基于数据挖掘和不同机器学习模型预测食品中金黄色葡萄球菌的种群行为

张琪果 陈一博 钟紫旋*

重庆第二师范学院生物与化学工程学院，重庆,400067

摘要：金黄色葡萄球菌是一种重要的病原菌。目前，针对食品中金黄色葡萄球菌的检测主要采用传统的培养方法和分子生物学技术。然而，这些方法往往耗时费力，且面临检测准确性不足、存在系统误差等问题。随着机器学习和数据挖掘技术的飞速发展，预测微生物学在预测食品中金黄色葡萄球菌的种群行为上展现出巨大的潜力。本研究利用数据挖掘和机器学习技术，建立预测食品中金黄色葡萄球菌种群行为的模型，首先对食品中金黄色葡萄球菌的相关数据进行搜集，涵盖初始微生物浓度、温度、pH 值、水分活度等诸多影响因素。随后，运用数据挖掘技术对所收集的数据进行预处理和特征提取，以发掘与金黄色葡萄球菌种群行为紧密相关的特征。在此基础上，采用机器学习算法构建预测模型，用以预测金黄色葡萄球菌的种群行为。在研究末尾，对所构建的模型进行验证与择优，确保其预测准确性和泛化能力。最后，对最优模型的不足之处和问题进行了展望。

关键词：金黄色葡萄球菌；机器学习模型；预测微生物学

DOI:10.69979/3041-0673.24.3.046

引言：

随着现代食品工业的快速发展，食品安全问题日益凸显，成为公众最为关注的话题之一。食品安全作为一项基本人权，可以挽救生命并促进个人和群体健康。食物本质上是生物的，为微生物提供了生长的环境，这些微生物可能成为食源性疾病的潜在来源[1]。

食源性疾病，无论是与微生物病原体还是其他食物相关，都是发展中国家和发达国家的重大健康挑战。据世界卫生组织估计，在工业化国家，食源性疾病的报告病例不足 10%，而在发展中国家可能更低，不到 1%。这意味着全球食源性疾病和死亡的数量可能十分庞大[2]。

近年来，细菌性食源性疾病的发病率呈现出持续上升的趋势。这种疾病的发病情况受到不同地区和国家饮食习惯的影响，导致感染细菌的种类和菌株具有一定的地区性和差异性[3]-[4]。

金黄色葡萄球菌（Staphylococcus aureus）是一种广受研究的病原菌，其能够引发一系列疾病，包括表面感染和多器官组织炎症，具有一定的流行病学特征，在公共卫生领域中被视为一个重要的研究课题[5]。除感染外，金黄色葡萄球菌还与毒素引起的食物中毒密切相关，这一特性使其成为食源性疾病研究的重要对象,农场和食品加工厂的环境、工人、牲畜组织黏膜等均有污染金黄色葡萄球菌的风险。

食用受金黄色葡萄球菌污染的食物后，由于产肠毒素的金黄色葡萄球菌菌株在食物中形成葡萄球菌肠毒素（SE），个体可能在极短的时间内（1-6 小时）出现严重的毒素介导疾病，包括胃肠炎、恶心、呕吐、腹泻和腹痛等症状。耐甲氧西林金黄色葡萄球菌（MRSA）作为金黄色葡萄球菌的一种重要致病菌株，已在多项研究中被证实可在奶酪、猪肉、牛肉等食品中引发疾病-。在我国，零售即食食品作为 MRSA 谱系传播的潜在载体，已引起学术界的广泛关注，并被视为一个严重的公共卫生风险。

因此，针对于金黄色葡萄球菌的种群预测十分重要，即研究金黄色葡萄球菌在不同环境条件下，细菌数变化和外部环境因素之间的响应关系。食品预测微生物学在其中能发挥重大作用。食品预测微生物学是食品微生物学中的一个新兴研究领域，旨在提供数学模型来预测食品环境中的微生物生长情况。预测模型的不同类型能够预测不同环境条件下食品中细菌的生长、滞后生长和死亡概率-。

机器学习 （ML）是人工智能（AI）的一个子领域，旨在从大规模不同数据中寻找特征。它的底层逻辑是使用算法来解析数据，自动分析数据中的模式，然后利用

这些模式对现实世界的事件做出预测和决策。通过跨尺度和复杂微生物群落的整合以及多组学的整合,ML 可用于系统地呈现微生物群落之间或与宿主的相互作用。从大型数据集生成的高维数据集中降维并提取空间特征的工作流程有助于探索微生物的功能潜力并扩大微生物技术应用的研究。与传统的机器学习相比，深度学习（DL）的维度更高，能够适用于庞大的序列数据集，并且有针对性地捕获原始数据中尽可能多的、完整的关系。它通过具有多层处理单元的深度网络对数据进行建模。

目前将深度学习应用于利用环境条件参数来预测目标微生物种群情况以及针对于金黄色葡萄球菌的种群预测研究的研究较少。本研究将采用五种机器学习模型贝叶斯模型（Bayes）、深度学习（DL）、逻辑回归模型（LR）、线性岭回归（Ridge 回归）、支持向量机（SVM），其中深度学习模型主要采用 CNN 卷积神经网络，来对从 Combase 数据库（www.combase.cc）中下载的金黄色葡萄球菌在猪肉、牛肉、奶酪以及培养基中的数据进行建模，并引入四大评价标准：决定系数（Coefficient of determination）R2、均方根误差（Root mean square error）RMSE、偏差系数（Bias factor）Bf、精度系数（Accuracy factor）Af，来进行模型评价。最终通过比较新型的机器学习与深度学习回归方法根据所提供的环境参数来预测金黄色葡萄球菌的生存和生长。

1. 材料与方法

1.1 收集数据

本研究数据来源于 Combase 数据库，该数据库汇集了众多研究机构和论文中的近 6 万个微生物量化记录，涵盖了微生物环境参数，如“记录 ID”、“微生物名称”、“食物类别”、“食物名称、”“温度”、“时间”、“ph 值”、“水分活度”、“处理方式”、“活菌数量”。通过对这些数据的分类与讨论，本研究能够对数据集进行分类以及分别讨论，并利用机器学习模型来预测金黄色葡萄球菌的种群行为。

本研究从 Combase 数据库中提取了四组数据集，分别为猪肉（336 个数据点）、牛肉（321 个数据点）、奶酪（2117 个数据点）以及培养基（7207 个数据点），总计包含 9981 个数据点用于模型的开发与评估。然而，部分数据在温度、时间、水分活度（Aw）等参数上存在缺失。在模型开发过程中，对于仅缺失一个参数值的数据点采用近似值进行补齐；而对于缺失两个及以上参数值的数据点，则直接予以舍弃。

1.2 数据预处理

在研究过程中，由于数据来源于公共数据库，各项数据难免有缺失值。在考量各个缺失值对于机器学习模型的影响之后，筛选出“初始微生物浓度 CO”、“温度 Temp”、“PH 值”、“水分活度 Aw”四个变量作为特征量，以“最终活菌数 LogCs”作为回归预测值。

首先读取变量与 LogCs 进行时间序列长度对比，如果是单一值，就处理为这段时间序列的恒定值，如果是多个值，就检查时间序列，并与之相对应，若有缺失值则使用临近差值法（选用与其最近的两个数值求平均值）或者经验值进行填充。

将所有数据集进行变量的重要性分析，以确定某个变量对金黄色葡萄球菌的生长影响最大。

1.3 建模

1.3.1 贝叶斯模型：

贝叶斯模型（Bayes）是一种利用贝叶斯公式（1）和假设条件独立性的基于概率统计分类算法，能提供很高的分类精度。并且由于贝叶斯的训练时间与训练实例的数量和属性的数量都是线性的，因而很适合采用时间序列进行预处理的数据集。在模型中，首先计算出每个类别的先验概率，再计算每个类别下各个的特征似然度，即每个特征在已知类别的情况下出现的概率，再根据贝叶斯公式计算后验概率。

$$P(B|A) = P(A|B) \times \frac{P(B)}{P(A)} \quad (1)$$

1.3.2DL 模型：

深度学习模型（Deep Learning Model，简称 DL 模型）是一种基于人工神经网络的深度学习模型，通常包括了 DNN 全链接神经网络、CNN 卷积神经网络等。其中 CNN 即卷积神经网络虽然主要用于二维图像数据的特征提取和处理，但也可用于处理一维数据（1-D CNN）,1-D CNN 通常用于处理一维数据，其特点是从整个数据集的固定长度片段中提取特征，而特征的位置不影响其有效性。这种特性使得一维 CNN 在处理时间序列预测和信号识别等任务上非常有效。Han 等人将一维 CNN 应用于短期公路交通流量预测。他们利用一维 CNN 来捕获交通流量的空间特征，并结合时间特征进行交通流量的预测。

1.3.3LR 模型：

逻辑回归模型（Logistic Regression Model，简称 LR 模型）是一种用于描述定性因变量（这类变量仅能取某些离散值）与自变量之间关系的统计模型。该模型主要应用于分析预测变量对分类结果的影响，其输出结果通常为二元形式。当模型中仅包含一个预测变量时，称之为简单逻辑回归模型。逻辑回归还具备分析特定自变量对控制混杂因素影响的能力，尤其是那些可能与结果及其它自变量存在相互作用的变量（即混杂变量）。

1.4 对模型质量的评估

为了比较各个模型的表现优良程度，本研究引入四个评价标准对它们进行评价，以便筛选出在金黄色葡萄球菌预测中表现最好的模型。分别是：决定系数（Coefficient of determination）R2、均方根误差（Root mean square error）RMSE、偏差系数（Bias factor)Bf 即对数平均误差(Mean Logarithmic Error)、精度系数（Accuracy factor）Af 即对数平均绝对百分比误差（Mean Absolute Percentage Error (MAPE)）。（3）-（6）

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n}\left(y_{S.aureus} - y_{pre}\right)^2}{\sum_{i=1}^{n}\left(y_{S.aureus} - \overline{y_{S.aureus}}\right)^2}\right] \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{S.aureus} - y_{pre}\right)^2}{n}} \quad (4)$$

$$B_f = \frac{1}{n}\sum_{i=1}^{n} 10 \cdot \log_{10}\left(\frac{y_{pre,i}}{y_{S.aureus,i}}\right) \quad (5)$$

$$A_f = \frac{1}{n}\sum_{i=1}^{n} 10 \cdot \left|\log_{10}\left(\frac{y_{pre,i}}{y_{S.aureus,i}}\right)\right| \quad (6)$$

式中，$y_{S.aureus}$为实验细菌金黄色葡萄球菌，y_{pre}为预测值，$\overline{y_{S.aureus}}$为种群计数的平均值，n 为样本数量。

通过上述四个指标来对五个模型进行综合评判（以 R2 和 RMSE 为主）；为防止过拟合，对各个模型再进行预测值和真实值的对比。以此筛选出最优模型，并进行模型外部验证。

2. 结果与讨论

2.1 数据收集及讨论

从提取的金黄色葡萄球菌生长数据集中所有数据点，包括水分活度 Aw、微生物初始浓度 C0、温度 Temp、PH 值以及最终活菌数 LogCs。将四个数据集(牛肉 Beef、猪肉 Pork、奶酪 Cheese 和培养基 Culture medium）分别绘制频数直方统计图（图 2-5），以便观察。并绘制了每个变量的最小值、最大值和标准差（σ）（表 1）。此外，对所有数据集进行了变量重要性探讨，并绘制成图（图 6）。

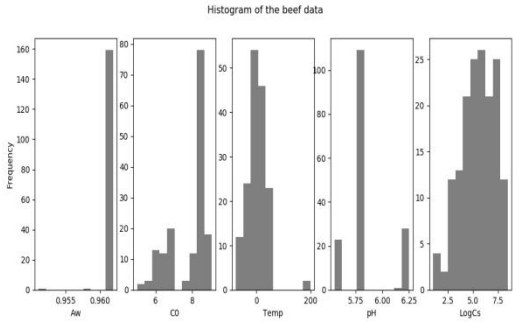


图 2 Beef 数据的频数直方统计图

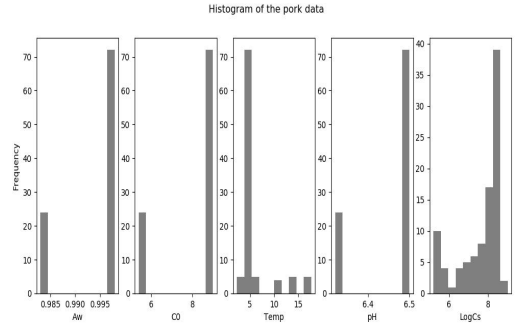


图 3 Pork 数据的频数直方统计图

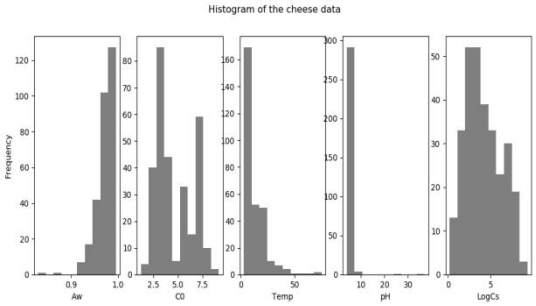


图 4 Cheese 数据的频数直方统计图

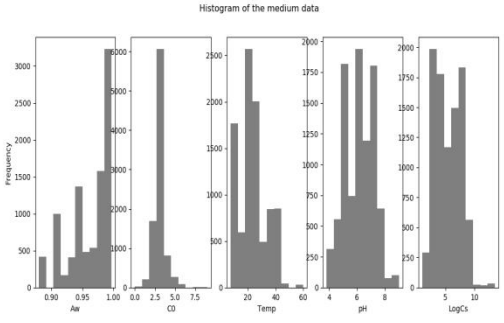
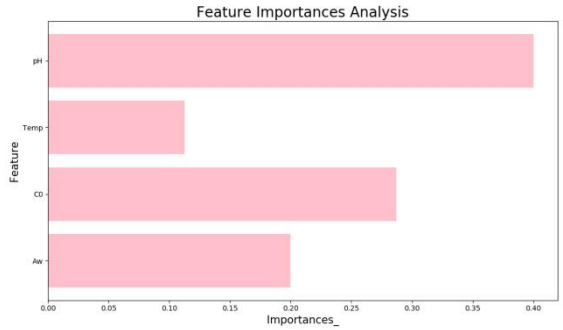


图 5 Culture Medium 数据的频数直方统计图

表 1 关于数据条件的全面细节

食品	温度（℃）			水分活度（Aw）			PH		
	最小	最大	σ	最小	最大	σ	最小	最大	σ
牛肉 Beef	4.00	26.00	7.74	0.89	0.96	0.01	5.58	6.90	0.29
猪肉 Pork	0.00	23.00	7.02	0.81	0.99	0.03	4.68	6.93	0.46
奶酪 Cheese	3.00	97.00	17.34	0.33	0.99	0.03	2.50	7.75	0.59
培养基 Culture Medium	7.50	60.00	10.18	0.86	0.99	0.03	3.80	9.00	0.99



2.2 建模与模型评估

利用五种机器学习模型：Bayes 模型、DL 模型、LR 模型、Ridge 模型和 SVM 模型，它们的模型质量用四种指标进行评估（表 2-5）。绘制图形并展示五种模型实际值和预测值的差异图（图 7）。以及它们的拟合效果（图 8）。

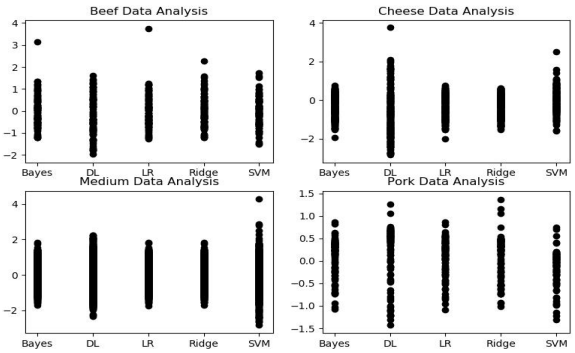


图 7 各数据集中不同模型的预测值和真实值的差异

表 2 牛肉 Beef 数据集中的模型参数

模型	R2	RMSE	Bf	Af
Bayes	0.95	0.73	-0.01	0.51
DL	0.92	0.93	0.22	0.63
LR	0.95	0.76	-0.14	0.64
Ridge	0.96	0.71	0.03	0.47
SVM	0.96	0.67	0.09	0.42

表 3 猪肉 Pork 数据集中的模型参数

模型	R2	RMSE	Bf	Af
Bayes	0.98	0.42	-0.05	0.21
DL	0.96	0.66	-0.05	0.35
LR	0.98	0.43	-0.05	0.21
Ridge	0.98	0.48	-0.03	0.24
SVM	0.98	0.42	0.09	0.17

表 4 奶酪 Cheese 数据集中的模型参数

模型	R2	RMSE	Bf	Af
Bayes	0.99	0.65	0.42	0.68
DL	0.94	1.40	1.13	1.49
LR	0.99	0.66	0.41	0.69
Ridge	0.99	0.63	0.49	0.68
SVM	0.99	0.56	0.10	0.59

表 5 培养基 Culture Medium 数据集中的模型参数

模型	R2	RMSE	Bf	Af
Bayes	0.95	0.63	-0.01	0.44
DL	0.92	0.77	-0.01	0.52
LR	0.95	0.63	-0.01	0.45
Ridge	0.95	0.63	-0.01	0.44
SVM	0.92	0.77	0.08	0.51

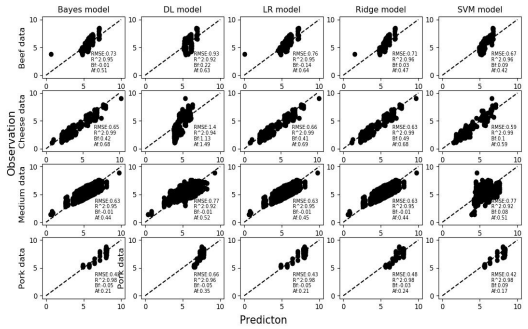


图 8 各模型在不同数据集上的回归效果

Bf 表示了预测值和实际值之间的对数平均误差，衡能够量预测值和实际值之间的对数差，可以表示预测值偏离实际值的程度，包括方向（高估或低估）。在本研究拟合数据图表中大量 Bf 为负值，代表预测值低估了实际值，即预测值小于实际值的程度，如 Bayes 模型在牛肉 Beef 数据集中，Bf=-0.01（在转换百分比中需要添加绝对值进行转换），表示预测值低估实际值 1%。

Af 表示了预测值和实际值之间的对数平均绝对百分比误差。它衡量的是预测值和实际值之间的相对误差，能够评估预测模型的准确性。例如 Bayes 模型在猪肉 Pork 数据集中 Af=0.21，表示预测值和实际值相差 21%

根据以上结果，特别是图 8 中可以直观的发现 Bayes 模型在各个数据集的误差上精度表现高且密集，

证明Bayes在预测金黄色葡萄球菌在食品上的生长数量具有良好的效果。其中针对于部分数据集的R2数值为0.95-0.98，RMSE为0.42-0.73（表6），在未参与比对（Tarlak F等人未研究）的奶酪数据集中R2达到0.99，RMSE为0.65。在Bf、Af的数值上，Bayes表现为-0.01<Bf<0.42，0.21<Af<0.68。故，本研究选择Bayes模型作为最优模型。

表 6 本研究与 Tarlak F 等人的研究的研究对比

	Tarlak F 等人的研究			本研究		
	牛肉 Beef	猪肉 Pork	培养基 culture medium	牛肉 Beef	猪肉 Pork	培养基 culture medium
数据点	282	595	4315	321	336	7207
R2	0.973	0.861	0.938	0.95	0.98	0.95
RMSE	0.326	0.968	0.6	0.73	0.42	0.63
Bf	0.6%	5.2%	1.9%	1%	5%	1%
Af	8.6%	40.8%	18.5%	51%	21%	44%

2.3 模型验证

为了进一步验证贝叶斯模型的稳定性和准确性，搜集数据集之外的独立实验数据是非常重要的。因此，本研究搜集了牛肉Beef、猪肉Pork、奶酪Cheese、培养基的独立实验数据与采用贝叶斯模型回归的预测值进行比对，结果用Af、Bf表示。并将贝叶斯外部验证模型回归效果绘制图形（图9）。

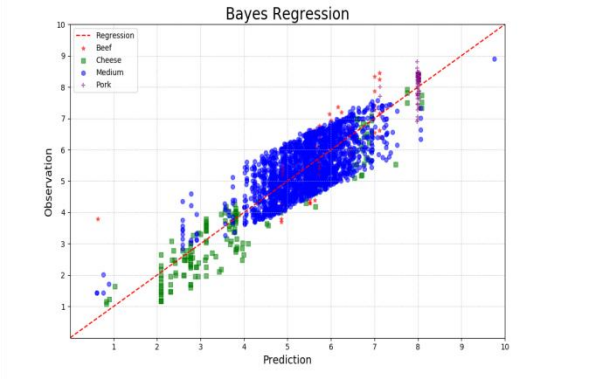


图 9 贝叶斯模型外部验证回归效果

经回归比对后，Af=0.21，Bf=-0.01。证明预测值与实际值平均偏差21%，模型平均低估1%。而回归拟合效果在图上表现较为集中密集，拟合效果较好，因此贝叶斯模型具有一定的稳定性。

3. 结论

本研究采用不同的基于机器学习的回归方法（贝叶斯模型、深度学习模型、岭回归模型、线性回归模型和支持向量机模型）来预测金黄色葡萄球菌的计数。在各种食品（牛肉、猪肉和奶酪）和培养基中。所有回归算法的性能都令人满意，但贝叶斯回归的估计能力最好。为了进一步验证其预测能力，利用文献中的外部数据对该算法进行了验证，结果依然比较稳定。尽管随机森林回归对每种食品都有良好的预测能力，但对培养基类别的最准确的估计。

产生这种现象的问题可能在于培养基类型的数据量最大，并且贝叶斯模型在Af精度系数的表现上差强人意，尽管不能以此证明其有过拟合的现象，但仍待加强。相比于传统的实验建模方法，机器学习模型在食品预测生物中具有时间成本和经济成本低的优点，因此提高机器学习模型的准确度和模型多样性是非常重要的。经本研究结果表明，贝叶斯模型能为机器学习在食品微生物预测中的应用提供新思路，并取得不错的效果。

参考文献

[1]Fung F, Wang H S, Menon S. Food safety in the 21st century[J]. Biomedical journal, 2018, 41(2): 88-95.

[2]Satcher D. Food safety: a growing global health problem[J]. Jama, 2000, 283(14): 1817-1817.

[3]Gill C J, Hamer D H. Foodborne illnesses[J]. Current treatment options in gastroenterology, 2001, 4: 23-38.

[4]Baird-Parker A C. Foods and microbiological risks[J]. Microbiology, 1994, 140(4): 687-695.

[5]Kallen A J, Hageman J, Gorwitz R, et al. Characteristics of Staphylococcus aureus community-acquired pneumonia during the 2006–2007 influenza season[J]. Clinical infectious diseases, 2007, 45(12): 1655-1655.

本研究受到重庆第二师范学院大学生科研立项项目资助，项目编号：KY20240041