DOI: 10.19812/j.cnki.jfsq11-5956/ts.20251207001
引用格式: 余雪梅. 基于多维时空风险特征工程的食用农产品抽检结果预测模型研究[J]. 食品安全质量检测学报, 2026, 17(10): 293–301.
YU XM. Research on predictive model for edible agricultural product inspection results based on multidimensional spatiotemporal risk feature engineering [J]. Journal of Food Safety & Quality, 2026, 17(10): 293–301. (in Chinese with English abstract).

基于多维时空风险特征工程的食用农产品抽检结果预测模型研究

余雪梅*

(三明市检验检测中心, 三明 365000)

摘 要: 目的 建立基于多维时空风险特征工程的机器学习预测模型, 对食用农产品监督抽检结果进行预测。**方法** 基于 2022—2024 年福建省地级市食用农产品监督抽检数据, 通过系统性地整合数据的食用农产品品种、抽样地、抽样场所、生产日期等信息, 引入多维时空风险特征工程方法处理数据集, 在原始数据的基础上构建 6 种新特征组成新数据集, 同时运用 3 种主流的预测模型对处理后的数据进行食用农产品抽检结果预测。**结果** 在 3 种主流模型上, 多维时空风险特征工程方法的预测性能均优于传统采样方法。其中, 多维时空风险特征工程与轻量级梯度提升机(light gradient boosting machine, LightGBM)组合对测试集样本的抽检结果进行预测, 该模型 AUC 值(0.845)、Recall (0.881)、G-mean 值(0.741)、业务成本(8561), 综合性能最佳。**结论** 多维时空风险特征工程相较于传统采样方法, 增强了模型对复杂风险模式的捕捉能力, 使模型的预测性能得到全面提升。基于多维时空风险特征工程的 LightGBM 模型能够较为准确地预测食用农产品的抽检结果, 能够为差异化、动态化的食用农产品监督抽检方案制定及食用农产品安全风险预警提供技术支撑。

关键词: 食用农产品; 多维时空风险特征工程; 轻量级梯度提升机; 预测模型

Research on predictive model for edible agricultural product inspection results based on multidimensional spatiotemporal risk feature engineering

YU Xue-Mei*

(*Sanming Inspection & Testing Center for Product Quality*, *Sanming* 365000, *China*)

ABSTRACT: Objective To establish a machine learning predictive model based on multidimensional spatiotemporal risk feature engineering for forecasting the results of supervisory sampling inspections of edible agricultural products. **Methods** Using supervisory sampling data for edible agricultural products from Fujian's prefecture-level cities between 2022 and 2024, information on product type, sampling location, sampling site and production date were systematically integrated. Multidimensional spatiotemporal risk feature engineering was applied to process the dataset, constructing 6 kinds of new features to form a novel dataset based on the original data. The 3

收稿日期: 2025-12-07
第一作者/*通信作者: 余雪梅(1988—), 女, 硕士, 工程师, 主要研究方向为食品安全抽检。E-mail: 649651911@qq.com

kinds of mainstream binary classification prediction models were applied to forecast inspection outcomes using the processed data. **Results** Across all 3 kinds of mainstream models, the multidimensional spatiotemporal risk feature engineering method demonstrated superior predictive performance compared to traditional sampling methods. Specifically, the combination of multidimensional spatiotemporal risk feature engineering with light gradient boosting machine (LightGBM) achieved the best overall performance on the test set samples, with an AUC of 0.845, Recall of 0.881, G-mean of 0.741 and misclassification costs of 8561. **Conclusion** Compared to traditional sampling methods, multidimensional spatiotemporal risk feature engineering enhances the model's ability to capture complex risk patterns, leading to a comprehensive improvement in predictive performance. The LightGBM model based on multidimensional spatiotemporal risk feature engineering can accurately predict the inspection results of edible agricultural products, providing technical support for formulating differentiated, dynamic supervision and inspection plans and risk early warning systems for edible agricultural products.

KEY WORDS: edible agricultural products; multidimensional spatiotemporal risk feature engineering; light gradient boosting machine; predictive model

0 引 言

农产品作为人类生存与发展的基本物质基础，其质量安全状况直接关系到下游食品产业链的安全水平，是保障公众健康和生命安全的首要屏障。然而，根据国家市场监督管理总局近 5 年公布的数据来看，食用农产品抽检不合格率持续处于较高水平，其中农药残留、兽药残留、重金属污染以及生物毒素污染是导致食用农产品质量不达标的主要因素[1−3]，反映出当前农产品质量安全形势依然严峻。食用农产品质量安全监管因供应链长、主体多元、储运难度大、风险类型复杂等因素，防控难度较高。传统监管方式仅依靠观察主体行为或简单分析抽检数据与典型案例，已难以精准判断监管重点与抽检方向。在此背景下，人工智能技术的快速发展为破解食用农产品质量安全监管难题提供了新的技术路径。

目前，已有诸多学者研究将数据挖掘、机器学习、深度学习等大数据和人工智能前沿技术融入食品安全预警监测体系中，其中机器学习因具备对大数据的强大分析能力与对多种特征间潜在联系的识别能力，非常适用于准确识别食品安全风险发生的多维度关联特征并预测风险。现阶段用于食品安全预警监测的机器学习模型主要包括贝叶斯网络(Bayesian network, BN)[4−5]、最邻近分类算法(K-nearest neighbor, KNN)[6−7]、反向传播神经网络(back propagation, BP)[8]、支持向量机(support vector machine, SVM)[9−10]、随机森林(random forest, RF)[11−12]、极端梯度提升树算法(extreme gradient boosting, XGBoost)[13−14]、轻量级梯度提升机(light gradient boosting machine, LightGBM)[15−16]、基于对称决策树算法的 GBDT 框架(categorical boosting, CatBoost)[17−18]、基于长短期记忆循环神经网络(long short-term memory, LSTM)[19−20]、Transformer[21]等。其中 RF、LightGBM、XGBoost 等主流的预测模型，在食品安全风险预测领域有良好的表现。我国当前已积累的大量食用农产品抽检数据，为应用机器学习模型进行食品安全风险预测提供了数据基础。然而，在食用农产品数据集中，各类别(尤其是合格与不合格品类)分布的高度不平衡，严重影响了模型的稳健性。这种失衡会扭曲模型的决策边界，导致其对稀有但关键的不合格样本识别能力不足，极大地削弱了模型在实际监管中的应用价值。

为应对类别不平衡带来的挑战，研究者目前主要聚焦于数据层面的处理技术[22]。其中，重采样方法是最为经典和广泛使用的策略之一[23]。过采样技术，如经典的合成少数类过采样技术(synthetic minority over-sampling technique, SMOTE)及其衍生算法，通过线性插值在少数类样本间生成人工合成样本，以增加少数类的代表性。与之相对的欠采样技术，如随机欠采样(random under sampling, RUS)、编辑最近邻法(edited nearest neighbors, ENN)等，则通过删除多数类样本来平衡数据集。而混合采样方法，如 SMOTE-ENN，则结合了过采样和欠采样的优势，试图在增加少数类样本的同时清理多数类中的噪声样本。杨鸿雁等[10]、宇盛好等[16]和周子文等[18]通过对比过采样、欠采样及混合采样策略与各类机器学习算法(如 RF、SVM、LightGBM、XGBoost)的组合，一致证实了采样方法是提升模型识别能力的关键因素，对食品抽检不平衡数据的预测性能具有显著影响。然而，这些采样方法主要从数据层面缓解类别不平衡问题，其固有局限性不容忽视：过采样可能因生成重复或相似样本而导致过拟合，欠采样则可能丢失多数类中的重要信息，且这些方法都未能从根本上增强特征空间的识别能力。

随着研究的深入，学者们逐渐认识到，不平衡数据问题的本质不仅在于样本数量的不均衡，更在于特征空间中类别判别信息的不足。因此，研究焦点开始从单纯调整数据分布转向通过特征工程构建判别性更强的新特征[24−25]。

这种方法的核心思想是: 通过创造能够更好地区分类别, 特别是能够凸显少数类独特模式的特征的算法, 从本质上提升模型的识别能力。特征工程方法不改变样本分布, 而是通过挖掘和构造更具代表性的特征, 为模型提供更深层次、更具判别力的信息, 从而使其能够从固有数据中更好地学习少数类模式。

在这一理念的指导下, 本研究提出了一种多维时空风险特征工程方法, 专门用于处理食品安全风险预测中的不平衡分类问题。(1)引入时间衰减风险特征, 基于指数衰减函数对历史不合格记录进行加权, 使近期事件比远期事件具有更高的权重, 从而捕捉风险的动态演化规律[26–28]。(2)构建空间综合风险特征, 在综合考虑本地风险影响的基础上, 同时利用地理信息系统中的邻接关系和空间相邻权重公式, 计算相邻地区的风险传播效应, 体现了风险在空间维度上的相关性[29–32]。(3)设计基于高风险品种的空间增强风险特征, 针对历史上表现出较高风险的特殊食品品类进行风险放大, 强化模型对重点监管对象的敏感度[33–35]。(4)引入场所偏置调整后的时间衰减风险特征, 根据不同抽样场所(如超市、农贸市场、网店等)的历史风险水平对基础风险进行校正, 反映不同抽样场所的风险差异[36–37]。

多维时空风险特征工程方法通过系统整合时间、空间、品种与场所 4 类风险维度, 对不合格样本进行精细化的风险刻画, 从而在不改变数据分布的情况下, 显著提升了模型对关键风险的识别能力。本研究将通过实证分析, 验证这一特征工程方法在食品安全风险预测中的有效性和优越性, 为不平衡数据分类问题提供新的解决思路, 同时也为食品安全智能监管提供技术支撑。

1　材料与方法

1.1　数据来源

本研究所用数据来源于福建省各地级市市场监督管理局官方网站公布的 2022—2024 年食用农产品监督抽检信息, 共 75763 条数据。数据包含食用农产品品种、生产日期、被抽样单位名称、抽样场所、抽样地、抽样环节以及抽检结果等关键字段, 其中按照食用农产品样品采集的种类分为蔬菜、肉、蛋等 7 个食用农产品大类以及 26 个食用农产品亚类; 按照食用农产品样品采集的场所分为流通环节(包括超市、农贸市场、生鲜零售门店等)、餐饮环节(包括餐饮店、单位食堂等)和网络环节(网店); 抽样地为各地市名称; 抽检结果为是否合格; 为本研究提供了真实、权威的样本基础。本研究以食用农产品亚类、抽样场所、规格型号、抽样环节、抽样地、生产日期 6 个代表性的字段为原始输入特征, 抽检结果(合格或者不合格)为样本标签。福建省各地市的基本地理要素的矢量数据来源于福建省自然资源厅官方发布的福建省标准地图服务系统。

1.2　预测模型构建

1.2.1　多维时空风险特征构建

(1)时间衰减风险特征构建

为了建立食用农产品的动态风险评估, 捕捉近期历史对当前风险的影响, 让模型更加关注近期的不合格风险, 本研究采用指数衰减权重公式来计算时间衰减风险[28]。首先对原始数据进行预处理, 将生产日期统一转换为距离基准日(2022 年 1 月 1 日)的天数, 同时将抽检结果二值化(不合格标记为 1, 合格标记为 0)。在此基础上, 按食用农产品亚类、抽样地、天数、标签构建特征子集。采用指数衰减函数为每个历史样本分配时间权重, 使得距离当前时间越近的样本具有越高的权重, 从而强化近期风险信号。计算公式见(1):

$$w_i=\exp[-\lambda\times(t_{\max}-t_i)] \tag{1}$$

式中: w_i 为该子集中样本 i 的时间衰减权重; t_i 为该子集中样本 i 距离 2022 年 1 月 1 日的天数; $t_{\max}$ 为该子集中最新的日期距离 2022 年 1 月 1 日的天数; λ 为衰减率, 取值 0.01。

针对每个(食用农产品亚类–抽检地)特征组合, 基于时间衰减权重, 对其包含的所有样本进行加权平均, 得到该特征组的近期风险值。这种方法能够有效捕捉风险随时间变化的动态特征, 近期的不合格记录对风险评分的贡献更大。计算公式见(2):

$$R_{\text{recent}}=\frac{\sum_{i=1}^{n}w_i\times y_i}{\sum_{i=1}^{n}w_i} \tag{2}$$

式中: R_{recent} 为相应(食用农产品亚类 – 抽检地)特征组的近期风险; y_i 为子集中该特征组第 i 个样本标签; n 为子集中该特征组的样本数。

考虑到某些特征组合可能样本数量有限, 直接使用近期风险容易出现估计偏差, 模型引入贝叶斯平滑技术进行校正。通过将近期风险与全局不合格率进行加权融合, 在样本量较小时更多地依赖全局先验信息, 在样本量充足时则更信任该特征组的实际观测值, 从而获得更加稳健的时间衰减风险评估结果。计算公式见(3):

$$R_{\text{decay}}=\frac{R_{\text{recent}}\times n+r_{\text{global}}\times\alpha}{n+\alpha} \tag{3}$$

式中: R_{decay} 为相应(食用农产品亚类-抽检地)特征组的时间衰减风险; r_{global} 为总不合格率; α 为平滑强度参数, 取值 50。

(2)空间综合风险特征构建

食品安全问题会通过供应链、人员流动等途径向周边地区传播, 为了实现地理位置相邻间的风险信息共享和预警, 本研究利用空间相邻权重公式来计算每个抽样地对应的空间相邻风险, 同时综合考虑本地风险影响, 构建空间综合风险特征[31–32]。

模型先计算出每个抽样地的不合格率, 然后从福建

省各地市的基本地理要素的矢量数据中获取相邻位置关系信息以及相应的距离，并根据反比例函数计算出相邻抽样地的距离权重，距离权重随着距离的增加而减小。计算公式见(4):

$$W_{i=}\frac{1}{d_i+10} \tag{4}$$

式中: W_i为抽样地的第 i 个相邻地的距离权重; d_i为抽样地到第 i 个相邻地相应的距离；常数 10 是为了防止当距离为 0 时分母为零的情况。

计算出相邻地的距离权重后，通过将权重进行归一化处理。同时加权每一个相邻地的风险，计算出抽样地的总体相邻风险。计算公式见(5)~(6):

$$W_i'=\frac{W_i}{\sum_{i=1}^{n'}W_i} \tag{5}$$

$$R_{\text{neighbor}}=\sum_{i=1}^{n'}W_i'\times r_i \tag{6}$$

式中: W_i'为归一化后距离权重; n'为相邻地的数量; r_i为第 i 个相邻地的不合格率; R_{neighbor}为抽样地的总体相邻风险。

为了综合考虑平衡本地实际情况与相邻区域环境影响，将抽样地本地风险(设置权重 0.6)和总体相邻风险(设置权重 0.4)融合，得到最终空间综合风险特征。计算公式见(7):

$$R_{\text{spatial}}=0.6\times r_{\text{loc}}+0.4\times R_{\text{neighbor}} \tag{7}$$

式中: R_{spatial}为抽样地的空间综合风险; r_{loc}为抽样地的不合格率; 0.6 为设置的抽样地的风险权重; 0.4 为抽样地的总体相邻风险权重。

(3)基于高风险品种的空间增强风险特征构建

为了提升模型对高风险品种的识别能力和预测性能，模型基于同质化传播效应和高风险入侵效应原理，通过评估高风险品种与抽样地风险之间的相似性以及高风险品种的影响，构建基于高风险品种的空间增强风险特征[34]。

根据历史数据识别出高风险品种(5 种食用农产品亚类: 根茎类和薯芋类蔬菜、豆类蔬菜、热带和亚热带水果、鳞茎类蔬菜、其他水产品)，并计算每种高风险品种与抽样地的不合格率相似性。通过计算高风险品种不合格率与抽样地不合格率的绝对差异，得到风险相似性指标，并据此确定传播权重。当高风险品种风险高于当地基础水平时，赋予 1.5 倍的增强系数，强化风险传播效应。计算公式见(8)~(9):

$$S=1-|r_{\text{variety}}-r_{\text{loc}}| \tag{8}$$

$$W_{\text{propagation}}=\begin{cases}1.5\times S, & \text{if } r_{\text{variety}}>r_{\text{loc}}\\ 1.0\times S, & \text{if } r_{\text{variety}}<r_{\text{loc}}\end{cases} \tag{9}$$

式中: S 为高风险品种与抽样地风险相似度; r_{variety}为高风险品种的不合格率; $W_{\text{propagation}}$为风险传播权重。

引入风险警戒阈值(θ=0.15)机制，当高风险品种不合格率超过此阈值时，触发高风险入侵效应。通过计算高风险品种不合格率与警戒阈值的比值，构建增强因子，实现对风险传播强度的非线性放大。计算公式见(10):

$$E=1+\frac{r_{\text{variety}}}{\theta} \tag{10}$$

式中: E 为风险增强因子; θ 为风险警戒阈值，取值 0.15。

以抽样地空间综合风险为基准，结合风险传播权重和风险增强因子，通过乘积融合生成最终的基于高风险品种的空间增强风险特征。该特征既反映了空间风险的基本水平，又融入了高风险品种的特异性风险信号。计算公式见(11):

$$R_{\text{enhance}}=W_{\text{propagation}}\times R_{\text{spatial}}\times E \tag{11}$$

式中: R_{enhance}为抽样地的空间增强风险。

(4)基于抽样场所因果去偏的时间衰减风险特征构建

在大量的抽样样本中，抽样场所数据分布不均衡，有可能样本量的差异，导致模型预测过程中出现较大偏差。为了避免这种统计陷阱，本研究根据监管强度和样本量的不同对各类抽样场所设置相应的场所风险偏置系数，从而修正历史风险[37]。具体而言，比如超市由于监管相对严格，抽样样本量大，可能由于这样的强监管而发现更多的问题，其实际风险需要向下修正的力度较大，应赋予其更高的偏置系数；农贸市场由于其监管困难，监管相对薄弱，抽样样本量较小，风险发现不够充分，其实际风险需要修正的力度较小。根据实际业务知识为不同的抽样场所设置了不同的偏置系数(超市: 1.0; 生鲜零售门店: 0.2; 农贸市场: 0.1; 餐饮店: 0.15; 单位食堂: 0.3; 网店: 0.25; 其他: 0.1)。计算公式见(12):

$$R_{\text{debiased}}=R_{\text{decay}}\times(1-0.3\times B_{\text{place}}) \tag{12}$$

式中: R_{debiased}为样本基于抽样场所因果去偏后的时间衰减风险; B_{place}为抽样场所偏置系数; 0.3 为修正系数。

将以上通过多维时空风险特征工程方法形成的新特征(R_{recent}、R_{decay}、R_{spatial}、$W_{\text{propagation}}$、R_{enhance}、R_{debiased})与原数据集的样本特征合并，形成新样本数据集。新样本数据集与原数据集的数据量相同，共 75763 条，新样本数据特征包含食用农产品亚类、抽样场所、规格型号、抽样环节、抽样地、生产日期、R_{recent}、R_{decay}、R_{spatial}、$W_{\text{propagation}}$、R_{enhance}、R_{debiased}、抽检结果(样本标签)。对新样本数据集进行数据标准化处理和分层抽样，分割数据集(训练集 80%，测试集 20%)。与此同时，本研究设计了对比试验，以无采样(原始数据集不做处理)、过采样(SMOTE)及混合采样(SMOTE-ENN)处理 3 种数据形态作为对照，与多维时空风险特征工程方法处理的新数据集同时进行模型的训练与预测，以确保评估的全面性与可靠性。

1.2.2 模型准备

本研究从 Python 的机器学习库中调取了在处理不平衡数据方面表现良好的 3 种主流模型(RF、LightGBM 和 XGBoost)进行后续的训练和预测。所有模型均在 4 种数据集[无采样数据、过采样(SMOTE)数据、混合采样(SMOTE-ENN)数据及多维时空风险特征数据]上进行训练与预测。模型的参数设置基于相关领域的先验研究与初步实验的经验调校确定，其核心参数设置如表 1 所示。

表 1　3 种模型的参数设置

Table 1　Parameter configurations of the 3 kinds of models

模型	参数设置
RF	*n* estimators=200; max depth=15; min samples split=10; min samples leaf=5
LightGBM	*n* estimators=200; max depth=10; learning rate=0.05; num leaves=31; min child samples=20
XGBoost	*n* estimators=200; max depth=8; learning rate=0.05; min child weight=5

1.3　模型评估指标

数据分类问题的评估指标对机器学习算法的研究起着重要的指导作用, 评估指标是否合理直接影响分类模型的可靠性和实用性。在传统分类问题中, 一般采用准确率为算法模型性能的评估指标, 但其对于评估不平衡数据分类性能并不适用, 准确率不能很好反映样本数据的整体性能。在不平衡数据分类任务中, 少数类样本数量远小于多数类样本, 需要在兼顾整体性能的基础上, 重点关注少数类的分类正确率, 尽可能地把所有少数类样本都识别出来。所以本研究选择了一些更适合不平衡数据分类问题的评估指标, 包括 AUC 值、Recall、G-mean 值[38]。其中, AUC 值是受试者工作特征曲线(receiver operating characteristic curve, ROC 曲线)下方的面积, 能够衡量模型整体区分能力; Recall 是不合格样本的召回率, 重点评估不合格样本的检出效能; G-mean 值是召回率(对不合格样本的识别能力)和特异度(对合格样本的识别能力)的几何平均, 能够评估模型平衡正负例的分类性能。另外, 本研究创新性地引入了业务成本指标[39], 设定假阳性(误判合格样本)成本为 1, 假阴性(漏检不合格样本)成本为 50, 以业务成本驱动阈值优化。希望通过业务成本指标尽可能把不合格样本都检测出来, 以免不合格样本漏检造成的损失。计算公式见(13)~(15):

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

$$\text{G-mean}=\sqrt{\frac{\text{TP}}{\text{TP}+\text{FN}}\times\frac{\text{TN}}{\text{TN}+\text{FP}}} \tag{14}$$

$$\text{业务成本}=1\times\text{FP}+50\times\text{FN} \tag{15}$$

式中: TP 指真实样本为不合格, 预测结果为不合格的样本数; TN 指真实样本为合格, 预测结果为合格的样本数; FP 为真实样本为合格, 预测结果为不合格的样本数; FN 为真实样本为不合格, 预测结果为合格的样本数。

1.4　数据处理

运用 WPS Office 12.1.0 软件统计并分析食用农产品的抽检数据, 在 Python 3.12.6 环境下建立食用农产品抽检结果的预测模型。本研究中的描述性统计表格由 WPS Office 12.1.0 制作, 模型性能图基于 Python 3.12.6 的 Matplotlib 库生成。

2　结果与分析

2.1　食用农产品的抽检结果

2.1.1　不同品种食用农产品的抽检结果

2022—2024 年福建省地市级食用农产品共抽检 75763 批次, 其中不合格食用农产品 2558 批次, 总体不合格率为 3.38%。按照食用农产品亚类来分析, 根茎类和薯芋类蔬菜的抽检不合格率最高(14.48%), 其中主要不合格的根茎类和薯芋类蔬菜品种包括生姜、山药等; 豆类蔬菜的抽检不合格率排名第二(11.67%), 其中主要不合格的豆类蔬菜品种包含豇豆、食荚豌豆等。热带和亚热带水果的抽检不合格率排名第三(6.80%), 其中主要不合格的热带和亚热带水果的品种包含荔枝、橄榄、芒果等。水生类蔬菜、鲜食用菌、芸薹属类蔬菜未检出不合格, 见表 2。导致这些食用农产品品种不合格率较高的原因有: 重金属污染、滥用禁用农药、常规农药残留超标、违规使用食品添加剂等。

表 2　不同食用农产品品种的抽检结果

Table 2　Sampling inspection results of different varieties of edible agricultural products

食用农产品大类	食用农产品亚类	抽检总批次数/批	不合格批次数/批	不合格率/%
蔬菜类	豆类蔬菜	2691	314	11.67
	根茎类和薯芋类蔬菜	4862	704	14.48
	瓜类蔬菜	1189	5	0.42
	鳞茎类蔬菜	2906	182	6.26
	茄果类蔬菜	6196	202	3.26
	水生类蔬菜	66	0	0
	叶菜类蔬菜	8168	181	2.22
	芸薹属类蔬菜	979	0	0
	豆芽	1435	13	0.91
	鲜食用菌	374	0	0
水果类	柑橘类水果	5045	54	1.07
	瓜果类水果	510	1	0.20
	核果类水果	699	1	0.14
	浆果和其他小型水果	2399	71	2.96
	热带和亚热带水果	5354	364	6.80
	仁果类水果	1416	1	0.07
肉类	畜肉	7022	12	0.17
	禽肉	4215	14	0.33
	其他禽畜副产品	763	5	0.66

表 2(续)

食用农产品大类	食用农产品亚类	抽检总批次数/批	不合格批次数/批	不合格率/%
水产类	贝类	2555	19	0.74
	淡水产品	3907	109	2.79
	海水产品	5563	151	2.71
	其他水产品	1172	63	5.38
豆类	豆类	134	1	0.75
生干坚果籽类	生干坚果籽类	1055	50	4.74
蛋类	鲜蛋	5088	41	0.81
合计		75763	2558	3.38

2.1.2 不同抽样月份食用农产品的抽检结果

食用农产品抽检覆盖了 1—12 月份, 4—9 月份食用农产品抽检的高峰期, 抽检总批次数较多。其中, 3 月份和 12 月份的食用农产品抽检不合格率相对较高, 分别达到了 4.18%和 4.28%, 4 月份和 9 月份的食用农产品不合格率相对较低, 分别为 2.99%和 2.83%, 见表 3。

表 3 不同月份食用农产品抽检结果

Table 3 Sampling inspection results of edible agricultural products in different months

月份/月	抽检总批次数/批	不合格批次数/批	不合格率/%
1	1124	35	3.11
2	1502	58	3.86
3	5138	215	4.18
4	7426	222	2.99
5	8914	283	3.17
6	8723	281	3.22
7	9561	366	3.83
8	9485	302	3.18
9	8211	232	2.83
10	7260	242	3.33
11	4754	165	3.47
12	3665	157	4.28
合计	75763	2558	3.38

2.1.3 不同抽样场所食用农产品的抽检结果

食用农产品的抽检涉及流通环节、餐饮环节和网络环节。在流通环节上抽检的食用农产品批次数最多(60854 批), 占总抽检批次的 80.32%, 且流通环节总体不合格率最高(3.65%)。从抽样场所分析, 农贸市场抽检的食用农产品不合格率最高(4.20%), 其次是生鲜零售门店(3.99%)。抽检不合格率相对较低的分别是餐饮环节中的单位食堂(1.37%)和网络环节中的网店(2.73%), 见表 4。食用农产品的抽检在流通环节上出现较高的不合格率可能是由于流通环节的食用农产品相对其他环节来说, 来源更加分散, 且供应商众多, 其监管难度更大。

表 4 不同抽样场所食用农产品的抽检结果

Table 4 Sampling inspection results of edible agricultural products from different sampling locations

抽样环节	抽样场所	抽检总批次数/批	不合格批次数/批	不合格率/%
流通	超市	39658	1378	3.47
	生鲜零售门店	15369	613	3.99
	农贸市场	4952	208	4.20
	其他	875	25	2.86
	小计	60854	2224	3.65
餐饮	单位食堂	7083	97	1.37
	餐饮店	5627	177	3.15
	小计	12710	274	2.16
网络	网店	2199	60	2.73
合计		75763	2558	3.38

2.2 基于多维时空风险特征的预测模型结果

本研究系统评估了无采样、过采样、组合采样和多维时空风险特征工程 4 种数据处理方法在食用农产品抽检预测机器学习模型上的性能。综合考量模型的综合性能(AUC 值)、关键风险识别能力(Recall)、不平衡数据稳健性(G-mean 值)及业务成本, 并对实验结果进行深度剖析。

如图 1 所示, 在 3 个模型中, 无采样方法虽在 AUC 值上表现尚可(0.821~0.841), 但由于食用农产品抽检结果数据是极端不平衡的数据类型, 在原始数据未做处理的情况下, 其对多数类(合格样本)的过度偏向牺牲了对少数类(不合格样本)的识别。过采样与组合采样方法在提升 Recall 的同时, 普遍导致了 AUC 值的显著下降(例如, 过采样处理下 RF 的 AUC 值从 0.821 降至 0.767, 下降了 6.58%)。然而, 多维时空风险特征工程方法成功打破了这一困境, 在 3 种模型上均取得了最高的 AUC 值(0.843~0.845), 不仅显著超越过采样和组合采样方法, 甚至较无采样方法亦有提升。这证明多维时空风险特征工程处理方法在有效应对类别不平衡的同时, 增强了模型的综合泛化能力和区分能力。

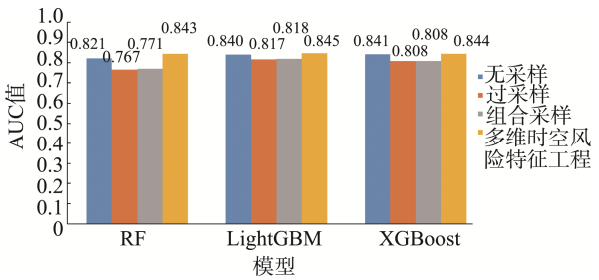


图 1 不同数据处理方法下 3 种模型的 AUC 值

Fig.1 AUC values of 3 kinds of models under different data processing approaches

如图 2 所示, 无采样方法 Recall 极低(均低于 0.2), 意味着超过 80%的风险样本被模型遗漏, 在实际应用中是不可接受的。过采样和组合采样将 Recall 提升至 0.574~0.766,

虽有改善，但仍有较高风险未被发现。相比之下，多维时空风险特征工程方法在 3 个模型上都有良好的表现。特别是 XGBoost 模型中，其 Recall 达到了 0.887，相较于过采样(Recall=0.670)和组合采样(Recall=0.629)有显著的提升，升幅分别高达 32.4%和 41.0%。此外，多维时空风险特征工程方法与 LightGBM 模型结合时，其 Recall 亦达到 0.881，同样表现出良好的性能。多维时空风险特征工程方法在 3 个模型上都能较大程度上降低不合格样本的漏检率。

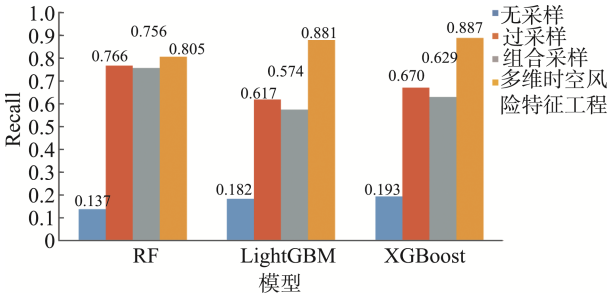


图 2　不同数据处理方法下 3 种模型的 Recall

Fig.2　Recall of 3 kinds of models under different data processing approaches

如图 3 所示，无采样方法因 Recall 极低，导致 G-mean 值也处于低水平(小于 0.44)。过采样和组合采样方法通过提升 Recall 改善了 G-mean 值(0.671~0.731)，但未能达到最优。多维时空风险特征工程方法在此指标上同样表现卓越，其中多维时空风险特征工程方法与 RF 组合的 G-mean 值达到了 0.764，为所有组合中的最高值。其次是多维时空风险特征工程方法与 LightGBM 组合，其 G-mean 值为 0.741，表现亦较为优异。这表明多维时空风险特征工程方法能最好地平衡对不合格样本的召回率与对合格样本的准确率，有效避免了在提升 Recall 时可能引发的误报激增问题。

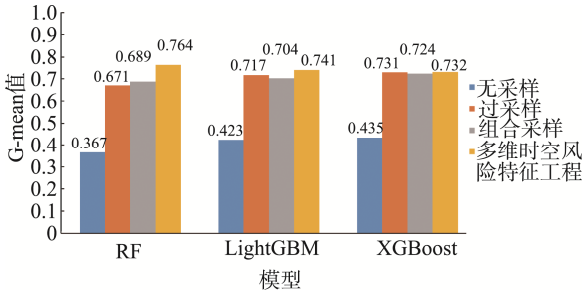


图 3　不同数据处理方法下 3 种模型的 G-mean 值

Fig.3　G-mean values of 3 kinds of models under different data processing approaches

如图 4 所示，多维时空风险特征工程方法在 3 个模型中业务成本均实现了明显下降。无采样方法因极高的漏检率，导致业务成本高昂(大于 20000)。过采样和组合采样方法将业务成本控制在 11401~12891 之间。然而，多维时空风险特征工程方法在 3 个模型中的业务成本压降至 9025 及以下。其中，多维时空风险特征工程方法与 LightGBM 组合以 8561 的成本最低，相较于其他方法中的最优值(过采样与 XGBoost 组合的 11401)，成本再度降低了 2840，降幅达 24.9%。

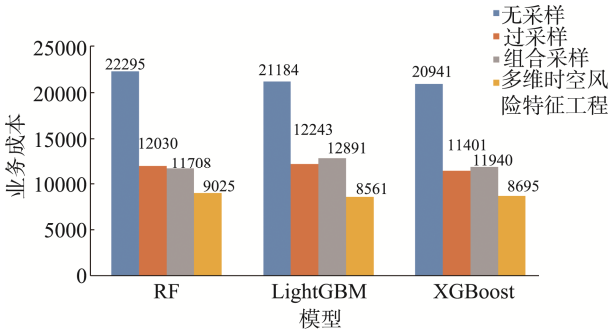


图 4　不同数据处理方法下 3 种模型的业务成本

Fig.4　Misclassification costs of 3 kinds of models under different data processing approaches

综合分析表明，相较于无采样、过采样和组合采样方法，多维时空风险特征工程在 AUC 值、Recall、G-mean 值及业务成本 4 项核心指标上均实现了全面而均衡的领先。另外，在模型优选方面，图 5 混淆矩阵直观呈现了模型在测试集上的分类细节：在全部测试样本中，多维时空风险特征工程与 LightGBM 组合模型正确识别出 451 件不合格样本，同时对合格样本保持较高的识别准确率，体现了其在降低漏检风险与控制误报成本之间的有效平衡。图 6 ROC 曲线则以假阳性率为横轴、真阳性率为纵轴，系统呈现了模型在各类识别阈值下的综合性能。由图 6 可见，多维时空风险特征工程与 LightGBM 组合模型的 ROC 曲线在绝大部分识别阈值下均位于其他模型曲线上方，表明该组合模型对食用农产品抽检数据具有更加优异的整体识别能力。同时综合其他指标，多维时空风险特征工程与 LightGBM 组合模型在 AUC 值(0.845)和业务成本(8561)上均达到最优水平，且该组合的 Recall 与 G-mean 值分别为 0.881 和 0.741，整体表现亦较为优异，是实际应用场景中的优选方案。

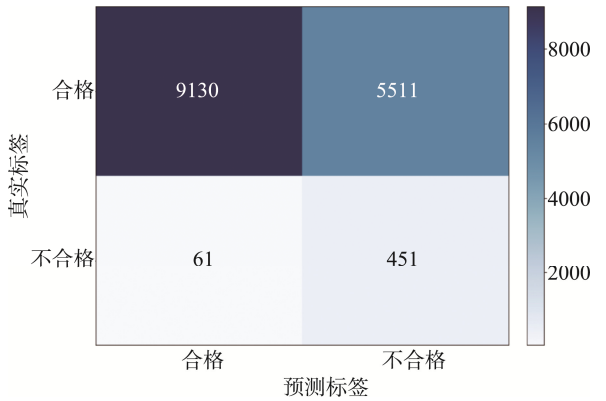


图 5　多维时空风险特征工程与 LightGBM 模型组合的混淆矩阵

Fig.5　Confusion matrix for the combination of multidimensional spatiotemporal risk feature engineering and LightGBM model

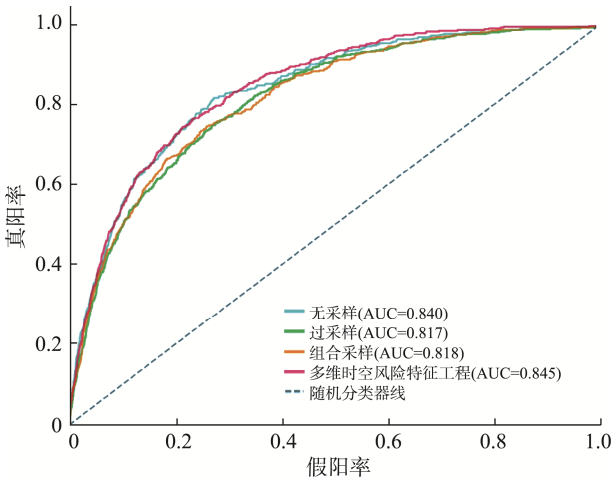


图 6 不同数据处理方法与 LightGBM 模型组合的 ROC 曲线对比图

Fig.6 Comparison diagram of ROC curves for combinations of different data processing methods and the LightGBM model

3 结论与讨论

本研究统计分析了福建省地级市 2022—2024 年 75763 条食用农产品监督抽检数据，发现数据总体的不合格率为 3.38%。其中，根茎类和薯芋类蔬菜的抽检不合格率最高(14.48%)。从抽检月份分析，3 月份和 12 月份的食用农产品抽检不合格率相对较高，分别达到了 4.18%和 4.28%。从抽样场所分析，农贸市场抽检的食用农产品不合格率最高(4.20%)。本研究创新性地引入了多维时空风险特征工程的方法，综合运用指数衰减权重、空间邻近权重与因果去偏策略等多元数学手段，在原始数据特征基础上，从时间、空间、品种及场所等维度构建表征复杂风险演化规律的新特征变量，最后将重构后的数据集应用于 LightGBM 模型上训练和预测。通过与无采样、过采样及组合采样方法的对比，结果表明，多维时空风险特征工程显著增强了模型对不合格样本的识别能力，有效提升了食用农产品抽检结果的预测性能。这一方法突破了传统采样技术仅从数据层面缓解不平衡问题的局限，为面向高度不平衡的食品安全抽检数据建模提供了从特征层面入手的新思路。

基于多维时空风险特征工程的 LightGBM 预测模型的 AUC 值(0.845)、Recall (0.881)、G-mean 值(0.741)、业务成本(8561)，表明该模型预测性能良好，能够较为准确地预测食用农产品抽检结果中的不合格样本。运用该模型，食品监管部门可以输入与食用农产品抽检相关的特征信息，预测食用农产品的抽检结果是否合格。另外，模型内嵌的时空风险评估逻辑，能动态适应不同地区和时期的波动规律，再加上 LightGBM 模型固有的训练效率优势，使基于多维时空风险特征工程的 LightGBM 预测模型能够精准、高效地满足食用农产品抽检预测的实际业务需求，能够为差异化、动态化的食用农产品监督抽检方案制定及食用农产品安全风险预警提供技术支撑。未来，研究可进一步探索特征工程的自动优化机制，并开发面向流式数据的在线学习版本，以持续提升模型在实时监管场景中的适应能力与决策支持水平。

参考文献

[1] 赵红磊, 祁立新, 刘欣, 等. 2021—2023 年国家食品安全监督抽检不合格结果分析[J]. 食品安全导刊, 2025(1): 45–47.
ZHAO HL, QI LX, LIU X, *et al*. Analysis on unqualified results of national food safety supervision and sampling inspection from 2021 to 2023 [J]. Journal of Food Safety Guide, 2025(1): 45–47.

[2] 国家市场监督管理总局通告. 市场监管总局关于2024年第三季度市场监管部门食品安全监督抽检情况的通告(2024 年第 26 号)[EB/OL]. (2024-11-01) [2024-11-07]. https://www.samr.gov.cn/spcjs/xxfb/art/2024/art_f66f2bd0b72144329220963657ae5c6b.html
Circular of the State Administration for Market Regulation. Notice of the State Administration for Market Regulation on the food safety supervision and sampling inspection situation of market regulation departments in the third quarter of 2024 (No.26 of 2024) [EB/OL]. (2024-11-01) [2024-11-07]. https://www.samr.gov.cn/spcjs/xxfb/art/2024/art_f66f2bd0b72144329220963657ae5c6b.html

[3] 国家市场监督管理总局通告. 市场监管总局关于2024年第一季度市场监管部门食品安全监督抽检情况的通告(2024 年第 16 号)[EB/OL]. (2024-04-30) [2024-05-07]. https://www.samr.gov.cn/spcjs/xxfb/art/2024/art_87f716b6a1634c70b8b7d6b10dd42eaa.html
Circular of the State Administration for Market Regulation. Notice of the State Administration for Market Regulation on the food safety supervision and sampling inspection situation of market regulation departments in the first quarter of 2024 (No.16 of 2024) [EB/OL]. (2024-04-30) [2024-05-07]. https://www.samr.gov.cn/spcjs/xxfb/art/2024/art_87f716b6a1634c70b8b7d6b10dd42eaa.html

[4] 董曼, 李洁, 尹佳, 等. 基于贝叶斯网络的酱卤肉制品安全预警研究[J]. 食品安全质量检测学报, 2021, 12(1): 27–33.
DONG M, LI J, YIN J, *et al*. Research on safety early warning of sauced meat products based on Bayesian network [J]. Journal of Food Safety & Quality, 2021, 12(1): 27–33.

[5] LIU N, BOUZEMBRAK Y, BULK LM, *et al*. Automated food safety early warning system in the dairy supply chain using machine learning [J]. Food Control, 2022, 136: 108872.

[6] LI K, ONODA K, KUMAZAKI T. Modeling taste score of rice by data mining from physicochemical parameters: AIP conference proceedings [C]. Melville: AIP Publishing, 2017.

[7] 唐昊, 梁江, 吴难, 等. 基于 Smote-KNN 的小麦 8 种真菌毒素共污染特征气候分类模型研究[J]. 中国食品卫生杂志, 2023, 35(6): 807–812.
TANG H, LIANG J, WU N, *et al*. Climate classification model of co-occurrence characteristics of 8 kinds of mycotoxins in wheat based on Smote-KNN [J]. Chinese Journal of Food Hygiene, 2023, 35(6): 807–812.

[8] 宋玉洁. 基于食品抽检数据的食品安全风险分析及预警系统构建研究[D]. 烟台: 烟台大学, 2024.
SONG YJ. Research on food safety risk analysis and early warning system construction based on food sampling data [D]. Yantai: Yantai University, 2024.

[9] 游清顺, 王建新, 张秀宇, 等. 基于支持向量机的食品安全抽检数据分析方法[J]. 软件工程, 2019, 22(2): 29–31.
YOU QS, WANG JX, ZHANG XY, *et al*. SVM-based analysis on food safety sampling and inspection data [J]. Software Engineering, 2019, 22(2): 29–31.

[10] 杨鸿雁, 田英杰. 机器学习在食品安全风险预警及抽检方案制订中的应用研究[J]. 管理评论, 2022, 34(11): 315–323.
YANG HY, TIAN YJ. Research on the application of machine learning in

food safety risk warning and sampling inspection plan formulation [J]. Management Review, 2022, 34(11): 315–323.

[11] 刘玉航, 曲媛, 蒋嘉铭, 等. 基于优化随机森林算法预测食品检验不合格指标[J]. 食品安全质量检测学报, 2021, 12(18): 7467–7472.
LIU YH, QU Y, JIANG JM, *et al*. Prediction of unqualified index of food inspection based on optimized random forest algorithm [J]. Journal of Food Safety & Quality, 2021, 12(18): 7467–7472.

[12] 周洁红, 魏珂, 金宇, 等. 基于机器学习的食品安全风险预测与监管政策启示——以生鲜水产品为例[J]. 农业经济问题, 2024(5): 4–19.
ZHOU JH, WEI K, JIN Y, *et al*. Food safety risk prediction and regulatory policy implications based on machine learning: Evidence from fresh aquatic products [J]. Issues in Agricultural Economy, 2024(5): 4–19.

[13] 王小艺, 王姿懿, 赵峙尧, 等. 集成改进 AHP 与 XGBoost 算法的食品安全风险预测模型: 以大米为例[J]. 食品科学技术学报, 2022, 40(1): 150–158.
WANG XY, WANG ZY, ZHAO ZY, *et al*. A food safety risk forecast model integrated with improved AHP and XGBoost algorithm: A case study of rice [J]. Journal of Food Science and Technology, 2022, 40(1): 150–158.

[14] 穆书敏, 陈锂, 尹佳, 等. 基于集成学习优化的肉制品安全风险等级预警分析[J]. 现代食品科技, 2023, 39(8): 273–286.
MU SM, CHEN L, YIN J, *et al*. Safety risk level early warning of meat products based on optimized ensemble learning [J]. Modern Food Science and Technology, 2023, 39(8): 273–286.

[15] 刘广杰. 基于数据驱动的食用农产品安全风险预警研究[D]. 保定: 河北大学, 2024.
LIU GJ. Research on data-driven safety risk warning for edible agricultural products [D]. Baoding: Hebei University, 2024.

[16] 宇盛好, 周子文, 姚烨, 等. 基于贝叶斯优化轻量级梯度提升机模型预测淡水鱼中恩诺沙星抽检结果[J]. 食品安全质量检测学报, 2024, 15(22): 301–309.
YU SH, ZHOU ZW, YAO Y, *et al*. Prediction of enrofloxacin sampling inspection results in freshwater fish based on Bayesian optimization light gradient boosting machine model [J]. Journal of Food Safety & Quality, 2024, 15(22): 301–309.

[17] 钱洁. 基于机器学习的食品添加剂安全性与稳定性虚拟评价研究[D]. 长沙: 中南林业科技大学, 2023.
QIAN J. Study on virtual evaluation of food additives safety and stability based on machine learning [D]. Changsha: Central South University of Forestry and Technology, 2023.

[18] 周子文, 范志仪, 彭少杰. 基于粒子群优化集成学习算法堆叠模型预测蔬菜中倍硫磷的抽检结果[J]. 食品安全质量检测学报, 2025, 16(5): 187–196.
ZHOU ZW, FAN ZY, PENG SJ. Stacked generalization model prediction of fenthion spot-check results in vegetables based on particle swarm optimization ensemble learning algorithm [J]. Journal of Food Safety & Quality, 2025, 16(5): 187–196.

[19] 陈晨, 尹佳, 董曼, 等. 基于灰色数据预处理的 WD-LSTM 模型对乳制品质量安全风险的预测预警分析[J]. 现代食品科技, 2023, 39(1): 300–310.
CHEN C, YIN J, DONG M, *et al*. Predictive early warning analysis of dairy product quality and safety risks based on grey data pre-processin using the WD-LSTM model [J]. Modern Food Science and Technology, 2023, 39(1): 300–310.

[20] WU M, LIU W, ZHENG S. Intelligent food safety: A prediction model based on attention mechanism and reinforcement learning [J]. Applied Artificial Intelligence, 2024, 38(1): 2379731.

[21] 罗晓清, 郭林, 杨雨萌, 等. 基于时空 Transformer 模型的食品监督抽检分类预测研究[J]. 中国食品学报, 2024, 24(11): 1–9.
LUO XQ, GUO L, YANG YM, *et al*. Research on classification prediction of food supervision and sampling based on spatiotemporal transformer model [J]. Journal of Chinese Institute of Food Science and Technology, 2024, 24(11): 1–9.

[22] 王乐, 韩萌, 李小娟, 等. 不平衡数据集分类方法综述[J]. 计算机工程与应用, 2021, 57(22): 42–52.
WANG L, HAN M, LI XJ, *et al*. Review of classification methods for unbalanced data sets [J]. Computer Engineering and Applications, 2021, 57(22): 42–52.

[23] ISLAM A, BELHAOUARI SB, REHMAN AU, *et al*. Knnor: An oversampling technique for imbalanced datasets [J]. Applied Soft Computing, 2022, 115: 108288.

[24] 段镁钰. 基于特征关联性的特征构造算法研究及应用[D]. 长春: 吉林大学, 2023.
DUAN MY. Research and application of feature construction based on inter-feature correlations [D]. Changchun: Jilin University, 2023.

[25] YADAV D, RANI D, VERMA OP. Impact of feature selection and feature engineering in prediction of cardiovascular diseases [J]. Computers in Biology and Medicine, 2025, 197: 111027.

[26] CHAPFUWA P, TAO C, LI C, *et al*. Calibration and uncertainty in neural time-to-event modeling [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 34(4): 1666–1680.

[27] ROY B, STEPISNIK T, VENS C, *et al*. Survival analysis with semi-supervised predictive clustering trees [J]. Computers in Biology and Medicine, 2022, 141: 105001.

[28] MICHELS K, SONG M, WILLETT W, *et al*. Latency estimation for chronic disease risk: A damped exponential weighting model [J]. European Journal of Epidemiology, 2020, 35(9): 807.

[29] RYAN A, AI C, FITZPATRICK C, *et al*. Crash proximity and equivalent property damage calculation techniques: An investigation using a novel horizontal curve dataset [J]. Accident Analysis & Prevention, 2022, 166: 106550.

[30] CHEN L, XIE M, LI G, *et al*. A spatial source-oriented and probability-based risk-assessment framework for heavy metal and PAH contamination of urban soils in Guangzhou, China [J]. Journal of Hazardous Materials, 2025, 482: 136500.

[31] VELICKOVIC P, CUCURULL G, CASANOVA A, *et al*. Graph attention networks: Proceedings of the 6th international conference on learning representations [C]. Vancouver: Open Review.net, 2018.

[32] KIPF TN, WELLING M. Semi-supervised classification with graph convolutional networks: Proceedings of the 6th international conference on learning representations [C]. Vancouver: Open Review.net, 2017.

[33] PASTOR-SATORRAS R, VESPIGNANI A. Epidemic spreading in scale-free networks [J]. Physical Review Letters, 2001, 86(14): 3200.

[34] SAHNEH FD, SCAGLIO C. Epidemic spread in human networks: 2011 50th IEEE conference on decision and control and european control conference [C]. Piscataway: IEEE, 2011.

[35] ZHANG C, SONG D, HUANG C, *et al*. Heterogeneous graph neural network: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining [C]. Piscataway: IEEE, 2019.

[36] JOHANSSON FD, SHALIT U, SONTAG D. Learning representations for counterfactual Inference [C]. New York: JMLR, 2016.

[37] 宋雨丹. 基于图神经网络的伪装识别欺诈检测方法研究[D]. 桂林: 广西师范大学, 2025.
SONG YD. Research on graph neural network fraud detection for camouflage recognition [D]. Guilin: Guangxi Normal University, 2025.

[38] 管维正. 基于不平衡数据分类算法研究及其应用[D]. 南京: 南京邮电大学, 2021.
GUAN WZ. Research and application of classification algorithm based on unbalanced data [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2021.

[39] HE H, GARCIA EA. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263–1284.

(责任编辑: 安香玉　韩晓红)