

非テキスト情報のみを用いた AROW による 効率的な CTR 予測モデルの構築

林 佑磨[†] 諏訪 晴士^{††} 山名 早人^{†††}

[†] 早稲田大学基幹理工学研究科 〒169-8555 東京都新宿大久保 3-4-1

^{††} 株式会社 ジーニー 〒106-0032 東京都港区六本木 4-1-4 黒崎ビル 3F

^{†††} 早稲田大学理工学術院 〒169-8555 東京都新宿大久保 3-4-1

E-mail: {yuma.hayashi,yamana}@yama.info.waseda.ac.jp, s.suwa@geniee.co.jp

あらまし オンライン広告の市場規模は近年急速に拡大している。広告効果を高めるためには、広告クリック率 (Click Through Rate: CTR) の予測が重要である。これまでに CTR 予測に関する研究は広く行われてきたが、その多くは独自の検索エンジンで取得したクエリのテキスト情報を利用するものである。しかし、広告配信サービスを提供する企業の多くは検索エンジンを持たない。そこで本研究では広告配信サービスから取得可能な非テキスト情報のみを用いて CTR 予測を行う。CTR 予測モデルには高精度かつ高速なオンライン型分類器の AROW (Adaptive Regularization of Weight Vectors) を利用する。既存手法との比較を行った結果、精度を下げることなく予測モデル構築時間を常に 300 倍以上、最大で 2665 倍短縮することに成功した。

キーワード オンライン広告, CTR 予測, AROW, 非テキスト情報

1. はじめに

オンライン広告の市場規模は近年急速に拡大している。広告主はメディアに掲載料を支払うことで、オンライン広告の掲示を行う。オンライン広告を掲載する主な目的は、1) ブランド認知度の向上や、2) 販売促進などである。これまでは広告主とメディアが、一定の広告単価で期間契約を行うことが一般的であった。しかし、近年 RTB (Real-Time Bidding) [1] 技術が確立したことで、広告主は各インプレッション (ユーザが広告を閲覧すること) 単位での入札が可能となった。

広告掲載の課金モデルも目的に応じて、インプレッション毎に課金を行う CPM (Cost Per Mille) や、ユーザが広告をクリックした場合にのみ課金を行う CPC (Cost Per Click) などいくつかの種類がある。多くの場合、広告主はユーザが広告をクリックすることを望むため CPC 課金モデルが用いられる [3]。CPC 課金モデルを用いた場合、総費用は「広告主が設定した広告の入札額 × クリック数」となる。したがって、高い精度での CTR (Click Through Rate) 予測は、予算を考慮した入札額の設定を可能とし、最終的な収益の予測にもつながるため重要な課題となる。

学術的にも、これまでに数多くの CTR 予測に関する研究がなされてきた [2] [3] [4] [5] [6] [7] [8] [9]。これらの研究はその多くが独自の検索エンジンから取得したクエリのテキスト情報やユーザの情報を利用するものである。しかし、広告配信サービスを提供する企業の多くは検索エンジンを持たない。そのため、広告配信サービスが取得可能な情報のみを用いて CTR 予測を行う必要がある。また、CTR 予測は多くの場合機械学習を用いて行われるが、膨大なサンプル数と素性数の扱いが課題とな

る。そのため、膨大なデータに対して、CTR 予測モデルを高速に構築することも大きな課題である。

そこで本稿では、実際の広告配信サービスから取得したログに含まれる非テキスト情報のみを用いて CTR 予測モデルの構築を行う。取得したログに含まれる具体的な情報は、広告主側の情報、メディア側の情報、ブラウザの ID や閲覧・クリックなどのイベントが発生した時間情報などである。これらの情報単体およびそれらの組み合わせを素性として利用する。

また、膨大なサンプル数に対応し高速に CTR 予測モデルを構築するため、本手法では高精度かつ高速なオンライン型分類器である AROW (Adaptive Regularization of Weight Vectors) [10] を利用する。なお、膨大な素性数に関しては Feature Hashing [11] による特徴量圧縮を適用することで対応する。現在最も広く用いられている Maximum Entropy Model [12] をベースラインとして比較を行うことで、提案手法の有用性を示す。

本稿の構成は以下の通りである。2 節で関連研究について説明し、3 節で CTR 予測モデルと利用する素性について説明する。同節では膨大な素性を扱う際の工夫についても述べる。4 節で評価実験およびその考察を行い、5 節でまとめと今後の課題について述べる。

2. 関連研究

オンライン広告における CTR 予測は、特定の Web サイト上で表示される広告が、それを閲覧したユーザにクリックされるか否かを判定する二値分類の問題である。以下では従来研究で用いられている機械学習モデルおよび入力として用いられているデータについてまとめる。

CTR 予測にはしばしば機械学習における線形オンライン学習モデルが用いられ、その中でも Maximum Entropy Model を用いているものが多く存在する [2] [4] [7] [8] [9]。特に, Chapelle ら [2] の研究においては, Maximum Entropy Model を用いたシンプルで大規模データを扱うことが可能な, 実用性の高い予測モデルが提案されている。Maximum Entropy Model は Logistic Regression としても広く知られており, モデルのシンプルさと実装の容易さ, そしてオンライン処理であるため大規模データの扱いが可能であることから CTR 予測モデルとして適している。

Maximum Entropy Model 以外にも, probit model を用いた研究 [13] や boosted trees を用いた研究 [6], Ling らが提案した Coupled Group Lasso (CGL) を用いた研究 [3] などが存在する。CGL を用いた研究では Maximum Entropy Model との比較実験が行われており, 精度も上回っている [3]。ただし, CGL はモデルが持つ 2 つのハイパーパラメータの調整を行わなければならない。それ以外の複雑なモデルに関しては, 大規模データを用いる際の計算コストが大きくなることから, システムへの適用は容易ではない。

また, オンライン広告における CTR 予測の研究は, その多くが検索連動型広告 (Sponsored Search) に関するものである [5] [6] [7] [8] [14]。これは, 検索連動型広告においては CPC 課金モデルがしばしば用いられるためである [14]。検索連動型広告を対象とする場合, ユーザが検索する際に用いたクエリのテキスト情報を用いることが可能である。CTR 予測やそれに似た課題である CVR 予測 [15] [16] を行う際に, クエリのテキスト情報が素性として特に有効なものであるということは Chapelle らにより示されている [2]。クエリ以外にも, CTR 予測モデルの構築に用いるための素性に関する調査は行われており [2] [9], 中には静止画像やフラッシュの広告に着目した研究も存在する [4]。

提案手法では, 大規模データを用いた場合でも高速なモデル構築を実現するため, チューニングが容易で高精度かつ高速なオンライン型の分類器として知られる AROW (Adaptive Regularization of Weight Vectors) を用いた効率的な CTR 予測モデルの構築を行う。また, 独自の検索エンジンを持たない場合, 検索エンジンのクエリログは一般的に入手困難である。豊富な素性を用いる場合もそれらの素性が利用できることが前提となるが, これらの情報が取得不可能な場合も多く存在する。そこで提案手法では, クエリのテキスト情報や豊富なマルチメディア情報などを利用せずに, 基本的に入手可能なログ情報のみを素性とした CTR 予測モデルの構築を行う。

3. CTR 予測モデル

本節ではまず 3.1 で CTR 予測の問題設定, および CTR 予測モデルの構築時に利用する素性の説明を行う。ついで, 3.2 で膨大な素性数を扱う際に用いる Feature Hashing という特徴量圧縮技術について説明を行う。CTR 予測モデルに関しては, 3.3 でベースラインとする Maximum Entropy Model について説明し, 3.4 にて提案手法で用いる Adaptive Regularization

表 1: イベントログに含まれる素性の情報

素性クラス	素性の詳細
広告主情報	広告主 ID, キャンペーン ID, キャンペーンカテゴリ, 広告グループ ID, 広告 ID
メディア情報	サイト ID, ゾーン ID, ゾーンカテゴリ
ブラウザ情報	ブラウザ ID
時間情報	イベント発生時のタイムスタンプ

of Weight Vectors (AROW) について説明する。

3.1 問題設定と素性について

CTR 予測は, 広告がある時刻に特定のウェブサイトでユーザにより閲覧された後に, クリックされるかどうかを予測する二値分類の問題として考えることができる。この二値分類を行う CTR 予測モデルの構築は, ユーザによる広告閲覧のイベントログから作成される素性ベクトル \mathbf{x} , およびクリックされたかどうかの正解ラベル c (クリックされれば +1, されなければ -1) を用いて行う。

イベントログは 1) 広告主情報, 2) メディア情報, 3) ブラウザ情報, 4) 時間情報の 4 つの素性クラスに含まれる情報から構成されている。各素性クラスに属する素性の詳細を表 1 に示す。ただし, ゾーンカテゴリとは広告が配信された枠に対するカテゴリのことであり, キャンペーンカテゴリとゾーンカテゴリはそれぞれ IAB によって公開されている 26 カテゴリ (Tier 1 Categories) [17] である。また, ブラウザ ID は Cookie に紐付けられたブラウザ単位の ID である。

3.2 Feature Hashing

素性ベクトル \mathbf{x} は表 1 に示した素性単体やそれらの組み合わせ (デカルト積) から作成される。このため, 素性ベクトルはイベントログの数が多くなるに伴い, 膨大な次元数の疎なベクトルとなる。この問題を解決し, メモリ消費量を削減するため, 本手法では Feature Hashing [11] を用いる。

Feature Hashing は Hashing Trick と呼ばれるベクトルの次元数を削減するシンプルなアルゴリズムである。Feature Hashing はユニークな素性に hash 関数を適用することで, 精度をほとんど下げることなしに, d 次元ベクトルを $m \ll d$ 次元で表現することを可能とする。以下に Feature Hashing の擬似コードを示す。ただし, ハッシュ関数 h と ξ はそれぞれ $h: f \in F \rightarrow \{1, \dots, m\}$, $\xi: f \in F \rightarrow \{\pm 1\}$ である。なお, 提案手法では Feature Hashing により素性ベクトルを 24bit ($m = 2^{24} \approx 1,600$ 万) に圧縮した。

Algorithm 1 Feature Hashing

入力: hash 適用対象の素性集合 F , 圧縮後の次元数 m , hash 関数 h, ξ
出力: 圧縮後の m 次元素性ベクトル \mathbf{x}'

```

1:  $\mathbf{x}' \leftarrow \mathbf{0}$  ▷  $\mathbf{0}$  は  $m$  次元の零ベクトル
2: for  $f \in F$  do
3:    $i \leftarrow h(f) \bmod m$ 
4:    $x'_i \leftarrow x'_i + \xi(f)$  ▷  $x'_i$  は  $\mathbf{x}'$  の  $i$  番目の要素
5: end for
6: return  $\mathbf{x}'$ 

```

3.3 Maximum Entropy Model

Maximum Entropy Model [12] は Logistic Regression としても広く知られる分類器であり、イベントログがクリックにつながる確率は以下の式 (1) として表される。

$$\mathbb{P}(c = 1 | \mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})} \quad (1)$$

ここで、 c はクリックされたかどうかを表す ± 1 いずれかの値であり、 \mathbf{x} は素性ベクトル、 \mathbf{w} は求めるべき重みベクトルである。Maximum Entropy Model では、素性ベクトルとクリックの有無に関する正解ラベルの対が集合 $\mathcal{E} = \{(\mathbf{x}_i, c_i)\}_{i=1}^n$ として与えられた際に、式 (2) のように L_2 正則化付きの交差エントロピー誤差 (負の対数尤度) を最小化することで最適な重みベクトル \mathbf{w}^* を求める。

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \log(1 + \exp(-c_i \mathbf{w}^T \mathbf{x}_i)) \quad (2)$$

ここで、 $C > 0$ は正則化項の度合いを調整するためのパラメータであり、 L_2 正則化項は過学習を抑えるために加えられている。

3.4 Adaptive Regularization of Weight Vectors

提案手法では、オンライン学習の二値分類器である AROW (Adaptive Regularization of Weight Vectors) [10] を用いて CTR 予測モデルの構築を行う。オンライン学習は新たな素性ベクトルが与えられるたびに重みベクトルの更新を行うため、少ないメモリで大規模なデータを扱うことが可能である。AROW は CW (Confidence-Weighted) [18] と呼ばれる手法の拡張であるため、以下ではまず CW について説明し、その後に CW の欠点を解決した AROW の説明を行う。

3.4.1 CW

CW は AROW の元となるオンライン学習の二値分類器である。オンライン学習のため、素性ベクトル \mathbf{x}_t が与えられる度に予測ラベル \hat{c}_t を求め、正解ラベル c_t と比較することでモデルの重みベクトル \mathbf{w} を更新する。CW では重みベクトル \mathbf{w} が平均 $\boldsymbol{\mu} \in \mathbb{R}^m$ 、分散 $\Sigma \in \mathbb{R}^{m \times m}$ の正規分布 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ に従うと仮定されている。重みベクトルの中で分散値が大きいパラメータに関しては、まだ自信 (confidence) があまりない状態と考え、大きくパラメータを更新する。逆に、分散値が小さなパラメータに関しては、頻出な特徴のためにもう既に十分な情報が得られていると考え、小さくパラメータを更新する。実際にラベルを推定する際には、重みベクトルの期待値 $\mathbb{E}[\mathbf{w}] = \boldsymbol{\mu}$ を用いて行う。

t 回目の学習において、学習用の素性ベクトル \mathbf{x}_t および正解ラベル c_t が与えられたとする。この際、CW は以下の式 (3) に示す制約付き最適化問題を解くことで、重みベクトルに関する分布を更新し、重みベクトルの新たな平均 $\boldsymbol{\mu}_t$ と分散 Σ_t を得る。

$$\begin{aligned} (\boldsymbol{\mu}_t, \Sigma_t) &= \arg \min_{\boldsymbol{\mu}, \Sigma} \text{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \parallel \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) \\ \text{s.t. } &\mathbb{P}_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)}[c_t(\mathbf{w} \cdot \mathbf{x}_t) \geq 0] \geq \eta \end{aligned} \quad (3)$$

ここで、 $\mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})$ は t 回目の学習による更新を行う前の重みベクトルに関する分布であり、 $\text{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \parallel \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1}))$ は $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ と $\mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})$ 間のカルバック-ライブラーダイバージェンスである。また、 $\eta \in (0.5, 1]$ は更新の度合いを制御するハイパーパラメータである。

式 (3) の制約条件において、 $\mathbb{P}_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)}[c_t(\mathbf{w} \cdot \mathbf{x}_t) \geq 0]$ は与えられた素性ベクトル \mathbf{x}_t に対して正しくラベル予測される確率を表している。したがって、CW は $\eta \in (0.5, 1]$ 以上の確率で正しく分類されるという条件を満たした上で、更新前の重みベクトルの正規分布に最も近い正規分布を求めることで学習を行う。この制約条件は、与えられた素性ベクトル \mathbf{x}_t を常に正しく分類出来るようにモデルを更新することを意味するため、ノイズデータに極めて弱く過学習を起しやすという欠点が存在する。

3.4.2 AROW

AROW は、上記の CW が持つ欠点を、制約条件を目的関数の一部に正則化項として持たせることにより解決した手法である。具体的には、以下の式 (4) に示す最適化問題を解くことで、重みベクトルに関する分布を更新する。

$$\begin{aligned} (\boldsymbol{\mu}_t, \Sigma_t) &= \arg \min_{\boldsymbol{\mu}, \Sigma} \text{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \parallel \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) \\ &\quad + \frac{1}{2r} \ell_{h^2}(c_t, \boldsymbol{\mu} \cdot \mathbf{x}_t) + \frac{1}{2r} \mathbf{x}_t^T \Sigma \mathbf{x}_t \end{aligned} \quad (4)$$

ここで、 $r > 0$ はモデルの更新を調節するハイパーパラメータである。式 (4) は 3 つの項から構成され、それぞれの項は以下の様な意味を持つ。

- (1) $\text{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \parallel \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1}))$
この項を小さくすることは、パラメータの更新を小さく抑え、更新前の重みベクトルの正規分布に最も近い正規分布を求めることを意味する。
- (2) $\frac{1}{2r} \ell_{h^2}(c_t, \boldsymbol{\mu} \cdot \mathbf{x}_t)$
 $\ell_{h^2}(c_t, \boldsymbol{\mu} \cdot \mathbf{x}_t) = (\max\{0, 1 - c_t(\boldsymbol{\mu} \cdot \mathbf{x}_t)\})^2$ は二乗ヒンジ損失である。この項を小さくすることは、現在与えられているデータに対する予測間違いをなるべく少なくするような重みベクトル \mathbf{w} の平均 $\boldsymbol{\mu}$ を求めることを意味する。ただし、この項には二乗ヒンジ損失関数以外の損失関数を適用することも可能である。
- (3) $\frac{1}{2r} \mathbf{x}_t^T \Sigma \mathbf{x}_t$
この項を小さくすることは、重みベクトル \mathbf{w} の各素性に関する分散 (自信のなさ) を、学習を進めるにつれて小さくしていくことを意味する。

以上から、AROW は 1) \mathbf{w} の分布を今までの正規分布になるべく近く、2) 現在の学習データを正しく分類し、3) \mathbf{w} の各素性に関する自信を少しずつ上げていくことで、CW の欠点であったノイズのあるデータに頑健なオンライン学習を実現している。

表 3: 各データセットに含まれるユニークな素性数

	$\gamma = 0.1\%$		$\gamma = 1.0\%$	
	学習データ	テストデータ	学習データ	テストデータ
広告主 ID	55.8 \pm 0.8	44.0 \pm 0.0	60 \pm 1	45 \pm 1
キャンペーン ID	71.4 \pm 0.9	49.0 \pm 0.0	76 \pm 2	50 \pm 1
広告グループ ID	273 \pm 5	110 \pm 1	428 \pm 4	115 \pm 1
広告 ID	1,602 \pm 4	1,148 \pm 5	2,030 \pm 4	1,261 \pm 2
サイト ID	683 \pm 3	491 \pm 3	772 \pm 4	607 \pm 3
ゾーン ID	1,861 \pm 2	1,104 \pm 8	2,193 \pm 8	1,522 \pm 10
ブラウザ ID	499,066 \pm 230	83,690 \pm 57	2,298,012 \pm 391	449,782 \pm 303

4. 評価実験

4.1 データセット

CTR 予測モデルの構築には、株式会社ジーニーが運営する Geniee SSP^(注1) により 2014 年 11 月 1 日から 2014 年 11 月 8 日までの間に配信された実際のシステムログの一部を利用した。なお、訓練データには前半 7 日間のログを利用し、テストデータには続く 1 日分のログを利用した。ただし、訓練データに関しては、7 日間分の時系列情報は考慮していない。また、オンライン学習ではデータの出現順に影響を受けるため、訓練・テストデータ共にシャッフルして用いた。訓練データとテストデータに含まれるイベントがその後クリックされたかどうかの正解情報に関しては、特別なトラッキング ID を用いて紐付けを行い、1 日以内にクリックを行ったかどうかで設定を行った。

本実験ではそれぞれのデータセットに対して、クリックされたログである正例は全て用い、クリックされなかったログである負例を $\gamma = 0.1\%$ および $\gamma = 1.0\%$ の割合でそれぞれランダムサンプリングを行うことで、最終的に学習およびテストで用いるデータセットを作成した。最終的なデータセットに含まれる正例と負例の数をそれぞれ表 2 に示す。

表 2: 各データセットに含まれる正例と負例の数

	$\gamma = 0.1\%$		$\gamma = 1.0\%$	
	学習データ	テストデータ	学習データ	テストデータ
正例数	265,987	39,205	265,987	39,205
負例数	479,566	69,736	4,814,080	699,982
合計	745,553	108,941	5,080,067	739,187

なお、負例サンプリングによる分散を考慮するため、各 γ に関して、それぞれ 5 回ランダムサンプリングを行い、データセットを作成した。本稿では以降、各 γ に対して、作成した 5 個のデータセットから得られる平均値と、 \pm 以降に信頼度 95% の t 分布により求めた誤差を表記する。

表 1 に示したカテゴリ情報と時間情報以外の素性に関して、各データセットに含まれるユニークな個数を表 3 に示す。

また、各 γ に対して、テストデータに出現したユニークな素性が、それ以前の過去 d 日間分の学習データに出現している割合は図 1 および図 2 のようになっている。図 1 と図 2 から、最

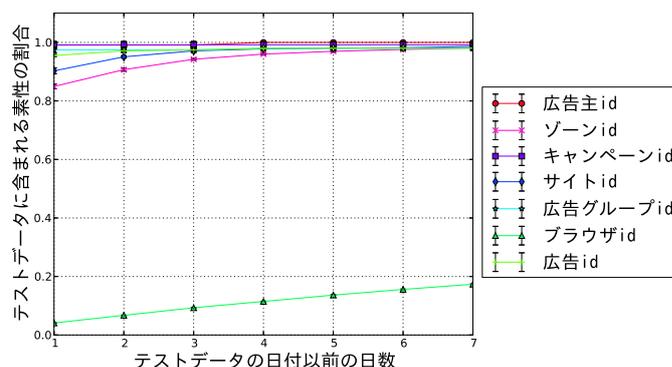


図 1: $\gamma = 0.1\%$ のテストデータに出現したユニークな素性において、テストデータに用いた日付から過去 d 日間の間に、学習データ内に出現しているユニークな素性の割合

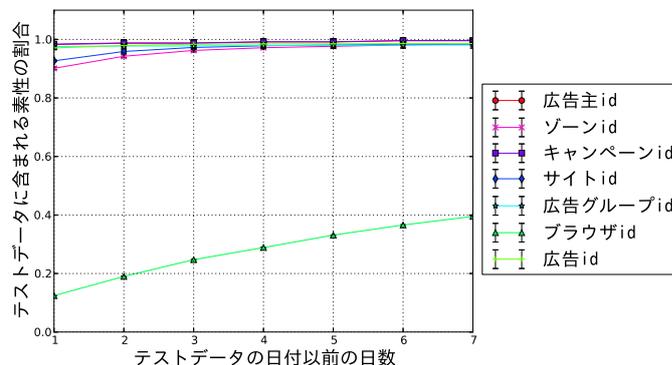


図 2: $\gamma = 1.0\%$ のテストデータに出現したユニークな素性において、テストデータに用いた日付から過去 d 日間の間に、学習データ内に出現しているユニークな素性の割合

も頻繁にユニークな新規 ID が増えているのはブラウザ ID で、過去 7 日分を学習データとして用いた場合でも、 $\gamma = 0.1\%$ で約 2 割、 $\gamma = 1.0\%$ で約 4 割の新規ブラウザ ID のみがテストデータに含まれている。一方、広告主 ID とキャンペーン ID はいずれもテストデータの日付の前日分のログにおいてほぼ全ての ID がカバーされており、最も変化していないことがわかる。評価実験においては、以下の (i)~(iv) の 4 つのルールで素性ベクトルを作成した。ただし、以下のルールにおいて素性の組み合わせとは、2 つの素性 f_1, f_2 のデカルト積 ($f_1 \times f_2$) として作られる組み合わせを新たな素性とするものである。具体的には、(広告主 ID \times サイト ID) などである。

(注1) : Geniee SSP: http://geniee.co.jp/publishers_pc.html

表 4: $\gamma = 0.1\%$ のデータセットにおける CTR 予測の結果

	NC24		NC4		NCNT		FC24	
	ME	AROW	ME	AROW	ME	AROW	ME	AROW
Accuracy (%)	73.54 \pm 0.04	74.04 \pm 0.03	73.51 \pm 0.05	74.03 \pm 0.04	73.50 \pm 0.03	74.03 \pm 0.04	73.63 \pm 0.05	73.65 \pm 0.06
Precision	0.6811 \pm 0.0008	0.6845 \pm 0.0008	0.680 \pm 0.001	0.685 \pm 0.001	0.6804 \pm 0.0008	0.684 \pm 0.001	0.679 \pm 0.001	0.676 \pm 0.001
Recall	0.4980 \pm 0.0008	0.5168 \pm 0.0006	0.498 \pm 0.002	0.5162 \pm 0.0006	0.497 \pm 0.001	0.5163 \pm 0.0004	0.507 \pm 0.001	0.5154 \pm 0.0006

表 5: $\gamma = 1.0\%$ のデータセットにおける CTR 予測の結果

	NC24		NC4		NCNT		FC24	
	ME	AROW	ME	AROW	ME	AROW	ME	AROW
Accuracy (%)	94.77 \pm 0.01	94.820 \pm 0.003	94.779 \pm 0.003	94.822 \pm 0.003	94.73 \pm 0.02	94.816 \pm 0.002	94.769 \pm 0.005	94.763 \pm 0.007
Precision	0.60 \pm 0.01	0.625 \pm 0.003	0.605 \pm 0.005	0.628 \pm 0.003	0.54 \pm 0.02	0.620 \pm 0.002	0.574 \pm 0.003	0.546 \pm 0.005
Recall	0.0448 \pm 0.0008	0.0582 \pm 0.0008	0.0449 \pm 0.0008	0.0584 \pm 0.0008	0.045 \pm 0.001	0.05823 \pm 0.0004	0.053 \pm 0.002	0.0758 \pm 0.0008

(1) 素性の組み合わせを行わないもの (No Conjunction: NC)

- (i) NC24: 1 日を 1 時間毎の 24 分割にして扱ったもの
- (ii) NC4: 1 日を 6 時間毎の 4 分割にして扱ったもの
- (iii) NCNT: 時間情報を利用しなかったもの

(2) 素性に、全ての組み合わせ素性を追加したもの (Full Conjunction: FC)

- (iv) FC24: 1 日を 1 時間毎の 24 分割にして扱ったもの

4.2 評価実験

本節では、ベースラインを Maximum Entropy Model (ME), 提案手法を AROW とし, 4.1 で説明した各サンプリング率 γ のデータセットに対する CTR 予測モデル構築とテストを行い評価をする. ただし, ME のハイパーパラメータ C と, AROW のハイパーパラメータ r は, それぞれ訓練データを 10 分割し 10-fold cross validation により設定を行った.

評価指標には Accuracy, および Precision と Recall を用いる. また, 効率性を検証するため, モデル構築時間の評価も行う. モデル構築に用いた計算機環境を表 6 に示す. ただしモデル構築時に使用したコア数は 1 つであり, 並列化は行っていない.

表 6: CTR 予測モデルの構築に用いた計算機環境

Processor	Intel Xeon E7-4850 2.00GHz \times 4
Memory	512GB

表 4 と表 5 に, 各サンプリング率 γ に対するそれぞれのデータセットに対して, AROW および ME を用いて CTR 予測を行った結果を示す.

これらの結果から, γ によらず, FC24 では Accuracy に関して AROW と ME で同等の精度が出ており, それ以外のデータセットでは AROW が ME を 0.1% ~ 0.5% 程度上回る結果となった. Precision と Recall に関しても, FC24 のデータセットとそれ以外のデータセットで少し異なる傾向が現れた. まず, FC24 以外のデータセットに関しては, AROW は Precision と Recall 両方において常に ME を上回っている. Precision では, $\gamma = 0.1\%$ の場合に平均 0.004, 最大 0.005 上回り, $\gamma = 1.0\%$ の場合に平均 0.043, 最大 0.080 上回った. Recall では, $\gamma = 0.1\%$

表 7: $\gamma = 0.1\%$ における CTR 予測モデルの構築時間 (秒)

	NC24	NC4	NCNT	FC24
ME (ベースライン)	356 \pm 75	329 \pm 79	297 \pm 35	3428 \pm 579
AROW (提案手法)	0.6 \pm 0.1	0.60 \pm 0.09	0.6 \pm 0.1	1.5 \pm 0.1

表 8: $\gamma = 1.0\%$ における CTR 予測モデルの構築時間 (秒)

	NC24	NC4	NCNT	FC24
ME	791 \pm 161	971 \pm 83	891 \pm 79	19188 \pm 2998
AROW	2.38 \pm 0.09	2.29 \pm 0.07	2.24 \pm 0.07	7.2 \pm 0.2

の場合に平均と最大ともに 0.019 上回り, $\gamma = 1.0\%$ の場合に平均と最大ともに 0.013 上回った.

FC24 のデータセットに関しては, いずれの γ の値においても, Precision は AROW の方が低く Recall は AROW の方が高いという結果が得られた. 特に, $\gamma = 1.0\%$ の場合は, AROW は Precision に関して ME を 0.028 下回り, Recall に関して ME を 0.023 上回る結果となり, FC24 以外のデータセットと比べると大きな差がでている.

以上から, AROW はクリックに至った正例の特徴を良く学習し, より多くの正例を正しく識別していることが分かる. ただし, 組み合わせ素性を用いた場合は, 豊富な素性から正例の特徴を過敏に学習し, 結果としてより多くのサンプルを正例に誤識別してしまった可能性が高い. また, 時間情報の有無は今回の結果にほとんど影響を与えていない. これは, 時間情報はクリックの有無を判断する上で本質的な特徴量ではなく, 不要な素性の追加でモデルが複雑になったため, あるいは時間情報の入れ方を工夫する必要があるためだと考えられる.

つぎに, 各サンプリング率 γ におけるそれぞれのデータセットに対して, ME と AROW それぞれを用いて CTR 予測モデルを構築した際にかかった時間を表 7 および表 8 に示す. この結果から, 構築時間の平均値で比較した場合, AROW は ME と比べいずれの場合も 300 倍以上速く, 最大で 2665 倍速くモデルの構築が行えていることがわかる.

また, それぞれの γ に対する各データセットに関して, AROW のハイパーパラメータ r に対する精度, 適合率と再現率, およびモデル構築時間の関係をそれぞれ図 3 から図 14 に示す.

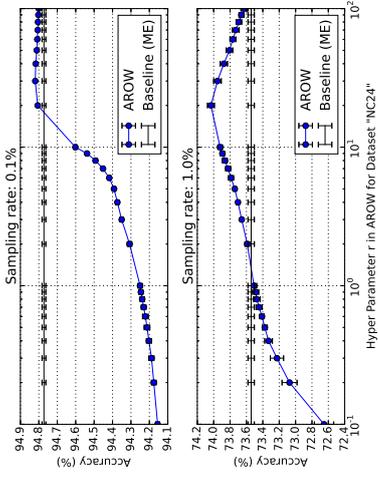


図 3: NC24 に対する r 毎の精度

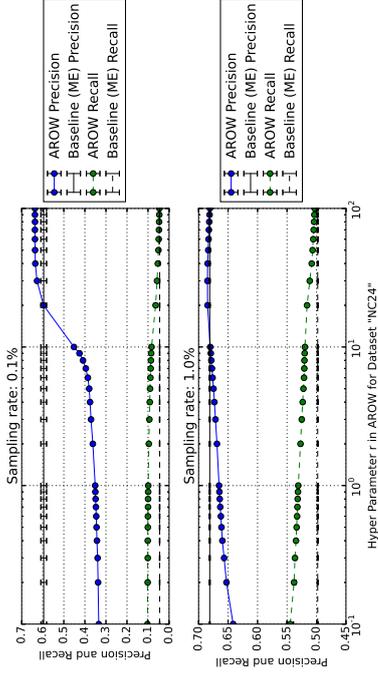


図 4: NC24 に対する r 毎の適合率と再現率

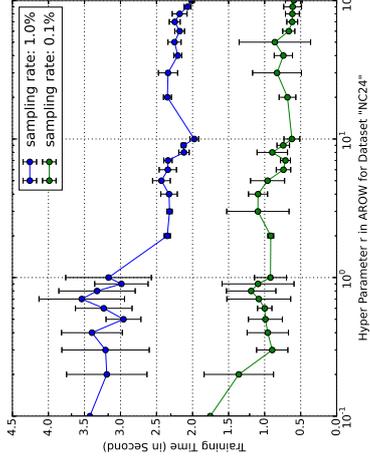


図 5: NC24 に対する r 毎のモデル構築時間

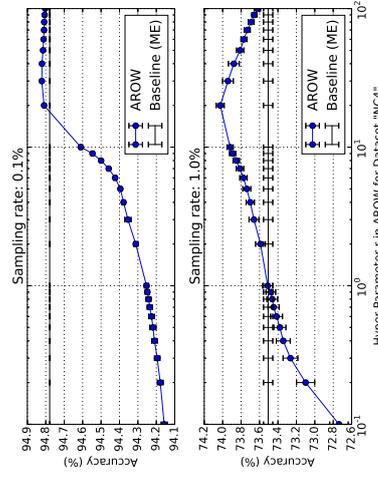


図 6: NC4 に対する r 毎の精度

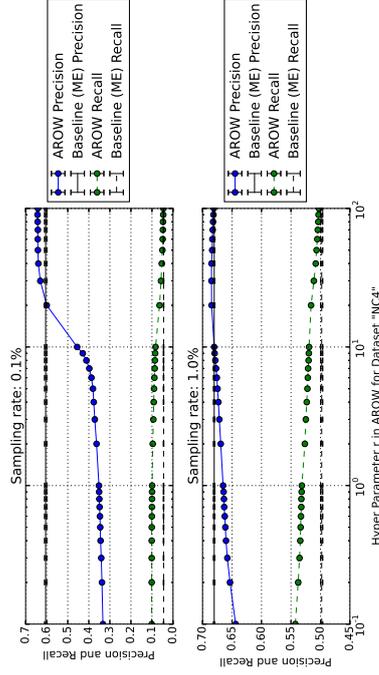


図 7: NC4 に対する r 毎の適合率と再現率

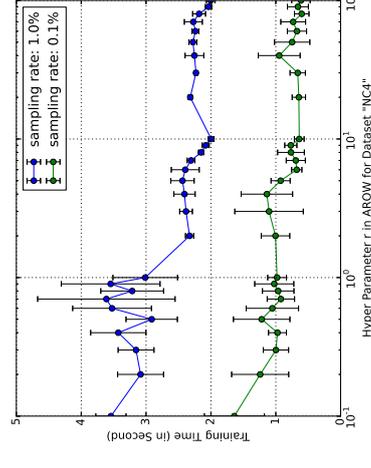


図 8: NC4 に対する r 毎のモデル構築時間

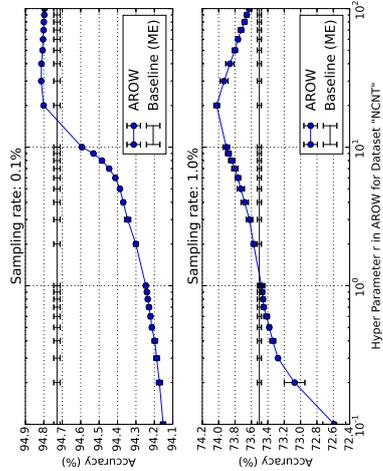


図 9: NCNT に対する r 毎の精度

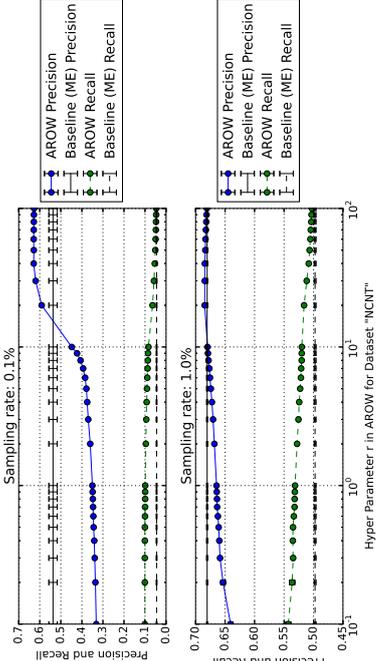


図 10: NCNT に対する r 毎の適合率と再現率

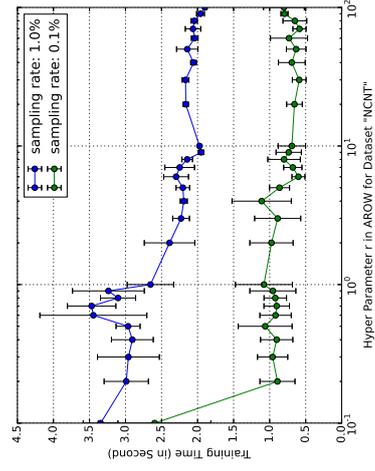


図 11: NCNT に対する r 毎のモデル構築時間

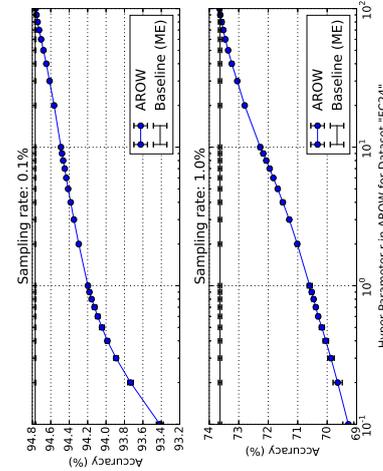


図 12: FC24 に対する r 毎の精度

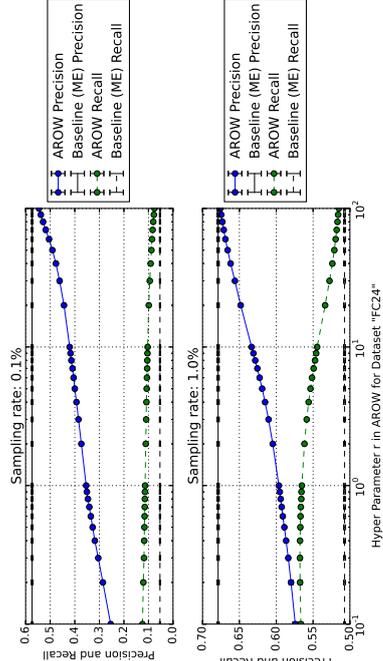


図 13: FC24 に対する r 毎の適合率と再現率

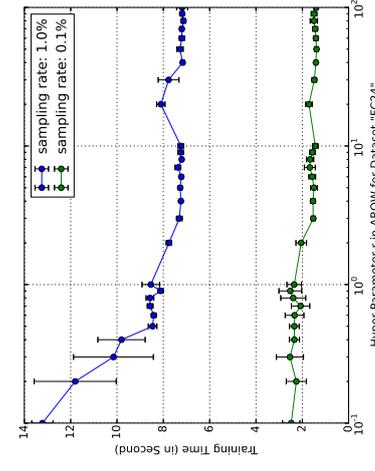


図 14: FC24 に対する r 毎のモデル構築時間

上図から、AROW のハイパーパラメータ r は 10 から 20 の間で Accuracy および Precision と Recall が大きく向上していることが分かる。 r は大きいほど、今までの識別器の状態を維持する。このため、素性数が膨大な FC24 では、 r が小さい場合は各サンプルに対して過敏に学習していると考えられる。

また、 r が大きな場合はより速く収束するため、モデルの構築時間もより速くなっている。 Accuracy および Precision と Recall を総合的に考慮すると、本データセットに関する CTR 予測では、AROW のハイパーパラメータ r は概ね $r = 40$ 前後でチューニングを行うと良い結果が得られることがわかった。ただし、組み合わせ素性を用いる場合のように、素性数が多い際には、 r を 100 以上と大きくした方が良い結果が得られる。

いずれの r においても、AROW によるモデル構築の時間は極めて短く、膨大なログデータに対してもチューニングを容易に行うことが可能であると言える。このことは、多くの機械学習モデルが、モデルの持つハイパーパラメータの調整が困難であり、時間を要する観点からも大きな利点であると言える。

5. まとめ

本稿では、CTR 予測で広く用いられてきた Maximum Entropy Model の代わりに、高速かつ高精度なオンライン学習型分類器の AROW を適用することで、効率的な CTR 予測モデルの構築を実現した。広告配信サービスから取得した非テキスト情報のみを用いて、既存手法との比較評価を行った結果、各データセットに関して既存手法と同等以上の Accuracy が得られ、最大で約 0.5% 上回った。その上で、提案手法では既存手法を用いた場合より常に 300 倍以上速く、最大で 2665 倍速く CTR 予測モデルの構築を行うことに成功し、精度およびモデル構築の速さの観点において有用性を示すことができた。

今後の課題としては、1) 最適な素性選択と、2) より効率的な識別器の適用が挙げられる。1) に関して、本実験では時間情報が精度に影響しなかったが、その他の素性と組み合わせることによって有益な特徴となる可能性がある。このため、素性に関する更なる実験が必要である。2) に関しては、AROW の発展型アルゴリズムとして知られる SCW (Exact Soft Confidence-Weight Learning) [19] などを適用し、より高速かつ高精度な CTR 予測モデルの構築を行うことが挙げられる。

文 献

- [1] Shuai Yuan, Jun Wang, and Xiaoxue Zhao: “Real-time Bidding for Online Advertising: Measurement and Analysis”, In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD’13, 2013.
- [2] Oliver Chapelle, Eren Manavoglu, and Romer Rosales: “Simple and scalable response prediction for display advertising”, *ACM Transactions on Intelligent Systems and Technology*, 2013.
- [3] Ling Yan, Wu-Jun Li, Gui-Rong Xue, and Dingyi Han: “Coupled Group Lasso for Web-Scale CTR Prediction in Display Advertising”, In *Proceedings of the 31st International Conference on Machine Learning*, ICML’14, 2014.
- [4] Haibin Cheng, Roelof van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou, and Vidhya Navalpakkam: “Multimedia features for click prediction of new ads in display advertising”, In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD’12, 2012.
- [5] H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Philips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica: “Ad Click Prediction: a View from the Trenches”, In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD’13, 2013.
- [6] Ilya Trofimov, Anna Kornetova, and Valery Topinskiy: “Using boosted trees for click-through rate prediction for sponsored search”, In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, ADKDD’12, 2012.
- [7] Jun Feng, Jiang Bian, Taifeng Wang, Wei Chen, Xiaoyan Zhu, and Tie-Yan Liu: “Sampling dilemma: towards effective data sampling for click prediction in sponsored search”, In *Proceedings of the 7th ACM international conference on Web search and data mining*, WSDM’14, 2014.
- [8] Haibin Cheng and Erick Cantú-Paz: “Personalized click prediction in sponsored search”, In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM’10, 2010.
- [9] 田頭 幸浩, 山本 浩司, 小野 真吾, 塚本 浩司, 田島 玲: 「オンライン広告における CTR 予測モデルの素性評価」, 第 7 回データ工学と情報マネジメントに関するフォーラム, DEIM’13, 2013.
- [10] Koby Crammer, Alex Kulesza, and Mark Dredze: “Adaptive regularization of weight vectors”, In *Advances in Neural Information Processing Systems*, NIPS’09, 2009.
- [11] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg: “Feature hashing for large scale multitask learning”, In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML’09, 2009.
- [12] Adam L. Berger, Vincent J. Della Pietra, Stephen A. Della Pietra: “A maximum entropy approach to natural language processing”, *Computational linguistics*, Vol. 22, No. 1, pp.39-71, 1996.
- [13] Thore Graepel, Joaquin Quiñero Candela, Thomas Borchert, and Ralf Herbrich: “Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft’s bing search engine”, In *Proceedings of the 27th International Conference on Machine Learning*, ICML’10, 2010.
- [14] Dawei Yin, Bin Cao, Jian-Tao Sun, and Brian D. Davison: “Estimating ad group performance in sponsored search”, In *Proceedings of the 7th ACM international conference on Web search and data mining*, WSDM’14, 2014.
- [15] Rómer Rosales, Haibin Cheng, and Eren Manavoglu: “Post-click conversion modeling and analysis for non-guaranteed delivery display advertising”, In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM’12, 2012.
- [16] Amr Ahmed, Abhimanyu Das, and Alexander J. Smola: “Scalable hierarchical multitask learning algorithms for conversion optimization in display advertising”, In *Proceedings of the 7th ACM international conference on Web search and data mining*, WSDM’14, 2014.
- [17] Interactive Advertising Bureau (IAB): “Open RTB API Specification Version 2.2”, http://www.iab.net/media/file/OpenRTBAPISpecificationVersion2_2.pdf, 2015 年 1 月 4 日アクセス.
- [18] Mark Dredze, Koby Crammer, and Fernando Pereira: “Confidence-weighted linear classification”, In *Proceedings of the 25th international conference on Machine learning*, ICML’08, 2008.
- [19] Jialei Wang, Peilin Zhao, and Steven C.H. Hoi: “Exact Soft Confidence-Weighted Learning”, In *Proceedings of the 29th International Conference on Machine Learning*, ICML’12, 2012.