Adaptive Designs for Optimizing Online Advertisement Campaigns

Andrey Pepelyshev, Yuri Staroselskiy and Anatoly Zhigljavsky

Abstract We investigate the problem of adaptive targeting for real-time bidding in online advertisement using independent advertisement exchanges. This is a problem of making decisions based on information extracted from large data sets related to previous experience. We describe an adaptive strategy for optimizing the click through rate which is a key criterion used by advertising platforms to measure the efficiency of an advertisement campaign. We also provide some results of statistical analysis of real data.

1 Introduction

Online advertisement is a growing area of marketing where advertisements can be personalized depending on user's behaviour. To determine user preferences, advertising platforms record data with visited webpages, previous impressions (i.e. ads shown), clicks, conversions, geographical information derived from IP address and then use these data to design strategies when, where and to whom to show some advertisements. Online advertisement has two main forms: one is related to leading technology companies like Google and another is processed by independent ad exchanges [12].

Ad exchanges use auctions with Real-Time Bidding (RTB), which is a magnificent way of delivering online advertising. As mentioned in [3], spending on RTB in the US during 2014 reached $10 billion. The participants of auctions are demand

Andrey Pepelyshev
Cardiff University, Sengennydd Road, Cardiff, UK, e-mail: pepelyshevan@cardiff.ac.uk

Yuri Staroselskiy
Crimtan, 1 Castle Lane, London, SW1E 6DR, UK, e-mail: yuri@crimtan.com

Anatoly Zhigljavsky
Cardiff University, Sengennydd Road, Cardiff, UK, and Lobachevskii State University of Nizhnii Novgorod, Gagarin av. 23, 603950, Russia, e-mail: ZhigljavskyAA@cardiff.ac.uk

partners, which are essentially advertising platforms whose core business is the design of bidding strategies for ad requests for delivery advertisements. Specifically, each time when an ad exchange sends information about a user visiting a webpage, the demand partner can identify the prospectiveness of the request depending on the parameters (e.g., user id, webpage visited, IP address, user browser agent) and behaviour data (e.g., track record of the user over the latest few months) and propose a bid to compete in the auction with other demand partners. Thus, RTB enables a demand side to find a favorable ad campaign and submit a bid for a request depending on parameters of the request and behaviour data. Supposedly, online advertising brings customers at lower cost which is achieved by targeting narrow groups of users.

The process of showing online advertisements through the RTB systems occurs billions of times every day and consists of the steps displayed in Figure 1.

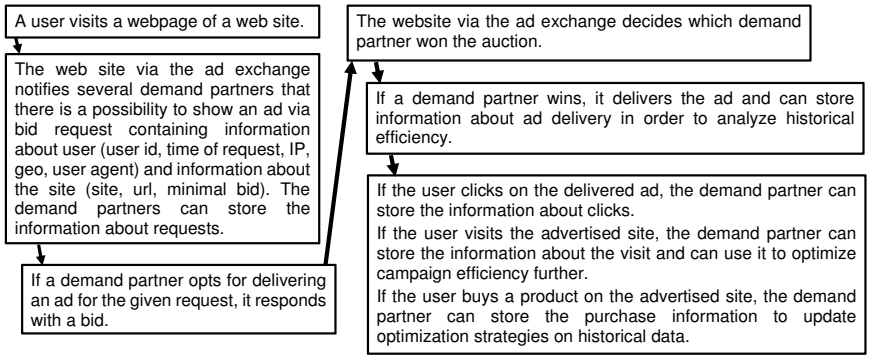

Fig. 1 Process of RTB and actions of a demand partner for delivering an ad.

The demand partner has to solve the problem of maximizing either the click through rate (CTR, i.e. the proportion of the number of clicks to the number of impressions) or the conversion rate (i.e. the proportion of the number of purchases to the number of impressions) by bidding on a set of requests under several constraints:

C1: Budget (total amount of money available for advertising);
C2: Number of impressions N_{total} (the total amount of ad exposures);
C3: Time (ad campaign is restricted to a certain time period).

In practice, the demand partner designs a strategy which cleverly chooses 5 – 500 million requests out of 50 billion available ones. To construct a good strategy, the demand partner has to use all log records.

General principles of adaptive designing are considered in [2, 4, 5, 6]. The design problem for optimizing the CTR has the following specifics compared to assumptions of the standard response surface methodology.

A1: The demand partner cannot choose requests with desired conditions but can leave an auction or suggest a bid for a user currently visiting a webpage.

A2: The design space is very complicated compared to typical $[0,1]^d$ and $\{0,1\}^d$ cases. Usually, the demand partner considers about 20 categorical factors; some factors (e.g. website, city, behaviour category) have hundreds of levels as well as other factors typically have about 10 levels.

A3: We observe the binary outcome but we have to consider the CTR as a function of the request.

The problem of adaptive targeting for ad campaigns was discussed in dozens of papers, see e.g. [8, 11, 15]. Some papers, for example [1, 14], use the look-alike idea implying that a new request will lead to the click/conversion if the new request is similar to (looks like one of) the previous successful requests. In 2014 two contests were organized at the Kaggle platform (www.kaggle.com), see [16] and [17], on algorithms for predicting the CTR using a dataset with subsampled non-click records so that the CTR for the dataset is about 20% while for a typical advertising campaign the CTR is about 0.4% or less. The algorithms, which were proposed by many teams are publicly available and give approximately the same performance with respect to the logarithmic loss criterion

$$\log(loss) = -\sum_{i=1}^{N} \Big\{ y_i \log(p_i) + (1-y_i)\log(1-p_i) \Big\} / N,$$

where N is the size of the data set, p_i is the predicted probability of a click for the i-th request, and $y_i = 1$ if the i-th request leads to a click and $y_i = 0$ otherwise. This criterion, however, does not look very sensible when the probabilities p_i are very small as it pays equal weights to type I and type II error probabilities (as noted above, typical values of p_i's are in the vicinity of 0.004 or even less).

In this paper, we provide a unified approach which comprises the popular methodologies, give a short review of these methodologies and make a comparison of several methods on real data.

2 Formal Statement of the Problem

Suppose that the advertisement we want to show is given and first assume that the price for showing a given ad is fixed; we shall also ignore the time constraint C3. Then the problem can be thought of as an optimization problem for a single optimality criterion which is the CTR. We discuss a generic adaptive targeting strategy which should yield the decision whether or not to show the ad to a request from a webpage visited by a user. If the strategy decides to show the ad, it then has to propose a bid.

An adaptive decision should depend on the current dataset of impressions and clicks which include all the users to whom we have shown the ad before and those who have clicked on the ad. Note that the dataset size N grows with time. We can increase the size of the dataset by including all our previous impressions of the

same or similar advertisements (perhaps applying some calibration to decrease the influence of past ad campaigns), so that N could be very large.

Denote the i-th request by $X_i = (x_{i,1}, \ldots, x_{i,m})$, $i = 1, \ldots, N$, where m is the number of features (factors); these features include the behavioural characteristics of the user, characteristics of the website, time of exposure, the device used (e.g. mobile telephone, tablet, PC , etc.), see the assumption A2. Let K be the number of the requests leading to a click on the ad, say, $X_{j_1}, \ldots, X_{j_K}$, where $1 \le j_1 < \ldots < j_K \le N$, among N requests of the current dataset of impressions. Note that K depends on N. The running performance criterion of the advertising campaign is the CTR defined by $P_N = K/N$. It is clear that the CTR changes as N grows.

We impose the following assumption of independence: if we choose a request $X = (x_1, \ldots, x_m)$ then the probability of a click is $p(X)$; different events ('click' or 'no click') are independent. The assumption of independence obviously fails on the set of users that have already made a click on the ad at an earlier time (these users comprise the set $L(0)$ defined in Section 4) but it seems a reasonable assumption for the general set of users.

We also assume that all possible vectors $X = (x_1, \ldots, x_m)$ belong to some set $\mathbb{X}$, which is either partly or fully discrete (see the assumption A2) and whose structure is difficult for determining a distance between different elements of $\mathbb{X}$. We also assume that for any two points X and $X' \in \mathbb{X}$, we can define a similarity measure $d(X, X')$ which does not have to satisfy mathematical axioms of the distance function.

If $\mathbb{X}$ is a discrete set with all possible requests $X = (x_1, \ldots, x_m) \in \mathbb{X}$ given on the nominal scale then we can use the Hamming distance

$$d(X, X') = \sum_{j=1}^{m} \delta(x_j, x'_j), \quad \delta(x_j, x'_j) = \begin{cases} 1 & x_j = x'_j, \\ 0 & x_j \neq x'_j, \end{cases}$$

or the weighted Hamming distance $d(X, X') = \sum_{j=1}^{m} w_j \delta(x_j, x'_j)$, where the coefficients w_j are positive and proportional to the importance of the j-th feature (factor), $j = 1, \ldots, m$. These weight coefficients can be chosen on the basis of the analysis of previous data of similar advertising campaigns, see Table 1 below.

Alternative ways of defining the similarity measure $d(X, X')$ are a logistic model for p_X (as is done in the so-called 'field-aware factorization machines' (FFM), see [13]) or to use sequential splitting of the set $\mathbb{X}$ based on the values of the most important factors of X ('gradient boosting machines' (GBM), see [7]). For FFM, the distance is defined on the space of parameters of the logistic model but in GBM $d(X, X')$ is small if $d(X, X')$ belongs to the same subset of $\mathbb{X}$ and it is large if the subsets which X and X' belong to have been split at early stages of the sequential splitting algorithm (that is, the values of the most influential features are very different).

2.1 Field-Aware Factorization Machines (FFM)

FFM describes the probability p_X by some sigmoidal parametric function, for example, the logistic function

$$p_X = \frac{1}{1+e^{-m(X,\theta)}},$$

where θ is a vector of parameters and $m(X,\theta)$ is linear in the parameters. For example, the second-order function $m(X,\theta)$ is given by

$$m(X,\theta) = \theta_0 + \sum_{i=1}^{m}\sum_{k=1}^{n_i}\theta_{i,k}\delta(x_i,l_{i,k}) + \sum_{i=1}^{m-1}\sum_{k=1}^{n_i}\sum_{j=i+1}^{m}\sum_{s=1}^{n_j}\beta_{i,k;j,s}\delta(x_i,l_{i,k})\delta(x_j,l_{j,s}),$$

where $\beta_{i,k;j,s} = \sum_{z=1}^{q}\theta_{i,k,z}\theta_{j,s,z}$ describes a factorization procedure, $l_{i,k}$ are all possible levels of the i^{th} factor, $i = 1,...,m$, $k = 1,\ldots,n_i$, $\delta(x_i,l_{i,k})$ equals 1 if $x_i = l_{i,k}$ and 0 otherwise. The vector of parameters θ consists of $\theta_0, \theta_{i,k}, \theta_{i,k,z}$ and is estimated by an iterative use of the gradient descent method for the logarithmic loss criterion, see [13].

A similar approach is the follow-the-regularized-leader (FTRL) methodology, where the function $m(X,\theta)$ has a simpler expression, see [10].

2.2 Gradient Boosting Machines (GBM)

GBM is a method of iterative approximation of the desired function p_X by a function of the form

$$p_X^{(k)} = \sum_{j=1}^{k}\alpha_j T(X,\theta_k),$$

where the vector θ_k is estimated at the k-th iteration, through minimizing the loss criterion [7]. Tree-based GBM considers the function $T(X,\theta)$ as the indicator function of the form

$$T(X,\theta) = \begin{cases} \theta_{in}, & \theta_{i,low} \le x_i \le \theta_{i,up},\ i = 1,\ldots,m, \\ \theta_{out}, & \text{otherwise}, \end{cases}$$

where $\theta = (\theta_{in}, \theta_{out}, \theta_{1,low}, \theta_{1,up}, \ldots, \theta_{m,low}, \theta_{m,up})$. Note that levels of categorical variables are encoded by integer numbers.

3 Generic Adaptive Strategy for Maximizing the CTR of an Advertising Campaign

The purpose of the strategy for maximizing the CTR is to employ the training set of past records for the new requests we will be showing the ad, to increase P_N as N increases.

We can always assume that N_{total} defined in the assumption C2 is very large. Mathematically, we can then assume that $N \to \infty$. If we assume that the bid price is the same (that is, we ignore C1) and there is no time constraint (here we ignore C3) then formally our aim becomes devising a strategy such that $\lim_{N\to\infty} P_N$ is maximum. This is simply an optimization problem of p_X, $x \in \mathbb{X}$. The algorithms solving this problem do this either in the parameter space (for the factorization machines) or in the original space $\mathbb{X}$ (for GBM and the look-alike strategies).

The main problems for applying these algorithms are as follows:

- Factorization machines: the number of parameters is of the order of billions. In practice, this number is reduced by confounding parameters.
- Gradient boosting: the number of observations with certain ranges of levels for several factors is small.
- Computational time grows very fast for all approaches as the size of training data increases. Consequently, in practice training data are often subsampled.
- All approaches have several tuning parameters which should be carefully chosen.

By the nature of the methods, the look-alike approach is applicable in practice if the number of observed clicks K is at least a few dozens, the GBM approach is applicable if K is at least several hundreds and the FFM approach is applicable if K is at least several thousands.

A generic adaptive strategy is an evolutionary one which chooses new requests in the vicinity of the requests that were successful previously; in marketing these kinds of methods are called look-alike methods. To define the preference criterion, for all N we need an estimator $\hat{p}_N(X)$ of the function $p(X)$, which is defined for all $X \in \mathbb{X}$. We do not need to construct the function $\hat{p}_N(X)$ explicitly; we just need to compute values of $\hat{p}_N(X)$ for a given X, where X is a request which is currently on offer for a demand partner. We hence suggest the following estimator $\hat{p}_N(X)$:

$$\hat{p}_N(X) = \frac{\sum_{k=1}^{K} \omega_{j_k} \exp\{-\lambda_N d(X, X_{j_k})\}}{\sum_{i=1}^{N} \omega_i \exp\{-\lambda_N d(X, X_i)\}} + \varepsilon_N, \tag{1}$$

where λ_N and ε_N are some positive constants (possibly depending on N) and ω_i is the weight of the i-th observation made after a calibration of the data is made (the possibility of making such calibration has been mentioned above). The sum in the numerator in (1) is taken over all users who have clicked on the ad. If all these (good) requests are far away from X then the value $\hat{p}_N(X)$ will be very close to zero. The constant ε_N is a regularization constant. As $\varepsilon_N > 0$ there is always a small probability assigned to each X, even if in the past there were no successful requests that were similar to X. Theoretically, as $N \to \infty$, we may assume that $\varepsilon_N \to 0$.

Note that an estimator $\hat{p}_N(X)$ for $p(X)$ is implicitly constructed in the factorization machines and in gradient boosting machines too. Using an estimator $\hat{p}_N(X)$, we can suggest how much the demand partner can offer for the request X in the bidding procedure (that is, we stop optimizing $p(X)$ and take into account the constraint C2). For example, the demand partner can offer larger bids if $\hat{p}_N(X) \geq p_*$, where p_* is the desired probability we want to reach. Another possible use of an estimator $\hat{p}_N(X)$ can be based on the following idea: the amount of money the advertising platform offers for X is proportional to the difference $\hat{p}_N(X) - K/N$, if this difference is positive, and a very small bid, if the difference is negative. For these strategies we do not obtain $\lim_N \hat{p}_N(X) = \max_X p(X)$ but we construct effective strategies which take into account not only the constraint C2 but also C1 and C3. Note in this respect that it is always a good idea to offer very small bids to the users with small values of $\hat{p}_N(X)$ for the following reasons: (a) learning about $p(X)$ in the subregions of X where we perhaps do not have much data, (b) the difference (ratio) between large values of probabilities $p(X)$ for 'good' X and 'bad' X can be smaller than the difference of the option prices for these 'good' and 'bad' X's, (c) the constraint C3 is easier to satisfy, and (d) by saving some funds on cheap X's we can afford higher prices on X's with large values of $\hat{p}_N(X)$.

4 Analysis of Real Data

In the present section we analyze an ad campaign which was executed by Crimtan from 2015-02-01 to 2015-02-17, the number of impressions is slightly above 3 millions and the number of clicks is slightly above 700, so that the CTR $\hat{p} \cong 2.4 \cdot 10^{-4}$, thus FFM approach is not applicable.

To investigate the performance of the strategies for the database of requests for the ad campaign, we split the database of impressions into 2 sets: the training set $\mathbb{X}_p(T)$ of past records with dates until a certain time T (where T is interpreted as the present time) and the test set $\mathbb{X}_f(T)$ of future records with dates from the time T. We now compare GBM and the look-alike approach by comparing the CTR for the samples of most favorable requests with the highest chances to click in Figure 2.

To form the sample of most favorable requests for the look-alike approach, we define the set

$$L(r) = \{X_j \text{ from } \mathbb{X}_f(T)\colon \min_{clicked\ \tilde{X}_i \in \mathbb{X}_p(T)} d(X_j, \tilde{X}_i) \leq r\};$$

that is, $L(r)$ is a set of requests where we have shown the ad and the minimal distance to the set of clicked requests from the set of past records is not greater than r. In other words, the set $L(r)$ is an intersection of the set of our requests with the union of balls of radius r centered around the clicked past requests. We consider X_j with 7 factors: website, ad exchange, city, postcode, device type, user agent, user behaviour category. In Figure 2 the points corresponding to the look-alike approach are given by $\big(\text{size of } L(r), \text{CTR for } L(r)\big)$, $r = 0, \ldots, 4$.

To form the sample of most favorable requests for the GBM approach, we construct the GBM model using the training set $\mathbb{X}_p(T)$ and then apply this model to predict the probability to click for each request from the test set $\mathbb{X}_f(T)$. Now we can sort the predicted probabilities and create samples of requests with highest predicted probabilities to click.

In Figure 2 we can see that the look-alike approach and the GBM approach have similar possibilities to increase the CTR for the considered ad campaign.

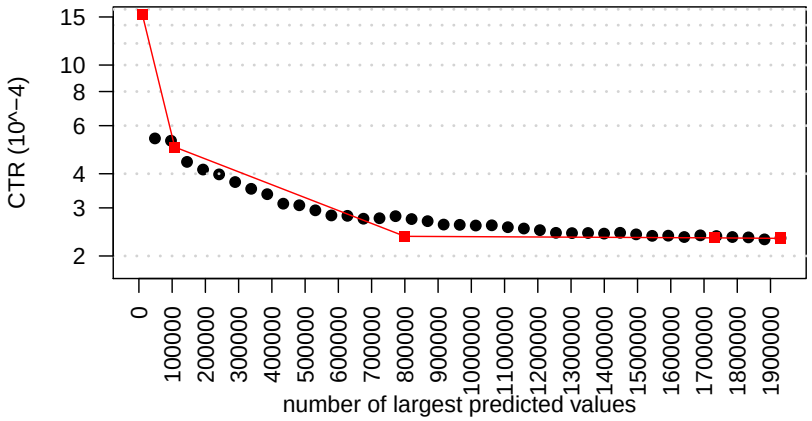

Fig. 2 The CTR for favorable samples of requests of certain sizes for the look-alike approach (squares) and the GBM approach (dots), T=2015-02-08, 7 factors are used.

Let us perform a sensitivity analysis of the CTR for the sets $L(r)$. In Table 1 we show the CTR for several sets $L(r)$ with T=2015-02-08 and different choices of factors, and the index of the influence of the ith factor

$$I_{f_i} = \sum_{r=0}^{2} \left(1 - \frac{\mathrm{CTR}[L(r)|f_1,\dots,f_{i-1},f_{i+1},\dots,f_m]}{\mathrm{CTR}[L(r)|f_1,\dots,f_m]}\right)^2$$

where $\mathrm{CTR}[L(r)|f_1,\dots,f_m]$ is the CTR for the set $L(r)$ with requests containing only factors $f_1,\dots,f_m$. We can observe that $I_{\mathrm{De}} = 0.0003$ and $I_{\mathrm{Ex}} = 0.09$; that is, the device type has no influence and the ad exchange has a small influence on the CTR; consequently such factors can be removed from the model (and computations). The postcode has no influence on the CTR for the set $L(0)$ but has some influence on the CTR for the set $L(1)$.

In contrast, the user agent, the user behaviour category, and the city are very influential factors. It is rather surprising that the postcode has no influence but the city has a big influence on the CTR for the set $L(0)$.

Acknowledgement. The paper is a result of collaboration of Crimtan, a provider of proprietary ad technology platform and University of Cardiff. Research of the third author was supported by the Russian Science Foundation, project No. 15-11-30022 "Global optimization, supercomputing computations, and application".

Table 1 The CTR multiplied by 10^4 for several sets $L(r)$ with T=2015-02-08 and different choices of factors. Abbreviation of factors are Be:behaviour category, We:website, Ex:ad exchange, Ci:city, Po:postcode, De:device type, Ag:user agent.

Set of used factors, S	CTR$[L(0)\|S]$	CTR$[L(1)\|S]$	CTR$[L(2)\|S]$	f_i	I_{f_i}
Be,We,Ex,Ci,Po,De,Ag	15.3	5.01	2.36		
We,Ex,Ci,Po,De,Ag	5.13	2.43	2.35	Be	0.71
Be, Ex,Ci,Po,De,Ag	11.69	2.81	2.35	We	0.25
Be,We, Ci,Po,De,Ag	12.29	3.89	2.31	Ex	0.09
Be,We,Ex, Po,De,Ag	7.62	2.46	2.32	Ci	0.51
Be,We,Ex,Ci, De,Ag	14.96	2.45	2.32	Po	0.26
Be,We,Ex,Ci,Po, Ag	15.27	5.09	2.38	De	0.0003
Be,We,Ex,Ci,Po,De	4.87	3.37	2.20	Ag	0.58

References

1. Aly, M., Hatch, A., Josifovski, V., Narayanan, V. K.: Web-scale user modeling for targeting. Proceedings of the 21st international conference companion on World Wide Web, 3–12, ACM (2012)
2. Box, G. E. P., Wilson, K.B.: On the experimental attainment of optimum conditions (with discussion). Journal of the Royal Statistical Society Series B 13(1): 1–45 (1951)
3. eMarketer: US programmatic ad spend tops $10 Billion this year, to double by 2016. http://www.emarketer.com/Article/US-Programmatic-Ad-Spend-Tops-10-Billion-This-Year-Double-by-2016/1011312 (2014)
4. Ermakov, S. M., Zhigljavsky, A. A.: Mathematical Theory of Optimal Design. Nauka, Moscow. (1987)
5. Fedorov, V. V., Hackl, P.: Model-Oriented Design of Experiments. Springer. (2012)
6. Fedorov, V. V., Leonov, S. L.: Optimal Design for Nonlinear Response Models. CRC Press. (2013)
7. Friedman, J. H.: Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232 (2001)
8. Jansen, B.J., Mullen, T.: Sponsored search: an overview of the concept, history, and technology. International Journal of Electronic Business 6(2), 114–131 (2008)
9. de Leeuw, J., Mair, P.: Multidimensional scaling using majorization: SMACOF in R. smacof. Journal of Statistical Software, 31(3), 1–30 (2009)
10. McMahan, H. B.: Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In International Conference on Artificial Intelligence and Statistics, 525–533 (2011)
11. McMahan, H. B., Holt, G., Sculley, D., et al.: Ad click prediction: a view from the trenches. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1222–1230 (2013)
12. Nicholls, S., Malins, A., Horner, M.: Real-time bidding in online advertising. http://www.gpbullhound.com/wp-content/uploads/2014/09/Real-Time-Bidding-in-Online-Advertising.pdf (2014)
13. Rendle, S.: Factorization machines. In Data Mining (ICDM), 2010 IEEE 10th International Conference, 995–1000 (2010)
14. Tu, S., Lu, C.: Topic-based user segmentation for online advertising with latent dirichlet allocation. Advanced Data Mining and Applications, 259–269. Springer Berlin Heidelberg (2010)
15. Yang, S., Ghose, A.: Analyzing the relationship between organic and sponsored search advertising: positive, negative or zero interdependence?, Marketing Science, 29 (4), 602–623 (2010)
16. https://www.kaggle.com/c/avazu-ctr-prediction Accessed September 12, 2015.
17. https://www.kaggle.com/c/criteo-display-ad-challenge Accessed September 12, 2015.