监督学习模型在中国期货市场中的应用

Fuquan Tang

School of Mathematics Science, Shanghai Jiao Tong University, 800 Dongchuan RD. Minhang District, Shanghai, China

arXiv:2303.04581v1 [q-fin.ST] 8 Mar 2023

摘要

本文基于中国期货市场的特点，构建了一个用于预测期货价格走势的监督学习模型，并根据预测结果设计了交易策略。分类问题的精确率、召回率和F1分数表明，就测试数据而言，我们的模型能够满足期货价格走势分类的准确性要求。回测结果显示，我们的交易系统收益曲线呈上升趋势，且资金回撤较低。

*Keyw*词语： 分数阶差分；三重屏障法；神经网络；监督学习 g

1. 引言

文艺复兴科技、老虎环球基金以及许多其他投资公司一直依靠量化技术取得成功，但它们的成功并非仅仅依赖计量经济模型[1]。过去几十年来，全球金融体系的规模、集中度和复杂性都有了显著增长，其复杂程度已超出传统量化方法的建模能力。此外，一些非常有用的数据集，例如卫星图像、语音录音或新闻情绪，也是计量经济学难以触及的[2]。近年来，许多对冲基金开始尝试机器学习（ML）方法。机器学习是人工智能的一个子领域，指利用机器从过往经验中学习[3]。传统编程要求开发人员预先考虑每一种可能的情况并据此编写程序，而机器学习能够根据数据灵活调整输出。机器学习算法实际上并不编写代码，而是构建现实世界的计算机模型，再用数据训练该模型。机器学习分为监督学习、无监督学习和强化学习。监督学习要求我们从历史数据中了解正确答案，并用它来预测未来结果。例如，可以利用过去的房价预测当前和未来的价格。通过有效运用基于试错的统计改进流程，机器根据监督者提供的一组数值进行测试，并依据测试结果逐步提高准确度[4, 5]。无监督学习则没有明确的正确答案，但我们希望从数据中发现新的规律。它最常用于对数据进行排序或分组。例如，音乐软件 Spotify 会对音乐进行分类，并推荐你可能想听的歌曲或专辑。随后，它们还会对听众进行分类，以判断他们更喜欢听 Radiohead 还是贾斯汀·比伯[6, 7]。强化学习不需要领域专家，但要求持续朝着预定目标取得进展。我们都知道，基于强化学习理论的 AlphaGo 在 DeepMind 中与自己进行了上百万场对弈，最终成为世界冠军[8, 9]。

由于每分钟都需要处理海量数据，机器学习已被证明有助于提升金融行业的服务与运营水平。如今，很难想象现代金融行业还能不使用机器学习。Kou等人[10]调查了结合机器学习技术评估和衡量金融系统性风险的现有研究与方法，

联系人：Fuquan Tang，邮箱：dongfang tang@163.com*本研究获国家自然科学基金项目（项目编号：11531001）支持。

包括大数据分析、网络分析和情感分析等。Babenko 等人[11] 回顾了机器学习（监督学习和无监督学习）的经典方法，举例说明了不同方法的应用，并探讨了有助于实证经济学研究的方法（基于乌克兰企业、银行和官方国家统计数据）。L´opez[12]回顾了机器学习从业者已提供更优解决方案的十个重要投资问题。这些投资问题包括：资产定价、风险管理与投资组合构建、异常值检测、下注规模、情感分析、特征重要性、信用评级与分析师建议、交易执行、大数据分析、控制效应和交互作用。与此同时，机器学习并非万能良方。机器学习技术虽然灵活有效，但也有其阴暗面。机器学习算法一旦遭到误用，就可能把统计偶然现象误认为规律。金融数据的信噪比较低，这几乎注定粗心的使用者会以越来越快的速度得出虚假发现。机器学习的失败率尤其高；Prado 和 L´opez[13]总结指出，原因可归结为七种常见错误：西西弗斯范式、整数阶差分、低效抽样、错误标记、非独立同分布样本加权、交叉验证泄漏和回测过拟合。

关于使用机器学习模型预测金融资产价格未来走势的研究文献非常丰富。Chiong 等人[14]提出了一种基于情感分析、用于金融市场预测的支持向量机。Henrique 等人[15] 提出了一种基于聚类的分类模型，用于预测金融危机。Wu 等人[16] 采用了一种名为“连续趋势标记”的新标注方法，对金融时间序列数据进行标注，并再次纳入价格趋势。Fleischer[17] 复现了 Marcos L´opez de Prado 教授的一些技术，并将其应用于美国股票市场，其中包括三重障碍法。不过，他没有使用分数阶差分，而是采用整数阶差分，并围绕新的 10-K 或 10-Q 报告提交时间对价格柱进行抽样，试图根据新信息的到来进行抽样。Borges 和 Rui[18] 提到 Marcos L´opez de Prado 教授的技术代表了最先进的水平，但最终并未采用这些技术，而是选择用分钟价格变动的累积和阈值来构建K线，这种方法更为直接。Galvis[19] 运用 Marcos L´opez de Prado 教授的数据分析技术，构建了一个用于预测比特币价格方向的深度神经网络。

为减少机器学习模型中的常见错误[13], ，本文复现了康奈尔大学 Marcos L´opez de Prado 教授针对中国期货市场采用的一些数据分析技术。首先，我们使用一种名为分数阶差分的数据转换方法，在尽可能保留记忆的同时确保数据的平稳性；其次，我们使用基于路径依赖的三重障碍法为金融数据打标签；最后，我们构建了一个神经网络模型，用于预测中国期货市场主力合约的价格走势。为获得更直观的结果，我们在测试数据上进行了交易模拟。

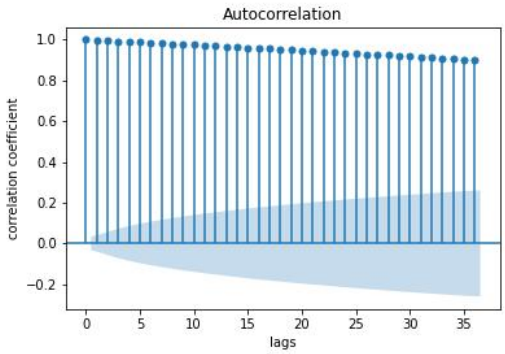
2. 数据分析

机器学习在金融领域的主要应用是利用市场交易数据（如价格和交易量）来预测资产价格走势。根据文献[20], ，股价波动具有非线性和随机性。具体而言，交易活动在一天、一周或一个月内很少保持稳定。随着宏观经济数据发布、行业政策文件或公司公告等信息流的变化，市场交易行为也会随之改变。下图 1–2 展示了中国期货市场螺纹钢主力合约收盘价数据的自相关函数（ACF）和偏自相关函数（PACF）图。这些图表明该时间序列是非平稳的。

我们的目标是构建监督学习模型，以预测金融时间序列的价格走势。在数据分析部分，我们主要完成两项任务：首先，通过分数阶差分变换将时间序列转换为近似平滑的时间序列；其次，使用三重屏障法为样本数据打标签。

2.1. 分数阶差分特征

众所周知，金融序列表现出较低的信噪比[21] ，且金融数据不稳定[22,23]。对于数据分析和机器学习而言，必须将其转换为平稳序列。通常的做法是先



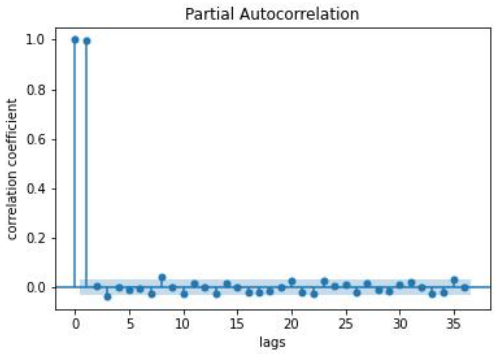


图 1：螺纹钢合约的自相关　图 2：螺纹钢合约的偏自相关

对时间序列数据取对数，然后进行整数阶差分，但这存在问题。为了确保时间序列平稳，会丢弃时间序列的记忆。正如 Marcos L´opez de Prado[24]所指出的，价格之间存在记忆，即今天的价格基于之前的价格。本文应用分数阶差分特征[24] ，对期货价格序列进行平稳化转换。

考虑后移算子 B 作用于实值特征矩阵 $\{X_t\}$，其中 $B^k X_t = X_{t-k}$ 对任意整数 $k \geq 0$ 均成立。例如，$(1-B)^2 = 1 - 2B + B^2$，其中 $B^2 X_t = X_{t-2}$，因此 $(1-B)^2 X_t = X_t - 2X_{t-1} + X_{t-2}$。

注意，$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$，其中 n 为正整数。对于实数 d，$(1+x)^d = \sum_{k=0}^{\infty} \binom{d}{k} x^k$ 是一个二项级数。在分数阶模型中，指数 d 可以取实数，其形式二项级数展开如下：

$$\begin{aligned}(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k &= \sum_{k=0}^{\infty} \frac{\prod_{i=0}^{k-1}(d-i)}{k!}(-B)^k \\ &= \sum_{k=0}^{\infty} (-B)^k \prod_{i=0}^{k-1} \frac{d-i}{k-i} \\ &= 1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \cdots .\end{aligned} \tag{2.1}$$

下面我们分析 d 阶分数阶差分如何保留记忆。这个算术级数由一个点积级数组成，点积级数由一个点积组成

$$\tilde{X}_t = \sum_{k=0}^{\infty} \omega_k X_{t-k} \tag{2.2}$$

权重为 ω

$$\omega = \left\{1, -d, \frac{d(d-1)}{2!}, -\frac{d(d-1)(d-2)}{3!}, \ldots, (-1)^k \prod_{i=0}^{k-1} \frac{d-i}{k!}, \ldots\right\} \tag{2.3}$$

值为 X

$$X = \{X_t, X_{t-1}, X_{t-2}, X_{t-3}, \ldots, X_{t-k}, \ldots\} \tag{2.4}$$

当 d 是正整数时，$\prod_{i=0}^{k-1} \binom{d-i}{k} ! = 0$，$\forall k > d$，并且该点之后的记忆会被取消。例如，$d = 1$ 用于计算收益，其中 $\prod_{i=0}^{k-1} \binom{d-i}{k} ! = 0$，$\forall k > 1$，且 $\omega = \{1, -1, 0, 0, \ldots\}$。

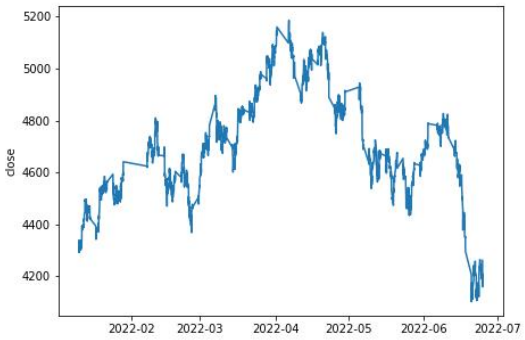


图3：螺纹钢合约的收盘价曲线

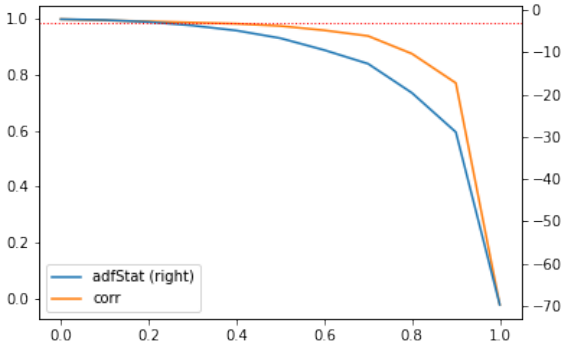


图4：螺纹钢合约的ADF统计量曲线

观察权重序列 ω，我们可以看出，对于 $k=0,\ \cdots,\ \infty$，当 $\omega_0=1$ 时，可以通过迭代生成权重：

$$\omega_k=-\omega_{k-1}\frac{d-k+1}{k} \tag{2.5}$$

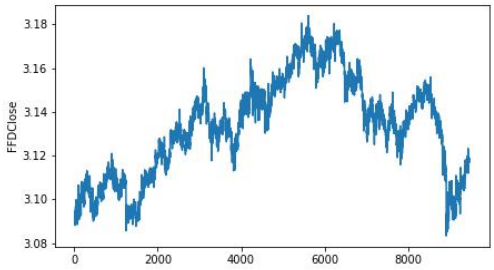
我们来考察权重的收敛性。根据上述结果可以看出，对于 $k>d$，如果 $\omega_{k-1}\neq 0$，那么 $\left|\frac{\omega_k}{\omega_{k-1}}\right|=\left|\frac{d-k+1}{k}\right|<1$；否则 $\omega_k=0$。因此，由于单位圆内因子的无限乘积，权重渐近收敛于零。此外，对于正的 d 和 $k<d+1$，我们有 $\frac{d-k+1}{k}\geq 0$，这使得初始权重的符号交替变化。对于非整数 d，一旦 $k\geq d+1$，ω_k 在 int $[d]$ 为偶数时为负，否则为正。总之，当 int $[d]$ 为偶数时，$\lim_{k\to\infty}\omega_k=0^-$（从左侧收敛至零）；当 Int $[d]$ 为奇数时，$\lim_{k\to\infty}\omega_k=0^+$（从右侧收敛至零）。在特殊情形 $d\in(0,1)$ 下，这意味着 $-1<\omega_k<0,\forall k>0$。这种权重符号交替对于使 $\{X_t\}_{t=1,\ldots,T}$ 平稳是必要的，因为随着时间推移，记忆会逐渐衰减或被抵消。分数阶差分有两种实现方法：标准的“扩展窗口”方法，以及一种我称为“固定宽度窗口分数阶差分”（FFD）[24]的新方法。

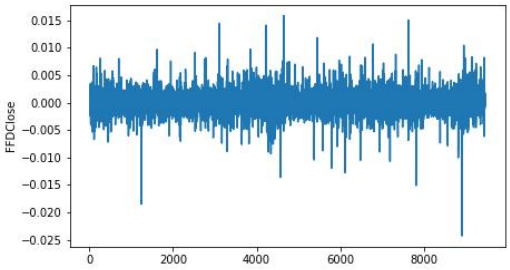
在我们的算法框架中，我们采用固定宽度窗口分数阶差分法。当权重的模 ($|\omega_k|$) 低于给定阈值 (τ) 时，我们便将其舍弃；这等同于找出第一个l^*，使得 $|\omega_{l^*}|\geq\tau$ 且 $|\omega_{l^*+1}|\leq\tau$。随后，我们得到一个新变量 $\tilde{\omega}_k$

$$\tilde{\omega}_k=\begin{cases}\omega_k & \text{if } k\leq l^*\\ 0 & \text{if } k>l^*\end{cases} \tag{2.6}$$

以及 $\tilde{X}_t=\sum_{k=0}^{t^*}\tilde{\omega}_k X_{t-k}$，对于 $t=T-l^*+1,\ \ldots,\ T$。这一过程的优点是，在对 $X_t,\ t=l^*,\ \cdots,\ T$ 的所有估计中，都使用相同的权重向量，从而避免扩展窗口新增权重所导致的负漂移。结果如预期，是水平加噪声的无漂移混合。由于记忆带来的偏度和超额峰度，分布不再是高斯分布，但仍然平稳。

图3展示了螺纹钢合约的收盘价曲线，图4展示了基于不同差分阶数的螺纹钢期货ADF统计量曲线。在图4中，右侧纵轴绘制了根据螺纹钢期货对数价格计算出的增广迪基-福勒（ADF）统计量。横轴显示用于生成计算ADF统计量所依据序列的 d 值。原始序列的ADF统计量为-0.3387，而收益序列的ADF统计量为-46.9114。在95%的置信水平下，检验的临界值为-2.8618。ADF统计量在 $d=0.3$附近越过该阈值。左侧纵轴绘制了原始序列（$d=0$）与不同d值下差分后序列之间的相关性。在 $d=0.3$处，时间序列的相关系数为0.988。这证实，FFD在保留大部分记忆的同时实现了





F图5：螺纹钢合约的 FFD，d=0。图6：螺纹钢合约的 FFD，d=1

平稳性。相比之下，原始序列与收益序列之间的相关性仅为0.022，因此标准整数阶差分几乎完全抹去了序列的记忆。几乎所有金融论文都试图通过应用整数阶差分 $d = 1 \gg 0.3$来恢复平稳性，这意味着大多数研究对序列进行了过度差分，去除了远超满足标准计量经济学假设所需的记忆。

图5—6展示了主力螺纹钢合约在 $d = 0.3$ 和 $d = 1$下的 FFD。 $d = 0.3$下的 FFD 与图3中的原始时间序列更为相似，而 $d = 1$ 下的 FFD 则更接近白噪声，与原始时间序列的差异也更大。

表1展示了中国期货市场60个主要期货合约指数经过 FFD(d) 处理后的 ADF统计量。我们知道，满足单位根检验的 ADF 值为 -2.8623，表格中单元格的黄色部分表示通过了单位根检验。从表中可以看出，$Shanghai\,Lead$、 $Bean\,One$、 $Polypropylene$和 $No.20\,rubber$无需进行差分即可满足单位根检验；大多数期货指数合约只需进行 $d = 0.3$ 阶差分即可获得平稳序列，只有 $Shanghai\,Tin$、 $Apple$和 Pig需要进行 $d = 0.5$阶差分才能满足单位根检验。在所有情况下，用于计算收益的标准 $d = 1$ 都意味着过度差分。

2.2. 标注

关于金融时间序列数据的标注，大多数监督学习文献采用固定时间窗口法（FHM）为观测值打标签。该方法可描述如下。考虑一个特征矩阵 X，包含 I行，$\{X_t\}_{t=1,\ldots,I}$，取自一组索引为 $t = 1, \cdots T$ 的价格柱，其中 $I \leq T$。观测值 X_i被赋予标签 $y_i \in \{-1, 0, 1\}$。

$$y_i = \begin{cases} -1 & \text{if } r_{t_{i,0},t_{i,0}+h} < -\tau \\ 0 & \text{if } \left|r_{t_{i,0},t_{i,0}+h}\right| \leq \tau \\ 1 & \text{if } r_{t_{i,0},t_{i,0}+h} > \tau \end{cases} \tag{2.7}$$

其中， τ 是预先定义的常数阈值，$t_{i,0}$ 是 X_i 发生后紧接着的价格柱的索引，$t_{i,0} + h$ 是 $t_{i,0}$之后第 $h-$ 个价格柱的索引，并且 $r_{t_{i,0},t_{i,0}+h}$ 是价格在 h这一价格柱时间跨度内的收益率，

$$r_{t_{i,0},t_{i,0}+h} = \frac{p_{t_{i,0}+h}}{p_{t_{i,0}}} - 1 \tag{2.8}$$

由于文献几乎总是使用时间柱， h 意味着固定时间窗口。FHM 有两个主要问题：首先， τ 是一个常数，而标的资产的波动率往往会随时间变化。当标的资产波动率较大时，被标记为零的样本很少；而当标的资产波动率较小时，大多数样本会被标记为零。换言之，样本仅根据数值的绝对大小而非统计显著性进行标注，这显然不合理。此外，FHM 没有考虑止损和止盈问题，而是

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
铁矿石	-0.374	-1.651	-2.246	-3.139	-4.309	-6.040	-8.691	-11.774	-18.417	-27.614	-78.621
螺纹钢	-2.079	-2.403	-3.049	-4.156	-5.632	-7.860	-11.267	-15.182	-23.465	-34.486	-78.921
热轧卷板	-1.690	-2.060	-2.771	-3.889	-5.361	-7.540	-10.852	-14.656	-22.709	-33.473	-78.978
不锈钢	-1.830	-2.071	-2.638	-3.495	-4.699	-6.501	-9.281	-12.537	-19.680	-29.717	-93.928
硅铁	-1.966	-2.168	-2.672	-3.488	-4.623	-6.331	-8.979	-12.024	-18.528	-27.166	-62.703
锰硅	-2.240	-2.558	-3.257	-4.305	-5.767	-7.911	-11.203	-14.944	-22.643	-32.366	-62.569
沪铜	-2.354	-2.592	-3.158	-4.099	-5.417	-7.428	-10.519	-14.148	-21.988	-32.780	-89.515
沪铝	-1.744	-1.939	-2.368	-3.073	-4.070	-5.584	-7.928	-10.714	-16.764	-25.439	-90.908
沪锌	-1.906	-2.233	-2.919	-3.962	-5.374	-7.461	-10.657	-14.355	-22.326	-33.259	-89.992
沪铅	-4.584	-5.550	-7.381	-10.119	-13.667	-18.750	-26.154	-34.291	-48.889	-64.613	-92.167
沪镍	-1.487	-1.759	-2.327	-3.196	-4.363	-6.101	-8.768	-11.897	-18.720	-28.326	-93.281
沪锡	-1.564	-1.586	-1.722	-2.055	-2.572	-3.415	-4.769	-6.387	-10.017	-15.297	-86.878
国际铜	-2.266	-2.505	-3.070	-3.995	-5.297	-7.274	-10.315	-13.888	-21.606	-32.258	-89.528
上海黄金	-2.458	-2.966	-4.002	-5.573	-7.638	-10.680	-15.31	-20.631	-31.752	-46.377	-101.248
豆一	-3.791	-4.311	-5.351	-7.042	-9.362	-12.810	-18.013	-23.870	-35.332	-48.882	-80.202
豆二	-2.107	-2.314	-2.852	-3.708	-4.895	-6.694	-9.497	-12.740	-19.811	-29.479	-77.838
豆油	-1.827	-2.019	-2.511	-3.279	-4.351	-5.986	-8.480	-11.395	-17.791	-26.665	-77.745
豆粕	-1.937	-2.203	-2.800	-3.760	-5.017	-6.932	-9.885	-13.310	-20.677	-30.741	-79.534
植物油	-1.858	-2.030	-2.430	-3.118	-4.092	-5.589	-7.922	-10.653	-16.647	-25.071	-78.483
植物粕	-2.092	-2.345	-2.897	-3.826	-5.059	-6.958	-9.893	-13.292	-20.630	-30.646	-78.871
棕榈油	-1.623	-1.771	-2.155	-2.791	-3.685	-5.056	-7.173	-9.644	-15.128	-22.870	-77.838
Corn	-1.765	-2.046	-2.640	-3.570	-4.824	-6.695	-9.579	-12.925	-20.128	-30.000	-80.550
玉米淀粉	-2.273	-2.572	-3.234	-4.312	-5.765	-7.969	-11.349	-15.259	-23.566	-34.661	-80.752
Eggs	-2.652	-2.947	-3.682	-4.894	-6.521	-8.947	-12.566	-16.684	-24.942	-34.955	-61.599
一号棉花	-1.459	-1.544	-1.820	-2.295	-2.991	-4.050	-5.728	-7.712	-12.121	-18.468	-77.492
棉纱	-1.813	-1.913	-2.278	-2.876	-3.755	-5.084	-7.199	-9.701	-15.203	-23.017	-77.862
白糖	-2.850	-3.024	-3.550	-4.526	-5.885	-8.001	-11.299	-15.096	-23.256	-34.188	-79.838
苹果	-0.491	-0.733	-1.177	-1.803	-2.579	-3.697	-5.378	-7.307	-11.504	-17.424	-61.124
红枣	-1.229	-1.393	-1.758	-2.353	-3.151	-4.355	-6.222	-8.409	-13.178	-19.858	-62.742
花生	-1.395	-1.646	-2.128	-2.870	-3.859	-5.344	-7.615	-10.254	-15.930	-23.642	-60.799
粳米	-0.942	-1.222	-1.775	-2.597	-3.632	-5.161	-7.499	-10.210	-16.114	-24.446	-81.155
Pig	-1.966	-1.760	-1.702	-1.784	-2.242	-2.868	-3.975	-5.281	-8.243	-12.613	-60.838
Fuel	-1.428	-1.615	-2.019	-2.704	-3.618	-5.016	-7.184	-9.724	-15.279	-23.158	-79.110
液化石油气	-1.353	-1.570	-2.036	-2.769	-3.750	-5.213	-7.478	-10.109	-15.865	-23.974	-78.700
乙二醇	-2.809	-3.234	-4.156	-5.650	-7.603	-10.516	-14.920	-19.927	-30.077	-42.728	-79.186
聚乙烯	-2.615	-3.069	-4.063	-5.545	-7.515	-10.414	-14.779	-19.733	-29.808	-42.343	-78.727
PTA	-1.728	-1.870	-2.268	-2.961	-3.392	-5.391	-7.680	-10.369	-16.236	-24.498	-79.117
聚氯乙烯	-1.502	-1.905	-2.792	-4.053	-5.670	-8.015	-11.542	-15.581	-24.040	-35.189	-79.259
聚丙烯	-3.028	-3.635	-4.872	-6.724	-9.145	-12.659	-17.878	-23.733	-35.080	-48.368	-78.921
苯乙烯	-2.438	-2.555	-3.124	-4.137	-5.491	-7.554	-10.763	-14.514	-22.418	-33.132	-79.563
纯碱	-2.377	-2.384	-2.543	-2.966	-3.692	-4.866	-6.771	-9.041	-14.088	-21.360	-77.330
Urea	-2.068	-2.300	-2.748	-3.541	-4.648	-6.345	-8.960	-11.991	-18.450	-27.080	-62.846
橡胶	-2.720	-3.325	-4.504	-6.262	-8.566	-11.913	-16.925	-22.584	-33.727	-47.110	-80.432
20号胶	-4.229	-4.997	-6.617	-9.088	-12.283	-16.876	-22.508	-30.773	-43.653	-57.311	-80.153
纸浆	-1.998	-2.178	-2.657	-3.442	-4.548	-6.236	-8.869	-11.921	-18.590	-27.822	-78.333
短纤	-2.565	-2.876	-3.644	-4.885	-6.526	-8.997	-12.783	-17.152	-26.248	-38.022	-78.730
焦煤	-1.676	-1.840	-2.240	-2.919	-3.860	-5.297	-7.543	-10.172	-15.957	-24.110	-78.567
Coke	-1.866	-2.188	-2.858	-3.906	-5.304	-7.382	-10.562	-14.237	-22.057	-32.601	-79.257
CSI	-1.706	-2.020	-2.642	-3.608	-4.899	-6.806	-9.714	-13.073	-20.109	-29.431	-66.374
沪深	-1.142	-1.420	-1.971	-2.791	-3.854	-5.415	-7.795	-10.552	-16.467	-24.594	-67.359
SSE	-1.239	-1.511	-2.051	-2.865	-3.931	-5.503	-7.908	-10.693	-16.678	-24.879	-67.436
十年期国债	-1.654	-1.836	-2.302	-3.054	-4.101	-5.657	-8.075	-10.890	-16.922	-25.262	-67.843
五笔债务	-1.687	-1.858	-2.328	-3.080	-4.141	-5.704	-8.136	-10.976	-17.042	-25.401	-67.324
第二笔债务	-1.550	-1.769	-2.293	-3.127	-4.248	-5.911	-8.479	-11.462	-17.814	-26.502	-68.633
上海白银	-1.724	-2.337	-3.494	-5.141	-7.226	-10.234	-14.766	-19.955	-30.795	-45.103	-101.618
原油	-1.398	-1.521	-1.833	-2.389	-3.161	-4.350	-6.203	-8.392	-13.252	-20.301	-100.094
低硫原油	-0.759	-1.015	-1.490	-2.197	-3.083	-4.385	-6.367	-8.658	-13.660	-20.799	-78.810
石油沥青	-1.114	-1.440	-2.068	-2.987	-4.164	-5.885	-8.497	-11.516	-18.017	-27.019	-78.242
甲醇	-2.150	-2.532	-3.357	-4.626	-6.293	-8.753	-12.482	-16.762	-25.667	-37.210	-78.386
玻璃	-0.802	-1.117	-1.701	-2.545	-3.612	-5.167	-7.524	-10.256	-16.172	-24.481	-79.247

表 1：中国期货交易品种合约指数的最优 FradDiff 平稳性。

这只是对时间区间内总体收益的大致指示。对于触发止损或头寸被强制平仓的情况，基于总收益的加价通常无法反映真实情况。Marcos L´opez de Prado[24] 提出了一种创新的样本标签标注方法——三重障碍法。这是一种路径依赖方法，因此能有效解决上述止损问题。该方法的具体细节如下。

2.2.1. 三重屏障法

我称之为三重障碍法，是因为它根据标签来观察价格是否触及三道障碍。首先，我们设置两道水平障碍和一道垂直障碍。两道水平障碍由止盈和止损定义限定，并且是基于估算波动率（已实现波动率或隐含波动率）的动态函数；第三道障碍则由自此之后经过的价格柱数量（即到期限制）定义。若价格先触及上方障碍，则将该观测值标记为 1；若先触及下方障碍，则标记为 -1；若先触及垂直障碍，则标记为 0。总体而言，水平标签定义了哪些价格水平将被归类为正或负，而垂直标签则决定了标签的回溯范围。

$$Y_i = \begin{cases} -1 & \text{if lower barrier is first touched ,} \\ 0 & \text{if no previous barriers were reached ,} \\ 1 & \text{if upper barrier is first touched ,} \end{cases} \tag{2.9}$$

其中，上方障碍和下方障碍定义为

$$upperBarrie = (1 + EMA(volatility) * upfactor) * closingPrice,$$

$$lowerBarrie = (1 + EMA(volatility) * lowerfactor) * closingPrice,$$

其中，upfactor 和 lowerfactor 是预先设定的波动率敏感度参数。这显然是一种路径依赖的方法，因为我们需要判断在整个时间区间内的某个时刻是否触及了这三个障碍。图7展示了三重障碍法的配置。如图所示，价格

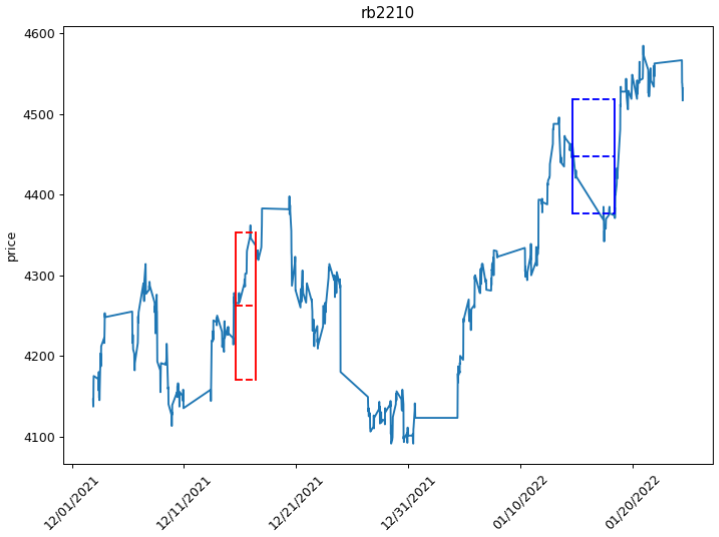


图7：使用三重障碍法进行标注的图表

红色区域应标记为 1，因为顶部边界最先被触及，而且图表的这一部分处于上升趋势；蓝色价格区间应标记为 -1，因为底部边界最先被触及，而且图表的这一部分处于下降趋势。

3. 自动交易系统

我们分四个步骤建立交易系统:

- **差分：** 我们对价格数据应用分数阶差分法，以确保时间序列平稳。
- **标注：** 我们使用三重屏障法为数据标注。
- **建模：** 我们使用 PyTorch 的 Module 类构建用于多分类的神经网络模型。
- **交易：** 利用模型的预测结果，我们使用量化交易平台 VeighNa 搭建交易系统。

前三个步骤用于训练神经网络模型，第四个步骤主要用于进行回测模拟，并基于预测标签构建交易策略。

3.1. 神经网络模型训练

模型训练分为三个步骤：第一步是对数据进行预处理，第二步是训练模型，第三步是进行模型验证。

任何机器学习模型的基础都是用于训练和测试的数据。关于数据预处理，我们以上海期货交易所的螺纹钢期货合约为例，目前的主力合约是10月到期合约 rb2210。我们的目标是构建一种日内交易策略，选取包含开盘价、收盘价、最高价、最低价和成交量的10分钟周期价格数据。首先，我们对数据进行平稳性变换。收盘价原始序列的ADF统计量为-0.3387，在95%的置信水平下，检验的临界值为-2.8618，因此收盘价是非平稳的。借助分数阶差分理论，对收盘价数据进行 $d = 0.3$ 阶差分后，得到的ADF统计量为-3.049，满足平稳性要求；差分后的收盘价数据与原始收盘价数据的皮尔逊相关系数达到0.988。其次，我们使用基于 Python 的技术分析软件包 TA-Lib 计算技术分析指标，共得到81个指标，例如 MACD、RSI、KDJ、布林带等。这些技术指标从六个方面刻画价格数据的变化：重叠指标、动量指标、成交量指标、周期指标、价格变换、波动率指标等。随后，我们使用三重屏障法为数据打标签。

将数据输入模型之前，我们需要对样本特征数据的每一列进行归一化。归一化的目的是消除不同变量的尺度差异对参数学习的影响，例如价格和成交量之间的尺度差异，这可能会在训练网络时造成潜在困难[25]。我们通过将 -1、0 和 1 分别改为 0、1 和 2，向标签向量添加一个由 1 组成的向量 Y，这样做只是为了便于独热编码。独热编码是机器学习中的一种处理过程，用虚拟变量表示分类变量。随后将数据随机划分为训练数据（80%）和测试数据（20%）。训练数据用于逐步拟合模型，而测试数据则用于评估模型的性能[25], ，此举是为了避免模型过拟合。

Torch.nn 是专为神经网络设计的模块化接口。nn 构建于 Autograd 之上，可用于定义和运行神经网络。我们基于 torch.nn 构建神经网络模型。网络最初的初始层由两个线性层和一个非饱和激活函数组成；网络最初的残差层由两个线性层和两个非饱和激活函数组成。下图展示了我们的网络模型结构。图8为神经网络模型结构图。对于优化，我们选用交叉熵损失函数 nn.CrossEntropyLoss()，并选择 Adam 作为优化器。图9为监督学习模型结构图。我们的神经网络模型经过数据训练和参数校正后，得到了最终的预测模型。

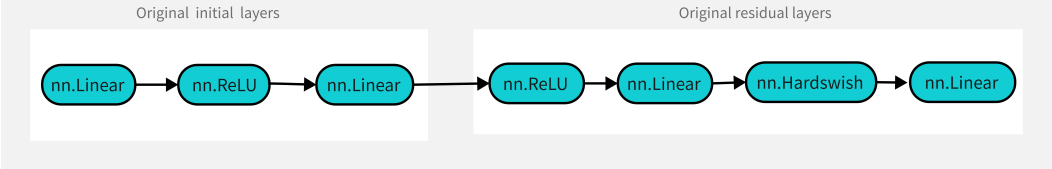


图8：神经网络模型结构图

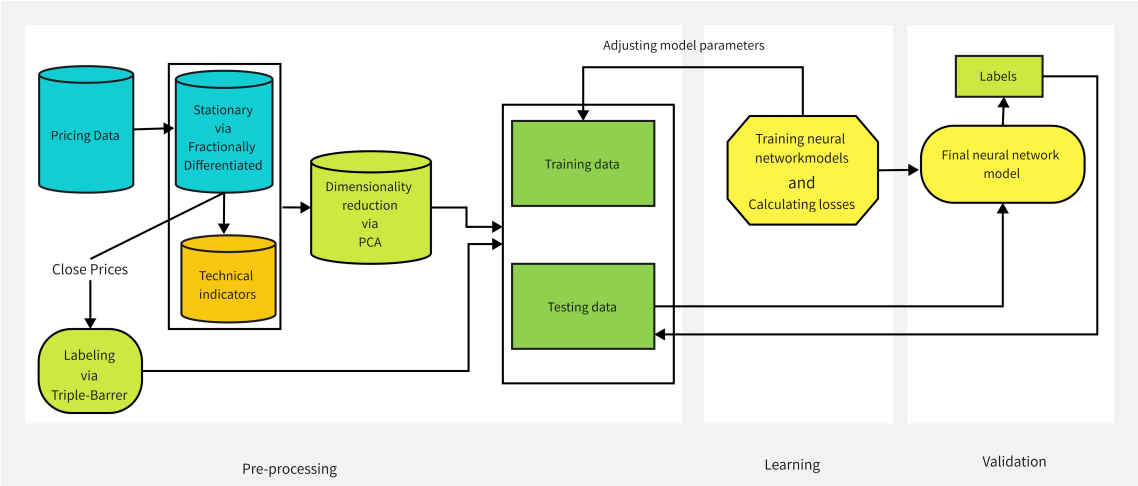


图9：监督学习模型结构图

3.2. 交易策略

为了获得更直观的结果，我们根据最终神经网络模型预测的标签，使用基于 Python 的开源量化交易系统 VeighNa 构建交易策略。交易规则设计如下：当期货头寸为零时，如果预测标签为 2，则买入 1 手期货；如果标签为 0，则做空 1 手；如果标签为 1，则不进行操作。下单时，我们根据收盘价的波动率设置止盈和止损条件。

持有期货多头订单时：

- **止盈：** 收盘价×(1+pa×波动率)。
- **止损：** 收盘价×(1-pb×波动率)。

持有期货空头订单时：

- **止盈：**收盘价×(1-pc×波动率)。
- **止损：**收盘价×(1+pd×波动率)。

上述交易规则中的超参数 pa、pb、pc、pd 是乘数系数，我们通过嵌入 VeighNa 的遗传算法在回测中对其进行优化，以获得最优解。图10为交易规则的结构图。

3.3. 回测

我们选择螺纹钢主力合约指数的十分钟周期数据进行回测分析。通过模拟发现，数据周期过短时，由于数据中的噪声过多，系统会提示频繁交易，最终因交易成本和交易滑点而导致交易结果不佳。我们希望构建一种日内交易次数较少的日内交易策略

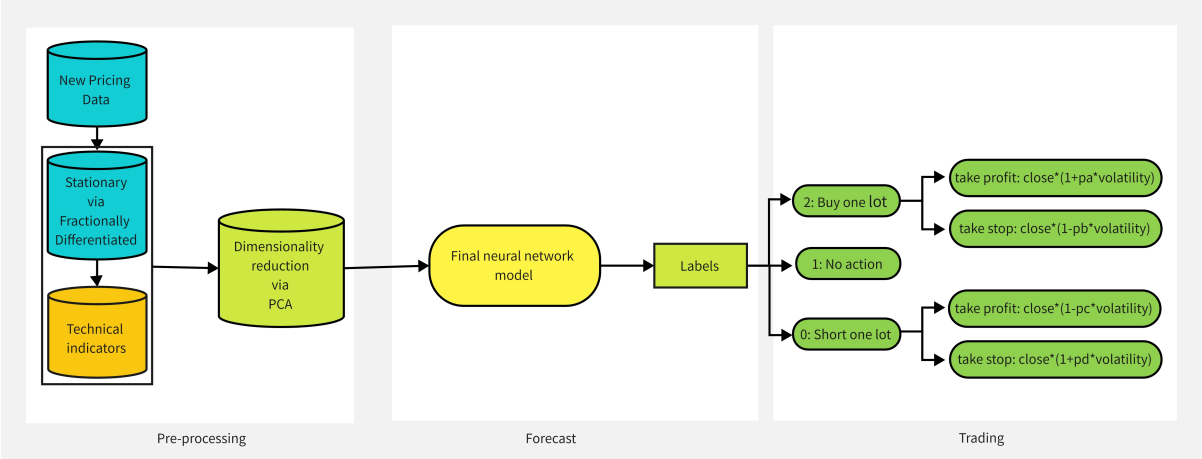


图10：交易规则结构图

；如果周期过长，则交易机会会过少。交易数据的时间范围为2022年12月15日至2022年5月24日，其中前80%的数据用于训练，后20%用于测试。首先，我们合成十分钟周期数据，然后使用 $d = 0.3$进行分数阶差分变换；对差分后的序列进行增强迪基-富勒检验后，拒绝了零假设，并得出不存在单位根的结论，因此得到了平稳序列。通过技术分析软件包 TA-Lib 计算技术指标，并结合价格数据，使每个样本数据成为一个86维向量；再借助主成分分析（PCA）将样本数据降维，使输入数据成为一个16维向量。使用三重障碍法为数据打标签时，我们将时间限制窗口设为12个周期，并将上下障碍的乘数均设为3。交易策略中的超参数经过优化后确定如下： $[pa,pb,pc,pd]=[5,2,5,2]$。

我们使用分类模型评估指标——精确率、召回率和 F1分数，来描述模型在测试集上的预测性能。预测结果指标分为三类：0、1、2。对于 $i \in \{0,1,2\}$，精确率表示被正确预测为指标 i 的指标数量占所有被预测为指标 i 的指标总数的比例；召回率表示被正确预测为指标 i 的指标数量占实际属于指标 i 的指标总数的比例。F1分数是精确率与召回率的调和平均值；支持度是每个分类测试集中的样本总数；宏平均是所有标签结果的平均值；加权平均是所有标签结果的加权平均值。精确率和召回率报告及示意图如图11所示。由此可见，所提出模型对做多、不下注和做空的精确率分别为 80%、76% 和 72%。图12展示了模型在测试集上的预测结果混淆矩阵。由于分类器在各个类别上均具有较高的精确率和召回率，因此分类器的泛化能力得到了充分保证，这意味着它足以对期货价格走势的各个阶段进行分类。

图13展示了回测结果分析。CtaBacktester 中的回测配置如下：我们约定回测资金为人民币 200,000 元，每笔开仓订单为 1 手，佣金率为 0.00005，交易滑点为 1.0。回测结果显示，在 25 个交易日中，有 11 个交易日盈利、9 个交易日亏损，最大回测亏损为 -1.07%，日均交易 2 笔，总收益率为 2.37%，年化收益率为 22.77%，夏普比率为 3.47。

4. 结论

本文基于监督学习模型构建了一个期货交易系统。我们首先根据分数阶差分方法理论对价格数据进行平滑处理，然后使用 Triple

```
Train set for fold 0: Average train_loss: 0.0000, Accuracy: 3296/3296 (100.00000%)
Test set for fold 0:  Average test_loss:  3.7715, Accuracy: 628/825 (76.12121%)

labels in prediction: [0 1 2]
```

	precision	recall	f1-score	support
long	0.80	0.73	0.77	225
no bet	0.76	0.81	0.78	392
short	0.72	0.71	0.71	208
accuracy			0.76	825
macro avg	0.76	0.75	0.75	825
weighted avg	0.76	0.76	0.76	825

图11：测试数据的分类报告

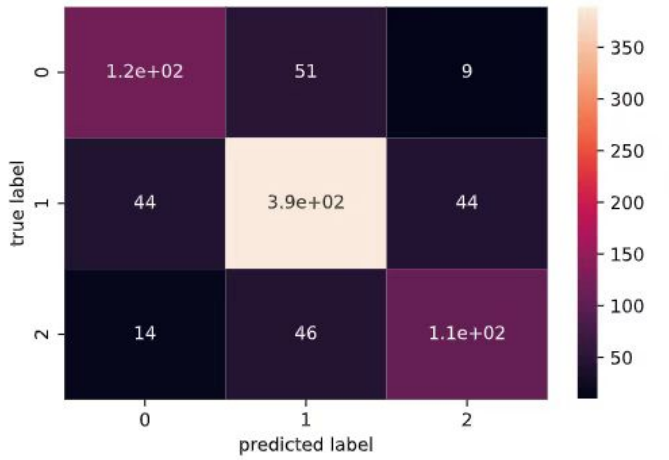


图12：测试数据的混淆矩阵



图13：交易系统的回测分析图

使用障碍法将期货收盘价标注为三类：上涨、下跌和震荡。我们构建了监督学习模型，并根据模型的预测结果设计交易策略。模型的回测结果证明了该交易系统的有效性。后续研究将从以下方面展开：首先，进一步优化数据柱的选择；其次，根据样本数据的结构特征，进一步研究神经网络模型的构建；第三，进一步优化交易策略的设计。

参考文献

[1] De Prado, M.L.. 特邀编者评论：数学与经济学：现实检验。2016. [2] L
´
3365282 opez
2019
de Prado, M.. 超越计量经济学：迈向金融机器学习的路线图。可见 SSRN；. [3] Dunjko, V., Briegel, H.J.. 量子领域中的机器学习与人工智能：近期进展综述。《物理学进展报告》2018;81(7):074001. [4] Ng, A., Deisenroth, M.. 用于伦敦住房价格预测的机器学习移动应用程序。伦敦帝国学院 2015;. [5] Zhang, Y., Rahman, A., Miller, E.. 使用机器学习和时间回归对住房价格进行纵向建模。《国际住房市场与分析》2022;(即将刊出). [6] Lin, D., Jayarathna, S.. 从个人音乐库自动生成播放列表。载于：2018年IEEE信息重用与集成国际会议（IRI）。IEEE；2018，第217–224页。[7] Schedl, M., Knees, P., McFee, B., Bogdanov, D.. 音乐推荐系统：技术、应用场景与挑战。载于：《推荐系统手册》。施普林格；2022，第927–971页。[8] Danesi, M.. 实验数学：概述及教学意义。《信息时代的数学（教育）》2020;：113–125. [9] Kim, D., Alber, M., Kwok, M.W., Mitrovi´C, J., Ramirez-Atencia, C., P´Erez, J.A.R., 等。革新专利法之前，厘清关于人工智能的假设。《GRUR国际》2022;71(4):295–321. [10] Kou, G., Chao, X., Peng, Y., Alsaadi, F.E., Herrera-Viedma, E.. 金融部门系统性风险分析的机器学习方法。《经济的技术与经济发展》2019;25(5):716–742. [11] Babenko, V., Panchyshyn, A., Zomchak, L., Nehrey, M., Artym-Drohomyretska, Z., Lahotskyi, T.. 经济学研究中的经典机器学习方法：宏观与微观层面的示例。《WSEAS商业与经济学汇刊》2021;18:209–217. [12] Lopez de Prado, M.. 金融机器学习的十种应用。可见 SSRN 3365271 2019;. [13] Lopez de Prado, M.. 机器学习基金失败的7 3031282 2017个主要原因（演示幻灯片）。可见 SSRN；. [14] Chiong, R., Fan, Z., Hu, Z., Adam, M.T., Lutz, B., Neumann, D.. 一种基于新闻披露的情感分析机器学习方法，用于金融市场预测。载于：遗传与进化计算会议附属会议论文集。2018，第278–279页。[15] Henrique, B.M., Sobreiro, V.A., Kimura, H.. 文献综述：应用于金融市场预测的机器学习技术。《应用专家系统》2019;124:226–251. [16] Wu, D., Wang, X., Su, J., Tang, B., Wu, S.. 一种基于趋势的金融时间序列预测标注方法。《熵》2020;22(10):1162. [17] Fleischer, D.. 使用机器学习的自动交易系统。硕士论文；挪威斯塔万格大学；2019. [18] Borges, T.A., Rui, F.N.. 采用不同数据重采样方法的加密货币投资机器学习算法集成。《应用软计算》2020;90(8):106187. [19] Galvis, D.E.A.. 深度学习分类模型中的数据分析：比特币的金融应用。博士论文；2021.[20] Dhifaoui, Z.. 对数收益率和条件波动率的确定性与非线性行为：26个股票市场的实证分析。《南亚宏观经济学与公共财政杂志》2022;11(1):69–94. [21] De Prado, M.L.. 实证金融学的未来。《投资组合管理杂志》2015;41(4):140–144. [22] Gadanecz, B., Jayaram, K.. 金融稳定性指标综述。《欧文·费雪委员会通报》2008;31(1):365–383.[23] Mam´edio, D., Rocha, C., Szczepanik, D., Kato, H.. 战略联盟与动态能力：系统性综述。《战略与管理杂志》2019;. [24] De Prado, M.L.. 金融机器学习进阶。约翰·威立父子出版公司；2018. [25] Rai, B.. 《使用R进行高级深度学习》：掌握如何使用R设计、构建和改进高级神经网络模型。Packt出版社；2019.