

形状制約モデルによる 顧客の商品選択行動の予測

西村 直樹, 鮎川 矩義, 高野 祐一, 岩永 二郎

機械学習のモデリングにおいて、学習データに対する過剰適合や多重共線性により、推定されたパラメータの解釈が困難であったり、予測精度が悪化したりする場合がある。本稿では、電子商取引サイトにおける商品選択行動予測の問題を取り上げ、結果の解釈可能性と予測精度の向上を目的に、商品閲覧と商品選択の関係を表す事前知識を形状制約として課し、モデルのパラメータを推定する二つの形状制約モデルについて紹介する。

キーワード：形状制約モデル、解釈可能性、商品選択行動、電子商取引サイト、ページ遷移データ

1. はじめに

インターネットの普及に伴い、現在では商品販売やサービスをウェブサイト上で提供する EC (Electronic Commerce: 電子商取引) サイトを多くの企業が運営するようになった。EC サイトを運営する企業にとって、顧客の商品選択の過程を把握することはマーケティング上の重要な課題であり、さまざまな情報を利用した商品選択予測の研究が行われている [1-3]。

取得されたデータをもとに素朴に機械学習モデルを学習すると、データに対する過剰適合や多重共線性のため、推定されたパラメータが事前知識に反し結果の解釈が困難であったり、予測精度が悪化したりする場合がある [4]。過剰適合や多重共線性を避ける方法としては、正則化などの手法 [5] が知られているが、解釈可能性の担保を目的として設計されたものではないため、必ずしも事前知識と整合するとは限らない。

本稿では、EC サイトにおける商品選択予測の結果の解釈可能性と予測精度の向上のために、顧客の商品閲覧の指標と商品選択の関係を表す事前知識を形状制約として課し、モデルのパラメータを推定する 2 次元確率表モデル [6] と閲覧数列モデル [7] の二つの形状制約モデルについて紹介する。

2. 2 次元確率表モデル

例として、表 1 のように 3 人の顧客 $u_1 \sim u_3$ が商品 $v_1 \sim v_4$ を閲覧したとする。表 1 の 1 行目では、顧客 u_1 が商品 v_2 を 6 月 1 日に 1 回、6 月 3 日に 1 回閲覧し、6 月 4 日に選択 (再閲覧や購入) したことを表している。このような顧客と商品に関する商品閲覧と商品選択の履歴が得られたとき、どのように商品選択確率が推定できるだろうか。

2.1 2 次元確率表

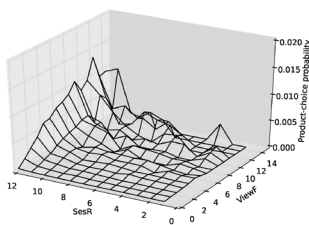
2 次元確率表モデル [6] では顧客分析の手法である RFM 分析 [8] の最新度 (Recency) と頻度 (Frequency) の指標を活用する。顧客 u の商品 v に対する「最新度」は、顧客 u が商品 v のページを閲覧した時点と現時点の近さを表し、「頻度」は顧客 u が商品 v のページを訪れた回数や時間の程度を表す。たとえば、最新度を最終閲覧日の日数の近さとし、頻度を閲覧回数とすると、顧客と商品の組合せに対する 6 月 4 日時点の最新度と頻度は表 1 の (最新度, 頻度) の列のように算出できる。

最新度と頻度の 2 種類の特徴量に対して商品選択確率を推定した表を「2 次元確率表」とよぶ。最新度と頻度を表す有限な正の整数の集合をそれぞれ $R = \{1, 2, 3\}$

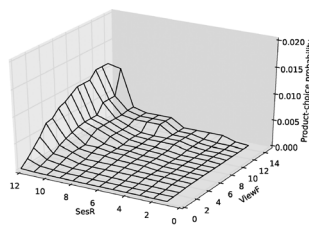
表 1 顧客の商品閲覧と商品選択の例

顧客	商品	商品閲覧回数			商品選択	(最新度, 頻度)	閲覧数列
		6/1	6/2	6/3	6/4		
u_1	v_2	1	0	1	1	(3, 2)	(1, 0, 1)
u_1	v_4	0	1	0	1	(2, 1)	(0, 1, 0)
u_2	v_1	3	0	0	0	(1, 3)	(0, 0, 3)
u_2	v_3	0	0	3	1	(3, 3)	(3, 0, 0)
u_2	v_4	1	1	1	0	(3, 3)	(1, 1, 1)
u_3	v_2	2	0	1	0	(3, 3)	(1, 0, 2)

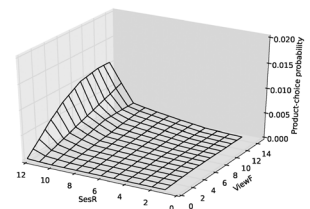
にしむら なおき
株式会社リクルートライフスタイル
〒100-6640 東京都千代田区丸の内 1-9-2
すけがわ のりよし
東京理科大学工学部情報工学科
〒125-8585 東京都葛飾区新宿 6-3-1
たかの ゆういち
筑波大学システム情報系
〒305-8573 茨城県つくば市天王台 1-1-1
いわたが じろう
株式会社エルデシュ
〒222-0033 神奈川県横浜市港北区新横浜 2-3-12 14F



(a) 実績確率表



(b) 単調性制約



(c) 単調性+凸性+凹性制約

図 1 2次元確率表の形状

と $F = \{1, 2, 3\}$ とし、最新度が r で頻度が f の顧客と商品の組について商品が顧客から選択される確率の実績値を $\hat{x}_{rf}$ とする。このとき、表 1 に基づく実績確率表 $(\hat{x}_{rf})_{(r,f) \in R \times F}$ は、欠損値をゼロとして、以下の行列として算出できる：

$$(\hat{x}_{rf})_{(r,f) \in R \times F} = \begin{pmatrix} \frac{0}{0} & \frac{0}{0} & \frac{0}{1} \\ \frac{1}{1} & \frac{0}{0} & \frac{0}{0} \\ \frac{0}{0} & \frac{1}{1} & \frac{1}{3} \end{pmatrix} \approx \begin{pmatrix} 0 & 0 & 0 \\ 1.00 & 0 & 0 \\ 0 & 1.00 & 0.33 \end{pmatrix} \quad (1)$$

このように過去のページ遷移データから、最新度と頻度の組 $(r, f) \in R \times F$ ごとに標本サイズとそのうちの商品選択数を集計し、商品選択確率の過去実績値である実績確率表を作成することは容易である。しかし、集計に用いる標本サイズが少ない場合、商品選択確率 $\hat{x}_{rf}$ に対する信頼性は低くなる。

図 1 では、実際のファッション EC サイトにおける閲覧履歴と購買履歴をもとに、最新度をセッション数に基づく最終閲覧の新しさ (SesR)、頻度を閲覧回数 (ViewF) として式 (1) のように計算した確率表 [6] を可視化した。実績確率表 (図 1(a)) には単調性を満たさず不自然な凹凸が生じている部分が確認できる。

2.2 最新度と頻度に関する形状制約

2次元確率表モデル [6] では、集計により推定された2次元確率表の信頼性の低い部分について最適化問題を解いて補正することを考える。

Fader et al. [9] は、過去の購買の最新度と頻度が将来の購買に対して正の相関があることを実証した。これは、頻度が同じ値のとき最新度が大きいほど購買確率は高く、また最新度が同じ値のとき頻度が大きいほど購買確率が高くなることを示唆している。そこで、2次元確率表に対して以下の単調性を仮定する：

$$x_{rf} \leq x_{r+1,f}, \quad (2)$$

$$x_{rf} \leq x_{r,f+1}. \quad (3)$$

また、最新度が低い場合には、最新度の減少による商品選択確率の減少幅が比較的小さいことが想定される。同様に、頻度が高い場合には、頻度の増加による商品選択確率の増加幅が比較的小さいことが想定される。つまり、最新度が大きくなるにつれて商品選択確率の増分が大きくなる凸性と、頻度が大きくなるにつれて商品選択確率の増分が小さくなる凹性を仮定する：

$$x_{r+1,f} - x_{r,f} \leq x_{r+2,f} - x_{r+1,f}, \quad (4)$$

$$x_{r,f+1} - x_{r,f} \geq x_{r,f+2} - x_{r,f+1}. \quad (5)$$

これらの最新度と頻度に関する形状制約の下で、実績確率表との誤差を最小化する最適化問題を解いた結果を利用して商品選択確率を推定する。単調性制約により補正した図 1(b) では確率表の凹凸がなくなり、さらに凸性と凹性の制約を課して補正した図 1(c) では確率表が滑らかになる。

2.3 2次元確率表モデルの利用

これらの形状制約を課した確率表を利用することで、十分な量のデータが取得できない状況においても顧客の商品選択確率の概形を捉えることが可能になり、商品閲覧の回数に対する商品選択確率の増加傾向や、最終閲覧以降のセッション数に対する商品選択確率の減少傾向に関して、実務観点の示唆を得ることもできる。たとえば、商品選択確率が低下した顧客に対して割引クーポンなどのインセンティブを提供することで、顧客離反を抑制する施策などが考えられる。また予測精度の観点では、これらの制約条件を課した確率表を利用することで、同じ最新度と頻度の特徴量を利用したサポートベクトルマシンとロジスティック回帰よりも商品選択の予測精度が向上することが報告されている [6]。

また EC サイトで扱う商品の異質性を考慮するために、過剰適合を抑制しつつ商品類型ごとに2次元確率表を推定する手法 [10] や、潜在クラスモデルを2次元確率表に適用する手法 [11] が提案されている。そして、2次元確率表を利用して推薦システムの精度を向

上させる手法 [12] も提案されている。

3. 閲覧数列モデル

2次元確率表モデル [6] は、商品閲覧履歴から各顧客の閲覧商品に対する最新度と頻度を数量化し、形状制約のもとで最新度と頻度の組に対して商品選択確率を推定する手法であった。しかし、この手法では顧客の閲覧履歴が最新度と頻度の2次元に縮約されるため、商品閲覧に関する多くの情報が失われてしまう。

たとえば、表1の4~6行目の顧客と商品の組について最新度と頻度を計算すると、いずれの組も、6月3日が最終閲覧日なので「最新度3」、閲覧回数が3回なので「頻度3」となり、顧客の商品に対する異なる閲覧行動を区別することができない。

そこで本節では、閲覧履歴を二つの特徴量に縮約することなく、各期間の商品閲覧数の時系列（閲覧数列）に対して商品選択確率を推定する閲覧数列モデル [7] を紹介する。

3.1 閲覧数列

顧客と商品の組に対して、 i 期前の閲覧数を v_i 、考慮する期間数を d とするとき、閲覧履歴は d 次元ベクトル $\mathbf{v} := (v_1, v_2, \dots, v_d)$ により表される。これを「閲覧数列」とよぶ。また、1 期間の閲覧数の上限を $v_{\max}$ とするとき、可能なすべての閲覧数列の集合は $\Gamma := \{0, 1, \dots, v_{\max}\}^d$ となる。

このように定義した閲覧数列によって、最新度と頻度の指標では区別できなかった閲覧行動を表1の閲覧数列の列のように区別して扱うことができるようになる。区別可能なパターン数を比較すると、2次元確率表モデルは $dv_{\max}$ 個、閲覧数列モデルは $(1 + v_{\max})^d$ 個となり、閲覧数列モデルは非常に表現力が高いことがわかる。一方で、閲覧数列のパターン数は期間数に対して指数的に増加するため、それに伴い各パターンに割り当てられる標本サイズも小さくなってしまう。

そこで閲覧数列モデル [7] でも、閲覧数列に対して形状制約を導入することで商品選択確率を補正する。このために、最新度と頻度に基づく2次元確率表モデルの単調性を多次元に拡張する。

3.2 閲覧数列の順序関係

閲覧数列の集合 Γ 上の順序関係を $\succeq$ とし、次の二つの観察に基づいて閲覧数列の順序関係を定義する：
閲覧数列の最新度に関する単調性：閲覧数の総和が等しい二つの閲覧履歴を比較した場合、より新しい期間に多く閲覧されているもののほうが商品選択確率が高くなることが想定される。形式的には、

$$\begin{aligned} & (\dots, v_i, \dots, v_j, \dots) \\ & \rightarrow (\dots, v_i + 1, \dots, v_j - 1, \dots) \end{aligned} \quad (6)$$

なる操作（閲覧数列の成分の値をより新しい期間へ1単位移動させること）を繰り返して $\mathbf{v}$ から $\mathbf{v}'$ へ変換できるとき、 $\mathbf{v}' \succeq \mathbf{v}$ とする。たとえば $(2, 1, 0) \succeq (1, 2, 0) \succeq (1, 1, 1)$ の関係が成り立つ。

閲覧数列の頻度に関する単調性：二つの閲覧履歴において、ある一つの期間の閲覧数のみ異なる場合、その期間に多く閲覧されているもののほうが商品選択確率が高くなることが想定される。形式的には、

$$(\dots, v_i, \dots) \rightarrow (\dots, v_i + 1, \dots) \quad (7)$$

なる操作（閲覧数列のある成分の値を1単位増加させること）を繰り返して $\mathbf{v}$ から $\mathbf{v}'$ へ変換できるとき、 $\mathbf{v}' \succeq \mathbf{v}$ とする。たとえば $(1, 2, 1) \succeq (1, 1, 1) \succeq (0, 1, 1)$ の関係が成り立つ。

本研究では以上の観察を踏まえ、 Γ 上の順序関係を $\succeq$ を次のように定義する：

$$\mathbf{v}' \succeq \mathbf{v} \iff \begin{cases} \text{(a) } \mathbf{v} = \mathbf{v}', \text{ または,} \\ \text{(b) 式 (6), (7) の操作を繰り返して,} \\ \quad \mathbf{v} \text{ を } \mathbf{v}' \text{ へ変換可能.} \end{cases}$$

このように定義された $\succeq$ は Γ 上の半順序となる [7]。閲覧数列モデルでは、 $x_{\mathbf{v}}$ を閲覧数列 $\mathbf{v}$ の商品選択確率として、順序関係 $\mathbf{v}' \succeq \mathbf{v}$ が成り立つ場合には、商品選択確率の単調性 $x_{\mathbf{v}'} \geq x_{\mathbf{v}}$ を課す。

3.3 閲覧数列モデルのパラメータの推定

閲覧数列モデルでは順序関係 $\succeq$ に基づく商品選択確率の単調性制約の下で、閲覧数列 $\mathbf{v}$ の学習データ数 $n_{\mathbf{v}}$ によって重み付けられた残差2乗和が最小となるように、商品選択確率 $x_{\mathbf{v}}$ を推定する。この問題は以下のように凸2次最適化問題として定式化できる：

$$\begin{aligned} & \text{最小化} \quad \sum_{\mathbf{v} \in \Gamma} n_{\mathbf{v}} (x_{\mathbf{v}} - \hat{x}_{\mathbf{v}})^2 \\ & \text{制約条件} \quad \begin{aligned} & x_{\mathbf{v}'} \geq x_{\mathbf{v}} \quad (\mathbf{v}, \mathbf{v}' \in \Gamma \text{ かつ } \mathbf{v}' \succeq \mathbf{v}), \\ & 0 \leq x_{\mathbf{v}} \leq 1 \quad (\mathbf{v} \in \Gamma). \end{aligned} \end{aligned} \quad (8)$$

ただし $\hat{x}_{\mathbf{v}}$ は商品選択確率の実績値である。

最適化問題 (8) は、素朴に実装すると閲覧数列間の順序関係の制約条件数が膨大になってしまう。そこで半順序構造のハッセ図（半順序の推移簡約表現）を構成し、冗長な制約条件を削除することを考える。

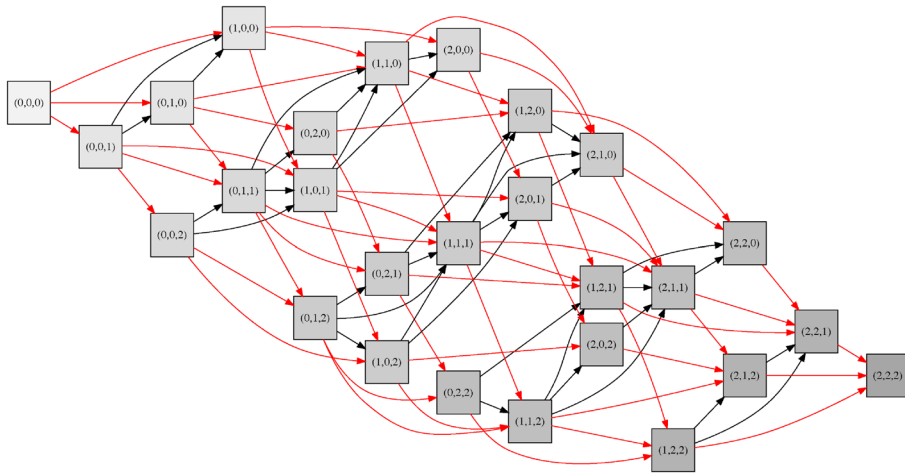


図2 式 (6), (7) の操作による閲覧数列の順序関係

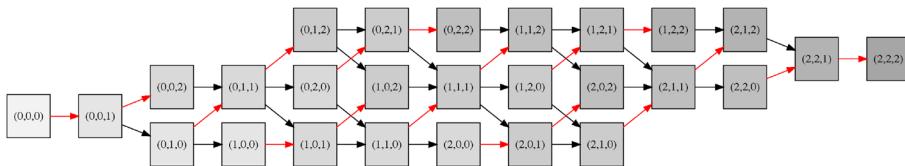
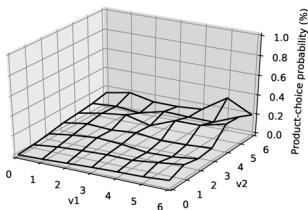
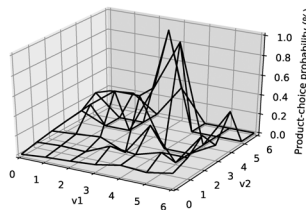


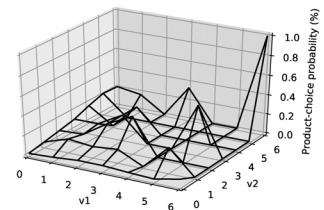
図3 閲覧数列の順序関係のハッセ図



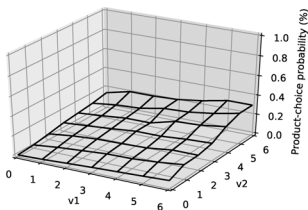
(a) 実績確率表: $v_3 = 0$



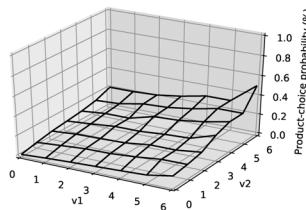
(b) 実績確率表: $v_3 = 1$



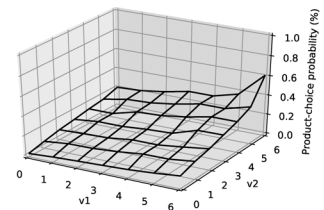
(c) 実績確率表: $v_3 = 2$



(d) 補正後の確率表: $v_3 = 0$



(e) 補正後の確率表: $v_3 = 1$



(f) 補正後の確率表: $v_3 = 2$

図4 閲覧数列の確率表の形状

図2と図3は $(d, v_{\max}) = (3, 2)$ とし、節点を閲覧数列、枝を順序関係とした有向グラフである。

図2では、式(6)、(7)の適用による推移を有向枝(枝の先が上の順序に対応)で表す。これらの枝に対応する制約条件をすべて課すと、その数は90組になる。一方、冗長な制約条件を削除することで図3のように

42組まで減らすことができ、このハッセ図で隣接する閲覧数列の組に対してのみ単調性制約を課せば十分である。期間数 d や閲覧数上限 $v_{\max}$ が大きくなるほど、この方法による計算効率化の効果は大きくなる。

図4は、中国のECサイトTMallのページ遷移データ[13]から作成した $(d, v_{\max}) = (5, 6)$ の閲覧数列の

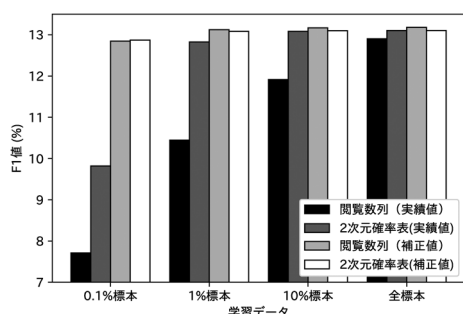


図5 2次元確率表モデルと閲覧数列モデルの予測精度

実績確率表と、その補正後の確率表を図示したものである。簡単のため、 v_4, v_5 を 0 に固定し、 $v_3 = 0, 1, 2$ の各値に対して v_1 と v_2 を 2 軸とした $(v_1, v_2, v_3, 0, 0)$ の形の閲覧数列の確率表を図示した。

計算の結果、 $v_3 = 0, 1, 2$ の順に標本サイズが小さくなるため、それに伴い実績確率表に凹凸が生じることがわかる。一方で、補正後の確率表では、閲覧数列の最新度と頻度に関する単調性に基づく補正がなされている。このように形状制約のもとで推定値を補正することで、2次元確率表モデルと同様に、結果の解釈可能性が向上する。

4. 形状制約モデルの予測精度

本節では、商品選択予測の精度検証 [7] について紹介する。検証では、前述の中国の EC サイト TMall のページ遷移データ [13] を用いた。

4.1 2次元確率表モデルと閲覧数列モデルの比較

本節では、各モデルの標本サイズの変化に対する頑強性についても確認するため、学習データの標本サイズを 0.1% 標本、1% 標本、10% 標本、全標本と変化させて検証した。検証では、各顧客に対して 3 商品を選択し、各商品が再閲覧された割合を表す F1 値を評価指標とした。

図 5¹ から、最適化問題を解き補正した 2次元確率表モデルと閲覧数列モデルについては、0.1% 標本のときは 2次元確率表モデルの予測精度が高いが、1% 標本、10% 標本、全標本のときは閲覧数列モデルの予測精度が高いことがわかる。また補正による予測精度の改善幅は閲覧数列モデルのほうが大きいことがわかる。このことから、標本サイズが十分に確保できる場合には表現力の優れた閲覧数列モデルの予測精度が高く、一方で標本サイズが少ない場合には 2次元確率表モデル

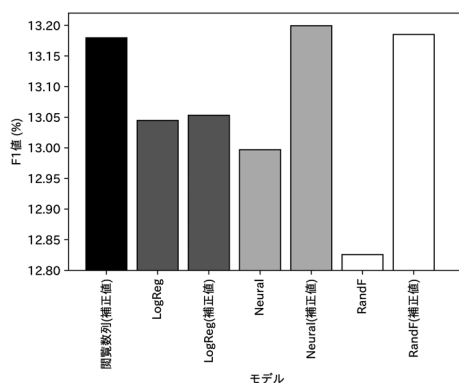


図6 機械学習モデルに対する予測値補正の効果

の予測精度が高い場合もあることが示唆される。

4.2 閲覧数列モデルによる予測値の補正

つづいて、各期の閲覧回数の特徴量として利用した機械学習モデルと予測精度を比較する。図 6 では、L2 正則化つきロジスティック回帰: LogReg, 人工ニューラルネットワーク: Neural, ランダムフォレスト: RandF, そして機械学習モデルの予測値を最適化問題 (8) の $\hat{x}_v$ に代入して補正した推定値と比較する。

図 6 から、閲覧数列モデルは補正なしの機械学習モデルと比べて予測精度が高いことがわかる。また閲覧数列の順序関係を考慮して補正することで、機械学習モデルの予測精度が向上している。これらのことから、対象データに対して商品選択の予測精度向上の観点では、閲覧数列の単調性が有効な事前知識であったと考えることができる。

5. おわりに

本稿では、EC サイトにおける顧客の商品選択行動の予測の問題を取り上げ、結果の解釈可能性と予測精度の向上を目的とした、2次元確率表モデル [6] と閲覧数列モデル [7] の二つの形状制約モデルを紹介した。

機械学習ライブラリの拡充などによりデータに基づく予測手法が広く利用されるようになった一方で、モデルの予測結果の解釈困難性が実利用の妨げとなる場合もあると感じる。4.2 節では、ページ遷移データに対する事前知識を表す形状制約を機械学習モデルの後処理として課すことで、最終的な予測結果が事前知識と整合することを保証する例を挙げた。このような補正は、予測結果に対して説明が求められるような状況において特に有用である。

今回紹介した二つのモデルに限らず、対象データに対して適切な事前知識を形状制約として課したモデルを用いることで、予測手法が利活用可能な領域がさら

¹ 図 5 と図 6 では基線が 0 でないことに留意されたい。

に広がるであろう。

参考文献

- [1] W. W. Moe and P. S. Fader, “Dynamic conversion behavior at e-commerce sites,” *Management Science*, **50**, pp. 326–335, 2004.
- [2] J. Qiu, Z. Lin and Y. Li, “Predicting customer purchase behavior in the e-commerce context,” *Electronic Commerce Research*, **15**, pp. 427–452, 2015.
- [3] Y. Zhang and M. Pennacchiotti, “Predicting purchase behaviors from social media,” In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1521–1532, 2013.
- [4] E. E. Altendorf, A. C. Restificar and T. G. Dietterich, “Learning from sparse data by exploiting monotonicity constraints,” In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 18–26, 2005.
- [5] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media, 2009.
- [6] J. Iwanaga, N. Nishimura, N. Sukegawa and Y. Takano, “Estimating product-choice probabilities from recency and frequency of page views,” *Knowledge Based Systems*, **99**, pp. 157–167, 2016.
- [7] N. Nishimura, N. Sukegawa, Y. Takano and J. Iwanaga, “Predicting online item-choice behavior: A shape-restricted regression approach,” *arXiv preprint*, arXiv:2004.08519, 2020.
- [8] A. M. Hughes, *Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-based Marketing Program*, McGraw-Hill, 2000.
- [9] P. S. Fader, B. G. Hardie and K. L. Lee, “RFM and CLV: Using iso-value curves for customer base analysis,” *Journal of Marketing Research*, **42**, pp. 415–430, 2005.
- [10] 西村直樹, 鮎川矩義, 高野祐一, 岩永二郎, 水野眞治, “EC サイトの商品特性を考慮した 2 次元確率表による購買予測,” *オペレーションズ・リサーチ: 経営の科学*, **60**, pp. 69–74, 2015.
- [11] N. Nishimura, N. Sukegawa, Y. Takano and J. Iwanaga, “A latent-class model for estimating product-choice probabilities from clickstream data,” *Information Sciences*, **429**, pp. 406–420, 2018.
- [12] J. Iwanaga, N. Nishimura, N. Sukegawa and Y. Takano, “Improving collaborative filtering recommendations by estimating user preferences from clickstream data,” *Electronic Commerce Research and Applications*, **37**, No. 100877, 2019.
- [13] M. Ludewig and D. Jannach, “Evaluation of session-based recommendation algorithms,” *User Modeling and User-Adapted Interaction*, **28**, pp. 331–390, 2018.