インターネット広告上の高次元でスパースな ユーザー行動ベクトルからのユーザー特徴獲 得モデルの研究

河本 哲

44

1 1.研究の背景

- 2 1994 年にオンラインマガジン HotWired にてバナー広
- 3 告が誕生して以来、情報技術の発展とともにコンテンツマ
- 4 ッチ配信や Web 訪問履歴などを利用したターゲティング
- 5 配信といった様々な広告配信手法が発展してきた。
- 6 特にユーザーの趣味嗜好に合わせた広告配信が可能に
- 7 なるターゲティング広告は、広告効果の高い配信手法とし
- 8 て注目されているが、その実現のためにはユーザー特徴を
- 9 適切なベクトルで表現する必要がある。しかし、Web上の
- 10 端末数は億のオーダーになり、広告及びメディアの数は数
- 11 十万の規模に上るため、一般的な手法ではベクトルの次元
- 12 が非常に大きなものとなる。加えて、広告がクリックされ
- 13 る確率はおおよそ 0.1 パーセント程度であり、広告をクリ
- 14 ックした人が商品購入などの成果地点に到達するのはさ
- 15 らにその数パーセント程度であるため、ベクトルのスパー
- 16 ス性が高くなってしまう問題がある。そのため効果の高い
- 10 八日が同くなりでしまプロ趣がある。でのため効木の同く
- 17 広告配信を実現するためには高次元でスパースなベクト
- 18 ルを効果的に低次元化するモデルが必要になる。
- 19 本研究ではユーザーの広告クリック情報および広告成
- 20 果測定地点への訪問 (コンバージョン) 情報を用いて、精
- 21 度と汎化性能の高い特徴量を獲得する手法を提案した。具
- 22 体的には、TF-IDF (Term Frequency and Inverse
- 23 Document Frequency) を用いてユーザーベクトルの重み
- 24 付けを行い、後述のスパースベクトル許容型オートエンコ
- 25 ーダを用いて、重み付けされたベクトルの次元を圧縮した。
- 26 また、獲得特徴量を用いてユーザー分類を行い、広告効果
- 27 の高いユーザー群が明確に検出されることを示した。

29 2.提案モデル

28

- 30 (1)TF-IDF によるユーザーベクトルの重み付け
- 31 前処理として、Web 上のユーザーを文書とみなし、広告
- 32 ID を単語インデックスとみなした TF-IDF によるユーザ
- 33 ーベクトルの重み付けを行った。例えば、あるユーザー ν
- 34 が広告 a を $x_{v,a}$ 回クリックして、広告 b にて $y_{v,b}$ 回コン
- 35 バージョンしたとする。TF-IDF による変換を掛ける前の
- 36 ユーザーベクトル \mathbf{X} は次のように表現される。

37
$$\mathbf{X}_{v} = (x_{v,1}, \dots, x_{v,N}, y_{v,1}, \dots, y_{v,M})$$

38 ここに、N をクリック実績のある広告数としM をコ

39 ンバージョン実績のある広告数とする。 \mathbf{X} に TF-IDF を

40 掛けたベクトル \mathbf{Y}_{ν} は、一般的にクリックされやすい広告

41 の重みが小さくなり、ユーザーを特徴付けるような広告へ

42 のアクションの重みが大きくなる。 $old Y_{_{\scriptscriptstyle arphi}}$ を後述のスパース

43 ベクトル許容型オートエンコーダを用いて次元圧縮する。

45 (2)スパースベクトル許容型オートエンコーダ

46 オートエンコーダは、入力層と出力層に同じデータを与

47 えることで中間層にデータの特徴を蓄積させるニューラ

48 ルネットワークであるが、欠点として、ユニット数が多く

49 なると計算時間が大きくなる問題がある。また \mathbf{Y} はその

50 成分のほとんどがゼロであるため、「任意の入力に対して

51 ゼロを返す」ような特徴学習をしてしまう問題点も存在す

52 る。これらの2点の問題点を解決するため、図1に示すよ

53 うに \mathbf{Y}_{ν} のゼロ成分を確率的に欠損値と見做してスキップ

54 させるモデルを提案する。これをスパースベクトル許容型

 $\mathbf{55}$ オートエンコーダと定義した。 $\mathbf{Y}_{\!\scriptscriptstyle V}$ のゼロ成分を全て欠損

56 値(NA)とせずに、一部明示的にゼロを与えることで「広告

57 A をクリックしたユーザーは広告 B をクリックしにくい」

58 といった非共起現象も学習させることが可能になる。

59 本研究では、第1段階としてユーザーベクトルを 28958

60 次元から 500 次元に圧縮し、第 2 段階でユーザーベクトル

61 を 500 次元から 146 次元に圧縮するような積層型のスパ

62 ースベクトル許容型オートエンコーダを用いた。また、入

63 力層と中間層の活性化関数はシグモイド関数とし、出力層

64 の活性化関数は恒等写像を用いた。

3.22 0(NA) 0 0 1 0(NA) 0(NA) 0(NA)

図1スパースベン トル許容型オートエンコーダ

67 68

66

65

69 (3)K-means によるユーザー分類

70 スパースベクトル許容型オートエンコーダを用いるこ

とにより、Web 上の各ユーザーの特徴をベクトル表現でき 71

るようになった。このベクトルを用いて K-means により

73 ユーザー分類をし、広告効果の高いユーザー群が検出され

74 るかどうかを評価した。

75

88

90

92

76 3.シミュレーションの構成と結果

77 (1)シミュレーションの概要

78 訓練データ(2016/6/1~2016/6/21のクリックログおよび

79 コンバージョンログ)を用いて、ユーザーベクトルを

80 TF-IDFによって重み付けし、スパースベクトル許容型オ

81 ートエンコーダによって次元圧縮後、圧縮された情報表現

82 をもとに K-means によりユーザーを分類する。ここで、

83 広告効果の高いユーザー群が抽出できているかどうかを、

84 テストデータ(2016/6/22~2016/7/12 のクリックログおよ

85 びコンバージョンログ)を用いて評価した。広告効果は

86 CVR(各ユーザー群のコンバージョン件数÷クリック件数)

を指標として評価した。 87

89 表1.訓練データおよびデストデータの統計量

	/ 1/10年1里	
項目	訓練データ	テストデータ
クリック実績のある広告数 N	23339	25646
コンバージョン実績のある広告数 M	5619	5552
クリック総数	5644230	5801623
コンバージョン総数	32020	38547
ユーザー数	3452491	3571259

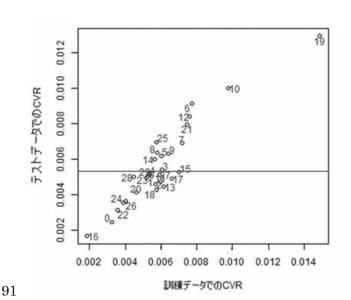


図2. 各ユーザー群のCVR

93 (2)シミュレーション結果

図2はユーザー群を30分類し、各ユーザー群の訓練デ 94

95 ータにおける CVR とテストデータにおける CVR を示して

96 いる。すなわち CVR の高い群は、なんらかのジャンルの

97 広告で、商品購入などの成果に到達しやすいユーザー群で

ある。訓練データで高い CVR を示したユーザー群はテス 98

99 トデータでも CVR が高く、提案モデルによって広告効果

100 の高いユーザー群が抽出出来ていることを示している。

102 4.結論と課題

101

114

103 本研究では、ユーザーがクリックおよびコンバージョン 104 した広告情報の次元を圧縮し、汎化性能の高い特徴を獲得 105 するモデルの考案と検証を行った。提案モデルによって、

106 広告効果の高いユーザー群と効果の低いユーザー群を明

107 確に分離することが可能になった。また、ニューラルネッ

108 トワークの計算量を軽減出来たことも意義が大きい。

109 今後は単純に広告効果の高いユーザー群を発見するだ 110 けではなく、各ユーザーの趣味嗜好の存在を検証すること

111 が重要な課題として挙げられる。また、ユーザーのサイト

訪問情報や滞在時間といった情報を加えてユーザー特徴

113 を獲得することも課題となる。

ザー行動がどのような特徴を持っているのかを詳しく知りたいと考えていました。また、普段の企業業務では取り組み辛い新規性の高い研究に取り組みたいと思ったことが、放送大学大学院への入学のきっかけでした。ゼミではユニークな研究テー

大学院在籍中は、秋光先生およびゼミの皆様に大変お世話に

2