3F1-01

順序のクラスタリング

Clustering Orders

神嶌 敏弘*1
Toshihiro Kamishima

*1産業技術総合研究所
National Institue of Advanced Industrial Science and Technology (AIST)

We propose a method of using clustering techniques for partitioning a set of orders, which are sequences of objects sorted according to some property, such as size, preference, or price. These orders are useful for, say, carrying out a sensory survey. We propose a method called the k-order means (k-o'means) method. We compared our method with the traditional clustering methods, and analyzed its characteristics. We also applied our method to questionnaire survey data on people's preferences in types of *sushi*.

1. はじめに

クラスタリングとは，内的結合と外的分離が達成されるようにデータ集合を分割する手法で，重要なデータ解析手段の一つである [神嶌 03]．多くのクラスタリング手法は，属性ベクトルや類似度行列で記述された対象しか扱えないが，本論文では順序で記述されたデータを扱う手法を提案する．

ここでいう順序とは，何らかの特徴に従って整列された対象の系列である．三つの対象 x^1，x^2，及び x^3 があるとき，ある人がこれらの対象を好きなものから順に並べた $x^3 \succ x^1 \succ x^2$ は順序の一例である．この順序を解析する手法は，アンケート調査などの主観的なデータの解析に有効である．例えば，幾つかの食べ物を被験者に示し，それらを被験者が好きな順番に並べてもらう．複数の被験者に同様の質問をして集めた順序をクラスタリングすることにより，類似した嗜好を持つ被験者のクラスタを発見できる．従来，この種の調査には，Semantic Differential（SD）法が用いられてきた [中森 00]．この方法では，被験者の嗜好は次のような，両端を対義語で表した物差しによって計測される．

好き　5 4 3 2 1　嫌い

この SD 手法では，解析手法の制限から，被験者が想定している物差しの両端や間隔が，全ての被験者間で共有されているという非現実的な仮定がなされている．このような仮定は，絶対的な物差しを用いる代りに，各被験者の相対的な嗜好の度合いを順序を用いて獲得することで回避できる．しかし，順序を扱うクラスタリング手法は開発されていないので，本論文ではこのための k-o'means 法を新たに提案する．

2. 節では順序のクラスタリング問題の定式化，3. 節では提案手法，4. 節では実験結果，5. 節ではまとめについて述べる．

2. 順序のクラスタリング

本節では順序のクラスタリング問題の定式化を行う．順序とは，大きさ，嗜好の度合い，価格といった何らかの特性に従って対象を整列したものである．対象 x^a とは整列される個体であり，対象全集合 X^* とは全ての対象を含む集合である．順序は $O{=}x^1 \succ x^2 \succ \cdots \succ x^3$ のように記し，$x^1 \succ x^2$ を「x^1 は x^2 より前にある」と言い表す．同一の順序内では推移律が成立する．すなわち，$x^1 \succ x^2$ かつ $x^2 \succ x^3$ ならば $x^1 \succ x^3$ である．$X_i \subseteq X^*$ は順序 O_i に現れる全ての対象を含む対象集合を表す．集合 A の大きさを $|A|$ で表すと，$|X_i|$ は順序 O_i の長さと一致する．

順序のクラスタリング問題を以下に述べる．入力として対象集合 $S = \{O_1, O_2, \ldots, O_{|S|}\}$ が与えられ，この集合中の順序をサンプル順序と呼ぶ．ここで，$X_i \neq X_j$ $(i \neq j)$ であったり，順序 O_i では $x^1 \succ x^2$ だが順序 O_j では $x^2 \succ x^1$ であってもよい．クラスタリングの目的は，分割 $\pi = \{C_1, C_2, \ldots, C_{|\pi|}\}$ に S を分けることである．ただし，C_j をクラスタといい，網羅的で互いに素であるものとする．すなわち，$C_i \cap C_j = \emptyset, \forall i, j, i \neq j$ かつ $S = C_1 \cup C_2 \cup \cdots \cup C_{|\pi|}$．分割は，同じクラスタ内では似ていて（内的結合），違うクラスタでは似ていない（外的分離）ように生成される．

3. k-o'mean 法

代表的なクラスタリング手法である k-means 法を，順序を扱えるように修正した k-o'means 法について述べる．

3.1 順序間の類似度

二つの順序の類似性を測るために Spearman の順位相関係数 ρ [Kendall 90] を用いた．ρ とは対象の順位の相関である．順位 $r(O, x)$ は，順序 O 中で対象 x が現れる先頭からの位置を示す基数である．例えば，順序 $O{=}x^1 \succ x^3 \succ x^2$ では，$r(O, x^1){=}1$ や $r(O, x^2){=}3$ である．二つの順序 O_1 と O_2 が，同じ対象集合で構成される（$X_1{=}X_2$）とき，ρ は次式で定義される．

$$\rho = \frac{\sum_{x \in X_1} \big(r(O_1, x) - \bar{r}_1\big)\big(r(O_2, x) - \bar{r}_2\big)}{\sqrt{\sum_{x \in X_1} \big(r(O_1, x) - \bar{r}_1\big)^2} \sqrt{\sum_{x \in X_1} \big(r(O_2, x) - \bar{r}_2\big)^2}}$$

ただし，$\bar{r}_i = (1/|X_1|) \sum_{x \in X_1} r(O_i, x)$．また，同順位の対象が無い場合には，次式で簡単に計算できる．

$$\rho = 1 - \frac{6 \times \sum_{x \in X_1} \big(r(O_1, x) - r(O_2, x)\big)^2}{|X_1|^3 - |X_1|}$$

ρ は二つの順序が一致するときには 1 に，互いに逆順であれば -1 になる．二つの順序が同じ対象集合で構成されていない場合は，共通の対象だけを抽出したあと，もとの前後関係を保存するように新たな順序中での順位をこれらの対象に再び与えたあと ρ を求めるものとする．共通の対象が無い場合は無相関，すなわち，$\rho = 0$ として扱う．

連絡先: 神嶌 敏弘，E-mail: mail@kamishima.net,
Homepage: http://www.kamishima.net/

Spearman の ρ は [Mooney 99] などで順位付けの評価などに用いられ，さらに，ランダムな二つの順序の間について，$\rho\sqrt{(|X|-2)/(1-\rho^2)}$ が自由度 $|X|-2$ の t 分布に従うという便利な特性も備えているので，この尺度を採用した．他に，順位の類似性の尺度として代表的な Kendall の τ もあるが，ρ が $O(|X|)$ の計算量であるのに対し，τ は $O(|X|^2)$ であり，実用的には顕著な差がないので，これを採用しなかった．

クラスタリングでは，類似度よりも非類似度を用いることが多いので，非類似度を次式で定義しておく．

$$d(O_1, O_2) = 1 - \rho \tag{1}$$

ρ の範囲は $[-1, 1]$ なので，この非類似度の範囲は $[0, 2]$ になる．

3.2 順序平均

次に，順序平均（order mean）について述べる．k-means ではクラスタの中心を，クラスタ中の対象からの非類似度の総和を最小にする点に設定する．この概念を順序に適合するように拡張する．すなわち，式 (1) を損失関数に用いて，クラスタ C の順序平均 $\bar{O}$ を次式で定める．

$$\bar{O} = \arg\min_{O_j} \sum_{O_i \in C} d(O_i, O_j) \tag{2}$$

この順序平均は，クラスタ C 中のいずれかの順序に含まれる全ての対象で構成される順序になる．よって，$\bar{X} = \cup_{O_i \in C} X_i$.

残念ながら，この順序平均を求める問題は離散最適化なので困難である．そこで，既存の順序の統合手法を人工データに適用し，式 (2) のエラーを小さくする手法を実験的に探した．その結果，次の Thurstone の一対比較法を採用した．

Thustone の比較判断の法則 (Thurstone's law of comparative judgment) の case V [Thurstone 27] とは順序の生成モデルである．このモデルでは，各対象にスコアを割り当て，このスコアの順に対象を整列することで順序が生成される．スコアは，各対象ごとに異なる平均 μ_a と全対象で共通の σ をパラメータとする正規分布に従う．このとき，対象 x^a が x^b より前になる確率は次式で表される．

$$\begin{aligned}\Pr[x^a \succ x^b] &= \int_{-\infty}^{\infty} \phi(\frac{t-\mu_a}{\sigma}) \int_{-\infty}^{t} \phi(\frac{u-\mu_b}{\sigma}) du\, dt \\ &= \Phi\left(\frac{\mu_a - \mu_b}{\sqrt{2}\sigma}\right)\end{aligned} \tag{3}$$

ただし，$\phi(\cdot)$ と $\Phi(\cdot)$ は正規分布の密度関数と分布関数である．μ_a に任意の単調変換を適用しても得られる順序は不変なので，$\sqrt{2}\sigma$ で割って，原点を μ_a の平均にする変換をした $\bar{\mu}_x$ を考えると，$\bar{\mu}_a - \bar{\mu}_b$ は分散 1 で平均 0 の正規分布に従う．このことを用いて，次の 2 乗残差の最小化によって $\bar{\mu}_x$ を推定する．

$$\sum_{x^a \in \bar{X}} \sum_{x^b \in \bar{X}} \left(\Phi^{-1}\big(\Pr[x^a \succ x^b]\big) - (\bar{\mu}_a - \bar{\mu}_b) \right)^2$$

この残差は次式で最小になり，得られた $\bar{\mu}_x$ で対象を整列すれば統合された順序，すなわち，順序平均の近似が得られる．

$$\bar{\mu}_a = \frac{1}{|\bar{X}|} \sum_{x^b \in \bar{X}} \Phi^{-1}\big(\Pr[x^a \succ x^b]\big) \tag{4}$$

あとは，クラスタ C 中の順序から $\Pr[x^a \succ x^b]$ を求める方法があれば $\bar{\mu}_a$ が計算できるが，この方法について述べる．このクラスタ中の順序 $O \in C$ について，この順序の中で x^a が x^b より前にあるような対象の対 (x^a, x^b) を全て抽出する．例えば，$O = x^3 \succ x^1 \succ x^2$ からは，対象の対 (x^3, x^1)，(x^3, x^2)，及び (x^1, x^2) を抽出する．これらの対をクラスタ中の $|C|$ 個の全ての順序から抽出し，それらを集めて集合 P_C を生成する．確率 $\Pr[x^a \succ x^b]$ の推定量には，0 にならないようにするため Dirichlet 分布を事前分布に用いた次式を用いた．

$$\Pr[x^a \succ x^b] = \frac{|x^a, x^b| + 0.5}{|x^a, x^b| + |x^b, x^a| + 1}$$

ただし，$|x^a, x^b|$ は P_C 中での対象の対 (x^a, x^b) の数．

アルゴリズム k**-o'means(**S**,** k**,** $maxIter$**)**
$S = \{O_1, \ldots, O_{|S|}\}$: 順序の集合
k: クラスタの数
$maxIter$: 反復回数の上限
1) 初期分割: S をランダムに分割 $\pi = \{C_1, \ldots, C_k\}$
　$\pi' := \pi$, $t := 0$.
2) $t := t + 1$, もし $t > maxIter$ ならステップ 6 へ
3) 各クラスタ $C_j \in \pi$ について
　順序平均 $\bar{O}_j$ を 3.2 節の方法で求める
4) S 中の各順序 O_i を次のクラスタに割り当て:
　$\arg\min_{C_j} d(\bar{O}_j, O_i)$.
5) もし $\pi = \pi'$ ならば ステップ 6 へ
　でなければ $\pi' := \pi$, ステップ 2 へ
6) π を出力

図 1: k-o'means 法

3.3 k-o'means 法

k-o'means 法は，クラスタの中心に順序平均を，非類似度に式 (1) を用いること以外は k-means 法と同じである．アルゴリズムを図 1 に示す．最初に，S をランダムに分割して初期分割を得る．クラスタの順序平均の再計算と，順序のクラスタへの再割り当ての二つのステップを反復することで，クラスタは改良される．反復回数が上限 $maxIter$ をこえるか，分割に変化が無かった場合に停止して，現在の分割を出力する．k-means と同様に，このアルゴリズムは局所最適解にしか収束しないため，初期分割を変えて数回繰り返し，式 (2) のエラーの全クラスタについての総和が最小になるものを選択する．

全体の計算量は $O(|X^*|^2|S| + |X^*||S|k)$ になり，サンプル数 $|S|$ やクラスタ数 k については k-means と同じ計算量だが，対象の総数 $|X^*|$ については 2 乗とやや多い．

4. 実験

ここでは，人工データと嗜好調査の実データを対象にした実験結果を示す．人工データでは，既存の階層的手法に式 (1) の非類似度を用いた場合と k-o'means 法による場合とを比較する．また，人工データの生成条件を変えることで k-o'means 法の特性を調査する．寿司の嗜好に対する調査データでの実験では，k-o'means 法を用いて解析を行い，これが有効な解析手段となることを示す．

4.1 人工データに対する実験

4.1.1 評価基準

まず，実験結果の評価尺度について述べる．本論文では，基準となる分割 π^* と推定した分割 $\hat{\pi}$ の比較基準として次の purity と RIL を用いた．

表 1: 人工データの生成パラメータ

1) 対象の総数: $|X^*| = 100$
2) サンプル順序の数: $|S| = 1000$
3) 順序の長さ: $|X_i| = 10$
4) クラスタ数: $|\pi| = \{2, 5, 10, 50\}$
5) 順序平均の交換回数: {a:∞, b:230000, c:120000}
6) クラスタの大きさの最小/最大の比率: $\{1/1, 1/2, 1/5, 1/10\}$
7) サンプル順序の交換回数: {a:0, b:30, c:72}

purity は幅広く用いられている尺度である．まず，クラスタ $C_i^* \in \pi^*$ 中の要素は，真のラベル i をもつクラスに分類されているとする．このとき，クラスタ $\hat{C}_i \in \hat{\pi}$ 中の全ての対象が，多数を占める真のクラスに分類されたと考えたときの正解率が purity である．形式的には次式で定義される．

$$\text{purity} = \frac{1}{|S|} \sum_{\hat{C}_i \in \hat{\pi}} \left(\max_{C_j^* \in \pi^*} |\hat{C}_i \cap C_j^*| \right) \tag{5}$$

purity の範囲は $[0, 1]$ で，二つの分割が一致するとき 1 になる．

この purity には，その下限が π^* に依存して変化するので，複数の分割の比較結果の平均値を求める場合には，本来は正規化すべきであるという問題がある．そこで，情報損失量（Ratio of Information Loss; RIL）による評価も行った．RIL は正しい分割を推定するために必要とされる情報量のうち，獲得できなかった情報量の割合を示す．ここで，順序 O_i と O_j が，分割 π 中で同じクラスタの要素であるとき 1 をとり，そうでないとき 0 をとる関数 $I((O_i, O_j), \pi)$ を導入し，a_{st} を，全ての順序の対の中で $I((O_i, O_j), \pi^*)$=s かつ $I((O_i, O_j), \hat{\pi})$=t を満たす順序対の数とする．このとき，RIL は次式で定義される．

$$\text{RIL} = \frac{\sum_{s=0}^{1} \sum_{t=0}^{1} \frac{a_{st}}{a_{\cdot\cdot}} \log_2 \frac{a_{\cdot t}}{a_{st}}}{\sum_{s=0}^{1} \frac{a_{s\cdot}}{a_{\cdot\cdot}} \log_2 \frac{a_{\cdot\cdot}}{a_{s\cdot}}} \tag{6}$$

ただし，$a_{\cdot t} = \sum_{s=0}^{1} a_{st}$，$a_{s\cdot} = \sum_{t=0}^{1} a_{st}$，$a_{\cdot\cdot} = \sum_{s=0}^{1} \sum_{t=0}^{1} a_{st}$ である．RIL の範囲も $[0, 1]$ だが，二つの分割が一致するとき 0 になる．

4.1.2 人工データの生成手順

実験に用いた人工データの生成手順について述べる．テストデータは次の 2 段階で生成する：第 1 段階では，k 個の順序平均を生成する．まず X^* の全ての対象を含む順序を生成し pivot とし，これから他の $k-1$ 個の順序平均を生成する．この生成手順は，pivot 中で隣接している対象を均一分布に従いランダムに選択し，それらを交換することを，一定回数だけ繰り返す．この交換回数によってクラスタ間の近さを調節する．第 2 段階では，各クラスタの平均順序からサンプル順序を生成する．順序平均から $|X_i|$ 個の対象をランダムに選択し，順序平均と無矛盾な順序で並べる．ここで，再び隣接する対象をランダムに選び交換することを，一定回数だけ行う．この交換回数によって，クラスタ内のまとまりを調節する．こうして得られたサンプル順序を集めて対象集合とする．

データ生成のパラメータを表 1 にまとめた．パラメータ 1〜3 は全てのデータについて共通である．パラメータ 4 はクラスタ数で，これが増えるとクラスタが小さくなるため，分割の復元は難しくなる．パラメータ 5 は，第 1 段階での交換回数で，a，b，c の順にクラスタが互いに近くなるので分割が困難になる．交換回数は pivot との間の ρ が，それぞれ平均 0.0，0.1，0.3 となるように定めた．パラメータ 6 は，クラスタの大きさの均一性を変える．既存の k-means 法は大きさが均一であるクラスタを抽出する傾向があるので，クラスタの大きさが均一の 1/1 の場合が最も分割が容易であると予測する．最後のパラメータは第 2 段階での対象の交換回数である．a，b，c の順にクラスタ内のまとまりが小さくなるので分割が困難になる．パラメータが a，b，c のとき，pivot とサンプル順序の間の ρ は，それぞれ，平均 1.0，0.715，0.442 になる．これは，サンプル順序より，ランダムな順序が順序平均に近くなる確率が 0.0，0.01，0.10 となるように定めた．

表 2: 階層的クラスタリング手法との比較

	purity	RIL
KOM	0.561 (0.3631)	0.705 (0.4095)
AVE	0.466 (0.2966)	0.909 (0.1671)
MIN	0.315 (0.2663)	0.998 (0.0043)
MAX	0.371 (0.2430)	0.995 (0.0116)

パラメータの組み合わせの総数は 144．各パラメータ設定ごとに 100 個の対象集合を生成し（よって対象集合の総数は 14,400 個），purity や RIL の平均を求めた．

4.1.3 階層的クラスタリング手法との比較

k-o'means 法と代表的な階層的手法である最短距離法，最長距離法，群平均法を比較する．これらの階層的手法では，式 (1) を非類似度として用いることで，順序をクラスタリングすることができる．これら 4 種類の手法を 144 種の人工データに，適用して求めた purity と RIL の平均（括弧内に標準偏差）を表 2 に示す．ただし，正しいクラスタ数は与えた．表中の KOM，AVE，MIN，及び MAX はそれぞれ，k-o'means 法，最短距離法，最長距離法，群平均法を示す．明らかに，k-o'means はどの階層的手法よりも正確にクラスタを復元している．さらに，t 検定を行ったところ，危険率がたとえ 0.1% であってもその差は有意であった．k-o'means 法が既存手法より良い理由は以下のとおりであると考える．まず，対ごとに非類似度を求めると，共通する対象が無い場合は全て無相関の 1.0 となってしまう．それに対し，k-o'means では順序平均の概念によって，より多数の順序をまとめて考慮できるので，同じ対象が幾つかの順序に現れる確率は大きくなり，有意な非類似度を計算できる．言い換えれば，階層的手法は局所的な特徴だけに基づくのに対し，本手法ではより大域的な特徴を考慮できるといえる．さらに，最短距離法や最長距離法にはチェイニングなどの性質により，外乱に弱いという性質もある [神嶌 03].

4.1.4 パラメータの影響

次に，表 1 のパラメータ 4〜7 の変化に伴う，k-o'means の特徴について調査した．表 3 には，特定のパラメータが同じグループについて purity と RIL の平均を示した．例えば，表 3(a) のラベルが "2" の列には，パラメータ 4 が 2，すなわち，$|\pi| = 2$ である 36 種のデータについての平均を示した．全体的に，パラメータ 4 と 7 はクラスタリングの結果に影響するが，他の影響は少ない．以下，詳細な検討を行う．

パラメータ 4: クラスタ数の増加に伴い推定精度は低下している．クラスタ数が 50 にもなると，クラスタ内のサンプル順序が順序平均と全く無矛盾な場合（パラメータ 7 が 0）でも，ほとんどクラスタを復元できない．考えられる原因を以下に示す．可能な順序平均の個数は $|X^*|!$ であるが，この中から一つを選ぶには $\log_2(|X^*|!) \approx 525$ ビットの情報量が必要になる．一方，$|X_i|$ 個の対象の順列の数は $|X_i|!$ なので，順

表 3: データの生成パラメータのクラスタ復元精度への影響

(a) パラメータ 4: クラスタ数

	2	5	10	50
purity	0.910	0.704	0.493	0.139
RIL	0.528	0.597	0.698	0.999

(b) パラメータ 5: クラスタ間の近さ

	a:∞	b:230000	c:120000
purity	0.567	0.565	0.553
RIL	0.694	0.699	0.723

(c) パラメータ 6: クラスタの大きさの均一性

	1/1	1/2	1/5	1/10
purity	0.544	0.546	0.569	0.586
RIL	0.683	0.690	0.711	0.738

(d) パラメータ 7: クラスタ内のまとまり

	a:0	b:30	c:72
purity	0.782	0.531	0.371
RIL	0.279	0.843	0.994

表 4: 寿司の嗜好についてのクラスタの要約

	C_1	C_2
被験者数 $\|C\|$	628	397
味のこってり度	0.3958	−0.1523
食べる頻度	−0.6480	−0.5766
価格	−0.4723	−0.0079
店舗にある頻度	−0.4407	−0.2501

序ひとつあたり約 $\log_2(|X_i|!)$ ビットの情報が得られる．よって，クラスタ数が 50 のときに，クラスタ全体で得られる情報量は $|C|\log_2(|X_i|!)\approx(|S|/|\pi|)\log_2(|X_i|!)\approx$436 ビットしかなく，情報が不足する．よって，十分に正確な順序平均が得られず，元のクラスタを復元できない．

パラメータ 5: クラスタ間の近さを変えても，クラスタの復元精度にあまり影響は無かった．正確な理由は不明だが，非常に多数の対象を含んでいる順序平均は，偶然に一致する確率が非常に低いので，多少のノイズがあっても十分に区別可能ではないかと考える．

パラメータ 6: 既存の k-means はクラスタの大きさに差があると推定精度が低下するが，k-o'means ではそのような減少は見られなかった．これも正確な理由は不明だが，上記と同様に順序平均が容易に区別可能であることが関係していると考える．

パラメータ 7: クラスタ内の類似性が低い場合にはクラスタの分離は困難である．これは，順序平均と比べて，サンプル順序は非常に短く，低いレベルのノイズでも大きな影響を受けるためであると考える．

4.2 嗜好調査データに対する実験

主観的な量の計測には，順序による解析に適しているので，k-o'means 法を寿司の嗜好調査データに適用した．WWW から 25 軒の寿司店のメニューを抽出し 100 種の対象（=寿司）を選んだ．ここから，メニューへの出現頻度に比例する確率で 10 種の対象を，被験者ごとに選んで提示し，好きなものから順に並べるよう指示した．また，対象を提示する順序による影響を避けるため，提示順序も被験者ごとにランダムに変更した．これらは商用の WWW アンケート調査を利用して実施した．全部で 1039 件の回答を得たが，回答時間が短すぎたり長すぎるデータは信頼性が低いと考え，1025 件の順序を選んだ．

k-o'means を探索的な解析手法として利用し，データを二つのクラスタに分割した．各クラスタについての要約を表 4 にまとめた．これは 20 回の試行で最もエラー総和が最小の結果である．表の第 1 行は，各クラスタ内の順序の数（=被験者の数）で，クラスタ C_1 の方が多数を占めている．表の残りの 4 行は，各クラスタの順序平均と，対象のある属性によって対象を整列した順序との間の ρ を示した．例えば，第 4 行は，寿司を価格順に並べた順序と，各クラスタの順序平均との ρ である．この相関は各クラスタの被験者の嗜好と各属性の関連を示すので，各クラスタを特徴づける属性がこの相関によって分析できる．以下に各属性についての詳細に議論する．なお，2 と 3 行目の属性は，被験者への SD 法による質問によって，残りの属性はメニューのデータを解析することで得た．

2 行目の属性は，寿司の味が「こってり」か「さっぱり」かを表し，正の相関はこってり味への嗜好を示す．クラスタ C_1 の被験者は，よりこってりした寿司を好むことが分かる．3 行目の属性は，被験者がその寿司を食べる頻度を表し，正の相関はふだんは食べないものを好むことを示す．どちらのクラスタの被験者もふだん食べる寿司を好み，二つのクラスタに差は見られない．4 行目の属性は寿司の価格に対する影響で，正の相関は安価な寿司を好むことを示す．クラスタ C_1 の被験者は高価な寿司を好むが，C_2 の被験者にはそのような傾向はない．5 行目の属性は，寿司店でその寿司が提供される頻度を表す．正の相関は，定番の寿司を好むことを示す．C_2 の方の相関がいくぶん高いが，その差は統計的に有意ではない．まとめると，クラスタ C_1 の被験者は，C_2 の被験者に比べて，こってりした高価な寿司を好むことが分かる．

5. まとめ

本研究では，順序をクラスタリングする k-o'means 法を提案した．本手法が既存の手法より高精度でクラスタを復元できることを示し，また，本手法の特性について調査した．さらに，寿司の嗜好調査データを解析し，本手法が有効な解析手段であることを示した．今後は，対象の総数 $|X^*|$ に対する 2 乗の計算量を減少させる手法について研究したい．

謝辞： 本研究は科研費萌芽研究 (14658106) の助成を受けた．

参考文献

[神嶌 03] 神嶌 敏弘：データマイニング分野のクラスタリング手法 (1) — クラスタリングを使ってみよう！ —, 人工知能学会誌, Vol. 18, No. 1, pp. 59–65 (2003)

[Kendall 90] Kendall, M. and Gibbons, J. D.: *Rank Correlation Methods*, Oxford University Press, fifth edition (1990)

[Mooney 99] Mooney, R. J. and Roy, L.: Content-Based Book Recommending Using Learning for Text Categorization, in *ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation* (1999)

[中森 00] 中森 義輝：感性データ解析 — 感性情報処理のためのファジィ数量分析手法, 森北出版 (2000)

[Thurstone 27] Thurstone, L. L.: A Law of Comparative Judgment, *Psychological Review*, Vol. 34, pp. 273–286 (1927)