|  |  |  |  |
| --- | --- | --- | --- |
| 文件编号 | 2012-RDD-DP-0001 | 作　　者 | 王曼 |
| 文档版本 | V1.92 | 最后修改日期 | 2012-09-26 |

**版本号V1.92**

**RTB实时竞价算法**

**编 写 人： 王曼**

**编写时间： 2012年6月**

**修订控制页**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **编号** | **文档版本** | **修订章节** | **修订原因** | **修订日期** | **修订人** |
| 1 | V1.1 |  | 纠正原文错误 | 2012-07-13 | 苏文魁 |
| 2 | V1.2 |  | 完善逻辑回归算法细节，公式推导 | 2012-07-20 | 王曼，苏文魁 |
| 3 | V1.3 |  | 输入输出表 | 2012-07-23 | 苏文魁 |
| 4 | V1.4 |  | 补充算法流程图及伪代码 | 2012-07-26 | 苏文魁，王曼 |
| 5 | V1.5 |  | 完善输入输出，特别是历史统计信息,增加回头客和商品关联属性 | 2012-07-30 | 王曼，苏文魁 |
| 6 | V1.6 |  | 完善数据预处理，以及根据历史数据调整竞价策略方法，添加了部分名词解释 | 2012-08-01 | 王曼，苏文魁 |
| 7 | V1.7 |  | 调整部分符号，并添加说明 | 2012-08-09 | 王曼 |
| 8 | V1.8 |  | 根据产品需求增加输出见6.3章输入输出表及6.2算法总流程 | 2012-08-16 | 苏文魁，王曼 |
| 9 | V1.9 |  | 修改损失函数 | 2012-09-03 | 王曼 |
| 10 | V1.91 |  | 修改竞价调整参数 | 2012-09-25 | 王曼，苏文魁 |
| 10 | V1.92 |  | 修改竞价系数 | 2012-09-26 | 苏文魁 |

目录

[1. 算法简介 4](#_Toc332372541)

[1.1 算法用途 4](#_Toc332372542)

[1.2 算法概述 4](#_Toc332372543)

[2. 算法详细说明 5](#_Toc332372544)

[2.1 广告活动筛选 5](#_Toc332372545)

[2.2 实时竞价算法 6](#_Toc332372546)

[2.3 广告活动的选择 12](#_Toc332372547)

[3. 算法改进方向 13](#_Toc332372548)

[3.1 特征属性的选择 13](#_Toc332372549)

[3.2 预期点击率(eCTR)预测 13](#_Toc332372550)

[3.3 竞价调整方式 13](#_Toc332372551)

[4. 补充问题 13](#_Toc332372552)

[4.1 无历史数据阶段 13](#_Toc332372553)

[4.2 利用历史数据阶段 13](#_Toc332372554)

[4.3 展示的分布控制 13](#_Toc332372555)

[5. 参考文献 15](#_Toc332372556)

[6. 附件 16](#_Toc332372557)

[6.1 名词 16](#_Toc332372558)

[6.2 算法总体流程图 17](#_Toc332372559)

[6.3 输入输出表 17](#_Toc332372560)

[6.4 数据预处理流程 19](#_Toc332372561)

[6.5 计算权重算法流程 21](#_Toc332372562)

[6.6 模型检验 24](#_Toc332372563)

## 算法简介

### 算法用途

RTB（Real-Time Bidding）实时竞价，是一种利用第三方技术在数以百万计的媒体上针对每一个用户展示行为进行评估以及出价的竞价技术。与大量购买投放频次不同，实时竞价规避了无效的受众到达，只针对有意义的用户进行购买。它的核心是DSP平台（需求方平台），在DMP数据的支持下根据媒体，广告特点和人的属性进行定向投放。RTB对于媒体来说，可以带来更多的广告销量、实现销售过程自动化及减低各项费用的支出。而对于广告商和代理公司来说，最直接的好处就是提高了效果与投资回报率。

RTB算法有几种常见的策略：展现优化（针对品牌推广）目标，点击率（CTR）优化目标和ROI(投资回报率)优化目标。其核心都是要做到合适的广告以合适的价格展现给合适的人，找到新客，找回老客。ROI优化是最直接能体现广告主的投资收益率的指标，但是目前情况下受RTB业界技术限制和公司数据支持情况的限制，点击率优化是现实可行的RTB竞价方法。虽然该方法不能直接体现在用户的投资收益率指标上，但是从统计学分析来看优良的点击率很大程度上会带来良好的收益(ROI)。

目前针对RTB算法中的点击率预测有很多种方法，如：逻辑回归，两阶段广义线性模型，非线性规划模型，典型方程等。选用逻辑回归的主要原因在于该算法成熟，对输入变量要求较低，输出结果较稳定可靠，比较符合现有数据的情况。

本算法主要以点击率（CTR）为优化目标，并选择逻辑回归作为主模型进行设计。

### 算法概述

首先对RTB的竞价流程总结如下图所示：

图1 RTB竞价流程图

从上图中可以看出，实时竞价算法的输入主要包括以下信息：

1. Exchange端提供的广告位信息；
2. 用户id， DMP收集的用户属性信息；
3. 广告主发布的活动信息及竞价策略；
4. 竞价历史（媒体信息，广告位，获胜竞价，用户id等）。

表1: RTB算法输入之竞价历史

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 广告 | 媒体信息 | | | | 用户ID | 时间戳 | 是否点击 | 是否成功 | 报价 | 成交价格 |
| 创意 | 类型 | 权重 | 尺寸 | …… |
|  |  |  |  |  |  |  |  |  |  |  |

实时竞价算法的输出主要是获胜的(广告活动信息，竞价)组。

实时竞价算法主要由两部分构成，一是广告活动的匹配，二是根据媒体，广告活动及用户属性等进行出价。



图2 RTB竞价算法流程

## 算法详细说明

### 广告活动筛选

广告活动筛选的目的是为了选出满足媒体广告位要求，以及符合用户属性特征的广告活动集合。首先，根据请求中的媒体广告位信息对广告活动进行初步筛选，过滤掉一些无效的广告活动，如：不满足广告位要求的广告活动等。然后，利用DMP系统中查找到的用户特征信息，以及广告活动的投放人群属性定向再次筛选广告活动。广告活动的投放人群属性定向主要包括访客所在地区，年龄，性别，婚姻状况，教育程度，兴趣爱好，购买偏好等。

主要包括以下步骤：

第一步，将请求中的媒体广告位信息与DSP系统中的广告活动信息进行初步匹配，筛选出满足条件的广告活动集合；如果不存在广告活动，则不参与此次请求的竞价。

第二步，利用DMP系统，根据用户id，获取用户特征信息。

第三步，将筛选出的广告活动集合与用户特征进行匹配，进一步筛选广告活动集合。



图3 广告匹配流程

表2 广告活动匹配需要考虑的信息

|  |  |  |  |
| --- | --- | --- | --- |
|  | **广告位信息** | **广告活动** | **用户属性** |
| 属性1 | 广告位宽 | 投放媒体偏好 | 性别 |
| 属性2 | 广告位高 | 访客所在地区 | 年龄 |
| 属性3 | 不允许的属性id集合 | 访客年龄 | 地域 |
| 属性4 | 允许的广告技术类别 | 访客性别 | 婚姻 |
| 属性5 | 排除的敏感类别 | 访客婚姻状况 | 教育程度 |
| 属性6 | 排除的产品类别 | 访客教育程度 | 兴趣 |
| 属性7 | 媒体提供的限制列表 | 访客兴趣爱好 |  |
| 属性8 |  | 访客购买偏好 |  |
| …… | …… | …… | …… |

关于匹配度的计算问题涉及到商业策略及广告本身估值，用户的估值，媒体的估值及DSP端广告集合的分析等情况，在综合分析的情况下，才能给出在特定商业策略上的匹配度，该部分计算放在后续进行。

### 实时竞价算法

通过广告匹配算法能够筛选出符合媒体信息和用户特征的广告活动，接下来需要对筛选出的广告活动进行竞价，并排序，返回一系类(广告活动，竞价)结果组。

广告活动的投放类型一般有CPC和CPM两种，目前只考虑CPC投放类型，如果是投放

类型是CPM，可以将CPM转化为CPC进行竞价，转换原则为：

CPM = CPC\*CTR\*1000 （2.1）

总体来说，实时竞价算法可以分为以下两个步骤：

1. 获得每次广告展示的价值，即在给定的相关展现水平和需求数据的情况下，对该次广告展现预期成本，表示形式如下：

 （2.2）

其中表示每次展现的预期点击率，是广告的预期点击价值，以广告主的CPC出价为基础(如果不存在，采用默认值)。

预测是一个关键的步骤：

 （2.3）

其中条件中的publisher表示与该次展示机会相关的媒体属性信息；ad campaign表示广告活动的相关属性信息，user表示用户属性信息。

2．实时竞价的调整

根据竞价策略和历史统计信息调整预期点击价值。

3. 结果合理性分析



当广告主广告以CPC方式投放时，为预期点击率，为预期的点击价值（价格）；

当广告主广告以CPM方式投放时，为常数1，此时用户的选择由人群的筛选完成，为预期的展示价值（价格）。

**推广：**

当广告主广告以CPA、CPS等方式投放时，eTR为预期收益的能力（概率），ePC为市场条件下获取该收益时需要付出的价格水平。这些方式需要对action和sale等行为进行收集，才能根据收集的数据来预测收益能力。

**价格修正：**

考虑到实际点击率与算法模型计算出的点击率之间的关系，为使得出价在可控制的范围内，在出价时对价格进行修正。

 （2.4）

其中：为平均点击率，应该是可配置参数，默认为0.001。则公式转化为：

 （2.5）

特别说明：

对于CPC模式的广告：

 （2.6）

对于CPM模式广告：

 （2.7）

#### 广告展示估值

为了准确地评估每次广告展示的价值，需要对该广告的点击率进行预测。初次使用系统时，由于缺乏获胜的历史竞价信息，可以使用默认的点击率，如：该品牌往常的点击率或其他平台媒体经验值等。随着历史数据的增加，需要根据用户和媒体等信息，对每次展示机会预测点击率。

当获得足够多的历史数据时，可以利用媒体特征信息（如：广告位置等），用户特征（如：年龄，兴趣等），广告活动特征（如：创意，关键字等）构建点击率预测模型。由于特征属性信息非常多，首先需要进行属性特征提取，获得有价值的属性特征。然后根据历史数据，构建点击率预测模型。当获得一个展示机会的时候，针对每个广告活动，根据用户和媒体的属性值进行点击率预测。



图4 点击率预测流程

##### 特征提取

在一个广告系统中，每一次出价和服务事件，包括广告机会，拍卖的获胜者，以及浏览该广告展现的用户都能够被记录。每一个广告机会通过其属性进行描述，包括url，出版商，广告位置以及用户统计信息，geo位置信息。获胜信息包括获胜的竞价值，实际的花费等。通常，有很多的目标属性值。它们中的大多数是复杂的布尔逻辑规则来匹配广告商想作为目标的广告库存，作为竞价需要考虑的因素也不应过多，造成算法的复杂程度过大，难以满足实时竞价的时间要求。因此，特征选择是系统中至关重要的一个因素。在选择特征的时候，有两个方面的方法可以选用。

1. 经验法。通过对行业的了解及属性的分析，对属性进行筛选。
2. 采用适当的算法对属性对竞价的影响进行分析，计算每个属性对竞价的影响程度。选择影响因素大的属性作为特征属性。

表3 需要考虑的特征属性信息

|  |  |  |  |
| --- | --- | --- | --- |
|  | **广告位信息** | **广告活动** | **用户属性** |
| 属性1 | 媒体id | 投放类型 | 性别 |
| 属性2 | 广告所属频道 | 活动创意 | 年龄 |
| 属性3 | 广告位宽 | 广告主id | 地域 |
| 属性4 | 广告位高 | 广告类型 | 兴趣 |
| 属性5 | 广告位的可视性 | 活动人群属性 | 消费水平 |
| 属性6 | 地理信息 | 投放媒体偏好 |  |
| …… | …… | …… | …… |

我们采用基于过滤的Fast-Correlation(FCBF)方法选择特征子集[1]，该方法对处理特征数目较大时非常有效。其基本原理描述如下：设数据集D有n条记录，且每条记录由m个非目标特征和一个目标特征C来刻画。如果非目标特征与目标特征之间的相关性过低（给定阈值），则将该特征作为不相关特征去除，如果两个非目标特征之间的相关性过大，超过了这两个特征与目标特征的相关性时，则认为两个特征之间存在冗余，这两种情况均需要进行删除。

FCBF是一个确定性算法来消除与目标值相关度较低的属性或非目标属性过度相关的属性冗余。它能使我们在很大程度上消减特征搜索时间。它依据对称不确定性计算特征和目标值的相关性，定义如下：



其中是信息增益（Information Gain），是熵（Entropy）。的值是给带来的信息增益，并且。是的一个归一化值。通过对目标属性与非目标属性及非目标属性之间的计算，在给定的阈值基础上，进行属性选择。

##### 预期点击率(eCTR)预测

在特征子集选择的基础上，对点击率进行预测。

把(媒体，广告活动，用户)属性值看成是一个多元组，针对每一个多元组，有一个预测目标，是否点击。该问题可以看成是一个典型的逻辑回归问题。给定一个媒体，广告活动，以及用户，利用逻辑回归模型计算预期点击率如下：



其中表示属性个数，表示从多元组获得的第个属性的值，**w**关于它的权重。

假设有个训练样本集，，其中表示由多元组（媒体，广告活动，用户）属性值构成的一个维向量， 是相应的分类标签(+1，点击，或者0：没有点击)。给定训练样本集合，模型通过减少数据中的总损失计算权重向量**w**，公式如下：



可以用L-BFGS算法[3]求解这种大规模的凸优化问题。具体方法如下。

**L-BFGS算法**步骤如下：

Step1:选初始点，允许误差，存储最近迭代次数m（一般取6）；

Step2:；

其中：令，则。

Step3:如果，则返回最优解，否则转Step4；

(注：)

Step4:计算本次迭代的可行方向：；

Step5:计算步长，对下面式子进行Backtracking线性搜索：

；

Step6:更新权重：；

Step7:如果 ，只保留最近次的向量对，需要删除；

Step8: 计算并保存：，；

Step9:用two-loop recursion算法近似计算；

k=k+1，转Step3。

**two-loop recursion算法:**

令****，

步骤1：对，循环以下运算：，；

步骤2：；

步骤3：对，循环以下运算：，。

其中：

，

，

，





**Backtracking线性搜索算法：**

任选，令，

重复直到 ，



结束(重复)

终止 。

#### 竞价调整

为了获得实时竞价调整参数值，利用历史数据建立线性规划模型，通过对偶理论将其转化为一个低维数形式，求解得到竞价调整参数的初值，然后利用控制论方法，基于一些误差函数调整竞价与满足限制条件的水平。

##### 线性规划模型的建立

为了得到在线竞价算法的调整参数，首先建立展示广告优化的线性规划模型，其中，展示机会是分别估价和分配的，需求方约束依据的是展示投放目标。定义符号标记如下：

：展示机会(impressions)，

**：广告活动系列(campaigns)，

：展示机会分配给广告活动**的CTR

：广告活动**的CPC

：分配方案的期望值eCPI，

：广告活动j的展示投放目标（有时有可能会用预算代替）

：决策变量，展示机会是不是分配给广告**，（0,1变量）

根据给定的符号标记，构建原始的线性规划问题如下：



其对偶问题可以表示如下：



对偶变量可以解释如下：表示活动**获取一个额外的目标展示机会的经济值（如：通过增加预算），表示出版商购买一个额外展示机会的经济值（如：通过吸引更多访问者）。

通过解对偶问题可以获得最优解，但还存在一些问题，比如在线情况下，怎么通过对偶问题的最优解获得原问题的最优解。

##### 竞价调整

**调整方案1**

基础算法中假定展示机会是平稳的，这就意味着给定足够多的历史数据，通过解离线对偶问题可以获得最优竞价。实际上，由于市场是动态的，展示机会的到达不是平稳的，比如：供应随季节变化。另一方面，需求方的估值也是一个不固定的过程，比如：旧的广告过期，新的广告开始。这些不固定性不满足完备松弛条件，从而历史最优的对于未来并不一定是最优的。为了解决这个问题，利用历史数据的离线最优解初始化，然后利用一种控制方法在线更新，以满足动态的供应方和需求方约束水平。

在实践中，由于在线计算的效率和展示机会到达的离散性，时间不需要立即被追踪。令表示足够小的时间区间，T表示整个在线竞价时期内的区间个数，在每一个区间内更新一次。受Water-Level算法[4]启发，采用更新公式如下：



其中表示广告活动在时间区间内获得的展示数目，指数因子是一个调节参数，控制算法对误差的响应速度。如果初始的对未来也是最优的，可以令。

更进一步，基于Water-Level的更新公式有一个很好的链接属性：



**调整方案2**

1. 如果是回头客和商品关联，在计算出的价格上乘以一个系数，以增加展示成功率，调整公式如下：

，为经验值。

注：对于回头客和商品关联属性数值不同，应采用配置文件的方式方便更改。目前均采用0.05。

1. 按照广告维度和媒体维度对价格分布和点击率进行统计，得出点击率和价格分布的分布情况（需要根据数据情况确定是用哪个维度的统计信息作为用户数据缺失时的eCTR的预测值）。
2. 根据上一天的竞价成功率与目标成功率之间的差异对价格进行调整（预期成功率和实际成功率进行比较，增加或减少一个系数）。调整公式如下：



其中：，

，。

注：由于系统前期广告数量较少，每个广告活动的预期成功率给定一个固定值，如20%，该参数需要采用配置文件的方式方便进行更改。实际成功率取自**竞价历史统计（计算与分析用-广告维度）**中**“成功率”**属性。

**调整方案选择**

根据系统现有数据收集和处理周期情况，选择调整方案2作为目前的调整方案。

### 广告活动的选择

在线情况下，每一次展示机会到达的时候，计算筛选出的广告活动组（）的竞价值：，并按照一定规则进行排序，给出竞价列表。

取表**竞价历史统计（计算与分析用-广告维度）**，表中“平均价格（成功）“价格作为平均千次展现成本（），竞价指数表示为：

)

取多支广告中该指数高者出价（系统运行前期，该指数为负值暂不做不出价条件，后期模型准确的得以验证后再作为不出价条件）。

## 算法改进方向

### 特征属性的选择

可以用基于过滤的优化的Fast-Correlation(FCBF)方法，属性相关性的特征加权等方法来评估属性特征。.

### 预期点击率(eCTR)预测

有多种方式可以预测eCTR，如：利用对数线性模型LMMH，两阶段广义线性模型，非线性规划模型，典型方程等。

### 竞价调整方式

还可以利用统计学方法进行竞价调整，通过获取展现的成功竞价的历史数据，根据统计学方法描述出竞价的分布情况，然后在满足限制条件下，为达到最大可能竞价成功做出竞价调整。

## 补充问题

### 无历史数据阶段

算法中eCPI和调整参数的计算均依赖历史数据，在系统运行的初始阶段由于缺少历史数据的积累，运行前一段时间采用经验数据进行竞价。建议eCPI=CTR\*CPC，给定CTR进行竞价，也不做调整。

### 利用历史数据阶段

利用历史数据阶段采用本文中算法计算eCPI和调整参数。由于eCPI的计算量比较大，可以采用较大的周期进行计算，推荐计算周期为1天，的计算量较小，可以实时或每小时计算一次。

### 展示的分布控制

为有效控制展示的分布情况，需要在算法中加入调节机制，防止展示过分集中的发生。最初可以采用每小时展示数量作为调节阈值进行控制。如某个广告的展示目标为每个月展示/点击1万次，则可以将这个点击数量根据流量情况或客户要求分布到每个小时中去，使得展示可以满足要求。

## 参考文献

[] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), pages 856-863, 2003.

[2] A. L. Berger and V. J. D. Pietra. A maximum entropy approach to natural language processing. Computational Linguistics, 22:39–71, 1996.

[3] D. C. Liu and J. Nocedal, “On the limited memory BFGS method for large scale optimization,” Mathematical Programming, vol. 45, no. 3, pp. 503–528, 1989.

[4] D. Charles, M. Chickering, N. R. Devanur, K. Jain, and M. Sanghi. Fast algorithms for finding matchings in lopsided bipartite graphs with applications to display ads. Proceedings of the 11th ACM Conference on Electronic Commerce (EC 2010), 2010.

[5] Y.Chen, P.Berkin, and B.Anderson. Real-Time Bidding Algorithms for Performance-Based Display Ad Allocation. KDD’11, 2011.

[6] D. Charles, M. Chickering, N. R. Devanur, K. Jain, and M. Sanghi. Fast algorithms for finding matchings in lopsided bipartite graphs with applications to display ads. Proceedings of the 11th ACM Conference on Electronic Commerce (EC 2010), 2010.

## 附件

### 名词

1. RTB: real time bidding 实时竞价
2. CTR: click-through rate 点击率
3. CPC: cost-per-click每点击成本
4. CPM：cost per one thousand impression 千次展现成本
5. ：expected click through rate 预期点击率
6. ：每次点击的预期成本
7. eCPI：expected cost per impression每次广告展现预期成本，定义为：



1. CPM与CPC的转换公式推导：CPM = CPC\*CTR\*1000

CPM表示的是Cost Per 1000 Impression，即：CPM = Cost / Impressions \* 1000……(1)

CPC表示的是Cost Per Click，即：CPC = Cost / Click….(2)

CTR表示的是Click-Through Rate，即CTR = Click/Impression…(3)

将(3)式代入(2)式：CPC = Cost / (CTR \* Impression) = (Cost / Imp.) \* (1 / CTR)

### 算法总体流程图

eCPC计算方法：根据默认值，设定eCPC初始值，计算过程中采用eCPC\*（1+β2）进行计算。VP与BP的计算方法见“2.2实时竞价算法”中“**价格修正：**”段内公式（（2.4）-（2.7））。

VP

BP

FC

W向量

### 输入输出表

**eCPI计算**输入参数表：

媒体属性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 名称 | 类型  (频道) | 权重 | 广告  位宽 | 广告  位高 |
| 重要度 | \*\*\*\*\* | \*\*\*\*\* | \*\*\* | \*\*\* |
| 算法类型 | 枚举 | 连续值 | 枚举 | 枚举 |

广告活动属性

|  |  |
| --- | --- |
| 名称 | 类型（创意） |
| 重要度 | \*\*\*\*\* |
|  |  |
| 算法类型 | 枚举 |

用户属性

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 名称 | 消费偏好 | 消费偏好权重 | 消费  频率 | 消费  次数 | 消费总额 | 线上消费习惯 | 消费等级 | 兴趣  爱好 | 兴趣爱好权重 | 平均消费水平 |
| 重要度 | \*\*\*\*\* | \*\*\*\*\* | \*\*\* | \*\* | \*\*\*\*\* | \*\*\*\* | \*\* | \*\*\* | \*\*\*\*\* | \*\*\*\* |
| DMP类型 | String | Int | Double | Int | Double | Int | Int | String | Int | Double |
| 算法类型 | 枚举 | 连续值 | 连续值 | 连续值 | 连续值 | 枚举 | 枚举 | 枚举 | 连续值 | 连续值 |

竞价历史（流水数据）

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 广告 | 媒体信息 | | | | 用户ID | 竞价方式 | 是否成功 | 是否点击 | 报价 | 成交价格 | 时间戳 |
| 创意 | 类型 | 权重 | 宽 | 高 |
|  |  |  |  |  |  |  |  |  |  |  |  |

（在计算时，取用竞价成功数据）

计算周期可以以天为单位

**α计算**所需输入：（此部分这期不做计算）

指定广告指定时间段内的点击次数（以小时为间隔）

**竞价历史统计（计算与分析用-广告维度）**：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 广告ID | 平台ID | 竞价次数 | 成功率 | 平均价格  （成功） | 平均价格  （出价） | 最后一次成功价格 | 最后一次出价 | 点击率 |

**竞价历史统计（计算与分析用-媒体维度）**：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 媒体ID | 平台ID | 竞价次数 | 成功率 | 平均价格  （成功） | 平均价格  （出价） | 最后一次成功价格 | 最后一次出价 | 点击率 |

**需要提供的默认（/初始）值（媒体维度）：**

|  |  |  |  |
| --- | --- | --- | --- |
| 媒体ID | 平均（/初始）出价 | 最高出价 | 平均点击率 |

**需要提供的默认（/初始）值（广告维度）：**

|  |  |  |  |
| --- | --- | --- | --- |
| 广告ID | 平均（/初始）出价 | 最高出价 | 平均点击率 |

**在线计算输入：**

用户属性，媒体属性，广告活动属性，是否回头客，是否商品关联。

**算法输出：**

离线输出：各个属性的权重向量。见6.2 算法总流程图中“W向量”

在线输出：用户价值（ValuePrice）VP 见6.2 算法总流程图中“VP”

竞价价格（Bidprice） BP 见6.2 算法总流程图中“BP”

预测点击率(FcCTR%) FC% 见6.2 算法总流程图中“FC”

### 数据预处理流程



数据预处理增加中间变量的生成，目前中间变量有两个：媒体类型与用户购买兴趣关联度及媒体类型与用户兴趣关联度。这两个变量的计算方法将在下文中给出。

#### 离线数据预处理

##### 数据选取

需要选取竞价成功的数据作为训练数据和验证数据，由于模型主要跟媒体，广告活动，用户三方面的属性值有关，因此做应尽可能选取包含这三方面属性值的数据，不选取用户属性完全缺失的数据。

统计竞价成功的数据总数，选取前80%作为训练数据，其余数据作为验证数据。

##### 缺失值填充

如果单条记录中的广告活动创意类型缺失，则需要采用该列的众数值进行填充(注：众数表示的是整列变量中出现次数最多的数据，如果有多个，采用最先出现的数据)；如果其它属性缺失，默认为0。

##### 中间变量的生成

**1) 消费属性预处理**

由于用户消费属性可能有多个值，并且只凭借用户消费属性不能确定其对点击率预测的影响，需要把它与媒体类型关联起来进行考虑。首先通过一定的转换，建立媒体类型和用户消费属性之间对应关系表，然后，通过对应关系将用户消费属性转换为[0,1]区间的数。具体方法如下：

假设媒体类型为，对应的权重为，用户的消费属性为，如果和具有一定的对应关系，对应值用表示，计算某一媒体类型与用户消费属性的关系如下：



其中，，表示用户消费属性对应的权重值，表示用户消费属性权重的最大值。

计算所有媒体类型与用户消费属性的关系值并进行加权求和，就可以把用户的消费属性转换为一个确定的[0,1]区间的数。公式如下：



其中，表示第个媒体类型与用户消费属性的关系值，通过上式计算得出，表示第个媒体类型的权重值。

注：无历史数据之前，对应关系表可以人工给出。具有了一定的历史数据后，可以通过统计信息计算出媒体类型与用户消费属性之间的对应值。

媒体属性与用户属性对应关系表形式如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 消费属性  媒体属性 | 票务旅游 | 运动户外 | … |
| Sports | 0 | 1 | … |
| Travel | 1 | 0.6 | … |
| … | … | … | … |

**2) 兴趣属性预处理**

同上建立媒体类型与用户兴趣属性的对应表，将用户兴趣属性转换为[0,1]区间的数，需要注意的是有可能不同。

##### 广告位预处理

通过对广告位尺寸的分析，可以按广告位宽高将其分为大、中、小三种情况，划分原则如下：假设宽高的最大值为A，如果A的值小于300，则认为是一个小广告位；如果A的值大于700，则认为是大广告位；其它情况下认为是中。

##### 数据标准化

如果属性值是枚举变量，需要引入哑变量，再进行Logistic回归分析。引入哑变量可使问题描述更简明，一个方程能达到多个方程的作用，而且接近现实。 如果某个属性值有n种互斥的取值，则将其用哑变量引入模型时，要设置n-1个哑变量，以避免完全的多重共线性。如：某个属性有四个取值（A/B/C/D），这时需要引入三列哑变量，比如D2,D3,D4,如果变量值是B，则D2=1，否则取0，如果是C，则用D3=1，否则取0，如果是D，则D4=1，否则取0。可以采用如下的矩阵方式进行存储。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | B | C | D |
| D2 | 0 | 1 | 0 | 0 |
| D3 | 0 | 0 | 1 | 0 |
| D4 | 0 | 0 | 0 | 1 |

如果属性值是连续型变量，则需要对变量进行标准化处理，以消除因数据单位不同而引起的计算结果的偏差，一般采用z-score标准化方法，方法如下：





经过变换后，有可能经过预处理的属性值不在[0, 1 ]区间，这时需要将其调整到[0, 1]区间，调整方法如下：如果标准化后的属性值大于1，调整其为1，如果属性值小于0，将其调整为0。最终各变量落入[0, 1]区间。

#### 在线数据预处理

##### 缺失值填充

如果广告活动创意类型缺失，则用离线部分记录的属性值进行填充。如果用户属性中消费偏好和兴趣爱好都不存在，通过模型预测的点击率准确度不高，只能采用媒体或广告活动默认的点击率，其它情况下，默认为0。

##### 中间变量的生成

与离线部分的预处理方式相同。

##### 广告位预处理

与离线部分的广告位预处理方法相同。

##### 数据标准化

根据离线预处理记录对属性值进行标准化，如果属性值是枚举型变量，通过哑变量矩阵进行转换。如果属性值是连续型变量，通过预处理记录查找到标准差和均值，标准化后的变量。需要注意的是由于在线数据多样性，有可能经过预处理的属性值不在[0, 1 ]区间，这时需要将其调整到[0, 1]区间，调整方法如下：如果标准化后的属性值大于1，调整其为1，如果属性值小于0，将其调整为0。

### 计算权重算法流程

采用L-BFGS算法计算逻辑回归预测模型中的权重向量，算法流程图如下：



注1：；一般取6

注2：为阶数和输入向量相同的单位阵，

其中：；

注3：的计算参见Backtracking线性搜索算法

注4：的计算参见two-loop recursion算法

#### Backtracking线性搜索算法

任选，令，

重复直到 ，



结束(重复)

终止 。



注1：初始值取1，取0.5，c取1

#### two-loop recursion算法

令****，。

步骤1：对，循环以下运算：，；

步骤2：；

步骤3：对，循环以下运算：，。

其中：

，

，

，

，

。

流程图：



### 模型检验

建立预测模型之后，计算验证数据的预期点击率，并进行比较，如果准确性达到A%，则认为预测模型合理，可以用来进行预期点击率预测。否则，采用竞价历史中的点击率或默认点击率。(准确度A需要根据实际情况确定)。