

# 基于机器学习的开放创新社区创意识别方法研究

薛水晶<sup>1</sup>, 郭伟<sup>1,2</sup>, 连文涛<sup>1</sup>, 张静<sup>2</sup>

(1.天津大学 装备设计与制造技术天津市重点实验室, 天津 300350;

2.天津仁爱学院, 天津 301636)

**摘要:** **目的** 探索基于机器学习的开放创新创意识别方法, 解决创意识别过程中存在的耗时长、效率低、成本高等问题。**方法** 从用户特征、用户参与度和创意内容特征三个方面构建评估模型, 以 OpenIDEO 社区为研究对象, 采集数据并进行数据清洗和数据转化映射, 最后进行多种机器学习算法的参数优化, 并以  $F1$  值为选择标准, 选择分类效果最佳的算法作为分类模型。**结果** 运用 KNN、SVM、决策树、随机森林四种机器学习算法分析 OpenIDEO 数据, 随机森林算法通过参数优化取得了最大的  $F1$  值(0.919 09), 同时对于验证数据, 该算法同样可以取得较好的分类效果。**结论** 应用机器学习方法对开放式创新社区中的创意进行识别, 具有较高的可行性和有效性, 可以大大降低社区在创意筛选中的投入, 提高创新效率, 优化社区生态。

**关键词:** 开放创新; 创新社区; 创意识别; 机器学习

**中图分类号:** TB472 **文献标识码:** A **文章编号:** 1001-3563(2023)08-0072-09

**DOI:** 10.19554/j.cnki.1001-3563.2023.08.007

## Machine Learning-based Method of Creative Identification in Open Innovation Community

XUE Shui-jing<sup>1</sup>, GUO Wei<sup>1,2</sup>, LIAN Wen-tao<sup>1</sup>, ZHANG Jing<sup>2</sup>

(1.Tianjin Key Laboratory of Equipment Design and Manufacturing Technology, Tianjin University,  
Tianjin 300350, China; 2.Tianjin Ren'ai College, Tianjin 301636, China)

**ABSTRACT:** The work aims to explore a machine learning-based method of creative identification in open innovation to solve a number of problems in creative identification, including long time consumption, low efficiency and high costs. An evaluation model was built at first in light of user characteristics, degree of user participation and creative content characteristics, followed by focusing research on OpenIDEO Community, collecting data, conducting data cleansing and data conversion mapping. Finally, parameter optimization of multiple machine learning algorithms was made with  $F1$  value as selection criterion to select the algorithm with the best classification effect as classification model. Four machine learning algorithms, including KNN, SVM, decision tree and random forest, were adopted to analyze OpenIDEO data, among which the random forest algorithm generated the maximum  $F1$  value (0.919 09) through parameter optimization. For data validation, such algorithm could also produce good classification effects. The adoption of machine learning method is highly feasible and effective for creative identification in open innovation community. It can greatly reduce community's input in creative screening, improve innovation efficiency and optimize community ecology.

**KEY WORDS:** open innovation; innovation community; creative identification; machine learning

收稿日期: 2022-11-26

基金项目: 天津市教委科研计划项目 (2021KJ082)

作者简介: 薛水晶 (1994—), 女, 博士生, 主要研究方向为开放创新社区群智涌现机理及度量方法。

通信作者: 张静 (1984—), 女, 博士生, 副教授, 主要研究方向为开放创新社区用户需求挖掘与创意识别方法。

在互联网发达的知识经济时代,“引入外部创新能力,充分利用外部及互联网用户资源进行创新”的开放式创新(Open Innovation)模式<sup>[1]</sup>,不断融合新的技术与的形式,逐渐成为企业创新的主导模式。目前,国内外众多知名企业基于互联网平台构建开放创新社区,聚集广泛的群体智能资源,通过“用户评论”“设计大赛”等形式聚集信息、知识,并通过各种激励方式推动用户创新<sup>[2]</sup>。开放创新社区中有价值的用户创意能被企业识别、采纳,甚至转换为成功的商业产品,能有效降低企业的研发成本,增强企业的创新能力。因此,如何从大量用户创意中高效准确地识别出潜在的、有价值的创意,是企业提高开放创新绩效的关键。受用户创新角色的随机性与不确定性、用户创新行为的复杂性与动态性、用户创意产出内容的丰富性与多样性等的影响,大多数开放创新社区中的用户创意思识与价值评估存在耗时长、效率低、无明确评估审核标准等<sup>[3]</sup>问题。为解决这一问题,本文基于用户特征、用户参与度和创意内容特征,构建创意思识识别指标体系,探究基于机器学习的用户创意思识方法,为企业发掘开放创新社区中潜在的有价值、高质量的创意提供借鉴。

## 1 开放创新社区创意思识与评估相关研究

开放创新社区中的群智创意思识与评估是企业解决技术难题、寻找优秀创意、发掘商业机会的关键步骤。基于开放创新社区的设计活动往往会收集到大量的、质量参差不齐的用户创意,为精准识别符合企业创新需求、契合市场发展的有价值的创意方案,开放创新社区在用户创意思识方面需要建立系统性、合理性的评估指标与审核机制。

目前,为有效识别、评估开放创新社区的用户创意,大多数学者研究开放创新社区创意采纳的影响因素。王楠等<sup>[3]</sup>基于详尽可能性模型,考虑用户社区地位、用户贡献行为、社区认可度和创意内容质量,构建了用户创新社区四维度创意采纳模型,分析了各维度中不同因素对创意采纳的影响作用;刘晴晴<sup>[4]</sup>从信息源可信度理论和信息论证质量的视角,基于开放创新社区 Salesforce Success Community 的用户及创意数据,研究创意评论数、创意评论内容、创意内容长度、创意表达方式等对创意采纳的影响;Hossain 等<sup>[5]</sup>通过分析戴尔 Idea Storm 平台的公开数据,研究创意发布时间、创意投票、评论内容、想法总数等属性对创意采纳的影响,发现提交的想法总数、创意评论、创意收到的投票数与创意实现显著正相关,有价值创意的方案可能会藏在创意评论中;Li 等<sup>[6]</sup>构建了用户创意被采纳、实施的影响因素模型,并得出各因素对用户创意实施的影响程度。也有一些学者研究用户创意的价值及其影响因素,如王楠等<sup>[7]</sup>基于创新重组理论,运用多群组结构方程模型与同感评估技术,以小

米、华为、苹果员工及小米 MIUI 社区用户、华为社区、苹果社区用户数据为例,研究分析内、外部领先用户创意价值的差异;郭伟等<sup>[8]</sup>基于需求满足理论,以 LocalMotors 平台创新活动为例,研究了用户反馈的数量、质量和情感对用户个体创新贡献的影响,得出了各因素对创新贡献的影响方向与影响程度,对用户创新价值具有指导作用。

现有研究较少关注如何提高创意方案识别与评估的效率,尚未深入研究如何从大规模创意设计方案中挖掘出有潜在价值创意的方法。

## 2 开放创新社区创意思识与评估模型

### 2.1 创意思识与评估模型构建

开放式创新的本质是获取与整合企业内、外部知识以提高创新绩效<sup>[9]</sup>,群智创意思识与评估实质上是对创意知识的识别与评估。从组织知识创新的角度来分析,创新参与用户是知识创新的主体,用户创新行为反映了知识交互、转移、吸收、整合与创新的过程,知识创新的结果以创意内容体现出来,故认为创新参与用户、用户创新参与行为及知识创新产出内容是影响创意质量和价值的关键因素。由此,基于现有的文献研究基础,综合考虑创意产生的过程与影响创意质量及价值的关键因素,建立了一个包括用户特征、用户过程参与度和创意内容特征三个维度的创意思识与评估模型,见图 1。

### 2.2 评价指标及解析

根据以上模型,构建了开放创新社区创意设计方案的评价指标体系,详细的指标及解析如下。

#### 2.2.1 用户特征指标

##### 2.2.1.1 用户等级、特征标签与用户评分

用户标签是对用户信息和特征的抽象概况,大多数开放创新社区会综合考虑社区用户的知识能力、行为习惯、社交情况等,结合用户参与创新活动的历史行为数据,包括提交创意数量、获奖创意数量、评论数量等,从多个维度对用户的各种能力与行为进行综合评估,为用户划分等级,并标记特征标签,甚至对用户综合能力进行评分。用户等级、特征标签与用户评分可以反映用户创新能力,创新能力突出的用户能更快地识别创新需求,更有效地提出有价值的创意<sup>[10]</sup>,且创意被采纳的机会更大。因此,用户等级、特征标签与用户评分等在一定程度上可以反映创意的潜在价值以及被采纳的可能性。

##### 2.2.1.2 用户提交创意数量

Hossain 等<sup>[5]</sup>在研究中发现,用户提交的创意越多,其创意被采纳的可能性就越大;张静等<sup>[11]</sup>在研究中发现,用户在创新社区中发布的创意数量越多,社区中设计知识认知扩散的广度越大,可以促进社区用

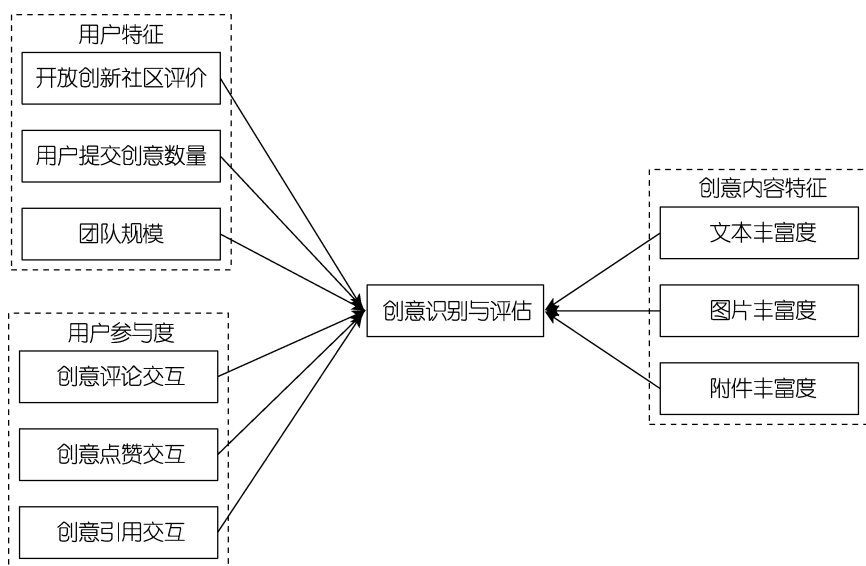


图 1 开放创新社区创意识别与评估模型

Fig.1 Creative identification and evaluation model of open innovation community

户快速形成设计认知,进一步提高用户的创新质量。此外,用户在开放创新社区中通过提交创意分享知识,为社区成员间的知识交流提供了情境空间,有利于用户间的知识交互,可以促进社区用户间的非正式学习,帮助用户形成多元化知识体系<sup>[12]</sup>,进而提高用户的创新能力。因此,用户提交创意数量可以反映用户创新活跃度与创新状态,间接影响用户创意产出的质量与价值。

#### 2.2.1.3 团队规模

开放创新社区基于互联网的建立,在线用户遵循一定的群体结构和运行机制,自愿贡献体现个体智能的可衡量产出物,并通过群智协同产出高质量的知识成果或产品<sup>[13]</sup>。故开放创新社区的创新过程,既包含体现个体智慧的创新认知、推理与决策行为,又包含体现创新团队的群智协同与知识交互行为<sup>[14]</sup>。陈文春等<sup>[15]</sup>对创新团队的异质性进行研究,指出知识异质性高的创新团队认知结构更多样化,知识技能的多样性更强,成员之间提出的创意想法越丰富,更有利于促进团队的学习行为。团队规模是影响团队结构的重要变量<sup>[16]</sup>,团队规模过大或过小都会降低创新绩效;在规模较大的团队中,成员组成更多样化<sup>[17]</sup>,团队知识、技能、经验更多样,会影响团队的学习能力,激发创新灵感,增强团队创造力<sup>[15,18-19]</sup>。因此,在开放式创新社区中,团队规模是观测社区用户创新状态的重要指标。

### 2.2.2 用户参与度指标

#### 2.2.2.1 创意评论交互

常见的开放创新社区会将征集到的创意对所有用户可见,因此,在线用户会浏览大量其他人的创意,并选择性地发表评论。根据资源基础观,当开放创新社区中的在线用户需要从其他用户那里获得资源或者他们具有其他用户所需的资源时,便产生了评论互

动。此时不同的知识源聚集在评论区而产生大量的相关知识,既可以深化用户对创新活动的认识,又可以促进创意的完善与改进<sup>[20]</sup>。同时,一个创意收集到的评论越多,其获得完善与改进的机会就越大,并且越有可能引起社区的关注<sup>[21]</sup>,从而增加创意被采纳的机会。此外,创意的评论也会帮助社区创意评估者理解创意内容,这也有助于社区评估者发现创意的价值<sup>[22]</sup>。因此,用户的创意评论交互会对创意的价值产生直接的影响。

#### 2.2.2.2 创意点赞交互

在线用户往往会根据自己的兴趣、知识背景等对发表在开放创新社区上的创意形成自己的理解与看法,并在社区上进行交互反馈,为创意“点赞”就是最普遍的一种交互方式。“点赞数”在一定程度上反映了创意的受关注度与认可度,而一个创意的认可度通常被认为是它在未来市场中被潜在接受的指标<sup>[23]</sup>。因此,创意是否具有潜在价值、是否被市场认可,在一定程度上可以用创意所获得的点赞数量来衡量,优质创意往往能够获得较多的点赞。

#### 2.2.2.3 创意引用交互

在开放创新社区中,在线用户可以浏览已经发布的创意,并可以学习、借鉴其他用户已发布的内容进行创新。在一些开放创新社区中,用户在其他创意的基础上进行二次创新时,会被要求标注出对其他创意的引用借鉴情况,可以帮助社区创意评估者更多地了解该创意对其他创意的参考情况。知识创新是量变到质变的过程<sup>[24]</sup>,创意的引用可以反映用户在创新时知识的积累情况,用户知识积累得越多越有可能创造出更有价值的创意。

### 2.2.3 创意内容特征指标

#### 2.2.3.1 文本丰富度

研究表明,文本越长,信息量越大,所包含的有

用信息越多, 所表达的意义越完整<sup>[25]</sup>。在开放式创新社区中, 用户发布的创意文本内容越丰富, 有用性可能越高, 越有可能成为潜在的有价值创意。本文认为, 文本长度是文本丰富度的重要体现, 故将文本长度作为文本丰富度的评价依据。

### 2.2.3.2 图片丰富度

Li 等<sup>[6]</sup>认为, 在创意想法的描述中添加图片, 可以增加创意的说服力。张艳辉等<sup>[26]</sup>认为, 通过上传图片, 可以增加信息的丰富度, 增强内容的有用性。大多数开放创新社区允许用户上传的创意中包含图片, 本文认为在创意中上传的图片丰富度越高, 反馈的有用信息越多, 创意的有用性越高, 本文将图片的数量作为评价图片丰富度的重要依据。

### 2.2.3.3 附件丰富度

除了文本和图片之外, 有些创意还有可能包含其他附件, 如 PDF 文件、PPT 文件等, 再次对创意进行补充说明。相关研究表明, 创意内容的质量可以通过创意链接的附件数量来衡量<sup>[6]</sup>。因此, 附件丰富度可以作为创意内容丰富度的一项评价指标, 高质量的、有潜在价值的创意往往具有较高丰富度的附件。

## 3 研究设计

### 3.1 研究对象与样本数据集描述

随着“互联网+”的不断推进, 开放创新平台得到了大力发展, 越来越多的企业借助互联网创新社区

开放创新流程, 利用互联网外部资源提升企业创新水平。成立于 2010 年的 OpenIDEO 开放创新平台, 目前已经收集到了超过 20 000 个想法, 解决了 70 余个社会热点问题。OpenIDEO 平台主要的创新活动是挑战赛 (Challenge), 平台收集了来自世界各地社区组织或企业的挑战项目, 以挑战赛方式展示给社区用户, 用户接受挑战并进行交互反馈, 提出解决项目问题的创意方案。OpenIDEO 平台会对收集到的创意进行多次筛选, 选择出最有价值的创意, 并在创新社区进行推广或实施。本文以 OpenIDEO 平台在 2018 年发布的“2018 BRIDGE BUILDER CHALLENGE”挑战赛为研究对象, 探究开放创新社区的创意思识别方法。在该挑战赛中, OpenIDEO 社区用户共提交了 685 个创意, 其中有 55 个创意被评定为“潜在有价值创意”。在 OpenIDEO 官方网站上, 用户可以查看每一条创意的详细内容以及创意发布者的个人信息。

依据本文构建的创意思识别与评估模型, 利用 Python 网络爬虫获取每个创意的详细内容及其发布者的个人信息, 并将其存储在 MySQL 数据库的两张表中, 其中名为“user”的表存放用户的个人信息, 见图 2; 名为“idea\_content”的表存放创意的相关信息, 见图 3。

### 3.2 模型变量与评价依据

对于用户特征, 获取“总分数”(Total Score, 平

| name               | teams | point | contribution | url                                                                 |
|--------------------|-------|-------|--------------|---------------------------------------------------------------------|
| Roshni Rides       | 0     | 44    | 1            | https://challenges.openideo.com/profiles/58f2f78bb5b4bala1574aa0b14 |
| robert hakiza      | 0     | 1,016 | 7            | https://challenges.openideo.com/profiles/1124643841150291806291     |
| Agueda Gon         | 0     | 19    | 1            | https://challenges.openideo.com/profiles/5ale39a044d142e4a482122615 |
| Nadia Naviwala     | 0     | 41    | 1            | https://challenges.openideo.com/profiles/grantsteam                 |
| PYXERA Global      | 0     | 20    | 1            | https://challenges.openideo.com/profiles/5b15aalab226a3256240ef2815 |
| Amelia Phillips    | 0     | 64    | 4            | https://challenges.openideo.com/profiles/10412538822092330200114756 |
| Demotech Design    | 0     | 17    | 1            | https://challenges.openideo.com/profiles/11201487046300036565115281 |
| Anil Pattni        | 0     | 23    | 1            | https://challenges.openideo.com/profiles/10909121203300315506315063 |
| Beta Hatch         | 3     | 52    | 1            | https://challenges.openideo.com/profiles/5b0dc5d178621071c17c101615 |
| Marco Antonio Roch | 0     | 15    | 1            | https://challenges.openideo.com/profiles/1021174757742067152709593  |
| Ryan Schlieff      | 0     | 32    | 1            | https://challenges.openideo.com/profiles/ryan3                      |
| Tara               | 2     | 14    | 1            | https://challenges.openideo.com/profiles/5b15832facf9fa50306ee37415 |
| David Ezra Jay     | 2     | 177   | 2            | https://challenges.openideo.com/profiles/10880190207635327123615251 |
| Celine Steer       | 7     | 45    | 1            | https://challenges.openideo.com/profiles/11727289405068347405415276 |
| Riya Bhartia       | 0     | 16    | 1            | https://challenges.openideo.com/profiles/10156289976623771152763815 |
| Valerie            | 0     | 47    | 1            | https://challenges.openideo.com/profiles/c29e3d9d91594a40931c3bca9e |
| Kate null          | 0     | 17    | 1            | https://challenges.openideo.com/profiles/67flac961f374c8ab37e48cff8 |
| Sneha Menon        | 6     | 38    | 1            | https://challenges.openideo.com/profiles/11656477792894489640315276 |

图 2 SQLyog 可视化的“user”  
Fig.2 Visual “user” table of SQLyog

| name            | idea         | comm... | userl... | inspir... | attac... | Explain y... | Who ar...  | How is your ... | Idea Prop... | Tell ...     | Expert... |
|-----------------|--------------|---------|----------|-----------|----------|--------------|------------|-----------------|--------------|--------------|-----------|
| Roshni Rides    | Roshni Rides | 17      | 20       | 0         | 1        | Ro...        | 1K O. 302B | There...        | 762B Pr...   | 110B <.      | 1K 1.. 9B |
| robert hakiza   | Helping F    | 28      | 21       | 0         | 0        | Ar...        | 1K O. 399B | We th...        | 855B In...   | 117B Y. 328B | 7.. 8B    |
| Agueda Gon      | miDistrit    | 3       | 4        | 0         | 0        | Th...        | 1K <.      | 2K It co...     | 438B In...   | 117B W. 270B | 5.. 9B    |
| Nadia Naviwala  | The 10 tr    | 13      | 5        | 0         | 0        | Ar...        | 1K <.      | 2K Even ...     | 926B In...   | 117B T. 288B | 7.. 8B    |
| PYXERA Global   | Partners     | 2       | 6        | 0         | 0        | Th...        | 2K C. 995B | PYXER...        | 1K In...     | 117B T. 222B | 5.. 9B    |
| Amelia Phillips | Kustawi f    | 3       | 9        | 0         | 0        | Il...        | 2K <.      | 6K There...     | 886B Ma...   | 153B <.      | 1K 5.. 9B |
| Demotech Design | Water fr     | 3       | 3        | 0         | 0        | Th...        | 1K -.      | 542B New a...   | 940B Pr...   | 110B P. 334B | 7.. 8B    |
| Anil Pattni     | www.Tiny     | 11      | 9        | 0         | 0        | Bu...        | 1K C. 155B | At th...        | 684B Ea...   | 166B <.      | 1K 7.. 8B |
| Beta Hatch      | Insect f     | 25      | 13       | 0         | 0        | By...        | 1K <.      | 7K Feedi...     | 977B Ea...   | 166B <.      | 4K 7.. 8B |
| Marco Antonio   | RECIFROC     | 0       | 5        | 0         | 0        | Th...        | 1K O. 996B | The i...        | 938B In...   | 117B T. 431B | 1.. 9B    |
| Ryan Schlieff   | Involve      | 19      | 11       | 0         | 1        | Im...        | 1K <.      | 2K IAP a...     | 993B Pi...   | 86B <.       | 2K 7.. 8B |
| Tara            | Bridging     | 3       | 4        | 0         | 0        | Wh...        | 2K T. 953B | Our p...        | 996B In...   | 117B W. 495B | 7.. 8B    |
| David Ezra Jay  | GreenStar    | 56      | 12       | 1         | 0        | Ou...        | 1K <.      | 1K We ha...     | 657B Pi...   | 86B W. 267B  | 7.. 8B    |
| Celine Steer    | Good Emp     | 23      | 16       | 0         | 0        | Go...        | 1K L. 995B | While...        | 992B Pr...   | 110B <.      | 1K 7.. 8B |
| Riya Bhartia    | Cosmos C     | 7       | 4        | 0         | 2        | Co...        | 2K <.      | 5K Why t...     | 945B In...   | 117B <.      | 2K 1.. 9B |
| Valerie         | Stories      | 7       | 19       | 0         | 0        | St...        | 1K <.      | 2K We ar...     | 672B Pr...   | 110B S. 247B | 1.. 9B    |

图 3 SQLyog 可视化的“idea\_content”  
Fig.3 Visual “idea\_content” table of SQLyog

台从研究、想法、评价、协作四个角度对用户进行评分的总和)作为用户等级与特征的综合评分;在个人信息页面获取“最新贡献”(Recent Contributions, 用户最近在平台上发表的创意数量)直接表征用户提交的创意数量;获取创意发布者所在团队的成员数量作为团队规模的度量指标。

对用户过程参与度而言,在创意详情页面获取创意获得的点赞数量来衡量点赞交互情况;获取创意评论数来衡量评论交互情况;获取该创意引用他人创意数量反映创意的引用交互情况。

对创意内容特征而言,在创意详情页面获取创意内容,OpenIDEO 社区的创意内容能以问答的方式结构化呈现出来,因此可以利用 Python 来抓取创意内容。本研究共抓取到 15 条问答,用户的“回答”内

容即创意内容,对“回答”文本内容进行分词,统计文本长度并求和,得到用户创意文本总长度,以此作为创意文本丰富度的衡量标准;在创意内容中,某些用户加入了图片用来对创意进行进一步的解释,通过统计创意内容对应的网页源码中名为“img”的 HTML 标签出现的次数,得到创意中图片的数量,并以此来衡量创意图片丰富度;此外,部分创意还带有附件(PDF 文件、PPT、图片等)对创意进行补充,在创意详情页获取创意附件数量,以此来衡量创意附件丰富度。

综上所述,得到如表 1 所示的开放创新社区潜在有价值创意的评价依据。

将爬取的数据按照评价依据全部数值化,得到表 2 所示的信息。

对表 2 的数据进行描述性统计,见表 3。

表 1 模型变量及评价依据  
Tab.1 Model variables and evaluation basis

| 变量类型 | 一级指标    | 二级指标               | 评价依据                |
|------|---------|--------------------|---------------------|
| 自变量  | 用户特征    | 用户等级、特征标签与用户评分(a1) | OpenIDEO 平台对用户的综合评分 |
|      |         | 用户提交创意数量(a2)       | 用户在平台提交的创意总数        |
|      |         | 团队规模(a3)           | 用户所在团队人数            |
|      | 用户过程参与度 | 创意评论交互(b1)         | 创意获得的评论数量           |
|      |         | 创意点赞交互(b2)         | 创意获得的点赞数量           |
|      |         | 创意引用交互(b3)         | 引用他人创意的数量           |
|      | 创意内容特征  | 创意文本丰富度(c1)        | 创意文本长度              |
|      |         | 创意图片丰富度(c2)        | 创意图片数量              |
|      |         | 创意附件丰富度(c3)        | 创意附件数量              |
| 因变量  |         | 创意状态(d1)           | 是否为潜在有价值创意          |

表 2 数据数值化后的信息  
Tab.2 Information after data digitization

| 编号  | 用户名称     | 用户特征 |     |     | 用户过程参与度 |     |     | 创意内容特征 |     |     | 是否为潜在有价值创意(d1) |
|-----|----------|------|-----|-----|---------|-----|-----|--------|-----|-----|----------------|
|     |          | a1   | a2  | a3  | b1      | b2  | b3  | c1     | c2  | c3  |                |
| 1   | Roshni   | 44   | 1   | 0   | 17      | 20  | 0   | 971    | 6   | 1   | 0(否)           |
| 2   | Vidhya   | 103  | 1   | 7   | 25      | 14  | 10  | 1 146  | 41  | 0   | 1(是)           |
| 3   | Valerie  | 47   | 1   | 0   | 7       | 19  | 0   | 1 105  | 3   | 0   | 0(否)           |
| ... | ...      | ...  | ... | ... | ...     | ... | ... | ...    | ... | ... | ...            |
| 685 | Chidozie | 19   | 1   | 0   | 4       | 5   | 0   | 234    | 0   | 0   | 0(否)           |

表 3 变量的统计性描述(N=685)  
Tab.3 Statistical description of variables (N=685)

| 统计变量         | Min | Max   | Mean   | Std.dev |
|--------------|-----|-------|--------|---------|
| 开放创新社区评价(a1) | 0   | 7 920 | 89.51  | 354.258 |
| 用户提交创意数量(a2) | 1   | 53    | 1.91   | 3.973   |
| 团队规模(a3)     | 0   | 21    | 1.09   | 2.379   |
| 创意评论交互(b1)   | 0   | 172   | 8.36   | 13.654  |
| 创意点赞交互(b2)   | 0   | 125   | 6.07   | 7.667   |
| 创意引用交互(b3)   | 0   | 44    | 0.42   | 2.812   |
| 创意文本丰富度(c1)  | 15  | 2 426 | 986.67 | 307.714 |
| 创意图片丰富度(c2)  | 0   | 67    | 6.02   | 8.716   |
| 创意附件丰富度(c3)  | 0   | 9     | 0.77   | 1.307   |



4 研究实验与结果

针对 OpenIDEO 社区用户创意数据, 本文分别运用 KNN、SVM、决策树、随机森林四种机器学习算法对其进行分类, 并通过调节参数, 寻找可以识别高质量的、潜在有价值创意的最优分类算法。

4.1 潜在有价值创意思识别算法对比研究

4.1.1 KNN 参数优化研究

KNN (K-Nearest Neighbor) 算法是机器学习算法中比较基础、简单的算法之一, 它既能用于分类, 也能用于回归。其原理是通过测量不同特征值之间的距离来进行分类。

影响 KNN 算法分类效果的三个基本要素包括  $K$  值、距离度量方法以及分类决策规则, 其中最主要的是  $K$  的取值, 因此本文将基于 Minkowsik 距离和多数表决的分类规则研究  $K$  值变化对分类结果的影响。

图 4 中的多项式拟合曲线表明,  $F1$  值随  $K$  值的增大先上升、后下降, 下降的原因可能是  $K$  值变大, 增大了模型的复杂度, 进而出现过拟合现象。结合实际的分类准确率数值, 当  $K$  取 5 时, 分类结果的  $F1$  值达到最大 0.879 59。

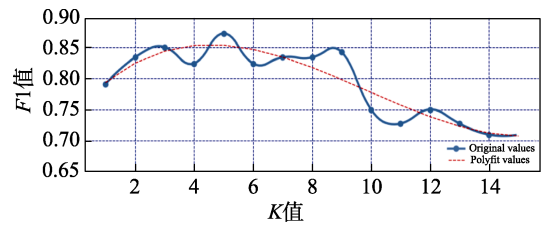


图 4 分类准确率随  $K$  值的变化  
Fig.4 Change of classification accuracy with  $K$  value

4.1.2 决策树参数优化研究

决策树 (Decision Tree) 是一种基本的分类与回归方法。它预测速度快, 可以处理类别型 (离散型) 数据和数值型 (连续型) 数据。决策树将数据的不同特征作为分类依据, 决策树的学习通常包含特征选择、决策树的生成和决策树的剪枝三个步骤。

决策树算法的典型调节参数包括: 决策树的最大深度 (Max Depth)、拆分内部节点所需的最小样本数 (Min Samples Split)、叶子节点最小样本数 (Min Samples Leaf)、最大特征数量 (Max Features)。对于多参数的调优问题, 本文采用网格搜索和交叉验证的方法, 搜索范围及参数的最佳取值如表 4 所示。

当选取最佳参数进行分类时, 精准率、召回率和  $F1$  值相较于默认参数都有上升, 其中,  $F1$  值达到了 0.904 8, 上升明显。

4.1.3 随机森林参数优化研究

随机森林 (Random Forest) 是集成学习算法的一种, 是一种利用多棵决策树对样本进行训练和预测的

表 4 决策树算法的参数调节区间及最佳取值  
Tab.4 Parameter adjustment interval and optimal value of decision tree algorithm

| 参数                | 参数变化区间 | 参数最佳取值 |
|-------------------|--------|--------|
| Max Depth         | [1,15] | 12     |
| Min Samples Split | [1,10] | 5      |
| Min Samples Leaf  | [1,10] | 2      |
| Max Features      | [1,9]  | 7      |

分类器。

在随机森林算法中, 以下 3 个参数可以提高模型的分类能力:

1) Max\_features, 表示在单个决策树生成时考虑的最大特征数。一般来说该值越大, 单个决策树的准确性越大, 最终模型的性能越好。本研究更注重模型准确性, 因而选择全部特征。

2) N\_estimators, 表示模型中决策树的个数。该值越大, 分类性能越好, 但模型训练时间会增加。在研究中需结合具体的实验环境与目的, 合理选择决策树的数量。

3) Min\_sample\_leaf, 表示决策树叶节点的最小样本数。参数值较小时, 模型对训练数据中的噪音比较敏感, 容易产生过拟合。

当 Min\_sample\_leaf 为默认值 1 时, 随机森林算法中决策树的数量对分类性能的影响如图 5 所示, 当决策树的数量为 30 时,  $F1$  值达到最大, 之后决策树的数量继续增加, 但是  $F1$  值几乎不再提升, 所以本文选择的 N\_estimators 值为 30。

当 Max\_features 选择全部特征、N\_estimators 为 30 时, Min\_sample\_leaf 在 [1,30] 区间内的取值对随机森林算法分类性能的影响如图 6 所示。由图 6 可知,

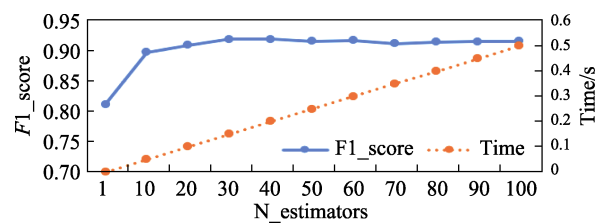


图 5 随机森林算法中树的数量对分类结果的影响  
Fig.5 Effect of the number of trees in random forest algorithm on classification results

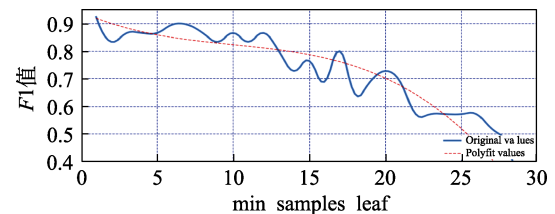


图 6 随机森林算法中 Min\_sample\_leaf 对分类结果的影响  
Fig.6 Effect of Min\_sample\_leaf on classification results in Random Forest algorithm

Min\_sample\_leaf 的值在 1 附近时分类性能最好,之后 F1 值逐渐下降。经过进一步实验,可以确定 Min\_sample\_leaf 取值为 1 时,随机森林算法分类效果最佳,此时 P 值为 0.908 32, R 值为 0.930 13, F1 值为 0.919 09。

4.1.4 SVM 参数优化研究

SVM 的全称是 Support Vector Machine,即支持向量机,它的基本模型是定义在特征空间上的间隔最大的线性分类器,主要用于解决模式识别领域中的数据分类问题。

SVM 算法通过使用核函数,将低维的数据映射到更高维的空间,使高维空间中的数据线性可分。因此 SVM 算法的分类性能很大程度依赖于核函数的选择。在目标函数的惩罚系数 C 取 1.0 时,本研究对不同核函数开展实验,不同核函数的 SVM 分类结果如表 7 所示。

由表 5 可以看出,在 C 取 1.0 时,高斯核函数(rbf)的分类效果明显优于其他核函数,因此选定高斯核函数(rbf)作为 SVM 算法的核函数。

表 5 不同核函数的 SVM 分类结果  
Tab.5 SVM classification results with different kernel functions

| 核函数     | F1       |
|---------|----------|
| rbf     | 0.852 60 |
| linear  | 0.750 00 |
| poly    | 0.823 52 |
| sigmoid | 0.087 00 |

对惩罚系数 C 与核函数参数 Gamma 进行调参。选定 Gamma 为默认数值,F1 值随 C 值的变化见图 7,随着 C 值不断增大,F1 值逐渐下降,因此惩罚系数 C 取 1。

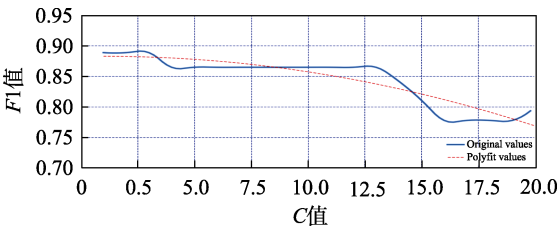


图 7 SVM 算法中 C 值对 F1 值的影响  
Fig.7 Effect of C value on F1 value in SVM algorithm

由表 5 可以看出,在 C 取 1.0 时,高斯核函数(rbf)的分类效果明显优于其他核函数,因此选定高斯核函数(rbf)作为 SVM 算法的核函数。

对惩罚系数 C 与核函数参数 Gamma 进行调参。选定 Gamma 为默认数值,F1 值随 C 值的变化见图 7,随着 C 值不断增大,F1 值逐渐下降,因此惩罚系数 C 取 1。

在确定 C 的取值之后,对 Gamma 进行调参,见图 8, F1 值随着 Gamma 的增大先增大,然后减小, F1 最大时对应的 Gamma 值为 0.08。此时 F1 的值为 0.887 93,与默认参数时相比有较明显的提升。

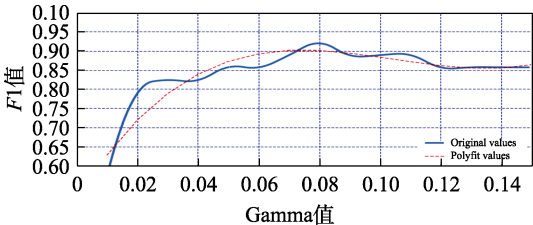


图 8 SVM 算法中 Gamma 值对 F1 值的影响  
Fig.8 Effect of Gamma value on F1 value in SVM algorithm

4.2 分类算法综合评价与模型验证

针对 OpenIDEO 社区数据分别对各分类算法进行参数调节,与默认参数相比,除 KNN 算法外,各分类算法的性能均得到了不同程度的提升,表 6 是参数调优后各算法的分类效果。

表 6 调参后各分类算法的分类效果  
Tab.6 Classification effect of each classification algorithm after adjusting parameters

| 分类算法 | Precision | Recall   | F1       | 默认参数     |
|------|-----------|----------|----------|----------|
| KNN  | 0.825 67  | 0.941 05 | 0.879 59 | 0.879 59 |
| 决策树  | 0.819 92  | 0.844 98 | 0.832 26 | 0.731 76 |
| 随机森林 | 0.908 32  | 0.930 13 | 0.919 09 | 0.894 82 |
| SVM  | 0.876 60  | 0.899 56 | 0.887 93 | 0.862 65 |

由表 6 可知,随机森林算法由于其集成了多个决策树分类器,分类性能得到了进一步的提升,在本文的四个分类器中具有最好的分类效果。因此,针对 OpenIDEO 社区的用户创意数据,本文选用随机森林算法来构建创意思识别与评估模型。

由表 6 可知,随机森林算法由于其集成了多个决策树分类器,分类性能得到了进一步的提升,在本文的四个分类器中具有最好的分类效果。因此,针对 OpenIDEO 社区的用户创意数据,本文选用随机森林算法来构建创意思识别与评估模型。

为验证通过随机森林方法构建的潜在有价值创意思识别模型的可用性,本文收集了 OpenIDEO 社区于 2020 年发布的名为“NextGen Cup Challenge”的挑战赛数据,共计 231 条创意,并按照相同的方法进行数据预处理和数据的转换映射。利用训练好的分类器进行分类,结果如表 7 所示。F1 值达到了 0.904 8,表

表 7 验证数据分类结果  
Tab.7 Classification results of validation data

| 分类算法 | Precision | Recall  | F1      |
|------|-----------|---------|---------|
| 随机森林 | 0.863 6   | 0.950 0 | 0.904 8 |

明该模型对 OpenIDEO 社区中其他挑战赛,可以较好地识别出社区中的潜在有价值创意,也说明该模型具有高可用性。

### 4.3 开放创新社区创意思识别方法总结

本文所研究的开放创新社区潜在有价值创意思识别方法主要分为四个步骤,见图9。首先构造创意思识别与评估模型,该模型要具备从多个维度来评估创意思价值的能力;然后依据模型指标,利用网络爬虫采集开放创新社区数据,并存储在数据库中;之后对数据进行预处理,根据数据类型与指标的映射关系进行数值转化;最后利用处理好的数据训练机器学习模型,对多种机器学习算法进行参数优化调节,选择分类效果最好算法作为分类器。对于该社区的其他创意,可以利用训练好的分类器来进行潜在有价值创意思的识别,达到提高创意思识别效率、降低成本、优化开放创新生态的目的。

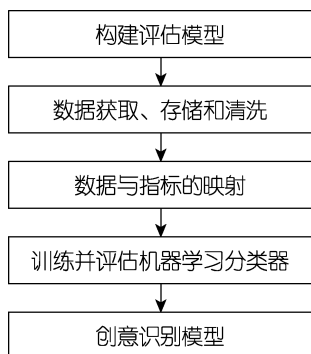


图9 开放创新社区创意思识别与评估技术路线

Fig.9 Technical route to identify and evaluate ideas for open innovation communities

## 5 结语

在知识经济时代,开放创新社区已成为企业获取外部创新资源的主要途径,促进了企业与外部资源间的开放与合作,加速了开放式创新模式的发展。但在创新实践中,现阶段开放创新社区中的创意筛选机制不完善,缺乏系统性的自动化筛选方法与工具,往往需要借助社区或企业专家对大量创意方案进行评估,导致筛选效率低下,缺乏准确性和时效性,进而导致开放创新社区生态恶化。本文通过分析创意思质量的影响因素,从用户特征、用户过程参与度和创意思内容特征三个维度出发,构建了开放创新社区潜在有价值创意思识别与评估模型,并以 OpenIDEO 开放创新平台为研究对象,通过社区创新数据,对数据进行转换映射,然后运用四种机器学习算法分别对实验数据进行分类,优化算法的参数,得到了分类效果最优的算法,完成了开放创新社区潜在有价值创意思识别模型的构建。

本文基于创意思价值影响因素,构建了一个较为全面的创意思识别与评估模型,可以直接帮助开放创新社

区从大量的创意中筛选出潜在有价值创意,可以有效降低企业评估外部创意思的成本,提高创意思评估的时效性和准确率,有利于增强外部用户以开放创新模式参与企业创新的动力,提高创新效率,优化开放社区生态。

同时本文的研究也存在局限性:

1) 本文以 OpenIDEO 社区的创意数据为研究对象,具有一定的局限性。现存的开放创新社区有很多,如戴尔 Ideastorm、海尔 Hope 等,未来将以多个开放创新社区为基础来构建研究模型,以提高模型的普适性。

2) 本文建立的研究模型所包含的评价指标仍有待完善,未来将更全面地挖掘影响创意思质量与价值的因素,完善创意思评价指标,以提高模型的准确性。

### 参考文献:

- [1] CHESBROUGH H. The Market for Innovation: Implications for Corporate Strategy[J]. California Management Review, 2007, 49(3): 45-66.
- [2] 石丽雯,郭伟,张静,等. 群智项目创新潜力的影响因素研究[J]. 包装工程, 2021, 42(24): 13-21, 72.  
SHI Li-wen, GUO Wei, ZHANG Jing, et al. Factors Impacting Innovation Potential in Crowd Intelligence Project[J]. Packaging Engineering, 2021, 42(24): 13-21, 72.
- [3] 王楠,陈详详,祁运丽,等. 基于详尽可能性模型的用户创新社区创意思采纳影响因素研究[J]. 中国管理科学, 2020, 28(3): 213-222.  
WANG Nan, CHEN Xiang-xiang, QI Yun-li, et al. The Research on Influence Factors of User Innovation Community Idea Adoption Based on Elaboration Likelihood Model[J]. Chinese Journal of Management Science, 2020, 28(3): 213-222.
- [4] 刘晴晴. 开放式创新社区创意思采纳的影响因素实证研究[D]. 济南: 山东大学, 2017.  
LIU Qing-qing. Empirical Study on Factors Influencing Idea Adoption in Open Innovation Community[D]. Jinan: Shandong University, 2017.
- [5] HOSSAIN M, ZAHIDUL ISLAM K M. Ideation through Online Open Innovation Platform: Dell IdeaStorm[J]. Journal of the Knowledge Economy, 2015, 6(3): 611-624.
- [6] LI Ming-guo, KANKANHALLI A, KIM S H. Which Ideas are more Likely to be Implemented in Online User Innovation Communities? an Empirical Analysis[J]. Decision Support Systems, 2016, 84: 28-40.
- [7] 王楠,陈详详,孙百惠,等. 谁的创意思更具价值——内外部领先用户的对比研究[J]. 南开管理评论, 2021, 24(5): 51-62, 72, 63.  
WANG Nan, CHEN Xiang-xiang, SUN Bai-hui, et al. Whose Idea is more Valuable? A Comparative Study Based on Internal and External Lead Users[J]. Nankai Business Review, 2021, 24(5): 51-62, 72, 63.
- [8] 郭伟,王洋洋,梁若愚,等. 开放式创新社区中用户交互反馈对个体创新贡献度的影响[J]. 科技进步与对



- 策, 2018, 35(3): 146-152.
- GUO Wei, WANG Yang-yang, LIANG Ruo-yu, et al. Study on Factors of Individual Innovation Contribution Degree Based on User Feedback in Open Innovation Communities[J]. Science & Technology Progress and Policy, 2018, 35(3): 146-152.
- [9] 蒋旋, 孟庆良, 徐信辉, 等. 双边视角下用户参与众包创新的知识获取前因及其对创新绩效的影响[J]. 运筹与管理, 2021, 30(2): 225-231.
- JIANG Xuan, MENG Qing-liang, XU Xin-hui, et al. User Participation in Crowdsourcing from a Bilateral Perspective: Antecedents of Knowledge Acquisition and Their Impact on Innovation Performance[J]. Operations Research and Management Science, 2021, 30(2): 225-231.
- [10] 王磊, 马龙江, 彭巍, 等. 群智创新社区用户创新能力分析[J]. 科技进步与对策, 2018, 35(18): 42-47.
- WANG Lei, MA Long-jiang, PENG Wei, et al. Research on the Innovation Ability of Online Users in Crowd Sourcing Innovation Community[J]. Science & Technology Progress and Policy, 2018, 35(18): 42-47.
- [11] 张静, 郭伟, 王磊, 等. 用户交互对产品知识认知扩散的影响及社区激励研究[J]. 科技进步与对策, 2018, 35(16): 7-15.
- ZHANG Jing, GUO Wei, WANG Lei, et al. The Influence of User Interaction on the Cognition and Diffusion of Products Knowledge and Community Incentive Research[J]. Science & Technology Progress and Policy, 2018, 35(16): 7-15.
- [12] KOHN N W, SMITH S M. Collaborative Fixation: Effects of Others' Ideas on Brainstorming[J]. Applied Cognitive Psychology, 2011, 25(3): 359-371.
- [13] THOMAS W MALONE M S B. Handbook of Collective Intelligence[M]. Cambridge, MA: The MIT Press, 2015.
- [14] JI Yan, LI Yue-en, LIU Na. A Study on Collaborative Innovation for Product Design[C]//Proceedings of the 2018 International Conference on Service Robotics Technologies. Chengdu, China. New York: ACM, 2018: 73-77.
- [15] 陈文春, 张义明. 知识型团队成员异质性对团队创造力的影响机制[J]. 中国科技论坛, 2017(9): 178-185.
- CHEN Wen-chun, ZHANG Yi-ming. Effects of Team Heterogeneity on Team Creativity: The Mediating Role of Team Learning Behavior and the Moderating Role of Team Identity[J]. Forum on Science and Technology in China, 2017(9): 178-185.
- [16] 刘惠琴. 团队异质性、规模、阶段与类型对学科团队创新绩效的影响研究[J]. 清华大学教育研究, 2008, 29(4): 83-90.
- LIU Hui-qin. The Effects of Team Heterogeneity, Size, Stages and Type on Academic Team's Innovation Performance[J]. Tsinghua Journal of Education, 2008, 29(4): 83-90.
- [17] 张春彩. 团队异质性对不同规模团队的影响研究[D]. 西安: 西安电子科技大学, 2011.
- ZHANG Chun-cai. The Research of Team Diversity Effect on Team under Different Team Size[D]. Xi'an: Xi-dian University, 2011.
- [18] 王洋洋. 群智创新过程中用户行为与创新项目协同演进研究[D]. 天津: 天津大学, 2018.
- WANG Yang-yang. Research on User Behavior and the Co-evolution of Innovative Project in the Crowd Intelligence Innovation Process[D]. Tianjin: Tianjin University, 2018.
- [19] TANG Chao-ying, YE Lin-na. Diversified Knowledge, R&D Team Centrality and Radical Creativity[J]. Creativity and Innovation Management, 2015, 24(1): 123-135.
- [20] CHOI S W, PARK S H, JEONG H S, et al. Improvement of Formability for Fabricating Thin Continuously Corrugated Structures in Sheet Metal Forming Process[J]. Journal of Mechanical Science and Technology, 2012, 26(8): 2397-2403.
- [21] DI GANGI P M, WASKO M. Steal my Idea! Organizational Adoption of User Innovations from a User Innovation Community: A Case Study of Dell IdeaStorm[J]. Decision Support Systems, 2009, 48(1): 303-312.
- [22] DI GANGI P, WASKO M, HOOKER R. Getting Customers' Ideas to Work for You: Learning from Dell how to Succeed with Online User Innovation Communities[J]. MIS Quarterly Executive, 2010, 9(4): 213-228.
- [23] HUANG Yan, SINGH P V, SRINIVASAN K. Crowdsourcing New Product Ideas under Consumer Learning[J]. Management Science, 2014, 60(9): 2138-2159.
- [24] 任亮. 开放式创新社区知识协同创新研究[D]. 长春: 吉林大学, 2020.
- REN Liang. Research on Knowledge Collaborative Innovation of Open Innovation Community[D]. Changchun: Jilin University, 2020.
- [25] WONG T C, CHAN H K, LACKA E. An ANN-Based Approach of Interpreting User-Generated Comments from Social Media[J]. Applied Soft Computing, 2017, 52: 1169-1180.
- [26] 张艳辉, 李宗伟. 在线评论有用性的影响因素研究: 基于产品类型的调节效应[J]. 管理评论, 2016, 28(10): 123-132.
- ZHANG Yan-hui, LI Zong-wei. Analysis of the Factors that Influence Online Reviews Helpfulness: Based on the Regulating Effect of Product Type[J]. Management Review, 2016, 28(10): 123-132.

责任编辑: 马梦遥