(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 110689368 A
(43)申请公布日 2020.01.14

(21)申请号 201910780066.X

(22)申请日 2019.08.22

(71)申请人 北京大学(天津滨海)新一代信息技术研究院
地址 300452 天津市滨海新区中心商务区于家堡金融区双创大厦25层

(72)发明人 刘譞哲 马郓 吕广利 陈震鹏 陆璇

(74)专利代理机构 北京辰权知识产权代理有限公司 11619
代理人 刘广达

(51)Int.Cl.
G06Q 30/02(2012.01)

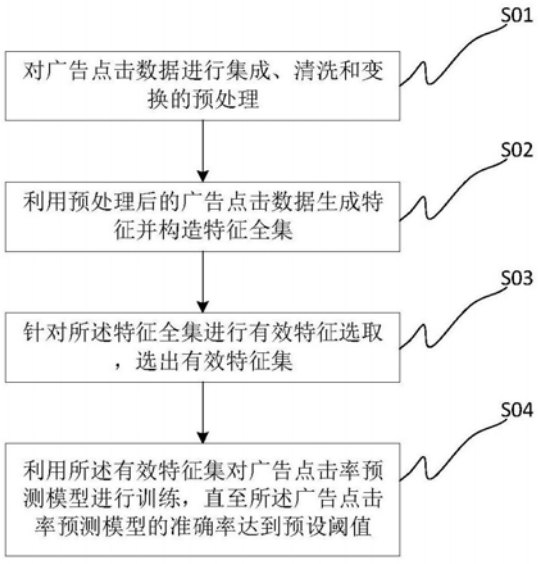
权利要求书2页 说明书19页 附图3页

(54)发明名称

一种移动应用内广告点击率预测系统设计方法

(57)摘要

本发明公开了一种移动应用内广告点击率预测系统设计方法,包括:对广告数据进行集成、清洗和变换的预处理;利用预处理后的广告数据生成特征并构造特征全集;针对所述特征全集进行有效特征选取,选出有效特征集;利用所述有效特征集对广告点击率预测模型进行训练。本发明实施例提供的移动应用内广告点击率预测方法,对广告数据中的长尾数据按照相似性进行归类,按照数据取值频次进行归类,克服了现有技术无法有效利用隐含在长尾数据中的信息的缺陷,充分利用了长尾数据中的信息提升了预测效果。



CN 110689368 A

1.一种移动应用内广告点击率预测系统设计方法,其特征在于,包括:

对广告数据进行预处理,包括集成、清洗和变换;

利用预处理后的广告数据生成特征并构造特征全集;

针对所述特征全集进行有效特征选取,选出有效特征集;

利用所述有效特征集对广告点击率预测模型进行训练。

2.根据权利要求1所述的方法,其特征在于,所述广告数据包括广告展示记录数据、广告点击记录数据和第三方应用程序信息;

所述对广告数据进行集成、清洗和变换的预处理包括:

将具有相同识别匹配标记的所述广告展示记录数据和所述广告点击记录数据合并为同一次广告投放活动;

根据应用程序的包名与展示点击记录来匹配合并相同的第三方应用程序信息。

3.根据权利要求2所述的方法,其特征在于,所述对广告数据进行集成、清洗和变换的预处理还包括:

处理数据缺失值;

去除同一次广告投放活动数据的重复数据;

对异常数据进行检测、判断及删除。

4.根据权利要求1所述的方法,其特征在于,所述对广告数据进行集成、清洗和变换的预处理还包括:提取隐含属性;提取连续数据。

5.根据权利要求1所述的方法,其特征在于,所述利用预处理后的广告数据生成特征并构造特征全集,包括:

对经过预处理的广告数据中的连续数据进行离散化操作,生成连续数据离散化特征;

对经过预处理的广告数据中的长尾数据进行归类,生成长尾数据归类特征;

将经过预处理的广告数据中的离散数据直接作为离散数据特征,与所述连续数据离散化特征和所述长尾数据归类特征共同构造特征全集。

6.根据权利要求5所述的方法,其特征在于,所述对经过预处理的广告数据中的连续数据进行离散化操作,生成连续数据离散化特征,包括:枚举所有类型的连续数据;

分别利用每种类型的连续数据生成特征集;

调用梯度提升树模型训练,得到经过验证的梯度提升树模型后,提取梯度提升树的所有树的所有内部结点的分裂值,组成该种类型数据的分箱数组。

7.根据权利要求5所述的方法,其特征在于,所述对经过预处理的广告数据中的长尾数据进行归类,生成长尾数据归类特征,包括:

定义取值频次相同或相近的用户为同一类型用户,将取值频次相同或相近的用户分组;

设置特征取值频次阈值,将每组用户中的取值频次低于阈值的长尾数据丢弃,使每组用户中的取值频次大于或等于阈值的长尾数据直接进入特征集。

8.根据权利要求1所述的方法,其特征在于,所述针对所述特征全集进行有效特征选取,选出有效特征集,包括:

步骤(1)对所述特征全集的所有特征进行评估,筛选并标记所有无益特征,并将对广告点击率预测模型影响最大的无益特征从所述特征全集中删除得到新的特征集,再利用所述

特征集更新所述特征全集；

步骤(2)对所述更新后的特征集内的所有无益特征进行评估，筛选并标记该次评估产生的新无益特征，取消其他所述无益特征的标记，并将对所述广告点击率预测模型影响最大的所述新无益特征从所述更新后的特征集中删除，再次更新所述特征集；

若未产生新无益特征，则停止操作，得到的特征集为有效特征集；

若产生新无益特征，则迭代执行步骤(2)，直至未产生新无益特征。

9.根据权利要求1所述的方法，其特征在于，所述利用所述有效特征集对广告点击率预测模型进行训练，包括特征编码、模型训练和与样本不均衡采样；

所述特征编码包括：

对所述有效特征集中的所有特征的特征值进行统计计数，对于取值频次小于频次阈值的特征值，将其丢弃或者将其设置为稀疏特征值，从而完成特征过滤；

对完成特征过滤后的所有特征进行排序，对排序后的所有特征的特征值建立特征向量，完成特征编码；

所述模型训练包括：

将特征编码得到的特征向量与对应的广告数据一同用于广告点击率预测模型训练，得到模型参数。

10.根据权利要求9所述的方法，其特征在于，所述样本不均衡采样包括：对负样本进行采样或增加正样本的权重。

一种移动应用内广告点击率预测系统设计方法

技术领域

[0001] 本发明涉及互联网技术领域，具体涉及一种移动应用内广告点击率预测系统设计方法。

背景技术

[0002] 近年来，随着移动互联网的迅速发展，由开发者、应用市场、用户、广告商等角色构成的移动生态系统逐渐形成。在这一生态系统下，移动应用内广告是移动应用的主要收入来源之一，而广告的精准投放是提高广告收入的关键。提高广告点击率的重点在于构造有效的广告点击率预测模型，为用户寻找最为匹配的广告。作为广告精准投放的关键技术之一，广告点击率预测得到了学术界和产业界的广泛关注。

[0003] 目前，广告点击率预测模型通常采用逻辑回归算法，基于多维特征的线性组合来预测广告点击率。为充分利用大量数据中的有效信息，提高广告点击率预测的准确性，需要进行复杂的特征设计，然而现有方法存在以下三个主要问题：

[0004] 广告点击率预测模型通常将连续数据进行离散化处理，现有的处理方法大多通过人工寻找连续数据取值的临界点，按临界点将其离散化，效率低且易出错。因此，对于连续数据离散化中的临界点选择问题，需要在保证准确性的同时，尽可能提升自动化程度；

[0005] 广告点击率预测模型往往涉及大量的长尾数据。在模型训练中，长尾数据对效果的增益相对较小，甚至可能导致模型参数较多。现有的处理方法往往直接舍弃长尾数据，但是，这会导致隐含在长尾数据中的信息无法得到有效利用，降低预测效果。因此，需要根据长尾数据的特点设计特征，以充分利用其中的信息；

[0006] 为了更好地表达特征之间的非线性关系，现有的方法是对特征进行两两组合，生成一系列组合特征。但是，现有技术的点击率预测模型使用的特征往往数目较大，导致穷举后产生的候选特征数量过多。因此，需要从全体候选特征当中进行高效的特征选择。

发明内容

[0007] 本发明的一个目的是提供一种移动应用内广告点击率预测系统设计的新的技术方案。为了对披露的实施例的一些方面有一个基本的理解，下面给出了简单的概括。该概括部分不是泛泛评述，也不是要确定关键/重要组成元素或描绘这些实施例的保护范围。其唯一目的是用简单的形式呈现一些概念，以此作为后面的详细说明的序言。

[0008] 根据本发明实施例的一个方面，提供一种移动应用内广告点击率预测方法，包括：

[0009] 对广告数据进行预处理，包括集成、清洗和变换；

[0010] 利用预处理后的广告数据生成特征并构造特征全集；

[0011] 针对所述特征全集进行有效特征选取，选出有效特征集；

[0012] 利用所述有效特征集对广告点击率预测模型进行训练。

[0013] 进一步地，所述广告数据包括广告展示记录数据、广告点击记录数据和第三方应用程序信息；

[0014] 所述对广告数据进行集成、清洗和变换的预处理包括：

[0015] 将具有相同识别匹配标记的所述广告展示记录数据和所述广告点击记录数据合并为同一次广告投放活动；

[0016] 根据应用程序的包名与展示点击记录来匹配合并相同的第三方应用程序信息。

[0017] 进一步地，所述对广告数据进行集成、清洗和变换的预处理还包括：

[0018] 处理数据缺失值；

[0019] 去除同一次广告投放活动数据的重复数据；

[0020] 对异常数据进行检测、判断及删除。

[0021] 进一步地，所述对广告数据进行集成、清洗和变换的预处理还包括：提取隐含属性；提取连续数据。

[0022] 进一步地，所述利用预处理后的广告数据生成特征并构造特征全集，包括：

[0023] 对经过预处理的广告数据中的连续数据进行离散化操作，生成连续数据离散化特征；

[0024] 对经过预处理的广告数据中的长尾数据进行归类，生成长尾数据归类特征；

[0025] 将经过预处理的广告数据中的离散数据直接作为离散数据特征，与所述连续数据离散化特征和所述长尾数据归类特征共同构造特征全集。

[0026] 进一步地，所述对经过预处理的广告数据中的连续数据进行离散化操作，生成连续数据离散化特征，包括：枚举所有类型的连续数据；

[0027] 分别利用每种类型的连续数据生成特征集；

[0028] 调用梯度提升树模型训练，得到经过验证的梯度提升树模型后，提取梯度提升树的所有树的所有内部结点的分裂值，组成该种类型数据的分箱数组。

[0029] 进一步地，所述对经过预处理的广告数据中的长尾数据进行归类，生成长尾数据归类特征，包括：

[0030] 定义取值频次相同或相近的用户为同一类型用户，将取值频次相同或相近的用户分组；

[0031] 设置特征取值频次阈值，将每组用户中的取值频次低于阈值的长尾数据丢弃，使每组用户中的取值频次大于或等于阈值的长尾数据直接进入特征集。

[0032] 进一步地，所述针对所述特征全集进行有效特征选取，选出有效特征集，包括：

[0033] 步骤(1)对所述特征全集的所有特征进行评估，筛选并标记所有无益特征，并将对广告点击率预测模型影响最大的无益特征从所述特征全集中删除得到新的特征集，再利用所述特征集更新所述特征全集；

[0034] 步骤(2)对所述更新后的特征集内的所有无益特征进行评估，筛选并标记该次评估产生的新无益特征，取消其他所述无益特征的标记，并将对所述广告点击率预测模型影响最大的所述新无益特征从所述更新后的特征集中删除，再次更新所述特征集；

[0035] 若未产生新无益特征，则停止操作，得到的特征集为有效特征集；

[0036] 若产生新无益特征，则迭代执行步骤(2)，直至未产生新无益特征。

[0037] 进一步地，所述利用所述有效特征集对广告点击率预测模型进行训练，包括特征编码、模型训练和与样本不均衡采样；

[0038] 所述特征编码包括：

[0039] 对所述有效特征集中的所有特征的特征值进行统计计数，对于取值频次小于频次阈值的特征值，将其丢弃或者将其设置为稀疏特征值，从而完成特征过滤；

[0040] 对完成特征过滤后的所有特征进行排序，对排序后的所有特征的特征值建立特征向量，完成特征编码；

[0041] 所述模型训练包括：

[0042] 将特征编码得到的特征向量与对应的广告数据一同用于广告点击率预测模型训练，得到模型参数。

[0043] 进一步地，所述样本不均衡采样包括：对负样本进行采样或增加正样本的权重。

[0044] 本发明实施例提供的技术方案可以包括以下有益效果：

[0045] 本发明实施例提供的移动应用内广告点击率预测方法，对广告数据中的长尾数据按照相似性进行归类，按照数据取值频次进行归类，克服了现有技术无法有效利用隐含在长尾数据中的信息的缺陷，充分利用了长尾数据中的信息提升了预测效果。

[0046] 进一步地，本发明实施例提供的移动应用内广告点击率预测方法，采用双向式特征选择的方式对特征集进行选择筛选，降低了迭代次数，不需要再对特征全集进行迭代，可以得到较大的模型提升效果，特征选择工程时间复杂度低，特征选择耗时短，工作效率高，可以很好地满足实际应用的需求。

附图说明

[0047] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述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中记载的一些实施例，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。

[0048] 图1为本申请一个实施例的流程图；

[0049] 图2为GBDT算法编码图；

[0050] 图3为长尾数据分布示意图；

[0051] 图4为GBDT寻找临界点的原理示意图；

[0052] 图5为数据预处理的流程示意图；

[0053] 图6为逻辑回归图。

具体实施方式

[0054] 为了使本发明的目的、技术方案及优点更加清楚明白，下面结合附图和具体实施例对本发明做进一步说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。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。

[0055] 本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语(包括技术术语和科学术语)，具有与本发明所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是，诸如通用字典中定义的那些术语，应该被理解为具有与现有技术的上下文中的意义一致的意义，并且除非像这里一样被特定定义，否则不会用理想化或过于正式的含义来解释。

[0056] 如图1所示，本申请的一个实施例提供一种移动应用内广告点击率预测系统设计方法，所述方法包括：

[0057] 步骤S01、对广告数据进行集成、清洗和变换的预处理；

[0058] 步骤S02、利用预处理后的广告数据生成特征并构造特征全集；

[0059] 步骤S03、针对所述特征全集进行有效特征选取，选出有效特征集；

[0060] 步骤S04、利用所述有效特征集对广告点击率预测模型进行训练，直至所述广告点击率预测模型的准确率达到预设阈值。

[0061] 在一些实施例中，所述对广告数据进行集成、清洗和变换的预处理包括：

[0062] 将具有相同识别匹配标记的广告展示记录数据和广告点击记录数据合并为同一次广告投放活动；

[0063] 根据包名与展示点击记录来匹配合并第三方数据。

[0064] 在一些实施例中，所述广告数据包括广告展示记录数据、广告点击记录数据和第三方应用程序信息；

[0065] 所述对广告数据进行集成、清洗和变换的预处理包括：

[0066] 将具有相同识别匹配标记的所述广告展示记录数据和所述广告点击记录数据合并为同一次广告投放活动；

[0067] 根据应用程序的包名与展示点击记录来匹配合并相同的第三方应用程序信息。广告数据主要来自广告投放系统的日志收集系统，以及第三方数据库描述信息。

[0068] 在一些实施例中，所述对广告数据进行集成、清洗和变换的预处理还包括：

[0069] 处理数据缺失值；去除同一次广告投放活动数据的重复数据；对异常数据进行检测、判断及删除。

[0070] 在一些实施例中，所述对广告数据进行集成、清洗和变换的预处理还包括：提取隐含属性；提取连续数据。

[0071] 在一些实施例中，所述利用预处理后的广告数据生成特征并构造特征全集，包括：

[0072] 对经过预处理的广告数据中的连续数据进行离散化操作，生成连续数据离散化特征；

[0073] 对经过预处理的广告数据中的长尾数据进行归类，生成长尾数据归类特征；

[0074] 将经过预处理的广告数据中的离散数据直接作为离散数据特征，与所述连续数据离散化特征和所述长尾数据归类特征共同构造特征全集。

[0075] 在一些实施例中，所述对经过预处理的广告数据中的连续数据进行离散化操作，生成连续数据离散化特征，包括：枚举所有类型的连续数据；

[0076] 分别利用每种类型的连续数据生成特征集；

[0077] 调用梯度提升树模型训练，得到经过验证的梯度提升树模型后，提取梯度提升树的所有树的所有内部结点的分裂值，组成该种类型数据的分箱数组。

[0078] 在一些实施例中，所述对经过预处理的广告数据中的长尾数据进行归类，生成长尾数据归类特征，包括：

[0079] 定义取值频次相同或相近的用户为同一类型用户，将取值频次相同或相近的用户分组；

[0080] 设置特征取值频次阈值，将每组用户中的取值频次低于阈值的长尾数据丢弃，使

每组用户中的取值频次大于或等于阈值的长尾数据直接进入特征集。

[0081] 在一些实施例中，所述利用预处理后的广告数据生成特征并构造特征全集，包括：

[0082] 利用广告数据生成的离散数据特征、连续数据离散化特征和长尾数据归类特征构造特征集。

[0083] 在一些实施例中，所述连续数据离散化特征的获取方法包括：

[0084] 枚举所有类型的连续数据；

[0085] 分别利用每种类型的连续数据生成特征集；

[0086] 调用梯度提升树模型训练，得到经过验证的梯度提升树模型后，提取梯度提升树的所有树的所有内部结点的分裂值，组成该种类型数据的分箱数组。

[0087] 在一些实施例中，所述长尾数据归类特征的获取方法包括：

[0088] 定义取值频次相同或相近的用户为同一类型用户，将取值频次相同或相近的用户分组；

[0089] 设置特征取值频次阈值，将每组用户中的取值频次低于阈值的长尾数据丢弃，使每组用户中的取值频次大于或等于阈值的长尾数据直接进入特征集。

[0090] 在一些实施例中，所述针对所述特征全集进行有效特征选取，选出有效特征集，包括：

[0091] 步骤(1)对所述特征全集的所有特征进行评估，筛选并标记所有无益特征，并将对广告点击率预测模型影响最大的无益特征从所述特征全集中删除得到新的特征集，再利用所述特征集更新所述特征全集；

[0092] 步骤(2)对所述更新后的特征集内的所有无益特征进行评估，筛选并标记该次评估产生的新无益特征，取消其他所述无益特征的标记，并将对所述广告点击率预测模型影响最大的所述新无益特征从所述更新后的特征集中删除，再次更新所述特征集；

[0093] 若未产生新无益特征，则停止操作，得到的特征集为有效特征集；

[0094] 若产生新无益特征，则迭代执行步骤(2)，直至未产生新无益特征。

[0095] 可选地，所述对所述特征全集的所有特征进行评估，筛选并标记所有无益特征，包括：

[0096] 用所述特征全集对广告点击率预测模型进行训练和测试，得到评估参考值；

[0097] 对所述特征全集中的每个特征进行缺省检验，获取每个特征的评估值；

[0098] 将每个所述评估值分别与所述评估参考值进行对比，若评估值优于所述评估参考值，则将对应的特征标记为无益特征。

[0099] 可选地，所述对所述更新后的特征集内的所有无益特征进行评估，筛选并标记该次评估产生的新无益特征，包括：

[0100] 用所述更新后的特征集对所述广告点击率预测模型进行训练和测试，得到新的评估参考值；

[0101] 对所述更新后的特征集中的每个无益特征进行缺省检验，获取对应于特征集中每个无益特征的评估值；

[0102] 将对应于所述特征集中每个无益特征的每个评估值分别与所述新的评估参考值进行对比，若评估值由于所述新的评估参考值，则保留其无益特征标记；

[0103] 可选地，所述取消其他所述无益特征的标记，包括：

[0104] 若评估值劣于或等同于所述评估参考值，则取消对应的无益特征的无益特征标记。

[0105] 可选地，所述进行缺省检验的步骤包括：

[0106] 单独删除特征集合(可以为特征全集或特征集)中的某个特征，形成与该被删除的特征相对应的特征集，利用所述与该被删除的特征相对应的特征集对所述广告点击率预测模型进行训练和测试，得到对应于该被删除的特征的评估值。

[0107] 在一些实施例中，所述利用所述有效特征集对广告点击率预测模型进行训练，包括：特征编码、模型训练和与样本不均衡采样。

[0108] 在一些实施例中，所述特征编码包括：

[0109] 对所述有效特征集中的所有特征的特征值进行统计计数，对于取值频次小于频次阈值的特征值，将其丢弃或者将其设置为稀疏特征值，从而完成特征过滤；

[0110] 对完成特征过滤后的所有特征进行排序，对排序后的所有特征的特征值建立特征向量，完成特征编码。

[0111] 在一些实施例中，所述模型训练包括：

[0112] 将特征编码得到的特征向量与对应的广告数据一同用于广告点击率预测模型训练，得到模型参数。

[0113] 在一些实施例中，所述样本不均衡采样包括：对负样本进行采样或增加正样本的权重。

[0114] 在一些实施例中，所述利用所述有效特征集对广告点击率预测模型进行训练，还包括：在对负样本进行采样后，对所述广告点击率预测模型的常数项进行修正。

[0115] 在一些实施例中，所述广告点击率预测模型包括逻辑回归算法模型。

[0116] 本实施例还提供一种电子设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述程序，以实现上述的移动应用内广告点击率预测系统设计方法。

[0117] 本实施例还提供一种非临时性计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行，以实现上述的移动应用内广告点击率预测系统设计方法。

[0118] 本申请的另一个实施例提供的一种移动应用内广告点击率预测系统设计方法，包括：

[0119] 步骤S10、对广告数据进行预处理；

[0120] 步骤S20、对预处理后的广告数据进行特征生成操作，生成有效特征集；

[0121] 步骤S30、利用所述有效特征集对广告点击率预测模型进行训练，直至所述广告点击率预测模型的准确率达到阈值。

[0122] 如图5所示，步骤S10包括：

[0123] S101数据集成；S102数据清洗；S103数据变换。

[0124] S101数据集成的步骤包括：S1011广告投放数据合并；S1012第三方数据合并。

[0125] 数据集成即将多个数据源合并，对每次展示请求生成一条完整的数据会话。数据源来自日志收集模块和第三方数据源。日志收集模块收集的包括广告的展示记录和点击记录。第三方数据源用于描述和app相关的信息，以及IP库等。数据集成需要不同数据源具有相同的key才可以合并。广告展示记录和广告点击记录都有一个字段用于识别匹配，被称为

Session Id,具有相同Session Id 的广告展示记录和广告点击记录属于同一次广告投放。app的相关信息如 category、download和rating等信息来自于Google Player Store等第三方可信数据库,这些信息可以通过app的包名与展示点击记录来匹配。数据集成会导致很多重复数据和异常数据,如展示数据缺失但是点击数据存在,展示数据时间戳大于点击数据时间戳等。这就需要对数据进行数据清洗。

[0126] S102数据清洗的步骤包括:S1021数据缺失值处理、S1022数据去重以及 S1023异常数据检测。

[0127] 缺失值处理方式多样,最简单的方式是丢掉该条数据。但是这种方式会丢掉大量数据,因为很多情况下都会有数据缺失,尤其在现在隐私权保护越来越受重视的情况下。用户可以主动管理手机端的权限,所以大部分和手机信息相关的数据都可能难以收集,比如地理位置信息。第三方数据源也可能更新不及时,一些不在Google Play Store上架的应用就无法获得关于这个应用的下载量,评分等。以上因素会导致的普遍现象就是几乎每条数据都会有字段缺失的情况发生,因此简单丢掉该条数据不是合适的选择。另一种方式是补充缺失值。补充缺失值是常见的做法,针对不同类型的数据,补充的方式也有很多种。

[0128] 本实施例将数据分为三种情况,字符串类型的数据可以分为离散数据和长尾数据,两者的区别在于该数据的不同数据取值的总数和每种数据取值的取值频次分布。对于离散数据,一般具有较少且固定的个数的数据取值。如国家信息属于离散数据,特点是特征值个数有限且固定。对于长尾数据,数据取值个数不固定,且不同数据取值个数较多,很多数据取值频次都很少。如User Id, 可达百万甚至上亿的特征值,但是大部分特征值都出现较少的次数。数字类数据即连续数据,如用户评分等信息。对于离散特征,可以填充缺失值为unknown,也可以填充该维度取值频次最多的特征值,即众数值,对于连续特征,则可以填充平均值,众数,中位数,甚至0值。对于如User Id类型的长尾数据,只能填充unknown值。

[0129] 数据去重是指由于网络问题等原因导致的数据重复。同一次广告投放活动,会有一个惟一的Session Id,但是由于网络问题等原因,会出现两条或者多条数据包含同一个Session Id,这时就只能保留一条数据,其他数据是这条数据的重复。

[0130] 异常数据的检测是指有些数据属于不合理数据,需要判断是否保留。如点击率超高,已经完全超出了合理范围,可以判定为误点击率高,或者强制用户点击的数据。另一种情况经常存在于预点击的广告位,有些广告位为了更快速地响应用户的点击行为,在请求到广告后,甚至在广告还未展示前就在后台加载了广告的点击链接。这样如果用户发生点击行为就可以迅速跳转到最终落地页。但是这种预点击行为会给机器学习带来很大困扰,因为实际点击行为有可能并没有发生。这类数据显然不合理,需要去除。

[0131] 还有些缺失了展示日志但是却有点击日志的数据,这种数据也不能进入训练集,应该删除。因为这类数据很可能是虚假点击数据,即广告甚至并没有展示给用户,但是后台却悄悄加载了点击链接,这是广告欺诈的一种。也可能是因为展示请求由于网络问题丢失了。另外,还有具有统一Session Id的广告点击记录和广告展示记录的时间戳相差较大,如差距达到一个小时甚至更高,这种情况的广告点击记录也需要筛除,但是广告展示记录可以保留,因为这种 Session中,广告真实的展示给了用户,但是点击延迟太高,不可信,有可能是因为广告被客户端缓存了很长时间。然而广告主并不希望这种情况发生,每次请求的广告会设置一个过期时间,达到过期时间的广告不会再响应点击行为。

[0132] 除此之外，还有一类异常数据，如广告点击时间戳小于广告展示时间戳，即广告点击事件先于广告展示事件的发生，这类数据一般是虚假数据。这可能是客户端bug或者一些第三方SDK工具的广告拦截行为造成的。这种情况需要去掉整个Session。

[0133] S103数据变换的步骤包括：S1031隐含属性提取；S1032连续数据提取。

[0134] 数据变换是指一些数据不能直接使用，需要提取这类数据的隐含属性。如时间戳，是连续数据，但是这个值在逻辑回归算法中几乎无法使用。这个值是非线性的，但是时间戳隐含的属性却很重要，需要转化为线性数据。如时间戳可以提取出小时信息，日期信息，星期信息，是否是周末信息等。甚至于可以结合时区信息从时间戳提取本地时间信息。

[0135] 另外IP数据也不能直接用，因为IP是固定的。但是使用IP的用户却可以变化，而且IP数据量很大，导致每个IP取值频次都有限。IP数据属于长尾数据，但是一般不会直接使用，而是会根据IP数据来提取地理位置信息，如根据 IP可以提取国家信息、城市信息等，时区信息也可以通过IP数据提取。

[0136] 另一方面，数据变换里一个重要的数据提取就是CTR信息，即广告点击率。遍历数据集来生成各种维度的CTR信息也是数据的二次加工。而CTR数据是描述数据特点的重要数据，同时这个值也是一个反馈数据。广告点击率预测系统预测广告的点击率，会导致广告投放发生变化，广告的CTR发生变化，变化的 CTR又会进入训练集，反馈到点击率预测系统。CTR数据属于全局数据，必须通过遍历数据集来获得，每个Session都需要查询CTR表来获取当前某个数据维度的CTR信息。一个常见的做法是通过数据集成操作将CTR信息保存到所有的 Session中，每个Session只保留和自己相关的CTR信息。虽然这样会导致数据冗余，数据存储开销增大，但是方便后续的处理。

[0137] S20特征生成包括：S201离散特征生成、S202长尾数据归类和S203双向式特征选择的步骤。特征生成部分主要内容是连续数据离散化、长尾数据归类和特征组合。特征生成依赖于离散特征生成算法、长尾数据归类算法和双向式特征选择算法这三种特征生成算法生成的特征配置文件。

[0138] 本实施例的广告点击率预测是指在特定环境下对用户点击广告概率的预测。具体而言，在特定上下文环境与特定用户的环境下，预测不同广告的点击率。基于这一预测，广告投放系统进一步结合广告价值等信息来决定展示广告的顺序。为了精准预测广告的点击率，充分利用大数据中的有效信息，需要进行复杂的特征设计。

[0139] S201离散特征生成：离散特征生成算法针对的是连续数据离散化效率低下且准确度不足的问题。

[0140] 为什么要对连续数据进行离散化，首先要分析离散化特征有哪些优势。离散特征的内积乘法运算快、效率高、容易扩展。而将连续数据离散化主要考虑的因素有以下几点：

[0141] 一是相近的数值对模型的影响相近，离散化是为了将相近的连续值合并，使得他们在模型中有相同的表现。但是在临界值附近的数据取值就会出现摇摆，即划分到哪一个箱都可以，这就对离散化方法有很高的要求。

[0142] 二是离散化可以减少异常值对模型训练的干扰。比如当某个广告位的 CTR＝1时，所有广告展示行为都带来了点击，那么这个广告位的数据取值对预测广告点击概率有很大影响。之所以CTR这么高，除了数据不合理之外，还有一种可能就是广告曝光量太少。所以不能直接删除。同理还有CTR＝0的情况。通过离散化可以减轻这两种情况对模型训练造成的

干扰。

[0143] 如何处理连续值是点击率预测模型常见的一个问题。连续值在点击率预测系统中重要性不言而喻。大多数连续值是从数据集得到的统计结果，因此具有反馈意义。比如广告的历史点击率，是描述广告吸引用户程度的数据。

[0144] 一般情况下，连续数据的离散化都是由有经验的人提出多种合理的划分方式，然后通过实验来确定哪一种划分方式对机器学习最终的效果提升更多，来决定采用哪一个或者几个最有效的划分方式。然而这种划分方式不仅要求开发人员掌握相关数据划分的经验，而且这些划分方法很难避免临界值附近的摇摆问题。数据集一旦改变，尤其在广告投放系统中，随着数据的积累和反馈，很多连续数据的分布会发生变化，又需要重新提出多种划分，重复测试。即使对于经验丰富的程序员来说，寻找连续数据的临界点也是很艰难的事情。

[0145] 现有技术中处理连续数据时一般就是使用这种方式的。这种方式需要统计连续数据的分布，然后尝试将每个点作为分割点，去判断将数据集分割为两个部分后的方差，将方差较小的点当做临界点数组，然后通过实验验证这些点的某部分组合是否合适，这种划分方式的缺点是依靠人工划分临界点组合很难覆盖到最优划分。通过这种方式寻找一个可以接受的划分大概需要2-3天的时间。

[0146] 本实施例通过分析发现，在寻找临界点时，通过分析这个数据划分数据集为两部分的方差之和来判断是否属于临界点，这和GBDT算法逻辑很相似。GBDT （梯度提升树，Gradient Boosting Decision Tree，简称GBDT)算法的树在分裂子结点时，也是通过判断分裂后的子结点方差之和最小来选择分裂阈值。因此本实施例结合Facebook的工作，提出了一种程序自动化寻找连续数据临界点的方法。Facebook的做法是将数据集传递给GBDT模型训练，得到每条数据在GBDT 的每颗树的叶子节点编号作为新的编码，参考图2所示，然后使用经典的逻辑回归模型做训练。

[0147] 本实施例同样参考这一方法进行连续数据离散化。GBDT算法可以有效地处理连续数据，其主要原因是树结构可以对连续数据进行二分。因此，只要使用 GBDT算法对连续值划分，就可以得到连续数据候选的临界点。本实施例获取的是GBDT的树的内部结点的值。GBDT树的内部结点描述的是划分的规则，对于连续数据，内部结点的值则是一个临界点，使用这个点进行连续数据的划分，可以得到最大的方差增益。因此这个点可以用来作为本实施例的连续数据候选临界点。那么只需要使用单个连续数据来训练模型，GBDT树的所有内部结点就全部都是候选的临界点。只要多训练几棵树，就可以把所有候选的临界点都提取出来。当然，这种方式得到的结果也需要验证。如图3，可以得到最多6个值的临界点用作分箱，分别为：(node1.value，node2.value，node3.value，node4.value，node5.value，node6.value) 这些点需要去重，因为可能会有重复值。

[0148] 离散特征生成算法思路：枚举所有类型的连续数据。对于每种连续数据，生成训练集，训练集的数据只包含这种数据，标签为正负样本值0或者1。调用GBDT模型训练，得到经过验证的GBDT模型后，提取GBDT的所有树的所有内部结点的分裂值threshold，组成数组。这个数组就是这种类型数据的分箱数组。为了简化模型以及便于计算，可以为threshold指定小数精确度。在提取 GBDT树的节点时，每个树的权重不一样，树在GBDT模型的树数组中下标越大，树的权重越小，同一颗树的不同层的节点重要性也不同。因此可以得到前i个临

界点组成的分箱数组(bin_0,…bin_i),然后验证i的取值从1到n,这些划分哪个对模型训练的提升最大,来决定i的取值。

[0149] 使用GBDT算法来分析连续数据临界点的方式相较于人工寻找临界点不容易出错。人工寻找临界点,所有临界点都是基于数据全集的划分。而GBDT算法是每次划分后基于划分的子集寻找临界点。而且每个临界点的重要性是已经排序的。因此只需要数次验证即可得到较为合理的划分方式,在数个小时内即可完成。该方法将离散特征生成时间从天级别降低到小时级别。把连续型数据切分为若干“段”,也称bin。对于连续数据,本实施例采用GBDT算法进行寻找连续数据的bin值。首先,构造连续数据集合,然后对每类数据生成训练数据集,经过训练得到GBDT模型。将模型内每棵树节点的threshold合并去重排序,即得到每个连续数据的候选临界点。可以优化的地方是第18行,如果不做去重排序的话,就可以对最终的Bin进行重要度划分,因为越靠前的Bin值,权重越高,越有意义。另外一个优化是针对Bin的精确度做优化,我们得到的Bin 值由于计算机精确度的原因,很可能是十几位的小数,但是最终想要的精确度一般精确到小数点后三位就可以很好地区分点击率数据了,因此,可以对最终的Bin值进行有效数字位数处理,也可以事先对数据集的s类型数据进行精确度处理,使得最终生成的Bin值有效数字位数固定。

[0150] 本实施例设计离散特征生成算法是为了快速对连续数据离散化。本实施例认为,通过GBDT算法寻找连续数据的临界点具有重要性排序,可以更快速的寻找临界点组合。不需要像人工寻找临界点一样完全依靠经验来判断。相对于经验划分,本实施例只需要对每个连续数据执行一次GBDT算法即可得到通过所有可能的划分方式,减少了开发者寻找连续数据临界点及进行排列组合的工作量。同时本实施例的离散特征生成算法可以寻找到很多开发人员分析数据难以考虑到的临界点。

[0151] S202长尾数据归类:

[0152] 长尾数据归类算法设计针对的是长尾数据难以被有效利用的问题。

[0153] 广告投放数据中包含广告Id,广告组Id,用户Id等数据,这类数据统称为长尾数据。这些数据的一个重要特点就是数量多,甚至多达百万维,呈现长尾分布。长尾分布如图4。

[0154] 长尾数据在个性化推荐上很重要。例如针对用户做个性化推荐,如果不利用用户Id特征,就无法做个性化匹配;利用用户Id数据,就需要考虑长尾部分数据,因为长尾效应一个特点就是长尾数据累积量超过了流行数据。广告投放里,搜索推荐类广告的价值高于普通广告位,就是因为搜索类广告的投放是个性化投放,完全针对搜索词来投放,可以理解为针对搜索相同词的用户做个性化推荐。根据这一点,可以拓展到针对同一类型用户做推荐。那么问题的关键在如何界定同一类型用户。

[0155] 目前大多数研究对长尾数据的处理方式都是精细到数据取值,这种方式在大数据集上理论是可行的,只要数据集足够大,足够覆盖到每个数据取值,并且每个数据取值出现的次数足够多,就可以学习到合适的机器学习模型。但是实际数据集很难满足这个要求。尤其是用户数据,有很多长尾数据,用户浏览广告次数分布不均匀。为了考虑用户使用的体验,大多数应用对同一用户展示广告的频次都不会很高。目前机器学习的研究大都是研究改进算法,然而现在并没有针对长尾数据效果很好的算法,针对长尾数据都是先过滤低频次数据后直接进行one-hot编码,这种方式精度虽高,但是由于数据的局限性,都很容易过

拟合。

[0156] 本实施例借鉴了搜索类广告的特点。搜索类广告点击率高，并不是细分到用户做推荐，而是对同一类型用户做推荐，其定义同一类型用户的关键就是搜索词一致的用户。本实施例所做广告推荐需要从其他维度来划分同一类数据取值。针对长尾数据的稀疏性，本实施例设计了长尾数据归类算法。具体地，如果数据取值频次足够进入训练集且不造成过拟合，则保留该数据取值；反之，将这部分数据取值合并归类，使其可以进入训练集。

[0157] 本实施例从用户的取值频次出发，定义取值频次相同或相近的用户视为同一类型用户，将取值频次相同或者相近的用户分组，这样每组用户的取值频次就足够多到能够进入训练集。这种做法主要是针对长尾数据有效，使得更多的长尾数据参与到训练集。如果直接进行one-hot编码，那么取值频次少的长尾数据或者有很大的权重导致过拟合，或者因为设置了特征取值频次阈值而被过滤掉，都不是好的选择。

[0158] 本实施例设置了特征取值频次阈值index_threshold，即数据取值频次低于阈值将会被丢弃。但是数据取值频次少的长尾数据通过相似归类，就有可能进入训练集。取值频次大于阈值的长尾数据占比小，这部分长尾数据取值可以直接进入训练集。这样进行长尾数据归类，相较于不进行长尾数据归类，能够引进更多的长尾数据，有利于个性化广告推荐。

[0159] 数据集会出现另外一种情况，即取值频次少的长尾数据占比会很大，因为大部分长尾数据取值都是很小，简单使用长尾数据归类算法，会导致取值频次很少的数据取值进入训练集。但是取值频次不是特别少，却低于index_threshold 的数据即使合并后也未必能进入训练集。这时直接使用数据取值频次等长归类就不是最好的选择，可以采用数据取值频次对数归类。具体使用哪一种更合适，可以通过实验来决定。除此之外，也可以考虑针对长尾数据，单独设置 index_threshold。

[0160] 长尾数据归类算法并不会从理论上对模型有较大的提升。然而通过充分利用长尾数据，对超参数的调试以及定义归类规则，有望使得模型效果得到一定的提升。因为长尾数据是个性化数据，这类数据只要能利用上且不给模型带来负影响，就可能在实际的个性化推荐时发挥作用。从这个角度，长尾数据具有高于流行数据的价值。对推荐算法而言，个性化推荐是关注的重点，因此这部分数据对真实场景的个性化推荐可能有很大的影响，进而带来点击率的提升。

[0161] S203双向式特征选择:

[0162] 双向式特征选择算法设计针对的是特征两两组合导致的候选特征数量过于庞大，难以有效选择特征的问题。

[0163] 为了更好地表达特征之间的非线性关系，现有的做法是对特征进行两两组合，生成一系列组合特征。对于广告投放中的某些属性来说，单个离散特征有时候难以单独描述清楚，这时可以考虑将不同离散特征进行组合。比如不同的广告发布者拥有多个广告位，且对广告位没有统一的命名规定，因此需要将广告发布者的其他属性如应用包名和广告位组合，才能具体定位到广告展示的真实位置。然而，点击率预测模型使用的特征数目巨大，特征两两组合会导致候选的特征数量过多，难以进行高效的特征选择。

[0164] 经典的特征组合选择方式，是将组合特征和原始离散特征放到一起，作为训练集的特征全集，然后每次遍历全集筛掉一个特征，直到找不到可以筛除的特征。

[0165] 本实施例对特征工程的优化思路是，首先在每一轮迭代时得到对模型效果有较大正面影响的特征，记为特征集H；接着，从剩余特征中去除对模型效果负面影响最大的特征，最终剩余特征记为特征集I，进入下一轮迭代；最后，特征集H与特征集I组成特征全集S，用于下一轮模型训练。这种优化方法会降低迭代的特征集合，不再对全集进行迭代，只对上述特征集I进行迭代。

[0166] 组合特征对模型效果的影响不确定，本实施例假设产生更好效果的概率符合二项分布。假设离散特征有n个，那么组合特征有

[0167] $$m = \frac{n*(n-1)}{2}$$

[0168] 个，理论上有一半的组合特征会提升模型效果，因此有一半的特征即m/2 个特征是需要从迭代集删除的，则共需要

[0169] $$m + \left(\frac{m}{2} - 1\right) * \frac{\left(\frac{m}{2} - 1 + 1\right)}{2}$$

[0170] 近似于$m^2/8$次训练。而传统的特征选择工程需要

[0171] $$\frac{m}{2} * \frac{m + m/2}{2}$$

[0172] 近似于$3m^2/8$次训练。本实施例设计的算法理论上可以将平均搜索时间减少66％。

[0173] 优化后的双向式特征选择算法流程如下

[0174] 1.构造特征集合全集，包括离散特征，连续数据的离散化特征，Id类特征，特征全集为S，迭代集合为I＝S，重要特征集合为H＝S-I＝0。

[0175] 2.使用全集S作为特征集，进行模型的训练和测试，得到评估值e_0。

[0176] 3.从训练集I去掉一个特征I_i，然后训练和测试，得到评估值e_i，重复这个过程，直到I中每个特征都被迭代一次。

[0177] 4.选择评估值相对于e_0提升最大的I_i，如果最大的提升为正值，即相对于使用全集S为特征集，有提升，从S集合去掉特征I_i，同时将特征集合I中去掉某一特征后模型效果变差的特征添加到结合H中，并从I 中删除该特征，否则不更新。

[0178] 5.重复步骤2、3、4，直到步骤4中模型效果不再有提升，此时的S集合即为有效特征集合。

[0179] 可选地，双向式特征选择的过程包括以下步骤：

[0180] 步骤S31、利用所述离散数据特征、所述连续数据离散化特征和所述长尾数据归类特征构造特征集；

[0181] 步骤S32、使用所述特征集对广告点击率预测模型进行训练和测试，得到评估参考值；

[0182] 步骤S33、对特征集中的每个特征进行缺省检验，获取对应于特征集中每个特征的缺省检验评估值，简称评估值；

[0183] 对特征集中的某个特征进行缺省检验的步骤包括：

[0184] 单独删除特征全集中的某个特征，形成对应于该被删除的特征的训练集，利用该训练集对广告点击率预测模型进行训练和测试，得到对应于该被删除的特征的评估值；

[0185] 步骤S34、将所述对应于每个特征的评估值分别与所述评估参考值进行对比，根据对比结果判断与每个评估值相对应的特征对于广告点击率预测模型性能的影响好坏；

[0186] 如果评估值优于评估参考值，则表明利用对应于该被删除特征的训练集对广告点击率预测模型进行训练和测试能够使所述逻辑回归模型性能变好，说明去掉该特征对于广告点击率预测模型性能是有利的，也就是说该特征对广告点击率预测模型性能的影响是坏的，则将该特征标记为无益特征；对所有的无益特征对模型性能影响程度进行排序，将所有的无益特征中对模型性能影响程度最大的特征称为最差无益特征(即对应于该无益特征的评估值相对于评估参考值来说是最差的)；

[0187] 如果评估值劣于评估参考值，则表明利用对应于该被删除特征的训练集对广告点击率预测模型进行训练和测试会导致所述逻辑回归模型性能变差，说明去掉该特征对于广告点击率预测模型性能是不利的，也就是说该特征对广告点击率预测模型性能的影响是好的，则将所对应的特征标记为有益特征；

[0188] 如果评估值等同于评估参考值，则表明利用对应于该被删除特征的训练集对广告点击率预测模型进行训练和测试对所述逻辑回归模型的性能无影响，说明去掉该特征对于广告点击率预测模型性能是无影响的，也就是说该特征对广告点击率预测模型性能是无影响的，则将该特征标记为一般特征；

[0189] 具体地，在一些实施例中，对于评估值与评估参考值的比较方法，可以设定阈值a，如果评估值减去评估参考值所得到的差大于阈值a，则认为评估值优于评估参考值；如果评估参考值减去评估值所得到的差大于阈值a，则认为评估值劣于评估参考值；如果评估参考值减去评估值所得到的差等于阈值a，则认为评估值等同于评估参考值；当然也还有其他的一些比较方法，根据具体算法的不同而有所区别；

[0190] 步骤S35、从特征集中删除最差无益特征并更新特征集；

[0191] 步骤S36、利用更新后的特征集对广告点击率预测模型进行训练和测试，得到新的评估参考值；

[0192] 步骤S36、对更新后的特征集中的每个无益特征进行缺省检验，获取对应于特征集中每个无益特征的新的评估值；

[0193] 步骤S37、将步骤S36中的每个新的评估值分别与所述新的评估参考值进行对比，根据对比结果判断与所述每个无益特征的评估值相对应的无益特征对于广告点击率预测模型性能的影响好坏；

[0194] 将对广告点击率预测模型性能是影响好的或无影响的原无益特征分别改标记为有益特征或一般特征；新的评估值优于所述新的评估参考值则表示新的评估值所对应的原无益特征对于广告点击率预测模型性能是影响好的；新的评估值等同于所述新的评估参考值则表示新的评估值所对应的原无益特征对于广告点击率预测模型性能是无影响的；

[0195] 保留对模型性能影响坏的无益特征的无益特征标记；从对模型性能影响坏的无益特征中选出影响最大的无益特征并将该影响最大的无益特征从所述更新后的特征集中删除，并再次更新特征集，然后转向步骤S36；新的评估值劣于所述新的评估参考值则表示新的评估值所对应的原无益特征对于广告点击率预测模型性能是影响坏的；

[0196] 若没有产生新的无益特征则停止操作，此时的特征集为有效特征集。所述的新的无益特征即与所述新的评估参考值进行对比得到的对于广告点击率预测模型性能的影响

坏的无益特征。将该有效特征集应用于广告点击率预测，效果较好。

[0197] 双向式特征选择的一个优点是会降低迭代次数，不需要再对全集进行迭代，可以得到较大的模型提升效果。特征组合对非线性特征转化为线性特征有帮助，另外特征组合增加了数据集的特征维度，能够更完善地描述数据集。

[0198] 本实施例期望双向式特征选择算法可以得到较大的模型提升效果。因为特征组合对非线性特征转化为线性特征有帮助，另外特征组合增加了数据集的特征维度，能够更完善地描述数据集。

[0199] 特征配置文件需要指明：

[0200] 1)所有连续数据的离散化临界点数组，以及每种连续数据的离散化方式；

[0201] 2)长尾数据的归类方式是等长划分还是对数划分，若为等长划分，其步长是多少；

[0202] 3)进行两两组合的离散特征，以及他们两两组合后是否保留原来的离散特征。

[0203] 接下来，本实施例将结合特征配置文件具体阐述特征生成模块涉及的三个步骤：

[0204] 1.连续数据离散化：在连续数据的配置中，需要指明使用哪些分箱数组。我们需要根据分箱配置，对连续数据进行离散化。如对于年龄数据，进行三段划分：(min-17,18-23,24-max)，这样就可以将年龄连续值用一个只有三个特征值的离散特征来表示。连续数据可以被分解为多个特征。如年龄数据可以分解为三分类特征(min-17,18-23,24-max)，也可以分解为五分类特征 (min-12,13-18,19-25,26-30,31-max)。这样一个年龄数据就成为了两个离散特征，这两个离散特征看似有一定的关联，但是却又并不互斥。在传统的特征工程里，因为不能确定如何对连续数据进行划分合适，往往会有多个分箱法组合起来，对连续数据进行描述。因为大多数时候一个依靠经验得到的分箱法很难描述清楚连续数据的特点。

[0205] 2.长尾数据归类：对于长尾数据，需要指定长尾数据归类的规则，本实施例设计归类规则为，按照取值频次分箱。长尾数据的一个重要特点就是长尾效应。对数据归类并不是说所有特征值都要归类，流行数据可以直接使用，这部分数据不会出现过拟合。但是长尾数据因为特征值出现频次低，很难直接放到训练集，因此需要对其进行归类，然后对合并归类后的特征再尝试放入训练集。这样做相对于现有的处理方式就是考虑了长尾数据，这部分数据的累加效果甚至大于流行数据的效果。对于长尾数据，只需要按照特征值的取值频次归类，取值频次相近的特征值视为同一个特征值，就可以大大压缩长尾数据的特征值个数，同时一定程度上保留了特征值包含的信息。

[0206] 3.特征组合：根据特征配置文件，这类特征只需要对每条数据生成一个新的特征列，这个特征列为多个离散特征的组合即可。为了格式上的统一，一般会设计为多个特征名称的字符串拼接加上多个特征值的字符串拼接。

[0207] S30模型训练包括：S301特征编码、S302模型训练与S303样本不均衡采样的修正的步骤。

[0208] 在进行模型训练前，需要对特征进行one-hot编码，中文名独热编码。逻辑回归模型最适合的特征向量是0-1编码的向量，这种向量的好处是训练迭代快，向量的每个位的重要程度是一样的。所谓one-hot编码，即一个特征具有多少个特征值，就编码为多少维的0-1向量，除了那个特征值所在的向量位为 1，其他位均为0。

[0209] 编码前，需要做特征过滤。因为逻辑回归很容易过拟合，所以对于取值频次较少的特征值需要预先处理，否则逻辑回归算法会对取值频次较少的特征值会赋予很高的权重，

导致模型过拟合。特征值取值频次阈值index_threshold 为正整数，首先对所有特征值进行统计计数，取值频次小于index_threshold 的特征值会被特殊处理。

[0210] 一种做法是直接丢弃，不进行编码，这种做法会导致某些数据的某个特征的one-hot编码所有维全为0。另一种处理方式是将取值频次少于 index_threshold的特征值归类到一起，设置为稀疏特征值。具体选用哪种方式需要实验验证两者的效果，效果好的方式会被采用。如果两者表现相同，可以采用直接丢弃的方式，能够减少计算量、减小配置文件大小和降低算法复杂度。

[0211] 假设进行特征值过滤后，剩余N个特征值。对于这N个特征值，进行从小到大排序。对排序后的特征值建立特征向量，特征向量的维度即为N，每一位的初始取值为0。如果某条数据包含特征向量中的第i个特征值，则其对应的特征向量的第i位置为1。这就是one-hot编码过程。

[0212] 编码得到的特征向量与对应的点击数据一同用于逻辑回归算法训练。训练完成后可以得到该算法的参数，包括W和b，其中W为N维的浮点型数组，b为浮点型数字。逻辑回归算法公式为

[0213] $$p = \frac{1}{1 + e^{-(W*X+b)}}$$

[0214] 其函数图像如图6所示：点击率预测问题是一个典型的分类问题。即预测结果是0或者1。0代表用户不点击，1代表用户点击。但是线上预测并不是预测用户是否点击，而是预测用户点击的概率，因此可以将点击率预测问题转化为回归问题。逻辑回归算法很适合处理这类问题。设特征向量为X，参数向量为W，则：

[0215] $X=(x_0,x_1,\ldots,x_{n-1})$

[0216] $W=(w_0,w_1,\cdots,w_{n-1})$

[0217] 逻辑回归公式如下：

[0218] $$y = \frac{1}{1 + e^{-W*X}}$$ 其中，$$h(z) = \frac{1}{1 + e^{-z}}$$

[0219] 称为sigmoid函数，也可简写为sigmod，该函数是典型的S型函数(如图 6所示)，可以将自变量映射到值域为(0,1)的区间。sigmoid函数的导数可以用自身表示，即

[0220] $h'(z)=h(z)(1-h(z))$

[0221] 逻辑回归算法是广告点击率预测中应用最广泛的算法之一。它计算简单高效，特征模型可解释性强，模型训练快。因此，本实施例采用该算法作为点击率预测系统的模型算法。

[0222] 机器学习里关于样本不均衡分布的问题，在广告投放数据里表现得很明显。在广告投放数据里，表现为展示数据多，点击数据少，点击率偏低。一般来说，广告的点击率在1%上下。这种数据集直接进行机器学习，训练出来的模型会严重偏向于负样本，得到的预测CTR偏低，甚至全部预测为不点击，也可以得到 99%的正确率。因此需要对正负样本不均衡问题进行处理。

[0223] 一种方法是增加正样本的权重，另一种方法是对负样本进行采样。两种方法处理后训练的模型相较于不处理模型会偏向正样本。进行上述方法处理后需要对模型进行修正。以负样本采样为例，负样本采样后训练的模型是对采样后数据的拟合。因此需要对训练

后的模型修正,修正公式推导过程如下:

[0224] 预测的点击率为pCTR,真实点击率为rCTR,正样本数为p,负样本数为n, 负样本采样率为r,训练出来的模型的权重为PW,截距为pb,真实的模型为W,截距为b,我们想要求出真实的W。对于真实数据来说,理论上应该满足以下公式

[0225] $$rCTR = \frac{p}{p+n} \quad , \quad rCTR = \frac{1}{1+e^{-(W*X+b)}}$$

[0226] 由上述公式可得

[0227] $$e^{-(W*X+b)} = \frac{p+n}{p} - 1 = \frac{n}{p}$$

[0228] 继而两边同时取对数

[0229] $$-(W*X+b) = ln\frac{n}{p}$$

[0230] 对于采样后的模型来说,可以得到以下公式:

[0231] $$pCTR = \frac{p}{p+r*n}$$

[0232] $$pCTR = \frac{1}{1+e^{-(PW*X+pb)}}$$

[0233] 由上述公式可得

[0234] $$e^{-(PW*X+pb)} = \frac{p+r*n}{p} - 1 = \frac{r*n}{p}$$

[0235] 两边同时取对数

[0236] $$-(PW*X+pb) = ln\frac{r*n}{p}$$

[0237] 则对上述两个推导公式两边做差可得

[0238] (W*X+b)-(PW*X+pb)=ln r

[0239] 即W*X+b=PW*X+(pb+ln r)

[0240] 即整个采样过程只对最后训练结果的权重的截距有影响。因此只需要对采样后的数据集训练后对逻辑回归模型的常数项进行修正即可。即对于采样后的训练出来的模型来说,只需要令b=pb+ln r

[0241] 即可达到对模型修正的效果。

[0242] 模型训练后需要验证模型的效果,记录评估值,这样可以对比评估值和线上效果是否正相关。模型的效果评估一般来说可以使用AUC或者Logloss。

[0243] 常见做法是用最大似然估计来推导逻辑回归的损失函数,首先根据逻辑回归的分类,得到

[0244] P(y=1|x)=h(x)

[0245] P(y=0|x)=1-h(x)

[0246] 即h(x)表示结果为1的概率,可以合并上述表达式

[0247] $P(y|x)=h^{y}(x)*(1-h(x))^{1-y}$

[0248] 其中y取值为0或者1,目标是最大化似然概率

[0249] $$max\prod_{i=1}^{n}P(y|x)$$

[0250] 将P(y|x)带入最大化似然概率公式后,可以对其取对数,取对数不改变其单调性,则目标函数为

[0251] $$Logloss=\frac{1}{N}\sum_{i=1}^{n}\left(y_i*log\,h(x)+\ (1-y_i)*log\left(1-h(x)\right)\right)$$

[0252] AUC是指ROC曲线下面的面积。ROC曲线即受试者工作特征曲线(receiver operating characteristic curve)。

[0253] 在点击率预测机器学习中,首先定义如下几个概念:

[0254] • TP(真正类),实例为正类,且被预测为正类。

[0255] • FN(假负类),实例为正类,且被预测为负类。

[0256] • FP(假正类),实例为负类,且被预测为正类。

[0257] • TN(真负类),实例为负类,且被预测为负类。

[0258] • TPR(真正类率)=TP/(TP+FN),即预测的正类中实际正实例占所有正实例的比例。

[0259] • FPR(负正类率)=FP/(FP+TN),即预测的正类中实际负实例占所有负实例的比例。

[0260] 随着真正类率的提高,即想要预测出更多的正实例,负正类率也会提高,即会导致更多的负类被预测为正类。

[0261] ROC曲线的横轴是负正类率,纵轴是真正类率,一般来说,一个表现好的模型,AUC面积是大于0.5的,如果小于0.5只需要将预测结果反转过来即可。

[0262] Logloss是逻辑回归损失函数,也叫对数似然损失,设测试集条数为N,每条数据的预测值p_i,真实值为y_i,其公式为

[0263] $$Logloss=-\frac{1}{N}\sum(y_i*log\,p_i+(1-y_i)*log(1-p_i))$$

[0264] 本实施例最终采用的是Facebook提出的评估值NE(Normalized Entropy),其中p为测试集的平均CTR,公式如下:

[0265] $$NE=\frac{Logloss}{-(p*log\,p+(1-p)*log(1-p))}=\frac{\frac{1}{N}\sum(y_i*log\,p_i+(1-y_i)*log(1-p_i))}{p*log\,p+(1-p)*log(1-p)}$$

[0266] 该公式相较于Logloss多了一个验证集的点击率对数损失,这样可以对不同数据集进行对比,这也是在对模型进行迭代更新时相较于AUC和Logloss都更好的一种评估方式,因为该方式考虑了测试集的CTR。AUC和NE的评估方式各有优点。AUC偏向于分类效果,NE偏向于点击率分布。结合两者来评估模型的好坏,如NE为主,AUC为辅,可以减少因数据集划分误差导致的过拟合。

[0267] 线下实验调参时使用Logloss即可,因为对于同一个数据集,NE公式的分母部分是一个常数,不影响评估效果的好坏。但是在点击率预测系统的迭代更新过程中,数据集一直在变化,因此使用NE来判断参数调参以及特征配置是否依旧正常工作比使用Logloss更合

理。

[0268] 除此之外本实施例还使用了Precision-Recall曲线面积来研究本实施例的系统对数据集预测的点击率的效果的提升。Precision-Recall曲线面积简写做AUCPR,其计算公式如下：

[0269] $$AUCPR = \sum_{i=0}^{n-1} (R_{i+1} - R_i) * P_{i+1}$$

[0270] AUCPR可以看做是加权了的准确率,即描述的是模型对数据集点击事件预测的准确率的期望。

[0271] AUCPR在描述正负样本不均衡的数据集的模型效果上,比ROC曲线面积更有效。因为ROC曲线面积对正负样本分布不敏感,无论正负样本分布如何,ROC 曲线面积始终保持在0.5以上。然而广告投放数据中正样本占比仅有百分之一,此时用ROC曲线面积来作为衡量广告投放数据的模型效果指标不是一个很好的选择。而使用AUCPR就可以很明显的观察到模型效果的微弱变化。

[0272] AUCPR对于模型在测试集上的点击率预测准确率的变化很敏感,对于广告预测准确率即使是很微弱的提升,使用AUCPR也可以很明显的观察到,尤其在样本不均衡的数据集。

[0273] 需要说明的是：

[0274] 在此提供的算法和显示不与任何特定计算机、虚拟装置或者其它设备固有相关。各种通用装置也可以与基于在此的示教一起使用。根据上面的描述,构造这类装置所要求的结构是显而易见的。此外,本发明也不针对任何特定编程语言。应当明白,可以利用各种编程语言实现在此描述的本发明的内容,并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0275] 在此处所提供的说明书中,说明了大量具体细节。然而,能够理解,本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中,并未详细示出公知的方法、结构和技术,以便不模糊对本说明书的理解。

[0276] 类似地,应当理解,为了精简本公开并帮助理解各个发明方面中的一个或多个,在上面对本发明的示例性实施例的描述中,本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而,并不应将该公开的方法解释成反映如下意图:即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说,如下面的权利要求书所反映的那样,发明方面在于少于前面公开的单个实施例的所有特征。因此,遵循具体实施方式的权利要求书由此明确地并入该具体实施方式,其中每个权利要求本身都作为本发明的单独实施例。

[0277] 本领域那些技术人员可以理解,可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件,以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外,可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述,本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代

替。

[0278] 此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0279] 本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本发明实施例的虚拟机的创建装置中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序（例如，计算机程序和计算机程序产品）。这样的实现本发明的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

[0280] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。

[0281] 应该理解的是，虽然附图的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本实施例中有明确的说明，这些步骤的执行并没有严格的顺序限制，其可以以其他的顺序执行。而且，附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，其执行顺序也不必然是依次进行，而是可以与其他步骤或者其他步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0282] 以上所述实施例仅表达了本发明的实施方式，其描述较为具体和详细，但并不能因此而理解为对本发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进，这些都属于本发明的保护范围。因此，本发明专利的保护范围应以所附权利要求为准。

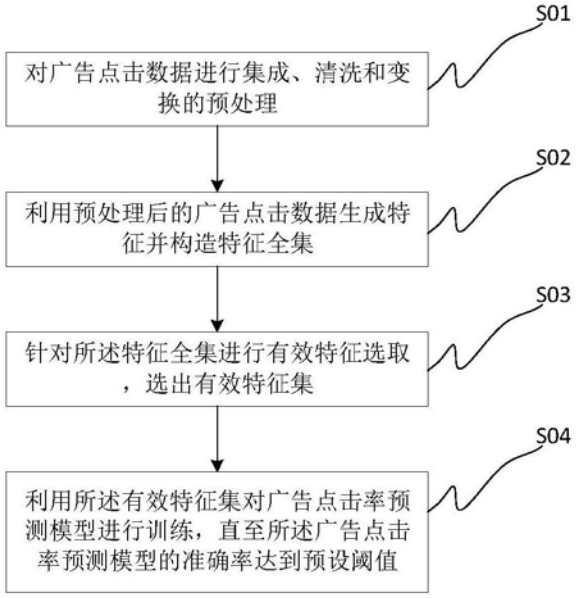

S01
对广告点击数据进行集成、清洗和变换的预处理
S02
利用预处理后的广告点击数据生成特征并构造特征全集
S03
针对所述特征全集进行有效特征选取，选出有效特征集
S04
利用所述有效特征集对广告点击率预测模型进行训练，直至所述广告点击率预测模型的准确率达到预设阈值

图1

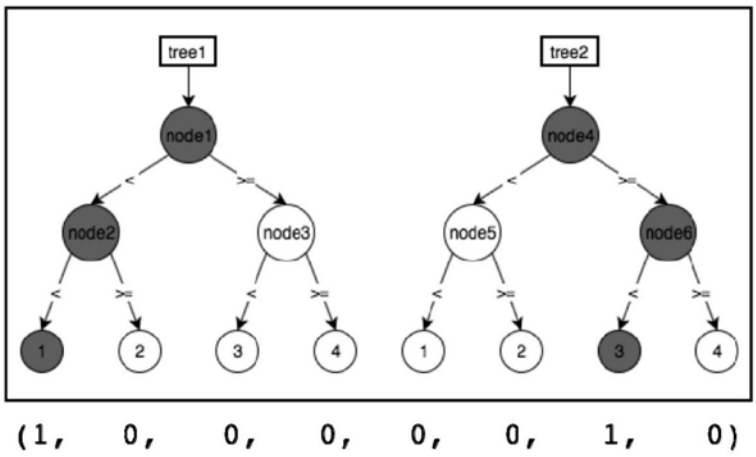

tree1
node1
node2
node3
1
2
3
4
tree2
node4
node5
node6
1
2
3
4
(1, 0, 0, 0, 0, 0, 1, 0)

图2

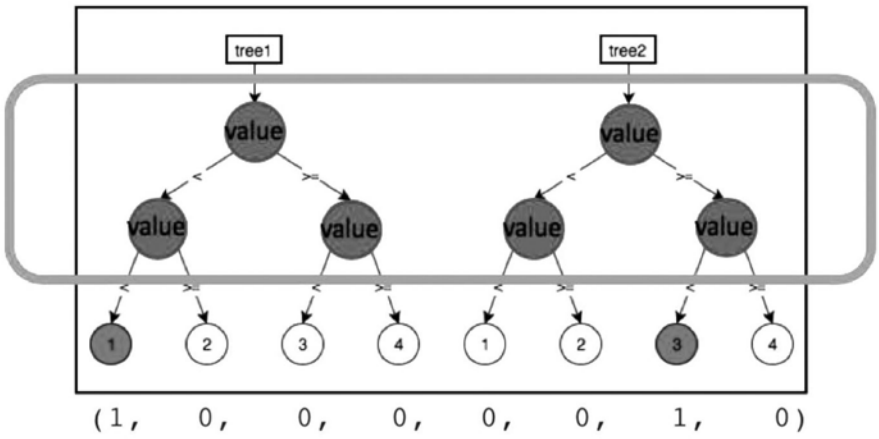

tree1
tree2
value
value
value
value
value
value
1
2
3
4
1
2
3
4
(1, 0, 0, 0, 0, 0, 1, 0)

图3

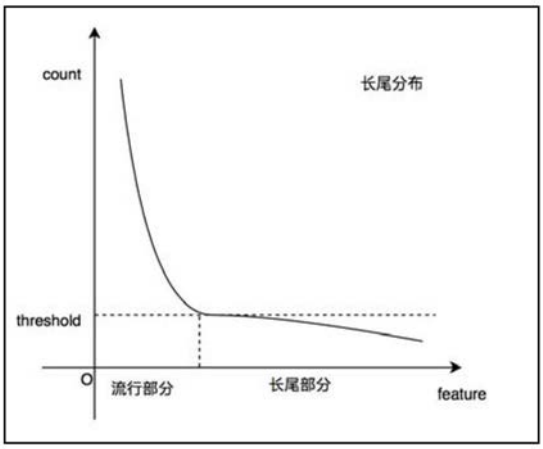

count
长尾分布
threshold
O
流行部分
长尾部分
feature

图4

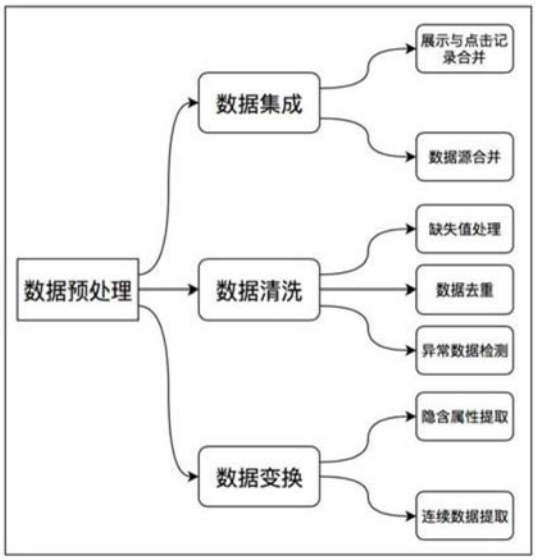

数据预处理
数据集成
数据清洗
数据变换
展示与点击记录合并
数据源合并
缺失值处理
数据去重
异常数据检测
隐含属性提取
连续数据提取

图5

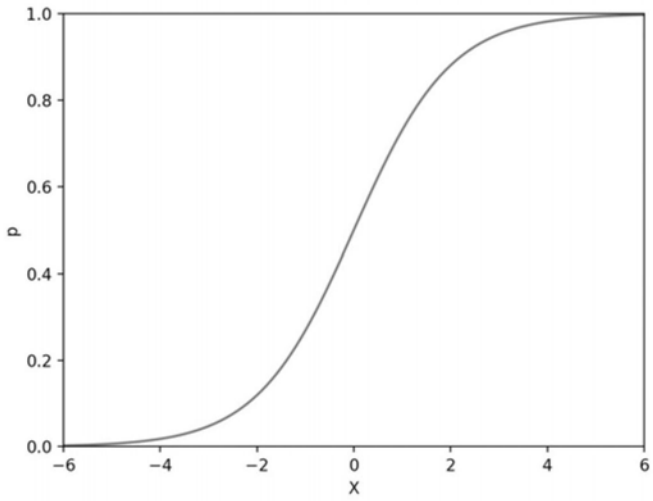

1.0
0.8
0.6
0.4
0.2
0.0
p
−6
−4
−2
0
2
4
6
X

图6