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

陈晨[1]，尹佳[2]，董曼[2]，穆书敏[1]，陈锂[2]，郭鹏程[2]，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）

摘要：乳制品是人们日常生活中一种重要的营养食品，为了提高对乳制品质量安全风险预测的准确性，保障乳制品质量安全，本文基于检测产品和检验数据的随机性、模糊性以及信息不完全性，将所得不同地区的乳制品检测数据通过改进的 softmax 公式进行等级划分，并按自然日进行分箱处理，通过风险权重等比例映射法得到风险等级，充分利用了乳制品灰色数据，对检验合格数据中的潜在风险进行挖掘。采用小波分解（Wavelet Decomposition，WD）和长短期记忆神经网络（Long Short-Term Memory，LSTM）模型结合的方式，对不同地区的乳制品检测数据进行风险预测。结果表明，该组合模型的平均准确率达 97.54%，标准偏差为 0.03，与经验模态分解（Empirical Mode Decomposition，EMD）-LSTM 模型和有选择性重构且间隔为 2 的 WD-LSTM 模型相比准确率更高，稳定性更好，可实现对乳制品质量风险的预测和防控，能为乳制品的风险监管提供有利参考和技术支撑。

关键词：乳制品；风险预测；风险等级划分；小波分解；长短期记忆神经网络

文章编号：1673-9078(2023)01-300-310 DOI: 10.13982/j.mfst.1673-9078.2023.1.0291

Predictive Early Warning Analysis of Dairy Product Quality and Safety Risks Based on Grey Data Pre-processing using the WD-LSTM Model

CHEN Chen[1], YIN Jia[2], DONG Man[2], MU Shumin[1], CHEN Li[2], GUO Pengcheng[2], WEN Hong[2*], GUI Yufeng[1*]

(1.School of Science, Wuhan University of Technology, Wuhan 430070, China)

(2.Hubei Provincial Engineering and Technology Research Center for Food Quality and Safety Test, 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, Wuhan 430075, China)

Abstract: Dairy products are important nutritious food items in people's daily life. To improve the analysis of dairy product quality and safety risk prediction, the randomness, fuzziness, and incomplete information of detection products and inspection data obtained during dairy product evaluation in different regions were divided into grades using an improved softmax formula and processed in boxes according to the natural day. The risk grades were determined from the risk weight and other proportional mapping methods, and the dairy gray data were used to determine the potential risks in the qualified inspection data. Wavelet decomposition combined with the long short-term memory (LSTM) model was used to predict the risk of dairy product detection data in different regions. The results showed that the combined model exhibited an average accuracy of 97.54% and a standard deviation of 0.03, indicating higher accuracy and better stability compared to those of the empirical

引文格式：

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 Chen, YIN Jia, DONG Man, et al. Predictive early warning analysis of dairy product quality and safety risks based on grey data pre-processing using the WD-LSTM model [J]. Modern Food Science and Technology, 2023, 39(1): 300-310

收稿日期：2022-03-16

基金项目：国家重点研发计划项目（2018YFC1603602）

作者简介：陈晨（1998-），女，硕士研究生，研究方向：食品安全风险预测，E-mail：chen0626chen@163.com

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，教授，研究方向：统计分析与建模、数据挖掘，E-mail：guiyufeng@whut.edu.cn

mode decomposition-LSTM model and wavelet decomposition-LSTM model with selective reconstruction and an interval of 2. Thus, risks associated with dairy product can be predicted and prevented. These results provide a reference and technical support for supervising risks associated with dairy products.

Key words: dairy product; risk prediction; risk classification; wavelet decomposition; long and short-term memory neural network

乳制品富含营养物质，可促进机体营养均衡、调节人体免疫机能。在疫情爆发初期，国家卫健委发布的《新型冠状病毒感染的肺炎防治营养膳食指导》[1]指出，科学的营养膳食和每日合理的乳制品摄入是提高机体抵抗力、预防与救治新冠肺炎的有效途径。我国人均乳制品消费呈上升趋势，在行业迅速发展的同时，还存在部分企业重产量而忽视质量管控的现象，如何加强对乳制品质量安全风险的识别，提高生产企业对质量安全的控制能力，已成为保障我国乳制品行业健康发展迫切需要解决的问题。因此，对问题产品或可能存在的风险发出及时预警，实现乳制品综合性、动态性的监管和控制，提供靶向性监管技术支持是非常有必要的[2]。

当前，专家学者们针对乳制品质量安全风险预警从不同方向开展了有关研究。如 Tian 等[3]基于主成分分析对生乳质量安全指标体系风险进行了评估；Zhang 等[4]构建了乳制品质量安全追溯系统，使供应环节可追溯；部分学者通过乳品供应链环节构建了乳制品质量风险预警指标体系[5,6]；陈嘉惠等[7]从三个层面分别对乳制品中的危害因素进行风险评估。此外，也有学者重点研究预警方法，将机器学习引入到食品风险的预测中，结合深度径向基函数[8]、集成极限学习机[9]、层次分析法[10]、BP 神经网络[11]、LSTM 模型[12]等新型预警方法，对乳制品进行深度层次预警建模，在一定程度上实现了对乳制品安全风险预警的预测和防控。

上述研究成果为我国乳制品质量安全预警的实践提供了良好的理论基础和方法依据。但目前针对海量抽检数据的风险预警研究还鲜有涉及，主要利用传统的数理统计、典型病例通报等手段，对历史抽检数据进行食品安全状况的评价和风险警示，该方法是对食品安全状况的事后分析，缺少深度的分析与应用[13-15]。我国已积累海量的乳制品检测数据，乳制品按照分类不同和每年食品安全状况的调整，检测项目存在差异，且并非每天都进行抽样检测，同时数据中存在缺失检测结果的大量空值。现有的乳制品检测数据中包含众多灰色数据[16,17]，这种情况下，对数据进行预处理，从风险因素中挖掘分析，提炼出有价值的信息尤为重要。

因此本文利用我国乳制品历史抽检信息为数据源，依据国家标准对检测结果中的灰色数据进行去量纲化处理，采用 softmax、数据分箱等方法进行数据预处理，通过小波对数据进行分解，对分解后不同细节的分量采用 LSTM 模型进行预测，并通过 symmetric 模式重构，输出最终的预测风险等级。通过测试集对本文构建的 WD-LSTM 组合模型预测准确度进行验证，该模型与同类模型相比有明显提高，可以为我国乳制品食品质量安全风险预警提供有力支持和参考。

1 材料与方法

1.1 实验材料

1.1.1 数据类型

本文选取 2015-2020 年对外公开以及检测机构内部自行检测获得的 543 336 条乳制品检测信息作为数据源，对原始数据进行分析可得，不同产品类别的检测信息存在差异，不同年份的检测信息也存在差异，为了更加全面的得到乳制品存在的风险预警，将所有项目都考虑在内，建立了乳制品风险预警的检验项目指标体系。指标体系共包括 12 个项目类别，76 个检验项目，见表 1。

由于获取的乳制品类别、年份以及检测项目的结果单位不同，存在数据属性类别多且格式杂乱，检验结果中信息不完全、不充分以及数据的多样化问题[18]，使其无法按照统一的规则转换为风险等级。此类灰色数据的高复杂度特点也提高了风险分析的难度，若直接将原始数据划分训练集和测试集，带入模型训练，所得到的结果可能存在较大的误差，因此需要对检测数据进行分类、去量纲化、数据分级等预处理。部分乳制品检测信息如表 2 所示。

1.1.2 灰色数据预处理

对于上述缺省数据多且容易受到多种噪声污染的灰色数据，通常需要进行数据清洗、集成、变换等预处理。数据清洗主要是按照一定的规则和标准对存在缺失、奇异值和离群点等问题的数据剔除；数据集成则是将混杂的数据按照一定的特征相互匹配，以提高数据的统一性；数据变换是将原始数据转换为满足一定的条件数据，主要包括运用分箱、聚类等进行数据光滑、将数据集中汇总进行数据聚集、使用高级概念代替低级概念的数据概化、将原始数据按特征缩放规范、构造新的特征并汇合到原本特征集中[19]。

表 1 乳制品风险预警的检验项目指标体系

Table 1 Index system of inspection items for risk warning of dairy products

序号	项目类别	检验项目
1	非食用物质	三聚氰胺、玉米赤霉醇、硫氰酸钠、舒巴坦、β-内酰胺酶、羟脯氨酸、
2	禁用兽药	诺氟沙星、氧氟沙星、氯霉素、氨苄青霉素
3	农药残留	狄氏剂
4	其他微生物	大肠菌群、酵母、菌落总数、商业无菌、霉菌计数
5	其他污染物	亚硝酸盐、高氯酸盐、氯酸盐
6	生物霉素	黄曲霉毒素 M1
7	食品添加剂	N-[N-(3,3-二甲基丁基)]-L-α-天门冬氨-L-苯丙氨酸 1-甲酯、山梨酸及其钾盐（以山梨酸计）、糖精钠（以糖精计）、维生素 E、苯甲酸及其钠盐（以苯甲酸计）、胭脂红、苋菜红、β-胡萝卜素、日落黄、柠檬黄、诱惑红、亮蓝、三氯蔗糖、纳他霉素、甜蜜素、阿斯巴甜、乙酰磺胺酸钾
8	重金属等元素污染物	总砷（以 As 计）、铅（以 Pb 计）、铬（以 Cr 计）、镉（以 Cd 计）、总汞
9	兽药残留	苄青霉素、氨苄青霉素、邻氯青霉素、地塞米松、达氟沙星、恩诺沙星、磺胺类、金霉素、沙拉沙星、四环素、土霉素
10	有机污染物	氨基甲酸乙酯、多氯联苯、壬基酚
11	质量指标	乳酸菌数、蛋白质、非脂乳固体、复原乳酸度、钙
12	致病性微生物	单核细胞增生李斯特氏菌、沙门氏菌、金黄色葡萄球菌、蜡样芽孢杆菌

表 2 部分乳制品检测信息

Table 2 Partial detection information of dairy products

序号	检验项目	检验结果	最小允许限	最大允许限	单位	项目分类
1	脂肪	3.05	3.1	/	g/100 g	质量指标
2	黄曲霉毒素 M1	未检出	/	0.5	μg/kg	生物毒素
3	酸度	13.5	12(牛乳)/6(羊乳)	18(牛乳)/13(羊乳)	°T	质量指标
4	金黄色葡萄球菌	0,0,0,0,0	-	*n*=5, c=0, m=0	/25 g	致病性微生物
5	苯甲酸	0.0395	/	/	g/kg	食品添加剂不符合规定
6	大肠菌群	(1)890;(2)1 200; (3)1 400; (4)1 200;(5)1 200	/	*n*=5, c=2, m=1, M=5	CFU/g	其他微生物
7	糖精钠(以糖精计)	未检出	/	不得使用	g/kg	食品添加剂不符合规定
8	地塞米松	<0.2(定量限 0.2 μg/L)	/	0.3	μg/kg	兽药残留
9	商业无菌	商业无菌	-	商业无菌	-	其他微生物
10	β-内酰胺酶	阴性	/	/	/	非食用物质

1.1.2.1 数据去量纲化处理

根据检测结果结合国家标准进行去量纲化处理。对于有最大允许限的项目 X_i 和有最小允许限的项目 Y_i，分别使用公式 1、2 对其进行标准化和去量纲化。

$$X_i=\begin{cases}\dfrac{x_i}{x_{standard}}, x_i \text{和} x_{standard} \text{均为单个数值时}\\ 1, x_{standard} \text{为不得检出而} x_i>0\end{cases} \quad (1)$$

$$Y_i=\frac{y_i}{y_{standard}} \quad (2)$$

式中:

X_i 和 Y_i——预处理后的检验数值;

$x_{standard}$ 和 $y_{standard}$——标准允许限的值;

x_i 和 y_i——标准化数值。

1.1.2.2 数据分级处理

将去量纲化后的数据，根据检验项目类别的不同，将检验项目划分为四部分，分别是有最大允许限的项目 X_i，有最小允许限的项目 Y_i，有限定范围允许限的项目 R_i 和检验结果为 5 个数值的项目 Z_i。该风险等级划分难以采用技术方法进行定量分析，故采用专家打分法进行风险等级的划分，邀请十位专家通过无记名投票的方法，得到专家确定的等级，使用加权评价法得到最终的评价结果，进行评判。结合检验项目风险等级划分标准和专家打分法将乳制品检验项目划分为 5 个风险等级，1 级为安全无风险，2 级为轻微风险，3 级为轻度风险，4 级为中度风险，5 级为不合格产品。

其中 1~4 级风险是符合国家标准的，但风险系数不同，而 5 级为不符合国家标准。具体划分标准见表 3。

经过初步的数据预处理，去掉因条件缺失无法判别的数据后，共518 640条乳制品项目风险等级数据，其中 1 级 499 371 条，2 级 14 054 条，3 级 3 993 条，4 级 1 008 条，5 级 214 条。分析 2015~2020 年抽检数据，前 5 年数据的检测项目基本一致，2020 年根据以往的检测结果，对风险较大和较少发现问题的项目进行了增减，致使 2020 年食品数据检测项目与前 5 年不一致，同时乳制品又分亚类、次亚类、细类，即使细类也包括了不同产品标准，其要求的项目也不同，最终造成即使同一细类产品中也存在项目不同的问题，使得用于分析的数据存在同类产品中项目缺失、同一标准产品中不同年度项目缺失问题。

表 3 检验项目的风险等级划分标准

Table 3 Risk classification standard of inspection items

序号	功能名称	描述（5 级最严重）
1	X_i	1 级：0≤标准化数值<0.10 MRL; 2 级：0.10 MRL≤标准化数值<0.30 MRL; 3 级：0.30MRL≤标准化数值<0.70 MRL; 4 级：0.70 MRL≤标准化数值<1.00 MRL; 5 级：标准化数值≥1.00 MRL. 限量值为不得检出的项目，有检出即按最高等级计算，未检出即按 1 级计算
2	Y_i	1 级：标准化数值≥1.00 MRL; 5 级：标准化数值<1.00 MRL.
3	R_i	1 级：下限≤检测值≤上限; 5 级：检测值<下限；检测值>上限.
4	Z_i	菌落总数 大肠菌群 金黄色葡萄球菌 1级：c=0或全部未检出 2级：检测值 ≤ m 3级：c=1, m<检测值 ≤ M 4级：c=2, m<检测值 ≤ M 5级：c＞2且m<检测值 ≤ M 或c ≤ 2且检测值＞M，即不合格项 沙门氏菌 1级：未检出 5级：检出即不合格

注：MRL 为标准允许限量，c 为最大可允许超出 m 值的样品数，m 为致病菌指标可接受水平限量值，M 为致病菌指标的最高安全限量值。

针对此类处理后的灰色异构数据（区间灰数、离散灰数等），不同产品因所属食品类别不同而导致检验项目存在差异，故仅对有检测结果的项目风险赋予权重，对缺失项目予以忽略。由于低风险等级的数据占绝大多数，若直接采用简单的加权平均来获得最终的产品风险等级，会导致整体风险等级偏低，不能反应真实的风险。在食品安全风险等级预警中，风险等级高的数据对最终的风险等级影响更大，故应该有更大的权重，风险等级低的数据权重应该较低，且如果在某一产品中存在一个不合格项目，则该产品综合风险等级应直接划分为 5 级。为体现权重的变化，采用改进的 softmax 函数来计算产品的综合风险等级（公式 3），通过 softmax 函数中指数权重的变化来调节风险等级的权重。

$$level=\begin{cases}5, \text{该产品中有不合格的检测项目}\\ \operatorname{argmax}\left(\omega_i * e^{i}\right)+1\end{cases} \quad (3)$$

式中：

Level——该产品的综合风险等级;

I——该检测项目的风险等级;

ω_i——该风险等级在该产品中的占比。

1.1.2.3 数据分箱

乳制品检测数据的样品生产日期存在不连续，同一天生产日期样品数量也不相同，因此从时序序列考虑，数据存在不均匀分布，存在缺失和稠密性差异，需要对经过预处理的检测数据进行分箱处理后再带入模型进行预测研究。数据分箱即是将一定时间段的数据划分为一个数据集，并对分箱数据选择合适的方法处理，得到各分箱数据集的综合等级。本文采用每个

自然日作为一个分箱，忽略缺失日期数据后进行时间压缩，并通过风险权重等比例映射的方法计算各分箱数据的综合等级。

1.2 风险预测方法与模型

1.2.1 小波分解（Wavelet Decomposition，WD）

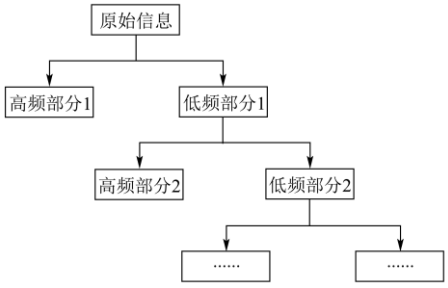


图 1 小波分解示意图

Fig.1 Wavelet decomposition diagram

小波分解是一种信号时频分析方法。它将一个波形分解成 N 个低频部分和 M 个高频部分的和，只针对信号的低频部分，即信号的趋势部分做进一步分解，而对于高频部分，也就是信号的噪声部分，不再继续分解，低频部分能够反映原始数据在平稳条件下本身的变化规律，高频部分包含原始数据的波动性和非线性等细节，所以小波变换可以对以低频信息为主要成分的信号做很好的表征[20]。图 1 为小波分解的示意图。

由于食品类检测数据的随机性和不确定性，所得到的乳制品风险等级是一个非平稳的离散时间序列，若直接使用 LSTM 模型对该数据进行预测，其噪声会导致学习曲线复杂，且预测精度受到影响。经典的傅里叶变换（公式 4）尽管能对信号的整体内涵进行反映，但噪声会使其频谱复杂化；短时傅里叶变换可以部分定位时间，但由于窗口的大小是固定的，故仅对频率波动小的平稳信号适用。小波变换既保留了局部变换的思想，又将无限长的三角函数基换成了有限长的会衰减的小波基（公式 5），能从不同尺度上对信号进行分解，按照频率自动调整窗口大小，提取非平稳信号的局部特征，是一种可以进行多分辨率分析的自适应时频分析方法[21]。

傅里叶变换公式：

$$F(w)=\int_{-\infty}^{\infty} f(t)*e^{-iwt}dt \tag{4}$$

小波变换公式：

$$\begin{cases} WT(a,\tau)=\dfrac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(t)*\psi\left(\dfrac{t-\tau}{a}\right)dt \\ \psi(t)=\dfrac{1}{\sqrt{a}}\psi\left(\dfrac{t-\tau}{a}\right) \end{cases} \tag{5}$$

式中：

t——时刻；

w——频率；

α——尺度因子，控制小波函数的伸缩；

τ——平移因子，控制小波函数的平移。

小波变换是对原始信号和小波基函数以及尺度函数做内积运算，因此一个小波基和一个尺度函数就能够确定一个小波变换。小波分解中使用到的小波函数具有多样性，同一个小波基函数可以通过平移和缩放生成不同的小波基，故对同样的问题，不同的小波基会产生不一样的结果。

根据本文所用数据波动性大，在时间上具有连续性的特征，选择小波分解中的一维多阶次离散小波分解，即 WaveDec 算法，该算法是采用离散小波变换（Discrete Wavelet Transformation，DWT）得到原始信号的低频部分和高频部分，再将经过 DWT 变换后的低频成分再进行 DWT 变换，循环次数由分解层数决定。常用的小波族有很多种，每个小波族又有多种系数可供选择，其中 Daubechies 小波函数由法国著名的小波分析学者 Inrid Đaubechies 提出，简称为 dbN，其中 N 代表小波的阶数[22]。dbN 是非线性相位，没有固定的核函数，通常情况下，Daubechies 族中消失矩的阶数越大，小波越光滑。结合数据特征选择了光滑性比较好的 db8 作为小波函数[23]，按照输入序列的复杂情况分解为频率不同的子序列，各个子序列包含原序列中不同频率的信息，且其长度不发生改变，提取小波分解系数对其进行分析，各子序列带入模型得到预测结果后再通过 symmetric 模式进行重构。

1.2.2 长短期记忆神经网络模型（Long Short-Term Memory，LSTM）

LSTM 是基于传统循环神经网络 RNN 的一种改进，不仅能学习时间规律，还可以适应非线性的复杂数据。LSTM 在 RNN 的基础上新增了一个间隔多个时间步长来传递信息的被称为“门”的内部机制，可以调节信息流，循环结构之间保持一个持久的单元状态不断传递下去[24]。“门”结构中包括激活函数 sigmoid，与 tanh 函数将值压缩到-1~1 之间不同，sigmoid 函数会把值压缩至 0~1，更加有利于“门”对信息的保存或遗忘。

1.2.3 WD-LSTM 组合模型

本研究在预测乳制品风险等级时，使用的是 WD-LSTM 组合模型，具体流程见图 2。该模型在单个 LSTM 模型的基础上，增设能够适应非平稳信号的小波分解，非线性、非平稳且波动性强的原始序列通过小波分解得到各分量，再将各分量分别代入 LSTM 模型，模型根据输入序列计算其对后面的综合风险等级的影响，同时考虑到后面的综合风险等级对该序列

的影响，前后影响值的大小决定了保留或遗忘多大程度，并且通过单元状态实时更新到下一步的预测。各分量预测结果经过 symmetric 模式重构，得到最终的预测结果。

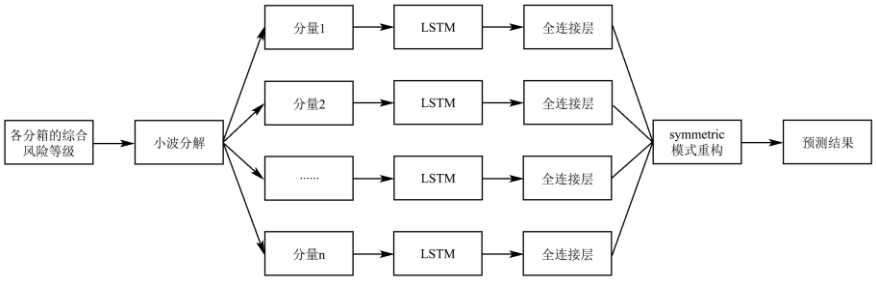


图 2 WD-LSTM 组合模型流程图

Fig.2 Flow chart of WD-LSTM combined model

1.2.4 模型参数的搭建和设置

为实现 LSTM 神经网络的双向构造，方便模型训练，需预先确定网络结构。本文构建的是一个 4 层神经网络，将待预测的前 20 个乳制品综合风险等级作为神经网络的输入，即输入层的神经元个数为 20；待预测的乳制品综合风险等级作为网络的输出，即输出层的神经元个数为 1；中间设置了一个 LSTM 层和一个全连接层作为两个隐藏层，其中全连接层在整个网络卷积神经网络中起到“特征提取器”的作用，结点数设定为 16。依据本文所用的数据集和实际目标需求，确定相关参数的调整方向，采用能更好反映预测值误差的实际情况的平均绝对误差（Mean Absolute Error，MAE）作为损失函数，优化器使用能基于训练数据迭代地更新神经网络权重的 Adam 优化算子，数据集按照 2:1 的比例划分为训练集和测试集，一次训练所选取的样本数为 64，训练轮次定为 100。

1.2.5 经验模态分解（Empirical Mode Decomposition，EMD）

经验模态分解可以对非线性非平稳信号的进行分析处理，能依赖信号本身的特征做自适应分解，无需事先设定基函数，也克服了基函数存在的无自适应性问题；分解后得到的各层信号分量，即为一系列的固有模态函数（Intrinsic Mode Functions，IMF），任何信号都可以被分解成若干个 IMF 之和，各分量分别代表原始信号中各频率分量，按照由高到低的频率顺序依次排列，可以反映原始信号的局部特征[25]。

1.2.6 数据分析

本文使用编程语言 Python 3.7.0，利用 Tensorflow 作为搭建平台。采用改进的 softmax 和数据映射方式对灰色数据进行预处理，将分箱数据集的综合等级时间序列输入到建立的 WD-LSTM 组合模型，进行风险预测预警分析，通过 matplotlib 画图软件包绘制预测各级分量和风险预测示意图，预测准确率作为评估模型优劣的指标。

2 结果与讨论

2.1 灰色数据分箱及等级划分

2.1.1 分箱时间间隔的选择

分箱处理的时间间隔会直接影响数据集个数，从而影响预测结果的准确性，因此，选择合适的时间间隔至关重要。本文分别采用了 1、4、7、15、30 d 为一个数据集进行分箱处理，计算综合等级。经过对比，若采用 7 d 及 7 d 以内进行分箱，间隔较短会导致缺失值过多，需要插值的数据过多而影响真实性，且使学习曲线更加复杂；而采用太长的时间间隔，则会导致数据集过小，导致模型学习过程太短，预测误差变大。结合实际情况和模型的预测效果，最终选择采用每个自然日作为一个分箱，对缺失数据的日期予以跳过处理。

2.1.2 分箱数据综合等级划分

试验中分别采用 5 种不同的综合风险等级公式，对数据分箱计算风险等级。

$$Y_1 = argmax(\omega(i) * e^{i}) + 1 \tag{6}$$

$$Y_2 = argmax(\omega(i) * e^{0.5i}) + 1 \tag{7}$$

$$Y_3 = argmax(\omega(i) * e^{\sqrt{i}}) + 1 \tag{8}$$

$$Y_4 = \omega(i) * e^{i} \tag{9}$$

$$Y_5 = \sum_{i=1}^{5} \omega_i * e^{i} \tag{10}$$

$$level = \begin{cases} 1, Y_5 \le 2.72 \\ 2, 2.72 \le Y_5 < 8 \\ 3, 8 \le Y_5 < 12 \\ 4, 12 \le Y_5 < 40 \\ 5, Y_5 \ge 40 \end{cases} \tag{11}$$

式中:

i——风险等级;

$\omega(i)$——风险等级 i 的占比。

表 4 部分数据集不同风险等级公式对比

Table 4 Comparison of formulas for different risk levels of some dataset

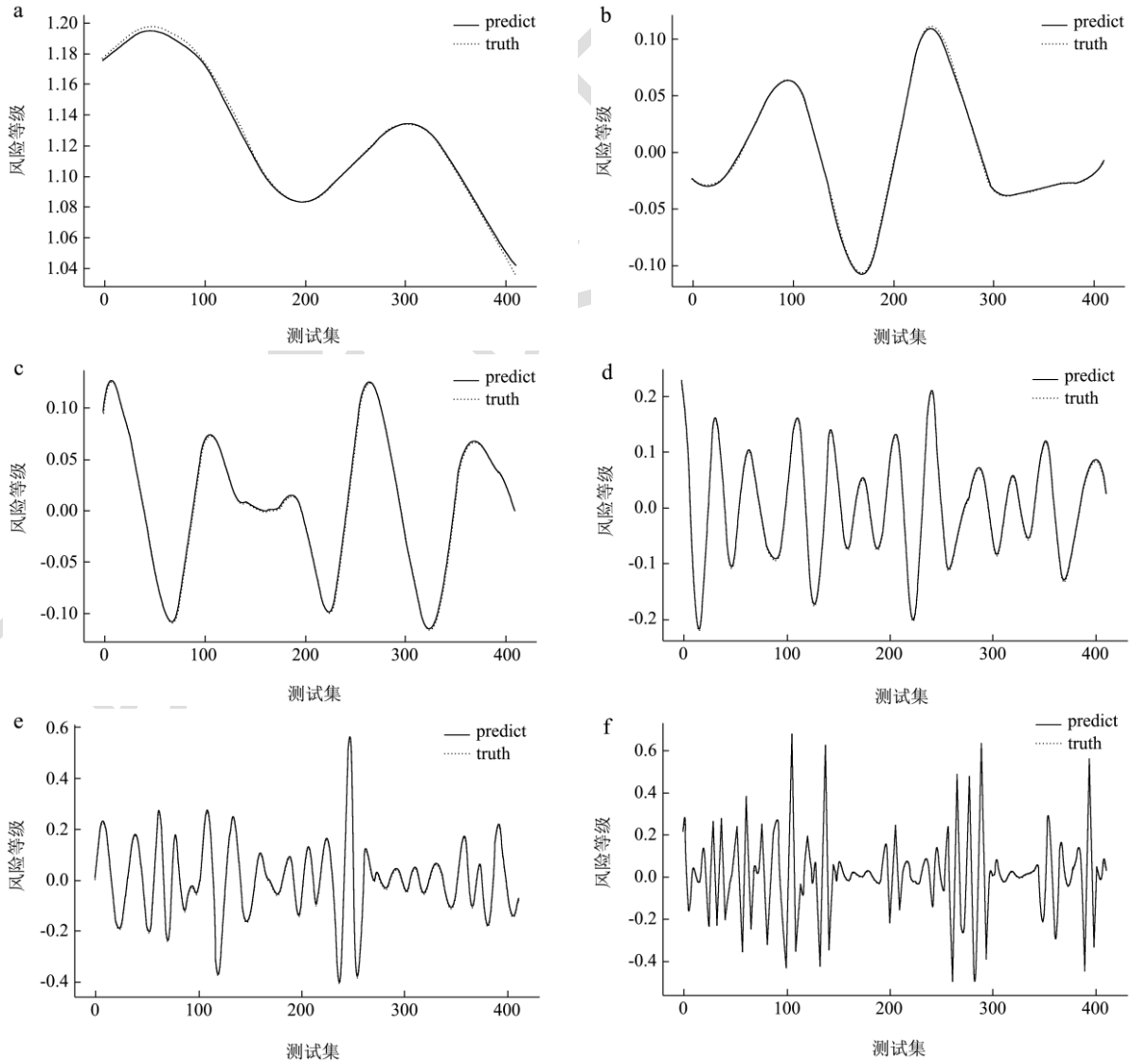
数据集	Y_1	Y_2	Y_3	Y_4	Y_5
[1, 1, 1, 1, 4]	4	4	1	3	4
[1, 1, 4]	4	4	4	3	4
[1, 4]	4	4	4	3	4
[1, 1, 1, 1, 3, 1, 1]	3	1	1	2	2
[1, 4, 1, 1, 1, 1]	4	1	1	3	3
[1, 1, 5, 1, 1]	5	5	1	3	4
[4, 1, 1, 1, 1, 1, 1, 1, 1]	4	1	1	3	3
[1, 1, 1, 1, 1, 1, 1, 1, 5, 1]	5	1	1	3	4

Y_4 和 Y_5 当数据集中只含有一种风险等级时，该风险等级即为该分箱的综合风险等级，数据集中的风险等级不唯一时，Y_4 通过公式 9 计算每个风险等级的权重，对权重最大的两个风险等级求平均值，若平均值为小数则采用向上取整；Y_5 通过公式 10 计算出该数据集的综合风险，根据产品原始的各风险等级占比，使用风险权重等比例映射的方法，按照相应的比例使用公式 11 对综合风险进行划分。部分数据集不同计算公式的风险等级对比见表 4。

经过对比，认为公式 6 会导致对风险等级高的产品赋予过大的权重；公式 7 和公式 8 对公式 6 的指数进行了调节，但导致高风险等级权重过小，难以确定合适的权重；Y_4 采用了平均法，无法体现对风险等级的侧重；通过得到的风险等级与原始数据的风险程度比较，公式 10 更符合实际风险的划分。因此，本文采用公式 10 结合公式 11 计算风险等级，共得到 14 037 条综合风险等级，其中 1 级 13 171 条，2 级 49 条，3 级 151 条，4 级 542 条，5 级 124 条。

2.2 结果分析

2.2.1 模型训练



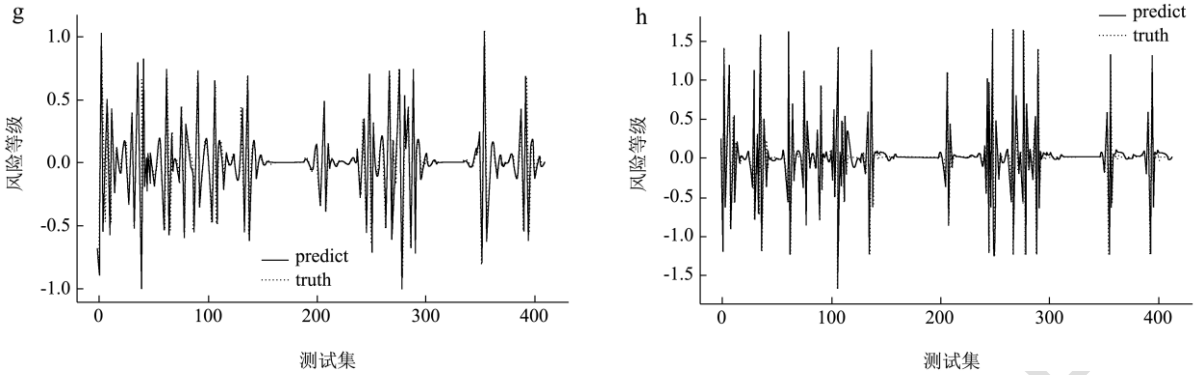


图 3 N1 地区乳制品数据 WD 各级分量预测示意图

Fig.3 Schematic diagram of WD components of dairy products data in N1

注：a~h 依次为 N1 地区 1~8 级分量的预测结果。

在本文中，将数据分箱后的综合风险等级输入到建立的组合模型，其中前 2/3 作为训练集，1/3 作为测试集，对其进行小波分解，再通过长短期记忆神经网络对小波分解得到的各个分量进行预测，将各分量重构后输出最终的预测结果。其中测试集用来验证该模型的精确度。图 3 为 N1 地区的乳制品数据经小波分解后各级分量预测示意图。橙线为各分量的真实值，蓝线为各分量的预测值。

2.2.2 有效性分析

由于本文构建的 LSTM 模型初始权重的随机性，在每轮预测时可能会存在误差，为验证该模型的稳定性，连续将该模型运行 5 次，得到该模型的平均误差为 0.03，波动较小，因此该模型的运行结果是可靠的。为了全面验证模型的有效性和适用性，将 29 个地区的风险等级序列经小波分解后带入 LSTM 模型进行预测，采用平均绝对值误差（Mean Absolute Error，MAE）和平均绝对百分比误差（Mean Absolute Percent Error，MAPE）衡量该模型的误差（公式 12、13），该值越大表明误差越大，当预测值与真实值完全吻合时等于 0。该模型在 29 个地区中预测的最大 MAE 为 0.07，最大 MAPE 为 2.71%，整体 MAE 和 MAPE 的平均值为 0.02 和 0.83%。通过公式 14，可以计算出该模型预测的准确率，准确率最低为 86.49%，其余均在 92.45%以上，整体平均准确率为 97.54%，标准偏差为 0.03。该结果表明，本文建立的 WD-LSTM 模型可以对乳制品质量安全风险等级有较好的预测。29 个地区的预测结果见表 5。

表 5 29 个地区乳制品风险等级预测结果

Table 5 Prediction results of risk grade of dairy products in 29 regions

序号	城市名	样品数/个	预测准确率/%	MAE	MAPE/%	序号	城市名	样品数/个	预测准确率/%	MAE	MAPE/%
1	J1	103	99.03	0.07	2.22	16	N1	1289	99.03	0.01	0.28
2	S1	908	98.94	0.01	0.29	17	H3	858	98.13	0.01	0.44
3	A	523	100.00	0.00	0.00	18	Z	476	97.86	0.01	0.60
4	H1	954	97.66	0.00	0.11	19	H4	352	100.00	0.00	0.00
5	B	388	98.18	0.03	1.21	20	Y	374	92.45	0.03	1.10
6	S2	369	98.08	0.02	0.80	21	G3	395	95.80	0.02	0.74
7	H2	413	95.80	0.03	1.26	22	F	168	100.00	0.00	0.00
8	J2	619	95.72	0.01	0.53	23	S4	332	97.83	0.03	1.27
9	T	535	98.11	0.02	0.94	24	G4	167	86.49	0.03	1.35
10	C	139	100.00	0.00	0.00	25	X	669	99.51	0.01	0.37
11	H5	287	98.70	0.01	0.43	26	G1	486	99.30	0.01	0.58
12	S5	469	97.08	0.03	1.28	27	S3	684	98.56	0.01	0.52
13	L	579	95.40	0.05	2.30	28	N2	475	97.12	0.04	1.62
14	J3	174	100.00	0.00	0.00	29	Q	316	96.51	0.02	1.16
15	G2	536	97.50	0.06	2.71						

MAE 的计算公式：

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \quad (12)$$

MAPE 的计算公式：

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i-\hat{y}_i\right|}{y_i} \quad (13)$$

$$A=\frac{B}{C} \quad (14)$$

式中:

A——预测准确率;

B——预测正确的样本数量;

C——测试集的样本数量。

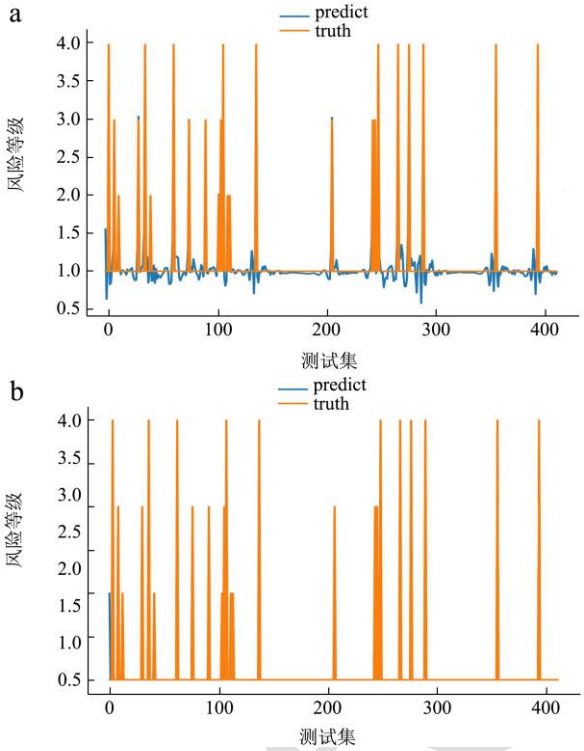


图 4 N1 地区乳制品风险预测示意图

Fig.4 Schematic diagram of risk prediction of dairy products in N1

以 N1 地区乳制品质量安全预测结果为例，图 4 中，橙线为分箱数据集的综合风险等级，蓝线为 WD-LSTM 模型得到的预测风险等级（图 a 为预测风险等级，图 b 为取整后的预测风险等级）。由图中两种颜色的线段重合度可以看出，二者吻合度较高，说明该模型预测的准确性较好。

2.2.3 模型比较与分析

本次研究中，还分别构建了 EMD-LSTM 模型和有选择性重构且间隔为 2 的 WD-LSTM 模型，通过对数据采用不同的分解方法和选取不同的间隔来验证本文所使用的WD-LSTM模型在乳制品灰色数据上的拟合效果，表 6 为不同模型的预测准确率对比。

模型 1 是 EMD-LSTM 组合模型。对 29 个地区的 2015~2020 年乳制品检测数据做同样的预处理后，模型 1 将分箱数据带入 EMD 模型进行分解，将得到的各分量 IMFs 输入 LSTM 模型，预测结果表明，准确率最低仅为 29.73%，整体准确率仅为 86.97%，标准偏差为 0.14。总体上看，模型 1 的准确率与小波分解-LSTM 模型相比明显降低，且预测结果差距较大，不够稳定。平均 MAE 和 MAPE 分别为 0.27 和 12.95%，且最大 MAE 和 MAPE 为 1.95 和 54.91%，均明显高于WD-LSTM模型。由于EMD的模态混叠现象严重，会导致特征提取、模型训练、模式识别变得困难，IMF 的特征不再是单一尺度[26]。因此，经过 EMD-LSTM 模型分解后得到的各个分量IMFs波动仍然较为强烈，预测误差变大，从而导致重构后的模型预测误差较大。而小波变换频带是固定的，在带入模型预测前采用了具有更好的光滑性的 db8 小波基，有效的减小了各分量变化趋势的复杂性，分解后得到更光滑的的各分量也使得 LSTM 模型预测的准确度更高[27]。

模型 2 与本文建立的 WD-LSTM 模型类似，也是一个小波分解后将各分量代入LSTM预测的组合模型，对各分量有选择性的进行重构，重构后的序列再通过 LSTM 模型进行预测。在本文中有选择性重构所选择的间隔为 2，以验证间隔大小对该模型产生的影响。该模型与对原始序列进行平滑处理类似，会对部分细节信息有所损失，预测精度也有所降低，J1 地区的准确率仅为 66.67%，整体准确率为 92.42%，标准偏差为 0.07，平均 MAE 和 MAPE 分别为 0.09 和 4.83%。故对比表 6，在整体预测精度和误差上，本文所用的 WD-LSTM 模型均优于模型 1 和 2。

表 6 不同预测模型的准确率对比

Table 6 Comparison of accuracy of different prediction models

序号	城市名	WD-LSTM 模型	模型 1 EMD-LSTM 模型			模型 2 选择间隔为 2 的分量重构		
		准确率/%	准确率/%	MAE	MAPE/%	准确率/%	MAE	MAPE/%
1	J1	99.03	80.00	0.27	23.33	66.67	0.13	6.67
2	S1	98.94	86.97	0.20	12.21	89.08	0.13	6.87
3	A	100.00	100.00	0.00	0.00	100.00	0.00	0.00

续表 6

序号	城市名	WD-LSTM 模型	模型 1 EMD-LSTM 模型			模型 2 选择间隔为 2 的分量重构		
		准确率/%	准确率/%	MAE	MAPE/%	准确率/%	MAE	MAPE/%
4	H1	97.66	95.99	0.07	5.24	95.65	0.05	2.84
5	B	98.18	98.18	0.05	5.45	96.36	0.08	5.45
6	S2	98.08	75.96	0.45	21.96	95.19	0.08	3.53
7	H2	95.80	84.03	0.34	16.81	90.76	0.11	5.32
8	J2	95.72	71.12	0.47	24.67	86.63	0.17	9.80
9	T	98.11	79.25	0.17	7.65	98.74	0.02	1.57
10	C	100.00	100.00	0.00	0.00	100.00	0.00	0.00
11	H5	98.70	93.51	0.23	9.31	96.10	0.04	1.95
12	S5	97.08	94.16	0.09	8.03	91.97	0.09	4.74
13	L	95.40	79.31	0.43	24.33	89.08	0.14	6.99
14	J3	100.00	100.00	0.00	0.00	100.00	0.00	0.00
15	G2	97.50	91.25	0.09	5.83	92.50	0.09	5.00
16	N1	99.03	88.32	0.15	12.00	88.56	0.15	7.79
17	H3	98.13	90.64	0.15	7.47	94.38	0.07	4.24
18	Z	97.86	91.43	0.12	5.83	97.86	0.04	3.21
19	H4	100.00	100.00	0.00	0.00	100.00	0.00	0.00
20	Y	92.45	86.79	0.31	17.20	83.96	0.17	7.70
21	G3	95.80	73.45	0.17	8.11	87.61	0.08	3.98
22	F	100.00	100.00	0.00	0.00	100.00	0.00	0.00
23	S4	97.83	92.39	0.82	33.70	86.96	0.16	8.51
24	G4	86.49	29.73	1.95	54.91	86.49	0.30	18.92
25	X	99.51	95.59	0.09	4.98	93.14	0.08	4.17
26	G1	99.30	95.80	0.16	8.93	93.71	0.07	3.50
27	S3	98.56	95.22	0.11	7.50	92.82	0.10	4.86
28	N2	97.12	76.26	0.42	21.16	90.65	0.12	6.12
29	Q	96.51	76.74	0.51	29.07	95.35	0.13	6.40

3 结论

针对目前备受关注的乳制品质量安全问题，本文对近六年具有“贫信息”且类型多样性的乳制品灰色数据进行了充分的预处理，按检测项目性质的不同划分为四部分，结合专家打分法得到各检测项目的风险等级后分别代入改进的 softmax 公式，并根据产品中风险等级的占比对数据分箱划分区间。将 29 个地区的检测数据转换为综合风险等级后带入构建的 WD-LSTM 模型，得到整体准确率为 97.54%，标准偏差为 0.03，MAE 和 MAPE 的平均值为 0.02 和 0.83%，而本文设置的对比模型 1、2 的整体准确率分别为 86.97%和 92.42%，标准差分别为 0.14 和 0.07，平均 MAE 分别为 0.27 和 0.09，平均 MAPE 分别为 12.95%和 4.83%。该预测结果意味着本文构建的 WD-LSTM 模型预测准确性较好，且在精度和稳定性方面均优于类似的相关模型，说明该模型对乳制品质量安全预测是准确且有效的，可以起到对乳制品质量安全中潜在的风险防控和监督的作用，并在日常检测的过程中提供技术支持。对于未来的工作，可以从以下两个方向进行改善：一是通过优化模型算法，调整参数，使模型在其他类别的产品得以推广使用；二是研究如何对长时间序列的内在关联性和数据严重不平衡使用更好的处理方法。

参考文献

[1] 新型冠状病毒感染的肺炎防治营养膳食指导[J].中国食品卫生杂志,2020,32(1):61,98

[2] WANG Wenjie. Comparative analysis and enlightenment of food safety supervision system in advanced countries [J]. IOP Conference Series Earth and Environmental Science, 2020,

512(1): 012064

[3] TIAN Debin, LI Cuixia. Risk assessment of raw milk quality and safety index system based on primary component analysis [J]. Sustainable Computing: Informatics and Systems, 2019, 21: 47-55

[4] ZHANG Jiaying, ZUO Min, ZHANG Qingchuan, et al. Research on the whole chain traceability system of dairy products based on consortium blockchain [J]. MATEC Web of Conferences, 2022, 355

[5] Kollia I, Stevenson J, Kollias S. AI-Enabled efficient and safe food supply chain [J]. Electronics, 2021, 10(11): 1223

[6] 刘璐.基于供应链的乳制品安全风险识别及风险评估研究[D].北京:北京化工大学, 2020

[7] 陈嘉惠,杨巧玲,钮冰,等.乳制品质量安全风险评估及预警的研究进展[J].自然杂志,2020,42(6):494-498

[8] GENG Zhiqiang, SHANG Dirui, HAN Yongming, et al. Early warning modeling and analysis based on a deep radial basis function neural network integrating an analytic hierarchy process: A case study for food safety [J]. Food Control, 2019, 96: 329-342

[9] GENG Zhiqiang, ZHAO Shanshan, TAO Guangcan, et al. Early warning modeling and analysis based on analytic hierarchy process integrated extreme learning machine (AHP-ELM): Application to food safety [J]. Food Control, 2017, 78: 33-42

[10] MABo, HAN Yongming, CUI Shiying, et al. Risk early warning and control of food safety based on an improved analytic hierarchy process integrating quality control analysis method [J]. Food Control, 2020, 108(C): 106824-106824

[11] 白宝光,朱洪磊,范清秀.BP 神经网络在乳制品质量安全风险预警中的应用[J].中国乳品工业,2020,48(7):42-45

[12] 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

[13] Van der Fels-Klerx H J, Van Asselt E D, Raley M, et al. Critical review of methods for risk ranking of food-related hazards, based on risks for human health. [J]. Critical Reviews in Food Science and Nutrition, 2018, 58(2): 178-193

[14] LI Ying, LIANG Guoxin, ZHANG Lei, et al. Development and application of a comparative risk assessment method for ranking chemical hazards in food [J]. Food Additives and Contaminants: Part A, 2021, 38(1): 1-14

[15] Hernandez Jover Marta, Culley Fiona, Heller Jane, et al. Semi-quantitative food safety risk profile of the Australian red meat industry. [J]. International Journal of Food Microbiology, 2021, 353: 109294

[16] ZENG Bo, LI Chuan, ZHOU Xueyu, et al. Prediction model of interval grey numbers with a real parameter and its application [J]. Abstract and Applied Analysis, 2014, 2014(1): 1-12

[17] HAN Yongming, CUI Shiying, GENG Zhiqiang, et al. Food quality and safety risk assessment using a novel HMM method based on GRA [J]. Food Control, 2019, 105: 180-189

[18] LIU Sifeng, Jeffrey Forrest, YANG Yingjie. A brief introduction to grey systems theory [J]. Grey Systems: Theory and Application, 2012, 2(2): 89-104

[19] 周党生.大数据背景下数据预处理方法研究[J].山东化工,2020,49(1):110-111,122

[20] Mallat S G. A theory for multiresolution signal decomposition: the wavelet representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(7): 674-693

[21] WANG Li, ZHENG Weiguang, MA Xiaojun, et al. Denoising speech based on deep learning and wavelet decomposition [J]. Scientific Programming, 2021, 1-10

[22] Daubechies I. The wavelet transform, time-frequency localization and signal analysis [J]. IEEE Transactions on Information Theory, 1990, 36(5): 961-1005

[23] 李敏,刘岩,马然,等.一种用于海水 DOC 微弱信号去噪处理的小波多阈值算法研究[J].传感技术学报,2021,34(1):75-79

[24] Greff K, Srivastava R K, Koutn ḱ J, et al. LSTM: A search space odyssey [J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(10): 2222-2232

[25] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis [J]. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903-995

[26] LIU Xiaohan, SHI Guangfeng, LIU Weina. An improved empirical mode decomposition method for vibration signal [J]. Wireless Communications and Mobile Computing, 2021, 2021(1): 1-8

[27] Mbatha N, Bencherif H. Time series analysis and forecasting using a novel hybrid LSTM data-driven model based on empirical wavelet transform applied to total column of ozone at buenos aires, argentina (1966-2017) [J]. Atmosphere, 2020, 11(5): 457