A-025

顧客セグメンテーションを目的とした潜在クラスモデルによる購買行動分析に関する一考察

Analysis of customer purchase behavior based on page transitions by latent class model for customer segmentation

松嵜 祐樹 † **三川 健太** ‡ **後藤 正幸** †
Yuki matsuzaki Kenta Mikawa Masayuki Goto

1 研究背景・目的

近年，インターネット上の EC サイトを通じた商品の購買が頻繁に行われるようになり，その市場規模は増大する傾向にある．このような背景のもと，膨大に蓄積された顧客の購買行動データを分析し，顧客ごとの特徴を考慮したマーケティング施策の重要性が高まっている．施策を実施する際には，どのような顧客層に対してどのような施策が効果的であるかを検討する必要がある．このように適切なターゲットに的確な施策を実施するためには，顧客を意味のあるグループに分ける顧客セグメンテーションが重要となる．この顧客セグメンテーションに関して，Aspect Model [1] (以下，AM) のような確率モデルの利用によるセグメンテーションの有効性が広く認識されている．AM では，主として顧客と商品のペアに対して，潜在的なクラスを仮定し，顧客の嗜好の異質性や商品の類似性を考慮した購買行動のモデル化を行っている．

一方，EC サイトのアクセスログデータには，顧客の購買履歴に加えて，購買までにどのページを閲覧したのかという閲覧履歴も含まれている．ページの閲覧履歴は，購買，または離脱に至るまでのプロセスを示すものであり，購買行動の特徴を強く反映したものであると考えられる．すなわち，顧客の購買履歴に加えて，どのような閲覧行動の後に購買が起こったのかをモデルとして考慮することができれば，より詳細に顧客の購買行動を記述することが可能となる．

そこで本研究では，顧客の購買有無に加えて，どのような遷移でページを閲覧したのかといった閲覧履歴を考慮したモデルの構築を行う．モデル化に際しては，ニュースサイトの閲覧行動に対して潜在的なクラスを仮定した Dias らのモデル [2] を拡張することで購買行動の分析を可能とする新たなモデルを構築する．また，大手総合通販カタログサイトのアクセスログデータに対して，提案モデルを適用しすることで，モデルの有用性を示す．

2 準備

2.1 データ概要

本研究では，EC サイトのアクセスログ解析を行う企業である Emotion Intelligence 社が保有するデータ (大手総合通販カタログサイトのアクセスログデータ) を分析対象とする．サイトに訪問してから購買，離脱するまでの 1 つ閲覧行動にセッション ID が付与されており，この ID に閲覧履歴と購買有無が紐付いている．また，1 ページあたり閲覧時間，ページ遷移回数，施策の実施有無なども蓄積されている．EC サイトの各ページにはそれぞれ item ページや category ページといったようなページタイプが付与されており，本研究ではこのページタイプを閲覧履歴として使用する．なお，顧客プライバシー保護の観点から，デモグラフィックデータは分析対象としない．

2.2 先行研究

顧客の閲覧行動にマルコフ性を仮定し，潜在クラスモデルとして表現したものに Dias らのモデル [2] がある．いま，I 個からなるセッション集合を $\mathcal{S} = \{s_i : 1 \leq i \leq I\}$，$J$ 種類からなるページタイプ集合を $\mathcal{K} = \{k_j : 1 \leq j \leq J\}$，$s_i$ の長さ T_i の閲覧履歴系列を，$\boldsymbol{x}_i = (x_0^i, x_1^i, \cdots, x_{T_i}^i)$ と表す．ただし，s_i の t 番目の閲覧ページタイプは，$x_t^i \in \mathcal{K}$ を満たす．ここで，t 番目の閲覧ページタイプは，$t-1$ 番目の閲覧ページタイプのみに依存するという，1 次のマルコフ性を仮定すれば，ある潜在クラス v_i に所属するセッション s_i の閲覧履歴 $\boldsymbol{x}_i$ の生起確率は以下の式 (1) のように表される．ただし，s_i が所属する潜在クラス v_i は，$v_i \in \mathcal{Z} = \{z_l : 1 \leq l \leq L\}$ を満たす．

$$P(\boldsymbol{x}_i, v_i) = P(x_0^i|v_i) \prod_{t=1}^{T_i} P(x_t^i|x_{t-1}^i, v_i) \tag{1}$$

また，式 (1) において，$P(x_0^i|v_i)$ は初期分布であり，$P(x_t^i|x_{t-1}^i, v_i)$ は s_i において $t-1$ 番目にページタイプ x_{t-1}^i を閲覧したのちに，t 番目にページタイプ x_t^i を閲覧する確率である．

3 提案モデル

本研究では，Dias らのモデルを拡張することによって，ページの閲覧履歴と購買行動の双方を考慮したモデルの提案を行う．

3.1 提案モデルの定式化

Dias らのモデルで考慮している閲覧履歴に加えて，購買行動の有無を考慮できるように拡張を行う．Dias らのモデルと同様に，I 個からなるセッション集合を $\mathcal{S} = \{s_i : 1 \leq i \leq I\}$，$J$ 種類からなるページタイプ集合を $\mathcal{K} = \{k_j : 1 \leq j \leq J\}$，$T_i$ 個ある s_i の閲覧履歴を $\boldsymbol{x}_i = (x_0^i, x_1^i, \cdots, x_{T_i}^i)$ と定義する．これに加えて，s_i の購買行動の有無を表す変数 w_i 以下のように定義する．

$$w_i = \begin{cases} 1 & (s_i\text{で購買が起こる場合}) \\ 0 & (s_i\text{で購買が起こらない場合}) \end{cases} \tag{2}$$

ここで，s_i の閲覧履歴 $\boldsymbol{x}_i$，購買行動 w_i に対応する潜在変数を v_i とすれば，i 番目の完全データは $(\boldsymbol{x}_i, w_i, v_i)$ と表される．ただし，$v_i \in \mathcal{Z} = \{z_l : 1 \leq l \leq L\}$ を満たす．よって，i 番目の完全データ $(\boldsymbol{x}_i, w_i, v_i)$ についての確率モデルは以下のように表される．

$$P(\boldsymbol{x}_i, w_i, v_i) = P(v_i)P(\boldsymbol{x}_i|v_i)P(c|v_i)^{w_i}P(\bar{c}|v_i)^{1-w_i} \tag{3}$$

ただし，c は，「購買が起きる」という事象，$\bar{c}$ はその余事象である．ここで，$P(\boldsymbol{x}_i|v_i)$ に対し，Dias らのモデルと同様に閲覧ページタイプに 1 次マルコフ性を仮定すれば，式 (3) は以下のようになる．

$$\begin{aligned} P(\boldsymbol{x}_i, w_i, v_i) &= P(v_i)P(\boldsymbol{x}_i|v_i)P(c|v_i)^{w_i}P(\bar{c}|v_i)^{1-w_i} \\ &= P(v_i)P(x_0^i|v_i)\left\{\prod_{t=1}^{T_i} P(x_t^i|x_{t-1}^i, v_i)\right\}P(c|v_i)^{w_i}P(\bar{c}|v_i)^{1-w_i} \end{aligned} \tag{4}$$

†早稲田大学
‡湘南工科大学

さらに，潜在変数 $v_i = z_l$ の下での閲覧履歴 $\boldsymbol{x}_i$，購買履歴 w_i の生起確率 $P(\boldsymbol{x}_i, w_i|v_i = z_l)$ は以下のようになる．

$$\begin{aligned} &P(\boldsymbol{x}_i, w_i|v_i = z_l) \\ &= P(\boldsymbol{x}_i|v_i)P(c|v_i)^{w_i}P(\bar{c}|v_i)^{1-w_i} \\ &= P(x_0^i|v_i)\prod_{t=1}^{T_i} P(x_t^i|x_{t-1}^i, v_i)P(c|v_i)^{w_i}P(\bar{c}|v_i)^{1-w_i} \\ &= \prod_{j=1}^{N} \lambda_{lj}^{\delta(x_0^i=k_j)} \prod_{j=1}^{N}\prod_{m=1}^{K} a_{ljm}^{n_{ijm}} \gamma_l^{w_i}(1-\gamma_l)^{1-w_i} \end{aligned} \quad (5)$$

ただし，$\lambda_{lj} = P(x_0^i = k_j|v_i = z_l)$，$\delta(x_0^i = k_j)$ は $x_0^i = k_j$ のとき 1 となるインジケータ関数，$a_{ljm} = P(x_t = k_j|x_{t-1} = k_m, z_l)$，$n_{ijm}$ はページタイプ k_j から k_m への遷移回数，$\gamma_l = P(c|v_i = z_l)$ である．

3.2 パラメータ推定

このモデルは観測できない変数である潜在変数を含むため，EM アルゴリズムを用いてパラメータ π_l, λ_{lj}, a_{ljm}, γ_l の推定を行う．以下に E-step, M-step それぞれの更新式を示す．

【E-step】

$$\begin{aligned} P(z_l|\boldsymbol{x}_i, w_i) &= \frac{\pi_l P(\boldsymbol{x}_i, w_i|z_l)}{\sum_r^S \pi_r P(\boldsymbol{x}_i, w_i|z_l)} \\ &= \alpha_{il} \end{aligned} \quad (6)$$

【M-step】

$$\pi_l = \frac{1}{n}\sum_i^I \alpha_{il} \quad (7)$$

$$\lambda_{lj} = \frac{\sum_i^I \alpha_{il}\delta(x_0^i = k_j)}{\sum_i^I \alpha_{il}} \quad (8)$$

$$a_{ljm} = \frac{\sum_i^I \alpha_{il} n_{ijm}}{\sum_l^K \sum_i^I \alpha_{il} n_{ijm}} \quad (9)$$

$$\gamma_l = \frac{1}{n\pi_l}\sum_i^I \alpha_{il} w_i \quad (10)$$

EM アルゴリズムでは，式 (11) で表される完全データの対数尤度が収束するまでパラメータの更新を行う．なお，$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_I)$，$\boldsymbol{W} = (w_1, \ldots, w_I)$，$\boldsymbol{V} = (v_1, \ldots, v_I)$ である．

$$\log P(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{V}) = \sum_i \log P(\boldsymbol{x}_i, w_i, v_i) \quad (11)$$

更新前と更新後の式 (11) の差分によって収束の判定を行う．

4 提案モデルを用いた実験

提案モデルの有用性を検証するため，大手総合カタログ通販サイトにおける閲覧履歴，及び，購買履歴データを用いた実験を行った．対象とする EC サイトでは，リアルタイムに割引クーポンを発行するという施策を実施しており，クーポン発行の対象となったセッションに対して，実際にクーポンを発行する A 群と発行しない B 群に分ける AB テストを行うことで効果の検証を行っている．本実験では，顧客の閲覧，購買有無を提案モデルを用いてモデル化することで，現在実施している施策の効果の検証を行う．

4.1 実験条件

実験には，2016 年 4 月 1 日から 7 日までの 7 日間で蓄積された閲覧，購買履歴データを用いる．1 セッションを 1 データとし提案モデルの学習を行う．学習データの総セッション数は 386,671，総閲覧ページ数は 6,031,916 である．なお，閲覧されるページの性質を考慮し，閲覧端末が PC であるデータのみを用い，潜在変数の数は事前実験より 8 とした．

4.2 結果・考察

購買確率を表すパラメータ $P(c|z)$ の値や A 群 (クーポン発行グループ) と B 群 (クーポン非発行グループ) の違いに着目し，特徴的な値をとっていたクラスについて詳細な分析を行った．その結果を表 1 に示す．なお，A 群と B 群はクーポン対象となったセッションからランダムに振り分けられており，購買割合 (A/B 群) はクーポンが発行された/されないセッションのうち，購買に至ったセッションの割合である．また，クーポン効果は，A 群と B 群のクーポン購買割合の比であり，クーポン発行による購買割合の変化を表す．

表 1．実験結果

潜在クラス	z_2	z_4	z_6
混合比	0.060	0.082	0.270
所属人数	23029	31836	104281
購買確率	0.092	0.481	0.002
クーポン対象割合	3.3%	6.2%	4.0%
購買割合 (A 群)	25.2%	68.8%	0.8%
購買割合 (B 群)	16.7%	66.3%	0.9%
クーポン効果 (A 群/B 群)	1.51	1.09	0.89

表 1 より，最も多くのクーポンが発行されているのは潜在クラス z_4 であることが分かる．この潜在クラスでは，購買確率が 0.481 と非常に高くクーポン効果も 1.09 と高くないことから，クーポン発行による効果が薄いことが分かる．つまり，この潜在クラスに属する顧客の購買行動はクーポンを発行せずとも購買が行われるものが多いと考えられる．一方で潜在クラス z_2 では，購買確率が 0.092 と比較的低く，クーポン効果 1.51 と最も高くなっていることが分かる．このことから，この潜在クラスに属する顧客の購買行動は割引クーポンによりよく反応するものであると考えられる．よって，潜在クラス z_4 よりもよりクーポン発行の効果が高いと考えられる潜在クラス z_2 に，より多くのクーポンを発行するべきであると言える．また，潜在クラス z_6 は購買確率は 0.02，クーポン効果 0.89 とどちらも低い値をとっている．よってこの潜在クラスに対してはクーポンを発行すべきでないと言える．

5 まとめ・今後の課題

本研究では，EC サイトにおける閲覧，購買履歴の双方を考慮した確率的潜在クラスモデルの構築を行った．また，実際の閲覧，購買履歴と施策の実施結果を用い，提案モデルによる実験を行った．その結果から，今後 EC サイトの運営者が取るべき方向性を示すことで，提案モデルの有用性を示した．今後の課題として，複数の EC サイトおよび異なる施策についてのさらなる実験や検証が挙げられる．

参考文献

[1] T. Hoffman,“Probabilistic Latent Semantic Indexing,”*Proc. the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.

[2] J. G. Dias and J. K. Vermunt,“Latent class modeling of website users’ search patterns : Implications for online market segmentation,” *Journal of Retailing and Consumer Services 14*, pp. 359–368, 2007.