社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
PRMU2003–83　(2003-09)

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　〒 305–8568 茨城県つくば市梅園 1–1–1 産総研つくば中央 2
E-mail: †mail@kamishima.net, ††jun-fujiki@aist.go.jp

あらまし　順序の集合をクラスタリングする k-o'means 法について報告する．順序とは，嗜好や大きさなどの特徴に従って整列した対象の系列である．この解析手法は官能検査など主観的なデータの解析に有用である．k-o'means 法ではクラスタを順序平均で代表するが，この順序平均の近似精度を検証する実験結果を示す.
キーワード　データマイニング，クラスタリング，順序，アンケート調査，官能検査

Clustering Orders — About the Optimality of Order Means

Toshihiro KAMISHIMA† and Jun FUJIKI†

† National Institute of Advanced Industrial Science and Technology (AIST)
AIST Tsukuba Central 2, Umezono 1-1-1, Tsukuba, Ibaraki, 305-8568 Japan
E-mail: †mail@kamishima.net, ††jun-fujiki@aist.go.jp

Abstract　We propose a method of using clustering techniques to partition a set of orders. We define the term *order* as a sequence of objects that are 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-o'means method, which is a modified version of a k-means method, adjusted to handle orders. In this Paper, we will present experimental results in terms of the optimality of the order means.
Key words　Data Mining, Clustering, Order, Questionnaire Survey, Sensory Survey

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サンプル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 [1]．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かった．そこで，順序で記述されたデータを扱う k-o'means 法を提案したが [2], [3]，この手法の性質をさらに調査する．

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 [4]．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．一方，絶対的な物差しを用いる代りに，順序を用いて各被験者の相対的な嗜好の度合いを計測すれば，このような仮定は回避できる．

順序はこうした主観的な変量を測定するのに有用だが，順序を扱うクラスタリング手法は開発されていなかったため，k-o'means 法を提案した．この手法ではクラスタを順序平均なるもので代表させるが，この順序平均の性質を調査する．

2. 節では順序のクラスタリング問題と k-o'means 法について，3. 節では順序平均について，4. 節では実験結果について，5. 節ではまとめについて述べる．

1.1 関連研究

時系列データをクラスタリングする手法 [5]〜[7] が提案されている．観測量の系列を分類する点は類似しているが，時系列では同じ観測量が系列中に複数回現れてもよいのに対し，順序では現れてはならない点が異なる．よって，これらの手法は順序の処理には適さない．

近年，順序に関する研究が多く見られるようになっている．

Cohen ら [8] と Joachims [9] は対ごとの順序関係から，属性ベクトルで記述された対象を順序付けする手法を提案した．神嶌と赤穂 [10] や賀沢ら [11] は順序付けされた対象集合からの学習問題について研究した．Mannila と Meek [12] は順序の集合から特徴的な順序構造を見つける手法を提案した．Sai ら [13] は順序変量に対する相関ルールを提案した．しかし，順序のクラスタリング手法は研究されていない．

2. 順序のクラスタリングと k-o'means 法

ここでは順序のクラスタリング問題と k-o'means 法について，文献 [2], [3] の概略を述べる．

2.1 順序のクラスタリング

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．$X_i{\subseteqq}X^*$ は順序 O_i に含まれる全ての対象からなる対象集合を表す．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．O_i は，$X_i = X^*$ なら完全順序，$X_i \subset X^*$ なら部分順序であるという．

順序のクラスタリング問題を以下に述べる．入力としてサンプル集合 $S = \{O_1, O_2, \ldots, O_{|S|}\}$ 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．

2.2 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 [14] 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．

アルゴリズム k-o'means(S, k, $maxIter$)
$S = \{O_1, \ldots, O_{|S|}\}$: サンプル集合
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

Spearman の ρ は [15] などで順位付けの評価などに用いられている．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，また，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．

$$d(O_1, O_2) = 1 - \rho \tag{3}$$

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．

2.3 k-o'means 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．

k-o'means 法は，クラスタの中心として 3. 節で述べる順序平均を，非類似度として式 (3) 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（ステップ 3）と，順序のクラスタへの再割り当て（ステップ 4）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，各サンプル順序から順序平均までの非類似度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．

2.4 k-o'means 法の既存手法に対する優位性

文献 [2], [3] では，従来の階層的手法と比較し，k-o'means 法の優位性を示したが，その概要をまとめる．

群平均法は，式 (3) を非類似度として用いることで，順序を

表 1　人工データに対する群平均法と k-o'means の RIL の平均

クラスタ内のまとまり		クラスタ数 2	5	10	50
	強	0.001	0.013	0.099	0.998
		0.484	0.643	0.868	0.995
		0.597	0.783	0.993	1.000
		0.947	0.990	0.999	0.999
	弱	0.977	0.999	1.000	1.000
		0.996	1.000	1.000	0.999

各欄の上段が k-o'means 法，下段が群平均法

表 2　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 $\|C\|$	607	418
味のこってり度	0.4016	−0.1352
食べる頻度	−0.6429	−0.6008
価格	−0.4653	−0.0463
店舗にある頻度	−0.4488	−0.2532

クラスタリングできるので，これと k-o'means 法を比較した．順序に基づくクラスタ構造をもった人工データを 4.1 節と同様の方法で生成し，元のクラスタ数は与えて，これら二手法を適用した．全データに対する RIL の平均は，k-o'means が 0.705 であるのに対し，群平均法が 0.910（パラメータの組み合わせ 144 種それぞれに対し 100 回の試行）であり，k-o'means 方がすぐれ，さらにその差は危険率が 0.1% でも有意であった．

さらに影響の大きかったデータ生成のパラメータのクラスタ内のまとまりとクラスタ数を変えた場合の RIL（付録 1. 参照）の試行 100 回の平均を表 1 に示す．各欄の上段に k-o'means 法の結果を，下段に群平均法のそれを示した．実験条件の詳細は文献 [2], [3] を参照されたい．クラスタ数が多く，クラスタ中のサンプルが少ない場合は，順序平均の推定精度が低下するので，クラスタはどちらもほとんど復元できなかった．しかし，順序平均が正確に求められる場合では，k-o'means はクラスタを復元できるが，群平均法ではできなかった．クラスタ内のまとまりが悪くなる，すなわち，クラスタ内のサンプル順序に矛盾（$r(O_1, x^a) > r(O_1, x^b)$ だが $r(O_2, x^a) < r(O_2, x^b)$ である場合）が生じる割合が多少増えても k-o'means はある程度の復元が可能だが，群平均法は急速に復元できなくなった．

k-o'means 法が既存手法より良い理由について考察する．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非類似度が無相関の 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．

2.5　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．100 種の対象（= 寿司）から 10 種の対象を，被験者ごとにランダムに選んで提示し，好きなものから順に並べるよう指示した．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 2 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．

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，定番でない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．

ここで，この実験について，クラスタ数に関する考察を加えておく．最適なクラスタ数は，クラスタリングした結果の利用目的に依存するため一般的には定められない．しかし，均一に分布するデータを分割するこは不適当だと考えられる．これを検証するため，最も近い順序平均の対が区別可能かどうかを検定する．クラスタの順序平均を $\bar{O}_1, \ldots, \bar{O}_{|\pi|}$ と，サンプル集合 S 全体の順序平均を $\bar{O}^*$ と記す．ρ_{ab} は $\bar{O}_a$ と $\bar{O}_b$ の間の順位相関，ρ^*_a は $\bar{O}^*$ と $\bar{O}_a$ の順位相関とする．まず，全ての順序平均の対のなかで，$\rho_{\alpha\beta}$ を最大にする対を見つけ，それらを $\bar{O}_\alpha$ と $\bar{O}_\beta$ とする．この最も類似した二つのクラスタ，C_α と C_β が併合されるべきかを検定するために，二つの順位相関 ρ^*_α と ρ^*_β の差について統計的検定を行った．この場合，次の統計量は，自由度 $|X^*|-3$ の t 分布に従うこと知られている：

$$t = (\rho^*_\alpha - \rho^*_\beta)\sqrt{\frac{(|X^*|-3)(1+\rho_{\alpha\beta})}{2(1-{\rho^*_\alpha}^2-{\rho^*_\beta}^2-{\rho_{\alpha\beta}}^2+2\rho^*_\alpha\rho^*_\beta\rho_{\alpha\beta})}}$$

もし仮説 $\rho^*_\alpha{=}\rho^*_\beta$ が棄却されなければ，これら二つのクラスタは併合され，クラスタ数を減らすべきである．上記の嗜好調査のデータを二つのクラスタに分割した場合，t 値は 9.039 となり，危険率 1% で，こららのクラスタは併合すべきでないといえる．だが，三つに分割したとき $t = 1.695$ となり，これらのクラスタは併合すべきと結論付けられる．この方法は，データ

を解析するときに k を定めるのに役立つが，文献 [16] でも指摘されているように，どのクラスタ数の選択規準も万能ではないことに注意すべきである．なぜなら，この最適性はクラスタリング結果の利用目的やデータの性質に依存するからである．例えば，k-o'means により得られた分割をキャッシュの目的で用いた [17] の研究では，k が 2 より大きな場合に最適な性能が得られている．

3. 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 (3) 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．

$$\bar{O} = \arg\min_{O_j} \sum_{O_i \in C} d(O_i, O_j) \tag{5}$$

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，$\bar{X}=\cup_{O_i \in C} X_i$.

3.1 サンプル順序が完全順序の場合

クラスタ C 中の全ての順序が全て完全順序である，すなわち，$X_i = X^*, \forall X_i \in C$ であるとき，順序平均は Borda の規則により求められる．これは，選挙による意思決定を扱う社会選択の分野で 18 世紀に de Borda が示し，後に論理学者 Dodgson（ペンネームの Lewis Carroll でも著名）によって Borda の規則と呼ばれるようになったもので，次のアルゴリズムと等価である．
1) X^* 中の各対象 x^a について，次の値を計算：

$$\tilde{r}^*(x^a) = \frac{1}{|C|} \sum_{O_i \in C} r(O_i, x^a) \tag{6}$$

2) $\tilde{r}^*(x^a)$ によって昇順に対象を整列し，C の順序平均とする．ただし，$\tilde{r}^*(x^a) = \tilde{r}^*(x^b)$, $x^a \neq x^b$ となる場合には，$x^a \succ x^b$ と $x^b \succ x^a$ のどちらに並べてもよい．

このアルゴリズムの最適性の証明を以下に示す．まず，制限を緩和した順位を考える．厳密な順位は，$1, \ldots, |X|$ のうちの一つの値をとるが，緩和した順位 $\tilde{r}(x)$ は実数をとり，次の条件だけを満たす：

$$\sum_{x \in X} \tilde{r}(x) = \sum_{i=1}^{|X|} i$$

明らかに，厳密な順位はこの条件を満たす．全ての $|X_i|$ は等しいので，式 (3) は，二つの順序の順位の差の二乗和に比例する．それゆえ，緩和された最適な順位は次式の最小化によって求められる：

$$\sum_{O_i \in C} \sum_{x \in X^*} \left(\tilde{r}(x) - r(O_i, x)\right)^2$$

これは，式 (6) の $\tilde{r}(x) = \tilde{r}^*(x), \forall x$ で最小になる．次に，このエラーを最小にする厳密な順序 O_j を見つける．式 (5) の最小化は次のエラーの最小化に等しい：

$$\sum_{O_i \in C} \sum_{x \in X^*} \left(r(O_j, x) - r(O_i, x)\right)^2$$

$$= \sum_{O_i \in C} \sum_{x \in X^*} \left(r(O_i, x) - \tilde{r}^*(x)\right)\right)^2 + |C| \sum_{x \in X^*} \left(\tilde{r}^*(x) - r(O_j, x)\right)^2$$

第 1 項は最小化されているので，第 2 項を最小化する厳密な順序 O_j が順序平均となる．次に，$\tilde{r}^*(x)$ の順に整列した順序 $\tilde{O}^*$ が順序平均に等しいことを背理法により示す．$\bar{O}$ と $\tilde{O}^*$ の間に，順序が一致しない対象の対が少なくとも一つ存在すると仮定する．形式的には，次式を満たす対象の対 x^a と x^b が存在する：

$$\left(\tilde{r}^*(x^a) - \tilde{r}^*(x^b)\right)\left(r(\bar{O}, x^a) - r(\bar{O}, x^b)\right) < 0$$

d_1 を，この場合のエラーの第 2 項とする．さらに，順序平均の方だけで，x^a と x^b を入れ替えたときの，この第 2 項の値を d_2 とすると，

$$d_1 - d_2 = -2|C|(\tilde{r}^*(x^a) - \tilde{r}^*(x^b))(r(\bar{O}, x^a) - r(\bar{O}, x^b)\big) > 0$$

これはエラーが対象の交換によって減ったことになり，$\bar{O}$ が，エラーを最小になる順序，すなわち，順序平均であることと矛盾する．よって，順序平均と $\tilde{O}^*$ とは，全ての対象対で順序が一致しなくてはならず，上記のアルゴリズムで順序平均が求められる．

3.2 サンプル順序が部分順序の場合

残念ながら，サンプル順序が全て完全順序である実問題はまれで，前節の手法は適用できない．例えば，100 個の対象を被験者に好きなものから順に並べさせるといったことは非現実的である．よって，部分順序に適用できる手法が必要だが，この問題は離散最適化なので困難である．そこで，既存の準最適な手法を人工データに適用して実験的に良い手法を探した．その結果，次の Thurstone の一対比較法を採用する．

Thustone の比較判断の法則 (Thurstone's law of comparative judgment) [18] 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．

$$\begin{aligned}\Pr[x^a \succ x^b] &= \int_{-\infty}^{\infty} \phi(\frac{t-\mu_a}{\sigma}) \int_{-\infty}^{t} \phi(\frac{u-\mu_b}{\sigma}) du\, dt \\ &= \Phi\left(\frac{\mu_a - \mu_b}{\sqrt{2}\sigma}\right)\end{aligned} \tag{11}$$

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 $\bar{\mu}_a$ を推定する．

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

この残差は次式で最小になり，得られた $\bar{\mu}_a$ 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．

$$\bar{\mu}_a = \frac{1}{|\bar{X}|} \sum_{x^b \in \bar{X}} \Phi^{-1}\left(\Pr[x^a \succ x^b]\right) \tag{13}$$

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．

4. 実　　験

3.2 節の方法は近似的に順序平均を求めるものなので，真の順序平均を用いた場合と比べて，クラスタの復元能力はどれほど低下するか疑問が生じる．しかし，部分順序について式 (5) の最小化は困難であり，実験できないので，代わりに，サンプル順序が全て完全順序である人工データを生成する．そして，3.1 節の方法で順序平均を求めるクラスタリング手法（最適 k-o'means と記す）と，3.2 節のものを利用する方法（単に k-o'means と記す）を適用し，性能を比較する．

さらに，順序 O_i 中の対象の順位を属性とする順位ベクトルを導入する．完全順序 O の順位ベクトルとは $(r(O, x^1), \ldots, r(O, x^{|X^*|}))$ である．この順位ベクトルをサンプルと見なして，通常の k-means 法を適用する方法と，最適 k-o'means 法とは類似している．すなわち，最適 k-o'means では，3.1 節のアルゴリズムのステップ 2 で，$\tilde{r}^*(x)$ で整列し，順位を自然数にするのに対し，k-means 法では $\tilde{r}^*(x)$ そのものでクラスタが表される．上記の 2 手法に加えこの方法（単に k-means と記す）も比較対象とした．

4.1 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サンプル集合とする．

データ生成のパラメータを表 3 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．

表 3　人工データの生成パラメータ

1) 対象の総数: $\|X^*\| = 10$
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:1000, b:204, c:107}
6) サンプル順序の交換回数: {a:0, b:15, c:30}

表 4　RIL の平均と標準偏差（初期分割が 3 手法で等しい場合）

k-means	k-o'means	最適 k-o'means
0.414 (0.2777)	0.518 (0.3319)	0.468 (0.3017)

表 5　最終分割のクラスタ数の平均（初期分割が 3 手法で等しい場合）

初期クラスタ数	2	5	10	50
k-means	2.000	4.471	8.630	39.826
k-o'means	1.847	3.817	8.012	39.633
最適 k-o'means	1.861	4.002	8.384	43.040

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．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 1.0，0.847，0.715 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 0.0，0.001，0.01 となるように定めた．

4.2 実験結果と考察

パラメータの組み合わせの総数は 36．各パラメータ設定ごとに 100 個のサンプル集合を生成（よってサンプル集合の総数は 3,600 個）し，RIL の平均を求めた結果を表 4（括弧内は標準偏差）に示す．k-means，最適 k-o'means，k-o'means の順に復元能力が高く，またどの対の差でも危険率 0.1% で有意である．k-means と最適 k-o'means は非常に類似した手法だが復元性能には大きな差が生じた．調査したところ次の例のような場合に，最適 k-o'means はクラスタの併合を生じることが多いことが主な原因と考えられる．対象集合が $\{x^1, x^2, x^3\}$ の場合に，次のサンプル順序集合を 2 分割する．(順位ベクトルで表した)

$$O_1 : (0, 1, 2), \quad O_2 : (1, 0, 2), \quad O_3 : (2, 0, 1)$$

初期分割が $C_1 = \{O_1, O_3\}$ と $C_2 = \{O_2\}$ であったとすると，C_1 の中心は，k-means では $(1, 0.5, 1.5)$ で，最適 k-o'means の順序平均は $(1, 0, 2)$ になる．一方，C_2 の中心はどちらも $(1, 0, 2)$ になる．すると，k-means では中心が異なるのでクラスタの併合は生じないが，最適 k-o'means では中心が同じになり併合されやすい．このことは，表 5 の獲得された分割のクラスタ数の平均からも分かる．初期分割にはデータ生成時のクラスタ数（列のラベル）を与えたが，途中でクラスタの中心が一致して併合が生じるので，最終分割では減少しており，その度合いは最適 k-o'means の方が大きい．

この最適 k-o'means の問題点は初期分割を変えることで多くは回避できる．よって，表 4 の実験では 3 手法とも同じ初期分割で実験したが，今度は各サンプル集合ごとに 50 種の異なる初期分割に 3 手法を適用し，平均とサンプルのエラーの総和が

表 6　RIL の平均と標準偏差（50 種回の試行で最良の分割を選択した場合）

k-means	k-o'means	最適 k-o'means
0.337 (0.2915)	0.362 (0.2921)	0.354 (0.2916)

表 7　最終分割のクラスタ数の平均（50 種回の試行で最良の分割を選択した場合）

初期クラスタ数	2	5	10	50
k-means	2.000	4.979	9.437	42.696
k-o'means	1.970	4.740	9.269	44.453
最適 k-o'means	1.969	4.812	9.314	44.440

最小になるものを選んだ結果を表 6 と 7 に示す．表 4 と比較すると，この複数回の試行により，どちらの手法も復元性能が改善された．また，最適 k-o'means の改善の方が大きく，二つの手法の差が小さくなっている．表 5 と 7 を比較すると，クラスタ数の減少の度合いが小さくなっており，このことが復元精度の向上に貢献していることが分かる．しかし，復元精度の差はまだ統計的に有意である．これは，k-means の平均は実数値をとり得るのに対し，順序平均は有限離散値しかとり得ないので，クラスタの中心が一致する頻度はやはり高いためであると考える．しかし，データの生成パラメータとの関連を見ると，差が大きいのは順序平均とサンプル平均が無矛盾（パラメータ 6 が a）な場合である．このような場合は実データではまれなので，影響は大きくはないと考える．定性的には，最適 k-o'means は，式 (3) の順位相関に基づく非類似度に基づいて分類するのに対し，k-means の類似度の意味付けは不明確であるという相違点もある．上記の例で，同じクラスタに分類されている O_1 と O_3 は式 (3) の非類似度では，最も離れた順序対であり，この観点では k-means の獲得する分割は適切でないともいえる．

最適 k-o'means と 3.2 節の方法を用いた k-o'means を比較すると，k-o'means は部分順序にも適用できる近似手法であるためさらに推定精度が低い．詳細に調査すると，差が大きいのは順序平均とサンプル平均が完全に無矛盾な場合であり，この場合が Thurstone モデルで $\sigma = 0$ である特異な場合であることが関係していると考える．

5. ま と め

以前に提案した順序をクラスタリングする k-o'means 法について，その順序平均の最適性を検証する実験を行った．今後は，対象の総数 $|X^*|$ に対する 2 乗の計算量を減少させた手法を考案したい．また，部分順序にも適用でき，かつ，さらに推定精度の高い，順序平均の推定手法を開発できれば，クラスタの復元能力を大きく改良できると考えている．

謝辞： 本研究は科研費萌芽研究 (14658106) の助成を受けた．

文　献

[1] 神嶌：“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, **18**, 1, pp. 59–65 (2003).

[2] 神嶌：“順序のクラスタリング”, 人工知能学会全国大会（第 17 回）論文集, 3F1–01 (2003).

[3] T. Kamishima and J. Fujiki: “Clustering orders”, Proc. of The 6th Int'l Conf. on Dsicvery Science (2003). (in press).

[4] 中森：“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).

[5] E. J. Keogh and M. J. Pazzani: “An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback”, Proc. of The 4th Int'l Conf. on Knowledge Discovery and Data Mining, pp. 239–243 (1998).

[6] I. V. Cadez, S. Gaffney and P. Smyth: “A general probabilistic framework for clustering individuals and objects”, Proc. of The 6th Int'l Conf. on Knowledge Discovery and Data Mining, pp. 140–149 (2000).

[7] 中本, 山田, 鈴木：“動的時間伸縮法に基づく平均時系列生成による時系列データの高速クラスタリング”, 人工知能学会論文誌, **18**, 3, pp. 144–152 (2003).

[8] W. W. Cohen, R. E. Schapire and Y. Singer: “Learning to order things”, Journal of Artificial Intelligence Research, **10**, pp. 243–270 (1999).

[9] T. Joachims: “Optimizing search engines using clickthrough data”, Proc. of The 8th Int'l Conf. on Knowledge Discovery and Data Mining, pp. 133–142 (2002).

[10] T. Kamishima and S. Akaho: “Learning from order examples”, Proc. of The IEEE Int'l Conf. on Data Mining, pp. 645–648 (2002).

[11] 賀沢, 平尾, 前田：“Order SVM: 一般化順序統計量に基づく順位付け関数の推定”, 電子情報通信学会論文誌 D-II, **J86-D-II**, 7, pp. 926–933 (2003).

[12] H. Mannila and C. Meek: “Global partial orders from sequential data”, Proc. of The 6th Int'l Conf. on Knowledge Discovery and Data Mining, pp. 161–168 (2000).

[13] Y. Sai, Y. Y. Yao and N. Zhong: “Data analysis and mining in ordered information tables”, Proc. of The IEEE Int'l Conf. on Data Mining, pp. 497–504 (2001).

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

[15] R. J. Mooney and L. Roy: “Content-based book recommending using learning for text categorization”, ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation (1999).

[16] G. W. Milligan and M. C. Cooper: “An examination of procedures for determining the number of clusters in a data set”, Psychometrika, **50**, 2, pp. 159–179 (1985).

[17] T. Kamishima: “Nantonac collaborative filtering: Recommendation based on order responses”, Proc. of The 9th Int'l Conf. on Knowledge Discovery and Data Mining (2003).

[18] L. L. Thurstone: “A law of comparative judgment”, Psychological Review, **34**, pp. 273–286 (1927).

[19] T. Kamishima and F. Motoyoshi: “Learning from cluster examples”, Machine Learning, **53**, pp. 0–0 (2003).

付　録

1. Ratio of Information Loss (RIL; 情報損失量)

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，情報損失量（Ratio of Information Loss; RIL）[19] を用いた．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．

$$\mathrm{RIL} = \frac{\sum_{s=0}^{1}\sum_{t=0}^{1} \frac{a_{st}}{a_{\cdot\cdot}} \log_2 \frac{a_{\cdot t}}{a_{st}}}{\sum_{s=0}^{1} \frac{a_{s\cdot}}{a_{\cdot\cdot}} \log_2 \frac{a_{\cdot\cdot}}{a_{s\cdot}}} \qquad \text{(A·1)}$$

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．$[0, 1]$ の値をとり，分割が一致するとき 0 になる．