

# 应用统计 与数据科学

Applied Statistics and Data Science



ART AND DESIGN PRESS INC.

(626 810 4480)

119 S Atlantic Blvd, Suite 300D

Monterey Park, CA 91754

Copyright © 2026 by ART AND DESIGN PRESS INC.

Complimentary Copy



## 编委会名单

### 主 编



王国长 (Guochang Wang), 暨南大学 (Jinan University),  
邮箱: wanggc023@amss.ac.cn  
研究兴趣: 函数型数据分析、时间序列、充分性降维、统计与机器学习  
functional data analysis, time series, sufficiency dimension reduction, statistics and machine learning

### 执行副主编



张兴发 (Xingfa Zhang), 广州大学 (Guangzhou University),  
邮箱: xingfazhang@gzhu.edu.cn  
研究兴趣: 金融统计、环境统计、机器学习  
financial statistics, environmental statistics, machine learning

### 编 委



胡志勇 (Zhiyong Hu), 广州大学 (Guangzhou University)  
邮箱: zhyhu65@163.com  
研究兴趣: 大数据, 人工智能, 财务与会计  
big data, artificial intelligence, finance and accounting



雷田礼 (Tianli Lei), 深圳职业技术大学 (Shen Zhen Polytechnic University)  
邮箱: ltl@szpu.edu.cn  
研究兴趣: 数量经济、高职数学教育  
auantitative economics, higher vocational mathematics education



舒连杰 (Lianjie Shu), 澳门大学 (University of Macau),  
邮箱: ljshu@um.edu.mo  
研究兴趣: 量化金融, 统计学习, 质量控制及管理  
quantitative finance, statistical learning, quality control and management



王纬 (Wei Wang), 山东财经大学 (Shandong University of Finance and Economics)  
邮箱: wangwei\_0115@sdufe.edu.cn  
研究兴趣: 统计机器学习、应用统计、计量经济  
statistical machine learning, applied statistics, econometrics



杨凯 (Kai Yang), 长春工业大学 (Changchun University of Technology)  
邮箱: yangkai@ccut.edu.cn  
研究兴趣: 时间序列分析、高维数据分析、贝叶斯分析  
time series analysis, high-dimensional data analysis, bayesian analysis



周彦 (Yan Zhou), 深圳大学 (Shenzhen University)  
邮箱: zhouy1016@163.com  
研究兴趣: 生物统计, 数据科学  
Biostatistics, data science



李永明 (Yongming Li), 上饶师范学院 (Shangrao Normal University)  
邮箱: lym1019@163.com  
研究兴趣: 非参数统计, 金融统计, 极限理论与统计大样本性质, 小波方法  
nonparametric statistics, financial statistics, limit theory and statistical large sample theory, wavelet method



刘照德 (Zhao de Liu), 广东财经大学 (Guangdong University of Finance & Economics)  
邮箱: lzhaode@163.com  
研究兴趣: 经济统计、多元统计、计量分析  
economic statistics, multivariate statistics, econometric analysis



王小刚 (Xiaogang Wang), 北方民族大学 (North Minzu University)  
邮箱: wongone@163.com  
研究兴趣: 经济与社会统计  
economic and social statistics



严晓东 (Xiaodong Yan), 西安交通大学 (Xi'an Jiaotong University)  
邮箱: yanxiaodong@xjtu.edu.cn  
研究兴趣: 统计机器学习、智能体统计学  
statistical machine learning, agent statistics



朱柯 (Ke Zhu), 香港大学 (The University of Hong Kong)  
邮箱: mazhuke@hku.hk  
研究兴趣: 统计建模、金融时间序列分析、计量经济、金融大数据  
statistical modeling, financial time series analysis, econometrics, financial big data

# 应用统计与数据科学

Applied Statistics and Data Science

第2卷 第1期 2026年1月刊

主办单位: ART AND DESIGN PRESS INC

(United States)

承办单位: 广州统计师事务所

指导单位: 广东省现场统计学会、广州市统计学会

主 编: 王国长

执行副主编: 张兴发

编 委: 胡志勇、李永明、雷田礼、刘照德、  
舒连杰、王小刚、王 纬、严晓东、  
杨 凯、朱 柯、周 彦

地 址: 119 S Atlantic Blvd, Suite 300D Monterey  
Park, CA 91754

网 址: <https://www.artdesignp.com>

## 本刊说明:

凡向本刊所投稿件, 全体作者需签署论文著作权  
转让声明书和论文发表承诺书, 声明、承诺及相关事  
项如下:

- 作者将论文的复制权、发行权、网络传播权、翻  
译权、汇编权、信息网络传播权、改编权等著作  
权在世界范围内免费转让给本刊。
- 论文不侵犯他人著作权和其他权利, 否则作者将  
承担由此产生的全部责任, 并赔偿由此给出版单  
位造成的全部损失。
- 论文署名作者享有该作品的完全著作权, 署名作  
者的身份真实。
- 论文未曾以任何形式公开发表过。
- 作者所投本刊稿件, 本刊编辑部拥有修改权。



## 管理统计 | MANAGEMENT STATISTICS

- 001 国企债权管理透明化建设与资金流动性优化路径探讨 张悠佳  
Transparent Construction of State-Owned Enterprise Credit Rights  
Management and Optimization Paths for Fund Liquidity Zhang Youjia
- 006 战略导向下全面预算管理体系的优化路径与实践 张嘉琪  
Optimization Path and Practice of a Strategy-Oriented Comprehensive Budget  
Management System Zhang Jiaqi
- 010 中小企业智慧财务系统上线对财务的影响 谭梓颖  
Impact of the Implementation of Intelligent Financial Systems on the Financial  
Management of Small and Medium-sized Enterprises Tan Ziyang

## 统计与教育教学 | STATISTICS AND EDUCATION TEACHING

- 014 基于回归分析与随机森林的初中数学成绩影响因素分析 李冰  
Determinants of Middle-School Mathematics Achievement: A Regression-Random-  
Forest Hybrid Analysis Li Bing

## 数据科学技术、方法与应用 | DATA SCIENCE TECHNOLOGY, METHODS AND APPLICATIONS

- 019 基于 TF-IDF 和 DeBERTa 混合模型的数据目录语义检索系统研究 谷剑芳  
Research on a Hybrid TF-IDF and DeBERTa Model for Semantic  
Retrieval in Data Catalog Systems Gu Jianfang
- 025 融合舆情的黄金期货价格预测:  
RFECV 特征选择与 DNN 模型 杨紫帆, 赵艺晴, 徐浩然, 潘晨宇, 肖尧  
Gold Futures Price Forecasting Incorporating Public Sentiment: RFECV Feature  
Selection and DNN Model Yang Zifan, Zhao Yiqing, Xu Haoran, Pan Chenyu, Xiao Yao
- 030 基于 TREC 真实邮件数据集的朴素贝叶斯分类教学  
案例构建与应用效果实证研究 曹寒问, 陈锦文, 车金星, 张毓华  
Empirical Study on the Construction and Application Effectiveness of a  
Naive Bayes Classification Teaching Case Based on the TREC Authentic  
Email Dataset Cao Hanwen, Chen Jinwen, Che Jinxing, Zhang Yuhua
- 035 线性秘密共享中的自对偶结构 林群  
On the Self-Dual Structure in Linear Secret Sharing Lin Qun
- 038 基于深度学习的恶意代码检测算法设计与实现 张家华  
Design and Implementation of Malicious Code Detection Algorithm  
Based on Deep Learning Zhang Jiahua

## 统计方法及应用 | STATISTICAL METHODS AND APPLICATIONS

- 042 处理效应的异质性的识别  
——基于因果机器学习的仿真和估计 晏发发, 颜丽金, 陈铮浩  
Identification of Treatment Effect Heterogeneity—Simulation and  
Estimation Based on Causal Machine Learning Yan Fafa, Yan Lijin, Chen Zhenghao
- 048 一类自适应权重下的混频逻辑回归模型及其应用研究 呼可可, 刘梦丽  
A Type of Mixed-frequency Logistic Regression Model under  
Adaptive Weights And Its Application Hu Keke, Liu Mengli

|     |   |   |
|-----|---|---|
| 053 | <b>基于多模型的胎儿染色体异常判定的探究</b><br>Research on the Determination of Fetal Chromosomal Abnormalities Based on Multiple Models                                      | <b>魏婧, 刘昱贤, 皮哲豪, 李乃医</b><br>Wei Jing, Liu Yuxian, Pi Zhehao, Li Naiyi   |
| 057 | <b>基于 Kruskal-Wallis 算法的卫生巾市场消费现状与需求分析</b><br>Analysis of Consumption Status and Demand in the Sanitary Napkin Market Based on the Kruskal-Wallis Algorithm | <b>朱金桂, 高坚国, 陈心语, 詹富婕, 刘芸, 周大镛, 林培琪</b><br>Zhu Jingui, Gao Jianguo, Chen Xinyu, Zhan Fujie, Liu Yun, Zhou Dazhuo, Lin Peiqi |

# 国企债权管理透明化建设与资金流动性优化路径探讨

张悠佳

贵州蔬菜集团有限公司, 贵州 贵阳 550001

DOI:10.61369/ASDS.2026010001

**摘 要：** 在当前复杂的经济环境下，债权资产管理与资金流动性安全已成为影响国有企业可持续发展的关键因素。本文以区域性具有一定民生保障职能的企业 S 集团为研究对象，通过深入分析其债权管理与资金流动现状，揭示了因债权管理粗放导致的应收账款规模膨胀、资金使用效率低下等问题。研究基于战略管理理论，提出了将债权管理与公司战略深度融合的解决方案，构建了包含组织体系、数据库建设、流程优化和信息集成的系统性管理框架。该体系通过建立透明化的债权管理机制，实现了对债权风险的精准识别和有效管控，同时显著改善了企业资金流动性状况。本文的研究成果为同类国有企业提供了具有实践价值的参考范式。

**关 键 词：** 债权管理透明化；资金流动性；风险管控；业财融合；战略管理

## Transparent Construction of State-Owned Enterprise Credit Rights Management and Optimization Paths for Fund Liquidity

Zhang Youjia

Guizhou Vegetable Group Co., Ltd., Guiyang, Guizhou 550001

**Abstract：** In the current complex economic environment, the management of credit assets and the security of fund liquidity have become critical factors influencing the sustainable development of state-owned enterprises. This study takes Enterprise S Group, a regional entity with certain public welfare functions, as the research object. Through an in-depth analysis of its credit management and fund flow status, the study reveals issues such as the expansion of accounts receivable scale and inefficient use of funds caused by extensive credit management. Based on strategic management theory, the research proposes a solution that deeply integrates credit management with corporate strategy, constructing a systematic management framework that includes organizational systems, database development, process optimization, and information integration. By establishing a transparent credit management mechanism, this system enables precise identification and effective control of credit risks while significantly improving the enterprise's fund liquidity. The findings of this study provide a practical reference model for similar state-owned enterprises.

**Keywords：** transparent credit management; fund liquidity; risk control; industry-finance integration; strategic management

## 引言

随着市场经济环境的深刻变化和国有企业改革进程的深入推进，国有企业在实现规模扩张的同时，面临着日益严峻的财务风险挑战。特别是在当前经济增速放缓的背景下，债权资产质量下降和资金链紧张已成为制约国有企业高质量发展的突出问题。本研究采用案例研究法与数据分析法相结合的研究范式，以 S 集团为典型案例（业务覆盖团餐配送、农产品贸易、生产基地建设，客户含政府部门、农业合作社等，受季节、政策回款影响显著），该集团作为省属重点民生保障企业，近年来业务规模快速扩张，但在营收持续增长的同时，也暴露出应收账款增速过快、资金使用效率低下、经营活动现金流持续承压等管理困境<sup>[1]</sup>。这些问题的深层次原因在于传统的债权管理模式已无法适应企业发展的需要，如此，建立一套与现代企业治理结构相适应的债权管理体系的重要性就凸显出来了。本文基于 S 集团的管理实践，从战略高度系统探讨债权管理透明化建设与资金流动性优化的协同路径，以期同类企业提供有益借鉴。

作者简介：张悠佳（1993.12-），湖南邵阳人，会计师，研究方向：国有企业集团财务管控体系优化研究、基于现金流安全的企业全面预算与融资策略协同研究、业财融合视角下的国有企业债权债务全流程治理研究。

## 一、S集团债权管理与资金流动性现状及问题分析

### （一）企业战略定位与业务发展特征

S集团作为区域性民生保障的重要力量，承担着保障基础民生供应的重要使命。公司以“团餐配送、蔬菜贸易、生产基地建设”三大业务板块为核心，形成了覆盖全产业链的业务格局。随着“十四五”规划的实施，公司进入了快速发展阶段，资产规模与营业收入均保持较快增长。然而，在规模扩张的背后，公司的资产结构却呈现出明显的不均衡态势。从趋势看，公司应收账款规模增长迅猛。2024年末余额（40,205.15万元）较2023年末（43,845.40万元）虽因结算等因素略有下降，但2025年预算显示将大幅反弹。与此同时，营业收入从2024年的285,392.21万元预算至2025年的322,463.45万元（同上来源），营收增长对债权资产的依赖度值得关注。

**行业特性与债权管理挑战：**作为以团餐配送（2024年收入占比54.85%）和农产品贸易为核心业务的省属民生保障企业，S集团的债权结构具有鲜明特点。决算分析明确指出，“团餐项目应收占比重大，主要因为政府核算中心核算流程较长，且当地教育局资金紧张，部分县（市）回款周期较长”。这表明，政策性、制度性回款障碍是导致应收账款高企的重要外部原因，非单纯的内部管理问题。

### （二）债权管理体系存在的深层次问题

通过对S集团管理实践的深入剖析，其债权管理存在多个维度的系统性问题。首先，在战略层面，债权管理目标与公司整体发展战略严重脱节。公司在制定业务扩张计划时，未能同步建立与之匹配的债权风险管控机制，导致业务规模增长以应收账款快速积累为代价。其次，在组织架构方面，债权管理责任体系存在明显缺陷。当前的管理模式将债权管理简单地界定为财务部门的专业职能，业务部门仅关注合同签订和业务执行，缺乏对应收账款回收的责任意识<sup>[2]</sup>。这种“业财分离”的管理模式直接导致了前端的业务决策与后端的资金回收严重脱节。

例如，在快速扩张的团餐配送板块（2025年收入预算175,927.42万元），业务部门为抢占市场份额、完成营收目标，可能与回款保障不足的地区教育局签约，而财务部门在合同签订前难以介入信用评估，导致后端形成大规模政府类应收账款。在农产品贸易板块（2025年收入预算182,800万元），为达成贸易额，可能放松对交易对手的资信审查，埋下坏账风险。

此外，在管理基础方面，债权管理信息化建设严重滞后。各子公司使用独立的业务系统，债权数据分散且标准不一，集团层面难以获取准确、及时的全球数据。这种信息碎片化状态使得管理层无法对债权风险进行有效识别和预警。最后，在激励机制方面，现有的绩效考核体系过于侧重收入和利润指标，未能将债权质量纳入关键绩效指标，导致业务单元缺乏管控债权风险的内在动力。

### （三）债权问题对资金流动性的传导效应

债权管理不善已对S集团的资金流动性产生严重的负面影响。最直接的表现是经营活动现金流持续恶化。根据公司预算数据，2025年经营活动现金流量净额预计为-28,821.34万元，这种“有

利润无现金”的经营状态充分暴露了公司在资金管理方面的深层次问题。更为严重的是，为弥补日常运营的资金缺口，公司不得不大幅增加外部融资规模<sup>[3]</sup>。融资结构分析显示，短期融资占比高达82.53%，这种期限结构错配不仅推高了融资成本，更增加了企业的流动性风险，其具体的预算融资结构如图1所示。2025年预算资产负债率预计上升至54.71%，较2024年提高7.35个百分点，这一变化趋势值得高度警惕。

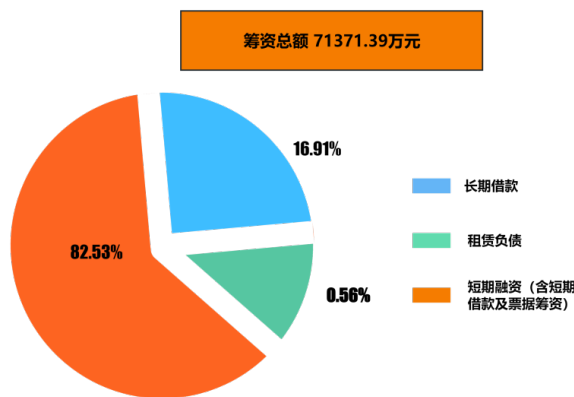


图1：S集团2025年预算融资结构饼状图

## 二、债权管理透明化与资金流动性优化的协同策略构建

为解决S集团面临的债权管理粗放与资金流动性紧张并行的困境，构建了一套系统性的协同管理策略，该策略以“四位一体”为核心框架，涵盖从顶层设计到落地实施的完整链条，旨在通过债权管理的透明化，从根本上优化企业的资金流动性状况。

### （一）核心理念：构建“四位一体”的协同管理框架

为实现债权管理与资金流动性的有效协同，核心在于确立“四位一体”的协同管理框架：即以战略为引领，以债权透明化为基石，以资金流动性优化为目标，以信息系统为引擎。其内在逻辑是一个闭环的价值创造过程，如图2所示：战略导向确保方向正确；债权透明化通过精准的风险识别和资产盘活，为资金流动性优化提供输入；资金状况的改善又通过预算与考核机制，反馈并约束前端的业务策略与债权发生；而信息系统则作为底层支撑，赋能整个循环高效、自动运转<sup>[4]</sup>。

### （二）战略引领：将债权与现金流安全嵌入公司顶层设计

为扭转业务扩张与资金回笼脱节的局面，S集团需要将债权与现金流安全理念深度嵌入公司战略。依据战略管理理论中的“资源基础观”，企业可持续竞争优势来源于其稀缺、有价值、难以模仿的资源。S集团将债权资产视为关键财务资源，通过客户全生命周期价值管理实现资源的优化配置——对回款记录良好的优质客户给予信用额度倾斜（资源优先投放），对长期拖欠的高风险客户限制合作（资源风险规避），从而在业务扩张与资金安全之间建立战略平衡。其目标分解与衔接的逻辑可如图3所示。

1. 目标分解与预算衔接：可行的路径是将“优化资产结构、



保障现金流安全”的总体战略目标，分解为“应收账款周转率”、“经营性净现金流”等关键绩效指标（KPI），并通过全面预算管理，将这些 KPI 转化为各业务单元必须遵守的刚性财务约束。例如，S 集团在 2025 年预算中已计划将融资计划纳入全面预算管理（见预算说明书“负债管控情况”），这正是战略引领下财务约束的具体体现。未来可进一步将“应收账款周转率”、“经营性净现金流”等从预算中的管控指标，升级为与业务部门绩效考核强关联的战略执行指标。

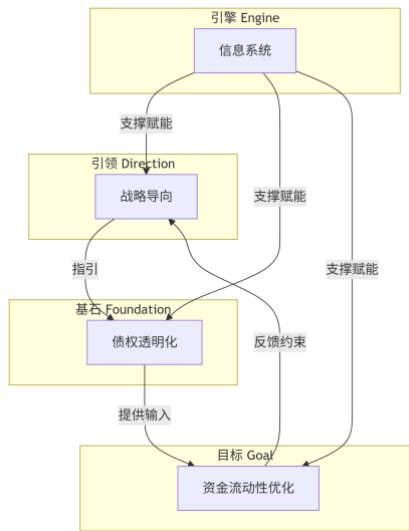


图2：“四位一体”协同管理框架图

2. 引入客户全生命周期价值管理：此举旨在引导业务部门从追求单一订单收入，转向追求客户的长期现金流价值。建立客户信用档案，对回款记录良好的优质客户给予更多支持；对长期拖欠、风险高昂的客户，即使其订单额大，也需采取审慎策略，从源头上控制不良债权的产生。

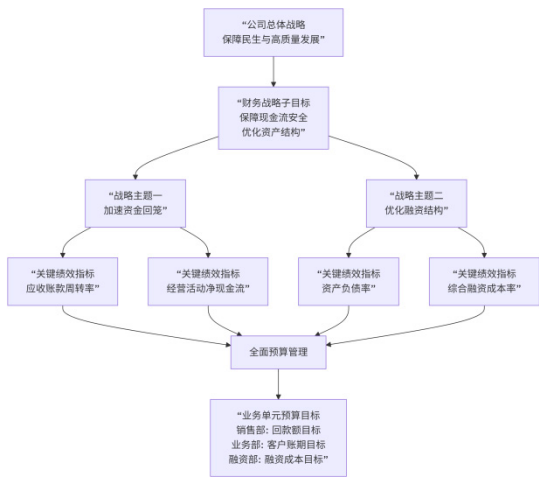


图3：战略目标传导与绩效指标联动图

（三）组织保障：构建跨职能、全流程的责任共同体

有效的管理离不开清晰的组织和责任。S 集团应构建一个从决策层到执行层的三级责任体系。

1. 决策层（债权管理委员会）：由集团董事长或总经理牵头，作为最高决策机构，负责审批信用政策、债权管理目标及重大风险处置方案。

2. 管理层（风险管理部门）：由风险管理部门定期对债权进行追踪盘点，对逾期债权进行催收、函证、起诉，明确责任人，相关风险管理部门对逾期债权的相关责任人进行追踪，如发生相关责任人需离职的情况，必须签字确认无未完结事项后方可离职；

3. 执行层（业务专员）：各业务人员对自身业务负责，并与绩效挂钩，作为连接业务与财务的纽带，负责客户资信初评、合同条款审核、应收账款日常跟进与催收，将管理触点延伸至业务最前端。为进一步强化执行合力，建议在各业务单元设立“债权管理小组”，由业务负责人、财务专员、法务专员共同组成，负责本单位债权动态监控、逾期预警与联合催收。

考核与能力建设：相关执行层的绩效考核指标包括“客户信用评估准确率（目标≥90%）”，“对逾期债权清收进度达80%可发全额绩效、40%-80%之间按比例发放绩效，40%以下停发绩效”等方案。其次，岗位胜任能力要求具备财务基础知识、客户沟通能力与风险识别能力，培训机制可采用“财务部集中授课+业务部门轮岗实践”双轨制。

（四）数据驱动：建设统一、实时、可溯源的债权数据中台

超越简单的数据库概念，建设企业级的债权数据中台，是实现管理透明化的技术基石。对集团实际情况而言，短期投入压力较大。建议在第一阶段试点时，暂不开发完整中台，而是利用现有司库系统（已全级次推广）的模块功能，新增“债权管理”子模块（如客户信用档案、回款提醒功能）该中台应涵盖的核心数据域及其管理用途如表1所示<sup>[6]</sup>。

1. 全口径数据治理：制定统一的数据标准，将应收账款、合同资产、预付款项等全部纳入管理范围，明确数十个核心字段的定义与采集规范。

2. 生命周期可视化：通过系统集成，记录每一笔债权从“合同签订 -> 发货 / 提供服务 -> 开票 -> 回款 / 逾期”的全流程数据，实现完整溯源。

3. 分析能力服务化：在中台内置客户信用评分、账龄分析、逾期预测等数据分析模型，为各业务部门、风控及管理层提供标准化的数据服务与决策支持。

表1：债权数据中台核心数据域与用途

| 核心数据域   | 包含关键信息                         | 主要管理用途              |
|---------|--------------------------------|---------------------|
| 客户主数据   | 注册资本、股权结构、信用评级、历史回款记录、司法风险     | 客户准入评估、信用额度审批、差异化定价 |
| 合同数据    | 合同编号、产品 / 服务、付款条件、账期、违约责任、结算方式 | 合同合规评审、生成收款计划、履行监控  |
| 交易与账务数据 | 发票金额、开票日期、记账科目、账龄、逾期状态、催收记录    | 风险分类、计提坏账准备、制定催收策略  |
| 行为数据    | 客户沟通记录、满意度调查结果、市场舆情            | 客户关系维护、风险早期预警、优化服务  |

信息系统建设的资源保障：S 集团已在推进司库系统建设（见预算“司库建设工作情况”），此为债权数据中台的重要组成部分

分。参考2025年研发费用预算（500.96万元，较2024年激增）的模式，在后续年度专项预算中明确数据中台的建设经费。初步估算，一个覆盖全级次26户企业的债权数据中台，包含定制开发、数据治理与迁移、三年运维，总投资约在800–1200万元区间，可分三年投入。

（五）流程再造：建立端到端的风险导向管理流程

建议S集团建立基于风险识别的、贯穿业务始终的管理流程。团餐配送业务涉及生鲜食材采购，上游对接农户/合作社、下游对接学校/机关，上游农户预付款可与政府农业补贴政策联动（如凭补贴文件缩短预付款回收周期），下游政府类应收账款可绑定“民生保障资金预算拨付节点”，可实时跟踪民生保障资金（如团餐补贴、农产品采购补贴）的拨付进度。其中，对已形成债权的风险分类与应对是流程的关键环节，具体的策略矩阵如表2所示。

- 事前防范：建立严格的客户信用评估与合同评审制度。所有重大合同必须经过业务、财务、法务三方会审，重点关注付款条款。
- 事中监控：系统根据合同条款自动生成“收款日程表”，并提前向业务负责人推送提醒。对即将到期和已逾期账款，系统自动触发预警，升级管理措施。
- 事后处置：根据风险评估结果，对债权进行五级分类，并采取差异化的处置策略。

表2：债权风险五级分类与应对策略矩阵

| 风险等级 | 核心特征       | 预警信号 | 主要应对策略               | 责任主体   |
|------|------------|------|----------------------|--------|
| 正常   | 未到期，客户经营正常 | 无    | 常规跟踪，维护关系            | 业务人员   |
| 关注   | 逾期30天以内    | 黄色   | 电话/邮件提醒，了解原因         | 业务主管   |
| 次级   | 逾期31–90天   | 橙色   | 发送正式催收函，上门沟通，暂停新业务   | 风控部门   |
| 可疑   | 逾期91–180天  | 红色   | 启动法律程序，申请财产保全，考虑债务重组 | 法务与高管  |
| 损失   | 逾期180天以上   | 黑色   | 计提足额坏账准备，启动核销程序      | 管理层与审计 |

（六）资金协同：建立基于债权预测的动态资金管理模式

将债权管理的产出，直接应用于资金管理决策，形成协同效应。

- 现金流滚动预测：根据集团实际情况，部分业务受季节影响显著，应收账款规模易随季节波动（如开学季团餐订单激增导致应收短期上涨）。可建立“季节性现金流预测模型”，结合历史季节数据（如近3年团餐应收增速、农产品贸易回款率，优化滚动预测的准确性，避免因季节因素导致资金缺口误判。编制未来13周（一个季度）的周度现金流滚动预测，极大提升资金可视性。
- 融资策略动态调整：当预测显示未来有大量现金流入时，可提前规划偿还高成本短期贷款；当预测出现资金缺口时，可提前与金融机构沟通，争取优惠融资条件。
- 供应商付款优化：下游客户因“食材损耗超标”拒付部

分款项（如学校因蔬菜腐烂扣减5%货款），易形成“争议性应收”。在进销存系统中新增“损耗争议记录”，关联对应的应收金额，并制定“损耗争议处理流程”（如24小时内现场核实、与客户协商折让比例、签订补充协议），避免争议演变为坏账。同时，在“效益评估”的“不良债权比率”指标中，单独统计“损耗争议类应收占比”，针对性监控。将自身的应收账款回收预测，与对供应商的付款账期进行动态匹配，在维护供应链稳定的前提下，平滑现金流出，优化营运资本。

（七）实施路线图：分阶段推进体系建设

为确保变革平稳落地，建议S集团采用“小步快跑、迭代优化”的策略，分三个阶段推进。详细的实施时间线与里程碑规划如图4所示。



图4：S集团债权管理体系实施路线图

- 第一阶段：基础建设与试点期（1–6个月）  
核心目标：搭建框架、验证流程。  
关键任务：成立领导小组与工作组；完成核心制度设计；完成数据中台核心模块开发与试点部署。具体为：逐步推广试点，因集团子公司业务差异大（如涉及学生营养餐配送、蔬菜贸易、基地种植、人才培养等），第一阶段（基础建设与试点期），优先选择“团餐业务占比超60%、政府客户占比超70%”的子公司（学生团餐）试点，此类子公司应收问题最突出，试点效果更易显现；第二阶段再推广至农产品贸易、生产基地类子公司，避免资源分散导致推进缓慢。
- 第二阶段：全面推广与深化期（7–12个月）  
核心目标：全域覆盖、深化应用。  
关键任务：制度全集团强制推行；系统全级次上线；完成全员培训；将核心KPI初步纳入业绩考核。具体为：正式推行分类适配的债权管理制度，完成数据中台在全级次子公司（包括信息化基础较弱的农业生产单元）的全覆盖部署，并开展针对业务与财务人员的分层实战培训。同时，将应收账款周转天数等核心指标按权重纳入各单位年度考核，并设置半年过渡期，以确保管理体系扎实落地、稳步见效。
- 第三阶段：优化与智能化期（13–18个月）  
核心目标：固化成效、价值创造。  
关键任务：基于数据反馈持续优化流程；深化信用评分与预测模型；实现与司库、预算等系统的深度集成；全面严格地将考核结果与激励约束机制挂钩。具体为：针对前期积累的全量数据，重点对农产品贸易的价格波动结算风险、种植基地的长周期合同履行风险等业务痛点，建立专项分析模型，驱动流程优化。同时，将信用模型从针对核心客户，扩展至覆盖供应商、渠道商等全部合作伙伴，实现风险动态监控。通过打通与司库系统的数据接口，实现应收账款预测与资金调拨计划的自动联动，构建管理闭环。最终，考核将从子公司负责人层面穿透至业务团队与关键岗位个人，并设立“风险准备金计提与追索”机制，使风险收



益相匹配，彻底固化风控文化。

三、效益评估、挑战与持续优化路径

（一）构建多维效益评估体系以衡量改革成效

为科学评估债权透明化管理体系建设的成效，S 集团应建立一套涵盖风险、资金、运营和客户等多维度的效益评估指标体系，用于持续跟踪与考核。

表 3：债权管理透明化建设效益评估指标体系

| 评估维度        | 核心指标             | 计算方法 / 说明                               | 目标值（示例）       | 考核频率 |
|-------------|------------------|---|---------------|------|
| 风险管控<br>维度  | 不良债权<br>比率       | (账龄 90 天以上应<br>收账款 / 应收账<br>款总额) × 100% | 降至 5% 以下      | 月度   |
|             | 重大风险债<br>权发生率    | 新发生的“重大<br>风险”级债权笔<br>数                 | 0             | 季度   |
| 资金效益<br>维度  | 应收账款周<br>转率      | 营业收入 / 平均<br>应收账款余额                     | 提升 25% 以上     | 季度   |
|             | 经营活动现<br>金流量净额   | 来自现金流量表                                 | 由负转正并持续<br>改善 | 月度   |
|             | 带息负债规<br>模       | 来自资产负债表                                 | 保持稳定或稳步<br>下降 | 季度   |
| 运营效率<br>维度  | 平均回款周<br>期 (DSO) | 应收账款总额 /<br>(营业收入 / 365)                | 缩短 15 天以上     | 月度   |
|             | 债权数据准<br>确率      | (1 - 数据错误笔<br>数 / 总数据笔数)<br>× 100%      | ≥ 98%         | 月度   |
|             | 逾期账款回<br>收率      | 本期收回的逾期<br>金额 / 期初逾期<br>总额              | ≥ 30%         | 季度   |
| 客户与战<br>略维度 | 优质客户留<br>存率      | (1 - 优质流失客<br>户数 / 总优质客<br>户数) × 100%   | ≥ 95%         | 年度   |
|             | 业财融合满<br>意度      | 通过业务与财务<br>部门双向调研评<br>分                 | 显著提升          | 半年度  |

“供应链应收周转效率” 指标：计算“(团餐应收 + 贸易应收) / (生鲜食材采购成本 + 配送成本)”，该指标反映应收与供应链核心成本的匹配度——若指标下降，说明应收占用资金未随供应链成本增长同步增加，符合蔬菜集团“以供应链效率为核心”的管理目标。建议将其纳入“运营效率维度”，考核频率为半年度。

该指标体系不仅用于事后评价，更应用于过程管理。管理层可通过指标看板实时监控运行状态，及时发现偏差并采取纠偏措施。同时，将评估结果与绩效激励紧密关联，能够有效保障管理

体系的生命力与持续优化。

（二）实施过程中可能面临的挑战

在体系建设过程中，S 集团需要清醒认识到可能遇到的困难。首先是系统整合的技术挑战，新旧系统的对接和数据迁移需要专业的技术支持和充足的实施周期。其次是组织变革的阻力，新的管理模式要求改变各部门长期形成的工作习惯，可能遭遇隐性抵制。再次是专业能力的短板，特别是业务人员的财务风险意识和业务人员的专业素质都需要全面提升。最后是管理成本的增加，系统建设、人员培训等都需要投入大量资源，短期内可能增加企业的运营成本。

（三）持续优化的方向与路径

针对上述挑战，S 集团需要制定系统的应对策略。在技术层面，采取分步实施的策略，优先解决最紧迫的管理痛点，确保系统平稳过渡。在组织变革方面，加强宣导和培训，通过试点先行的方式积累经验，逐步推广。在能力建设方面，建立分层级的培训体系，重点加强业务人员的财务知识和财务人员的业务理解。在资源保障方面，做好投入产出分析，确保资源投入与预期效益相匹配。同时，要建立持续改进机制，定期评估体系运行效果，根据内外部环境变化动态优化管理策略<sup>[6]</sup>。

（四）试点实施与初步成效

本方案的核心思路已在 S 集团 2025 年预算编制与管控措施中得到部分体现。例如，集团已明确将加强“两金”管控，从事前（《客商授信管理办法》）、事中（合同台账与预警）、事后（《债权管理办法》）建立全流程管理（见预算说明书）。同时，司库系统已在全级次推广，为实现债权数据的统一归集与动态监控奠定了技术基础。建议在 2026 年，可选取农产品贸易板块（农发供应链公司）或新增的团餐配送公司作为全面试点单位，将文中提出的“风险五级分类与应对策略矩阵”等工具进行实地运行，一个季度后评估对回款周期和现金流的影响，为 2026 年全面推广积累经验、优化模型。

本方案基于 S 集团（民生保供型贸易流通企业）的实践，尤其适用于具有以下特征的国有企业：一是业务增长快，应收账款对流动资金占用大；二是客户结构中包含政府部门、事业单位或大型国企，回款受外部程序影响大；三是业务板块多元，存在“业财分离”管理痛点。对于纯生产制造类国企，需强化“存货”与“应收账款”的联动管理；对于平台类国企，需更关注政府回购款项的合同管理与预算衔接。S 集团的实践表明，将债权管理与全面预算、司库系统建设相结合，是此类企业优化资金流动性的可行路径。

参考文献

[1] 王春燕. 国有企业做好债权管理及清收工作的策略探究 [J]. 投资与创业, 2025, 36(11): 104–106.  
[2] 李晓光. 基于公司战略的 RD 集团债权管理体系构建 [J]. 财务与会计, 2022, (06): 41–44.  
[3] 余妮娜. 企业债权债务管理中的常见问题及优化措施 [J]. 商业 2.0, 2025, (12): 40–42.  
[4] 马凯. 国有企业资金流动性风险管理策略与优化路径研究 [J]. 环渤海经济瞭望, 2025, (07): 170–173.DOI: 10.16457/j.cnki.hbhjllw.2025.07.019.  
[5] 王欢欢. 风险管理在国有企业资金流动性管理中的应用 [J]. 中国市场, 2025, (31): 100–103.DOI: 10.13939/j.cnki.zgsc.2025.31.025.  
[6] 刘佳. 企业资金管控与流动性风险防范的协同策略 [J]. 纳税, 2025, 19(27): 79–81.

# 战略导向下全面预算管理体系的优化路径与实践

张嘉琪

东莞市石龙粤龙实业有限公司，广东 东莞 523320

DOI:10.61369/ASDS.2026010002

**摘 要：** 本文以 Y 集团的全面预算管理实践为研究对象。Y 集团作为镇属综合性投资运营平台，其业务兼具政策性与市场性，传统的预算编制与执行模式难以有效承接“聚焦重点产业、优化资本布局”的战略意图。本文旨在剖析集团当前预算管理与战略脱节、执行刚性不足、考评激励弱化等现实问题，进而构建一个以战略目标为起点、涵盖编制、执行、分析、考评全流程的优化体系。论文将重点阐述如何将战略解码为预算指标、如何强化过程监控、如何将预算执行与绩效管理深度融合，并设计一套符合镇属企业特点的、分阶段的实施路径，旨在为同类企业提升预算管理效能、保障战略目标落地提供一套行之有效的解决方案。

**关 键 词：** 全面预算管理；战略导向；优化路径；镇属企业；绩效挂钩

## Optimization Path and Practice of a Strategy-Oriented Comprehensive Budget Management System

Zhang Jiaqi

Dongguan Shilong Yuelong Industrial Co., Ltd., Dongguan, Guangdong 523320

**Abstract：** This paper takes the comprehensive budget management practice of Y Group as its research subject. As a town-owned comprehensive investment and operation platform, Y Group's business incorporates both policy-driven and market-oriented characteristics. Its traditional budgeting and execution models struggle to effectively support the strategic intent of "focusing on key industries and optimizing capital allocation." This paper aims to analyze practical issues within the group, such as the disconnection between current budget management and strategy, insufficient execution rigidity, and weakened evaluation incentives. Subsequently, it constructs an optimized system that starts with strategic objectives and encompasses the entire process of formulation, execution, analysis, and evaluation. The paper will focus on explaining how to decode strategy into budget indicators, how to strengthen process monitoring, how to deeply integrate budget execution with performance management, and will design a phased implementation path tailored to the characteristics of town-owned enterprises. The goal is to provide a set of effective solutions for similar enterprises to enhance budget management effectiveness and ensure the achievement of strategic objectives.

**Keywords：** comprehensive budget management; strategy-orientation; optimization path; town-owned enterprises; performance linkage

### 引言

在当前深化国资国企改革、推动经济高质量发展的宏观背景下，地方镇属企业的转型升级已进入关键阶段。传统的镇属企业正从单一的资产管理者和融资平台，向兼具政策使命与市场化运营能力的综合性实体转变。这一转变对其内部管理水平，尤其是资源配置的核心机制——全面预算管理，提出了前所未有的更高要求。然而，现实情况是，许多镇属企业的预算管理体系仍停留在传统模式，难以有效支撑战略目标的实现，其升级优化显得尤为迫切。

### 一、镇属企业战略落地呼唤全面预算管理升级

镇属企业普遍面临着战略执行层面的挑战。战略规划往往由

高层制定，但在向下传导的过程中，容易与年度预算编制和日常经营决策脱节，形成“战略归战略，预算归预算”的“两张皮”现象<sup>[1]</sup>。预算编制时常沦为场简单的数字汇总与讨价还价，而非基

于战略目标进行资源精准配置的管理活动。其后果是，有限的资本无法高效聚焦于战略重点领域，影响了企业核心竞争力的培育与整体运营效率的提升。同时，在预算执行过程中，普遍存在监控乏力、分析滞后、调整随意等问题，预算的刚性约束作用未能有效发挥，无法对经营风险进行及时预警和过程控制。此外，预算执行结果与组织及个人的绩效考核关联度弱，使得预算管理缺乏足够的权威性和激励性，最终导致战略目标在执行环节层层衰减。

因此，探索如何构建一套以战略为导向的全面预算管理优化路径，具有重要的现实意义。其核心在于，将预算管理从传统的成本费用控制工具，提升为贯穿战略规划、资源配置、执行监控和绩效评价全过程的战略落地工具。这要求企业的预算管理必须实现根本性的转变：从被动的事后核算，转向主动的事前预测和事中控制；从财务部门的单向指令，转向业务部门与财务部门的深度协同<sup>[9]</sup>；从与战略脱节的静态数字游戏，转向驱动战略实现的动态管理流程。

本研究旨在基于镇属企业转型升级的现实需求，深入剖析其预算管理体系中存在的 key 问题，并构建一个系统性的优化框架。研究将重点阐述如何将战略目标有效解码为预算指标，如何强化预算的执行刚性监控，以及如何实现预算与绩效的紧密挂钩，从而为提升镇属企业战略执行力与价值创造能力提供一套具有实践指导意义的解决方案。

二、Y 集团预算管理现状与战略脱节问题剖析

要构建有效的优化路径，必须首先对研究对象有清晰的认识。深入剖析 Y 集团作为镇属企业所承担的战略使命，并以此为标尺，检视其现行全面预算管理体系中存在的突出问题和薄弱环节，明确优化改进的精准靶向<sup>[9]</sup>。

Y 集团的战略定位对其预算管理提出了独特且更高的要求。作为镇属综合性运营平台，其业务呈现出明显的二元性：一方面需稳健经营现有资产，确保保值增值，履行政策性职能；另一方

面需积极投身产业孵化与重点项目建设，承担较高的市场风险以谋求长远发展。这种复杂的业务生态要求其预算管理体系必须具备高度的战略协同性、资源的前瞻配置能力以及对风险的动态管控能力。预算不应仅是成本费用的控制工具，更应成为承接镇域经济发展战略、引导资源向重点领域倾斜、评估项目投资效益的核心管理机制。

然而，对标上述要求，Y 集团现行的预算管理体系在实践中暴露出若干深层次问题，导致其战略支撑作用未能充分发挥<sup>[4]</sup>。首要问题在于战略传导机制失灵，预算与业务计划“两张皮”。集团的战略规划与年度预算编制在时序和逻辑上衔接不畅。预算编制往往由财务部门主导，倾向于在上年基数上进行调整，或由业务部门基于本位主义上报，缺乏将宏观战略目标有效解码、细化为各业务单元可执行、可量化的预算指标的具体路径。其结果便是，预算方案无法清晰反映“聚焦重点产业”的战略优先级，资源分配可能偏离战略方向。

其次，过程监控流于形式，预算执行刚性不足。预算批复后，其管理重心便转向事后核算，而非事中控制。虽然定期有财务报表反映预算执行情况，但分析往往停留在数字对比层面，缺乏对偏差背后业务动因的深入追溯。对于预算外支出或重大偏差，缺乏有效的预警和快速的纠偏机制，预算调整的随意性较大。这使得预算对经营活动的约束力大打折扣，无法在项目执行过程中及时识别成本超支、进度滞后等风险，容易导致“预算决算化”，失去了过程管理的意义。

最后，绩效考评关联弱化，预算管理闭环未能形成。目前，预算完成情况与业务部门及其负责人的绩效考核关联度较低，或权重不足。干好干坏一个样，预算执行缺乏有效的激励与约束机制<sup>[5]</sup>。这直接导致了业务部门在编制预算时倾向于“讨价还价”、预留水分，在执行时对成本控制缺乏内在动力。预算管理因此难以形成从编制、执行到考核、激励的完整闭环，其权威性和严肃性被严重削弱。

表 1：Y 集团预算管理主要问题、表现及后果分析

| 核心问题   | 具体表现   | 导致的负面后果  |
|--------|--|--|
| 战略传导失灵 | 1. 战略目标未能分解为可衡量的预算指标<br>2. 预算编制与业务计划、市场预测脱节<br>3. 资源分配无法有效体现战略重点 | 1. 资源错配，战略重点项目投入不足<br>2. 预算无法成为战略落地的操作工具<br>3. 各部门目标与公司整体发展方向不一致 |
| 过程监控虚化 | 1. 预算执行分析滞后，重核算轻控制<br>2. 缺乏动态预警和有效的偏差分析机制<br>3. 预算调整随意，刚性约束力弱    | 1. 经营风险无法及时发现和应对<br>2. 成本失控，经营效率低下<br>3. 预算失去过程管理价值              |
| 绩效考评脱钩 | 1. 预算执行结果与部门 / 个人绩效考核关联弱<br>2. 缺乏基于预算完成情况的激励与问责                  | 1. 业务部门缺乏预算管控的内在动力<br>2. 预算管理闭环断裂，权威性不足<br>3. 预算目标难以得到有效执行       |

三、战略导向下全面预算管理优化体系的构建

构建一个以战略为导向、逻辑严谨且可执行的全面预算管理优化体系，核心目标在于打通战略与预算之间的壁垒，使预算真正成为引导资源配置、落实战略意图、管控经营风险的核心管理工具，而非单纯的财务计算工作。优化体系的构建遵循系统性原则，涵盖从目标设定到执行反馈的全过程<sup>[6]</sup>。

优化体系的总体框架是一个动态的闭环管理系统。该体系以

集团的战略规划为逻辑起点，通过科学的战略解码工具将宏观战略转化为具体的预算目标；在此基础上，推动业财深度融合的预算编制，确保预算方案既符合战略方向又切合业务实际；预算执行过程中，强调动态监控与刚性控制，建立快速响应机制；最终，通过将预算考核结果与绩效激励紧密挂钩，完成管理闭环，并为进一步的战略调整提供决策依据。这一“战略 - 预算 - 执行 - 考核 - 战略”的循环，确保了预算管理与公司战略保持持续一致。



（一）优化的关键在于建立战略解码与预算指标的有效链接

为解决战略与预算“两张皮”的问题，需引入战略地图、平衡计分卡等管理工具，将集团层面的战略目标，按照财务、客户、内部流程、学习与成长四个维度，逐层分解至各业务单元和职能部门，形成一套相互关联、可量化的关键绩效指标。这些指标将构成预算编制的直接依据。例如，对于“提升资产运营效率”这一战略主题，可分解为“租金收缴率”、“物业空置率”等具体预算控制指标，并明确责任部门<sup>[7]</sup>。这一过程确保了每一笔预算支出都能追溯到具体的战略目标。

（二）预算编制流程必须实现从财务主导到业财协同的根本转变

预算编制不应是财务部门闭门造车，也不应是业务部门各自为政。应建立由财务部门搭建模型、提供专业指导，业务部门基于市场预测和详细业务计划提报预算需求，双方共同审议、反复磋商的联合编制机制。特别是在产业投资和项目孵化领域，财务人员需提前介入，参与项目可行性研究，对投资回报率、现金流

等关键财务指标进行审慎评估，确保项目预算的科学性和合理性，这种协同确保了预算数据既承接战略，又根植于业务实际。

（三）必须强化预算执行的动态监控与刚性控制

预算批复后，管理重心应立即转向事中控制。建议建立月度或季度预算执行分析会制度，由财务部门不仅汇报执行数据，更重点分析重大偏差背后的业务原因，并向管理层提出预警和行动建议。同时，须严格规范预算调整的审批权限和流程，保持预算的严肃性。对于战略性重点项目，更应实行专项跟踪管理，确保资源投入与项目进度相匹配。

（四）实现预算绩效与考核激励的硬挂钩是保障体系落地的关键

将分解到各部门的预算关键绩效指标的完成情况，作为其年度绩效考核的核心组成部分，并赋予较高权重。通过业绩合同等形式，明确奖惩标准，使预算目标的实现程度直接关系到部门与个人的利益，从而激发各部门主动、有效地进行预算管理的内部动力，形成良性循环。

表2：全面预算管理优化体系核心环节对比

| 核心环节      | 优化前状况                | 优化后设计                     | 关键改进点  |
|-----------|----------------------|---------------------------|--|
| 战略传导与目标设定 | 战略规划与预算编制脱节，缺乏量化衔接机制 | 采用战略地图工具，将战略目标分解为可量化的预算指标 | 1. 建立战略 - 预算联动机制<br>2. 制定关键绩效指标 (KPI) 体系<br>3. 实现战略目标的具体化、量化 |
| 预算编制流程    | 财务部门主导编制，业务部门参与度低    | 推行业财协同的联合编制模式，建立分级审核制度    | 1. 业务部门深度参与预算编制<br>2. 财务部门提供专业指导和支持<br>3. 确保预算既符合战略又贴近业务实际   |
| 执行过程监控    | 侧重于事后核算，缺乏动态监控机制     | 建立“月度分析 + 季度评估”的动态监控体系    | 1. 实施预算执行实时跟踪<br>2. 建立重大偏差预警机制<br>3. 强化过程控制和及时纠偏             |
| 绩效考核应用    | 预算完成情况与绩效考核关联较弱      | 将预算指标纳入绩效考核体系，强化结果应用      | 1. 预算执行与部门绩效直接挂钩<br>2. 建立基于预算的激励机制<br>3. 增强预算管理的权威性和约束力      |

四、Y 集团预算管理优化的实施路径与保障措施

构建科学的优化体系是前提，而能否有效落地则取决于周密的实施安排与坚实的资源保障。鉴于全面预算管理优化是一项涉及观念转变、流程再造和组织调整的系统工程，切忌“一刀切”的激进改革。设计一套分阶段、可操作的实施路径，并明确确保改革成功的关键支撑要素是非常必要的。

（一）为确保改革稳步推进、降低变革阻力，建议采用“试点先行、全面推广、系统固化”的三阶段渐进式实施策略

第一阶段为试点突破与经验积累期。此阶段的核心目标是验证优化方案的可操作性，积累实战经验，树立成功样板。应选取战略重点清晰、业务流程相对规范、且管理层支持度高的业务单元作为试点，例如资产租赁业务或某个明确的产业孵化项目。在此阶段，首要任务是成立跨部门的项目小组，针对试点业务完成战略目标解码、细化预算指标、设计业财协同编制模板，并试运行新的预算编制、监控与汇报流程。通过小范围的实践，发现并解决流程中存在的具体问题，形成可在更大范围复制的标准化操作指南和成功案例，为后续推广凝聚共识、增强信心。

第二阶段为总结推广与制度固化期。在试点成功的基础上，对方案进行优化完善后，分批次、有步骤地将优化后的预算管理

模式推广至产业投资、项目建设等集团所有核心业务领域。此阶段的关键任务是将前期实践证明有效的做法，通过制度形式固化下来。必须系统性地修订和完善《全面预算管理办法》、《绩效考核管理办法》等一系列核心管理制度，明确各环节的权责、流程和标准。同时，需组织面向全集团的大规模培训与宣贯，确保各级管理人员深刻理解并掌握新体系的要求，实现从“要我做”到“我要做”的转变。

第三阶段为系统固化与持续优化期。当新的预算管理模式稳定运行后，应将其信息化建设提上日程。将优化后的预算编制流程、控制节点、审批权限、数据标准和分析模型等，作为核心需求融入到集团未来规划的企业资源计划（ERP）或业财一体化系统中。通过信息技术实现预算数据的自动采集、流程的线上流转、执行情况的实时反馈以及异常指标的自动预警，最大限度地提升预算管理的效率和刚性，减少人为干预，为实现数据驱动的精准管理和持续优化奠定坚实基础。

（二）实施路径的顺利推进，有赖于以下四个方面的有力保障

1. 组织保障是前提。必须成立由集团最高管理者挂帅的预算管理委员会，成员涵盖财务、战略、业务及各职能部门负责人。该委员会负责审议预算目标、协调重大资源分配、仲裁预算分

歧，并监督执行情况。高层领导的深度参与和强力推动，是打破部门壁垒、克服改革阻力的决定性力量。

2. 制度保障是基石。优化成果必须通过制度得以巩固。要及时制定和修订相关的管理制度与实施细则，明确战略解码、预算编制、执行控制、分析调整、绩效考核等各环节的规则、权限和责任，使新的预算管理体系有章可循、有法可依。

3. 技术保障是引擎。长远来看，预算管理的效率和精度提升离不开信息技术的支撑。应提前规划，将预算管理的信息化需求纳入集团整体数字化蓝图，通过系统实现流程自动化、数据一体化和管理精细化。

4. 人才与文化保障是长效支撑。加强对各级人员，特别是业务部门和项目人员的财务知识培训，同时对财务人员进行业务知识培训，培养复合型人才。更重要的是，通过持续的宣传和引导，在全集团范围内培育“全员参与、战略导向、重视绩效”的预算管理文化，为体系的持续有效运行营造良好的软环境。

综上所述，通过清晰的阶段划分和坚实的保障措施，Y 集团能够稳步推进全面预算管理体系的优化升级，最终将其打造成为

支撑战略落地、提升运营效率、防范经营风险的核心管理工具<sup>[8]</sup>。

## 五、结论与展望

本研究通过分析 Y 集团全面预算管理现状，构建了以战略为导向的预算管理优化体系。研究显示，传统预算模式难以满足镇属企业转型升级需求，必须建立战略 - 预算 - 绩效联动机制，实现预算管理与战略目标的深度融合。优化方案通过战略解码、业财协同、过程监控和绩效挂钩四个关键环节，有效提升了预算管理的战略支撑能力，提出了分阶段实施路径和全方位保障体系，为镇属企业预算管理优化提供了可行方案。通过试点验证、全面推广和系统固化的三阶段实施，以及组织、制度、技术和人才四维保障，确保优化方案稳步落地。未来研究可进一步探讨不同类型镇属企业的预算管理特点，加强数字化技术在预算管理中的应用，深化预算管理与绩效提升的内在机制研究，为镇属企业高质量发展提供更多实践指导。

## 参考文献

- [1] 刘日成. 基于战略导向的企业全面预算管理体系优化探究 [J]. 中国产经, 2024, (20): 111-113.DOI:CNKI:SUN:KJHU.0.2024-20-032.
- [2] 彭小娟. 基于战略导向的企业全面预算管理体系构建 [J]. 大众投资指南, 2024, (27): 137-139.DOI:CNKI:SUN:DATZ.0.2024-27-051.
- [3] 刘东进, 宋志鹏, 陈永煌, 等. 基于战略价值导向的全面预算管理体系构建与实践 [J]. 财务与会计, 2024, (11): 17-22.DOI:CNKI:SUN:CWKJ.0.2024-11-006.
- [4] 上官士全. 战略导向下集团公司全面预算管理体系的建设与优化 [J]. 财会学习, 2024, (13): 74-76.DOI:CNKI:SUN:CKXX.0.2024-13-025.
- [5] 李卓成. 基于战略导向的 C 公司全面预算管理体系优化研究 [D]. 长江大学, 2024.DOI: 10.26981/d.cnki.gjhsc.2024.001341.
- [6] 桑明丽. 基于战略导向的企业全面预算管理体系 [J]. 大众投资指南, 2024, (07): 114-116.DOI:CNKI:SUN:DATZ.0.2024-07-045.
- [7] 苏盛梅. 战略导向的全面预算管理体系在企业中的应用 [J]. 市场瞭望, 2024, (02): 138-140.DOI:CNKI:SUN:SHLW.0.2024-02-040.
- [8] 易云平. 战略导向下全面预算管理体系在企业中的应用分析 [J]. 会计师, 2023, (07): 19-21.DOI:CNKI:SUN:KJSZ.0.2023-07-007.



# 中小企业智慧财务系统上线对财务的影响

谭梓颖

万灵帮桥医疗器械（广州）有限责任公司，广东，广州 510000

DOI:10.61369/ASDS.2026010003

**摘 要：**在数字经济时代，智慧财务系统的应用已成为中小企业重塑管理架构、实现财务转型的关键驱动力。本文深入探讨智慧财务系统上线对中小企业财务工作产生的双重影响及应对策略。研究发现，在积极影响方面，该系统数据自动化处理机制，构建业务与财务深度融合的数据闭环体系，有效破除传统财务手工核算模式长期面临的流程割裂、效率低下等瓶颈。利用实时监控与预警机制，强化资金、税务、业务多维度风险管控，切实增强企业核心竞争力，并有效推动财务职能从基础核算向管理决策支持转型，提升业务与财务融合水平。然而，系统实施过程中亦伴随着不容忽视的负面冲击，主要表现为财务人员技能滞后与思维转变带来的压力、建设初期与运营成本高造成的资金负担、开放网络环境下数据安全风险敞口的扩大，以及标准化系统逻辑与中小企业业务流程之间的适配冲突。针对上述挑战，本文提出强化财务人员复合型能力培训、遵循成本效益原则科学规划投入、构建技术与管理并重的数据安全防线，以及建立动态优化的流程适配体系等策略，旨在助力中小企业趋利避害，充分释放智慧财务系统的管理效能，实现财务管理体系的现代化升级。

**关 键 词：**中小企业；智慧财务系统；财务；影响

## Impact of the Implementation of Intelligent Financial Systems on the Financial Management of Small and Medium-sized Enterprises

Tan Ziying

Wanlingbang Bridge Medical Equipment (Guangzhou) Co., Ltd., Guangzhou, Guangdong 510000

**Abstract：** In the era of the digital economy, the application of intelligent financial systems has become a key driving force for small and medium-sized enterprises (SMEs) to restructure their management architecture and achieve financial transformation. This paper conducts an in-depth analysis of the dual impacts—both positive and negative—of the implementation of intelligent financial systems on the financial work of SMEs, as well as the corresponding coping strategies. The study finds that, on the positive side, such systems, through automated processing and multi-source data integration, significantly break through the efficiency bottlenecks of traditional manual accounting, strengthen risk control over funds and taxation by means of real-time monitoring and early-warning mechanisms, and effectively promote the transformation of financial functions from basic bookkeeping to management decision support, thereby deepening the integration of business and finance. However, the implementation process is also accompanied by non-negligible negative impacts, mainly manifested in the transformation pressure caused by the lag in skills adaptation among financial personnel, the financial burden arising from high initial construction and operation-maintenance costs, the expansion of data security risk exposure in an open network environment, and the misalignment between standardized system logic and the flexible business processes of SMEs. In response to these challenges, this paper proposes strategies such as strengthening the training of financial personnel to develop compound competencies, planning investments scientifically in line with the cost-benefit principle, building a data security defense system that gives equal weight to technology and management, and establishing a dynamically optimized process adaptation mechanism. These measures aim to help SMEs leverage advantages while mitigating risks, fully unlock the managerial effectiveness of intelligent financial systems, and realize the modernization and upgrading of their financial management systems.

**Keywords：** and medium-sized enterprises (SMEs); intelligent financial system; financial management; impact

# 引言

随着数字化趋势的不断发展，企业财务管理模式正经历着重大的变革。智慧财务系统作为信息技术和财务业务相结合的成果，日益成为企业提高管理效率的重要手段。中小企业在国民经济体系里处于重要位置，是市场活力和创新动力的关键提供者，其财务管理的水平直接关系到企业的运营状况和发展潜力。中小企业的财务工作大多处于基础核算和事后记录的阶段，存在着流程复杂、信息传递不及时、决策支持不够等情况，对企业整体管理效率的提升形成阻碍<sup>[1]</sup>。智慧财务系统借助流程的重新构建和数据的整合，希望推动财务职能从核算类型向管理类型转变。智慧财务系统的上线是否能够切实提升财务工作的效率、优化资源的配置、增强风险管控的能力，是中小企业在推进财务数字化过程中需要迫切解决的实际问题。基于这样的背景，针对智慧财务系统上线给中小企业财务工作带来的影响进行探讨，对于明确其应用价值、指导企业的转型实践具有一定的参考作用。

## 一、中小企业智慧财务系统上线对财务的正面影响

### （一）提升财务工作运转效率

在传统财务管理方式中，中小企业的财务工作对手工操作依存度较高，像凭证录入、账务处理、报表编制等基础事务会耗费大量时间和精力，工作效率难以得到有效提升<sup>[2]</sup>。智慧财务系统的投入使用，从本质上扭转这一局面。在日常核算环节，系统依托自动化技术，对发票识别、凭证生成、账簿登记等重复性流程实施批量处理，大幅压缩重复性工作耗时，推动财务人员脱离繁杂手工操作，集中精力开展核心财务管理工作。业务数据与财务数据存在口径不一、数据滞后及手工录入误差等问题。系统依托业务与财务数据接口的贯通，实现信息实时自动化更新，有效降低传统模式下因数据滞后、手工录入误差引发的反复核对时间损耗及数据差错风险。在报告生成环节，系统能够按照预先设定的规则，自动生成财务报表和分析图表，明显缩短月末结账和报告编制的周期。

### （二）强化财务风险管控能力

中小企业经营灵活但抗风险能力相对薄弱，资金链断裂、应收账款逾期、成本失控等问题时有发生，故对财务风险管控提出较高要求<sup>[3]</sup>。而传统管理模式以下手工操作为主，数据处理效率显著低下，准确性难以得到充分保障；业务与财务协同存在脱节现象，数据滞后问题突出，难以有效支撑业务决策；风险管控能力偏弱，合规性隐患凸显；财务分析维度单一、深度不足，价值赋能能力有限等问题<sup>[4]</sup>。智慧财务模式呈现出业务与财务深度融合、数据驱动决策实施、智能高效管控落地、价值主动创造赋能的核心特质。系统是业务与财务深度融合的核心支撑，助力构建一体化协同体系，实现全流程贯通。业务流程触发后，业务数据自动流转同步至财务模块，形成管理闭环，保障数据源头唯一、口径统一。系统大幅降低手工操作占比，将财务人员从重复核算中解放，聚焦价值创造型管理；通过设置标准化凭证模板降差错率，标准化结账流程使周期从15天缩至10天内，显著提升财务效率。

资金管理层面，实时监控销售、库存、往来账款等核心数据，强化资金全生命周期管控。针对大额支付、异常资金流动等场景自动预警，降低资金挪用、坏账风险等因素，推动财务管理从“事后反馈”向“事前预警、事中管控”转型，提升资金管理

安全性与可控性<sup>[5]</sup>。

依托数据分析技术，具备多维度分析与决策支撑，可按产品、客户等多维度分析盈利与成本，实现趋势预测与风险预警。基于历史数据挖掘风险点，提前推送预警信息（如应收账款逾期催收、原材料涨价前置采购建议），生成企业专管理报表，为决策提供支撑。

内控与风险防控方面，系统通过规则固化与智能监控，将内控要求转化为分级审批流程，杜绝人为漏洞，契合企业内控规范。监控存货周转率等核心指标，实现供应链风险联动预警，推动风险防控前置，达成“合规管控+风险预警”双重目标，强化企业风险抵御能力。

系统为全员协同平台，业务人员需规范录入单据、参与预算编制执行，承担数据源头责任，强化成本与效率意识；管理层可通过实时查看、远程审批，打破时空限制，提升决策效率，推动企业经营管理水平提升。

### （三）促进财务职能转型升级

长期以来，中小企业财务部门的主要职责集中于记账、报税、对账等事后核算工作，极少参与事前预测、事中管控等管理环节。其工作主要围绕企业负责人及外部监管机构需求，主动为业务部门提供支撑的意愿与能力均相对薄弱。智慧财务系统的应用为财务职能的转变提供有利条件。一方面，该系统承担大量基础的核算工作，让财务人员从繁琐事务中解脱出来，有更多时间和精力投入到预算管理、成本分析、绩效评价等管理会计相关工作中<sup>[6]</sup>。另一方面，系统拥有的数据整合以及可视化分析效能，增强财务部门探寻业务内在逻辑、发现经营过程中所存问题的能力，能够为管理层提供更具参考性的决策建议。财务工作的核心要点正从对以往状况的记录向未来的计划安排延伸，从被动地回应需求向主动地赋予能力转变。这一进程并非仅仅是技术工具的更新换代，更是财务角色定位的重新构建。财务部门正逐步从后台的支撑部门走向前端参与业务活动，从成本中心向价值中心演进，在企业战略的落地实施以及资源配置的优化等方面发挥更为主动的作用。

### （四）提高财业协同发展水平

中小企业在传统模式下，财务模块与业务模块往往处于相互独立、数据割裂、财务滞后的状态，导致财务数据与业务实际情

况相脱节，难以精准反映企业真实的经营状况。智慧财务系统作为连接企业内部各个环节的核心纽带，能够有效打破采购、生产、销售、库存与财务之间的壁垒，实现业务流、资金流和信息流的深度融合统一。借助系统整合，每笔业务都能及时转化为财务方面的会计信息，实现财务对业务流程的全程管理。这种业务与财务高度融合，不仅能够解决部门之间的信息不对称问题，还能推动财务管理能够向前延伸到业务前端。比如在合同管理或者项目定价过程中，财务数据可以及时反馈成本构成和利润空间，为业务部门制定更科学的市场策略提供指导<sup>[7]</sup>。反过来，业务数据的及时反馈也让财务预算的编制更加准确真实。智慧财务系统的投入使用，从根本上改变企业内部的协作机制，推动财务与业务在目标、流程和数据上的深入协同。

## 二、中小企业智慧财务系统上线对财务的负面影响

### （一）加大财务人员适配压力

智慧财务系统的投入使用对中小企业财务工作人员的专业水平和技术能力提出更高的要求。在此过程中出现的适配压力，成为限制系统功能发挥的关键因素。在知识架构方面，传统财务人员长期依靠人工操作和基础软件进行账务处理，其知识体系主要包括会计准则与核算方法以及财务基础工作规范等内容。智慧财务系统要求财务人员具备数字化工具应用与数据处理能力、业财融合知识、系统规则配置与流程优化技能以及风险管控与战略决策相关知识等多方面的综合能力。这种知识架构的断层，使财务人员承受着较重的学习压力和转型困难。从心理层面来看，部分具备丰富工作经验的财务人员对新科技存在抵触心理，担心自己的工作岗位被智慧财务系统替代，从而产生职业焦虑感。这类消极心理不仅会削弱对个人开展工作的主动性产生负面影响，还会对整个财务队伍的稳定性造成不良影响。此外，中小型企业普遍缺乏完善的内部培训体系，难以向财务人员提供针对性的技能提升支持，使得人员能力和系统要求之间的差距很难在短时间内缩小<sup>[8]</sup>。

### （二）增加财务运营成本负担

企业引入一套完整的智慧财务系统，所需投入可划分为前期投入成本与运营过程中的持续投入成本两大类。其中，前期投入成本主要涵盖软件授权费、硬件适配费、实施服务费、咨询费、人员培训费及前期调研规划费等相关开支。对于中小企业而言，前期投入成本与企业规模、业务复杂程度及部署模式选择直接挂钩，企业规模越大、业务复杂程度越高、部署模式要求越严格，对应的前期投入成本就越高；且该类成本为一次性支付款项，短期内易占用企业核心运营资金，对核心业务的资金流转造成不利影响，进而加重企业的财务运营成本负担。此外，随着企业自身业务的发展、财务核算规则的调整以及税收政策的更新，智慧财务系统需同步跟进企业业务发展节奏与外部政策调整方向，持续推进二次开发、系统迭代及版本升级工作，同时承担数据安全保障与合规管理相关成本。上述各项工作均需企业持续投入资金，进一步加剧了企业的财务运营成本负担。为确保系统持续稳定运行，企业需通过聘请外部专业技术顾问或组建内部信息技术维护

团队的方式，保障系统日常平稳运转，这将直接导致人工成本与咨询服务费用攀升，加重企业财务运营压力。若智慧财务系统上线后实际效果未达预期，无法快速产生经济效益以覆盖前期及持续投入成本，这种高投入、低回报的财务压力，将严重拖累企业整体经营效能，进一步凸显财务运营成本负担过重对企业发展的制约作用。

### （三）引发财务数据安全风险

智慧财务系统凭借网络与云端平台实施搭建操作，及时共享数据，此特征可有效提升工作效率。同时，它会使原本处于封闭状态的安全环境发生变化，让财务数据处于风险之中。针对不具备专业信息安全防护体系的中小型企业，在网络安全层面存在极大的隐患问题。倘若遭遇黑客入侵、邪恶病毒攻击情形或者系统漏洞被利用现象，成本构成、定价策略、客户名单信息以及银行账户数据等核心财务机密等面临泄露危机<sup>[9]</sup>。智慧财务系统具备的高度集成特性让账号权限管理工作变得更加复杂。权限设置不恰当或者审批程序存在逻辑层面的缺陷等情况，均可能致使未获取授权的人员接触到敏感数据信息。当中小型企业遭遇数据灾难事件时，若缺失备份机制与应急恢复体系，核心财务数据产生的损坏情况将会对企业造成灾难性的后果。

### （四）造成财务流程适配难题

当前市场主流的智慧财务系统，多以标准化管理思维与规范化程序架构为核心搭建而成，其设计核心侧重于流程的严密性与管控的制约性。中小型企业的运营则具备业务形式灵活多变、决策链路短、响应速度快、可快速适配市场供需变化等特征，这些特征与智慧财务系统的设计逻辑之间存在冲突。在智慧财务系统上线过程中，系统与业务之间的不契合状况尤为突出：前者强调流程统一规范、数据标准化流转与风险刚性约束，后者则需随市场变化快速调整业务流程与财务核算规则；前者通过严密的权限分级与程序制约实现管控目标，后者则更注重运营效率与决策敏捷性。这种本质性的逻辑冲突，直接导致智慧财务系统在中小企业的落地应用过程中，面临流程适配、功能兼容与价值发挥等多重现实挑战。

## 三、中小企业智慧财务系统上线应对财务负面影响的策略

### （一）强化财务人员能力培训

针对智慧财务系统上线运行后财务人员存在的业务技能不匹配问题，中小企业应构建完善的人才培养策略。该策略消除财务人员抵触心理、助力财务人员掌握系统基础操作与树立正确财务观念、推动定期开展财务政策与系统适配学习等方面展开，进而提升财务人员专业工作能力。通过引入管理会计知识体系，增强财务团队的数据敏感度与数据分析能力，推动财务人员从单纯的核算记录角色向数据分析师与价值创造者转变，使其能够依托系统生成的报表为企业经营决策提供支撑。同时，鼓励财务人员深入业务一线实践，培育业财融合的综合思维，助力财务人员从业务场景中理解财务数据，更精准地适配智慧财务系统对业务流程



管控的需求。

（二）合理管控财务投入成本

考虑到中小企业资金能力有限，在推进智慧财务系统建设工作时，需以精准选型、分阶段实施、优化资源利用为核心原则，通过科学合理的规划安排减轻企业资金压力。企业在智慧财务系统选型阶段，应优先结合自身业务特点与企业规模，选取可优先配置核心模块、延后部署非必要模块的系统解决方案，并依据实际能力制定科学的预算计划。在实施策略方面，可采用分阶段推进模式分摊资金压力，优先落地核心模块；待核心模块稳定运行、企业产生明确扩展需求且相关人员配备到位后，再逐步拓展功能范围，以此持续缓解企业资金流压力。同时，企业应全面梳理现有财务流程，通过优化内部业务流程、优先采用系统标准功能、严格管控二次开发及后期维护成本等方式，规避不必要的高额支出。在人员培训环节，应充分借助系统配套的免费线上培训资源（如操作视频、直播课程、知识库等），既可组织财务人员开展自主学习，亦可编制企业内部系统操作问题管理手册，以便快速解决日常操作难题，替代部分外部付费培训服务，进而节约培训投入，保障从“手工核算”向“智慧财务”的平稳过渡。最终目标是使智慧财务投入转化为可量化的效率提升与价值创造，而非成为企业的运营负担。

（三）筑牢财务数据安全防护线

在技术层面，应聚焦“源头防控”与“全链路保护”核心要求，通过部署企业级防火墙、入侵检测系统及采用加密传输协议等技术手段，阻断外部黑客攻击与病毒入侵路径。应优先遴选具备合规认证的云服务平台或本地化部署方案，重点核查确认系统是否内置数据加密、访问控制、异常监测等核心安全功能；针对财务核心数据（含营收数据、成本数据、税务数据等），需实施传输加密与存储加密双重防护机制，借助数据脱敏技术对非必要展示的敏感信息进行匿名化处理。在管理方面，需建立严密的权限管理制度，根据岗位职责严格设定系统操作权限，遵循不相容职务分离原则，杜绝账号混用和越权操作现象，防范内部人员导

致的敏感信息泄露<sup>[10]</sup>。另外，构建完善的容灾备份机制作为最后一道防线。企业应制定严格的数据备份计划，采取本地和异地双重备份策略。定期开展数据恢复演练，确保在遭遇病毒攻击或系统崩溃等极端情况时，能够迅速恢复核心财务数据，保障企业经营的连续性，将数据灾难带来的财务损失降至最低。

（四）优化财务流程适配体系

企业在应对智慧财务系统标准化和中小企业灵活业务模式的矛盾时，适合运用流程优化的动态适配方式。在智慧财务系统正式上线以前，企业应当安排财务领域与业务领域的核心人员对现行的业务流程开展全方位的梳理和诊断工作。在业务流程再造的进程中，去除冗余且低效的环节，保留企业的灵活性优势。在开展智慧财务系统配置工作时，需要预先留出一定的弹性空间。针对紧急业务或者特殊交易事项，在遵循合规要求的基础上，设置绿色通道或者例外审批流程。除此之外，在智慧财务系统运行的初始阶段，财务部门需要积极主动地收集业务部门的反馈信息，定期对流程节点的运行效率展开评估，针对业务开展造成阻碍的环节实施有针对性的参数调整或者适度的二次开发。通过加强业务与财务之间的沟通交流，让智慧财务系统流程能够符合市场的需求。

四、结束语

在数字经济演进的过程中，中小型企业为强化核心竞争力，运用智慧财务系统成为必然趋势。该系统在提升核算工作效率、强化风险防范治理、推进职能转化以及促使业务与财务融合等方面，展现出显著的成效。智慧财务系统在资金投入规模、人员配置适配、数据安全维护和流程协调契合等方面，给企业带来挑战。所以，智慧财务系统的上线不只是技术工具的更新替换，更是一场管理理念革新和组织架构重构的系统性变革，需要企业以理智的心态看待并慎重地应对。

参考文献

[1] 段玉峰. 国有企业财务系统切换的困境突破与实践智慧——以 N 集团为例 [J]. 西部财会, 2025, (09): 27-30.

[2] 吴威. 数字化转型驱动下烟草企业智慧财务生态系统构建研究 [C]// 江西省汽车工程学会, 江西省工程师联合会. 工程技术与新能源经济学术研讨会论文集 (二). 红河烟草 (集团) 有限责任公司乌兰浩特卷烟厂, 2025: 643-645.

[3] 王宏利, 张耀杰, 王彦博. 智慧财务系统对企业运营模式的影响研究 [J]. 财务管理研究, 2021, (09): 156-160.

[4] 汪夏青. 基于业财融合的智慧财务系统建设分析 [J]. 活力, 2025, 43 (09): 22-24.

[5] 李雪儿, 关梦琦, 李烨, 等. 数字化背景下企业智慧财务系统建设存在的问题及对策 [J]. 中国市场, 2025, (04): 125-128.

[6] 王宏利. 企业财务管理数字化转型研究——智慧财务系统的构建 [J]. 产业创新研究, 2024, (13): 11-13.

[7] 马春蕾. 基于“大智移云”的智慧财务系统构建 [J]. 财会通讯, 2023, (06): 159-164.

[8] 陆秀芬, 蒋丽君. 基于区块链的企业智慧碳财务系统构建研究 [J]. 财会通讯, 2023, (06): 170-176.

[9] 王晓宇. 依托 ERP 系统提升企业财务管理智慧化水平的方案研究 [J]. 质量与市场, 2022, (23): 7-9.

[10] 韩雪丽. 智慧财务: 企业经济高质量发展的助推器 [J]. 商业会计, 2021, (22): 20-23.

# 基于回归分析与随机森林的初中数学成绩影响因素分析

李冰

天津市梧桐中学, 天津 300200

DOI:10.61369/ASDS.2026010004

**摘 要 :** 为探究影响八年级初中生数学成绩的关键因素, 基于回归分析模型与随机森林方法研究了某中学八年级84名学生期末数学成绩的影响因素, 系统分析了各解释变量与成绩间的相关性及影响机制。结果表明: 期中成绩是核心预测指标, 月考成绩、两次测试成绩与期末成绩呈现强正相关关系, 期末前作业完成质量呈中等正相关, 出勤天数对期末成绩影响较弱, 性别因素存在显著差异, 班级因素无显著影响。研究结果可为初中数学个性化教学与学业提升策略制定提供实证依据。

**关 键 词 :** 回归分析; 随机森林; 数学成绩; 影响因素

## Determinants of Middle-School Mathematics Achievement: A Regression–Random-Forest Hybrid Analysis

Li Bing

Tianjin Wutong Middle School, Tianjin 300200

**Abstract :** To identify the key influencing factors of eighth-grade mathematics achievement, we analyzed math scores for 84 students using a regression model and random-forest method. Our study systematically examined the strength and mechanism of each predictor's association with performance. Midterm score emerged as the dominant predictor; monthly-quiz and two additional test scores displayed strong positive correlations with the final result, whereas pre-final homework quality showed a moderate positive link. Attendance exerted only a weak influence, and a significant gender gap was detected; class membership had no discernible effect. These findings furnish an empirical foundation for tailoring middle-school mathematics instruction and designing targeted academic-improvement strategies.

**Keywords :** regression analysis; random forest; mathematics score; influencing factors

### 引言

初中阶段的学习是夯实学生基础知识、培养综合素养的关键环节, 数学是重要载体之一, 它能培养学生的逻辑思维、抽象推理与问题解决能力, 数学成绩的高低会直接影响学生的理科学习路径与综合素养发展<sup>[1]</sup>。学生在这个阶段需要完成知识跨越(从直观认知发展至逻辑证明), 学业难度与认知要求都显著提升, 数学成绩不仅反映出学生的知识掌握水平, 也能折射出学生的学习态度、习惯等深层素养。在初中数学教学过程中, 学生成绩出现分化的情况非常普遍, 而成绩差异受多重因素交织影响。

国内外学者对学业成绩影响因素的研究已形成丰富成果<sup>[2]</sup>, 初中生数学学业成绩受学习行为<sup>[3]</sup>、个体特征层面、教学因素、个体心理与情绪<sup>[4]</sup>等多维度因素交互作用, 其中阶段性练习(月考、期中)、出勤次数、平时作业效果也是关键作用变量。在个体特征层面, 部分研究认为性别差异会影响数学学习, 男性在空间推理、逻辑运算上表现更优, 女性在细节处理、基础运算上更具优势<sup>[5]</sup>, 也有研究指出, 随着教育公平推进, 性别对数学成绩的影响逐渐减弱<sup>[6]</sup>。在学习行为层面, 出勤作为课堂参与的基础, 其规律性与学业成绩呈正相关, 缺课会导致知识断层, 显著降低成绩表现<sup>[7]</sup>; 作业作为课堂知识的延伸, 其完成质量直接反映知识掌握程度, 高质量作业能强化知识内化, 提升学业成效<sup>[8]</sup>。Smith等研究发现阶段性练习(月考、期中)是学业评价的重要环节, 月考成绩能及时反馈学生阶段性知识漏洞, 其与期末成绩的相关系数达0.72, 是重要预测指标<sup>[9]</sup>。王颖对初中学生的研究表明, 期中成绩因覆盖知识范围与期末高度契合, 相关性最强, 可作为期末成绩的核心预测变量<sup>[10]</sup>。

现有研究虽已关注多因素对数学成绩的影响, 但存在三方面不足: 一已有研究多聚焦单一维度, 或对影响因素的分析较为零散, 缺乏对性别、出勤、作业完成度、阶段性练习等多指标的系统整合研究; 二是针对八年级关键学段的实证研究较少; 三是部分研究采用传统统计方法, 对非线性关系的捕捉能力较弱。因此本文结合多种量化分析方法研究期末成绩与性别、出勤率、平时练习、期中成绩、作业得分等因素的相关性, 为教学实践提供更精准的参考。

作者简介: 李冰, 女, 天津市梧桐中学, 二级教师, 主要从事初中数学教学与研究工作。



一、数据来源与统计方法

（一）数据来源

研究选取了八年级 A 班（46 人）和 B 班（38 人），共 84 名学生为研究对象，学生年龄分布为 13–14 岁，均接受统一数学课程教学，排除特殊教育需求学生。响应变量选取了学年期末数学成绩，解释变量考虑了性别、月考成绩、期中成绩、平时测试成绩、出勤天数及平时作业得分等，变量具体定义见表 1。

| 表 1：变量定义 |             |      |                    |
|----------|-------------|------|--------------------|
| 变量类型     |             | 变量含义 | 定义与测量方式            |
| 响应变量     | Y           | 期末成绩 | 期末数学质量检测成绩         |
|          | X1          | 班级   | 分类变量（A 班、B 班）      |
|          | X2          | 性别   | 分类变量（1= 男生，2= 女生）  |
| 解释变量     | X3          | 月考成绩 | 月考数学成绩（满分 100 分）   |
|          | X4          | 期中成绩 | 上学期期中数学检测成绩        |
|          | X51、X52、X53 | 测试成绩 | 三阶段的测试成绩           |
|          | X61、X62、X63 | 出勤天数 | 三阶段的实际出勤天数         |
|          | X71、X72、X73 | 作业得分 | 三阶段的数学作业得分（满分 3 分） |
|          |             |      |                    |

注：如无特殊说明，成绩满分是 100 分。

（二）统计分析方法

为系统探究各解释变量对期末数学成绩的影响方向与程度，揭示变量间的潜在关联规律。本研究整合多维度数据分析方法，构建

由表及里、逐层递进的分析框架，从数据特征刻画、关联强度探究到模型构建与优化，形成完整的分析链路，具体方法如下：

描述性统计分析：测算各核心变量的均值、标准差、中位数、极值及偏度等关键统计量，通过定量指标精准刻画数据的整体分布特征、离散趋势与形态特征，为后续深入分析奠定数据基础。

相关性分析：剖析各自变量与期末数学成绩的线性关联强度，构建相关性矩阵以可视化呈现变量间的关联模式，初步筛选对期末成绩存在潜在影响的关键变量。

多元线性回归：构建递进式回归模型，通过引入不同层次的自变量，结合假设检验方法，依次验证各因素对期末成绩影响的统计学显著性，明确变量间的线性作用关系。

随机森林模型：基于集成学习算法构建随机森林模型，评估各输入变量的特征重要性排序，有效弥补传统线性回归模型难以捕捉变量间复杂非线性关系的局限，挖掘影响期末成绩的核心驱动因素。

LASSO 回归：依托 L1 正则化机制，对多元线性回归模型进行优化，通过系数压缩与变量筛选，剔除冗余变量，降低模型过拟合风险，进一步提升模型的泛化能力与预测效能。

二、数据分布特征

（一）描述性统计

变量的详细描述性统计分析见表 2。

表 2：变量的描述性统计分析

| 变量名 | 变量含义   | 均值    | 标准差   | 中位数   | 最小值  | 最大值 | 偏度    |
|-----|--------|-------|-------|-------|------|-----|-------|
| Y   | 期末成绩   | 56.10 | 25.10 | 58.00 | 10.0 | 100 | -0.07 |
| X3  | 月考成绩   | 67.95 | 22.97 | 72.00 | 17.0 | 100 | -0.42 |
| X4  | 期中成绩   | 76.30 | 18.60 | 81.50 | 21.0 | 100 | -0.97 |
| X51 | 测试成绩 1 | 67.52 | 23.83 | 72.00 | 0.0  | 100 | -0.55 |
| X52 | 测试成绩 2 | 68.10 | 28.28 | 76.00 | 0.0  | 99  | -1.31 |
| X53 | 测试成绩 3 | 52.79 | 27.50 | 51.00 | 0.0  | 100 | -0.18 |
| X61 | 出勤天数 1 | 15.39 | 4.98  | 12.00 | 6.5  | 21  | 0.08  |
| X62 | 出勤天数 2 | 17.68 | 1.59  | 18.00 | 10.0 | 19  | -1.84 |
| X63 | 出勤天数 3 | 17.58 | 3.55  | 17.00 | 1.5  | 21  | -1.53 |
| X71 | 作业得分 1 | 1.91  | 0.75  | 2.00  | 0.0  | 3   | -0.65 |
| X72 | 作业得分 2 | 1.99  | 0.90  | 2.20  | 0.0  | 3   | -0.83 |
| X73 | 作业得分 3 | 1.79  | 0.93  | 2.09  | 0.0  | 3   | -0.65 |

由表 2 可知，响应变量期末成绩（Y）的均值 56.10 分，低于 60 分及格线，标准差 25.10，均值较低且离散程度大，反映学生成绩分化明显。期末成绩主要集中在 40–60 分区间（约 25 人），呈近似对称分布，偏度为 -0.07，表明近似呈现正态分布。解释变量中期中成绩（X4）均值为 76.30 分，高于月考成绩（67.95 分），说明学生中期学习效果提升。三阶段的出勤天数均值分别为 15.39、17.68、17.58 天，相差不太大，但第二次出勤天数最高，其偏度 -1.84，呈左偏分布，说明多数学生二阶段的出勤天数较高，少数学生缺勤较多。三次作业得分均值在 1.79–1.99 之间，标准差较小（0.75–0.93），说明学生作业完成质量差异不大。

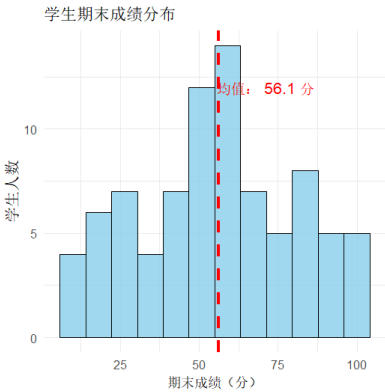


图 1：学生期末成绩直方图

为了清晰展示学生期末数学成绩的分布特征，绘制了响应变量的直方图，见图1。由图1可知，期末数学成绩主要集中在25至75分之间，该区间人数最多，75至100分区间人数较少。整体成绩偏低，均值仅为56.1分，表明多数学生成绩未达到及格水平，成绩分布呈现右偏形态，高分段学生极少，整体学习效果可能不理想，需关注学生学习中重要的影响因素。

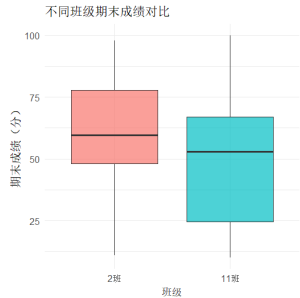


图2：两个班成绩的箱线

两个班的成绩对比箱线图如图2所示。由图2可得，两个班成绩对比来看，A班（红色）成绩中位数高于B班（绿色），且A班箱子厚度比B班小，说明A班成绩较为集中，A班整体成绩更优，B班成绩分化较严重。

### （二）相关性分析

由于期末成绩与期中成绩具有较高的相关性，因此绘制了两个班成绩的散点图及拟合曲线，见图3。

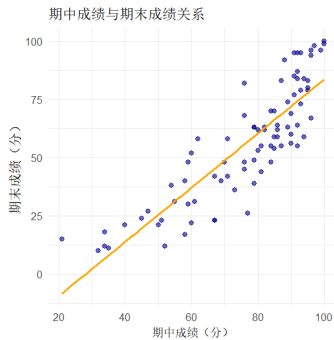


图3：期中成绩与期末成绩关系散点图

图3显示趋势线向上，表明期中成绩与期末成绩呈现强正相关关系。期中成绩高于80分的学生，期末成绩多大于70分，反之，期中成绩低于50分的学生，期末成绩多小于50分，表明期中成绩是期末成绩核心预测指标。

其他变量的相关性矩阵热力图见图4。

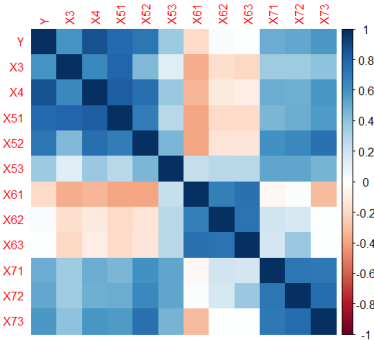


图4：变量的相关性矩阵热力图

第一次测试成绩（X51）相关性0.79，第二次测试成绩（X52）相关性0.75，均为强正相关，第三次作业得分（X73）相关性0.62，为中等正相关，期中成绩、测试成绩、平时作业得分与期末分数的相关性说明了数学成绩是一个具有连续性、可积累的结果，日常学习的过程性表现能直接反映并影响最终的学业水平。出勤天数（X61-X63）与期末成绩相关性均小于0.2，影响较弱，说明数学分数的呈现比较依赖基础知识是否牢固，不受偶尔缺勤的影响，也存在人虽然到了课堂，但可能上课走神、不参与课堂互动、不跟着老师的思路思考问题，这种“无效出勤”也无法转化为成绩的提升，也会呈现出出勤天数和成绩之间没有明显关联的结果。

## 三、回归与随机森林模型分析

### （一）多元线性回归分析

为了精准探究各解释变量对学生的影响关系，通过多元线性回归模型进行分析，结果见表3。

表3：基于多元线性回归的成绩的影响因素分析

| 估计                | 模型1       | 模型2       | 模型3       |
|-------------------|-----------|-----------|-----------|
| 截距                | -32.83*** | -50.22*** | -36.25*** |
| 水平一层变量            |           |           |           |
| X3(月考成绩)          |           | 0.11      | 0.12      |
| X4(期中成绩)          | 1.17***   | 0.94***   | 0.92***   |
| test_mean(综合测试)   |           | 0.19      | 0.23*     |
| attend_mean(综合出勤) |           | 0.19      |           |
| 水平二层变量            |           |           |           |
| 班级(11班)           |           |           | 5.78      |
| 性别(2)             |           |           | -5.39*    |
| R <sup>2</sup>    | 0.746     | 0.773     | 0.782     |

注：\*、\*\*、\*\*\* 分别表示显著性水平为0.05、0.01、0.001。

由表3可得如下结论：

模型1仅含相关系数最大的期中成绩，R<sup>2</sup>为0.746，说明期中成绩单独可解释期末成绩74.6%的变异，期中成绩每提高1分，期末成绩预期提高1.17分。在模型1的基础上，在加入水平一层的

解释变量，构成了模型2的结果，R<sup>2</sup>有所提升，达到0.773，期中成绩系数降至0.94（p<0.001），仍高度显著，表明期中成绩仍然是解释期末成绩的关键影响因素。月考成绩、综合测试、综合出勤的系数均不显著，说明这些变量在控制期中成绩后，独立贡献有限。在模型2基础上再加入班级、性别等因素构成模型3，R<sup>2</sup>进

一步提升至0.782，期中成绩系数稳定在 0.92（ $p<0.001$ ），综合测试成绩变得显著，表明综测成绩每提升1分，期末成绩能有效增加0.23分。性别因素显著，相较于男同学，女同学的成绩低5.39分。班级因素无显著差异。

（二）随机森林方法

为避免样本划分的随机性，让评估结果更稳定可靠。通过模拟生成了不同样本划分下的超参数选择，以RMSE（均方根误差）、 $R^2$ （判定系数）与MAE（平均绝对误差）作为衡量预测模型精度的核心指标为评估指标。RMSE、MAE值越小代表模型越好， $R^2$ 则越大越好。通过5折重复交叉验证，对12种超参数组合进行筛选，最优超参数选择见表4。

表4：随机森林超参数选择

| mtry | splitrule  | RMSE     |           | MAE      |
|------|------------|----------|-----------|----------|
| 2    | variance   | 13.63175 | 0.7303722 | 11.34675 |
| 2    | extratrees | 14.57091 | 0.7109171 | 11.87205 |
| 4    | variance   | 13.00682 | 0.7441912 | 10.64079 |
| 4    | extratrees | 13.64631 | 0.7349649 | 11.09301 |
| 6    | variance   | 12.85370 | 0.7475450 | 10.37788 |
| 6    | extratrees | 13.34373 | 0.7427058 | 10.90067 |
| 8    | variance   | 12.77588 | 0.7490694 | 10.26280 |
| 8    | extratrees | 13.21776 | 0.7436431 | 10.75150 |
| 10   | variance   | 12.79958 | 0.7453868 | 10.28632 |
| 10   | extratrees | 13.06458 | 0.7495955 | 10.64378 |
| 13   | variance   | 12.85875 | 0.7422460 | 10.36590 |
| 13   | extratrees | 12.97155 | 0.7529319 | 10.57561 |

由表4可知，特征随机选择数为8，叶子节点最小样本数为5，决策树数量选择了800棵。在测试集和训练集学生成绩数据进行评估，评估结果见表5。

表5：模型性能评估

| 数据集 | 样本量 | RMSE   | MAE   | $R^2$ |
|-----|-----|--------|-------|-------|
| 训练集 | 68  | 5.535  | 4.408 | 0.950 |
| 测试集 | 16  | 10.862 | 9.707 | 0.820 |

由表5， $R^2$ 越接近1，说明模型能解释的因变量（期末成绩）变异比例越高，解释力越强。模型 $R^2=0.82$ 说明在期末成绩的变化中有82%可以由模型的自变量（比如作业得分、期中成绩、测试分数等）解释，剩下变化影响可能来自模型未纳入的因素（比如考试心态、临场发挥等），模型泛化性能良好，未出现明显过拟合。

为了更清晰展示解释变量的重要性，绘制了解释变量重要性程度图，见图5。

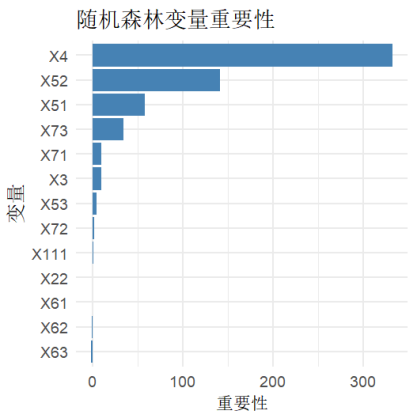


图5：随机森林变量重要性排序

由图5可知，期中成绩（X4）为最重要变量，对期末成绩影响最大，说明期中是阶段知识的综合验收，其覆盖范围、考查深度、评估维度最接近期末，且直接反映学生对中期核心知识的掌握程度，而这部分知识是期末备考的基础；X52、X51（测试成绩）次之；作业得分3（X73）排名第4，说明期末前的作业完成质量对成绩提升有针对性作用；出勤天数变量（X61-X63）重要性较低，班级与性别影响可忽略，表明成绩差异主要来自学生个体学习表现。

（三）LASSO 回归

为避免模型过于简单而欠拟合，又防止模型过度复杂而出现过拟合，使用LASSO回归选择解释变量，通过交叉验证确定两个关键参数，最小化交叉验证误差 $\lambda_{\min}=0.206$ 以及最小误差一个标准差内的最简模型 $\lambda_{1se}=2.108$ 。变量筛选结果见表6。

表6：变量筛选结果（ $\lambda_{1se}$ 模型）

| 变量          | 系数      |
|-------------|---------|
| (Intercept) | -41.816 |
| 性别 X2       | -1.99   |
| 月考成绩 X3     | 0.006   |
| 期中成绩 X4     | 0.796   |
| 测试1成绩 X51   | 0.154   |
| 测试2成绩 X52   | 0.080   |
| 出勤 X62      | 1.221   |

由LASSO变量筛选结果可知，最重要的解释变量为性别、月考成绩、期中成绩、测试1成绩、测试2成绩与出勤X62。期中成绩每提高1分，期末成绩预计提高0.796分。出勤天数每增加1天，期末成绩预计提高1.221分。两次测试成绩对期末成绩均有正向影响，性别差异也会对期末成绩产生影响。

四、结论与启示

（一）研究结论

考虑数学知识具有层层递进、环环相扣的特点，比如“全等三角形”是“轴对称”的基础，“一次函数”又会关联“反比例函数”，那么月考、期中这类阶段性检测内容是覆盖了一个阶段的模块知识，分数高、成绩好则意味着学生已经梳理明白了这一模块的知识逻辑，掌握了扎实的基础知识；而反过来，分数较低，成绩偏低的学生并未掌握知识，知识存在漏洞，如果不及及时补上，会影响后续新知识的学习，最终体现在期末成绩上。相比之下，作业、练习更偏向“单点知识巩固”，对知识体系的检验力度远不如阶段性检测作业完成质量对期末成绩的积极作用。因此如研究结果显示期末成绩的最有效的预测指标是期中成绩，期中成绩与期末成绩相关性最强，月考成绩和两次阶段性测试成绩也呈强正相关。

出勤天数对期末成绩影响较弱，因为出勤只是学习的“必要非充分条件”，它无法反映知识掌握程度，是出勤过程中有效的学习行为在影响成绩，因此出勤天数只是一个“表面指标”，和成绩的关联自然很弱。

性别因素存在显著差异，“显著”不是“差距大”，是

指这种差异并非由随机抽样的误差而导致的,是真实存在的规律,因为不同性别在数学学习上的思维特点、学习习惯会有所不同(比如男生擅长逻辑推理类题型,女生擅长计算和规范答题类题型)。

班级因素无显著影响,说明成绩差异主要源于学生个体学习表现,班级教学环境差异并不会引起显著影响,而是考虑期末成绩的核心影响因素的解释力很强(例如阶段性检测成绩、作业得分等),当模型中纳入了这些核心变量后,班级因素的影响就被“稀释”了——学生的成绩好坏,主要由他自身的阶段性表现决定,与学生所在的哪个班关系不大。

成绩分化的话,B班较A班更严重,说明B班学生在月考、期中这类练习后的补漏效果分层明显,基础扎实的学生能及时查漏补缺巩固知识,基础薄弱的学生知识漏洞本来就多,因此漏洞越积越多,也就导致差距逐渐拉大,自律的学生能通过作业和检测持续提分,而基础弱、学习习惯差的学生则跟不上,叠加之下

会进一步加剧整体分化。

## (二) 教学启示

从研究结果来看,提示教学时重视强化阶段性检测的反馈:重视期中、月考等阶段性测试的诊断功能,针对测试中暴露的知识漏洞及时开展专项辅导,尤其关注期中成绩低于50分的学生,提前干预避免成绩进一步下滑。

优化作业设计与反馈:提高作业针对性,尤其加强期末前的知识巩固类作业,通过高质量作业强化知识内化;同时关注作业完成质量差异较小的特点,设计分层作业满足不同学生需求。

实施个性化教学:针对B班等成绩分化严重的班级,建立分层教学机制,对学困生加强基础辅导,对优等生提供拓展性学习任务,缩小成绩差距。重视性别因素带来的成绩差异,因材施教,为不同性别学生提供适配的学习指导策略,同时避免因班级标签忽视个体发展需求。

## 参考文献

- [1] 教育部. 义务教育数学课程标准(2022年版)[S]. 北京:北京师范大学出版社,2022.
- [2] 李娟. 初中数学学业成绩影响因素的实证研究[J]. 教育学报,2020,16(3):89-96.
- [3] 刘伟. 项目式教学对八年级数学成绩的影响研究[J]. 课程·教材·教法,2021,41(5):112-117.
- [4] 张明. 学习兴趣对初中生数学成绩的影响及培养策略[J]. 数学教育学报,2019,28(2):45-50.
- [5] Hyde J S. Gender Differences in Mathematics Performance: A Meta-Analysis[J]. Psychological Bulletin,2005,131(2):109-135.
- [6] 王静. 教育公平视角下初中生数学成绩的性别差异研究[J]. 教育研究,2022,43(7):68-75.
- [7] 陈阳. 初中生课堂出勤与学业成绩的相关性分析[J]. 基础教育参考,2020(12):34-37.
- [8] 刘敏. 数学作业质量对初中生学业成绩的影响机制研究[J]. 数学通报,2021,60(8):23-28.
- [9] Smith A, Jones B. The Role of Monthly Tests in Predicting Final Mathematics Achievement of Middle School Students[J]. Journal of Educational Measurement,2020,57(3):412-428.
- [10] 王颖. 初中生期中与期末数学成绩的相关性研究[J]. 中国教育月刊,2022(S1):156-158.

# 基于 TF-IDF 和 DeBERTa 混合模型的数据目录语义检索系统研究

谷剑芳

河南省政务大数据中心, 河南 郑州 450000

DOI:10.61369/ASDS.2026010005

**摘要：** 本文以政务数据目录智能化搜索为研究目标，针对传统关键词匹配方法存在的语义理解不足、同义词覆盖不全等问题，提出了一种基于 TF-IDF 与 DeBERTa 算法的混合模型的轻量级语义搜索的解决方案。通过整合政务数据目录中的目录名称、申请依据、应用场景、标签等多维度数据构建语料库，利用中文分词、动态权重调整、领域术语扩展等技术，建立融合关键词精确匹配与语义相似度计算的双层检索机制。通过实践，本研究解决了语义识别场景中的三个核心问题：一是利用 TF-IDF 与 DeBERTa 的有机融合，发挥两个算法在关键词匹配和短句语义理解的优势，提高文本搜索的召回率；二是面向政务服务领域应用，建立丰富的知识库，建立专业术语与民间表述之间的映射桥梁，解决语义鸿沟跨越问题；三是通过对搜索结果的二次过滤，解决过度泛化所产生的“语义漂移”，进一步提升搜索的精确度。

**关键词：** 语义分析；TF-IDF；DeBERTa；混合模型；动态权重；数据目录

## Research on a Hybrid TF-IDF and DeBERTa Model for Semantic Retrieval in Data Catalog Systems

Gu Jianfang

The Henan Provincial Government Big Data Center, Zhengzhou, Henan 450000

**Abstract：** This study focuses on intelligent search for government data catalogs. To address the limitations of traditional keyword matching methods—such as inadequate semantic understanding and incomplete synonym coverage—we propose a lightweight semantic search algorithm based on a hybrid model integrating TF-IDF and BERT algorithms. This research aims to build a lightweight intelligent search model for government service scenarios, addressing three core issues. First, it leverages the organic integration of TF-IDF and BERT to utilize the advantages of both algorithms in keyword matching and short sentence semantic understanding, thereby improving the recall rate of text search. Second, it establishes a rich knowledge base for government service applications, creating a mapping bridge between professional terms and folk expressions to overcome the semantic gap. Third, through secondary filtering of search results, it addresses over-generalization and potential "semantic drift" from Word2Vec, further enhancing the precision of government data directory search and providing an efficient intelligent search solution for government data resource sharing scenarios.

**Keywords：** semantic analysis; TF-IDF; DeBERTa; hybrid model; dynamic weight; data catalog

## 引言

在数字化转型背景下，政务数据共享已成为提升政府治理能力的关键路径。截至2023年底，全国一体化政务服务平台使用总量超过888亿人次，其中，证照共享服务体系持续优化，已汇聚全国31个地区和26个国务院部门58亿条目录，累计提供电子证照共享服务97亿余次，持续推动“减证便民”<sup>[1]</sup>。本研究致力于构建轻量级政务数据目录智能搜索系统，实现以下目标：建立跨部门术语知识库，覆盖“社会保险”“行政审批”等12个高频领域的586个同义词对设计动态权重调节机制，使 TF-IDF 与 BERT 的混合权重随查询特征自适应变化开发面向政务场景的查询扩展算法，解决“个人养老金→基本养老保险待遇”等专业映射问题。本研究提出的混合搜索模型，通过融合 TF-IDF 的精准匹配优势与 BERT 的语义泛化能力，经过实验证明，有效提升跨部门目录检索的准确率。特别在社保、卫生、教育、交通、公积金、房地产等高频领域，解决术语标准不统一带来的搜索盲区问题，为构建全国一体化政务大数据体系提供技术支撑。

作者简介：谷剑芳，女，河南省政务大数据中心数据治理部，副部长，研究方向：数字政府建设、政务数据治理、数据要素流通；邮箱：13525568581@sina.com.cn。



## 一、相关工作

### （一）数据目录检索系统的现状

政务数据目录作为政府数据资源的管理核心，具有以下特征：

1. 结构化与非结构化数据并存：目录名称、标签类型等字段具有明确的结构化特征，而应用场景、申请依据等字段多为自然语言描述的非结构化文本。

2. 领域术语密集性：数据目录和应用场景包含大量政策文件特有的专业术语（如“容缺受理”“一网通办”等）和部门专属概念（如“社保基数核定”“税控云监管”等）。

3. 动态更新需求：随着政务服务业务的扩展，跨部门业务协同的丰富，数据目录会持续增加、变更。如：2024年，随着国办推进“高效办成一件事”，大部分一件事都需要跨部门协同办理，新的目录对应新的业务，要求目录搜索算法同步提升对文本的推理能力。

以中部K省为例，政务数据目录超过两千条，目录名称平均长度13.7个字符，包含专有名词占比12.5%。应用场景文本平均长度44.3字符，每句覆盖专用术语：0.11个，专业术语占比81.68%。申请依据文本平均长度75.6字符，文件引用占比53.95%，标签体系包含11类58个。这些特征决定了传统关键词匹配方法难以满足需求，需引入分词、词向量和语义分析等技术。现行目录检索系统主要采用关键词匹配或分词匹配的方法，从语义精准匹配的角度，普遍存在三大痛点：其一，传统关键词匹配对“企业-公司”“养老-退休金”等政务场景高频同义词缺乏映射能力；其二，缺乏跨领域术语差异解析能力，目录名称与应用场景的语义割裂导致跨部门检索效率低下；其三，政策术语快速迭代得静态词库难以适应“一件事一次办”“免申即享”等新型服务模式的搜索需求。

随着政务服务领域“三融五跨”业务的快速增长，数据共享也越来越高频。2025年，国家数据局发布《数字中国发展报告（2024年）》公布<sup>[2]</sup>：近5年，各地区政务服务平台数据共享累计超过5400亿次。数据目录作为数据共享的驱动，为用数单位找数、供数单位发布服务，提供了基本工作界面和载体。近年来，随着政务数据资源目录和数据共享业务的不断增长，数据目录查询量不断增加，传统关键词检索机制暴露两大核心短板：其一，难以精准识别“生存认证/死亡记录”等业务场景同义词关联；其二，缺乏跨领域术语差异解析能力（如“税控云、成品油监管”等专业表述辨识困难）。本研究是为了促进数据共享，通过算法实现智能化数据目录搜索，为用数单位找到需要的数据提供便利。

### （二）TF-IDF 模型概述

1. TF-IDF 的基本原理。Luhn<sup>[3]</sup>的 TF-IDF 加权策略通过逆文本频率强化专业术语表征，TF 是词频 (Term Frequency)，IDF 是逆文本频率指数 (Inverse Document Frequency)，其核心思想是如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力<sup>[4]</sup>。该方法适合用来分类，组成要素包括词频、逆文本频率和 TF-IDF 值。词频，是词语在单个网页中出现的频率，反映词语

对当前网页的重要性。逆文本频率，用于衡量词语在整个语料库中的稀有程度，稀有词具有更高的区分能力。TF-IDF 值越高，词语对网页的代表性越强。

2. TF-IDF 在信息检索中的应用。TF-IDF 在处理中文政务数据时面临两大挑战：一是政务场景特有的缩略语（如“一网通办”“跨省通办”）导致特征稀疏；二是政策文件的多义表述造成语义漂移。Mikolov<sup>[5]</sup>提出的 Word2Vec 模型开创了分布式词向量新范式，Lee<sup>[6]</sup>将其扩展为 Doc2Vec 实现文档级语义表征。在政务领域，Zhang<sup>[7]</sup>构建的政务词向量库 GOV2VEC 包含 5.7 万个政策术语，但未解决新政策术语的动态更新问题。

### （三）DeBERTa 模型概述

1. DeBERTa 的前身：BERT 模型的基本原理。2018 年 Google AI 研究院提出 BERT 模型。与传统的单向语言模型不同，BERT 利用 Transformer 架构中的多层 Encoder 结构，实现了对输入序列中每个位置的双向依赖建模，这使得它能够同时考虑输入文本的上下文信息。这种双向性显著增强了模型对复杂语义关系的理解能力，进而提高了在各种自然语言处理任务中的表现。

BERT 在自然语言处理中的应用。翁克瑞<sup>[8]</sup>等结合 BERT 预训练模型与改进的 TextCNN 架构，通过分析社交媒体（如抖音）评论数据，识别公众对新能源汽车、碳中和等政策的意愿倾向，分类准确率达 82.0%–86.4%，显著优于传统深度学习方法。刘青<sup>[9]</sup>等利用 BERT 语义模型 + 企业知识图谱，从专利摘要中识别劳动节约型技术创新（如自动化设备专利），实现量化劳动力成本上升对企业技术创新的促进作用。采用 BERT 模型实现语义搜索，其 768 维向量的计算复杂度进一步提升语义搜索的准确性，同时也会增加内存开销。

2. DeBERTa: 解码增强的 BERT 与解耦注意力大模型。DeBERTa 具有解耦注意力机制和增强的掩码解码器<sup>[10]</sup>。在解耦注意力机制方面，DeBERTa 模型则采用双向量表示法，分别编码词的内容和位置信息，并通过基于内容和相对位置的解耦矩阵计算词间注意力权重。在增强型掩码解码器方面，与 BERT 类似，DeBERTa 采用掩码语言建模 (MLM) 进行预训练，通过分析掩码词周围的上下文词汇来预测被遮蔽的词语。通过解耦注意力机制捕捉上下文词汇的内容特征和相对位置，通过 MLM 预训练有效处理词汇的绝对位置信息，两者结合，能够更有效发现词语之间的关联关系。

## 二、混合模型设计

### （一）需求分析

随着政务服务领域“三融五跨”业务的快速增长，数据共享也越来越高频。数据目录作为数据共享的驱动，为用数单位找数、供数单位发布服务，提供了基本工作界面和载体。近年来，随着政务数据资源目录和数据共享业务的不断增长，数据目录的搜索量不断增加，传统关键词检索机制暴露两大核心短板：一是难以精准识别“生存认证/死亡记录”等业务场景同义词关联；二是缺乏跨领域术语差异解析能力（如“税控云、成品油监管”等专

业表述辨识困难)。为了促进数据共享,通过算法实现智能化数据目录搜索,为用数单位找到需要的数据提供便利。确定以下业务逻辑:

1. 明确分析对象。以数据共享订阅历史记录为搜索对象,匹配的字段包括目录名称、应用场景,以及与目录有关联的标签、场景分类、提供部门等字段。

2. 建立权重参数。按照直接和间接表达目录实际用途的价值,为目录名称含义、应用场景、申请依据、场景分类、标签等字段建立相似度计算权重参数。

3. 建立语义识别知识库。语义识别具有场景化的特点,每个应用领域都有行业术语。为了让大模型能够理解行业术语,精准识别文本含义,建立政务服务领域字典和同义词词典,为了过滤语言中的助词、语气词、副词,建立停用字词词典。

(二) 系统架构

1. 总体设计。政务数据目录智能化搜索系统的核心架构采用分层设计理念(如图1),通过数据预层、算法层、融合层、增强功能四个核心模块的协同工作,实现从原始数据到精准搜索结果的完整处理流程。构建 TF-IDF 与 DeBERTa 的双通道语义理解体系,突破传统单一算法的局限性。通过 TF-IDF 向量空间(精确匹配目录名称等结构化字段)与 DeBERTa 语义识别的并行计算,实现关键词匹配与语义理解的优势互补。在技术实现上,采用 hstack 矩阵拼接方式融合两种向量表示,并创新性地引入动态权重调整机制:通过将等参数权重的相似度统计特征(均值、标准差、极值)构建状态向量,输入强化学习模型,模型通过预测动作概率分布输出动态权重分配方案,利用加权 MSE 损失函数结合用户反馈奖励进行梯度下降优化,最终实现权重参数的在线自适应调整,显著提升搜索结果中语义相关性与用户意图的匹配精度。

2. 模块划分。主要功能方法包括数据层、算法层、融合层、增强功能、分析模块和接口层,见图1。



图1: 整体模型架构

(三) 数据预处理

1. 数据清洗与标准化。将对读取的目录名称、申请依据等字段内容去除异常字符,按权重系数拼接。

2. 语料准备。

(1) 中文分词:采用结巴分词工具进行细粒度切分。支持 n-gram (1, 2), 增强分词效果,提高文本特征的表征能力。

(2) 停用词过滤:移除“的”“是”等无意义词汇。

(3) 领域术语注入:加载政务专属词典。

(四) TF-IDF 模型实现

1. 词频与逆文档频率计算。TF-IDF (Term Frequency-Inverse Document Frequency) 通过统计词频衡量词语重要性,其中,  $f_{t,d}$  为词  $t$  在文档  $d$  中的出现次数,  $N$  为总文档数,  $n_t$  为包含词  $t$  的文档数。在本系统实现中,通过 TF-IDF 对文本的向量化实现以下优化:

(1) 双向 n-gram:设置 ngram\_range=(1,2),捕获“社保\_缴费”等复合词。

(2) 动态剪枝: min\_df=0.001 过滤低频噪声词, max\_df=0.95 去除泛化停用词。

(3) 权重增强:目录名称字段在语料构造时重复3次,提升关键字段权重。

在文本处理流程中,通过双向 n-gram 解析设置 ngram\_range=(1,2) 构建弹性词窗,精准捕获“社保缴费”类复合语义单元;结合动态剪枝机制,运用 min\_df=0.001 智能过滤低频噪声词汇并借助 max\_df=0.95 精准剔除泛化停用词;同时采用语义权重强化策略,在语料构造时对“目录名称”字段内容复制3份,相当于进行了3次语义强化,实现关键字段的战略级权重提升。

2. TF-IDF 向量化。TF-IDF 的优势是对关键词的匹配,但是缺乏语义关联分析能力:

(1) 语义盲区:无法识别“企业-公司”等同义词,导致“省直国企名录”无法匹配“省级国有公司登记信息”。

(2) 语境缺失:对“养老”一词,难以区分社保场景中的“养老保险办理”与民政场景中的“养老机构监管”。

(3) 长尾效应:IDF 逆文档率算法,让政策文件中的一些低频专业术语(如“非税收入划转”)IDF 值过高,从而产生过度加权。

TF-IDF 算法虽在关键词精准匹配层面展现优势,却存在语义关联解析的系统性局限:其机械的字面匹配机制无法辨识“企业-公司”等语义等价关系,导致“省直国企名录”与“省级国有公司登记信息”形成检索断层;对“养老”等多义词汇缺乏上下文感知能力,难以区分社会保障领域的“养老保险办理”与民政管理范畴的“养老机构监管”应用场景;更因 IDF 逆文档频率算法的特性,使文本中“非税收入划转”等低频专业术语产生异常加权,引发算法权重失衡。实证研究表明,基于 TF-IDF 的单一模型在标准测试集上相似度较低,超过 50% 的相关目录无法检索出来,语义鸿沟导致的跨领域匹配失效构成主要误差源。

(五) DeBERTa 模型实现

1. DeBERTa 模型选择与微调。采用 deberta-v3-base 预训

练模型（12层 Transformer，768维隐藏层），通过以下策略实现计算效率与语义精度的平衡：

文档向量预计算：系统初始化时离线计算全部目录的 DeBERTa 向量，存储为 numpy 数组。

批处理优化：采用 16 条 / 批的小批量处理，避免内存溢出。

缓存机制：将预计算结果保存为外部文件，后续直接加载，避免重复计算。

向量提取策略：使用 [CLS] 标记的向量作为文档级语义表征，相比传统词向量平均法更能保留上下文信息。`vector = outputs.last_hidden_state[0, 0, :].numpy()`

## 2. 语义嵌入生成

### 第一步：语料增强处理

（1）停用词过滤算法：建立多级过滤规则，包括 163 个通用停用词，如“查询”“获取”等操作动词，根据 TF-IDF 权重自动发现低频噪声词。

（2）启用领域术语：构建包含三级结构的领域知识库（如图 1），包括核心结构（政务词典基础同义词库、200+ 领域术语库）、重点领域（社保、税务、市场监管等重点政务领域术语）、特色应用（智慧城市、电子证照、区块链存证等），支持“医保→医疗保险→医疗保障局”的递进式扩展。

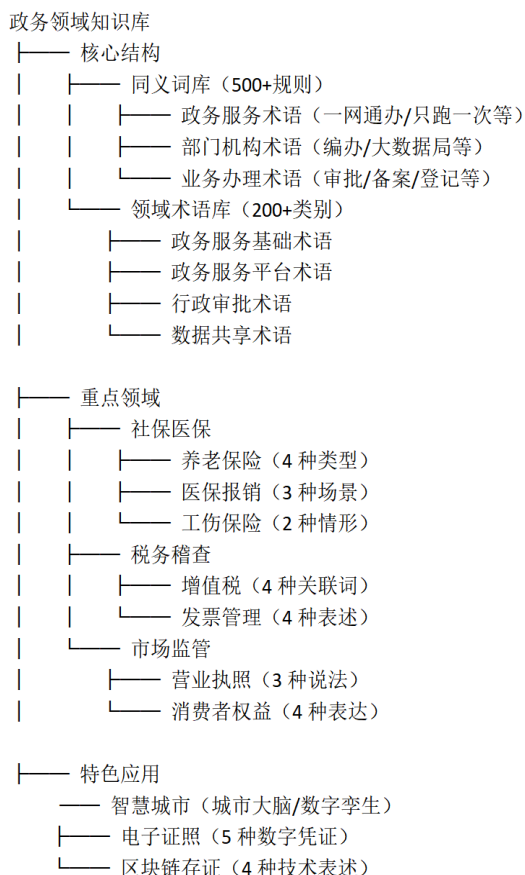


图 2：三级结构的领域知识库

第二步：采取混合向量化算法。建立双通道向量空间模型：建立 TF-IDF 通道，生成稀疏特征矩阵。建立 DeBERTa 通道：训练词向量模型并计算文档向量。

（1）优化设置 TF-IDF 算法参数：包括设置次线性词频缩放参数（Sublinear TF=true），避免高频词因绝对频次过高而主导权重，从而更合理地反映词语在文档中的重要性。设置动态 DF 阈值（min\_df=0.001）过滤长尾词。

（2）DeBERTa 模型训练：捕捉语义关联，解决一词多义问题，平均词向量生成文档表示，多线程并行计算显著提升 BERT 训练效率。

（3）动态策略：引入了强化学习（RL）来动态调整，影响算法精准度的参数权重：TF-IDF、应用场景、目录名称、申请依据、标签类型、标签名称、语义。

（4）用户参与：根据用户反馈的无效目录来过滤无关结果，提升搜索质量。

### 第三步：动态权重调节

基于相似度分布的动态权重算法，根据各字段相似度统计量动态调整权重  $w_i$ ，实现表现好的字段权重高，波动大的字段权重低的智能分配。

（1）波动抑制：当某字段相似度方差过大时自动降低权重。

（2）重点突出：对稳定高匹配字段给予权重奖励。

（3）归一化处理：保证权重总和为 1。

### 第四步：对检索结果实施二次过滤

为强化关键词匹配效果，采取 TF-IDF 硬过滤，经过多次测算，且判断条件为：目录名称相似度 >0.15 或应用场景 >0.15。

## （六）系统实现

### 1. 开发环境与工具

（1）硬件平台：华为笔记本电脑（32GB 内存，INTEL CPU）

（2）软件环境：Python 3.8、WINDOWS 操作系统

（3）模型：

模型 1：TF-IDF 模型（n-gram 分词范围：1-2）

模型 2：Word2Vec 模型（向量维度 80）

模型 3：TF-IDF 混合 Word2Vec 模型（向量维度 80、n-gram 分词范围：1-2）

模型 4：TF-IDF 混合 BERT 模型（向量维度 768、n-gram 分词范围：1-2）

模型 5：TF-IDF 混合 DeBERTa 模型（向量维度 768、n-gram 分词范围：1-2）

### 2. 系统功能模块

#### （1）主要功能

采用 deberta-v3-base 预训练模型（12 层 Transformer，768 维隐藏层），使用 [CLS] 标记的向量作为文档级语义表征，相比传统词向量平均法更能保留上下文信息。模型要解决三个问题：一是利用 TF-IDF 与 deberta-v3-base 的有机融合，发挥两个算法在关键词匹配和语义理解的优势，提高文本搜索的召回率。二是面向政务服务领域应用，建立丰富的知识库，建立专业术语与民间表述之间的映射桥梁，解决语义鸿沟跨越问题。三是通过对搜索结果的二次过滤，解决过度泛化产生的“语义漂移”，进一步提升政务数据目录搜索的精确度。通过以下策略实现计算



效率与语义精度的平衡。

文档向量预计算：系统初始化时，离线计算全部目录的 DeBERTa 向量，存储为 numpy 数组。

批处理优化：采用 16 条 / 批的小批量处理，避免内存溢出。

缓存机制：将预计算结果保存为 npy 文件，后续直接加载。

## （2）用户输入与查询处理

首先使用自定义的词典对语料进行 jieba 分词，然后对分词结果进行向量化计算，计算词向量的平均值作为文档向量，其中空文档用零向量表示。

## （3）检索结果生成

以查询“个人养老金”为例，直接用模型计算，取相似度值前 15 个。15 个结果中，相似度最高的第 4 项是“机关事业单位工勤岗位登记电子证照”，与养老金没有任务关系，第 7 项“城乡居民基本养老保险待遇领取信息”属于个人养老金，但是排序靠后。

## 3. 性能优化

### （1）优化检索速度

①领域语料注入：定义政务领域同义词和术语。本案例设置了覆盖教育、医疗、税务、社保、就业、卫生、交通、海关、消防、金融、职称、自然资源等政务行业的 128 组同义词和 83 组覆盖智慧城市、数据共享、特色政府服务等领域术语。

②混合相似度计算：通过矩阵拼接实现 TF-IDF 稀疏矩阵与 DeBERTa 稠密向量的协同计算。

（2）提升结果准确性。DeBERTa 算法核心是双向 Transformer 编码器，通过多层自注意力机制和双向上下文学习生成动态词嵌入，基于词向量简单平均的文档级语义表征方法存在明显局限，其聚合过程会损失词汇序列的位置信息，从而产生语义漂移，例如，搜索“个人养老金”，检索出与“个人”相关，但与“养老金”无关的住房贷款、强制执行等数据目录。避免因同义词扩展导致的过度泛化问题，解决算法在语义解读方面产生的“语义漂移”，即避免语义相关性分析关联到与查询词相关性不大的目录，确保结果与查询存在实质性关联，对搜索结果进行停用词二次过滤，并过滤掉目录名称相似度小于 0.1 或应用场景相似度小于 0.1 的结果。经过二次过滤，均与个人养老金有直接关系的四条数据目录：离退休人员养老保险退休信息、城镇企业职工基本养老保险待遇领取信息、机关事业单位养老保险待遇领取信息、城乡居民基本养老保险待遇领取信息，排序在 1 至 4，有效数据召回显著提升。

### （3）建立动态权重

构建参数权重的深度学习模型，针对影响相似度识别的 7 个参数 RL\_WEIGHT\_FIELDS（TF-IDF、应用场景、目录名称、申请依据、标签类型、标签名称、DeBERTa 语义），计算每个字段相似度的均值、标准差、最大值、最小值，组合成状态向量。将 RL\_WEIGHT\_FIELDS 参数的动态权重状态向量，输入强化学习模型，模型通过预测动作概率分布输出动态权重分配方案，利用加权 MSE 损失函数结合用户反馈奖励（如过滤无关结果惩罚 -1/-2 分）进行梯度下降优化，最终实现权重参数的在线自适应调整，显著提升搜索结果中语义相关性与用户意图的匹配精度。工程化流程实现如下：

## 动态权重生成过程

①状态特征提取：对每个查询，系统实时计算 RL\_WEIGHT\_FIELDS 中各字段（如 TF-IDF 整体、场景相似度、语义等）的相似度统计特征，包括均值、标准差、最大值、最小值，形成长度为  $7 \times 4 = 28$  维的状态向量（7 个参数  $\times$  4 种统计量）。

②模型推理：将状态向量输入 PyTorch 神经网络（包含 128 单元隐藏层和 Softmax 输出层），输出各字段的初始权重概率分布（如 [0.3, 0.2, 0.1, ..., 0.15]）。

③归一化处理：通过 Softmax 归一化确保权重总和为 1，例如将 [0.3, 0.2, 0.1, 0.15, 0.05, 0.1, 0.1] 转换为 [0.25, 0.17, 0.08, 0.13, 0.04, 0.08, 0.07]。

④在线调整：根据用户反馈（如标记无关目录）动态更新模型参数，例如若 BERT 权重过高导致语义偏差，后续查询会降低 BERT 的权重分配。

## LR 学习过程（强化学习训练）

①损失函数设计：采用加权 MSE 损失函数，公式为：

$$\text{Loss} = \text{MSE}(\text{predicted\_weights}, \text{true\_weights}) \times (1 + \text{Reward})$$

其中 Reward 根据用户行为动态调整（如过滤无关结果时 Reward=-2，点击相关结果时 Reward=+1）。

②探索与利用策略：通过  $\epsilon$ -greedy 策略平衡探索（随机采样权重）与利用（按模型预测分配权重），初始  $\epsilon = 0.2$ ，每轮搜索后衰减至 0.99 倍。

③梯度更新：使用 Adam 优化器（学习率  $1e-4$ ）进行反向传播，重点优化导致低奖励的动作参数，例如对产生高惩罚权重的神经元增大反向传播梯度。

④经验回放机制：将状态-动作-奖励三元组存入经验池（rl\_training\_data），当积累 100 条数据后进行批量训练，提升样本利用效率。

## 4. 实验与评估

### （1）实验设计

①数据集选择。本实验采用政务数据资源共享平台的实际业务数据，构建包含 18,456 条目录记录的测试数据集。数据字段包括目录名称（平均长度 13.7 字符）、应用场景（平均长度 44.3 字符）、申请依据（平均长度 75.6 字符）等关键字段。实验数据集经过严格预处理：统一数字格式和字段内容。使用相同的分词词典、同义词扩展、停用词、弃用词。通过同一组测试输入，取相似度前 10 个结果，检验不同模型的检索结果与业务的相关性。

②实验指标。选择高频使用的数据目录内容：企业社保缴费，能够体现个人行为 and 生存状态的信息，包括养老、工伤和医保缴费，死亡，火化，住院和门诊结算等，查询公积金贷款相关目录，文化旅游政务服务和企业开办一件事相关目录。

### （2）实验结果分析

①多模型运行效率对比。从 1115 条检索结果中，针对上述三个模型从评价指标（检索精度、查全率和排序质量）和热词两个方面进行效果评估。从检索结果的热词分布看，WORD2VEC 模型的热词虽然多，但是与检索输入的相关性不高，TF-IDF 与

DeBERTa 混合模型检索结果热词比 TF-IDF 模型更优，不仅与检索输入的相关性高，输出的有效内容更多。以搜索“企业社保缴费”为例，TF-IDF 模型的高频热词是“企业”“养老保险”和“职工”，WORD2VEC 模型的高频热词是“民企”“企业”和“政策”，TF-IDF 与 Word2Vec 混合模型的高频热词是“缴费”“企业”和“养老保险”，TF-IDF 与 DeBERTa 混合模型的高频热词是“缴费”“社会保险”“职工参保”。可以看出，TF-IDF 与 DeBERTa 混合模型的搜索结果更符合业务期望。

②混合模型的性能评估

定义三个评价指标：精确率、召回率、F1和排序质量<sup>[11]</sup>，计算值见表1，TF-IDF、Word2Vec、TF-IDF 混合 Word2Vec、TF-IDF 混合 BERT 和 TF-IDF 混合 DeBERTa 五个模型的效果比对比示意图见图3。定义和计算方法分别如下：

精确率 (Precision)

定义：强调预测的准确性，不受负类样本数量的直接影响，适合不平衡数据集评估。

公式：Precision = TP / (TP + FP)

举例：例如，top\_k=10，有3个相关结果，则 precision@10=0.3。

召回率 (Recall)

定义：系统检索到的相关文档占有所有相关文档的比例。

公式：recall@k = 检索到的相关数 / 总相关数

举例：总共有5个相关文档，检索到3个，则 recall@k=0.6。

MRR 排序质量 (RankQuality)

定义：首个相关文档在结果中的排名的倒数，未找到则为0。平均倒数排名，取值范围：

0,1，值越大表示排序质量越好，主要衡量模型返回结果的排序合理性。

公式：MRR = 1 / rank\_of\_first\_relevant\_doc（如果存在）

举例：第一个相关文档在第3位，则 MRR=1/3 ≈ 0.333。

F1 综合性能（精确率 + 召回率调和平均数）

定义：分类模型评估中一个综合性的指标，用于衡量模型在精确率和召回率之间的平衡能力。

公式：F1=2 × (Precision × Recall)/(Precision+Recall)，综合

反映模型的整体性能。

通过业务专家标注五个模型的搜索结果并建立评估模型，从评估结果看，TF-IDF 混合 DeBERTa 的 MRR 和精确率最优，展现的结果是相同搜索对象，该模型的搜索排序和结果最优，该模型的指标 F1 综合性能、召回率和准确率排名第二。

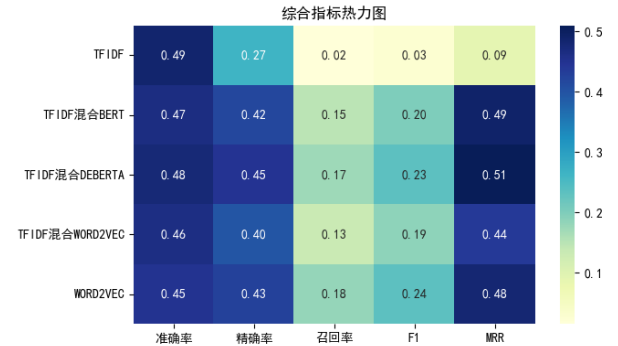


图3：五种模型综合热力图

三、总结

本研究提出 TF-IDF 与 DeBERTa 的动态混合模型，在政务数据目录搜索场景中实现更具有智能化、更贴近用户意图的数据目录。通过向量预计算、缓存等优化技术，使系统响应速度持续提示，支持万级数据目录的文本语义检索。通过语义分析，发现跨部门数据目录的相关性，为挖掘政务服务领域更多具有并联串联办理的“一件事”场景提供线索。未来，需要在三个方面探索政务数据目录智能化搜索的能力：一是构建政务领域预训练语料库。引入主动学习优化同义词库，提升词语推理的理解能力，从而提高语义匹配的准确性。二是持续完善政务语言资源库。包含术语体系、专用词典、同义词库、停用词表及弃用词等核心要素，强化语义解析的领域适应性。三是采取搜索的兜底措施。针对词向量语义分析的不确定性，某些情况下无法正确返回结果甚至结果为空情况，采用分词比对的方法进行兜底，为用户提供有效的结果。

参考文献

[1] 国家互联网信息办公室. 国家信息化发展报告（2023年）[EB/OL].[https://www.cac.gov.cn/2024-09/06/c\\_1727308607362592.htm](https://www.cac.gov.cn/2024-09/06/c_1727308607362592.htm).  
[2] 国家数据局. 数字中国发展报告（2024年）[EB/OL].[https://www.nda.gov.cn/sjj/zhuanti/sjzgzxd/szzgbg/0605/20240830180401077761745\\_pc.html](https://www.nda.gov.cn/sjj/zhuanti/sjzgzxd/szzgbg/0605/20240830180401077761745_pc.html).  
[3] Luhn, H.P. The automatic creation of literature abstracts[J].1958.《IBM Journal of Research and Development》第2卷，159-165.  
[4] 百度百科“tf-idf”词条[EB/OL].<https://baike.baidu.com/item/TF-IDF/8816134>.  
[5] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.  
[6] Lee J, Kim S, Song Y. Doc2Vec-based semantic document retrieval in big data environments[J]. Future Generation Computer Systems, 2020, 112: 997-1005.  
[7] Zhang Y, Li J, Song Y. GOV2VEC: A domain-specific word embedding model for government documents[C].Proceedings of the 2021 IEEE International Conference on Big Data. IEEE, 2021: 1023-1032.  
[8] 翁克瑞, 周雅洁, 於世为. 基于 BERT 的多层次特征融合的舆情文本政策意愿识别模型研究 [J]. 中国地质大学学报 ( 社会科学版 ), 2025, 25(01): 131-140.  
[9] 刘青, 肖柏高. 劳动力成本与劳动节约型技术创新——来自 AI 语言模型和专利文本的证据 [J]. 经济研究, 2023, 58(02): 74-90.  
[10] Engcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen.《DeBERTa: Decoding-enhanced BERT with Disentangled Attention》[EB/OL].<https://arxiv.org/abs/2006.03654>.  
[11] 王国霞, 刘贺平. 个性化推荐系统综述 [J]. 计算机工程与应用, 2012, 48(07): 66-76.



# 融合舆情的黄金期货价格预测：RFECV 特征选择与 DNN 模型

杨紫帆<sup>1\*</sup>, 赵艺晴<sup>1</sup>, 徐浩然<sup>2</sup>, 潘晨宇<sup>1</sup>, 肖尧<sup>1</sup>

1. 上海对外经贸大学 统计与数据科学学院, 上海 201620

2. 上海对外经贸大学 国际经贸学院, 上海 201620

DOI:10.61369/ASDS.2026010006

**摘 要：** 本文提出了一个面向黄金期货价格预测的“情绪指标增强 + 特征筛选 + 深度学习”的统一建模框架，重点探讨将多平台中文评论转化为可计算情绪特征，并与传统市场变量融合以提升预测精度。首先，基于情感词典与分词结果构建情绪强度，并进一步给出情绪波动率与7日滚动相关的一致性系数，使文本信息以时间序列形式进入模型输入。其次，引入以 SVR 为基学习器的 RFECV 进行递归特征消除与交叉验证筛选，保留关键价格变量、技术指标及情绪特征，降低维度并抑制噪声。最后，构建并对比 CNN- 线性回归混合模型与 DNN 模型进行预测评估，结果表明加入情绪指标后两类模型的误差指标整体改善，且对价格局部起伏与拐点刻画更贴近真实走势。该模型的作用与优点在于：可将非结构化文本信号标准化为通用可迁移特征，结合特征筛选与深度模型提高稳健性与泛化能力，适用于多类金融时间序列的预测任务。

**关 键 词：** 深度神经网络；卷积神经网络；RFECV 递归特征消除；支持向量回归；黄金期货价格预测

## Gold Futures Price Forecasting Incorporating Public Sentiment: RFECV Feature Selection and DNN Model

Yang Zifan<sup>1\*</sup>, Zhao Yiqing<sup>1</sup>, Xu Haoran<sup>2</sup>, Pan Chenyu<sup>1</sup>, Xiao Yao<sup>1</sup>

1.School of Statistics and Data Science, Shanghai University of International Business and Economic, Shanghai 201620

2.School of International Business, Shanghai University of International Business and Economic, Shanghai 201620

**Abstract :** This paper proposes a unified modeling framework for gold futures price forecasting, integrating “sentiment indicator enhancement + feature selection + deep learning.” It focuses on converting multi-platform Chinese comments into computable sentiment features and fusing them with traditional market variables to enhance prediction accuracy. First, sentiment intensity is constructed based on sentiment dictionaries and word segmentation results. Furthermore, a consistency coefficient is derived for the rolling 7-day correlation between sentiment volatility and price movements, enabling textual information to enter the model as time series data. Second, RFECV with SVR as the base learner is introduced for recursive feature elimination and cross-validation screening, retaining key price variables, technical indicators, and sentiment features to reduce dimensionality and suppress noise. Finally, a CNN-linear regression hybrid model and a DNN model are constructed and compared for prediction evaluation. Results demonstrate that incorporating sentiment indicators improves overall error metrics for both models and better captures local price fluctuations and inflection points aligned with actual market movements. The model's utility and advantages lie in: standardizing unstructured text signals into universal transferable features; enhancing robustness and generalization through feature selection and deep learning; and its applicability to diverse financial time series forecasting tasks.

**Keywords :** Deep Neural Network(DNN); Convolutional Neural Network (CNN); Recursive Feature Elimination with Cross-Validation (RFECV); Support Vector Regression (SVR); gold futures price forecasting

作者简介：

赵艺晴，上海对外经贸大学统计与数据科学学院，本科生，研究方向为经济统计学；

徐浩然，上海对外经贸大学国际经贸学院，本科生，研究方向为数字经济下的企业升级；

潘晨宇，上海对外经贸大学统计与数据科学学院，本科生，研究方向为数字经济与大数据统计应用；

肖尧，上海对外经贸大学统计与数据科学学院，本科生，研究方向为数据科学与大数据技术。

通讯作者：杨紫帆，上海对外经贸大学统计与数据科学学院，本科生，研究方向为统计学与机器学习。

## 引言

本文聚焦于金融时间序列在噪声强、驱动因素复杂条件下的短期预测，旨在构建一个可同时利用市场数值信息与网络文本信号、并具备可迁移性的算法模型框架<sup>[1]</sup>。首先，从通用预测任务出发指出：仅依赖价格等结构化变量的模型往往难以反映突发信息带来的预期变化，而仅用文本信号又容易受平台差异与时序对齐影响，特征稳定性不足<sup>[2]</sup>。其次，本文提出“情绪特征构建—特征筛选—深度预测”的流程：将多源中文评论转化为可滚动更新的情绪强度与一致性指标，并与技术指标共同作为输入；再通过递归特征消除与交叉验证筛去冗余变量，降低噪声干扰<sup>[3]</sup>。最后，在统一输入与筛选机制下构建并对比不同深度模型，以黄金期货价格为例展示框架的通用性与可扩展性<sup>[4]</sup>。实验结果表明，该框架在波动刻画与整体误差上均优于仅使用传统变量的基线方法。

## 一、理论方法介绍

### （一）情绪强度指标构建

情绪强度用于衡量文本情感倾向的方向与强弱程度：正值表示积极，负值表示消极，绝对值越大说明情绪表达越强。本文对单条评论的情绪得分采用平均情感权重法（Average Sentiment Weighting, ASW），步骤如下<sup>[5]</sup>：

文本分词：使用 jieba 分词对中文评论进行分词，得到词语序列；

情感词匹配：遍历词语，若该词出现在情感词典中，则提取其情感得分；

情感得分计算：将匹配到的情感词得分取算术平均，作为该条评论的最终情绪得分。其表达式为：

$$S = \frac{1}{N} \sum_{i=1}^N \text{sentiment}(W_i) \quad (1)$$

其中： $S$  为评论的情绪得分， $N$  为评论中匹配到的情感词数量， $\text{sentiment}(W_i)$  为词语  $W_i$  在情感词典中的得分。若评论中未出现任何情感词，则令  $S = 0$ 。

采用平均情感权重法的主要原因是它更关注“情绪表达”的强度水平，而不是情感词出现的数量，从而减少评论长短差异带来的偏差。将评论情绪得分按时间汇总后，可以把市场心理转化为可建模的量化序列。在实际市场中，情绪强度往往会在重大事件前后出现更明显的变化，并可能与波动放大和阶段性趋势有关。

### （二）情绪波动率的构建

情绪波动率用于衡量市场情绪在短期内变化的幅度，反映投资者情绪的不稳定性与突变特征<sup>[6]</sup>。设情绪强度时间序列为  $s_t, t=1^T$ ，其中  $s_t$  表示第  $t$  日的情绪强度值。本文用相邻两期情绪强度的相对变化幅度定义情绪波动率，公式为：

$$v_t = \left| \frac{s_t - s_{t-1}}{s_{t-1}} \right| \times 100\% \quad (2)$$

当  $s_{t-1} = 0$  时，可采用数值平滑技术，避免分母为 0 造成计算不稳定。

情绪波动率序列  $v_{t=2}^T$  作为相对指标，可以削弱情绪强度量纲的影响，便于不同阶段进行比较；同时该指标恒为非负，通常更容易呈现右偏分布，这也更贴近金融时间序列中“波动”的常见统

计形态。

### （三）情绪一致性的构建

情绪一致性是衡量情绪强度序列在短期内是否“延续”或“反转”的关键指标。它关注前一日的情绪对后一日情绪的影响方向与程度，可看作在滚动窗口内对一阶相关结构的刻画，从而识别情绪变化的持续性特征或反转性特征<sup>[7]</sup>。

#### 1. 数学模型与计算框架

设情绪强度时间序列数据集为有序集合： $D = (d_i, s_i)_{i=1}^T$ 。其中  $d_i$  为时间索引（通常为日期）， $s_i$  为对应日期的情绪强度值，且  $d_1 < d_2 < \dots < d_T$ 。

对于每个足够靠后的时间点  $t \geq 7$ ，定义一个长度为 7 的滚动窗口  $W_t$ ，包含当前及前 6 天的情绪强度： $W_t = \{s_{t-6}, s_{t-5}, s_{t-4}, s_{t-3}, s_{t-2}, s_{t-1}, s_t\}$ 。从向量表示角度，窗口也可写为  $W_t = [s_{t-6}, s_{t-5}, s_{t-4}, s_{t-3}, s_{t-2}, s_{t-1}, s_t]^T \in R^7$ 。窗口长度取 7 的主要考虑两点因素：一是 7 天的窗口能覆盖一周的情绪变化节奏，包含了周末信息与情绪的累积效应；二是样本量足够用于相关计算，又能保持对近期变化的敏感度。

#### 2. 一致性系数的计算过程

对每个窗口  $W_t$ ，将其中 7 个观测值依次记为  $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ ，其中  $x_i = s_{t-7+i}$ （ $i=1, 2, \dots, 7$ ）。在此基础上构造两个长度为 6 的序列：

原始子序列（ $X$ ）：

$$X = [X_1, X_2, X_3, X_4, X_5, X_6] = [x_1, x_2, x_3, x_4, x_5, x_6]；$$

滞后一期子序列（ $Y$ ）：

$$Y = [Y_1, Y_2, Y_3, Y_4, Y_5, Y_6] = [x_2, x_3, x_4, x_5, x_6, x_7]。$$

两者对应 6 组相邻日配对观测： $(X_k, Y_k) = (x_k, x_{k+1})$ ， $k=1, 2, \dots, 6$ ，每一组配对都表示相邻两日的情绪强度关系。

窗口  $W_t$  的一致性系数  $c_t$  定义为序列  $X$  与  $Y$  的样本皮尔逊相关系数：

$$c_t = r_{XY} = \frac{\sum_{k=1}^6 (X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^6 (X_k - \bar{X})^2} \cdot \sqrt{\sum_{k=1}^6 (Y_k - \bar{Y})^2}} \quad (3)$$

其中  $\bar{X} = \frac{1}{6} \sum_{k=1}^6 X_k$ ， $\bar{Y} = \frac{1}{6} \sum_{k=1}^6 Y_k$  分别为两个序列的样本均值。

这一过程本质上是在每个滚动窗口内刻画情绪强度的“一阶相关”程度，用来反映相邻日期情绪变化是否呈现稳定的线性

依赖。

### 3. 系数的统计解释

一致性系数  $C_t$  的取值范围为  $[-1, 1]$ ：

当  $C_t$  接近 +1 时，说明窗口  $W_t$  内情绪具有较强延续性：前一日偏高（偏低）时，后一日往往也偏高（偏低）。

当  $C_t$  接近 0 时，说明情绪变化缺乏稳定模式，更接近随机波动。

当  $C_t$  接近 -1 时，说明情绪更偏向反转：前一日偏高（偏低）往往对应后一日回落（回升）。

最终可得到新的时间序列  $C_{t=7}^T$ ，其中每个  $C_t$  对应日期  $d_t$  的情绪一致性度量，为分析情绪的短期传导、惯性与修正特征提供时间维度上的依据。

综上，情绪波动率和情绪一致性从“变化幅度”和“变化规律”两方面刻画投资者情绪。情绪波动率关注相邻时点情绪强度的相对变动大小，用来识别情绪剧烈波动与相对平稳的阶段，反映市场对信息冲击的即时反应程度。情绪一致性则基于滚动窗口内相邻日情绪的相关性，判断情绪是延续还是反转，从而体现情绪的惯性与自我修正特征

### （四）RFECV 方法介绍

RFECV（Recursive Feature Elimination with Cross-Validation）是一种嵌入式特征选择方法。其核心思路是先用基学习器训练模型，根据特征重要性逐步剔除贡献较小的特征；每删一轮，就用交叉验证评估当前特征子集的效果，最后选择验证表现最好的那一组特征及其数量<sup>[8]</sup>。

设训练数据集为  $D = (x_i, y_i)_{i=1}^N$ ，其中  $x_i \in R^p$  表示包含  $p$  个特征的输入变量， $y_i$  为目标变量。RFECV 在每次迭代中以预设的基学习器  $f(\cdot)$  训练模型，并利用交叉验证得分函数

$$S_k = \frac{1}{K} \sum_{j=1}^K s(f_{-j}, D_j), \quad (4)$$

表 1：情绪指标描述性分析

| 指标     | 样本数  | 均值     | 标准差    | 最小值     | 中位数    | 最大值     | 偏度    | 峰度     |
|--------|------|--------|--------|---------|--------|---------|-------|--------|
| 情绪强度   | 2023 | 0.400  | 0.061  | 0.198   | 0.401  | 0.633   | 0.086 | 0.349  |
| 情绪一致系数 | 2021 | -0.156 | 0.362  | -1.000  | -0.162 | 0.909   | 0.195 | -0.560 |
| 情绪波动率  | 2022 | 2.107  | 21.129 | -55.582 | -0.538 | 105.432 | 0.680 | 0.956  |

从表 1 可以看出：

情绪强度整体分布较为平稳。均值约为 0.400，说明样本期内情绪整体偏正向；标准差为 0.061，波动不大。偏度 0.086 接近 0，整体较为对称；峰度 0.349 略高于正态，意味着在少数时点可能出现相对“更极端”的情绪值，但不算突出。

情绪一致系数的均值为 -0.156，整体呈现轻微的“反转倾向”，即情绪在相邻时段并不总是延续。标准差 0.362，说明不同时期差异明显。偏度 0.195 为轻微右偏；峰度 -0.560 较为平坦，极端值的集中程度不高。取值范围从 -1.000 到 0.909，说明样本中既有延续阶段，也有反转阶段，情绪演化并不单一。

情绪波动率的波动最强。均值约 2.107，但标准差高达 21.129，说明该指标本身非常不稳定；偏度 0.680 右偏、峰度 0.956 呈尖峰厚尾，更符合金融数据中“平时平稳、关键时点跳

其中  $K$  为交叉验证折数， $s(\cdot)$  表示模型评价指标（如  $R^2$ 、MSE 等）。通过比较不同特征子集下的交叉验证得分，可以得到更合适的特征组合，从而在模型效果与特征维度之间取得平衡。

### （五）卷积神经网络介绍

卷积神经网络（CNN）是一类常用于处理网格状或具有局部结构数据的深度学习模型。其核心设计思想源于对生物视觉机制的模拟，通过局部连接、权重共享和空间下采样等机制，能够高效地从输入数据中逐层提取从低级到高级的抽象特征。

### （六）深度神经网络介绍

深度神经网络（DNN）是一种由多层结构组成的人工神经网络模型，其核心思想在于通过多层非线性映射与特征变换，从复杂的数据中自动提取出具有代表性的高层特征。

## 二、实证分析

### （一）情绪指标的构建

本研究基于 BosonNLP 情感词典，对多平台中文评论文本进行情绪量化，并构建日度情绪指标<sup>[9]</sup>。流程如下：

对收集到的非结构化评论文本进行清洗与预处理，使用 jieba 分词得到规范化词序列；

将分词结果与 BosonNLP 情感词典匹配，识别情感词并提取对应得分；对单条评论中命中的情感词得分取算术平均，作为该评论的情绪强度。若未识别到情感词，则情绪得分记为 0；

以日期为索引，对各平台、各关键词下的评论情绪得分进行汇总，通过计算日度平均值，得到反映当日整体情绪强度的综合指标。

在此基础上，进一步构建情绪一致系数与情绪波动率。然后对三类情绪指标做描述性统计与分析。

#### 1. 情绪指标描述性分析

动很大”的特征。最小值为 -55.582、最大值为 105.432，进一步体现极端波动时点的存在，往往对应重大事件冲击或情绪集中转向。

#### 2. 情绪指标时间序列动态特征

情绪强度序列存在明显的波动聚集现象，在部分阶段会持续走高或走低，说明市场情绪并非随机波动，而是带有一定的阶段性特征。尤其在重大市场事件前后，情绪强度往往出现更突出的峰值，反映投资者对关键信息的集中反应。

情绪一致性（7 天一致性系数）在样本期内起伏较大，且正负值频繁切换，说明情绪在“延续”与“反转”之间不断转换。滚动均值结果也支持这一点：有些阶段一致性保持为正，意味着情绪更容易延续，前一日情绪对后一日具有较明显的正向带动；而在另一些阶段，一致性下降甚至为负，更像是情绪出现了回调与

修正, 市场情绪的自我调整更明显。

情绪波动率(日变化率)的序列同样呈现典型的波动率聚集特征, 高波动期与低波动期交替出现。从30天滚动均值与滚动标准差来看, 波动率在部分阶段会快速抬升, 通常对应市场不确定性上升或信息冲击更频繁的时期; 滚动标准差区间较宽也说明该指标本身不稳定, 容易在特定时点被外部冲击拉大。

### 3. 情绪指标间相关性结构

为考察三类情绪指标之间的关系, 本文进一步进行了相关性分析。结果显示: 情绪强度与情绪波动率呈微弱负相关(约-0.15), 说明情绪达到极端时, 波动率未必同步处于最高水平, 可能存在“情绪冲击后短暂平静”的阶段。情绪强度与情绪一致性相关性较弱(约0.09), 表明情绪水平的高低与情绪是否延续并不存在明显对应关系。情绪波动率与情绪一致性呈负相关(约-0.28), 说明情绪变化越剧烈, 越难形成稳定的延续结构, 一致性更容易被打破。

## (二) RFECV 特征变量筛选

在黄金期货价格预测建模中, 为提升训练效率并减少无效特征带来的干扰, 本文在深度学习建模前引入基于支持向量回归(SVR)的递归特征消除交叉验证(RFECV)进行特征筛选<sup>[10]</sup>。

实验中选用线性核 SVR 作为基础学习器。该模型结构简单、结果更便于解释, 可为特征重要性评估提供相对稳定的依据。考虑到黄金期货数据具有明显的时间序列属性, 筛选过程中采用时间序列交叉验证而非随机划分, 以保证训练集与验证集在时间上保持前后顺序, 降低信息泄露风险。

目标变量设为“次日收盘价”, 其余变量作为候选输入特征参与筛选。通过 RFECV 的逐步迭代, 保留了一组核心特征用于后续建模。

根据筛选结果, 本文将“期货收盘价(连续): 黄金”“期货开盘价(连续): 黄金”“期货最高价(连续): 黄金”“期货最低价(连续): 黄金”“黄金储备: 中国”“现货价: 黄金”“RSI12”“SMA30”等特征变量, 与构造的三个情绪指标一并作为模型输入, 用于后续预测实验。

总体来看, RFECV 在本文中主要起到两方面作用: 一是减少输入维度, 提高 DNN 与 CNN- 线性回归混合模型的训练效率与收敛速度; 二是降低冗余与噪声特征的影响, 缓解过拟合, 使模型在金融时序数据下的表现更稳定。

## (三) 模型实验阶段

在黄金期货价格预测的实证研究中, 本文构建并比较两类模型: CNN- 线性回归混合模型与深度神经网络(DNN), 用于评估不同结构在预测任务中的表现差异。

### 1. CNN+ 线性回归混合模型

该模型将局部特征提取与回归预测结合起来: 首先通过一维卷积层(Conv1d)提取局部模式, 设置64个滤波器、卷积核大小为3, 并采用 padding=1 保持维度; 卷积层后接 ReLU 激活, 再通过最大池化层(MaxPool1d)降维(池化核大小为2)。为抑制过拟合, 在卷积模块后加入 Dropout(丢弃率0.3)。卷积

输出的特征图展平后输入全连接网络, 全连接部分包含三层线性层, 隐藏层维度为100与50, 层间使用 ReLU 与 Dropout(丢弃率0.2), 最终输出单值预测。训练时采用 MSE 损失与 Adam 优化器, 学习率0.001, 并加入 L2正则化(权重衰减 $1e-4$ )与梯度裁剪(最大范数1.0)提高训练稳定性; 学习率调度使用 ReduceLROnPlateau, 当验证损失进入平台期超过20轮时学习率减半。训练300轮, batch size 为32。

### 2. 深度神经网络(DNN)模型

DNN 采用四个隐藏层, 维度依次为256、128、64、32, 形成逐层递减的结构。每个隐藏层后加入批归一化以加速训练并提升泛化能力, 激活函数使用 LeakyReLU(负斜率0.1)以降低“神经元失活”风险; 各层设置 Dropout(丢弃率0.2)。权重使用 Kaiming 正态初始化以匹配 LeakyReLU。损失函数采用 SmoothL1Loss, 提高对异常值的鲁棒性; 优化器为 AdamW, 学习率0.001, 权重衰减 $1e-3$ , 并结合余弦退火热身重启调整学习率( $T_0=50$ ,  $T_{mult}=2$ )。训练时设置早停: 连续30轮无改善则提前结束, 最大训练轮数为500。

### 3. 测试集表现与区间差异

从测试集拟合效果看, 两种模型在不同价格区间的表现存在差异。DNN 在480-550区间拟合较好, 预测曲线与真实价格更接近; CNN+ 线性回归模型在520-580的中位价格区间表现相对稳定, 卷积层提取的局部特征有助于识别短期价格形态。

但在价格极端区域, 两者都出现明显偏差: 当价格突破650后, 预测值普遍低于实际值约5%-8%, 形成持续的负向偏差, 不过仍能较好跟随价格波动形态。一个可能的原因是训练数据在高价区样本较少, 模型在学习绝对价格水平时更偏保守, 导致高价阶段系统性低估。

## (四) 模型对比

为比较 CNN+ 线性回归混合模型与深度神经网络(DNN)在黄金期货预测任务中的效果, 本文采用 MAE、RMSE、 $R^2$  与 MAPE 四项指标进行评估, 结果见表2。

表2: 两模型指标对比

| 评估指标     | CNN+ 线性回归 | 深度神经网络 |
|----------|-----------|--------|
| MAE      | 20.913    | 17.425 |
| RMSE     | 24.075    | 21.07  |
| $R^2$    | 0.941     | 0.954  |
| MAPE (%) | 0.954     | 2.688  |

从结果看, 两种模型侧重点不同。DNN 在 MAE、RMSE 和  $R^2$  上更优, 说明其对绝对误差的控制更好, 整体拟合能力也更强; CNN+ 线性回归模型在 MAPE 上更低, 说明其在相对误差控制方面更占优势。也就是说, 如果更关注“误差的绝对大小”, DNN 更合适; 如果更关注“相对误差比例”, CNN+ 线性回归模型表现更稳。

## (五) 情绪指标贡献

为检验情绪信息的增量价值, 本文对比了“加入情绪指标”和“未加入情绪指标”两种情况下的模型表现, 结果如表3所示。



| 表 3：两模型有无情绪指标模型指标对比 |           |        |        |
|---------------------|-----------|--------|--------|
| 评估指标                | 模型类型      | 有情绪指标  | 无情绪指标  |
| MAE                 | CNN+ 线性回归 | 15.011 | 20.913 |
|                     | 深度神经网络    | 17.425 | 29.536 |
| RMSE                | CNN+ 线性回归 | 24.075 | 26.344 |
|                     | 深度神经网络    | 21.07  | 37.616 |
| $R^2$               | CNN+ 线性回归 | 0.941  | 0.93   |
|                     | 深度神经网络    | 0.954  | 0.857  |
| MAPE                | CNN+ 线性回归 | 0.954  | 2.46   |
|                     | 深度神经网络    | 2.688  | 4.729  |

先看总体趋势：对 DNN 来说，引入情绪指标后 MAE 与 RMSE 明显降低， $R^2$  提升，MAPE 也下降，说明情绪信息确实增强了模型对价格波动的刻画能力，预测更贴近真实走势。对 CNN+ 线性回归模型而言，RMSE、 $R^2$  与 MAPE 的变化同样显示出情绪指标带来的改善。

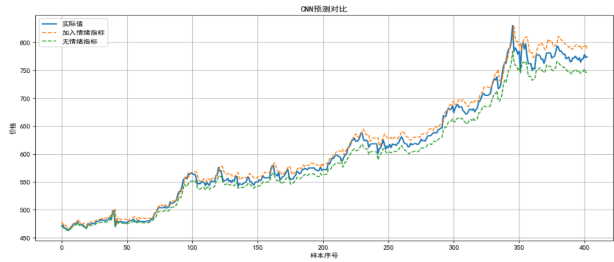


图 1：CNN+ 线性模型有无情绪指标预测结果对比

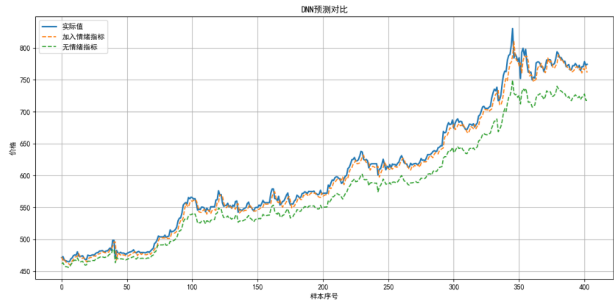


图 2：DNN 有无情绪指标预测结果对比

由图 1 和图 2 可知，情绪指标对模型预测的改进不仅体现在数值精度上，还表现在对价格波动形态的刻画上，在加入情绪特征后，模型对局部起伏、拐点变化的刻画更贴近真实走势，证明了情绪因素作为价格先行指标的有效性。

综上，情绪指标的引入能够降低预测误差、提高了模型拟合优度，还增强了模型对不同市场环境的适应能力，显著提升了 DNN 和 CNN 两类模型在黄金期货价格预测中的综合性。这也意味着，将情绪信息作为补充特征，有助于提高模型输出在实际量化分析与策略辅助中的可用性与稳定性。

### 三、结束语

本文提出了一个融合多源舆情情绪特征与市场结构化变量的黄金期货价格预测框架，能够将行为金融中的“情绪冲击”转化为可计算信号，并提升预测的稳定性与泛化性。首先，通过情绪强度、情绪一致系数与情绪波动率刻画市场预期变化，其中情绪波动率标准差达 21.129、区间为 [-55.582, 105.432]，对突发波动更敏感。其次，采用以 SVR 为基学习器的 RFECV 进行特征筛选，减少冗余与噪声，使输入更精炼、迁移更方便。最后，在统一管线下对比深度模型，DNN 取得 MAE=17.425、RMSE=21.07、( $R^2=0.954$ )，显示出更好的整体拟合与波动刻画能力。未来研究可进一步融合事件级文本与跨市场信息以增强鲁棒性。

### 参考文献

[1] 范彩云, 童君逸, 程俊彦, 等. 基于 ML-DMA 的黄金期货价格预测研究 [J]. 数理统计与管理, 2024, 43(03): 541-558. DOI: 10.13860/j.cnki.sltj.20240509-009.

[2] 杨文清. 基于文本挖掘的投资者情绪对股票收益率的影响 [D]. 哈尔滨工业大学, 2021. DOI: 10.27061/d.cnki.ghgdu.2021.001613.

[3] 孙晶. 基于 RFECV 特征选择和随机森林预测模型的应用与优化 [J]. 数字通信世界, 2024, (09): 114-116.

[4] 陈湘源, 秦伟, 刘晏驰, 等. 融合卷积神经网络与线性回归的带式输送机托辊故障音频识别方法 [J]. 煤炭科学技术, 2025, 53(S1): 389-398.

[5] 黄奇. 基于多模态信息融合的情感分析机制研究与实现 [D]. 南京邮电大学, 2022. DOI: 10.27251/d.cnki.gnjdc.2022.001457.

[6] 吕成双, 王彤, 孙浩然. 融合情绪指标的股价波动率预测研究 - 基于微调大语言模型与 GAT-TCN 网络 [J/OL]. 运筹与管理, 1-8[2025-12-09]. <https://link.cnki.net/urlid/34.1133.G3.20250106.1142.002>.

[7] 朱晓彤. 基于 Transformer 模型的金融时间序列特征分析与预测方法研究 [D]. 山东财经大学, 2025. DOI: 10.27274/d.cnki.gsdjc.2025.000832.

[8] 陈亦真, 宋旭峰, 李超顺. 基于 RFECV-RF 和 PCA-LightGBM 的水电机组劣化趋势评估 [J]. 水电能源科学, 2025, 43(03): 196-199+172. DOI: 10.20040/j.cnki.1000-7709.2025.20240361.

[9] 肖璐, 陈果, 刘继云. 基于情感分析的企业产品级竞争对手识别研究 - 以用户评论为数据源 [J]. 图书情报工作, 2016, 60(01): 83-90+97. DOI: 10.13266/j.issn.0252-3116.2016.01.012.

[10] 郭佳伦, 钟浩琨, 赵俊博, 等. 基于支持向量回归 (SVR) 的马尾松木材脱脂率预测 [J]. 北京林业大学学报, 2025, 47(03): 151-161.

# 基于 TREC 真实邮件数据集的朴素贝叶斯分类教学 案例构建与应用效果实证研究

曹寒问, 陈锦文, 车金星, 张毓华  
江西水利电力大学 理学院, 江西 南昌 330099  
DOI:10.61369/ASDS.2026010007

**摘 要 :** 人工智能技术的迅猛发展对概率统计学科教学提出了前所未有的革新要求。当前传统课程面临理论与应用脱节、教学案例缺乏应用、课堂缺乏实践, 难以培养学生将概率模型转化为解决实际问题的能力。针对这一挑战, 本研究基于 TREC Public Corpus 包含的 75,419 封真实邮件数据集, 系统构建朴素贝叶斯分类教学案例。借助 Python 工具链实现邮件解析、文本预处理、特征工程至概率决策的全流程教学转化, 使学生能够动态修改邮件内容并实时观测朴素贝叶斯后验概率变化。结合人工智能技术, 学生得以了解理论知识的具体实践应用场景。

实证研究结果表明: 模型在 22,607 封测试邮件中整体准确率较高, 其垃圾邮件识别和正常邮件的识别精确度均表现优异。特征重要性分析揭示“pill”在垃圾邮件中出现概率显著高于正常邮件, 而“per”和“desjardin”等商业词汇构成关键判别模式。教学实验中, 学生通过添加“meeting”等工作词汇, 成功降低测试邮件的垃圾概率, 直观验证先验分布与似然概率的协同决策机制。并通过对 223 名学生分为案例教学组和传统教学组开展对比研究, 独立样本 t 检验结果显示, 案例教学组的期末成绩显著优于传统教学组, 两组差异达到统计显著水平, 平均成绩提升 5.30 分, 及格率提高 16.8 个百分点。

该案例将条件概率、全概率公式等抽象理论转化为可操作的实践载体, 显著提升学生构建概率模型解决复杂问题的能力, 突破传统教学中公式记忆和机械演算的认知局限, 实现理论向应用层面的跃迁。

**关 键 词 :** 概率统计; 朴素贝叶斯; 条件概率; 理论实践融合

## Empirical Study on the Construction and Application Effectiveness of a Naive Bayes Classification Teaching Case Based on the TREC Authentic Email Dataset

Cao Hanwen, Chen Jinwen, Che Jinxing, Zhang Yuhua  
School of Science, Jiangxi University of Water Resources and Electric Power, Nanchang, Jiangxi 330099

**Abstract :** The rapid advancement of artificial intelligence technology has imposed unprecedented demands for innovation in probability and statistics education. Current traditional curricula face challenges such as disconnect between theory and application, lack of practical teaching cases, and insufficient hands-on classroom activities, making it difficult to cultivate students' ability to translate probability models into solutions for real-world problems. To address this challenge, this study systematically constructs a Naive Bayes classification teaching case based on the TREC Public Corpus dataset comprising 75,419 authentic emails. Utilizing Python toolchains, the entire teaching process—from email parsing and text preprocessing to feature engineering and probabilistic decision-making is fully implemented. This enables students to dynamically modify email content and observe real-time changes in the Naive Bayes posterior probability. By integrating artificial intelligence technology, students gain insight into the concrete practical applications of theoretical knowledge.

Empirical research findings indicate that the model demonstrated high overall accuracy across 22,607 test emails, with outstanding precision in identifying both spam and legitimate messages. Feature importance analysis revealed that “pill” appears significantly more frequently in spam emails than in legitimate ones, while commercial terms like “per” and “desjardin” form key discriminative patterns. In the teaching experiment, students successfully reduced spam probability by adding work-related vocabulary like “meeting,” intuitively validating the collaborative decision-making mechanism

基金项目: 江西省教育厅高等学校教学改革研究省级重点课题 (AI 赋能《概率论与数理统计》的个性化教学方式探索, 编号: JXJG-24-18-5)。

作者简介:

曹寒问, 江西水利电力大学理学院, 硕士, 副教授, 研究方向: 数据分析, 22099166@qq.com;

陈锦文, 江西水利电力大学理学院, 本科生, 专业: 应用统计学, 1283671376@qq.com;

车金星, 江西水利电力大学理学院, 博士, 教授, 研究方向: 人工智能与水电能源统计, jinxingche@163.com;

张毓华, 江西水利电力大学理学院, 博士, 副教授, 研究方向: 能源经济统计分析, 1984zhangyuhua@163.com。

between prior distribution and likelihood probability. A comparative study involving 223 students divided into a case-based teaching group and a traditional teaching group was conducted. Independent samples t-test results showed that the case-based teaching group achieved significantly higher final scores than the traditional teaching group, with the difference reaching statistical significance. The average score improved by 5.30 points, and the pass rate increased by 16.8 percentage points.

This case study transforms abstract theories such as conditional probability and the law of total probability into practical tools, significantly enhancing students' ability to construct probability models for solving complex problems. It breaks through the cognitive limitations of traditional teaching methods focused on formula memorization and mechanical calculations, achieving a leap from theory to application.

**Keywords :** probability and statistics; naive bayes; conditional probability; theory-practice integration

## 引言

全球 AI 融合应用正加速推进社会变革，据市场分析预测，2025 年人工智能市场规模将突破 4.8 万亿美元，技术渗透率预计超过 68%。在此背景下，概率统计学科需应对技术发展带来的教学革新需求。根据教育部《高等学校人工智能创新行动计划》的战略部署，基础数学课程与人工智能技术的深度整合已成为高等教育改革的关键路径。尽管产业界对数据分析人才的需求呈现指数级增长，现行概率统计课程却暴露出教学内容与技术前沿的严重脱节问题——这种脱节既反映在教学案例的时效性不足，更体现在数学理论与算法实现之间的认知壁垒，最终制约学生将数学概念转化为实际问题解决能力的发展<sup>[1]</sup>。

概率统计课程教学案例体系目前存在三个问题：内容更新滞后、理论实践断层和真实训练缺失。教育部大学数学课程群虚拟教研室 2025 年调研数据显示，全国 72.3% 高校的概率统计课程仍在沿用工业化时期的经典案例，这些案例已难以匹配现代数据分析需求；在贝叶斯公式教学中，普遍采用的简化版疾病检测模型既弱化了先验概率设定的复杂性，又模糊了似然函数估计的技术难点，致使真实决策过程无法完整呈现；更值得关注的是，2023 年全国 12 所高校课程评估报告表明，仅有 29.6% 的课程作业涉及真实数据集分析，这种设计缺陷直接导致教学陷入“公式记忆-机械演算”的恶性循环。显然，现有教学范式既无法有效培养学生基于真实数据构建概率模型的能力，也难以满足人工智能时代对创新人才核心素养的培养要求，迫切需要开发整合真实数据集与现代技术场景的新型教学载体<sup>[2]</sup>。

本研究基于 TREC Public Corpus 真实邮件数据集开发朴素贝叶斯分类教学案例，该数据集包含 75, 419 封真实邮件，案例设计以传统统计学中的条件概率理论为根基，将朴素贝叶斯算法转化为可授课的教学载体，通过结构化解析邮件文本内容、结合文本清洗技术与词干归并处理，构建适合课堂演示的特征工程流程。教学方法重点阐释词频统计与传统概率估计的内在联系，引导学生掌握机器学习算法对经典统计方法的拓展逻辑。研究借助 Python 教学工具演示贝叶斯决策过程，建立传统统计理论与智能算法实践的教学桥梁。

## 一、朴素贝叶斯分类方法与数学原理

### （一）垃圾邮件分类问题描述

文本分类作为机器学习领域的经典任务，核心目标在于将文档自动归类至预定义类别体系。伴随数字化信息交流的蓬勃发展，电子邮件已成为关键沟通载体，其重要性不言而喻。但垃圾邮件的泛滥严重破坏着这一渠道的高效与安全，引发信息过载、隐私泄露及网络诈骗等多重风险。为应对此挑战，机器学习技术被引入电子邮件过滤领域，其中朴素贝叶斯分类算法凭借简洁高效的优势脱颖而出。该算法以贝叶斯定理为理论基础，通过数据集统计规律学习实现分类决策。“朴素”命名源于特征独立性假设，此假设显著降低计算复杂度，并为复杂分类问题构建基础框架。在垃圾邮件过滤场景中，基于邮件文本特征的深度分析，该

算法能够精准识别垃圾邮件与正常邮件，大幅提升用户信息管理的效率<sup>[3]</sup>。

### （二）朴素贝叶斯算法原理

朴素贝叶斯分类算法的理论根基深植于概率论体系，其核心机制在于最大化后验概率<sup>[4]</sup>。朴素贝叶斯分类器的构建旨在精准计算每个类别  $C_i$  在给定特征向量  $X$  下的后验概率  $P(C_i|X)$ ，依据贝叶斯定理，该概率可严谨推导为：

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

其中， $P(C_i)$  是通过训练数据集中类别  $C_i$  样本出现的相对频率估计得出的先验概率； $P(X|C_i)$  是在类别  $C_i$  确定的条件下，特征向量  $X$  出现的条件概率，即似然概率； $P(X)$  则是特征向量  $X$  在整

个样本空间中出现的边缘概率。在实际运算中，鉴于 $P(X)$ 对于所有类别而言保持恒定，因此在比较不同类别对应的后验概率时，可将其视为常数项予以简化处理。朴素贝叶斯算法通过精确计算每个类别的后验概率，并依据最大后验概率准则确定样本类别，从而实现高效精准的分类决策。

多项式朴素贝叶斯垃圾邮件分类器的工作流程可以清晰地总结为图1。

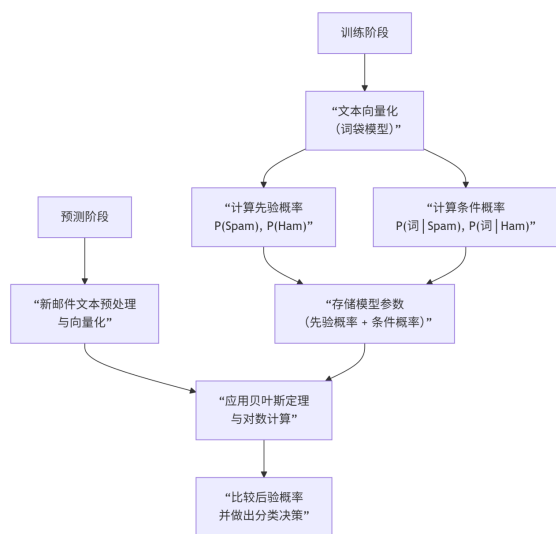


图1：贝叶斯垃圾邮件分类器工作流程图

Figure 1: Workflow Diagram of the Bayesian Spam Classifier

## 二、教学案例设计

### (一) 数据准备及预处理

研究采用 TREC Public Corpus 提供的 75,419 封真实邮件作为数据集。邮件解析过程借助 Python 邮件模块处理原始文件<sup>[5]</sup>，有效应对多部分邮件结构与多样编码格式的解析需求。数据预处理涵盖三个核心环节：文本清洗环节移除 HTML 标签、URL 链接、邮箱地址及非字母字符；停用词过滤环节基于 NLTK 英语停用词表剔除“the”“and”等无意义词汇<sup>[6]</sup>；词干提取环节采用 PorterStemmer 算法将词形变体归并为统一词根。标签数据通过独立索引文件加载实现邮箱精确标记，其中垃圾邮件占总量的 66.5%，一共 50,126 封。据此，我们通过带入先验概率计算公式计算各类别邮件在总体样本中的频率，得到了朴素贝叶斯分类器所需的先验概率：

垃圾邮件 (Spam) 的先验概率  $P(\text{Spam}) = 50,126 / 75,419 \approx 66.5\%$ 。

正常邮件 (Ham) 的先验概率  $P(\text{Ham}) = 25,293 / 75,419 \approx 33.5\%$ 。

### (二) 模型训练及结果展示

模型训练基于多项式朴素贝叶斯分类器，核心任务聚焦两类关键概率参数估计。先验概率计算获得： $P(\text{Spam})=66.5\%$ ，

$P(\text{Ham})=33.5\%$ ，该结果准确反映数据集真实分布。条件概率参数基于特征词频统计进行估计。特征词概率分布呈现显著差异，作出表1，其数据显示：商业推广词汇“pill”在垃圾邮件中的概率达0.015，“price”在垃圾邮件中概率为0.0089；相比之下，工作相关词汇“offic”在正常邮件中出现概率更高。特征重要性分析进一步揭示“anatrium”、“cialis”等词汇具有显著判别力，此类差异构成分类决策的数学基础<sup>[7]</sup>。

表1：最重要的垃圾邮件特征词  
Table 1: Most Significant Spam Feature Words

| 排名 | 特征词       | 垃圾概率     | 正常概率     | 重要性      |
|----|-----------|----------|----------|----------|
| 1  | pill      | 0.015147 | 0.000135 | 0.015012 |
| 2  | per       | 0.01213  | 0.000636 | 0.011494 |
| 3  | desjardin | 0.010682 | 0.000001 | 0.010681 |
| 4  | price     | 0.00892  | 0.001744 | 0.007176 |
| 5  | item      | 0.007833 | 0.000746 | 0.007087 |
| 6  | save      | 0.007351 | 0.000803 | 0.006547 |
| 7  | product   | 0.006641 | 0.001262 | 0.005379 |
| 8  | votr      | 0.005221 | 0.000003 | 0.005219 |
| 9  | viagra    | 0.004886 | 0.000014 | 0.004872 |
| 10 | onlin     | 0.005584 | 0.000798 | 0.004786 |
| 11 | transact  | 0.004852 | 0.000176 | 0.004676 |
| 12 | vou       | 0.004579 | 0.000001 | 0.004569 |
| 13 | anatrium  | 0.004123 | 0.000001 | 0.004123 |
| 14 | retail    | 0.003755 | 0.000117 | 0.003638 |
| 15 | cialis    | 0.003478 | 0.000005 | 0.003473 |
| 16 | men       | 0.003541 | 0.000169 | 0.003372 |
| 17 | buy       | 0.003652 | 0.000409 | 0.003242 |
| 18 | money     | 0.003602 | 0.000429 | 0.003173 |
| 19 | qualiti   | 0.003251 | 0.00025  | 0.003001 |
| 20 | adob      | 0.002984 | 0.00007  | 0.002914 |

对邮件中的关键词特征重要性和特征概率对比作图2。



图2：TOP 垃圾邮件特征词分析

Figure 2 : Analysis of TOP Spam Keywords

图2特征重要性分析结果呈现词汇判别能力分布规律，前三关键特征词——“pill”（药物类）、“per”（促销类）、“desjardin”（品牌类）的重要性值均超过0.01。这些词汇在垃圾邮件中的出现概率较正常邮件高出数十倍至百倍。商业推广词汇明显构成垃圾邮件的核心标识，而工作相关词汇则与正常邮件显著关联。随后



对高频词的频率做统计并且可视化作出图3。

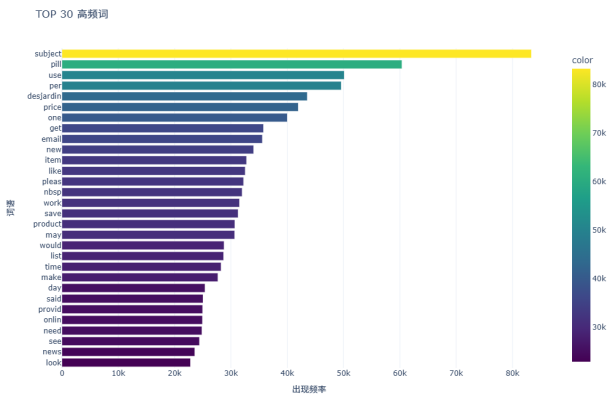


图3：邮件内容 TOP30 高频词分布

Figure 3: Distribution of the TOP 30 High-Frequency Words in Email Content

高频词统计图3呈现邮件文本的基础构成特征。主题词“subject”以超8万次的绝对频次居首位，体现邮件基础结构的普遍性。需关注的是，高判别力特征词“pill”同时位列高频榜前五，其出现频率达常规词汇“news”的7倍。该双重特征印证特征选择策略的有效性——高频商业词汇天然携带强判别力，应作为过滤规则优化的核心目标。

课堂交互演示环节，输入测试邮件后模型以100%置信度判定其为垃圾邮件。决策分析表明“offer”、“discount”等词汇引发高条件概率响应，此类词汇在特征重要性排名中均居前20位，直观揭示特征词对分类决策的作用机制。

测试阶段采用22,607封邮件构成测试集，正常邮件7,566封，垃圾邮件15,041封。作出混淆矩阵如图4显示：正常邮件正确识别率达83.5%，7,515封正确分类，51封误判为垃圾邮件；垃圾邮件成功拦截13,554封，漏判1,487封。垃圾邮件识别存在显著漏判现象，此特性表明模型优先保障正常邮件的准确传递（符合“避免误拦重要邮件”的实际需求），同时需重点识别垃圾邮件的隐蔽变体。

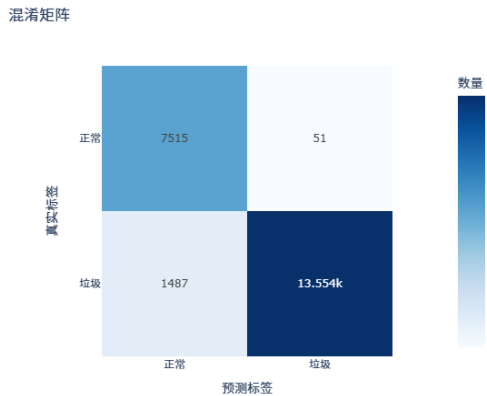


图4：朴素贝叶斯分类器混淆矩阵

Figure 4: Confusion Matrix of the Naive Bayes Classifier

整理出模型的关键性能指标作出表2。

表2：分类性能指标

| Table 2 Classification Performance Metrics |       |       |       |         |
|--|-------|-------|-------|---------|
| 类别   | 精确率   | 召回率   | F1 分数 | 样本量     |
| 正常邮件                                       | 83.5% | 99.3% | 0.91  | 7, 566  |
| 垃圾邮件                                       | 99.6% | 90.1% | 0.95  | 15, 041 |

垃圾邮件识别精确率达99.6%，召回率为90.1%；正常邮件召回率高达99.3%，但精确率仅83.5%。该差异反映模型对垃圾邮件的判定策略较为保守，而对正常邮件的识别相对宽松，分类结果体现朴素贝叶斯模型在垃圾邮件过滤中的特性：垃圾邮件判定精确度极高达到99.6%，但正常邮件识别精确率为83.5%。这种非对称性能源于实际需求平衡——系统选择允许部分垃圾邮件进入收件箱，以近乎100%的保障重要邮件送达。此设计契合邮件服务的核心原则：优先确保通信可靠性，其次提升过滤效率。

### 三、教学效果实证分析

#### （一）实验设计与研究方法

本研究选取2022，2023，2024三个连续学年度的应用统计学专业教学班级作为研究对象，所有班级均由同一教师授课，统一学习概率论课程，确保教学条件的一致性。教学改革组包含2023级（传统教学结合基础练习，n=85）与2024级（案例教学法，n=59）两个班级，共计144名学生；传统教学组为2022级学生群体（n=79），采用纯理论讲授模式。在教学干预实施过程中，教学改革组强调将概率统计理论知识与真实世界问题解决相结合，通过朴素贝叶斯分类器在垃圾邮件识别中的具体应用，将条件概率、贝叶斯公式等抽象数学概念转化为可操作的实践工具；而传统教学组则维持以公式推导与例题讲解为主的理论传授模式<sup>[8]</sup>。

研究采用独立样本t检验探究教学模式（教学改革 vs 传统教学）对期末考试成绩影响的统计学显著性，检验的原假设设定为两种教学模式下的学生平均成绩无显著差异。以评估教学改革的实际成效。

#### （二）实验结果分析

对数据做描述性统计结果作出表3，教学改革组学生的平均成绩为63.69分，传统教学组学生的平均成绩为58.39，两组间存在5.30分的均值差异。

表3：不同教学模式学生成绩描述性统计对比  
Table 3: Descriptive Statistics Comparison of Student Performance Across Different Teaching Modes

| 组别    | 样本量 | 均值    | 标准差   | 最小值  | 25% 分位 | 中位数  | 75% 分位 | 最大值  |
|-------|-----|-------|-------|------|--------|------|--------|------|
| 传统教学组 | 79  | 58.39 | 16.66 | 26.0 | 47.0   | 60.0 | 70.0   | 98.0 |
| 教学改革组 | 144 | 63.69 | 16.11 | 14.0 | 51.0   | 64.0 | 75.0   | 93.0 |

为了直观展示实验分析过程，对数据进行可视化作出图5。

图5两组学生成绩分布比较直观展示了教学改革组与传统教学组的成绩分布特征。从箱线图可见，教学改革组的中位数位置更高，四分位距分布更为集中；小提琴图则进一步揭示了教学改革组在中等分数区间的密度更高，表明该教学方法对中等水平学生的提升效果尤为明显。均值比较与及格率分析结果清晰显示，教

学改革组在平均成绩与及格率两个维度均显著优于传统教学组。均值比较图中的置信区间表明两组差异的估计精度，而及格率比较则凸显了教学改革在降低不及格率方面的实际成效，这一发现具有重要的教学实践意义。

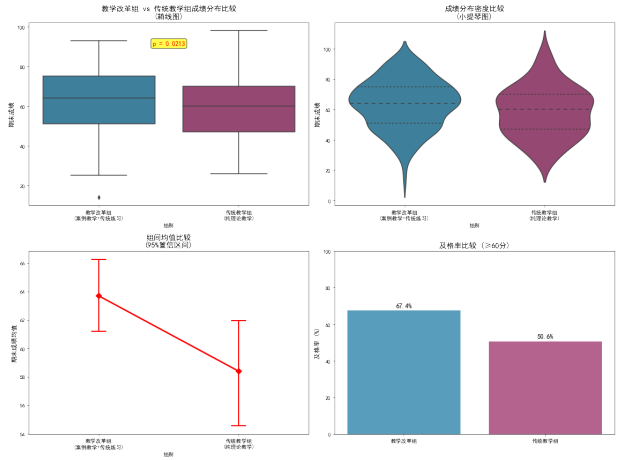


图5：教学效果检验可视化

Figure 5: Visualization of Teaching Effectiveness Evaluation

为验证组间差异的统计学显著性，研究进行了独立样本 t 检验。分析结果整理成表 4，发现教学改革组与传统教学组的成绩差异达到统计显著水平  $t=2.319$ ，因此我们拒绝原假设 ( $H_0$ )，接受备择假设，即教学改革组与传统教学组的学生平均成绩存在显著差异。且  $p=0.021<0.05$  结果显著<sup>[9]</sup>。

表 4: 教学改革效果推断统计检验结果

Table 4: Statistical Test Results for Inferring the Effects of Teaching Reform

| 指标项 | t 值   | 自由度 | p 值    | 均值差异 | 95% CI      | Cohen' s d |
|-----|-------|-----|--------|------|-------------|------------|
| 数值  | 2.319 | 221 | 0.0213 | 5.30 | 0.78 - 9.81 | 0.325      |

## 四、结论与展望

本研究基于 TREC Public Corpus 真实邮件数据集构建融合概率统计理论与机器学习实践的朴素贝叶斯分类教学案例。通过对 75,419 封邮件的系统化文本特征解析，完成从条件概率理论至贝叶斯决策算法的教学创新，证实传统概率论知识在现代人工智能应用中的核心价值。案例教学表明：

理论实践融合：贝叶斯公式、全概率公式等核心概念通过邮件分类任务具象化，学生能直观理解条件概率估计与先验分布的协同决策机制<sup>[10]</sup>；

算法可解释性：特征重要性分析，将抽象的概率计算转化为可观测的语言特征规律；

教学有效性：测试集高度的精确率证明，基于真实数据集的案例设计显著提升学生解决复杂问题的能力，突破传统教学中只有公式记忆，机械演算的认知局限。将理论层面提高到了应用层面<sup>[11]</sup>。

通过持续迭代教学案例库与实验工具链，推动概率统计课程实现从理论传授到”数据驱动－算法实现－决策优化”的能力培养范式转型，为人工智能时代培养兼具数学素养与工程能力的创新人才<sup>[12]</sup>。

## 参考文献

[1] 肖睿, 王峰, 黄文彬. 人工智能赋能教育的发展态势与未来路径 [J]. 现代教育技术, 2023, 33(1): 12-20.

[2] 陈希儒, 刘乐平. 新时代统计学教育改革的方向与路径 [J]. 统计研究, 2022, 39(4): 145-156.

[3] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2021: 123-135.

[4] 李航. 统计学习方法 [M]. 第 2 版. 北京: 清华大学出版社, 2021: 58-72.

[5] 张良均, 王靖, 刘名军. Python 数据挖掘与机器学习实战 [M]. 北京: 人民邮电出版社, 2022: 89-105.

[6] 宗成庆. 统计自然语言处理 [M]. 第 2 版. 北京: 清华大学出版社, 2021: 67-82.

[7] 黄文彬, 徐健. 基于特征重要性的文本分类模型可解释性研究 [J]. 计算机研究与发展, 2023, 60(3): 567-578.

[8] 温忠麟, 刘红云. 教育实证研究中的统计分析方法 [M]. 北京: 北京师范大学出版社, 2020: 156-170.

[9] 张厚粲, 徐建平. 现代心理与教育统计学 [M]. 第 5 版. 北京: 北京师范大学出版社, 2021: 245-260.

[10] 王陆, 刘菁. 人工智能时代案例教学法的创新路径研究 [J]. 电化教育研究, 2023, 44(2): 78-85.

[11] 任友群, 李锋. 面向人工智能时代的中小学计算思维培养 [J]. 中国电化教育, 2022(5): 1-8.

[12] 祝智庭, 魏非. 教育数字化转型的现实路径与发展趋势 [J]. 华东师范大学学报 (教育科学版), 2023, 41(1): 1-15.

# 线性秘密共享中的自对偶结构

林群

韩山师范学院数学与统计学院, 广东 潮州 521041

DOI:10.61369/ASDS.2026010008

**摘 要 :** 线性秘密共享方案 (LSSS) 是现代密码学中支撑安全多方计算与密码协议的关键基础。本文旨在系统性地构建一个基于线性码的 LSSS 理论框架。首先, 形式化线性秘密共享方案, 阐述了份额生成与秘密重构的算法流程。其次, 阐明了线性码的理论基础, 明确了生成矩阵与校验矩阵的核心作用。本文的贡献在于探讨了自对偶码的数学性质, 并通过一个具体的二元域实例加以验证。自对偶码因其内在的对称性和优美的结构, 为构建高效安全的线性秘密共享方案提供了理论工具, 在信息安全及相关领域中具有重要的应用价值。

**关 键 词 :** 线性秘密共享; 访问结构; 生成矩阵; 自对偶码

## On the Self-Dual Structure in Linear Secret Sharing

Lin Qun

Institute of Mathematics and Statistics, Hanshan Normal University, Chaozhou, Guangdong 521041

**Abstract :** Linear secret sharing schemes (LSSS) are a fundamental cornerstone of modern cryptography, underpinning secure multi-party computation and cryptography protocols. This paper aims to systematically construct a theoretical framework for LSSS based on linear codes. First, we formalize LSSS and elaborates on the algorithmic processes for share generation and secret reconstruction. Second, we clarifies the theoretical foundations of linear codes, emphasizing the central role of generator and parity-check matrices. The contribution of this work lies in exploring the mathematical properties of self-dual codes and validating them through a concrete example over the binary field. Owing to their inherent symmetry and elegant structure, self-dual codes provide a theoretical tool for constructing efficient and secure LSSS, holding significant application value in information security and the related fields.

**Keywords :** linear secret sharing; access structure; generator matrix; self-dual code

## 引言

完美保密的秘密共享是现代密码学的核心技术之一, 为信息的安全分布式存储与计算提供了根本保障。其核心思想是将秘密值分解为多个份额并分发给一组参与者, 确保只有经过授权的参与者集合能够联合恢复该秘密, 而任何未授权的集合则无法获取关于秘密的任何信息<sup>[1-5]</sup>。在众多秘密共享方案中, 线性秘密共享方案因其线性同态特性备受关注, 它最初由 Jackson 与 Martin 提出<sup>[6]</sup>。该类方案中, 份额的计算与授权集合中秘密的恢复仅依赖于线性映射和线性方程组的求解<sup>[7-9]</sup>, 因此具有计算高效的特点。此外, 其同态特性使得该方案能够支持安全多方计算、门限密码等高级密码协议<sup>[10-12]</sup>, 从而在隐私计算和数据安全领域发挥着至关重要的作用。

线性秘密共享方案的构造由生成矩阵与访问结构共同定义。生成矩阵不仅通过线性变换将秘密与随机数映射为各参与者的份额, 还内在刻画了方案的访问结构, 即明确哪些参与者子集能够重构秘密<sup>[13]</sup>。这种将访问结构嵌入线性代数框架的表示方式, 深刻揭示了秘密共享与编码理论之间的内在联系<sup>[14, 15]</sup>。具体而言, 线性秘密共享方案可视为一类特殊的线性码, 其中有效的码字对应由合法秘密生成的合法份额集合。

在编码理论中, 自对偶码因其独特的对称性与优美的数学结构而受到广泛关注。自对偶码要求线性码与其对偶码完全重合, 这一强约束条件使得其生成矩阵与校验矩阵具有同一性, 并衍生出一系列非平凡性质。这种内在的高度对称性, 使自对偶码成为构建具备高效性与安全性的秘密共享自对偶结构的理想工具。

本文旨在系统阐述基于线性码 (尤其是自对偶码) 的线性秘密共享方案框架。首先, 将介绍线性秘密共享方案的形式化定义、份额生成与秘密重构算法, 并深入分析其线性性质与同态特性。随后, 阐述线性码的基本理论, 重点说明生成矩阵与校验矩阵在描述线性码及其对偶关系中的核心作用。最后, 聚焦于自对偶码, 详细分析其数学性质, 并结合具体实例加以说明, 同时进一步阐明该类方案在安全多方计算等高级密码协议中的应用潜力及其与其它研究领域的深刻联系。

基金项目: 潮州市科技计划项目 (2025ZC29); 韩山师范学院理科重点项目 (XN202028)。

作者简介: 林群, 韩山师范学院数学与统计学院, 讲师, 研究方向: 密码学与信息安全。

## 一、预备知识

### (一) 线性秘密分享方案的形式化描述

定义 1. (线性秘密分享方案)

一个在域  $F_p$  上关于参与者集合  $P=\{U_1, \dots, U_m\}$  的线性秘密分享方案 (LSSS) 由以下两个核心组件完全定义<sup>[8]</sup>:

1. 生成矩阵  $M_{k \times m}$ : 其中 1 (行数) 指秘密向量  $v=(s, r_2, r_3, \dots, r_l)$  的维度,  $s$  是真正的秘密,  $r_2, r_3, \dots, r_l$  是随机数, 用于隐藏秘密。 $m$  (列数) 指参与者的数量, 也等于生成的份额数量。

2. 单调访问结构  $\Gamma \subseteq 2^P$ : 这是一个参与者子集的集合, 定义了哪些参与者组合可以合法地重构出秘密。它必须满足单调性: 即若  $A \subseteq B$  且  $A \in \Gamma$ , 则必有  $B \in \Gamma$ 。

### (二) 份额生成算法

算法 1. LSSS 份额生成

– 输入: 秘密  $s \in F_p$ ,

– 输出: 份额向量  $(s_1, s_2, \dots, s_m) \in F_p^m$ 。

步骤:

1. 构造秘密向量  $v=(s, r_2, r_3, \dots, r_l) \in F_p^l$ , 其中  $r_2, r_3, \dots, r_l$  为均匀随机选择的随机数。

2. 计算份额  $(s_1, s_2, \dots, s_m) = v \cdot M$ , 并将份额  $s_i$  分配给参与者  $U_i$  ( $i=1, 2, \dots, m$ )。

### (三) 访问结构的矩阵表征

生成矩阵  $M$  不仅定义了如何生成份额, 还隐式地定义了访问结构 (即哪些参与者集合可以恢复秘密)。

定义 2. (授权集<sup>[8]</sup>) 对于生成矩阵  $M \in F_p^{k \times m}$ , 参与者子集  $A \subseteq P$  称为授权集, 当且仅当存在系数向量  $c=(c_1, c_2, \dots, c_{|A|}) \in F_p^{|A|}$ , 使得  $M_A \cdot c^T = (1, 0, 0, \dots, 0)^T$ . 其中  $M_A$  表示矩阵  $M$  中对应于集合  $A$  的列的子矩阵。访问结构  $\Gamma$  即为所有授权集的集合  $\Gamma = \{A \subseteq P \mid \exists c \in F_p^{|A|} \text{ 满足 } M_A \cdot c^T = (1, 0, 0, \dots, 0)^T\}$ 。

### (四) 秘密重构算法

算法 2. LSSS 秘密重构

– 输入: 授权集  $A \in \Gamma$  及其份额  $\{s_i\}_{i \in A}$ ,

– 输出: 秘密  $s \in F_p$ 。

步骤:

1. 求解线性方程组  $M_A \cdot c^T = (1, 0, 0, \dots, 0)^T$ , 得到重构系数  $c=(c_1, c_2, \dots, c_{|A|})$ 。

2. 计算秘密:

$$s = \sum_{i \in A} c_i s_i$$

即秘密可以通过将授权集中的份额进行线性组合恢复出来。

### (五) 线性性质

LSSS 满足以下线性性质<sup>[8]</sup>:

1. 份额生成线性: 每个份额  $s_i (i=1, 2, \dots, m)$  是秘密  $S$  和随机数  $r_j (j=2, \dots, l)$  的线性函数。

2. 秘密重构线性: 秘密  $S$  可表示为授权份额  $\{s_i\}_{i \in A}$  的线性组合, 其中  $A \in \Gamma$ 。

3. 同态性: 假定两个秘密  $s, s'$  的对应份额分别为  $\{s_i\}_{i \in A}, \{s'_i\}_{i \in A}$ , 则  $\{s_i + s'_i\}_{i \in A}$  是秘密  $s + s'$  对应的有效份额。

## 二、基于线性码的 LSSS 框架

定义 3. (线性码<sup>[8]</sup>) 设  $F_p$  是一个包含  $p$  个元素的有限域, 其中  $p$  为素数。一个参数为  $[n, k]$  的线性码  $C$  是向量空间  $F_p^n$  的一个  $k$  维线性子空间。其中  $n$  称为码的长度, 即码字的长度。 $k$  称为码的维数, 即子空间的维度。码中的码字总共有  $|C| = p^k$  个。

一个线性码  $C$  可以由以下两种矩阵之一完全描述:

### (一) 生成矩阵

设  $\{g_1, g_2, \dots, g_k\} \subset F_p^n$  是子空间  $C$  的一组基。将其作为行向量, 构成一个  $k \times n$  的矩阵  $G$ , 称为码  $C$  的生成矩阵, 即

$$G = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_k \end{pmatrix} = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{k1} & g_{k2} & \cdots & g_{kn} \end{pmatrix}$$

码  $C = \{u \cdot G \in F_p^n \mid u \in F_p^k\}$  可以由其生成矩阵  $G$  线性生成, 即信息向量  $u$  通过线性变换  $G$  被编码为码字  $c = u \cdot G$ 。

### (二) 校验矩阵

作为一个  $k$  维子空间, 线性码  $C$  在  $F_p^n$  中的对偶空间 (或零化空间) 是  $n-k$  维的。这个对偶空间本身也是一个线性码, 记为  $C^\perp$ , 称为  $C$  的对偶码<sup>[8]</sup>。设  $C^\perp$  的一组基为  $\{h_1, h_2, \dots, h_{n-k}\}$ , 将这些基向量作为行, 形成一个  $(n-k) \times n$  的矩阵  $H$ , 称为码  $C$  的校验矩阵。即

$$H = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_{n-k} \end{pmatrix}$$

码  $C$  可以等价地定义为校验矩阵  $H$  的零空间:

$$C = \{c \in F_p^n \mid H \cdot c^T = 0 \in F_p^{n-k}\}$$

这种对偶关系在秘密共享中有重要应用: 若线性码  $C$  定义一个线性秘密分享方案  $S$ , 则  $C^\perp$  对应  $S$  的对偶方案  $S^*$ 。

### (三) 生成矩阵与校验矩阵的关系

如果一线性码  $C$  有一个  $k \times n$  的生成矩阵  $G$ , 那么它就存在一个  $(n-k) \times n$  的校验矩阵  $H$ , 使得  $H \cdot G^T = 0$  或者  $G \cdot H^T = 0$ 。

这意味着生成矩阵  $G$  的所有行向量都与校验矩阵  $H$  的所有行向量正交。

结论: 1. 校验矩阵  $H$  其实就是对偶码  $C^\perp$  的生成矩阵。

2. 码  $C$  由  $G$  的行张成, 而对偶码  $C^\perp$  由  $H$  的行张成。

## 三、自对偶码

定义 4. (自对偶码<sup>[8]</sup>) 如果一个码  $C$  满足  $C = C^\perp$ , 则称  $C$  为一



个自对偶码。这意味着：一个向量是  $C$  的一个有效码字，当且仅当它与  $C$  中的每一个码字都正交，即点积为零。

从定义  $C = C^\perp$  可以推导出以下性质：

性质1：自对偶码的码字长度  $n$  必须是偶数，并且其维度  $k = \frac{n}{2}$ 。

证：设一个码  $C$  的长度是  $n$ ，维度是  $k$ ，则它的对偶码的维度是  $n - k$ 。

由于  $C = C^\perp$ ，它们的维度必须相等，即  $k = n - k$ ，

从而推出  $k = \frac{n}{2}$ ， $n$  必为偶数。

性质2：生成矩阵与校验矩阵本质一致。

证：由于  $C = C^\perp$ ，生成矩阵  $G$  能张成  $C$ ，也必能张成  $C^\perp$ ，

而  $C^\perp$  的生成矩阵是原码  $C$  的校验矩阵  $H$ 。

因此，生成矩阵  $G$  和校验矩阵  $H$  在本质上是同一个矩阵，它们张成相同的行空间。所以，通常使用同一个矩阵  $G$  来同时担任生成和校验的角色，即  $G = H$ 。

性质3：生成矩阵  $G$  必须满足正交性  $GG^T = 0$ 。

证：生成矩阵  $G$  和校验矩阵  $H$  满足  $GH^T = 0$ ，又由性质2，知  $G = H$ ，从而推出  $GG^T = 0$ 。

例：在二元域 构造生成矩阵  $G = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$ ，

有  $GG^T = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}^T = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ 。

给定  $F_2$  上向量  $u = (u_1, u_2)$ ，计算码字  $c = u \cdot G$ ，推出

$$C = \{u \cdot G \mid u \in F_2^2\} = \{(0, 0, 0, 0), (1, 0, 0, 1), (0, 1, 1, 0), (1, 1, 1, 1)\}$$

可以检验  $C$  中的码字两两正交，所以  $C$  是自对偶码。

对于一个自对偶的 LSSS，其访问结构  $\Gamma$  满足：一个集合是授权集，当且仅当它的补集是禁止集。

自对偶码具有强烈的内在对称性，这使其在理论和应用上成为一个非常强大的工具。它具有重要的应用价值：

1. 优美的对称性：它具有极其优美的数学结构，这使其更容易分析和实现。

2. 在密码学中的应用：它是构建自对偶线性秘密分享方案的基础，并能提供最优的效率和安全性，所以经常应用在安全多方计算和其它高级密码协议中。

3. 与其它领域的联系：自对偶码与群论、组合设计、甚至量子计算中的稳定子码都有着深刻的联系。

## 四、结束语

本文系统阐述了线性秘密分享方案（LSSS）的形式化定义、基于线性码的理论框架，以及自对偶码所蕴含的完美对称性与重要价值。这不仅为密码方案的设计提供了坚实的数学基础，也为其在安全多方计算等前沿密码协议中的高效实现开辟了道路。这一理论体系充分体现了代数方法在构建安全高效的可计算秘密共享机制中的核心作用，也为后续研究更复杂的访问结构以及跨领域的应用衔接提供了清晰的理论视角。

## 参考文献

- [1]Gharahi M ,Khazaei S .Optimal linear secret sharing schemes for graph access structures on six participants[J].Theoretical Computer Science,2019,7711–8.DOI:10.1016/j.tcs.2018.11.007.
- [2]Jafari A ,Khazaei S .On Abelian Secret Sharing: duality and separation.[J].IACR Cryptology ePrint Archive,2019,2019575.
- [3]Gharahi M ,Dehkordi H M .The complexity of the graph access structures on six participants[J].Designs, Codes and Cryptography,2013,67(2):169–173.DOI:10.1007/s10623-011-9592-z.
- [4]Kaboli R ,Khazaei S ,Parviz M .On Ideal and Weakly-Ideal Access Structures[J].IACR Cryptol. ePrint Arch. 2020, 2020:483.DOI:10.3934/AMC.2021017.
- [5]M á t é G ,P é t e r L .On the information ratio of graphs without high-degree neighbors[J].Discrete Applied Mathematics,2021,30455–62.DOI:10.1016/J.DAM.2021.07.011.
- [6]Jackson W A ,Martin K M .Geometric secret sharing schemes and their duals[J].Designs Codes & Cryptography , 1994, 4(1):83–95.DOI:10.1007/BF01388562.
- [7]Padr ó C ,V á zquez L ,Yang A .Finding lower bounds on the complexity of secret sharing schemes by linear programming[J].Discrete Applied Mathematics,2013,161(7–8):1072–1084.DOI:10.1016/j.dam.2012.10.020.
- [8]Padr ó C .Lecture Notes in Secret Sharing.[J].IACR Cryptology ePrint Archive,2012,2012674.
- [9]Farras O ,Kaced T ,Martin S ,et al.Improving the Linear Programming Technique in the Search for Lower Bounds in Secret Sharing[J].IEEE Transactions on Information Theory,2020,PP(99):1–1.DOI:10.1109/tit.2020.3005706.
- [10]Gharahi M ,Khazaei S .Reduced access structures with four minimal qualified subsets on six participants.[J].Advances in Mathematics of Communications,2018,12(1):199–214.DOI:10.3934/AMC.2018014.
- [11]Csirmaz L .Secret sharing and duality.[J].IACR Cryptology ePrint Archive,2019,20191197.
- [12]Jafari A ,Khazaei S .Partial Secret Sharing Schemes[J].IEEE Transactions on Information Theory, 2023, 69(8):5364–5385.DOI:10.1109/TIT.2023.3265093.
- [13]Mart í –Farr é J ,Padr ó C .On Secret Sharing Schemes, Matroids and Polymatroids.[J].IACR Cryptology ePrint Archive,2006,200677.
- [14]Xing C ,Yuan C .Evolving Secret Sharing Schemes Based on Polynomial Evaluations and Algebraic Geometry Codes[J].IEEE Transactions on Information Theory, PP[2025–12–10].DOI:10.1109/TIT.2024.3379278.
- [15]Abram D ,Roy L ,Scholl P .Succinct homomorphic secret sharing.[J].In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. Cham: Springer Nature Switzerland, 2024: 301–330.

# 基于深度学习的恶意代码检测算法设计与实现

张家华

北京鼎石国际学校, 北京 101318

DOI:10.61369/ASDS.2026010009

**摘 要 :** 针对恶意代码检测中的特征表示不足与检测精度有限问题, 提出一种基于多模态融合的深度学习检测方法。该方法首先从静态与动态两个维度设计并提取恶意代码的多模态特征表示, 构建包含静态特征处理子网络与动态特征处理子网络的深度学习模型架构, 通过特征融合模块实现跨模态信息的高效整合。进一步详细阐述模型的训练策略与优化细节, 并完成检测系统的模块化实现与部署, 实现了从样本预处理、特征提取、模型推理到结果可视化的全流程自动化。

**关 键 词 :** 恶意代码检测; 多模态特征; 深度学习

## Design and Implementation of Malicious Code Detection Algorithm Based on Deep Learning

Zhang Jiahua

Keystone Academy, Beijing 101318

**Abstract :** To address the issues of insufficient feature representation and limited detection accuracy in malicious code detection, a deep learning detection method based on multimodal fusion is proposed. This method first designs and extracts multimodal feature representations of malicious code from both static and dynamic dimensions, and constructs a deep learning model architecture that includes a static feature processing sub-network and a dynamic feature processing sub-network. Through a feature fusion module, efficient integration of cross-modal information is achieved. Further, the training strategy and optimization details of the model are elaborated in detail, and the detection system is modularly implemented and deployed, achieving full automation from sample preprocessing, feature extraction, model inference to result visualization.

**Keywords :** malicious code detection; multimodal features; deep learning

## 引言

随着网络攻击手段日趋复杂, 恶意代码数量呈指数级增长, 传统的基于特征代码与启发式规则的检测方法已经很难应对高层次持续威胁与未知变种。深度学习凭借其强大的特征学习能力和模式识别能力, 给恶意代码检测带来了一场革命。通过对代码二进制结构、行为序列和可视化图像的分析, 可以从海量数据中挖掘出深层的恶意模式, 提高检测的准确性和泛化能力。本研究旨在系统设计并实现高效的深度学习检测算法, 这对构建主动防御体系、保障关键信息基础设施安全具有重大理论价值与现实意义, 是当前网络安全领域的迫切需求。

## 一、多模态恶意代码特征表示与提取设计

### (一) 总体设计思路

面对恶意代码种类繁多、规避技术不断演化, 单模态特征描述已不能充分刻画恶意代码的恶意本质。本设计的核心思想是建立一种协同融合的多模态特征表达框架, 从静态和动态两个正交互补的角度实现对恶意代码的全方位画像<sup>[1]</sup>。整体架构遵循“特征级融合”的先进思想, 并不是简单的“拼接”, 而是在高层抽象层次上对齐和整合。首先, 需要建立统一的中间表示规范, 保证

不同模态特征能够在同一个向量空间中被度量和关联。本研究着重于特征的鲁棒性与解释性, 引入注意力机制等深度学习模块, 对不同模态、不同特征向量在判别任务中的贡献进行自动评估与权重, 实现在保持原有语义信息的前提下提取最具判别性的恶意模式。

### (二) 静态特征表示设计

静态特征分析主要研究非运行状态下恶意代码的结构、语法和统计特性。其核心思想是突破传统基于规则的人工特征工程, 采用深度学习模型对原始数据进行分层表达。对于可执行程序, 如 PE, 采用嵌入层或者一维卷积神经网络对二进制序列进行处

理,从而获得编码序列的语义和结构信息。同时,将文件反汇编得到的指令流视为一种文本序列,应用基于 Transformer 的预训练语言模型进行建模,以理解代码片段的功能语义。此外,将二进制文件可视化方法被进一步深化,引入密集连接的卷积网络提取其全局与局部的纹理、形状及空间分布特征<sup>[2]</sup>。这些来自不同视角的静态表示将通过一个特征投影网络映射到共享子空间,形成统一的静态特征向量,其优势在于能够有效揭示代码的固有意图与潜在威胁,且分析效率极高。

### （三）动态特征表示设计

动态特征表示旨在捕捉恶意代码在受控沙箱环境中运行时的行为语义,其设计重点在于对系统调用序列、API 调用图、网络流量、文件与注册表操作等运行轨迹进行结构化建模。将系统调用序列及其参数视为时间序列数据,采用门控循环单元或时序卷积网络捕获其长程依赖关系及顺序模式。针对较为复杂的应用程序接口依赖关系,将其抽象为有向图结构,利用图神经网络对节点和拓扑进行高层表示,揭示恶意行为的协同和触发机制。然后利用编码器将网络流量及行为记录转换成特征矢量<sup>[3]</sup>。动态特征提取的关键挑战在于行为轨迹的冗余与噪声,设计中引入了轨迹切片和关键行为筛选机制,并利用序列到序列的自动编码器进行降维与去噪,最终凝练出表征恶意代码核心动态行为的鲁棒特征向量。

## 二、多模态融合深度学习检测模型的设计

### （一）模型总体架构设计

模型总体架构采用分层分流的混合深度学习框架,其核心是一个双通道并行处理网络与一个中心融合决策模块。该架构以特征级融合为主导策略,旨在实现动、静态信息的有效协同利用。静态特征处理子网络和动态特征处理子网络是两条独立的特征抽取骨干,分别从代码内在属性与运行时行为中学习深层的判别性表征<sup>[4]</sup>。每一个子网络中都包含了多个层次的非线性变换,它们负责将原始的多模态特征映射到高维的语义空间中。两个子网络输出分别输入到特征融合模块中,实现特征的拼接和加权求和,并利用跨通道注意机制和张量融合技术挖掘多个模态间的深层次关联和互补信息。

### （二）静态特征处理子网络设计

静态特征处理子网需要同时处理多源异质数据,如代码语义、结构和可视化图像。对于运算码序列和反汇编文本,该子网络采用层次化的特征抽取策略。底层采用内嵌层将离散指令转换成密集矢量,再叠加一维卷积网络和门控环路单元层。卷积层用来捕捉本地的指令模式和语法结构,同时循环层对长距离的上下文依赖进行建模。针对二进制文件可视化图像,采用轻量级深度残差网络提取特征,通过残差连接有效缓解梯度丢失问题。采用全连通层对齐和降维,引入批规范化和 Dropout 层提高稳定性,防止过拟合。最后,在该子网络中输出一个具有统一、紧凑和信息量丰富的静态特征表达矢量。

### （三）动态特征处理子网络设计

针对系统调用序列、API 调用流等时序数据,本网络以双向

记忆网络为核心部件,通过双向结构实现对行为逻辑前后向的同步理解,并有效捕获恶意行为的因果关系。对于图结构的数据,如关系、进程树等,子网络与图卷积网络或者图注意力网络相结合。该网络利用消息传递机制聚集相邻节点的信息,学习其功能角色和结构重要性,实现非欧几何图数据向量表示的转换<sup>[5]</sup>。引入自关注层,自动评价不同时间阶段、不同行为事件在恶意判断中的作用权重,突出恶意行为关键片段。所有的动态特征在经过各自处理路径后,在融合层上聚合、压缩,得到动态特征向量,以反映行为语义。

### （四）多模态特征融合模块设计

多模态特征融合技术突破了单纯的矢量拼接和元素叠加的局限,是实现模型性能提升的关键。该方法首先从静态和动态子网络中获取特征矢量,然后利用共享的全连通层将其映射到统一的特征空间。采用基于交叉注意理论的张量融合方法对核进行融合<sup>[6]</sup>。具体地,通过计算静态和动态特征间的交互注意力权重矩阵,实现对不同维度特征交互强度和相关性的建模。然后,通过外积操作得到双模交互张量,将特征之间的高阶组合关系显式表示出来。这张数据经过多层感知机的压缩和提炼,提取出最有区别的跨通道协作特征。最后,在提取的协同特征基础上,采用小神经网络动态产生权重。

### （五）模型训练细节设计

模型的训练遵循深度学习标准范式,但针对多模态与安全领域的特殊性进行了精细化设计。该算法采用自适应矩估计方法,结合余弦退火学习速率调度策略,实现初始阶段的快速收敛和后期的精细调整。为避免过拟合问题,在网络结构中引入 Dropout、批量标准化等方法,同时针对多模态数据,研究多模态增强策略,如静态字节序列随机掩码、动态行为序列随机丢弃等,提高模型的鲁棒性。在训练过程中,采用分步学习的策略,对动、静两个子网络分别进行独立的预训练,并在训练过程中引入辅助标记,使其特征抽取能力更加稳定。通过端到端联合训练,对网络参数尤其是融合模块的权重值进行调整。利用检验集进行早停监测与超参调整,保证模型对未知数据保持最优的泛化性能。

## 三、检测系统的实现与部署

### （一）系统需求分析与整体架构

该系统旨在实现一个面向高吞吐量、低延迟环境的自动化恶意代码检测平台。核心需求包括支持对 PE、ELF、文档宏等多种文件格式的自动化分析;具有静、动态双方面的分析技能;将训练好的多模态深度学习模型用于实时或批处理;提供交互式测试报告。整个体系结构采用了 Microservice 设计模式,保证了模块的松耦合性、高内聚性和独立的可伸缩性。系统分为四层:数据录入与调度层,核心分析引擎层,模型服务层和用户界面层。数据输入层负责接收原始样本文件,将样本文件分发到静态和动态分析队列中。核心分析引擎由两个独立的服务(样本预处理和特征抽取)组成,分别部署在独立的计算环境中。模型服务层采用容器技术对深度学习模型进行封装,并提供高性能 REST 式 API



接口。用户界面层提供带有结果可视化仪表盘的网络管理接口。各层采用消息队列和 API 网关通讯，保证了系统的异步处理能力，可靠，可扩展性强。

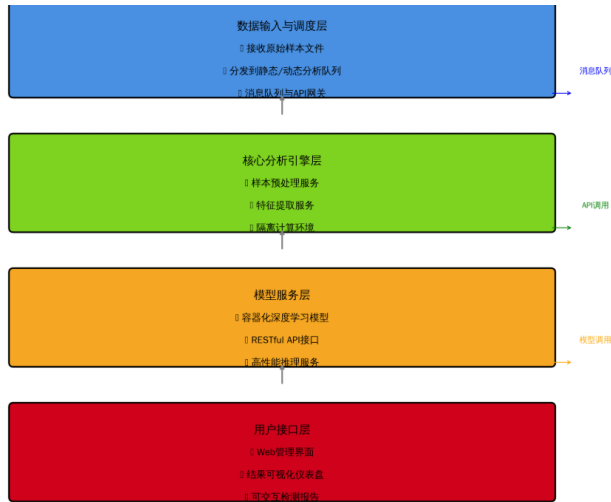


图1：整体架构图

（二）核心模块实现

1. 样本预处理模块实现

样本预处理模块主要负责接收原始文档，对原始文档进行初步的标准化和分流。在实现过程中，首先利用文件的魔数、扩展名等信息对文件格式进行快速地识别，并对文件的完整性进行检查。针对可执行程序，利用开源 LIEF 库对其进行解析，提取结构元数据，如文件头信息，输入导出表等，并对其进行初步识别和记录。然后，这个模块会启动一个轻量级的静态分析子过程，生成带有控制流图的拆装器代码。同时，对于需要进行动态分析的样本，该模块可以自动生成适用于沙箱或自定义沙盒环境的概要文件，并将其提交到动态分析任务队列中。通过前处理流程设计成流水式流程，对关键步骤进行日志记录，生成统一的样本元数据 JSON 描述文件，为后续特征抽取提供规范的输入接口和上下文信息。

2. 特征提取模块实现

基于多模态特征表达，特征抽取模块设计了静态和动态两个独立的处理流程。静态特征管线从预处理模块输出的解编文本、二进制字节和结构信息进行处理。将操作代码序列和 API 调用序列数字化，生成相应的嵌入矢量。可视化影像的产生是通过把二进制文件重新塑造成具有一定宽度的灰阶影像矩阵。动态特性管道对沙盒的执行环境进行监控，捕捉系统调用顺序，网络活动、注册表、文件操作等行为。该系统调用序列通过进程和线程 ID 进行会话划分与清洗。图结构特征从进程树和 API 调用依赖中构建邻接矩阵。所有原始特征经过标准化后，由相应的特征提取器转化为预设维度的特征向量。该模块的核心在于一个特征注册与管理中心，它统一调度各类特征提取器，确保特征生成的一致性和可复现性，并将最终的多模态特征向量存入高性能的向量数据库中以供模型推理调用。

3. 深度学习模型推理模块实现

该服务采用高性能深度学习推理引擎对模型进行加载和优

化。服务显示从特征抽取模块接收特征矢量的统一 gRPC 或 REST 风格 API 端点。在界面内，服务首先验证输入的静态和动态特性矢量，并对其进行维度校准。然后调用优化后的模型计算图用于正向传输。推理过程充分利用了 GPU 强大的并行运算能力，并通过批量处理技术提高了系统的吞吐率<sup>[7]</sup>。为了保证系统的稳定性和性能，系统采用了动态批处理机制、请求队列管理机制和自动扩展机制。对推理请求和结果进行了详细的日志记录，并将结果反馈给系统的监测部件。

4. 结果生成与可视化模块实现

结果生成和可视化模块完成了预处理、特征提取和模型推理等各个环节的融合，生成结构化的检测报告，并提供可视化的分析界面。首先将模型输出的原始概率得分与显著性分析结果相结合，利用可配置的判决引擎，将恶意判断标签和置信度转换为最终的恶意判断标签和置信度。报表产生器会汇编样本元数据，静态和动态分析总结，关键的恶意特性度量和模型推断细节，并以结构化的 JSON 或者 HTML 格式输出<sup>[8]</sup>。可视化界面则基于 Web 技术开发，提供交互式图表展示样本的行为序列图、API 调用关系图以及模型关注度的热力图。此外，该模块维护一个历史检测数据库，支持对检测结果的聚合统计、趋势分析与关联挖掘。

表1：可视化模块关键输出数据

| 字段名称                  | 数据类型          | 描述                   |
|-----------------------|---------------|----------------------|
| Sample_ID             | String        | 样本唯一标识符 (MD5/SHA256) |
| Final_Verdict         | String        | 最终判定结果（恶意/良性）        |
| Confidence_Score      | Float         | 模型判定置信度 (0-1)        |
| Top_Static_Features   | Array[String] | 权重最高的静态特征描述列表        |
| Top_Dynamic_Behaviors | Array[String] | 权重最高的动态行为描述列表        |
| Report_Path           | String        | 完整分析报告存储路径           |
| Visualization_URL     | String        | 交互式可视化界面访问链接         |

（三）系统集成与部署

采用容器和编排技术实现了系统的集成，所有的核心模块都以独立的容器镜像形式封装，并由 Kubernetes 统一编排、部署和管理。通过配置映射和安全对象来管理具有密钥的系统配置信息。数据持久层采用高性能分布式文件系统对样本文件进行存储；消息队列是整个系统的中枢，它将各个微服务连接在一起，实现异步解耦。整个部署体系结构采用前、后端分离的方式，前端网络接口通过 API 网关和后端服务集群进行通讯。该系统具有完备的监测报警系统，利用 Prometheus 采集系统的性能指标，通过 Grafana 在仪表盘上展示，实时监测服务健康状况、资源利用率和监测业务指标。

表2：系统部署资源配置

| 服务组件    | 容器副本数 | CPU 资源请求 / 限制 | 内存资源请求 / 限制 | 存储卷挂载 | 网络策略 |
|---------|-------|---------------|-------------|-------|------|
| 预处理调度器  | 2     | 0.5 / 2核      | 1Gi / 4Gi   | 无     | 内部服务 |
| 静态特征提取器 | 3     | 1 / 4核        | 2Gi / 8Gi   | 样本缓存卷 | 内部服务 |
| 动态沙箱集群  | 5     | 2 / 8核        | 4Gi / 16Gi  | 隔离数据卷 | 严格出站 |
| 模型推理服务  | 2     | 申请 GPU        | 4Gi / 16Gi  | 模型存储卷 | 内部服务 |



|                |   |            |             |        |       |
|----------------|---|------------|-------------|--------|-------|
| 消息队列           | 3 | 0.5 / 1 核  | 512Mi / 2Gi | 持久化数据卷 | 内部服务  |
| API 网关与 Web 前端 | 2 | 0.25 / 1 核 | 256Mi / 1Gi | 无      | 面向互联网 |

四、结语

本研究设计并实现了一种基于多模态融合深度学习的恶意代

码检测方案。构建静态和动态双模态特征表达体系，设计特征处理子网和融合模块，提高模型的识别准确率和泛化能力。通过系统的集成和部署，证明了所提出的方法是可行和实用的。未来工作将集中于引入更丰富的代码表征模态、探索更高效的自适应融合机制，并研究模型在对抗性样本攻击下的鲁棒性增强方法，以应对日益复杂和隐蔽的恶意代码威胁。

参考文献

[1] 李梦, 刘万平, 黄东. 基于特征融合的恶意代码检测 [J]. 计算机工程与设计, 2024, 45(12): 3568–3574.

[2] 蒋应瑞, 黎秋玲. 一种基于卷积神经网络的恶意代码检测模型 [J]. 江苏通信, 2024, 40(06): 102–105.

[3] 熊其冰, 苗启广, 杨天, 等. 一种基于混合量子卷积神经网络的恶意代码检测方法 [J]. 计算机科学, 2025, 52(03): 385–390.

[4] 蒲经纬, 张辉, 唐斌. 基于深度学习的恶意代码检测技术研究 [J]. 网络安全技术与应用, 2024, (10): 39–43.

[5] 宋亚飞, 张丹丹, 王坚, 等. 基于深度学习的恶意代码检测综述 [J]. 空军工程大学学报, 2024, 25(04): 94–106.

[6] 张晓良, 柴艳玉, 吴克河, 等. 一种基于增量学习的恶意代码检测方法 [J]. 计算机与数字工程, 2024, 52(07): 2141–2145+2220.

[7] 靳黎忠, 薛慧琴, 段明博, 等. 基于多频特征学习的恶意代码变种分类 [J]. 计算机工程与设计, 2024, 45(07): 1934–1940.

[8] 尚承翔, 李梓宇, 李瀚洋, 等. 基于深度迁移学习的恶意代码可视化检测 [J]. 网络安全技术与应用, 2024, (03): 37–39.

# 处理效应的异质性的识别

## ——基于因果机器学习的仿真和估计

晏发发<sup>1,2</sup>, 颜丽金<sup>1</sup>, 陈铮浩<sup>1\*</sup>

1. 广州理工学院经济管理学院, 广东 广州 510540

2. 广东特色金融与高质量发展研究中心, 广东 广州 510540

DOI:10.61369/ASDS.2026010010

**摘 要 :** 既有因果机器学习的相关文献所提出的用于识别异质群体的平均处理效应 (GATE) 的方法未能考虑在协变量变化的情况下解释不同群体间处理效应的异质性。为了解决该问题, 本文提出基于无偏机器学习 (DML) 的平衡组平均处理效应 (BGATE) 来衡量具有预先确定的协变量特定分布的组平均处理效应 (GATE), 通过计算两个 BGATE 之间的差值来比较两个 GATE 的值, 从而更好地识别因果效应的异质性, 最终将由协变量不同分布所导致的差异与由解释变量所导致的差异区分开来。该估计量在标准条件下具有  $\sqrt{N}$ -一致性和渐近正态的性质。通过对比 DML、自动无偏机器学习 (Auto-DML) 和重新加权 (Reweighting Approach) 三种估计方法的仿真结果可知: 如果已知 DML 没有性能问题, 如当倾向得分很极端时, 建议采用 DML 估计量, 其模拟表现最好; 如果已知 DML 有性能问题, 建议采用 Auto-DML 估计量或重新加权方法。

**关 键 词 :** 调节效应; 无偏机器学习; 重新加权方法; 异质性; 平衡组平均处理效应

## Identification of Treatment Effect Heterogeneity — Simulation and Estimation Based on Causal Machine Learning

Yan Fafa<sup>1,2</sup>, Yan Lijin<sup>1</sup>, Chen Zhenghao<sup>1\*</sup>

1. School of Economics and Management, Guangzhou Institute of Science and Technology, Guangzhou, Guangdong 510540

2. Guangdong Research Center of Characteristic Finance and High-Quality Development, Guangzhou, Guangdong 510540

**Abstract :** Existing studies on causal machine learning proposes methods for identifying the Average Treatment Effect for heterogeneous groups (GATE), failed to account for the heterogeneity of treatment effects across different groups under varying covariates. To address this issue, this paper proposes the Balanced Group Average Treatment Effect (BGATE) based on Double Machine Learning (DML) to measure the Group Average Treatment Effect (GATE) with predetermined specific distributions of covariates. By calculating the difference between two BGATE values, the GATE values of two groups can be compared, thereby better identifying the heterogeneity of causal effects. Ultimately, this approach distinguishes between differences caused by varying distributions of covariates and those caused by explanatory variables. The estimator exhibits  $\sqrt{N}$ -consistency and asymptotic normality under standard conditions. A comparison of simulation results from three estimation methods—the DML, the Auto-DML, and the Reweighting approach—shows that if DML is known to have no performance issues (e.g., when the propensity score is not extreme), the DML estimator is recommended, as it performs best in simulations. If DML is known to have performance issues, the Auto-DML estimator or the reweighting approach is recommended.

**Keywords :** moderation effect; debiased machine learning; reweighting approach; heterogeneity; balanced group average treatment effect

基金项目: 2024年广州理工学校级项目“数字信贷对服务业创业的影响效应和机制——基于广东县域面板数据的经验证据”(2024KY052); 2024-2025年度广州理工学院校级教学质量与教学改革工程项目“多语言程序设计+数字经济专项人才培养计划(实验班)”(2024XZLGC39); 2025年广州理工学校级项目“数字经济对中国经济不平衡的影响效应与影响路径研究”(2025KYQ003); 2025年广州理学院工作室项目铭心博士工作室; 2025年广东省教育科学规划课题(高等教育专项)“产学研赛+混合编程+数字经济跨学科人才培养探究”(2025GXJK0660); 2025年广东省教育科学规划课题(高等教育专项)“AI虚拟助教助力高校国贸专业互动式教学的实践探索”(2025GXJK0664); 2025-2026年度广东省教育评估协会一般课题“学生视角下高校课程思政实施效果评价体系构建研究”(BDPG25100); 2025年广东省普通高校特色创新类项目哲学社会科学类“广东省低空经济产业发展的路径优化: 基于多维系统关联的研究”(2025GXJK0664)。

作者简介:

晏发发, 广州理工学院经济管理学院, 副教授, 研究领域: 计量经济学理论和应用、宏观模型构建和应用、农业经济、数字经济、教育经济、资产定价、机制设置、博弈论等;

颜丽金, 广州理工学院经济管理学院, 本科生在读, 研究领域: 计量经济学理论和应用、宏观模型构建和应用、会计学等;

通讯作者: 陈铮浩(第三作者), 广州理工学院经济管理学院, 本科生在读, 研究领域: 计量经济学理论和应用、宏观模型构建和应用、会计学等。

## 引言

一个特定的变量是否导致处理效果的差异称为因果调节变量。除了平衡所有协变量混淆组间调节效应的分布外，还必须进行其他假设来解释因果关系的处理效应的差异，只有这样组变量才可以被认为是一个不混杂的调节因素。机器学习方法不仅可以融入传统计量方法来改进因果推断计量方法，还可以采用数据驱动以考虑广泛的模型，从而避免模型设定错误。因此，机器学习在因果推断经济学中的应用有着重要现实意义。

本文聚焦于无混淆设定下组间处理效果差异的估计方法研究。首先，提出平衡组平均处理效果（BGATE），该参数是基于预先确定的协变量特定分布的组平均处理效应（ $\Delta$ GATE），通过计算两个 BGATE 的差值（同样记作 BGATE），能够有效实现组间处理效果的对比。本文展示 BGATE 它如何与两个组平均处理效应的差异（ $\Delta$ GATE）相关，讨论了它的识别方法，并提出了不同的估计量来描述离散调节因素（为了简单起见，本文将发现异质处理效应的变量称为调节因素）和离散处理。关于无偏机器学习的估计量（Debiased Machine Learning, DML）（Chernozhukov 等，2018）<sup>[1]</sup>，本文证明了该估计量在真实值处渐近正态分布，是一个有效的统计量。此外，本文还将 DML 方法与自动无偏机器学习（Auto-DML）和重新加权方法（Reweighting Approach）估计  $\Delta$ BGATE<sup>[2]</sup> 的结果进行对比。DML 依赖于使用一个基于倾向得分的双重稳健评分函数，如果倾向得分极端，这可能会出现（Lechner 和 Mareckova，2024<sup>[3]</sup>；Frolich，2004<sup>[4]</sup>）。因为 Auto-DML 使用 Riesz 表示而不是倾向得分，所以 Auto-DML 不会受到极端倾向得分的问题影响（Chernozhukov 等，2021<sup>[5]</sup>；Chernozhukov 等，2018<sup>[1]</sup>）；重新加权方法允许用户使用任何他们偏好的方法来估计  $\Delta$ BGATE，因为  $\Delta$ BGATE 等于重新加权数据上的  $\Delta$ GATE。这两种估计方法是有效的（郭峰和陶旭辉，2023）<sup>[6]</sup>。

本文其余章节安排如下：第一节对相关领域文献进行综述，第二节定义了因果调节框架；第三节给出了 DML 估计量的渐近性质；第四节报告了样本规模为 500 的蒙特卡罗仿真设计和结果；第五节对全文进行总结归纳，得出研究结论。

## 一、文献综述

既有研究探讨了如何在组水平上检测（低维）异质性，称为“组平均处理效应”（GATEs 是一种条件平均处理效），目前已经开发了几种方法来估计 GATEs。例如，Abrevaya 等（2015）<sup>[7]</sup>展示了如何在无混杂条件下非参数识别 GATEs，并使用逆概率加权估计量对其进行估计。Lechner（2018）<sup>[8]</sup>改进了一种随机森林算法，以调整混淆因素从而来估计异质性处理效应。Semenova 和 Chernozhukov（2021）<sup>[9]</sup>使用 DML 框架来寻找基于线性模型的异质性。Zimmert 和 Lechner（2019）<sup>[10]</sup>以及 Fan 等（2022）<sup>[11]</sup>开发了一个两步估计量，允许非参数估计 GATEs。第一步是使用机器学习方法估计，第二步是应用非参数局部常数回归。

本文通过提出  $\Delta$ BGATE 估计量以刻画不同组间处理效果差异，即，在平衡一些协变量的分布的同时，比较两组的处理效果。若研究目的聚焦于识别处理结果中赢家与输家的特征，该方法具有较高适配性；然而，当核心诉求在于量化某一变量对两组处理效果差异的影响程度时，该方法则更具针对性与有效性。

DML 框架由 Chernozhukov 等（2018）<sup>[1]</sup>提出，该框架使用机器学习的方法来进行因果分析。机器学习算法可能会引入两种偏差：正则化偏差和过拟合偏差。DML 的主要思想是，通过使用奈曼正交得分函数（Neyman-orthogonal score functions）以克服正则化偏差。既有文献已经将根据不同的设置调整了通用 DML 框架用于中介分析（Farbmacher 等，2022）<sup>[12]</sup>。本文通过使用这个框架来估计因果调节效应来为该领域的文献做出贡献。

## 二、因果调节框架

### （一）定义

本文使用的因果调节框架是基于（Rubin，1974）<sup>[13]</sup>的潜在结果框架。因果效应被定义为两个潜在结果之间的差异，而对于一个个体，本文只能观察到两个潜在结果中的一个。因此，找到一个可信的反事实是很棘手的。我们观察到来自独立随机变量  $H_i = (D_i, Y_i, Z_i, X_i)$  的  $N$  个独立同分布的样本，它们以未知概率的分布  $P$  来抽取。重点在于处理变量  $D_i$  和调节变量  $Z_i$ 。为简单起见，假设它们是二元的（处理变量的结果为  $d \in \{0, 1\}$ ，调节变量的结果为  $z \in \{0, 1\}$ ）。与既有研究中的设定一样，潜在结果由处理变量索引： $(Y_i^0, Y_i^1)$ 。最后，一组  $k \in \{0, \dots, p\}$  协变量  $X_{i,k}$  可能同时影响处理和潜在的结果，其中  $X_i = (X_i^0, X_i^1, Z_i^0, Z_i^1)$ 。此外，潜在的协变量和潜在的调节因子被定义为： $X_i = (X_i^0, X_i^1, Z_i^0, Z_i^1)$ 。

由于只能观察到一种潜在结果的实现，因此我们永远无法一致地估计个体处理效应（ITE） $\xi_i = Y_i^1 - Y_i^0$  的实现。然而，在适当的假设下，平均处理效应（ATE） $\theta = E[Y_i^1 - Y_i^0]$  的识别是可能的（Imbens 和 Wooldridge，2009）<sup>[14]</sup>。另外，为了进一步研究  $\xi_i$  异质性的不同方面，可以通过条件平均处理效应（CATE）来捕捉。CATE 测量了一组（子）协变量  $x_i$  条件下的平均处理效果。个体化平均处理效果（IATE）和组平均处理效果（GATE）都是特异性的 CATEs。IATE 在最精细的聚合水平上测量处理效果。也就是说，它比较了所有个体的平均处理效果与所有相关协变量的特定值。在形式上，IATE 被定义为

$$\tau(x, z) = E[Y_i^1 - Y_i^0 | Z_i = z, X_i] \quad (1)$$

GATE 测量组间处理效果, 即在比 IATE 更聚合的水平上且比 ATE 更精细。GATE 定义为

$$\theta^G(z) = E[Y^1 - Y^0 | Z_i = z] = E[\tau(X_i, 0) | Z_i = 0] \quad (2)$$

如果仅仅刻画处理效应的异质性, 那么采用 IATEs 和 GATEs 就足够了。如果识别两组之间的处理效果差异, 那么需要采用两组 GATE 之间的差异 ( $\Delta$  GATE), 即

$$\begin{aligned} \theta^{\Delta G} &= E[Y^1 - Y^0 | Z_i = 1] - E[Y^1 - Y^0 | Z_i = 0] \\ &= [\tau(X_i, 1) | Z_i = 1] - [\tau(X_i, 0) | Z_i = 0] \end{aligned} \quad (3)$$

然而, 因为两组在协变量  $x_i$  的分布上可能存在差异, 通常该效应难以解释。因此, 本文引入平衡组平均处理效应 (BGATE) 以平衡本文想要相互比较的各 (由不同  $Z_i$  值定义) 中其他变量的分布。用于平衡 GATEs 的变量表示为  $W_i$ 。  $W_i$  是  $X_i$  的一部分。如果  $W_i$  为空集, 或  $W_i$  独立于  $Z_i$ , 则 BGATE 减少为 GATE。因此, 新的感兴趣的参数, 即用  $\theta^B(z)$  表示, 被定义为

$$\theta^B(z) = E[E[Y_i^1 - Y_i^0 | Z_i = z, W_i]] = E[E[\tau(X_i, z) | Z_i = z, W_i]] \quad (4)$$

以及其差异

$$\begin{aligned} \theta^{\Delta B} &= E[E[Y_i^1 - Y_i^0 | Z_i = 1, W_i]] - E[E[Y_i^1 - Y_i^0 | Z_i = 0, W_i]] \\ &= E[E[\tau(X_i, 1) | Z_i = 1, W_i]] - E[E[\tau(X_i, 0) | Z_i = 0, W_i]] \end{aligned} \quad (5)$$

其中,  $\Delta$  BGATE ( $\theta^B$ ) 表示两组之间的差异, 将两组中协变量 ( $W_i$ ) 的分布调整为总体人群分布。 $\Delta$  BGATE 通常显示出调节效应。平衡组平均处理效应 ( $\Delta$  BGATE) 作为调节因果效应。

## (二) 效应分解

为了明确  $\Delta$  GATE 和  $\Delta$  BGATE 之间的区别,  $\Delta$  GATE 被分解为两个组成部分:  $\Delta$  BGATE, 代表调节变量的直接效应, 以及由各组间  $W_i$  分布的差异引起的组合效应组, 组合成分根据其相对重要性对两个  $Z_i$  子样本进行加权, 以捕捉  $W_i$  分布的差异:

$$\begin{aligned} & \frac{E[Y_i^1 - Y_i^0 | Z_i = 1] - E[Y_i^1 - Y_i^0 | Z_i = 0]}{\Delta \text{GATE}} \\ &= \underbrace{E[E[Y_i^1 - Y_i^0 | W_i, Z_i = 1] - E[Y_i^1 - Y_i^0 | W_i, Z_i = 0] | Z_i = 0]}_{\text{直接效应: } \Delta \text{BGATE}} \\ &+ \underbrace{\frac{P(Z_i = 0)}{P(Z_i = 1)} E[E[Y_i^1 - Y_i^0 | W_i, Z_i = 1]] - E[E[Y_i^1 - Y_i^0 | W_i, Z_i = 1] | Z_i = 0]}_{\text{组合效应(1)}} \\ &- \underbrace{\frac{P(Z_i = 1)}{P(Z_i = 0)} E[E[Y_i^1 - Y_i^0 | W_i, Z_i = 0]] - E[E[Y_i^1 - Y_i^0 | W_i, Z_i = 0] | Z_i = 1]}_{\text{组合效应(2)}} \end{aligned} \quad (6)$$

这种分解用于将两组之间结果差异分解为两个部分: 一部分是由于协变量分布的差异所致, 另一部分是由于感兴趣的变量所致。本文关注总体而非特定部分。虽然  $\Delta$  BGATE 是使用协变量  $W_i$  的无条件分布来定义的, 但是  $\Delta$  BGATE 也可以使用条件分布  $W_i | Z_i = 1$  和  $W_i | Z_i = 0$  来定义。

## (三) 识别

为了在不混杂的环境中识别 GATE、BGATE、 $\Delta$  GATE 或  $\Delta$  BGATE, 通常需要识别假设 (例如, Imbens, 2004)<sup>[15]</sup>, 具体如下:

假设 1

(1) 条件独立性:

$$(Y_i^1, Y_i^0) \perp D_i | X_i = x, Z_i = z, \forall x \in X, \forall z \in Z$$

(2) 共同支持:  $0 < P(D_i = d | X_i = x, Z_i = z) < 1$ ,

$\forall d \in \{0, 1\}, \forall x \in X, \forall z \in Z$ ,

(3) 混杂因子的外生性:  $X_i^0 = X_i^1, Z_i^0 = Z_i^1, (d)$ ,

(4) 稳定的单位处理值假设:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0 \quad (7)$$

定理 1

在假设 1 下, 参数

$$\theta^B(z) = E[E[Y_i^1 - Y_i^0 | Z_i = z, W_i]] \quad (8)$$

被识别为

$$E[E[\mu_1(Z_i - X_i) - \mu_0(Z_i - X_i) | Z_i = z, W_i]] \quad (9)$$

其中

$$\mu_d(z, x) = E[Y_i | D_i = d, Z_i = z, X_i = x] \quad (10)$$

因此

$$\theta^{\Delta B} = E[E[Y_i^1 - Y_i^0 | Z_i = 1, W_i]] - E[E[Y_i^1 - Y_i^0 | Z_i = 0, W_i]] \quad (11)$$

已识别为

$$E[E[\mu_1(Z_i - X_i) - \mu_0(Z_i - X_i) | Z_i = 1, W_i]] - E[E[\mu_1(Z_i - X_i) - \mu_0(Z_i - X_i) | Z_i = 0, W_i]] \quad (12)$$

## 三、基于 DML 的因果调节效应的估计

由于本研究关注的是处理效应的异质性, 因此本文采用 DML 估计策略侧重于  $\Delta$  BGATE。

### (一) 估计方法

无偏机器学习方法采用三步估计策略。为了灵活处理潜在高维协变量向量, 可以选用 Chernozhukov 等 (2018)<sup>[1]</sup> 提出的 DML 方法。第一步, 估计通常的双鲁棒评分函数。第二步, 评分函数回归到由调节变量  $Z_i$  的协变量定义的两个指标变量上。最后, 计算由调节变量所划分的两组之间的差异, 并对平衡变量  $W_i$  求平均值。估计的双鲁棒评分函数由以下表达式:

$$\begin{aligned} \hat{\phi}^{\Delta B}(h; \theta^{\Delta B}, \hat{\eta}) &= \hat{g}_1(w) - g_0(w) + \frac{z(\hat{\delta}(h) - \hat{g}_1(w))}{\hat{\lambda}_1(w)} \\ &- \frac{(1-z)(\hat{\delta}(h) - \hat{g}_0(w))}{1 - \hat{\lambda}_1(w)} - \theta^{\Delta B} \end{aligned} \quad (13)$$

$$\begin{aligned} \delta(h) &= \mu_1(z, x) - \mu_0(z, x) + \frac{d(y - \mu_1(z, x))}{\pi_1(z, x)} \\ &- \frac{(1-d)(y - \mu_0(z, x))}{1 - \pi_1(z, x)} \end{aligned} \quad (14)$$

$$\lambda_z(w) = P(Z_i = z | W_i = w) \quad (15)$$

$$g_z(w) = E[\delta(H_i) | Z_i = z, W_i = w] \quad (16)$$

$$\mu_d(z, x) = E[Y_i | D_i = d, Z_i = z, X_i = x] \quad (17)$$



$$\text{和 } \pi_d(z, x) = P(D_i = d | Z_i = z, X_i = x) \quad (18)$$

其中,  $\hat{\delta}(h)$ ,  $\hat{\mu}_d(z, x)$ ,  $\hat{\lambda}_z(w)$ ,  $\hat{\pi}_d(z, x)$  分别代表了  $\delta(h)$ ,  $\mu_d(z, x)$ ,  $\lambda_z(w)$ ,  $\pi_d(z, x)$  的估计值。另外, 需注意  $\hat{\delta}(h)$  对  $Z_i$  和  $W_i$ ,  $\hat{g}_z(w) = \hat{E}[\delta(H_i) | Z_i = z, W_i = w]$  的回归是  $\hat{\delta}(h)$  对  $Z_i$  和  $W_i$  的相应回归  $\hat{g}_z(w) = \hat{E}[\delta(H_i) | Z_i = z, W_i = w]$ 。因此,  $E[\dots]$  表示一个一般的回归估计量, 它可以是线性的, 也可以是非线性的, 这取决于假定的数据生成过程。最后, 估计的干扰参数为  $\hat{\eta} = (\hat{\mu}_d(z, x), \hat{\lambda}_z(w), \hat{g}_z(w))$ 。

如上所述, 必须估计分数函数  $\delta(h)$ 。 $\delta(h)$  的干扰函数误差的乘积必须收敛速度快于或等于  $\sqrt{N}$ , 并且必须使用与 K 折交叉验证 (K-folds) (K > 1) 的交叉拟合。在第二步估计中, 是干扰项的乘积。数误差必须以  $\sqrt{N}$  的速度收敛, 并必须使用与 J 折交叉验证 (J-folds) (J > 1) 交叉拟合。如果满足某些条件, 则估计量是  $\sqrt{N}$  一致且渐近正态的 (见前文所述)。因为  $E[\phi^{AB}(H_i; \theta^{AB}, \eta)] = 0$ , 所以  $\theta^{AB}$  的方差为

$$\text{Var}(\hat{\theta}^{AB}) = \text{Var}(\phi^{AB}(H_i; \theta^{AB}, \eta)) = E[\phi^{AB}(H_i; \theta^{AB}, \eta)^2] - E[\phi^{AB}(H_i; \theta^{AB}, \eta)]^2 = E[\phi^{AB}(H_i; \theta^{AB}, \eta)^2] \quad (19)$$

其中,  $S_J$  是第二次估计步骤中的一个随机折叠, 具体步骤参考 (Chernozhukov 等, 2018) [1] 的研究。

## (二) 渐近性质

在以下假设条件下, 本文研究了估计量的渐近性质:

假设 2

(交叠性): 倾向得分  $\lambda_z(w)$  和  $\pi_d(z, x)$  都不趋于 0 和 1:

$\kappa < \lambda_z(w), \pi_d(z, x) < 1 - \kappa \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$ , 对于某些常数  $\kappa > 0$ 。

假设 3

(一致性): 扰动函数估计量是上确界范数一致的:

$$\begin{aligned} (1) \quad & \sup_{w \in \mathcal{W}} |\hat{\lambda}_z(w) - \lambda_z(w)| \xrightarrow{p} 0, \\ (2) \quad & \sup_{w \in \mathcal{W}} |\hat{g}_z(w) - g_z(w)| \xrightarrow{p} 0, \\ (3) \quad & \sup_{x \in \mathcal{X}, z \in \mathcal{Z}} |\hat{\mu}_d(z, x) - \mu_d(z, x)| \xrightarrow{p} 0, \\ (4) \quad & \sup_{x \in \mathcal{X}, z \in \mathcal{Z}} |\hat{\pi}_d(z, x) - \pi_d(z, x)| \xrightarrow{p} 0. \end{aligned}$$

假设 4

(风险衰减性): 结果和倾向得分函数的估计误差的乘积衰减为:

$$E[(\hat{g}_z(W_i) - g_z(W_i))^2] E[(\hat{\lambda}_z(W_i) - \lambda_z(W_i))^2] = o_p\left(\frac{1}{N}\right) \quad (20)$$

$$E[(\hat{\mu}_d(Z_i, X_i) - \mu_d(Z_i, X_i))^2] E[(\hat{\pi}_d(Z_i, X_i) - \pi_d(Z_i, X_i))^2] = o_p\left(\frac{1}{N}\right) \quad (21)$$

其中, 期望是指  $X_i$  和  $Z_i$  或  $W_i$  取  $\hat{\mu}_d$ ,  $\hat{\pi}_d$ ,  $\hat{g}_z$  和  $\hat{\lambda}_z$ 。

如果两个扰动参数都以  $\sqrt{N}$  一致速率的估计, 则误差的乘积是  $o_p\left(\frac{1}{N}\right)$  同阶项。因此, 参数的估计是  $N^{\frac{1}{2}}$  一致就足够了。

假设 5

(条件方差的平稳性): 结果和倾向得分函数的条件方差是

有界的:

$$(1) \quad \sup_{w \in \mathcal{W}} \text{Var}(\delta(H_i) | Z_i = z, W_i = w) < \hat{\delta}_{z1} < \infty,$$

$$(2) \quad \sup_{w \in \mathcal{W}} \text{Var}(\delta(H_i) | Z_i = z, W_i = w) < \hat{\delta}_{z0} < \infty,$$

$$(3) \quad \sup_{x \in \mathcal{X}, z \in \mathcal{Z}} \text{Var}(Y_i | D_i = d, Z_i = z, X_i = x) < \hat{\delta}_d < \infty \text{ 对于某些常}$$

数  $\hat{\delta}_{z1}, \hat{\delta}_{z0}, \hat{\delta}_d > 0$ 。

假设 6

(收敛性): 第二步回归估计量  $E[\dots]$  在定义域内室平稳的。

假设 2 至 5 是 DML 的标准假设 (Chernozhukov 等, 2018) [1], 唯一的区别是这些假设只适用于第一个和第二个估计步骤。假设 4 是必要的, 以确保扰动项函数误差的乘积收敛速度比  $\sqrt{N}$  更快或与  $\sqrt{N}$  相等。当稀疏性条件得到满足时, 各种机器学习方法的  $L_2$  收敛速率符合这些性质。Belloni 和 Chernozhukov (2013) [10] 表明, 在近似稀疏的情况下, 这意味着排序系数的绝对值迅速衰减, Lasso 估计量的误差是阶数

$$O\left(\sqrt{\frac{\log(\max(p, N))}{N}}\right)$$

其中,  $p$  是回归变量,  $s$  是真实系数的未知数量。根据这些假设, 本文可以得出主要的理论结果 (见定理 1):

定理 1

在假设 2—假设 6 下, 所提出的  $\Delta$ BGATE 估计策略满足  $\sqrt{N}(\hat{\theta}^{AB} - \theta^{AB}) \xrightarrow{d} N(0, V^*)$  和  $V^* = E[\phi^{AB}(H_i; \theta^{AB}, \eta)^2]$ 。该估计量是  $\sqrt{N}$  一致且渐近正态的。

## 四、因果调节效应的仿真

### (一) 数据生成过程

本文从用  $p = 10$  开始模拟一个  $p$  维协变量矩阵  $X_{i,p}$ 。前两个协变量来自均匀分布的  $X_{i,0}$ ,  $X_{i,1} \sim U[0, 1]$ , 其余的协变量来自正态分布的  $X_{i,2}, \dots, X_{i,p-1} \sim N(0.5, \sqrt{1/12})$ 。所有协变量的平均值均为 0.5, 标准差为  $\sqrt{12}$ 。调节变量  $Z_i$  来自于一个有概率的伯努利分布

$$P(Z_i = 1 | X_{i,0}, X_{i,1}) = (0.1 + 0.8\beta_{2,4}(X_{i,0} \times X_{i,1})) \quad (22)$$

其中,  $\beta_{2,4}(X_{i,0} \times X_{i,1})$  表示形似参数为 2 和 4 的 Beta 分布的累积分布函数。处理变量  $D_i$  是从下面伯努利分布中抽取的

$$P(D_i = 1 | X_{i,0}, X_{i,1}, X_{i,2}, X_{i,3}, Z_i) = (0.2 + 0.6\beta_{2,4}(\frac{X_{i,0} \times X_{i,1} \times X_{i,2} \times X_{i,3} \times Z_i}{5})) \quad (23)$$

本文据此指定了处理和非处理下的反应函数, 以及调节变量的两种状态。高度非线性的非处理响应函数由

$$\mu_0(X_i) = \sin(\pi \times X_{i,0} \times X_{i,1}) + (X_{i,2} - 0.5)^2 + 0.1X_{i,3} + 0.3X_{i,5}$$

处理下的反应函数依赖于  $Z_i$ , 并被定义为

$$\mu_1(0, X_i) = \mu_0(X_i) + \sin(1.4X_{i,0}) + \sin(6X_{i,1}) + 0.6X_{i,2}^2 + 0.3X_{i,5} \text{ 和}$$

$\mu_1(1, X_i) = \mu_0(X_i) + \sin(4.9X_{i,0}) + \sin(2X_{i,1}) + 0.7X_{i,2}^4 + 0.4X_{i,3} + 0.2$ , 它们的选择使得  $\Delta$ BGATE 与  $\Delta$ GATE 不同。最后, 本文将潜在的结果仿真为  $Y_i^d(z) = \mu_d(z, X_i) + e_{i,d,z} \quad \forall z \in \{0, 1\}$ , 综上所述, 这些数据由一个可观测的四维数据组成  $(y_{i,r}, d_{i,r}, z_{i,r}, x_{i,2})$ 。

## (二) 仿真和估计

本文对  $\Delta$  GATE 和  $\Delta$  BGATE 进行仿真和估计，二者的计算公式如下：

$$\theta^{\Delta GATE} = E[E[Y_i^1 - Y_i^0 | Z_i = 1] - E[Y_i^1 - Y_i^0 | Z_i = 0]] \quad (24)$$

$$\theta_{x_{i,0}}^{\Delta BGATE} = E[E[Y_i^1 - Y_i^0 | Z_i = 1, X_{i,0}] - E[Y_i^1 - Y_i^0 | Z_i = 0, X_{i,0}]] \quad (25)$$

$$\theta_{x_{i,2}}^{\Delta BGATE} = E[E[Y_i^1 - Y_i^0 | Z_i = 1, X_{i,2}] - E[Y_i^1 - Y_i^0 | Z_i = 0, X_{i,2}]] \quad (26)$$

其中， $X_{i,0}$  在两个调节组之间不平衡，而  $X_{i,2}$  是平衡的。因此， $\theta_{x_{i,0}}^{\Delta B}$  不同，而  $\theta_{x_{i,2}}^{\Delta B}$  等于  $\theta^{\Delta G}$ ，本文生成了  $N=500$  的仿真样本。

在两个估计步骤中均使用了两折 ( $K=2, J=2$ )。DML 估计策略中的所有扰动函数都是使用随机森林来估计的，调整用于估计扰动参数的学习算法是至关重要的 (Bach 等, 2024)<sup>[17]</sup>。本文在两个估计步骤中再次使用了两折叠 ( $K=2, J=2$ )。该算法是利用 Chernozhukov 等 (2021)<sup>[5]</sup> 提出的神经网络实现的。最后，通过估计效应重新加权数据，使平衡变量的分布与调节变量定义的组相同。重新加权算法平衡数据后，用 DML 估计量来估计  $\Delta$  GATE。

## (三) 仿真结果

表 1 给出了样本量为 500 时的效应和估计量的仿真结果。从不同估计量的表现来看，DML 在几乎所有情况下的均方根误差 (RMSE) 都最小，其次是 Auto-DML。与预期一致，重新加权策略通常会导致标准差 (STD) 略微升高，从而导致 RMSE 升高，它的覆盖率有时较差，原因在于其效应偏差或者标准误差过大。此外，使用神经网络的样本量应该足够大，而这对于 500 个样本量可能不是这样的情况，这或许就是 Auto-DML 的覆盖率对  $\Delta$  GATE 的估计较差的一个原因。最后，尽管 DML 估计量在模拟中表现最好，但这一发现可能不适用于已知 DML 有性能问题的场景，例如当倾向得分很极端时，建议采用 Auto-DML 估计量或重新加权方法。

表 1：不同效应和估计量的仿真结果 ( $N=500$ )

| 效应                        | 估计值         | 偏差                 | 标准差      | 均方根误差    |
|---------------------------|-------------|--------------------|----------|----------|
| $\theta_{x_0}^{\Delta B}$ | DML         | $2.442491e^{-17}$  | 0.166694 | 0.166527 |
| $\theta_{x_0}^{\Delta B}$ | Auto-DML    | $1.154632e^{-17}$  | 0.178052 | 0.177874 |
| $\theta_{x_0}^{\Delta B}$ | Reweighting | $-5.329071e^{-18}$ | 0.168738 | 0.168569 |
| $\theta_{x_2}^{\Delta B}$ | DML         | $5.585810e^{-18}$  | 0.003011 | 0.003008 |
| $\theta_{x_2}^{\Delta B}$ | Auto-DML    | $-4.847226e^{-19}$ | 0.003829 | 0.003825 |
| $\theta_{x_2}^{\Delta B}$ | Reweighting | $-3.157197e^{-18}$ | 0.002978 | 0.002975 |

注：数据由作者自行整理。

接下来，本文比较了不同参数下三种仿真的结果，其中，图 1—图 3 描述了为中的误差的分布  $\hat{\theta}_{x_0}^{\Delta B}$  和  $\hat{\theta}^{\Delta G}$  如果利息的影响是  $\theta_{x_0}^{\Delta B}$ ，则估计  $\hat{\theta}^{\Delta G}$  导致不同的结果，因为变量  $x_{i,0}$  在两组  $Z_i$  中不平衡。图中展示了与标准差为  $\theta_{x_0}^{\Delta G}$  与  $\theta^{\Delta G}$  对比的正态分布的标准差。图 1 显示了 DML 的结果，图 2 显示了 Auto-DML 的结果，图 3 显示了 Reweighting 方法的结果。

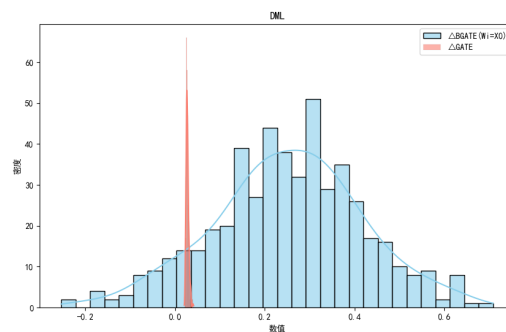


图 1：DML 仿真结果

注：横坐标的数值是指  $(\hat{\theta}^{\Delta G} - \theta^{\Delta G})$  或  $(\hat{\theta}^{\Delta B} - \theta^{\Delta B})$ 。

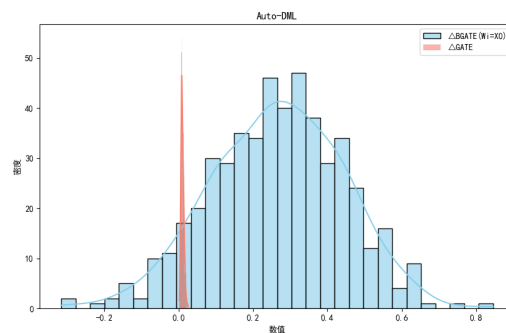


图 2：Auto-DML 仿真结果

注：横坐标的数值是指  $(\hat{\theta}^{\Delta G} - \theta^{\Delta G})$  或  $(\hat{\theta}^{\Delta B} - \theta^{\Delta B})$ 。

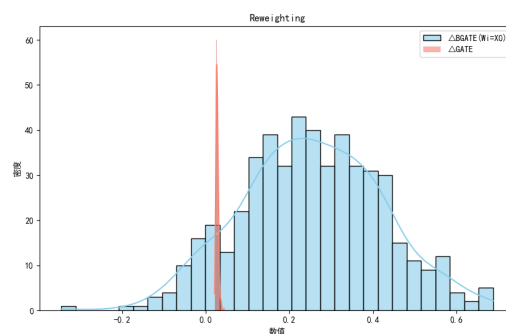


图 3：Reweighting Approach 仿真结果

注：横坐标的数值是指  $(\hat{\theta}^{\Delta G} - \theta^{\Delta G})$  或  $(\hat{\theta}^{\Delta B} - \theta^{\Delta B})$ 。

在图 1—图 3 中，两种效应的重叠分布体现在两种颜色的混合的效果。与预期一致，估计  $\hat{\theta}^{\Delta G}$  得出的结果与估计  $\hat{\theta}_{x_2}^{\Delta B}$  相同，因为  $x_{i,2}$  已经平衡。因此，如果协变量（或多个协变量）不平衡，则区分这两种效应非常重要。

## 五、结论

本文提出了  $\Delta$  BGATE 估计量以衡量不同组间处理效果差异，并考虑协变量差异，该方法适用于无混杂环境的处理效应的异质性。同时，提出 DML 估计量，并证明该估计量在标准条

件下的一致性和渐近正态性，同时提出两种替代估计策略：一是 Auto-DML，可避免 DML 估计量对极端估计倾向分数的非稳健性问题；二是 Reweighting 方法，可结合适用于重新加权数据 GATE 的一致估计量使用。仿真研究验证了这些估计策略的有效性，为理解政策决策的差异化影响提供支持。

未来研究可从多方面拓展：将估计方法拓展至连续处理和调

节场景；探索在处理变量、调节变量或两者均存在工具变量的设置下应用该方法；深入研究 RieszNet 的调整方式以及 Auto-DML 和重新加权估计量的渐近性质；开展更广泛的模拟研究，以全面了解估计量在有限样本条件下的特性。

## 参考文献

- 
- [1] Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters[J]. The Econometrics Journal, 2018, 21(01): C1–C68.
- [2] 钱浩祺, 龚嫒然, 吴力波. 更精确的因果效应识别：基于机器学习的视角[J]. 计量经济学报, 2021, 1(04): 867–891.
- [3] Lechner M, Mareckova J. Comprehensive Causal Machine Learning, arXiv preprint arXiv:2405.2024:10198.
- [4] Frölich M. Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators[J]. The Review of Economics and Statistics, 2004, 86(01): 77–90.
- [5] Chernozhukov V, Newey W, Quintas-Martinez V M, et al. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests[J]. In International Conference on Machine Learning, 2022, : 3901–3914. PMLR.
- [6] 郭峰, 陶旭辉. 机器学习与社会科学中的因果关系：一个文献综述[J]. 经济学(季刊), 2023, 23(01): 1–17.
- [7] Abrevaya J, Hsu Y C, Lieli R P. Estimating Conditional Average Treatment Effects[J]. Journal of Business and Economic Statistics, 2015, 33(4): 00–00. DOI: 10.1080/07350015.2014.975555.
- [8] Lechner M. Modified Causal Forests for Estimating Heterogeneous Causal Effects[J]. Michael Lechner, 2018. DOI: 10.13140/RG.2.2.13540.22405.
- [9] Semenova V, Chernozhukov V. Debiased Machine Learning of Conditional Average Treatment Effects and Other Causal Functions[J]. The Econometrics Journal, 2021, 24(02): 264–289.
- [10] Zimmert M, Lechner M. Nonparametric estimation of causal heterogeneity under high-dimensional confounding[J]. Papers, 2019. DOI: 10.48550/arXiv.1908.08779.
- [11] Fan Q, Hsu Y C, Lieli R P, et al. Estimation of Conditional Average Treatment Effects With High-Dimensional Data[J]. Journal of Business And Economic Statistics, 2022, 40(01): 313–327.
- [12] Farbmacher H, Huber M, Langen H, et al. Causal mediation analysis with double machine learning[J]. The Econometrics Journal, 2022, 25, 277–300.
- [13] Rubin D B. Estimating causal effects of treatments in randomized and nonrandomized studies[J]. Journal of Educational Psychology, 1974, 66, 688.
- [14] Imbens W G, Wooldridge M J. Recent Developments in the Econometrics of Program Evaluation[J]. Journal of Economic Literature, 2009, 47(01): 5–86.
- [15] Imbens W G. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review[J]. The Review of Economics and Statistics, 2004, 86(01): 4–29.
- [16] Belloni A, Chernozhukov V. Least squares after model selection in high-dimensional sparse models[J]. Bernoulli, 2013, 19(02): 521–547.
- [17] Bach P, Schacht O, Chernozhukov V, et al. Hyperparameter tuning for causal inference with double machine learning: A simulation study[J]. In Causal learning and reasoning, 2024: 1065–1117. PMLR.

# 一类自适应权重下的混频逻辑回归模型及其应用研究

呼可可, 刘梦丽

广州大学 经济与统计学院, 广东 广州 510006

DOI:10.61369/ASDS.2026010011

**摘 要 :** 混频数据在经济预测中具有重要价值, 但传统处理方法往往因同频化操作而损失高频信息。本文研究了混合频率数据下的逻辑回归建模问题, 为了提升对高频数据的信息利用率, 我们提出了一种迭代加权优化算法, 模拟结果表明, 优化权重可以显著提高回归系数的估计精度, 降低估计系数的偏差和标准差。实证结果显示, 优化权重对经济衰退风险的识别灵敏度更强。本文为混频数据下的二分类预测提供了新的方法参考。

**关 键 词 :** 混频数据; 逻辑回归模型; 自适应权重; 经济风险评估

## A Type of Mixed-frequency Logistic Regression Model under Adaptive Weights And Its Application

Hu Keke, Liu Mengli

School of Economics and Statistics, Guangzhou University, Guangzhou, Guangdong 510006

**Abstract :** Mixed frequency data has important value in economic forecasting, but traditional processing methods often lose high-frequency information due to co-frequency operation. In order to improve the information utilization of high-frequency data, we propose an iterative weighted optimization algorithm, and the simulation results show that the optimization weight can significantly improve the estimation accuracy of the regression coefficient, and reduce the deviation and standard deviation of the estimation coefficient. The empirical results show that the optimization weight has a stronger sensitivity in identifying recession risks. This paper provides a new method reference for binary prediction under mixed frequency data.

**Keywords :** mixed-frequency data; logistic regression model; adaptive weight; economic risk assessment

### 引言

经济预测中常面临数据频率不一致的问题, 如季度因变量与月度自变量, 传统回归模型基于同频数据研究, 对混频数据需要进行同频化处理, 简单的等权聚合往往会导致大量高频信息损失, 由此带来了关于混频数据的影响模式探讨和精准预测研究。混频数据抽样模型的提出, 为直接使用原始混频变量进行建模分析提供了可能。Logistic 回归是一种成熟的、功能强大的二元结果分类方法, 我们在逻辑回归中引入混频数据模型, 提出了一种基于 Logistic 回归的新的混频数据驱动的加权方法。本文旨在构建一种自适应权重的 Logistic-MIDAS 模型<sup>[6]</sup>, 通过数据驱动方式动态优化权重, 提升模型在经济风险评估中的预测精度。

### 一、模型概述

#### (一) MIDAS 模型

Ghysels 等<sup>[1]</sup>首次提出能够直接对原始混频数据进行建模分析的 MIDAS 模型, 该模型通过参数化的权重多项式, 将高频解释变量直接应用到线性模型的构建、估计与预测。假设  $y_t$  表示第  $t$  期低频被解释变量,  $x_t^{(m)}$  表示高频解释变量, 即在  $t$  至  $t+1$  时间

隔内存在  $m$  个观测值,  $m$  为高频变量与低频变量的频率倍差, 如果  $y_t$  是年度数据,  $x_t^{(m)}$  是月度数据, 则有  $m=12$ ,  $q$  是高频变量的滞后阶数, 则单变量 MIDAS 回归模型的基本形式可以表示为:

$$y_t = \beta_0 + \beta_1 W(L^{1/m}, \theta) x_t^{(m)} + \varepsilon_t$$

其中  $W(L^{1/m}, \theta) = \sum_{k=0}^q w(k, \theta) L^{k/m}$ ,  $w(k, \theta)$  为多项式权重函数  $L^{1/m}$  为高频滞后算子,  $L^{1/m} x_t^{(m)} = x_{t-1/m}^{(m)}$ ,  $\varepsilon_t$  是模型误差项。

作者简介:

呼可可, 广州大学经济与统计学院, 硕士研究生;

刘梦丽, 广州大学经济与统计学院, 硕士研究生。



传统的 MIDAS 回归模型通常依赖于参数化的权重函数，这大大增加了模型的复杂性。为了解决这个问题，我们的研究建立在这些基础上，提出了一个简化的权重函数，直接结合了观测数据的特征。我们采用了一种非参数的、数据驱动的方法来选择权重，这使得混合频率数据的建模更加精确和自适应。因此，我们改进了上述公式中的模型，该模型具有无参数的权重函数，公式如下：

$$y_t = \beta_0 + \beta_1 W(L^{1/m}) x_t^{(m)} + \varepsilon_t$$

## (二) Logistic 模型

逻辑回归模型为一种非线性回归分析模型，采用逻辑函数研究事件发生概率，判断因变量与自变量之间影响关系，在预测二分类问题方面发挥了重要作用<sup>[2]</sup>。

模型预测器部分：

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta'x$$

$\beta = (\beta_0, \beta_1, \dots, \beta_n)'$  为参数向量， $x = (1, x_1, x_2, \dots, x_n)'$  为包含常数项的特征向量。

逻辑函数转换机制（sigmoid 函数）：

$$\pi x = \frac{1}{1 + \exp(-\beta'x)}$$

通过 sigmoid 函数将线性预测器  $\eta$  映射到 (0,1) 区间，得到事件发生的概率。对于输入样本  $(x, y)$ ，利用 Logit 模型可以得到它属于某一类别的概率：

$$p(y|x) = \pi(x)^y (1 - \pi(x))^{1-y}$$

这种概率计算方式通过逻辑函数将线性组合映射到区间，使得输出结果能被解释为事件发生或不发生的概率，从而为分类预测提供基础。

## (三) 多项式权重函数类型

MIDAS 模型的核心思想是通过引入多项式权重函数  $w(k, \theta)$  减少待估参数的数量，而权重函数的设定形式不同会对 MIDAS 模型的结果产生不同的影响。以下介绍六种权重为混频建模中常用的传统权重，将作为后续模拟和实证研究中，与本文提出的自适应权重对比的基准模型<sup>[3]</sup>。

### 1. 等权重

等权重为所有滞后项分配相同的权重，假设各滞后期对当前值的影响程度完全相同。无参数需要估计，权重函数为常数，这是最简单的权重分配方式。权重表达式为：

$$w_i = \frac{1}{m}, i = 1, 2, 3, \dots, m$$

### 2. 指数 Almon 权重

指数 Almon 滞后多项式是多项式分布滞后模型的现代扩展。该方法的核心思想是通过指数变换的多项式函数来构造权重，确保权重非负且灵活可变。指数变换具有双重作用：一是保证权重非负，二是增加函数非线性度以捕捉复杂模式。指数 Almon 权重广泛应用于混频数据采样模型中，用于处理高频解释变量对低频

被解释变量的影响。具体形式可以为：

$$w(k; \theta) = \frac{\exp(\theta_0 + \theta_1 k + \theta_2 k^2 + \cdots + \theta_p k^p)}{\sum_{k=1}^K \exp(\theta_0 + \theta_1 k + \theta_2 k^2 + \cdots + \theta_p k^p)}$$

其中  $p$  为多项式阶数（通常  $p=2$  或  $p=3$ ）， $\theta_0, \theta_1, \dots, \theta_p$  是多项式的参数。

常用二阶形式（ $p=2$ ）：

$$w(\theta_1, \theta_2) = \frac{\exp(\theta_0 + \theta_1 k + \theta_2 k^2)}{\sum_{k=1}^K \exp(\theta_0 + \theta_1 k + \theta_2 k^2)}$$

### 3. Step 权重

Step 权重是一种非参数化或半参数化的权重分配方法，它将整个滞后区间划分为若干个区间，在每个区间内权重为常数。这种权重设定允许滞后效应在不同滞后阶段有不同但恒定的影响，适用于滞后效应呈现阶段性的情况，例如政策变化前后、经济周期不同阶段等。在计量经济学中，Step 权重可以看作分段常数函数的离散近似。具体形式可以表示为：

$$w(i, \theta_1, \theta_2, \theta_3) = \begin{cases} \theta_1, 0 \leq i < k \\ \theta_2, k_1 \leq i < k_2 \\ \theta_3, k_2 \leq i < i^{max} \end{cases}$$

其中  $\theta_1 > \theta_2 > \theta_3 > 0, \sum_{j=1}^3 \theta_j = 1$

### 4. Beta 权重

Beta 权重方法<sup>[4]</sup>是基于标准化 Beta 概率密度函数构造的权重，利用 Beta 分布丰富的形状灵活性来刻画各种滞后模式。Beta 分布定义在 [0,1] 区间上，恰好对应归一化的滞后时间指标。一般在金融市场波动的预测和分析中使用较多，Beta 多项式的具体形式可以表示为：

$$w(k; \theta_1, \theta_2) = \frac{f(k/K, \theta_1, \theta_2)}{\sum_{k=1}^K f(k/K, \theta_1, \theta_2)}$$

其中  $f(x, a, b) = \frac{x^{a-1}(1-x)^{b-1} \Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ ， $\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$

### 5. 非零 Beta 权重

非零 Beta 权重是在标准 Beta 权重的基础上进行修改，以确保所有权重都严格大于零，避免某些滞后项被赋予零权重。这在某些模型中可以提高数值稳定性或满足理论要求。其具体形式与 Beta 相似。

### 6. Legendre 权重

Legendre 权重基于 Legendre 正交多项式来构造权重函数。Legendre 多项式是在区间 [-1,1] 上定义的正交多项式，可以用于逼近任意光滑函数。通过将滞后指标映射到 [-1,1] 区间，并使用 Legendre 多项式的线性组合作为权重函数<sup>[4]</sup>，我们可以获得高度灵活的权重形式。Legendre 权重函数的形式为：

$$w(k; \theta) = \sum_{i=0}^p \theta_i p_i(k)$$

其中  $w(k; \theta)$  是第  $k$  期权重,  $k$  是滞后时间期数,  $p_i(k)$  是第  $i$  阶 Legendre 多项式,  $\theta_0, \theta_1, \dots, \theta_p$  是多项式的参数,  $p$  是多项式阶数, Legendre 多项式的递推公式为:

$$p_0(k) = 1, p_1(k) = k, p_{i+1}(k) = \frac{(2i+1)kp_i(k) - ip_{i-1}(k)}{i+1}$$

## 二、基于自适应权重的 Logistic-MIDAS 模型

基于以上简单逻辑回归模型和 MIDAS 模型的介绍, 我们给出 Logistic-MIDAS 模型<sup>[5]</sup>以及自适应权重迭代优化算法。设响应变量  $y_t$  为二值变量, 观测频率较低, 协变量  $X_t^{(m)}$  观测频率较高, 其中  $m$  表示高频期数。

Logistic-MIDAS 基本形式:

$$p(y_t = 1|x) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 W(L^{1/m})x_t^{(m)}))}$$

线性预测项:  $\eta_t = \beta_0 + \beta_1 W(L^{1/m})x_t^{(m)}$ , 其中  $\beta_0$  为截距项,  $\beta_1$  为高频协变量的系数,  $W(L^{1/m})$  为权重多项式,  $L^{1/m}$  为滞后算子。

逻辑变换后的概率预测:

$$\pi_t = \frac{\exp(\eta_t)}{1 + \exp(\eta_t)} = \sigma(\eta_t), \text{ 其中 } \sigma(z) = \frac{1}{1 + e^{-z}} \text{ 是逻辑函数。}$$

$$\pi x = \frac{1}{1 + \exp(-\beta'X)}, \quad p(y|x) = \pi(x)^y (1 - \pi(x))^{1-y}$$

对于样本  $(X_i, y_i)$ , 其似然函数可以表示为:

$$L(\beta) = \prod_{i=1}^m \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

对数似然函数:

$$l(\beta) = \log L(\beta) = \sum_{i=1}^m y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i))$$

给出设计矩阵

$$X_t^* = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{1,t-\frac{1}{m}} & x_{1,t-\frac{2}{m}} & \cdots & x_{1,t-\frac{q}{m}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p,t-\frac{1}{m}} & x_{p,t-\frac{2}{m}} & \cdots & x_{p,t-\frac{q}{m}} \end{pmatrix}, \quad w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_q \end{pmatrix}, \quad X_t = X_t^* w = \begin{pmatrix} 1 \\ x_{1t} \\ \vdots \\ x_{pt} \end{pmatrix}$$

最小化目标函数:

$$-l(\beta) = -\sum_{i=1}^m y_i \ln \pi(X_i) + (1 - y_i) \ln(1 - \pi(X_i))$$

权重需要满足约束条件  $\sum_{i=1}^k w_i = 1$  和  $w_i \geq 0$

该模型的参数估计为带约束的优化问题, 权重需要满足约束

条件  $\sum_{i=1}^k w_i = 1$  和  $w_i \geq 0$ , 通过交替迭代最小化负对数似然函数求解, 具体流程如下:

i: 初始化: 给定一个初始权重  $w^j$ , 满足  $\sum_{i=1}^k w_i = 1$  和  $w_i \geq 0$  和初始系数  $\beta^j$ ;

ii: 交替迭代:

步骤 1: 固定当前系数  $\beta^j$ , 通过梯度下降算法最小化负对数似然函数, 得到优化权重  $w^{j+1}$ ;

步骤 2: 固定新权重  $w^{j+1}$ , 最小化逻辑回归负对数似然函数,

得到更新值  $\beta^{j+1}$

iii: 收敛判断: 若  $||\beta^{j+1} - \beta^j|| < \varepsilon$  且  $||w^{j+1} - w^j|| < \varepsilon$  ( $\varepsilon$  为某个预设阈值), 迭代终止, 输出最优参数  $\beta^{new}$  和权重  $w^{new}$ 。

## 三、模型评估指标

### (一) 偏差和标准差

偏差衡量模型估计量的系统性偏离程度, 它可以帮助我们判断估计量是否存在系统性偏差; 标准差衡量预测概率的离散程度或波动性, 体现了估计量的变异性, 较小的样本标准差意味着更稳定的估计性能。

### (二) 均方误差

均方误差是衡量预测概率与真实结果之间差异的综合指标, 同时考虑了偏差和方差的影响, 在比较不同模型时, 较低的 MSE 通常意味着更好的整体预测性能。

### (三) AUC(Area Under the ROC Curve)

混淆矩阵

| 预测值 \ 真实值     | $y = 0$ | $y = 1$ |
|---------------|---------|---------|
| $\hat{y} = 0$ | TN      | FP      |
| $\hat{y} = 1$ | FN      | TP      |

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}$$

真阳性率 (True Positive Rate, TPR) 与假阳性率 (False Positive Rate, FPR)

ROC 曲线是通过变化分类阈值以 FPR 为横坐标、TPR 为纵坐标绘制出的曲线 AUC 则是 ROC 曲线下的面积。

### (四) 风险组 (Risk Group, RG)

风险组是指在预测二分类问题时, 按照预测概率从高到低排序后, 取前  $\alpha$  比例的观测样本构成的群体。RG<sup>[6]</sup>是一种针对性评估指标, 特别适用于类别不平衡的场景。在预测“失业率对 GDP 的负面冲击”的研究中, 风险组是指根据预测模型给出的概率值, 从所有经济观测样本中筛选出预测概率最高的前  $\alpha$  比例的观测单元。这些单元被认为最可能在未来特定时期内出现“高失业率伴随 GDP 显著下降”的复合负面经济事件。 $RG_\alpha$  越高, 表明模型识别极端风险事件的灵敏度越强。若  $RG_{10\%} = 0.8$  则意味着模型仅通过关注其判定的风险最高的前 10% 的时期, 就能捕捉到历史上 80% 的经济衰退。  $RG_\alpha = \frac{\text{风险组中实际发生负面事件的观测数字}}{\text{总负面事件观测书数}}$ 。

### (五) 风险组下面积 (Area Under Risk Group, AURG)

是评估模型在风险组内部对负面经济事件排序质量的指标。它通过绘制“风险组规模扩大过程中捕捉到的负面事件比例”曲线, 并计算该曲线下面积占最大可能面积的比例, 来反映模型在高风险群体中的分辨能力。计算方法如下:

从最小的风险组规模开始, 逐步扩大至目标  $\alpha$  比例。在每个规模点, 计算当前风险组中实际发生负面事件的样本数占总负面事件样本数的比例 (即该规模下的 RG 值)。以“风险组规模 (横轴, 从 0 到  $\alpha\%$ )”和“捕捉到的负面事件比例 (纵轴, 从 0 到

100%)”绘制曲线。计算该曲线下的面积  $S_{shaded}$ ，将  $S_{shaded}$  除以整个矩形面积（即  $\alpha \times 100\%$ ），得到  $AURG_{\alpha} = \frac{S_{shaded}}{\alpha \times 100\%}$ 。

四、模拟

为了阐述自适应权重相较于其他权重在提取有效信息方面的优越性，本文构建了 Logistic-MIDAS 模型，并运用不同权重函数开展对比分析。我们采用第 2 章介绍的等权重、指数 Almon 权重、Beta 权重、Step 权重、非零 Beta 权重、Legendre 权重作为权重滞后多项式，将这六种权重以及我们构建的自适应权重函数代入到 Logistic-MIDAS 模型中去进行比较分析。蒙特卡罗模拟<sup>[6]</sup>设计基于 Logistic-MIDAS 回归模型的数据生成过程。

我们假设高频单变量遵循以下 ARCH(1,1) 过程：

$$x_{t/m}^{(m)} = \sqrt{\left(\sigma_{t/m}^{(m)}\right)^2} e_{t/m}^{(m)}, \left(\sigma_{t/m}^{(m)}\right)^2 = \alpha_0 + \alpha_1 \left(x_{t-1/m}^{(m)}\right)^2$$
$$\alpha_0 = 0.25, \alpha_1 = 0.85, e_{t/m} \sim N.i.i.d.(0,1)$$

在模型中，我们设定  $\beta_0 = 2, \beta_1 = 4, m = 5$ ，考虑样本量大小为 400，800，1000 的情形，给出六种不同权重和自适应权重的模拟结果。 $w_{equ}, w_{exp}, w_{beta}, w_{step}, w_{non-zero}, w_{legendre}$ ，分别代表以上六种权重， $w_{opt}$  代表我们提出的自适应优化权重。相应的系数估计分别表示为  $\hat{\beta}_{equ}, \hat{\beta}_{exp}, \hat{\beta}_{beta}, \hat{\beta}_{step}, \hat{\beta}_{non-zero}, \hat{\beta}_{legendre}, \hat{\beta}_{opt}$ ，计算出系数估计值的偏差和标准差来评估他们的有限样本性能，同时为了评估相对效率<sup>[7]</sup>，我们给出六种不同权重基于自适应权重的均方误差比，结果见表 1。

表 1：偏差、标准差、均方误差比

| n    | BS1          | BS2          | BS3           | BS4           | BS5               | BS6               | BS7    |
|------|--------------|--------------|---------------|---------------|-------------------|-------------------|--------|
| 400  | -0.5835      | -0.3713      | -0.5835       | -0.5594       | -0.5654           | -0.3738           | 0.1316 |
| 800  | -0.6305      | -0.4367      | -0.6305       | -0.6271       | -0.6205           | -0.4515           | 0.0343 |
| 1200 | -0.6400      | -0.4518      | -0.6400       | -0.6364       | -0.6236           | -0.4686           | 0.0176 |
| n    | SD1          | SD2          | SD3           | SD4           | SD5               | SD6               | SD7    |
| 400  | 0.4652       | 0.5118       | 0.4652        | 0.5981        | 0.5989            | 0.4965            | 0.5544 |
| 800  | 0.3328       | 0.3771       | 0.3328        | 0.4889        | 0.4834            | 0.3703            | 0.3933 |
| 1200 | 0.2623       | 0.2999       | 0.2623        | 0.4128        | 0.3981            | 0.2946            | 0.3215 |
| n    | MSEequ / opt | MSEexp / opt | MSEbeta / opt | MSEstep / opt | MSEnon-zero / opt | MSElegendre / opt |        |
| 400  | 1.7163       | 1.2316       | 1.7162        | 2.0668        | 2.0904            | 1.1901            |        |
| 800  | 3.2640       | 2.1376       | 3.2640        | 4.0594        | 3.9720            | 2.1891            |        |
| 1200 | 4.6190       | 2.8380       | 4.6190        | 5.5535        | 5.2831            | 2.9570            |        |

从以上数据可以看出，随着 n 值增加，偏差标准差在逐渐减小；自适应加权方法的偏差和标准差始终小于等权重（ $w_{equ}$ ）、指数权重（ $w_{exp}$ ）等六种不同权重方法，且均方误差比值都显著大于 1，证明了该方法在精度和偏差方面的优越性。

此外为进一步证明该权重在分类性能上的优化效果，我们给出更多评估指标，以下使用 AUC，RG 作为评估指标并给出模拟结果，如表 2 所示：

表 2：AUC、RG 评估

| n    | AUC1   | AUC2   | AUC3   | AUC4   | AUC5   | AUC6   | AUC7   |
|------|--------|--------|--------|--------|--------|--------|--------|
| 400  | 0.8286 | 0.8347 | 0.8347 | 0.8301 | 0.8288 | 0.8346 | 0.8635 |
| 800  | 0.8284 | 0.8329 | 0.8329 | 0.8280 | 0.8275 | 0.8330 | 0.8612 |
| 1200 | 0.8285 | 0.8326 | 0.8326 | 0.8276 | 0.8274 | 0.8327 | 0.8609 |
| n    | RG1    | RG2    | RG3    | RG4    | RG5    | RG6    | RG7    |
| 400  | 0.0632 | 0.0633 | 0.0632 | 0.0633 | 0.0633 | 0.0633 | 0.0635 |
| 800  | 0.0632 | 0.0633 | 0.0632 | 0.0633 | 0.0633 | 0.0632 | 0.0635 |
| 1200 | 0.0631 | 0.0632 | 0.0631 | 0.0632 | 0.0632 | 0.0632 | 0.0635 |

五、实证研究

本研究使用了两种类型的宏观经济数据<sup>[8]</sup>：被解释变量为二值变量<sup>[9]</sup>，即低频季度 GDP 衰退指标，核心解释变量为高频月度失业率。选择季度 GDP 衰退指标作为低频被解释变量（二值变量：1= 衰退，0= 非衰退），符合 Logistic 模型的二分类预测场景；选择月度失业率作为高频解释变量，因为失业率与经济增长呈显著负相关，即失业率上升往往预示经济下行，且高频数据能及时捕捉经济波动信号，为 GDP 衰退预测提供时效性支撑。数据从 1970 年 1 月 1 日到 2020 年 12 月 1 日，涵盖了美国多个经济周期，

数据来源于美联储经济在线数据（FRED）。由于高频数据和低频数据的频率不匹配，传统的回归方法无法直接应用<sup>[10]</sup>。为充分利用高频失业率数据中的信息，以及对二值变量进行处理，本研究采用 Logistic-MIDAS 方法构建预测因子，使用六种不同权重与自适应权重进行对比，结果如表 3：

表 3：实证结果

| Weight    | Metrics    | Value  |
|-----------|------------|--------|
| $w_{equ}$ | $RG_{0.1}$ | 0.1000 |
|           | AUC        | 0.6943 |

|                |            |        |
|----------------|------------|--------|
| $W_{exp}$      | $RG_{0.1}$ | 0.2000 |
|                | AUC        | 0.7299 |
| $W_{beta}$     | $RG_{0.1}$ | 0.3000 |
|                | AUC        | 0.7528 |
| $W_{step}$     | $RG_{0.1}$ | 0.3000 |
|                | AUC        | 0.7361 |
| $W_{non-zero}$ | $RG_{0.1}$ | 0.3000 |
|                | AUC        | 0.7356 |
| $W_{legendre}$ | $RG_{0.1}$ | 0.3000 |
|                | AUC        | 0.6722 |
| $W_{opt}$      | $RG_{0.1}$ | 0.4000 |
|                | AUC        | 0.7809 |

六、结论

针对传统混频数据建模中高频信息损失、参数化权重复杂的问题，本文将逻辑回归模型与混频数据相结合，提出了带自适应权重的 Logistic-MIDAS 模型。通过蒙特卡罗模拟对比6种传统权重与自适应权重的性能，发现自适应权重能显著降低估计偏差和标准差，且 MSE 比值均大于1，验证了其精度优势；基于美国 1970–2020 年宏观经济数据的实证研究表明，该模型在 AUC(0.7809) 和  $RG_{0.1}$ (0.4000) 指标上优于传统权重模型，能更精准识别经济衰退风险。本文提出的模型为经济风险评估提供了新的有效工具，可为政策制定者提前部署干预措施提供参考。

参考文献

[1]Andreou, E., Ghysels, E. Regression models with mixed sampling frequencies[J]. Journal of Econometrics, 2010.

[2] 江丽. 基于日内高频数据的 Logit 回归模型及其应用研究——以螺纹钢期货为例 [D]. 广州大学, 2025.

[3] 刘莹. 中国经济增长的高维混频短期预测与精度提升机理研究 [D]. 吉林大学, 2023.

[4] 许敏. 基于自适应权重函数的 MIDAS-GARCH 模型及其应用研究 [D]. 广州大学, 2025.

[5]Audrino, F., Kostrov, A., and Kostrov, J. Predicting U.S. Bank Failures with MIDAS Logit Models[J]. Journal of Financial and Quantitative Analysis, 2019.

[6] 祝子逸, 朱敏, 杨爱军等. 基于稀疏组惩罚混频数据抽样模型的中国宏观经济预测 [J]. 数理统计与管理, 2024.

[7] 卓杏轩. 高维非线性混频数据模型及应用研究 [D]. 合肥工业大学, 2019.

[8] 于洋. 混频数据回归模型的建模理论、分析技术研究 [D]. 东北财经大学, 2016.

[9] 刘汉. 中国宏观经济混频数据模型的研究与应用 [D]. 吉林大学, 2013.

[10]Ghysels, E. and Qian, H. Estimating MIDAS regressions via OLS with polynomial parameter profiling[J]. Econometrics and Statistics, 2019.



# 基于多模型的胎儿染色体异常判定的探究

魏婧, 刘昱贤, 皮哲豪, 李乃医\*

广东海洋大学 数学与计算机学院, 广东 湛江 524000

DOI:10.61369/ASDS.2026010013

**摘 要 :** 本研究针对女胎染色体异常判定中模型易受性别因素干扰、泛化性能不足的问题, 构建多模型分析框架。首先对比 LDA、QDA、GNB 模型的决策边界判定效果。然后分析 DNN、改进型 MLP 及融合注意力机制的 Att-MLP 模型的阈值判定结果。最后创新性引入神经网络迁移学习方法, 将非线性注意力机制与迁移学习结合。研究表明, LDA 与 Att-MLP 模型准确率均超 90%, 且 Att-MLP 模型特异度达 97.25%, 有效避免了误判问题。迁移学习虽效果有限, 但验证了性别间异常特征的可迁移潜力, 为后续研究提供了方向。

**关 键 词 :** 线性判别分析; 全连接神经网络; 改进多层感知机; 注意力机制; 迁移学习

## Research on the Determination of Fetal Chromosomal Abnormalities Based on Multiple Models

Wei Jing, Liu Yuxian, Pi Zhehao, Li Naiyi\*

School of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang, Guangdong 524000

**Abstract :** This study addresses the issues of gender factor interference and insufficient generalization performance in the determination of chromosomal abnormalities in female fetuses by constructing a multi-model analysis framework. Firstly, the decision boundary determination effects of LDA, QDA, and GNB models are compared. Then, the threshold determination results of DNN, improved MLP, and the attention mechanism integrated Att-MLP model are analyzed. Finally, the neural network transfer learning method is innovatively introduced, combining the nonlinear attention mechanism with transfer learning. The research shows that the accuracy rates of both LDA and Att-MLP models exceed 90%, and the specificity of the Att-MLP model reaches 97.25%, effectively avoiding misjudgment problems. Although the effect of transfer learning is limited, it verifies the transfer potential of abnormal features between genders, providing a direction for subsequent research.

**Keywords :** linear discriminant analysis; fully connected neural network; improved multi-layer perceptron; attention mechanism; transfer learning

## 引言

目前, 无创产前检测 (NIPT) 已广泛应用于胎儿染色体异常早期筛选方面<sup>[1-3]</sup>。然而, 女胎染色体异常的判定因异常样本稀缺、特征维度与男胎存在显著差异等原因, 仍面临数据不足、判定精度欠佳等问题。传统统计方法常依赖线性假设, 难以捕捉女胎样本中相关指标间复杂的非线性特征关系。近年来, 机器学习技术在医学领域展现出强大潜力<sup>[4]</sup>, 注意力机制的融合使模型能够聚焦关键特征。然而, 将线性判别分析与迁移学习结合的研究尚不多见<sup>[5]</sup>, 且女胎异常判定问题缺乏系统性方法探究。

为此, 本文针对女胎异常样本稀缺的痛点, 通过多模型比较、融合注意力机制的新模型构建、迁移学习探索及提出“初筛 + 复核”分级决策策略, 为女胎异常判定提供一种高精度、高稳健性的方法, 并尝试突破女胎数据稀缺对检测性能的制约问题。

基金项目: 2022 年度国家自然科学基金项目“基于广义相依删失数据小波估计的构建与统计推断”(12161075); 2024 年广东省自然科学基金项目“排序集抽样下相依删失数据经验似然推断”(2024A1515011258)。

作者简介:

魏婧, 广东海洋大学数学与计算机学院, 本科生;

刘昱贤, 广东海洋大学数学与计算机学院, 本科生;

皮哲豪, 广东海洋大学数学与计算机学院, 本科生。

通讯作者: 李乃医, 广东海洋大学数学与计算机学院, 博士, 教授, 硕士生导师, 研究方向为数理统计及其应用。

一、相关研究

以往关于女胎染色体异常的判定研究，多集中于判定模型的开发，强调了染色体异常判定对影响染色体疾病的筛查与早期诊断的重要性。早期研究通过传统线性判别模型进行判定，但该模型往往难以捕捉多维数据之间存在的复杂非线性关系。

线性判别分析模型与神经网络迁移学习的结合是一项重要突破，可充分发挥线性判别分析模型的强可解释性与神经网络捕捉复杂特征的能力，通过跨性别数据的迁移学习方法，有效弥补女胎数据较少、样本特征学习不充分等不足，从而提高预测精度。

二、数据预处理

本研究的实证基础是2025年全国大学生数学建模竞赛公开数据集。该数据集包含的男女胎数据样本均有27个不同的特征属性，本文系统地对数据进行了数据预处理工作。

缺失值处理：孕妇 BMI 有缺失情况，通过身高与体重补全缺失的 BMI 数值。

异常值处理：为防止测试质量出现问题，删除 GC 含量远低于40% 的数据。

数据增强采用“基于噪声添加的自定义数据增强 +RandomOverSampler 过采样处理 + 样本权重调整”的组合策略：少数类增强倍数设为1，过采样处理增加少数类样本数量，初步平衡类别分布；自动平衡样本权重强化模型对异常类样本的关注。

数据划分采用模型差异化策略：传统判别模型按8:2比例随机分层划分训练集与测试集。神经网络相关模型按7:2:1比例随机分层划分训练集、测试集与验证集。

三、建模与求解

（一）传统判别模型

本文通过比较线性判别分析、二次判别分析、高斯朴素贝叶斯三种模型的性能指标（如表1）<sup>[6]</sup>，发现 LDA 模型表现更优，选择该模型作为后续分析的基准模型。

| 表1：不同模型对比表   |          |        |
|--------------|----------|--------|
| 模型           | 交叉验证平均得分 | 测试集准确率 |
| 线性判别分析（LDA）  | 0.9120   | 0.9091 |
| 二次判别分析（QDA）  | 0.9021   | 0.8843 |
| 高斯朴素贝叶斯（GNB） | 0.8821   | 0.8926 |

由表2可知，LDA 模型训练集和测试集的准确率差距极小，均在0.9以上，说明模型泛化能力稳定，且训练集和测试集的AUC 值均高于0.8，说明模型分类效果好。

| 表2：线性判别分析评估指标 |        |        |
|---------------|--------|--------|
| 评估指标          | 训练集    | 测试集    |
| 准确率           | 0.9170 | 0.9091 |
| 精确率           | 0.7895 | 0.8333 |
| F1 分数         | 0.4286 | 0.4762 |
| AUC-ROC       | 0.8113 | 0.8277 |

为验证模型的正确性，图1和图2分别从可视化和量化角度展示了 LDA 模型的判定表现。由图1、图2可知模型预测准确率为0.9154，特异度为532/537=99.07%，说明模型整体预测效果较好。

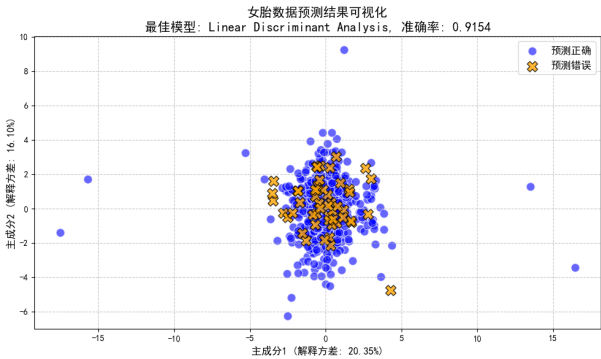


图1：女胎数据预测结果可视化

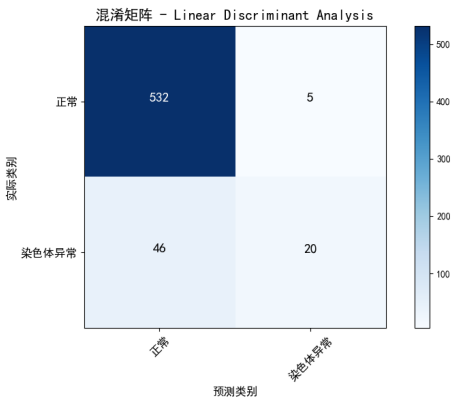


图2：女胎数据预测结果的混淆矩阵热图

（二）神经网络模型

针对染色体异常的判定问题，本文还建立了神经网络模型，分别为全连接神经网络模型、针对性改进多层感知机与融合注意力机制的针对性改进多层感知机。

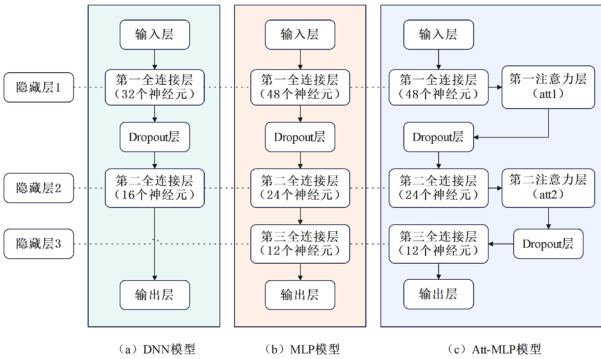


图3：三种神经网络模型结构对比图

1. 全连接神经网络（DNN）模型

采用 Sequential 线性堆叠结构，包含输入层、2个全连接层（均使用 ReLU 激活函数）、1个 Dropout 层与输出层。该模型结合 L2 正则化抑制权重过大、Droupout 层随机失活 20% 节点避免过拟合。DNN 模型阈值为 0.3680。

| 表3：全连接神经网络（DNN）模型评估指标 |        |        |        |        |
|-----------------------|--------|--------|--------|--------|
|                       | 准确率    | 灵敏度    | 特异度    | F1 分数  |
| DNN 模型                | 0.9007 | 0.6250 | 0.9817 | 0.7407 |

2. 针对性改进多层感知机（MLP）

针对DNN模型的不足，MLP模型采用Sequential顺序结构，包含输入层、3个全连接层、1个Dropout层与输出层。全连接层激活函数由ReLU替换为GeLu，以解决ReLU激活函数的“神经死亡”问题与“梯度消失”风险。MLP模型阈值为0.3911。

表4：针对性改进多层感知机（MLP）评估指标

|       | 准确率    | 灵敏度    | 特异度    | F1分数   |
|-------|--------|--------|--------|--------|
| MLP模型 | 0.9078 | 0.6562 | 0.9817 | 0.7636 |

3. 融合注意力机制的针对性改进多层感知机（Att-MLP）

注意力机制源于生物视觉系统，在视觉上减少对冗余信息的关注，以聚焦更多的关键信息。通过分配特征权重将关键数据分配更多的权重，以获取更多有效信息<sup>[7]</sup>。

针对MLP模型的不足，Att-MLP模型采用“全连接层+残差式注意力”的组合架构，包含输入层、3个全连接层、2个残差注意力层、2个Dropout层与输出层。

该模型核心点在于双层注意力机制。att1、att2分别作用于浅层原始特征与深层抽象特征。“特征重要性评估+残差特征增强”的残差式设计，模拟了临床医生先关注核心病理指标、后综合多指标异常关联规律的诊断逻辑，提升了对染色体异常判定的准确性与临床适配性。Att-MLP模型阈值为0.3478。

表5：融合注意力机制的针对性改进多层感知机（Att-MLP）评估指标

|           | 准确率    | 灵敏度    | 特异度    | F1分数   |
|-----------|--------|--------|--------|--------|
| Att-MLP模型 | 0.9128 | 0.7500 | 0.9725 | 0.8219 |

表6：Att-MLP 特征权重分配表

| 指标            | 特征权重   |
|---------------|--------|
| 13号染色体的 Z 值   | 0.0162 |
| 18号染色体的 Z 值   | 0.0176 |
| 21号染色体的 Z 值   | 0.0176 |
| X 染色体的 Z 值    | 0.0213 |
| .....         | .....  |
| 13号染色体的 GC 含量 | 0.0166 |
| 18号染色体的 GC 含量 | 0.0236 |
| 21号染色体的 GC 含量 | 0.0182 |
| X 染色体浓度       | 0.0220 |

对比三大模型评估指标结果（表3、表4、表5），DNN模型识别异常样本能力较弱；MLP模型较DNN模型能更有效地捕捉异常特征，灵敏度有所提高；Att-MLP模型引入注意力机制，使模型自动聚焦关键特征，灵敏度显著高于DNN、MLP模型。

其中，Att-MLP模型的准确率达0.9128，特异度为0.9725，表明该模型在识别正常样本方面具有高可靠性，可有效减少怀正常胎儿的孕妇的检查流程。且该模型为多个关键特征赋予了合理权重，既体现了模型权重的分配符合医学中重点关注的染色体非整倍体类型，又验证了模型构建的科学性与合理性。

（三）神经网络模型的迁移学习方法

目前，我们已经构建了LDA、Att-MLP两大模型，均展现出优异的分类效能，考虑到“染色体非整倍体”的异常特征可能存在“跨性别共性”，且女胎数据少，本文提出采用迁移学习方法，即利用男胎异常判定知识提升女胎异常判定性能的方法。

为验证“跨性别共性”的合理性与“迁移学习方法”的可行性，本文将LDA模型运用在男胎数据上，将预测值与实际值进行比较，同样从可视化和量化角度展示了LDA模型的判定表

现。由图4、图5可知，该模型预测准确率为0.9122，特异度为948/956=99.16%，说明LDA模型在男胎数据上预测效果同样较好。

结果表明“跨性别共性”具有一定的合理性，所以选择使用结合了双模型优势的神经网络迁移学习方法。迁移学习精简了上述Att-MLP模型，通过对模型结构的优化来提高模型的泛化能力和运行效率，从而更适合处理目标数据少的小样本二分类。

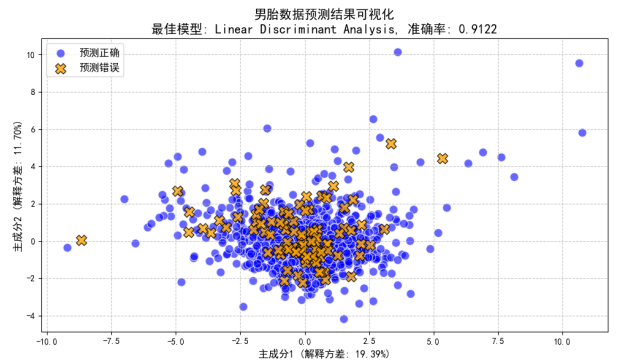


图4：男胎数据预测结果

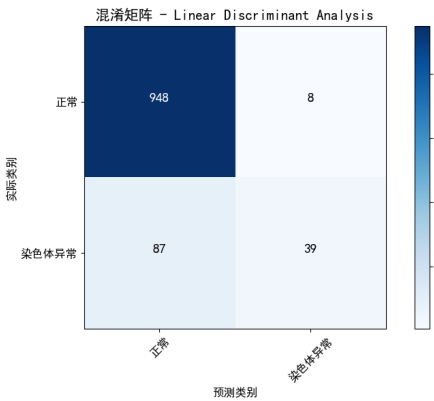


图5：男胎数据预测结果的混淆矩阵热图

本文设计先利用男胎数据训练模型。

表7：迁移学习特征权重分配表

| 指标            | 特征权重   |
|---------------|--------|
| 13号染色体的 Z 值   | 0.0416 |
| 18号染色体的 Z 值   | 0.0418 |
| 21号染色体的 Z 值   | 0.0307 |
| X 染色体的 Z 值    | 0.0226 |
| .....         | .....  |
| 13号染色体的 GC 含量 | 0.0259 |
| 18号染色体的 GC 含量 | 0.0264 |
| 21号染色体的 GC 含量 | 0.0380 |
| X 染色体浓度       | 0.0278 |

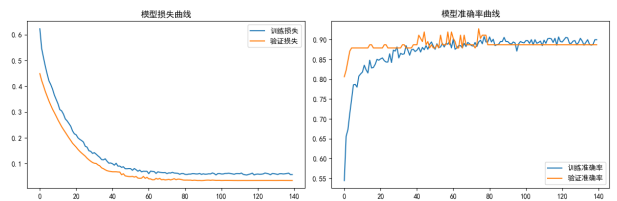


图6：染色体异常检测模型训练损失与准确率曲线

表7中,该模型对核心染色体(13号、18号、21号)的Z值等与染色体异常密切相关的特征赋予较高权重,符合医疗逻辑。图6是模型在训练阶段的监控指标,用于评估模型在男胎数据上的学习效果。模型在整个训练过程中表现出良好的收敛特性。且验证集曲线始终与训练集保持相似趋势,表明模型学习过程稳定,泛化能力可靠。

综上所述,基于男胎数据进行训练的迁移学习确实学到了染色体异常的核心特征等可复用的有效知识。本研究针对迁移学习模型的核心参数进行了针对性微调后<sup>[8]</sup>再将该模型运用于女胎数据实现异常判定,以提升预训练模型在女胎数据上的适应性。

由图7可直观看出模型预测特异度为339/537=63.13%,灵敏度为56.06%,漏诊率较高,风险较大。对比LDA模型,迁移学习准确率也较低,仅62.35%。

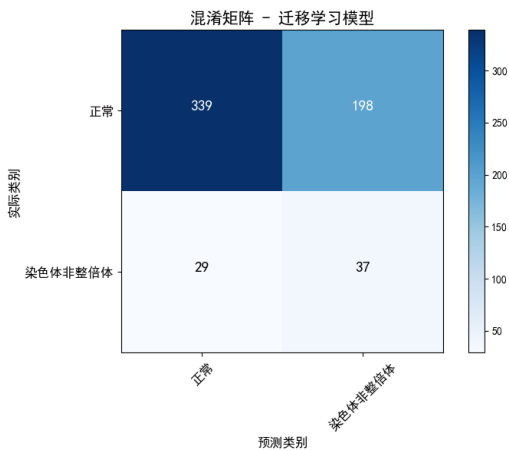


图7: 迁移学习女胎数据预测结果的混淆矩阵

针对该问题,本文尝试通过优化迁移策略提升模型性能,采用渐进式两阶段迁移学习策略,第一阶段冻结特征提取器,仅对分类器层进行重新训练,第二阶段解冻特征提取器,实现目标域特征的适配。结果表明,模型准确率提升至71.48%,特异度达77.65%,但灵敏度下降至21.21%,存在显著类别偏置性。核心问题与具体改进方向为:

数据分布差异:男胎与女胎的染色体特征仍存在本质差异,迁移学习模型未能完全适配女胎数据特性。可引入领域自适应网络,通过对抗训练<sup>[9]</sup>实现特征空间分布对齐,增强模型跨性别泛

化能力。

迁移学习策略不足:未对模型进行完善的训练结构优化与参数微调,导致模型在小数据集场景下收敛速度慢。可采用渐进式分层微调策略,通过分层解冻参数,使模型学习到目标域的特异性特征。

样本类别不平衡:女胎数据中异常样本仅占10.95%,类别失衡问题突出,导致模型无法聚焦于异常样本的特征提取。可采用生成对抗网络,使用生成合理的异常样本,再通过判别器区分真实数据与生成数据。

四、结论与讨论

本文通过多模型探究,发现LDA、Att-MLP模型体现了模型特性的差异,McNemar检验p值为0.0023,远小于显著性水平 $\alpha=0.05$ ,说明两模型间的性能差异显著。

表8: LDA与Att-MLP模型性能与统计检验结果

| 模型      | 准确率    | 灵敏度    | 特异度    | 检验统计量  | p 值    |
|---------|--------|--------|--------|--------|--------|
| LDA     | 0.9154 | 0.3030 | 0.9907 | 9.3333 | 0.0023 |
| Att-MLP | 0.9128 | 0.7500 | 0.9725 |        |        |

数据显示,LDA模型特异度高、可解释性强,适用于高特异度需求、误判容忍度低的临床初筛场景;Att-MLP模型灵敏度与综合性能更优,适用于高灵敏度需求的复核场景。针对构建胎儿异常判定的临床辅助系统,本文提出“初筛+复核”的分级决策策略:通过LDA模型快速扫描样本实现初筛,再通过Att-MLP模型处理可疑样本。该策略需应对计算复杂度相对较高、难以快速获得临床医生的信任与采纳等核心问题。

因此,我们提出以下建议:在模型架构方面,可引入轻量型Transformer变体<sup>[10]</sup>,降低临床设备的算力需求。在数据增强方面,可采用SMOTE-ENN混合采样法,实现合成样本的生成、噪声和冗余样本的剔除,以平衡类别分布。在验证体系拓展方面,可收集不同医院数据,验证“初筛+复核”策略的稳定性与模型的适配性。

为应对女胎数据稀缺所带来的建模挑战,本研究首次将非线性注意力机制与迁移学习结合,未来应持续探索迁移学习等跨性别别共性挖掘方法。

参考文献

[1] 蒋丽雅, 卢劭侃, 杜佳恩, 等. 无创产前检测技术的发展与应用 [J]. 临床医学研究与实践, 2025, 10(23): 191-194. DOI: 10.19347/j.cnki.2096-1413.202523047.

[2] 于丹丹, 李奉瑾, 姚欣雨, 等. NIPT在不同慎用人群胎儿常见染色体非整倍体异常检测中的效能分析 [J]. 中国计划生育和妇产科, 2025, 17(03): 72-76+82.

[3] 鞠爱萍, 孟祥荣, 覃燕龄, 等. 无创产前检测在筛查胎儿染色体拷贝数变异中的应用价值 [J]. 实用心电与临床诊疗, 2025, 34(05): 665-671. DOI: 10.13308/j.issn.2097-5716.2025.05.008.

[4] Alonso E, Beristain A, Burgos J, et al. Comparison of Machine Learning Algorithms to Predict Down Syndrome During the Screening of the First Trimester of Pregnancy [J]. Applied Sciences (2076-3417), 2025, 15(10). DOI: 10.3390/app15105401.

[5] 车志勇. 基于线性判别分析(FDA)的迁移学习方法 [D]. 广东工业大学, 2019. DOI: 10.27029/d.cnki.ggdgu.2019.000284.

[6] 林宏瞻. 基于朴素贝叶斯、线性判别、二次判别分类算法的选股实证研究 [D]. 山东大学, 2018.

[7] 赵文海. 基于自注意力机制和互通双流MLP的点击率预估模型研究 [D]. 电子科技大学, 2025. DOI: 10.27005/d.cnki.gdzku.2025.004745.

[8] 王柯惟. 基于深度学习的染色体异常检测研究 [D]. 中南大学, 2023. DOI: 10.27661/d.cnki.gzhnu.2023.001790.

[9] 雷雨佳. 基于卷积神经网络的染色体异常检测 [D]. 湖南师范大学, 2020. DOI: 10.27137/d.cnki.ghusu.2020.001208.

[10] 赵秋博. 基于Transformer特征关联融合小目标检测算法研究 [D]. 西安电子科技大学, 2024. DOI: 10.27389/d.cnki.gxadu.2024.002573.



# 基于 Kruskal-Wallis 算法的卫生巾市场消费现状与需求分析

朱金桂, 高坚国, 陈心语, 詹富婕, 刘芸, 周大镛, 林培琪

惠州学院 数学与统计学院, 广东 惠州 516007

DOI:10.61369/ASDS.2026010014

**摘要：** 在当前居民收入水平不断提升与健康消费理念持续升级的背景下，结合《“健康中国2030”规划纲要》的政策引导，本文聚焦于卫生巾市场在趋于饱和的背景下仍频现质量安全问题的现实，围绕行业实现高质量发展这一目标展开探讨。通过实施问卷调查与实地走访，研究构建了多维度分析体系，采用多元非线性回归、Borda 排序法、K-means 聚类及随机森林模型进行综合解析。研究结果显示，卫生巾市场已进入稳定阶段，增长动力主要来自产品升级，消费结构呈现高端化转变趋势，而安全监管仍是当前的核心难点。据此，本文提出应强化卫生巾质量标准的完善与关键指标的监督、促进产品创新与环保材料应用、推进精细化市场划分与精准营销策略、并提升社会对女性健康议题的关注程度等建议。

**关键词：** 卫生巾市场；K-means 聚类；推荐模型；Kruskal-Wallis 算法

## Analysis of Consumption Status and Demand in the Sanitary Napkin Market Based on the Kruskal-Wallis Algorithm

Zhu Jingui, Gao Jianguo, Chen Xinyu, Zhan Fujie, Liu Yun, Zhou Dazhuo, Lin Peiqi  
School of Mathematics and Statistics, Huizhou University, Huizhou, Guangdong 516007

**Abstract：** With the growth of residents' income and the upgrade of health consumption, coupled with the policy support from the "Healthy China 2030" plan, this paper focuses on the high-quality development of the sanitary napkin market against the backdrop of market saturation and frequent quality and safety incidents. Through questionnaire surveys and field visits, this paper constructs a multidimensional analysis framework and employs multiple nonlinear regression, Borda count, K-means clustering, and random forest models for analysis. The results indicate that the sanitary napkin market size is stabilizing, growth relies on product upgrades, consumer structure is shifting towards high-quality products, and safety regulation remains a key challenge. Based on these findings, this paper recommends strengthening the standardization of sanitary napkin quality and the regulation of key indicators, promoting product innovation and the use of eco-friendly materials, implementing refined market segmentation and precision marketing, and enhancing social attention to women's health issues.

**Keywords：** sanitary napkin market; K-means; recommendation model; Kruskal-Wallis algorithm

## 引言

随着我国居民收入提升与健康消费理念升级，城镇居民人均可支配收入年均增长7.14%，医疗保健消费支出复合增长率达14.04%，健康消费规模预计“十四五”末期突破18万亿元，为卫生巾市场奠定坚实基础。我国女性人口达6.89亿，卫生巾市场渗透率近100%，叠加国家政策推动行业标准化发展，行业增长潜力显著。但2024年以来，卫生巾“集体塌房”、翻新售卖等质量问题频发，长度虚标、pH值不达标等乱象引发广泛健康焦虑，市场供给与女性对安全高品质产品的需求严重脱节。

本文通过调研消费现状与需求特征，剖析行业痛点与优化路径，为产品研发、标准完善及政策制定提供数据支撑，助力健康中国建设，保障女性健康权益，推动行业高质量发展。

### 作者简介：

朱金桂，惠州学院数学与统计学院本科在读，研究方向为数学与应用数学；  
高坚国，惠州学院数学与统计学院本科在读，研究方向为应用统计学；  
陈心语，惠州学院数学与统计学院本科在读，研究方向为数据科学与大数据；  
詹富婕，惠州学院数学与统计学院本科在读，研究方向为应用统计学；  
刘芸，惠州学院数学与统计学院本科在读，研究方向为应用统计学；  
周大镛，惠州学院数学与统计学院，教授，硕士生导师，研究方向为数据挖掘；  
林培琪，惠州学院数学与统计学院，辅导员，研究方向为思想政治。

一、文献综述

卫生巾市场的研究主要涉及产品性能、安全标准、消费者心理及健康影响等方面。国内研究多聚焦于渗透率、功能改进及有害物质检测，探索产品创新与安全性保障。学者甘陆卓睿与何峰（2025）指出，卫生巾凭借高渗透性与便利性，已成为女性生活必需品<sup>[1]</sup>。李燕京（2024）认为，尽管市面上有宣称抑菌、香氛、草本等功能的卫生巾，但这些未必具备实际健康效益，消费者应更多关注吸收性能与皮肤友好性<sup>[2]</sup>。

在安全性研究方面，邢飞、颜怀玉等（2023）通过 HS-GC-MS 检测卫生巾中55种挥发性有机化合物（VOCs），建立了完善的定量分析体系，为有害物质监测提供了技术基础<sup>[3]</sup>。王琦等（2023）发现部分产品仍含微量芳香族化合物，提示在环保方面亟需改进<sup>[4]</sup>。郑靖（2020）警示，使用劣质卫生巾可能导致细菌滋生、瘙痒、异味等健康问题，甚至影响生殖健康<sup>[5]</sup>。这些研究揭示了卫生巾安全性与女性健康之间的紧密关系。

随着消费者健康意识的提高，市场逐渐从功能竞争转向健康与舒适导向。陈华蕾等（2022）强调，热湿舒适性能是评价卫生巾舒适性的重要指标，并通过产品测试提出改进建议<sup>[6]</sup>。我国卫生巾的标准体系也在持续完善。GB/T 8939-2018新增甲醛与荧光物质检测指标，优化了吸收性能和背胶测试方法<sup>[7]</sup>；GB 15979-2024《一次性使用卫生用品卫生要求》也正在加速实施<sup>[8]</sup>。目前，主要包括强制性国标 GB 15979-2024和推荐性国标 GB/T 8939-2018，体现了我国在生产规范方面的不断优化<sup>[9]</sup>。

国际研究更多关注公共健康与可持续性問題。Van Eijk 等（2021）综述了44项研究，发现可重复使用卫生巾（RMP）在低收入和中等收入国家逐渐普及，并在防渗漏与皮肤刺激率方面优于传统布料产品，但其推广仍受清洁条件的限制<sup>[10]</sup>。Blignaut 等（2025）从材料与环境角度指出，一次性卫生巾中的非降解塑料与邻苯二甲酸酯对环境与女性健康构成潜在威胁，呼吁发展可降解材料，以平衡性能与环保性<sup>[11]</sup>。

现有研究多从有害物质检测、国家标准适用性及宏观政策建议角度进行探讨，然而与女性实际健康需求的结合尚显不足。由于卫生巾是直接接触女性敏感部位的日常必需品，其质量安全直接关系到女性健康。从健康视角出发，研究其市场消费现状和消费者需求，具有重要现实意义。

在市场规模持续扩大的背景下，回应女性消费者的真实需求，提升产品质量与行业创新水平，已成为推动卫生巾市场发展的核心议题。本文旨在通过健康导向的市场调研，精确对接消费需求，并提出发展建议，推动行业的可持续发展。

二、调查方案设计及实施

（一）调查内容

本调研聚焦于女性消费者的个体特征、消费行为、需求偏好及市场反馈四个核心维度。在个体特征层面，涵盖15-49岁女性的年龄段、职业类型、所在地区等基础信息，并整合初潮年龄、月经周期规律性、经期不适症状等生理指标，为细分群体识别奠定基础；消费行为层面重点收集卫生巾更换频率、常用品牌、产品类型及购买渠道等数据，以揭示消费习惯规律；需求偏好层面

通过量表量化消费者对吸收性、舒适度、材质安全性等8项关键属性的重视程度，同时调研对特殊功能（如防过敏、有机材料）的需求及支付意愿；市场现状部分则关注产品质量痛点、现有产品不足及未来发展趋势，全面映射市场供需矛盾。

（二）调查方法

本调研采用多源数据融合策略，通过文献调查、问卷调查与深度访谈相结合的方式提升数据可信度。文献调查法即在前期系统梳理国内外卫生巾行业报告、政策文件等，明确研究框架与变量定义。问卷调查法作为核心手段，团队采用线上电子问卷与线下纸质问卷并行发放的形式，覆盖学生、职场女性、自由职业者等多元群体；抽样设计基于配额抽样原则，参照国家统计局人口结构比例分配样本，在95%置信度、5%极限误差下，经设计效应调整及回答率修正，最终确定样本量为550份（有效回收527份，有效率95.8%），确保地域、年龄与职业的均衡代表。深度访谈法则选取惠州惠城区市大型商超、社区及高校作为实地调研点，通过访谈记录店员、消费者对产品体验的深层反馈，辅以现场观察记录购买行为与货架陈列，弥补问卷数据的局限性，形成三角验证。

（三）问卷设计

问卷设计遵循“模块化、逻辑化、用户友好”原则，并在预调查阶段（发放100份，回收93份）进行初步测试，依据反馈优化题目表述与选项设置，如增加注意力测试题以提升数据质量。正式问卷分为四个模块：基本信息模块通过分类变量与矩阵量表收集人口学数据与月经生理指标；使用行为模块通过多选和频率量表收集更换习惯、品牌选择等数据；需求偏好模块通过 Borda 排序法与李克特量表量化产品需求；市场评价模块结合开放题和矩阵题收集产品质量反馈与改进建议。

（四）问卷信效度检验

预调查信度检验结果如表1所示，各维度的  $\alpha$  系数均高于0.7 结果表明问卷内部一致性良好。

| 表1：预调查信度检验 |                        |    |
|------------|------------------------|----|
| 变量         | Cronbach's $\alpha$ 系数 | 项数 |
| 健康意识与风险认知  | 0.711                  | 11 |
| 消费行为与品牌偏好  | 0.795                  | 22 |
| 产品需求与改进方向  | 0.892                  | 33 |
| 质量与价格关系    | 0.732                  | 2  |
| 总体         | 0.868                  | 68 |

正式调查信度检验结果如表2所示。相比于预调查，正式调查多项类别 Cronbach's  $\alpha$  系数大于0.8，表示本问卷有较好的信度，具有较高的内部一致性。

| 表2：正式调查信度检验表 |                        |    |
|--------------|------------------------|----|
| 类别           | Cronbach's $\alpha$ 系数 | 项数 |
| 使用体验         | 0.847                  | 8  |
| 购买行为         | 0.715                  | 5  |
| 产品认知与期望      | 0.821                  | 12 |
| 总体           | 0.873                  | 25 |

本调查的信度检验通过因子分析验证结构效度，预处理后进行 KMO 测度与 Bartlett 球形检验。

预调查信度检验结果如表3所示，KMO 值为0.838 (>0.8)，Bartlett 球形检验 p 值为0.000\*\*\* (\*\*\* 代表10% 的显著性水平)，说明变量间存在显著相关性，支持因子分析，问卷具有较好的结构效度。

| 表3：预调查效度检验结果表  |       |          |
|----------------|-------|----------|
| KMO 值          | 0.838 |          |
| Bartlett 球形度检验 | 近似卡方  | 227.108  |
|                | df    | 28       |
|                | P     | 0.000*** |

注：\*\*\*、\*\*、\* 分别代表 1%、5%、10% 的显著性水平

正式调查效度检验结果如表 4 所示，KMO 值进一步提升至 0.875，验证了问卷结构效度优异，能够准确测量目标变量。

| 表 4：正式调查效度检验结果表 |       |          |
|-----------------|-------|----------|
| KMO 值           | 0.875 |          |
| Bartlett 球形度检验  | 近似卡方  | 1262.239 |
|                 | df    | 326      |
|                 | P     | 0.000*** |

综上，正式调查具有更好的结构效度，问卷设计科学合理，信效度指标均符合学术标准，为后续数据分析提供可靠基础。

### 三、卫生巾市场发展及消费现状

#### （一）基于非线性回归的市场规模预测

##### 1. 建立卫生巾价格以及销量的回归预测模型

为了分析卫生巾行业在成熟期的增长趋势，本研究基于 2010 至 2025 年中国卫生巾市场的核心年度数据，构建了两条非线性回归预测路径：分别描述价格与销量的变化趋势。具体而言，因变量包括卫生巾的平均单价（单位：元 / 片）与年销量（单位：亿片），自变量则涵盖了影响价格与销量的关键因素，如女性人口总量（单位：亿人）、国际原油均价（单位：美元 / 桶）、木浆（漂白阔叶浆 / 针叶浆）价格（单位：元 / 吨）等成本与需求侧的变量。此外，为了验证模型的有效性，本文同时构建了线性基准模型，并采用 2010 至 2022 年数据作为训练集，2023 至 2025 年数据作为预测集，报告了模型的拟合优度指标，包括  $R^2$  与 RMSE 等。

模型设定价格模型与销量模型均采用含交互项的非线性形式：

$$P = \gamma_0 + \sum_{i=1}^3 \gamma_i Z_i + \sum_{i=1}^3 \gamma_{i+3} Z_i^2 + \sum_{1 \leq i < j \leq 3} \gamma_k Z_i Z_j \quad (1)$$

$$Q = \beta_0 + \sum_{i=1}^3 \beta_i Z_i + \sum_{i=1}^3 \beta_{i+3} Z_i^2 + \sum_{1 \leq i < j \leq 3} \beta_k Z_i Z_j \quad (2)$$

其中， $P$  为卫生巾平均价格， $Q$  为卫生巾销量  $Z_1$  为木浆产量， $Z_2$  为石油价格， $Z_3$  为女性人口。交互项系数  $k$  按  $i, j$  组合顺序编号。

| 表 5：回归模型系数估计表 |                   |                   |
|---------------|-------------------|-------------------|
| 系数项           | 价格模型 ( $\gamma$ ) | 零售量模型 ( $\beta$ ) |
| 截距项           | 0.6743            | 116980.1698       |
| $Z_1$         | 0.0350            | 3313.2227         |
| $Z_2$         | -0.0034           | -1394.4506        |
| $Z_3$         | 0.0271            | 1629.1369         |
| $Z_1^2$       | 0.0433            | 4168.8817         |
| $Z_2^2$       | 0.0076            | 239.8618          |
| $Z_3^2$       | -0.0054           | -840.5300         |

|           |         |            |
|-----------|---------|------------|
| $Z_1 Z_2$ | 0.0040  | -450.3201  |
| $Z_1 Z_3$ | -0.0027 | -1167.0108 |
| $Z_2 Z_3$ | -0.0005 | 359.2940   |

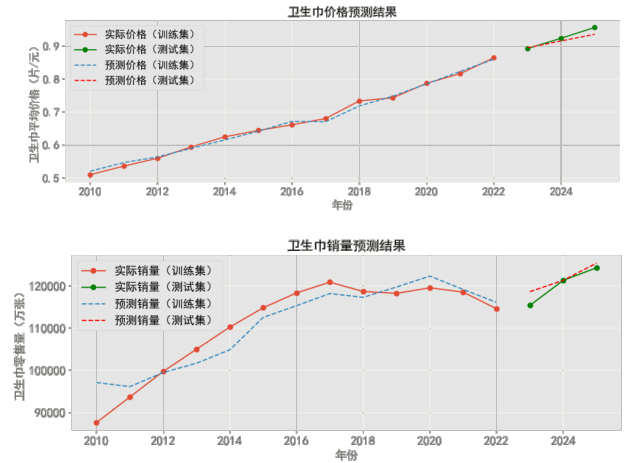


图 1：卫生巾销量及单价预测对比（2010–2025）

##### 2. 拟合优度与基准对比

与线性回归模型相比，非线性回归模型在价格方程和销量方程上均表现出显著的拟合优度提升。具体来说，非线性回归模型的  $R^2$  值分别达到了 0.89 和 0.85，而 RMSE 则分别降至 0.04 和 150（单位与因变量一致），显著优于线性基准模型。这表明，非线性模型在捕捉历史波动和市场结构变化的过程中，具有更强的稳健性和精确度。具体拟合优度比较结果见表 6。

| 表 6：非线性与线性模型的拟合优度比较 |            |            |         |         |
|---------------------|------------|------------|---------|---------|
| 模型                  | 价格方程 $R^2$ | 销量方程 $R^2$ | 价格 RMSE | 销量 RMSE |
| 线性回归                | 0.72       | 0.68       | 0.08    | 280     |
| 非线性回归               | 0.89       | 0.85       | 0.04    | 150     |

基于非线性回归模型的估计结果，卫生巾行业的增长动力呈现出“成本—价格—结构升级”的多层次耦合效应。上游原油和浆价通过非线性传导作用，影响终端价格，并在不同市场阶段展现出“压制 / 放大”效应。在女性人口趋于稳定的背景下，销量对价格和结构升级的敏感度逐渐降低，进而形成“价格温和上升、销量小幅波动”的市场成熟期格局。

#### （二）基于 Mann-Whitney 检验的消费者画像与偏好差异

本研究通过 Mann-Whitney U 检验，基于“购买决策关键属性”（吸收性能、舒适度、材质、透气性、价格、品牌知名度、环保性、长度 / 厚度），分析了不同地区、年龄与职业群体之间的偏好差异。使用 Kruskal-Wallis 秩和检验方法对群体差异进行了初步判定，若存在显著差异，再通过 Mann-Whitney U 检验进行两两比较，并报告了相应的  $p$  值。

##### 1. 总体差异

通过检验发现，不同群体之间在关键属性上的偏好存在显著差异。具体而言：首先，不同地区消费者在品牌知名度的重视程度上存在显著差异 ( $p = 0.02$ )。其次，不同年龄段群体对环保性（如有机材质）的关注度存在显著差异 ( $p = 0.03$ )。最后，不同职业群体在透气性上的偏好也存在显著差异 ( $p = 0.025$ )。其他

属性在相应分组下未出现显著差异，具体见表7。

表7：不同群体在关键属性上的总体差异（KruskalWallis p 值）

| 特征        | 位置（p）       | 年龄（p）       | 职业（p）       |
|-----------|-------------|-------------|-------------|
| 吸收性能      | 0.067611346 | 0.700832097 | 0.062058096 |
| 舒适度       | 0.69893057  | 0.441957935 | 0.339451807 |
| 材质（纯棉、网面） | 0.433017565 | 0.447073122 | 0.742347635 |
| 透气性       | 0.14680651  | 0.628160277 | 0.025312953 |
| 价格        | 0.316361288 | 0.978759398 | 0.409097162 |
| 品牌知名度     | 0.02359599  | 0.845514156 | 0.525034272 |
| 环保性       | 0.206419096 | 0.038273449 | 0.833061974 |
| 长度、厚度     | 0.66333947  | 0.882780138 | 0.866244589 |

## 2. 结果解读与行业建议

秩和检验结果表明，消费者在职业、地区和年龄等维度上对关键属性的偏好存在显著差异，反映出消费情境和身份特征对产品选择的影响。首先，自由职业者更关注卫生巾的透气性，产品应强调透气性和舒适性，并通过量化测试降低体验不确定性。其次，一线城市消费者依赖品牌知名度和口碑效应，而非一线城市的消费者则更注重性价比和产品效果验证。因此，企业应在一线市场加强品牌信任，通过权威认证和专家评测，在二三线市场则应聚焦高性价比和真实用户评价。最后，年轻群体对环保性和可持续材料更为敏感，偏好有机材质和可回收包装。针对这一趋势，企业应突出环保特性，并加强供应链透明度，以提升产品的可信度。

### （三）基于集成学习的用卫生巾推荐模型

在搭建推荐模型之前，数据处理工作已经完成。考虑到本次问卷收集的数据属于中等规模，团队计划首先进行数据标准化，同时利用 UMAP 来进行降维，将数据精准化处理，并运用 K-means 聚类算法完成用户分群操作，核心目的是依据用户的行为特征，划分出具有相似偏好的群体，最后使用随机森林模型完成推荐模型的搭建。

对应分析法的具体步骤如下：

采用 Z-score 标准化对原始数据进行处理，此方法将不同量纲的特征统一到均值为 0、标准差 1 的分布中，公式为：

$$Z=\frac{\mathbf{k}-\mathbf{\acute{o}}}{\mathbf{\acute{o}}}\tag{3}$$

其中， $\mathbf{\acute{i}}$  为特征均值， $\mathbf{\acute{o}}$  为标准差。通过标准化达到消除数值范围差异对模型的影响。

为了从问卷数据中提取关键特征、降低多重共线性带来的噪声并提升后续模型精度，团队采用无监督降维方法 UMAP 代替传统 PCA 模型。UMAP 兼顾数据局部与全局结构保留，且适配大规模问卷数据的特征筛选需求，通过量化特征对降维后数据结构的影响程度，筛选出核心关键特征，最终提取结果如下表8。

表8：主要特征和其贡献值

| 特征       | 贡献度      |
|----------|----------|
| 虫卵       | 0.989192 |
| 有异味      | 0.968716 |
| 价格与质量的改善 | 0.953347 |
| 职业类型     | 0.947433 |

|                                  |          |
|----------------------------------|----------|
| 假冒伪劣产品                           | 0.946978 |
| 您是否知晓使用不合格卫生巾会导致皮肤过敏、细菌滋生感染、妇科疾病 | 0.946651 |
| 订阅制服务与精准配送                       | 0.946283 |
| 形状与内裤不合                          | 0.944397 |

基于 K-Means 聚类算法，团队将用户划分为 0-9 共 10 个类群，并把算法输出的 0-9 簇标签作为新特征列融入原始数据，以此建立用户分群与行为特征之间的关联。下图展示了每个用户簇对各款卫生巾的评分偏好情况。

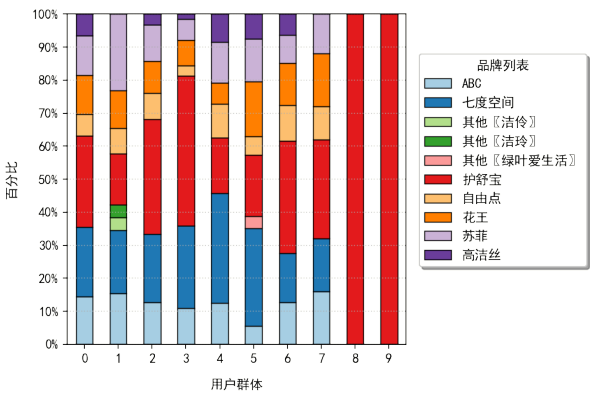


图2：不同用户群体对于卫生巾品牌使用偏好

结合前文分析所得的用户行为特征与聚类分群结果，团队通过预测用户对卫生巾产品的满意度，进而实现个性化推荐目标。本部分选用集成学习框架下的随机森林模型，将聚类标签纳入特征体系，以此增强模型的解释性与预测性能。

将数据集按 8:2 的比例划分为训练集和测试集。将使用原有数据以及加入聚类标签的数据作为特征输入模型，并做出仅使用原始数据来进行对比训练，结果如下表9所示，结果显示使用聚类标签的效果更佳。

表9：加入聚类标签前后准确率对比

| 数据集    | 准确率    |
|--------|--------|
| 原始数据   | 25.58% |
| 加入聚类标签 | 75.58% |

在本次模型构建过程中，团队将推荐对象划分为 9 类消费群体。这 9 类消费者多维度均呈现出显著的差异化特征，为后续精准营销落地与产品迭代优化提供了坚实的数据支撑。具体量化详情如下表 10 所示。

表10：消费者群体量化表

| 消费者群体    | 年龄标准 | 预算标准 | 渠道标准 | 问题标准 | 需求标准 |
|----------|------|------|------|------|------|
| 价格敏感型学生党 | 2    | 2    | 2    | 6    | 6    |
| 职场品质追求者  | 6    | 8    | 6    | 5    | 7    |
| 环保健康倡导者  | 7    | 6    | 8    | 5    | 7    |
| 经期问题困扰者  | 4    | 4    | 7    | 8    | 8    |
| 传统品牌忠诚用户 | 8    | 4    | 2    | 6    | 6    |



|         |   |   |   |   |   |
|---------|---|---|---|---|---|
| 特殊需求专业户 | 7 | 8 | 7 | 5 | 7 |
| 大流量焦虑群体 | 5 | 4 | 4 | 7 | 7 |
| 健康风险规避者 | 6 | 8 | 8 | 6 | 6 |
| 科技尝鲜先锋派 | 3 | 8 | 9 | 5 | 7 |

#### 四、结论与建议

##### （一）结论

###### 1. 市场增长放缓，消费群体趋于稳定

当前卫生巾市场面临的主要问题之一是市场增长放缓。女性总人口增长的放缓导致了新增消费者的不足，无法依赖人口红利来推动市场扩展。特别是在2018至2023年期间，中国女性人口的复合年增长率不足0.4%，使得卫生巾行业的消费群体趋于稳定，整体市场规模的扩张受到了限制。

###### 2. 市场渗透率接近饱和，新增用户空间有限

另一个突出的问题是市场渗透率已接近饱和，几乎所有适龄女性都已开始使用卫生巾。根据《中国生活用纸年鉴》2017年的数据，卫生巾的渗透率接近100%，行业几乎没有新增用户的空间。因此，企业只能依赖现有消费者来维持市场份额，难以通过用户拓展来驱动进一步的 growth。

###### 3. 消费结构向高端倾斜，低端产品面临生存压力

随着消费者生活水平的提升，市场的消费结构逐渐向高端产品倾斜。高端卫生巾的市场份额从2012年的15% 上升至2022年的35%，这导致了传统低端产品的需求逐渐减少，品牌竞争也愈加激烈。中低端企业面临着生存压力，必须寻找新的竞争策略来适应市场的变化。

###### 4. 质量安全问题频发，行业信任危机加剧

质量安全问题的频繁曝光加剧了公众对卫生巾行业的健康焦虑，行业的信任危机也日益凸显。荧光增白剂残留、活虫污染等

问题的出现，以及监管缺失和标准滞后，导致消费者对产品的安全性产生了质疑。这种情况不仅影响了品牌形象，也加剧了消费者的警惕性，进一步削弱了市场的信任度。

##### （二）建议

###### 1. 推动产品升级与消费频次提升

针对市场增长放缓的问题，企业应推动产品的升级与消费频次的提升。通过多元回归预测模型优化供需平衡，预计2023年至2029年市场零售规模的增长率将保持在0.2% 至0.5% 的区间。企业可以通过提升现有产品的附加值和满足消费者个性化需求，促进消费者的复购，从而在稳定的市场环境中实现小幅增长。

###### 2. 加强质量监管与标准化，保护消费者权益

为了应对市场渗透率接近饱和的问题，行业应加强质量监管与标准化，尤其要配合新国标（GB 15979-2024）的实施，进一步细化生产标准，如内吸收层长度、PH 值等方面。企业应从单纯的利润导向转向更加注重消费者权益的方向，确保产品的高质量和安全性，这样能够在饱和市场中保持竞争力。

###### 3. 推动产品创新与个性化发展，满足多样化需求

为应对消费结构向高端倾斜的趋势，企业应推动产品创新与个性化发展。针对不同用户群体（如价格敏感型学生、职场女性的品质追求者、环保健康倡导者等），推出定制化产品。同时，开发更多天然环保材料，满足日益多样化的消费者需求。通过这些创新，企业能够在高端市场中获得更大份额，并解决低端市场逐渐萎缩的问题。

###### 4. 加强市场细分与精准营销，提升品牌形象

针对行业信任危机，企业应加强市场细分与精准营销。根据不同消费者群体（如学生群体的价格敏感性、职场白领对品质的需求等），进行定向推广。同时，企业应加大对女性月经卫生问题的关注，推动“月经贫困”和“月经羞耻”话题的讨论，提升社会责任感。通过提供健康环保的产品，企业不仅能提升品牌形象，还能在竞争激烈的市场中建立消费者的信任。

#### 参考文献

[1] 甘陆卓睿, 何峰. 基于女性消费心理的卫生用品情感化包装设计研究 [J]. 中国包装, 2025, 45(01): 49-52.

[2] 李燕京. 这些“加料”的卫生巾值得买吗 [N]. 中国消费者报, 2024-08-22(003).

[3] 邢飞, 颜怀玉, 罗佳, 等. 顶空-气相色谱-质谱法测定卫生巾中55种挥发性有机物的含量 [J]. 理化检验-化学分册, 2023, 59(03): 315-320.

[4] 王琦, 颜怀玉, 孙中岩, 等. 顶空气相色谱-质谱联用法测定卫生巾中3种二甲苯同分异构体 [J]. 山东纺织科技, 2023, 64(04): 25-28.

[5] 郑靖. 预防妇科疾病, 从科学护理私处开始 [J]. 幸福家庭, 2020, (06): 35.

[6] 陈华蕾, 王林林, 肖爱民. 卫生巾热湿舒适性评价与预测 [J]. 产业用纺织品, 2022, 40(05): 29-36.

[7] 洁. 《卫生巾(护垫)》等4项生活用纸制品国家标准发布 [J]. 福建轻纺, 2018, (06): 6.

[8] 王小月. 卫生巾市场不平静相关标准正在修订 [N]. 中国消费者报, 2024-11-29(003).

[9] 牛煜辉. 新国标实施, 告别“卫生巾焦虑” [N]. 中国人口报, 2025-08-20(007).

[10] Van Eijk A. M., Jayasinghe N., Zulaika G., et al. Exploring menstrual products: A systematic review and meta-analysis of reusable menstrual pads for public health internationally[J]. PLOS ONE, 2021, 16(9): e0257610.

[11] Blignaut J., Visser H. G., Erasmus E., et al. Sanitary pads—composition, regulation, and ongoing research to address associated challenges[J]. Journal of Materials Science, 2025, 60(31): 13109-13155.

