计算机科学
COMPUTER SCIENCE

众包平台用户价值识别与细分:基于改进的 RFM 模型

陈丹红, 彭张林, 万德全, 杨善林

引用本文

陈丹红, 彭张林, 万德全, 杨善林. 众包平台用户价值识别与细分:基于改进的 RFM 模型[J]. 计算机科学, 2022, 49(4): 37-42.

CHEN Dan-hong, PENG Zhang-lin, WAN De-quan, YANG Shan-lin. Identification and Segmentation of User Value in Crowdsourcing Platforms:An Improved RFMModel[J]. Computer Science, 2022, 49(4): 37-42.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

空间众包任务的路径动态调度方法

Dynamic Task Scheduling Method for Space Crowdsourcing

计算机科学, 2022, 49(2): 231-240. https://doi.org/10.11896/jsjkx.210400249

基于树增益朴素贝叶斯网络的服务定价策略

TAN-based Service Pricing Strategy

计算机科学, 2021, 48(6A): 203-. https://doi.org/10.11896/jsjkx.200900024

供需匹配中的非诚信行为预防

Prevention of Dishonest Behavior in Supply-Demand Matching

计算机科学, 2021, 48(4): 303-308. https://doi.org/10.11896/jsjkx.200900090

基于改进狮群进化算法的面向空间众包平台的多工作者多任务路径规划方法

Multi-worker and Multi-task Path Planning Based on Improved Lion Evolutionary Algorithm forSpatial Crowdsourcing Platform

计算机科学, 2021, 48(11A): 30-38. https://doi.org/10.11896/jsjkx.201200085

基于区块链的去中心化众包技术综述

Survey of Crowdsourcing Applications in Blockchain Systems

计算机科学, 2021, 48(11): 12-27. https://doi.org/10.11896/jsjkx.210600152

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800255

众包平台用户价值识别与细分:基于改进的 RFM 模型

陈丹红　彭张林　万德全　杨善林
合肥工业大学管理学院　合肥 230009
过程优化与智能决策教育部重点实验室　合肥 230009
(alicealice_hi@163.com)

摘　要　在众包平台上,不同类型的用户在参与意愿、工作动机、业务能力等方面具有多样性和差异性的特征,在平台上产生的价值也不同。基于用户价值度量对用户进行细分,是更好地洞察用户价值和需求、对用户进行个性化和精细化管理的关键。同时,选择众包用户价值衡量维度也是目前需要解决的问题。因此,该研究首先基于 RFM 模型并结合众包平台及众包用户的特性,将用户信用纳入用户价值模型,提出并构建了众包用户价值衡量模型 RFMC(Recency,Frequency,Monetary,Credit);然后,结合"一品威客"平台获取所需的实验数据,运用 GBDT 算法完成众包用户分类;最后,比较了 Nave Bayes,Multinomial Logistic Regression 与 GBDT 算法的分类效果,并比较了不考虑用户信用的传统模型与 RFMC 模型的分类效果。结果表明,所提模型适用于众包用户且具有较好的实验效果。

关键词:众包;用户细分;RFM 模型;GBDT 算法
中图法分类号　C934

Identification and Segmentation of User Value in Crowdsourcing Platforms:An Improved RFM Model

CHEN Dan-hong,PENG Zhang-lin,WAN De-quan and YANG Shan-lin
School of Management,Hefei University of Technology,Hefei 230009,China
Key Laboratory of Process Optimization and Intelligent Decision,Ministry of Education,Hefei 230009,China

Abstract　On the crowdsourcing platform,different types of users have diversity and differences in participation intention,work motivation,business ability and other aspects,and the value they generated on the platform is also different. The segmentation of users based on user value measurement is the key to better insight into user value and needs for personalized and refined management of users. At the same time,the choice of crowdsourcing user value measurement dimension is also a problem to be solved. Therefore,based on the RFM model,combined with the characteristics of crowdsourcing platform and crowdsourcing users,this paper firstly incorporates user credit into the user value model,proposes and constructes a crowdsourcing user value measurement model—RFMC. Secondly,combined with the required data obtained on the platform of "Yipinweike",using GBDT algorithm to complete the crowdsourcing user classification. Finally,the classification performance of Nave Bayes,Multinomial Logistic Regression and GBDT are compared. Also,the classification performance of RFMC model is compared with that of traditional model without considering user credit. Evaluation indicators show that the proposed model is suitable for crowdsourcing users and has good experimental results.

Keywords　Crowdsourcing,User segmentation,RFM model,GBDT algorithm

1　引言

作为一种基于社会网络形成的工作平台,众包有别于以数据和流程为核心的流程支持信息技术,它是一种以用户为核心的新一代信息沟通支持技术,能够快速突破信息沟通壁垒,降低创新成本[1],有利于个体之间的实时协作、知识共享和共同生产[2]。在众包模式下,用户作为众包任务的接包方,与众包任务的发包方以及互联网众包平台共同构成了一个双边市场机制。随着众包模式的不断发展与广泛应用,一方面,众包用户群体的数量不断增加,众包平台的竞争越发激烈;另一方面,众包平台及参与用户也呈现出一些新的特点,例如:从众包用户的参与行为来看,用户逐渐呈现出不同层次的

到稿日期:2021-08-30　返修日期:2021-12-09
基金项目:教育部人文社会科学研究基金(16YJC630093);国家自然科学基金(72071060,71601066,71901086)
This work was supported by the Humanities and Social Science Research Foundation of Ministry of Education(16YJC630093) and National Natural Science Foundation of China(72071060,71601066,71901086).
通信作者:彭张林(pengzhanglin@hfut.edu.cn.)

参与深度、不同频次的参与强度等特点，体现出用户需求的差异性；从众包平台来看，平台具有服务特性和社交特性，其不仅关注业务的流转，还提供用户交流、用户评价等功能模块，对于不同类型的用户，平台需要满足其差异化的需求，为其提供个性化的服务。用户个体的差异性使得他们在众包平台上创造的价值不同。因此，基于用户价值计算和衡量对用户进行精准细分，以更好地洞察用户需求、了解用户的参与意愿和忠诚度[3]，并根据用户类型的差异在众包任务设计、平台服务策略、奖励激励机制以及平台功能等方面调整和优化管理策略，是支撑众包平台实施个性化、精细化管理，提升用户参与强度，最大化用户价值，增强众包平台创新效率的关键。

用户细分是依据一定的规则将大量的用户划分为不同子类的过程，处于同一子类内的用户一般具有相同或相似的特征。目前，相关研究主要从用户个体的人口统计学信息、用户线上或线下行为数据等方面来识别和提取用户特征。例如，Carbajal[4]从用户的坐标角度开展研究，基于用户轨迹特征实现了用户在购物中心周围活动的自动分类问题；Nilashi 等[5]提取社交网站中旅行者过去的评分和在线评论中的特征来预测用户的选择偏好；Lin 等[6]利用用户在汽车论坛的中心度、回帖频次、帖子的平均长度、粉丝量等特征为企业识别领先用户；Lu 等[7]基于会话分析理论，从会话交互内容和会话交互关系两个维度揭示了虚拟学术社区的用户交互特征，并提出了用户交互策略；Xiao 等[8]介绍了分组数据处理方法（Group Method of Data Handling），提出了一种基于 GMDH 的半监督特征选择算法，并将其应用于客户特征选择。由于用户的特征获取缺少理论模型的支持，基于用户指标体系构建好用户模型后，可开展用户细分研究。目前，许多经典的分类方法及相关衍生方法都被用于用户细分研究。Shi 等[9]基于 Folksonomy 的数字资源系统中用户的共性，提出了利用用户标签对用户进行分类的方法，并对 Flikcr 网站进行了实证研究，证明其可行性；Zhu 等[10]建立了政府数据开放平台用户体验的要素体系，运用 Kano 模型的问卷处理方法、传统分类和 Better-Worse 分类方法对用户体验要素进行归类；Hu 等[11]对用户兴趣数据库信息进行聚类分析，构建了基于用户兴趣分类的书籍自动推荐系统；Li 等[12]针对传统 K-means 聚类算法的局限性，利用量子粒子群优化算法全局搜索能力强、收敛速度快的优势，提出了一种基于改进量子粒子群的 K-means 聚类算法，并将该方法应用于我国鲜食葡萄市场的客户分类中；Shang 等[13]设计了一个基于改进支持向量机（Support Vector Machine，SVM）的互联网用户分类模型，该模型通过模拟互联网用户的浏览轨迹，基于改进支持向量机实现了对互联网用户的分类；Shi 等[14]基于 BP 和 AdaBoost 算法，提出了一种新的物流企业客户分类方法，该方法利用 AdaBoost 算法将传统 BP 网络进行线性组合，产生了强分类器并将其应用于客户分类中。综上，如何提高分类效果是用户细分研究的重点之一。

本文主要参考了客户关系管理领域中用于衡量客户价值的经典模型，即 RFM 模型。RFM 模型因其变量少、特征显性等特点被广泛应用。例如，Anitha 等[15]基于 RFM 模型对消费者的消费历史和购买行为进行分析，结果表明，RFM模型在预测消费者购买行为和相关模式方面具有显著的洞察力；Chen 等[16]基于 RFM 指标相对学习结果的变化趋势，提出了一种基于 SOM 的客户分类方法。但 RFM 模型也存在一定局限性，传统的 RFM 模型的关注点在于用户对具有周期性、快速消费类产品的消费情况，而目前互联网平台上提供的服务与产品存在周期过长或数据稀疏等问题，传统的 RFM 模型并不完全适用。因此，许多学者对模型进行改进以满足不同领域的研究。如 Zhang 等[17]借鉴了 RFM 模型的思路，以人工智能领域的专利数据为例，提出了一种识别特定领域内技术机会的新方法，并发现了该方向的趋势性知识元素；Ji 等[18]构建了改进的 RFMT 模型，用于衡量电子商务客户的价值，将客户群划分为流失客户和新客户群体。

综上所述，聚焦于众包领域的用户细分研究较少，且多是基于用户兴趣、用户能力指标等，未形成较为成熟的用户模型。因此，构建合理的用户价值衡量模型对众包平台用户进行用户细分研究具有重要意义。本文以众包平台上的接包方为研究对象，基于 RFM 模型并结合众包平台及众包用户的特性，将用户信用纳入用户价值模型中，提出并构建了众包用户价值衡量模型——RFMC，基于此模型，结合 GBDT、朴素贝叶斯、多项式逻辑回归等分类算法，进行众包用户细分。

2 研究框架

本文的研究框架分为以下 3 个部分：模型构建、价值计算和用户细分。1）模型构建部分，首先，对传统 RFM 模型进行深入理解，了解从传统 RFM 模型的应用领域延伸至改进的 RFM 模型在不同领域的应用情况，然后结合众包平台及其用户特点，识别出众包用户的关键特征，并将其纳入用户价值模型中，以构建 RFMC 模型；2）价值计算部分，基于 RFMC 模型，获取相关数据项，形成实验数据集，并对指标进行归一化处理，然后使用层次分析法确定各指标的权重，根据用户价值计算公式来确定用户价值；3）用户细分部分，使用 GBDT 算法对预处理后的数据集进行用户分类，并与其他算法进行对比，对提出的 RFMC 模型进行验证。本文的研究框架如图 1 所示。

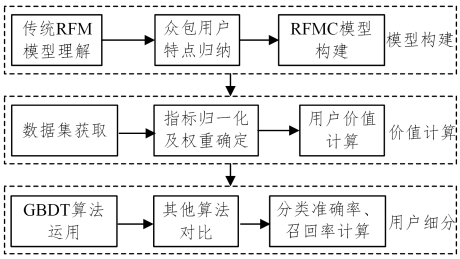

图 1 研究框架

Fig. 1 Research framework

2.1 改进的 RFM 模型

RFM 分析模型由 Hughes[19]于 1994 年提出，其中 3 个变量的详细定义如下：R 表示用户最近一次消费行为发生到现在的时间间隔，间隔越短，R 值越大；F 表示某一段时间内用户的交易次数，频率越大，F 值越大；M 代表购买金额，指某一段时间内用户所消费的货币数量，金额越大，M 值越大。

目前，RFM 模型已经被广泛应用于传统零售行业、电商领域等，但在众包领域的研究较少。RFM 模型通过构建关键指标来识别重要价值对象，该思想可以为众包领域用户细分研究提供重要借鉴。

原来的众包模式主要是发包方在平台上发布任务，有意愿的接包方接收该任务并在规定时间内完成，接包方评估任务结果并将相应酬金给予该结果的产出者。而现在，众包模式逐渐向接包方提供服务、发包方购买服务并给予评价的形式转化，与电商平台消费模式趋近。分析众包平台的相关特性可知，与电商平台相似，众包平台具有服务平台的特性：作为用户信息和任务信息的聚集地，众包平台的数据包括各接包方属性、各发包方属性、接包方与发包方的交易情况、各类任务的属性及执行状态等。同时，与电商平台相似，众包平台还具有社交特性：发包方对任务需求进行描述，用户间可以进行交流，发包方对接包方及任务结果进行评价等。基于电商领域和众包领域的相似性，本文将 RFM 改进模型用于众包领域的用户价值分析，形成 RFMC 模型以实现众包用户细分，具体指标对应到众包平台有其特定的含义。RFMC 模型中各指标的详细定义如下。

(1)R：用户最近完成任务到现在的时间间隔。在众包平台中，用户使用平台的时间间隔反映了用户对平台的粘性，长时间不使用平台的用户为平台带来的增益很小。相对而言，经常借助平台完成任务的用户价值较高。由于平台上未涉及用户最近完成任务的时间，本文以用户被评价的最新时间为替代数据项，因此 R 值的计算式如下：

$$R_i = T_N - T_E \tag{1}$$

其中，R_i 表示用户 i 最近完成任务到现在的时间间隔，T_N 表示当前时间，T_E 表示用户 i 所收到的评价中最新评价的时间。

(2)F：某一段时间内用户的交易次数。在众包平台中，用户发生一次交易即完成一个众包任务，在固定时间内完成任务数量多的用户一般具有较高的技能水平，这样的用户会为平台带来增益且能吸引其他发包方加入。因此 F 值对于描述众包用户特征具有切实意义。需要说明的是，本文的研究平台“一品威客”未提供固定时间段内用户的中标量数据，因此，本文定义的时间段为当前时间与用户信用证书更新的时间间隔，交易次数即为用户中标总次数。每位用户的时间间隔不同，平台中定义的用户信用证书更新时间与用户参与频次强相关，即用户参与频次越多，用户信用证书更新时间离现在越近，时间间隔就越短。因此，本文的 F 值放大了用户频次的效用，F 值的计算式如下：

$$F_i = \frac{N_W}{T_N - T_C} \tag{2}$$

其中，F_i 表示用户 i 的平均交易次数，N_W 表示用户 i 的中标总次数，T_N 表示当前时间，T_C 表示用户 i 的证书更新时间。

(3)M：某一段时间内用户通过众包任务收获的金额。在众包平台中，任务标定的金额往往反映了任务的难度，这种难度体现在任务内容本身或是完成任务所需时间的限制。因此，M 值描述了用户接受任务的难度及其对外部激励的追求。同上，本文定义的时间段为当前时间与用户信用证书更新的时间间隔。因此，M 值计算公式如下：

$$M_i = \frac{P_W}{T_N - T_C} \tag{3}$$

其中，M_i 表示用户 i 的平均获得金额，P_W 为用户 i 通过完成众包任务获得的总金额，T_N 表示当前时间，T_C 表示用户 i 的证书更新时间。

(4)C：众包用户信用。在互联网环境中，信用机制是降低交易中信息不对称和保护各方利益等一系列规范交易主体行为的规则和契约[20]。从众包平台及用户特点的角度考虑，在平台中，用户与用户间可构成社交网络。众包用户的信用分数是基于平台建立的一套信用体系，是用户资质、真实性认证、可信交易行为、口碑评价等数据的集合。在平台的社交网络中，某一接包用户的信用分数是平台和发包方对该用户的共同评价，且这种评价会传播至其他用户，该接包方是否按时完成任务并达到客户要求对用户本人及平台声誉会产生重要影响，也是雇主选择接包方的参考依据。因此，众包用户的信用是衡量其价值的参考因素之一。本文将这种信用定义为用户信用等级证书上的信用分数，C 值的计算式如下：

$$C_i = S_C \tag{4}$$

其中，S_C 表示用户 i 信用等级证书的信用分数。

2.2 价值计算

传统的 RFM 模型通过 3 个变量来识别用户类别，至于如何对这 3 个变量进行权重选择，不同学者有不同的看法。Hughes[19] 认为这 3 个变量的重要性相当，因此 3 个变量的权重相同；Stone 等[21] 指出，由于行业的特点，这 3 个变量在重要性上有所不同，因此 3 个变量的权重不相等。考虑到本文的 RFMC 模型是基于众包领域特点的，4 个指标的权重应各不相同，因此实验利用层次分析法和 yaahp 工具，使用 AHP 方法对各个指标的权重进行确定，而后更新各个指标的数值。

层次分析法由美国运筹学家 Saaty 于 20 世纪 70 年代提出，该方法将与决策相关的要素分成目标层、准则层、方案层来进行定量和定性的分析。本研究按照 4 个主要步骤来实现层次分析法。首先，建立层次结构模型，将确定 RFMC 各指标权重的问题梳理为目标层、准则层、方案层，本文的层次结构模型涉及 4 个维度和 4 个指标；其次，邀请了 5 位众包领域的专家作为决策者，对各指标的重要性进行标度，并计算各层次的判断矩阵；然后，对层次单排序的判断矩阵进行一致性检验，验证其判断的合理性；最后，对层次总排序的判断矩阵进行一致性检验，并最终确定 RFMC 各项指标的权重。

2.3 用户细分

决策树是一种直观的树状模型，被广泛应用于机器学习领域。当运用决策树解决分类问题时，决策树被称为分类树，具有分类规则易理解、速度相对快、准确率相对高的特点[22]；当运用决策树解决回归问题时，决策树被称为回归树。目前典型的决策树算法有 ID3 算法、C4.5 算法以及 CART 算法等。

GBDT，即梯度提升决策树（Gradient Boosting Decision Tree），是一种以决策树为基础学习器的提升算法。相比单一学习器，GBDT 作为一种集成学习方法具有更好的泛化性。利用 GBDT 算法进行分类时，使用 CART 回归树作为弱学习器，使用均方差损失函数的负梯度拟合回归树，对每轮训练

得到的弱分类器进行加权求和得到最终的分类器。

GBDT 的训练过程如图 2 所示。输入训练数据集 $\{(X^{(1)},Y^{(1)}),(X^{(2)},Y^{(2)}),\cdots,(X^{(n)},Y^{(n)})\}$，其中 n 为众包用户的数量；$X^{(i)}$ 为第 i 个样本的特征变量值；$Y^{(i)}$ 为第 i 个样本的类别标签，$Y^{(i)}\in\{1,2,3,4,5,6,7,8\}$。以 $\mu=\frac{1}{n}\sum_{i=1}^{n}Y^{(i)}$ 为样本预测值得到残差，将 $\{(X^{(1)},\mathrm{d}Y^{(1)}),(X^{(2)},\mathrm{d}Y^{(2)}),\cdots,(X^{(n)},\mathrm{d}Y^{(n)})\}$ 作为第一棵树的根节点训练出一个分类器，得到第一棵决策树。根据第一棵决策树获得每个样本的预测值，更新样本残差 $\mathrm{d}Y_i=\mathrm{d}Y^i-\partial_k*p(t_k,X^{(i)})$，$\partial_k$ 为学习率，$p(t_k,X^{(i)})$ 为第 k 棵决策树对第 i 个样本的预测值。第二棵树对更新后的残差 $(\mathrm{d}Y_1,\mathrm{d}Y_2,\cdots,\mathrm{d}Y_n)$ 进行学习，如此重复迭代，直至损失函数趋于收敛。那么样本 j 的分类预测值是前面所有训练树的预测值累加所得。

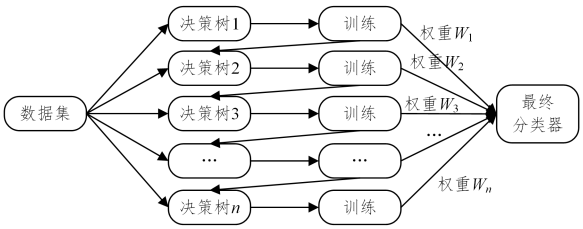

图 2 GBDT 训练过程

Fig. 2 GBDT training process

3 实验及结果分析

3.1 实验过程

本实验数据来自一品威客平台。一品威客是国内领先的创意众包服务平台和知识技能共享服务平台，聚集了包括设计、开发、策划、营销等品类的千万级服务商。

基于前文构建的模型，本文通过 Python 获取所需数据，实验数据主要来自该网站的 logo 设计板块。实验中涉及的数据项如表 1 所列。

表 1 数据项

Table 1 Data items

数据项	解释
用户名称	即店铺名称
信用分数	即威客信用等级证书显示的信用分数
最新评论时间	即店铺的交易评价中的最新评价时间
证书更新时间	即威客信用等级证书的更新时间
总金额	即用户在平台上的交易总金额
中标总次数	即用户在平台上的中标总次数

(1)数据预处理

首先，利用 Python 在一品威客网站的 logo 设计板块爬取了 2359 条用户记录，包含用户名称、信用分数、最新评论时间、证书更新时间、总金额和中标总次数。

其次，去除总金额和中标总次数为 0 的用户记录，去除最新评论时间为缺失值的用户记录，筛选得到 2 276 条用户记录。

然后，将最新评论时间和证书更新时间转化为相同形式，方便后面计算时间间隔；将不同用户的总金额中¥和万转化为统一单位。

最后，分别按照式(5)—式(8)进行归一化处理，计算得到 RFMC 模型中各个指标更新后的值。

$$R1_i=\frac{\max\{R_1,\cdots,R_n\}-R_i}{\max\{R_1,\cdots,R_n\}-\min\{R_1,\cdots,R_n\}} \tag{5}$$

$$F1_i=\frac{F_i-\min\{F_1,\cdots,F_n\}}{\max\{F_1,\cdots,F_n\}-\min\{F_1,\cdots,F_n\}} \tag{6}$$

$$M1_i=\frac{M_i-\min\{M_1,\cdots,M_n\}}{\max\{M_1,\cdots,M_n\}-\min\{M_1,\cdots,M_n\}} \tag{7}$$

$$C1_i=\frac{C_i-\min\{C_1,\cdots,C_n\}}{\max\{C_1,\cdots,C_n\}-\min\{C_1,\cdots,C_n\}} \tag{8}$$

其中，R 为负向指标，F，M 和 C 为正向指标。

(2)权重的确定

层次分析法(Analytic Hierarchy Process，AHP)是一种多属性决策方法[23]。该方法在影响因素、内在关系和复杂决策等多个属性或指标的基础上，选择最佳备选方案以进行方案排序。

本文利用 yaahp 工具将研究中确定各个指标权重的问题划分为目标层、准则层和方案层，在进行指标对比时选择了 5 位众包领域的专家进行比较，最终得出判断矩阵一致性为 0.0370。各指标的权重如表 2 所列。

表 2 各指标权重

Table 2 Weight of each indicator

指标	w_1	w_2	w_3	w_4
权重	0.3469	0.2901	0.1047	0.2583

(3)用户价值计算

基于 AHP 方法确定的各指标权重，利用式(9)计算得到用户价值，如表 3 所列。

$$V_i=R_{1i}*w_1+F_{1i}*w_2+M_{1i}*w_3+C_{1i}*w_4 \tag{9}$$

其中，V_i 表示用户 i 的用户价值，R_{1i}，F_{1i}，M_{1i}，C_{1i} 分别代表归一化后的 R_i，F_i，M_i，C_i 值，w_1，w_2，w_3，w_4 分别代表 R_{1i}，F_{1i}，M_{1i}，C_{1i} 的权重。

表 3 处理后的数据集(部分)

Table 3 Processed data set(part)

ID	R_1	F_1	M_1	C_1	用户价值
0	0.996246	−0.99817	−0.99611	−0.27083	−0.11822
1	0.999725	0	0	0	0.346804
2	0.985683	−0.99998	−0.99997	−0.63194	−0.21609
3	0.999399	−0.98305	−0.94935	−0.34028	−0.12578
4	0.999625	−0.98760	−0.97485	−0.37500	−0.13866
5	0.998974	−0.98943	−0.89032	−0.40972	−0.13954
6	0.996246	−0.99948	−0.99886	−0.47917	−0.17270
7	0.996721	−0.99317	−0.99399	−0.48611	−0.17199
8	0.996095	−0.99731	−0.99506	−0.58333	−0.19863
9	0.99970	−0.98831	−0.98081	−0.59028	−0.19507
10	0.999549	−0.99718	−0.97721	−0.65278	−0.21346
…	…	…	…	…	…

(4)用户类别标注

因为本研究模型包括了 R，F，M，C 这 4 个维度，所以用户分类至少应形成 16 个类别。但从实际管理角度考虑，对 16 个类别的用户差异度进行观察难度较大，且对于平台而言，管理成本也较大。因此，本研究对用户价值进行排序，并结合传统 RFM 模型的 8 个类别的用户定义来进行用户类别标注，分别为高价值用户、一般价值用户、重点发展用户、一般发展用户、重点保持用户、一般保持用户、重点挽留客户和

一般挽留客户，它们分别对应用户标签 1，2，3，4，5，6，7，8。各类用户的特征如表 4 所列。

表 4 各类用户的特点

Table 4 Characteristics of various users

	用户类型	特点
1	高价值用户	用户在众包平台上完成任务的间隔时间较短，完成任务的频率较高，任务收益较高，信用分数较高
2	一般价值用户	用户在众包平台上完成任务的间隔时间较短，完成任务的频率较高，任务收益较高，信用分数较低
3	重点发展用户	用户在众包平台上完成任务的间隔时间较短，完成任务的频率较高，任务收益较低，信用分数较高
4	一般发展用户	用户在众包平台上完成任务的间隔时间较短，完成任务的频率较高，任务收益较低，信用分数较低
5	重点保持用户	用户在众包平台上完成任务的间隔时间较长，完成任务的频率较低，任务收益较高，信用分数较高
6	一般保持用户	用户在众包平台上完成任务的间隔时间较长，完成任务的频率较低，任务收益较高，信用分数较低
7	重点挽留用户	用户在众包平台上完成任务的间隔时间较长，完成任务的频率较低，任务收益较低，信用分数较高。
8	一般挽留客户	用户在众包平台上完成任务的间隔时间较长，完成任务的频率较低，任务收益较低，信用分数较低

(5)GBDT 算法的运用

本研究利用 Python 编程实现 GBDT 算法，将数据集的 75%划分为训练集、25%划分为测试集。利用网格搜索法进行参数寻优，得出最终分类结果和分类评价指标，分类评价指标为准确率和召回率。表 5 列出了部分用户的分类结果。

表 5 分类结果(测试集)

Table 5 Classification results (test set)

用户标签	用户数量(合计 569)	用户类型
1	2	高价值用户
2	2	一般价值用户
3	0	重点发展用户
4	16	一般发展用户
5	47	重点保持用户
6	31	一般保持用户
7	290	重点挽留客户
8	181	一般挽留客户

从表 5 可以看出，在众包平台用户中，重点挽留用户和一般挽留用户占多数，说明平台用户的活跃度整体不高，多数用户参与频率较低，参与行为涉及的金额较小。对于重点挽留用户和一般挽留用户占多数的情况，众包平台应重点关注并分析导致该结果的原因。一方面，这类用户可能是新用户，因进入众包社区时间较短，对众包业务整体流程不熟悉，导致其参与程度低；另一方面，这类用户可能是由其他类型用户转化而来，因平台中存在的问题降低了其对平台的满意度。对于新用户，平台应该着重提供前期引导与培训，使该类用户快速了解平台规则与整体流程，进而提高参与频率；对于对平台满意度较低的用户，平台应该积极调查用户的使用感受，深入了解其需求，对需要改进的平台模块进行重点优化。总体而言，对不同类型用户采取相应管理措施是提高众包任务投标转化率、促进平台稳健发展的关键。

3.2 实验效果对比

为了验证所提改进模型的性能，本文对比了改进模型和未改进模型在 GBDT 算法下的分类准确率和召回率。分类准确率是算法分类正确的样本个数与样本总数的比值，分类准确率的值越大，表明分类模型的分类效果越好；分类召回率是正确分类的用户数与应有分类用户数的比值，分类召回率的值越大，表明分类模型的分类效果越好。RFM 模型与 RFMC 模型的分类准确率、召回率如表 6 所列。

表 6 RFM 模型与 RFMC 模型的对比

Table 6 Comparison between RFM and RFMC models

(单位：%)

指标	RFM(GBDT)	RFMC(GBDT)
准确率	0.5835	0.9982
召回率	0.2227	0.8750

表 6 中，准确率为分类正确的用户数/用户总数；召回率为分类正确用户数/实际分类用户数。

除了利用 GBDT 算法进行用户分类，本文还将朴素贝叶斯算法和多项式逻辑回归分类算法用于众包用户细分的对比实验。基于这 3 种算法的分类结果，使用分类准确率和召回率来衡量分类效果。

从表 7 可以看出，基于本研究提出的 RFMC 模型在 GBDT 算法下具有较高的准确率与召回率。相比朴素贝叶斯算法和多项逻辑回归分类算法，GBDT 算法呈现出明显优势，表明其更适用于本文构建的用户细分模型。因此，基于传统 RFM 模型进行改进得出的 RFMC 模型能够较好地结合众包平台及众包用户的特点，有效识别细分用户群体，并为众包平台的用户管理提供决策支持。

表 7 GBDT 与其他算法的对比

Table 7 Comparison between GBDT and other algorithms

(单位：%)

指标	GBDT	NB	MLR
准确率	0.9982	0.5378	0.5905
召回率	0.8750	0.7141	0.1642

结束语 本文以众包平台及接包用户为研究对象，基于众包平台的特点和平台用户的特征，对众包用户细分进行了研究。针对众包平台面临的无法有效识别重要价值用户的问题和相关理论研究的不足，本文提出了一种基于 RFM 改进的众包用户细分模型 RFMC；明确了改进模型各指标的具体含义，并获取了众包网站的真实数据进行实验；采用不同算法进行用户分类，并对部分用户类别提出平台管理的建议；最后从准确率和召回率两个指标比较了不同算法的分类效果。实验结果表明，本文提出的 RFMC 模型在 GBDT 算法下呈现出了较好的分类效果。

参 考 文 献

[1] FENG S，LIU J P，JIANG H Y，et al. Attribute Value Extraction Method Based on Machine Reading Comprehension Model and Crowdsourcing Verification [J]. Computer Engineering，2021，47(5)：97-103.

[2] SUN Y，HE S J，SHANG R A，et al. Research on the influence of social work platform on employees' improvisation ability：Based on the perspective of online social network [J]. Management World，2019，35(3)：157-168.

[3] LU X Y，LONG D Z，CHEN Y. An analysis of the influence factors of customer engagement intention based on loyalty in

crowdsourcing mode [J]. Journal of Management Science, 2016, 13(7): 1038-1044.

[4] CARBAJAL S G. Customer Segmentation through Path Reconstruction[J]. Sensors, 2021, 21(6): 1-17.

[5] NILASHI M, SAMAD S, MINAEI-BIDGOLI B, et al. Online Reviews Analysis for Customer Segmentation through Dimensionality Reduction and Deep Learning Techniques[J]. Arabian Journal for Science and Engineering, 2021, 46(9): 8697-8709.

[6] LIN J, YANG Z J. User Network Behavior characteristics and Professional Knowledge level: An empirical Study based on autohome Registered Users [J]. Management Review, 2021, 33(5): 1-10.

[7] LU H, ZHANG X X, ZHANG L M, et al.. Research on User Interaction Behavior in Virtual Academic Community from the Perspective of Conversation Analysis [J]. Library and Information Service, 2020, 64(13): 80-89.

[8] XIAO J, CAO H, JIANG X, et al. GMDH-based semi-supervised feature selection for customer classification [J]. Knowledge-Based Systems, 2017, 132: 236-248.

[9] SHI H, LI H J, LAI W, et al. User classification based on Folksonomy tags [J]. Library and Information Service, 2011, 55(2): 117-120.

[10] ZHU H C, HU X, LI S L. Research on User Experience Factor Classification of Government Data Open Platform Based on Kano Model [J]. Modern Information, 2018, 38(12): 13-21.

[11] HU Z X, DU Y, LIU X Y. Design of automatic book recommendation System based on User interest Classification [J]. Modern Electronic Technique, 2021, 44(6): 58-62.

[12] LI Y, MU W S, CHU X Q, et al. K-means Clustering Algorithm based on improved quantum Particle Swarm and its application [J]. Control and Decision, 2021, 13(2): 1-10.

[13] SHANG H. Classification of Internet Users Based on Improved SVM [J]. Application of Computer Systems, 2021, 30(4): 266-270.

[14] SHI X, LI G, LI K, et al. Customer Classification Method of Logistics Enterprises Based on BP-AdaBoost[J]. Journal of Physics, 2020, 1670(1): 12-18.

[15] ANITHA P, PATIL M M. RFM model for customer purchase behavior using K-Means algorithm[J]. Journal of King Saud University-Computer and Information Sciences, 2019(12): 1-11.

[16] CHEN B C, LIANG B, ZHOU Y B, et al. An Application of Self-organizing mapping Neural Network (SOM) in Customer Classification [J]. Systems Engineering theory & Practice, 2004(3): 8-14.

[17] ZHANG Z G, LUO T Y. Technology Opportunity Identification Based on RFM model and Random Actor-oriented Model [J]. Journal of Information Science, 2021, 40(1): 53-61.

[18] JI H J, NI F, LIU J, et al. Customer value Classification of Ecommerce based on Grey Correlation degree and K-Means++ [J]. Application of Computer Systems, 2020, 29(9): 249-254.

[19] HUGHES A M. Strategic database marketing[M]. NewYork: McGraw-Hill Publishing Company, 2005.

[20] YAN H L. Gradient Effect and mechanism of e-commerce platform credit mechanism and user embedment [J]. Business Research, 2020(4): 1-12.

[21] STONE B, JACOBS R. Successful direct marketing methods [M]. NewYork: McGraw-Hill Publishing Company, 2007.

[22] LIU H Y, CHEN J, CHEN G Q. A Review of Data Classification Algorithms in Data Mining[J]. Journal of Tsinghua University (Science & Technology), 2002(6): 727-730.

[23] YOU T H, GAO M L. An Interval Number Multi-attribute Decision Making Method Based on Error Analysis [J]. Journal of Systems Management, 2014, 23(2): 224-228.

CHEN Dan-hong, born in 1996, postgraduate. Her main research interests include crowdsourcing users and cluster analysis.

PENG Zhang-lin, born in 1984, Ph. D, associate professor. His main research interests include information resource management, swarm intelligence and so on.

（责任编辑：李亚辉）