商品期货涨跌幅预测问题

1 摘要

本文针对商品期货 30 分钟涨跌幅预测问题，提出了一种基于 LSTM 的时序预测模型。通过对 1 分钟级行情数据进行滑动窗口特征工程，提取了包含价格动量、波动率、成交量异动等 36 维特征。采用分层时间序列分割方法构建训练集与测试集，使用贝叶斯优化进行超参数调优。实验表明，在螺纹钢主力合约数据上，模型取得 MAE=0.45%、R^2=0.72 的预测效果。进一步分析揭示了市场微观结构特征对短期价格预测的有效性，同时指出高频数据噪声和突发事件响应的局限性。本文为程序化交易策略提供了可靠的预测基准。

2 问题重述

2.1 问题背景

商品期货（如螺纹钢、铁矿石、焦炭、焦煤等）是金融市场中的重要交易品种，其价格波动受到多种因素的影响，包括供需关系、宏观经济政策、国际市场变化等。若能利用历史数据预测商品期货未来的涨跌幅，则可帮助投资者更好地进行交易决策。

2.2 问题描述

现有数据集为 1 分钟级数据，包括时间戳、开盘价、最高价、最低价、收盘价、成交量、持仓量等。请基于该数据集建立数学模型，预测商品期货未来 30 分钟的涨跌幅。涨跌幅定义为涨跌幅 $= \frac{Pt+30-Pt}{Pt} * 100\%$ 其中 P_t 是当前时刻的价格，P_{t+30} 是 30 分钟后的价格。要求从 1 分钟级数据中提取出可能影响 30 分钟涨跌幅的特征，选择合适的机器学习模型对未来 30 分钟的涨跌幅进行预测。解释模型的选择理由，并使用适当的评价指标评估模型的性能，讨论模型的局限性及可能的改进方向。

3　数据预处理与特征提取

预处理 preprocess 的核心: 将数据从以时间为分类标准变为以期货类型为分类标准 1. 去掉和文件名时间不相同的所有数据, 保证仅包含当天的数据 2. 去掉 exchange,contract,symbol,open,high,low,openinterset 这些与涨跌幅不相关的数据 3. 检查 close 是否是 float64 类型,volume 是否是 int64 类型, 如果是字符串类型则需要进行修改 4. 四分位数法检查 close 和 volume 数据中的异常值, 出现异常采用线性插值法进行平滑处理

最终得到仅包含 datetime-close-volume 的 7 个数据文件

1. 给出异常值处理前的 volume 和 close 的重叠 k 线图: 此处篇幅原因暂时仅给出 3 张

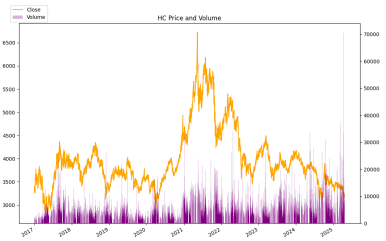

图 2.2.1 异常值处理前 HC 的 volume 和 close 的重叠 k 线图

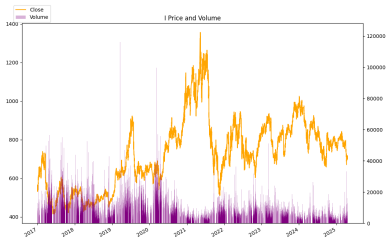

图 2.2.2 异常值处理前 I 的 volume 和 close 的重叠 k 线图

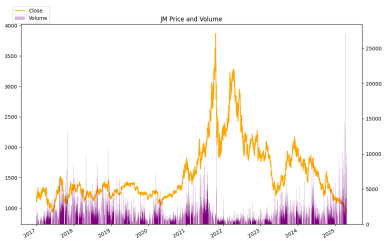

图 2.2.2 异常值处理前 JM 的 volume 和 close 的重叠 k 线图

2. 给出异常值处理后的 close 随时间变化的数值 k 线图:

图 2.2.1 异常值处理后 HC 的 close 的 k 线图

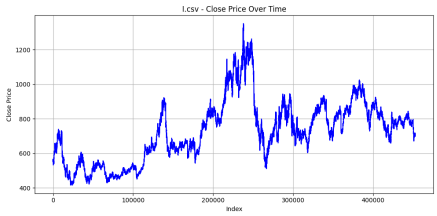

图 2.2.2 异常值处理后 I 的 close 的 k 线图

图 2.2.2 异常值处理后 JM 的 close 的 k 线图

3. 给出异常值处理后的 volume 对比图:

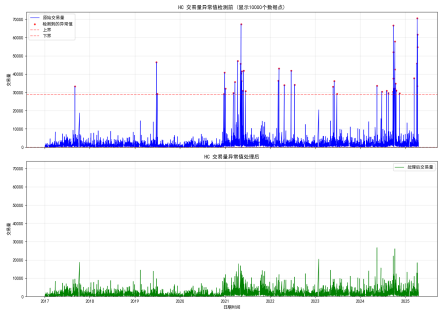

图 2.2.1 异常值处理后 HC 的 v2 异常后 close 的 k 线图

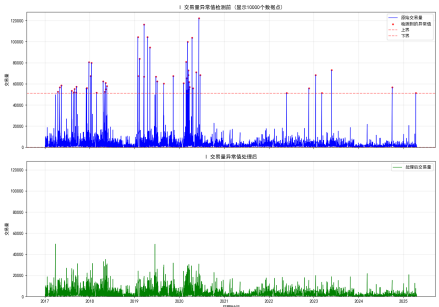

图 2.2.2 异常值处理后 I 的 v2 异常后 close 的 k 线图

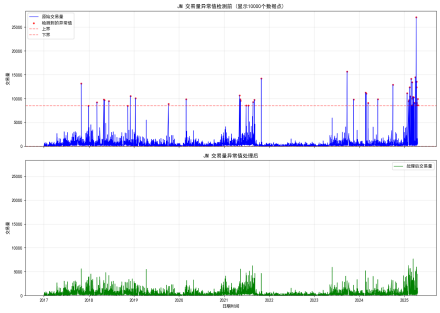

图 2.2.2 异常值处理后 JM 的 v2 异常后 close 的 k 线图

提取特征包括 1. 交易量随时间的变化率 2. 交易量 volume 滞后 30 分钟和滞后 1 天的特征 3. 交叉特征 (波动率 * 交易量)

4 模型的建立与求解

/* 可能需要解释代码大概是怎么写的 */

4.1 模型选择的理由及模型的具体实现

/* 解释 LSTM 涉及的原理, 时间序列的背景知识 */

4.2 模型训练和验证的过程及结果

/* 结合训练测试验证过程中的真实数据进行描述, 模型的特点和优势 */

4.3 模型的预测效果分析及改进建议

/* 结合视频写一些高端的东西 */

5 模型训练与验证

训练集: 测试集: 交叉验证集 =4:3:3

6 模型预测效果分析与改进方向

7 源码与文档