル集合 $S = \{O_1, \cdots, O_{|S|}\}$

順序のクラスタリングの目的

S 中の順序をクラスタ$C_1 \cdots C_{|\pi|} \subseteq S$に分割

・同一クラスタ内の順序は類似し，違うクラスタでは類似していない

k-means アルゴリズム

- 数値ベクトルをクラスタリングする代表的手法
- 初期分割に対し，次の2ステップを反復的に適用
  - クラスタ中のベクトルの中心を計算
  - 最も類似した中心へ，各ベクトルを再分類

k-o'means アルゴリズム

- 中心と類似度の概念を順序に適合するよう修正
  - 類似度 ➡ Spearmanの ρ に基づく類似度
  - 中心 ➡ 順序集合を代表する順序平均

順序の非類似度

Spearmanの順位相関 ρ：対象の順位の相関係数

値域は $[-1,1]$ で，1 なら一致，-1 なら逆順

- 順位の例：順序 $O_1 = x^1 \succ x^3 \succ x^2$ で対象 x^2 の順位は **3**
- O_1 と $O_2 = x^1 \succ x^2 \succ x^3$ の対象 $[x^1, x^2, x^3]$ の順位は，それぞれ $[1,3,2]$ と $[1,2,3]$．これを数値ベクトルと見なしてPearsonの相関を計算

問題点：二つの順序に含まれている対象の集合が一致していなくては計算できない ➡ 共通する対象だけを取り出して計算

順序平均

k-means の中心

クラスタ中のベクトルと中心ベクトルの類似度の総和が最大

k-o'means の中心も，この性質を備えるようにする

$$\text{順序平均 } \overline{O} = \arg\min_{O_i} \sum_{O_j} \left(1 - \rho\left(O_i, O_j\right)\right)$$

最適な $\overline{O}$ を求めるのは離散最適化で困難

➡ Thurstoneの一対比較法を利用した手法で代用

Thurstoneの一対比較を用いた手法

$\Pr[x^a \succ x^b]$の推定

対象集合の順序をアイテム対の順序に分解して

例：$x^1 \succ x^2 \succ x^3$ ➡ $x^1 \succ x^2$，$x^1 \succ x^3$，$x^2 \succ x^3$

これらの対の頻度から確率を推定

Thurstoneの一対比較の法則を用いて整列

$$\mu_a = \sum_{x^b \in C} \Phi^{-1}\left(\Pr\left[x^a \succ x^b\right]\right)$$

※ Φ は正規分布の分布関数

対象$x^1 \cdots x^{|C|}$を，そのμ_aの値で整列 ➡ 順序平均

人工データでの実験（内容）

k-o'means の性能を人工データで検証

- 類似度にSpearmanの ρ を用いた階層的クラスタリング手法と復元性能を比較
- 4種類のパラメータを変えた人工データを生成
  1. クラスタ数
  2. クラスタ間の分離性
  3. クラスタの大きさの均一性
  4. クラスタ内のまとまり

人工データでの実験（結果）

クラスタ数とクラスタ内のまとまりの影響が大きかったのでそれらについての結果のみ示す

数値は情報損失量(RIL)で，小さいほどよく，0なら完全に復元できた

赤字：*k*-o'means　　青字：群平均法(階層的手法の一つ)

クラスタ内まとまり	少 ← クラスタ数 → 多			
強 ↑	0.001 0.479	0.016 0.651	0.101 0.880	0.998 0.995
	0.606 0.919	0.776 0.991	0.992 0.999	1.000 0.999
↓ 弱	0.976 0.991	0.999 1.000	1.000 1.000	1.000 0.999

k-o'meansの復元性能は，従来手法より非常に良い

嗜好調査データの解析（内容）

WWW上の寿司の嗜好調査

- 寿司は全部で100種
- 各被験者ごとに，10種の寿司を，その提供頻度に応じてランダムに選択
- それら寿司を，好きなものから順に並べさせる
- サンプル数 1025

嗜好調査データの解析（結果1）

k-o'means法によって，被験者を2クラスタに分類

各クラスタの被験者の嗜好と寿司の属性との相関係数から，嗜好に影響する属性を検出

	C1		C2
被験者数	628		397
あっさり味よりこってり味が好き	+0.396	$>$	−0.152
いつもは食べない寿司が好き	−0.648	$\fallingdotseq$	−0.577
安い値段の寿司が好き	−0.472	$<$	−0.008
店にあまり無い寿司が好き	−0.441	$\fallingdotseq$	−0.250

クラスタC1の被験者はこってり味の高価な寿司が好き

本文の訂正：4章の最後から4行目「定番の寿司」→「定番ではない寿司」

嗜好調査データの解析（結果2）

クラスタリングの前後で，順序平均中で順位変動が大きな寿司を抽出

C1	好き	コハダ +33	アンキモ +31	ヒモキュウ巻 +27	カモ +27	トリガイ +23
		カニミソ +20	サバ +19	ウニクラゲ +19	バサシ +18	ウニ +15
	嫌い	カズノコ -41	イナリ -33	ナットウ巻 -30	タマゴ -28	ウメ巻 -21
		サザエ -19	アカガイ -19	メンタイコ巻 -18	サラダ巻 -18	ホッキガイ -17
C2	好き	カッパ巻 +61	ナス +56	タクワン巻 +55	ウメシソ巻 +52	カンピョウ巻 +48
		カイワレ +44	イカナットウ +43	カキ +36	カラスミ +35	ホヤ +35
	嫌い	ウニ -77	アナゴ -73	ハマチ -72	アジ -59	シマアジ -52
		カンパチ -47	サバ -47	ウナギ -47	カニミソ -40	カツオ -35

赤字：好き(順位が上がった)　　青字：嫌い(順位が下がった)

数字は，100種の寿司の中で上下した順位の差を表す

まとめと今後の課題

まとめ

- 順序をクラスタリングするk-o'means法を提案
  - 従来手法よりクラスタの復元性能が優れる
  - 嗜好調査データに適用した

今後の課題

- 計算量が，サンプル数とクラスタ数に対してはk-means法同じ線形のオーダだが，対象の種類の総数に対しては2乗のオーダ．この計算量を減らしたい.