

基于集成学习优化的肉制品安全风险等级预警分析

穆书敏¹, 陈锂², 尹佳², 郭鹏程², 陈晨¹, 董曼², 赵锦³, 徐晴雪³, 文红^{2*}, 桂预风^{1*}

(1. 武汉理工大学理学院, 湖北武汉 430070) (2. 湖北省食品质量安全监督检验研究院, 湖北省食品质量安全检测工程技术研究中心, 国家市场监管重点实验室(动物源性食品中重点化学危害物检测技术), 湖北武汉 430075) (3. 石首市公共检验检测中心, 湖北荆州 434200)

摘要: 该研究依据 2013~2017 年肉制品抽检数据构造了 5 个安全风险等级, 使用特征构造及独热编码进一步关联与肉制品安全相关的影响因素, 构建极端梯度提升树算法 (Extreme Gradient Boosting, XGBOOST) 研究食品生产过程各类因素对于食品安全风险等级的影响程度, 并使用多个指标评价模型。此外通过上采样解决样本不平衡问题、贝叶斯优化调节超参数, 来提高模型性能及分类效果。相较于模型决策树 (Decision Tree, DT) 和随机森林 (Random Forest, RF), XGBOOST 模型在肉制品安全风险等级分类中的表现效果最佳。研究结果表明, 食品生产过程环节错综复杂, 使用 one-hot encoding 处理后的模型能够有效判断出各类因素对于食品安全风险等级的影响程度, 集成模型中 RF 的学习效果比较稳定, XGBOOST 经过参数调节后准确率等指标得到有效的提升且优于 RF。不同采样下 XGBOOST 的平均精确率均能达到 89.14%, 平均 F1 值为 88.59%, 说明 XGBOOST 在肉制品安全风险等级预警中适用性, 为日常抽检提供技术指导。

关键词: 食品安全风险; 独热编码; 决策树; 集成学习; 极端梯度提升树; 随机森林

文章编号: 1673-9078(2023)08-273-286

DOI: 10.13982/j.mfst.1673-9078.2023.8.0878

Safety Risk Level Early Warning of Meat Products Based on Optimized Ensemble Learning

MU Shumin¹, CHEN Li², YIN Jia², GUO Pengcheng², CHEN Chen¹, DONG Man², ZHAO Jin³, XU Qingxue³, WEN Hong^{2*}, GUI Yufeng^{1*}

(1.School of Sciences, Wuhan University of Technology, Wuhan 430070, China) (2.Hubei Provincial Institute for Food Supervision and Test, Key Laboratory of Detection Technology of Focus Chemical Hazards in Animal-derived Food for State Market Regulation, Hubei Provincial Engineering and Technology Research Center for Food Quality and Safety Test, Wuhan 430075, China) (3.Shishou City Public Inspection Test Centre, Jingzhou 434200, China)

Abstract: Five safety risk levels were established based on the detection data of meat products sampled between 2013 and 2017. Feature construction and one-hot encoding were used to further correlate factors relevant to meat product safety. An extreme gradient boosting (XGBOOST) model was established to study the influence levels of various factors during food production on the safety risk level; subsequently then multiple indices were used to evaluate the model. In addition, sample imbalance problem was solved by upsampling, and the hyperparameters were adjusted by Bayesian optimization to improve the model performance and classification results. Simultaneously, the model constructed was compared with the decision tree (DT) and random forest (RF) methods to evaluate their classification performance. The XGBOOST outperformed others in classifying the safety risk levels of meat products. Food production processes are complex, and this study shows that model processing with one-hot encoding could effectively

引文格式:

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 Shumin, CHEN Li, YIN Jia, 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

收稿日期: 2022-07-13

基金项目: 国家重点研发计划项目 (2018YFC1603602)

作者简介: 穆书敏 (1996-), 女, 硕士研究生, 研究方向: 大数据分析与管理, E-mail: shuminmu@whut.edu.cn

通讯作者: 文红 (1971-), 女, 硕士, 教授级高级工程师, 研究方向: 食品安全风险评估与预警, E-mail: 604962461@qq.com; 共同通讯作者: 桂预风 (1963-), 男, 博士, 教授, 研究方向: 应用统计与机器学习, E-mail: guiyufeng@whut.edu.cn

identify the influence levels of various factors on food safety. Moreover, the result suggested that XGBOOST performs better in terms of total accuracy and other indices after parameter adjustment, compared to the RF model, while RF had the most stable learning performance. The average of accuracy and F1 score can reach 89.14% and 88.59%, respectively, under different sampling. The results suggest that XGBOOST can be applied to determine safety risk levels of meat products and provide technical support for daily supervision.

Key words: food safety risk; one-hot encoding; decision tree; ensemble learning; extreme gradient boosting; random forest

随着我国全面建成小康社会的目标实现,人民的生活质量不断提高,人民群众对食物的需求也随之从温饱转为安全和营养。近年频发的食品安全问题事件不仅严重影响群众身体健康,还关乎社会稳定,食品安全问题已成为我国的民生问题^[1]。食品安全风险管理是食品安全工作的主要内容,涉及原材料提供、加工、生产、销售、运输等多个环节,加强各环节的管控是降低食品安全风险的有效途径^[2-4]。据调查,肉制品在我国居民日常食品消费中所占的比重越来越大,作为具有高营养价值的重要日常消费食品,也是具有较高安全风险和隐患的食品之一,保障肉制品质量安全满足消费者的需求刻不容缓^[5]。但当前肉制品供应链环节复杂多变,经手的企业公司较多,且各企业之间无法相互监控,信息不对称使得肉制品安全问题缺乏可溯源性,难以有效的定位和管理^[6]。囤积的肉制品检测数据具有数据量多信息量少的特征,挖掘该历史数据中的信息研究肉制品检测数据的食品安全风险评价预警反馈机制具有重要意义。目前数据挖掘工作受到学者的广泛关注,该研究对保障人民身体健康、实现肉制品的源头防控和主动预防具有重要意义。

大数据时代的到来推动了机器学习的发展,机器学习被广泛应用于各领域并取得了良好效果^[7-9]。根据目前食品安全风险评估和预警研究现状,将经典方法与机器学习模型进行组合往往取得较好的效果^[10-14]。而集成学习考虑综合不同机器学习算法的优点,将多种简单的机器学习模型结合起来,通过一定的策略形成一个集成学习模型^[15,16],在食品生产以及其他领域有广泛应用^[17-20]。荷兰食品和消费品安全局在 2015 年起草了关于红肉链的 DT 分类模型, Van Asselt 等^[21]使用三个 DT 分别对红肉中化学物质(禁用物质、污染物、兽药)进行分类,研究表明 DT 有助于化学物质的分类,可作为风险监测的计划工具。王小艺等^[22]使用层次分析法(AHP)结合 XGBOOST 算法对大米安全风险进行预测,研究对比了支持向量机(SVM)、最邻近分类算法(KNN)、长短期记忆人工神经网络(LSTM),XGBOOST 显示出较好的平稳性和准确性。Xu 等^[23]使用 RF 对家禽农场和加工环境中的菌落进行预测,该研究有效的帮助牧场家禽种植场指定食品安全控制策略。

本文基于全国肉制品抽检数据,采用独热编码对不平衡数据进行数据处理,以决策树为基础,分别构建决策树(DT)、极端梯度提升树算法(XGBOOST)和随机森林(RF)集成学习算法,建立食品安全风险评价预警模型,并通过准确率、精确度、召回率等一系列指标说明模型的应用效果。该研究有助于监管部门提高对于肉制品安全的监管效率,有助于构建更加完善的肉制品质量安全监督管理体系,也为集成学习在食品安全问题中的进一步应用提供了参考。

1 材料与方法

1.1 数据处理

1.1.1 数据来源

本文数据来源市场监督管理局^[24]公开的 2013~2017 年全国(除港澳台地区)肉制品抽检数据和 HBQT 检验检测数据。原始数据的信息包括抽样地点、单位地点、被抽样单位名称、抽样编号、样品名称、生产日期、样品规格、保质期、检验项目、检验结果等 79 项内容,涵盖了熟肉制品和预制肉制品等多种肉制品类别,共计约 4 万条数据。本文以肉制品数据检验结果的综合风险等级为研究对象,通过与各产品的生产信息相关联来构造对于肉制品质量安全风险等级预警机制。但原始数据中如“商标”、“检验机构”等多列与研究无关,且部分列不能直接作为变量使用,需要对数据所含信息的内容进行整合、提取、清洗等处理后确定变量。

1.1.2 数据清洗

数据的质量和特征决定了机器学习的上限,而模型和算法只是在逼近这个上限^[8]。本研究中肉制品数据涵盖维度信息少,因此数据处理部分尤为关键。数据预处理的首要步骤为数据清洗,数据清洗主要包括删除重复值、统一规格、修正逻辑、数据压缩、补足缺失或空值、丢弃异常值等,该过程能有效地提高数据质量,保证数据的完整性、唯一性和合法性^[25]。然而数据清洗工作同时存在复杂繁琐、耗时长,且影响结果分析的问题需要解决。综上所述,本文对肉制品数据处理工作主要通过计算机程序化解决,部分结合人工处理。本研究具体数据处理流程如图 1 所示。

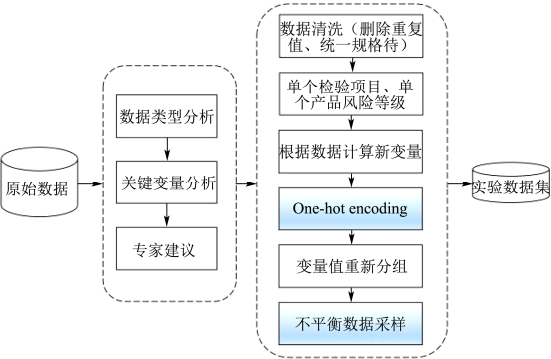


图 1 数据处理流程图

Fig.1 The flow diagram of data processing

选择与本研究中与肉制品安全风险等级预警有关的食物品种、生产省份、抽样省份、抽样地点、生产日期、保质期、抽样日期、是否大型生产企业、是否大型经营企业、是否进口、检验项目、检验结果作为数据源。其中食品品种、抽样地点、是否大型生产企业、是否大型经营企业、是否进口划分类别明确，且无缺失值，可直接作为输入变量。产品的生产日期和抽样日期记录形式有多种，需要进行统一，并将前者拆分为年（Year）、月（Month）两个变量便于特征重要性分析确定模型的输入变量。保质期中含有信息多、内容复杂，有 30 d、6 个月、12 个月（25℃以下真空包装）、1 年等多种记录，本研究中将保质期转化为以

天为单位并去温度、包装完整性等条件，如 12 个月（25℃以下真空包装）转化为 720 d，作为输入变量。检验项目的检验结果按照前期研究^[26]的处理方式进行标准化和归一化进行去量纲，用于划分及计算综合风险等级，作为模型的输出变量。

异常值检测中，一个产品的抽检时间为 2005 年 4 月 8 日，对此判断为录入时的人工错误，将其年份修改为 2015 而不作删除。数据处理过程严格遵循完整性、合法性、唯一性原则。原始数据 40 052 条经处理后为 38 315 条，利用率达到 96%。

1.1.3 获取等级数据

将抽样产品的检测结果进行分级区分有利于提高对于食品安全的敏感度和精准判断^[27,28]。本文研究对象为肉制品综合风险等级，该等级通过两步得到：以抽样产品编号作为每个产品的唯一编号，第一步将去量纲后的检测结果通过专家打分划分为 5 个等级，1~5 级分别表示安全、无警、轻度警、中度警、重度警；第二步将每一个产品的所有检测项目通过改进的 softmax 函数获取该产品的综合风险等级。综合风险等级用 level 表示，计算方法见公式（1），其预警程度和说明与检验项目一致。肉制品的项目分类和检验项目见表 1，检验项目等级划分标准见表 2。

表 1 肉制品检验项目

Table 1 Inspection items of meat products

项目分类	检验项目
质量指标	蛋白质、淀粉、过氧化值、三甲胺氮、食盐、酸价、脂肪
非食用物质	苏丹红Ⅰ、苏丹红Ⅱ、苏丹红Ⅲ、苏丹红Ⅳ、酸性橙Ⅱ
禁用兽药	克伦特罗、莱克多巴胺、氯霉素、沙丁胺醇
污染物	镉（以 Cd 计）、铬（以 Cr 计）、铅（以 Pb 计）、砷（以 As 计）、苯并[a]芘、N-二甲基亚硝胺
致病性微生物	大肠埃希氏菌、单核细胞增生李斯特氏菌、金黄色葡萄球菌、沙门氏菌
其他微生物	大肠菌落、菌落总数、商业无菌
食品添加剂	糖精钠、胭脂红、诱惑红、日落黄、赤藓红、柠檬黄、苋菜红、新红、苯甲酸、丁基羟基茴香醚、二丁基羟基甲苯山梨酸、特丁基对苯二酚、脱氢乙酸及其钠盐（以脱氢乙酸计）、亚硝酸盐残留量（以亚硝酸钠计）、山梨酸及其钾盐（以山梨酸计）、防腐剂混合使用时各自用量占其最大使用量的比例之和

表 2 检验项目风险等级划分标准

Table 2 Standard for risk classification of inspection items

风险等级	检验项目标准化、归一化后的值	预警程度	说明
1 级	[0,0.1]	安全	检验中无污染，各项指标均达标
2 级	(0.1,0.3]	无警	有少量危害物质
3 级	(0.3,0.7]	轻度警	危害物质在可接受范围内
4 级	(0.7,1)	中度警	危害物质一定程度影响健康
5 级	1	重度警	危害物质含量严重损害身体健康

注：表中[]为闭区间，包含端点值；（）为开区间，不包含端点值。

$$level = \operatorname{argmax}(w_i e^i) \quad (1)$$

式中:

i ——风险等级值, 取值为 $i=1, 2, 3, 4, 5$;

w_i ——每个产品中检验项目等级为 i 的占比。

值得注意的是, 综合风险等级加大了对严重损害身体健康的检验项目的惩罚力度, 若一个产品的任一检验项目风险等级为 5, 则产品综合风险等级为 5 级, 而不使用公式 (1) 计算。

1.1.4 特征构造

由于原数据中模型可用的特征较少, 故利用清洗后数据集中的日期及省份变量进行特征构造。研究表明^[18]通过多个字段构造新特征作为输入变量, 新特征能够包含无法直接获取的重要影响因素。结合生产日期、抽样日期和保质期重新构造新特征“保鲜度”, 用 $lasting$ 表示, 简记为 l , 具体计算方法见公式 (2)。

$$l_i = \begin{cases} 0, & \alpha_i = \beta_i \text{ 且 } period_i > 1 \\ 1, & \alpha_i = \beta_i \text{ 且 } period_i = 1 \\ (date_{\alpha_i} - date_{\beta_i}) / period_i, & \text{其他} \end{cases} \quad (2)$$

式中:

l_i ——自定义的第 i 个产品保鲜度;

α_i ——第 i 个产品的抽样日期;

β_i ——第 i 个产品的生产日期;

$period_i$ ——第 i 个产品的保质期。

保鲜度值分布在 $[0, 1]$ 区间时肉制品在保质期内, 当保鲜度大于 1 时为过期产品。

根据产品生产省份和抽样省份信息构造新特征“省份一致性”, 用 $province_consistance$ 表示, 若生产省份与抽样省份一致, 则记为 yes, 若不一致则记为 no。

1.1.5 独热编码 (One-Hot Encoding)

为了进一步探究因素中的不同取值对食品安全风险的影响, 采用 one-hot encoding 处理变量。one-hot encoding 又称一位有效编码, 可以处理非连续性数据特征, 不但能一定程度上扩充特征维度, 还可以处理缺失数据, 是将类别变量转换为机器学习算法易于利用的一种形式的过程^[29]。目前多用 one-hot encoding 处理特征作为输入变量来提高模型的分类精度、决策质量以及计算时间^[30-32]。本文采用 one-hot encoding 对变量特征进行处理, 不但能扩充数据特征的维度, 还能具体比较各取值对肉制品等级分类的影响程度, 但使用过程中需要慎重选择扩充的变量, 盲目扩充所有变量易造成维度灾难, 严重影响模型效果。为探究季节性因素对肉制品安全风险等级的影响, 使用 one-hot encoding 对月份 (month) 变量处理将其转化为二进制输入形式, 然后转化为稀疏 0, 1 矩阵作为输入变量。

处理后月份变量由 1 维增加到 12 维, 程序中默认从 0 开始编码, 故一月为 month_0, 二月为 month_1, 其余月份以此类推。表 3 给出 one-hot encoding 编码前后月份变量的比较。

1.1.6 数据集变量说明

结合肉制品数据特征和研究目的, 决策树模型具有较好的适用性, 本文优化模型均以决策树为基础构造。但部分特征的变量种类过多, 若按照原始类别建树易导致树过大、枝过多以及过拟合现象, 为了提高构建的基础树质量, 将部分变量的值重新分组后作为模型的输入变量, 将各生产省份根据国家地理区域将其分为八大类记作地区 (Areas) 变量, 其中生产地非中国的抽样产品一律归为境外类别。抽样地区以实际情况对餐馆、生产线等进行合并记作位置 (Location) 变量。分组后变量的取值均用英文字符表示。数据经过处理变换后对相应的符号含义以及模型中的变量符号相关说明见表 4。

1.1.7 数据重采样

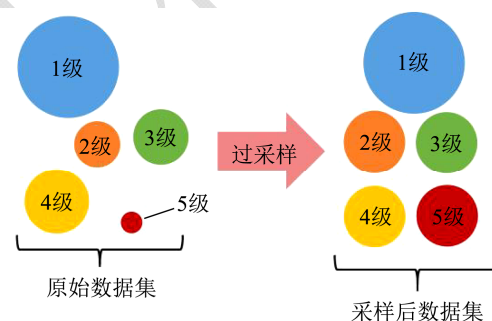


图 2 采样示意图

Fig.2 Demonstration of sampling

经上述一系列处理后得到模型的输入和输出变量, 可使用该数据集进行分类模型学习。分类模型基本假设样本类是平衡的, 而本文研究数据严重偏离平衡。不平衡样本下训练模型往往为达到更高的精确率, 偏向于多类别的样本而忽略少类别的样本^[33]。抽样肉制品数据中 1 级产品较多 (安全), 但需要重点关注对象为含有风险的产品, 特别是 5 级 (严重损害身体健康), 以不平衡数据建模缺乏实际意义且影响模型分类效果, 因此要解决不平衡数据问题。一般不平衡数据的处理方法有欠采样、过采样、阈值移动^[34]。但由于阈值移动过于复杂, 欠采样会损失大量样本信息, 故选择过采样。过采样又称上采样, 在本研究中即对除 1 级以外各等级的产品在原始数据中重复抽样以增大数据量, 使其能够充分挖掘少量样本中的信息。图 2 为本研究的采样示意图。

表 3 独热编码处理前后月份变量比较

Table 3 Comparison of monthly variables before and after one-hot encoding		
处理	输入变量	变量取值
不使用 one-hot encoding	month	一月、二月、三月、四月、五月、六月、七月、八月、九月、十月、十一月、十二月
使用 one-hot encoding	month_0、month_1、month_2、month_3、month_4、month_5、month_6、month_7、month_8、month_9、month_10、month_11	0、1

表 4 变量及符号说明

Table 4 Description of variables and symbols			
变量	符号	含义	类别/说明
地区 (areas*)	J	京津冀三省市	河北、天津、北京
	DB	东北三省一区	辽宁、吉林、黑龙江、内蒙古
	Z	中部六省	湖南、江西、山西、河南、湖北、山东
	XN	西南五省市	贵州、重庆、四川、云南、西藏
	XB	西北五省市	甘肃、新疆、宁夏、陕西、青海
	N	南部沿海四省市	广西、海南、广东、宁夏
	T	长三角四省市	安徽、上海、浙江、江苏
	W	境外	所有其他国家一律为境外
位置 (location*)	L1	菜市场	菜市场、农贸市场、批发市场
	L2	餐馆	大型餐馆、特大型餐馆、快餐店、小吃店、小型餐馆、中型餐馆
	L3	超市	小食杂店、超市、商场
	L4	成品库	成品库 (待检区)、成品库 (已检区)
	L5	生产线	半成品库、生产线、原辅料库
	L6	网购	网购
	L7	其他	除以上类别之外的均为其他
保鲜度 (lasting*)		保鲜度	(抽样日期-生产日期)/保质期
保质期		单个产品的保质期	1 天、4 天、7 天、180 天等
是否大型生产企业			是、否
是否大型经营企业			是、否
生产月份 (month_0、month_1、month_2、month_3、month_4、month_5、month_6、month_7、month_8、month_9、month_10、month_11*)		One-hot encoding 处理	一月、二月、三月、四月、五月、六月、七月、八月、九月、十月、十一月、十二月
食品品种		肉制品的种类	熏煮香肠、火腿制品、酱卤肉制品、腌腊肉制品、调理肉制品、熟肉干制品、熏烧烤制品、发酵肉制品、其他肉制品
省份一致性 (province_consistance*)		生产省份与 抽样省份一致性	yes、no
综合风险等级 (level*)		产品的综合风险等级	1、2、3、4、5 级
	RF	Radom Forest	集成学习 Bagging 模型
	XGBOOST	Extreme Gradient Boosting	集成学习 Boosting 模型

注：变量列中带有*注释的变量为根据选择后的数据源构造或计算的新变量，英文命名的变量为经相应处理后变作为模型的输入或输出变量。此表中变量为优化模型中的全部输入变量，另有变量选择时涉及的是否进口、年份 (Year) 和月份 (Month) 未列在此表中。

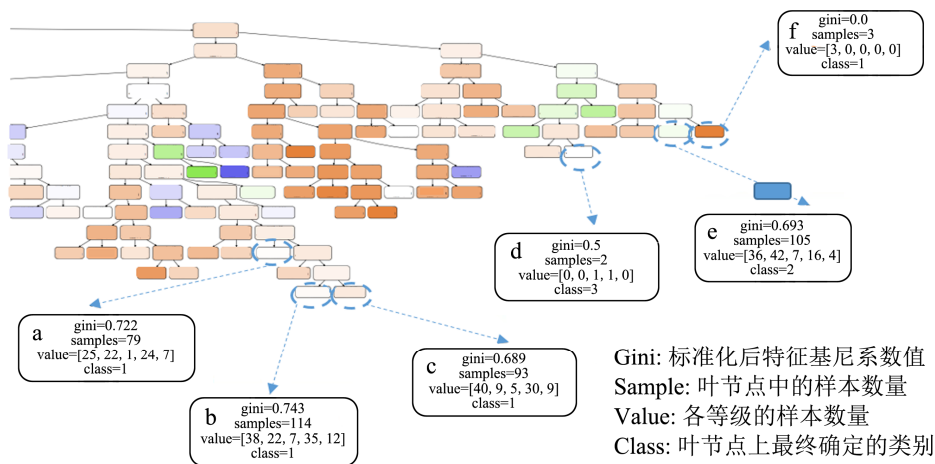


图 3 决策树部分可视化结果

Fig.3 A part of visualization results of DT

经清洗处理后的数据集中含 38 315 条数据，风险等级为 1 级的产品为 29 770 条，占中总体的 77.70%，2 级、3 级、4 级、5 级的数据分别为 1 001 条、1 686 条、4 713 条、1 145 条，分别占总体的 2.61%、4.40%、12.30%、2.99%。可见样本存在严重的数据不平衡现象。图 3 为不平衡样本数据集在决策树模型上的分类可视化结果（部分），由于 1 级产品数量远大于其他等级产品数量，图中显示多数分类结果为 1 级，失衡样本下模型的分类结果无法学习到有风险的产品，更难以挖掘具有严重安全风险食品与影响因素之间的关联，显然缺乏实际研究意义。综上所述，使用上采样改善不平衡数据，1 级数量保持不变，将 1 级数量与 2、3、4、5 级产品数量分别按照 8:2、7:3、6:4 扩充数据量少数类的样本。表 5 为各比例上采样后的数据量。

表 5 上采样数据量说明

Table 5 Description of volumes of oversampling					
肉制品综合风险等级	1 级	2 级	3 级	4 级	5 级
原始数据量	29 770	1 001	1 686	4 713	1 145
8:2	29 770	7 443	7 443	7 443	7 443
7:3	29 770	12 758	12 758	12 758	12 758
6:4	29 770	19 846	19 846	19 846	19 846

1.2 模型的构建

集成学习（Ensemble Learning）通过构建并结合多个学习器来完成学习任务，也被称为多分类器系统。相比于单一学习器，将多个学习器结合能获得显著优越的泛化性^[34,35]。根据个体学习器生成方式集成学习分为 Boosting 和 Bagging 两类，前者个体学习器之间存在强依赖关系，必须串行依次生成新的学习器；后者个体学习器之间不存在强依赖关系，可同时并行生成多个学习器。图 4 为集成学习模型的构建示意图。

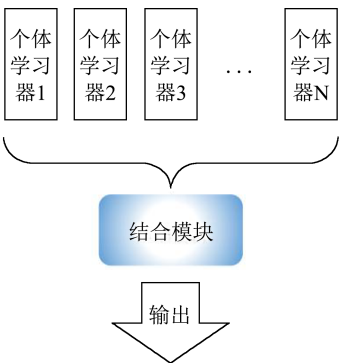


图 4 集成学习示意图

Fig.4 Sketch map of Ensemble Learning

其中个体学习器又称基学习器，相同个体学习器之间的集成称“同质”的集成，不同个体学习器之间的集成称“异质”的集成。本文以决策树为基学习器，分别构建 XGBOOST 和 RF 两个同质集成模型用于肉制品安全风险等级预警研究。

1.2.1 决策树（Decision Tree，DT）

决策树是常见的机器学习方法，其构建主要包括特征选择、决策树的生成、决策树的修剪^[34]。特征选择依据信息论的相关理论进行，这一过程对应特征空间的划分和决策树的生成，而决策树的修剪利用验证数据集对已生成的树进行剪枝并选择最优子树。其中特征选择是决策树算法的核心，选择划分的特征依据该属性含有信息量，1948 年香农提出的“信息熵”可以对信息量进行量化度量^[36]。式（3~7）为决策树中信息熵（Ent）、信息增益（Gain）、信息增益率（Gain_ratio）、基尼系数（Giin）的计算方法。

Ent(D)=−∑k=1|y|pklog2pk(3)

Gain(D,A)=Ent(D)−∑v=1V|Dv||D|Ent(Dv)(4)

$$Gain_ratio(D, A) = \frac{Gain(D, A)}{IV(A)} \quad (5)$$

$$IV(A) = - \sum_{v=1}^V \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|} \quad (6)$$

$$Gini(D) = 1 - \sum_{i=1}^n p_k^2 \quad (7)$$

式中:

D ——样本集,第 k 类样本所占比例为 $P_k, k=1, 2, 3, \dots, |v|$;

A ——离散属性;

V ——属性 A 包含的可能 V 个取值 $A_1, A_2, \dots, A_v$;

D^v —— D 中包含的所有属性 A 取值为 A_v 的样本, $IV(A)$ 称为属性 A 的固有值,属性 A 的可能取值数目越多(即 V 越大),则 $IV(A)$ 的值越大。

决策树经典算法有 ID3、C4.5 和 CART, 分别用信息增益、信息增益比和基尼系数来进行特征选择。决策树计算复杂度低, 输出结果易于理解, 对缺失值不敏感, 可以处理不相关数据, 但容易产生过拟合的问题, 对此引入的集成学习模型, 秉承了决策树算法优点的同时缓和了过拟合现象, 能有效提高模型的性能。

1.2.2 极端梯度提升树算法 (Extreme Gradient Boosting, XGBOOST)

XGBOOST 是 Boosting 算法的经典代表之一, 基模型按分步前向加进行, 训练基模型的训练集加入了正则项的损失函数每次都进行一定的转化, 在迭代完成之后才进行下一次迭代, 这不仅能够提高运行速度和准确率还能够有效的抑制过拟合^[22,34]。本研究中的基模型采用 CART 树, 通过串行的方式对多个树进行迭代, 最终输出结果为产品综合风险等级。图 5 为 XGBOOST 模型的构造过程。

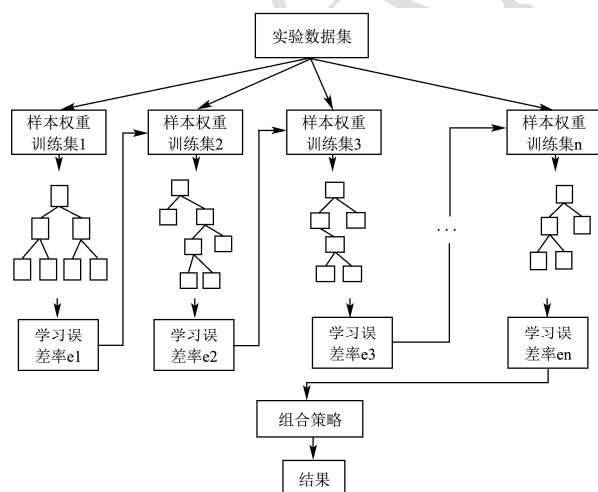


图5 XGBOOST 原理示意图

Fig.5 Schematic diagram of XGBOOST

1.2.3 随机森林 (Random Forest, RF)

RF 是 bagging 的扩展变体, 以 DT 为基学习器构

建 bagging 集成, 进一步在决策树训练过程中引入随机属性选择^[34]。RF 准确率高, 泛化能力好, 能够有效处理大数据集, 在众多现实任务中展现出强大的性能^[18]。本研究将处理后的实验数据集作为输入变量, 首先随机选择包含 k 个属性的子集, 其次以包含不同子集的 N 个子数据集分别构建决策树, 最后通过投票机制确定最终肉制品的风险等级类别。图 6 为 RF 模型的构建过程。

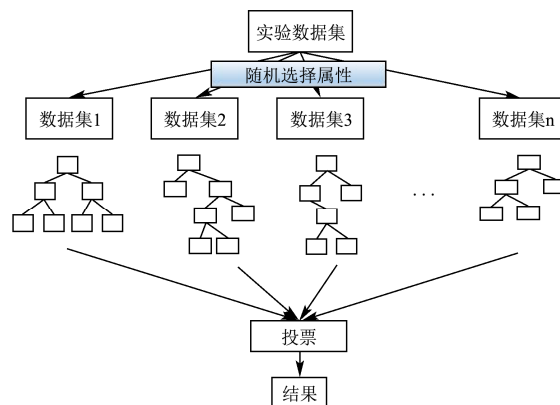


图6 随机森林原理示意图

Fig.6 Schematic diagram of random forest

1.2.4 最优模型技术路线图

本文以 DT 为基础, 将集成学习引入肉制品安全风险等级预警模型的应用中, 构建了 XGBOOST 和 RF 两个模型, 并以肉制品风险等级分类的结果以及评价指标来比较说明模型效果。经特征构造、数据重采样、超参数优化等优化处理后 XGBOOST 在肉制品安全风险等级预警中效果最佳(对比分析见 2.2.2 节模型比较分析)。图 7 为本文最优 XGBOOST 模型的构建框架。

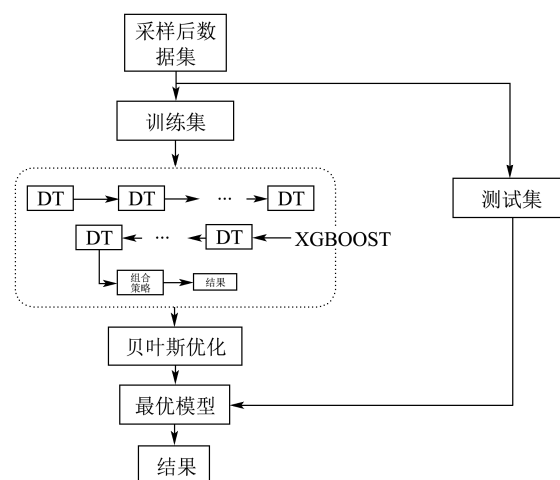


图7 肉制品安全风险等级预警的最优模型

Fig.7 Optimal model for risk level warning of meat products

1.3 模型的评价

1.3.1 特征重要性

基于树的集成模型,特征的重要性是在所有单颗树上该特征重要性的一个平均值,单颗树上特征重要性是该特征进行分裂后平方损失的减少量的求和,即把这个变量在所有树上出现的节点数累计求和,若该特征在所有树上出现的次数越多则越重要。

模型的输入变量为 1.1.6 节中处理后的变量,每个变量称为特征 x_j ,其中 j 取值为 areas、location、lasting、保质期、食品品种、是否大型生产企业、是否大型生产企业、是否进口、month/(month_1, ..., month_12)。 J^2 记作重要程度,特征 x_j 在整个模型中的重要程度:

$$\hat{J}_j^2 = \frac{1}{M} \sum_{m=1}^M \hat{J}_j^2(T_m) \quad (8)$$

特征 x_j 在单独一棵树上的特征重要程度:

$$\hat{J}_j^2(T_m) = \sum_{t=1}^{L-1} \hat{I}_t^2(v_t = j) \quad (9)$$

式中:

M ——模型中的树的数量;

T_m ——第 m 棵树,满足 $m \in \{1, 2, \dots, M\}$;

$L-1$ ——树中非叶子节点数量;

v_t ——在内部节点 t 进行分列式选择的特征;

$\hat{I}_t^2$ ——内部节点 t 分裂后平方损失的减少量。

本文将 one-hot encoding 处理后变量作为集成模型的输入变量,再通过计算特征重要性排序,能够更精确对比不同特征取值对于肉制品安全风险等级分类的影响程度。

1.3.2 性能度量指标

表 6 混淆矩阵

真实值	预测值	
	正类	负类
正类	TP	FN
负类	FP	TN

注: TP (True Positive) 真的正样本; FN (False Negative) 假的负样本; FP (False Positive) 假的正样本; TN (True Negative) 真的负样本。

分类模型的评价指标通常为分类准确率,以准确率作为评价指标便于理解,但对于多分类问题来说相对单一^[37,38]。为了更全面地选择最佳模型、评估模型性能,使用混淆矩阵和多个评价指标综合比较分析模型的效应。本文使用的评价指标有准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 值(Precision 和 Recall 的调和平均数)。表 6 和公式(10~13)为混淆矩阵以及各指标计算方法,本文对于模型的整体效果主要以 F1 综合判断,辅助参考 Accuracy、Precision

和 Recall 指标。

$$Accuracy = \frac{TP + TN}{FP + TN + TP + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \quad (13)$$

1.4 数据分析

本文的实验环境为 64 位 Windows10 计算机处理器和操作系统,相应配置的 CPU: Intel 第 11 代处理器,处理器型号为 i5-1135G7,主频 2.40 GHz,内存 16 GB。Anaconda 作为开源的 Python 发行版本,包含大量工具包且能够并存多个版本 Python。本研究代码基于 Anaconda 4.10.3 中的 Jupyter Notebook 平台通过 Python 3.7 实现,数据处理使用 numpy 和 pandas 包,集成学习模型的构建通过调用 sklearn 模块中的 Random Forest Classifier 和 xgboost 模块中的 XGBClassifier,展示分类结果的混淆矩阵图使用 matplotlib 包绘制,模型的超参数优化借助于 Bayesian Optimization 库。将 1.1.7 节重采样后的数据集的 areas 等特征作为输入变量,产品综合风险等级 level 作为输出变量,按照 8:2 将数据分为训练集和测试集,设定划分数据集的种子为 66 以保证不同模型下训练集和测试集的一致性。

2 结果与分析

2.1 分类结果分析

2.1.1 混淆矩阵分析

本文将 2013-2017 年的肉制品监测数据经处理采样后分别在决策树、XGBOOST 和 RF 模型中进行分类综合评价,得到各模型下测试集分类的混淆矩阵。图 8 为以 8:2 采样的编码和超参数调节前后生成的混淆矩阵,本文肉制品数据类别为 5 类,当对少数类以 8:2 进行采样时,安全肉制品(1 级)与有风险肉制品(2 级、3 级、4 级和 5 级)的数量比例恰好为 1:1,相对均衡(其他比例采样以及模型的混淆矩阵不展示)。在混淆矩阵中右侧的颜色深浅可直观看到各类的数量,颜色越深数量越多,左侧坐标 0~4 分别表示实际等级 1 级、2 级、3 级、4 级和 5 级,底部坐标 0~4 分别表示预测的等级 1 级、2 级、3 级、4 级和 5 级,对角线上的值表示实际类别和预测类别一致的数量。

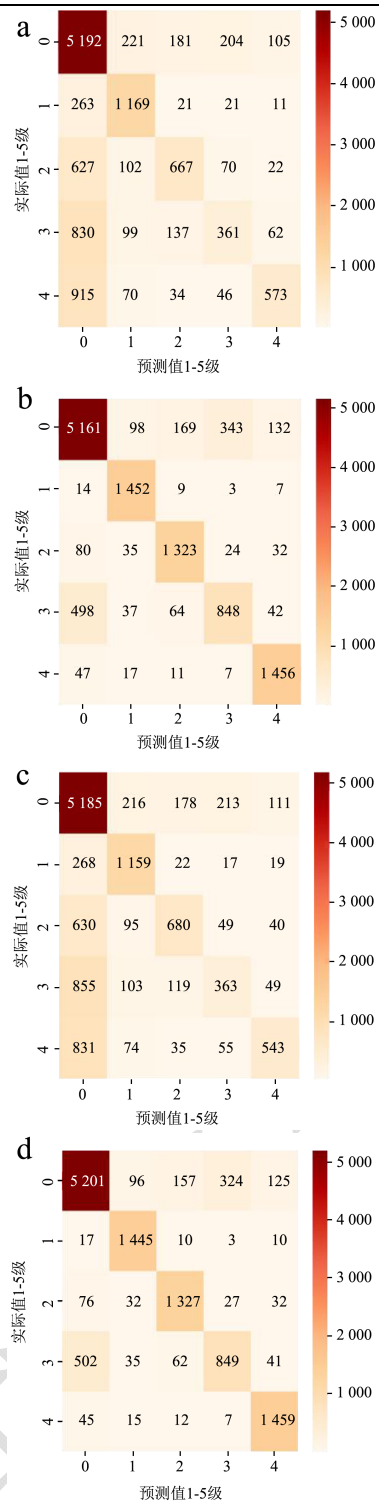


图8 测试集混淆矩阵

Fig.8 Confusion matrix of test set

注: (a) 为未进行优化处理的原始混淆矩阵图, (b) 在 (a) 基础上作了贝叶斯参数调节, (c) 为 (a) 的基础上经过 one-hot encoding 处理月份, (d) 同时用了 one-hot encoding 和超参数优化。

由图8混淆矩阵, 首先对比 (a) 和 (b), 参数优化后对角线上的颜色明显加深, 最左边一列的颜色变浅, 说明经过超参数优化后的预测效果明显提升; 其次对比 (a) 和 (c), 对角线上前2个颜色较深, 其余

均较淡, 左边第一例中后3个颜色较深, 说明1级2级分类效果好, 而3、4、5级分类较差, 但经过处理后 (c) 对角线上以外区域颜色比较淡, 说明错分类别有显著的改善, 从方格中标注数据量也能够证明效果的提升; 最后 (d) 展示出来颜色对比比较为鲜明, 说明了双重处理下比单一处理更有效。

2.1.2 评价指标分析

混淆矩阵简单直观展示了分类的大致趋势, 为了进一步说明肉制品模型分类的有效性, 表7给出 XGBOOST 模型 (决策树和 RF 模型的处理方式相同, 本文不展示决策树和 RF 各类别分类结果数据, 仅在 2.2.2 节中给出综合指标进行比较) 不同采样情况下测试集的超参数调节以及 one-hot encoding 处理前后各肉制品风险等级分类的具体指标。

由表7中结果可得: 在 XGBOOST 模型中, (1) 首先对比独热编码处理, 该编码后模型的肉制品风险等级分类的各指标有小幅度提高, 使用 one-hot encoding 提高 F1 值的同时能够更加具体分析月份取值的影响程度; (2) 其次分析超参数优化, 结果显示, 经贝叶斯优化后各指标上升效果显著, Precision 平均提高了 0.18、Recall 平均提高了 0.22、F1 值平均提高了 0.2; (3) 最后比较各个风险等级肉制品的总体分类效果在不同处理下的情况, 1 级 Precision 最高, 2 级和 5 级 Recall 最高, 2 级和 5 级 F1 值最高, 综合看 2、5 级分类效果最好, 其次是 1、3 级, 相比之下 4 级的分类效果稍差。综上, XGBOOST 模型在肉制品安全风险等级分类中的应用效果较好, 且随着 one-hot encoding 处理和超参数调节, 稳定性进一步得到了提高, 能够较好的实现肉制品风险等级的分类预警。

2.1.3 特征重要性分析

利用特征重要性分析对原始特征进行变量选择并确定时间维度的变量, 以及 one-hot encoding 处理后的各因素分析。由于编码后变量较多, 仅保留前十个重要特征分析, 图9为变量的特征重要性结果。

图9a为原始特征的部分特征重要性排序, 位居前三位的分别是 lasting、month、保质期。Lasting 对食品安全分级影响最大, 而是否进口、是否大型企业、是否大型经营企业分类的影响程度很小, 结合原始数据情况, 进出口数据较少使得对于风险等级的分类不够明显。是否是大型企业、是否是大型经营企业的分类影响最小, 由于原始数据中对于生产企业、经营企业的规模化分不够具体, 该分类辨识度较低, 没有中小型等企业, 类别划分过于粗略对风险等级分类有一定程度的影响。原始特征的重要性排序显示出变量月份对肉制品安全风险分级的重要性, 故而对月

份进行 one-hot encoding，再根据特征重要性比较月份取值的影响程度。此外还对其他变量进行了 one-hot encoding 处理，经过 XGBOOST 模型验证 F1 值没有提高，仅有月份处理后有了一定的提高，故本文仅对月份进行扩充，同时避免了维度灾难。

表 7 测试集各级产品评价指标

Table 7 Test set product evaluation indexes at all levels

采样比	处理	未调参				超参数调节			
		类别	Precision	Recall	F1	Precision	Recall	F1	support
未采样	month	1 级	0.80	0.98	0.88	0.82	0.93	0.87	5 982
		2 级	0.54	0.14	0.23	0.40	0.20	0.27	183
		3 级	0.43	0.06	0.11	0.29	0.14	0.19	332
		4 级	0.50	0.12	0.20	0.35	0.22	0.27	948
		5 级	0.50	0	0.01	0.23	0.09	0.13	218
8:2	One-hot	1 级	0.67	0.88	0.76	0.89	0.88	0.89	5 903
		2 级	0.70	0.78	0.74	0.89	0.97	0.93	1 485
		3 级	0.60	0.46	0.54	0.85	0.89	0.87	1 494
		4 级	0.52	0.24	0.33	0.70	0.57	0.63	1 489
		5 级	0.71	0.35	0.47	0.88	0.95	0.91	1 538
	month	1 级	0.67	0.88	0.76	0.89	0.87	0.88	5 903
		2 级	0.70	0.79	0.74	0.89	0.98	0.93	1 485
		3 级	0.64	0.45	0.53	0.84	0.89	0.86	1 494
		4 级	0.51	0.24	0.33	0.69	0.57	0.62	1 489
		5 级	0.74	0.37	0.49	0.87	0.95	0.91	1 538
7:3	One-hot	1 级	0.64	0.80	0.71	0.92	0.86	0.89	5 855
		2 级	0.75	0.84	0.79	0.92	0.98	0.95	2 576
		3 级	0.68	0.57	0.62	0.88	0.93	0.90	2 608
		4 级	0.59	0.43	0.49	0.82	0.79	0.81	2 573
		5 级	0.77	0.66	0.66	0.91	0.97	0.94	2 549
	month	1 级	0.64	0.79	0.71	0.92	0.85	0.89	5 855
		2 级	0.75	0.86	0.80	0.92	0.98	0.95	2 576
		3 级	0.67	0.56	0.61	0.87	0.93	0.90	2 608
		4 级	0.57	0.42	0.49	0.82	0.79	0.80	2 573
		5 级	0.76	0.55	0.64	0.91	0.98	0.94	2 549
6:4	One-hot	1 级	0.64	0.69	0.66	0.95	0.82	0.88	5 902
		2 级	0.76	0.90	0.83	0.91	1.00	0.95	3 978
		3 级	0.72	0.65	0.68	0.91	0.94	0.93	4 069
		4 级	0.61	0.50	0.55	0.85	0.87	0.86	3 978
		5 级	0.76	0.74	0.75	0.92	0.98	0.95	3 904
	month	1 级	0.65	0.68	0.66	0.96	0.82	0.89	5 902
		2 级	0.76	0.89	0.82	0.92	1.00	0.95	3 978
		3 级	0.71	0.64	0.67	0.91	0.94	0.92	4 069
		4 级	0.60	0.50	0.55	0.85	0.87	0.86	3 978
		5 级	0.74	0.76	0.75	0.92	0.98	0.95	3 904

注：未采样为表 5 中的原始数据量，One-hot 表示独热编码处理月份之后的数据集，month 为未使用独热编码的数据集，support 表示测试集中的各等级样本总数。

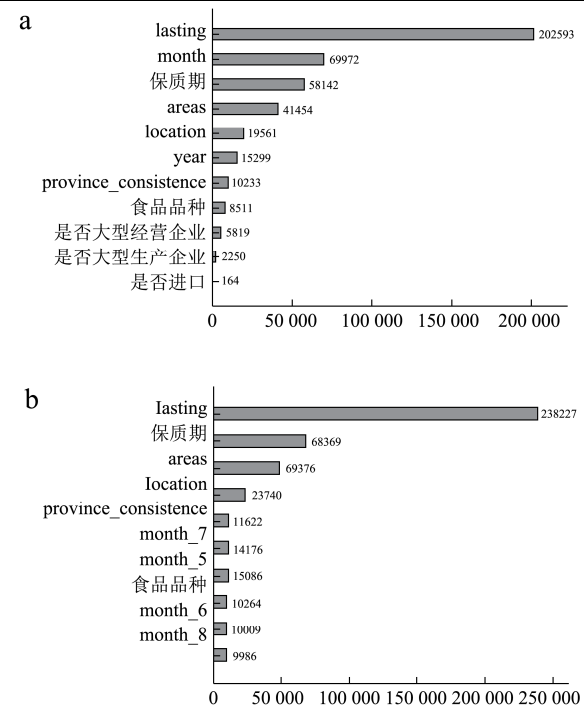


图9 变量选择和特征重要性分析

Fig.9 Variable selection and feature importance analysis

注：（a）为原始特征，（b）为 one-hot encoding。

图 9b 中处理后的特征重要性可分为三个层次。
（1）lasting；该指标直接反应抽样产品的新鲜程度也间接包含了存储的温度要求、包装等难以获取的复杂因素，该指标所处位置一定程度反映了食品安全中的不确定风险包含因素多。（2）保质期、areas；两者相差不大，保质期长短对于保鲜剂以及存储条件有不同限制，在不同环境下也有可能引起食物出现安全问题，该环节的影响也要重视。（3）食品品种新、location、province_consistence、month_7、month_6、month_8、month_5；食品种类和抽样地点表明不同产品生产条件等情况的差异，对食品安全风险的影响较大。province_consistence 对食品安全风险等级的分类影响

同样较高，该指标表示生产省份和抽样省份是否一致，包含了产品抽检前阶段经历的运输时长和路程的信息量，结果显示对食品安全的影响较为重要。使用 one-hot encoding 能具体了解各个月份对于食品安全风险等级分类的影响，位列前面的分别是 6 月、4 月、5 月和 7 月，月份作为季节性因素直接表现变量，包含了整年度外部温度环境，季节性需求、季节性生产量等难以收集的复杂因素，处理前后月份的重要性充分说明了其中包含的信息量多，对肉制品安全风险影响之大。

2.2 模型比较

2.2.1 模型参数优化

由于各算法参数配置存在差异，对不同模型下的重要参数进行调节，使用贝叶斯优化（Bayesian Optimization）对超参数调节，该方法比 sklearn 中的网格搜索（Grid Search CV）调节效率更高，运行速度更快且更加稳健^[39]。本文通过贝叶斯优化调节 DT、XGBOOST 和 RF 的相关超参数，在训练数据集中使用 5 折交叉验证调整优化。

表 8 中展示的为三个模型重点调节的超参数。max_depth 为树的最大深度，min_samples_leaf 表示叶子节点最少样本数，max_feature 决定单个决策树使用特征的最大数量，min_samples_split 内部节点划分需要的最小样本数，learning_rate 可提高模型的鲁棒性，n_estimators 为迭代次数，表示生成树的数量，subsample 控制每棵树随机采样的比例，可调节过拟合和欠拟合。consample_bytree 表示每棵树随机采样特征占比，min_child_weight 表示最小叶子节点权重和。每个模型下通过贝叶斯优化寻找最优的参数组合进而优化模型性能，结合分类各指标结果表明经过超参数调节之后的模型性能得到了进一步提升。

表 8 不同模型的参数配置

Table 8 Parameter configuration of different models

模型	调节参数
DT	max_depth、min_samples_leaf、min_samples_split、max_feature
XGBOOST	learning_rate、n_estimators、max_depth、subsample、colsample_bytree、min_child_weight
RF	n_estimators、max_depth、min_samples_leaf、min_samples_split、max_feature

2.2.2 模型比较分析

本文研究了 DT、XGBOOST 和 RF 目前广泛应用的三个模型，其中集成学习模型 XGBOOST 和 RF 分别以 DT 为基础的改进，通过肉制品抽检数据的分类来对食品安全风险等级进行预警，不同采样情况的训练集和测试集随机种子设定为 66 以保证数据集的一

致性和可比性，表 9 为不同采样情况下三个模型的最优结果对比。

由各模型下的训练集测试集准确率以及 F1 值得出，（1）在采样之后模型的分类效果 F1 值均能达到 80%，整体效果较好，模型的测试集准确率达到 94%，学习效果较好且各模型的训练效果没有显著区

别。但根据测试集的分类准确率反映出不同模型的泛化能力差别，测试集和训练集的准确率的平均误差中 DT、XGBOOST 和 RF 分为 0.09、0.05 和 0.06，以综合指标 F1 值来看 XGBOOST 得分最高，相比之下集成模型中 XGBOOST 更适合肉质品抽检数据的安全风险评价分析。(2) 根据不同采样情况分析得出，随着小样本比例的提高，模型的学习能力有了显著的提高。在样本严重不平衡情况下，三个模型的训练集准确率均高于采样之后能够达到 95%，测试集准确率在 70%~80%之间，若只凭借准确率来判定分类效果均在可接受范围内，但从 F1 得分(综合 Precision 和 Recall)

得分看效果显然较差，F1 平均值仅有 35.20%，说明在不平衡样本中仅凭准确率作为模型评价标准的可信度较低。采样后的 F1 值至少提高 45%，当采样比例由 8:2 增加到 6:4 模型的 F1 值平均升高了 7.73%，DT、XGBOOST 和 RF 分别提高了 9.13%、7.06%和 7.02%，虽然 DT 整体提升效果最高，但综合来看 XGBOOST 的提升 F1 值最高。综合分析得到集成学习在肉制品安全风险分类中的效果较好，其中 XGBOOST 模型在肉制品安全风险评价分析中的应用效果优于单棵 DT 和 RF，且随着不平衡类别的处理，模型的效果得到显著改善。

表 9 不同模型评价指标对比

Table 9 Comparison of evaluation indexes of different model

过采样比例	模型	训练集准确率/%	测试集准确率/%	F1 得分/%
未采样	DT	95.19	71.26	33.91
	XGBOOST	94.99	76.30	34.50
	RF	95.18	76.66	37.20
8:2	DT	94.17	81.98	80.59
	XGBOOST	94.16	86.33	84.44
	RF	94.16	84.78	83.78
7:3	DT	94.32	86.50	86.99
	XGBOOST	94.32	89.76	89.84
	RF	94.32	88.64	88.89
6:4	DT	94.60	89.46	89.72
	XGBOOST	94.58	91.34	91.50
	RF	94.60	90.69	90.80

注：未采样为表 5 中的原始数据量。

3 结论

针对目前食品安全检测数据囤积的情况，本文基于全国范围内肉制品抽检数据提出了一种基于集成学习的肉制品安全风险等级评价的方法，以抽样产品的综合风险等级为目标变量，并采用 3 种上采样改善不平衡数据，以 DT 为基础分别构建 2 个集成学习模型，联合使用 one-hot encoding 和贝叶斯优化进行指标细化和模型优化，有效的弱化了不平衡数据影响、细化了影响因素、提升了模型效果。实验结果表明 6:4 重采样结合 XGBOOST 的集成优化模型在肉制品检测数据中显示出最佳效果。研究表明集成模型比决策树单一模型更适用于肉制品食品安全监测，通过 one-hot encoding 进一步反应出变量取值对于食品风险的影响程度，为日常监测中的食品抽检重点方向提供参考，同时为今后集成学习在食品安全风险评预警的应用提供了思路。对于未来的研究工作，将从以下两个方面改进：一是对不平衡数据的处理深入讨论。本文随机

重采样一定程度缓解了不平衡中的过拟合和欠拟合，但有优化提升的空间，下一步通过研究不平衡数据采样方法的组合进一步改善过拟合和欠拟合。二是特征构造的变量一定程度弥补了数据的不足，但仍有多种复杂的因素需要挖掘。后续研究将针对此类检测数据继续挖掘影响肉制品安全风险的因素以提高分类的效率，深入分析各因素的影响程度。

4 建议

本文以肉制品检测数据的风险等级分类，通过提取的基本信息反应出各环节对于食品安全风险等级的影响。食品安全抽检部门以及相关食品安全管理局可根据各环节的影响程度调整抽检力度和重点关注环节，提高工作效率。食品安全涉及环节多，如运输时长、路程、温度、存储条件等信息难以有效记录，本研究的特征重要性分析结果显示这些信息对食品安全的影响较为重要。肉制品原始数据具有数量大、信息密度低的特点，建立更加完善的信息录入系统提高食

品数据的信息质量,不但有利于对食品安全的监管,还为今后在该领域开展深入的研究奠定了数据基础。

参考文献

- [1] 杨定登.以层次分析法构建餐饮业的食品安全监管量化分级评价模式[J].食品安全导刊,2020,36:99-100.
- [2] 赵可金,郎昆.中美竞争下的供应链安全研究[J].东北亚论坛,2022,31(2):19-39,127.
- [3] 彭兆伟,彭兆彤.供应链视角下食品安全风险防控研究[J].食品安全导刊,2021,35:159-161.
- [4] 苏亚松,张旻昱,王宗水,等.我国风险评估研究的内容、方法与趋势-基于文献分析的视角[J].科技管理研究,2019,39(9): 205-214.
- [5] 李笑曼,臧明伍,赵洪静,等.基于监督抽检数据的肉类食品安全风险分析及预测[J].肉类研究,2019,33(1):42-49.
- [6] Goldberg David M, Khan Samee, Zaman Nohel, et al. Text mining approaches for postmarket food safety surveillance using online media [J]. Risk Analysis, 2020, 42(8): 1749-1768.
- [7] 金剑,齐思源,彭亚拉.神经网络法在食品安全预警中的应用[J].食品科技,2011,36(11):306-309.
- [8] 徐钢,黎敏,吕志民,等.基于机器学习的产品质量在线智能监控方法[J].工程科学学报,2022,44(4):730-743.
- [9] Yu L, Wu Y, Tang L, et al. Investigation of diversity strategies in RVFL network ensemble learning for crude oil price forecasting [J]. Soft Computing, 2021, 25(5): 3609-3622.
- [10] Rona-TasAkos, Cornuejols Antoine, Blanchemanche Sandrine, et al. Enlisting supervised machine learning in mapping scientific uncertainty expressed in food risk analysis [J]. Sociological Methods and Research, 2019, 48(3): 608-641.
- [11] Niu B, Zhang H, Zhou G, et al. Safety risk assessment and early warning of chemical contamination in vegetable oil [J]. Food Control, 2021, 125(2): 107970.
- [12] Feng H, Wang W, Chen B, et al. Evaluation on frozen shellfish quality by blockchain based multi-sensors monitoring and SVM algorithm during cold storage [J]. IEEE Access, 2020, 8: 54361-54370.
- [13] Bouzembrak Y, Marvin H J P. Impact of drivers of change, including climatic factors, on the occurrence of chemical food safety hazards in fruits and vegetables: A bayesian network approach [J]. Food Control, 2019, 97: 67-76.
- [14] 江志英,李宇洋,李佳桐,等.基于层次分析的长短期记忆网络(AHP-LSTM)的食品安全网络舆情预警模型[J].北京化工大学学报(自然科学版),2021,48(6):98-107.
- [15] WANG Kunkun, LIU Xianda, ZHAO Jianming, et al. Application Research of Ensemble Learning Frameworks [C]// 2020 Chinese Automation Congress (CAC). Shanghai, Peoples R China: IEEE, 2020: 5767-5772.
- [16] Dasarathy B V, Sheela B V. A composite classifier system design: Concepts and methodology [J]. Proceedings of the IEEE, 1979, 67(5): 708-713.
- [17] 刘颂,赵亚迪,张振,等.基于集成学习的高炉压差预报模型研究[J].电子测量技术,2022,45(2):31-38.
- [18] 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.
- [19] Nguyen C N, Phan Q T, Tran N T, et al. Precise sweetness grading of mangoes (*Mangifera indica* L.) based on random forest technique with low-cost multispectral sensors [J]. IEEE, Access, 2020, 8: 212371-212382.
- [20] Wang L Y, Niu D T, Wang X Y, et al. A novel machine learning strategy for the prediction of antihypertensive peptides derived from food with high efficiency [J]. Foods, 2021, 10(3): 550.
- [21] Van Asselt E D, Noordam M Y, Pikkemaat M G, et al. Risk-based monitoring of chemical substances in food: Prioritization by decision trees [J]. Food Control, 2018, 93: 112-120.
- [22] 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.
- [23] Xu X R, Rothrock M J, Mohan A, et al. Using farm management practices to predict *Campylobacter* prevalence in pastured poultry farms [J]. Poultry Science, 2021, 100(6): 101122.
- [24] 市场监督管理局[EB/OL].<http://news.foodmate.net/shiyao>
- [25] Cheng Hongju, Feng Danyang, Shi Xiaobin, et al. Data quality analysis and cleaning strategy for wireless sensor networks [J]. Eurasip Journal on Wireless Communications and Networking, 2018, 1: 1-11.
- [26] 尹佳,陈翔,董曼,等.基于小波分解-长短期记忆网络预测模型的酱卤肉制品安全预测分析[J].食品科学,2022,43(3): 121-128.
- [27] 王芳,孙晓红,陶光灿.中国食品安全风险分级研究进展[J].食品科学,2021,42(21):271-277.
- [28] 陈锂,邹礼华,孟可欣,等.长短期记忆神经网络在肉制品中铅含量风险预警的应用[J].现代食品科技,2020,36(8):317-324,372.
- [29] YU Lean, ZHOU Rongtian, CHEN Rongda, et al. Missing

- data preprocessing in credit classification: one-hot encoding or imputation? [J]. *Emerging Markets Finance and Trade*, 2022, 58(2): 472-482.
- [30] 梁杰,陈嘉豪,张雪芹,等.基于独热编码和卷积神经网络的异常检测[J].*清华大学学报(自然科学版)*,2019,59(7):523-529.
- [31] Qiao Y, Yang X, Wu E. The research of BP neural network based on one-hot encoding and principle component analysis in determining the therapeutic effect of diabetes mellitus [C]// *IOP Conference Series-Earth and Environmental Science*. Guangzhou, Peoples R China: IOP Publishing, 2019, 267(4): 042178.
- [32] Nogales Alberto, Diaz-Moron Rodrigo, Garcia-Tejedor Alvaro J. A comparison of neural and non-neural machine learning models for food safety risk prediction with European Union RASFF data [J]. *Food Control*, 2022, 134: 108697.
- [33] 向鸿鑫,杨云.不平衡数据挖掘方法综述[J].*计算机工程与应用*,2019,55(4):1-16.
- [34] 周志华.机器学习[M].北京:清华大学出版社,2016.
- [35] Gutierrez-Espinoza L, Abri F, Namin A S, et al. Ensemble Learning for Detecting Fake Reviews [C]// 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC). Electr Network: IEEE, 2020: 1320-1325.
- [36] Shannon C E. A mathematical theory of communication, 1948 [J]. *Bell System Technical Journal*, 1948, 27(3): 3-55.
- [37] SONG Bo, SHANG Kefan, HE Junliang, et al. Impact assessment of food safety news using stacking ensemble learning [C]// *Advances in Transdisciplinary Engineering*. Electr Network: IOS Press, 2020: 353-362.
- [38] Wu Liya, Weng Sungshun. Ensemble learning models for food safety risk prediction [J]. *Sustainability*, 2021, 13(21): 12291.
- [39] Liu J F, Jiang C, Zheng J. Batch bayesian optimization via adaptive local search [J]. *Applied Intelligence*, 2021, 51(3): 1280-1295.