(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108182597 A
(43)申请公布日 2018.06.19

(21)申请号 201711439302.9

(22)申请日 2017.12.27

(71)申请人 银橙(上海)信息技术有限公司
地址 201414 上海市奉贤区青村镇南奉公路3878号8幢203室

(72)发明人 彭文元 周小强 申晓宏

(74)专利代理机构 上海愉腾专利代理事务所(普通合伙) 31306
代理人 唐海波

(51)Int.Cl.
G06Q 30/02(2012.01)
G06N 99/00(2010.01)

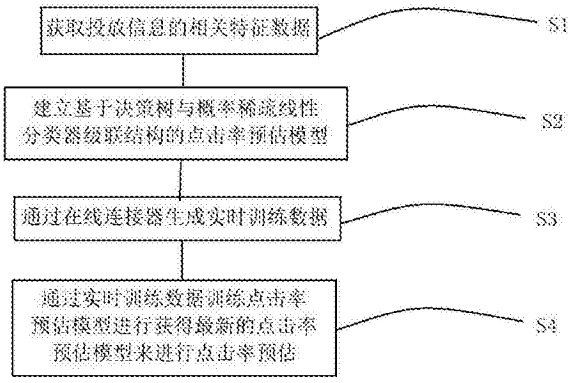
权利要求书1页 说明书9页 附图3页

(54)发明名称

一种基于决策树和逻辑回归的点击率预估方法

(57)摘要

本发明公开了一种基于决策树和逻辑回归的点击率预估方法,包括以下步骤:获取投放信息的相关特征数据;建立基于决策树与概率稀疏线性分类器级联结构的点击率预估模型;通过在线连接器生成实时训练数据;通过实时训练数据训练点击率预估模型进行获得最新的点击率预估模型来进行点击率预估;提出了一个基于决策树与概率稀疏线性分类器级联结构的模型体系结构,它还包含了一个在线学习层,并公开了在线连接器,它是一个在线学习层中非常关键的组成部分,可以将训练数据转换成实时的流式数据;本发明所述的基于决策树和逻辑回归的点击率预估方法,相较于现有的点击率评估方法至少有10%的效果提升。



CN 108182597 A

1.一种基于决策树和逻辑回归的点击率预估方法，其特征在于，所述基于决策树和逻辑回归的点击率预估方法包括以下步骤：

获取投放信息的相关特征数据；

建立基于决策树与概率稀疏线性分类器级联结构的点击率预估模型；

通过在线连接器生成实时训练数据；

通过实时训练数据训练点击率预估模型进行获得最新的点击率预估模型来进行点击率预估。

2.根据权利要求1所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，所述在线连接器的工作为：在数据中加入标签并以在线方式训练输入的数据，将投放信息展现和投放信息点击通过请求ID进行连接，每次用户使用时都会生成一个唯一的请求ID，通过这个ID将展现和点击进行匹配。

3.根据权利要求2所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，所述通过在线连接器生成实时训练数据包括以下步骤：

用户访问网站或者app，用户的相关信息会传递到系统中；

系统通过排序将相关的投放信息返回给用户的设备上；

将上述过程产生的数据记录在展现数据流中；

当用户点击他所看到的投放信息时，这个点击数据记录在点击数据流中；

当时间窗口期过后，在线连接器就会把连接好的展现数据发送到训练数据集中。

4.根据权利要求3所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，在通过在线连接器生成实时训练数据过程中，需要建立异常检测机制。

5.根据权利要求1所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，使用在线学习方法训练线性分类器。

6.根据权利要求1所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，使用增强决策树来对特征进行转换。

7.根据权利要求6所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，所述增强决策树包括：每棵单独的树都为一个分类特征，它的值就是树叶的索引值。

8.根据权利要求6所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，所述增强决策树训练数据的方式是以批量形式进行训练的。

9.权利要求6所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，对每个特征都加入了特征权重，在每个树节点结构中，选择并分割一个最佳特征，一旦一个特征在多棵树中使用时，每个特征的重要性会通过将整棵树全部的损失值相加计算得出。

10.根据权利要求1至9之一所述的基于决策树和逻辑回归的点击率预估方法，其特征在于，所述基于决策树和逻辑回归的点击率预估方法包括：使用抽样方法处理大量训练数据。

一种基于决策树和逻辑回归的点击率预估方法

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种基于决策树和逻辑回 归的点击率预估方法。

背景技术

[0002] 数字广告是一个价值数十亿美元的产业，并且每年还在持续增长。大部分的在线广告平台都是动态分配广告的，根据用户的反馈信息做 出调整，进而向用户展现其感兴趣的广告。机器学习在向用户展现哪个 广告中扮演着一个很重要的角色，使用这种类似推荐的模式也会提升 广告的投放效率。

[0003] 2007年的一篇由Varian和Edelman等人创作的论文介绍了一种 按点击付费的竞价模式，该竞价模式的效果取决于预估点击的准确性。在平常的竞价中产生的数据是非常大量的，而且会有很多新的特征或 元素加入，所以预估系统需要良好的适应性和处理大量数据的能力。

[0004] 在搜索广告系统当中，用户所查询的数据就会成为选取候选广告 的依据，但是在广告投放系统中，用户并不会主动去输入任何东西，所 以在向用户展现广告时，就会有大量的广告会匹配上用户的所定向的 一些条件，比如地理位置、兴趣属性、身份信息等。但要从这些大量的 广告当中选取一个最合适的广告，这时就需要借助机器学习来对每个 广告进行点击率(CTR，Click-Through-Rate)预估，进而选取点击率 最高的广告展现给用户。

发明内容

[0005] 鉴于目前存在的上述不足，本发明提供一种基于决策树和逻辑回 归的点击率预估方法，提出了结合了决策树和逻辑回归的预估模型，提 升了预估效果。

[0006] 为达到上述目的，本发明的实施例采用如下技术方案：

[0007] 一种基于决策树和逻辑回归的点击率预估方法，所述基于决策树 和逻辑回归的点击率预估方法包括以下步骤：

[0008] 获取投放信息的相关特征数据；

[0009] 建立基于决策树与概率稀疏线性分类器级联结构的点击率预估模 型；

[0010] 通过在线连接器生成实时训练数据；

[0011] 通过实时训练数据训练点击率预估模型进行获得最新的点击率预 估模型来进行点击率预估。

[0012] 依照本发明的一个方面，所述在线连接器的工作为：在数据中加入 标签并以在线方式训练输入的数据，将投放信息展现和投放信息点击 通过请求ID进行连接，每次用户使用时都会生成一个唯一的请求ID，通过这个ID将展现和点击进行匹配。

[0013] 依照本发明的一个方面，所述通过在线连接器生成实时训练数据 包括以下步骤：

[0014] 用户访问网站或者app，用户的相关信息会传递到系统中；

[0015] 系统通过排序将相关的投放信息返回给用户的设备上；

[0016] 将上述过程产生的数据记录在展现数据流中；

[0017] 当用户点击他所看到的投放信息时，这个点击数据记录在点击数 据流中；

[0018] 当时间窗口期过后，在线连接器就会把连接好的展现数据发送到 训练数据集中。

[0019] 依照本发明的一个方面，在通过在线连接器生成实时训练数据过 程中，需要建立异常检测机制。

[0020] 依照本发明的一个方面，使用在线学习方法训练线性分类器。

[0021] 依照本发明的一个方面，使用增强决策树来对特征进行转换。

[0022] 依照本发明的一个方面，所述增强决策树包括：每棵单独的树都为 一个分类特征，它的值就是树叶的索引值。

[0023] 依照本发明的一个方面，所述增强决策树训练数据的方式是以批 量形式进行训练的。

[0024] 依照本发明的一个方面，对每个特征都加入了特征权重，在每个树 节点结构中，选择并分割一个最佳特征，一旦一个特征在多棵树中使用 时，每个特征的重要性会通过将整棵树全部的损失值相加计算得出。

[0025] 依照本发明的一个方面，所述基于决策树和逻辑回归的点击率预 估方法包括：使用抽样方法处理大量训练数据。

[0026] 本发明实施的优点：本发明所述的基于决策树和逻辑回归的点击 率预估方法，包括以下步骤：获取投放信息的相关特征数据；建立基于 决策树与概率稀疏线性分类器级联结构的点击率预估模型；通过在线 连接器生成实时训练数据；通过实时训练数据训练点击率预估模型进 行获得最新的点击率预估模型来进行点击率预估；提出了一个基于决 策树与概率稀疏线性分类器级联结构的模型体系结构，它还包含了一 个在线学习层，并公开了在线连接器，它是一个在线学习层中非常关键 的组成部分，可以将训练数据转换成实时的流式数据；本发明所述的基 于决策树和逻辑回归的点击率预估方法，相较于现有的点击率评估方 法至少有10％的效果提升。

附图说明

[0027] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例 中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅 仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创 造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0028] 图1为本发明所述的一种基于决策树和逻辑回归的点击率预估方 法示意图；

[0029] 图2为本发明所述的训练数据的新鲜度测试结果示意图；

[0030] 图3为本发明所述的修改学习率进行模型的训练试验结果示意图；

[0031] 图4为本发明所述的不同的特征数量对结果的影响示意图；

[0032] 图5为本发明所述的均匀抽样训练结果示意图；

[0033] 图6为本发明所述的负样本抽样训练结果示意图。

具体实施方式

[0034] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方 案进行清楚、完

整地描述，显然，所描述的实施例仅仅是本发明一部分 实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技 术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属 于本发明保护的范围。

[0035] 在本实施例中，我们归一化熵(NE)和校准作为我们的主要评判 指标。NE的分子部分其实是交叉熵，也是LR的代价函数。(y是样 本label，取1或者-1；pi为样本i的点击预测概率)；分母部分 是原样本的信息熵(p为正样本的概率，或者准确的说是频率)，即 原始样本的不确定性。假设给定的训练数据集包含N条数据，每条数 据都有一个标签$y_i \in \{-1,+1\}$和预估的点击率p_i，其中i＝1,2,…N，记 实验的平均CTR为p，那么NE可表示为

[0036] $$NE = \frac{-\frac{1}{N}\sum i = 1(\frac{1+y_i}{2}\log(p_i) + \frac{1-y_i}{2}\log(1-p_i))}{-(p*\log(p) + (1-p)*\log(1-p))}$$

[0037] NE是计算相对信息增益(RIG)的一个根本组件，并且RIG＝1-NE，它帮助我们消除了样本的不确定性。当我们没有模型帮助的时候，样本 正负的不确定性会大，我们不是很容易判断样本正负；但有了模型帮助 之后，我们会得到一个预测的点击率，在这个帮助下我们就可以更容易 的去判断样本正负，这时候不确定性就下降了。如图1所示，一种基于决策树和逻辑回归的点击率预估方法，所述基于决策树和逻辑回归的 点击率预估方法包括以下步骤：

[0038] 步骤S1：获取投放信息的相关特征数据；

[0039] 所述步骤S1获取投放信息的相关特征数据的具体实施方式可为：采用决策树模型，首先对所述广告数据及用户数据形成的特征进行筛 选组合，生成区分度高更具代表性的强分类特征，即交叉特征。由此，一方面可以大大降低特征向量的维数，加快机器学习的收敛过程，提高 评估效率；另一方面由于采用更高区分度的特征进行广告点击率的评 估，可以得到更精确的评估值。

[0040] 获取预定历史时间段内特定历史投放广告的相关特征信息；所述 历史投放广告具体指在所述预定历史时间段内已经投放的广告，以各 种形式展示到用户界面，如搜索引擎的搜索列表，应用程序的消息栏提 示界面，应用程序的对话框界面等。所述预定历史时间段为预设时间内 维持特定历史广告不更新的时间段。获取所述预定历史时间段内的特定历史投放广告的相关特征信息，具体的，所述特定历史投放广告指进 行点击率预估的当前投放广告。其中，所述历史投放广告的相关特征信 息具体包括但不限于如下任意一项或多项：广告所属行业、广告尺寸、广告文本、广告图片、广告历史展现次数、广告历史点击次数、广告位 置归一化后的点击率。

[0041] 所述广告所属行业特征通过广告投放时注册的信息获取或通过其 内容简介等信息提取相应关键字获取；所述广告尺寸通过显示的尺寸 大小获取；所述广告文本直接通过其发布的信息获取；所述广告图片具 体为表征其图像特征的描述值，如特征向量，通过相应的图像特征提取 算法提取所述图片的相应特征；所述广告历史展现次数具体指统计的获取的特定历史时间段内展示给用户的次数；所述广告历史点击次数 指广告被展示后用户的点击次数；所述广告位置归一化后的点击率具 体指广告所显示位置经过一定算法计算后，选择最优位置进行展示后 用户的点击次数。

[0042] 获取目标用户的个性化特征信息；所述个性化特征信息具体指与 目标用户相关

的，表征其本身属性的特征信息。在具体实施例中，所述 目标用户个性化特征信息包括但不限于如下任意一项或多项：

[0043] 性别、省份、职业、收入、学校、年龄、学历、血型、星座、联网 方式、联网时间、偏好、婚恋情况。

[0044] 步骤S2：建立基于决策树与概率稀疏线性分类器级联结构的点击 率预估模型；

[0045] 所述步骤S2建立基于决策树与概率稀疏线性分类器级联结构的点 击率预估模型的具体实施方式可为：提出一种模型结构，增强决策树与 概率稀疏线性分类器的级联结构。

[0046] 在实际应用中，本实施例使用的在线学习模型是基于Stochastic Gradient Descent(SGD)算法，在特征转换之后，一个广告创意就 会由一个结构化向量组成：$x=(e_{i_1},e_{i_2},\cdots,e_{i_n})$，其中$e_i$表示第i个单元 向量，$i_1,i_2,\cdots,i_n$表示的是第n个输入特征的值，在训练阶段，我们 使用了二分标签$y\in\{+1,-1\}$来表示是否点击。当给定标签化的广告创 意(x,y)，那么权重的线性组合可以表示为：

[0047] $$s(y,x,w)=y\cdot w^{T}x=y\sum_{j=1}^{n}w_{j,i_j}$$

[0048] 其中w表示线性点击得分的权重向量。

[0049] 在贝叶斯在线学习模式中，其中两个关键因素，似然函数和优先度 的表达方式分别为：

[0050] $$p(y|x,w)=\emptyset(\frac{s(y,x,w)}{\beta})$$

[0051] 和

[0052] $$p(w)=\prod_{k=1}^{N}N(w_{k;}\mu_k,\sigma_k^2)$$

[0053] 其中$\emptyset(t)$表示的是标准正态分布的累积分布函数，N(t)表示的是标 准正态分布的密度函数，它的在线训练是通过期望匹配和矩匹配来实 现的，该模型由加权向量w近似后验分布的均值和方差组成，因此，我 们可以将上述公式更改为：

[0054] $$u_{i_j}\leftarrow u_{i_j}+y\cdot\frac{\sigma_{i_j}^2}{\Sigma}\cdot v(\frac{s(y,x,\mu)}{\Sigma})$$

[0055] $$\sigma_{i_j}^2\leftarrow\sigma_{i_j}^2\cdot[1-\frac{\sigma_{i_j}^2}{\Sigma 2}\cdot w(\frac{s(y,x,\mu)}{\Sigma})]$$

[0056] $$\Sigma^2=\beta^2+\sum_{j=1}^{n}\sigma_{i_j}^2$$

[0057] 其中v(t)：=N(t)/∮(t)，w(t)：=v(t)·[v(t)+t]。

[0058] 然而SGD算法中的似然函数的表达式为：

[0059] $p(y|x,w)=\mathrm{sigmoid}(s(y,x,x))$

[0060] 其中sigmoid(t)＝exp(t)/(1+exp(t))，我们通常把它称为逻辑回 归(LR)，该模型推理出了对数似然的衍生物并且将每一个坐标定步长 的梯度方向表示为：

[0061] $$w_{i_j} \leftarrow w_{i_j} + y \cdot \eta_{i_j} \cdot g(s(y,x,w))$$

[0062] 其中g是所有非零特征的对数似然梯度值，可表示为$g(s) := [\frac{y(y+1)}{2} - y \cdot \mathrm{sigmoid}(s)]$。

[0063] 具体的，所述生成决策树模型的过程简述如下：设数据样本集为S，首先根据某种策略选择一个属性，如用户年龄，依照该属性进行划分，如年龄30为分界，大于30岁的样本分为一个集合，小于30岁的样本 分为一个集合。具体的，用户各个性化特征作为一个属性，如性别、省 份、职业、收入、学校、年龄、学历、血型、星座、联网方式、联网时 间、偏好、婚恋情况等特征，分别基于一定的量化值进行划分，同时特 定历史投放广告的相关特征也分别作为一个属性，如广告所属行业、广 告尺寸、广告文本、广告图片、广告历史展现次数、广告历史点击次数、 广告位置归一化后的点击率等特征，分别基于相应的量化值进行进一 步划分，直到不能划分为止，从而生成决策树的不同叶子节点，所述每 个叶子节点表征一个交叉特征。

[0064] 在实际应用中，为了提高准确度，有两种简单的方法来改变线性分 类器的输入特征。对于连续特征，学习非线性变换的一个简单方法是把 特征放到一个bin集合中，然后将该bin当作一个分类特征。线性分 类器有效地学习了一个分段的常数非线性映射，学习有用的bin边界 是很重要的，并且有许多方法可以实现这一工作。第二种简单且有效的 转换方式是构建元组输入特征，对于分类特征而言，最笨的方法就是使 用笛卡尔乘积，但它有一个缺点就是不能将没用的组合进行修正，如果 输入特征都是连续的，则可以进行联合绑定，例如使用k-d树。

[0065] 增强决策树是一种强大且非常方便的方法可以实现我们刚才描述 的非线性和元组转换。我们将每棵单独的树都视为一个分类特征，它的 值就是树叶的索引值。例如，假设一颗决策树有2颗子树，其中一颗子 树有3个叶子节点，另一颗有2个叶子节点，这时有一条数据在子树1 的第2个叶子节点和子树2的第1个叶子节点结束，那么我们可以将 二分向量[0,1,0,1,0]作为线性分类器的输入值，其中前3个值代表的 是子树1的叶子节点，后两个值代表的是子树2的叶子节点。我们使 用的增强决策树遵循了梯度提升机(GBM)，在这里使用了经典的L2- TreeBoost算法，在每次学习迭代中，都会创建一颗新树对之前的树的 残余进行建模，基于决策树的转换是一种受监督的特征编码，它将实值 向量转换成一个紧凑的二进制值向量，从根节点到叶子节点的遍历其 实就是某些特征的规则，在二进制向量上拟合线性分类器，本质上就是 为一组规则学习权重，与其他方式不同的是增强决策树训练数据的方 式是以批量形式进行训练的，这大大可以节省训练时间。

[0066] 在实际应用中，我们进行了一些实验来展示将树的特征作为线性 模型的输入带来的影响，在该试验中我们比较了两个逻辑回归模型，其 中一个包含了特征转换逻辑，另一个直接使用的原始特征，之后我们也 把增强决策树进行了对比。对比结果如下表：

[0067]

模型	NE值
逻辑回归+增强决策树	96.58%

逻辑回归	99.43%
增强决策树	100%

[0068] 从表中可以看出使用了特征转化后NE值降低了近3%，这是非常明 显的效果提升。表中显示逻辑回归与决策树相结合的方式带来了更大 的提升。

[0069] 在实际应用中，为了使数据能保持最大的新鲜度，我们使用了在线 学习线性分类器的方式。

[0070] 评估不同的学习率对基于SDG的逻辑回归产生的影响。要实现该 目的，我们做了如下一些处理：

[0071] 1.对于在第t次迭代中特征i的学习率可以表达为

[0072] $$\eta_{t,i}=\frac{\alpha}{\beta+\sqrt{\sum_{j=1}^{t}\nabla_{j,i}^{2}}}$$

[0073] 其中α，β都是可调参数。

[0074] 2.每个权重的平方根学习率：

[0075] $$\eta_{t,i}=\frac{\alpha}{\sqrt{n_{t,i}}}$$

[0076] 其中$\eta_{t,i}$表示的是特征i迭代到第t次后的所有训练实例之和。

[0077] 3.每个权重的学习率：

[0078] $$\eta_{t,i}=\frac{\alpha}{\eta_{t,i}}$$

[0079] 4.全局学习率：

[0080] $$\eta_{t,i}=\frac{\alpha}{\sqrt{t}}$$

[0081] 5.即时学习率：

[0082] $\eta_{t,i}=\alpha$

[0083] 前三个等式针对每个特征设置了学习率，后两个等式所有的特征 的学习率都是一样的。其中可调的参数都是通过网格搜索的形式进行 优化的，具体优化值如下表：

[0084]

学习率等式	参数
1	$\alpha=0.1$，$\beta=1.0$
2	$\alpha=0.01$
3	$\alpha=0.01$

[0085]

4	$\alpha=0.01$
5	$\alpha=0.0005$

[0086] 通过上面几种方式来修改学习率进行模型的训练，实验的结果如 图3所示，可以看出，第1种方式有最优的NE值，第3种方式表现最 差，第2种方式和第5种方式的结果相似，第4种方式表现差的主要 原因可能是每个特征上的实例数量的不平衡导致的，因为每个训练实 例会包含不同的特征，这时就会出现有些特征含有更多的训练实例。在 使用第4种方式时，含有少量实例的特征的学习率就会急剧下降，并会 防止权重收敛到最优。虽然第3种方式没有这样的问题，但是因为它将 所有特征的学习率都降低了所以表现依然很差，这样就会导致在模型 收敛到一个非最优点的时候，训练就终止了，这也解释了为什么这种方式表现最差。

[0087] 步骤S3：通过在线连接器生成实时训练数据；

[0088] 点击预估系统通常是部署在一个动态的竞价环境里的，所以数据 分布会随着时间而变化，我们发现训练数据的新鲜度很大程度上会影 响到预测的性能。为了验证这个结论，我们使用了特定一天的数据作为 训练，然后将模型应用在接下来连续的几天竞价里。最终测试的结果如 图2所示，图中横坐标代表的是测试数据与训练数据相隔的天数，纵坐标表示NE值。从图中可以很明显的看出随着相隔天数的增加NE值也 相应的增加，所以在一段时间(不超过7天)过后需要重新训练最新的 数据以使模型保持最优，我们使用一个定时任务来实现这一目的，训练 增强决策树的时间取决于多方面的因素，包括树的数量，每棵树叶子节 点的数量，cpu，内存等等，在单cpu的情况下可能需要超过24小时的 时间来训练出一颗增强决策树。但在生产环境中，我们需要使用多核、足够内存的机器来并发的训练这样一棵树。

[0089] 越新的训练数据会提高预测的准确度，它还提供了一个简单的模 型体系结构，其中线性分类器层是在线训练的。

[0090] 在实际应用中，本实施例提出了一种实验系统，该系统可以生成实 时训练数据，并通过在线学习训练线性分类器。我们把这个系统称为 “在线连接器”，因为它的关键操作是加入标签(点击/不点击)并以在 线方式训练输入的数据(广告创意)。在投放过程中点击标签是可以实 时获取到的，但由于数据的延迟和网络原因我们并不能实时的知道该 用户是否未点击该广告，所以要知道广告创意是否被点击，必须在一定 的时间窗口期内对创意进行标签的设置，问题是这个时间窗口期到底 该设置多大呢。

[0091] 当该窗口期设置过长那么就需要更多的内存来缓存创意信息以等 待点击时间的发生，当设置过短则会丢失一些正常的点击数据。这会带 来“点击覆盖”的问题，所有点击的分数都成功地加入到这次展现中了， 因此，在线连接系统必须在重新连接和点击覆盖之间取得平衡。

[0092] 没有完整的点击覆盖意味着实时训练集将会有偏见：实验的CTR往 往会比真实的要低。这是因为，如果等待时间足够长的话，一小部分被 标记为“不点击”的数据就会被标记为“点击”。然而，在实践中，我 们发现在等待窗口不断变化的情况下，很容易将这种偏差减小，从而将 内存需求变得可控。此外，这种小偏差也可以被测量和纠正。在线连接 器的主要工作就是将广告展现和广告点击通过请求ID进行连接，每次 用户在银橙竞价系统中竞价都会生成一个唯一的请求ID，所以就可以 通过这个ID将展现和点击进行匹配。在线连接器的一个大致流程为：用户访问网站或者app，用户的相关信息会传递到银橙的竞价系统中，竞价系统通过排序将相关的广告返回给用户的设备上，这个过程产生 的数据会被

记录在展现数据流中，当用户点击他所看到的广告时，这个 点击数据就会记录在点击数据流中，当时间窗口期过后，在线连接器就 会把连接好(加入点击或未点击标签)的展现数据发送到训练数据集中。通过这种方式训练者就可以持续不断地生成最新的模型了。最终机器 学习模型形成了一个紧密的封闭循环，在这个模型中，特征分布或模型 性能的变化可以被捕获、学习并在短时间内得到纠正。

[0093] 在使用生成实时训练数据系统时，需要考虑的一个重要考虑因素 是要建立保护机制来防止可能破坏在线学习系统的异常现象。比如，当 点击数据流由于某些原因导致其中的数据都是旧的数据时，那么在线 连接器产生的训练数据就会变得非常小，这会导致实时的训练者训练 产生的模型预估出来的点击率变得很低，进而使竞价系统的广告展现 数降低。这时异常检测机制就可以帮助我们避免这类问题，比如当实时 训练数据分布突然改变，就可以自动断开在线连接器的在线训练。

[0094] 步骤S4：通过实时训练数据训练点击率预估模型进行获得最新的 点击率预估模型来进行点击率预估。

[0095] 在实际应用中，模型中的树越多，预测的时间就越长。在这部分，我们研究了增加树的数量对预估准确性的影响。我们将树的数量从1增 加到2000，训练的数据集是一整天的数据，测试数据是之后一天的数 据。测试后发现树的数量从0增加到500时NE值下降的比较明显，但 在之后NE值基本保持不变。所以并不是树越多效果越好，在训练过程 中往往会在某一个地方达到拟合。

[0096] 特征数量是另一个影响预估准确度和计算性能的因素，为了能更 好的理解特征数量的影响，我们对每个特征都加入了特征权重。在每个 树节点结构中，选择并分割一个最佳特征，以最大限度地减少平方误差，一旦一个特征在多颗树中使用时，每个特征的重要性会通过将整棵树 全部的损失值相加计算得出。

[0097] 根据经验，通常只有少量的特性会对模型产生较大的影响，而其他 大部分特性对模型的影响可以忽略不计。我们也针对该发现做了实验，只保留其中的10,20,50,100,200个特征时，然后评估不同的特征数量 对结果的影响，结果如图4所示，从图中可以看出在10-50这个区间 里，NE值下降的比较明显，而50-200NE值下降幅度较小，从而验证了 对模型影响较大的特征数量往往占很小的比例。

[0098] 在实际应用中，处理大量训练数据时，我们给出两种抽样数据的方 法并评估他们的优劣，这两种方法是：均匀抽样和负样本数据抽样。我 们将使用含有600颗树的增强决策树来作对比。

[0099] 对训练数据进行均匀抽样是一种很常用的方法，因为它实现简单 而且不需要修改样本数据就可以使用新生成的模型。在这部分中，我们 队不同的抽样率进行了评估，对于每一组样本数据，我们都会使用增强 模型进行训练，实验结果如图5所示，从图中可看出数据越多，模型效 果越好，并且使用10％的训练数据时，NE值比使用全部训练数据时只低了0.02，所以在做实验时我们不需要将所有数据进行训练。

[0100] 到目前为止，已经有很多研究人员对类不平衡的问题进行了大量 的研究，结果表明，这种不平衡会对学习模式的表现产生很大的影响，下面我们会使用负样本采样来解决类不平衡问题。同样地，我们将数据 使用多种采样率来进行效果对比，对比结果如图6所示，从图中可以看 出采样率在0.025时模型效果最好。

[0101] 本发明实施的优点:本发明所述的基于决策树和逻辑回归的点击 率预估方法,包括以下步骤:获取投放信息的相关特征数据;建立基于 决策树与概率稀疏线性分类器级联结构的点击率预估模型;通过在线 连接器生成实时训练数据;通过实时训练数据训练点击率预估模型进 行获得最新的点击率预估模型来进行点击率预估;提出了一个基于决 策树与概率稀疏线性分类器级联结构的模型体系结构,它还包含了一 个在线学习层,并公开了在线连接器,它是一个在线学习层中非常关键 的组成部分,可以将训练数据转换成实时的流式数据;本发明所述的基 于决策树和逻辑回归的点击率预估方法,相较于现有的点击率评估方 法至少有10%的效果提升。

[0102] 以上所述,仅为本发明的具体实施方式,但本发明的保护范围并不 局限于此,任何熟悉本领域技术的技术人员在本发明公开的技术范围 内,可轻易想到的变化或替换,都应涵盖在本发明的保护范围之内。因 此,本发明的保护范围应以所述权利要求的保护范围为准。

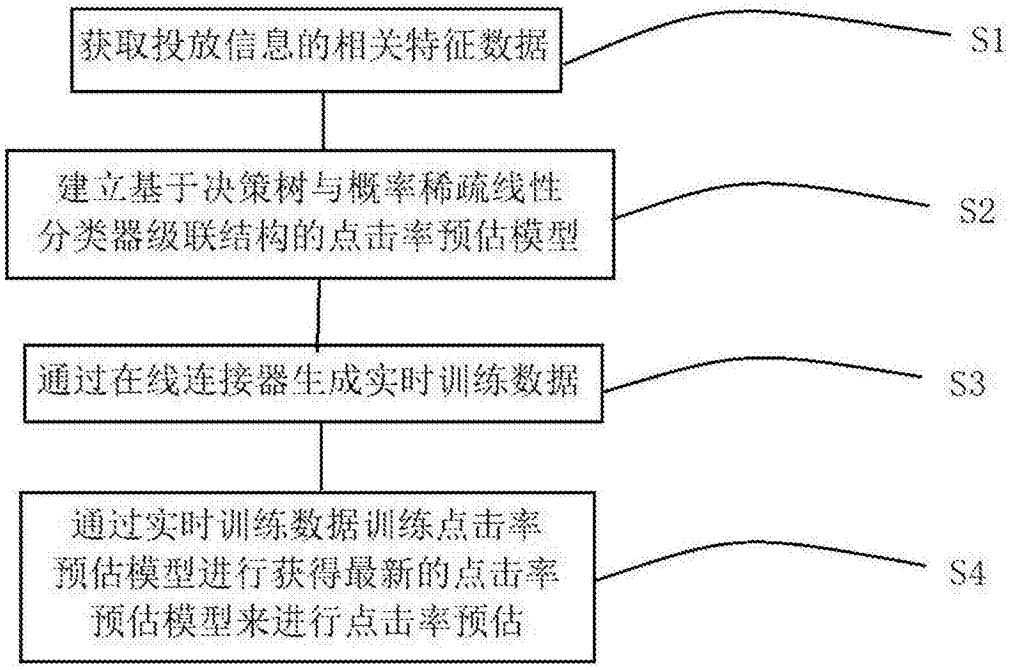

获取投放信息的相关特征数据
S1
建立基于决策树与概率稀疏线性分类器级联结构的点击率预估模型
S2
通过在线连接器生成实时训练数据
S3
通过实时训练数据训练点击率预估模型进行获得最新的点击率预估模型来进行点击率预估
S4

图1

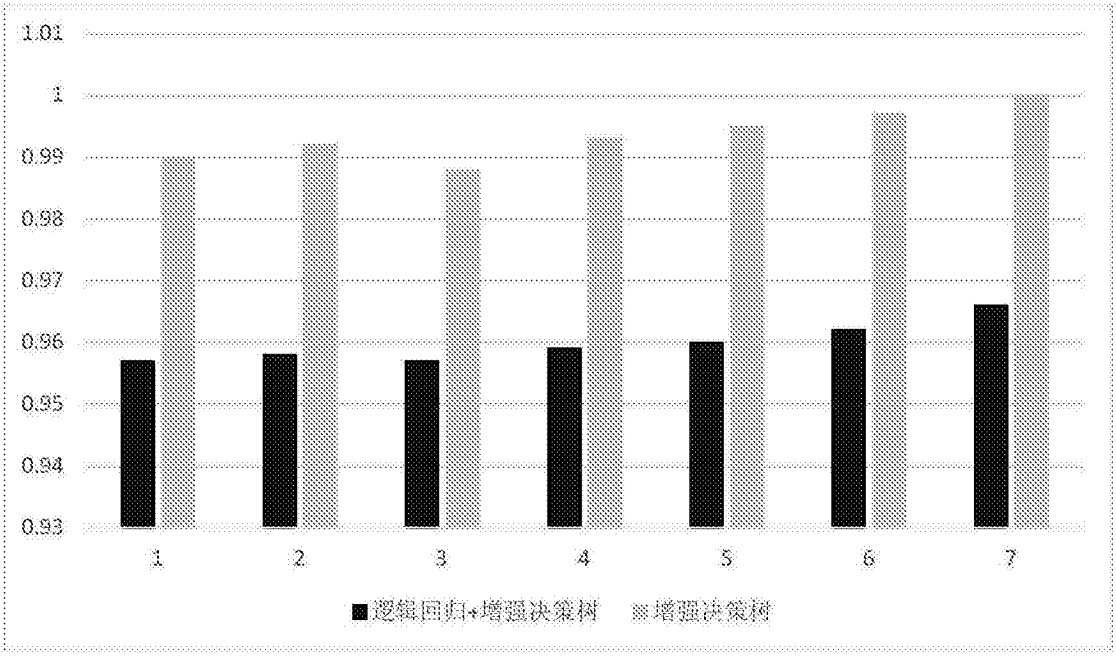

1.01
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
1
2
3
4
5
6
7
逻辑回归+增强决策树
增强决策树

图2

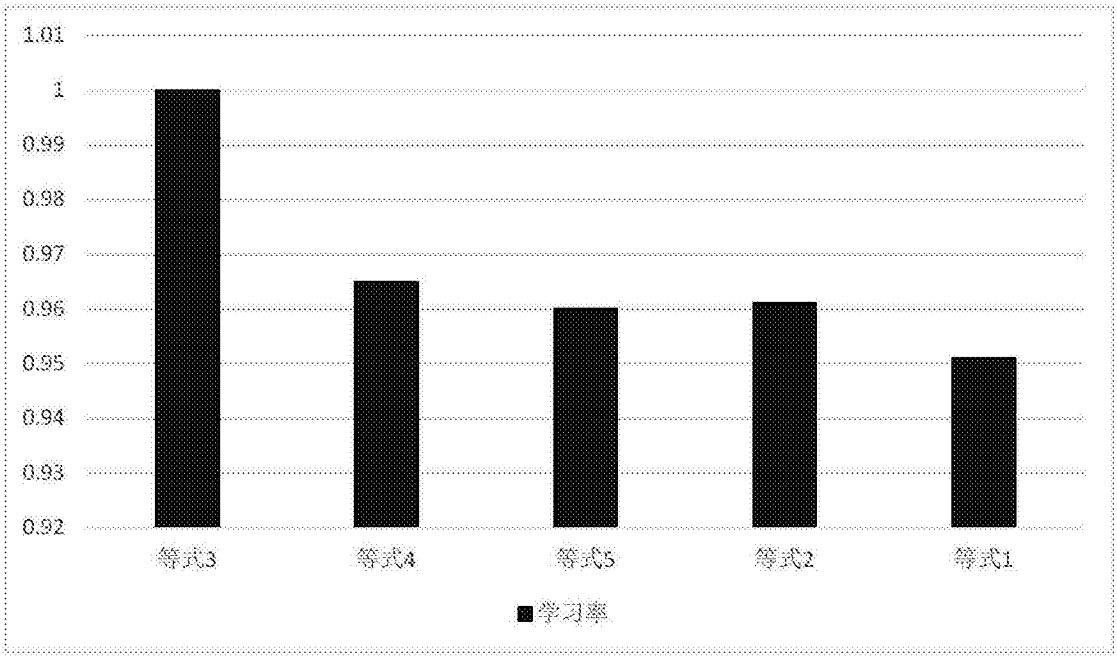

1.01
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
等式3
等式4
等式5
等式2
等式1
学习率

图3

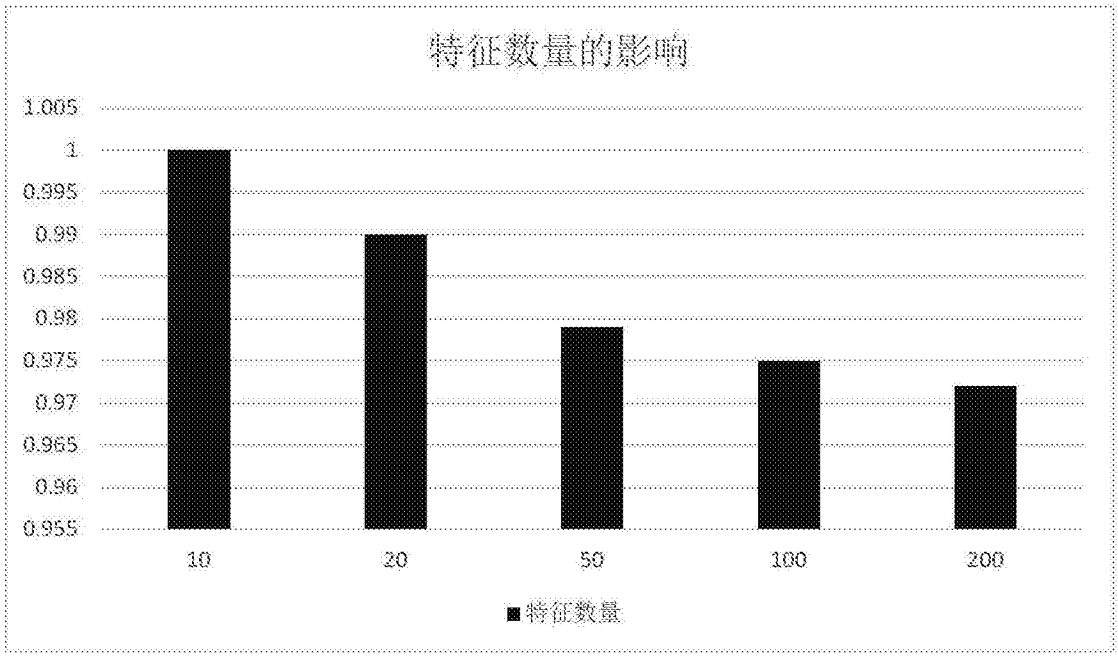

特征数量的影响
1.005
1
0.995
0.99
0.985
0.98
0.975
0.97
0.965
0.96
0.955
10
20
50
100
200
特征数量

图4

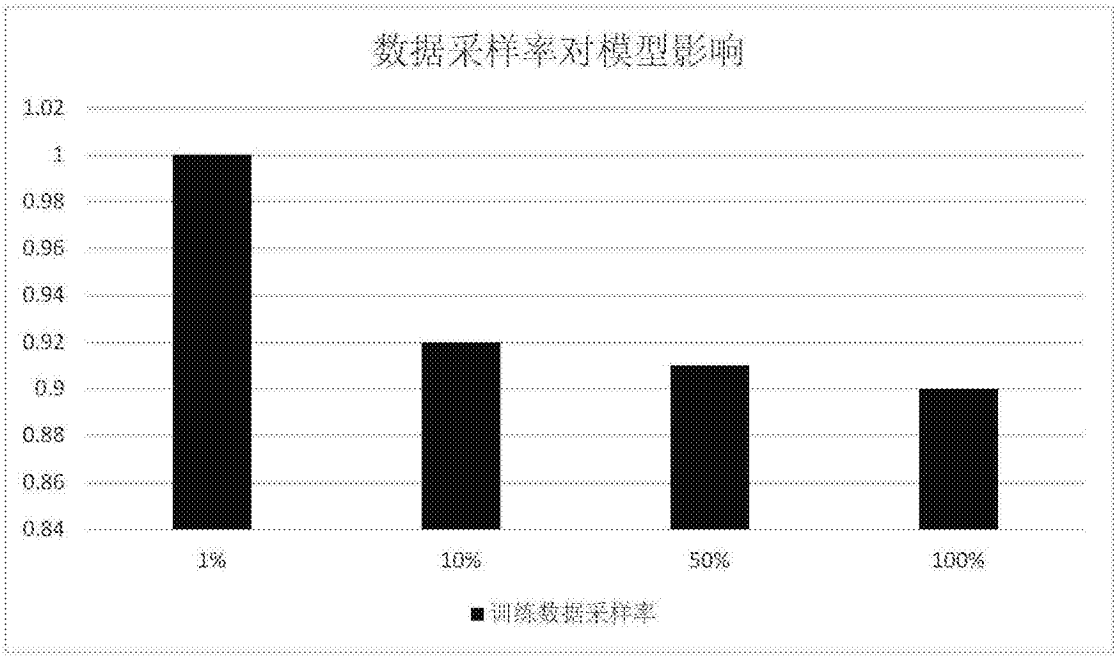

数据采样率对模型影响
1.02
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
1%
10%
50%
100%
训练数据采样率

图5

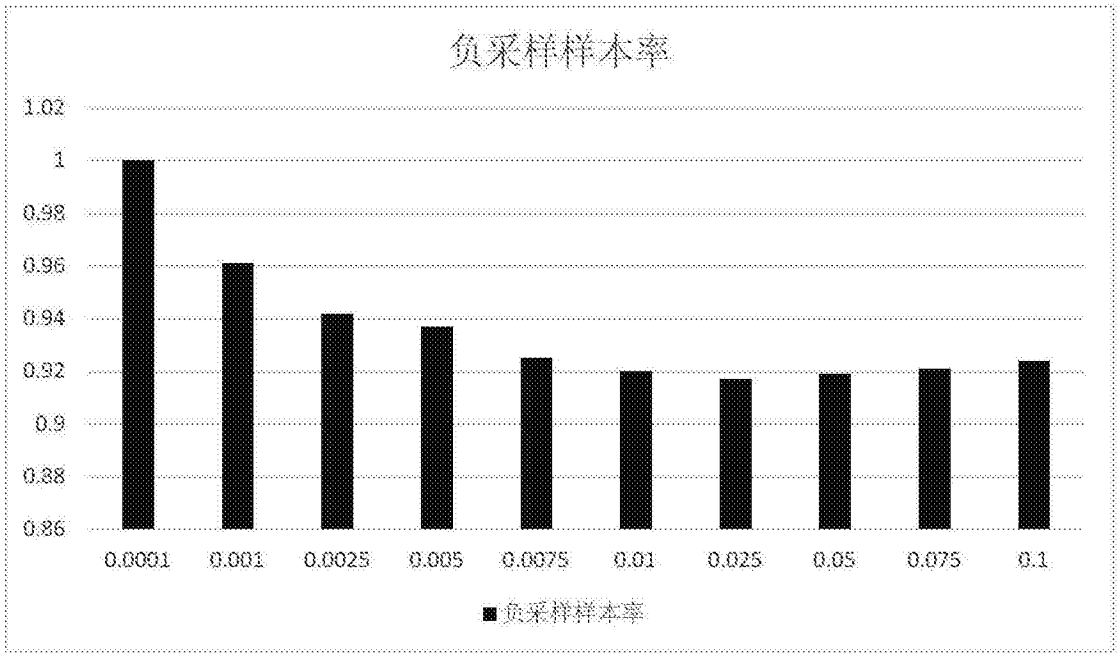

负采样样本率
1.02
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.0001
0.001
0.0025
0.005
0.0075
0.01
0.025
0.05
0.075
0.1
负采样样本率

图6