一种基于LightGBM机器学习算法的用户年龄及性别预测方法

A Prediction Method of User Age and Sex Based on LightGBM Machine Learning Algorithms

高　洁，张　涛，程新洲，关　键（中国联通网络技术研究院，北京 100048）
Gao Jie，Zhang Tao，Cheng Xinzhou，Guan Jian（China Unicom Network Technology Research Institute，Beijing 100048，China）

摘　要：
随着智能手机的普及，几乎在任何时间、任何地点，移动用户都可以用手机浏览网页、聊天、玩游戏。手机的上网数据可以反映每个用户的属性特征和行为偏好。通过机器学习算法可以精准地预测用户的属性特征（例如性别和年龄）以及在精准营销中常使用的特征标签。提出了一种基于开源的监督学习算法来预测用户性别与年龄的方法。

关键词：
大数据；数据挖掘；机器学习；预测算法
doi：10.12045/j.issn.1007-3043.2019.09.008
中图分类号：TP274
文献标识码：A
文章编号：1007-3043（2019）09-0036-04

Abstract：
With the popularity of smart phones，mobile users can browse the Web page，chat and play games with their mobile phones at almost any time and anywhere. The mobile phone's online data can reflect the attributes and behavior preferences of each user. Machine learning algorithm can accurately predict user attributes（such as gender and age）and feature labels commonly used in precision marketing. An open source supervised learning algorithm is proposed to predict users'gender and age.

Keywords：
Big data；Data mining；Machine learning；Prediction algorithm

引用格式：高洁，张涛，程新洲，等. 一种基于LightGBM机器学习算法的用户年龄及性别预测方法[J]. 邮电设计技术，2019(9)：36-39.

0 引言

随着移动网络和智能手机的迅速发展，几乎每个人都离不开手机。咨询公司的报告显示，在近5年的时间里，智能手机在移动市场的渗透率已经从2014年的50%上升到2019年的80%，到2019年底，预计将达到85%。在日常生活中，人们几乎每天都在使用手机浏览网页、聊天和网上购物，手机的上网数据可以直观地反映用户的属性特征和行为偏好。因此，运营商可以通过智能网管平台采集移动用户终端APP安装列表、APP使用记录、终端类型和终端价格等数据，再结合GiHtub上开源的机器学习算法，便可以开展移动用户的精准画像工作，例如预测用户的年龄、性别等信息，这些在精准营销中是非常重要的客户标签属性。它不仅可以帮助互联网公司了解用户的行为特征，迭代开发产品，还可以帮助企业提高广告投放的精准度，从而节约广告投资成本。

收稿日期：2019-06-13

1 相关机器学习算法

在机器学习算法领域，监督学习算法中最常用的2类算法为回归（Regression）算法和分类（Classification）算法。回归和分类的算法区别在于输出变量的

类型，定量输出或者连续变量预测称为“回归”；定性输出或者离散变量预测称为“分类”。而对移动用户年龄和性别的预测过程是一个典型的分类问题，因此，可以利用分类算法对移动用户的年龄和性别进行精准预测。

目前比较流行的分类算法包括经典的决策树、集成学习Boosting算法中的梯度提升树（GBDT——Gradient Boosting Decision Tree）算法和极端梯度提升（XGBOOST——eXtreme Gradient Boosting）算法。其中，GBDT算法通过多轮迭代，每轮迭代产生一个弱分类器，后续每个分类器在上一轮分类器的残差基础上进行训练，如图1所示。

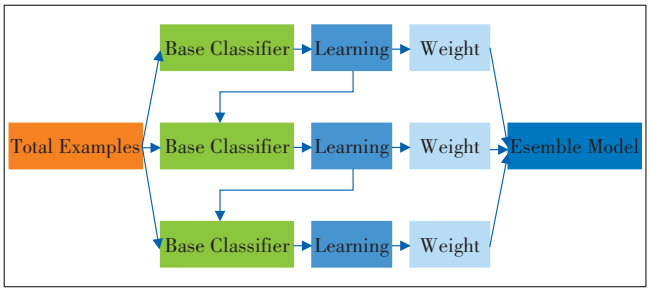

图1　GBDT的模型训练过程

XGBOOST算法是GBDT的改进，它是大规模并行boosted tree的工具，是目前最快最好的开源boosted tree工具包。在XGBOOST之后，微软公司又提出了一种LightGBM算法来增强GBDT的性能。LightGBM是一个实现GBDT算法的框架，支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率以及支持分布式海量数据处理的能力。首先它抛弃了GBDT算法使用的按层生长的决策树生长策略，而使用了带有深度限制的按叶子生长算法，以加速训练过程，减少内存使用。因此，基于以上集成机器学习算法优劣势比较，提出了一种基于LightGBM机器学习算法的预测用户年龄和性别的方法。

2　一种用户年龄及性别的预测方法

为了提高算法模型预测的准确性，笔者采用了LightGBM机器学习算法和交叉验证的训练方式。该模型算法的整体流程框架如图2所示，整个模型训练过程可分为5个步骤：数据收集、特征工程、模型训练、交叉验证和精度评价。

2.1　数据采集

运营商通过智能网管平台可实时采集到用户的

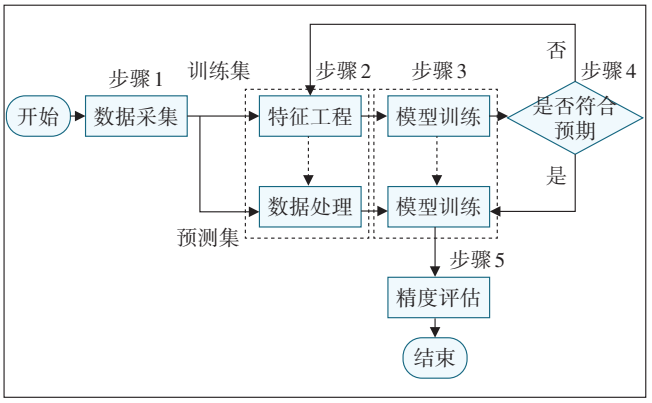

图2　用户年龄及性别预测算法的整体流程框架

基本属性，如用户标识、终端品牌、终端子品牌、终端价格、性别、年龄以及用户的业务信息（APP的安装列表、APP所属类型、APP所属子类型、APP的访问时间和结束时间记录）。当然，用户的年龄和性别信息只在训练集中存在，在预测集中是目标预测结果。本次研究收集了近73 000名Android用户的基本属性和业务信息，其中有5万名用户提供了性别和年龄信息，其中男性用户32 324人（占64.6%），女性用户17 676人（占35.4%）。

表1和表2展示了智能网管平台采集到的用户基本属性和业务信息，在后续的流程中，将用这些数据去训练预测模型。

表1　用户基本属性

用户ID	终端品牌	子品牌	终端价格	性别	年龄
f69cdc	Samsung	GT-I9507V	2 538	1	3

表2　用户业务信息

用户ID	APP名称	APP类型	APP子类型	开始时间	结束时间
10125	支付宝	金融	支付	2018-06-01 08:00:00	2018-06-01 08:00:00

2.2　特征工程

特征工程是机器学习研究课题中最重要的部分。在这一过程中需要找到最能反映分类本质的特征来完成原始数据的分类工作。总之，特征工程的研究是否精细，会直接影响到模型的预测性能。因此，首先对有5万用户的训练数据集进行大数据挖掘分析。

将男性和女性用户分别表示为1和2，同时，将每个用户群体的年龄按10年为一个段进行划分。例如，在25~30组的使用者会在年龄特征中以3表示，如表3所示。

通过对用户基本属性以及业务信息进行分析，发

表3　年龄、性别的分组映射关系表

性别	标识	年龄	标识	年龄	标识	年龄	标识
男性	1	0~10	0	30~35	4	50~60	8
		10~20	1	35~40	5	60~70	9
女性	2	20~25	2	40~45	6	70~80	10
		25~30	3	45~50	7		

现了不同性别和不同年龄段对终端品牌的倾向分布规律，分别如图3和图4所示。从图3可以看出华为、小米、三星占领了安卓智能手机的主要市场份额，其中，使用小米手机的男女性别占比分别为20.5%和18.7%。从图4可以看出对于1~5年龄组的用户来说，小米、三星品牌呈现“齐头并进”的趋势，而8~10年龄组又呈现出“反转”的走势，这主要是受样本数量的影响。

接下来对用户APP安装列表进行了分析，不同性别安装APP类型的分布情况如图5所示。对于图5中的男性用户来说，最受欢迎的3类应用是社交、购物和应用管理。而对于图5的女性用户而言，社交、购物和健康类APP深受关注。当然，这种市场规律的分布是由不同的生活习惯、思维方式所决定的。

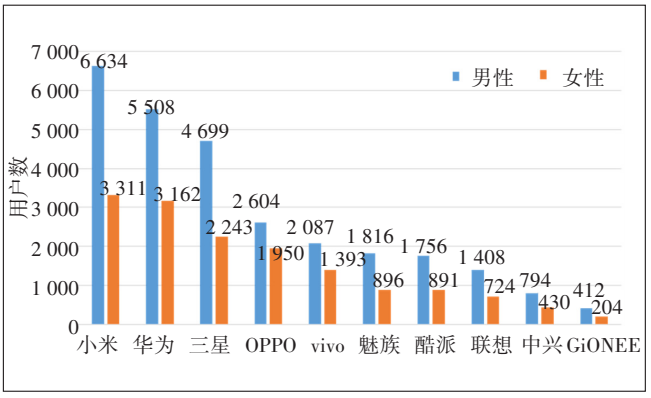

图3　不同性别对终端品牌的倾向性分析

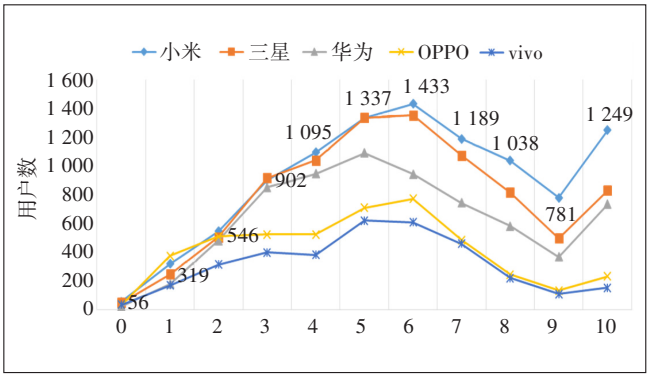

图4　不同年龄对终端品牌的倾向性分析

2.3 模型训练

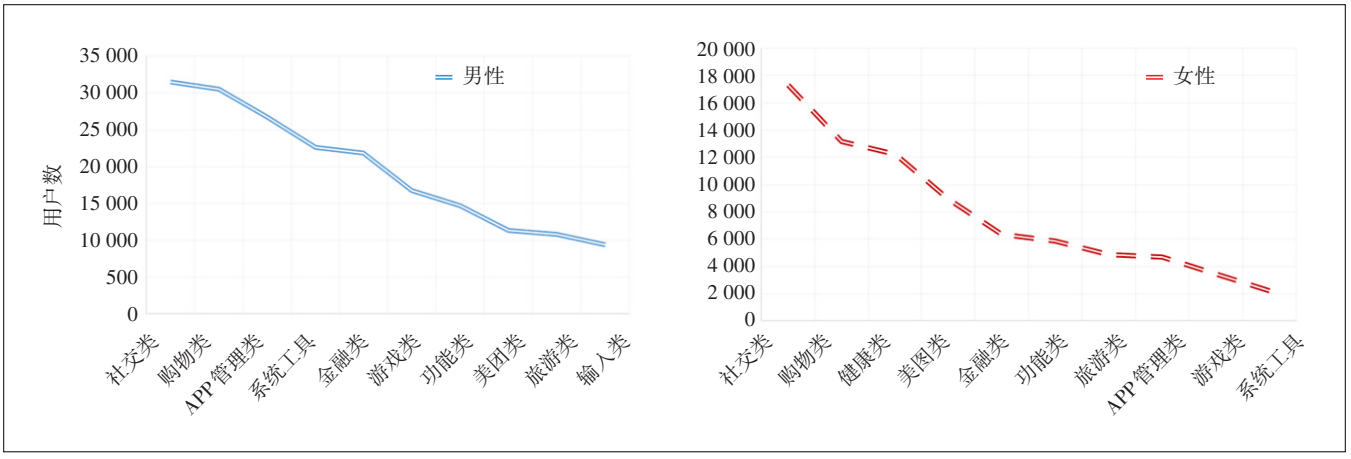

图5　不同性别对APP类型的倾向性分析

本文研究了包括GBDT、XGBOOST和LightGBM 3种最常用的机器学习算法的区别和特点，通过比较预测精度和复杂度，最终选择了LightGBM作为整个模型的核心算法，并通过Python实现数据处理和模型训练。其中，Python中的LightGBM参数设置如表4所示。

表4　LightGBM的参数设置

Boosting类型	最大深度	评估函数	最大分类	结果类型	随机状态
GBDT	3	Multi_logloss	22	多分类	333

2.4 交叉验证

在此步骤中，采用交叉验证的方式来优化LightGBM算法迭代分类过程中特征间的不同权重情况。将训练数据集分为5份，其中的1份用于模型训练，其余4份数据用于验证模型精度。

2.5 精度评估

为了明确模型训练后的精准性，采用损失函数评估其分类精度。损失函数如式(1)所示。

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{22} y_{ij}\ln(p_{ij}) \quad (1)$$

其中，N是预测集中的用户数，j是按性别和年龄划分的不同用户组数，y_{ij}是一个布尔值，表示用户是否属于这一个年龄—性别组，p_{ij}是由模型计算出的该用户属于这一年龄—性别组的概率。在模型训练过程中，通过损失函数计算出该次训练的分类精度。

3 模型对比分析

本文对实际用户的基本属性和业务信息等数据进行了分析挖掘，提出了一套基于机器学习算法的用户年龄及性别的预测方法，并通过Python数据挖掘工具实现整体流程。通过调用Sklearn工具包中的机器学习模型，对比不同继承算法下的模型的精准性，结果如图6所示。

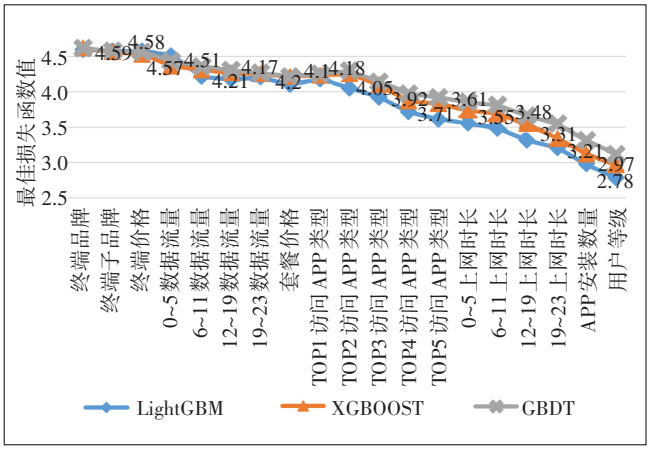

图6　不同集成算法下模型精准性的差异

从图6可以看出，整体模型算法的精准性将随着特征的不断积累而得以提升，而且图5所列举的特征都是对分类有增益的，对结果有干扰的特征已经被排除。通过不断的特征迭代训练，使用LightGBM算法的最佳损失函数可以控制在2.78左右，是整个研究过程中能达到的最佳结果。

4 总结

本文提出了一种基于大数据分析和机器学习算法的用户性别和年龄的预测方法。该算法框架基于集成算法对移动用户的年龄和性别进行预测，最佳结果可以将损失函数控制在2.78左右。同时，如果在后续的研究中引入更多、更丰富的数据，整体模型的精准性还可以进一步提升。该模型算法可以丰富用户的标签信息，不仅可以帮助互联网公司了解用户的行为特征，迭代开发产品，还可以帮助企业提高广告投放的精准度，从而节约广告投资成本。

参考文献：

[1] 董润莎，徐争莉，袁明强，等. 基于机器学习的用户离网预测研究[J]. 邮电设计技术，2018(10)：1-5.
[2] 艾达，罗爱平. 移动通信重入网用户识别算法分析研究[J]. 西安邮电大学学报，2012，17(3)：30-33.
[3] JAKIR K，FENIL A，MITHILA S. Different approaches and methods for targeted advertisements by predicting user's behavioral data and next location[C]// Conference ICISC，2018：1345-1350.
[4] CHEN T，CARLOS G. Xgboost：A scalable tree boosting system[C]// 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2016：785-794.
[5] KE G，MENG Q，FINLEY T. LightGBM：A Highly Efficient Gradient Boosting Decision Tree[C]// 31st Conference on Neural Information Processing System. NIPS，2017：342-353.
[6] BI B，SHOKOUHI M，KOSINKI M，et al. Inferring the demographics of search users：Social data meets search queries[C]// 22nd international conference on World Wide Web，2013：131-140.
[7] AARTHI S，BHARANIDHARAN S，SARAVANAN M，et al. Predicting Customer Demographics in a Mobile Social Network[C]// International Conference on Advances in Social Networks Analysis & Mining. IEEE Computer Society，2011：553-554.
[8] CHEN J，WANG C，HE K，et al. Semantics-Aware Privacy Risk Assessment Using Self-Learning Weight Assignment for Mobile Apps[J]. IEEE Transactions on Dependable and Secure Computing，2018：1-1.
[9] 赵慧，刘颖慧，崔羽飞，等. 机器学习在运营商用户流失预警中的运用[J]. 信息通信技术，2018(1).
[10] 王建仁，李妮，段刚龙. 基于信息融合的电信客户流失预测研究[J]. 计算机工程与应用，2016，52(10)：64-70.
[11] 陈素香. 数据挖掘技术在福建移动经营分析中的应用[J]. 电脑知识与技术，2010，6(35)：9932-9933.
[12] 王琴，张炯. 数据挖掘在移动客户投诉分析中的应用研究[J]. 湖南邮电职业技术学院学报，2018，17(4)：29-31，46.
[13] 张淑云. 数据挖掘在渠道偏好用户识别中的应用——以某市移动网上营业厅为例[D]. 杭州：浙江工商大学，2017.
[14] 冀鸣，朱江，杨志成，等. 数据挖掘在移动通信用户行为分析的应用研究[J]. 信息系统工程，2017(5).

作者简介：
高洁，毕业于北京邮电大学，工程师，硕士，主要从事大数据分析技术及行业创新应用产品的研究工作；张涛，毕业于北京邮电大学，工程师，硕士，主要从事物联网大数据分析及行业创新应用产品的研究工作；程新洲，中国联通网络技术研究院大数据研发中心总监，教授级高级工程师，主要从事大数据研究与应用工作；关键，毕业于北京邮电大学，工程师，硕士，主要从事大数据分析技术及行业创新应用产品的研究工作。