互联网广告点击率预估模型中特征提取方法的研究与实现*

田嫦丽，张 珣†，潘 博，杨 超，许彦茹
（北京工商大学 计算机与信息工程学院，北京 100048）

摘 要：互联网广告是一个具有上千亿元规模的市场，广告的点击率（CTR）是互联网广告投放效果的重要指标。在广告点击率预估模型中，特征提取是关键因素，特征的好坏直接影响到最终模型的效果。针对如何提高广告点击率预估效率问题，在 Hadoop 大数据平台环境中，提出了基于梯度提升决策树（gradient boost decision tree，GBDT）模型的多维特征提取方法。该方法利用原始数据构建多维基础特征库，并将基础特征库中除 ID 类特征以外的其余特征输入 GBDT 模型进行特征刷选，得到高层特征，进一步进行分类。该方法的使用不仅减少了特征提取的人工成本和时间成本，也在很大程度上提升了模型的精度。

关键词：CTR 预估；特征提取；互联网广告；Hadoop 大数据平台；GBDT
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2017)02-0334-05
doi:10.3969/j.issn.1001-3695.2017.02.003

Research and implementation of feature extraction methods on Internet CTR prediction model

Tian Changli, Zhang Xun†, Pan Bo, Yang Chao, Xu Yanru
(*School of Computer & Information Engineering, Beijing Technology & Business University, Beijing* 100048, *China*)

Abstract: Internet advertising is a hundreds of billions of dollars of market. CTR(click-through-rate) is an important indicator of the effectiveness of Internet advertising. In the CTR prediction model, features are used to be a key factor to the success or failure of many machine learning projects and the characteristics of the feature will directly affect the final model. In order to make the Internet advertisement CTR prediction model can be more accurate, this paper put forward a GBDT-based multi-dimensional feature extraction method which ran on the Hadoop big data platform. This method used raw data to build a multi-dimensional feature library and put all the basic features into GBDT model for feature selection except for ID features, in order to get high level features for further classification. This method not only reduces labor costs and time costs in feature extraction stage, but largely enhances the accuracy of the CTR prediction model.

Key words: CTR prediction; feature extraction; Internet advertising; Hadoop big data platform; GBDT

0 引言

互联网广告点击率（click through rate，CTR）是指在给定网页和用户的情况下，估计所投放的广告被点击次数占展示总次数的比例[1]。

随着新一代信息技术的飞速发展，大数据平台技术已成为技术发展的重要支撑之一。近年来，互联网、物联网、云计算、三网融合等 IT 与通信技术迅猛发展，数据的快速增长成为了许多行业共同面对的严峻挑战和宝贵机遇。社会各行业当前对基于数据的应用愈加重视，大数据的兴起引发了各行业研究大数据、应用大数据的热潮。互联网广告点击率对于搜索引擎服务供应商和广告商都是一个重要的量化指标，互联网广告点击率预估是计算广告领域的关键问题之一。大数据平台下互联网广告点击率预估的实现具有很强的理论研究价值和实际应用价值。而广告由过去“粗放式”投放正在向“精准化”投放转变，以数据驱动的广告精准投放已成为在线推广的主流趋势。在广告需求方平台（demand site platform，DSP）的程序化购买和搜索广告投放的过程中，都需要评估用户对广告的偏好程度，而衡量这一偏好程度的重要指标就是广告的点击率。互联网广告点击率预估对广告的后续投放具有非常重要的指导意义。图 1 中的灰色模块描述了一个广告点击率预估系统的工作流程。

预估模型的使用大大方便了点击率的计算，建立模型的过程中数据的预处理与特征提取为很重要的一环。本文的主要目的与意义在于，针对已有的广告数据集，重点研究特征提取的方法以使广告点击率预估具有更高的效率和更好的模型精度。

收稿日期：2016-01-29；**修回日期**：2016-03-08 **基金项目**：北京市自然科学基金重点项目 B 类（KZ201410011014）；2015 年研究生科研能力提升计划资助项目；北京市自然科学基金青年项目（9164025）；国家教育部人文社会科学研究青年基金资助项目（15YJCZH224）

作者简介：田嫦丽（1990-），女（土家族），湖南人，硕士，主要研究方向为数据挖掘、机器学习；张珣（1986-），男（通信作者），讲师，主要研究方向为 GIS 软件技术、智能信息处理、商业地理分析（zhangxun@btbu.edu.cn）；潘博（1994-），男，硕士研究生，主要研究方向为智能信息处理、大数据挖掘；杨超（1993-），女，主要研究方向为数据挖掘；许彦茹（1992-），女，主要研究方向为数据挖掘.

1　问题提出

1.1　特征工程的主要内容

在机器学习中从原始信息中生成和选择特征被称为特征工程（feature engineering）或者特征抽取（feature extraction）。人对事物进行分类主要依据事物之间共同的特点和差别，同样，分类器要作出正确的分类也依赖能对事物作出联系和区分的描述信息，这些信息就是依靠事物的特征。特征生成就是从各种角度和侧面来刻画事物。华盛顿大学教授 Domingos 在文献[2]中讲到，使用什么特征是很多机器学习项目成败的关键因素，特征工程也是机器学习项目中最花费时间的部分。

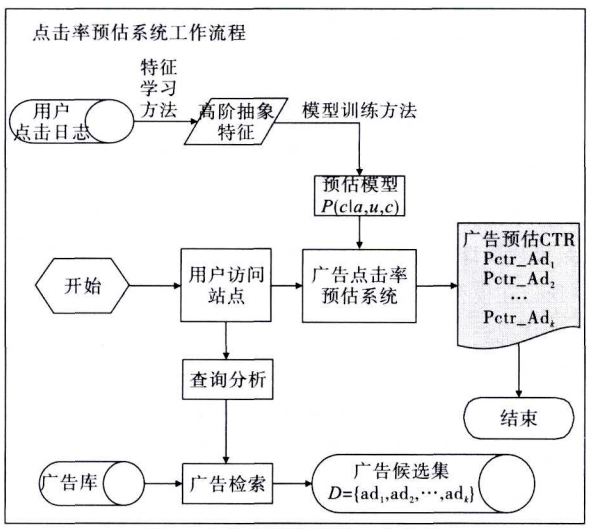

图1　点击率预估系统工作流程

特征工程的主要内容如图 2 所示。

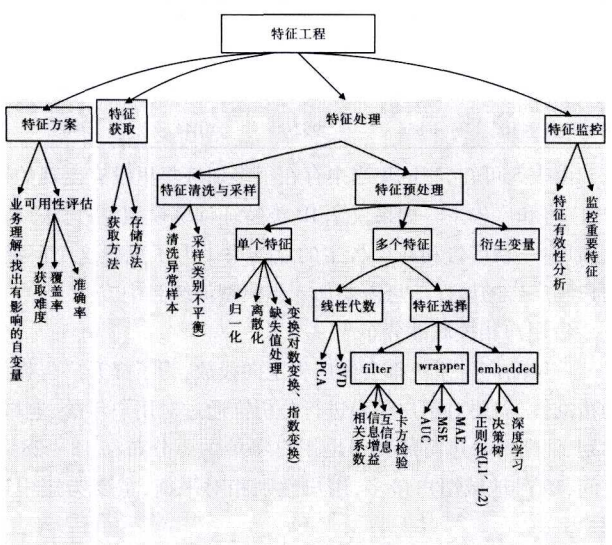

图2　特征工程主要内容

特征工程主要包括特征方案的制定、特征获取、特征处理和特征监控。

特征方案的制定需要根据对业务的理解，尽可能找出对因变量有影响的自变量，并评估获取这些自变量的难度，以及其覆盖率和准确率。特征获取涉及到特征获取的方法和特征存储方法。特征监控需要对特征进行有效性分析，并监控重要特征的变化情况，防止引入质量差的特征影响模型效果，必要时可调整特征的权重。

特征处理在特征工程中占有重要地位，其主要包括对特征进行清洗采样和预处理。清洗的目的是为了清除噪声数据；采样是为了解决类别分布不均衡或数据量过大等问题。对于单个特征，特征预处理涉及到对该特征进行归一化、离散化、缺失值处理或作相应变换，如将时域特征变换为频域特征，将数值作对数或指数变换。衍生变量指的是由现有的变量生成新的更有意义的变量，如本文中由广告曝光量和点击量得到的广告历史 CTR。对于多个特征，特征预处理涉及到降维，而降维的一些最常用的方法是使用线性代数技术，如主成分分析（principal components analysis，PCA）[3] 和奇异值分解（singular value decomposition，SVD）[4]。降维的另一种方法是仅使用特征的一个子集，即需要对特征进行选择。特征选择的理想方法是将所有可能的特征子集作为感兴趣的数据挖掘算法的输入，然后选取产生最好结果的子集。然而，由于涉及 n 个属性的子集多达 $2n$ 个，这种方法在大多数情况下行不通，所以需要其他的策略[5]。有三种标准的特征选择方法：过滤、包装和嵌入。

本文正是在嵌入思想的范畴下提出了基于 GBDT 模型的多维特征提取方法，该方法利用原始数据构建多维特征库，并将特征库中除 ID 类特征以外的其余特征输入 GBDT 模型进行特征刷选，得到高层特征。

1.2　数据预处理与特征提取总流程

数据预处理与特征提取是数据建模中很重要的一环，基本上百分之八十的时间都在进行数据处理和特征提取，特征的好坏将会直接影响到最终模型的效果。根据现有数据集和数据建模目标，本文在对数据和目标充分理解的基础上做了初步的数据处理和特征提取工作。

数据预处理与特征提取流程图如图 3 所示。

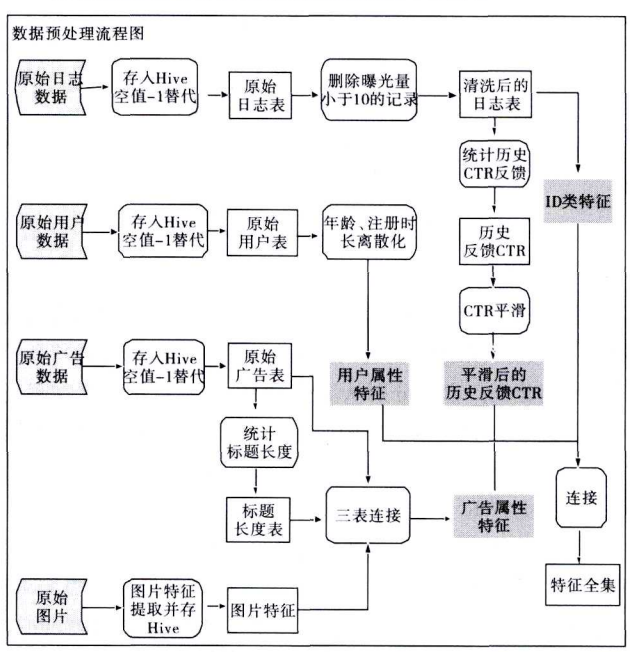

图3　数据预处理与特征提取流程

图 3 中浅色部分代表原始数据文件，深色部分代表最终生成的特征数据，其余部分代表对数据进行的处理和生成的中间数据。该数据预处理涉及到空值的处理、噪声数据的清洗以及多表连接。特征提取涉及到从原始素材图片中提取每张图片的主要颜色、颜色种类和亮度特征，从清洗后的日志中提取广告、

用户和广告的历史反馈 CTR，从原始用户表提取用户性别、年龄、地域等属性特征，从原始广告数据表提取广告标题长度、广告所属类别等特征以及从清洗后的日志中提取 ID 类特征。这些特征组成了基础特征库，具体请见下一章基础特征库的构建。

2 基础特征库的构建

2.1 数据集

本文的数据集来源于腾讯多媒体展示广告数据集，主要包含会话日志、用户信息、广告信息三方面内容。其中会话日志包含表 1 所示字段。

表 1 会话日志字段

序号	字段	说明
1	userID	用户 ID
2	adID	广告 ID
3	advertiserID	广告主 ID
4	creativeID	素材 ID
5	impression	这个字段为 1，表示该广告(adID)被展示给该用户(userID)，此样本为曝光样本
6	click	这个字段为 1，表示该用户(userID)点击了该广告(adID)，此样本为点击样本

用户信息包含表 2 所示字段，可供建模使用。

表 2 用户信息字段

序号	字段	说明
1	userID	用户 ID
2	gender	用户性别，枚举类型
3	age	用户年龄，整数类型
4	registration age	用户注册至今时长，整数类型
5	location	用户所在地区，枚举类型

广告信息包含表 3 所示字段和素材图片，可供建模使用。

表 3 广告信息字段

序号	字段	说明
1	adID	广告 ID
2	title	广告标题，字符串类型
3	creativeID	素材 ID

2.2 数据探索

在特征提取和建模之前需要对数据进行全面的探索以充分地认识数据，根据数据集具体的分布情况进行合理的特征提取与预测建模。本文分别从历史 CTR 分布，点击样本与曝光样本类别分布，用户年龄、性别、地域、注册时长分布，新增新用户、广告、广告主数量等方面进行数据探索。

1)历史 CTR 分布

历史 CTR 的高低可以评估一个广告平台的质量。根据每个广告的历史展示量和点击量可以得到该广告的历史 CTR。通过对训练集进行统计得到如图 4 所示的 CTR 直方图(除去点击率为 0 的广告和点击率异常过高的广告)。

由图 4 可知互联网展示广告同样服从广告点击的长尾特性，这种分布特性使得处在长尾部分的广告预测准确度较差。通过对历史 CTR 进行统计，发现有的广告 CTR 异常过高，最高达到了 0.755 555 57；有的广告展示量太少，点击量为 0。为了提高模型精度和减少计算复杂度，本文提前删除了这些高 CTR、低展示量的广告记录。

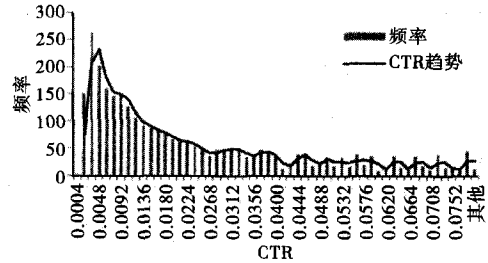

图4 CTR直方图

2)类别分布

通过对训练集和测试集分别统计其点击样本(正例)和曝光样本(负例)得到如表 4 所示的结果。

表 4 训练集、测试集类标签统计

类标签	训练集	测试集
正例(1)	746 056	97 530
负例(0)	19 141 712	2 634 174

由表 4 可知，训练集和测试集当中正负标签严重分布不平衡，负样本比正样本多出两个数量级。如果直接进行分类预测，结果数据肯定有偏，大多数甚至全部测试数据将会标为负类。所以在模型的选择和训练过程中需要考虑到类别分布不平衡的情况。

3)新增用户、广告、广告主

在互联网广告领域时刻都存在着用户、广告和广告主的动态变化，这在很大程度上导致了 CTR 预估的不准确性，成为广告 CTR 预估的一个极大挑战。通过对数据集进行统计得到如表 5 所示的用户、广告、广告主的数量分布。

表 5 用户、广告、广告主的数量分布

字段	总数	训练集	测试集	测试集—训练集
用户 ID	19 904 290	16 297 053	2 630 994	1 807 644
广告 ID	13 799	8 846	10 792	2 422
广告主 ID	1 056	962	1 044	94

由表 5 可知，测试集当中存在很多训练集里没有出现的用户、广告和广告主。如果仅拿 ID 类特征训练模型，将无法预测出新用户、新广告和新广告主的点击率。为了克服这个问题，用户和广告的属性信息需要加入到模型训练当中。

4)用户相关属性分布

一般的广告系统都会做广告定向投放，为了提升广告投放的精准性，需要对用户受众进行分析。通过对用户年龄、地域、性别、注册时长进行统计得出用户年龄主要分布在 18 ~ 28 岁之间，19 号地域用户最多，用户性别相对平衡，主要为短中期用户。

2.3 特征库设计

根据原始数据集的分布特点，本文中的特征主要分为如下几类：

a)ID 类特征。

ID 类特征在 CTR 预估中是非常重要的特征，本文所涉及的 ID 有用户 ID、广告 ID、广告主 ID、素材 ID。这些 ID 实质上是对用户、广告、广告主、素材的一种映射，所以其本身包含了一定的信息，在已有用户和广告上具有显著的预测能力。本文把 ID 类特征二元化，将其进行独热编码。例如本文有 19 904 290 位用

户，userID 这个特征将被扩展为 19 904 290 维的特征向量，只有当某个用户 ID 出现的时候该维上的值为 1，否则为 0，即每一次只有一个维度上的值为 1，其余均为 0。这样处理的原因是，ID 在数值上是递增的，即有序，而实质上 ID 类特征是标称类特征，特征值之间不存在顺序关系，如果将 ID 类特征直接输入到分类器，算法会将输入的数字当成具体的值进行处理。采用独热编码使特征值不再有序并且特征变得非常稀疏，但是 ID 类特征值一般都很大，这会产生上亿维的稀疏特征向量。为了降低特征维度本文删除了曝光次数较少的广告记录。

另外，由表 5 可知，测试集当中存在很多训练集里没有出现的用户、广告和广告主。如果仅拿 ID 类特征训练模型，将无法预测出新用户、新广告和新广告主的点击率。为了克服这个问题，用户和广告的属性信息需要加入到模型训练当中，当 ID 类特征没有预测能力的时候，用户和广告的属性信息将会发挥主要作用。

b）用户特征。

本文包含的用户特征有用户年龄、性别、地域和注册时长。由于年龄和注册时长为连续型特征，为了统一独热编码，需要将其离散化。根据数据探索部分的年龄和注册时长分布图将年龄分为少年、青年、中年、老年四个阶段，注册时长分为短期、中期、长期三个阶段。

c）广告特征。

本文包含的广告特征主要有广告类别、标题长度、素材图片特征。

素材图片特征在很大程度上影响用户的点击，从直观上来说那些颜色亮丽、质感好的图片会更加符合用户的喜好。根据素材图片的特点，本文提取了图片的主要颜色、颜色种类和平均亮度作为素材图片特征。

d）历史反馈特征。

历史反馈特征的意思是当前正在投放的广告，之前已经投放了一部分，已投放部分的点击率基本可以认为是这个广告的点击率，也可以认为是这个广告的质量的一个体现。同样，用户和广告主也存在历史反馈 CTR，作为用户和广告主的一个质量体现。本文的历史反馈特征包括广告、广告主历史 CTR。

但是由于大量的广告未产生点击实例，将导致历史反馈 CTR 普遍为 0。有的广告曝光次数太少，使得 CTR 偏高，不能反映真实的 CTR，即存在数据有偏的现象。为了克服历史反馈 CTR 的数据有偏性，本文采用贝叶斯平滑[6]对 CTR 进行修正处理。

2.4　高层特征提取

GBDT 在 1999 年由 Friedman[7]提出，是一种常用的非线性模型。它基于集成学习中的 Boosting 思想[8]，每次迭代都在减少残差的梯度方向新建立一棵决策树，迭代多少次就会生成多少棵决策树。GBDT 的思想使其具有天然优势，可以发现多种有区分性的特征，决策树的路径可以作为预估模型输入特征使用，省去了人工寻找特征的步骤。图 5 为使用 GBDT 前后的特征选择示意图。融合前人工寻找有区分性的特征和特征组合，融合后直接通过黑盒子（GBDT 模型）进行特征的自动选择和自动抽象，形成高级特征。

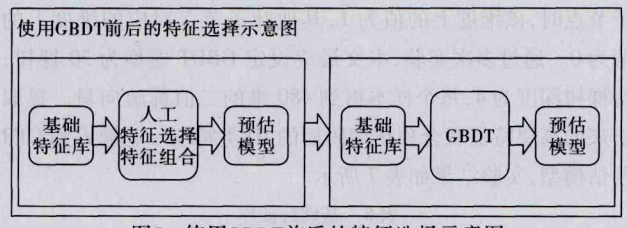

图5　使用GBDT前后的特征选择示意图

GBDT 算法的特点正好可以用来发掘有区分度的特征，减少特征工程中人力成本，所以本文采用 GBDT 算法进行特征的自动选择以及特征的非线性转换。通过 Hadoop 大数据平台分布式训练 GBDT 算法模型，加快其收敛速度，提高算法效率。本文在实验实施部分将通过对比实验验证 GBDT 所提取出来的高层特征的有效性。

3　实验实施

3.1　实验集群环境设置

本文的实验平台为基于 Hadoop 的大数据处理平台，集群规模为 1 个 master 节点，3 个 slave 节点。集群配置及网络拓扑如图 6 所示。

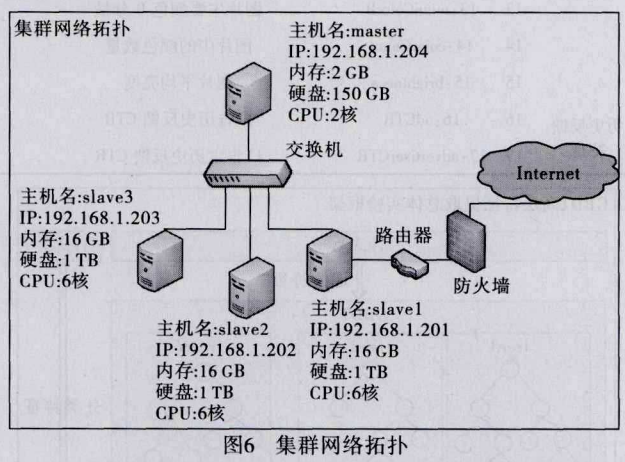

图6　集群网络拓扑

3.2　原始特征提取

根据基础特征库的设计，本文利用 Hive[9]和 MapReduce[10]从 2.1 节介绍的原始数据集中提取到如表 6 所示的特征。

ID 类特征在广告 CTR 预估模型中是很重要的一项特征，其进行独热编码之后会形成高维稀疏数据，可加快模型的训练速度。为了将整个模型训练由密集矩阵计算转换到稀疏矩阵计算，连续类特征需要进行稀疏转换。GBDT 模型不仅能进行特征的自动选择和非线性转换，也可以通过高层特征提取，将连续特征转换成二值稀疏特征。

3.3　高层特征提取实验

为了验证本文提出的基于 GBDT 模型的多维特征提取方法是否有效，本文利用因子分解机（factorization machine，FM）模型[11]和逻辑回归（logistic regression，LR）模型[12]进行了多组实验，实验总体框架如图 7 所示。

从原始数据中采样之后形成最终的训练集，将该训练集按连续特征和分类特征分开存储。连续特征输入 GBDT 模型，得到 $N \times M$ 维二值稀疏向量。其中 N 为 GBDT 模型树的数量，M 为每棵树的叶子节点个数。当一个样本落到某棵树的某个叶

子节点时，该维度上的值为 1，其他叶子节点对应的维度上的值为 0。通过多次实验，本文最终设定 GBDT 模型为 30 棵树，每棵树深度为 4，每个样本得到 480 维的二值稀疏向量。提取出来的高层特征联合独热编码后的 ID 类特征一起输入到 CTR 预估模型，实验结果如表 7 所示。

表 6 基础特征库

特征大类	特征编号	特征内容	说明
原始 ID 类特征	1	1:userID	清洗后的用户数为 16 275 742
	2	2:adID	清洗后的广告数为 6 365
	3	3:advertiserID	清洗后的广告主数为 857
	4	4:creativeID	清洗后的素材数为 5 362
用户属性特征	5	5:gender	1:男 2:女
	6	6:age	1:少年 2:青年 3:中年 4:老年
	7	7:registration age	1:短期 2:中期 3:长期
	8	8:location	取值为 1 ~ 34，代表不同的省份
广告属性特征	9	9:adCategrory	取值为 1 ~ 11，代表不同的类别
	10	10:titleLength	广告标题长度
	11	11:mianColorR	图片主要颜色 R 分量
	12	12:mianColorG	图片主要颜色 G 分量
	13	13:mianColorB	图片主要颜色 B 分量
	14	14:colorCount	图片中的颜色数量
	15	15:brightness	图片平均亮度
历史反馈特征	16	16:adCTR	广告历史反馈 CTR
	17	17:advertiserCTR	广告主历史反馈 CTR

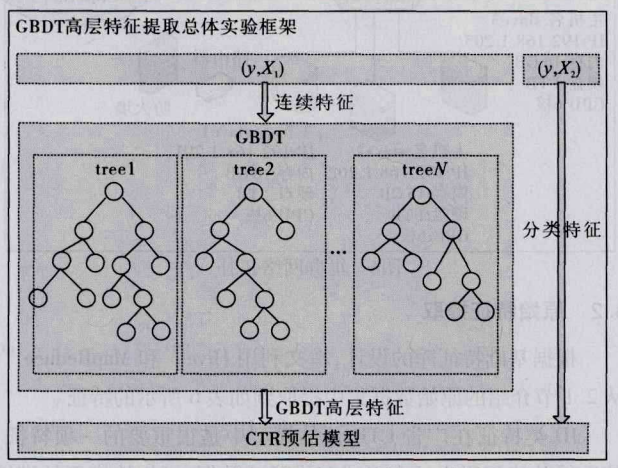

图7 GBDT高层特征提取总体实验框架

表 7 实验结果记录

模型	特征	AUC
FM	ID 类特征	0.752 4
FM	ID 类特征 + GBDT 高层抽象特征	0.790 1
LR	ID 类特征	0.691 4
LR	ID 类特征 + GBDT 高层抽象特征	0.720 4

表 7 中，LR（逻辑回归）为 CTR 预估的常用线性模型，其简单、易扩展、易在线更新；FM（因子分解机）基于因子分解，能拟合特征之间的相关性，能很好地处理高维稀疏数据；AUC（area under roc curve）[13] 为 CTR 预估模型的常用评价指标，越接近于 1 代表模型效果越好。

从表 7 可以看出，本文提取的 GBDT 高层抽象特征对 FM 和 LR 模型均有效，不仅能提升模型的效果，且可以减少特征工程的成本和时间。不需要将特征库中的特征进行一一的验证和刷选，也不需要将它们进行各种各样的组合，GBDT 算法可以用来发掘有区分度的特征，且可以对特征进行非线性转换，消除噪声的干扰。

4 结束语

本文首先分别从历史 CTR 分布，分类类别分布，用户年龄、性别、地域、注册时长分布，新增新用户、广告、广告主数量等方面进行了全面的数据探索。在对数据和目标充分理解的基础上做了初步的数据处理和特征提取工作，建立了基础特征库，利用基础特征库里的特征和 GBDT 模型提取了高层抽象特征。通过实验证明了这些特征能够在很大程度上提高模型的精度。

特征提取在互联网广告点击率预估当中是很费时费力的一项工程，且具有重要地位。因为特征的好坏直接影响到预估模型的整体效果，所以本文在前人的基础上，从特征提取的角度提出了基于 GBDT 模型的多维特征提取方法。该方法在减少建模人工成本和时间成本的同时，能提升模型的精度。

参考文献：

[1] 周傲英，周敏奇，宫学庆. 计算广告：以数据为核心的 Web 综合应用[J]. 计算机学报，2011，34(10)：1805-1819.

[2] Domingos P. A few useful things to know about machine learning[J]. Communications of the ACM，2012，55(10)：78-87.

[3] Boutsidis C，Garber D，Karnin Z，*et al*. Online principal components analysis[C]//Proc of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms. New York：ACM Press，2015：887-901.

[4] Liesen J，Mehrmann V. The singular value decomposition[M]//Linear Algebra. [S. l.]：Springer International Publishing，2015：295-302.

[5] Tan Pangning，Steinbach M，Kumar V. Introduction to data mining [M]. Boston：Pearson Addison Wesley，2006.

[6] Yang Hongxia，Ormandi R，Tsao H Y，*et al*. Estimating rates of rare events through a multidimensional dynamic hierarchical Bayesian framework[J]. Applied Stochastic Models in Business and Industry，2016，32(3)：340-353.

[7] Friedman J H. Greedy function approximation：a gradient boosting machine[J]. Annals of Statistics，2000，29(5)：1189-1232.

[8] Zhou Zhihua. Ensemble methods：foundations and algorithms[M]. [S. l.]：CRC Press，2012.

[9] Thusoo A，Sarma J S，Jain N，*et al*. Hive：a warehousing solution over a Map-Reduce framework[J]. Proceedings of the VLDB Endowment，2009，2(2)：1626-1629.

[10] Dean J，Ghemawat S. MapReduce：simplified data processing on large clusters[J]. Communications of the ACM，2008，51(1)：107-113.

[11] Rendle S. Factorization machines[C]//Proc of the 10th IEEE International Conference on Data Mining. 2010：995-1000.

[12] Harrell F. Regression modeling strategies：with applications to linear models，logistic and ordinal regression，and survival analysis[M]. 2nd ed. New York：Springer-Verlag，2015.

[13] Lobo J M，Jiménez-Valverde A，Real R. AUC：a misleading measure of the performance of predictive distribution models[J]. Global Ecology and Biogeography，2008，17(2)：145-151.