文章编号：2095-2163(2024)01-0214-10 中图分类号：F272 文献标志码：A

基于多模型融合的开放式创新社区内容质量特征挖掘

杨汶静，汪明艳
（上海工程技术大学 管理学院，上海 201620）

摘 要：开放式创新社区高质量用户生成内容特征对企业精准获取技术创新意见具有重要作用。本文构建了开放式创新社区用户生成内容质量多维评价体系，提出一种融合 5 种算法的特征选择方法，在 3 种分类模型评估中得出最优特征子集，挖掘重要因素与高质量用户生成内容之间的关系。集成特征子集在模型上计算时间平均节约 54.54%，比单一特征选择算法得到的特征子集预测准确率平均提高 10.47%。基于多模型融合算法能够客观全面评估开放式创新社区用户生成内容质量，让企业能够精准识别高质量用户生成内容，促进企业开放式创新。

关键词：开放式创新社区；用户生成内容；多模型融合算法；特征选择

Mining content quality features in open innovation community based on multi-model fusion

YANG Wenjing, WANG Mingyan
(School of Management, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: Extract high-quality user-generated content from open innovation communities to provide companies with community management suggestions. This article constructs a multi-dimensional evaluation system for the quality of user-generated content in an open innovation community, propose a feature selection method that integrates five algorithms, evaluate the optimal feature subset from three classification models, and explore the relationship between key factors and high-quality user-generated content. The integrated feature subset save 54.54% of the calculation time on average, which is an average increase of 10.47% compared with the prediction accuracy of feature subset obtained by single feature selection algorithm. The multi-model fusion method can objectively and comprehensively evaluate the quality of user-generated content in open innovation community, allowing companies to accurately identify high-quality user-generated content and promote open innovation.

Key words: open innovation community; user-generated content; multi-model fusion algorithm; feature selection

0 引 言

科技的快速更新、资源的迅速流动给企业带来便利的同时也对企业的创新能力提出了挑战。2003 年 Chesbrough 提出开放式创新理论，企业应该与外部形成开放的合作交流模式，整合内外部创新资源以提升企业科技创新力。开放式创新社区正是企业实施开放式创新战略的一步，而今国内外也有许多知名的开放式创新社区，如 Salesforce、My Starbucks Idea、Apple Developer、小米 MIUI 社区、华为花粉俱乐部、海尔 HOPE 等。

在开放式创新社区中，用户通过发布内容帮助企业改进产品和服务，开发新产品和新业务，积极参与企业创办的创新活动。随着社区规模不断扩大，用户数量逐渐增多，用户发布的内容数量巨大且质量不等，导致企业面对大量用户生成内容时不能高效获取优质内容，增加企业创新成本。

如何在众多用户生成内容中识别出高质量内容是值得关注的问题。本文从用户、内容、社区 3 个维度构建了一个普适的开放式创新社区用户生成内容质量评价指标体系，运用投票算法融合 F 检验过滤法、皮尔森相关系数过滤法、逻辑回归嵌入法、随机森林嵌入法、包装法 5 种单一特征选择算法，得出影响用户生成内容质量的重要因素，使用 SHAP 值探

基金项目：国家社科基金一般项目(17BGL159)；上海市科学技术委员会软科学重点项目(22692104700)。

作者简介：杨汶静(1998-)，女，硕士研究生，主要研究方向：商务统计、数据分析。
通讯作者：汪明艳(1975-)，女，博士，教授，主要研究方向：信息管理、数据分析、电子商务。Email：wmy61610@126.com

收稿日期：2023-02-28

究关键因素与高质量用户生成内容之间存在的关系，为企业开放式创新社区提供综合管理意见。

1　文献综述

1.1　开放式创新社区研究现状

开放式创新社区（Open Innovation Community，OIC）的概念起源于两个重要理论[1]。其一是开放式创新理论，Chesbrough[2] 提出开放式创新（Open Innovation），认为开放式创新是企业融合内部、外部想法并且通过内外市场路径推进公司技术的范式；其二是用户创新理论，由麻省理工学院教授 Von Hippel[3] 提出，强调用户在参与创新的过程中起着重要的作用，甚至在某些行业里用户是创新的主要源泉。继两位学者的开创性工作之后，有不少学者在此基础上展开了对开放式创新社区的深入研究。

现阶段 OIC 研究方向主要分为用户管理、知识管理、创新管理 3 个方面。在用户管理方面，领先用户识别是现研究领域的热点，Pajo[4] 等提出特征提取技术进行自动识别在线主要用户；Yang[5] 基于用户的创新能力、专业能力、影响能力和主动能力 4 个维度的用户创新价值评估体系，进一步构建了包含主题、创新价值和创新阶段的三维用户分类模型框架。在知识管理方面，张海涛等[6] 总结出 OIC 用户之间的知识协同交互过程是由一小部分领先用户带动社区中其他活跃度的一般用户协同参与到知识创新；任伶[7] 基于开放创新社区的创新开放特性，从创新个人、创新社区和创新平台 3 个层面分析 OIC 知识共享的影响因素。在创新管理角度，Wu 等[8] 认为用户创新行为（即发帖）与用户交互行为（即评论和查看）相比，前者对企业创新绩效的影响更大；刘静岩等[9] 以小米 MIUI 社区为例进行实证分析，研究表明用户产出的创新知识点质量越高，企业可利用转化的外部创新知识点越多，进而促进企业创新绩效。

1.2　用户生成内容质量研究现状

用户生成内容（User Generated Content，UGC）是伴随着 Web2.0 时代到来应运而生。2007 年经济合作与发展组织（OECD）[10] 界定了 UGC 的 3 个特征，即互联网上公开可用的内容、内容的创新性、由非专业用户的创作。在企业 OIC 中的 UGC 主要表现为用户发布的评论、文章、创意和提交的问题回答等。

研究成果发现，OIC 中 UGC 质量尤为重要，李奕莹等[11] 通过系统动力学模型得到证实：提高创意（即 UGC）质量可以帮助企业更好地吸收和利用 OIC 中的 UGC，促进企业感知能力、吸收能力和创新能力的提升。UGC 质量的研究成果主要分为 3 个方向：一是 UGC 质量问题，主体多元化、媒介市场化与政治化及其他因素，比如用户线上与线下的不一致性、垃圾数据问题等造成了 UGC 数据质量的不均衡[12]。二是 UGC 质量评价，Fu 等[13] 结合先前研究，建立了基于用户评价标准和数据特征的 UGC 质量评价指标体系。在用户层面上，评估内容特性、认知程度、效用、信息来源、外部因素、社会情感。在数据层面上，选用文本长度、文本结构、文本风格、用户专业性、评审特性 5 个特征指标。以往大多研究多是从用户内在因素调研，存在研究者设定评价标准的主观性、评分人员的主观随意性。进一步，学者们选用客观数据，建立机器学习模型，如 Jain [14] 比较各类监督机器学习方法，发现逻辑回归算法在 UGC 质量预测中表现效果最好；阮光册等[15] 将隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）主题模型应用到高质量 UGC 的识别中，从语义层面挖掘高质量 UGC 所具有的特征。但这些指标测度范围过小、不具有普适性，不能真正为企业解决 UGC 冗余。三是 UGC 质量管控是针对 UGC 质量现存在的问题提出应对的管控措施。Liu 等[16] 发现经济激励与贡献者的参与及其总内容量不是正相关关系。

本文设计出从用户个体、内容、社区 3 个维度评估 UGC 质量的评价指标体系，运用投票算法融合 F 检验（联合假设检验）过滤法、皮尔森相关系数过滤法、逻辑回归嵌入法、随机森林嵌入法、包装法 5 种特征选择算法的优点，在朴素贝叶斯、决策树、梯度提升决策树（Gradient Boosting Decision Tree，GBDT）3 种分类模型评估出最优特征子集，保证计算得出的 UGC 质量重要影响因素的客观准确性。

2　开放式创新社区 UGC 质量多维评价体系

本文基于三元交互决定论构建了开放式创新社区用户生成内容质量评价指标体系。三元交互决定论是由美国著名心理学家阿尔伯特·班杜拉在 20 世纪 70 年末提出的社会认知理论中的重要内容[17]。该理论将环境、行为、个人三者当作相对独立、两两之间又具有交互决定作用对的理论实体。本文从用户个体、发布内容、社区环境 3 个维度细分 UGC 质量评价指标。

2.1　用户个体

用户个体评价标准主要从用户个人信息提取，如用户等级、用户个人信息完成度、发布 UGC 数量等。Li 等[18] 在研究中总结出用户评分方面主要包

含发布内容、追随者、影响力等。在个人特征方面注重研究用户的经验性、接受性[19]。结合本文研究对象 OIC 中用户信息展示的特点,归纳出 9 个特征指标:发布内容数量、粉丝数、阅读量、点赞数、排名、等级、技能数、个人信息完成度、关注数。

针对本文研究对象,在 OIC 中某些特征数据不能直接获得,经过如下转化:

用户技能数:统计用户选择技能标签数;

个人信息完成度:用户所填数与社区用户个人信息总共填写数量的比值;

用户发布所有内容的数量:在 OIC 中 UGC 主要表现为问题、问答等。本研究中统计用户所有发布的问题、提问、回答内容数量的总和。

2.2 内容特征

内容特征评价标准主要分为发布内容特征和内容对应话题特征。内容特征是可以直接从 UGC 中提取到的特征数据,例如 UGC 的字数等。内容对应的话题与 UGC 本身的内容信息是相关的,所有话题相关的特征分析也必不可少。

在质量评价数据特征分析方面,可以从语言特征、文本风格、用户信息、审阅信息、信源可信度 5 个维度出发,对 UGC 和话题各自的 UGC 字数、图片数、链接数统计分析[20]。OIC 中的话题在 OIC 中其他用户浏览产生内容交互,研究中选用内容点赞数、收藏数、评论数 3 个指标评估。在海量的 UGC 中存在数据冗余的问题,而一般用户浏览内容会停留在相对靠前 UGC 上,所以话题下的 UGC 排序也影响着内容质量,内容的专业性也影响着 UGC 质量。

本文对不能直接获取数据的指标做出如下解释:

内容专业性:统计获取话题所属标签与用户技能标签的匹配数量;

话题所属范围大小:话题所属标签个数;

话题热度:即该话题 UGC 发布数量、查看数量和评论数之和。

2.3 社区环境

Yu 等[21]从环境因素的角度分析得到公平、认同和开放三者与社区共享文化构成线性相关关系;同时 Yang 等[22]也证实了用户外部因素的激励条件和认同可以促进用户知识共享。本文选取 4 个社区环境的特征指标:社区认证、社区推荐次数、额外奖励荣誉次数、首位推荐次数。

综上所述,整合用户个体、内容、社区 3 个维度的特征指标得到了表 1 所示的开放式创新社区用户生成内容质量评价指标体系。

表 1 开放式创新社区用户生成内容质量评价体系

Table 1 User-generated content quality evaluation system in open innovation communities

维度	特征指标
个体	用户等级
	用户技能数
	用户个人信息完成度
	用户关注数
	用户粉丝数
	用户排名
	用户获得点赞数
	用户发布内容的阅读总量
	用户发布所有内容的数量
内容	内容字数
	内容图片数
	内容链接数
	内容中提供的代码量
	内容点赞数
	内容收藏数
	内容评论数
	内容专业性
	内容在话题下展示的排序
	话题标题的字数
	话题所属范围大小
	话题下所有的内容总数
	话题关注度
	话题阅读量
	话题热度
	话题详情字数
	话题介绍中包含的图片数量
	话题介绍中的链接个数
社区	用户是否通过社区认证
	受到社区推荐次数
	社区额外奖励荣誉次数
	社区首位推荐次数

3 多模型融合算法框架

本文特征选择算法通过集成投票算法融合 F 检验过滤法、皮尔森相关系数过滤法、逻辑回归嵌入法、随机森林嵌入法、包装法 5 种单一特征选择算法,得到不同效果的集成特征子集,基于多模型融合的特征选择算法框架如图 1 所示。选取朴素贝叶斯、决策树、GBDT 3 种分类模型,测试集成特征子集与样本特征集、单一特征子集在不同分类模型上表现效果。

3.1 单一特征选择算法

在机器学习算法中,特征选择是提高性能以及在某些情况下降低计算成本的关键过程之一。一个数据集中有很多特征与目标值是不相关的,而不相关的特征会严重影响学习过程,因此选择最适合的特征可以提高高维数据集中模型精度,降低内存和计算成本。特征选择主要有 3 种方法:过滤法、嵌入法、包装法。

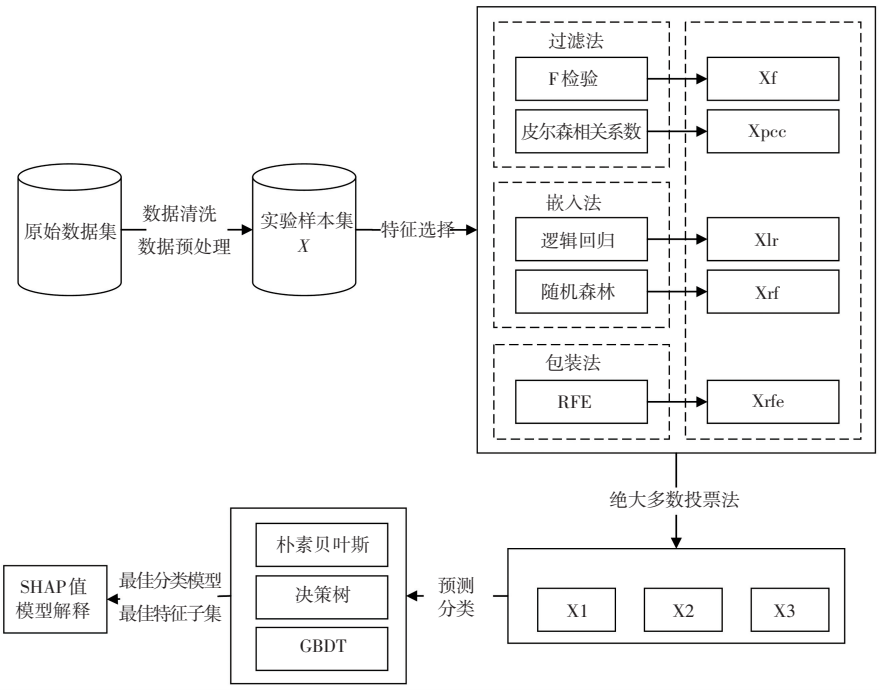

图 1　基于多模型融合的特征选择算法框架

Fig. 1　Feature selection algorithm framework based on multi-model fusion

3.1.1　过滤法

过滤法完全独立于任何机器学习算法，与学习过程无关。过滤法根据各种统计检验计算结果以及相关性的各项指标进行特征选择。常用的方法有方差过滤法、卡方过滤法、皮尔森相关系数、F 检验和互信息法。本文选取 F 检验和皮尔森相关系数两种过滤法。

(1)F 检验：F 检验(F-test)又称方差齐性检验(ANOVA)，是用来计算每个特征与标签之间的线性关系的过滤方法。F 检验的本质是寻找两组数据之间的线性关系，如果某特征 p 值小于 0.05，则表示该特征与标签是显著相关的；若 p 值大于 0.05，则认为该特征与标签没有显著线性关系，应当剔除。

(2)皮尔森相关系数：皮尔森相关系数(Pearson Correlation Coefficient，PCC)衡量两个变量之间线性相关关系，相关系数 r 取值范围是[-1，1]。相关系数的绝对值越接近 1，表示两变量之间的相关性越强；相关系数越接近 0，相关性越弱，相关系数计算如式(1)所示：

$$r = \frac{\mathrm{cov}(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}} \tag{1}$$

其中，x_1，x_2 分别表示两个变量；$\mathrm{cov}(x_1, x_2)$ 表示 x_1，x_2 的协方差；σ_{x_1}，σ_{x_2} 分别表示 x_1，x_2 的标准差。

通常情况下通过表 2 取值范围判断变量的相关强度。

表 2　皮尔森相关系数判断

Table 2　Pearson correlation coefficient judgment

相关系数范围	关系强度
$\lvert r\rvert > 0.8$	表示变量间存在极强相关关系
$0.6 < \lvert r\rvert \leqslant 0.8$	表示变量间存在强相关关系
$0.4 < \lvert r\rvert \leqslant 0.6$	表示变量间存在中度相关关系
$0.2 < \lvert r\rvert \leqslant 0.4$	表示变量间存在弱相关关系
$\lvert r\rvert \leqslant 0.2$	表示变量间存在极弱相关或无相关关系

本文选取皮尔森相关系数计算特征两两之间线性相关关系，若两个特征之间存在极强相关关系，产生特征冗余，浪费模型计算成本，故使用过滤法剔除高度相似的特征。

3.1.2　嵌入法

嵌入法依赖于模型评估完成特征子集选择，集特征选择和算法训练同时进行。使用嵌入法需要用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，权值系数代表了特征对模型的贡献或某种重要性，再根据权值系数从大到小进行特征选择。本文 UGC 研究对象，大多时候高质量用户生成内容是极少部分，所以采集到的数据集具有稀疏性。逻辑回归是常用的稀疏预测模型，使用逻辑回归分类器的 $L1$ 范数作为惩罚项，选择系数不为 0 的特征，以达到减少特征维度的目的。随机森林作为当前主流的机器学习方法具有较好泛化能力。

(1)逻辑回归：逻辑回归(Logistic Regression，LR)是线性回归分析模型的一种，运用逻辑函数(Sigmoid 函数)对线性回归模型的结果进行转化。

LR 模型可以处理二分类问题,也可以用 softmax 函数处理多分类问题。本文选择用 $L1$ 范数实现特征的自动选择,把没有信息的特征权重设置为 0,实现最优特征子集的选择。

(2)随机森林:随机森林(Random Forest, RF)是集合多棵决策树评估器的算法,这个方法是结合 Breimans 的“Bootstrap aggregating”想法和 Ho 的“random subspace method”以建造决策树的集合。在分类问题和回归问题中,都可以使用随机森林建立多个决策树,每棵决策树得出各自的分类结果,然后将其合并在一起以获得更准确稳定的预测结果,所以随机森林的预测效果会比单棵决策树要好。

3.1.3 包装法

包装法与嵌入法相似,特征选择和算法训练也是同时进行,也依赖于算法自身选择。而不同的是会使用一个目标函数进行特征选择。包装法在初始特征集上用分类模型对选定的子集进行评估属性或获取每个特征的重要性,然后从当前特征集中选出最优的特征子集。递归重复该过程,修剪特征子集,直到获取所需数量的最优特征子集。通过基于学习模型中选择特征,包装法比过滤器方法有更好的最终结果,但也存在过拟合和计算复杂度高的风险。

包装法最经典的目标函数是递归特征消除(Recursive Feature Elimination,RFE)。通过 RFE 反复构建模型,并在每次迭代后保留最佳特征,下一次迭代会使用上一次建模中未被选中的特征构建下一个模型,直到所有特征都被选取为止,最后对留下的特征进行排名,选取最佳特征子集。

3.2 集成算法

集成学习(Ensemble learning)主要是通过构建多个个体学习器并使用某种策略将其组合完成学习任务,有时也被称为多分类器系统(Multi-Classifier System)、基于委员会的学习(Committee-Based Learning)等。根据个体学习器的生成方式,目前的集成学习方法大致可分为两大类:一是个体学习器间不存在强依赖关系、可同时生成并行化方法称为 Bagging;二是个体学习器间存在强依赖关系、必须串行生成的序列化方法称为 Boosting。

本文选用 5 种特征选择算法,F 检验和皮尔森相关系数两种过滤法,逻辑回归和随机森林两种嵌入法,基于支持向量机的递归特征消除包装法。三大类特征选择方法都存在局限性:过滤法评价标准独立于模型特定算法,模型预测准确率会比其他两种方法低;嵌入法依赖最终模型算法表现;包装法比嵌入法计算时间更长,不适用于较大数据集。为了克服各类特征选择方法的缺陷,本文选用集成学习中的投票法(Voting)选取最优的特征子集,平衡各类特征选择方法的不足。

投票法将学习器 h_i 在样本 x 上预测输出为一个 N 维向量 $(h_i^1(x);h_i^2(x);\cdots;h_i^N(x))$,其中 $h_i^j(x)$ 是 h_i 在类别标记 c_j 的输出。本文选取绝对多数投票法(Majority voting),特征子集选取机制如式(2)所示:

$$H(x)=\begin{cases}c_j & \text{if } \sum_{i=1}^{T} h_i^j(x) > 0.5\sum_{k=1}^{N}\sum_{i=1}^{T} h_i^k(x)\\ \text{reject} & \text{otherwise}\end{cases} \tag{2}$$

即若某标记得票过半,则预测为该标记;否则拒绝预测。

本文使用 5 种特征选择方法对原始特征集 X 进行筛选,得到 5 个特征子集 $X_f,X_{pcc},X_{lr},X_{rf},X_{rfe} \in X$;再运用绝对多数投票法进行评估,获得 3 个特征子集 $X_1,X_2,X_3 \in X$, 特征子集 X_1 中 $x_j(j=1,\cdots,31)$ 至少被 3 种算法选择,特征子集 X_2 中 $x_j(j=1,\cdots,31)$ 至少被 4 种算法选择,特征子集 X_3 中 $x_j(j=1,\cdots,31)$ 通过 5 种算法选择。

3.3 模型算法

3.3.1 朴素贝叶斯

朴素贝叶斯(Naive Bayesian,NB)模型是以贝叶斯定理为基础,采用“特征之间独立性假设”的方法。通过给定的训练集,以特征之间相互独立作为假设前提,构建联合分布模型,基于训练完成的模型进一步预测,求出后验概率最大的输出值。

3.3.2 决策树

决策树(Decision Tree,DT)是一种非参数的有监督学习方法,以树结构的方式呈现出特征集和标签中总结出的决策规则。通常决策树是递归地选择最优特征,并根据该特征对训练数据进行分割,使得各个子数据集有一个最优分类的过程。

3.3.3 GBDT

梯度提升决策树(Gradient Boosting Decision Tree,GBDT)由多棵决策树组成,迭代集成多个预测结果的算法。该算法能够自适应地调整后续弱评估器拟合的目标,具有较强的泛化能力,能够自动发现特征之间的高阶关系。

3.3.4 模型可解释性

使用集成算法提高学习性能,由于模型复杂度提高,模型可解释性会降低。Lundberg 等 2017 年提出了 SHAP 方法,是一类统一的可解释机器学习方法。SHAP 值来源于 Shapley 在 1953 年提出的

Shapley 值，这是一个来自合作博弈论（coalitional game theory）的方法，根据玩家对总支出的贡献来为玩家分配支出，玩家在联盟中合作并从这种合作中获得一定的收益。SHAP 值是 Shapley 值解释表示为一种可加特征归因方法，将模型的预测值解释表示为输入特征的累积归因值之和，式（3）：

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j' \tag{3}$$

其中，g 表示可解释模型；M 是特征数目；ϕ_i 是每个特征的归因值；ϕ_0 表示所有样本的预测均值；$z' \in \{0,1\}^M$ 表示模型中能否识别该特征，在结构数据中的每个样本特征都是可以被观察到，该公式可简化为式（4）：

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{4}$$

4　实证研究

4.1　实验数据

本文通过 python 调用 web 自动化工具 selenium 库，从腾讯云社区（https://cloud.tencent.com/developer/ask）中直接抓取所需要的客观数据，整合了网站 2021 年 11 月 26 日-2021 年 12 月 3 日内历史数据，包括 5 755 位用户信息，33 944 条用户生成内容。本文用“是否受到推荐或采纳”标准界定用户生成内容质量标签，若此条内容被推荐并被采纳，标签为 2；若此条内容被推荐或采纳，标签为 1；若此条内容未被推荐或采纳，标签则为 0。

4.2　实验环境

实验环境：AMD Ryzen 7 4800U with Radeon Graphics 1.80 GHz；操作系统是 Windows 10 家庭中文版 64 位，基于 PyCharm 专业版 2021.2、Python3.8 编写数据获取清洗、模型计算程序。

4.3　研究结果

4.3.1　数据预处理

在数据爬取过程中，由于部分用户账号存在异常无法访问个人信息，该条用户生成内容特征向量存在缺失值，故剔除此类数据，最终得到特征集的维度为（25 423，31）。

方差是反映一组数据离散程度的度量。如果某一个特征方差为 0 或可变性不大，则表示该特征对模型学习没有较大意义。故数据预处理中，第一步计算特征集 $X = (x_1, x_2, \cdots, x_{31})$ 中特征向量 $\boldsymbol{x}_j^{\mathrm{T}}(j = 1,2,\cdots,31) = (x_{1j}, x_{2j}, \cdots, x_{mj}, m = 25\ 423)$ 的方差，31 个特征向量的方差均大于 0，所以保留 31 个特征进入下一步特征选择。

4.3.2　特征选择

本文运用了 5 种不同的特征选择方法，即 F 检验过滤法（F）、皮尔森相关系数过滤法（PCC）、逻辑回归嵌入法（LR）、随机森林嵌入法（RF）、包装法（RFE）。F 检验过滤法剔除 4 个特征后构成的特征子集 X_f 维度为（25 423，27），皮尔森相关系数过滤法剔除 2 个特征后所得到特征子集 X_{PCC} 的维度是（25 423，29），逻辑回归嵌入法筛选得到维度为（25 423，7）特征子集 X_{IR}，随机森林嵌入法筛选得到维度为（25 423，11）特征子集 X_{RF}，最后利用递归特征消除法得到维度为（25 423，15）的特征子集 X_{RFE}。5 种特征选择方法的筛选特征结果见表 3，其中“1”表示特征保留，“0”表示特征剔除。

表 3　单一特征选择方法的特征子集

Table 3　Feature subsets of single feature selection method

特征指标	F	PCC	RF	LR	RFE	选择次数
内容字数	1	1	1	0	1	4
内容图片数	1	1	0	0	0	2
内容链接数	1	1	0	1	0	3
内容中提供的代码量	0	1	0	0	0	1
内容点赞数	1	1	1	0	1	4
内容收藏数	1	1	1	1	1	5
内容评论数	1	0	1	0	1	3
内容在话题下展示的排序	1	1	1	1	1	5
话题标题的字数	1	1	0	0	1	3
话题所属范围大小	1	1	0	0	0	2
话题下所有的内容总数	1	1	0	0	0	2
话题关注度	1	1	0	0	0	2
话题阅读量	1	0	1	0	1	3
话题热度	1	1	1	0	1	4
话题详情字数	0	1	0	0	1	2
话题介绍中包含的图片数量	0	1	0	0	0	1
话题介绍中的链接个数	0	1	0	0	0	1
内容专业性	1	1	0	0	0	2
用户等级	1	1	0	1	0	3
用户是否通过社区认证	1	1	0	0	0	2
用户技能数	1	1	0	0	0	2
用户个人信息完成度	1	1	0	1	1	4
用户关注数	1	1	0	0	1	3
用户粉丝数	1	1	1	0	1	4
用户排名	1	1	0	0	0	2
用户获得点赞数	1	1	1	1	1	5
用户发布内容的阅读总量	1	1	0	0	0	2
社区首位推荐次数	1	1	0	0	0	2
受到社区推荐次数	1	1	1	0	1	4
社区额外奖励荣誉次数	1	1	0	1	0	3
用户发布所有内容的数量	1	1	1	0	1	4

进一步结合集成投票算法的结果，从 31 个特征中剔除 14 个特征，根据特征被不同算法选择的次数，得到 3 个特征子集 X_1, X_2, X_3。特征子集 X_1 包含 17

个特征,特征子集 X_2 包含 10 个特征,特征子集 X_3 包含 3 个特征。3 个特征子集所包含的特征见表 4,“1”表示该特征子集中包含此特征。

表 4 集成特征选择方法的特征子集

Table 4 Feature subsets of integrated feature selection method

特征指标	X_1	X_2	X_3
内容收藏数	1	1	1
内容在话题下展示的排序	1	1	1
用户获得点赞数	1	1	1
内容字数	1	1	
内容点赞数	1	1	
话题热度	1	1	
用户个人信息完成度	1	1	
用户粉丝数	1	1	
受到社区推荐次数	1	1	
用户发布所有内容的数量	1	1	
内容链接数	1		
内容评论数	1		
话题标题的字数	1		
话题阅读量	1		
用户等级	1		
用户关注数	1		
社区额外奖励荣誉次数	1		

4.3.3 特征子集评估

4.3.3.1 原始特征集与集成特征子集

评估基于集成特征选择的效果,将原始特征集 X 和 3 个集成特征子集 X_1,X_2,X_3 分别输入朴素贝叶斯(NB)、决策树(DT)、梯度提升决策树(GBDT)3 种不同的分类模型,经过十折交叉验证得出计算结果;纵向比较特征选择前后模型预测准确率、模型计算时间,见表 5;经对比分析结果可以得出,特征选择后在 3 个分类模型上的分类预测准确率都有不同程度的提高,模型计算时间大幅下降。在朴素贝叶斯模型上,特征子集 X_1,X_2,X_3 准确率分别提高了 17.7%、20.80%、20.14%,模型计算时间平均缩短 59.48%。在决策树模型上,特征子集 X_1,X_2 准确率与原始特征集预测准确率大致相同,但计算时间平均缩短 56.48%,大幅节约了运行成本。在 GBDT 模型上,特征子集 X_1,X_2 预测准确率有微弱的提升效果,而计算时间平均缩短 53.60%。特征子集 X_3 特征数量大幅减少,影响了在 DT、GBDT 模型上的表现效果。

再进一步横向对比 3 个分类模型,输入同一特征子集,DT 分类准确率最高且在使用集成特征子集 X_2 时模型学习效果最佳。与 NB 模型比较,分类准确率平均提高 9.04%,计算时间平均增加 0.23 s。与 GBDT 模型比较,分类准确率平均提高 0.1%,但计算时间平均减少了 73.34 s。DT 模型准确率高且学习计算时间短,说明决策树比较适用于本文的分类问题,为今后开放式创新社区 UGC 质量预测研究提供了一条新的思路。此外,集成特征子集 X_2 与原始特征集 X 比较,在 3 个分类模型上的预测准确率平均提高 9.35%,计算时间缩短了 54.54%。

表 5 原始特征集与集成特征子集在各类分类模型上的表现效果

Table 5 The performance of the original and integrated feature subsets on three classification models

特征集	NB	NB 计算时间	DT	DT 计算时间	GBDT	GBDT 计算时间
X	0.742 3	0.222 1	0.960 0	0.612 1	0.958 9	123.237 9
X_1	0.920 0	0.125 0	0.960 0	0.439 1	**0.959 1**	92.332 6
X_2	**0.950 3**	0.089 0	**0.960 2**	0.291 1	**0.959 1**	60.078 6
X_3	0.943 7	0.056 0	0.957 4	0.069 0	0.956 7	19.123 3

4.3.3.2 集成特征子集与单一特征子集

为了进一步研究集成特征子集的融合效果,本文将 3 个集成特征子集 X_1,X_2,X_3 分别与单一特征子集 X_F、X_{PCC}、X_{IR}、X_{RF}、X_{RFE} 在 3 个分类模型上纵向比较,评估集成特征子集的筛选效果。

首先,比较特征子集 X_1、X_F、X_{PCC}、X_{IR}、X_{RF}、X_{RFE} 在 3 个不同分类模型上的表现。特征子集 X_1 中共有 17 个特征,由于特征子集 X_{IR}、X_{RF}、X_{RFE} 特征数量分别为 7、11、15,不满足维度为(25 423,17)的特征子集,故在特征子集 X_F、X_{PCC} 中选出前 17 个特征与特征子集 X_1 比较,将 3 个特征子集分别输入 3 个分类模型,计算预测准确率和运行时间,见表 6。根据分析可以得出,与单一特征子集 X_F、X_{PCC} 相比,集成特征子集 X_1 在 NB、DT、GBDT 3 种分类模型中分类准确率分别平均提高 56.76%、0.73%、0.57%。

表 6　特征子集(特征数量为 17)在各类分类模型上的表现效果

Table 6　The performance of feature subsets (the number of features is 17) on various classification models

特征集	NB	NB 计算时间	DT	DT 计算时间	GBDT	GBDT 计算时间
X_1	**0.920 0**	0.125 0	**0.960 0**	0.439 1	**0.959 1**	92.332 6
X_F	0.904 78	0.140 0	0.956 7	0.252 0	0.957 3	55.231 2
X_{PCC}	0.434 3	0.143 0	0.949 4	0.303 1	0.950 1	73.421 3

其次，在模型上比较特征子集 X_2、X_F、X_{PCC}、X_{IR}、X_{RF}、X_{RFE} 的学习效果。特征子集 X_2 中有 10 个特征，由于特征子集 X_{IR} 中只有 7 个特征，故在特征子集 X_F、X_{PCC}、X_{RF}、X_{RFE} 中排名靠前的 10 个特征与特征子集 X_2 纵向比较，效果见表 7。可以发现，与 X_F、X_{PCC}、X_{RF}、X_{RFE} 相比，特征子集 X_2 在 NB、DT、GBDT 3 种分类模型中分类准确率分别平均提高 30.65%、0.40%、0.35%。

表 7　特征子集(特征数量为 10)在各类分类模型上的表现效果

Table 7　The performance of feature subsets (the number of features is 10) on various classification models

特征集	NB	NB 计算时间	DT	DT 计算时间	GBDT	GBDT 计算时间
X_2	**0.950 3**	0.089 0	**0.960 2**	0.291 1	**0.959 1**	60.078 6
X_F	0.926 3	0.089 0	0.957 6	0.164 0	0.957 1	31.445 1
X_{PCC}	0.442 7	0.089 7	0.949 8	0.249 1	0.950 4	50.195 5
X_{RF}	0.924 8	0.090 0	0.960 0	0.326 1	0.958 5	68.806 4
X_{RFE}	0.926 2	0.089 0	0.958 2	0.378 1	0.957 2	79.762 9

最后，比较特征子集 X_3，X_F，X_{PCC}，X_{IR}，X_{RF}，X_{RFE} 在 3 个分类模型上的表现，特征子集 X_3 中只有 3 个特征，在特征子集 X_F，X_{PCC}，X_{IR}，X_{RF}，X_{RFE} 中选出前 3 个特征，输入 NB、DT、GBDT 模型计算预测准确率和运行时间，结果见表 8。在各个分类算法中，与单一特征子集 X_F，X_{PCC}，X_{IR}，X_{RF}，X_{RFE} 相比，特征子集 X_3 在 NB、DT、GBDT 3 种分类模型中分类准确率分别平均提高了 44.37%、0.59%、0.52%。

表 8　特征子集(特征数量为 3)在各类分类模型上的表现效果

Table 8　The performance of feature subsets (the number of features is 3) on various classification models

特征集	NB	NB 计算时间	DT	DT 计算时间	GBDT	GBDT 计算时间
X_3	**0.943 7**	0.056 0	**0.957 4**	0.069 0	**0.956 7**	19.123 3
X_F	0.941 9	0.059 0	0.950 6	0.079 0	0.950 7	21.159 1
X_{PCC}	0.926 8	0.072 0	0.950 8	0.095 0	0.950 8	22.061 9
X_{IR}	0.447 6	0.059 0	0.953 9	0.063 0	0.953 7	17.768 0
X_{RF}	0.940 0	0.058 0	0.953 7	0.087 0	0.953 6	23.175 2
X_{RFE}	0.452 4	0.060 0	0.949 8	0.166 0	0.949 8	39.908 9

综上所述，本文得出 3 个集成特征子集 X_1，X_2，X_3 比单一特征子集在不同分类模型上的表现都更好。这一纵向对比说明了集成特征选择是有效的，可以弥补各类单一算法特征选择的缺点，集合各自优势找到与预测目标最相关的特征，更好地进行模型学习，能够更加准确地识别出开放式创新社区中的高质量 UGC。在 3 个特征子集 X_1，X_2，X_3 中，3 个分类预测模型的准确率平均为 93.80%、95.92%、95.83%，特征子集 X_2 的平均分类预测准确率分别比 X_1，X_3 高出 2.12%、0.09%。虽然特征子集 X_2 与 X_3 的准确率相差不大，但在模型计算时间上平均节约了 56.91 s，大幅提高了模型的计算能力，故本文选择特征子集 X_2 的所有特征作为影响开放式创新社区用户生成内容质量的关键因素，根据开放式创新社区用户生成内容质量评价指标体系的个体、内容、社区 3 个维度归纳得到开放式创新社区用户生成内容质量影响因素见表 9。

表 9　开放式创新社区用户生成内容质量影响因素

Table 9　Factors of user-generated content quality in open innovation communities

维度	特征指标
个体	用户获得点赞数
	用户个人信息完成度
	用户粉丝数
	用户发布所有内容(文章、提问、回答)的数量
内容	内容收藏数
	内容在话题下展示的排序
	内容字数
	内容点赞数
	话题热度
社区	用户获得点赞数

4.3.4　结果分析

本文的分类模型比较中，决策树表现出更好的学习效果，故选用决策树作为特征解释的分类模型。设定决策树模型训练参数最大树深为 4，将特征子集 X_2 输入决策树中，分类准确率为 96.02%，运行计算时间为 0.291 1 s。因为特征子集 X_2 是通过集成 5

个特征选择结果得到的最优特征子集,而在集成学习的过程中模型复杂度提高,模型可解释性会降低,所以在结果分析中选用了 SHAP 值进行解释说明。

运用 SHAP 值计算得到不同特征对于模型影响的绝对平均值,即各个特征对于模型影响的重要度。在开放式创新社区中 UGC 展示排序、UGC 点赞数、UGC 收藏数、社区推荐次数 4 个特征明显区分了 UGC 质量,而话题热度对 UGC 是否只受到推荐几乎没有影响,用户发布内容总数、UGC 文本字数不是高质量 UGC 判断标准。如果 UGC 排位靠前、受到用户更多的收藏,发布内容的用户此前受到社区更多的推荐,那么 UGC 就更容易受到推荐;如果 UGC 排位靠前、获得更多的点赞数,则 UGC 就更容易受到采纳。

为了挖掘出影响开放式创新社区高质量用户生成内容的重要因素,帮助企业或社区发现优质的用户生成内容以更好地管理社区、促进企业开放式创新。本文进一步结合 SHAP 值研究分析得出,开放式创新社区中高质量用户生成内容的三大重要特征:UGC 点赞数、UGC 收藏数和社区推荐次数。话题下 UGC 的排位次序首要影响着 UGC 质量。由于在本文研究对象腾讯云社区中,用户可以不断修改自己的发布内容,内容排序越靠前越能提高内容质量。故社区可以考虑开放内容编辑修改功能,从而提高 UGC 质量。对于 UGC 点赞数和社区推荐次数,样本都存在两极分化分布。一般地,UGC 点赞数、发布内容的用户此前受到社区推荐次数越多,表示 UGC 的质量越好;反之,UGC 点赞数、社区推荐次数越少会对其质量产生负面影响。但也存在少部分高质量 UGC,点赞数和社区推荐次数较少。此外,UGC 收藏数与内容质量高低存在正相关关系。UGC 文本字数和用户发布内容总数对大多数高质量 UGC 没有明显影响,只有小部分样本反映出 UGC 字数越少,反而会提高 UGC 质量;用户发布内容总数越多,代表 UGC 质量越好。

5 结束语

本文选取“腾讯云+社区”作为全新的研究对象,并基于三元交互决定论从用户个体、用户内容、社区环境 3 个维度构建了评价指标体系,完善了以前仅从用户或内容层面构建指标的不足,使得 UGC 质量评价指标体系更加全面。

融合 F 检验、皮尔森相关系数、逻辑回归嵌入法、随机森林嵌入法、递归特征消除法 5 种单一特征选择算法,选取 UGC 质量重要影响因素,在朴素贝叶斯、决策树、GBDT 3 种经典机器学习模型上评估特征选择结果,本文还进一步考虑了特征与特征相互联系,避免单个模型评估结果的不稳定性,得出的影响指标具有客观准确性,可以有效解决“维度灾难”问题,减小模型复杂度的同时模型准确率提升。

基于最优特征子集和分类效果最佳的模型,进一步分析了重要影响因素对高质量 UGC 的作用效果,发现高质量用户生成内容的三大关键特征:UGC 点赞数、UGC 收藏数和社区推荐次数。针对本研究成果为企业 OIC 管理提出如下建议:

首先,支持 UGC 修改功能以提升排序。用户在 OIC 的话题下可以不断修改发布内容,社区可以设计明显修改功能提示或内容排序激励方案,通过 UGC 排序竞争鼓励用户发布高质量内容;

其次,UGC 收藏数是 UGC 质量重要评判标准,UGC 收藏数与质量呈正相关关系,用户大多会在阅读完内容后选择是否收藏内容,这也等同于是对内容质量单向评价;

最后,企业 OIC 中需要注意 UGC 字数越多往往不代表其属于高质量内容,尽量避免空文本出现;对于发布 UGC 的用户,若以往发布内容总数越多,不代表在此问题下有高质量内容贡献,还要注意内容本身。

本文为 OIC 管理人员或企业提供社区管理决策意见,也为今后 UGC 研究方向提供一个思路。在今后的研究中,可以更进一步聚类探究不同的用户类型对质量的影响,以提高分类准确率。

参考文献

[1] 秦敏. 企业开放式创新社区研究探索与展望[J]. 江西师范大学学报(哲学社会科学版),2014,47(5):21-26.

[2] CHESBROUGH H W. Open innovation the new imperative for creating and profiting from technology[M]. Harvard Business School Press, 2003:157-169.

[3] HIPPEL V E. Lead users: a source of novel product concepts[J]. Management Science,1986, 32(7):791-805.

[4] PAJO S J, PAUL-ARMAND V, DENNIS V, et al. Fast lead user identification framework[J]. Procedia Engineering, 2015, 131: 1040-1145.

[5] YANG J J. A framework of user classification model of online user innovation communities based on user innovation value[J]. Open Journal of Social Sciences,2020,8(5):232-244.

[6] 张海涛,刘伟利,任亮,等. 开放式创新社区的用户知识协同交互机理及其可视化研究[J]. 情报学报,2021,40(5):523-533.

[7] 单晓红,王春稳,刘晓燕,等. 开放式创新社区领先用户识别——知识基础观视角[J]. 数据分析与知识发现,2021,5(9): 85-96.

[8] WU B, GONG C Y. Impact of open innovation communities on enterprise innovation performance: a system dynamics perspective [J]. Sustainability,2019,11(17):1-27.

[9] 刘静岩,王玉,林莉. 开放式创新社区中用户参与创新对企业社

区创新绩效的影响——社会网络视角[J]. 科技进步与对策，2020,37(6):128-136.
[10]VICKERY G, WUNSCH V S. Participative web and user created content: Web 2.0, Wikis and Social Networking [M]. Paris: Organization for Economic Cooperation and Development (OECD),2007
[11]李奕莹,戚桂杰. 基于系统动力学的企业开放式创新社区中用户生成内容管理研究[J]. 情报杂志,2017,36(4):112-117,129.
[12]陈峥. 不平行的空间:用户生成内容大数据质量探析[J]. 图书馆,2021(3):90-98.
[13]FU H Y, SANGHEE O. Quality assessment of answers with user-identified criteria and data-driven features in social Q&A [J]. Information Processing & Management, 2018, 56(1): 14-28.
[14]JAIN P K, PAMULA R, ANSARI S. A supervised machine learning approach for the credibility assessment of user-generated content [J]. Wireless Personal Communications, 2021(3):1-17.
[15]阮光册,夏磊. 高质量用户生成内容主题分布特征研究[J]. 图书馆杂志,2018,37(4):95-101.
[16]LIU Y W, FENG J. Does money talk? the impact of monetary incentives on user-generated content contributions[J]. Information Systems Research,2021,32(2):394-409.
[17]BANDURA A, CERVONE D. Differential engagement of self-reactive influences in cognitive motivation [J]. Organizational Behavior & Human Decision Processes,1986,38(1):92-113.
[18]LI L, HE D Q, JENG W, et al. Answer quality characteristics and prediction on an academic Q&A site: a case study on research gate [C]// Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). Association for Computing Machinery, New York, NY, USA, 2015:1453-1458.
[19]DAHANAYAKE A. An approach to improve the quality of user-generated content of citizen science platforms [J]. ISPRS International Journal of Geo-Information,2021,10(7):434
[20]沈旺,李世钰,刘嘉宇,等. 问答社区回答质量评价体系优化方法研究[J]. 数据分析与知识发现,2021,5(2):83-93.
[21]YU T K, LU L C, LIU T F. Exploring factors that influence knowledge sharing behavior via weblogs[J]. Computers in Human Behavior,2009,26(1):32-41.
[22]YANG C, YANG Y Y, SHI W D. A predictive model of the knowledge-sharing intentions of social Q&A community members: a regression tree approach [J]. International Journal of Human-Computer Interaction,2022,38(4):324-338.

(上接第 213 页)

路径系数较大(2.024),这表明乘客对车站是否熟悉将对无障碍电梯的选择造成显著影响。因此,地铁车站应采取相关措施,使乘客更了解地铁车站设施布局。如:张贴地铁车站布局图、循环广播设施位置等,使有无障碍电梯需求的乘客可以高效、便利地找到并使用无障碍电梯,进而有效提升乘客的出行体验,降低站台拥挤程度。

3　结束语

本文以地铁乘客无障碍电梯选择行为做为研究对象,构建指标体系并提出关于乘客无障碍电梯选择行为的假设关系。基于实地调研得到的 112 份有效调查问卷,选择结构方程模型进行研究,验证了指标体系和假设路径的合理性,具体结论如下:

(1)通过 AMOS 26.0 软件验证了指标体系及问卷量表的信度和效度。计算结果表示,信度指数结果理想,近似均方根误差、规范拟合指数、规范拟合指数接近判别标准。说明模型的拟合优度和适配度较好,能较为真实的反映实际情况。

(2)通过模型路径图得到各潜变量之间的关系,显变量对潜变量的影响情况。心理潜变量、环境因素都对乘客无障碍电梯的选择行为造成正面影响,人因因素会造成负面影响,系数分别为:1、0.08、-2.79。

(3)通过中介效应和结构方程,对乘客无障碍电梯选择行为影响因素的作用机制展开研究,其中乘客负重程度、乘客对车站是否熟悉、无障碍电梯可视性指数 3 条路径的路径系数最大,分别为 2.417、2.024、1.855,表明这 3 个因素将会对乘客无障碍电梯选择行为造成最大的影响。

参考文献

[1] 耿亚宁,胡华,孟艳丽,等. 基于视频识别的地铁站突发大客流智能预警方法研究[J]. 智能计算机与应用,2022,12(2):170-173,177.
[2] 张凌波,郝妍熙,胡华. 基于 ATS 与 AFC 数据的地铁乘客出站走行时间估计方法[J]. 智能计算机与应用,2020,10(12):164-169.
[3] 梁妍娇,袁振洲. 基于 Logit 模型的综合交通枢纽乘客出站设施选择[J]. 交通信息与安全,2014,32(4):36-40.
[4] 刘剑锋,孙福亮,柏赟,等. 城市轨道交通乘客路径选择模型及算法[J]. 交通运输系统工程与信息,2009,9(2):81-86.
[5] DELL'ORCO M, MARINELLI M. Modeling the dynamic effect of information on drivers' choice behavior in the context of an Advanced Traveler Information System [J]. Transportation Research Part C Emerging Technologies, 2017, 85:168-183.
[6]刘建荣,刘志伟,任倩. 考虑出行者异质性的高铁站到达方式选择[J]. 华南理工大学学报(自然科学版),2019,47(9):47-52.
[7] 陈立扬,宋瑞,李志杰,等. 基于 Anylogic 的地铁站站厅层设施布置仿真研究[J]. 交通信息与安全,2013,31(5):19-24.
[8] 谷鑫鑫,赵胜川,罗欢欢. 考虑共享汽车的高铁站接驳交通出行方式选择影响因素[J]. 交通运输研究,2021,7(4):10-17.
[9] 景鹏,隽志才,查奇芬. 考虑心理潜变量的出行方式选择行为模型[J]. 中国公路学报,2014,27(11):84-92,108.
[10] AJZEN I. The theory of planned behavior [J]. Organizational Behavior and Human Decision Processes, 1991, 50(2):179-211.
[11]熊文真,徐建新,张娅莉. 基于结构方程模型的白酒满意度影响因素分析[J]. 昆明理工大学学报(自然科学版),2023,48(4):186-193.