

[](http://news.cau.edu.cn/upload/2005/caunews_20050711181514.jpg)

**本科生毕业论文（设计）**

面向乳品加工不确定性酸碱清洗数据的回归算法研究

**A Study on Regression Algorithm for Acid-Base Cleaning Data Uncertainty in Dairy Processing**

学生姓名： 吴镔芮

指导教师： 胡标 副教授

合作指导教师：

专业名称： 机械电子工程

所在学院： 工学院

2024年 5 月

**原创性声明**

本人郑重声明： 所呈交的毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文（设计）不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明。

本人签名：　　　　　　　　　　 2024 年 5 月 25 日

**摘 要**

为保证乳制品的生产质量和食品安全，有必要对其生产环境进行清洗工作。由于生产管道、罐体的密封性，需要采用酸碱清洗液与罐体管道内的残余物发生化学反应溶解，从而实现清洁功能。同时，随着生活水平提高和国家相关政策出台，对于乳制品清洗精细化提出了更高要求，极需要解决中国乳品工厂清洗工艺落后，设备清洗效果差等问题，因此提出使用原位清洗系统（Clean in the Place，缩写为CIP）。对于CIP系统的整体设计和控制系统搭建国内外已经实现和应用，但清洁程序都是基于经验制定，缺乏可靠的监控和预测作为科学辅助决策。因此，基于对酸碱清洗液用量预测和评估，进而反馈得到清洗环节是否异常以及方便工厂进行成本规划，具有实用价值。

本文提出以乳制品行业使用的CIP系统中酸碱清洗液用量为研究对象，通过对每日各个单元酸碱用量的跟踪采集获得基础数据，将其用于预测建模和效果评估。首先完成对工厂数据的预处理工作得到用于训练和预测的数据集合，通过深入分析每日酸碱使用量的相关数据特性，并全面评估了多元线性回归、贝叶斯岭回归、高斯过程回归以及支持向量回归机等多种模型方法用于预测酸碱使用量，通过滚动测试以及重采样调整数据频率对进行数据集和预测模型进行优化，最后通过预测集进行测试和评价。实验结果表明回归分析算法相较于时间序列在酸碱用量预测部分表现更好，预测结果准确度达到87.8363%，相较于时间序列预测提高328.786%，误差同比下降83.485%，证明模型有很强的泛化能力和预测意义，具有实际价值。

**关键词：**CIP清洗，乳制品，回归算法，酸碱用量，预测

**Abstract**

To guarantee the production quality and safety of dairy products in terms of food consumption, it is necessary to carry out cleaning work in their production environment. On account of the sealing of production pipes and tanks, acid and alkali cleaning liquid solution is usually used to chemically react to the residues in the tanks and pipes, so as to achieve the cleaning function. With the improvement of living standards and the introduction of national policies for dairy processing and other refined industry put forward higher requirements, in order to solve the Chinese dairy plant cleaning process backwardness, poor cleaning effect of the equipment and other issues, including the use of in-situ cleaning system (Clean in the Place, CIP). For the overall design of the CIP system and control system construction at home and abroad has been realised and applied, but the cleaning procedures are based on experience, the absence of dependable surveillance and forecasting as a scientific adjunct for decision-making. Therefore, based on the predicted value and evaluation of acid and alkali cleaning fluid usage, it is of practical value to get feedback on whether the cleaning process is abnormal or not as well as to facilitate cost planning in the factory.

we propose to take the acid and alkali usage in the CIP system as the research object, and obtain the basic data by tracking and collecting the daily acid and alkali usage in each unit, which will be used for predictive modelling and effect evaluation. Firstly, the preprocessing of the plant data is completed to obtain the data set for training and prediction, the data characteristics of the daily acid-alkali usage are analysed in depth, and multiple linear regression, Bayesian Ridge Regression, Gaussian Process Regression, and Support Vector Regression are comprehensively evaluated for the prediction of acid-alkali usage, and the data set and prediction model are optimised by rolling tests and resampling to adjust the data frequency, and finally the prediction set is optimised by the prediction model. The data set and prediction model were optimised through rolling tests and resampling to adjust the data frequency, and finally the prediction set was tested and evaluated. The experimental results show that the regression analysis algorithm performs better than time series in the prediction of acid and alkali usage, the accuracy of forecasting results reaches 87.8363%, which is 328.786% higher than that of the time series prediction, and the error decreases by 83.485% year-on-year, which proves that the model has a strong generalisation ability and significance of prediction, and has practical value.

**Key words:** CIP Cleaning, Dairy Products, Regression Algorithm, Acid And Base Usage, Prediction

**目 录**

[第一章 绪论 1](#_Toc168156665)

[1.1问题的提出与研究意义 1](#_Toc168156666)

[1.2国内外研究现状 2](#_Toc168156667)

[1.2.1国内外CIP系统研究现状 2](#_Toc168156668)

[1.2.2酸碱用量对乳制品清洗品质影响 3](#_Toc168156669)

[1.2.3基于回归分析算法的酸碱用量预测 4](#_Toc168156670)

[1.3研究目标与研究内容 7](#_Toc168156671)

[1.3.1研究目标 7](#_Toc168156672)

[1.3.2研究内容 7](#_Toc168156673)

[1.4研究方法与技术路线 8](#_Toc168156674)

[第二章 预测数据集建立 9](#_Toc168156675)

[2.1预测的基本原理 9](#_Toc168156676)

[2.1.1 预测概念 9](#_Toc168156677)

[2.1.2 预测步骤 9](#_Toc168156678)

[2.1.3 预测效果的评价 10](#_Toc168156679)

[2.2酸碱清洗液用量数据集建立 11](#_Toc168156680)

[2.2.1 数据集选择 11](#_Toc168156681)

[2.2.2 历史数据预处理 12](#_Toc168156682)

[第三章 回归分析预测问题理论 13](#_Toc168156683)

[3.1回归分析的基本原理 13](#_Toc168156684)

[3.2回归分析方法预测 14](#_Toc168156685)

[3.2.1多元线性回归分析 14](#_Toc168156686)

[3.2.2高斯过程回归分析 16](#_Toc168156687)

[3.2.3贝叶斯岭回归分析 18](#_Toc168156688)

[3.2.4支持向量回归机分析 19](#_Toc168156689)

[第四章 酸碱清洗液预测模型对比构建 21](#_Toc168156690)

[4.1基于ARIMA的时间序列预测模型应用 21](#_Toc168156691)

[4.1.1ARIMA预测模型机理 21](#_Toc168156692)

[4.1.2基于ARIMA的预测模型构建 21](#_Toc168156693)

[4.2基于回归分析算法的预测模型应用 24](#_Toc168156694)

[4.2.1酸碱用量的多元线性回归预测 24](#_Toc168156695)

[4.2.2酸碱用量的高斯过程回归预测 26](#_Toc168156696)

[4.2.3酸碱用量的贝叶斯岭回归预测 27](#_Toc168156697)

[4.2.4酸碱用量的支持向量回归预测 29](#_Toc168156698)

[4.2.5预测结果和性能分析 30](#_Toc168156699)

[4.3预测模型优化 32](#_Toc168156700)

[4.3.1酸碱用量数据重采样 32](#_Toc168156701)

[4.3.2滚动预测 34](#_Toc168156702)

[4.4预测模型总体评价 38](#_Toc168156703)

[第五章 研究结论与展望 40](#_Toc168156704)

[5.1研究结论 40](#_Toc168156705)

[5.2研究展望 40](#_Toc168156706)

[参考文献 41](#_Toc168156707)

[致 谢 43](#_Toc168156708)

[作者简介 44](#_Toc168156709)

**图目录**

[图1-1 CIP系统组成 1](#_Toc168156848)

[图1-2 乳制品原位清洗系统（CIP）工作流程图 3](#_Toc168156849)

[图1-3 技术路线图 8](#_Toc168156850)

[图3-1 梯度减小不同学习率影响 16](#_Toc168156851)

[图3-2 支持向量回归机原理 19](#_Toc168156852)

[图4-1 酸碱用量时间序列数据 22](#_Toc168156853)

[图4-2 模型定阶图 23](#_Toc168156854)

[图4-3 BIC信息准则定阶热力图 23](#_Toc168156855)

[图4-4 基于ARIMA模型的预测结果 24](#_Toc168156856)

[图4-5 多元线性回归模型预测效果可视化 25](#_Toc168156857)

[图4-6 基于不同核函数高斯过程回归模型预测效果可视化 27](#_Toc168156858)

[图4-7 基于不同参数贝叶斯岭回归模型预测效果可视化 28](#_Toc168156859)

[图4-8 基于不同核函数支持向量回归模型预测效果可视化 29](#_Toc168156860)

[图4-9 不同回归模型预测效果可视化 30](#_Toc168156861)

[图4-10 不同回归模型准确度差异 31](#_Toc168156862)

[图4-11 模型评价对比 32](#_Toc168156863)

[图4-12 重采样原理 33](#_Toc168156864)

[图4-13 时间划分情况 33](#_Toc168156865)

[图4-14 滑动窗口原理 35](#_Toc168156866)

[图4-15 固定滑动窗口下的滚动测试原理 35](#_Toc168156867)

[图4-16 不同模型下预测变化 36](#_Toc168156868)

[图4-17 基于BRR滚动预测变化 36](#_Toc168156869)

[图4-18 基于GPR滚动预测变化 37](#_Toc168156870)

[图4-19 基于SVR滚动预测变化 37](#_Toc168156871)

[图4-20 预测模型的优化过程效果展示 38](#_Toc168156872)

**表目录**

[表1-1 酸碱洗涤液对污垢清洗的区别 4](#_Toc168156364)

[表1-2 预测常见模型对比 6](#_Toc168156365)

[表2-1 清洗数据CIP报表 11](#_Toc168156366)

[表4-1 ADF检验数表 22](#_Toc168156367)

[表4-2 白噪声检测部分数组 22](#_Toc168156368)

[表4-3 ARIMA模型评估准则 24](#_Toc168156369)

[表4-4 多元线性回归的酸碱清洗液用量预测表 25](#_Toc168156370)

[表4-5 多元线性回归模型评价指标 26](#_Toc168156371)

[表4-6 高斯过程回归参数设置表 26](#_Toc168156372)

[表4-7 基于不同核函数的高斯过程回归模型评价指标 27](#_Toc168156373)

[表4-8 贝叶斯岭回归参数设置表 28](#_Toc168156374)

[表4-9 基于不同核函数的贝叶斯岭回归模型评价指标 28](#_Toc168156375)

[表4-10 支持向量回归参数设置表 29](#_Toc168156376)

[表4-11 基于不同核函数支持向量回归模型评价指标 30](#_Toc168156377)

[表4-12 基于不同时间频率回归模型评价指标 33](#_Toc168156378)

[表4-13 基于训练窗口回归模型评价指标 37](#_Toc168156379)

[表4-14 基于测试窗口回归模型评价指标 38](#_Toc168156380)

# 第一章 绪论

## 1.1问题的提出与研究意义

乳制品作为高营养消费品，含有蛋白质、脂肪、糖类、酶类、矿物质等成分，也因此成为大多数消费者的日常购买需求[1]。随着生活水平提高，我国乳制品产业得到飞速发展，2023年中国乳制品总产量实现3054.6万吨，同比增长3.1%，中国奶业产值超过千亿元[2]。但在加工生产过程中，与乳制品接触的装置内表面会产生部分残留或者生成沉积物附着于表面。一方面残留的营养物质会造成微生物大量繁殖，进而影响产品质量；另一方面沉积物可能会影响设备的传热效率，造成乳制品生产蒸发水分效率降低，造成乳制品固体损失[3]。因此在乳制品加工企业，在设备生产运行结束之后，需要进行彻底地的清洗，从而保证产品的卫生质量。

在过去10-15年的乳制品加工行业，主要采用人工手动清洗，他们需要拆卸设备并对设备和管道进行擦洗和冲刷[4]，导致了大量的人力浪费，同时存在清洗液不能回收、缺少清洗过程记录等问题。经过发展，目前均使用原位清洗系统CIP作为乳制品加工配套设施。由于CIP清洗过程不需拆卸组装，通过清洗液的化学反应、物理反应（热溶解）以及液体流动造成的机械冲洗，即实现清洗功能[5]，因此得到良好的应用和研究。CIP系统配置主要包括清洗罐体、各类控制阀和仪器仪表等部件，具体如图1-1。

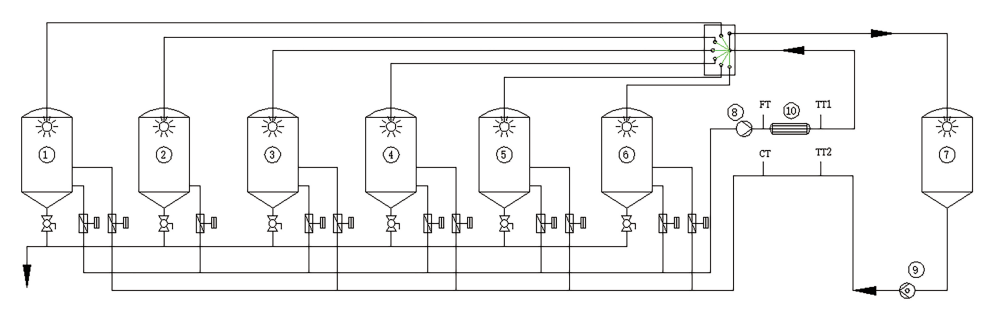


图1-1 CIP系统组成

注：①消杀罐；②清水罐；③热水罐；④酸罐；⑤碱罐；⑥回收水；⑦清洗目标

目前，对于CIP着重于控制系统和工艺流程设计，对于影响清洗效果的因素的清洗液用量缺乏探究，导致清洁程序都是基于经验制定，缺乏可靠的监控和预测作为科学辅助判断。由于影响清洗效果的清洗用量具有短期时间依赖和强烈的不确定性等特性，因此迫切需要建立对清洗效果的预测评估工具，为清洗环节决策提供帮助。

酸碱清洗液用量作为评价清洗浓度的直观体现，反映了清洗效果好坏。因此通过预测酸碱清洗液用量可以实现监控乳制品生产是否出现异常，即清洗并未达到效果或者清洗液是否造成浪费等问题；同时，酸碱用量预测也方便企业根据可能值进行成本规划，节约生产成本，也方便质监部门对企业进行管理和监督。因此我们希望能有效地利用过去数据集找到潜在联系从而自动推导出清洗用量的变化趋势描述，并且具有强大的泛化能力能够帮助我们进行预测。回归分析是预测模型构建的重要方式之一，旨在探讨多个变量之间的相互依存关系，并依据这些关系构建预测模型，即通过建立变量间的回归方程来实现预测[6]。基于以上考虑，我将根据已有的车间数据（包括影响用量的CIP清洗使用设备、具体时间、酸碱用量详细消耗记录），提取酸碱用量和生产单元关系的特征参数，并将回归分析类方法用于酸碱用量预测模型的构建，进行数据拟合和未来酸碱清洗液用量的预测值输出，并对比评价各种回归模型对于预测的准确性，得到较好的模型输出，为未来酸碱清洗液购入和监测清洗过程是否发生异常提供科学判断依据。

## 1.2国内外研究现状

### 1.2.1国内外CIP系统研究现状

近年来，围绕乳制品CIP系统研究集中于CIP机理讨论和融合自动化生产的系统设计开发，集中于通过生产周期控制和清洗剂品种控制达到更高的清洗效果。国内CIP在线清洗系统由于造价昂贵或国内外工艺不同导致CIP设备无法达到良好的清洗效果，从而不能完全适应于国内产业要求。

周大志等[7][8]等提出一种基于PLC的新型CIP自控系统，针对乳品企业独特工艺，实现高性价比功能完善的在线清洗自控系统构建，完成部分酸碱清洗液的回收利用，减少了酸碱清洗液的排放量，一定程度上降低了企业的运营成本。姜丽等[9]将这种无拆卸循环清洗的CIP系统实际应用于某电场基建锅炉，得到较好的清洗效果。张仁华等[10]深入分析了乳制品加工结垢的机理，并结合CIP系统分析其对于清洗效果的影响因素，但对于实际不确定的酸碱清洗液对于清洗效果的影响仍缺乏实际探究。虽然国内在乳制品在线清洗系统控制研发上有所尝试，但停留于系统组成模仿和基本机理的探究，清洁程序都是基于经验缺乏可靠的监控和预测作为科学辅助判断，对于清洗效果的影响因素缺乏控制。

国外对于CIP系统的研究不仅仅停留于实现生产系统的清洗自动化，而是开始对于清洗机制和清洗效果评估进行深入探索。Aramouni等[11]使用在线过程分析技术与光学成像分析结合，研究原位清洗中过程环境参数对清洗效果的影响，确定了酸碱清洗液的化学性质和清洁时间，量化了每个参数对整体清洁效果的相对贡献。Van Asselt等[12]通过多类传感器获得阀门设置、温度、电导率等参数评估CIP过程，进行“监测”清洁效果，但问题是设备仍是离线监控，即对已经清洗作用结束后的清洗效果进行评估和异常检测，因此并不适用于生产实际。Simeone[13]综合上述两类研究，提出了多传感器检测CIP系统清洗的视觉和声学系统，通过紫外荧光成像和超声波声学传感器，通过评估污垢厚度判断清洗效果，但仍然是一个无反馈检测。

可以看到目前对于乳制品的原位清洗尽管实现了初步尝试，但系统参数和清洗效果仍停留于经验，或者清洗过程开环控制，缺乏科学依据进行辅助判断清洗过程是否发生异常、清洗用量是否合理，同时企业也缺乏对于清洗液用量的规划，无法降低这部分生产成本。

1.2.2酸碱用量对乳制品清洗品质影响

沉积生成奶垢在乳制品生产中是无法避免的问题，其主要成分来自于牛乳本身，含有的脂肪、碳水化合物、矿物质等共同组成复合沉积物[14]，这些成分经过蛋白质变性、磷酸钙热沉淀反应使结构更加紧凑复杂，并吸附在设备的不锈钢表面。原位清洗（CIP）作为全球乳制品加工行业中最常用的清洁和卫生系统，通过利用附着在物体表面的沉积物与酸碱溶液的接触，产生溶解、机械冲刷、界面活化、化学反应以及热能效应等多重作用，进而有效地实现沉积物的去除。典型的CIP系统包括带有混合单元和加热器的罐体、泵、阀门等组成的管道系统和控制单元[15]，因此使用合适温度和正确步骤循环酸碱清洗液可确保卫生过程的全面控制和最大有效性。CIP系统典型工艺主要分为五步，如图1-2所示[16]。

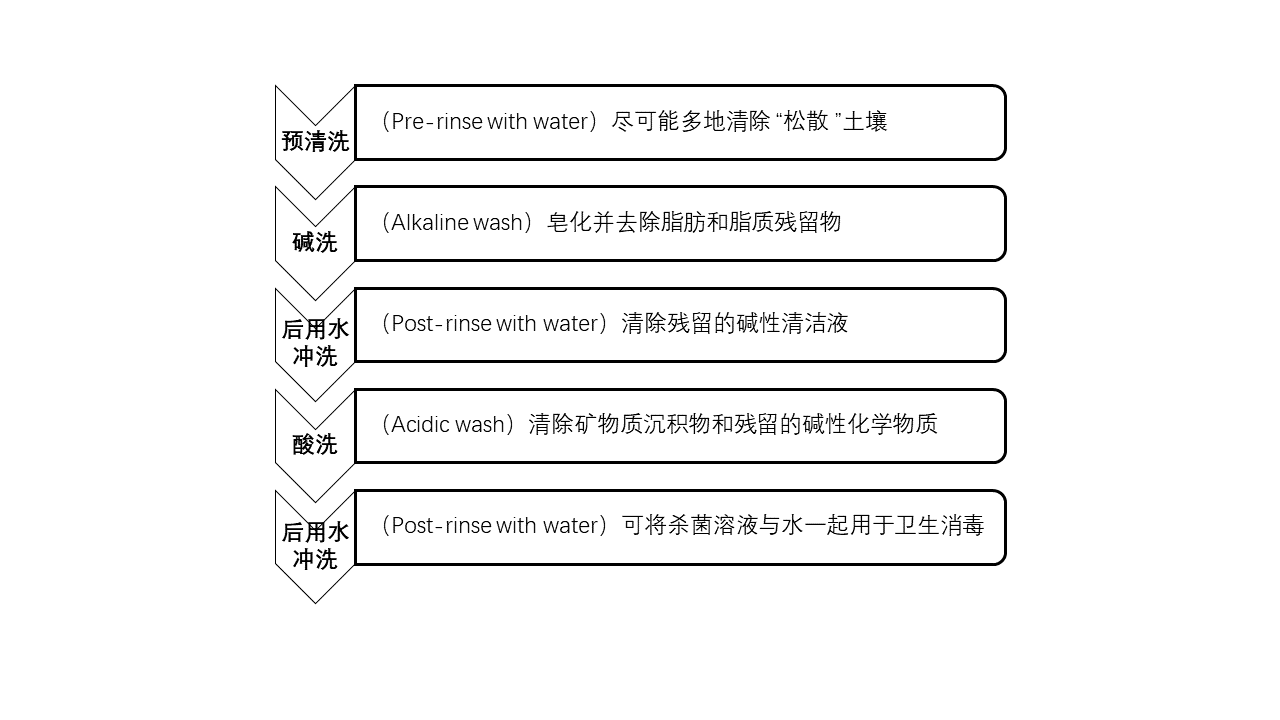


图1-2 乳制品原位清洗系统（CIP）工作流程图

根据传统CIP清洗方法，主要涉及到预冲洗、碱洗、中间水洗、酸洗、最终水洗等基本步骤，常规水冲洗需要在20℃下进行，碱性清洗液在70-80℃下使用，酸性清洗液通常在55-80℃下使用。根据碱性洗涤液溶解脂肪和蛋白质、酸性洗涤液去除设备中的矿物质沉积物（乳石）等性质（详细化学反应和对应的工作参数如下表1-1），企业会购入对应酸碱性洗涤液用于清洗设备，从清洗剂的用量可以反应使用清洗液浓度大小。张微等[17]发现即便增加酸碱清洗液的使用量能够加速清洗液与设备内部沉积物的化学反应，进而提高清洗效果，但如果清洗液用量过多会导致成本上升、设备腐蚀、蛋白质变性等反作用，导致清洗结果下降。因此在清洗中存在清洗剂临界浓度和用量，通过用量大小可以判断是否异常存在清洗异常[18]。

表1-1 酸碱洗涤液对污垢清洗的区别

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类型 | 使用清洗剂 | 参数 | | 功能 |
| 碱性清洗液 | 氢氧化钾（KOH），碳酸钠（Na2CO3），氢氧化钠（NaOH），硅酸钠（Na2SiO3） | 0.15%-5%的碱性清洗液在70-80°C下使用10-30分钟，具体取决于污染程度 | | 使脂肪发生皂化，可以很容易地用水去除。 |
| 酸性清洗液 | 硝酸（HNO3），磷酸（H3PO4），盐酸（HCl），硫酸（H2SO4） | 0.5%-1% 的酸性清洗液在55-80°C 下使用 5-20 分钟，具体取决于污染程度。 | | 去除残留的碱性洗涤剂，去除矿物质和抑菌作用。 |
| 消毒剂 | 氯基消毒剂（次氯酸盐或有机氯释放消毒剂），碘基消毒剂，过氧化氢（H2O2），过乙酸（C2H4O3） | 消毒剂所需的浓度、温度和接触时间取决于所使用的消毒剂和工厂的要求。 | 破坏病原体的营养细胞，并显着减少不需要的微生物的数量。 | |

目前对于酸碱清洗液用量对于清洗效果的影响仍停留于研究和实验参考，导致实际工厂在清洗液用量规划时仍按照研究浓度和经验配比，忽略了清洗设备污垢的时间性和随机性，必然导致清洗效果与预期存在较大误差。为了更精确捕捉清洗用量变化趋势和为未来达到清洗品质所需要的清洗用量提供科学参考，需要对历史数据进行拟合预测。

1.2.3基于回归分析算法的酸碱用量预测

酸碱清洗液用量一方面与参与清洗的单元、不同环境有关，另一方面在一定程度上受到上一次清洗残留清洗液的影响，具有一定的时间依赖性。通过研究数据的发展趋势，综合多方面信息对数据特征进行分析提取，进而建立这种相关关系的模型，拟合用量变化，实现在历史数据的经验基础上预测未来数据[19]。在实际生产应用中，对于预测类模型构建主要有以下三种方式：

1. 传统时间序列模型：

预测问题作为数理统计的分支之一，主要通过研究历史数据揭示系统内在变化结构和规律，其基础是概率论和数理统计。在这里任何预测问题都归因于三部分的叠加：趋势项+周期项+随机噪声[20]，通过分解就可以依据有限训练长度建立输入输出相互依存的数学模型，通过统计模型的数学特征进行分析映射关系实现拟合预测的目的。在基于数理统计部分模型主要有：Broersen[21]提出自回归模型（AutoRegression Model，AR）、Eswarappa[22]提出的自回归滑动平均模型（AutoRegressive Moving Average Model，ARMA）、Box[23]提出的差分自回归滑动平均模型（Autoregressive Integrated Moving Average Model，ARIMA），其利用时序图和相关系数来检验时间序列稳定性和非白噪声平稳性，此类方法计算量较小、预测速度快、容易理解且鲁棒性较差。

1. 机器学习（回归分析）：

机器学习作为一个重要的预测方法，不同于经典统计学习，其主要利用历史数据捕捉潜在的重要特征以及变量之间的关系用于数据分析和预测，凭借其对于数据进行提取处理并作出正确决策的能力[24]，广泛应用于医疗诊断、金融风险诊断、产品生产控制等领域。在实际生产情况分析中，影响决策的因素通常不只一个，因此需要针对大量样本和多种变量建立起解释模型，为了突破基于统计的多种限制条件，特别是要求序列的平稳性，研究人员提出了多种回归分析方法：Hui Zou等[25]提出弹性网络回归（Elastic Net Regression）通过结合L1和L2先验的正则化器来减小经验损失实现特征选择和防止过拟合；而Vapnik等[26]提出的支持向量回归（Support Vector Regression，SVR）通过引入不敏感损失函数ε和核函数技巧从而更好非线性问题；Johnstone[27]提出贝叶斯岭回归（Bayesian Ridge Regression，BRR）采用惩罚参数干预来解决在最大似然估计中模型复杂度无法确定的问题；最后是Williams等[28]提出高斯过程回归（Gaussian Process Regression, GPR）基于贝叶斯理论通过核函数进行无参数预测，实现为描述函数的分布情况实现具体时间序列的准确预测。相较于时间序列预测方法，使用这类机器学习算法需要通过数据处理对特征进行提取来提高预测准确性，但其计算成本相较更高，能够处理更加复杂的数据序列，使得预测结果更加准确。针对以上模型的不同表现能力，许多学者对其进行对比分析，Vining等[29]比较了高斯过程回归、支持向量回归等模型在营养物质消化率方面的预测，考虑各相应的相关性和过程经济性，并对其进行优化，结果显示基于GPR的预测技术有更好的表现能力。Del Castillo等[30]使用随机森林模型、高斯过程回归模型等预测蒸发皿的蒸发量大小，GPR仍有不错的表现。然而由于研究对象不同，其对应的特征和数据量不相同，因此在实际场景中对于多变量的预测任务，不是所有的回归分析模型都有良好的表现，因此本文重点研究酸碱清洗液用量的回归分析模型，并对于如何提高模型性能和精度进行探索。

1. 神经网络：

随着数据量的扩张，采集的时序数据具有强烈的动态非线性，因此将人工神经网络与数据预测相结合进行新的探索性研究。人工神经网络理论是基于人脑构造和运行机制进行组织建立，可以通过模拟人脑层次构建和预测的系统[31]，能够自动调节模型内部权值实现自主预测，具备出色的容错能力和自适应性，非常适用于处理众多非线性问题。在神经网络预测时间序列方面，最先是Hopfield[32]提出的循环神经网络（Recurrent Neural Network，RNN）在时间序列预测领域，提出RNN能够从时间序列数据中的不平稳数据中提取时间信息展现出优异的表现。Hochreiter[33]后提出的长短期记忆神经网络（Long Short Term Memory，LSTM），相比基础RNN模型加入隐藏层捕捉时间序列长期依赖的能力得到提升，同时，有效地解决了循环神经网络（RNN）在训练过程中出现的梯度爆炸和梯度消失问题。这些方法大部分基于循环神经网络建立了单元时间序列的时间依赖关系，在不构建函数和变量关系的情况下实现用量与其影响因素的映射过程。

表1-2 预测常见模型对比

|  |  |  |  |
| --- | --- | --- | --- |
| 预测  模型 | 模型特点 | 优势 | 缺陷 |
| 时间序列法 | 根据时序数据记录，我们探究各变量的演变趋势与内在规律 | 对于有明显周期性和趋势的数据，时间序列模型表现良好，计算成本低，模型易于解释 | 异常值敏感，要求数据平稳有时间周期性，无法考虑其他影响因素 |
| 回归分析法 | 构建特征与目标因变量之间的对应关系，以实现未来趋势的预测 | 可解决多类型输入变量，具有明确的统计意义，有大量成熟的理论支持 | 特征过于繁多导致过拟合问题的发生，在进行大规模数据预测增加计算时间 |
| 神经网络法 | 通过多层结构实现特征的自动学习和表示 | 可以通过增加层数和节点来提高模型的复杂度和预测能力 | 需要大量的计算资源，难以解释内部工作机制，数据量不足容易发生过拟合 |

表1-2对比三类预测方法中，预测问题早期使用一些数理分析方法，基于物理原理进行预测数学模型搭建，随着监控数据不断累积，基于数据驱动的方法逐渐得到广泛的应用，其中应用最为普遍的就是回归分析算法，其发展相对成熟和体系化，并且在多种实际的生产中如金融、医疗领域已经得到广泛的尝试应用。神经网络方法通过模仿人脑进行多层学习，原理机制较为模糊，但可以处理高复杂度数据。

对比模型并进行结果归因分析可以发现传统时间序列预测只对于平稳性较强的时间序列有良好的预测效果，而现实生产数据由于各种环境因素的影响，均是非平稳数据，在本文酸碱用量对象探索性尝试中也验证了其不满足非白噪声平稳条件，因此并不适宜本文研究，同时存在最大的问题是无法引入影响因素。而在酸碱用量分析中，清洗单元不同会极大影响到用量大小不同，因此使用此类模型进行预测精度很难提高；而对于神经网络算法虽然对于非线性数据有很好的表现，但需要大量的样本点投入支持预测需要，在本文研究用量对象样本个数相对较小并且不完善的数据集中难以应用，在较少数据量用于训练的情况下预测误差相对较大，不能很好捕捉数据的变化趋势。综合三种模型在本文研究对象上的预测表现，发现回归分析算法完善的理论基础和良好的预测表现都更适用于酸碱清洗液用量的拟合预测，因此本文重点研究基于回归分析算法预测酸碱清洗液用量。

## 1.3研究目标与研究内容

### 1.3.1研究目标

本文主要研究基于回归分析方法的乳制品中酸碱清洗液用量预测。进行采集预处理酸碱用量的历史工厂数据，生成带有标签和特征的集合，并划分测试集训练集，通过提取酸碱用量与生产单元来源的映射关系，识别酸碱用量变化趋势及特点进行拟合，构建合适的回归分析数学模型用于预测，并比较预测模型精度大小，检验预测模型算法的实际效果。后续继续改进预测模型，提高预测的精确度，探讨其有效性，并为工厂酸碱清洗液用量规划提供科学参考。

1.3.2研究内容

针对CIP系统中酸碱清洗液用量的不确定性预测构建需求，本文的主要研究内容如下：

（1）明确预测模型的输入输出：收集工厂从2023年1月份到2024年3月份酸碱清洗液使用情况统计表，并针对数据冗杂残缺和不匹配问题进行数据集成、数据清洗，据此构建训练和预测数据集。通过实际调研深度挖掘酸碱用量与清洗单元的关系，提取以清洗单元为主的有效的多个变量作为模型输入。

（2）进行相关理论方法的研究，构建了基于回归分析的多种预测模型进行对比实验：首先进行包括经典时间序列模型、回归分析模型的预测模型构建应用，并对预测模型效果进行对比理论分析：ARIMA模型对数据集时间依赖性进行评估得到酸碱用量数据具有很强的随机性，时序分析方法不能很好反映用量变化趋势，因此对于时间无很强依赖的数据不适用于ARIMA时间序列；通过回归分析方法组合构建预测模型，针对多单元多组成的混合数据集，通过对比各种回归分析预测模型的RMSE等指标，得到BRR、GPR、SVR模型预测未来酸碱用量经评估后预测精度得到很大提高，能够更好地反映酸碱用量与来源的映射关系。

（3）构建滚动分析对预测模型进行优化：考虑到酸碱用量短期预测属性和动态更新，采用滚动权值调整优化模型，舍弃早期对当前影响较小部分数据并给出实际应用效果，该方法实现预测模型的实时更新并有效地提高了预测的精度。

## 1.4研究方法与技术路线

本文学习乳制品产业原位清洗（CIP）的基础上，重点研究回归分析方法用于CIP清洗酸碱液使用量预测的过程。根据图1-3首先介绍了CIP系统概念，分析国内外研究现状，然后提出酸碱清洗液用量预测和回归预测模型的相关理论基础，并对搜集的数据进行处理、分析，提取酸碱用量数据与清洗单元的映射关系，并对回归分析方法进行对比预测，最后通过滚动预测和数据重采样提升模型的精确性，从而得到预测精度较高的回归分析模型。

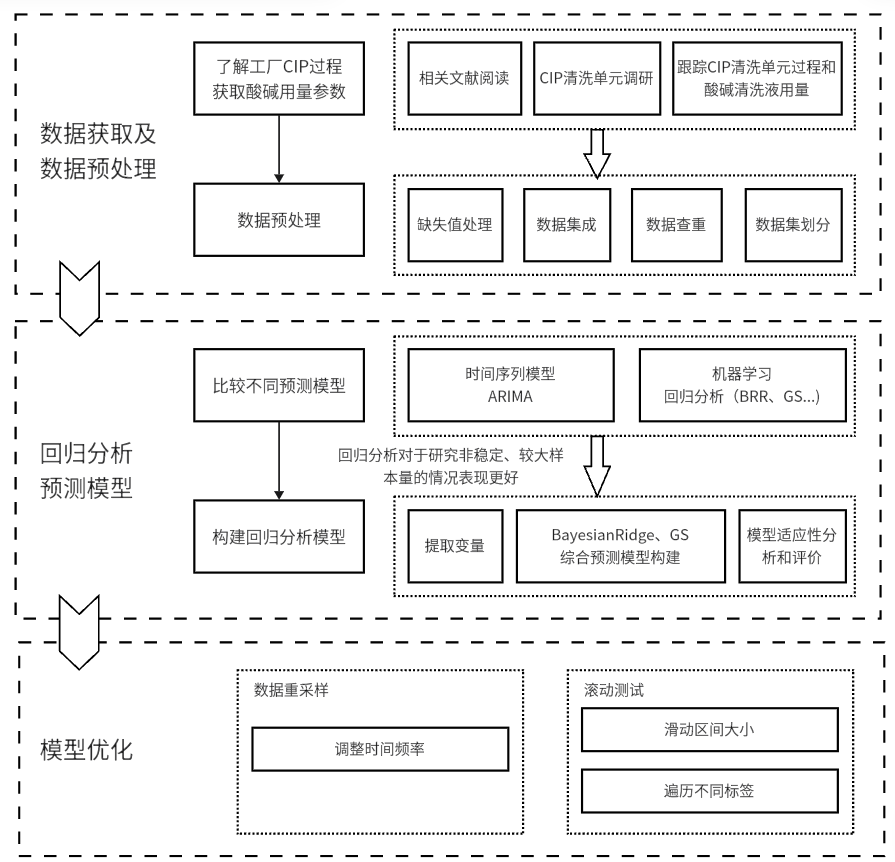


图1-3 技术路线图

# 第二章 预测数据集建立

乳制品产业CIP清洗系统的酸碱清洗液用量受到清洗过程本身的影响，因此本章通过梳理CIP清洗过程，记录清洗用量过程数据，进而探究酸碱用量的变化趋势和特点以及影响酸碱用量的指标，为后续建立预测模型提供支持。同时在本章完成对采集量化后的清洗数据预处理，得到用于预测模型构建数据集。

## 2.1预测的基本原理

### 2.1.1 预测概念

预测是通过依照事物发展所呈现的历史数据和特征，从过去的情况出发，综合考虑相关的各种影响因素，通过一定手段或技术对不可知、未出现、复杂的事物进行定性或者定量分析，从而揭示其发展规律并为人为决策提供科学依据和辅助决策。预测的三要素主要有：预测对象、预测方法和预测结果。历史数据是进行预测的基础，因此需要收集与预测对象相关的信息并对预测对象的特征进行深入剖析，进而选取恰当的预测手段进行未来趋势的预测，最后通过统一的标准进行评价预测是否满足要求。对于复杂多变不稳定的未知数据进行预测的可行性在于事物发展遵循以下几条原则：

1）连贯性原则：如果时间线上的发展不受到外界条件的干扰，这种发展规律将一直持续。这种连续性一方面是某一个标签会受到过去标签值影响，存在“惯性”，另一方面是系统结构在短期不会发生改变比较稳定，这种情况通常可以利用结果和特征的映射关系进行建立预测模型。

2）相对性原则：历史数据和未来数据在发展规律上存在相似性。

3）相关性原则：在看似复杂的系统中，影响变量与系统总体变化之间，以及各影响变量之间，存在着某种内在联系。

根据以上原则，我们发现尽管酸碱清洗液用量数据在表面上不具备上升或下降趋势比较随机，但会受到内部某些规律因素，而我们的目的就是发现这些规律并用适当的数学模型进行描述。

### 2.1.2 预测步骤

1）明确预测研究对象，确定变量的影响因素即用于模型构建的特征

2）搜集历史数据

3）针对特定的预测目标，应选取恰当的预测手段。运用多种预测手段进行预测及评估，进而甄选出最为适宜的预测结果。

4）以最新的历史数据为依据，探寻模型调整的可能性，旨在进一步提升预测的准确度

### 2.1.3 预测效果的评价

由于预测是对于未知的分析，需要统一的评价指标来判断模型优劣和算法性能，进而可以将预测结果用于现实决策或者为其提供科学辅助判断依据。机器学习中，一般采用对预测值和真实值y进行评价，通常采用以下常见指标来评价预测值和真实值的差距，从而来评价预测的准确性。在这里我们通常假设：

预测值为，真实值为，样本数量为n。

1）均方根误差（Root Mean Square Error, RMSE）：是衡量实际观测值与模型预测值之间差异的一种统计指标。其计算方式为，先对每一个观测值与对应的预测值之间的误差进行平方，然后将这些平方误差加总并求其平均值，最后对这个平均值开平方根。所得的RMSE值越小，意味着模型的预测准确度越高。



2）平均绝对误差（Mean Absolute Error, MAE）：是用于评估观测值与预测值之间误差大小的一种指标。其计算方法为取所有观测值与预测值偏差的绝对值的平均数。当MAE值较小时，表明模型的预测精度较高。利用绝对值来计算偏差，可以有效避免误差的正负相抵问题。但MAE对于数据中的异常值反应并不敏感。



3）可决系数（R2 Score，coefficient of determination）：用来衡量回归模型拟合优度的统计量。它表示模型解释的方差所占总方差的比例，即模型对观测值的变异性解释了多少比例。其数值范围介于0至1之间，数值越趋近于1，则显示出模型的拟合度越高，也就是说，模型能阐释的数据变异性更为丰富；相反，数值越接近0，则意味着模型的拟合效果不佳，无法解释数据的变异性。



其中RSS（Residual Sum of Squares）对应模型预测值和真实值的残差：



TSS（Total Sum of Squares）对应y的变化程度，与方差成正比：



4）可解释方差（Explained Variance）：在统计学中用于衡量模型解释数据变异性的量度。它表示模型能够解释的数据变异的比例，即模型对数据方差的解释能力本身是分散程度的对比，最终值越接近于1，表示预测和样本值的分散分布程度越相近。



## 2.2酸碱清洗液用量数据集建立

### 2.2.1 数据集选择

用于本文研究的数据来源某乳制品工厂原位清洗数据记录，其中使用到酸碱清洗液的过程主要分为四个：①罐、线、灌装机，②部分产品生产用到的10吨酸奶巴杀机，③对乳制品进行消毒的3吨超高温机，④20吨预巴杀使用机器，⑤不定期的用于研发车间的清洗液。在这里②-④均是一个CIP单元站，⑤是部分天数的部分用量排除，①这部分清洗按照清洗单元进行排列和划分，得到的每天清洗过程信息（不完整展示）如表2-1所示，每条数据包括4个标签，分别是：清洗单元、CIP ID、CIP程序、CIP结果，其中由于不同单元的清洗目的不同、不同程序使用的清洗液成分不同，我们主要关心使用的单元和程序以及是什么，“Caustic”、“Caustic Hot water”均会使用到碱性清洗液，“Caustic Acid”、“Caustic Acid Hot water”均会使用到酸性清洗液。包括394个样本，130个特征变量。

表2-1 清洗数据CIP报表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 单元 | CIP ID | CIP程序 | | CIP结果 |
| TM13 | 27731031126837 | Caustic | | Cleaned |
| M1L51 | 27711030915605 | Caustic Acid | | Cleaned |
| TM14 | 27731031180234 | Caustic | Cleaned | |
| PSL51 | 27751031173706 | Caustic Hot water | Sanitized | |
| FSL54 | 27780871505485 | Hot water | Sanitized | |
| TS34 | 27770871508185 | Caustic Acid | Cleaned | |
| PSL51 | 27751032194378 | Caustic Acid Hot water | Sanitized | |
| [TS31](#javascript:AddNewTab('CIP','CipEventLog) | 27770871508195 | Caustic | Cleaned | |
| [TS32](#javascript:AddNewTab('CIP','CipEventLog) | 27770871508205 | Caustic | Cleaned | |
| TS33 | 27770871508325 | Caustic | Cleaned | |
| H2L03 | 27741032146757 | Caustic | Cleaned | |

### 2.2.2 历史数据预处理

经过人为统计得到的酸碱清洗用量的原始数据中存在着多种问题，包括空缺值的存在、数据的不一致性、信息的冗余重复，以及较大的数据噪声等。在投入机器学习训练前需要一定的预处理，主要分为以下处理：

（1）数据清洗：由于环境条件和人为记录存在的错误，可能会出现异常值和缺失值的情况，将2023年1月-2024年3月数据中的缺失和明显偏离数据作为异常值处理，并对所有数据进行筛选。

（2）特征选择：通过参考文献和相关研究，除了清洗单元外，其他因素对于酸碱清洗液用量，还有一些无法量化分析和具体采集影响因素如人为操作等，考虑到这些因素的影响程度十分有限，难以系统考虑，此类影响因素统称为其他因素，在这里并不考虑进入模型预测的变量里面。因此本研究提取了各个清洗单元和其余CIP站清洗程序作为模型的输入和变量。

（2）数据集成：不同日期和月份使用CIP系统的清洗单元范围一致，其清洗程序均为“Caustic”“Caustic Hot water”“Caustic Acid”“Caustic Acid Hot water”在数据标签上保持一致，在数据预处理过程中，将不同日期的数据和标签进行匹配对应，得到以酸碱洗清洗单元为自变量，酸碱清洗液用量为因变量的数据表格。

（3）数值重采样：为了平衡环境因素造成的样本差异对预测结果的影响，将每天的酸碱用量数据进行合并组合以连续2天、连续5天……不同时间频率进行对比分析，这样就会得到多个数据集。

（4）将数据集拆分为训练集train和测试集test：由于酸碱清洗液用量会轻微受到上一个标签时间的剩余影响，会包含一定的时间信息，因此需要按照一定的滑动窗口大小和数据前后关系进行划分，实现通过训练集获得较好拟合模型用于预测独立的测试集，尽管模型性能在新测试集降低但能够实现较高精度的预测功能。

# 第三章 回归分析预测问题理论

在预测方法体系的划分下，我们学习了时序分析预测、回归模型预测以及人工神经网预测策略，明确界定了研究焦点和所收集的特定属性后，进而构建模型以获取相应结果。由于研究对象酸碱清洗液用量大部分会受到清洗单元这个重要因素的影响，而对于上一个标签时间的剩余影响较少，在时间特征上具有很强的随机性，时序分析方法和人工神经网络方法都不能很好反映用量变化趋势，因此本文的研究重点在于回归分析进行预测。

## 3.1回归分析的基本原理

对于研究对象与其相关影响因素的因果关系，通常存在两种可能：一个是存在完全确定的函数关系，另一个是变量存在统计相关，即存在一定的影响和制约但不能精准求解，在这里需要使用统计方法来研究变量之间的规律即回归分析。因此可以说回归分析就是利用研究对象与其相关影响因素之间的因果关系，确立响应Y和因素X之间的函数关系，寻找最佳拟合模型用于准确表达自变量和因变量相关性，从而作出分析与预测的方法。在回归分析中有几个重要的参数分别是：

1）因变量（响应或目标变量）是我们尝试预测或解释的结果。

2）自变量（也称为预测变量、特征或解释变量）是影响因变量的特征。

3）回归系数反映了自变量对因变量的影响程度，它揭示了在其它所有变量维持恒定的条件下，每当自变量增加一个单位时，因变量将产生的相应变化量。

4）截距表示当所有自变量都等于零时因变量的期望值，这些值是在模型拟合期间使用普通最小二乘法或最大似然估计技术来估计的。

在实践中，回归分析方法根据采用的损失函数计算方式、参数权重选择等不同有多种类型，例如线性回归、逻辑回归、多项式回归。由于不同模型的依赖原则不同，即需要满足一定假设才能准确有效进行预测，如对于线性和多项式回归需要满足残差应遵循正态分布、残差方差在自变量所有水平上应该恒定等条件。在本文研究对象为酸碱清洗液用量变化称为因变量，研究对象的影响因素清洗单元的工序作为自变量。贝叶斯岭回归和高斯过程回归、支持向量机回归在该实验中预测的精确度会存在差距，将其与简单的多元线性回归进行对比评估，研究不同回归模型在预测方面的表现。本章会详细讲解四类模型并在下一章对比具体实验效果。

## 3.2回归分析方法预测

### 3.2.1多元线性回归分析

在实际生产中，实物发展变化是受多种影响因素制约，因此需要考虑多个自变量进行组合来共同预测研究对象的变化，其中多元线性回归是回归分析研究最基本方法。在构建多元线性回归模型时，我们遵循的原则是最小化预测值与真实值之间的累积误差（即最小二乘法），以此求得回归方程中各回归系数的估计值。在该模型中，因变量是多个自变量和误差项的线性函数，在含有n（n≥2）个自变量的多元线性回归模型表达式如下：



在实际应用中通常会记录采集足够多的p组数据，，则多元线性回归表达式变为：



采用矩阵简写表示为：



其中是各因变量（响应）形成的p维列向量；是各自变量形成的矩阵包含不同变量和不同组数据特征；是回归系数矩阵；则是误差形成p维列向量，在这里的矩阵式具体阐述为：



由于作为实验涉及到的影响因素和响应的历史数据，通常是已知的，我们需要求解得到回归系数，即权重，可以反应特征和总体变量的函数关系。采用构建一个代价函数，通过最小化拟合值与真实值的累积误差来求解回归系数，在这里使用最小二乘法观测到残差平方和达到最小，确定对应的权重参数为回归参数，对于多元方程组采用矩阵求解，但过程中可能出现并不满秩的情况，在这里分开讨论：

1. 正规方程解法：

对于残差的定义为：



最小化残差平方和为准则函数：



对上述准则函数进行极小化，我们需要对系数β进行求导。当导数为0时，我们可以找到β的最优解。这个过程通常涉及到线性代数中的矩阵运算。具体来说，我们通过对准则函数关于β求偏导数，并令其等于0，从而得到一个线性方程组：



解得回归系数矩阵为：



但在实际中，因为样本量大小不够或者存在不满秩的情况，通常有两个解决方案，一个是采用梯度下降通过迭代逐步调整，一个是采用正则化实现特征选择。

2）梯度下降法：作为一种优化算法，用于最小化一个损失函数。该方法通过迭代地沿着损失函数的负梯度方向更新模型参数，以此逐步优化模型并持续降低损失函数的值。这种策略有助于模型在训练过程中逐渐逼近最优解，从而实现损失函数的最小化。具体而言其更新规则为：



其中是学习率，用于控制迭代中参数更新的步长大小，如图3-1所示，当学习率越小的时候，更新步长越小，算法收敛速度减慢，但更容易找到全局的最优解：

|  |  |
| --- | --- |
|  |  |
| (a)学习率较小时梯度减小方式 | (b)学习率较大时梯度减小方式 |
| 图3-1 梯度减小不同学习率影响 | |

3）通过L1正则化解决不满秩情况，这里会引入一种回归模型Lasso：即保留所有的特征变量，但是大量减小特征变量的数量级。对于准则函数新增一个惩罚项如，其中为正则化参数，作用是控制模型的复杂度即惩罚力度，对于过拟合的情况，惩罚项帮助损失函数中部分变量尽量小，从而减小损失函数。但为了修复过拟合的问题，会最小化即减少部分特征删除历史信息不符合实际情况，因此建议不用此回归模型进行实验。



### 3.2.2高斯过程回归分析

高斯过程回归模型是一种非参数的贝叶斯回归方法，其核心思想是将输出视为一个随机函数，其有限数量的输入点的集合服从一个多元高斯分布。通过高斯过程对该函数进行建模，可以从历史数据探索输入输出映射关系，通过给定的先验数据实现贝叶斯框架下的参数估计，并使用预测参数得到后验分布，输出预测的均值、方差、变量分布情况等信息。

高斯过程（Gaussian process, GP）指的是一种特定的随机过程集合，其中的随机变量是依据时间或空间进行索引的。在这个集合中，任意选取的有限个随机变量所构成的联合概率分布，均遵循多元高斯分布。一个高斯过程的特性可以完全由其均值函数与协方差函数来定义。换言之：



式中，使用x代指其中随机变量的函数分布的均值函数式：



协方差函数为：



在这里我们采用函数空间角度推导和解释高斯过程回归GPR：

我们给定数据集以及表达方式如下：



对任何输入函数均服从联合高斯分布，在传统回归模型中，鉴于高斯过程的先验分布是由其均值函数和协方差函数所确定的，我们可以据此推导出Y的先验分布：



其中几个关键的参数解释如下：

是噪声，通常假设为：均值为0，方差为，p为样本总数





对于测试集预测值预测：



观测值Y和预测值的联合先验分布服从高斯分布：



通过贝叶斯公式，得到的概率分布：



其中：



通过的均值方差便可得到其概率分布，通常我们取均值作为预测值。从上述推导公式可以发现，高斯过程回归预测不在于函数表达的准确性，而是通过概率拟合其分布，因此可以为预测提供不确定性估计。

### 3.2.3贝叶斯岭回归分析

贝叶斯岭回归是一种结合了贝叶斯统计学和岭回归（Bayesian Ridge Regression,BRR）思想的回归方法。岭回归通过在回归系数上施加L2正则化来减少过拟合，而贝叶斯模型的参数被视为随机变量，引入贝叶斯先验来实现参数估计（不确定性）的概率描述，从而解决最大似然估计中很难决定模型复杂程度的问题和过拟合问题，并增强模型的稳健性，我们的理论基石在于运用高斯分布作为模型参数的先验概率分布的基础。随后，借助贝叶斯公式，我们计算出参数的后验概率分布。进而，依据后验分布的统计特性（如均值和方差），我们可以精确地确定预测结果。具体的实施流程阐述如下：

1）采集具有相互独立变量描述的p组数据，，

2）使用的多元回归模型仍然采用：。为回归系数向量，是噪声，通常假设为，Y也服从高斯分布，其概率函数为：



3）贝叶斯岭回归通过将岭回归中L2正则化项解释为对回归系数的高斯先验来实现，回归系数高斯分布，均值为0、方差为，并引入L2正则项控制向量大小，避免过拟合问题，其概率密度的函数为式：



4）参数和的先验分布均取共轭分布，根据贝叶斯定理可以得到的后验分布密度函数，在这里是指训练集的似然函数，通过最大似然估计可以得到和取值：



同时可以得到回归系数的分布函数：



给定新的测试点用于预测，对应的因变量分布函数如下式



一般采取预测变量的均值作为预测值。上述推导可以发现BRR可以充分利用样本数据，采用贝叶斯原理对回归系数进行估计，在避免过拟合的同时完成预测参数，相较于普通多元线性回归其对拟合缺失数据的效果更优。

### 3.2.4支持向量回归机分析

支持向量回归机(Support vector regression，SVR)其核心思想是通过非线性映射将输入数据从低维空间映射到高维空间，并在高维空间中寻找一个最佳的回归超平面。其主要目的是平衡模型的复杂度和预测精度，从而提高泛化能力。通过在支持向量机（SVM）理论上进行延伸和拓展，保留最优超平面概念，差异在于SVM寻找超平面，使得离该超平面最近的样本点之间的距离最大化，从而便于进行分类；而SVR则寻求超平面，使得离该超平面最远的样本点之间的距离最小化，以此减少回归分析的损失，如图3-2所示。

|  |  |
| --- | --- |
|  |  |
| (a)SVM示意图 | (b)SVR示意图 |
| 图3-2 支持向量回归机原理 | |

SVR核心在于：在超平面函数两端构造“间隔带”，间距为（容忍偏差），所有位于间隔带内的样本均不纳入损失计算；仅当样本点位于间隔带外时，其产生的损失才会被计入损失函数中。模型的最优化是通过最小化间隔带的宽度以及总损失来实现的。

给定训练集为，采用基本回归模型来拟合数据，测量值与实际值存在差值，当误差小于预测正确，超出区间部分产生惩罚项，并考虑到最小化问题中：

根据SRM原则要求最小化风险：



通过最小化系数向量β的范数||β||以及综合考量预测误差，我们可以控制回归函数的复杂性，使其尽可能保持平稳，来进行模型优化：



惩罚系数：函数的经验风险和对于大于的偏差的平衡

表示置信区间来表达函数集的表达能力

体现了结构风险。

通常情况下，我们不能直接求解最小化问题，而是引入拉格朗日函数形成对偶问题，求解拉格朗日算子，并求出构造回归决策函数重要参数：



其中通过对偶问题求解：



并根据KKT条件计算得到b：



从而构造回归决策函数：



SVR一方面通过不敏感损失函数拒绝目标函数提供间隔外点的信息，保持了算法的稀疏性；另一方面通过引入松弛变量和对偶问题，找到一个在范围内偏差最小的函数，同时控制模型的复杂性。

# 第四章 酸碱清洗液预测模型对比构建

对于预测数据的大体思路可以分为：纵向分析与横向分析。其中，时间序列分析方法作为一种成熟的预测手段，适用于进行纵向的数据预测。考虑到酸碱清洗液用量的不同时间标签下的自相关性，尝试使用时间序列分析得到预测值。但根据酸碱用量时序图得到的数据，分散程度较高，缺少明显的周期性，因此本文重点将采用四种回归分析方法实现预测模型的构建：多元线性回归、高斯过程回归、贝叶斯岭回归、支持向量回归机，进行对比分析，以期获得符合企业生产需要的预测模型和更加准确的评估结果，为生产提供科学参考和辅助决策。

## 4.1基于ARIMA的时间序列预测模型应用

时间序列数据是指按照时间先后顺序排列的一系列观测值，这些观测值之间往往呈现出显著的自相关性。在预测领域，时间序列分析是一种广泛应用的方法，其中，Box和Jenkins提出的自回归积分滑动平均模型（ARIMA）是时间序列预测的一种常见且有效的方法，本文将使用ARIMA进行纵向预测并对预测结果进行评估。

### 4.1.1ARIMA预测模型机理

ARIMA模型是基于AR模型和MA模型进行合并训练，AR（自回归）模型可以捕获较长历史趋势的数据即主要取决于历史数据发展趋势，MA（移动平均）模型可以处理临时、突发变化的时间序列，因此ARIMA结合两者优点，提出核心思想是：某一时间t的标签值既受到过去一段时间内的标签值影响，也受过去一段时间内偶然事件的影响，可以将其表示为：



在这个公式中：

是我们正在预测的时间序列数据。

是AR模型参数，描述测试样本与过去p个时间数据的关系。

是MA模型参数，描述测试样本与过去q个时间误差的关系。

### 4.1.2基于ARIMA的预测模型构建

1）获取酸碱清洗液用量变化的时间序列数据，通过可视化得到图4-1。

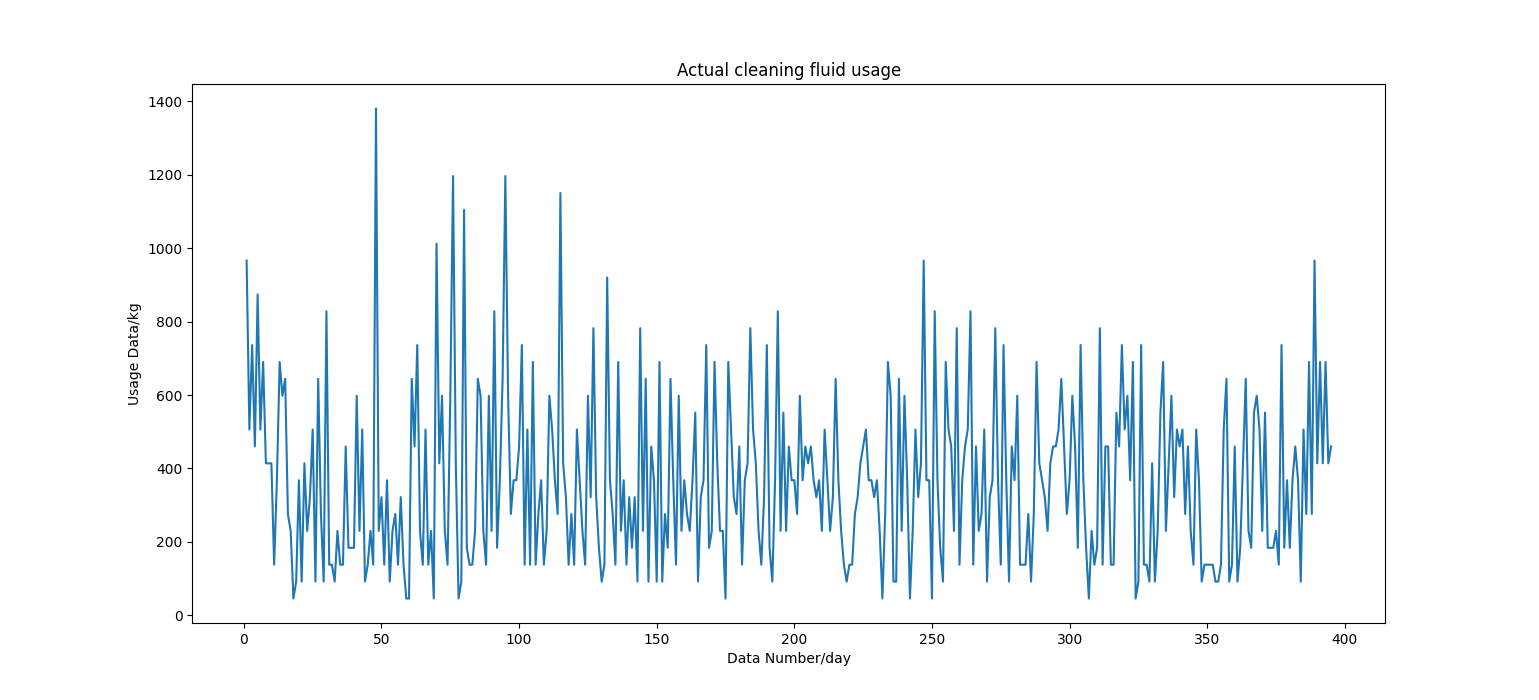


图4-1 酸碱用量时间序列数据

2）时间序列预处理

时间序列预处理主要包括两个方面检验：

1. 识别序列的平稳性

在比较1%、5%、10%的显著性水平下拒绝原假设的统计值与ADF测试结果时，若ADF测试结果同时小于上述三个统计值，并且其对应的P值非常接近于0，则表明原假设被拒绝。通过实验，我们得到了单位根检验的数据表表4-1，在本数据中adf的结果，满足拒绝原假设条件，即酸碱用量数据稳定：

表4-1 ADF检验数表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ADF Test result | P-value | 1%拒绝原假设 | 5%拒绝原假设 | 10%拒绝原假设 |
| -9.194598 | 2.079e-15 | -3.447186 | -2.868960 | -2.570723 |

②检验序列白噪声：

统计20阶延迟期数下LB和BP统计量的P值，选取部分结果列表进行判断，若p值远小于0.05，因此我们拒绝原假设（统计量的P值小于显著性水平0.05，则可以以95%的置信水平拒绝原假设，认为序列为非白噪声序列）。在这里得到的实际数据如下表4-2：在这里p值已经远大于0.05，因此已经可以判定酸碱用量序列是白噪声，不具有时间上的相关性，因此后面重点研究通过回归分析探究变量（用量标签值）与影响他的特征之间的映射关系。

表4-2 白噪声检测部分数组

|  |  |  |  |
| --- | --- | --- | --- |
| lb\_stat | lb\_pvalue | lb\_stat | lb\_pvalue |
| 0.231296 | 0.630564 | 2.978305 | 0.394982 |
| 0.235012 | 0.889135 | 12.986608 | 0.011341 |

3）模型定阶

根据自相关系数ACF与偏自相关系数PACF的截尾及拖尾特性，可以判定模型的阶数p与q，如图4-2所示。横坐标是不同的滞后程度，纵坐标是当前滞后程度下ACF和PACF的数值大小，蓝色部分代表95%的置信区间，在蓝色区域外认为序列之间的相关性彼此信任，即自相关性很强。可以看到这里PACF和ACF均为拖尾，则p≠0，q≠0，但大概1阶位置就开始基本落在置信区间，认为（p，q）取值为(1,1)。

|  |  |  |
| --- | --- | --- |
|  | | |
| (a)PACF定阶图 | (b)ACF定阶图 |
| 图4-2 模型定阶图 | | |

由于观察不准确，因此辅助热力图采用BIC信息准则定阶进行判断并生成热力图图4-3。n是观测数据量，k是模型中估计参数的数量，L是最大似然，BIC可以比较不同参数构成模型的选择标准。



当BIC越小模型拟合效果越好，得到p，q取值为（0,1）。

|  |
| --- |
|  |
| 图4-3 BIC信息准则定阶热力图 |

4）模型预测和评价

根据确定模型和阶数带入进行预测，采用前369个样本点的酸碱用量数据作为训练点、25个样本点作为参考的预测样本，得到的预测图谱如图4-4所示。

|  |  |
| --- | --- |
|  | |
| (a)预测结果 | （b）预测部分详细展示 |
| 图4-4 基于ARIMA模型的预测结果 | |

根据评估准则得到表4-3，在这里虽然准确度数值并不低但是可解释性方差和R2决定系数都小于0，说明模型预测效果和直接对训练集求平均用于预测还差，说明误差很大不具有特征因素，而且很难实现预测值与实际值匹配。

表4-3 ARIMA模型评估准则

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 可解释方差 | R2决定系数 | 均方根误差（RMSE） | 平均绝对误差（MAE） | 准确度（accuracy） |
| -0.199815 | -0.199832 | 228.972993 | 183.827721 | 0.204849 |

## 4.2基于回归分析算法的预测模型应用

从ARIMA模型预测中得到表现较差的数据拟合，同时发现，酸碱用量数据不满足非白噪声假设处于更加无序状态，与时间相关性不强，因此考虑使用横向预测，通过对其影响因素（清洗单元的使用）和输出目标的映射关系进行分析，并利用回归分析研究变量之间的特征并建立回归分析模型，从而得到预测值。在这里主要分析多元线性回归、高斯过程回归、贝叶斯岭回归、支持向量机回归四种回归模型的预测效果。

### 4.2.1酸碱用量的多元线性回归预测

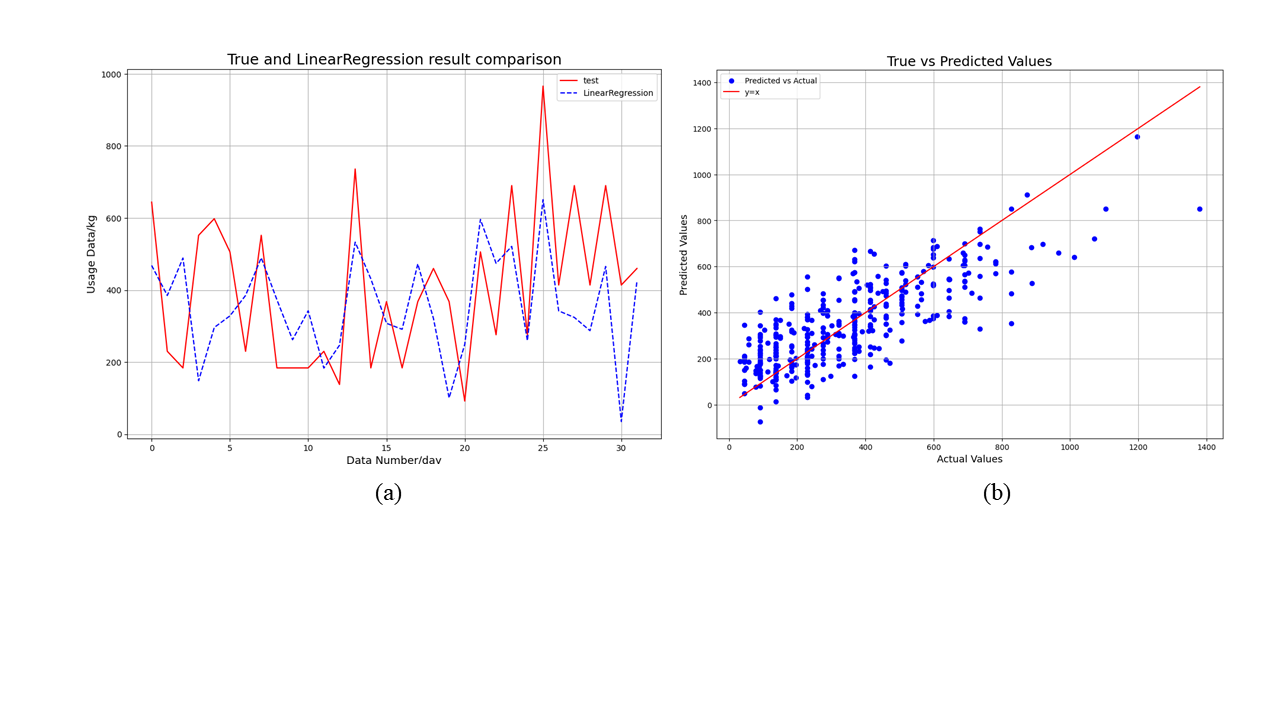
数据集处理见第二章，累计包括394个样本点，130个特征值，其中清洗单元包括如TF25、TM14、PSL54、H2L01等生产罐体管道和灌装机，以及10吨酸奶巴杀机设备、3吨超高温设备、20吨预巴杀设备和用于研发车间的酸碱用量作为特征量，酸碱清洗液用量作为输出特征。将预处理好的数据导入到Pycharm，按照23:2的比例进行划分训练集和测试集，通过对训练集进行训练可以得到最优参数，进行预测模型的构建，从而得到准确的预测值。

通过多元线性回归我们得到预测数据值大小如表4-4所示。

表4-4 多元线性回归的酸碱清洗液用量预测表

|  |  |  |
| --- | --- | --- |
| 序号 | （预测值） | 真实值 |
| 1 | 468.33124254 | 644.014 |
| 2 | 384.91918729 | 230.005 |
| 3 | 425.88131918 | 460.01 |
| 4 | 385.1507671 | 230.005 |
| 5 | 489.49285419 | 552.012 |
| 6 | 488.9299448 | 184.004 |
| …… | | |

表注：因为本文篇幅限制，因此只去部分预测值进行表示。



|  |  |
| --- | --- |
| （a）不同日期下预测结果与实际差距 | （b）预测实际残差图 |

图4-5 多元线性回归模型预测效果可视化

|  |  |
| --- | --- |
| （a）不同日期下预测结果与实际差距 | （b）预测实际残差图 |

图4-5(b)是对模型训练集数据特征拟合值和真实值评价，横轴表示酸碱用量数据的真实值，纵轴表示酸碱用量的模型拟合值，当预测与拟合越接近y=x这条线，回归分析模型越能解释用量的变化规律，可以看到模型的数据点较为集中，拟合效果较好；

|  |  |
| --- | --- |
| （a）不同日期下预测结果与实际差距 | （b）预测实际残差图 |

图4-5(a)展示了预测集部分真实值与预测值的变化和相对大小，当红线和蓝线越贴近其预测效果越好，可以看到虽然在部分时候模型的预测效果很好，但是预测趋势并不强烈匹配，因此模型的泛化能力较弱，不适用于酸碱用量预测。

为表达模型性能，我们采用可解释性方差、R2决定系数、均方根误差对模型进行综合评价，具体情况如表4-5所示。

表4-5 多元线性回归模型评价指标

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| explained\_variance | mae | rmse | r2 | accuracy |
| 0.059566 | 176.595913 | 203.489581 | 0.01258 | 0.268122 |

对于多元线性回归参数无需调整，其预测性能往往取决于数据本身。但和ARIMA模型相比，多元线性回归模型在趋势上更加贴合，同时误差进一步减小。

### 4.2.2酸碱用量的高斯过程回归预测

高斯过程回归是一种非参数贝叶斯方法，用于解决回归问题。它通过假设数据服从高斯过程来进行建模，具有灵活的非线性能力，相较于多元线性回归，不仅提供了预测值，更给出预测值的分布，即对于用量不不确定性提供了方案。在本章中仍采用394个样本点，130个特征值，数据集导入和数据集的划分同上保持一致，将预处理好的数据导入到Pycharm，按照23:2的比例进行划分训练集和测试集。

由于高斯过程采用核函数形式来表达协方差函数，实现对特征的高维映射，使得在高维空间上参数是线性相关的，本文分别采用“Rbf（径向基函数）”“C+Rbf（常数核函数+径向基函数）”“Matern”“RationalQuadratic（有理二次内核）”“WhiteKernel”五种核函数，结合网格搜索进行参数调整，建立了五个不同的高斯过程回归模型，具体的参数设置如表4-6所示，并得到不同模型下可视化预测结果如图4-6所示。

表4-6 高斯过程回归参数设置表

|  |  |  |  |
| --- | --- | --- | --- |
| Kernal | length\_scale | length\_scale\_bounds | n\_restarts\_optimizer |
| RBF | 1.0 | (1e-1, 10.0) | 10 |
| Constant+RBF | 100 | (1e-6, 1e6) | 10 |
| Matern | 1.0 | (1e-1, 10.0) | 10 |
| RationalQuadratic | 1.0 | (1e-1, 10.0) | 10 |
| WhiteKernel | 1.0 | (1e-5, 1e1) | 4 |

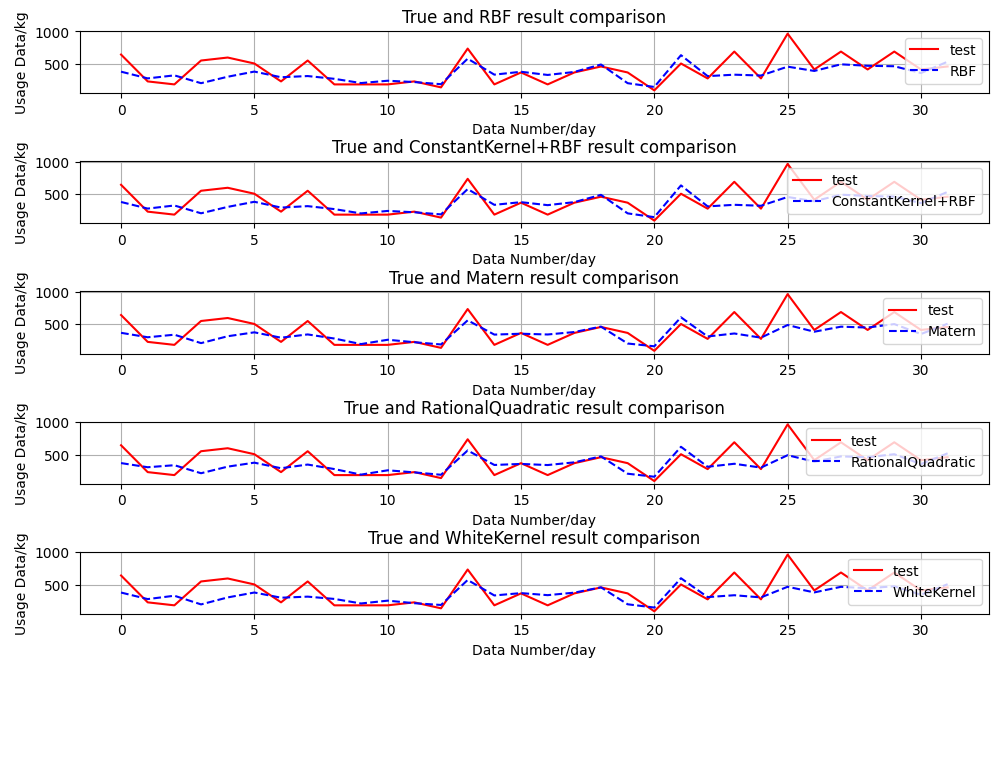


图4-6 基于不同核函数高斯过程回归模型预测效果可视化

为了刻画不同核函数影响下高斯过程回归模型的性能，分别计算其回归预测指标，如表4-7所示，其中使用Matern核函数得到的预测效果最好，和多元线性回归相比其误差更小拟合程度增大。

表4-7 基于不同核函数的高斯过程回归模型评价指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kernal | explained\_variance | mae | rmse | r2 | accuracy |
| RBF | 0.351808 | 129.906030 | 171.397949 | 0.299467 | 0.308549 |
| Constant+RBF | 0.352410 | 130.002030 | 171.314163 | 0.300152 | 0.309084 |
| RationalQuadratic | 0.386890 | 129.854832 | 166.669982 | 0.337582 | 0.313629 |
| Matern | 0.392040 | 128.911151 | 165.722160 | 0.345095 | 0.310015 |
| WhiteKernel | 0.366144 | 131.897843 | 170.508333 | 0.306720 | 0.323641 |

### 4.2.3酸碱用量的贝叶斯岭回归预测

贝叶斯岭回归是岭回归的贝叶斯版本。岭回归通过在最小二乘法中添加一个 ||L2||正则化项来解决多重共线性问题和过拟合问题，而贝叶斯岭回归进一步采用贝叶斯方法在模型参数的确定性和不确定性之间找到一个平衡，来对模型参数提供不确定性估计。在本章中仍采用394个样本点，130个特征值，数据集导入和数据集的划分同上保持一致，将预处理好的数据导入到Pycharm，按照23:2的比例进行划分训练集和测试集。

对于贝叶斯岭回归模型，其权重参数即影响因素大小并不是直接限定和改变的，而是通过给权重施加先验分布来隐式控制，其中先验分布的形状和范围由alpha\_1, alpha\_2, lambda\_1,和lambda\_2这些超参数控制，具体的参数设置如表4-8所示，并得到不同模型下可视化预测结果如图4-7所示。

表4-8 贝叶斯岭回归参数设置表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | alpha\_1 | alpha\_2 | lambda\_1 | lambda\_2 |
| 1 | 1e-6 | 1e-6 | 1e-6 | 1e-6 |
| 2 | 10 | 1e-6 | 1e-6 | 1e-6 |
| 3 | 1e-6 | 100 | 1e-6 | 1e-6 |
| 4 | 1e-6 | 1e-6 | 1e-2 | 1e-6 |
| 5 | 1e-6 | 1e-6 | 1e-6 | 1e-4 |
| 6 | 1e-6 | 100 | 1e-2 | 1e-4 |

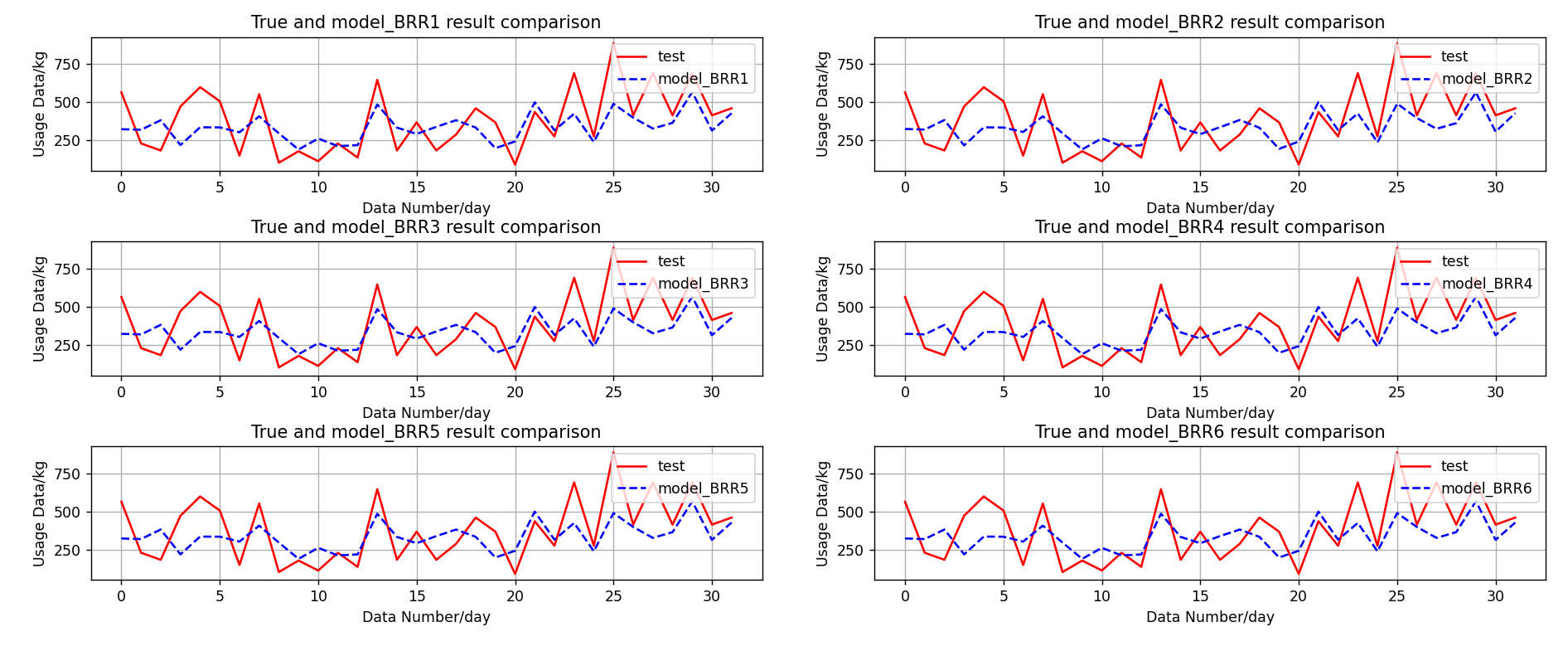


图4-7 基于不同参数贝叶斯岭回归模型预测效果可视化

为了刻画不同参数及先验分布影响下贝叶斯岭回归模型的性能，分别计算其回归预测指标，如表4-9所示，得到最佳的参数组合（alpha\_1=1e-6, alpha\_2=100, lambda\_1=1e-2, lambda\_2=1e-4），和多元线性回归相比其误差更小拟合程度增大。

表4-9 基于不同核函数的贝叶斯岭回归模型评价指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | explained\_variance | mae | rmse | r2 | accuracy |
| 1 | 0.360626 | 141.430831 | 170.107043 | 0.309979 | 0.332038 |
| 2 | 0.357567 | 142.364774 | 170.612859 | 0.305870 | 0.332317 |
| 3 | 0.360627 | 141.430512 | 170.106872 | 0.309981 | 0.332038 |
| 4 | 0.360647 | 141.424375 | 170.103594 | 0.310007 | 0.332036 |
| 5 | 0.360626 | 141.430831 | 170.107043 | 0.309979 | 0.332038 |
| 6 | 0.360648 | 141.424055 | 170.103423 | 0.310009 | 0.332036 |

### 4.2.4酸碱用量的支持向量回归预测

支持向量回归是一种基于支持向量机的回归方法。SVR通过在高维特征空间中寻找一个平面，使得大部分数据点能够落在平面上区间内，从而实现回归，通过核函数，SVR可以处理复杂的非线性关系，对于某些异常值和噪声数据，该方法展现出了良好的鲁棒性，并且其泛化能力也得到了显著提升。在本章中数据集导入和数据集的划分均保持一致，将预处理好的数据导入到Pycharm，按照23:2的比例进行划分训练集和测试集。

支持向量回归机（SVR）对参数的确定同高斯过程回归（GPR）一样，使用核函数来控制模型的偏倚和方差平衡，本文采用相同的三种核函数（“Linear”“Sigmoid”“Rbf”）结合网格搜索进行自动调参，建立三种模型，具体的参数设置如表4-10所示，并得到不同模型下可视化预测结果如图4-8所示。

表4-10 支持向量回归参数设置表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kernal | Gamma | Degree | C | ... | epsilon |
| Linear | 1.0 | 3 | 20.0 | ... | 0.1 |
| Sigmoid | 1.0 | 3 | 20.0 | ... | 0.1 |
| Rbf | 1.0 | 3 | 20.0 | ... | 0.1 |

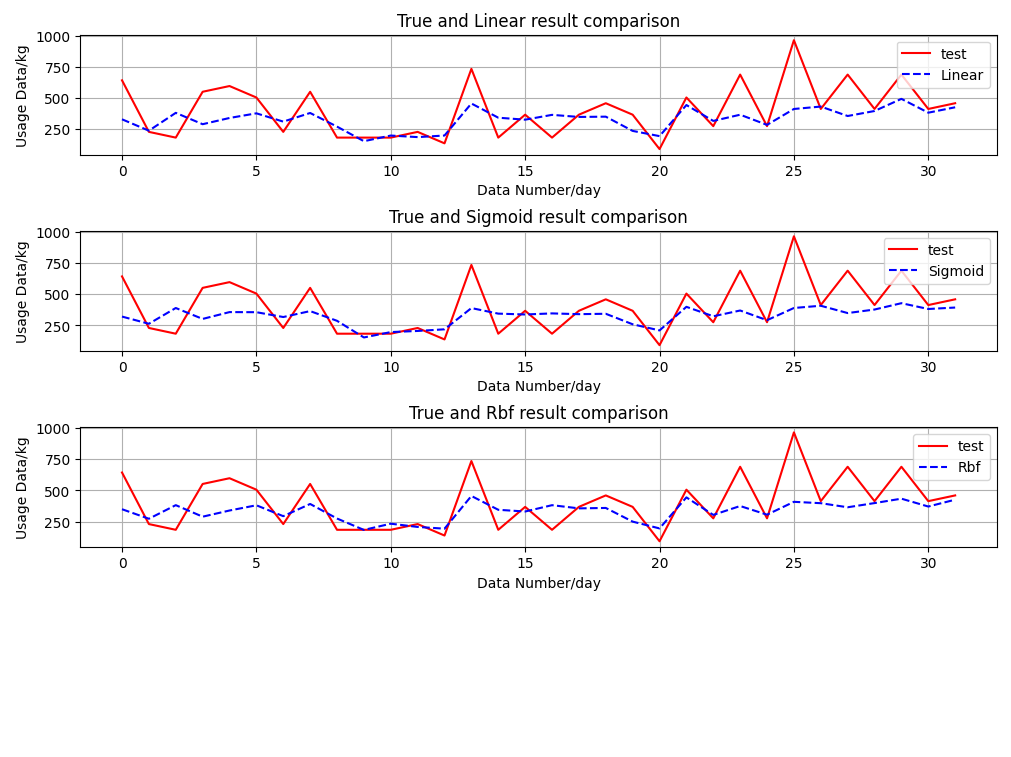


图4-8 基于不同核函数支持向量回归模型预测效果可视化

为了刻画不同核函数影响下支持向量回归模型的性能，分别计算其回归预测指标，如表4-11所示，使用简单的Linear核函数可以取得更好的预测效果，和多元线性回归相比其误差更小拟合程度增大。

表4-11 基于不同核函数支持向量回归模型评价指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kernal | explained\_variance | mae | rmse | r2 | accuracy |
| Linear | 0.368763 | 130.827254 | 174.136654 | 0.276901 | 0.359669 |
| Sigmoid | 0.293169 | 143.642525 | 185.167538 | 0.182388 | 0.379716 |
| Rbf | 0.356447 | 133.123316 | 175.282742 | 0.267351 | 0.364386 |

### 4.2.5预测结果和性能分析

为了形象比较多元线性回归、高斯过程回归、贝叶斯岭回归、支持向量机回归，本文采用预测值拟合图像以及预测准确度整合对比图像进行分析，在这里采用先前生成的数据进行对比，数据来源仍然是按照23:2的比例进行划分训练集和测试集，不同模型采用的方法和核函数取最好表达如下：多元线性回归（无参数调整），高斯过程回归（使用Matern核函数：length\_scale=1、length\_scale\_bounds=(1e-1, 10.0)、n\_restarts\_optimizer=10），贝叶斯岭回归（alpha\_1=1e-6, alpha\_2=100, lambda\_1=1e-2, lambda\_2=1e-4），支持向量回归（使用Linear核函数），具体模型效果对比如图4-9所示：

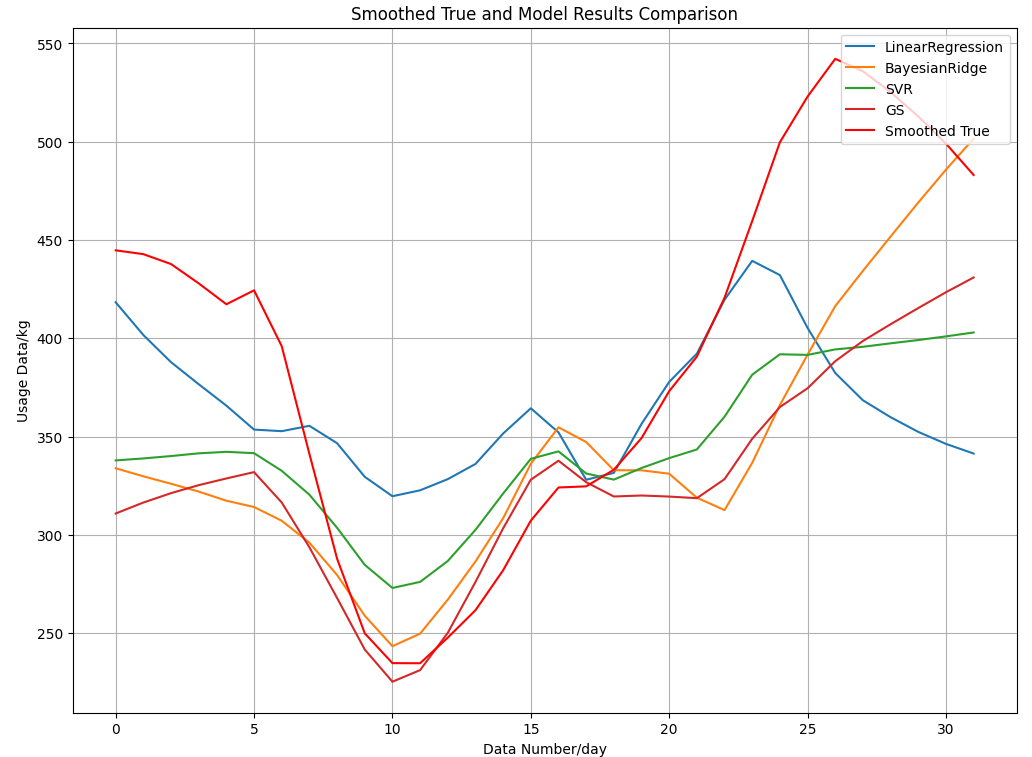


图4-9 不同回归模型预测效果可视化

通过整合四个模型预测值与真实值的差距可视化图像并对其进行局部加权平滑（Lowess），其中心思想为：以标签点为中心，对称选取长度frac（在这里我们取数据的0.35比例大小）的数据，对该值进行加权拟合，获得拟合值作为该点预测值，去除部分偏差，使得模型的预测趋势更加明显。通过**错误!未找到引用源。**对不同模型的预测对比可以看到，虽然LR在部分区域形成了高度耦合，但变化趋势并没有随着真实值变化，因此不能很好解释用量变化，而贝叶斯岭回归、高斯过程回归、支持向量回归机都相较于多元线性回归有了更好的表现。

|  |  |
| --- | --- |
|  |  |
| （a）四种模型平均准确度箱型图 | （b）箱型图注释 | |

图4-10 不同回归模型准确度差异

综合准确度箱型图4-10(a)，我们可以发现：对于四个回归分析模型，虽然不影响箱型图形状异常值仍然较多，即模型只能小范围解释预测用量变化；每组的数据盒子与箱须长度较短，说明其方差较为集中；四个模型中位数都更接近于上极限，是左偏分布更接近高值；对比四个模型，其中位数平均水平相近，根据下极限可以看到在这里多元线性回归函数的异常值和较小数据过多，分布更加离散，并没有良好反应影响因素（自变量）和预测值（因变量）的映射关系，不适用于酸碱清洗液用量的预测。根据已有数据初步判定支持高斯过程回归有更良好的表现。

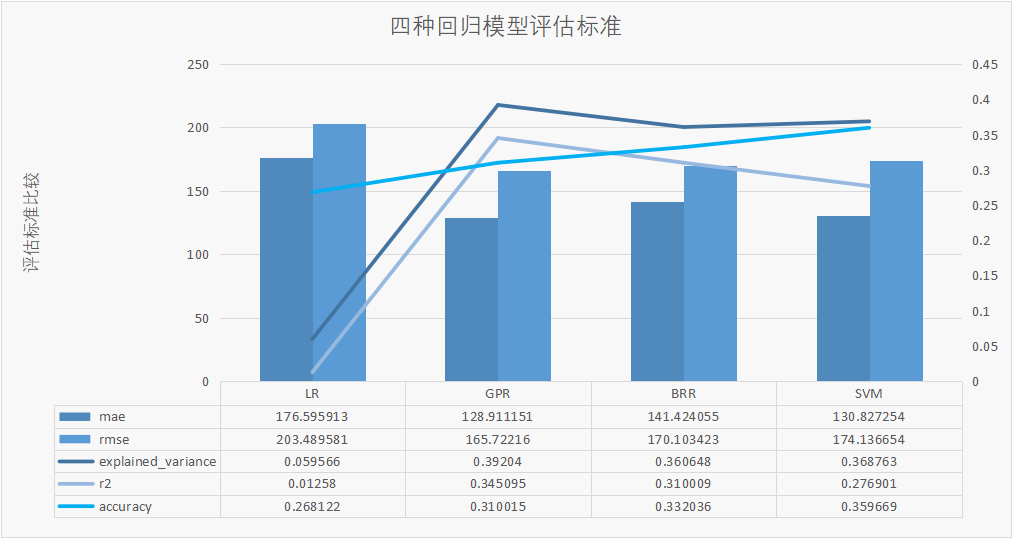


图4-11 模型评价对比

对于其他评估参数，这里采用柱状图和折线图组合进行符合对比，其中准确度在这里取预测集每单位预测准确度的均值，具体四个模型的综合评价指标如图4-11（这里均取模型最佳表现结果）使用高斯过程回归模型的均方根误差较ARIMA下降了27.624 %，较多元线性回归模型下降了18.560 %；使用贝叶斯岭回归模型的平均误差较ARIMA下降了25.710 %，较多元线性回归模型下降了16.407 %；使用支持向量回归模型的平均误差较ARIMA下降了23.949 %，较多元线性回归模型提升了14.425 %。预测效果均得到加强，在这里高斯过程回归总体优于其他模型。

## 4.3预测模型优化

在已有的模型预测中，虽然相比单纯的时间序列预测有了良好的提升，但回归分析模型本身的算法得到的预测精度仍然不足以满足实际工厂生产参考需要，因此需要对模型数据进行优化，在这里主要采用时间序列的重采样以及滚动测试的方法融合四个基本的预测模型进一步提高模型精度。

### 4.3.1酸碱用量数据重采样

在第二章数据预处理部分提到时间序列重采样，他通过改变时间序列数据的采样频率更改数据的时间间隔，减小或增大粒度，来平滑数据或调整数据的时间分辨率，从而提高预测模型的性能。其基本原理在于时间频率的选择上，如图4-12所示。

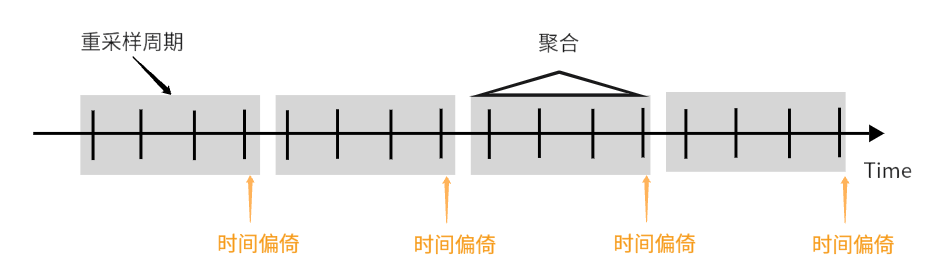


图4-12 重采样原理

因为时序数据采集时通常会存在时间标签界定与用于分析间隔所需要的时间不匹配的情况，因此在这里我们作为提高预测模型的手段，按照不同的时间间隔搜集数据，对于不同的频率进行建模（多元线性回归模型LR、高斯过程回归模型GPR、贝叶斯岭回归模型BRR、支持向量回归机SVR）和分析，如表4-12所示：在这里样本数量得到相应的减少，分别为394、393、390、385、380个样本点，按照23:2的比例进行划分训练集和测试集，具体划分情况如下图4-13：

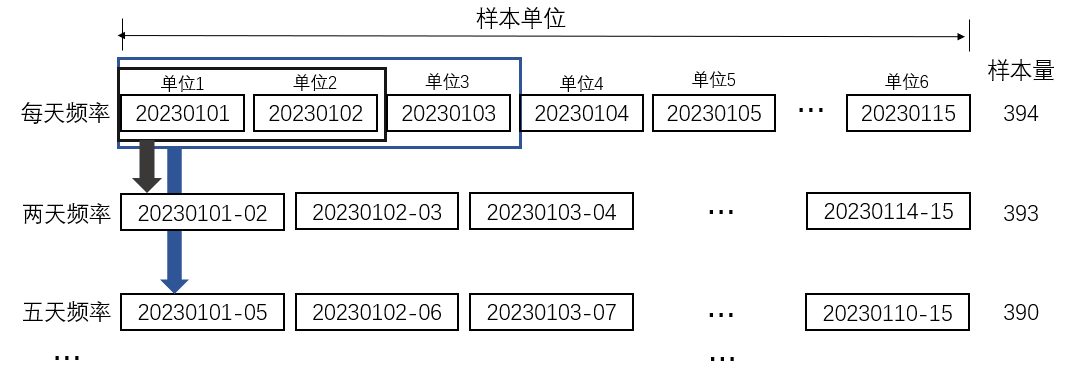


图4-13 时间划分情况

使用不同的时间频率得到的模型结果如下：

表4-12 基于不同时间频率回归模型评价指标

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 时间频率 | 回归类型 | explained\_variance | mae | rmse | r2 | accuracy |
| 每天频率 | LR | 0.126866 | 174.81591 | 202.488223 | 0.078531 | 0.343022 |
| GPR | 0.408217 | 127.245248 | 170.553894 | 0.346261 | 0.390950 |
| BRR | 0.378342 | 135.918493 | 174.346964 | 0.316860 | 0.410431 |
| SVR | 0.360992 | 134.045535 | 183.766423 | 0.241050 | 0.436012 |
| 连续两天 | LR | -0.201748 | 122.096340 | 156.736602 | -0.246949 | 0.525271 |
| GPR | 0.384221 | 93.667321 | 116.005114 | 0.316934 | 0.572587 |
| BRR | 0.390786 | 93.730138 | 119.298573 | 0.277598 | 0.590832 |
| SVR | 0.446732 | 88.220026 | 111.322850 | 0.370962 | 0.588881 |
| 连续五天 | LR | 0.096405 | 76.317850 | 109.961952 | 0.038104 | 0.716614 |
| GPR | 0.575547 | 62.016509 | 81.396644 | 0.472944 | 0.721930 |
| BRR | 0.159732 | 79.919600 | 111.185633 | 0.016577 | 0.744137 |
| SVR | 0.426449 | 68.919825 | 91.113597 | 0.339596 | 0.740924 |
| 连续十天 | LR | 0.518780 | 56.645252 | 71.352580 | 0.077880 | 0.803661 |
| GPR | 0.541879 | 40.835220 | 53.789620 | 0.475959 | 0.828034 |
| BRR | 0.530198 | 55.241244 | 74.750767 | -0.012044 | 0.826903 |
| SVM | 0.577750 | 41.858186 | 57.469659 | 0.401801 | 0.831790 |
| 连续十五天 | LR | 0.591641 | 29.347659 | 40.281777 | 0.438018 | 0.860120 |
| GPR | 0.468082 | 28.876322 | 39.365876 | 0.463283 | 0.865855 |
| BRR | 0.610947 | 35.850064 | 47.037439 | 0.233711 | 0.859956 |
| SVM | 0.618431 | 28.253703 | 38.445247 | 0.488093 | 0.873000 |

根据数据重采样处理之后减小数据频率的预测值误差明显得到降低，同时模型的准确度得到了极大的改善，当数据处理频率过低虽然准确度和拟合情况更好但用量数据接近于直线，丧失了很多特征并不适用于模型构建。因此选取连续五天的数据进行进一步探索，相较于模型本身在每天频率下的预测数据，预测误差同比减小56%左右，准确性提高到70%以上，即模型表现优良，足够反应影响因素和预测用量的关系，可以用于实际生产当中。

4.3.2滚动预测

考虑到数据可能存在的短期依赖问题，我们引入滚动测试（Rolling Forecast），即随着时间的推移，在每一个新的时间点上，使用最近的一段历史数据来更新和重新训练预测模型，从而进行下一时间点的预测，来确保预测的准确性和及时性。在数据分析和时序数据预测中：滚动预测能够实时更新和调整模型，使其能够适应最新的数据变化，保持预测的准确性，同时能够有效利用最新的数据，减少由于数据过时带来的预测误差，从而提高预测精度。

滚动预测的实现涉及到一个重要概念：滑动窗口（Sliding Window）。其最早出现在信号处理领域，用于滤波和信号分析。随后，在数据分析和时间序列预测中，滑动窗口方法是一种常用且有效的技术。它能够帮助模型捕捉数据中的短期依赖关系和变化趋势，从而提高预测精度。滑动窗口技术的引入为时间序列分析和预测提供了新的视角和工具，使得处理长序列数据变得更加高效。其核心机理是通过固定大小的窗口在数据序列上滑动，通过在每一步截取数据的一部分，形成一个新的数据子集。这些子集可以用来进行计算、分析或训练预测模型。通过滑动窗口如图4-14我们可以排除酸碱用量变化中的短期依赖关系。

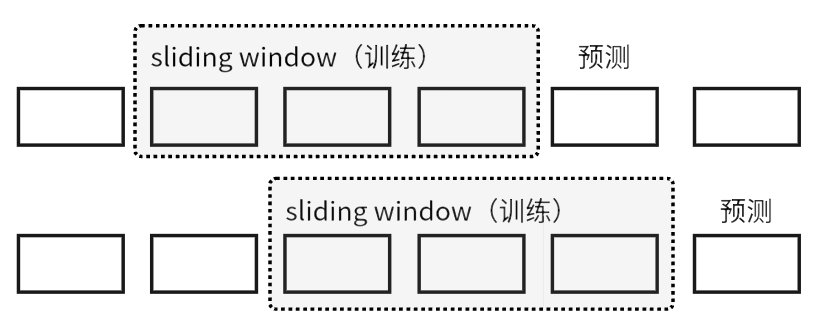


图4-14 滑动窗口原理

在这里我们分别对不同模型回归预测进行分析，考虑到模型的短期时间依赖性，如图4-16所示，在部分区域预测和实际重合度很高，说明存在局部变化趋势，为解决这部分问题，使用滑动窗口对用量数据进行子窗口数据预测和评估。具体步骤是：

1）对滑动窗口进行设置，其取值范围n从观测100天到观测390天数据，每隔20天确定一个滑动窗口大小（即滑动窗口训练集使用的天数为200、210、220...）如图4-15，根据以上四种回归分析模型进行预测，得到接下来5天的酸碱用量大小。

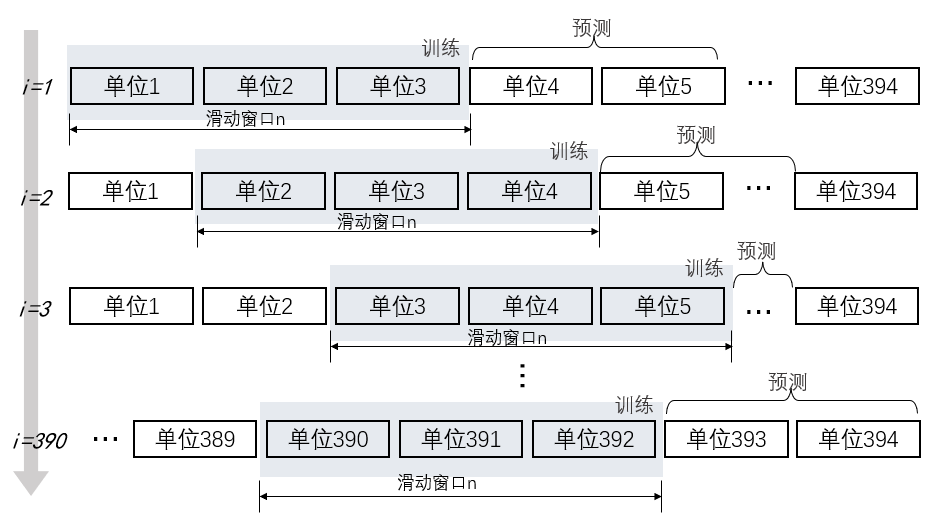


图4-15 固定滑动窗口下的滚动测试原理

2）固定滑动窗口大小即固定n值不变，按照时间顺序进行预测，这里序列从（0，n）的滑动窗口获得到一系列评估指标如RMSE、Accuracy，随着滑动窗口移动到（i，i+n），将评估指标进行平均，即可获得该滑动窗口下预测能力。

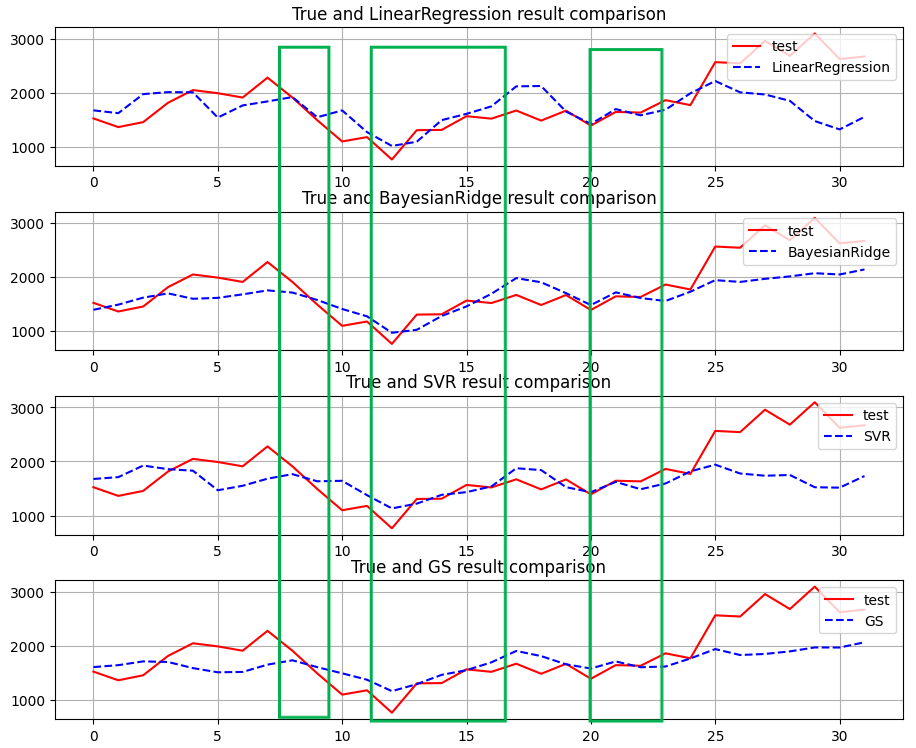


图4-16 不同模型下预测变化

对不同模型（在这里由于多元线性回归的不良好表现，并不对其进行进一步优化提升）的滑动窗口进行滚动比较，并得出预测最佳值，比较不同模型和滚动窗口大小对于准确度的影响：图4-17到图4-19中n代表滑动窗口的大小，Accuracy是在不同滑动窗口下进行滚动预测的分布情况。

|  |
| --- |
|  |
| 图4-17 基于BRR滚动预测变化 |
|  |
| 图4-18 基于GPR滚动预测变化 |
|  |
| 图4-19 基于SVR滚动预测变化 |

根据以上图像的变化趋势，进一步细化滚动测试的数据划分，在训练集大小区间为（100-300）之间，每隔5天取得一个滑动窗口大小，得到最佳模型预测效果及滑动区间的选取如表4-13所示。

表4-13 基于训练窗口回归模型评价指标

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 回归类型 | 最佳参数对应滑动窗口大小 | | rmse | accuracy |
| BRR | | 165 | 54.205271 | 0.786062 |
| GPR | | 135 | 53.715774 | 0.807799 |
| SVM | | 135 | 55.389292 | 0.805165 |

根据以上图表的反应，尽管在训练集数量达到350个之后基本无异常值的出现，但其准确度、误差曲线都表明350左右的数据发生突变且不符合发展规律，不分析这部分数据。可以看到高斯过程回归相对于贝叶斯岭回归更加稳定，改变预测窗口大小对其结果影响变化更加缓慢，而贝叶斯岭回归考虑更多样本影响，导致更少的异常值。而支持向量回归机，虽然相较于前两个模型受滑动窗口选取影响较大，但经过测试即使在样本量较少的情况下（在训练集只有20个样本的情况下，平均准确度能达到0.61），也可以进行较为准确的预测。

除此以外考虑到短期依赖问题，对于预测集窗口的选取也进行重新滚动测试，选取表现最好时间频率五天连续样本为实验内容，滑动区间的变化以一天、三天、五天样本点为测试集得到不同模型的预测值，探究不同模型对于短期预测，由于一天的测试集已经无法求取其R2决定系数和可解释性方差，因此，这里主要采用RMSE、MAE、Accuracy作为评估指标得到表4-14：

表4-14 基于测试窗口回归模型评价指标

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 回归类型 | 滑动预测窗口选择 | Mae | Rmse | Accuracy |
| BRR | 1 | 38.905608 | 38.905608 | 0.874971 |
| 3 | 44.583348 | 49.605597 | 0.821637 |
| 5 | 47.683143 | 54.205272 | 0.786062 |
| GPR | 1 | 37.813889 | 37.813889 | 0.878363 |
| 3 | 43.731236 | 49.326118 | 0.834973 |
| 5 | 46.462771 | 53.715774 | 0.807799 |
| SVM | 1 | 44.847780 | 44.847780 | 0.850536 |
| 3 | 47.019220 | 52.603648 | 0.823829 |
| 5 | 47.782850 | 55.389292 | 0.805165 |

## 4.4预测模型总体评价

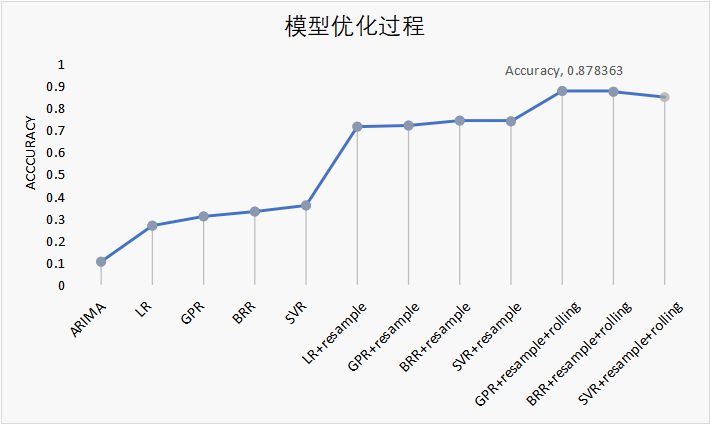


图4-20 预测模型的优化过程效果展示

通过纵向对比所做所有实验，得到不同模型的准确度优化过程图4-20。相较于未进行优化的预测模型和数据处理方式，对于高斯过程回归模型其准确度性能提升了183.33%，均方根误差减小了77.182%；对于贝叶斯岭回归模型其准确度提升了163.52%，均方根误差减小了77.770%；对于支持向量回归机其准确度提升了136.477%，均方根误差减小了74.246%。其准确度均达到80%以上，即预测模型表现较好，足够生产实际的需要。

# 第五章 研究结论与展望

## 5.1研究结论

本文针对国内乳制品行业使用的原位清洗系统（CIP）的工作机理进行探究，引出本文的主要研究对象乳品加工中酸碱清洗用量的预测，对比分析四类基本的回归分析方法并使用其进行预测模型构建，通过预测均方差、准确率等对模型效果进行评估，通过准确的预测结果为工厂成本规划和异常检测提供科学判断。针对研究目的，本文主要开展以下工作：

（1）本文数据来源工厂从2023年1月份到2024年3月份酸碱清洗液使用情况统计表，并完成数据预处理工作。构建训练和预测数据集，通过实际调研深度挖掘酸碱用量与清洗单元的关系，提取以清洗单元为主的有效的多个变量作为模型输入。

（2）本文提出了针对有部分时间依赖数据的预测方法，一种是时间序列预测，一种是回归分析预测。根据实验效果，在白噪声检测中发现，虽然酸碱清洗液用量数据满足平稳要求，有很强的随机性，因此相较而言，酸碱清洗液不具有较好的时间依赖性，无法采用时间序列的分析方法进行研究。而回归模型探究其影响因素（清洗单元的使用）和输出目标的映射关系进行分析，能够更好地解释酸碱用量变化趋势，对于某些异常值和噪声数据，展现出了良好的鲁棒性，并且其泛化能力也得到了显著提升。本文采用清洗单元数据为影响因素，酸碱用量为预测对象，通过对多元线性回归、BRR、GPR、SVR方法进行对比，验证了回归分析方法的有效性。

（3）基于四种基本的回归分析方法，针对短期依赖和异常值问题，采用数据重采样和滚动测试对预测模型进行优化，舍弃早期对当前影响较小部分数据并给出实际应用效果，实现模型实时调整并有效地提高了预测的精度。

## 5.2研究展望

本文研究了基于回归分析的酸碱清洗液用量预测，取得了一定的效果。但由于时间有限，仍有很多问题和方向值得考虑：

（1）本文选取了回归分析的特征为清洗单元的过程，但对于特征的选取还有其他的非量化的影响因素，可能会对数据造成一些影响，因此下一步需要得到更全面详细的观测数据。

（2）本文提出的回归算法主要是实现变量相关性的探究偏重横向分析，但对于酸碱用量时间依赖性并未更深入研究，由于数据集较小神经网络缺乏良好表现，下一步继续采集工厂样本，尝试将时序和变量因素共同考虑进入预测模型。

# 参考文献

1. 乳品工业手册编写组. 乳品工业手册[M]. 北京: 轻工业出版社, 1987, (12):428-436
2. 李胜利, 姚琨, 刘长全, 等. 2023年度奶牛产业与技术发展报告[J]. 中国畜牧杂志, 2024, 60(03):338-341
3. 张水成, 王沂, 张世卿. 食品工厂设备清洗系统CIP的原理与发展前景[J]. 中国农学通报, 2006(06):91-94
4. Memisi N, Moracanin S V, Milijasevic M, Babic J, Djukic D. CIP Cleaning Processes in the Dairy Industry[J]. The 58th International Meat Industry Conference (MeatCon2015), 2015, 5:184–186
5. 乔聚林, 蒲俊. 乳品厂原地清洗(CIP)设备的应用[J]. 包装与食品机械, 2001(06):25-30
6. 游士兵, 严研. 逐步回归分析法及其应用[J]. 统计与决策, 2017(14):31-35
7. 周大志. 一种基于PLC的新型在线清洗自控系统的研制[D]. 哈尔滨工业大学, 2018
8. 李湘伟, 刘尉. 基于PLC控制的全自动在线清洗系统的设计[J]. 食品与机械, 2015, 31(6):120-123
9. 姜丽, 马瑶瑶, 王亚敏, 等. 在线清洗工艺在某基建锅炉清洗中的运用[J]. 清洗世界, 2008, (07):11-14
10. 张仁华, 王成. 乳制品加工设备结垢与清洗综述[J]. 中国乳业, 2023(08):110-118
11. Aramouni N A K, Steiner-Browne M, Mouras R. Application of process analytical technology (PAT) in real-time monitoring of pharmaceutical cleaning process: Unveiling the cleaning mechanisms governing the cleaning-in-place (CIP)[J]. Process Safety and Environmental Protection, 2023, 177:212-222
12. Van Asselt A J, Van Houwelingen G, Te Giffel M C. Monitoring system for improving cleaning efficiency of cleaning-in-place processes in dairy environments[J]. Food and bioproducts processing, 2002, 80(4):276-280
13. Simeone A, Watson N, Sterritt I, et al. A multi-sensor approach for fouling level assessment in clean-in-place processes[J]. Procedia Cirp, 2016, 55:134-139
14. Guerrero-Navarro A E, Ríos-Castillo A G, Avila C R, et al. Development of a dairy fouling model to assess the efficacy of cleaning procedures using alkaline and enzymatic products. LWT - Food Science and Technology, 2019, 106:44–49
15. Bremer P, Seale B, Flint S, et al. Biofilms in dairy processing[J]. Biofilms in the food and beverage industries, 2009:396-431
16. Gugała M, Zarzecka K, Kapela K, et al. Analysis of energy consumption in the facilities of the Ciechanów Dairy Cooperative. Journal of Ecological Engineering, 2015, 16(3):139–142
17. 张微, 杨国栋, 袁雄雄, 等. 原位清洗系统在乳制品加工中的应用[J]. 中国乳业, 2019, (03):61-66
18. 张水成, 王沂, 张世卿. 食品工厂设备清洗及CIP系统[J]. 食品科技, 2006(08):167-170
19. 朱连章, 杨杰. 基于GA和ANN的随机规划模型对储层参数的预测[J]. 微计算机应用, 2005, 26(2):3
20. 夏丽. 基于ARIMA模型及回归分析的区域用电量预测方法研究[D]. 南京理工大学, 2013
21. BROERSEN P M. Finite-sample bias propagation in the yule-walker method of autoregressive estimation[J]. IFAC Proceedings Volumes, 2008, 41(2):2744-2749
22. ESWARAPPA C, HOEFER W. Autoregressive (ar) and autoregressive moving average (arma) spectral estimation techniques for faster tlm analysis of microwave structures[J]. IEEE Transactions on Microwave Theory and Techniques, 1994, 42(12):2407-2411
23. BOX G E, PIERCE D A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models[J]. Journal of the American statistical Association, 1970, 65(332):1509-1526
24. 陈慧灵. 面向智能决策问题的机器学习方法研究[D]. 吉林大学, 2012
25. Zou H, Hastie T. Regularization and variable selection via the elastic net[J].Journal of the Royal Statistical Society, 2005, 67(5):768-768
26. Vapnik V. The Support Vector Method of Function Estimation[J]. Nonlinear modeling: Advanced black-box techniques. Boston, MA: springer us, 1998:55-85
27. Johnstone I M, Silverman B W. Needles and Straw in Haystacks: Empirical Bayes Estimation of Possibly Sparse Sequences. The Annals of Statistics, 2004, 32(4):1594-1649
28. Williams C K I, Rasmussen C E. Gaussian processes for regression, advances in neural information processing systems[J]. MIT Press, 1996:514-520
29. Vining G G. A compromise approach to multiresponse optimization[J]. Journal of Quality Technology, 1998, 30(4):309-313
30. Del Castillo E, Montgomery D C, McCarville D R. Modified desirability functions for multiple response optimization[J]. Journal of quality technology, 1996, 28(3):337-345
31. 张利. 基于时间序列ARIMA模型的分析预测算法研究及系统实现[D]. 江苏大学, 2008
32. Hopfield J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proceedings of the national academy of sciences, 1982, 79(8):2554-2558
33. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780

# 致 谢

又是一年六月，我又一次踏上了告别。我与农大的故事开始与初秋，落幕于盛夏，忽而想起开始的迷茫困顿失措，我错过的那么多浓墨重彩的时刻，拥有了很多犹豫不决和筋疲力尽的时刻，可时间却滚滚不停，推着我向前跑，我甚至来不及回顾这步履匆匆的四年，因此我为这篇短暂的致谢回顾打过无数次腹稿，在每一个难熬和幸福的时刻、每一个团聚和分离的时刻，因此我想在致谢把四年的我留下来。

大学四年，感谢相遇的每一位老师，尽管我并不优秀也没能力收获到所有的知识，可那些谆谆教诲、大量的实验考试其实教会我不仅是理论还是为人道理亦或者是一种思维方式。感谢我的毕设指导老师胡标，何德何能在大学末期遇到这么负责的老师，无论是前期定题、阅读文献、采取模型构建方法、实验方法改进都提出的很多建议和帮助，因为老师我的实验每一步都有目标有方向，我学习到很多分析解决问题的方法，在此表示真挚的感谢。除了感谢还有抱歉，辅导一窍不通的我真的辛苦了。

感谢我的父母，我想如果我身上有任何值得为人所称赞的品德，一定是来自于他们的言传身教，而这十几年的学习生涯，给了我最大的支持，我时常有着各种不撞南墙不回头的执拗，是他们的尊重给了我选择的底气和可能，我可以去尝试所有的爱好和大胆的冒险。

感谢过去的每一次相遇，和他们的相遇像是在我贫瘠的土地上生出的玫瑰，灿烂耀眼、夺目可贵，人海茫茫、时间短暂，遇到很多真挚热情温暖的人，他们在不经意间塑造了现在的我，有幸能同行一程，我很满足。

感谢过去的自己，我记得努力伸出援手解救于煎熬的自己，不被瓦解撂倒，从未停止奔跑，在琐碎的生活里，我努力给自己创造着幸福。大学四年，我也尝试过很多喜欢的事情，去过很多想去的地方，见了耳机里很想见的人，在有限中自由地践行自己的信条，虽偶有害怕、退缩，但也努力完成，在22岁，我替18岁的自己完成了大部分梦想，那个曾在高中贴上心愿明信片的女孩，一步步走向了属于她的旷野。

自此，行文至此，落笔为终，论文的写作就告一段落，愿自己永远怀着最初的善良和真诚，无比坚定地继续相信自己，接受所有的自己。感恩相遇，期盼重逢，希望我们好运，万事胜意，所愿皆所得。

吴镔芮

2024年6月

# 作者简介

基本介绍（姓名、性别、出生日期、籍贯等）：

吴镔芮，女，汉族，中共党员，2002年4月出生于重庆市渝北区。

教育经历：

2020年考入中国农业大学工学院机械电子工程系

2024年保研至北京理工大学机械与车辆学院激光微纳制造理论与应用专业

本科期间主持/参与的科研项目：

校级创新创业训练项目《机械系统振动分析方法及试验研究》2021.1-2021.12

北京市创新创业训练项目《种用玉米杂穗智能分拣装置研制》2022.3-2023.3

北京市创新创业训练项目校际合作计划《秸秆变速粉碎还田系统装置的研究》2022.6-2023.6

本科期间获得的奖励和荣誉：

2022“高教杯”全国大学生先进成图大赛二等奖

2020-2021、2021-2022 学年三等学业奖学金、2022-2023学年学业二等奖学金

2022-2023学年大北农励志奖学金

2020-2021 学年校级优秀共青团干部

2022"外研社·国才杯"全国英语阅读大赛初赛二等奖

2021 年北京市大学生工程设计表达竞赛二等奖

2021-2022 学年校级三好学生

其他成果：

软件著作权《秸秆粉碎还田装置变速控制系统V1.0》