杨汶静, 汪明艳. 基于混合模型的开放式创新社区用户生成内容质量预测[J]. 智能计算机与应用,2024,14(5):179-185.
DOI:10.20169/j.issn.2095-2163.240524

基于混合模型的开放式创新社区用户生成内容质量预测

杨汶静, 汪明艳
(上海工程技术大学 管理学院, 上海 201620)

摘 要: 为解决开放式创新社区内容冗余导致高质量用户生成内容无法充分发挥其价值的问题,挖掘高质量 UGC 深层价值。首先采用随机过采样、SMOTE、ADASYN 解决 UGC 数据不平衡问题,然后构建支持向量机、朴素贝叶斯、决策树、随机森林、GBDT 分类模型并生成多种混合预测模型,进一步使用基于 Hard-voting、Soft-voting、Stacking 的采样方法和分类模型组合优化预测方法,比较选取最优的开放式创新社区 UGC 质量预测模型。采用随机过采样和 Stacking 的混合模型 *Accuracy*、*F*1 值和 *AUC* 分别平均提升了 3.85%、28.18%、12.30%。该方法能够精准识别创新社区高质量用户生成内容,帮助企业多维度管理社区、提高创新力。

关键词: 开放式创新社区; 用户生成内容; 过采样; 机器学习; 混合模型预测
中图分类号: TP391.2 **文献标志码**: A **文章编号**: 2095-2163(2024)05-0179-07

Quality prediction of User-Generated Content in open innovation community based on mixed model

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

Abstract: To solve the problem that high-quality UGC cannot give full play to its value due to the content redundancy of open innovation community, and excavate the deep value of high-quality UGC, random oversampling, SMOTE, ADASYN are used to solve the UGC data imbalance in this paper. Then, the classification models such as Support Vector Machine, Naive Bayes, Decision Tree, stochastic forest, GBDT are built and a variety of mixed prediction models are generated. Further, sampling methods based on Hard-voting, Soft-voting, Stacking and classification model combination optimization prediction method are used to compare and select the optimal UGC quality prediction model of open innovation community. *Accuracy*, *F*1 and *AUC* values are increased by 3.85%, 28.18% and 12.30% on average, respectively, with random oversampling and Stacking. This method can accurately identify high-quality User-Generated Content in innovation communities, correspondingly help enterprises manage communities in a multi-dimensional manner and improve innovation ability.

Key words: open innovation community; User-Generated Content; oversampling; machine learning; hybrid algorithm

0 引 言

互联网的快速发展推动着企业传统的封闭式创新逐渐转向开放式创新[1],更好地整合企业内、外部资源进行产品优化、创新。企业通过开放式创新社区(Open Innovation Community, OIC)让用户参与到知识、产品的创造过程中,这已成为企业聚集和吸引用户共享知识和提出产品相关想法的关键战略[2]。

目前,现阶段 OIC 研究方向主要分为用户管理、知识管理、创新管理三个方面。Pajo 等学者[3]提出特征提取技术进行自动识别在线主要用户。Yang[4]基于用户的创新能力、专业能力、影响能力和主动能力四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。张海涛等学者[5]以花粉俱乐部为案例研究发现核心用户或者结构洞用户在知识创新过程中起着关键推动作用。Wu 等学者[6]认为

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: 2023-03-23

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。刘静岩等学者[7]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。张宁等学者[8]认为创意信息熵、情感强度、支持量、企业回复信息熵、回复情感强度都正面影响创意采纳。吉海颖等学者[9]以 Lego Ideas 社区为例,基于意见的交互是促进用户高创意更新主要路径中的核心存在,"声音"比"行动"更能促进用户的创意更新持续贡献。

在现有的国内外研究成果中,虽然能够识别开放式创新社区的领先用户,但对用户生成内容质量评估的相关研究较少。随着互联网的渗透、开放式创新社区的推广,社区的用户、内容日益增多,但是也不可避免地出现了信息泛滥和信息超载的问题,从而形成马太效应。基于此,以腾讯云社区用户及其发布内容作为研究对象,建立多方面多维度开放式创新社区用户生成内容质量评价指标体系,运用机器学习方法训练各类模型。进一步,基于 Stacking 算法组合优化模型提取影响高质量用户生成内容的关键影响因素,对开放式创新社区用户生成内容的有用性识别进行实证分析。

1 文献综述

1.1 用户生成内容质量预测模型

随着 Web2.0 快速发展,用户生成内容是各大平台、网站中最重要的组成部分。近年来,在学术研究领域内 UGC 作为一种新的研究数据,研究人员运用不同的方法来研究、评估 UGC 相关的用户和内容。用户累积形成社区,社区和用户自身各方面因素影响着用户生成内容质量。根据现有高质量用户生成内容的相关数据,进一步识别未来社区中用户发布的优质内容,这样可以帮助社区发掘高质量内容、完善社区管理。在构建用户生成内容质量预测模型上,Jain 等学者[10]比较各类监督机器学习方法发现逻辑回归算法 UGC 质量预测中表现效果最好。Sahu 等学者[11]根据回答者的权威性特征构建各类机器学习分类模型,研究得出随机森林分类器优于其他分类算法。而 Li 等学者[12]从学术社交网站 ResearchGate 发现优化的支持向量机算法在准确性方面优于其他模型。金燕等学者[13]基于低质量用户生成内容的用户画像运用机器学习方法构建预测模型,通过重点监测异常行为来识别低质量 UGC。阮光册等学者[14]将 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。Lei 等学者[15]提出 UCCC(User Communities and Contents Co-ranking)算法是基于社区、用户以及二者之间相互联系网络的联合排名算法。王伟等学者[16]从 3 个维度构建答案质量组合特征向量,分别构建逻辑回归、支持向量机、随机森林二分类模型,准确率分别为 0.861、0.847、0.921。

1.2 数据不平衡问题

此外,开放式创新社区用户生成内容的真实数据往往存在不平衡的问题,该问题指多数类的样本数量远远超过少数类。这种不平衡数据集会使传统的机器学习方法构建预测模型往往会表现较差、性能降低。由于开放式创新社区中用户增加、内容泛滥,实际只存在少数量的高质量用户生成内容,从而造成了高质量用户生成内容与低质量用户生成内容比例存在严重的不平衡。解决不平衡数据集分类问题的主要目的是保证提高少数类分类准确率的同时不会使得多数类分类准确率下降太多,其方法主要分为 2 类。一是从数据层面,经过数据处理后改变数据分布情况;二是从算法或模型层面,通过多种机器学习算法平衡数据集[17]。从数据层面上看,分为过采样和欠采样两种方法。其中,过采样方法是指增加少数类样本以达到数据集平衡,欠采样方法则是指剔除多数类样本已到达数据集平衡。本研究数据集偏小,所以选用过采样方法。最简单的方法是随机过采样法(Random Oversampling),从少数类样本集中随机重复抽取得到同多数类样本量的样本[18],但此种方法容易造成过拟合。Chawla 等学者[19]提出了少数类过采样技术(Synthetic Minority Over-sampling Technique,SMOTE),对少数类样本的相邻点进行随机线性选择从而合成新样本。He 等学者[20]提出自适应合成采样(Adaptive Synthetic Sampling, ADASYN),对少数类样本赋予不同权重计算得出不同类别样本的数。

综上,开放式创新社区用户生成内容质量预测模型研究存在以下不足。一是在数据方面,仅从用户、内容、社区或任意 2 个维度选取特征变量,研究数据维度单一,不能全面地展现出开放式创新社区用户的真实使用状态。二是在模型方面,构建预测模型之前没有处理真实数据不同类别之间的不平衡问题,并且预测模型往往表现效果不佳。

因此,本研究从用户个体、内容、社区三个维度构建了开放式创新社区 UGC 质量的评价指标体系,

采用随机过采样、SMOTE、ADASYN 三种方法以解决开放式创新社区 UGC 数据不平衡问题，然后构建支持向量机、朴素贝叶斯、决策树、随机森林、GBDT 五种机器学习分类模型并生成多种组合预测模型，进一步使用基于 Hard－voting、Soft－voting、Stacking 的采样方法和分类模型组合优化预测方法，比较选取最优的开放式创新社区 UGC 质量预测模型，便于社区发掘优质用户内容，优化社区知识内容，为企业输送不竭动力。

2　模型与方法

2.1　模型构建

在开放式创新社区中，高质量的用户生成内容总是会占极少数的，所以社区样本数据的分类占比会处于不平衡的状态。本研究提出的组合预测模型框架如图 1 所示。这里将对此做阐释分析如下。

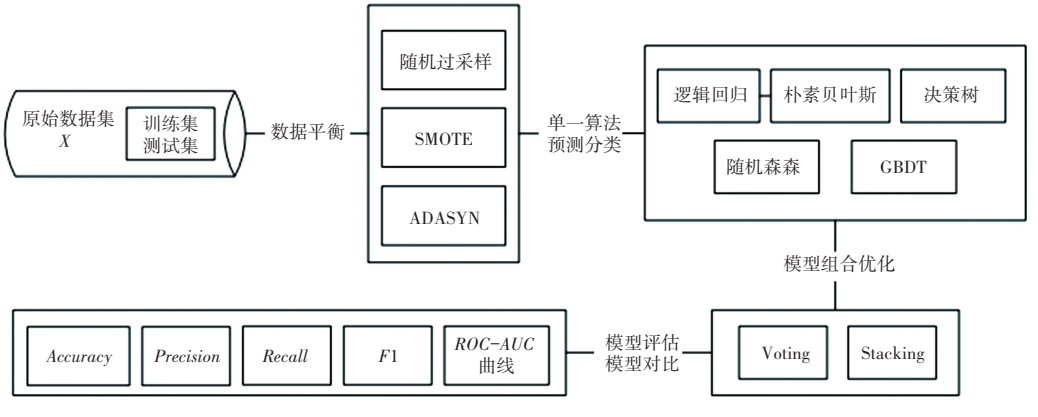

图 1　开放式创新社区用户生成内容质量组合预测模型框架

Fig. 1　A combined prediction model framework for User-Generated Content quality in open innovation communities

（1）将原始数据集划分为训练集、测试集；

（2）研究通过数据多样性（采样）和模型多样性（评估器）提高泛化能力；

（3）数据集采样和评估器组合构成多种预测模型：单一采样与单分类器、单一采样与多分类器；

（4）组合模型评估。通过模型评价指标筛选出适用于开放式创新社区高质量 UGC 的最优预测模型。

2.2　数据采样

通常情况下，模型的损失函数对所有样本权重其实是相同的。因此，数据不平衡会使得模型在少数类样本上毫无泛化性，或完全被多数类样本压制，从而学不到少数类样本的特点。因为即使模型只能识别多数类，也能使得总损失很低，而模型训练的过程仅仅是降低总损失。机器学习中对于不平衡样本集的常见采样方法一般分为过采样（Oversampling）和欠采样（Undersampling）。随机过采样是从少数类样本集中随机重复抽取得到同多数类样本量的样本；随机欠采样则相反，从多数类样本集中随机选取同少数类样本量的样本。直接随机采样虽然可以使样本数据平衡，但同时也衍生出一些其他问题。比如，过采样重复操作少数类样本，扩大了数据规模，增加了模型复杂度，容易产生过拟合问题。欠采样只选取部分多数类样本，模型只能学习到部分数据信息，降低了模型的有效性。

对于样本量偏少的情况，一般采用过采样方法。为了规避随机过采样方法（Random Oversampling，ROS）对模型产生的负面影响，采用一些新方法生成新样本。SMOTE 算法[18]（Synthetic Minority Oversampling Technique）的生成方法是从每个少数类样本最近邻中随机选取一点来生成新的少数类样本。ADASYN 算法[19]在少数类样本分布密度较低的空间中生成更多的少数类样本，而在分布密度较高的空间生成较少的少数类样本。

2.3　基分类器选取

为了保证模型组合可以产出更好的效果，构建模型多样性可以达此目的。分类器之间的差别越大，两两之间的独立性就越强。而在现实中完全独立的分类器是不存在的，因为不同算法在相同测试集进行相同预测，所以分类器不能完全独立。现今也有多种手段用来提升分类器多样性，让分类器之间相对独立，本研究聚焦在算法多样性方向。增加不同类型的算法，比如线性、概率、树、集成模型组合。基于以上考虑，研究选取的机器学习算法有逻辑回归（Logistic Regression，LR）、朴素贝叶斯（Naive Bayes，NB）、决策树（Decision Tree，DT）、随机森林

(Random Forest, RF)、GBDT (Gradient Boosting Decision Tree)。

2.4 模型组合优化

常见的模型组合方法有:均值法(Averaging)、投票法(Voting)、堆叠法(Stacking)、改进堆叠法(Blending)等。其中,均值法是将每个评估器的输出结果取平均值,适用于回归问题。投票法是按每个评估器的输出结果进行投票,适用于分类问题。堆叠法使用了一个或多个算法预测输出的结果作为下一个算法的训练数据。而 Blending 可以看作是特殊的 Stacking。本研究属于分类问题,故选取了 Voting 和 Stacking 两种方法组合以上 5 种算法。对此可展开剖析论述如下。

(1) Voting。投票法(Voting)分为硬投票和软投票两种。硬投票(Hard Voting)根据少数服从多数的原则预测结果。软投票(Soft Voting)将所有模型预测样本为某一类别的概率平均值作为标准,概率最高对应的类型即是最终的预测结果;

(2) Stacking。Stacking 是一个分层模型集成框架。第一层输入初始训练集训练基分类器,然后以第一层基学习器的输出结果作为特征进入第二层训练。若直接用基学习器训练集来产生第二层模型的训练集,这样存在过拟合风险较大的问题。因此一般是通过使用交叉验证或留一法,用基学习器未使用的样本来产生第二次模型的训练样本。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。其中,内容特征主要是可以直接从 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。

2.5 模型评估

为了更好地比较模型表现好坏,本文选取多个指标进行模型评估。混淆矩阵见表 1。

表 1 混淆矩阵

Table 1 Confusion matrix

真实类别	预测结果	
	正	负
正	*TP*	*TN*
负	*FP*	*FN*

3 实验与结果

3.1 实验数据

本研究从腾讯云社区(https://cloud.tencent.com/developer/ask)爬取实验相关数据,用于训练开放式创新社区 UGC 质量预测模型。该数据集包含了腾讯云社区的用户、内容、社区三大维度中 31 个特征指标见表 2。运用 selenium 库自动爬取腾讯云社区 2021 年 11 月 26 日~2021 年 12 月 3 日内历史数据,获取 5 755 位用户信息,33 944 条用户生成内容,剔除信息缺失严重样本,最后得到维度为(25 423,31)的特征矩阵。

表 2 开放式创新社区用户生成内容质量评价体系

Table 2 Quality evaluation system for User-Generated Content in open innovation communities

维度	特征指标	特征描述
个体	*user_level*	用户等级
	user_skill_num	用户技能数
	self_information	用户个人信息完成度
	user_ follow	用户关注数
	user_ followed	用户粉丝数
	user_rank	用户排名
	user_like	用户获得点赞数
	user_read	用户发布内容的阅读总量
	submit_ugc	用户发布所有内容的数量
内容	*ugc_words*	内容字数
	img_in_ugc	内容图片数
	url_in_ugc	内容链接数
	code_in_ugc	内容中提供的代码量
	ugc_like	内容点赞数
	ugc_mark	内容收藏数
	ugc_comment	内容评论数
	skill_tag	内容专业性
	ugc_order	内容在话题下展示的排序
	topic_title_words	话题标题的字数
	topic_tag_num	话题所属范围大小
	topic_ugc_num	话题下所有的内容总数
	topic_ follow	话题关注度
	topic_read	话题阅读量
	topic_hotpoints	话题热度
	topic_words	话题详情字数
	img_in_topic	话题介绍中包含的图片数量
	url_in_topic	话题介绍中的链接个数
社区	*user_verified*	用户是否通过社区认证
	community_commend	受到社区推荐次数
	user_honor	社区额外奖励荣誉次数
	first_commend	社区首位推荐次数

本研究采用“是否受到推荐或采纳”的标准评

判用户生成内容质量的优劣。若此条内容被推荐且被采纳，标签为 2，样本数量为 493；若此条内容被推荐或被采纳，标签为 1，样本数量为 785；若未被推荐或采纳，标签则为 0，样本数量为24 145。进一步得出，该实验数据正负样本比约为 2∶3∶100。

3.2　实验设计

为了评估分类器和组合方法的预测效果，进行了以下 3 组对照试验。

(1)比较不同采样方法下单一分类模型预测效果。对原始数据集使用不同的采样方法进一步得到随机过采样数据集、SOMTE 采样数据集、ADASYN 采样数据集，分别结合 5 个分类预测模型(SVM、NB、DT、RF、LGBM)，研究使用不同采样方法不同分类模型的预测效果；

(2)比较不同采样方法与单一分类模型组合预测效果。3 种采样方法(随机过采样法、SMOTE、ADASYN)和 5 种分类模型一一组合得到 15 种预测模型(ROS-LR、ROS-NB、ROS-DT、ROS-RF、ROS-GBDT、SMOTE - LR、SMOTE - NB、SMOTE - DT、SMOTE - RF、SMOTE - GBDT、ADASYN - LR、ADASYN-NB、ADASYN - DT、ADASYN - RF、ADASYN-GBDT)，研究不同采样方法与不同分类模型组合分类预测效果；

(3)比较不同采样方法下多分类模型组合预测效果。分别采用软投票、硬投票、Stacking 组合 5 种分类模型，研究使用不同采样方法不同组合分类模型的预测效果。

综合比较以上各类模型的预测效果，能够更好地预测开放式创新社区用户生成内容质量。

3.3　实验环境

本研究的实验环境：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。

3.4　结果分析

在第 1 组实验中，结果见表 3，使用采样方法明显地提高了分类模型预测效果，而树模型和集成模型在分类学习过程中表现更好，更适用于研究开放式创新社区用户生成内容质量问题。不论数据集是否使用采样方法，随机森林分类模型在 5 类评估指标中总体优于其他 4 种分类模型。

随机过采样方法中，逻辑回归模型的 *AUC* 值提升 26.54%，朴素贝叶斯模型的 *AUC* 值提升了 3.89%，决策树模型的 *AUC* 值提升了 35.26%，随机森林模型的 *AUC* 值提升了 43.60%，GBDT 模型的 *AUC* 值提升了 26.67%。

表 3　不同采样方法下单一分类模型预测效果

Table 3　Prediction effect of single classification model under different sampling methods

采样方法	分类器	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
/	LR	0.950 4	0.485 2	0.375 4	0.394 2	0.536 2
	NB	0.843 5	0.399 0	0.457 0	0.398 1	0.624 0
	DT	0.945 9	0.604 6	**0.628 8**	0.610 2	**0.734 3**
	RF	**0.969 5**	**0.944 4**	0.583 8	**0.678 2**	0.695 4
	GBDT	0.969 1	0.846 4	0.586 9	0.662 0	0.704 5
ROS	LR	0.570 9	0.569 9	0.571 4	0.564 1	0.678 5
	NB	0.532 3	0.646 0	0.531 3	0.503 1	0.648 3
	DT	0.991 0	0.991 2	0.990 9	0.991 0	0.993 2
	RF	**0.998 1**	**0.998 2**	**0.998 1**	**0.998 1**	**0.998 6**
	GBDT	0.856 6	0.857 0	0.856 5	0.856 2	0.892 4
SMOTE	LR	0.546 6	0.541 9	0.546 3	0.542 2	0.659 8
	NB	0.569 0	0.671 8	0.568 0	0.545 2	0.675 9
	DT	0.955 6	0.955 9	0.955 5	0.955 6	0.966 6
	RF	**0.989 9**	**0.989 9**	**0.989 8**	**0.989 8**	**0.992 4**
	GBDT	0.874 1	0.876 0	0.874 2	0.874 5	0.905 6
ADASYN	LR	0.518 9	0.517 2	0.519 2	0.509 5	0.639 3
	NB	0.545 2	0.639 0	0.544 3	0.515 9	0.658 0
	DT	0.958 7	0.958 9	0.958 6	0.958 6	0.969 0
	RF	**0.990 0**	**0.990 0**	**0.989 9**	**0.989 9**	**0.992 4**
	GBDT	0.863 0	0.866 0	0.863 1	0.863 8	0.897 2

SMOTE 采样方法中，逻辑回归模型的 *AUC* 值提升 23.05%，朴素贝叶斯模型的 *AUC* 值提升了 8.32%，决策树模型的 *AUC* 值提升了 31.64%，随机森林模型的 *AUC* 值提升了 42.71%，GBDT 模型的 *AUC* 值提升了 28.55%。

ADASYN 采样方法中，逻辑回归模型的 *AUC* 值提升 19.23%，朴素贝叶斯模型的 *AUC* 值提升了 5.45%，决策树模型的 *AUC* 值提升了 31.96%，随机森林模型的 *AUC* 值提升了 42.71%，GBDT 模型的 *AUC* 值提升了 27.35%。

在第 2 组实验中，结果见表 4。使用随机过采样的随机森林分类模型(ROS-RF)在 5 类评估指标中均计算得出最优值，是不同采样方法与单一分类模型组合的最佳模型。该模型比未采样随机森林模型分类准确率提高 2.95%，*Precision* 值提高 5.69%，*Recall* 值提高 70.97%，*F*1 值提高 41.93%，*AUC* 值提高 43.60%。从准确率来看，3 种采样方法降低了 LR 和 NB 模型评估 *Accuracy* 值，准确率在 50% ~

60%之间,GBDT 模型预测准确率下降 10%左右,而在 DT 和 RF 模型中提升幅度在 2%~4%左右。从 *AUC* 值指标看,经过采样处理数据集和不同分类模型组合能达到 3%~40%的提升幅度。不论采用何种采样方法,在随机森林分类模型上的提升效果最明显。从 *F*1 指标来看,在未采样情况下,各类分类模型 *F*1 值的提升在 40%~68%之间,而采用采样方法可以将 *F*1 值提高到 95%以上,增幅在 30%~60%上下。由此可见,使用采样方法可以解决开放式创新社区用户生成内容数据集不平衡的问题,采样处理后的组合模型也能够获得更好的针对开放式创新社区高质量用户生成内容的预测性能,大幅度地提升了模型预测效果。

表 4 混合组合分类模型预测效果

Table 4 Prediction performance of mixed combination classification model

采样方法	组合方法	*Accuracy*	*Precision*	*Recall*	*F*1	*AUC*
/	Soft Voting	0.963 0	0.737 8	0.595 8	0.650 3	0.707 3
	Hard Voting	0.962 2	0.733 1	0.586 4	0.641 3	0.699 3
	Stacking	**0.971 7**	**0.881 2**	**0.632 6**	**0.703 1**	**0.739 9**
ROS	Soft Voting	0.987 0	0.987 4	0.986 8	0.986 9	0.990 2
	Hard Voting	0.935 9	0.942 4	0.935 5	0.936 4	0.951 7
	Stacking	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
SMOTE	Soft Voting	0.961 3	0.962 8	0.961 0	0.961 3	0.970 8
	Hard Voting	0.924 6	0.931 5	0.924 3	0.925 4	0.943 3
	Stacking	**0.992 3**	**0.992 3**	**0.992 3**	**0.992 3**	**0.994 2**
ADASYN	Soft Voting	0.962 9	0.964 5	0.962 5	0.962 9	0.972 0
	Hard Voting	0.917 4	0.926 0	0.917 0	0.918 4	0.937 8
	Stacking	**0.992 2**	**0.992 2**	**0.992 2**	**0.992 2**	**0.994 0**

在第 3 组实验中,在使用 Stacking 组合优化策略下,混合模型的 *Accurary*、*Precision*、*Recall*、*F*1、*AUC* 值达到了最高值,模型几乎接近正确预测。在 3 种不同采样 Stacking 组合的情况下,准确率相比简单软投票法分别提升 1.31% 、3.23%、3.05%,相对于硬投票法分别提升约 6.84%、7.32%、8.16%;*F*1 值相比简单软投票法分别提升 1.32% 、3.32%、3.05%,相对于硬投票法分别提升约 6.79%、7.23%、8.03 %; *AUC* 值相比简单软投票法则分别提升 0.99%、2.41%、32.29%,相对于硬投票法分别提升约 5.08%、5.41%、6.01%。相较于单一模型最优结果(第 1 组实验无采样处理结果),*Auccrary*、*F*1 值和 *AUC* 分别平均提升了 3.85%、28.18%、12.30%;相较于其他组合模型最优结果(第 2 组实验),*Auccrary*、*F*1 值和 *AUC* 分别平均提升了 0.19%、0.19%、0.14%。此外,混合模型 *AUC* 指标约为 1,高于第 1、2 组实验中所有模型的 *AUC* 取值。说明结合 Stacking 组合优化策略下的混合模型具有优异的高质量 UGC 预测性能,并且 *AUC* 和 *F*1 指标的提升说明 Stacking 组合优化策略下的混合模型更能够有效地预测高质量用户生成内容,这也是建立开放式创新社区用户生成内容预测模型的核心目标。

4 结束语

本研究的目的是解决开放式创新社区内容冗余导致高质量用户生成内容无法充分发挥其价值的问题,提出了一个基于机器学习混合模型的优化方法,用于高质量用户生成内容预测,主要思想是使用 Stacking 算法将数据采样技术和机器学习预测模型进行策略组合优化,得到了开放式创新社区用户生成内容质量的较好预测性能。本文的主要意义如下:

(1)从用户个体、内容、社区三个维度更加全面地构建了开放式创新社区 UGC 质量特征矩阵,囊括了用户在社区已有以及衍生统计特征。使用真实的腾讯云社区用户生成内容数据集进行实验验证,更具说服力地为企业开放式创新社区提供有价值的参考思路;

(2)本研究结果显示,混合模型—随机过采样 Stacking 在开放式创新社区高质量用户生成内容预测结果取得了最优表现,能够有效提升开放式创新社区高质量用户生成内容的预测精度和总体性能。由于 UGC 数据的不平衡问题,如果仅仅按照以往在高质量用户生成内容预测研究上简单使用机器学习模型,将会出现预测精度偏低的结果。所以本文针对采取了在数据、模型两个层面的组合方法以提高高质量用户生成内容的预测性能。在数据层面,选择随机过采样、SMOTE 和 ADASYN 三种采样方法,以降低原始数据的样本不平衡性;在模型层面,基于机器学习组合优化考虑多种模型策略组合,进一步优化组合权重。结果显示,使用采样方法 Stacking 混合机器学习预测模型可以明显地解决数据不平衡性带来的问题;

(3)本文提出的针对高质量用户生成内容具有较高预测精度的模型具有实践价值,研究思路对开放式创新社区 UGC 研究有一定的参考意义。建立高质量用户生成内容预测模型的目标是识别具有价值的用户生成内容以及 UGC 所表现出的特征,进一步更好地运营开放式创新社区、生产高质量内容、帮

助企业提高创新力。

虽然本文在以往用户生成内容质量研究效果有一定的提升，但尚有不足之处。一是在方法选择的主观性，在选取采样方法和单个机器学习分类模型的时候偏向学习经验的主观意识。后面可以考虑数据预处理，采用机器学习中特征选择的思想，比如经常使用的 Filter 和 Wrapper 方法；二是开放式创新社区高质量用户生成内容数据量偏少，缺少开放式创新社区线上评估，这些问题都会影响整体模型的泛化能力。后续研究需在更大数据层面展开更加深入的研究，以进一步验证模型。

参考文献

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

[2] DIGANGI P M, WSAKO M. Steal my idea! Organizational adoption of user innovations from a user innovation community: A case study of Dell IdeaStorm[J]. Decision Support Systems, 2010, 48(1):303-312.

[3] PAJO S,VERHAEGEN P A, VANDEVENNE D, et al. Fast lead user identification framework[J]. Procedia Engineering, 2015, 131(Complete):1140-1145.

[4] YANG Jingjing. 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.

[5] 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.

[6] WU Bing, GONG Chunyu. Impact of open innovation communities on enterprise innovation performance: A system dynamics perspective [J]. Sustainability,2019,11(17):4794.

[7] 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.

[8] 张宁,赵文斐,庞智亮,等. 企业开放式创新社区创意采纳影响因素研究—价值共创视角[J]. 科技进步与对策,2021,38(16):91-100.

[9] 吉海颖,戚桂杰,梁乙凯. 行动比声音更有力量吗？—开放式创新社区用户交互与用户创意更新持续贡献行为研究[J]. 管理评论,2022,34(4):80-89.

[10] 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.

[11] SAHU T P, NAGWANI N K, VERMA S. Topical authoritative answerer identification on Q&A posts using supervised learning in CQA sites [C]//Proceedings of the 9th Annual ACM India Conference. India:ACM,2016:129-132.

[12] LI Lei, HE Daqing, JENG W, et al. Answer quality characteristics and prediction on an academic Q&A site: A case study on ResearchGate [C]// Proceedings of the 24th International Conference on World Wide Web (WWW'15 Companion). Association for Computing Machinery. New York, USA:ACM, 2015:1453-1458.

[13] 金燕,孙佳佳. 基于用户画像的 UGC 质量预判模型[J]. 情报理论与实践,2019,42(10):77-83.

[14] 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.

[15] LEI Li. LIN Xin, ZHAI Yue, et al. User communities and contents co-ranking for user-generated content quality evaluation in social networks[J]. International Journal of Communication Systems,2016,29(14):2147-2168.

[16] 王伟,冀宇强,王洪伟,等. 中文问答社区答案质量的评价研究：以知乎为例[J]. 图书情报工作,2017,61(22):36-44.

[17] 李昂,韩萌,穆栋梁,等. 多类不平衡数据分类方法综述[J]. 计算机应用研究,2022,39(12):3534-3545.

[18] BATISTA G, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1):20-29.

[19] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic Minority Over-sampling Technique [J]. Journal of Artificial Intelligence Research,2002, 16(1):321-357.

[20] HE H, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//2008 IEEE International Joint Conference on Neural Networks. IEEE World Congress on Computational Intelligence. Hong Kong: IEEE, 2008: 1322-1328.