

doi:10.19920/j.cnki.jmsc.2026.07.015

基于情感分析的期货价格高频鲁棒预测方法^①

刘金培^{1,2}, 罗 瑞¹, 陈华友¹, 余乐安³, 贾 宁^{4*}, 朱 宁⁵

(1. 安徽大学大数据与统计学院, 合肥 230601; 2. 安徽大学商学院, 合肥 230601; 3. 四川大学商学院, 成都 610065; 4. 天津大学管理与经济学部, 天津 300072; 5. 中国科学技术大学管理学院, 合肥 230026)

摘要: 商品期货价格高频数据具有高度的非线性、非平稳性和高波动性等特征, 这对期货价格有效预测提出了巨大的挑战. 当前研究通常通过提取新闻文本中的情感极性信息对商品期货价格进行预测, 从而有效提升预测的时效性. 然而, 由于海量的新闻文本通常存在较多的噪声信息, 严重影响了预测的精度和稳定性. 为此, 本研究对期货价格相关新闻文本进行情感分析, 并提出一种加权双井势曲面支持向量回归(WDWPS-SVR)模型来对商品期货价格高频数据进行有效预测, 同时能显著提升预测的鲁棒性. 该模型基于双井势(DWP)函数, 在原始特征空间内利用四次双井势曲面对数据进行回归, 来刻画高频期货数据中的复杂数据特征. 并且通过权重函数计算各训练样本点的权重值, 以衡量样本点在拟合过程中的相对重要性, 从而滤除新闻文本噪声对参数估计的干扰. 将该模型应用于大连商品交易所 2019 年—2022 年大豆期货收盘价预测, 并从预测性能和鲁棒性两个方面与现有模型进行对比分析. 研究结果表明: 1) 通过赋权方法降低新闻文本噪声对回归过程产生的影响, 能够增强模型的鲁棒性, 进而可提升预测结果的稳定性. 2) WDWPS-SVR 模型具有较强的非线性特征学习能力, 能够实现对期货价格高频数据细节波动的有效拟合, 进而可提高预测方法的准确率. 3) 新闻文本情感变量能够有效解释期货价格的高频波动, 显著提高了预测时效性.

关键词: 期货价格预测; 无核支持向量回归; 鲁棒性; 情感分析; 噪声数据

中图分类号: F205 **文献标识码:** A **文章编号:** 1007-9807(2026)07-0221-18

0 引 言

自 2022 年以来, 受美联储加息和巴以冲突等多重因素影响, 兼具商品与货币双重属性的商品期货成为投资者进行风险分散的重要工具. 商品期货是以实物商品为标的物的期货合同, 因其在价格发现、资源配置和风险规避等方面的独特经济功能, 在宏观经济中发挥着举足轻重的作用^[1]. 商品期货具有资本流动性强、信息透明度高, 以及抵御通货膨胀和货币贬值等优势, 在国际

政治和经济环境剧烈动荡的时期仍能保持其战略投资价值, 因此被视为关键的风险对冲工具^[2]. 在此背景下, 商品期货的未来价格走势受到社会各界的高度关注, 并被用以实现资产配置平衡与稳健收益. 然而, 贸易融资套利、跨国资本流动和信息溢出效应等问题日益突出^[3], 加剧了商品期货价格的随机波动性和高度不确定性^[4]. 因此, 商品期货价格的精确预测已成为金融工程领域的关键问题, 而且对政府、企业和投资者具有重要意义^[5]. 政府可据此制定科学的监管政策和风

① 收稿日期: 2024-04-09; 修订日期: 2025-02-20.

基金项目: 国家自然科学基金资助项目(72471001; 72331007; 72071001; 72371001; 72122015); 安徽省高校优秀青年人才支持计划资助重点项目(gxyqZD2022001); 安徽省高校杰出青年基金资助项目(2023AH020009); 安徽省自然科学基金资助优秀青年项目(2408085Y035).

通讯作者: 贾 宁(1983—), 男, 山东德州人, 博士, 教授, 博士生导师. Email: jia_ning@tju.edu.cn

险预警体系,企业可规避价格波动风险、优化生产计划,投资者则能建立合理市场预期,制定多元化投资策略和风险管理计划。

在“互联网+”时代,市场参与者可以更加方便地获取商品期货市场的各类交易数据,从而显著提高了高频数据的可获取性和可操作性。作为期货市场微观结构的关键指标,高频数据蕴含丰富的日内交易信息,能够有效地反映商品期货交易价格的细节波动^[6]。同时,随着期货市场交易频率的加快和投资模式的多样化,高频数据在准确刻画期货价格短期风险特征方面的优势日益凸显,能够及时传递投资者交易行为变化的信号^[7, 8]。由此可见,在商品期货价格预测领域,高频数据能够为预测研究提供多维度的市场信息。然而,现有文献主要集中于期货价格低频数据的预测研究^[9],尚未充分考虑高频数据的实际研究价值^[6]。因此,亟需对商品期货价格高频数据的预测建模进行有效探究,以深入揭示期货市场的潜在波动规律。

传统商品期货价格预测模型主要依据线性或非线性假设,应用统计或人工智能方法分析期货价格与结构性关系变量之间的数量关系^[10-14]。与统计方法相比,人工智能方法能够更有效地提取期货价格的非线性和非平稳特征,但其对参数设置较为敏感,易导致“局部最优”或“过拟合”等问题。为了进一步提高预测精度,分解集成方法被广泛应用于商品期货价格预测^[15, 16]。例如, Fang 等^[15] 建立基于经验模态分解的方法,对原油价格序列进行分解后预测。Zeng 等^[16] 运用多种分解方法对农产品期货价格序列进行混合分解,提升了农产品期货价格的预测精度。分解集成方法基于“分而治之”的思想,将期货价格序列分解为若干更具规律性的子序列,以降低数据复杂度,然后对各子序列分别进行建模预测^[17]。因此,依据分解集成策略,本研究利用完全自适应噪声集合经验模态分解(complete ensemble empirical mode decomposition with adaptive noise, CEEMDAN)方法,将复杂的期货价格预测问题简化为若干个子问题。事实上,分解后子序列的建模预测是决定预测效果的关键。现有研究通常基于结构化数据对子序列进行预测。然而,结构化数据存在更新速度慢和信息过时等缺点,导致预测结果存在一定的滞

后性。

为了提高预测的及时性,随着自然语言处理(natural language processing, NLP)技术的快速发展,相关学者将研究视角转向非结构化文本数据^[18-20],提出基于文本情感分析的期货价格预测方法。例如, Wu 等^[18] 利用卷积神经网络(convolutional neural network, CNN)对新冠疫情相关新闻文本进行分析,探讨其对中国原油期货价格的影响。Li 等^[9] 提出一种基于双向长短期记忆模型(bidirectional long short-term memory, Bi-LSTM)文本情感分析的农产品期货价格预测框架。Li 等^[19] 探究了期货市场情感分值指标与原油期货价格波动的关系,并将情感分值用于对原油期货价格进行预测。上述研究表明,此类预测模型通过对文本数据的特征提取、情感分类以及情感极性研究等处理过程,能够将财经新闻文本量化为与商品期货市场相关的情感变量,从而提高预测信息的时效性。然而,由于海量的新闻文本存在较多的冗余信息,导致提取出的情感变量在反映投资者情感倾向或情感强度的变化时存在误差,从而使得情感变量含有大量噪声^[21],严重影响了预测的稳定性。因此,如何提出一种期货价格高频数据鲁棒预测方法,既具有较高的精确度和时效性,也能有效减少新闻文本噪声对预测结果的干扰,实现有效的鲁棒性预测,是一个亟待解决的关键科学问题。

为了提升预测的鲁棒性, Luo 等^[22] 研究表明支持向量回归(support vector regression, SVR)不易受噪声信号干扰,相较于其他预测模型具有更强的鲁棒性。其原因在于 SVR 模型的稀疏性特征,即回归曲线主要由少数支持向量决定,降低了异常值和噪声对其影响程度。在此背景下,核支持向量回归(kernel-based support vector regression, KSVR)模型被广泛应用于非线性预测^[23, 24, 27]。然而,核函数及其参数的选择缺乏通用的规则,导致回归拟合效果受到干扰。为了克服 KSVR 模型的缺点, Luo 等^[22] 和 Ye 等^[25] 提出了无核的二次曲面支持向量回归(quadratic surface support vector regression, QSSVR)模型。该模型通过在原始特征空间直接生成非线性二次曲面,有效拟合期货价格序列的周期性波动和趋势变化。然而,它却难以有效刻画商品期货价格高频数据中的高度非线性

关系^[26]. 双井势(double-well potential, DWP)函数是一种四次多项式函数,相较于二次函数和三次函数,其具有较强的非线性拟合能力^[26]. 因此,本研究提出双井势曲面支持向量回归(double-well potential surface SVR, DWPS-SVR)预测模型,以利用DWP函数来提取商品期货价格高频数据的复杂数据特征.

综上所述,情感极性作为输入特征能够提高期货价格高频数据预测的准确性和及时性. 然而,新闻文本中的冗余信息导致情感极性含有大量噪声,严重影响了预测的稳定性. QSSVR模型对含有噪声的数据回归具有较强的鲁棒性,但难以有效拟合期货高频数据中的高度非线性关系. 因此,如何提出适用于复杂数据特征提取的强鲁棒性预测模型,以实现商品期货价格高频数据预测精确度和稳定性的双重提升,是一个极具挑战性的问题.

针对上述问题,基于“数据特征驱动建模”的思想,本研究构建了加权DWPS-SVR(weighted DWPS-SVR, WDWPS-SVR)模型,进而提出基于新闻文本情感分析的鲁棒预测方法,从微观层面对商品期货价格高频数据进行预测. 首先,基于Bi-LSTM模型对期货新闻文本进行情感分析,以反映投资者情绪对期货价格的影响. 其次,利用CEEMDAN方法对商品期货价格序列和投资者情绪变量进行多尺度分解. 然后,构建WDWPS-SVR模型,同时证明了模型解的存在性和唯一性. 最后,利用WDWPS-SVR模型对分解得到的各子序列进行鲁棒预测,进而集成得到最终预测结果.

本研究的主要贡献如下:1)为了提升商品期货价格预测结果的精确度和稳定性,提出基于新闻文本情感分析的期货价格高频数据鲁棒预测方法. 通过将新闻文本情感分析与WDWPS-SVR模型相结合,该方法能够准确刻画商品期货价格的高频波动,并且可以降低数据噪声对回归过程的干扰. 2)为了对商品期货价格高频数据的细节非线性波动进行有效拟合,构建WDWPS-SVR预测模型. 区别于现有SVR模型,该模型一方面通过生成以四次双井势曲面为核心的“回归器”,能够有效刻画高频期货数据的高度非线性数据特征. 另一方面,通过设计权重函数以评估输入数据点

的不同重要程度,进一步提升了模型的鲁棒性. 此外,本研究对WDWPS-SVR模型的数理性质进行深入研究,证明了模型解的存在性和唯一性. 3)本研究为将新闻文本情感分析应用于高频期货价格数据预测提供了新的思路. 通过对比分析,验证了新闻文本数据对期货价格波动具有显著的解释能力和建模能力,能够提高预测效果的时效性,也可为探究其他非结构化数据驱动的高频数据预测问题提供启发.

1 模型构建

1.1 新闻文本情感分析

相较于传统的结构化数据,新闻文本、在线评论、推特(Twitter)和论坛等非结构化文本数据具有更强的时效性,能够为预测提供与商品期货相关的实时事件信息. 由于新闻文本的写作方式相较于其他网络文本更加规范,而且权威新闻网站提供的新闻通常由领域内专业的记者或专家所撰写,其对相关事件发表的观点具有更高的可信度和更广泛的影响力^[31]. 因此,本研究选择新闻文本作为非结构化数据源,并利用基于机器学习的情感分析方法提取新闻文本潜在的情感极性,以作为期货预测的情感变量. 基于机器学习的情感分析方法主要包括三个步骤:新闻文本预处理、新闻文本向量化表征和情感语义解析与量化,其具体框架流程如图1所示.

步骤1 新闻文本预处理. 首先,使用网络爬虫技术从财经网站收集与商品期货相关的新闻文本及其时间. 然后,对文本进行数据清洗,并基于jieba分词器采用精准模式进行分词,最后去除停用词.

步骤2 新闻文本向量化表征. 首先,采用零填充策略统一文本长度,以最大字符长度作为标准维度,对短文本进行零值填充. 进而,基于Word2Vec模型将离散词汇映射至预定义的向量空间,生成具有高维语义特征的词汇向量,最终构建维度规整的句子矩阵. 该过程将非结构化文本转化为数值型张量结构,为情感计算任务提供数据基础.

步骤3 情感语义解析与量化. 首先,基于百

度 Senta 开源语料库^[9]构建迁移学习框架,该语料库包含大规模细粒度情感标注语料,通过领域自适应预训练优化 Bi-LSTM 模型的语义感知能

力. 然后,利用微调后的 Bi-LSTM 模型分析新闻文本的情感强度,得到情感分值,以建立文本情感信号与市场波动间的映射关系.

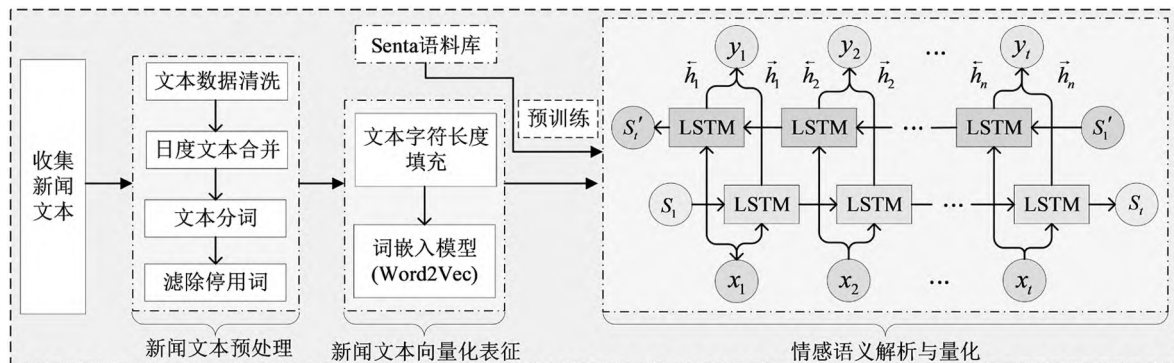


图 1 基于机器学习的情感分析方法流程图

Fig. 1 Machine learning-based sentiment analysis method flowchart

1.2 加权双井势曲面支持向量回归模型构建

针对新闻文本情感特征提取中的语义噪声干扰以及高频期货价格序列的多重非线性和非平稳

特征,本研究提出了 WDWPS-SVR 预测模型,并探讨其数理性质. 部分数学符号的含义如表 1 所示.

表 1 数学符号含义

Table 1 The meaning of mathematical notations

数学符号	含义
S^n	n 阶的实对称矩阵
e_n	元素均为 1 的 n 维列向量, 即 $e_n = (1, 1, \dots, 1)^T \in \mathbb{R}^n$.
I_n	n 阶单位矩阵
G^i	矩阵 G 的第 i 列
$0_{m \times n}$	元素均为 0 的 $m \times n$ 阶矩阵
$\text{vec}(A)$	矩阵 $A = (a_{ij})_{n \times n}$ 向量化, 即 $\text{vec}(A) = (a_{11}, \dots, a_{1n}, a_{21}, \dots, a_{2n}, \dots, a_{n1}, \dots, a_{nn})^T \in \mathbb{R}^{n^2}$
$\text{hvec}(A)$	将实对称矩阵 $A \in \mathbb{S}^n$ 的上三角元素转化为向量形式, $\text{hvec}(A) = (a_{11}, \dots, a_{1n}, a_{22}, \dots, a_{2n}, \dots, a_{n-1, n-1}, a_{n-1, n}, a_{nn})^T \in \mathbb{R}^{\frac{n(n+1)}{2}}$
$\text{hlvec}(a)$	向量 $a = (a_1, \dots, a_n)^T \in \mathbb{R}^n$ 的元素交叉项生成的新向量, $\text{hlvec}(a) = \left(\frac{1}{2}a_1a_1, a_1a_2, \dots, a_1a_n, \frac{1}{2}a_2a_2, a_2a_3, \dots, a_2a_n, \dots, \frac{1}{2}a_na_n \right)^T \in \mathbb{R}^{\frac{n(n+1)}{2}}$
$\text{lvec}(a)$	向量 $a = (a_1, \dots, a_n)^T \in \mathbb{R}^n$ 的元素两两交叉相乘生成新向量, $\text{lvec}(a) = \left(\frac{1}{2}a_1a_1, \frac{1}{2}a_1a_2, \dots, \frac{1}{2}a_1a_n, \dots, \frac{1}{2}a_na_1, \frac{1}{2}a_na_2, \dots, \frac{1}{2}a_na_n \right)^T \in \mathbb{R}^{n^2}$
$\text{llvec}(a)$	$\text{llvec}(a) = (\text{lvec}(a))^T, a^T)^T \in \mathbb{R}^{n(n+1)}$
$\text{mat}(a)$	向量 a 生成的矩阵, 其中 $\text{mat}(a) = \begin{pmatrix} a^T & 0_{1 \times n} & \dots & 0_{1 \times n} \\ 0_{1 \times n} & a^T & \dots & 0_{1 \times n} \\ \dots & \dots & \dots & \dots \\ 0_{1 \times n} & 0_{1 \times n} & \dots & a^T \end{pmatrix} \in \mathbb{R}^{n \times n^2}$
$\text{lmat}(a)$	$\text{lmat}(a) = (\text{mat}(a), I_n) \in \mathbb{R}^{n \times n(n+1)}$
$\ a\ _2$	向量 a 的 ℓ_2 范数, $\ a\ _2 = \sqrt{a^T a}$

1.2.1 双井势曲面支持向量回归模型

对于具有高度非线性、非平稳性和高波动性

特征的高频数据,为了有效学习输入特征值与目标值之间的复杂关系,在 KSVR 和 QSSVR 模型的

基础上,构建基于四次双井势函数的 DWPS-SVR 模型. 基于 DWP 函数的特点,将 $F(\mathbf{x}) + \varepsilon$ 与 $F(\mathbf{x}) - \varepsilon$ 的区域带间距转化为两个二次函数生成的区域带的间距和,如图 2 所示. 令 $y = F(\mathbf{x}) = F_1(\mathbf{z}) + F_2(\mathbf{x})$, 其中

$$\begin{aligned} F_1(\mathbf{z}) &= \frac{1}{2} \mathbf{z}^T \mathbf{W} \mathbf{z} \\ F_2(\mathbf{x}) &= \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + q \\ \mathbf{z} &= ((\text{hlvec}(\mathbf{x}))^T, \mathbf{x}^T, 1)^T \end{aligned} \quad (1)$$

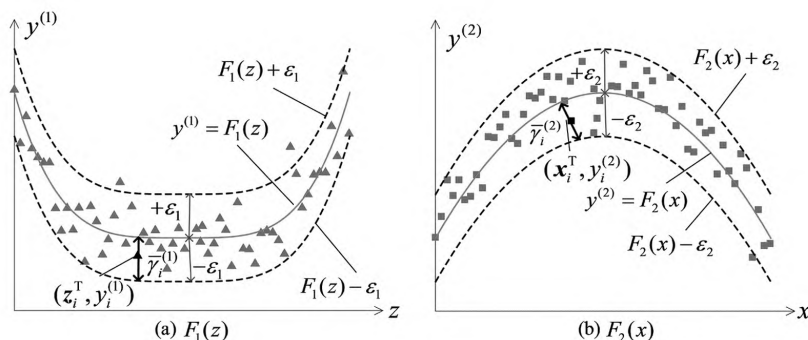


图 2 DWPS-SVR 模型的几何边距

Fig. 2 Geometrical margin of DWPS-SVR model

根据相对几何边距的定义^[22, 30], $y^{(1)} = F_1(\mathbf{z})$ 在 $(\mathbf{z}_i^T, y_i^{(1)})$ 处的相对几何边距为 $\varepsilon_1 / \|(\mathbf{W} \mathbf{z}_i, -1)\|_2$, $y^{(2)} = F_2(\mathbf{x})$ 在 $(\mathbf{x}_i^T, y_i^{(2)})$ 处的相对几何边距为 $\varepsilon_2 / \|(\mathbf{A} \mathbf{x}_i + \mathbf{b}, -1)\|_2$. 因此,建立初步的 DWPS-SVR 模型如下

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{A}, \mathbf{b}, q, \xi} & \sum_{i=1}^n \frac{1}{2} \|\mathbf{W} \mathbf{z}_i\|_2^2 + k_1 \sum_{i=1}^n \frac{1}{2} \|\mathbf{A} \mathbf{x}_i + \mathbf{b}\|_2^2 + k_2 \sum_{i=1}^n \xi_i \\ \text{s. t.} & \left| y_i - \left(\frac{1}{2} \mathbf{z}_i^T \mathbf{W} \mathbf{z}_i + \frac{1}{2} \mathbf{x}_i^T \mathbf{A} \mathbf{x}_i + \mathbf{b}^T \mathbf{x}_i + q \right) \right| \leq \varepsilon + \xi_i \\ & \xi_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (2)$$

其中 $k_1 > 0, k_2 > 0$ 和 $\varepsilon > 0$ 为自定义参数, $\mathbf{W} \in \mathbb{S}^{\frac{1}{2}(m+1)(m+2)}$, $\mathbf{A} \in \mathbb{R}^{m \times m}$, $\mathbf{b} \in \mathbb{R}^m$ 和 $q \in \mathbb{R}$ 为未知参数. 在模型(2)中, $\sum_{i=1}^n \frac{1}{2} \|\mathbf{W} \mathbf{z}_i\|_2^2 + k_1 \sum_{i=1}^n \frac{1}{2} \|\mathbf{A} \mathbf{x}_i + \mathbf{b}\|_2^2$ 为正则项,体现了结构风险最小化原则,其目的为提高模型的泛化能力,防止发生过拟合. 系数 k_1 可以调节回归函数中不同频率数据的重要性程度. $k_2 \sum_{i=1}^n \xi_i$ 最小化的目的是减少区域带之外样本点的回归误差损失.

1.2.2 加权双井势曲面支持向量回归模型

从新闻文本中提取投资者情绪变化和心理预期等实时信息时,海量新闻文本存在的冗余信息将导致情感极性数据包含大量的噪声,进而会对模型输入变量造成数据污染. 在此情况下, DWPS-SVR 模型的拟合精度会受到一定的影响. 为了解决该问题,进一步对各训练样本点进行赋权处理,

以反映数据间重要性程度的差异.

对于任意样本数据 $(\mathbf{x}_i^T, y_i)$ ($i = 1, 2, \dots, n$),先利用 QSSVR 进行拟合,得到回归函数 $\hat{y} = \frac{1}{2} \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$,再计算各数据点的回归误差绝对值

$r_i = |y_i - \hat{y}_i| = \left| y_i - \frac{1}{2} \mathbf{x}_i^T \mathbf{W} \mathbf{x}_i + \mathbf{b}^T \mathbf{x}_i + c \right|$ ($i = 1, 2, \dots, n$). 进一步,令 $\bar{r}$ 为 r_i 的中位数, M_e 为 $|r_i - \bar{r}|$ 的中位数,则可以根据式(3)计算样本数据 $(\mathbf{x}_i^T, y_i)$ 对应的权重 λ_i ,如下

$$\lambda_i = e^{-\mu_i}, i = 1, 2, \dots, n \quad (3)$$

其中 $\mu_i = |r_i - \bar{r}| / M_e$. 实际上,若某训练样本点存在数据噪声问题,则该点的 r_i 可能偏离于数据集整体的误差绝对值中位数,导致 μ_i 值增加,由式(3)可知,该点对应的权重 λ_i 则会变小.

为降低新闻文本中噪声的干扰,在模型(2)的基础上,进一步考虑样本中不同数据的权重差异,提出了 WDWPS-SVR 模型,如下

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{A}, \mathbf{b}, q, \xi} & \sum_{i=1}^n \frac{1}{2} \lambda_i \|\mathbf{W} \mathbf{z}_i\|_2^2 + k_1 \sum_{i=1}^n \lambda_i \frac{1}{2} \|\mathbf{A} \mathbf{x}_i + \mathbf{b}\|_2^2 + k_2 \sum_{i=1}^n \xi_i \\ \text{s. t.} & \left| y_i - \left(\frac{1}{2} \mathbf{z}_i^T \mathbf{W} \mathbf{z}_i + \frac{1}{2} \mathbf{x}_i^T \mathbf{A} \mathbf{x}_i + \mathbf{b}^T \mathbf{x}_i + q \right) \right| \leq \varepsilon + \xi_i \\ & \xi_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (4)$$

由模型(4)可知,若训练数据 $(\mathbf{x}_i^T, y_i)$ 的回归值 $\hat{y}_i$ 受到了存在噪声污染的数据的干扰,则该数据点对参数估计的贡献度较低. 根据式(3),能够

对该数据点赋予较小的权重,以降低其正则项 $\frac{1}{2} \|Wz_i\|_2^2 + k_1 \frac{1}{2} \|Ax_i + b\|_2^2$ 在目标函数中的重要性. 由此可见,WDWPS-SVR 模型通过对训练样本点进行赋权处理,能够减少数据噪声对回归过程的影响,进而增强模型鲁棒性.

1.2.3 模型的数理性质

在对 WDWPS-SVR 模型的形式进行化简的基础上,探讨其数理性质,并通过对偶问题转换,对该优化模型进行求解. 将 WDWPS-SVR 模型化简为如下形式

$$\begin{aligned} \min_{w_{WA}, q, \xi} \quad & \frac{1}{2} w_{WA}^T X w_{WA} + k_2 e_n^T \xi \\ \text{s. t.} \quad & |y - (R_{RS}^T w_{WA} + q)| \leq \varepsilon + \xi \\ & \xi \geq 0 \end{aligned} \quad (5)$$

其中 w_{WA}, q 以及 ξ 为待求的决策变量, $\xi = (\xi_1, \xi_2, \dots, \xi_n)^T \in \mathbb{R}^n$, $q = qe_n = (q, q, \dots, q)^T \in \mathbb{R}^n$, $w_W = \text{vec}(W)$, $w_A = ((\text{vec}(A))^T, b^T)^T$, $w_{WA} = (w_W^T, w_A^T)^T$. $k_2 > 0$ 和 $\varepsilon = \varepsilon e_n$ 为模型的给定参数.

在模型(5)中, $y \in \mathbb{R}^n$, $R_{RS} \in \mathbb{R}^{(l^2+m(m+1)) \times n}$, $X \in \mathbb{S}^{l^2+m(m+1)}$ 为已知系数. 针对样本数据 (x_i^T, y_i) 和权重 λ_i , 可知 $R = (\text{lvec}(z_1), \text{lvec}(z_2), \dots, \text{lvec}(z_n))$, $S = (\text{llvec}(x_1), \text{llvec}(x_2), \dots, \text{llvec}(x_n))$, $y = (y_1, y_2, \dots, y_n)^T$, $R_{RS} = (R^T, S^T)^T$. 令 $X_W = (\sqrt{\lambda_1} (\text{mat}(z_1))^T, \sqrt{\lambda_2} (\text{mat}(z_2))^T, \dots, \sqrt{\lambda_n} (\text{mat}(z_n))^T)^T$, $X_A = (\sqrt{\lambda_1} (\text{lmat}(x_1))^T, \sqrt{\lambda_2} (\text{lmat}(x_2))^T, \dots, \sqrt{\lambda_n} (\text{lmat}(x_n))^T)^T$, 则 $X = \begin{bmatrix} X_W^T X_W & 0_{l^2 \times m(m+1)} \\ 0_{m(m+1) \times l^2} & k_1 X_A^T X_A \end{bmatrix}$, 且

$k_1 > 0$, 即对于任意不为 0 的向量 $w_{WA} = (w_W^T, w_A^T)^T$, 有 $w_{WA}^T X w_{WA} = w_W^T X_W^T X_W w_W + k_1 w_A^T X_A^T X_A w_A \geq 0$, 可知 X 为半正定矩阵. 因此,WDWPS-SVR 模型是带有线性约束的凸二次规划问题.

定理 1 对于任意训练数据集 $\{(x_i^T, y_i), i = 1, 2, \dots, n\}$ 和 $k_2 > 0$, WDWPS-SVR 模型的目标函数值有界,且存在最优解.

证明 对任意 $((\hat{w}_{WA}^T), \hat{q}) \in \mathbb{R}^{l^2+m(m+1)+1}$, 令 $\tilde{\xi}_i = \max_{i=1,2,\dots,n} \{0, |y_i - ((R_{RS}^i)^T \hat{w}_{WA} + \hat{q})| - \varepsilon\}$, 则 $|y_i - ((R_{RS}^i)^T \hat{w}_{WA} + \hat{q})| - \varepsilon \leq \tilde{\xi}_i, \tilde{\xi}_i \geq 0$, 即满足约束条件,可见, $\hat{w}_{WA}, \hat{q}$ 和 $\tilde{\xi}_i$ 是规划问题的可行解.

由于目标函数具有连续性,可行域为线性不等式定义的闭凸集,且对于任意参数 $w_{WA}, \xi_i \geq 0$, 目标

函数值 $\frac{1}{2} w_{WA}^T X w_{WA} + k_2 \sum_{i=1}^n \xi_i \geq 0$, 所以 WDWPS-SVR 模型在有限的目标值中存在最优解, 证毕.

定理 2 对于任意训练数据集 $\{(x_i^T, y_i), i = 1, 2, \dots, n\}$, 如果 X 是正定矩阵, 则 WDWPS-SVR 模型中关于决策变量 w_{WA} 的最优解具有唯一性.

证明 设 $(\hat{w}_{WA}, \hat{q}, \hat{\xi})$ 和 $(\bar{w}_{WA}, \bar{q}, \bar{\xi})$ 均为 WDWPS-SVR 模型的最优解, 且 $\hat{w}_{WA} \neq \bar{w}_{WA}$. 在凸优化问题中, 对于 $\forall \beta \in (0, 1)$, $(w_{WA}^*, q^*, \xi^*) \triangleq \beta(\hat{w}_{WA}, \hat{q}, \hat{\xi}) + (1 - \beta)(\bar{w}_{WA}, \bar{q}, \bar{\xi})$ 也为优化问题的可行解. 因此

$$\frac{1}{2} (w_{WA}^*)^T X w_{WA}^* + k_2 e_n^T \xi^* \geq \frac{1}{2} \hat{w}_{WA}^T X \hat{w}_{WA} + k_2 e_n^T \hat{\xi} \quad (6)$$

$$\frac{1}{2} (w_{WA}^*)^T X w_{WA}^* + k_2 e_n^T \xi^* \geq \frac{1}{2} \bar{w}_{WA}^T X \bar{w}_{WA} + k_2 e_n^T \bar{\xi} \quad (7)$$

将式(6)两端乘以 β , 式(7)两端乘以 $(1 - \beta)$, 进而将两者相加得到

$$\begin{aligned} \frac{1}{2} (w_{WA}^*)^T X w_{WA}^* + k_2 e_n^T \xi^* \geq \\ \frac{\beta}{2} \hat{w}_{WA}^T X \hat{w}_{WA} + \frac{(1 - \beta)}{2} \bar{w}_{WA}^T X \bar{w}_{WA} + k_2 e_n^T (\beta \hat{\xi} + (1 - \beta) \bar{\xi}) \end{aligned} \quad (8)$$

根据 $w_{WA}^* = \beta \hat{w}_{WA} + (1 - \beta) \bar{w}_{WA}, \xi^* = \beta \hat{\xi} + (1 - \beta) \bar{\xi}$, 式(8)可转化为 $\frac{(1 - \beta)}{2} \bar{w}_{WA}^T X \bar{w}_{WA} + \frac{\beta}{2} \hat{w}_{WA}^T X \hat{w}_{WA} \leq \frac{1}{2} (\beta \hat{w}_{WA} + (1 - \beta) \bar{w}_{WA})^T X (\beta \hat{w}_{WA} + (1 - \beta) \bar{w}_{WA})$. 进一步, 得到 $\beta(1 - \beta)(\hat{w}_{WA}^T X \hat{w}_{WA} - 2\hat{w}_{WA}^T X \bar{w}_{WA} + \bar{w}_{WA}^T X \bar{w}_{WA}) \leq 0$. 所以, 有 $\beta(1 - \beta)(\hat{w}_{WA} - \bar{w}_{WA})^T X (\hat{w}_{WA} - \bar{w}_{WA}) \leq 0$. 当 X 为正定矩阵时, $\hat{w}_{WA} = \bar{w}_{WA}$, 与原假设相矛盾. 因此, 关于变量 w_{WA} 的最优解具有唯一性, 证毕.

定理 3 给定训练数据集 $\{(x_i^T, y_i), i = 1, 2, \dots, n\}$, 对于任意最优解 (w_{WA}, q, ξ) , 若 X 是正定矩阵, 则存在常数 $\underline{q}$ 和 $\bar{q}$, 使得 $q \in [\underline{q}, \bar{q}]$.

证明 令 (w_{WA}^*, q^*, ξ^*) 是 WDWPS-SVR 模

型的一组最优解, 则对于任意最优解 $(\mathbf{w}_{WA}, q, \xi)$,

$$\text{等式 } \frac{1}{2} \mathbf{w}_{WA}^T \mathbf{X} \mathbf{w}_{WA} + k_2 \sum_{i=1}^n \xi_i = \frac{1}{2} (\mathbf{w}_{WA}^*)^T \mathbf{X} \mathbf{w}_{WA}^* +$$

$k_2 \sum_{i=1}^n \xi_i^*$ 均成立. 根据定理 2, 当 $\mathbf{X}$ 是正定矩阵时,

$$\mathbf{w}_{WA}^* = \mathbf{w}_{WA}, \text{ 由此可得 } K = \sum_{i=1}^n \xi_i = \sum_{i=1}^n \xi_i^*, 0 \leq \xi_i \leq K.$$

依据该模型的约束条件, 得到 $y_i - (\mathbf{R}_{RS}^i)^T \mathbf{w}_{WA}^* - \varepsilon - \xi_i^* \leq q \leq y_i - (\mathbf{R}_{RS}^i)^T \mathbf{w}_{WA}^* + \varepsilon + \xi_i^*$, 由于 $0 \leq \xi_i \leq K$, 所以 $y_i - (\mathbf{R}_{RS}^i)^T \mathbf{w}_{WA}^* - \varepsilon - K \leq q \leq y_i - (\mathbf{R}_{RS}^i)^T \mathbf{w}_{WA}^* + \varepsilon + K$. 令 $\underline{q} = \min_{i=1,2,\dots,n} (y_i - \varepsilon - (\mathbf{R}_{RS}^i)^T \mathbf{w}_{WA}^* - K)$, $\bar{q} = \max_{i=1,2,\dots,n} (y_i - (\mathbf{R}_{RS}^i)^T \mathbf{w}_{WA}^* + \varepsilon + K)$, 则 $\underline{q} \leq q \leq \bar{q}$,

证毕.

2 期货价格高频数据预测方法

新闻文本能够为期货价格高频数据预测提供实时信息. 然而, 海量新闻文本数据含有大量的噪声, 并且高频期货数据具有高度非线性和非平稳性等特征. 这些预测因子复杂的本质特性和模式特征导致现有预测方法难以实现预测精确度和稳定性的双重提升. 因此, 依据“数据特征驱动建模”的思想, 在 1.1 节中详细阐述了新闻文本情感分析方法. 进而, 在 1.2 节中构建了具有强鲁棒性的 WDWPS-SVR 模型. 在此基础上, 本节进一步提出商品期货价格高频数据鲁棒预测方法, 如图 3 所示.

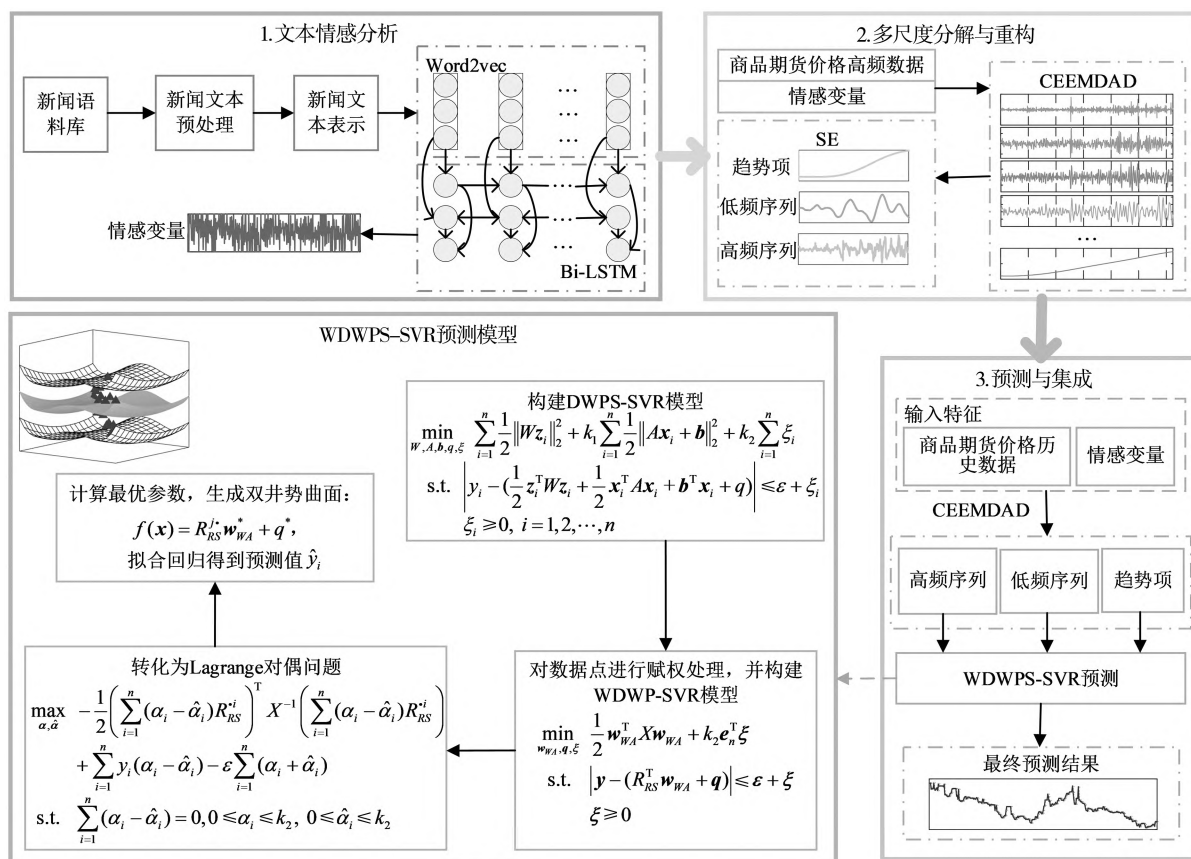


图 3 预测方法流程图

Fig. 3 Forecasting method flowchart

步骤 1 文本情感分析. 首先, 从财经网站爬取与商品期货相关的在线新闻, 并对其数据进行清洗和分词等数据预处理, 以生成低噪声的新闻语料库. 然后, 结合金融新闻词向量, 对新闻语料库中的文本进行特征提取, 将其转换为数值型数

据. 最后, 利用 Bi-LSTM 对新闻文本进行情感分析, 以得到情感变量.

步骤 2 多尺度分解与重构. 首先, 利用 CEEMDAN 方法分别将商品期货价格序列和情感变量分解为多个更加平稳的子序列. 然后, 运用样

本熵(sample entropy, SE)方法对若干子序列进行重构处理,分别得到高频率序列、低频率序列和趋势项,以减少预测误差累计.具体而言,SE方法通过计算信号中新模式生成的概率,衡量商品期货价格序列的复杂性^[40].新模式生成的概率越高,样本熵值越大,则序列的复杂性越强,反之亦然.

步骤3 预测与集成.将情感分值变量以及滞后三期的商品期货收盘价 y_{t-1} , y_{t-2} 和 y_{t-3} 作为预测模型输入值,当期商品期货收盘价 y_t 作为输出值.首先,根据权重函数式(3)对各训练样本点进行赋权.然后,基于样本数据及赋权结果构建WDWPS-SVR模型,并进行求解,以降低情感变量包含的噪声对估参过程的干扰.进而,根据WDWPS-SVR模型的求解结果分别对高频率序列、低频率序列和趋势项进行预测.最后,将各子序列的预测值相加集成,得到商品期货价格的最终预测

结果.

3 实证分析

3.1 数据来源与处理

大连商品交易所(Dalian Commodity Exchange, DCE)作为亚洲最大的大豆期货交易场所,其大豆期货的交易量和成交额均为世界领先.因此,本研究选择的DCE大豆期货价格具有良好的代表性.本研究主要收集了2017年1月4日—2024年3月20日期间DCE网站(<http://www.dce.com.cn/>)披露的日度大豆期货收盘价共1 751个数据作为研究对象(剔除周末和节假日).其中,将样本集前80%的1 401个数据作为训练集,后20%的350个数据作为测试集,具体划分情况如图4所示.

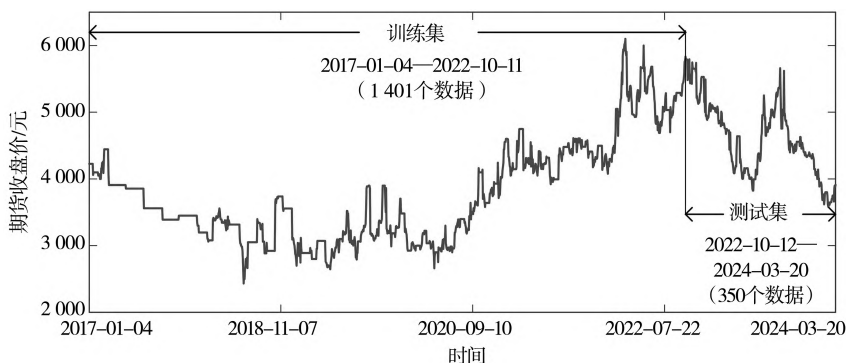


图4 大豆期货价格序列

Fig. 4 Soybean futures price series

本研究将新闻标题作为非结构化影响因素,并利用Bi-LSTM模型将其量化为情感分值,以作为大豆期货价格预测的情感变量.具体情感分析过程如下.

1) 首先,采用网络爬虫技术从大连商品交易所、和讯网、东方财富网和新浪财经等国内主流金融网站上收集2017年1月4日—2024年3月20日期间所发布的6 074条新闻标题.为了减少无关新闻文本信息的干扰,本研究以“大豆期货”和“大豆”等相关的大豆期货价格关键词作为索引,仅保留包含索引词汇的新闻标题,由此获取了5 094条新闻标题,平均每天约3条.然后,对所收集的文本进行数据清洗和分词处理,生成新闻语料库.

2) 以中文词向量语料库(<https://github.com/Embedding/Chinese-Word-Vectors>)

所涉及的金融模块词向量作为基准,利用Word2Vec模型,将每个词语映射到一个低维向量空间中,得到相应的稠密向量表示.进而,将新闻标题转化为词汇向量构成的句子矩阵.

3) 首先,利用百度开源的权威情感语料库Senta对Bi-LSTM模型进行预训练,以提高其对大豆期货新闻文本中潜在情感特征的学习能力.基于Bi-LSTM模型的情感分析结果的取值范围为 $[0, 1]$,数值越大,情感倾向越积极.然后,利用经过训练的Bi-LSTM模型,对新闻标题进行情感分析,得到情感分值.

为检验大豆期货价格序列和情感变量序列的数据特征,采用描述性统计、ADF检验、BDS检验三种方法对其进行统计分析.分析结果表明,大豆

期货价格序列和情感变量序列均呈右偏非正态分布,并且均具有非线性和非平稳数据特征.为检验将情感变量作为大豆期货价格预测的输入值的可行性,本研究对大豆期货价格数据与情感变量数据进行 Johansen 协整检验.协整检验结果显示,经过一阶差分处理后,迹统计量值为 10.275,超过了 5% 临界值 3.841.这意味着在 5% 的显著性水平上,拒绝了原假设“ $At\ most\ 1$ ”,说明大豆期货价格与情感变量之间至少存在一个长期协整关系.因此,从计量经济学的角度来看,将情感变量作为大豆期货价格预测的输入值具有可行性.

3.2 预测结果分析

根据 3.1 节的数据特征分析,可知大豆期货价格时间序列具有高度非线性和非平稳性特征.为降低大豆期货价格序列和情感分值变量的数据复杂度,利用 CEEMDAN 方法分别将两者分解为 10 个更具规律、更加平稳的子序列.然后,利用 SE 方法对各分量进行重构处理,阈值设置为 0.1 和 0.3.大豆期货价格序列的分解重构结果如图 5 所示,趋势项反映了 DCE 大豆期货收盘价的整体变化趋势,低频序列体现了大豆期货价格的季节性变化规律,高频序列是其他因素综合影响的结果.

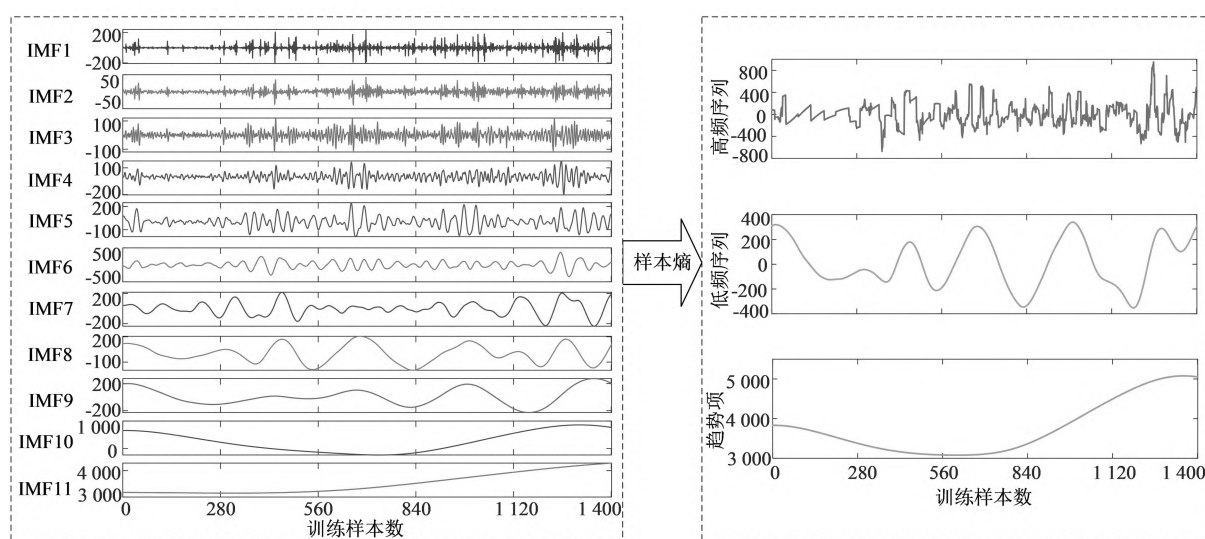


图5 大豆期货序列分解重构图

Fig. 5 Decomposition and reconstruction of soybean futures prices time series

在上述分解重构的基础上,利用构建的 WDWPS-SVR 模型分别对高频序列、低频序列和趋势项进行预测.然后,将各子序列预测值进行相加集成处理,得到大豆期货价格最终预

测结果.由图 6 可知,实际值与预测值之间的拟合度较高,说明本研究模型具有较高的预测精度,适用于基于新闻情感分析的商品期货价格预测.

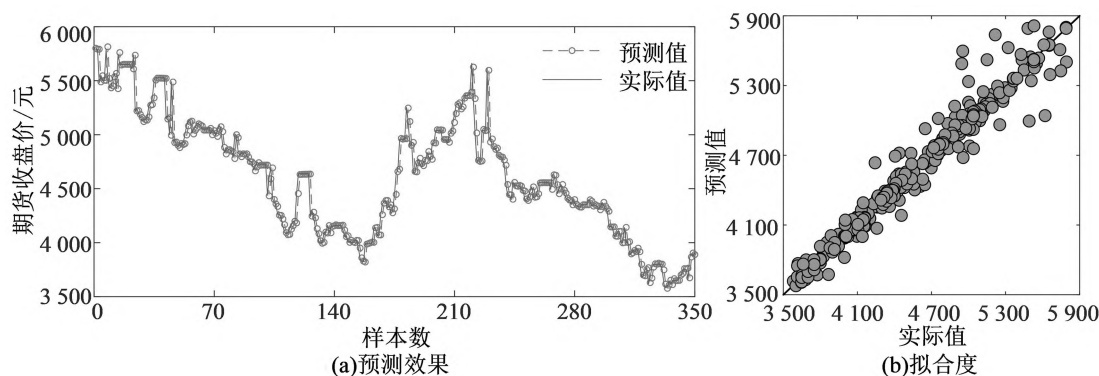


图6 预测效果对比图

Fig. 6 Comparison of forecasting performance

3.3 预测方法对比分析

本研究利用六种误差评价指标,即平均绝对误差(mean absolute error, MAE)、平均绝对百分比误差(mean absolute percentage error, MAPE)、加权绝对百分比误差(weighted absolute percentage error, WAPE)、对称平均绝对百分比误差(symmetric mean absolute percentage error, SMAPE)、绝对误差中位数(median of absolute percentage error, MdAPE)和 Diebold-Mariano (DM)检验作为实验结果的评价标准^[32-35],对三组实验进行对比分析.第一组实验的目的是检验 WDWPS-SVR 模型对期货价格高频数据的拟合能力.第二组实验旨在探究情感变量在期货价格预测中的作用.第三组实验旨在验证本研究预测方法在处理噪声数据时的鲁棒性.

3.3.1 模型高频数据预测性能检验

在本研究方法的基础上,不考虑情感变量的影响,将滞后三期的大豆期货价格作为模型的输入变量.进而,利用九种基准模型分别对大豆期货价格高频数据进行预测,以检验 WDWPS-SVR 模型的预测性能.基准模型分别为:QSSVR 模型^[22]、KSVR 模型^[28, 29]、迭代加权最小二乘法(iteratively re-weighted least squares, IRLS)^[36]、 L_1 正则化回归(L_1 regularized regression, L_1)模型^[36]、极端梯度提升(extreme gradient boosting, XGboost)

模型^[37]、整合移动平均自回归(autoregressive integrated moving average, ARIMA)模型^[10]、随机游走(random walk, RW)模型^[38]、三次指数平滑法(Holt-Winters, Holt)^[39]以及广义自回归条件异方差(generalized autoregressive conditional heteroskedasticity, GARCH)模型^[11].其中, KSVR 模型的核函数设定为高斯径向基函数, IRLS 模型的加权函数设置为“logistic”函数.对比结果如表 2 所示,可知 CEEMDAN-WDWPS-SVR 方法在五种指标上均表现出最优的预测精度,说明 WDWPS-SVR 模型能够有效提取大豆期货价格高频数据的复杂数据特征.为进一步检验上述模型的预测性能,以 CEEMDAN-WDWPS-SVR 作为目标方法,分别对其与其他五种方法两两进行 DM 检验.检验结果表明,在 10% 的显著性水平上,检验结果均拒绝原假设,即 WDWPS-SVR 模型的预测效果明显优于其他五种模型,验证了该模型的有效性和适用性.

在九种不同的高频数据预测对比方法中,相较于其他六种基于统计方法的预测结果,CEEMDAN-QSSVR、CEEMDAN-KSVR 和 CEEMDAN-XGboost 方法的预测误差较低.由此可见,传统统计方法不适用于大豆期货价格的预测,原因在于统计方法主要刻画序列的线性特征,而大豆期货价格序列具有非线性和非平稳性特征,使得统计方法难以达到预期的拟合效果.

表 2 模型预测性能检验结果对比

Table 2 Comparison of model forecasting performance test results

序号	预测方法	MAE	MAPE/%	WAPE	SMAPE/%	MdAPE/%
(1)	CEEMDAN-QSSVR	65.757	1.405	1.427	1.400	0.007
(2)	CEEMDAN-KSVR	65.603	1.402	1.424	1.398	0.008
(3)	CEEMDAN-IRLS	815.766	18.414	17.705	16.688	0.178
(4)	CEEMDAN- L_1	84.564	1.834	1.835	1.842	0.013
(5)	CEEMDAN-XGboost	71.320	1.545	1.548	1.539	0.010
(6)	CEEMDAN-ARIMA	73.231	1.565	1.589	1.560	0.010
(7)	CEEMDAN-RW	107.592	2.314	2.335	2.304	0.016
(8)	CEEMDAN-Holt	456.475	10.546	9.907	9.859	0.096
(9)	CEEMDAN-GARCH	304.865	6.265	6.617	6.555	0.060
(10)	CEEMDAN-WDWPS-SVR	63.222	1.345	1.372	1.341	0.007

3.3.2 情感变量解释能力检验

本节将基于 Bi-LSTM、CNN 和 SnowNLP 的情感分值分别作为 QSSVR、KSVR、IRLS 和 WDWPS-SVR 模型的辅助变量,对大豆期货价格高频数据进行预测.

由表 3 可知,依据不同文本情感分析方法,可以将对比预测方法划分为四类:不进行文本情感分析类、基于 Bi-LSTM 文本情感分析类、基于 CNN 文本情感分析类、和基于 SnowNLP 文本情感分析类.以情感分值作为输入变量的预测方法的

拟合精确度高于仅考虑历史价格的预测方法. 由此说明, 新闻文本等非结构化数据能够为大豆期货价格预测提供实时有效的增量信息, 改善了基于结构化数据预测模型存在的滞后性问题, 进而提高了预测的时效性. 需要注意的是, 基于情感分析的方法(14)和方法(19)的误差评价指标高于未考虑情感变量的方法(4), 且方法(17)的预测误差也高于方法(2), 原因在于情感变量含有大量的噪声, 从而影响了预测结果的稳定性. 在三种

基于不同文本情感分析的预测方法对比中, 与基于 Bi-LSTM 文本情感分析的预测方法相比, 基于 CNN 和 SnowNLP 文本情感分析的预测方法的预测准确性均较低. 由此可知, Bi-LSTM 可以依据上下文的长期依赖关系, 获取新闻文本的双重语义信息, 以更准确地识别新闻文本潜在反映的投资者情感极性. 另外, 在所有对比方法中, 本文预测方法(10)实现了最高的预测精度, 进一步说明本研究提出的期货价格高频数据预测方法具有较强的预测性能.

表 3 情感变量解释能力检验结果对比

Table 3 Comparison of test results for explanatory power of sentiment variables

情感分析方法	预测方法	MAE	MAPE/%	WAPE	SMAPE/%	MdAPE/%
不进行文本情感分析	(1) CEEMDAN-QSSVR	65.757	1.405	1.427	1.400	0.007
	(2) CEEMDAN-KSVR	65.603	1.402	1.424	1.398	0.008
	(3) CEEMDAN-XGboost	71.320	1.545	1.548	1.539	0.010
	(4) CEEMDAN- L_1	84.564	1.834	1.835	1.842	0.013
	(5) CEEMDAN-WDWPS-SVR	63.222	1.345	1.372	1.341	0.007
基于 Bi-LSTM 文本情感分析	(6) CEEMDAN-BiLSTM-QSSVR	62.297	1.327	1.352	1.322	0.006
	(7) CEEMDAN-BiLSTM-KSVR	63.214	1.349	1.372	1.345	0.007
	(8) CEEMDAN-BiLSTM-XGboost	69.560	1.497	1.510	1.493	0.009
	(9) CEEMDAN-BiLSTM- L_1	83.472	1.815	1.812	1.822	0.014
	(10) CEEMDAN-BiLSTM-WDWPS-SVR (本文预测方法)	61.484	1.310	1.334	1.306	0.006
基于 CNN 文本情感分析	(11) CEEMDAN-CNN-QSSVR	62.674	1.333	1.360	1.329	0.007
	(12) CEEMDAN-CNN-KSVR	64.662	1.376	1.403	1.374	0.008
	(13) CEEMDAN-CNN-XGboost	69.810	1.506	1.515	1.501	0.009
	(14) CEEMDAN-CNN- L_1	85.548	1.863	1.857	1.872	0.014
	(15) CEEMDAN-CNN-WDWPS-SVR	61.879	1.317	1.343	1.313	0.007
基于 SnowNLP 文本情感分析	(16) CEEMDAN-SN-QSSVR	63.611	1.354	1.381	1.351	0.007
	(17) CEEMDAN-SN-KSVR	66.328	1.415	1.440	1.414	0.008
	(18) CEEMDAN-SN-XGboost	70.102	1.512	1.521	1.507	0.009
	(19) CEEMDAN-SN- L_1	86.251	1.884	1.872	1.893	0.014
	(20) CEEMDAN-SN-WDWPS-SVR	62.257	1.328	1.351	1.324	0.007

3.3.3 本文预测方法鲁棒性检验

为了验证本文预测方法的鲁棒性, 对三组情感变量进行数据污染处理, 以模拟情感极性数据含有不同程度噪声信号的情况. 具体步骤: 首先, 计算情感变量序列的均值和标准差. 然后, 根据不同比率生成服从不同正态分布的噪声数据. 其次, 将各扰动项加入原始数据, 生成不同训练数据集, 作为预测模型的辅助变量. 本节利用 IRLS、 L_1 、XGboost、KSVR 和 QSSVR 作为预测模型, 分别记为: 预测方法 1、预测方法 2、预测方法 3、预测方法 4 和预测方法 5. 进而, 利用五种预测方法和本

文预测方法对大豆期货价格高频数据进行预测.

由表 4、表 5 和表 6 可知, 当情感变量均值和标准差的比率为 10%、20% 和 30% 时, 预测方法 3、预测方法 4、预测方法 5 和本文预测方法的三种误差评价指标均显示相似的结果, 原因在于机器学习模型具有良好的泛化性, 对于存在极少噪声的数据集, 均能表现出较好的预测性能. 同时, 随着比率的上升, 即均值和标准差的增加, 扰动项序列的波动幅度不断加大. 在这种情况下, 本文预测方法的预测精度明显高于其他对比方法, 该结果说明通过利用权重函数对数据点进行赋权

处理,本研究预测方法能够降低强随机性噪声对 拟合效果的干扰.

表 4 基于 Bi-LSTM 情感分析的预测方法鲁棒性检验结果对比

Table 4 Comparison of robustness test results of forecasting method based on Bi-LSTM sentiment analysis

情感变量均值 和标准差比率	评价指标	预测方法					
		预测方法 1	预测方法 2	预测方法 3	预测方法 4	预测方法 5	本文预测方法
10 %	MAE	352.129	82.929	66.293	62.654	62.054	62.203
	MAPE/%	7.511	1.801	1.425	1.339	1.320	1.323
	WAPE	7.642	1.800	1.439	1.360	1.347	1.350
	SMAPE/%	7.688	1.806	1.421	1.335	1.314	1.317
	MdAPE/%	0.069	0.013	0.008	0.006	0.006	0.006
20 %	MAE	342.528	78.645	65.592	62.868	64.519	64.730
	MAPE/%	7.328	1.675	1.406	1.340	1.375	1.381
	WAPE	7.434	1.707	1.424	1.364	1.400	1.405
	SMAPE/%	7.445	1.662	1.403	1.336	1.370	1.375
	MdAPE/%	0.070	0.010	0.008	0.007	0.007	0.007
30 %	MAE	392.018	86.688	70.568	68.037	66.972	64.631
	MAPE/%	8.368	1.871	1.534	1.455	1.430	1.378
	WAPE	8.508	1.881	1.532	1.477	1.454	1.403
	SMAPE/%	8.646	1.866	1.525	1.451	1.424	1.373
	MdAPE/%	0.082	0.012	0.010	0.009	0.008	0.007
40 %	MAE	389.335	115.346	71.774	70.712	71.243	68.016
	MAPE/%	8.294	2.549	1.559	1.518	1.527	1.455
	WAPE	8.450	2.503	1.558	1.535	1.546	1.476
	SMAPE/%	8.639	2.581	1.550	1.517	1.524	1.450
	MdAPE/%	0.083	0.020	0.009	0.010	0.010	0.008
50 %	MAE	348.332	106.013	72.050	72.682	75.077	68.694
	MAPE/%	7.474	2.250	1.544	1.557	1.604	1.468
	WAPE	7.560	2.301	1.564	1.577	1.629	1.491
	SMAPE/%	7.541	2.223	1.537	1.554	1.599	1.464
	MdAPE/%	0.071	0.017	0.009	0.010	0.010	0.009
60 %	MAE	359.039	87.272	79.873	78.215	80.304	73.461
	MAPE/%	7.697	1.901	1.755	1.699	1.734	1.574
	WAPE	7.792	1.894	1.734	1.698	1.743	1.594
	SMAPE/%	7.834	1.906	1.740	1.701	1.731	1.568
	MdAPE/%	0.073	0.012	0.012	0.011	0.010	0.010
70 %	MAE	329.850	154.246	83.797	82.352	83.286	71.106
	MAPE/%	7.068	3.455	1.855	1.760	1.804	1.524
	WAPE	7.159	3.348	1.819	1.787	1.808	1.543
	SMAPE/%	7.192	3.513	1.839	1.756	1.792	1.517
	MdAPE/%	0.069	0.026	0.012	0.012	0.012	0.010
80 %	MAE	492.604	217.555	87.408	86.553	87.212	77.267
	MAPE/%	10.553	4.893	1.931	1.880	1.891	1.655
	WAPE	10.691	4.722	1.897	1.878	1.893	1.677
	SMAPE/%	11.213	5.040	1.918	1.883	1.882	1.648
	MdAPE/%	0.113	0.043	0.013	0.013	0.014	0.010
90 %	MAE	396.000	192.775	95.151	92.594	90.777	82.091
	MAPE/%	8.498	4.063	2.100	2.007	1.993	1.760
	WAPE	8.595	4.184	2.065	2.010	1.970	1.782
	SMAPE/%	8.765	3.964	2.075	1.999	1.980	1.756
	MdAPE/%	0.086	0.033	0.015	0.016	0.015	0.010
100 %	MAE	316.234	228.972	100.843	97.002	95.796	84.969
	MAPE/%	6.817	4.888	2.259	2.141	2.127	1.819
	WAPE	6.863	4.969	2.189	2.105	2.079	1.844
	SMAPE/%	6.777	4.740	2.229	2.119	2.105	1.812
	MdAPE/%	0.062	0.046	0.017	0.016	0.016	0.011

表 5 基于 CNN 情感分析的预测方法鲁棒性检验结果对比

Table 5 Comparison of robustness test results of forecasting method based on CNN sentiment analysis

情感变量均值 和标准差比率	评价指标	预测方法					
		预测方法 1	预测方法 2	预测方法 3	预测方法 4	预测方法 5	本文预测方法
10 %	MAE	335.434	84.772	65.682	64.074	63.585	63.498
	MAPE/%	7.103	1.840	1.398	1.365	1.355	1.352
	WAPE	7.280	1.840	1.426	1.391	1.380	1.378
	SMAPE/%	7.261	1.842	1.396	1.362	1.351	1.348
	MdAPE/%	0.066	0.012	0.007	0.007	0.006	0.006
20 %	MAE	325.234	86.066	65.343	66.195	65.499	65.310
	MAPE/%	6.929	1.838	1.398	1.413	1.398	1.393
	WAPE	7.059	1.868	1.418	1.437	1.422	1.417
	SMAPE/%	6.997	1.826	1.395	1.411	1.394	1.389
	MdAPE/%	0.064	0.012	0.008	0.008	0.008	0.008
30 %	MAE	319.925	102.888	68.161	69.443	68.442	68.205
	MAPE/%	6.782	2.266	1.472	1.479	1.466	1.461
	WAPE	6.943	2.233	1.479	1.507	1.485	1.480
	SMAPE/%	6.913	2.282	1.465	1.479	1.462	1.457
	MdAPE/%	0.062	0.016	0.008	0.009	0.008	0.008
40 %	MAE	364.207	98.064	69.040	70.583	69.859	67.946
	MAPE/%	7.682	2.137	1.489	1.507	1.491	1.443
	WAPE	7.905	2.128	1.498	1.532	1.516	1.475
	SMAPE/%	7.958	2.143	1.483	1.505	1.490	1.439
	MdAPE/%	0.074	0.015	0.009	0.009	0.009	0.007
50 %	MAE	343.847	114.822	70.920	73.307	71.920	69.354
	MAPE/%	7.270	2.535	1.531	1.574	1.533	1.488
	WAPE	7.463	2.492	1.539	1.591	1.561	1.505
	SMAPE/%	7.478	2.550	1.522	1.573	1.533	1.485
	MdAPE/%	0.070	0.017	0.009	0.010	0.009	0.009
60 %	MAE	391.205	109.197	75.445	75.851	74.944	71.069
	MAPE/%	8.300	2.358	1.656	1.621	1.618	1.516
	WAPE	8.490	2.370	1.637	1.646	1.627	1.542
	SMAPE/%	8.497	2.350	1.644	1.617	1.611	1.511
	MdAPE/%	0.080	0.016	0.009	0.010	0.010	0.008
60 %	MAE	375.326	143.679	77.170	77.692	76.927	71.335
	MAPE/%	7.940	3.201	1.697	1.669	1.664	1.520
	WAPE	8.146	3.118	1.675	1.686	1.670	1.548
	SMAPE/%	8.209	3.240	1.686	1.668	1.655	1.516
	MdAPE/%	0.078	0.024	0.011	0.010	0.011	0.009
80 %	MAE	319.626	153.165	82.778	82.283	83.538	76.844
	MAPE/%	6.830	3.231	1.824	1.792	1.808	1.654
	WAPE	6.937	3.324	1.797	1.786	1.813	1.668
	SMAPE/%	6.860	3.168	1.809	1.783	1.802	1.653
	MdAPE/%	0.062	0.026	0.013	0.012	0.012	0.010
90 %	MAE	408.807	181.662	85.605	87.793	86.595	80.566
	MAPE/%	8.656	4.078	1.891	1.905	1.894	1.724
	WAPE	8.872	3.943	1.858	1.905	1.879	1.749
	SMAPE/%	9.110	4.172	1.876	1.897	1.880	1.722
	MdAPE/%	0.083	0.034	0.011	0.013	0.013	0.010
100 %	MAE	349.431	109.457	92.251	91.040	91.778	83.025
	MAPE/%	7.502	2.383	2.043	1.968	2.010	1.804
	WAPE	7.584	2.376	2.002	1.976	1.992	1.802
	SMAPE/%	7.492	2.380	2.018	1.948	1.996	1.805
	MdAPE/%	0.066	0.015	0.015	0.015	0.015	0.011

表 6 基于 SnowNLP 情感分析的预测方法鲁棒性检验结果对比

Table 6 Comparison of robustness test results of forecasting method based on SnowNLP sentiment analysis

情感变量均值 和标准差比率	评价指标	预测方法					
		预测方法 1	预测方法 2	预测方法 3	预测方法 4	预测方法 5	本文预测方法
10 %	MAE	376.808	88.234	63.477	64.037	64.592	62.408
	MAPE/%	8.330	1.926	1.356	1.365	1.379	1.329
	WAPE	8.178	1.915	1.378	1.390	1.402	1.354
	SMAPE/%	8.085	1.935	1.354	1.363	1.375	1.326
	MdAPE/%	0.074	0.013	0.007	0.007	0.007	0.006
20 %	MAE	372.534	91.690	65.471	66.419	65.399	64.293
	MAPE/%	8.215	2.001	1.400	1.426	1.395	1.372
	WAPE	8.085	1.990	1.421	1.442	1.419	1.395
	SMAPE/%	8.022	2.006	1.396	1.423	1.391	1.368
	MdAPE/%	0.072	0.013	0.008	0.008	0.007	0.007
30 %	MAE	390.129	95.775	69.795	71.270	68.630	68.305
	MAPE/%	8.636	2.049	1.495	1.537	1.470	1.466
	WAPE	8.467	2.079	1.515	1.547	1.490	1.482
	SMAPE/%	8.345	2.035	1.492	1.534	1.468	1.464
	MdAPE/%	0.076	0.015	0.008	0.009	0.008	0.009
40 %	MAE	433.058	167.346	73.118	76.856	75.086	72.732
	MAPE/%	9.529	3.738	1.589	1.650	1.616	1.557
	WAPE	9.399	3.632	1.587	1.668	1.630	1.579
	SMAPE/%	9.286	3.794	1.585	1.645	1.611	1.552
	MdAPE/%	0.085	0.028	0.010	0.010	0.010	0.009
50 %	MAE	516.027	152.739	78.166	82.459	78.927	76.243
	MAPE/%	11.494	3.219	1.721	1.784	1.718	1.643
	WAPE	11.200	3.315	1.696	1.790	1.713	1.655
	SMAPE/%	10.704	3.156	1.706	1.783	1.704	1.639
	MdAPE/%	0.096	0.025	0.010	0.011	0.010	0.009
60 %	MAE	404.705	197.705	84.433	87.377	86.302	82.459
	MAPE/%	8.950	4.161	1.854	1.886	1.875	1.770
	WAPE	8.783	4.291	1.832	1.896	1.873	1.790
	SMAPE/%	8.606	4.045	1.836	1.889	1.869	1.770
	MdAPE/%	0.076	0.036	0.012	0.012	0.013	0.011
70 %	MAE	415.086	123.562	88.000	90.443	87.098	83.809
	MAPE/%	9.156	2.715	1.929	1.960	1.897	1.802
	WAPE	9.009	2.682	1.910	1.963	1.890	1.819
	SMAPE/%	8.869	2.729	1.909	1.962	1.884	1.800
	MdAPE/%	0.083	0.020	0.013	0.013	0.014	0.011
80 %	MAE	441.623	167.542	97.440	100.278	94.922	89.612
	MAPE/%	9.767	3.566	2.146	2.179	2.091	1.934
	WAPE	9.585	3.636	2.115	2.176	2.060	1.945
	SMAPE/%	9.304	3.511	2.118	2.179	2.070	1.934
	MdAPE/%	0.082	0.028	0.017	0.013	0.015	0.012
90 %	MAE	453.126	234.189	99.690	109.740	98.076	89.858
	MAPE/%	9.954	5.320	2.175	2.402	2.161	1.936
	WAPE	9.834	5.083	2.164	2.382	2.129	1.950
	SMAPE/%	9.715	5.495	2.147	2.410	2.131	1.936
	MdAPE/%	0.093	0.044	0.016	0.017	0.015	0.014
100 %	MAE	425.924	256.526	101.770	111.429	105.079	92.401
	MAPE/%	9.324	5.748	2.258	2.419	2.336	1.968
	WAPE	9.244	5.567	2.209	2.418	2.281	2.005
	SMAPE/%	9.227	5.876	2.228	2.411	2.304	1.966
	MdAPE/%	0.085	0.047	0.017	0.019	0.018	0.014

为了更客观地评估不同波动程度噪声对六种对比方法预测效果的影响,以 MAPE 作为评估指标,对比分析各方法在噪声波动强度逐渐增大时预测误差的变化情况,其结果如图 7、图 8 和图 9 所示.其中图 7 展示了基于 Bi-LSTM 情感分析方法数据集的预测误差变化情况,结果显示预测方法 1、预测方法 2、预测方法 3、预测方法 4 和预测方法 5 和本文预测方法的误差评价指标增长率分别为 49%、192%、61%、60%、61% 和 37%.图 8 为基于 CNN 情感分析方法数据集的预测误差变化情况.可见,随着噪声波动幅度的增强,预测方法 1 的误差评价指标呈现了上升和下降交替的情况,表明其对噪声数据的抗干扰能力不稳定.

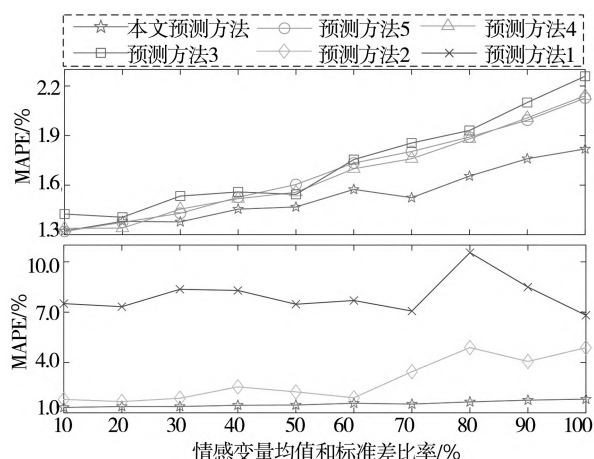


图 7 基于 Bi-LSTM 情感分析方法数据集的预测误差变化情况

Fig. 7 Variation of prediction error based on Bi-LSTM sentiment

analysis method dataset

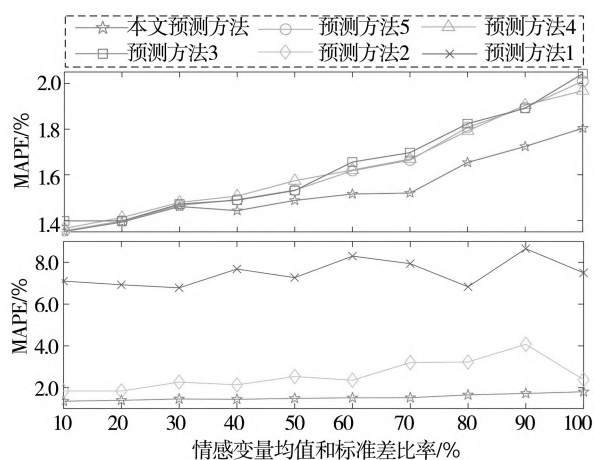


图 8 基于 CNN 情感分析方法数据集的预测误差变化情况

Fig. 8 Variation of prediction error based on Bi-LSTM sentiment

analysis method dataset

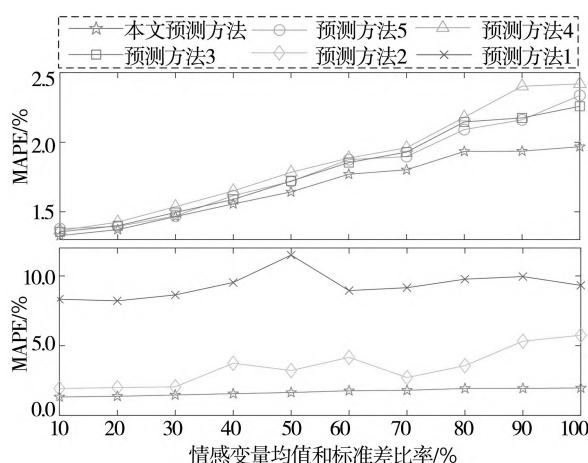


图 9 基于 SnowNLP 情感分析方法数据集的预测误差变化情况

Fig. 9 Variation of prediction error based on SnowNLP sentiment

analysis method dataset

另外,按照 MAPE 指标对其他五种预测方法的误差增长速度从高到低进行排序,得到预测方法 2 > 预测方法 5 > 基预测方法 3 > 预测方法 4 > 本文预测方法.图 9 显示了基于 SnowNLP 情感分析方法数据集的预测误差变化情况,结果表明预测方法 2 的误差评价指标增长幅度最大,高达 198%,而本文预测方法的误差评价指标的增幅则较低,仅为 48%.通过上述对比分析可知,本文预测方法具有较强的鲁棒性,能够有效解决预测过程中情感输入特征可能含有大量噪声的问题.

4 结束语

在商品期货市场中,期货交易的频率高,且价格变化快.传统的期货价格低频数据预测通常不能充分反映市场的快速变化.与低频数据相比,期货价格高频数据包含了更多交易信息,能够详细刻画期货市场的微观结构变化.考虑到高频期货数据具有高度非线性、非平稳性和高波动性等特征,本研究以提升期货高频数据预测的精确度和鲁棒性作为切入点,提出了基于新闻情感分析的期货价格高频数据鲁棒预测方法.首先,对期货相关新闻文本进行情感分析,得到情感分值变量.然后,分别对商品期货价格序列和情感分值变量进行多尺度分解,并利用 SE 方法对其进行重构处

理.其次,基于 DWP 函数,构建 WDWPS-SVR 模型,并利用其对各子序列进行预测.最后,将本研究应用于 2019 年—2022 年的 DCE 大豆期货收盘价预测,得出以下主要结论.

第一,本研究提出的基于新闻文本情感分析的商品期货价格高频预测方法具有较高的预测精度和强鲁棒性特征.通过设计权重函数,该方法在训练过程中能够有效降低海量新闻文本存在的噪声对参数估计的干扰,从而生成拟合优度更高的“回归器”.鲁棒性检验结果表明,随着数据噪声随机波动性的增强,该预测方法的拟合效果仍能保持良好的准确性和稳定性,适用于解决非结构化文本数据驱动的预测问题.

第二,WDWPS-SVR 模型具有较强的非线性特征学习能力,能够在原始空间内生成四次双井势曲面,从而对商品期货价格高频数据的复杂细

节波动进行非线性拟合.相比于现有 SVR 等机器学习模型,该模型在拟合高频期货数据时,表现出更优的预测性能,为金融领域高频数据复杂特征的提取提供了一种新的视角.

第三,新闻文本潜在反映了投资者行为偏好、情绪变化和心理预期.通过 NLP 技术,新闻文本被量化为情感变量,能够为商品期货价格走势预测提供实时信息.在预测过程中,情感变量对于期货价格波动具有显著的解释能力和建模能力,可提高预测效果的时效性.

未来,将从以下两方面进行深入探究.第一,针对非结构化数据的利用,进一步挖掘与商品期货相关的在线语音和视频报道等内容,以综合考虑不同来源的多类型实时在线信息.第二,在预测应用上,将本研究方法扩展至其他预测问题,以更好地满足其它实际应用的需求.

参 考 文 献:

- [1] 部 慧,汪寿阳.商品期货及其组合通胀保护功能的实证分析[J].管理科学学报,2010,13(9):26-36.
Bu Hui, Wang Shouyang. Empirical study of inflation-hedging characteristics of commodity futures and its portfolio in China[J]. Journal of Management Sciences in China, 2010, 13(9): 26-36. (in Chinese)
- [2] Basak S, Pavlova A. A model of financialization of commodities[J]. Journal of Finance, 2016, 71(4): 1511-1556.
- [3] 尹力博,杨清元,韩立岩.技术指标能够预测商品期货价格吗?来自中国的证据[J].管理科学学报,2018,21(6):99-109.
Yin Libo, Yang Qingyuan, Han Liyan. Can technical indicators forecast commodity prices? Evidence from China[J]. Journal of Management Sciences in China, 2018, 21(6): 99-109. (in Chinese)
- [4] 王静远,葛逸清,汤 珂,等.调整期货交易规则可以降低投资者杠杆吗?[J].管理科学学报,2021,24(2):99-110.
Wang Jingyuan, Ge Yiqing, Tang Ke, et al. Can adjustments of futures trading rules reduce investor's leverage? [J]. Journal of Management Sciences in China, 2021, 24(2): 99-110. (in Chinese)
- [5] Luo J, Klein T, Ji Q, et al. Forecasting realized volatility of agricultural commodity futures with infinite Hidden Markov HAR models[J]. International Journal of Forecasting, 2022, 38(1): 51-73.
- [6] Song Y, Cai C, Ma D, et al. Modelling and forecasting high-frequency data with jumps based on a hybrid nonparametric regression and LSTM model[J]. Expert Systems with Applications, 2024, 237: 121527.
- [7] Li Z, Wang T, Drapeau S, et al. Evolution of Chinese futures markets from a high frequency perspective[J]. Economics & Politics, 2024, 36(3): 1416-1449.
- [8] Gong X, Lai P, He M, et al. Climate risk and energy futures high frequency volatility prediction[J]. Energy, 2024, 307: 132466.
- [9] Li J, Li G, Liu M, et al. A novel text-based framework for forecasting agricultural futures using massive online news headlines[J]. International Journal of Forecasting, 2022, 38(1): 35-50.
- [10] Xu Z, Mohsin M, Ullah K, et al. Using econometric and machine learning models to forecast crude oil prices: Insights from economic history[J]. Resources Policy, 2023, 83: 103614.
- [11] Hasselgren A, Hou A J, Suardi S, et al. Do oil price forecast disagreement of survey of professional forecasters predict

- crude oil return volatility? [J]. *International Journal of Forecasting*, 2025, 41(1): 141–152.
- [12] Huang W, Gao T, Hao Y, et al. Transformer-based forecasting for intraday trading in the Shanghai crude oil market: Analyzing open-high-low-close prices[J]. *Energy Economics*, 2023, 127: 107106.
- [13] Purohit S K, Panigrahi S. Novel deterministic and probabilistic forecasting methods for crude oil price employing optimized deep learning, statistical and hybrid models[J]. *Information Sciences*, 2024, 658: 120021.
- [14] Zhang X, Xia Z, He F, et al. Forecasting crude oil prices with alternative data and a deep learning approach[J]. *Annals of Operations Research*, 2024: 1–27.
- [15] Fang T, Zheng C, Wang D. Forecasting the crude oil prices with an EMD-ISBM-FNN model[J]. *Energy*, 2023, 263: 125407.
- [16] Zeng L, Ling L, Zhang D, et al. Optimal forecast combination based on PSO-CS approach for daily agricultural future prices forecasting[J]. *Applied Soft Computing*, 2023, 132: 109833.
- [17] Yu L, Liang S, Chen R, et al. Predicting monthly biofuel production using a hybrid ensemble forecasting methodology[J]. *International Journal of Forecasting*, 2022, 38(1): 3–20.
- [18] Wu B, Wang L, Wang S, et al. Forecasting the US oil markets based on social media information during the COVID-19 pandemic[J]. *Energy*, 2021, 226: 120403.
- [19] Li Y, Jiang S, Li X, et al. The role of news sentiment in oil futures returns and volatility forecasting: Data-decomposition based deep learning approach[J]. *Energy Economics*, 2021, 95: 105140.
- [20] Lin W, Wei Y. Economic forecasting with big data: A literature review[J]. *Journal of Management Science and Engineering*, 2024, 9(2): 254–270.
- [21] Grinberg N, Joseph K, Friedland L, et al. Fake news on Twitter during the 2016 US presidential election[J]. *Science*, 2019, 363(6425): 374–378.
- [22] Luo J, Hong T, Gao Z, et al. A robust support vector regression model for electric load forecasting[J]. *International Journal of Forecasting*, 2023, 39(2): 1005–1020.
- [23] Jha N, Tanneru H K, Palla S, et al. Multivariate analysis and forecasting of the crude oil prices: Part I-Classical machine learning approaches[J]. *Energy*, 2024, 296: 131185.
- [24] 刘 岭, 王聚杰, 李建平. 融合新闻影响力衰减的国际原油价格预测研究[J]. *系统工程理论与实践*, 2022, 42(10): 2710–2720.
- Liu Ling, Wang Jujie, Li Jianping. International crude oil price forecasting with news influence index attenuation[J]. *Systems Engineering: Theory & Practice*, 2022, 42(10): 2710–2720. (in Chinese)
- [25] Ye J, Yang Z, Ma M, et al. Epsilon-kernel-free soft quadratic surface support vector regression[J]. *Information Sciences*, 2022, 594: 177–199.
- [26] Gao Z, Fang S C, Luo J, et al. A kernel-free double well potential support vector machine with applications[J]. *European Journal of Operational Research*, 2021, 290(1): 248–262.
- [27] Brooks J P. Support vector machines with the ramp loss and the hard margin loss[J]. *Operations Research*, 2011, 59(2): 467–479.
- [28] Wang W, Zhang Z, Wang L, et al. Forecasting the market value of power battery industry chain: A novel RRMIDAS-SVR model[J]. *Journal of Management Science and Engineering*, 2024, 9(4): 474–489.
- [29] Yu L, Ma Y, Ma M. An effective rolling decomposition-ensemble model for gasoline consumption forecasting[J]. *Energy*, 2021, 222: 119869.
- [30] Luo J, Yan X, Tian Y. Unsupervised quadratic surface support vector machine with application to credit risk assessment[J]. *European Journal of Operational Research*, 2020, 280(3): 1008–1017.
- [31] Bai Y, Li X, Yu H, et al. Crude oil price forecasting incorporating news text[J]. *International Journal of Forecasting*, 2022, 38(1): 367–383.
- [32] 刘 莉, 郝显峰, 王玉东. 基于时变稳健加权最小二乘法的股市收益率预测[J]. *管理科学学报*, 2024, 27(1): 141–158.
- Liu Li, Hao Xianfeng, Wang Yudong. Forecasting stock returns: A time-varying robust weighted least squares approach[J]. *Journal of Management Sciences in China*, 2024, 27(1): 141–158. (in Chinese)
- [33] 孙丽君, 李方方, 胡祥培. 基于离散数据流分割算法的预测方法[J]. *管理科学学报*, 2024, 27(9): 48–61.

- Sun Lijun, Li Fangfang, Hu Xiangpei. Forecasting method based on segmentation algorithm of discrete data stream[J]. Journal of Management Sciences in China, 2024, 27(9): 48–61. (in Chinese)
- [34] Zhang C, Tian Y X, Fan Z P. Forecasting sales using online review and search engine data: A method based on PCA-DS-FOA-BPNN[J]. International Journal of Forecasting, 2022, 38(3): 1005–1024.
- [35] Nie Y, Liang N, Wang J. Ultra-short-term wind-speed bi-forecasting system via artificial intelligence and a double-forecasting scheme[J]. Applied Energy, 2021, 301: 117452.
- [36] Luo J, Hong T, Fang S. Robust regression models for load forecasting[J]. IEEE Transactions on Smart Grid, 2018, 10(5): 5397–5404.
- [37] Runge J, Saloux E. A comparison of prediction and forecasting artificial intelligence models to estimate the future energy demand in a district heating system[J]. Energy, 2023, 269: 126661.
- [38] Chen J, Hong Z, Niu L. Forecasting interest rates with shifting endpoints: The role of the functional demographic age distribution[J]. International Journal of Forecasting, 2025, 41(1): 153–174.
- [39] Dinis D, Barbosa-Povoa A, Teixeira Â P. Enhancing capacity planning through forecasting: An integrated tool for maintenance of complex product systems[J]. International Journal of Forecasting, 2022, 38(1): 178–192.
- [40] Wang L, Wang X, Zhao Z. Mid-term electricity demand forecasting using improved multi-mode reconstruction and particle swarm-enhanced support vector regression[J]. Energy, 2024, 304: 132021.

Robust forecasting method for high-frequency futures price based on sentiment analysis

LIU Jin-pei^{1, 2}, LUO Rui¹, CHEN Hua-you¹, YU Le-an³, JIA Ning^{4*}, ZHU Ning⁵

1. School of Big Data and Statistics, Anhui University, Hefei 230601, China;
2. School of Business, Anhui University, Hefei 230601, China;
3. School of Business, Sichuan University, Chengdu 610065, China;
4. College of Management and Economics, Tianjin University, Tianjin 300072, China;
5. School of Management, University of Science and Technology of China, Hefei 230026, China

Abstract: Commodity futures prices exhibit high nonlinearity, non-stationarity, and volatility. The sentiment polarity extracted from news texts has significantly improved the timeliness and accuracy of forecasting. However, massive news texts contain a significant amount of noise information, which severely impacts the stability of forecasting. Therefore, sentiment analysis is conducted on news texts related to futures prices, and a weighted double-well potential surface support vector regression (WDWPS-SVR) model is proposed to improve prediction robustness for commodity futures prices. The model generates a quartic double well potential surface to conduct regression on the data in the original feature space, capturing the complex data characteristics in high-frequency data. Besides, the weight of each training sample are calculated using a weight function to assess their importance in the fitting process, reducing the interference of news texts noise in parameter estimation. The model is applied to predict soybean futures closing prices on the Dalian Commodity Exchange from 2019 to 2022. The experimental results show that: 1) Sentiment variables extracted from news texts effectively explain high-frequency price volatility, improving prediction timeliness; 2) The WDWPS-SVR model's strong nonlinear learning ability enhances prediction accuracy by fitting fine-grained price fluctuations; 3) A weighting approach reduces noise from news texts, boosting model robustness and prediction stability.

Key words: futures price forecasting; kernel-free support vector regression; robustness; sentiment analysis; noise data