順序のクラスタリング
～ 順序平均の最適性について ～

神嶌敏弘 藤木淳
産業技術総合研究所

開始

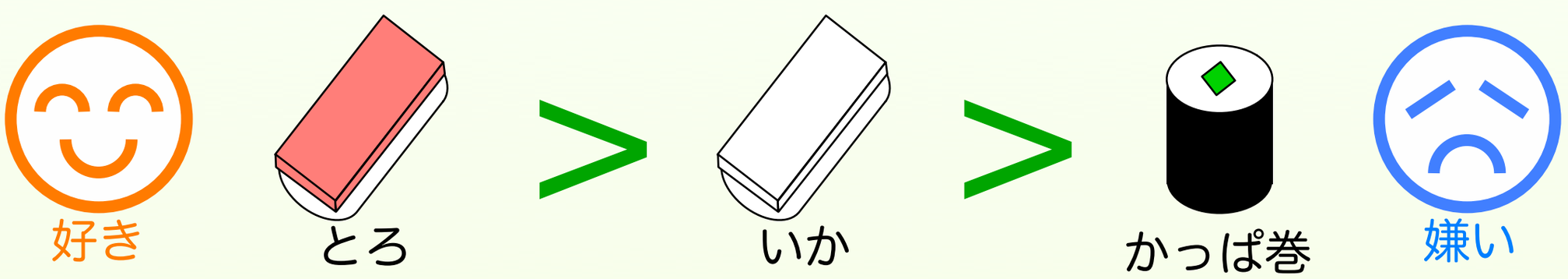
概要

順序をクラスタリングする k-o’means法 の提案

性能評価のための追加実験を行った

順序：何らかの基準で対象を整列したもの

例：Aさんが好き[基準]で寿司[対象]を整列



“いか” より “とろ” が好き

しかし，どれくらい好きかは不明

順序は主観的な変量の計測に適している

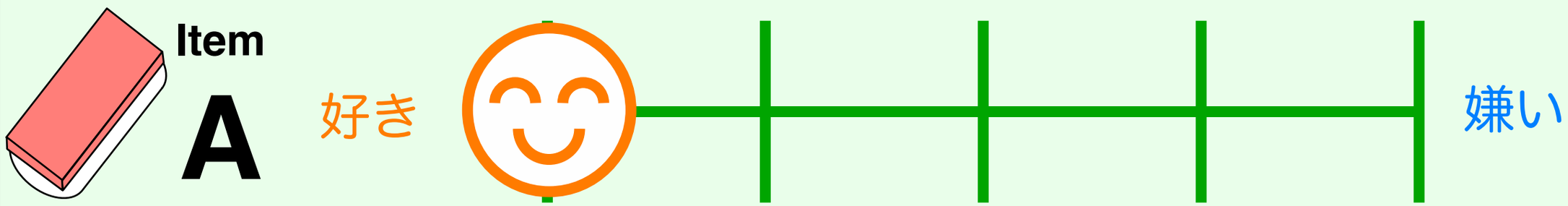
この順序をクラスタリングにより解析する

順序による計測

Semantic Differential 法 (SD法) [Osgood 57]

両端を対義語で表した尺度を用いる

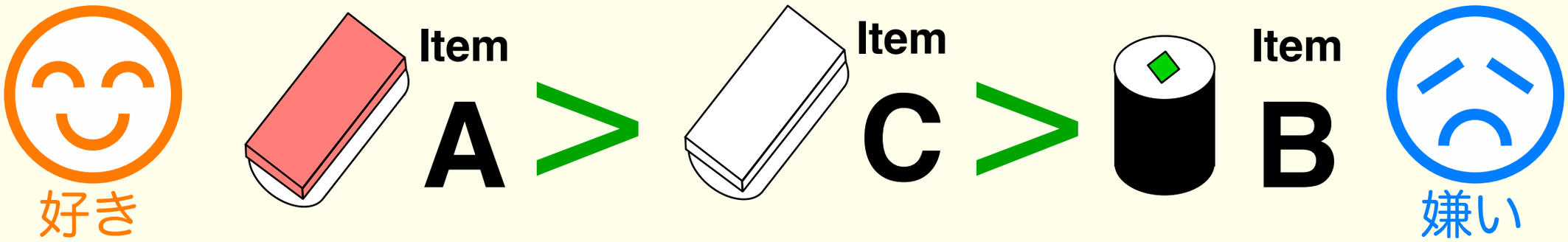
例：被験者が Item A を好きならば，尺度の「好き」を選ぶ



順位法

計測する度合いの強さの順に対象を整列

例：被験者は Item A が最も好きで，Item B が最も嫌い



SD法の問題点(1)

1．SD法の非現実的な仮定

- 全被験者は尺度の絶対値についての理解を共有

  二人の被験者AとBが共に「非常に好き」としていても，一般には，その好きな度合いは二人で共有されていない

  順位法では相対的にしか評価しないので，絶対値は無関係

- 尺度の目盛りは等間隔

  被験者が好きさの度合いを等間隔に分割するのは困難

  順位法では間隔は無視し，相対的な大小関係だけに注目

2．心理的な攪乱要因

例：Central Tendency Effect

尺度の中心に近い部分を使って回答する傾向

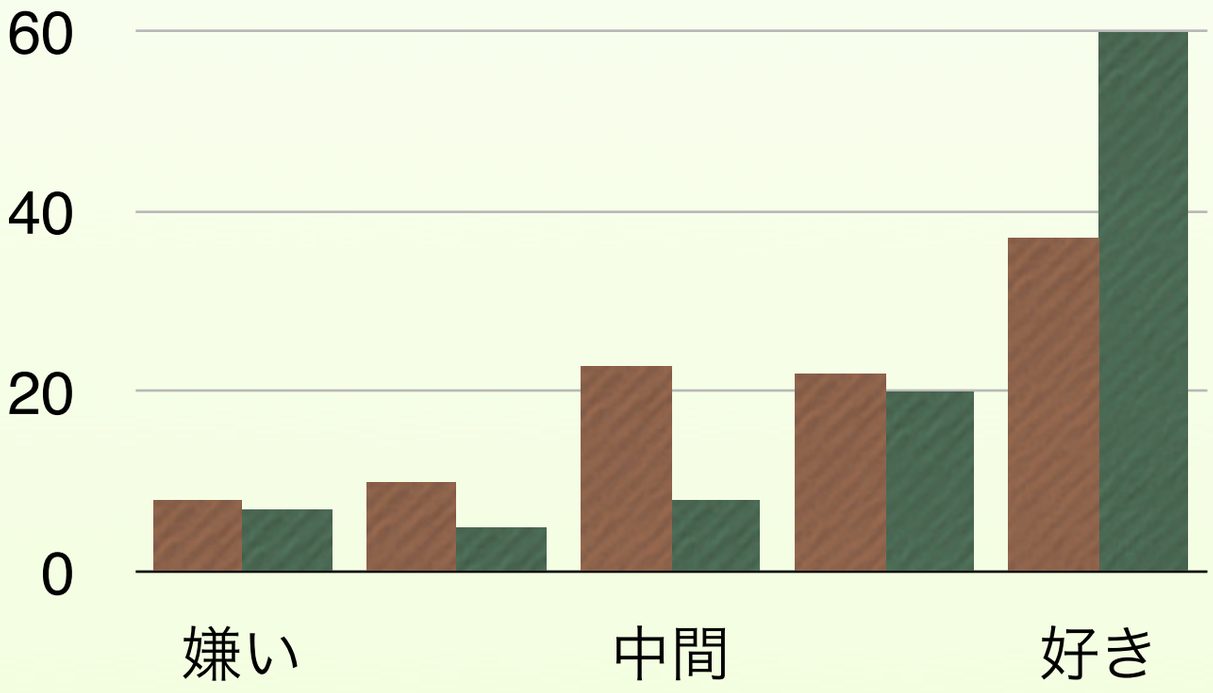
SD法の問題点(2)

3. 対義語の選択

対義語を決めにくい：エレガント ⇔ ポップ？ ワイルド？
対義語によって結果が変わる：きれい ⇔ 醜い or 汚い

5段階のSD法で，被験者が選択した値の分布の例

仮定と異なったかなり歪んだ分布



※ Amazonの書評のデータは
KDD2003のA.S.Weigendの
招待講演より引用

順序のクラスタリング

対象 $x^i \in X^*$: 整列される個体

対象の全集合 X^*

サンプル順序 $O_i = x^1 \succ \cdots \succ x^{|X_i|}$:
X^*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 $S = \{O_1, \ldots, O_{|S|}\}$

順序のクラスタリングの目的

S 中のサンプル順序をクラスタ$C_1 \cdots C_{|\pi|} \subset S$に分割

同一クラスタ内では順序は類似し，違うクラスタでは異なる

k-o'meansアルゴリズム

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，次の二つのステップを反復適用
  1. クラスタ中のベクトルの**中心**を計算
  2. 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に修正
  - **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：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 $(1, 3, 2)$

  これを数値ベクトルとみなしたPearson相関係数
- 二つの順序の対象集合が違うときは，共通する対象だけをとりだす

  例：$x^3 \succ x^4 \succ x^2$ と $x^1 \succ x^2 \succ x^3$ なら x^2 と x^3

順序平均 (概念)

k-means法での中心

クラスタ中のベクトルと中心ベクトルの非類似度の総和が最小

k-o'meansでの中心

k-meansと同様に次式の最小化が目標

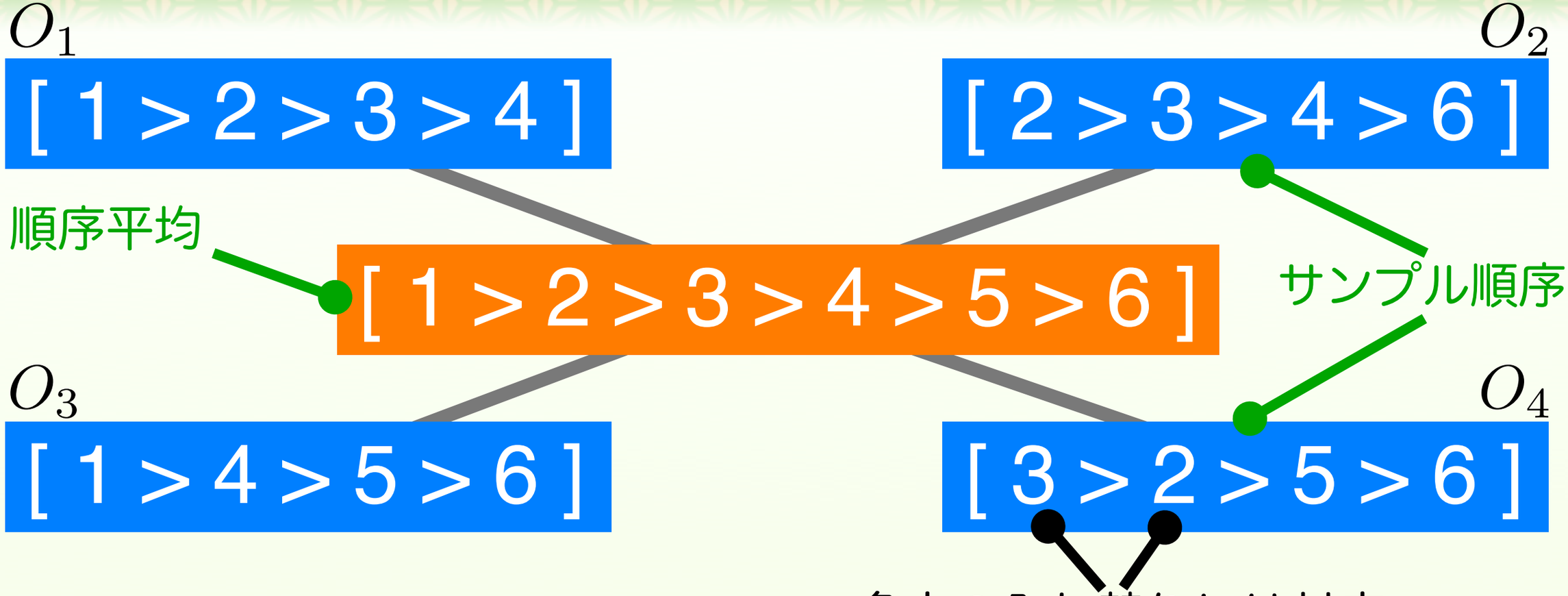
順序平均

$$\bar{O} = \arg\min_{O_i} \sum_{O_j} \left(1 - \rho(O_i, O_j)\right)$$

最適な順序平均の導出は離散最適化なので困難

➡ Thurstoneの一対比較法を用いた近似解法

順序平均 (例と長所)



※ 順序平均はサンプル順序中の全ての対象を含む

	共通対象が	類似度が
サンプル順序ー順序平均	多い	正確
サンプル順序ーサンプル順序	少ない	不正確

Thurstoneの一対比較法の利用

$\Pr[x^a \succ x^b]$の推定

対象集合の順序を対に分解する

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の一対比較法を用いて対象を整列

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

※Φは正規分布の分布関数

対象 $x^1 \ldots x^{|C|}$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)

クラスタの復元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化させた人工データ
  1. クラスタ数
  2. クラスタ間の近さ
  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）

k-o’meansの復元性能は，従来手法より非常によい

クラスタ内のまとまり	少	クラスタ数		多
弱	0.001 0.484	0.013 0.643	0.099 0.868	0.998 0.995
	0.597 0.947	0.783 0.990	0.993 0.999	1.000 0.999
強	0.977 0.996	0.999 1.000	1.000 1.000	1.000 0.999

数値は適切なクラスタへの近さ(RIL)で，小さいほど良い

赤字：*k*-o’means　　青字：群平均法(階層的手法の一つ)

嗜好調査データの解析（内容）

WWWでの寿司の嗜好調査データ

- 寿司は全部で100種
- 被験者ごとに，10種の寿司をランダムに抽出
- 被験者は寿司を，好きなものから順に並べる
- 総被験者数は1025

類似した嗜好をもつ被験者のクラスタ抽出して，各クラスタの特徴を解析

順位の入力フォーム

寿司の名前　　順位を指定

嗜好調査データの解析（結果1）

k-o’means法を用いて，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の順位相関を求めて，嗜好に影響する属性を検出

	C1	C2
被験者数	607	418
あっさり味よりこってり味が好き	+0.402 ≻	-0.135
いつもは食べない寿司が好き	-0.643 ≈	-0.601
値段が安い寿司が好き	-0.465 ≺	-0.046
店であまり売られていない寿司が好き	-0.449 ≈	-0.253

クラスタC1の被験者はこってり味の高価な寿司が好き

嗜好調査データの解析 (結果2)

クラスタリングの前後で，順序平均中の順位変動が大きかった寿司

C1	好き	コハダ +35	カモ +31	アンキモ +30	ヒモキュウ巻 +26	トリガイ +24
		カニミソ +20	サバ +19	ギュウサシ +17	ウニ +16	キャビア +16
	嫌い	カズノコ -44	イナリ -36	アカガイ -30	タマゴ -28	サラダ巻 -27
		メンタイコ -25	ホッキガイ -23	ウメ巻 -21	ウメシソ巻 -19	ナットウ -18
C2	好き	ナス +57	ウメシソ巻 +54	タクワン巻 +54	カイワレ +46	カキ +44
		イカナットウ +43	カッパ巻 +42	カンピョウ巻 +38	ホヤ +37	カズノコ +33
	嫌い	ウニ -78	アナゴ -72	アジ -59	ハマチ -55	シマアジ -52
		ウナギ -50	サバ -47	カニミソ -38	サンマ -38	カツオ -36

赤字：好き (順位があがった)　　青字：嫌い (順位が下がった)

数字は，クラスタリング前の順位から，後の順位を引いた値

順序平均の最適性

離散最適化なので，最適な順序平均は一般に困難

➡ Thurstoneの方法で近似

※ 非反復解法で実験的に性能が良かったので採用

近似によるクラスタの復元性能の低下の程度は？

サンプル順序が部分順序の場合では困難だが，全て

完全順序ならば最適な順序平均は計算可能

完全順序：サンプル順序で，全ての事例が順序付けられている i.e. $X_i = X^*, \forall O_i \in S$

部分順序：サンプル順序に含まれない対象がある

最適 k-o'means と k-o'means の比較

完全順序の場合に，最適な順序平均を求める方法

各対象のサンプル順序中での平均順位を求めてソート

例:

順序		順位ベクトル		平均順位		順序平均
		$x^1 x^2 x^3$				
[1 > 2 > 3]		(1, 2, 3)		$x^1 \quad x^2 \quad x^3$		
[1 > 2 > 3]	→	(1, 2, 3)	→	(1.0, 2.3, 2.6)	→	[1 > 2 > 3]
[1 > 3 > 2]		(1, 3, 2)				

人工データに対する適用結果 (RIL)

最適 k-o'means 上記の方法で順序平均を求める	k-o'means Thurstoneの方法で順序平均を求める
0.354	0.362

Thurstoneの近似は統計的に優位に悪い

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
  - 嗜好調査データにより解析手法の有効性を確認
- 順序平均の最適性はあまりよくない

今後の課題

- 対象の種類の総数に対する2乗の計算量を減らす
- 完全順序で最適になり，かつ，部分順序にも適用可能な順序平均の推定手法