Advances in Applied Mathematics 应用数学进展, 2020, 9(5), 791-797
Published Online May 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.95094

Hans汉斯

Conversion Rate Prediction Based on Combined Response Prediction Model

Xiao Zhang

School of Mathematics and Information Science, Guangxi University, Nanning Guangxi
Email: 330265136@qq.com

Received: May 1st, 2020; accepted: May 20th, 2020; published: May 27th, 2020

Abstract

In real-time bidding (RTB), conversion rate is an important index to measure the advertising effect. The conversion rate of real-time bidding advertising is very low, which results in the sparsity of advertising log data and makes it difficult to predict the conversion rate through historical data. In this paper, the CRPM (Combined Response Prediction Model) is constructed to predict the conversion rate, and the simultaneous equation of click rate and conversion rate is used to eliminate the endogenous problem. The Bayesian personalized ranking algorithm in the recommendation system is optimized to solve the data sparsity problem. The experimental results show that the combined response prediction model has a better prediction effect on the conversion rate prediction than the commonly used logistic regression prediction method.

Keywords

Real Time Bidding, Conversion Rate, Bayesian Personalized Ranking

基于组合响应预测模型的转换率预测

张　骁

广西大学，数学与信息科学学院，广西　南宁
Email: 330265136@qq.com

收稿日期：2020年5月1日；录用日期：2020年5月20日；发布日期：2020年5月27日

摘　要

在实时竞价广告(Real-time bidding, RTB)中，转化率是衡量广告效果的重要指标。因为实时竞价广告的转换率很低，造成了广告日志数据的稀疏性，给通过历史数据预测转换率带来了难度。本文通过构建组

文章引用: 张骁. 基于组合响应预测模型的转换率预测[J]. 应用数学进展, 2020, 9(5): 791-797.
DOI: 10.12677/aam.2020.95094

合响应预测模型CRPM (Combined Response Prediction Model)预测转换率，通过点击率和转换率的联立方程来消除内生性问题，并通过推荐系统中贝叶斯个性化排序算法进行了优化，解决了数据的稀疏性问题。实验结果显示，组合响应预测模型的相较于目前常用的逻辑回归预测方法对转换率的预测有着更好的预测效果。

关键词

实时竞价广告，转换率，贝叶斯个性化排序

1. 引言

实时竞价广告是一种新兴的在线广告投放方式，通过程序化拍卖方式实时销售互联网广告[1]。在实时竞价广告拍卖的流程中，广告商可以评估每个广告展示(impression)的价值和收益，从而判断是否进行购买，大大提高了广告交易的效率。作为连接需求方和供应方的重要媒介，需求方平台(Demand-Side Platform, DSP)在整个竞标流程中扮演了重要的角色[2]。作为面向广告主的广告投放管理平台，DSP 需要帮助广告商完成对投标信息的分析、投标效果的分析和广告的定价、出价等一系列重要的活动，其中包括对点击率和转换率的预测。但转换事件很少发生，据 ipinyou 公司的实际数据，转换事件发生的概率只有 0.02% [3]，这也造成了数据稀疏(Data Sparsity)的问题，使得对转换率的预测充满挑战。

在点击率和转换率预测模型中，逻辑回归模型以及各种广义线性模是最直观、最容易解释的一类模型，广泛的应用于 RTB 转换率或点击率预测。如 Agarwal [4]等人通过一种动态线性模型研究了实时竞价广告点击率预测问题。Richardson [5]等人在研究搜索广告点击率预测问题时提出了利用广告信息为特征建立逻辑回归模型来预测 RTB 广告点击率，基于特征建模的方法提高了广告系统的性能。Lee [6]等人将单个基于逻辑回归转换率估计器组合起来，以准确识别转换事件。最近相关的研究中，基于机器学习的实时竞价广告转换率预测方法也被广泛的应用。Zhang [7]等人提出基于神经网络的算法 FNN，思路是先使用 FM 计算出每个特征对应到的隐变量，然后将样本中的每个特征转化为对应的隐变量，各个隐变量进行内积或外积计算后，输入 DNN 模型。Gan [8]等人为了克服现有的研究忽略用户点击率行为的顺序特征这一限制，提出了一个名为最近循环神经网络(R-RNN)的预测模型，提高了预测的准确度。

在本文中，需要考虑到点击率和转换率两者的关联性在构建转换率和点击率的预测方程时，构建了联立预测模型。此外，针对数据的稀疏性问题造成的转换率难以预测的问题，此外本文使用贝叶斯个性化排序算法(Bayesian personalized ranking, BPR)对预测模型进行优化。将这种方法称为组合响应预测模型方法记作 CRPM (Combined Response Prediction Model)。实验结果表明，本文方法能显著提高转换率的预测效果。

2. 点击和转化的联立模型

DSP 平台的广告竞标历史日志数据可以表示为以下的形式：

$$D=\left\{\left(x_i,c_i\right)\mid i=1,\cdots,n\right\}$$

其中 x_i 表示每一次展示用户、广告商和发布商的所有信息，包括广告中的广告的各种属性、用户地理位置以及广告的成交价格等信息。表示关于转换率和点击率的信息，$c_i=0$ 表示没有转换和点击事件发生，

$c_i = 1$表示有转换和点击事件发生。n表示数据中展示的数量。

我们使用一个 logit 回归模型来研究在访问者点击进入公司登录页面的情况下转换的概率。

$$p_i^{con} = \frac{\exp\left(\alpha_i^{con} x_i + \varepsilon_i^{con}\right)}{1+\exp\left(\alpha_i^{con} x_i + \varepsilon_i^{con}\right)} \tag{1}$$

其中 p_i^{con} 表示转换的概率，x_i 表示数据信息中的变量，α_i^{con} 表示参数，ε_i^{con} 表示误差项。得到的预测公式表明了在获得用户的各项数据后，根据数据预测转换发生的概率。

在产生转换之前，用户首先需要点击相关网页，即有点击的行为。与转换类似，我们使用 logit 回归模型对点击率决策进行建模：

$$p_i^{cli} = \frac{\exp\left(\alpha_i^{cli} x_i + \varepsilon_i^{cli}\right)}{1+\exp\left(\alpha_i^{cli} x_i + \varepsilon_i^{cli}\right)} \tag{2}$$

其中 p_i^{cli} 表示点击的概率，x_i 表示数据信息中的变量，α_i^{cli} 表示参数，ε_i^{cli} 表示误差项。

根据 Ghose [9]和 Rutz [10]的研究表明，在线广告竞标的过程中，广告的转换率和点击率具有内生的关联性。因此，需要将点击率和转换率的预测模型联合建模，以描述点击率和转换率的相关性，因此本文通过将转换率模型和点击率模型的误差项联系起来得到：

$$\begin{pmatrix} \varepsilon_i^{con} \\ \varepsilon_i^{cli} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{con,con} & \sigma_{con,cli} \\ \sigma_{cli,con} & \sigma_{cli,cli} \end{pmatrix}\right]$$

其中 σ 表示待估的参数。

3. 贝叶斯个性化排序

为了消除广告日志数据存在稀疏性问题，本文采用了贝叶斯个性化排序方法对模型进行了优化。贝叶斯个性化排序是一种推荐系统算法，由 Rendle [11]等人所提出。贝叶斯个性化排序方法算法考虑到了用户对于不同项目的偏好排序关系，更加充分的利用了数据中的信息，一定程度上解决了由于数据稀疏性造成排序困难。

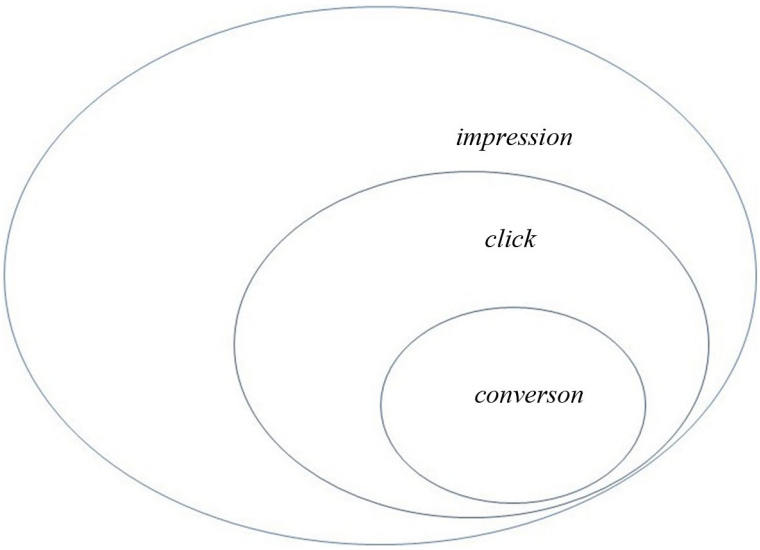

Figure 1. The relationship among impression, click and conversion
图 1. 展示，点击和转换的关系

对于一次广告展示，可能会产生三种结果：用户没有点击广告；用户仅点击并进入了广告页面，没有进一步操作；用户点击进入了广告页面，并且进行了进一步的操作，如在该页面进行预订或购买相关产品等，投送的广告产生了转换。对于广告商来说，产生了转换的广告使得广告商直接获得收益，因此广告商最偏重于购买可能产生转换的广告。产生了点击的广告可能表明消费者对于广告有着一定的兴趣，可能是潜在的客户，并且这类广告也会起到一定的宣传效果，相较于没有产生点击的广告，广告商更倾向于购买产生了点击的广告。对于广告商而言，不同的广告就有了不同的优先级，产生了转换的广告优于只产生了点击的广告，而产生了点击的广告优于没有产生点击的广告。如图 1 中所示根据广告展示后的三种不同状态，可以将历史数据 D 分为三个子集 D^{con} 、 D^{cli} 和 D^{imp} 。分别表示有转换的数据集、只有点击的数据集和没有点击的数据集，其中 $x_i \in D^{con}$ ， $x_j \in D^{cli}$ ， $x_k \in D^{imp}$ 分别表示三个子集的数据。为了表示上述三个子集之间的偏序关系，我们用 $x_i \succ x_j$ 和 $x_j \succ x_k$ 三个子集的偏序关系[12]。

我们可以按照贝叶斯个性化排序的思想进一步对它进行拓展，将成对排序关系拓展为一个三对排序的关系，根据用户对广告反馈可能出现的三种情况，构建新的似然函数，我们可以得到转换率的似然函数 M_{con}：

$$\begin{aligned} M_{con} &= \ln P\left(x_i \succ x_j \wedge x_j \succ x_k \mid \theta\right) P(\theta) \\ &= \ln P\left(x_i \succ x_j \mid \theta\right) P\left(x_j \succ x_k \mid \theta\right) P(\theta) \\ &= \sum_{(u,i,j)\in D} \ln \sigma\left(p_i^{cik} - p_j^{cli}\right) + \sum_{(u,j,k)\in D} \ln \sigma\left(p_j^{con} - p_k^{con}\right) - \lambda_\theta \|\theta\|^2 \end{aligned}$$

得到这种偏序关系后，可以训练新的预测模型参数 θ。根据 Rendle [11]的研究，可以使用基于梯度下降的描述的算法来最来最小化，并采用训练组的 bootstrap 抽样，不仅该训练方法的收敛速度有效，而且其性能比枚举所有组的情况下更稳定。通过这种方法，θ 梯度计算可表示如下：

$$\frac{\partial M_{con}}{\partial \theta} = \sum_{(u,i,j)\in D} \frac{\partial \hat{x}_{uij}}{\partial \theta} \cdot \frac{-\mathrm{e}^{-\left(p_i^{cik} - p_j^{cli}\right)}}{1+\mathrm{e}^{-\left(p_i^{cik} - p_j^{cli}\right)}} + \sum_{(u,j,k)\in D} \frac{\partial \hat{x}_{ujk}}{\partial \theta} \cdot \frac{-\mathrm{e}^{-\left(p_j^{con} - p_k^{con}\right)}}{1+\mathrm{e}^{-\left(p_j^{con} - p_k^{con}\right)}} - \lambda_\theta \theta \tag{3}$$

4. 实证分析

4.1. 实验数据

我们使用 2013 年由广告需求方平台公司 iPinyou 发布的全球竞价算法竞赛的数据集[3]来验证本文提出的算法。数据集共分为三个季度，每一个季度的数据都包括竞标、展示、点击和转换的数据，分为训练数据集和测试数据集两部分。由于第三季的数据较为完整，本文采用第三季的数据进行实验。

4.2. 评价指标

为了检验构建模型的效果，需要构建评价的指标对于模型进行评价，我们选取了几种在点击率预测和转换率预测常用的评价指标作为构建模型的评价指标。包括一些研究中使用的评价指标，如 AUC (Area Under Curve)、ROC 曲线(Receiver Operating Characteristic curve) [13]等，NDCG (Normalized Discounted Cumulative Gain) [12]，MAE (Mean Absolute Error) [14]。

4.3. 广告预测效果评估

在本文中，我们分别使用常见的逻辑回归和本文使用的组合响应预测模型，结果如表 1 所示。在表

格中逻辑回归为 LR，组合响应预测模型为 CRPM。

通过表 1 可以看出，传统的逻辑回归方法的得到的 AUC 值为 0.6933，本文提出的组合响应预测模型为 0.7635，有了显著的提升。从图 2 中 ROC 曲线中也可以看出组合响应预测模型的表现要由于传统的逻辑回归方法。显示了相较于传统的预测方法，组合响应预测模型预测的结果和显示数据拟合的更好，能够较好地解决在数据稀疏的情况下，较好地对实时竞价广告的转换率进行预测。此外，组合响应预测模型的 MAE 值相较于逻辑回归更低，说明模型预测产生的误差更小；与此同时，NDCG 值相较于逻辑回归更高，证明了组合响应预测模型获得了更多的收益，相较于逻辑回归模型的预测能力有了显著的提高。

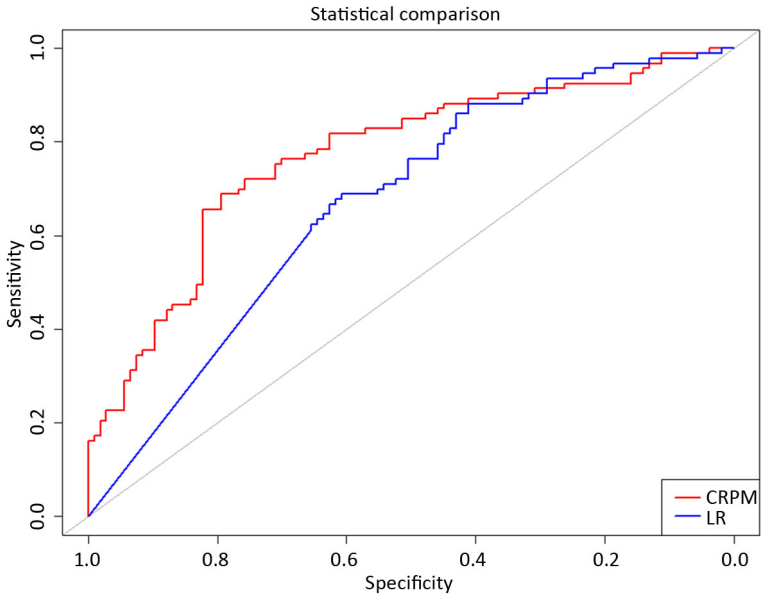

Figure 2. ROC curve comparison of conversion rate prediction
图 2. 转换利率预测的 ROC 曲线比较

Table 1. Comparison of prediction effect of advertising conversion rate
表 1. 广告转换率预测效果比较

方法	AUC	MAE	NDCG
LR	0.6933	0.5217	0.3017
CRPM	0.7635	0.2543	0.4218

此外，我们还进一步分析了两种方法对于广告点击率预测的影响，结果如表 2 所示。可以看出，无论在 AUC 还是 MAE 或是 NDGG 指标上，组合响应预测模型的表现要好于逻辑回归方法。由此可以看出，组合响应预测模型在转换率预测上依然具有优势。值得注意的是，逻辑回归和组合响应预测模型方法在点击率预测方面的效果都要好于对于转换率预测的效果，这也是与直观相吻合的。由于广告的转换一定会先发生点击事件，因此广告的点击数据也要远远地多于广告的点击数据。这就使得点击数据集的数据在数据的稀疏性上要好于转换数据集，使得逻辑回归的预测效果大大提升。组合响应预测模型在预测点击率上要优于预测转换率，但是效果提升较逻辑回归不明显，仅有小幅度的提高。这是因为组合响应预测模型的特性，即使在数据更为稀疏时，也能较好的预测广告的转换率，也从侧面说明了组合响应预测模型相较于传统常用的逻辑回归的优越性。

Table 2. Comparison of prediction effect of advertising click-through rate
表 2. 广告点击率预测效果比较

方法	AUC	MAE	NDCG
LR	0.7523	0.3135	0.4116
CRPM	0.7804	0.2347	0.4220

表 3 表示了点击率和转换率误差项的协方差矩阵，从表中可以看出，点击率和转换率的误差项的协方差是相关的，这也证明了我们的假设，点击率和转换率是相关的，需要对点击率和转换率进行联立建模。

Table 3. Covariance matrix of error term
表 3. 误差项协方差矩阵

	σ_{con}	σ_{cli}
σ_{con}	0.66	0.18
σ_{cli}	0.18	0.27

在实时竞价广告中，广告商更在意是否能够及时从众多客户中发现可能会带来收益的潜在客户，因此能够准确的识别潜在的用户，并能够付出较少的代价，也成为了广告商关注的焦点之一。广告商更倾向于减少预测误差，以降低潜在的损失。在本文中，我们使用平均预测误差从侧面衡量预测的准确度。

平均预测误差定义为：

$$e = \frac{\sum_1^n \hat{p}_i - p_i}{n} \tag{4}$$

其中，e 表示平均预测误差，$\hat{p}_i$ 表示点击率和和转换率的预测值，p_i 表示数据中真实的点击率和转换率的值，n 表示样本的数量。所得结果如表 4 所示：

Table 4. Comparison of average prediction errors
表 4. 平均预测误差比较

	转换率	点击率
LR	0.4243	0.3365
CPM	0.1784	0.1290

从表 4 中可以看出，无论是点击率还是转换率的平均预测误差误差，组合响应预测模型都要远小于逻辑回归，这也表明了组合响应预测模型预测效果较好。此外，由于在实时竞价广告中点击和转换都较少，大部分数据都是负类数据，即没有产生转换和点击的数据。因此平均预测误差很大程度商反映了模型对于假负类率(False Negative Rate)的预测，可以看出组合响应预测模型假负类率相较逻辑回归模型显著的小，这表明组合响应预测模型能够更准确的识别负类数据，使得广告商能够避免购买无法带来收益的广告，降低了广告商购买的成本。

5. 结论

在本文中，我们基于组合响应预测模型的实时竞价广告转换率预测对实时竞价广告市场中的转换率预测问题进行了研究，针对实时竞价广告数据中转换稀疏性的问题，构建点击率和转换率的连理方程来

解决转换率难以预测的问题，并通过推荐系统中贝叶斯个性化排序算法进行了优化。通过实验，我们发现，点击率和转换率具有内生相关性，这证实了我们连理点击率和转换率预测方程的合理性。此外，从实验中可以看出，相较于常用的逻辑回归方法，本文提出的在转换率和点击率的预测性能上有着显著提升。本文对于原始数据变量的处理、选取和设置上还不够深入，这也是未来可能的研究和改进方向。

参考文献

[1] Yuan, S., Wang, J. and Zhao, X. (2013) Real-Time Bidding for Online Advertising: Measurement and Analysis. *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, 1-8. https://doi.org/10.1145/2501040.2501980

[2] Yuan, Y., Li, J. and Qin, R. (2014) A Survey on Real Time Bidding Advertising. *Proceedings of* 2014 *IEEE International Conference on Service Operations and Logistics, and Informatics*, 418-423. https://doi.org/10.1109/SOLI.2014.6960761

[3] Liao, H., Peng, L., Liu, Z., *et al*. (2014) iPinYou Global RTB Bidding Algorithm Competition Dataset. *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 1-6. https://doi.org/10.1145/2648584.2648590

[4] Agarwal, D., Chen, B.-C. and Elango, P. (2009) Spatio-Temporal Models for Estimating Click-Through Rate. *Proceedings of the* 18*th International Conference on Worldwide Web*, 21-30. https://doi.org/10.1145/1526709.1526713

[5] Richardson, M., Dominowska, E. and Ragno, R. (2007) Predicting Clicks: Estimating the Click-Through Rate for New Ads. *Proceedings of the* 16*th International Conference on Worldwide Web*, 521-530. https://doi.org/10.1145/1242572.1242643

[6] Lee, K.-C., Orten, B., Dasdan, A., *et al*. (2012) Estimating Conversion Rate in Display Advertising from Past Erformance Data. *Proceedings of the* 18*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 768-776. https://doi.org/10.1145/2339530.2339651

[7] Zhang, W., Du, T. and Wang, J. (2016) Deep Learning over Multi-Field Categorical Data. *European Conference on Information Retrieval*, 45-57. https://doi.org/10.1007/978-3-319-30671-1_4

[8] Gan, M.X. and Xiao, K.J. (2019) R-RNN: Extracting User Recent Behavior Sequence for Click-Through Rate Prediction. *IEEE Access*, **7**, 111767-111777. https://doi.org/10.1109/ACCESS.2019.2927717

[9] Ghose, A. and Yang, S. (2009) An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets. *Management Science*, **55**, 1605-1622. https://doi.org/10.1287/mnsc.1090.1054

[10] Rutz, O.J., Bucklin, R.E. and Sonnier, G.P. (2012) A Latent Instrumental Variables Approach to Modeling Keyword Conversion in Paid Search Advertising. *Journal of Marketing Research*, **49**, 306-319. https://doi.org/10.1509/jmr.10.0354

[11] Rendle, S., Freudenthaler, C., Gantner, Z., *et al*. (2009) BPR: Bayesian Personalized Ranking from Implicit Feedback. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 452-461.

[12] Shan, L., Lin, L. and Sun, C. (2018) Combined Regression and Tripletwise Learning for Conversion Rate Prediction in Real-Time Bidding Advertising. *ACM/Sigir Proceedings*, 115-123. https://doi.org/10.1145/3209978.3210062

[13] Fawcett, T.J.M.L. (2004) ROC Graphs: Notes and Practical Considerations for Researchers. *Machine Learning*, **31**, 1-38.

[14] Pan, S.-M., Yan, N. and Xie, J.-K. (2017) Study on Advertising Click-Through Rate Prediction Based on User Similarity and Feature Differentiation. *Computer Science*, **44**, 283-289.