F-046

Webアクセス遷移に基づいた優良顧客度スコアリング
Trusted Customer Scoring based on the Web Access Transition

小阪 勇気† 広瀬 俊亮† 森永 聡† 藤巻 遼平† 山西 健司†
Yuki Kosaka Shunsuke Hirose Satoshi Morinaga Ryohei Fujimaki Kenji Yamanishi

1 はじめに

近年，EC（e-commerce）企業は，Webログ（顧客の行動履歴データ）を活用して，顧客一人一人を対象とするOne to Oneマーケティングを盛んに行っている．

ECサイトを訪れる顧客は，突然，購入するのではなく，1）ぼんやりと商品を閲覧し，2）購買意識が高まった後に，3）実際の購入に至ることが多い．そのため，顧客がどのレベルに属するかを見分けることは，EC企業にとって重要である．その中でも，2）に当たる，購入見込みのある顧客を発見して，積極的に売り込むことで効率よく売り上げを伸ばすことが期待できる．また，顧客の購入見込みは，高いときもあれば低いときもあるため，顧客の行動を常に追跡して，見込みが高くなったときを見逃してはならない．

本稿では，顧客ごとのコンテンツ閲覧ログ（クリックストリーム）をセッションに区切り，各セッションに対して，優良顧客度スコアを付与する手法を提案する．

購入見込み度合いを表す本優良顧客度スコアを基準にすることで，ぼんやりと商品を閲覧している顧客か，購買意識が高い顧客かを見分けることができるとともに，セッション単位にスコアを付与することで，見込みが高くなったときを見逃さない．

購買に至るセッションか否かを判定する問題には，Dimitrisら[1]が取り組んでいる．しかし，ECの実務上，「このセッションで購買に至らなくても，同一顧客が別のセッションで購買に至りそうかどうか判定する問題」を解くことが重要になる．この至りそうかどうかの度合いを本研究では，優良顧客度スコアリングと定義する．しかしながら，優良顧客度スコアリングは，「ぼんやりと商品を閲覧しているセッション」との判別が難しいと予想され，ほとんど研究されていない．

本稿では，優良顧客度スコアリング手法を提案し，ベンチマークデータを用いて本手法の有効性を検証する．

2 問題設定

本研究では，全セッション$\{\boldsymbol{y}_1,...,\boldsymbol{y}_M\} \in \mathcal{Y}$がWebログとして蓄えられているとする．Mは全セッション数を表す．ここで，j番目のセッション$\boldsymbol{y}_j$を$\boldsymbol{y}_j = (y_{j1} \rightarrow ... \rightarrow y_{jT_j}) \in U^{T_j}$とする．$U$はURL集合であり，例えば，$U$={/sports,/news,/shop,/cash,/thankyou,...}，/newsは，ニュースサイトのURLを表す．また，T_jはセッション長，$\rightarrow$は遷移を示す．

例えば，全セッションは，$\boldsymbol{y}_1$=(/sports → /news → /sports),...，$\boldsymbol{y}_M$=(/shop → /cash → /thankyou)となる．

† NEC共通基盤ソフトウェア研究所
NEC Common Platforms Software Research Labs.

全セッション集合$\mathcal{Y}$は，EC企業が保持する顧客のIDと，ECサイトでの購入経験の有無に関する情報とを紐付けることで，以下3つに分割できる．

A：優良顧客の購入に至るセッション集合：$\mathcal{Y}_A$

B：優良顧客の購入に至らないセッション集合：$\mathcal{Y}_B$

C：非優良顧客セッション集合：$\mathcal{Y}_C$

ここで，購入経験がある顧客を優良顧客と呼び，そのセッション集合を$\mathcal{Y}_T$とする．同様に，購入経験が無い顧客を非優良顧客と呼び，そのセッション集合を$\mathcal{Y}_{NT}$と定義すると，$\mathcal{Y}_T \cup \mathcal{Y}_{NT} = \mathcal{Y}$，$\mathcal{Y}_A \cup \mathcal{Y}_B = \mathcal{Y}_T$，$\mathcal{Y}_A \cap \mathcal{Y}_B = \emptyset$，$\mathcal{Y}_C = \mathcal{Y}_{NT}$，という関係が成り立つ．

$\mathcal{Y}_A$は，購入したことのある顧客（優良顧客）の実際の購入に至ったセッション集合，$\mathcal{Y}_B$は購入したことのある顧客（優良顧客）の成約ページに至らなかったセッション集合，$\mathcal{Y}_C$は購入したことがない顧客（非優良顧客）のセッション集合であり，購入に至ることはないが，購買意識が高い顧客の行動を含むセッションは，$\mathcal{Y}_B$と$\mathcal{Y}_C$にのみ存在することがわかる．そして，$\mathcal{Y}_B$は$\mathcal{Y}_C$に比べて，商品の購入を迷った行動を多く含み，$\mathcal{Y}_C$は閲覧しているだけの行動を多く含むと考えられる．

EC企業にとって，ある顧客のセッション$\hat{\boldsymbol{y}}$が，3つのどの状態に属するか特定できれば，顧客の状態に応じたきめ細かなコミュニケーション戦略が立てられるようになる．$\mathcal{Y}_A$に関しては，成約ページに至ったか否かを直接調べて判断できるが，購買意識が高い行動を多く含む$\mathcal{Y}_B$とただ閲覧しているだけの行動を多く含む$\mathcal{Y}_C$の，どちらに属するかの判別は難しい．

そこで，本研究では，BとCどちらに属するか分からない$\hat{\boldsymbol{y}}$に対して，優良顧客度スコア$s(\hat{y})$を算出することで，スコアが高ければBと判別し，低ければCと判別する，2値判別問題に取り組む．

3 優良顧客度スコアリング手法

3.1 アクセス遷移の学習

本稿では，セッション$\boldsymbol{y}_j$の確率的生起のモデルを混合隠れマルコフモデル（式（1），(2)）によって表現する[2]．式（2）は1次の隠れマルコフモデルを表す*．

$$P(\boldsymbol{y}_j|\theta) = \sum_{k=1}^{K} \pi_k P_k(\boldsymbol{y}_j|\theta_k) \quad (1)$$

*各$\boldsymbol{y}_j$がK個のクラスタからなる混合分布から発生していると仮定する．Kは固定．π_kはk番目のクラスタの生起確率．θは，$(\pi_1,...,\pi_{T_k},\theta_1,...,\theta_k)$．$x$は，隠れ変数で，$\gamma_k(x)$は，隠れ状態の初期確率．$a_k(.|.)$を対応する状態間の遷移確率とし，$b_k(.|.)$は，状態が与えられたときの，URLに対する条件付き確率を表す．パラメータθ_kは，$(\gamma_k(.), a_k(.|.), b(.|.))$．

$$P_k(\boldsymbol{y}_j|\theta) = \sum_{(x_1,..,x_{T_j})} \gamma_k(x_1) \prod_{t=1}^{T_j-1} a_k(x_{t+1}|x_t) \prod_{t=1}^{T_j} b_k(y_t|x_t) \tag{2}$$

T_j はセッション長を表し，可変とする．

3.2 尤度比によるスコアリング

本研究では，初めに，$\mathcal{Y}_\mathrm{B}$ と $\mathcal{Y}_\mathrm{C}$ を混合隠れマルコフモデルによって別々に学習し，$Model(\mathcal{Y}_\mathrm{B})$，$Model(\mathcal{Y}_\mathrm{C})$ を構築する．その後，B と C どちらに属するか分からない $\hat{\boldsymbol{y}}$ に対して，優良顧客度スコア $s(\hat{\boldsymbol{y}})$ を算出し，スコアが高ければ B，スコアが低ければ C と判別する問題を解く．そこで，

- 優良顧客の購入に至る前後のセッションモデル $Model(\mathcal{Y}_\mathrm{B})$ に対して尤もらしく，
- 非優良顧客セッションモデル $Model(\mathcal{Y}_\mathrm{C})$ に対して尤もらしくない，

$\hat{\boldsymbol{y}}$ に高いスコアを付けるように，式 (3) を用いて優良顧客度スコア $s(\hat{\boldsymbol{y}})$ を計算する．

$$s(\hat{\boldsymbol{y}}) = E^C(\hat{\boldsymbol{y}}) - E^B(\hat{\boldsymbol{y}}) = \frac{1}{\hat{T}} \log \frac{P^B(\hat{\boldsymbol{y}}|\theta^B)}{P^C(\hat{\boldsymbol{y}}|\theta^C)} \tag{3}$$

$$E^i(\hat{\boldsymbol{y}}) = -\frac{1}{\hat{T}} \log P^i(\hat{\boldsymbol{y}}|\theta^i) \tag{4}$$

ここで，$E^i(\hat{\boldsymbol{y}})$（式 (4)）は，$Model(\mathcal{Y}_\mathrm{i})$ に対する $\hat{\boldsymbol{y}}$ の異常度を表す．$E^i(\hat{\boldsymbol{y}})$ の値が小さい場合は，$Model(\mathcal{Y}_\mathrm{i})$ に対して尤もらしいことになる．$\hat{T}$ は $\hat{y}$ のセッション長で，θ^i を $Model(\mathcal{Y}_\mathrm{i})$ のパラメータとする．

式 (3) より，$\hat{\boldsymbol{y}}$ に対する優良顧客度スコア $s(\hat{\boldsymbol{y}})$ は，$Model(\mathcal{Y}_\mathrm{B})$ に対する $\hat{\boldsymbol{y}}$ の尤度 $P^B(\hat{\boldsymbol{y}}|\theta^B)$ と，$Model(\mathcal{Y}_\mathrm{C})$ に対する尤度 $P^C(\hat{\boldsymbol{y}}|\theta^C)$ の比を用いて算出する．

4 実験

ベンチマークデータを用いて本手法の精度を評価し，本手法の有効性を検証する．

実験データには，KDDCUP2000 のデータ [3] を使用した．本データは，2000/01/31-2000/03/31 に収集された，レッグウェアやレッグケア商品を販売する EC サイトの Web ログで，Click ログと Order ログの 2 種類提供されている．Click ログは，いつ，誰が，どのセッションで，どの URL にアクセスしたか分かるログで，Order ログは，いつ，誰が，どのセッションで購入したか分かるログである．ログに残る SessionID を使用してセッションデータを生成し，$\mathcal{Y}_\mathrm{A}$，$\mathcal{Y}_\mathrm{B}$，$\mathcal{Y}_\mathrm{C}$ の 3 種類に分けた．

本実験では，$\mathcal{Y}_\mathrm{B}$，$\mathcal{Y}_\mathrm{C}$ のデータを学習用とテスト用で 9:1 に分けて評価する，10-fold クロスバリデーションを行う．ここで，$\mathcal{Y}_\mathrm{A}$ は使用しない．データ数は，$\mathcal{Y}_\mathrm{B}$ のセッション数＝ 561，$\mathcal{Y}_\mathrm{C}$ のセッション数＝ 9141 とした．学習用データで，それぞれ $Model(\mathcal{Y}_\mathrm{B})$，$Model(\mathcal{Y}_\mathrm{C})$ を構築した後に，テスト用データに対して優良顧客度スコアを算出し，スコアが閾値を超えたら B，その他は C と判別する．評価指標には，ROC 曲線を用いた．

アクセス遷移に基づく本手法とアクセス遷移に基づかない単純な手法（ナイーブベイズモデル）を比較することで，アクセス遷移を考慮する本手法の有用性を調べる．

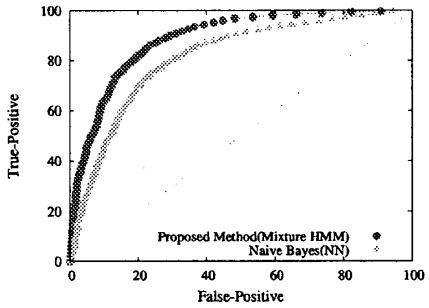

図 1: ROC 曲線

4.1 実験結果

実験結果を図 1 に示す．図 1 は，10 回実験を行ったときの平均値を示す．横軸は false positive rate，縦軸は true positive rate である．提案手法は，8%の false positive rate（FP）で 60%の true positive rate （TP）を達成している．

結果から，ある $\hat{\boldsymbol{y}}$ が $\mathcal{Y}_\mathrm{B}$，$\mathcal{Y}_\mathrm{C}$ どちらに近いか，本優良顧客度スコア $s(\hat{\boldsymbol{y}})$ を用いて高精度に判別可能であり，「このセッションで購買に至らなくても，同一顧客が別のセッションで購買に至りそうか否かを判定する」ことが可能であることを検証した．

さらに，提案手法は比較手法と比べて 9.19%精度が良いため（ROC 曲線の下側の面積の大きさを比較），アクセス遷移に基づいた提案手法は，遷移を考慮しない単純な手法に比べて，有用であることを検証した．

4.2 応用例：見込み顧客の発見

EC 企業は，優良顧客度スコアが閾値以上のセッションを持つ顧客を，このセッションで購買に至らなくても，近い将来購買に至りそうな見込み顧客と判断することで，この顧客に対して，積極的に売り込める．さらに，顧客一人ひとりのセッションごとにスコアを付与することで，購買意識が高まった瞬間を見逃さず，機会損失を減らせる．

5 おわりに

EC サイトを訪れる顧客のコンテンツへのアクセス遷移に基づき，各顧客に対して優良顧客度をスコアリングする手法を提案した．

Web ログの他に，POS データや電子マネーなどの購買履歴データにも適用可能である．

参考文献

[1] Dimitris Bertsimas, Adam J. Mersereau, and Nitin R. Patel. Dynamic classification of online customers. In *SDM*, pp. 107–118, 2003.

[2] 松永祐子, 山西健司. 情報理論的手法に基づく異常行動検出. 第 2 回情報科学技術フォーラム（FIT2003）予稿集 (情報. 技術レターズ), pp. 123–124, 2003.

[3] Ron Kohavi, Carla Brodley, Brian Frasca, Llew Mason, and Zijian Zheng. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explorations*, Vol. 2, No. 2, pp. 86–98, 2000. http://www.ecn.purdue.edu/KDDCUP.