The 16th Annual Conference of Japanese Society for Artificial Intelligence, 2002

3C1-01

順序例からの学習

Learning from Order Examples

神嶌 敏弘[*1] 赤穂 昭太郎[*1]
Toshihiro Kamishima Shotaro Akaho

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

In this paper, I advocate a new learning task that deals with the orders of items, and call this *Learning from Order Examples*(LOE) task. The term *order* means a sequence of items sorted according to some property, such as preference, size, cost and so on. The aim of the task is to acquire the rule that is used for estimating the appropriate order of a given unordered item set. The rule is acquired from a training example set that consists of examples that are pairs of an item set and its true order. The solution methods for the LOE task will serve, for example, a questionnaire survey for predicting popularity of products. I present several solution methods for this problem, and evaluate the performances and the characteristics of these methods based on the experimental results on the artificially generated data.

1. はじめに

本論文では，アイテムの順序を扱う新たな学習タスク『順序例からの学習』(Learning from Order Examples;LOE）を提案する．

ここでいう順序とは，嗜好の強さなどの基準で整列したアイテムの系列をさす．例えば，三つの車種：a，b，c を，ある人が好んでいる順で整列した $c \succ a \succ b$ は順序である（車種 c が一番好きで，b がそうでないことを示す）．LOE タスクの目的は，未整列のアイテム集合の真の順序を推定する規則を獲得することである．この規則を，アイテム集合とこの集合の真の順序の対である学習事例の集合から，獲得する．

この学習は嗜好の調査などへの応用が考えられる．好きな度合いを 5 段階中 4 にするか 3 するか答えにくい場合があるが，そのような場合でも，アイテムを好きな順序に並べられることがあることが，この種の調査で順序を利用する利点である．

本研究では，LOE タスクを形式的に定義し，その解法いくつかを示す．これらの解法を，人工データに適用して，それぞれの性質を明らかにする．

2. 節では，LOE タスクを形式的に定義し関連研究を示す．3. 節では LOE タスクの解法をいくつか示し，4. 節でこれらの解法を人工データに適用する実験を行う．5. 節ではまとめを述べる．

2. 順序例からの学習

2.1 LOE の定式化

ここでは順序例からの学習 (LOE) タスクを形式的に定義する．LOE のタスクは，図 1 にあるように，学習段階と整列段階の二つの段階に分けられる．図 1 左の学習段階では，整列用の規則を訓練事例集合から獲得し，右の整列段階では，獲得された規則を用いて，未整列のアイテム集合の真の順序を推定する．

アイテム I^x とは整列される物や対象で，属性ベクトル $A(I^x)=(a^1(I^x), a^2(I^x), \ldots, a^{\#A}(I^x))$（$\#A$ は属性数）で記述される．この論文では，全ての属性がカテゴリ属性である場合を扱い，s 番目の属性は $v^s_1, \cdots, v^s_{\#a^1}$ のうちのいずれかをとるものとする．アイテム全体の集合をアイテム全集合，$\{I\}_{All}$，と呼び，その部分集合を $\{I\}_i$ で表し，単にアイテム集合と呼ぶ．アイテム集合 $\{I\}_i$ の要素数は $\#I_i$ で表す．

順序とは，大きさ，嗜好の強さ，価格などのある特性で整列したアイテムの系列である．アイテム集合 $\{I\}_i = \{I^x, I^y, \ldots, I^z\}$ の順序を $O_i = I^x \succ I^y \succ \cdots \succ I^z$ と記す．$O_1 = I^9 \succ I^3 \succ I^7$ は，アイテム集合 $I_1=\{I^3, I^7, I^9\}$ の順序の一例である．また，二つのアイテムの間の順序が $I^1 \succ I^2$ であるとき，I^1 は I^2 の前であるという．

アイテム全集合には隠された順序があり，この順序を絶対順序 O^*_{All} と呼ぶ．事例は，アイテム集合 $\{I\}_i$ とこの集合の真の順序 O^*_i の対であり，ノイズがない場合，真の順序は絶対順序と無矛盾である．事例集合 EX は $\#EX$ 個の事例を含む集合である．

$$EX = \{(\{I\}_1, O^*_1), (\{I\}_2, O^*_2), \ldots, (\{I\}_{\#EX}, O^*_{\#EX})\}.$$

アイテム全集合に含まれていても，事例集合のどのアイテム集合にも含まれないアイテムが存在しうることに注意されたい．

LOE タスクの目的は，未整列のアイテム集合 $\{I\}_U$ の推定順序 $\hat{O}_U$ を求める規則を，訓練用の事例集合から獲得することである．ただし，$\{I\}_U$ は未整列だが，集合中のアイテムの属性値は既知とする．

真の順序と推定順序がどれだけ類似しているかの評価に，スピアマンの順位相関係数（Rank Correlation Index; RCI）[Kendall 90] を用いる．同順位が無い場合，アイテム集合 $\{I\}_i$

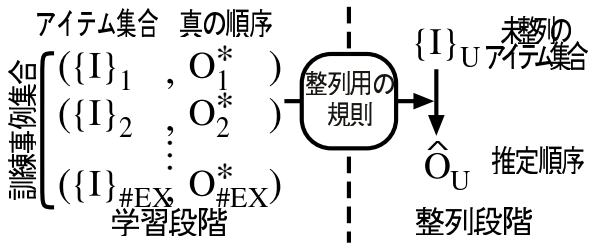

図 1: 順序例からの学習の概要

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

の順序 O_i^1 と O_i^2 の RCI は次式：

$$\mathrm{RCI}(O_i^1, O_i^2) = 1 - \frac{6 \times \sum_{I^x \in \{I\}_i} (r(O_i^1, x) - r(O_i^2, x))^2}{(\#I_i)^3 - \#I_i}$$

ただし，順位 $r(O_i, x)$ は，順序 O_i でアイテム I^x が $r(O_i, x)$ 番目に現れることを示す．この係数は二つの順序が完全に一致するときのみ 1 になり，完全に逆になるとき -1 になる．

2.2 関連研究

順序を扱う研究はいくつかあり，Cohen ら [Cohen 99] の研究はその一つである．タスクへの入力は，アイテムの対のどちらが前にあるかを示したものの集合で，この集合からアイテム I^x が I^y より前にある可能性を表す評価関数 $\mathrm{PREF}(I^x, I^y)$ を求める．その後，次式を最大にする順序を求める：

$$\sum_{x,y:I^x \succ I^y} \mathrm{PREF}(I^x, I^y) \tag{1}$$

Cohen の研究の事例はアイテム対の順序であるが，LOE の事例は任意のアイテム集合の順序である．この違いにより，推定エラーの評価に RCI を利用することが LOE では可能となっている．Cohen のものを含めた他のアイテムの整列を扱う研究では，順序の決定に用いる評価関数のエラーを小さくする努力がなされているが，このエラーが小さくても正しい順序が得られるとは限らない．例えば，評価関数 PREF のエラーが小さな値 ε であったとしても，関数の値の差 $|\mathrm{PREF}(I^x, I^y) - \mathrm{PREF}(I^y, I^x)|$ が ε より小さければ正しい順序は得られない．よって，順序のエラーが直接評価される必要がある．

その他，順位相関係数は官能検査の順位法 [佐藤 85] や，画像の照合 [流郷 01] に利用されているが，いずれも，順位の相関の検証にだけ用いられており，順序の推定を目的とはしていない．

3. LOE タスクの解法

大きく二種類に分類できる LOE の解法について述べる．

分類手法を用いた解法: Cohen の方法に類似した方法で，まず，訓練事例をアイテムの対に分解し，評価関数 $\mathrm{PREF}(I^i, I^j)$ を推定する．この評価関数を用いて，未整列のアイテム集合の推定順序を求める．

回帰手法を用いた解法: まず訓練事例中の順序を一つの全順序にまとめ，この全順序に回帰手法を適用してアイテムの順位を推定する関数を求める．この関数で推定された順位を基に，未整列のアイテム集合を整列する．

3.1 分類手法を用いた LOE の解法

図 2 に，分類手法を用いた LOE 解法の概要を示す．学習段階は L1 と L2 で，整列段階は S1 で構成される．

ステップ L1 では，事例 $(\{I\}_i, O_i^*)$ のアイテム集合 $\{I\}_i$ から，順序 O_i^* で I^x が I^y より前にあるような全てのアイテム対

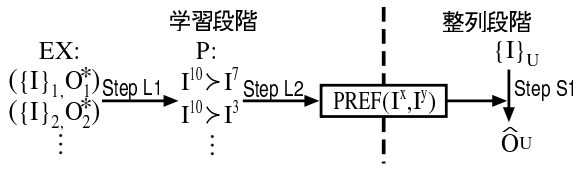

図 2: 分類手法を用いた LOE 解法の概要

(I^x, I^y) を取り出す．例えば，順序 $I^3 \succ I^1 \succ I^2$ からは，(I^3, I^1)，(I^3, I^2)，(I^1, I^2) の三つの順序対を取り出す．このアイテム対の取り出しは EX の全ての事例について行い，取り出したアイテム対集合を P で表す．

ステップ L2 では，この P から評価関数 $\mathrm{PREF}(I^x, I^y)$ を求める．この関数は，属性ベクトル $A(I^x)$ と $A(I^y)$ から，絶対順序で I^x が I^y より前にある可能性の度合いを返すもので，これを単純ベイズ分類器 [Mitchell 97] の手法を用いて求める:

$$\begin{aligned}\mathrm{PREF}(I^x, I^y) &= \mathtt{Pr}[I^x \succ I^y | A(I^x), A(I^y)] \\ &= \frac{\mathtt{Pr}[A(I^x), A(I^y) | I^x \succ I^y]}{\mathtt{Pr}[A(I^x), A(I^y) | I^x \succ I^y] + \mathtt{Pr}[A(I^x), A(I^y) | I^y \succ I^x]}\end{aligned} \tag{2}$$

$$\mathtt{Pr}[A(I^x), A(I^y) | I^x \succ I^y] \approx \prod_{s=1}^{\#A} = \mathtt{Pr}[a^s(I^x), a^s(I^y) | I^x \succ I^y]$$

$$\mathtt{Pr}[a^s(I^x), a^s(I^y) | I^x \succ I^y] = \frac{\#(a^s(I^x), a^s(I^y)) + 1/(\#a^s)^2}{\#P + 1}$$

ただし，$\#P$ は P 中の対の数で，$\#(a^s(I^x), a^s(I^y))$ は，$a^s(I^x) = a^s(I^z)$ かつ $a^s(I^y) = a^s(I^w)$ を満たすような P 中のアイテム対 (I^z, I^w) の数．また，$\mathtt{Pr}[I^x \succ I^y] = \mathtt{Pr}[I^y \succ I^x] = 1/2$ を仮定した．

整列段階では，$\mathrm{PREF}(I^x, I^y)$ を用いて $\{I\}_U$ の真の順序を推定する．ステップ S1 では次の 2 種類の戦略を用いた．

SumClass(SC): これは Cohen[Cohen 99] と同様の戦略である[*1]．式 (1) を最大化する順序を求めるのは $\#I_U$ が大きいときには実行不可能である．よって，最も前にあると推定されるものから順次，推定順序に加える，次の欲張り法を用いる．

1) 推定順序 $\hat{O}_U$ を空にする
2) $\{I\}_U$ から $\sum_{y:I^y \in \{I\}_U, x \neq y} \mathrm{PREF}(I^x, I^y)$ を最大にする I^x を見つける
3) $\{I\}_U$ から I^x を取り除き，$\hat{O}_U$ の末尾に加える
4) $\{I\}_U$ が空ならば $\hat{O}_U$ を出力して終了，でなければステップ 2 へ

ProductClass(PC): SumClass 方法と，最適性の評価関数以外は同じ方法．Cohen の式 (1) では総和を用いているが，なぜ総和とするのかは，ヒューリスティックであるとしか述べていない．そこで，次の理由により，評価関数値の総積を用いる：順序 $I^x \succ I^y \succ I^z$ が観測されるときは必ず，$I^x \succ I^y$，$I^x \succ I^z$，かつ $I^y \succ I^z$ となる．対ごとのこれらの事象が生じる確率はまさに関数 PREF で表されているので，もしこれらの対ごとの事象が独立なら，これらの積は順序 $I^x \succ I^y \succ I^z$ の発生確率を表す．実際には，対ごとの事象は独立ではないが，全くのヒューリスティックであるよりは合理的であると考える．

ProductClass は，SumClass のステップ 2 で最大化する関数が $\prod_{y:I^y \in \{I\}^{(t)}, x \neq y} \mathrm{PREF}(I^x, I^y)$ になる点以外は全く同じである．

3.2 回帰手法を用いた LOE の解法

図 3 は，回帰手法を用いた LOE 解法の概要である．学習段階は L1 で，整列段階は S1 で構成される．この方法を “R” と略記する．

ステップ L1 では，事例集合中の全てのアイテム集合を一つにまとめたアイテム集合 $\{I\}_C$ を生成する．そして，事例中の順序 O_i とできるだけ整合性のある，集合 $\{I\}_C$ の結合順序，

*1 $\mathrm{PREF}(I^x, I^y) = 1 - \mathrm{PREF}(I^x, I^y)$ でなければ，Cohen の方法とは一致しない

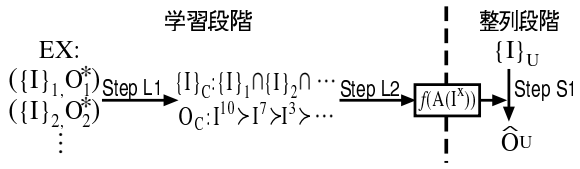

図 3: 回帰手法を用いた LOE 解法の概要

O_C，を求める．この順序を求めるために，前節の方法でアイテム対集合 P を生成し，次の評価関数を求める．

$$\mathrm{PREF}'(I^x, I^y) = \mathtt{Pr}[I^x \succ I^y] = \frac{\#(I^x, I^y) + 0.5}{\#P + 1}$$

ただし，$\#(I^x, I^y)$ は，アイテム対 (I^x, I^y) の数．この評価関数を利用して，ProductClass 手法によって結合順序を求める．SumClass や ProductClass の評価関数 PREF とは異なり，PREF′ は属性値を参照していないことに注意されたい．

ステップ L2 では，アイテムを表す属性ベクトルから，絶対順序中でアイテムがどれだけ前にあるかを予測する順位関数 $f(A(I^x))$ を計算する．この関数の計算には，説明変数がカテゴリ属性であり，目的変数が数値属性である場合の回帰分析である数量化 I 類 [竹内 89] を用いる．まず，$\{I\}_C$ 中の各アイテム I^x のについて，そのアイテムの属性値ベクトルと，アイテム I^x の順序 O_C における順位 $r(O_C, x)$ の対を生成する．この対の集合に数量化 I 類を適用して，アイテムの属性ベクトルから，そのアイテムの順位を予測する順位関数 $f(A(I^x))$ が求まる．

整列段階のステップ S1 では，未整列のアイテム集合 $\{I\}_U$ の各アイテムの属性値ベクトルからアイテムの予測順位を求め，その予測順位の順に整列することで推定順序 $\hat{O}_U$ が求められる．

4. 実験

ここでは，前節の手法を人工データに適用して各手法の特徴を解析する．

4.1 実験の対象と手順

実験の対象データと手順について述べる．

次の 6 種類の型のデータを生成した：1 種類目は，3 値をとる属性が 3 個の場合で (3,3) と表記する．残り 5 種類のデータは (属性数，属性値数) の形式で表すと，それぞれ (3,4)，(3,5)，(5,3)，(5,4)，及び (5,5) である．アイテムの絶対順序は線形のスコア関数を用いて定めた．例えば，アイテム I^7 の属性値ベクトルが $A(I^7)$=(v_1^1, v_3^2, v_1^3) であるとき，スコアは，重み $w(a^s)$ や $w(v_t^s)$ を用いて，$w(a^1)w(v_1^1) + w(a^2)w(v_3^2) + w(a^3)w(v_1^3)$ となる．絶対順序は，このスコアの順にアイテムを並べたものである．6 種類のデータの型それぞれについて，0 と 1 の間の重みをランダムに生成して，10 セットの異なる絶対順序を定めた．すなわち，60 種類のアイテム全集合と絶対順序の対を生成した．さらに，これらの対全てについて，アイテム数 $\#I_i$ が，3，5，及び 10 に，事例数 $\#EX$ を 10，30，及び 50 にして，9 種類の事例集合を生成した．こうして，全部で 540 個の事例集合を生成した．

この 540 個の事例集合に，2 種類の分類手法を用いた解法（SC と PC）と回帰手法を用いた解法（R）の計 3 種類の解法を適用した．評価方法には，$\#EX$ 分割の交叉確認法である，leave-one-out (LVO) 法を用いた．これは，最初の事例，$(\{I\}_1, O_1^*)$ を事例集合から取り出し，残りの事例を用いて整列用の規則を獲得する．最初に取り出した事例のアイテム集合 $\{I\}_1$ に規則を適用して推定順序 $\hat{O}_1$ を求める．この推定順序の損失，すなわち，2.1 節の RCI を求める．損失は推定順序と事例中の真の順序の間で求めるのが一般的だが，ここでは人工データに対する実験で絶対順序が分かっているので，絶対順序に対する損失を求める．この手続きを事例集合中の全ての事例について繰り返し，その平均をもってどれだけ適切に整列されているかを測る．

表 1: RCI の平均：アイテム数ごとの集計

	3	5	10	全体
SC	0.707	0.843	0.940	0.830
PC	0.706	0.844	0.940	0.830
R	0.652	0.849	0.949	0.817

表 2: RCI 間の t 値：アイテム数ごとの集計

	3	5	10	全体
PC-SC	−0.6980	1.1307	0.8914	−0.1283
R-SC	−4.6994	1.0254	3.5008	−2.9264
R-PC	−4.5418	0.9011	3.4233	−2.8887

4.2 実験 1：ノイズの無い場合

ここでは，ノイズの無い場合，すなわち，事例中の真の順序は絶対順序と無矛盾である場合の実験を行う．

表 1 には，540 個の訓練事例それぞれについて LVO で求めた RCI の平均を，アイテム数 $\#I_i$ が同じものごとに分けて，それぞれについて RCI の平均の平均を示した．全体的に，アイテム数の増加に伴い，RCI が 1 に近づき，よりよい推定がなされている．詳細な結果は省略するが，アイテム全集合の要素数が少ない方が，また，訓練事例数が多い方が，精度の良い推定がなされていた．

次に，より厳密な検討を行う．アイテム数が $\#I$ の二つのランダムな順序の間の RCI について，次の t 値は自由度 $\#I - 2$ の t 分布に近似的に従うことが知られている．[Kendall 90]

$$t = \mathrm{RCI}\sqrt{\frac{\#I - 2}{1 - \mathrm{RCI}^2}}$$

$\#I$ が 3，5，及び 10 のときの，99%点はそれぞれ，0.9995，0.9343，及び 0.7155 なので，もし RCI がこれらの値より大きければ，危険率 1%で有意な相関があるといえる．より厳密には，RCI 平均の分布を考慮する必要があるが，おおまかに言って，アイテム数が 10 の場合は絶対順序と有意に無矛盾な順序が推定されている．アイテム数が 5 や 3 の場合は明確な相関があるとは断定できないが，アイテム数が 5 の場合は，危険率を 5%まで緩和すれば，有意な相関があるといえた．

3 種類の手法を比較するために，対応のある場合での t 検定を行った．表 2 は RCI の差の t 値を示したもので，“PC−SC” は SC による RCI から PC による RCI を差し引いた場合で，正の値は PC を用いる方がよい推定ができていることを示す．危険率 1%の有意水準は，全体の場合で 2.332，その他の場合は 2.347 となる．全体としては，PC と SC の間に有意な差は

表 3: 順序の入れ替わりノイズがある場合の RCI の平均

	SC	PC	R
0%	0.830 (1.0000)	0.830 (1.0000)	0.817 (1.0000)
1%	0.829 (0.9985)	0.829 (0.9986)	0.817 (1.0007)
3%	0.826 (0.9949)	0.825 (0.9943)	0.810 (0.9918)
5%	0.823 (0.9907)	0.823 (0.9910)	0.809 (0.9911)
10%	0.814 (0.9806)	0.815 (0.9812)	0.801 (0.9804)

表 4: 属性値が変化するノイズがある場合の RCI の平均

	SC	PC	R
0%	0.830 (1.0000)	0.830 (1.0000)	0.817 (1.0000)
1%	0.825 (0.9942)	0.826 (0.9947)	0.816 (0.9986)
3%	0.819 (0.9862)	0.819 (0.9860)	0.818 (1.0013)
5%	0.818 (0.9848)	0.818 (0.9851)	0.809 (0.9912)
10%	0.805 (0.9696)	0.806 (0.9704)	0.789 (0.9665)

ないが，R は他の二つに劣っている．だが，アイテム集合の要素数別に見てみると，アイテム数の増加に伴い，R は他の 2 手法に対して有意によい推定ができており，特にアイテム数が 10 の場合では，他の手法より有意に優れている．回帰手法を用いる R では，はじめにアイテムの全順序を生成するが，訓練用の順序が短いとこれが困難であるなどの理由が推測ができるが，明確な理由は検証中である．

4.3 実験 3：順序が入れ替わるノイズのある場合

ここでは，順序中で隣接するアイテムが入れ替わるノイズの影響について調べる．表 3 に，訓練事例の順序で隣接するアイテムの 0%—10%が入れ替わった場合の RCI の平均を示す．括弧内はノイズが無い場合に対する相対値である．どの手法も，ノイズの影響による予測制度の低下は同等であるといえる．また，詳細なデータは紙面の都合上提示しないが，アイテム数が少ない方がノイズの影響による性能低下が大きかったが，事例数やアイテム全集合の要素数の違いによる影響は小さかった．また，その性能低下は，どの手法でもだいたい同等であった．

4.4 実験 3：属性値が変化するノイズがある場合

ここでは，アイテムの属性値が変化するノイズの影響について調査する．表 4, 訓練事例のアイテムの属性値が 0%—10%の確率で他の値に変わった場合の RCI の平均で，括弧内はノイズが無い場合に対する相対値である．手法 SC や PC を用いた場合は，ノイズの増加に伴い徐々に予測性能が低下している．それに対して，手法 R のの場合は，ノイズが 1%–5%の間はほとんど性能低下がなく，10%になってやっと低下する．属性値のノイズに対する頑健性では R 手法が他の手法を上回っているといえる．また，アイテム数，事例数，アイテム全集合の要素数の違いによる明確な傾向は見られなかった．

4.5 計算量に関する考察

以下，各手法の計算量について考察する．SC や PC の場合，学習段階では訓練事例集合中のアイテム対を数え上げればよいので，計算量は各事例集合のアイテム数の 2 乗の和のオーダー $\sum_i^{\#EX} \#{I_i}^2$ になる．一方，R の場合，アイテム対の頻度の調査に $\sum_i^{\#EX} \#{I_i}^2$ の時間，整列自体に $\#{I_C}^3$ の時間がかかるため，アイテムの整列には $\#{I_C}^3$ のオーダの時間がかかる．それに加え，数量化 I 類で逆行列を求めるために，属性数 $\#A$ と属性値数 $\#a$ に対して $(\#A\#a)^3$ のオーダの時間が必要である．よって，学習段階では，R 法の方が計算量が多い．

整列段階では，SC や PC の場合，順序にアイテムを追加するたびに，評価関数の値を全てのアイテム対について計算する必要があるため，整列に $\#{I_U}^3$ のオーダの時間がかかる．一方，R の場合は，スコア関数の値の順にソートする時間があればよいので，$\#I_U \log \#I_U$ のオーダの時間で済む．よって，整列段階では，R の方が計算量が少ない．

5. まとめ

この論文では，新しい学習タスクである順序例からの学習を提案し，その解法をいくつか示した．これらの方法を，人工データに適用して，各手法の特徴を解析した．

その結果，どの手法も，適切に順序を推定できることが分かった，分類手法に基づく 2 種類の方法は，回帰手法に基づく方法に対して，整列するアイテム数が少ないときに有利で，多いときに不利であった．順序が入れ替わるノイズに対する頑健性では，どの手法も差がみられなかったが，属性値が変化するノイズでは，回帰手法を用いた方法がすぐれいていた．

今後は，この手法を実問題に適用し，また，RCI を直接小さくできるような学習手法の開発を行いたい．

参考文献

[Cohen 99] Cohen, W. W., Schapire, R. E., and Singer, Y.: Learning to Order Things, *J. of Artificial Intelligence Research*, Vol. 10, pp. 243–270 (1999).

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

[Mitchell 97] Mitchell, T. M.: *Machine Learning*, The McGraw-Hill Companies (1997).

[流郷 01] 流郷, 金子, 五十嵐, 宮本, 亀和田：順位相関に基づくロバスト画像照合法とその地下透水係数推定への応用, 信学技報, Vol. PRMU 2001-26, pp. 47–52 (2001).

[佐藤 85] 佐藤信：統計的官能検査, 日科技連 (1985).

[竹内 89] 竹内啓（編）：統計学辞典, 東洋経済 (1989).