基于大数据视角的肉类食品安全抽检数据分析

王 博[1,2], 张锴旺[3], 陆逢贵[1], 刘登勇[1,4*], 曹振霞[1]

(1. 渤海大学食品科学与工程学院, 生鲜农产品贮藏加工及安全控制技术国家地方联合工程研究中心, 锦州 121013; 2. 渤海大学化学化工学院, 锦州 121013; 3. 河北联合轻工学院电气信息学院, 唐山 063000; 4. 肉类生产与加工质量安全控制协同创新中心, 南京 210095)

摘 要: 目的 基于大数据视角分析肉类食品安全抽检数据。**方法** 在大数据技术支持下以国家肉类食品抽检监测数据源为基础, 通过 Python 语言编程设计分类与预测实验, 并利用数据挖掘预测集实验结果与真实食品检验结果进行对比研究, 以验证该方法的可行性。**结果** 基于决策树+典型相关系数和二次指数平滑法相结合的数据挖掘方法分类效果较好, 预测准确性达到 98.26%。**结论** 通过预判不合格肉类食品的出现数量和分布情况, 可指导其安全抽检监测工作, 提高管理的效率和准确率, 有效预防肉类食品安全事故的发生。

关键词: 大数据; 数据挖掘; Python; 肉类食品

Study on meat food safety sampling data under the perspective of big data

WANG Bo[1,2], ZHANG Kai-Wang[3], LU Feng-Gui[1], LIU Deng-Yong[1,4*], CAO Zhen-Xia[1]

(1. *College of Food Science and Technology*, *Bohai University*, *Food Safety Key Laboratory of Liaoning Province*, *National and Local Joint Engineering Research Centre of Storage*, *Processing and Safety Control Technology for Fresh Agricultural and Aquatic Products*, *Jinzhou* 121013, *China*; 2. *College of Chemistry and Chemical Engineering*, *Bohai University*, *Jinzhou* 121013, *China*; 3. *Hebei United Light Industry College*, *Tangshan* 063000, *China*; 4. *Collaborative Innovation Center of Meat Production and Processing*, *Quality and Safety Control*, *Nanjing* 210095, *China*)

ABSTRACT: Objective To analyze meat food safety sampling data under the perspective of big data. **Methods** With the support of big data technology and based on the national meat food sampling inspection and monitoring data source, the classification and prediction experiment was designed by Python language programming, and the experimental results of the prediction set were compared with the real food inspection results to verify the feasibility of this method. **Results** The data mining method based on decision tree+typical correlation coefficient and quadratic exponential smoothing method had better classification effect, and the prediction accuracy reached 98.26%. **Conclusion** By predicting the quantity and distribution of unqualified meat foods, it can guide its safety sampling inspection and monitoring, improve the efficiency and accuracy of management, and effectively prevent the occurrence of meat food safety accidents.

KEY WORDS: big data; data mining; Python; meat food

基金项目: 辽宁省高等学校产业技术研究院重大应用研究项目(041804)、辽宁省重点研发计划指导计划项目(2017205003)

Fund: Supported by Major Applied Research Projects of Liaoning Institute of Industrial Technology (041804), and Key R&D Program of Liaoning Province (2017205003)

***通讯作者:** 刘登勇, 博士, 教授, 主要研究方向为肉品加工与质量安全控制。E-mail: jz_dyliu@126.com

***Corresponding author:** LIU Deng-Yong, Ph.D, Professor, Bohai University, No.19, Keji Road, New Songshan District, Jinzhou 121013, China. E-mail: jz_dyliu@126.com

1 引 言

随着国民经济的发展和人们消费观念的转型升级, 饮食与健康越来越受到人们的关注, 食品行业和有关部门也将关注重点聚焦在食品的质量与安全方面[1]。食品抽检工作既是保障食品质量与安全的主要环节和重要枢纽, 同时也是预防食品安全事故最根本、有效和可行的手段。食品安全抽检工作需要定期制定方案, 进行抽样和跟踪分析[2], 其中包括对微生物[3]、营养成分[4]以及添加剂等指标进行检测[5], 并以此为依据做出相应判断, 进而抑制不合格食品的流通与销售, 从而保障食品的质量与安全, 避免食品安全事故的发生。

食品安全抽检数据包括生产企业的名称、地址、被抽检样品的规格型号、生产日期以及检验结果等信息, 会被食品安全监管部门详实记录, 并以此为依据进行食品安全监管。目前国内外对于食品安全抽检数据研究却涉及较少, 主要集中在食品安全现状分析、食品风险分析, 以及食品生产与其他领域的交叉研究。Kleboth 等[6]构建了欧洲食物控制和审计系统框架, 并进行了风险分析; Lee 等[7]对相关数据进行存储和分析, 并划定食品安全级别, 消费者可以利用手机读取食品的安全级别; 李笑曼等[8]通过神经网络预测模型, 分析并预测了 2015~2017 年我国肉与肉制品主要安全现状与风险种类; 黄湘鹭等[9]通过国家食品药品监督管理总局发布的2016~2017年食品安全监督抽检结果, 对不合格项目进行归类分析; Lake 等[10]将食品安全数据与气候变化数据进行结合, 分析了高收入西方国家食源性疾病的诱因。常见的数据分析技术虽可以进行分析, 但在数据分析数量和非结构化数据处理能力等方面表现较差, 面对来源丰富、类型多样, 可能会存在大量非结构化数据的食品数据, 传统方式分析结果的准确性较低。因此需要能处理庞大数据量, 且数据容错率较高、来源丰富、整体数据展现能力强、非结构化数据处理能力强等优点的大数据挖掘方式。

鉴于此, 本文在大数据技术支持下, 采集“国家食品药品监督管理总局(China Food and Drug Administration, CFDA)”官方网站的肉类食品相关数据, 借助“决策树+典型相关系数(canonical correlation coefficient, CCA)”和二次指数平滑预测法等方法, 探索食品自身属性与检验结果之间的内在规律和关联模式, 进而利用已知的属性信息预测食品的检验报告是否合格, 以期将预测结果作为检测工作的先验知识, 为重点检测哪些项目、判断检验结论是否有误等提供决策指导, 进而提高抽检工作的效率, 达到预防食品安全事故的目的。

2 材料与方法

2.1 实验材料

本文以“国家食品质量安全监督检验中心”(http://www.cfda.com.cn/default.aspx)的抽检测试数据为基础, 辅以爬取的部分数据构建研究数据集。CFDA 网站提供的数据为 2014~2017 年(其中 2014 年数据无关值和缺失值数量较多, 影响数据分析结果, 因此在分类和预测过程中分析 2015~2017 年 3 年的数据)。CFDA 网站肉类食品数据标签为“肉及肉制品监督抽检合格/不合格产品信息”、“肉制品监督抽检合格/不合格产品信息”和“肉类合格/不合格”3 类, 以 2015 年和 2016 年数据为训练集, 2017 年数据为预测集, 预测 2017 年的数据变化, 用以检验预测集和真实数据之间的差异, 验证该方法的可行性。

2.2 设备与测试环境

本研究使用的是 Y7000P-拯救者笔记本电脑, 中国联想公司; 搭建的 Python 语言数据分析环境(硬件: CPU Intel Core I53320M, GPU GTX 740M - 1G; 系统 Windows 7), 主要使用了开源的第三方 Python3.7 扩展模块 matplotlib、numpy、pandas, 实现了回归、分类以及预测等多种数据挖掘方法[11,12]。

2.3 实验设计

利用 Python 软件的“(Beautiful Soup, bs4)”模块来编写定向爬虫程序。设置起始统一资源定位符(uniform resource locator, URL), 并通过 fiddler 采集“国家食品药品监督管理总局(CFDA)”网站的 Cookie 的主题网页信息。利用网页源码中带有链接的标记, 如
、<img>、<font>、<script>、<不合格数据>、<合格数据>等参数进行数据清洗, 最后对结构化的信息进行提取和保存[13]。

数据挖掘主要分为 2 个部分: 第一部分是数据的分类, 首先筛选特征变量, 再读取数据集中的所需信息, 进行分类处理, 将采集到的不合格样品属性信息作为输入, 已知的检验结果作为输出, 训练模型反映的是食品属性与食品检验结果之间的关系, 进而分析模型的性能; 第二部分是数据的预测, 即分析训练集的数据变化规律, 训练数据、并利用潜在关系预测, 最后与真实数据比较得到预测集的准确率。

2.4 数据挖掘方法

分类技术(classifier)也被称为分类器, 是指根据已有数据集的特征找出描述并区分数据类型的模型。在众多分类算法中, 决策树网络算法具有分类准确率较高、训练速度较快、算法性能较好、泛化性能较强等优点[14−18]。食品检验结果的分类, 本质为寻找和分割特征变量, 将具有多维属性, 且不同取值的食品数据准确的分类到相应的类别中, 这些特点与决策树数据挖掘方法的功能相吻合, 同时, 相对于其他数据挖掘算法, 决策树模型的大小与数据库大小无关, 具有分类精度高、生成模式简单、对噪声数据(如空值和错误值等)和体量较少数据集均有较好的健壮性优

势[19,20], 并且也符合“国家食品质量安全监督检验中心”公布的肉类食品安全抽检数据的特点。此外, 为避免仅采用决策树分类而导致效果较差的情况, 结合 CCA 的优点, 采用“决策树+CCA”对食品抽样检测数据进行分类。

样本分类后采用指数平滑法对食品抽样检测数据进行预测。指数平滑预测是时间序列分析的重要分支, 按照平滑次数的不同, 又可分为一次、二次和三次指数平滑预测法[21], 其中二次和三次指数平滑预测分别是对一次和二次指数平滑的再平滑, 但文章数据不具有适合 3 次平滑的变化趋势, 因此采用二次指数平滑法。二次指数平滑法用于预测不合格数据, 仅需选择一个模型参数, 对于相对稳定、有序的研究对象预测效果较好, 同时, 肉类食品安全抽检数据集的实际情况跨度较短, 数据量较少, 符合二次指数平滑法所需数据资料少的特点[22−25]。综上论述, 文章采用二次指数平滑法对食品抽样检测数据进行预测。

2.4.1　数据预处理

为了保证输入数据的质量, 并将形态调整为适于模型分析的形态, 需要对数据进行预处理。结合抽检数据的特点对训练数据集进行如下处理:

1) 数据集存在大量的空值(不占用字符空间)、零值(NULL)(占用字符空间)、“未提供”值, 因“决策树+典型相关系数”可以处理此类缺失值, 所以仅需将无关值删除后即可进行分类[26,27]。

2) “标称生产企业地址”属性值包括“省”、“市”等信息标签, 数据较为全面。将地名进行切片处理, 得到最关键的区域信息。如“湖北省荆州市公安县斗湖堤镇”切片保存为“湖北省”, “上海市徐汇区斜土路 1995 号”切片保存为“上海市”。

3) 原始数据中“不合格项目”的属性值包括“蛋白质”、“柠檬黄”、“莱克多巴胺(限畜肉)”、“过氧化值(以脂肪计)”等各种指标的含量, 需对部分重复测量指标进行合并, 去除多余信息, 提高分类的速率和准确度。共分为“营养指标”、“食品添加剂”、“药物残留”、“货架期”、“微生物指标”和“化学有害物质残留”6 类, 如“大肠菌群”“56000 CFU/g”“金黄色葡萄球菌”“930 MPN/100 g”统一为“微生物指标”, 将名词性属性进行整合。

2.4.2　数据分类

数据分类过程主要分为 3 个阶段, 第一阶段, 构建决策树, 将二维数据源列表(list)作为实现数据的结构, 省市为父节点, 行政区为子节点, 形成上下级父子关系; 第二阶段, 数据读取, 读取肉类食品安全数据表格, 切片处理清洗筛选后的数据, 分类得出一个省或者直辖市的数据, 以字典(dict)结构进行储存; 第三阶段, 分类输出, 值(keys)为各省或者直辖市, 对应的数值为各子类的统计数据, 重复遍历所有省份数据, 储存到 dict, 输入至 excel 进行记录。如父节点北京市包含东城区、西城区和崇文区等子节点, 在分类中确保数值与键值对应, 将数据反向读取并比较结果, 结果相同, 再存入表格, 并作为最终分类结果。该部分功能可以保证自动识别数据, 进行归类, 同时也为后续的数据分类和未来的数据库管理系统(Mysql)的构建提供可能。

采用决策树算法整理归类的数据使用 CCA 方法再次进行处理, 计算出每个不合格项目的指标平均值, 建立协方差矩阵, 综合考虑不同指标对之间的相关关系来反映各区域之间的整体相关性, 根据相关性作为判断分类结果的标准。

2.4.3　二次指数平滑法预测

大数据按照时间顺序依次进行分类归集, 以获取 3 个年份的不合格项目统计信息。第一步, 删除无关数据和不完整数据, 不给予预测以及评估; 第二步, 分类并计数整理后的数据, 以 dict 结构实现, 键值为不合格项目, 数值为 counts, 进行计算; 第三步, 反向读取计算机结果, 进行对比分析, 数据准确无误后计入表格, 为下一步评估与预测做准备; 第四步, 分别读取已知数据并集中 3 个年份某一类别的不合格项目, 进行指数 a 的评估, 并利用平滑指数进行预测, 与真实值比较并计算准确率。指数平滑法预测的核心是平滑参数 a 的获取, 本研究使用全局优化算法进行计算, a 从最小的可能性开始每次增加 0.01 至最优化的平滑参数, 具体过程如公式(1)所示; 二次指数平滑法基本公式如公式(2)所示。

$$yt+m=(2+am/(1-a))yt'-(1+am/(1-a))yt=(2yt'-yt)+m(yt'-yt)a/(1-a) \quad \text{式(1)}$$

式中, $2yt'-yt$--截距;

$(yt'-yt)a/(1-a)$--斜率;

a--平滑参数;

t--预测天数。

$$St=\alpha St+(1-\alpha)St-1\,Yt+T=at+btT\;at=2St-St\;bt=(\alpha/1-\alpha)(St-St)$$

$$St=aSt=(1-a)St-1Yt+T=at+btT\;at=2St-St\;bt=(a/1-a)(St-St) \quad \text{式(2)}$$

式中, St--第 t 期的一次指数平滑值;

$St-1$--第 t 期的二次指数平滑值;

α--平滑系数;

$Yt+T$--第 $t+T$ 期预测值;

T--由 t 期向后推移期数。

3　结果与分析

筛选、过滤后的预处理的部分数据集共 1014 条, 其结构如表 1 所示, 导入数据分析环境, 进行数据描述分析、分类和预测。

3.1　数据描述分析

文章数据一共收集了 26 个地区(省、直辖市)6 大类共 1014 条不合格项目, 统计整理不同地区不合格项目出现的频数, 用以描述该地区的肉类食品安全状况, 同时利用 ArcGis10.2 软件绘制出不合格项目分布地图, 以便于划分肉类食品安全等级, 结果如图 1 所示。

表 1 研究数据集
Table 1 Research dataset

序号	标称生产企业名称	标称生产企业地址	被抽样单位名称	被抽样单位地址	食品名称	规格型号	商标	生产日期	不合格项目	检验结果	标准值
1	上海云阳食品有限公司	上海嘉定马陆镇育绿东路609弄6号	上海联家超市有限公司徐汇店	上海市徐汇区斜土路1995号	五花腊肉	计量称重	云阳	2015-02-07	酸价	7.8 (KOH)/(mg/g)	≤4.0 (KOH)/(mg/g)
…	…	…	…	…	…	…	…	…	…	…	…
1014	江苏农大肉类食品有限公司	江苏市溧水区白马镇食品园大道11-3号(江苏白马农业高新技术产业园区)	亚马逊中国特产江苏馆	江苏市溧水区白马镇食品园大道	黄教授烧鸡	500 g/袋	黄教授	2017/7/22	单核细胞增生李斯特氏菌	检出; 0; 0; 0; 检出/25 g	*n*=5, *c*=0, *m*=0

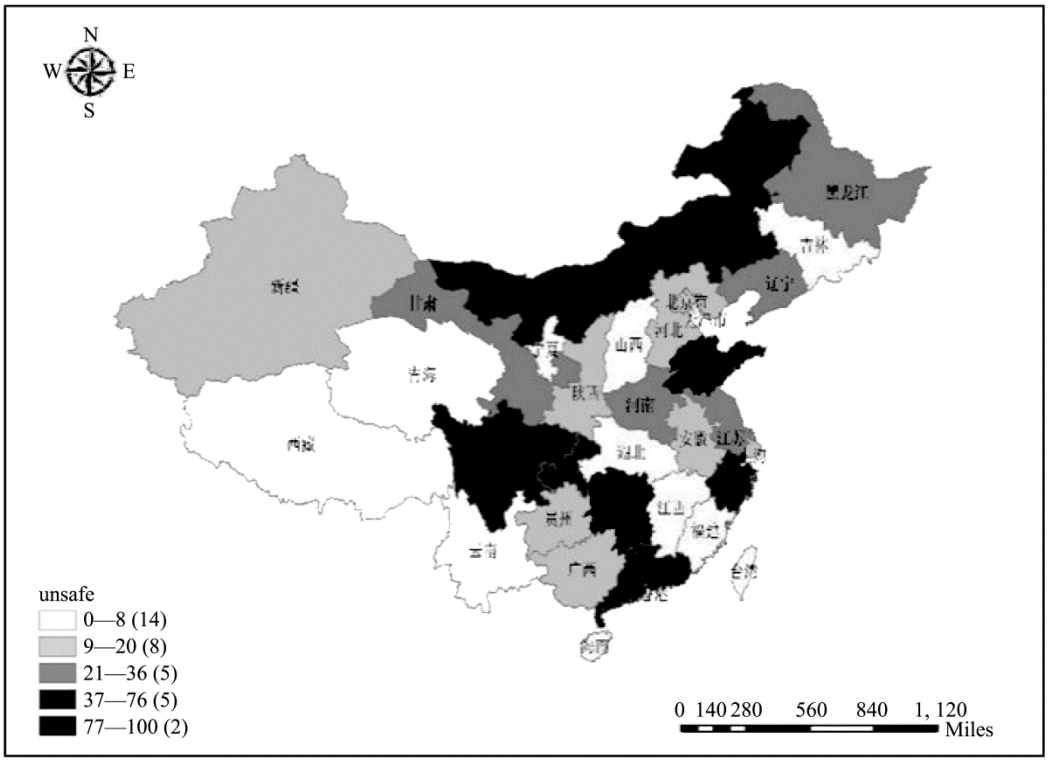

图 1 不合格项目分布图
Fig.1 Unqualified project distribution map

由图 1 可知, 根据不安全项目出现的次数, ArcGis 内置算法将我国划分为 5 类区域, 依次分布结果如表 2 所示, 第一类是肉类食品安全度最高的区域, 第五类是肉类食品安全度最低的区域, 以此类推。其中湖南省不合格项目出现总频数最高, 达 100 项, 占总数的 9.61%, 其他总频数排名较高的省份依次为山东省、内蒙古自治区、浙江省、重庆市、四川省, 而广东省占比在 6.15%~8.45%区间, 不合格频数较高。从图中可以看出: 除了内蒙古自治区、山东省和广东省之外, 其他 4 省市分别位于长江流域上游、中游和下游地区, 彼此之间经济和贸易互通往来频繁, 形成了较为密集的高危区域; 其次高危区域以湖南省和山东省为中心, 辐射周边城市, 导致整个区域肉类食品安全事件频发, 应予以重视。

为进一步明晰 2015-2017 年 26 省市的不合格数据的内在规律性, 以不合格项目为思考出发点, 观察其在各城市的分布规律, 生成雷达图, 结果如图 2 所示。

由图 2 可知, 矩形树状图的面积大小代表了项目出现的频率, 6 个项目面积依次为“微生物指标”>“食品添加剂”>“化学有害物质残留”>“营养指标”>“药物残留”>“货架期”, 说明“微生物指标”和“食品添加剂”是出现频率最高的

两个项目, 分别为 333 次和 280 次, 占总数的 72.89%, 是肉类食品最容易出现不合格的项目指标; 此外, 不同项目对应的高频区域的分布状况, “微生物指标”对应的内蒙古自治区不合格项目出现频率最高, “食品添加剂”对应的湖南省不合格项目出现频率最高, “化学有害物质残留”对应的甘肃省不合格项目出现频率最高, 安徽省对应的“营养指标”不合格项目出现频率最高, 广东省对应的“货架期”不合格项目出现频率最高, 对比发现其规律性并不明显, 较为分散。

3.2　数据分类

利用决策树对抽样数据进行处理, 并进行分类分析, 结果如图 3 所示。

图 3A 是将 6 个不合格项目标签当作一个维度为 6 的向量, 对 26 个向量做 PCA 降维, 映射到二维平面中形成的二维图像, 由图可以看出 26 个省市彼此之间距离较远, 没有明显出现类似聚类的效果, 说明各省市间相关性较差, 联系性不强; 图 3 B 为决策树聚类后的结果图, 由图可以看出共划分为四类, 且并没有团聚及聚类效果交叉现象; 图 3 C 和图 3D 是原始数据进行归一化后映射的二维图像和聚类效果图, 聚类效果较归一化前有所加强, 但分类效果仍较差, 规律不明显。因此本研究采用 CCA 方法对决策树处理后的数据进行进一步处理和完善, 部分结果见表 3、4 所示。

因研究数据分析呈现结果较多, 本文以食品添加剂为例进行分析, 表 3 中天津市和上海市的典型相关系数相同, 说明两个城市在该指标属性上相似, 联系性较强。以此方法对全部数据进行分析, 如表 4 所示, 结果表明食品添加剂、微生物指标和化学有害物质残留项目对应的区域能形成较强的联系, 并进而形成不同类型的分区, 在其他的指标下不同城市之间差别较小, 不易分类。根据分类结果可以指导后续的实际工作, 当某些区域之间关系密切, 当一个区域发生某种不合格项目数量过多情况时, 关系密切的其它区域也会发生类似事件, 比如重庆市食品添加剂的不合格项目过多, 广东省也有可能出现类似的情况, 可提高食品安全风险防护的准确度和效率。文章分类效果不理想的原因主要是由于数据集周期较短、不合格项目训练样本较少、数据质量较差导致的, 存在很大的改进空间。

表 2　不合格项目分布表

Table 2　Unqualified project distribution table

	第一类(8)	第二类(8)	第三类(5)	第四类(5)	第五类(2)
城市名称	吉林、天津、陕西、宁夏、青海、西藏、云南、海南、台湾、福建、江西、湖北香港、澳门	新疆、北京、河北、陕西、贵州、广西、安徽、上海	黑龙江、辽宁、甘肃、河南、江苏	内蒙古、四川、重庆、广东、浙江	湖南、山东

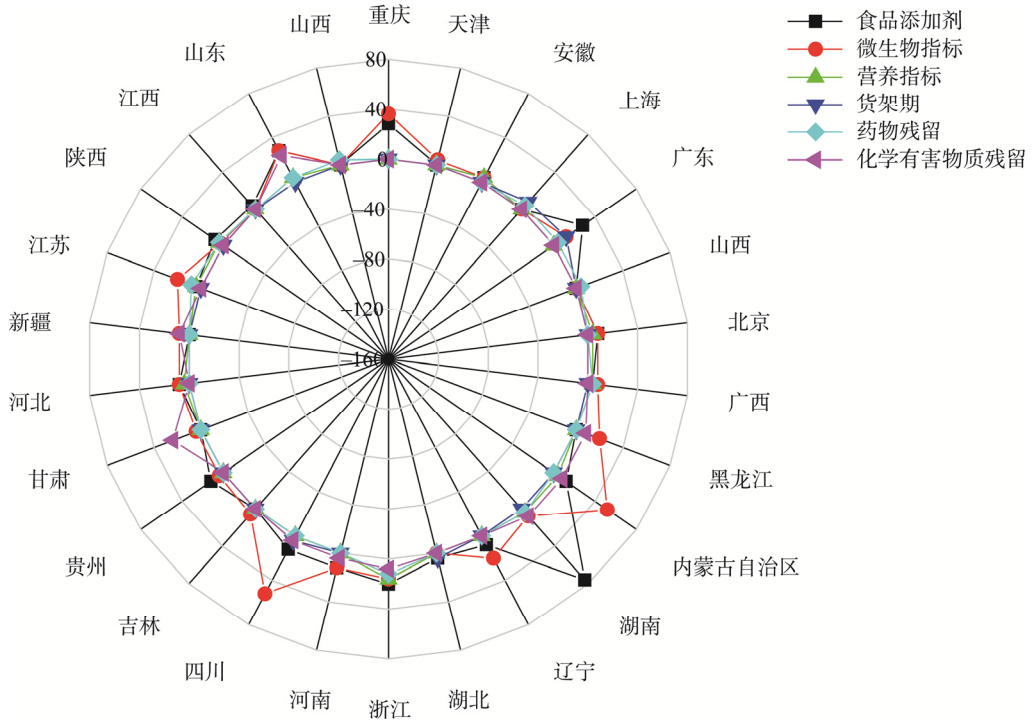

图 2　不合格项目雷达图

Fig.2　Radar map of unqualified projects

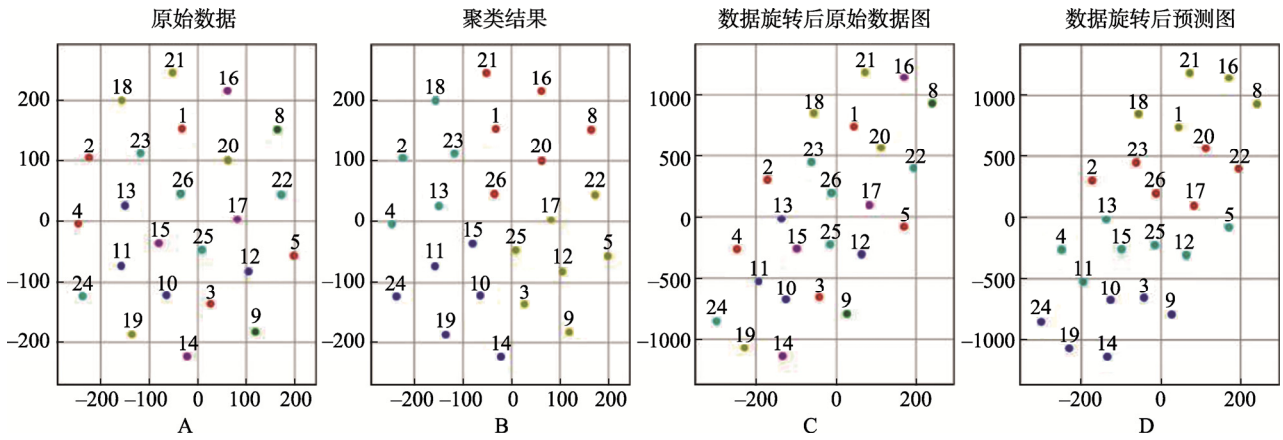

图 3 分类图示

Fig.3 Graph of classifier

表 3 模型评价指标

Table 3 Model evaluation indicators

地区	不合格项目											
	食品添加剂	典型相关系数	微生物指标	典型相关系数	营养指标	典型相关系数	货架期	典型相关系数	药物残留	典型相关系数	化学有害物质残留	典型相关系数
重庆	28	−74.0769	36	−134.308	0	−1	0	0	0	−1	0	−2.69231
天津	0	−28.8462	4	−19.2308	0	−1	0	0	0	−1	0	−2.69231
安徽	4	−11.3077	4	−19.2308	4	−1	0	0	0	−1	0	−2.69231
上海	0	−28.8462	0	−40.8462	0	−1	8	−11.6923	4	−1	0	−2.69231
…	…	…	…	…	…	…	…	…	…	…	…	…

表 4 分类结果

Table 4 Classification results

	不合格项目					
	食品添加剂	微生物指标	营养指标	货架期	药物残留	化学有害物质残留
聚类区	天津、上海、黑龙江、吉林、新疆	天津、安徽、贵州、甘肃	重庆、天津、安徽、上海、广东、山西	重庆、天津	重庆、天津、安徽、上海	重庆、天津
	内蒙古、河南、四川、贵州、甘肃	上海、山西、湖北、江西	北京、广西	安徽、山西	广东、山西、北京、广西	安徽、上海
	重庆、广东	北京、广西、湖南、河北、新疆	黑龙江、内蒙古、湖南	北京、广西	黑龙江、内蒙古、湖南	广东、山西
	安徽、广西、湖北、江苏	广东、河南	辽宁、湖北河南、四川	黑龙江、内蒙古、湖南、辽宁、河南、吉林、贵州、甘肃、河北、新疆、江苏、陕西、江西、山东	辽宁、湖北河南、四川	北京、广西
	北京、辽宁、河北		吉林、贵州		吉林、贵州	辽宁、湖北
	天津、上海、黑龙江、吉林、新疆	天津、安徽、贵州、甘肃	甘肃、河北		甘肃、河北	吉林、贵州
			新疆、江苏		新疆	河北
			陕西、江西		陕西、江西	陕西、江西
			山东		山东	山东
				四川、湖北		河南、四川
						黑龙江、内蒙古、湖南、浙江、新疆

3.3　数据预测

从每年的数据集中抽取各类别的不合格项目进行平滑指数 a 的评估, 并根据结果进行预测, 验证该方法的可行性。因目前仅有 3 年的有效数据, 因此使用 2015 年和 2016 年的监测数据预测 2017 年的监测数据, 并与 2017 年的监测数据真实值比较计算准确率。平滑指数评估结果和 2017 年各指标预测部分结果如表 5、6 所示.

以表 5 和表 6 中的 2017 年预测可知, 抽样工作抽取 2015 年和 2016 年的检测数据, 属性为: 不合格项目比如苯甲酸和菌落总数, 以及对应项目的出现的次数 4 和 50, 利用已训练好的二次平滑指数(分别为类平滑指数和年平滑指数, 其中年平滑指数为–1.858)预测出 2017 年可能出现的不合格项目过氧化值 3 项、脂肪 1 项和菌落总数 41 项, 检测人员参考预测结果需对该样品的过氧化值、脂肪和群落总数检验项目进行重点检测。研究方法预测值为 134 项, 2017 年实际不合格项目为 136 项, 预测准确率达到 98.26%。在该准确率下, 检测人员可以根据分类结果有选择的设置重点检验项目或者高发区域, 提前发布预检测结果并进行预警, 在节约时间和财力的同时, 起到监督和保障作用。

表 5　模型预测指标
Table 5　Evaluation indexes of model

项目	a 值	2017 年预测值	项目	a 值	2017 年预测值
苯并〔a〕芘熏烧烤工艺) 1.0	0.99998999999089700000	1.0000200000182033	大肠菌群	1.19999999998998000000	9.8999999999956
蛋白质	3.00019999998179000000	5.249992499962449	镉	9.99999954485186000000	2.00000000000000000000

表 6　预测实例
Table 6　Instances of prediction

2015 年	2016 年	2017 年
苯甲酸(4)	菌落总数(50)	过氧化值(3)
金黄色葡萄球菌(1)	柠檬黄(1)	脂肪(1)
酸价(4)	230 MPN/100g(1)	菌落总数(41)
……	……	……

4　结论与讨论

文章基于“决策树+典型相关系数(CCA)”和二次指数平滑法对 2015~2017 年 CFDA 抽检测试数据集进行数据挖掘, 结果如下:

(1) 现状分析结果表明湖南省是近年来肉类食品不安全事件高发区, 并且在内蒙古和我国中部、长江流域部分地区形成了高危区域; 研究涉及的检测指标 6 大类, 其中“微生物指标”和“化学有害物质残留”是肉类食品最易出现不合格项目的指标, 出现频率最高的城市分别是内蒙古自治区、湖南省、甘肃省、安徽省和广东省。结果说明食品安全不局限于一个省市的区域性问题, 因食品流通和经济往来, 一旦食品安全高发城市出现, 周围地区受其影响往往也会变成重灾区, 进而演变成较大范围内的食品安全问题, 应对类似的事件应予以重点关注, 提前进行预警判断。且还应该注意这种相互之间的辐射影响, 不单单是空间地理效应, 还包括贸易频繁导致的经济集聚效应。

(2) 分类过程中, 因仅使用决策树效果较差, 所以配合采用 CCA 方法处理数据。结果表明, 许多城市典型相关系数相同, 在该指标属性上表现为联系性较强, 可根据具体城市进行对应分析, 例如天津市和上海市在食品添加剂的方面联系性较强, 说明 2 个城市产品均大量检测出食品添加剂指标不合格; 食品添加剂、微生物指标和化学有害物质残留项目存在差异性显著, 形成了有效的三分类或者四分类, 而其他指标差异性较小, 说明各城市基本都具有了上述特征, 且围绕特征形成了不同的组团。

(3) 数据预测过程中, 利用 2015 年和 2016 年的检测数据作为训练集, 2017 年检测结果作为预测集, 采用二次平滑法预测准确率高达 98.26%。

(4) 文章基于大数据视角对肉类食品安全抽检数据开展挖掘研究, 结果表明, 在获得食品基本信息的情况下, 该方法可在实际检测工作之前, 准确预测出不合格项目的数量及分布状况, 对提高检测效率、增强食品安全的预防监测能力, 具有重要的意义。

参考文献

[1] 孙宝国, 王静. 中国食品产业现状与发展战略[J]. 中国食品学报, 2018, 18(8):1–7.
Sun BG, Wang J. The status of food industry in china and development strategy [J]. J Chin Instit Food Sci Technol, 2018, 18(8): 1–7.

[2] King T, Cole M, Farber JM, *et al*. Food safety for food security: Relationship between global megatrends and developments in food safety [J]. Trend Food Sci Technol, 2017, (68): 160–175.

[3] Singh BK, Trivedi P. Microbiome and the future for food and nutrient security [J]. Microbial Biotechnol, 2017, 10(1): 50–53.

[4] Allison, David B, Bassaganya-Riera, *et al*. Goals in nutrition science 2015-2020 [J]. Front Nutr, 2015, (2): 26.

[5] Hill AA, Crotta M, Wall B, *et al*. Towards an integrated food safety surveillance system: A simulation study to explore the potential of

combining genomic and epidemiological metadata [J]. Royal Soc Open Sci, 2017, 4(3):1–22.

[6] Kleboth JA, Luning PA, Fogliano V. Risk-based integrity audits in the food chain-A framework for complex systems [J]. Trend Food Sci Technol, 2016, (56): 167–174.

[7] Lee, Jongchan, Bahk GJ. Design of food management system using NFC tag [J]. J Korea Soc Comput Inf, 2018, 23(5): 25–29.

[8] 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 XM, Zang MW, Zhao HJ, *et al*. Analysis and prediction of meat product safety based on supervision and sampling data [J]. Meat Res, 2019, 33(1): 42–49.

[9] 黄湘鹭, 邢书霞, 吕冰峰, 等. 2016~2017 年我国食品安全抽检数据分析[J]. 食品安全质量检测学报, 2018, 9(17): 4746–4754.
Huang XL, Xing SX, Lv BF, *et al*. Analysis of national food safety supervision and sampling inspection in 2016–2017 [J]. J Food Saf Qual, 2018, 9(17): 4746–4754.

[10] Lake IR, Barker GC. Climate change, foodborne pathogens and illness in higher-income countries [J]. Current Environ Health Reports, 2018, 5(1): 191–214.

[11] Akossou AYJ, Attakpa EY, Fonton NH, *et al*. Spatial and temporal analysis of maize (zea mays) crop yields in Benin from 1987 to 2007 [J]. Agric Forest Meteorol, 2016, (220): 177–189.

[12] Clinton W, Brownley. Python 数据分析基.[M]. 北京: 人民邮电出版社, 2017.
Clinton W, Brownley. Foundations for analytics with Python [M]. Beijing: Posts and Telecommunications Press, 2017.

[13] 陈涛, 张旭, 崔扬, 等. Python 自然语言处理[M]. 北京: 人民邮电出版社, 2017.
Chen T, Zhang X, Cui Y, *et al*. Natural language processing with Python [M]. Beijing: Posts and Telecommunications Press, 2017.

[14] Teimouri N, Omid M, Mollazade K, *et al*. On-line separation and sorting of chicken portions using a robust vision-based intelligent modelling approach [J]. Biosystem Eng, 2018, (107): 8–20.

[15] Kim D, Hong S, Kim YT, *et al*. Metagenomic approach to identifying foodborne pathogens on Chinese cabbage [J]. J Microbiol Biotechnol, 2018, 28(2): 227–235.

[16] 李钦. 数据挖掘技术在热镀锌产品质量评定中的应用研究[D]. 兰州: 兰州理工大学, 2018.
Li Q. The application of data mining technology in the quality assessment of hot galvanizing products [D]. Lanzhou: Lanzhou University of Technology, 2018.

[17] 汪武裙. 基于决策树的煅烧工艺参数的研究与分析[D]. 北京: 北方工业大学, 2018.
Wang WQ. Process parameters of calcining based on decision tree research and analysis [D]. Beijing: North China University of Technology, 2018.

[18] 芦思雨. 数据挖掘中分类算法的比较分析[D]. 天津: 天津财经大学, 2016.
Lu SY. Comparing classifiers in data mining [D]. Tianjin: Finance and Economics University of Tianjin, 2016.

[19] Lu H, Du B, Liu J, *et al*. A kernel extreme learning machine algorithm based on improved particle swam optimization [J]. Memetic Comput, 2017, 9(2): 121–128.

[20] Manek AS, Shenoy PD, Mohan MC, *et al*. Aspect term extraction for sentiment analysis in large movie reviews using Gini index feature selection method and SVM classifier [J]. World Wide Web-internet Web Inf System, 2017, 20(2): 135–154.

[21] 童强, 张克功, 杜吉梁. 指数平滑预测法及其在经济预测中的应用[J]. 经济研究导刊, 2013, (4): 11–12.
Tong Q, Zhang KG, Du JL. Exponential smoothing forecasting method and its application in economic forecast [J]. Econ Res Guid, 2013, (4): 11–12.

[22] Thiel D, Vo TLH, Hovelaque V. Forecasts impacts on sanitary risk during a crisis: A case study [J]. Int J Logist Manag, 2014, 25(2): 358–378.

[23] Akossou AYJ, Attakpa EY, Fonton NH, *et al*. Spatial and temporal analysis of maize (zea mays) crop yields in Benin from 1987 to 2007 [J]. Agric Forest Meteorol, 2016, (220): 177–189.

[24] Ya B. AWNG-BP prediction technique study based on nonlinear combination-A case study of prediction of food supply chain in rural areas of Hubei province, China [J]. J Intellig Fuzzy System, 2018, 34(2): 761–770.

[25] Prasetyo SYJ, Agus YH, Dewi C, *et al*. Information system based on geospatial for early warning tracking and analysis agricultural plant diseases in central Java [J]. IOP Conf Series-Mater Sci Eng, 2017, (180): 1–9.

[26] 晁凤英, 杜树新. 基于关联规则的食品安全数据挖掘方法[J]. 食品与发酵工业, 2007, 33(4):107–109.
Chao FY, Du SX. Data mining technics for food safety based on association rules [J]. Food Ferment Ind, 2007, 33(4): 107–109.

[27] 王星云, 左敏, 肖克晶, 等. 基于 BP 神经网络的食品安全抽检数据挖掘[J]. 食品科学技术学报, 2016, 34(6): 85–90.
Wang XY, Zuo M, Xiao KJ, *et al*. Data mining on food safety sampling inspection data based on BP neural network [J]. J Food Sci Technol, 2016, 34(6): 85–90.

(责任编辑: 武英华)

作者简介

王　博, 主要研究方向为食品大数据分析。
E-mail: daqingwb@163.com

刘登勇, 教授, 主要研究方向为肉品加工与质量安全控制。
E-mail: jz_dyliu@126.com