数据驱动的科研项目绩效评估方法研究

李洪波，姚博文，闫鑫
（上海大学 管理学院，上海 200444）

摘要：科研项目绩效的有效评估有助于提升管理部门的决策水平与效率，促进科研项目管理的科学化。项目管理部门和各类学术数据库已经积累了大量的项目相关数据，而机器学习等人工智能技术的进步，为项目数据的自动化和智能化分析提供了重要支撑。鉴于此，本文提出了一个数据驱动的科研项目绩效评估方法，利用多分类监督学习算法，对已结题项目数据中隐含的关于项目绩效的信息进行有效挖掘，形成项目绩效评估模型；当需要对新项目的绩效进行评估时，将新项目的特征信息输入评估模型，即可自动得到新项目的绩效预测。在本文的评估方法中，利用网络爬虫自动化的从多个不同的网站收集项目相关的结构化和非结构化数据，在对数据进行处理与融合后，从中提取出有效的特征，基于一系列多分类算法构建模型，并利用多种指标对模型的性能进行估计。基于国家自然科学基金项目数据，对本文方法进行了验证。

关键词：科研项目；绩效评估；机器学习；数据驱动
中图分类号：G311　　**文献标志码**　A

A data-driven performance evaluation approach for scientific research projects

LI Hongbo, YAO Bowen, YAN Xin
(School of Management, Shanghai University, Shanghai 200044, China)

Abstract：The effective evaluation of the performance of scientific research projects is helpful to improve the decision-making level and efficiency of management departments and promote the scientific management of scientific research projects. Project management departments and various academic databases have accumulated a large number of project-related data. Meanwhile, the progress of artificial intelligence technologies such as machine learning has provided important support for the automation and intelligent analysis of project data. In terms of this, this paper proposes a data-driven evaluation method for the performance of scientific research projects. By using the multi-classification supervised learning algorithm, the project performance information implicit in the completed project data is effectively mined to form a project performance evaluation model. When it is necessary to evaluate the performance of the new project, the predictive performance of the new project can be obtained automatically by inputting the characteristic

基金项目：国家自然科学基金项目（71602106）；上海市“科技创新行动计划”软科学重点项目（20692192400）。
作者简介：李洪波（1985-），男，山东东营人，博士，上海大学管理学院副研究员，主要研究方向：项目调度、人工智能。
通讯作者：姚博文（1996-），男，辽宁锦州人，上海大学管理学院硕士研究生，主要研究方向：大数据与大数据技术。
闫鑫（1998-），女，山西运城人，上海大学管理学院硕士研究生，主要研究方向：大数据与大数据技术。

information of the new project into the evaluation model. In this paper's evaluation method, the use of web crawler has collected structured and unstructured projects' data from a number of different sites, and after the data processing and integration, effective features are extracted, which have been utilized to construct the model based on a series of multiple classification algorithm. Moreover, some indicators have been used to estimate the performance of the model. The method of this paper is verified based on the project data of National Natural Science Foundation of China.

Key words scientific research projects; machine learning; performance evaluation; data-driven

1 引言

科技创新对高质量发展具有显著的引领作用，我国各级政府部门为促进科技创新投入了大量资金，科研计划项目的资助规模逐年提高。以国家自然科学基金中的面上、青年和地区三类项目为例，从2015年到2019年，资助数量和资助金额分别增长了11.85%、13.13%（图1）。资助规模的增长，有力促进了我国基础科研的发展，但也加剧了科研管理部门项目评估工作的复杂性和工作量[1]。因此，科研项目评估工作需要更高效的量化评估方法，以提升管理部门的决策水平与效率，促进科研项目管理的科学化和数字化[2]。

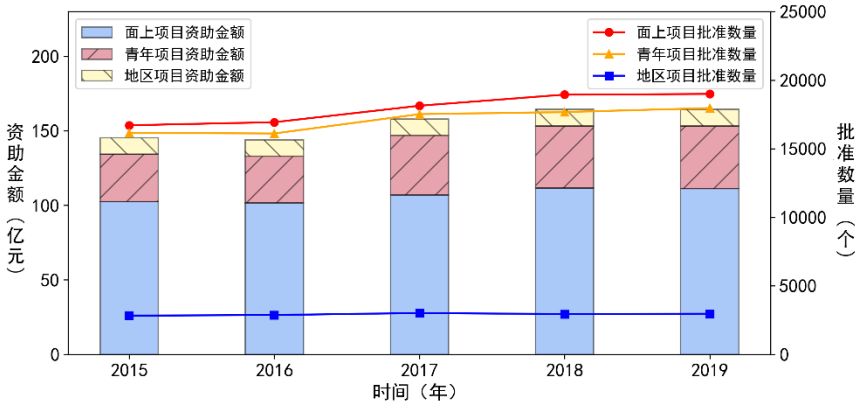

图 1 2015-2019年 NSFC“面青地”三类项目资助数量和金额情况

学者们对科研项目的绩效评估进行了大量深入的研究，提出了许多行之有效的评估方法[3-9]。专家估计法（同行评议法）是应用最为广泛的方法，这是一种主观评价为主的方法，专家根据项目的成果和各个指标完成情况对项目打分，但该方法容易存在学术或认识上的偏差[10]。数据包络分析法（DEA）也是一种被研究的较多的方法[11]，如Hsu & Hsueh（2009）使用DEA对我国台湾省过去20年的IT项目研发效率进行评价[12]；Johns（2008）使用DEA检验了2003和2004年中国109所大学科研项目成果效率，评价的变量涵盖人员、学生、资金和资源等指标[11]。

近年来，机器学习等人工智能技术的进步，为科研项目的绩效评估，带来了新的发展契机。郝平（2009）设计了一个基于数据挖掘技术的科技计划项目绩效评价系统，数据来自项目申请、立项文件、合同和自评报告等，系统中涵盖了聚类、分类和关联规则等常见的数据挖掘方法[13]。刘宝涛（2019）研究了人工智能的正态云模型在农业科研绩效评价中的应用[14]。Costantino（2015）使用人工神经网络（ANN）方法，将专家观点提取成隐含在神经网络中的知识，不需要专家参与到复杂的评判中来，而是将他们的感知翻译成模型的输入和输出，实现了机器学习模拟专家评估的目的[15]。Liu（2019）等人提出数据驱动的基于证据推理的推论模型，利用中国国家自然科学基金管理学部497个项目的数据，将评估水平（优良

中差）和资助建议（资助、不资助、优先立项）作为两大指标，用ER推论模型模拟了专家打分过程[16]。Jang（2019）使用韩国国家级资助的科研项目数据构建了机器学习模型，以项目的论著和专利数量作为项目产出水平的指标，基于项目类别、科研经费等维度估计项目产出水平[17]。

然而，在政府资助的科研项目领域，将机器学习方法用于项目评估的研究很少，以中国的科研计划项目为对象、基于机器学习的研究则更为匮乏。经过几十年的发展，我国各级项目管理部门积累了大量的科研项目的数据，并且其中很多数据已经公开在了网络上，如国家自然科学基金大数据知识管理服务门户（http://kd.nsfc.gov.cn）。鉴于此，本文提出了一个数据驱动的科研项目绩效评估方法，利用多分类监督学习算法，对已结题项目数据中隐含的关于项目绩效的信息进行有效挖掘，形成项目绩效评估模型；当需要对新项目的绩效进行评估时，将新项目的特征信息输入评估模型，即可自动得到新项目的绩效预测。

本文的主要贡献体现在：（1）数据驱动：本文构建了自动化数据采集方法，综合使用了网络爬虫、正则匹配和图片识别等数据获取和处理手段，形成了科研项目数据从获取到处理的完整框架体系。（2）智能评估：本文研究了基于机器学习的科研绩效评估方法，是人工智能和科研绩效评估两个领域的结合；本文系统性对比了11种机器学习算法，采用多种模型评价指标，构建了数据驱动的科研绩效评估模型，一方面减少了科研绩效评估过程中的人力耗费，另一方面提高了科研评估工作的效率，为科研管理的数字化创新转型提供了切实可行的思路。

2 问题的提出

评审专家对已结题的科研项目进行评估时，会从项目成果的水平、效益、质量等多个方面进行综合考虑。假设给定一个待评估项目i，从n个维度提取出反映项目绩效的特征，记为n维向量$\boldsymbol{x}_i = \left(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}, \dots, x_i^{(n)}\right)^T$，其中$x_i^{(m)}$表示项目$i$第$m$个维度的值。专家评审后，给出的项目$i$评估等级为$y_i \in Y$，本文也将$y_i$称为标记，$Y$为所有评估结果取值的集合。根据项目资助机构的要求，y_i可能为数值型取值（如80分、90分等），也可能为离散型取值（如优、良等），本文采用后者。

可见，科研计划项目绩效的评估，可以视作一个函数拟合的过程，即将影响科研计划项目绩效的各种因素作为特征，将专家评估得到的项目绩效作为标记，进而拟合出一个尽可能准确的函数f，实现从特征$\boldsymbol{x}_i$到标记y_i的有效映射。

从机器学习的角度看，上述项目绩效评估是一个多分类问题，即给定各种科研项目的特征，去预测科研项目绩效的好坏。因此，可以借助多分类预测算法进行科研项目最终绩效估计。本文将利用多分类算法，挖掘已完成评估项目的诸多特征与专家评估结果之间的隐含规律，形成尽可能准确的项目绩效评估模型。当需要对新的待评估项目的绩效进行评估时，将新项目的特征信息输入模型，即可自动得到新项目的绩效评估信息。希望利用本文的数据驱动的项目绩效评估方法，为评审专家提高有效的参考信息，减少其工作量，并提高项目评估结果的客观性与准确性。

基于以上论述，本文接下来的数据驱动的科研项目绩效评估方法，将主要解决如下3个核心问题：（1）如何自动化获取科研项目数据；（2）如何从项目数据中提取有效的特征信息；（3）如何构建有效的多分类预测模型。

3 数据驱动的科研项目绩效评估框架

本文提出了一个数据驱动的科研项目绩效评估框架，它主要分为项目数据获取、数据处理和评估模型构建三个部分（图2）。

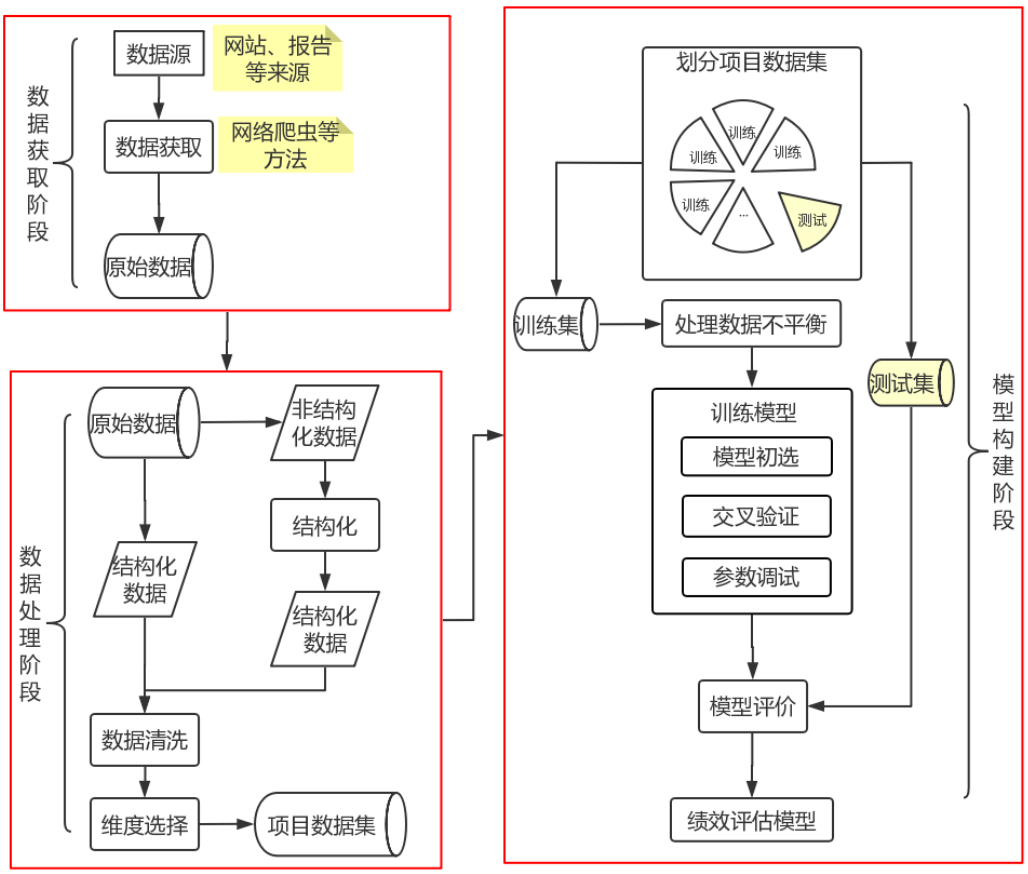

图 2 数据驱动的科研项目绩效评估框架

（1）数据获取部分实现项目相关数据的自动化收集，主要从项目管理部门的网站、学术数据库等多种途径利用网络爬虫收集数据。假设一共获取P个项目的数据，构建数据集$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_P, y_P)\}$，其中$n$维向量$\boldsymbol{x}_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}, \ldots, x_i^{(n)}\right)^T$表示第$i$个项目所有特征集合（在机器学习术语中，$\boldsymbol{x}_i$也称为样本、示例等），$y_i$是第$i$个项目的标记，也就是项目评估后的等级，$y_i \in \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{n}\}$，第4.1和4.2节将详细阐述本部分。

（2）在数据处理部分，对数据集D进行清洗与融合。去掉重复值、空白值、异常值等非正常数据。对部分维度进行标准化和归一化，从而消除不同特征之间的量纲影响以及特征之间对标记的权重影响。对于图片和文本数据等非结构化数据，分别使用图片识别和自然语言处理的方法，提取其中的信息、并将其转化为结构化数据，进而将整理好的数据集进行初步的描述性统计分析，第4.3和4.4节将详细阐述本部分。

（3）在评估模型构建部分，从数据集D中选择一部分数据作为训练集D^{tr}。针对数据集中存在的分布不平衡问题，利用SMOTE算法进行处理。之后，在训练集上使用预测算法构建n个机器学习模型$\{f_1, f_2, \ldots, f_m, \ldots, f_n\}$，并对$n$种模型初步评价，选择效果最好的模型进行进一步调试。将数据集D中剩余的数据作为测试集D^{te}，考察模型在测试集中的表现，探讨不同特征对评估模型的影响程度。第5节将详细阐述本部分。

本文构建的数据驱动的科研项目绩效评估框架，具有较好的普遍适用性和扩展性，其中的子流程很容易根据实际问题做出相应调整。例如，本框架既适用于中国国家自然科学基金项目数据，还可以应用到其他类型和层次的科研项目计划项目中，对于不同种类的项目，只需对数据格式相关的部分做少量修改。接下来将基于中国国家自然科学基金项目数据，对本文数据驱动的科研项目绩效评估框架进行系统性探讨。

4 项目数据获取与处理

4.1 自动化数据收集

本文采用python语言编写网络爬虫[18]（流程见图3），从国家自然科学基金委官网（http://www.nsfc.gov.cn）自动收集结题时间为2016年和2017年的管理科学部面上项目数据。本文爬虫利用项目批准号，生成每个项目对应的URL网址，然后调用selenium库自动化访问每个URL的网页[19]。在爬取数据过程中，为减小给网站带来的额外访问负载，使用selenium自动化测试工具时，设置了每次访问网站的间隔时间来模拟人工访问效果，以最大程度减轻服务器的压力。在获取网页源代码后，使用pyquery库解析html源代码，构造正则表达式匹配所需的信息，存入mongo数据库。

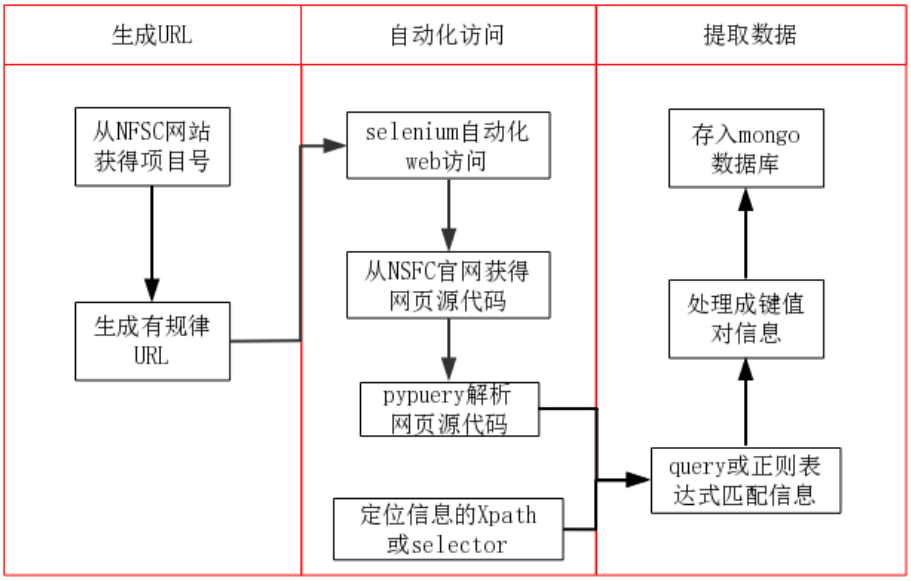

图 3 基于网络爬虫的自动化数据收集流程

基于上述爬虫，本文收集到的每个项目均包含如下信息：项目批准号、项目类别、金额、项目后评估结果、项目的成果信息等结构化数据，以及项目摘要文本和图片形式的结题报告等非结构化数据。本文特征的选取主要依据国家自然科学基金委项目后评估科研指标评价体系[20]。

本文还另外开发爬虫，从中国知网（www.cnki.net）、超星期刊网（http://qikan.chaoxing.com/）和中外文核心期刊查询系统（http://corejournal.lib.sjtu.edu.cn/）获取了常见中外文期刊的影响因子，通过与之前收集到的项目成果所在期刊名称进行匹配，计算得到每项成果影响因子的平均值，与前述结构化数据合并，形成本文的第一部分原始数据集D_1。

4.2 非结构化数据中的信息提取

本节对前述收集的摘要文本和结题报告图片这两类非结构化数据进行处理，从中提取结构化数据。

（1）摘要文本信息的提取

为了训练一个具有较好预测能力的模型，需要将文本转换成可以描述文本内容的数值型特征。使用词向量转化的方法将一个文档转化成一个固定维度的向量，每个向量的维度表示文档的一个特征[21][22]。本文采用了 TF-IDF 模型构建词向量，在对项目摘要进行词向量转化后，对于每个项目的摘要D_0，其对应的文本特征为维度较高的向量C。将一系列特征加入原

始数据集中，得到总数据集D_{text}。由于文本挖掘产生了维度较高的特征，在实际构建预测模型时带来较大运行负载，因此本文的文本挖掘作为构建模型的辅助方法，仅用于模型构建完成后提高预测能力的方法使用。

（2）图片信息的提取

前面爬取到的项目结题报告为图片格式，使用文字识别技术从中提取有用的信息。在提取时，使用图片文本识别库tesseract库[23]，tesseract库能够对单行文本格式图片中的文字进行提取。但国家自然科学基金项目的结题报告中，存在很多固定格式的表格，相关数据位于表格中，而tesseract不能识别表格文本，因此本文通过像素定位、裁剪切割表格的方法，实现了信息的提取[24]，随机抽样对比后发现该方法的识别结果十分准确。从结题报告中可以获得项目发表的论文数、论文的收录情况、人才培养情况和国际交流的情况等信息。将结题报告中获得的数据作为第二部分数据集D_2。

4.3 数据清洗与融合

对数据集D_1和D_2进行清洗，去掉重复值、空数据以及不符合常识的异常数据。以“项目批准号”为主键对D_1和D_2进行合并融合为最终数据集D。融合后数据集中评估结果$y_i \in Y =$ {特优，优，良，中}，为简便起见，将特优、优、良、中分别编码为4、3、2、1。

表 1 本文所用数据集中的特征、标记及其解释

特征名（括号内为代号）	解释
批准号（ID）	项目的唯一标识码，整张表的主键
项目经费（EXPENDITURE）	项目获批经费，连续型变量，以万元为单位
论文数（TOTAL_NUM）	项目成果中发表期刊的论文总数，连续型变量
外文期刊影响因子（AVER_INDEX_E）	项目发表外文论文所在期刊的影响因子平均值，连续型变量
中文期刊影响因子（AVER_INDEX_C）	项目发表中文论文所在期刊的影响因子平均值，连续型变量
加权影响因子（WEIGHTED_INDEX）	用SCI论文数量加权的影响因子，连续型变量
外文期刊论文数（ENG_NUM）	项目成果中发表外文期刊的论文数量，连续型变量
中文期刊论文数（CHI_NUM）	项目成果中发表外文期刊的论文数量，连续型变量
学术报告次数（ACA_REPORT）	项目学术报告数量，连续型变量
期刊论文数（JOURNAL）	发表期刊论文数量，连续型变量
会议论文数（MEETING_PAPER）	发表会议论文数量，连续型变量
SCI论文数（SCI）	发表SCI论文数量，连续型变量
EI论文数（EI）	发表EI会议论文数量，连续型变量
北大核心期刊论文数（BU_CORE）	发表北大核心期刊论文数量，连续型变量
CSSCI论文数（CSSCI）	发表CSSCI期刊论文数量，连续型变量
博士数量（PHD）	项目培养博士数量，连续型变量
硕士数量（MASTER）	项目培养硕士数量，连续型变量
国际会议次数（NUM_OF_IM）	项目参与国际会议次数，连续型变量
获奖（PRIZE）	项目所获奖项数目，连续型变量
负责人上次项目评估结果（LAST_E）	项目负责人上次申请项目的评估结果，离散型变量
项目摘要的一系列特征	通过文本挖掘得到的一系列特征
标记（括号内为代号）	**解释**
评级（ASSESS）	项目评估中获得的等级，包括特优、优、良、中、差五个级别

在数据融合时，发现数据集中存在一些方差小、相关性高和含0率高的特征。例如毕业

硕士数量和培养人才总数这两个特征，其相关系数>0.95，因此本文舍弃培养人才总数这一特征，保留毕业硕士数量这一特征。而含0率高的特征，则可能对数据的预测产生影响，例如重要报告和专利等特征，超过90%的项目在这些特征上取值为0，标记的取值在这一特征维度上差别不明显，因此去掉此类变量，以减少对预测效果的负面影响。表1给出了数据集D中的特征和标记（项目的绩效评估结果）。

表 1 中的特征可以归纳为 6 个方面[1]：

（1） 报告论著相关的特征：在学术期刊上发表的论文作为项目的成果，如果一个项目的论文发表的多，可以在一定程度上说明项目的成果也多。由于学术期刊种类不一，简单的数量判断不能表明成果的实际价值，因此选取了几种知名索引SCI、CSSCI、北大核心等来对发表的论文进行区分，另外，通过对发表论文的期刊影响因子爬取，可以大致评价项目发表论文的影响力。

（2） 人才培养相关的特征：一个项目的产出水平除了体现在发表论文外，还体现在对人才的培养上，表 1 展示的特征中，培养学术带头人、博士、硕士等的数量可以作为这一方面的衡量标准,一般来讲，如果项目培养更多的人才，那么这个项目就会更有价值。

（3） 学术创新相关的特征：表 1 中的获奖情况可以表示学术创新方面的成果，奖项包括国家级科技创新奖和省级科技创新奖等奖励。如果项目已经得到其他权威机构的认可，也足以说明项目成果十分优秀。

（4） 国际交流相关的特征：可以从与国际会议相关的变量和发表外文期刊或国际级别的期刊索引等特征中衡量该方面，此类特征显示了研究成果在国际中受到的认可程度以及国际影响力。

（5） 文本相关的特征：除了常规的项目评价指标外，本文引入了文本挖掘的方法，在项目摘要种找出可能会影响项目绩效评价的文本特征。

（6） 项目负责人上一次申请项目的评估结果：申请人上次完成的项目的后评估结果对本次项目的后评估可能有影响，因为之前评估获得过好的绩效水平意味着项目负责人有很好的项目经验，可能对这一次的项目完成有借鉴作用，当然如果之前完成的项目评估结果不好，则可能对本次评估结果有负面影响。对于这一特征为空的项目，也就是负责人之前没申请过项目或者之前的项目没参加评估的这些项目，取所有样本的众数，也就是“良”这一水平。

4.4 描述性统计

本文的数据集D中共包含 1199 个项目，表 3 给出了数据集的一些描述性统计数据。表 2 的第 2、3 列展示了不同评估结果项目的数量与比例信息。可以看出，大多数项目评估结果为“优”或“良”（其总数占总项目的 93.62%），且二者比例接近；其余项目的后评估结果为“特优”或“中”，没有评估结果为“差”的项目。表 2 还列出了几个代表性特征的平均取值，可以看到不同评估等级项目的特征在均值上存在较大差异。特优的项目，其在各个特征的平均值均好于其他项目，尤其在 SCI 发表论文数和发表外文期刊文章数这两个特征上。

表 2 不同类别部分特征平均值比较

评估结果	计数	频率	几个代表性特征的平均取值					
			论文数	外文期刊论文数	SCI 论文数	外文影响因子	博士数量	硕士数量
特优	51	4.25%	27.92	16.39	13.35	4.19	2.15	3.67
优	492	41.03%	23.43	7.61	5.12	1.80	1.53	4.71
良	632	52.71%	18.70	3.76	1.67	0.78	1.05	4.40
中	24	2.00%	8.87	1.50	0.13	0.29	0.58	4.33

图 4 项目摘要词云

5 评估模型构建

5.1 处理数据不平衡问题

从上一节表 2 可以发现，不同类别评估结果对应的项目数量存在巨大差异，因此数据集D中存在数据分布不平衡问题。如果不对这一问题做处理，将会导致多数分类算法的预测结果很差。为了解决这一问题，本文在对预测模型训练前，先利用 SMOTE（Synthetic Minority Oversampling Technique）算法对数据集D进行处理[25]，以降低类别不平衡对分类算法的负面影响[26]。SMOTE 方法通过最近邻算法产生新数据，增加到小类别样本中，使其数量上达到大类别样本的级别，这样训练的模型将会有更高的预测能力[25]。

5.2 模型性能评估指标

基于样本的真实标记值和模型f对标记的预测值，可以构造混淆矩阵$C = \left(c_{yy'}\right)_{4\times 4}$，其中混淆矩阵中的元素$c_{yy'}$表示数据集$D$中，所有真实标记（评估结果）为 y 的项目被预测为y'的总数，$y, y' \in Y = \{1,2,3,4\}$。在上述混淆矩阵的基础上，进一步利用如下指标对绩效预测模型的结果进行评估[27]：

（1）精度

精度A表示正确分类的样本个数占总样本的比例，模型f在数据集D上的预测精度按下式计算：

$$A(f;D) = \frac{1}{m}\sum_{y=1}^{4}\sum_{y'=1}^{4} I(y = y') \times c_{yy'} \tag{7}$$

其中m为数据集D中包含的项目数量，示性函数$I(\cdot)$表示如果其括号内计算结果成立，则该函数取值为 1，否则取 0。精度的取值越大，表明模型预测的准确率越高。

（2）查全率

查全率R是指被正确预测为某个类别的个体数量与数据集中该类别个体总量的比例，查全率越高，则说明模型评估等级和专家一致的项目样本数占该等级类别总数的比例越高。本文多分类问题的查全率为各单个类别查全率的平均，模型f在数据集D上的查全率计算如下：

$$\mathrm{R}(f;D)=\frac{1}{4}\sum_{y=1}^{4}\frac{\sum_{y'=1}^{4}I(y=y')\times c_{yy'}}{\sum_{y'=1}^{4}c_{yy'}} \tag{7}$$

（3）查准率

查准率P是指预测结果属于某一类个体实际属于该类的比例，查准率越高，表明模型评估为某一等级的项目个体中，与实际专家评审的等级类别一致的比例也越高。模型f在数据集D上的查全率为：

$$P(f;D)=\frac{1}{4}\sum_{y'=1}^{4}\frac{\sum_{y=1}^{4}I(y=y')\times c_{yy'}}{\sum_{y=1}^{4}c_{yy'}} \tag{7}$$

（4）$F1$值

查全率和查准率是一对矛盾的概念，通常，当查准率高的时候，查全率一般很低，反之也是如此。因此本文使用综合指标$F1$值，综合考虑查准率P和查全率R两大指标。$F1$即为查准率和查全率的调和平均数，当$F1$的值越大时，说明模型性能更优。$F1$指标计算如下：

$$F1(f;D)=\frac{2\times P\times R}{P+R} \tag{7}$$

5.3 模型选择

为了找到适用于本文数据的多分类算法，本文对 11 种常见的机器学习分类算法进行了计算实验。实验时，使用 python 语言调用 Scikit-Learn 库中的分类算法[28]。按照 7:3 的比例，将前面收集到的数据集 D 分为训练集和测试集。这样，训练集中包含 839 个项目的数据，测试集包含 360 个项目数据。基于训练集，用 11 种不同的分类算法训练出 11 个备选预测模型，这些模型在测试集上的混淆矩阵和预测结果分别如图 5 和表 3 所示。

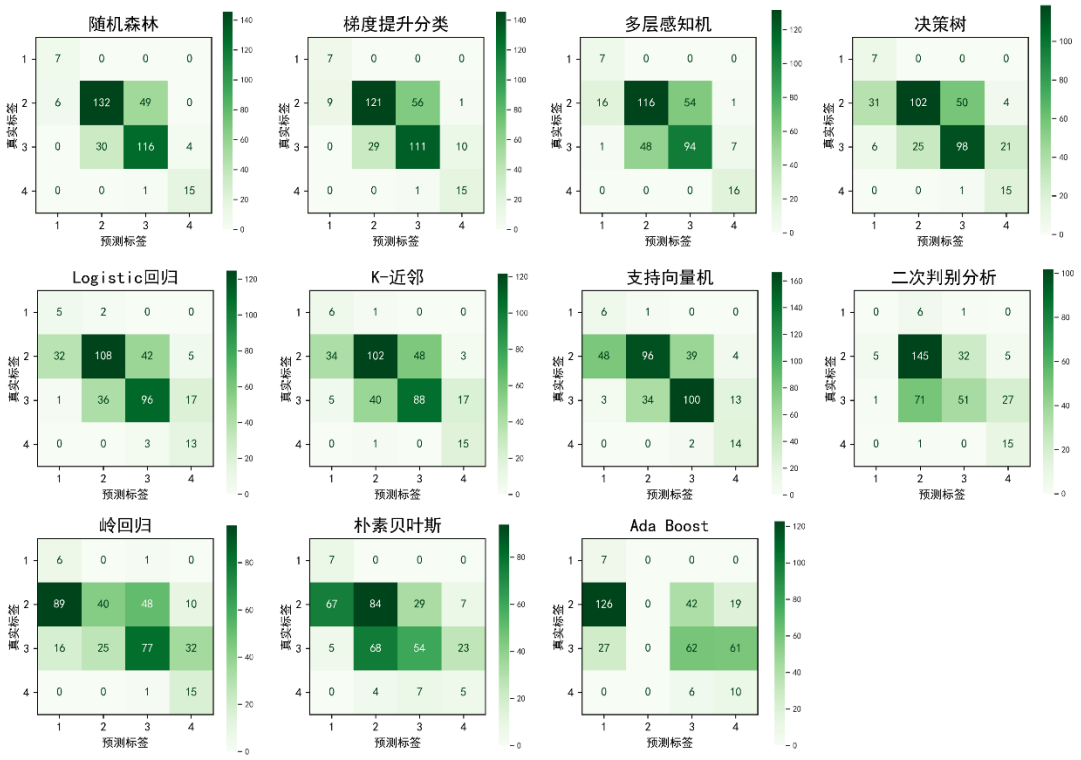

图 5 不同分类算法输出结果的混淆矩阵

表 3　不同分类算法的预测结果

序号	算法	精度A	查全率R	查准率P	$F1$值
1	随机森林	***0.75***	***0.85***	***0.71***	***0.76***
2	坡度提升分类	0.71	0.83	0.62	0.68
3	多层感知机	0.65	0.81	0.58	0.64
4	决策树	0.62	0.78	0.50	0.53
5	Logistic 回归	0.60	0.73	0.50	0.52
6	K 近邻	0.59	0.73	0.48	0.51
7	支持向量机	0.58	0.75	0.43	0.44
8	二次判别分析	0.59	0.51	0.39	0.40
9	岭回归	0.38	0.63	0.38	0.35
10	朴素贝叶斯	0.42	0.53	0.34	0.32
11	Ada Boost	0.22	0.51	0.18	0.19

图 5 中的每个子图，对应一种算法得到的预测结果的混淆矩阵。在每个混淆矩阵中，横坐标表示模型预测的结果，纵坐标表示项目的真实结果，每个单元格中的数字表示真实标签为y的项目被预测为y'的数量（因此矩阵主对角线的元素表示了模型预测正确结果的数量）。从图中可以看出，随机森林模型（图中第一行第一列）可以识别最多数量的正确类别（矩阵主对角线元素）。

观察表 3 和图 5 可以发现，随机森林分类器的各项指标均优于其他模型，因此接下来的分析将基于随机森林模型。同时，为了进一步提高随机森林的性能，本文使用网格搜索功能调节其参数，使用 10 折交叉验证保证模型有效性，其中，分类器的个数设置为从 300 到 1000，以 100 为间隔，最大深度设置为 40 到 110 之间，间隔为 10，同时也增加一个“None”表示模型默认值，叶节点最小样本数为 2、3、4、5、10，内部节点再划分最小样本数为 1、2、3、4，其余参数为模型默认。最终得到最佳模型f_{best}，其参数为：分类器为 900，叶节点最小样本数为 1，内部节点再划分最小样本数为 2，最大深度为默认。

6 结果与讨论

6.1 主要结果

最终优化后的随机森林模型f_{best}的预测结果为：精度A为 0.78，查全率R为 0.87　　查准率P为 0.73，$F1$值提升至 0.78，可见，随机森林模型f_{best}在四个评估指标上均体现出了最好的性能。

接下来进一步考察随机森林对每种类别的评估情况与专家估计的差别。按照标记的取值，将测试集分为 4 部分，随机森林在每部分的预测结果如图 6 所示。预测结果与实际结果相比，尽管每一个类别中评估结果正确的比例不相同，但是评估结果与专家相符的个数都占主导地位，效果较好的两个类别是“中”和“特优”，相对预测结果容易混淆的类别是“优”和“良”。

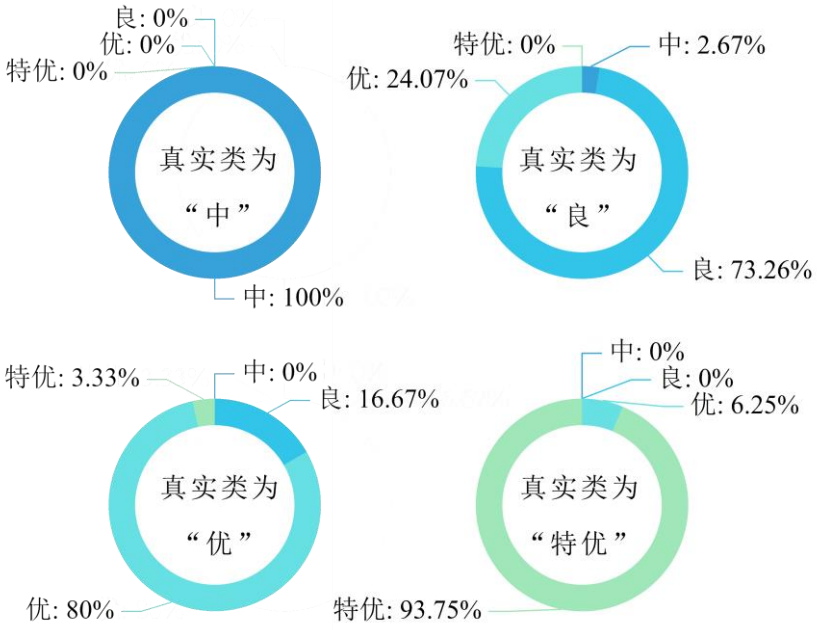

图 1 随机森林对不同类别数据的预测结果

随机森林对“优”和“良”的分类能力不及另两类的原因可以归结为：（1）模型使用的评判维度不及专家评判的维度丰富。本文模型采用的数据均为网络采集，是已经公开的数据，但是项目除了公开的数据外，还有很多没有公开的维度，这就导致本文提出的模型在一定程度上不及专家评审过程。（2）专家在部分项目的评判上存在其他主观因素。在 4.3 节描述性统计中可以看出，“特优”和“中”这两个较为极端的绩效等级在各个维度上都有自己的特点，其中，“特优”在各个维度的表现均为突出。“中”在各个维度上表现均明显落后于其他类别，而“优”和“良”这两个类别本身就处在四个类别中的中间两个位置，相互之间就会存在区分程度小的情况。对成果水平相近的两个项目而言，可能存在客观评分相近的情况下，专家认为其中一个项目具有更高的价值和意义，或者创新性更强，从而主观加分较多的情况。

6.2 特征重要性探讨

本节探讨不同的特征对预测结果的影响程度。给定一个特征，将该特征的值修改为一组噪音值，通过比较特征修改前后的模型预测结果，比较得出该特征的重要性[29]。具体而言，特征重要性的计算过程如下：

输入：训练好的预测模型f，以及模型用到的训练集D_{train}。
步骤 1：计算模型f参考评价指标，本文选择 $F1$ 指标。
步骤 2：对于训练集D_{train}上的每个特征j：
　步骤 2.1：重复K次如下操作，对于其中的第k次：
　　随机打乱该列特征产生噪音，从而获得该特征损坏的数据集$\widetilde{D_{k,j}}$
　　计算在损坏的数据集$\widetilde{D_{k,j}}$上模型的评价指标 $F1$：$F1_{k,j}$
　步骤 2.2：计算并输出特征j的重要性i_j：

$$i_j = F1 - \frac{1}{K}\sum_{k=1}^{K} F1_{k,j} \tag{7}$$

输出：对于每个数据集中的特征j，由高到低排列的特征重要性i_j

图 7 给出了重要性排名前五位的特征。由此可以推测，在机器学习绩效评估模型中，以 SCI 加权的影响因子、CSSCI 论文数所代表的学术成果的水平是考虑的重要因素，人才培养也是区分项目绩效评估的主要因素。此外，机器学习模型也认为该项目经费合申请人之前项目的评估结果也是较大的影响因素。

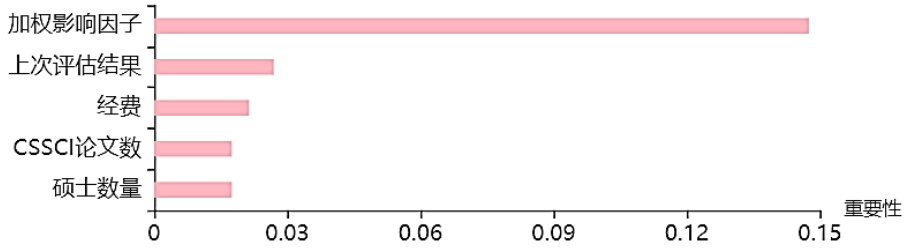

图 7 重要性排名前五位的特征

由于本文的数据来自以往已经结题的项目，所以上述结果也从侧面进一步印证了以往项目对 SCI 论文较为重视的现象[30]。2020 年，教育部和科技部联合印发《关于规范高等学校 SCI 论文相关指标使用 树立正确评价导向的若干意见》，对破除论文“SCI 至上”提出明确要求，应积极探索建立科学的评价体系，引导评价工作突出科学精神、创新质量、服务贡献。尽管 SCI 论文数量加权的影响因子在本文模型中是一个重要的指标，但是依然可以发现人才培养等指标也起到了关键的作用。由此可见，我国的科研评价体系正在逐渐摆脱“SCI 至上”等刻板的量化指标，逐渐建立起多方位的综合评价体系。

6.3 将项目摘要文本扩充到原数据集

本文在 4.1.2 节还构建了项目的摘要文本数据集D_{text}，但在前面的分析中，并没有将这些数据纳入D。接下来，将D_{text}与前面的数据集D合并，再次重复前面的数据分析过程，考察项目的摘要信息是否对预测模型的性能有提升作用。数据分析结果如表 4 所示，可见，加入文本特征后，各项指标或者没有改进，反而因为文本特征的噪声影响了原来的模型效果。总之，项目摘要的文本特征没有起到明显分类作用，加入文本特征与否，对本文的数据分析结论影响不大。

表 4 加入文本特征后的模型评估结果

随机森林使用的数据集	准确率A	查全率R	查准率P	$F1$值
D	0.78	0.87	0.73	0.78
$D_{text} \cup D$	0.76	0.86	0.71	0.76

总而言之，上述数据实验结果验证了本文提出的数据驱动的科研项目绩效评估框架的有效性，本文方法对项目评估结果具有较高的预测准确性。对科研管理部门而言，本文模型能够对项目的绩效水平做出预评估，可以自动化地在项目提交结题报告时做出一个大致评价，有助于减轻管理部门和评估专家的工作量，提高项目评估的工作效率。

7 结论与展望

本文提出了一个数据驱动的科研项目绩效评估方法，通过多分类监督学习算法实现项目绩效的有效估计。在本文的评估方法中，设计了网络爬虫自动化的从多个不同的网站渠道收集项目相关数据，进而对收集到的结构化和非结构化数据进行处理与融合，从中提取出有效的特征，基于一系列多分类监督学习算法实现评估模型的构建，并利用多种指标对模型的性

能进行估计。以国家自然科学基金项目数据为研究对象，对上述数据驱动的科研项目绩效评估方法进行了验证。结果显示，在 11 种分类算法中，随机森林算法的综合表现最好。SCI 加权的影响因子和申请人上次项目评估结果以及人才培养情况，在很大程度上决定了项目绩效评估的结果，而项目申请书中的摘要文本，对项目结题后的评估结果影响不大。

总之，本文所提的数据驱动的科研项目绩效评估方法为项目绩效的评估提供了一个统一且易于扩展的框架，本文的评估框架有助于提升项目评估工作的自动化与智能化水平，为科研项目管理部门提供数据化的决策支持。未来的研究工作将进一步融合更丰富数据，探索更为有效的基于新型机器学习算法的项目绩效评估方法。

参考文献

[1] 周寄中,杨列勋,许治.关于国家自然科学基金管理科学部资助项目后评估的研究[J].管理评论,2007,19(03):13-19+63.

[2] 王小霞.大数据时代科研管理信息化对策研究[J].科研管理,2019,40(10):282-288.

[3] 鲁晶晶,谭宗颖,万昊.关于科技项目成果评估研究内容的分析与思考[J].科学管理研究,2016,34(01):37-41.

[4] 姜群.基于 DEA-malmquist 的基础研究投入效率及其影响因素研究[J].科学管理研究,2019,37(03):35-41.

[5] 田人合,张志强,高志.基于分段线性模型的科研项目论文产出评价研究——以杰青基金地球科学项目为例[J].科技进步与对策,2019,36(01):142-151.

[6] 史童,杨水利,王春嬉,谭文俊.科技成果转化政策的量化评价——基于 PMC 指数模型[J].科学管理研究,2020,38(04):29-33.

[7] 黄璐,朱一鹤,陈丽,郑永和.科学基金资助 F0701 的科学计量分析[J].科学学研究,2019,37(06):977-985.

[8] 李瑛,孙涛.高校科研管理绩效评价——基于 2007—2010 年数据的实证研究[J].江西社会科学,2013,33(01):218-221.

[9] EILAT H, GOLANY B, SHTUB A. R&D project evaluation: An integrated DEA and balanced scorecard approach[J]. Omega, 2008, 36(5):895-912.

[10] CRAIG S. GALBRAITH, ALEX F. DENOBLE, SANFORD B. EHRLICH, et al. Can experts really assess future technology success? A neural network and Bayesian analysis of early stage technology proposals[J]. The Journal of High Technology Management Research, 2006, 17(2):125-137.

[11] JOHNES J, Li Yu. Measuring the search performance of Chinese higher education institutions using data envelopment analysis[J]. China Economic Review, 2008, 19(4): 679-696.

[12] HSU F M, HSUEH C C. Measuring relative efficiency of government-sponsored R&D projects: A three-stage approach[J]. Evaluation & Program Planning, 2009, 32(2): 178-186.

[13] 张华波,郝平,金永夫,郑国全.基于 DM 的科技计划项目绩效评价系统的设计[J].控制工程,2009,16(S3):114-116.

[14] 刘宝涛,刘帅.中国高等农业院校科研绩效评价研究——基于正态云模型的实证[J].中国农机化学报,2019,40(06):230-236.

[15] COSTANTINO F, DI GRAVIO G, NONINO F. Project selection in project portfolio management: An artificial neural network model based on critical success factors[J]. International Journal of Project Management, 2015, 33(8):1744-1754.

[16] LIU Fang, CHEN Yuwang, YANG Jianbo, et al. Solving multiple-criteria R&D project selection problems with a data-driven evidential reasoning rule[J]. International Journal of Project Management, 2019, 37(1):87-97.

[17] JANG H. A decision support framework for robust R&D budget allocation using machine learning and optimization[J]. Decision Support Systems, 2019, 121(7): 1-12.

[18] 周中华,张惠然,谢江.基于 Python 的新浪微博数据爬虫[J].计算机应用, 2014, 34(11): 3131-3134.

[19] 姜文,刘立康.基于 Selenium 的 Web 软件自动化测试[J].计算机技术与发展,2018,28(09):47-52+58.

[20] 陈晓田,黄海军,李若筠.绩效评估——切实加强科学基金面上资助项目后期管理的有效途径[J].中国科学基金,2004,18(03):60-62.

[21] SINGHAL A, KASTURI R, SRIVASTAVA J. Automating Document Annotation Using Open Source Knowledge[C]// The 2013 IEEE/WIC/ACM International Conference on Web Intelligence. ACM, 2013.

[22] AYUSH, SINGHAL, JAIDEEP, et al. Research dataset discovery from research publications using web context[J]. Web Intelligence & Agent Systems, 2017, 15(3):81-99.

[23] AKASH V PAVASKAR, AKSHAY S ACHHA, ANOOP R DESAI, et al. Information Extraction From Images Using Pytesseract and NLTK[J]. International Journal of Emerging Technologies and Innovative Research. 2017, 4(5): 83-84.

[24] 曾湘宁,沈兰生,任鲲鹏.印刷表格文本分析识别系统的研究[J].中文信息学报,1997,11(04):34-42.

[25] 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.

[26] SIMSEK S, KURSUNCU U, KIBIS E, et al. A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival[J]. Expert Systems with Application, 2020, 139(1):112863.1-112863.13.

[27] 周志华.机器学习[M].清华大学出版社:北京,2016:28-35.

[28] SWAMI A, JAIN R. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2013, 12(10):2825-2830.

[29] BREIMAN L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[30] 王芹,崔卫芳.从 SCIE 收录论文视角看高校基础研究影响因素分析[J].科学管理研究,2018,36(06):45-49