1H5-4

インターネット広告におけるコンバージョンに近いユーザの抽出方法の検討

Extracting Segments of Users with High Conversion Probabilities for Internet Advertising

原　淳史 *1　高野　雅典 *2　Roman Shtykh*2　川端　貴幸 *1
Atsushi Hara　Masanori Takano　Takayuki Kawabata

*1株式会社サイバーエージェント　アドテクスタジオ
AdTech Studio, CyberAgent Inc.

*2株式会社サイバーエージェント　技術本部
Technical Department, CyberAgent Inc.

Recognizing users with high conversion probability is an important task for Internet Advertising. Advertisers/DSPs usually rely on the data provided by Data Management Platforms (DMPs) to create segments of users who are likely to make a purchase. Access frequency, dwell time. and number of visits to Web pages that are in proximity to the goal (conversion) page are the attributes often utilized to create such segments. However, because of large amounts of access logs, user segment generation may easily become an extremely human resource-consuming activity. In this research, we propose a model for automatically generating segments of users with high conversion probabilities based on their Web access history. We show that the proposed model yields user segments that outperform manually created segments with real-world access log data.

1. はじめに

1.1 インターネット広告

ここ数年，インターネットやスマートフォンの普及に伴い，インターネット広告への費用が増加している．インターネット広告の特徴の一つとして，特定のユーザ層へのターゲティングが可能なことが挙げられる．テレビなどの従来メディアの広告では，不特定多数のユーザへの一斉配信が基本であり目的によっては効率が悪かった．インターネット広告では Cookie を利用してユーザのオンライン行動を収集，分析することで，より自社サービスに興味を持ちそうなユーザに絞って広告を配信することが可能である．一般的にこのような広告は行動ターゲティングと呼ばれる．

行動ターゲティング広告では，コンバージョン（期待されるユーザアクション，例えば，広告主のオンラインサービスで商品購入というアクション）しそうなユーザを識別し，適切なセグメントとしてまとめることが重要である．セグメントの設計は，広告主のオンラインサービスを利用したユーザのアクセスログや顧客情報 (CRM) などが用いられるが，多くの広告主は DMP(Data Management Platform) と呼ばれるツールを利用することが一般的となっている．

1.2 DMP の役割

DMP は，広告主が保有するデータや外部のデータを組み合わせて，広告を含むマーケティング施策を効果的に行うための基盤である．DMP の機能は様々あるが主なものは，広告主のサイトに訪問してきたユーザのアクセスログの蓄積，蓄積したユーザアクセスログの可視化・分析，分析結果に基づくユーザセグメントの作成とセグメント単位での広告配信である．

DMP でのセグメント設計は人手で与えたルールに基づくものが一般的である．例えば，“商品をカートに入れてから 3 日以内” や，“Top ページに 1 週間で 5 回以上訪問” など特定の条件を満たすユーザをまとめることでセグメントを作成している．

連絡先: 原　淳史, 株式会社サイバーエージェント アドテクスタジオ, Email: hara_atsushi@cyberagent.co.jp

1.3 研究目的

従来の DMP のようなルールベースによるセグメント設計にはいくつか問題点がある．一番大きな問題として，そもそも適切なルールを手動で設計することが困難であることが挙げられる．ユーザに紐づく行動データや CRM データは無数にあり，そこから得られる変数は数十万以上になることが普通である．これらの変数をさらに組み合わせたり，期間や回数などの閾値も組み合わせるためルールは無限に生成可能である．そのような組み合わせ爆発の中から，手動で最適なルールを抽出することは現実的ではない．本研究は，広告主のサイトでのユーザのアクセスログを用いて，コンバージョンに近いユーザを自動的に抽出する手法について提案する．

提案手法では，コンバージョンに至るまでの一定期間内にとった行動と，コンバージョンに至らない一定期間内にとられた行動を素性として学習した分類モデルを作成し，任意のユーザの直近の行動からコンバージョンに至る確率を予測する．この予測確率をある閾値で切ることで，閾値より高い予測確率を持つユーザをコンバージョンに近いユーザセグメントとして自動生成することが可能となる．

2. 関連研究

インターネット広告に関連する研究として，広告のクリック・コンバージョンを予測するモデルが多く提案されている [Agarwal 07, Lee 12, Rosales 12]．一般的にコンバージョン予測はコンバージョンの正例が少ないため，クリック予測よりもずっと難しい．コンバージョン予測でよく利用されているモデルはロジスティック回帰である [Rosales 12, Lee 12]．ロジスティック回帰がよく利用される背景としては，学習が容易で，実際に予測するときの計算速度が速いことが挙げられる．DSP (Demand Side Platform) などの広告システムでは，100ms 以内にレスポンスを返すことが要求されるため，計算コストが小さいモデルが好まれる．本研究も先行研究 [Rosales 12, Lee 12] と同様，コンバージョン予測の際，ロジスティック回帰を用いる．その上で，本研究では，第 3. 章で後述するように，正例・負例のサンプリングと素性の抽出を工夫することで精度向上を

図っている．

また，膨大なアクセスログと会員情報からユーザを顧客層ごとにクラスタリングをし，購買予測を行う手法も多く提案されている [山口 14, 久松 13] ．先行研究 [山口 14, 久松 13] では，オンラインサービスに訪問したユーザを顧客層ごとに分類する手段として，アクセスログだけでなく，ユーザの性別や年齢などの属性情報なども利用している．

しかしながら，必ずしもそのような属性情報が利用できるわけではないので，今回我々は，比較的容易に取得可能なオンラインサービスのアクセスログだけを用いたコンバージョン予測モデルを提案している．

3. 提案手法

本章では提案するユーザのアクセスログに基づいたコンバージョン予測モデルについて述べる．最初に予測モデルの素性とするユーザ属性を定義した後，その素性を使用したコンバージョン予測モデルについて述べる．次にそのモデルの精度を大きく左右するであろうパラメータとして「行動データ利用期間」（後述）について記述する．最後にトレーニングデータ・テストデータの作成方法について述べる．

3.1 素性

本研究では，幅広く取得可能なユーザのアクセスログのみからそのユーザのコンバージョン率を予測することを目的とする．そのためにアクセスログから各アクセス時の URL と，アクセス情報から生成した定量値（PV 数，階層の深さ，滞在時間，一定期間内のコンバージョン数，直近のコンバージョンからの経過時間）を素性として用いる（図 1）．

1	2	3	4	…	N	PV 数	階層の深さ	滞在時間	一定期間の CV数	直近CVからの経過日数
1	1	1	0		0	7	4	25	1	10

図 1: コンバージョン予測に利用する素性とそのサンプル値

ユーザの各アクセスと素性の関係を図 2 に示す．アクセスログにはユーザの 1 回のアクセスごとにユーザ ID，アクセスした時間，アクセスした URL が記録される．各アクセス間の時間差が 30 分以内である場合，セッションとしてまとめる．

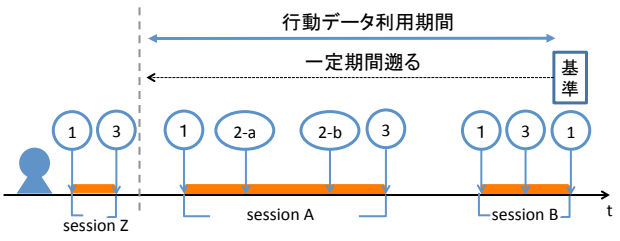

図 2: 素性の元になるユーザ行動の模式図．数直線は時間を表し，そこに向かう矢印はユーザの一回のアクセスを表す．各アクセスログにはユーザ ID，アクセスした URL，時間が記録される．行動データ利用期間外の session Z は素性の元にはならない．

- アクセス URL

  アクセス URL とはユーザがオンラインサービスにアクセスした際の URL である．これは各アクセスごとに記録される．アクセス URL に関する素性は利用する URL 種別分の長さのベクトルとし，各 URL の種別にアクセスした（1）か否か（0）を各次元の値とする（図 1 のカラム 1〜N を参照）．ただし，アクセスされたあらゆる URL を対象にすると URL の種別が膨大になるため，クエリパラメータを取り除いて扱った．また，1 日あたりの平均アクセス数が 1 回未満の URL は素性から削除した．図 2 の例では，2–a と 2–b はクエリパラメータのみが異なり同一とみなされるため，ユーザは 7 回のアクセスで URL 種別 1，2，3 の 3 種類の URL にアクセスしている．

- PV 数

  PV 数はユーザの行動データ利用期間における訪問した回数を表す．図 2 の例では，7 回のアクセスしているので PV 数は 7 である．

- 階層の深さ

  階層の深さとはユーザがオンラインサービスの利用の詳細さ・深さを表す指標である．一般に Web サイトは階層が深くなるほど，ディレクトリ構造も深くなるように設計されることが多いため，ここでは URL 中に含まれるスラッシュ（/）の数を階層の深さとみなす．本研究ではユーザの行動データ利用期間における最大の階層の深さを採用する．図 2 の例では，3 の URL が訪問した URL の中で最もスラッシュの数が多いため，階層の深さは 4 である．

- 滞在時間

  滞在時間は，ユーザの行動の中でオンラインサービスに訪問していた累計時間である．各セッションの最初のアクセスと最後のアクセス時間の差を各セッションの滞在時間として，それらの滞在時間の累計とした．図 2 の例では，sessionA に 20 分，sessionB に 5 分滞在しているので，滞在時間は 25 分である．

- 過去のコンバージョン数

  過去のコンバージョン数とは，オンラインサービスで過去にユーザがコンバージョンした回数である．

- 直近のコンバージョンからの経過日数

  直近のコンバージョンからの経過時間とは，ユーザがオンラインサービスで最後にコンバージョンしてから経過した時間である．

実際には，上記の素性の内の幾つかはカテゴリ化して学習に用いる．

3.2 行動データ利用期間の検討

本研究では，ユーザの行動を抽出するにあたり最新のアクセス時間から一定期間遡り，その期間から素性となる行動データを作成した（過去のコンバージョン数と直近のコンバージョンからの経過時間を除く）．本研究ではこの期間を「行動データ利用期間」と呼ぶ．

この遡る期間をどのように設定するかはコンバージョンするかどうかを予測する上で重要なパラメータである．なぜなら，様々なビジネスモデルや商品を持つ広告主が存在するため，その広告主のオンラインサービスを利用するユーザのコンバージョンのタイムスケールも異なると考えられるからである．例えば，日用品を扱う場合と不動産を扱う場合では，ユーザがオンラインサービスを利用開始して商品の購入を検討し，コンバージョンに至るまでの期間は大きく異ると考えられる．そのため，第 4. 章の実験では行動データ利用期間をパラメータとして変更し，予測精度への影響を考察した．

3.3 コンバージョン予測モデル

前述のユーザの行動データ（素性）からコンバージョンしやすさを予測するためにロジスティック回帰を利用する．目的変数はコンバージョンをしている場合は 1，コンバージョンしていない場合は 0 とし，素性は前節のものを使用する．本研究では，リッジ回帰によってパラメータを推定する．

3.4 データセット作成手法

ユーザのコンバージョンは頻度の高い行動ではないため正例と負例の比率には大きな偏りが存在する．そのため，本実験ではトレーニングデータを 正例 : 負例 = 1 : 3 になるようにネガティブサンプリングを行った．また効率よく満遍なく負例をサンプリングするため，正例と負例で異なったサンプリング方法を採用している（図 3）．トレーニングデータは 2ヶ月の期間からサンプリングした．

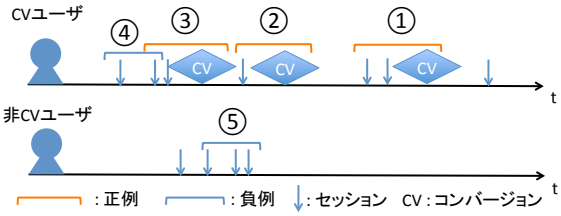

図 3: 正例，負例の例

- 正例

  正例とはコンバージョンしたユーザの行動データ（素性）を指す．具体的には各ユーザの行動データ利用期間の最終セッションにコンバージョン行動が存在する場合（図 3 の 1，2，3）に，その行動データ利用期間内のデータを正例とする．ただし，行動データ利用期間の範囲において複数回コンバージョンが発生した場合，直近のコンバージョン以降からコンバージョンまでの期間のデータを正例として用いる（図 3 の 2）．

- 負例

  負例とは，コンバージョンに至らなかった行動を指す．具体的には各ユーザの行動データ利用期間の最終セッションにコンバージョン行動が存在しない場合（図 3 の 4 と 5）に，その行動データ利用期間内のデータを負例とする．その結果，行動データ利用期間もその後もコンバージョンしていないデータ（図 3 の 5）だけでなく，行動データ利用期間の後にコンバージョンした「コンバージョンに近いもののしなかった」と言えるデータ（図 3 の 4）も負例に含まれる．また，PV 数が 1 しかないデータはコンバージョンしないことが容易に予測されるため PV 数が 2 以上のデータのみ使用した．

4. 実験結果

4.1 実験概要

提案手法の評価として 2 つの実験を行った．1 つは，行動データ利用期間がコンバージョン予測に与える影響を評価し，もう 1 つは手動のセグメントに対して，予測確率から作成されたセグメントの予測精度を比較評価する実験である．

各ユーザのコンバージョン率の予測はオンラインサービス内の最後のアクセスを基準にした行動を対象としている．行動データの取り方はトレーニングデータセットの作成方法と同じとする．ただし，一定期間の間にコンバージョンがある場合，基準のアクセスからそのコンバージョン直後のアクセスまでが行動データの対象となる．テストデータは 5 日間のアクセスログを使用して素性を作成し，正解はその直後 7 日間にコンバージョンの有無で判断した．

今回の実験では，5 社のオンラインサービス A〜E のアクセスログを用いた．これらのオンラインサービスは，一月あたり数十万–数百万のユニークユーザ，数百万–数千万のアクセス規模である．

4.2 行動データ利用期間による予測精度の違い

行動データ利用期間がコンバージョン予測に与える影響について評価した結果を示す．

行動データ利用期間を 1 日，3 日，5 日，7 日，14 日と変化させたときの各オンラインサービス A〜E のコンバージョン予測精度を Log Loss を用いて評価した．Log Loss の評価式は，

$$LogLoss = \frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})) \quad (1)$$

である．ここで，N は評価データ数，y_iは i 番目の評価データがコンバージョンした場合は 1，そうでない場合は 0，また,$\hat{y_i}$ はi 番目の評価データのコンバージョン予測確率を示す．

実験結果を図 4 に示す．横軸は行動データ利用期間を表し，縦軸は各オンラインサービスごとに最も予測値が悪かった行動データ利用期間を基準（100%）とした相対値で表している．例えば，オンラインサービス B では予測精度が最も低かったのは行動データ利用期間を 1 日としたときであり，そこから長くすることで予測精度は単調に上がり，7 日をピークに 14 日では予測精度が下がっていることが見てとれる．また，他のオンラインサービスでもそれぞれ，行動データ利用期間を変えたときに予測精度のピークが観測され，さらにその行動データ利用期間はサービスごとに異なることが分かる．

これらの結果から仮説通り，オンラインサービスごとに行動データ利用期間のパラメータを最適に決めることが，コンバージョン予測モデルにおいて重要と考えられる．ただし，オンラインサービス A だけは，5 日と 14 日と 2 つのピークが観測された．これは，オンラインサービス A に，複数の異なるコンバージョンのタイムスケールが存在したためと考えられる．その場合，コンバージョンの種別によりユーザをグルーピングし，グループごとに異なる行動データ利用期間を設定することで予測精度の改善が見込まれる．これについては今後の課題とする．

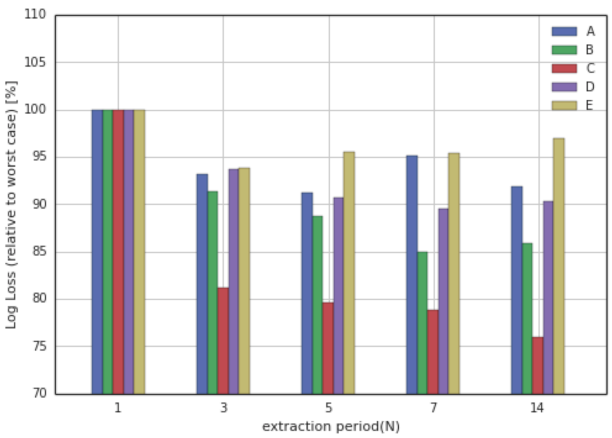

図 4: 行動の長さによるコンバージョン予測精度評価

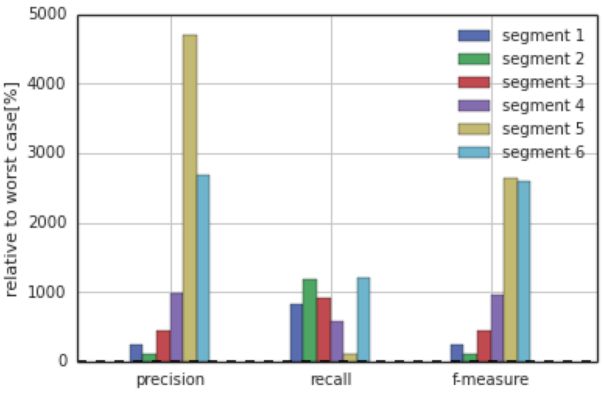

図 5: 手動セグメント（segment1–4）と自動セグメント（segment5–6）の精度（左），再現率（中），f 値（右）での比較.

4.3 手動セグメントと自動セグメントの比較

手動で設定したルールベースのセグメント（手動セグメント）と提案手法であるコンバージョン予測確率に基づいて作成したセグメント（自動セグメント）について比較評価した結果を示す.

通常，手動セグメントは一つのオンラインサービスに複数存在し，コンバージョンへの到達度を軸とした階層ごとに作成される．コンバージョンへの到達度が高いほど階層の深いセグメントになる．例えば，階層の深いセグメントの例として “商品をカートに入れて 3 日以内”，階層の浅いセグメントの例としては “Top ページに 1 週間以内に 1 回以上訪問” などである．手動セグメントと自動セグメントの比較実験は，前述のオンラインサービス D のアクセスログを用いて行った．実験結果を図 5 に示す．segment1–4 は手動セグメントを表し，segment1 が最も階層が浅く，segment4 が最も階層が深くなっている．また，segment5–6 は自動セグメントを表し，segment5 は，コンバージョン予測確率が 0.5〜1.0 のユーザを含めたセグメントであり，segment6 は，同じく 0.2〜1.0 のユーザを含めたセグメントとしている．このとき、行動データ利用期間のパラメータは前実験により最も予測精度の高かった 7 日を用いた．評価尺度としては，精度，再現率，f 値とし，前実験と同様，最も値が低かったものを基準とした相対値を縦軸に示している.

実験結果より，自動セグメントの方が f 値で 2 倍以上高い結果を示していることが分かる．特に，精度に関しては顕著であり，コンバージョンの予測モデルが効果的であったと言える．また，自動セグメントは，予測確率のどこで切り分けるかで精度と再現率がトレードオフとなるため，広告の予算に応じて決めることが望ましい.

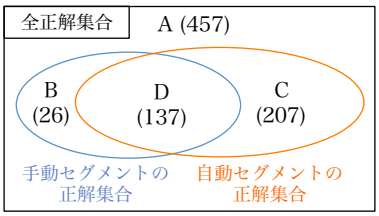

図 6: 手動セグメント (segment 4) と自動セグメント (segment 6) の正解ユーザ集合の関係．図内の括弧の数字は領域に含まれる正解ユーザ数を示す.

続いて，手動セグメントと自動セグメント間での正解ユーザの重複について考察をする．手動セグメントととして最も階層の深い segment4 と，自動セグメントで再現率の高い segment6 について，正解したユーザ集合の関係を図 6 に示す．ここから，自動セグメントは，手動セグメントの正解ユーザを 80.05% ($D/(B+D)$) 含んでいることが分かる．さらに，手動セグメントが正解できなかったユーザのうち 70.41% ($C/(A-B-D)$) をカバーすることできている．このことからも，コンバージョンの予測モデルが効果的であったと言える.

本実験により，提案手法であるコンバージョン予測確率に基づいて作成したセグメントは，手動で設定したルールベースのセグメントと比較して，コンバージョンに近いユーザを効率良く抽出できることを示した.

5. 終わりに

本研究において，自動的にセグメントの作成を可能にするために，オンラインサービス上のユーザのアクセスログから行動データの抽出方法とロジスティック回帰によるコンバージョン予測の手法を提案した．そして，提案手法を通じてオンラインサービスの特性によって行動データ利用期間が異なること，予測確率からセグメントを作成することで従来の手動で作成していたセグメントで予測していたコンバージョンユーザだけでなく，予測できていなかったコンバージョンユーザを予測できることを示した.

本研究では，行動データ利用期間のみをパラメータとした．この行動データ利用期間は，オンラインサービスの特質に影響されると考えられる．今後は，より予測精度を高める方法として，行動データ利用期間の最適な期間が決まる要因を明らかにしていき，オンラインサービスの特徴に応じて行動データ利用期間が最適に決まる方法を明らかにしていきたいと考えている.

参考文献

[Agarwal 07] Agarwal, D., Broder, A. Z., Chakrabarti, D., Diklic, D., Josifovski, V., and Sayyadian, M.: Estimating Rates of Rare Events at Multiple Resolutions, in *Proc. of KDD 2007*, pp. 16–25 (2007)

[Hoerl 70] Hoerl, A. E. and Kennard, R. W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, Vol. 12, No. 1, pp. 55–67 (1970)

[Lee 12] Lee, K., Orten, B., Dasdan, A., and Li, W.: Estimating Conversion Rate in Display Advertising from Past Performance Data, in *Proc. of KDD 2012*, pp. 768–776 (2012)

[Rosales 12] Rosales, R., Cheng, H., and Manavoglu, E.: Post-click Conversion Modeling and Analysis for Non-guaranteed Delivery Display Advertising, in *Proc. of WSDM 2012*, pp. 293–302 (2012)

[久松 13] 久松 俊道, 外川 隆司, 朝日 弓未, 生田目 崇：EC サイトにおける購買予兆発見モデルの提案, オペレーションズ・リサーチ：経営の科学, Vol. 58, No. 2, pp. 93–100 (2013)

[山口 14] 山口 景子：頻度の時間変化を考慮した階層ベイズモデルによるウェブサイト訪問行動の分析, マーケティング・サイエンス, Vol. 22, No. 1, pp. 13–29 (2014)

[川野 10] 川野 秀一, 廣瀬 慧, 立石 正平, 小西 貞則：回帰モデリングと L_1 型正則化法の最近の展開, 日本統計学会誌, Vol. 39, No. 2, pp. 211–242 (2010)