赵杰斌, 黄穗东, 孙远明, 等. 基于数据挖掘的江门市蔬菜食品安全风险分析与预测 [J]. 食品工业科技, 2023, 44(20): 281−288. doi: 10.13386/j.issn1002-0306.2022120006

ZHAO Jiebin, HUANG Suidong, SUN Yuanming, et al. Safety Risk Analysis and Prediction of Vegetable in Jiangmen City Based on Data Mining[J]. Science and Technology of Food Industry, 2023, 44(20): 281−288. (in Chinese with English abstract). doi: 10.13386/j.issn1002-0306.2022120006

· 食品安全 ·

基于数据挖掘的江门市蔬菜食品安全风险分析与预测

赵杰斌[1,2], 黄穗东[3], 孙远明[1], 徐振林[1,*]

（1.广东省食品质量安全重点实验室, 华南农业大学食品学院, 广东广州 510642;
2.广东省台山市市场监督管理局, 广东江门 529000;
3.广州市食品检验所, 广东广州 510410）

摘　要: 目的：对江门市 2016~2020 年蔬菜食品安全抽检数据进行分析，建立基于数据挖掘的食品风险预测模型。方法：以江门市辖区内农贸市场、批发市场、商场超市、餐饮服务单位等单位 10 个种类的蔬菜样本共 1928 份，分析其不合格样本和不合格项目的分布情况，并基于监测指标和样本信息，选取蔬菜种类、蔬菜品种、监测场所等 7 个属性为输入，结论属性为输出，利用反向传播（back-propagation，BP）神经网络构建蔬菜食品安全风险分析与预测模型。结果：风险分析显示，江门市芽菜类蔬菜、叶菜类蔬菜、根茎类和薯类蔬菜合格率分别为 81.7%、95.9%、96.3%，均低于总体合格率 96.6%；4-氯苯氧乙酸钠、毒死蜱和铅元素超标问题突出，不合格批次占比达 71.2%。经数据处理、最优参数筛选、数据训练和验证、模型优化等步骤构建出 3 层的 BP 神经网络模型，该模型总体精度为 96.3%，灵敏度为 96.8%，特异性为 83.9%。结论：该模型具有良好的预测准确度和性能，可为食品安全监管工作提供技术参考。建议可利用快检技术的大数据量优势与 BP 神经网络相结合，构建多算法组合模型，并加强样品信息登记的规范性，以构建出准确度更高，应用更广的风险分析与预测模型。

关键词: 反向传播（BP）神经网络, 数据挖掘, 蔬菜, 食品安全, 预测模型

中图分类号: TS201.6　　**文献标识码:** A　　**文章编号:** 1002−0306(2023)20−0281−08

DOI: 10.13386/j.issn1002-0306.2022120006

本文网刊:

Safety Risk Analysis and Prediction of Vegetable in Jiangmen City Based on Data Mining

ZHAO Jiebin[1,2], HUANG Suidong[3], SUN Yuanming[1], XU Zhenlin[1,*]

（1.Guangdong Provincial Key Laboratory of Food Quality and Safety, College of Food Science, South China Agricultural University, Guangzhou 510642, China;
2.Guangdong Taishan City Administration for Market Regulation, Jiangmen 529000, China;
3.Guangzhou Institute of Food Inspection, Guangzhou 510410, China）

Abstract: Objective: To establish a safety risk prediction model based on the analysis of food safety sampling data of vegetable in Jiangmen City from 2016 to 2020 and data mining. Methods: A total of 1945 samples of 10 kinds of vegetables from farmers' markets, wholesale markets, supermarkets and catering in Jiangmen City were collected and used to analyze the distribution of unqualified samples and unqualified items. Based on monitoring index and sample information, seven attributes including vegetable type, vegetable variety and monitoring place were selected as input and conclusion attribute was used as output. A risk analysis and prediction model of vegetable safety was established by back propagation (BP)

收稿日期: 2022−12−10

基金项目: 国家自然科学基金（31301467）; 广东省普通高校重点研究项目（2019KJDXM002）。

作者简介: 赵杰斌（1990−）, 男, 硕士, 助理工程师, 研究方向: 食品安全风险分析, E-mail: 123512204@qq.com。

*** 通信作者:** 徐振林（1982−）, 男, 博士, 教授, 研究方向: 食品安全检测与控制, E-mail: jallent@163.com。

neural network analysis. Results: The risk analysis showed that the qualified rates of sproutie vegetables, leafy vegetables, root vegetables and potato vegetables were 81.7%, 95.9% and 96.3%, respectively, lower than the averaged level. The excessive problems of sodium 4-chlorphenoxyacetate, chlorpyripyrix and lead were the main safety problems, 71.2% of the unqualified samples. A three-layer BP neural network model was constructed by data processing, optimal parameter screening and data training and validation, with an accuracy of 96.3%, a sensitivity of 96.8% and a specificity of 83.9%. Conclusion: The proposed model has good prediction performance, which can provide technical reference for food safety supervision. It is suggested that the multi-algorithm combination model can be built with the large data volume of rapid detection technology and BP neural network. Based on the the standardization of sample information registration, it is able to establish a risk analysis and prediction model with improved accuracy and broader application.

Key words: back-propagation (BP) neural network; data mining; vegetables; food safety; predictive model

蔬菜作为人民群众生活必需品，保障“菜篮子”产品的安全供应，事关民生福祉和社会稳定。近年来，随着生活质量的提升，人们更加注重营养均衡和膳食搭配，对于蔬菜的需求量在不断提升，根据相关研究，2020 年全国人均全年蔬菜消费量为 140 公斤，预计到达 2023 年消费量可达 166 公斤[1−2]。目前，威胁蔬菜质量安全的污染物主要以农药残留和重金属为主[3]，国内外相关国家和组织均对农产品质量安全建立风险预警体系，包括由欧洲食品安全局（European Food Safety Authority, EFSA）建立的食品与饲料快速预警系统（Rapid Alert System for Food and Feed, RASFF），美国疾病预防和控制中心（Centers for Disease Control and Prevention, CDC）建立的食源性疾病主动监测网络（Foodborne Disease Active Surveillance Network, FoodNet），国家食品安全风险评估中心（China National Center for Food Safety Risk Assessment, CFSA），均通过对农产品风险监测的相关数据进行分析，实现农产品安全隐患的提前预测与介入[4]。

目前，对于农产品安全风险预警分析主要是基于具体抽检数据，对不合格样本、食品种类、涉及场所等环节进行数据统计分析，并对存在的主要问题进行剖析的传统方法[5]。而利用数据挖掘技术，对现有数据间内在联系进行挖掘，可构建出对问题分析和预测的模型，人工神经网络（Artificial Neural Network, ANN）作为数据挖掘工具，在食品领域中国外已有相关的应用研究，如 Chen 等[6] 利用高光谱成像测定牡蛎中总挥发性碱性氮含量，以总挥发性碱性氮含量对牡蛎新鲜度进行评价，利用 BP 人工神经网络预测牡蛎在贮藏期间的新鲜度；Tarafdar 等[7] 以含水率、干燥效率和干燥速率为输出干燥参数，利用人工神经网络构建了蘑菇冷冻干燥应用模型预测生物材料的干燥过程，并与半经验模型进行比较发现，人工神经网络具有更优秀的预测效率；Mercie 等[8] 利用物理传热模型来计算货柜中的温度分布情况，并以此作为训练数据构建出可预测易腐食品在运输过程中温度变化的神经网络框架。国内在食品领域的应用研究中，范维等[9] 基于实时聚合酶链式反应法（Real-time PCR）对牛、羊肉串成分检测，并运用 BP 神经网络算法构建牛、羊肉串掺假的风险预测模型；陈锂等[10] 参照国家食品检验标准结合专家打分，对肉制品中铅含量分成 6 个风险等级，利用神经网络建立三层的时间序列风险预警模型；魏泉增等[11] 基于顶空固相微萃取结合气质联用（GC-MS）测定花生油的挥发性成分，采用人工神经网络对数据进行建模和预测，建立可用于鉴别不同工艺花生油的模型。相比较下，目前鲜见基于抽检数据利用人工神经网络构建农产品质量安全分析和预测模型的研究，有关江门市乃至广东省抽检数据的模型研究尚处于空白。

因此，本研究根据 2016~2020 年江门市全域范围内蔬菜抽检数据，运用 SPSS 软件对其质量安全状况进行分析，并运用 IBM SPSS Modeler 软件基于抽检数据的各项指标，利用 BP（Back-Propagation, BP）神经网络进行数据挖掘，构建出江门市蔬菜食品安全风险分析与预测模型，为监管部门进一步做好农产品监管提供技术参考。

1 材料与方法

1.1 样本来源

蔬菜样本　2016~2020 年，本研究从江门市四市三区的农贸市场、批发市场、商场超市、餐饮服务单位等单位对 10 个种类的蔬菜进行采样，每年各类蔬菜抽样情况如表 1 所示，共抽取蔬菜样本 1945 份，其中豆类蔬菜 148 份、叶菜类蔬菜 469 份、根茎类和薯类蔬菜 297 份、鳞茎类蔬菜 116 份、瓜类蔬

表 1　2016~2020 年各类蔬菜采样情况

Table 1　Samples of various vegetables from 2016 to 2020

样品种类	抽样数量（份）					样品总数（份）
	2016	2017	2018	2019	2020	
豆类蔬菜	16	16	41	17	58	148
叶菜类蔬菜	24	43	84	58	260	469
根茎类和薯芋类蔬菜	16	15	73	34	159	297
鳞茎类蔬菜	8	8	29	18	53	116
瓜类蔬菜	8	16	49	50	71	194
芸薹类蔬菜	16	32	35	26	87	196
茄果类蔬菜	24	25	88	51	89	277
芽菜类蔬菜	24	9	25	25	59	142
水生类蔬菜	3	3	2	4	7	19
食用菌	21	16	4	13	33	87
合计	160	183	430	296	876	1945

菜 194 份、芸薹类蔬菜 196 份、茄果类蔬菜 277 份、芽菜类蔬菜 142 份、水生蔬菜 19 份、食用菌 87 份。

1.2　检测和判定方法

依据 GB23200.113-2018《食品安全国家标准 植物源性食品中 208 种农药及其代谢物残留量的测定 气相色谱-质谱联用法》、GB 23200.121-2021《食品安全国家标准 植物源性食品中 331 种农药及其代谢物残留量的测定 液相色谱—质谱联用法》、BJS 201703《豆芽中植物生长调节剂的测定》、GB 5009.12-2017《食品安全国家标准 食品中铅的测定》、GB 5009.15-2014《食品安全国家标准 食品中镉的测定》等标准，对蔬菜中的氧乐果、毒死蜱、甲基异柳磷、克百威、水胺硫磷、氟虫腈、腐霉利、阿维菌素、噻虫嗪、灭蝇胺、4-氯苯氧乙酸钠 11 种农药残留，镉和铅 2 种重金属元素进行测定。根据 GB 2762-2017《食品安全国家标准 食品中污染物限量》、GB 2763-2021《食品安全国家标准 食品中农药残留最大残留量》和《国家食品药品监督管理总局、农业部、国家卫生和计划生育委员会关于豆芽生产过程中禁止使用 6-苄基腺嘌呤等物质的公告》(2015 年第 11 号)对检测结果进行判定。

1.2.1　蔬菜食品安全预测模型构建

1.2.1.1　BP 神经网络模型的构建　人工神经网络(ANN)是一种模仿人脑功能，基于生物神经系统结构的数据处理系统，通过人工神经元之间相互连接构建出一个非线性的自适应系统，可应用于数据分析为核心的数据挖掘领域，实现对数据的预测、分类等功能，为决策制定和问题分析提供技术参考[6,12]。BP 神经网络，又称反向传播(back-propagation,BP)神经网络，作为一种前馈、多层式网络，在反复向输入的样本学习的训练过程中，通过不断调整网络的权值来获得最小的误差，以实现网络输出无限逼近期望值的目标[13−14]。本研究将利用 2016 年至 2020 年江门市蔬菜食品安全监督抽检数据为样本，将多维度的不同抽检数据输入 BP 神经网络，通过反复的训练从而获得预测目标变量结果。

1.2.1.2　数据样本的预处理　数据自身的数量，格式和结构等特点是构建 BP 神经模型的关键基础，除了要保证有足够数据量作为支撑外，还需要选择数据的特征属性，以保证模型的准确度和实用性。原始的抽检数据中包括了受检地址、联系人、联系电话、型号规格、文字商标、检验机构等多个属性，这些仅能代表单一样本的属性，无法适用于以多个样本为基础的模型构建中[15]。由于本研究通过基于采样时间、采样地点和样本自身属性等多个维度，构建蔬菜食品安全预测模型，因此选取了年份、月份、行政区域、所属镇街、监测场所、蔬菜种类、蔬菜品种 7 个具有代表性属性作为输入变量，将结论作为输出变量(目标变量)。

此外，蔬菜样本在采集工作中，采样信息时受个人理解、地方方言等影响，造成同品种蔬菜会登记成不同名字，如结球甘蓝会登记为包菜、卷心菜、包心菜等，为保证数据的统一性，本研究将参考 GB 2763-2021《食品安全国家标准 食品中农药最大残留限量》中附录 A 来规范蔬菜品种名称。为了方便数据的导入，对输入数据按照属性进行整理，数据框类型见表 2。

表 2　BP 神经网络模型数据框类型

Table 2　Data frame type for BP neural network model

年份	月份	行政区域	所属街道	监测场所	蔬菜种类	蔬菜品种
2020	10	蓬江区	白沙街道	超市	叶菜类菜蔬	芹菜
2020	10	蓬江区	环市街道	农贸市场	叶菜类蔬菜	菠菜
2019	7	鹤山市	沙坪街道	中型餐馆	芸薹类蔬菜	菜薹
2019	8	台山市	台城街道	小型餐馆	茄果类蔬菜	辣椒
2018	9	开平市	三埠街道	批发市场	豆类蔬菜	豇豆
2018	3	开平市	长沙街道	食堂	芽菜类蔬菜	绿豆芽
2017	5	恩平市	恩城街道	食杂店	根茎类和薯芋类蔬菜	萝卜

1.2.1.3　建模流程　模型构建的流程见图 1，通过源节点对整理好的 excel 电子表格数据进行导入；通过字段选项节点的类型选项对数据值进行读取并进行角色调整；利用分区选项将输入的数据集分为训练集和验证集；由于食品安全抽检的结论为合格和不合格的样本量之间是不平衡的，为提高模型对于合格和不合格样本的预测准确性，需要通过记录节点的平衡选项对数据集进行平衡；最后通过神经网络节点对建模后，再利用分析节点和图形评估节点来对模型的准确度进行分析。

1.2.1.4　建模参数设置　a. 类型节点的设置：将数据导入后，类型节点将读取数据集的值，将各项属性设

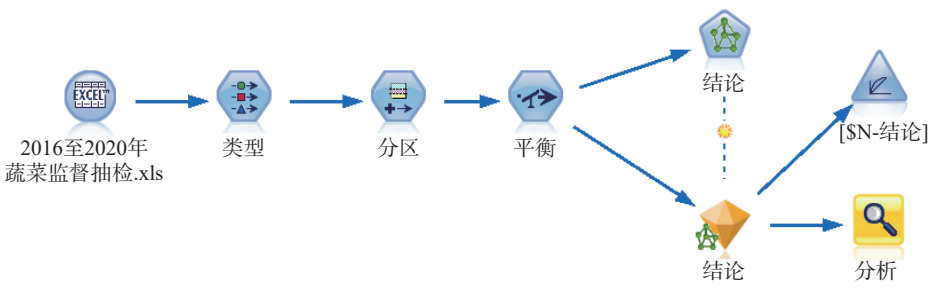

图 1　模型构建流程图

Fig.1　Flow chart of model establishment

为名义属性，除将检测结果设为输出角色，其他变量均设为输入角色。由于神经网络只能处理数值型输入变量，自动数据处理会将分类变量转化为取值为 0 或 1 的数值变量，对各变量进行二进制编码，符合神经网络输入的要求[16−17]。

b. 分区节点的设置：将数据集分成训练集和验证集，以提高模型的稳定性和延续性。本研究的分区节点的设置为：训练分区大小 70%，测试分区大小 30%。

c. 平衡节点的设置：本研究中共采集的蔬菜样本 1945 份，其中不合格样本 62 份，占比 3.2%，合格样本和不合格样本的分布比例过于悬殊，如果直接用于模型构建，会因小概率样本的预测结果较差，影响模型的准确率。参考 Linoff 的研究[18]，采用过抽样（Oversampling）或欠抽样（Undersampling）技术来调整两者样本的分布比例，Linoff 认为小概率样本的比率维持在 10%~50% 之间，会获得较好的效果。对于平衡节点的设定，将以模型的总体准确率为判定指标，选出最佳的小概率样本事件的比率。

d. 神经网络建模节点的设置：神经网络模型选择多层感知器（MLP），一种通过“误差反向传播算法”多层前向网络；模型停止规则为“无法进一步降低误差”，在模型训练环节中，模型在向样本学习的同时，通过权值进行不断修正，使得获得预测误差最低的期待模型；在模型中间隐藏层的神经元数量设定上，根据田兴国等人的研究[19]，每一个 BP 神经网络都有一个最优的中间神经元数量，通过经验计算公式确定隐藏层中神经元的数量合理范围。

$$n1 = \sqrt{n+m}+a \qquad 式（1）$$

式中：n1为隐藏层节点数，n为输入节点数，m为输出节点数，a介于 1~10 的常数。

1.3 数据处理

采用 Excel 2007 软件进行数据统计和建模数据样本预处理；SPSS 19.0 软件进行相关性分析和显著性差异分析；IBM SPSS Modeler 14.1 软件构建 BP 神经网络模型；采用 Origin 2019 软件进行绘图。

2 结果与分析

2.1 蔬菜食品安全风险分析

2.1.1 各年间总体合格率的分析 本研究抽检的 1945 份蔬菜样本，共有 65 份样本不合格，总体合格率为 96.6%，各年检测情况见图 2。采用 Excel 2007 和 SPSS 19.0 软件，对各年间整体合格率进行显著性差异分析，通过卡方检验显示 $\chi 2$=10.8，P<0.05，表明各年检整体合格率之间的差异显著，具有统计学意义。除 2017 年以外，均保持在 95.5% 以上，且维持在稳定水平。

2.1.2 不合格样本种类的分析 由图 3 可知，10 个种类蔬菜中，除瓜类蔬菜、水生蔬菜、食用菌以外，其余 7 个种类的蔬菜均检出不合格样本，其中合格率最低三个蔬菜种类分别是芽菜类蔬菜 81.7%、豆类蔬菜 95.9%、根茎类和薯类蔬菜 96.3%，均低于总体水平。

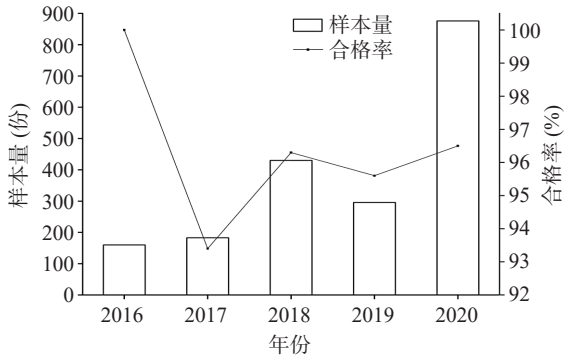

图 2 2016~2020 年各年间总体合格率情况

Fig.2 Overall qualified rate of each year from 2016 to 2020

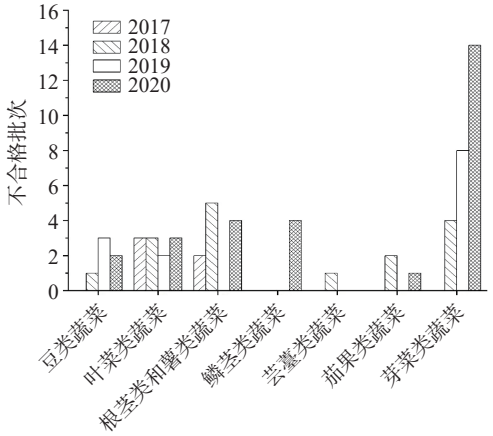

图 3 2017~2020 年抽检不合格蔬菜种类情况

Fig.3 Distribution of unqualified samples on different vegetables from 2016 to 2020

在不合格样本中芽菜类蔬菜最多，占不合格样本总数的 40%，其次是叶菜类蔬菜，根茎类和薯类蔬菜，分别占比 16.9% 和 21.5%。利用 SPSS 19.0 软件对上述 3 个种类蔬菜的历年采样总数和不合格总数进行皮尔逊（Person）相关性分析，显示 Person 相关系数 ρ=0.95，P<0.05，表明采样总数和不合格总数之间存在极强的相关，通过增加高风险品种蔬菜的采样量能有效地发现不合格样本。

2.1.3 不合格项目的分析 根据农业农村部公布的《禁限用农药名录》（2019 版）要求，毒死蜱、氟虫腈、甲基异柳磷、克百威 4 种农药被禁止用于蔬菜的种植种；根据原国家食药总局在 2015 年发布的《关于豆芽生产中禁止使用 6-苄基腺嘌呤等物质的公告》，将 4-氯苯氧乙酸钠列作农药登记管理，并禁止用于芽菜类蔬菜的种植种。由图 4 可知，2016 年至 2020 年间，江门市蔬菜食品安全问题可分为农药残留超标、重金属超标和植物生长激素超标三类问题，其中 4-氯苯氧乙酸钠、铅元素、毒死蜱不合格批次最多，分别占不合格项次比例 39.4%、21.2%、10.6%。可见，江门市蔬菜中禁限用农药残留超标、铅元素蓄积和违规使用植物生长激素的问题较为突出，结论与李培

武等[3] 和周辉等[20] 的研究相一致。

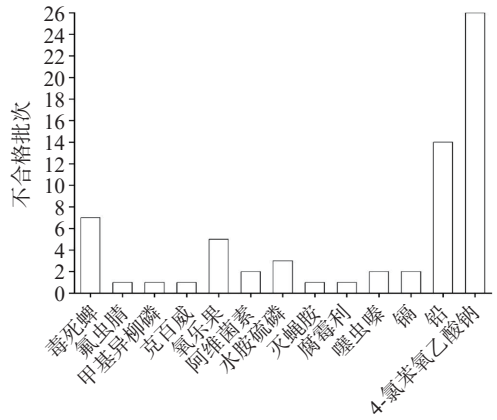

图 4　2016~2020 年抽检不合格项目情况

Fig.4　Distribution of unqualified sampling items from 2016 to 2020

造成上述问题的原因：a. 在豆芽生长过程中加入 4-氯苯氧乙酸钠能促进生产，提高产率，使其芽轴变长，根部变短[21]，加之 4-氯苯氧乙酸钠的无需指定销售或购买登记，造成部分种植户为提高豆芽产量违规使用；b. 除《禁限用农药名录》(2019 版)中 41 种禁止生产销售和使用的农药以外，其他农药均可登记销售，但对于购买后的施药范围和施药量缺乏监管，加之农户安全用药知识和意识缺乏，导致超范围和超限量使用农药问题的出现；c. 重金属元素在植物代谢旺盛的器官中蓄积量最大，同时根部作为最先接触土壤重金属的器官，造成植物根部相对于其他部位而言蓄积量更多，根茎类和薯类蔬菜可食用部分主要是其根部，导致此类蔬菜在抽检时较其他种类的蔬菜更容易出现铅元素超标的问题[22-23]。根据胡霓红等[24]，陈志良等[25] 对江门周边城市蔬菜重金属蓄积情况的研究结果显示，叶菜类蔬菜的重金属蓄积能力较其他种类蔬菜更强，但江门市实际情况与其研究结果不一致。

2.2　蔬菜食品安全数据挖掘与预测模型

2.2.1　不同平衡节点下模型准确度对比　参考 Linoff 等[18] 的研究，将平衡节点中的合格：不合格分别设置为 3.4∶1、7.6∶1、13.0∶1、20.2∶1、30.4∶1，其他节点按软件默认设置，所形成的模型概况如表 3 所示。

在模型的整体准确度方面，通过过抽样的方式增加不合格样本的比率，对于模型的总体合格率有明显的提升，但当合格：不合格达到 13∶1 时，再次提升不合格样本的比率对于整体合格率提升并不明显。在合格样本和不合格样本的准确度上，随着不合格样本的比率增加，对不合格样本的预测准确度也随着增加，并在合格∶不合格=13∶1 时达到 100%，与此同时对合格样本的预测准确度也在下降。因此，对于本研究模型的平衡节点设置采用合格∶不合格=13∶1。

2.2.2　不同隐藏层神经元模型下准确度对比　模型共有 7 个输入节点和 1 个输出节点，根据公式(1)计算，隐藏层节点数取值在 4 至 13，按照训练集∶验证集=7∶3，平衡节点合格∶不合格=13∶1 设置，根据不同隐藏层节点数所构成的 10 个模型准确度在 92.7% 至 96.1% 区间，合格样本的预测准确度在 89.4% 至 94.5% 区间，不合格样本的预测准确度在 91.3% 至 100% 区间，如图 5 所示。通过比较发现，当隐藏层节点设置为 5 个时，在模型总体准确度、合格样本和不合格样本预测准确度上均为最优，因此将其作为模型最优设置参数。

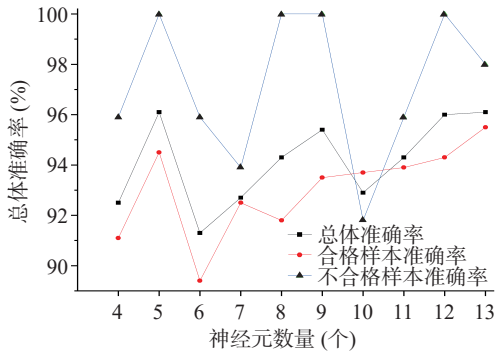

图 5　不同隐藏层神经元模型下准确度对比

Fig.5　Accuracy comparison of different hidden layer neuron models

2.2.3　模型优化和评价　为进一步提升模型的准确性，采用推进方法(Boosting)对模型进行优化。Boosting 是由 Schapire 在 1990 年提出的多项式级的算法[26]，其原理是通过产生一系列“成分模型”，每个会在整理数据上进行构建，在构建每个后续成分模型之前，会根据前一个模型的残差对记录进行加权。具有较大残差的会被给予较高的分析权重，下一个成分模型将更侧重于这些记录。这些成分模型共同构建一个整体模型，同时采用组合规则对新纪录进行评

表 3　不同平衡节点设置下模型准确度对比

Table 3　Accuracy comparison of different balance node settings

目标	结论									
模型	多层感知器									
所使用的停止规则	无法进一步降低误差									
平衡节点设置合格：不合格	3.4∶1		7.6∶1		13.0∶1		20.2∶1		30.4∶1	
总体准确度	91.3%		93.3%		96.4%		96.8%		96.9%	
观察/预测	不合格	合格	不合格	合格	不合格	合格	不合格	合格	不合格	合格
不合格	40.4%	59.6%	85.7%	14.3%	100%	0%	100%	0%	100%	0%
合格	1.9%	98.1%	4.4%	95.6%	5.0%	95.0%	5.3%	94.7%	5.9%	94.1

分，可用的规则将取决于目标的测量级别[27−28]。使用 Boosting 构建模型相对于标准模型而言，需要花费更长的构建和评分时间，但是模型的结果预测精确度会更高。结合“2.2.1”和“2.2.2”的最优设置参数，使用 Boosting 方法构建模型后，添加分析节点和评估节点，并与标准模型连接，以对比两个模型的准确性和性能。另外，还将通过交叉验证方法来对比两个模型的稳定性。

在模型的准确性上，本研究将采用灵敏度（sensitivity，sen）、特异度（specificity，spe）、精度（accurary，acc）三个参数，分析模型的准确性。

灵敏度，代表实际为正例被判断为正例的概率，当灵敏度越高时，反映模型对合格样本的预测不容易出现误判。

$$\text{sen}(\%) = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100 \qquad \text{式（2）}$$

式中：TP 代表模型正确预测为合格的样本数（true positive，TP），FN 代表错误预测为合格的样本数（false negative，FN）。

特异度，代表实际为负例被判断为负例的概率，当特异度越高时，反映模型对不合格样本的预测不容易出现漏判。

$$\text{spe}(\%) = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100 \qquad \text{式（3）}$$

式中：TN 代表模型正确预测为不合格的样本数（true negative，TN），FP 代表错误预测为不合格的样本数（false positive，FP）。

精度，代表预测正确的样本占总样本的比例，反映出模型总体分类的能力。

$$\text{acc}(\%) = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FN}+\text{FP}} \times 100 \qquad \text{式（4）}$$

从表 4 可知，使用 Boosting 构建的模型相比标准模型在精度上提升了 2.32%；在不合格样本预测的特异度方面，Boosting 模型被错判的样本比标准模型增加 4 个，不合格样本总量较少，即便被错判的样本数量相差较小，造成两者特异性差异较大，但 Boosting 模型特异度为 83.87% 仍可接受；在合格样本预测的灵敏度方面，Boosting 模型被错判的样本比标准模型减少了 74 个，灵敏度提升了 4.37%。综合分析，运用 Boosting 构建的模型比标准模型精确度有明显的提升。

表 4 标准模型与优化模型的准确性比较

Table 4 Accuracy comparison between normal model and optimized model

观察	Boosting模型		标准模型	
	预测		预测	
	不合格	合格	不合格	合格
不合格	52	10	56	6
合格	61	1816	135	1734
特异度/灵敏度	83.87%	96.75%	90.32%	92.38%
精度	96.34%		94.02%	

在模型性能评估上，分别选择增益和提升两种类型的图，结合基线与最佳线综合分析，在累积收益图中，一个良好的模型，收益线会向 100% 徒增，然后趋于平稳状态；在累积提升图中，累积线始于大于 1.0 的值，并向 1.0 靠近，良好模型的响应图，图表左侧会保持较高水平，在图表右侧曲线将迅速下降。从图 6 可知，增益和提升图表均显示 Boosting 模型和标准模型与最优模型想接近，对于不合格样本预测的总体性能较好，但两个模型之间对比，Boosting 模型明显优于标准模型，在提升图表中，前 1 个百分位不合格预测的性能前者比后者提升 15.4%。

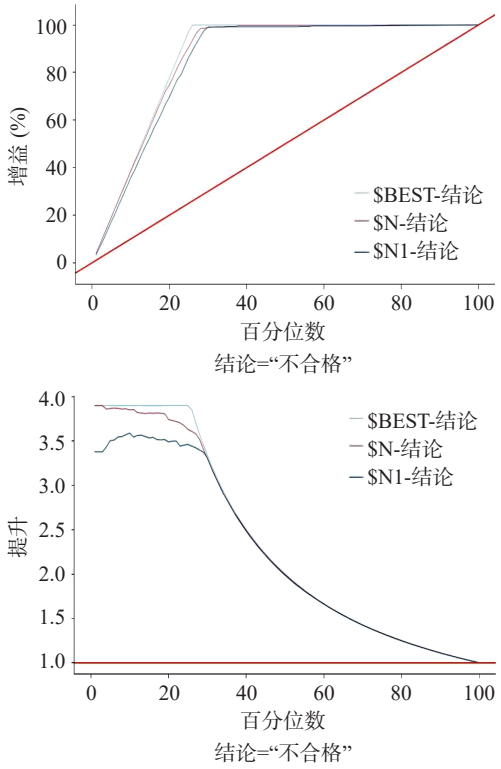

图 6 标准模型与优化模型性能评估图

Fig.6 Performance evaluation diagram of normal model and optimized model

在模型的稳定性评估上，采用十折交叉验证（ten-fold cross validation）对标准模型和 boosting 模型进行分析，利用 excel 将 1945 个样本数据随机分为 10 份，其中任意 9 份作为训练数据集训练模型，剩余 1 份作为测试数据集测试模型，重复 10 次试验。

如图 7 所示，在十折交叉验证中，boosting 模型训练集正确率保持在 97.8%~98.6%，方差为 0.1，测试集正确率保持在 91.5%~95.5% 之间，方差为 1.8；标准模型模型训练集正确率保持在 91.4%~97.1%，方差为 3.0，测试集正确率保持在 87.5%~95.5% 之间，预测正确率之间方差为 6.5。由此可见，运用 boosting 构建模型的稳定性优于标准模型。因此，本研究所构建模型的最优参数设置如表 5 所示。

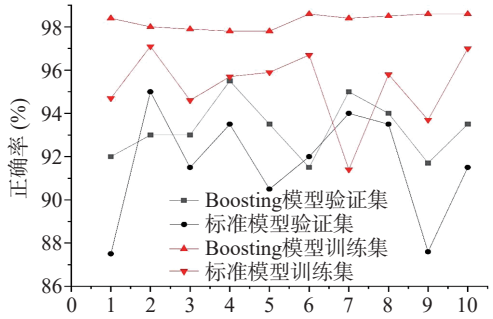

图 7　标准模型与优化模型十折交叉验证情况

Fig.7　Ten-fold cross validation of normal model and optimized model

表 5　模型最优参数设置

Table 5　Optimal parameter settings for the model

参数名称	参数值
模型	BP神经网络多层感知器(MLP)
使用最大训练时间（每个组建模型）	15 min
平衡节点设置（合格:不合格）	13.0:1
结构	7-5-1
训练方法	boosting
用于boosting的组件模型数量	10
过度拟合防止集合(%)	30

3　讨论与结论

3.1　讨论

本研究所构建的模型，是通过输入样本的属性来预测样本的结果，在实际食品安全抽检工作前作为参考，将样本的属性输入到已训练好的 BP 神经网络模型后，可根据模型预测的结果指定抽检方案并进行针对性检测，提升食品安全抽检的靶向命中率，同时对于降低部分合格率高品种的抽样量，节省人力、物力的投入，提升抽检工作的效率，具有较高的实用价值。另外，除了在法定检测中运用外，也可用结合日常快速检测工作使用。相对比法定检测而言，快速检测利用快速、简便、廉价等优势实现大面积筛查，可为模型构建提供大量的数据量，进而对优化模型。

但是模型构建过程中，样本属性信息登记不规范、不合格样本占比极低、样本输入属性选择等因素也会对模型的准确性和实用性造成较大的影响，因此在建模过程中对数据准确尤为重要，建议监管部门丰富抽检信息公示的内容，并按照规范性文件规范产品信息填写。此外，使用 BP 神经网络构建的模型虽然具备较高的分类精度，能自适应和自主学习，但其作为“黑箱模型”(Black box)弊端也得模型的解释性和稳定性较差，根据 Liu 等[29]、向晖[30]、王强等[31] 研究，将不同算法构建的模型进行组合，利用不同种类模型的优势互补，避免单体模型的弊端，同时又共同解决同一个问题。因此，在模型进一步优化中，可以此为方向，基于多种算法组合的农产品质量安全分析和预测模型。

3.2　结论

本研究对江门市 1945 份蔬菜样本的农药残留和重金属含量情况分析发现，芽菜类蔬菜、叶菜类蔬菜，根茎类和薯类蔬菜三类蔬菜总体合格率低于总体水平，其中以 4-氯苯氧乙酸钠、铅元素、毒死蜱问题较为突出。通过对上述检测数据进行深层挖掘，利用 IBM SPSS Modeler 14.1 软件构建蔬菜食品安全的 BP 神经网络预测模型。经过对平衡节点、隐藏层神经元、训练方法的优化，构建出 3 层神经模型的精度为 96.3%，对合格样本的灵敏度为 96.8%，对不合格样本的特异性为 83.9%，结合增益和提升两种评估图分析，该模型的预测效果良好，可为蔬菜食品安全抽检工作提供参考。建议在利用法定抽检数据的进行模型构建的基础上，结合快速检测的优势获得更大的数据量，同时通过不同算法构建的模型组合，利用各自的优势来构建准确度更高，应用面更广的预测模型。

参考文献

[1] 赵丽云, 刘素, 于冬梅, 等. 我国居民膳食营养状况与《中国食物与营养发展纲要 (2014 年-2020 年)》相关目标的比较分析[J]. 中国食物与营养, 2015, 21(8): 5−7. [ZHAO L Y, LIU S, YU D M, et al. Comparative analysis of dietary nutrition status of chinese residents and related targets of chinese food and nutrition development program (2014-2020)[J]. Food and Nutrition in China, 2015, 21(8): 5−7.]

[2] 沈辰, 梁丹辉, 王盛威, 等. 2014 年-2023 年中国蔬菜市场展望[J]. 农业展望, 2014, 12(10): 14−18. [SHEN C, LIANG D H, WANG S W, et al. China's vegetable market outlook for 2014-2023 [J]. Agricultural Outlook, 2014, 12(10): 14−18.]

[3] 李培武, 张奇, 丁小霞, 等. 食用植物性农产品质量安全研究进展[J]. 中国农业科学, 2014, 47(18): 3618−3632. [LI P W, ZHANG Q, DING X X, et al. A review of studies on quality and safety of edible vegetable agro-products[J]. Scientia Agricultura Sinica, 2014, 47(18): 3618−3632.]

[4] 周雪巍, 郑楠, 韩荣伟, 等. 国内外农产品质量安全风险预警研究进展[J]. 中国农业科技导报, 2014, 16(3): 1−7. [ZHOU X W, ZHENG N, HAN R W, et al. Research progress on risk early-warning for quality and safety of international and domestic agricultural products[J]. Journal of Agricultural Science and Technology, 2014, 16(3): 1−7.]

[5] 韩世鹤, 李红, 江逸楠, 等. 基于食品抽检数据的风险预警智能研究模型构建分析[J]. 食品安全质量检测学报, 2022, 13(10): 3172−3179. [HAN S P, LI H, JIANG Y N, et al. Constructing analysis of risk early warning intelligent research models base on food sampling data[J]. Journal of Food Safety and Quality, 2022, 13 (10): 3172−3179.]

[6] CHEN L P, LI Z J, YU F, et al. Hyperspectral imaging and chemometrics for nondestructive quantification of total volatile basic nitrogen in Pacific oysters *(Crassostrea gigas)*[J]. Food Analytical Methods, 2019, 12(3): 799−810.

[7] TARAFDAR A, SHAHI N C, SINGH A. Freeze-drying behaviour prediction of button mushrooms using artificial neural network and comparison with semi-empirical models[J]. Neural Computing and Applications, 2019, 31(11): 7257−7268.

[8] MERCIE S, UYSAL I. Neural network models for predicting

perishable food temperatures along the supply chain[J]. Biosystems Engineering, 2018, 171: 91−100.
[9] 范维, 高晓月, 董雨馨, 等. 基于数据挖掘建立北京地区牛、羊肉串掺假风险预测模型[J]. 食品科学, 2020, 41(20): 292−299. [FAN W, GAO X Y, DONG Y X, et al. Establishment of risk prediction model for adulterated beef and lamb kebabs in beijing by data mining[J]. Food Science, 2020, 41(20): 292−299.]
[10] 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, 41(20): 292−299. [CHEN L, ZOU L H, MENG K X, et al. Application of long short-term memory neural network in early warning of lead risk in meat products[J]. Modern Food Science and Technology, 2020, 41(20): 292−299.]
[11] 魏泉增, 李瑞, 张卓栋, 等. 人工神经网络在鉴别不同工艺花生油中的应用[J]. 粮食与油脂, 2021, 34(11): 46−51. [WEI Q Z, LI R, ZHANG Z D, et al. Identification of the different technology of peanut oil based on artificial neural network[J]. Cereals & Oils, 2021, 34(11): 46−51.]
[12] SAPPATI P K, NAYAK B, VANWALSUM G P. Thermophysical properties prediction of brown seaweed (Saccharina latissima) using artificial neural networks (ANNs) and empirical models [J]. International Journal of Food Properties, 2019, 22(1): 1966−1984.
[13] LING W, WANG G W, NING X J, et al. Application of BP neural network to prediction of coal ash melting characteristic temperature[J]. Fuel, 2020, 260: 1−8.
[14] 薛薇, 陈欢歌. SPSS Modeler 数据挖掘方法及应用[M]. 北京: 电子工业出版社, 2014. [XUE W, CHEN H G. SPSS Modeler data mining methods and applications[M]. Beijing: Electronic Industry Press, 2014.]
[15] 王若衡, 汪生, 王钰麟, 等. 肉制品 BP 神经网络风险预警模型构建研究[J]. 中国食品工业, 2022(6): 114−118,128. [WANG R H, WANG S, WANG Y L, et al. Construction of BP neural network risk warning model for meat products[J]. China Food Industry, 2022(6): 114−118,128.]
[16] DEVI B R, RAO K N, SETTY S P. et al. Disaster prediction system using IBM SPSS data mining tool[J]. International Journal of Engineering Trends and Technology, 2013, 4(8): 3352−3357.
[17] 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. [LI X M, ZANG M W, ZHAO H J, et al. Analysis and prediction of meat product safety based on supervision and sampling data[J]. Meat Research, 2019, 33(1): 42−49.]
[18] LINOFF G S, BERRY M J. 数据挖掘技术[M]. 巢文涵, 张小明, 王芳, 译. 北京: 清华大学出版社, 2013. [LINOFF G S, BERRY M J. Data mining technology[M]. CHAO W H, ZHANG X M, WANG F, Trans. Beijing: Tsinghua University Press, 2013.]
[19] 田兴国, 陈江涛, 吕建秋. 基于数据挖掘的兽药质量风险预测[J]. 现代食品科技, 2017, 33(11): 212−218. [TIAN X G, CHEN J T, LÜ J Q. Quality-risk prediction of veterinary drugs by data mining[J]. Modern Food Science and Technology, 2017, 33(11): 212−218.]
[20] 周辉, 张志转. 中国蔬菜农业污染现状、污染来源及污染防控[J]. 农业灾害研究, 2013, 3(5): 27−38,50. [ZHOU H, ZHANG Z Z. Agriculture pollution situation, source and countermeasures of chinese vegetable[J]. Journal of Agricultural Catastrophology, 2013, 3(5): 27−38,50.]
[21] 杨婕, 黄少文, 孙远明, 等. 4-氯苯氧乙酸钠对绿豆芽生长的影响及其残留分析[J]. 食品工业科技, 2015, 36(15): 104−108. [YANG J, HUANG S W, SUN Y M, et al. Analysis of sodium 4-chlorophenxyacetate on mung bean sprouts growth and residue[J]. Science and Technology of Food Industry, 2015, 36(15): 104−108.]
[22] KRZESLOWSKA M, RABĘDA I, BASIŃSKA A, et al. Pectinous cell wall thickenings formation-A common defense strategy of plants to cope with Pb[J]. Environmental Pollution, 2016 (214): 354−361.
[23] 李富荣, 李敏, 杜应琼, 等. 茄果类蔬菜对其产地土壤重金属的吸收富集与安全阈值研究[J]. 农产品质量与安全, 2018(1): 52−58. [LI F R, LI M, DU Y Q, et al. Solanaceous vegetables absorption and accumulation of heavy metal in soil of producing area and its safety threshold[J]. Quality and Safety of Agro-products, 2018(1): 52−58.]
[24] 胡霓红, 文典, 王富华, 等. 珠三角主要工业区周边蔬菜产地土壤重金属污染调查分析[J]. 热带农业科学, 2012, 32(4): 67−71. [HU M H, WEN D, WANG F H, et al. Investigation and analysis of heavy metals in vegetable producing soil around main industrial areas in the pearl river delta[J]. Chinese Journal of Tropical Agriculture, 2012, 32(4): 67−71.]
[25] 陈志良, 黄玲, 周存宇, 等. 广州市蔬菜中重金属污染特征研究与评价[J]. 环境科学, 2017, 38(1): 389−398. [CHEN Z L, HUANG L, ZHOU C Y, et al. Characteristics and evaluation of heavy metal pollution in vegetables in guangzhou[J]. Environmental Science, 2017, 38(1): 389−398.]
[26] VALIANT L. A Theory of the Learnable[J]. Communications of the Acm, 1984, 27(11): 1134−1142.
[27] AL-SALEMI B, AYOB M, NOAH S A M. Feature ranking for enhancing boosting-based multi-label text categorization[J]. Expert Systems with Applications, 2018, 113: 531−543.
[28] 冯曙明, 冯佳禹, 杨永成, 等. 基于改进的 Boosting 算法的仓库监控区域目标跟踪研究[J]. 微型电脑应用, 2020, 36(5): 76−79. [FENG S M, FENG J Y, YANG Y C, et al. Research on target tracking of warehouse monitoring area based on improved boosting algorithm[J]. Microcomputer Applications, 2020, 36(5): 76−79.]
[29] LIU T J, HU A Q. Model of combined transport of perishable foodstuffs and safety inspection based on data mining[J]. Food and Nutrition Sciences, 2017, 8: 760−777.
[30] 向晖. 个人信用评分组合模型研究与应用[D]. 长沙: 湖南大学, 2011. [XIANG H. Research on ensemble model for credit scoring and its application[D]. Changsha: Hunan University, 2011.]
[31] 王强, 冯玲然, 余晓斌. 基于 BP 神经网络和遗传算法优化番茄红素发酵培养基[J]. 食品与生物技术学报, 2019, 38(2): 111−119. [WANG Q, FENG L R, YU X W. Medium optimization for the production of lycopene based on BP neural network and genetic algorithms[J]. Journal of Food Science and Biotechnology, 2019, 38(2): 111−119.]