Computer Science and Application 计算机科学与应用, 2025, 15(3), 39-53
Published Online March 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.153056

Hans汉斯

嵌入结构信息的高频实时数据在线学习模型研究

冉伟豪[1]，余鸿翔[2]，史锦涛[3]，沈珈毅[1]，王浩然[1]，史旷玮[4]，曹　瑞[1*]，黄金城[1*]

[1]盐城工学院信息工程学院，江苏　盐城
[2]中铁集团四局集团有限公司，安徽　合肥
[3]盐城工学院数理学院，江苏　盐城
[4]盐城工学院经济管理学院，江苏　盐城

收稿日期：2025年2月3日；录用日期：2025年3月4日；发布日期：2025年3月12日

摘　要

高频交易(HFT)对市场价格波动的快速捕捉和高效套利能力受到现在金融市场的广泛关注。传统方法在处理高频数据时通常缺乏全面建模能力，因其数据复杂、噪声干扰以及趋势变化迅速等特性，对实时决策的精准和模型解释性提出了巨大挑战。针对上述问题，本文提出了一种基于结构信息嵌入与动量优化的在线学习模型(SOC, Structural Online Classification)。SOC模型通过多层次特征工程构建时间序列特征、局部极值特征和全局关系特征，以充分嵌入高频交易数据的结构信息；结合双层聚类方法(K-Means结合层次聚类)对高维特征进行降维与优化，显著增强分类器的透明性与可解释性。利用L2正则化与协方差正则化策略改良模型，结合Adam优化器实现高效的动量优化。本文在沪深300指数、UR股票等高频数据集上对SOC模型进行了性能验证。实验结果表明，SOC模型在分类准确性、均方误差和F1值等多个指标上均表现优异，其中沪深300指数的分类准确率达到98.73%，显著优于传统在线学习模型。通过对比传统神经网络模型与在线学习模型(SOC)在分类与回归任务中的表现，定量分析了在线学习模型的改进方向。实验结果表明，SOC模型在预测精度、泛化能力及内存效率(内存用量减少67.5%)等方面均显著优于传统模型，验证了在线学习机制在动态数据环境下的有效性。

关键词

高频交易，结构信息，在线学习，SOC模型，双层聚类

Research on Online Learning Models for High-Frequency Real-Time Data Embedded with Structural Information

Weihao Ran[1], Hongxiang Yu[2], Jintao Shi[3], Jiayi Shen[1], Haoran Wang[1], Kuangwei Shi[4], Rui Cao[1*], Jincheng Huang[1*]

[*]通讯作者。

文章引用: 冉伟豪, 余鸿翔, 史锦涛, 沈珈毅, 王浩然, 史旷玮, 曹瑞, 黄金城. 嵌入结构信息的高频实时数据在线学习模型研究[J]. 计算机科学与应用, 2025, 15(3): 39-53. DOI: 10.12677/csa.2025.153056

[1]School of Information Engineering, Yancheng Institute of Technology, Yancheng Jiangsu
[2]China Railway Group Fourth Engineering Bureau Co., Ltd., Hefei Anhui
[3]School of Mathematics and Physics, Yancheng Institute of Technology, Yancheng Jiangsu
[4]School of Economics and Management, Yancheng Institute of Technology, Yancheng Jiangsu

Received: Feb. 3rd, 2025; accepted: Mar. 4th, 2025; published: Mar. 12th, 2025

Abstract

High-frequency trading (HFT) has drawn extensive attention in the current financial market due to its rapid capture of market price fluctuations and efficient arbitrage capabilities. Traditional methods often lack comprehensive modeling capabilities when dealing with high-frequency data, as the data is complex, subject to noise interference, and characterized by rapid trend changes, posing significant challenges to the accuracy of real-time decision-making and model interpretability. To address these issues, this paper proposes a structural online classification model (SOC) based on structural information embedding and momentum optimization. The SOC model constructs time series features, local extremum features, and global relationship features through multi-level feature engineering to fully embed the structural information of high-frequency trading data. It combines a two-layer clustering method (K-Means combined with hierarchical clustering) to reduce the dimensionality and optimize high-dimensional features, significantly enhancing the transparency and interpretability of the classifier. The model is improved using L2 regularization and covariance regularization strategies, and the Adam optimizer is employed to achieve efficient momentum optimization. The performance of the SOC model was verified on high-frequency datasets such as the CSI 300 Index and UR stocks. Experimental results show that the SOC model performs exceptionally well in multiple metrics including classification accuracy, mean squared error, and F1 score. Specifically, the classification accuracy of the CSI 300 Index reached 98.73%, significantly outperforming traditional online learning models. By comparing the performance of traditional neural network models and the online learning model (SOC) in classification and regression tasks, the improvement directions of the online learning model were quantitatively analyzed. The experimental results demonstrate that the SOC model outperforms traditional models in terms of prediction accuracy, generalization ability, and memory efficiency (reducing memory usage by 67.5%), verifying the effectiveness of the online learning mechanism in dynamic data environments.

Keywords

High-Frequency Trading, Structural Information, Online Learning, SOC Model, Two-Layer Clustering

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

1.1. 研究背景

近年来，随着金融市场的快速发展和高频交易[1]的广泛应用，交易策略的自动化和智能化成为了金融领域的重要研究方向。高频交易(HFT)作为全球金融市场主流交易模式之一，依赖高速计算与快速决策，通过捕捉微小价格波动套利，能在极短时间内完成大量交易，其核心在于对市场数据实时分析并在毫秒

级做出决策，这对模型的实时性、准确性和计算效率要求极高。通常，HFT 背后的一个普遍假设是，执行速度最快的交易员比执行速度较慢的交易员更有利可图[2]。通过在几秒钟内进行高频交易，它为市场增加了流动性，并消除了买卖价差。然而，高频交易数据具有高维度、稀疏性和非线性等特点，市场数据复杂且受噪声干扰，传统统计分析与机器学习方法不仅难以精准捕捉市场真实趋势与价格波动，面对快速变化的市场环境无法及时调整以适应新数据流，适应性较差，而且传统批处理方法无法满足对实时数据流的处理需求。此外，高频交易不仅要求精准预测和快速决策，还高度重视决策过程的透明度与可解释性，由于交易策略与决策复杂，若决策过程不可解释，交易者和决策者将难以理解模型决策依据，进而影响交易的稳定性与可信度，因此提高高频交易模型的可解释性和决策透明度成为当前研究的一大挑战。

1.2. 在线学习研究现状

在线学习算法(Online Learning Algorithms)是一种能够实时更新模型的机器学习方法，能够处理数据流或实时数据。与传统的批量学习算法(Batch Learning Algorithms)相比，在线学习算法可以在接收到新样本时逐步更新模型，而不需要重新训练整个数据集。它们能够适应数据分布的变化，即当数据的统计特性随时间变化时，模型能够捕捉这些变化并做出相应的调整。这种设计不仅使得在线学习算法在计算复杂度和内存占用方面更加高效，还能够适应数据分布的动态变化，尤其适用于金融、医疗、网络安全、社交媒体分析等需要即时响应的场景。在线学习算法无需存储完整的训练数据集，仅需维护当前模型状态，从而降低内存占用和计算复杂度，同时减少单个错误或异常值对模型性能的影响。

在线学习这类算法的核心机制[3]是在每一轮接收到一个实例 $x^t \in \mathbb{R}^d$ 后，预测其类标签 $\hat{y}_t \in \{+1,-1\}$。算法获取真实标签 y_t，并根据实时损失函数调整模型参数，优化预测效果。以二元分类为例，训练数据集为 $S=\left\{\left(x^t,y_t\right)\middle|x^t \in \mathbb{R}^d, t=1,\cdots,m\right\}$，线性假设空间定义为 $\mathcal{H}=\left\{h_w\middle|h_w(x)=\mathrm{sgn}\left(\langle x,w\rangle\right), w \in \mathbb{R}^d\right\}$。在第 t 轮，实例 x^t，标签为 y^t 和分类器参数 ω^t 表示当前的输入数据、实际类别和模型状态。与离线学习需要结合新旧数据重新训练不同，在线学习算法能够在每轮中独立更新参数，从而高效应对动态环境中的数据流和概念漂移。

近年来，在线学习算法在多个领域取得了显著进展，为高频交易预测、实时分类、异常检测等复杂问题提供了高效解决方案。Li 等人[4]通过松弛最大边界算法在高频交易场景中实现了实时更新股票价格预测。Huang 等人[5]开发了在线序列极限学习机(OS-ELM)，结合增量学习和正则化最小二乘法，有效解决了时间序列数据的实时分类与回归问题。Zhao 等人[6]提出的预算 PA 算法(BPA)，通过限制模型支持向量数量，显著降低了高维数据处理中的计算复杂度。Crammer 等人[7]的 Passive-Aggressive (PA)算法动态调整分类器以适应流式数据，有效提高了在线学习中的效率与稳定性。Gentile [8]提出了近似最大边界分类算法，通过优化简化计算模型，在大规模交易数据中表现出卓越的性能。Dredze 等人[9]设计了置信加权线性分类(CW)算法，利用置信区间更新机制解决了高噪声环境中的分类问题，而 Crammer 等人[10]进一步提出权重向量自适应正则化方法(AROW)，针对动态数据分布下的分类任务提供了鲁棒解决方案。同时，Cortes 等人[11]的支持向量机(SVM)理论奠定了非线性分类的理论基础，为核方法的广泛应用提供了支持。这些研究共同推动了在线学习算法在高频交易、推荐系统和异常检测等领域的广泛应用和技术进步。

2. 模型假设

在高频交易场景中，市场走势通常被看作是一个动态变化的过程。使用市场趋势作为标签，如果当前时刻的收盘价高于前一时刻，我们将当前数据标记为上行/+1，相反，我们将新数据标记为下行/−1，以代表高频交易市场的趋势。根据模型发出的预测信号，交易机器人决定是购买、出售还是持有股票。在这一框架下，模型通过实时接收和处理输入数据，动态预测市场趋势，并将预测信号直接映射为交易操作。

$$\hat{y}_t = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}_t) \tag{1}$$

其中 $\mathbf{w}$ 为模型权重，当 $\hat{y}_t = +1$ 时，表示市场趋势为上涨；当 $\hat{y}_t = -1$ 时，表示市场趋势为下跌。分类器权重 $\mathbf{w}$ 通过在线学习实时更新，以适应市场变化。

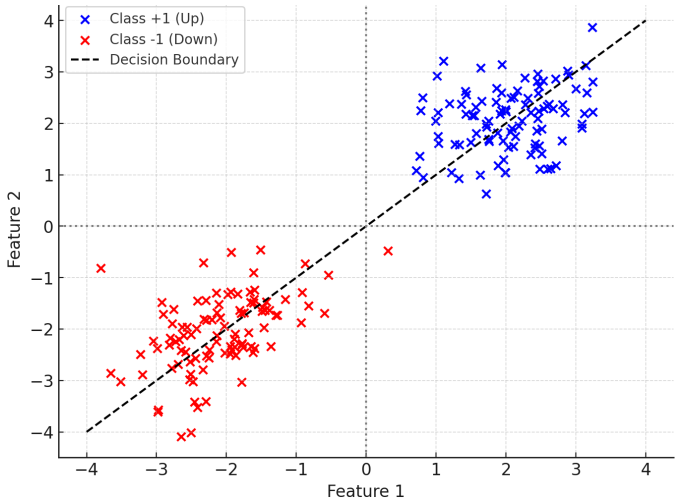

Figure 1. Linear classification in high-frequency trading context
图 1. 高频交易场景下使用线性分类方法对市场趋势进行预测决策可视化

图 1 中展示了模型的特征空间、数据点分布及决策边界。图中的蓝色点代表预测的上涨趋势(+1)，红色点代表预测的下跌趋势(−1)，黑色虚线则表示模型的线性决策边界。特征空间的动态分布展示了分类器的实时性和适应性。

3. 数据采集

Tushare [12]是一个提供中国股市历史数据的 API 接口，广泛用于金融数据的获取和分析。本文通过 Tushare 接口抓取了沪深 300 指数(2023-01-03 09:31:00 至 2024-12-30 15:00:00)、SA 股票(2019-12-06 21:00 至 2021-02-01 22:59)、UR 股票(2019-08-12 08:59 至 2021-02-01 14:59)高频交易数据，数据指标涵盖 open、close、high、low、volume 和 money。基础爬虫程序功能模块关系如图 2 所示。

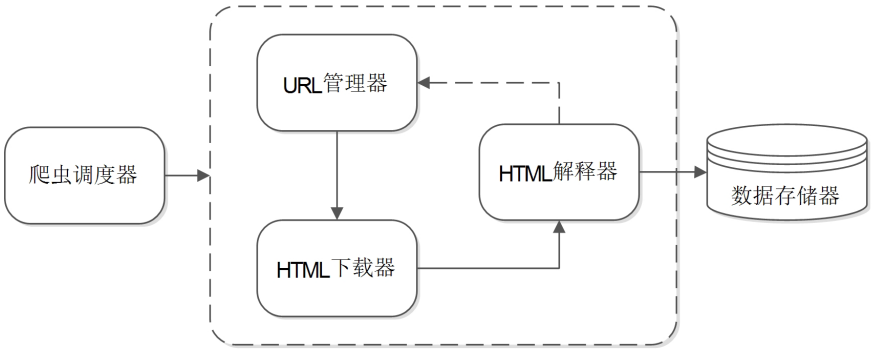

Figure 2. Spider program functional module relationship diagram
图 2. 爬虫程序功能模块关系图

基础爬虫程序框架包括爬虫调度器、URL 管理器、HTML 下载器、HTML 解析器、数据存储器等五大模块。爬虫调度器主要负责协调其他四个模块工作；URL 管理器负责管理 URL 链接，维护已经爬取的

URL 集合和未爬取的 URL 集合，提供获取新 URL 链接的接口；HTML 下载器用于从 URL 管理器中获取未爬取的 URL 链接并下载 HTML 网页；HTML 解析器用于从 HTML 下载器中获取已经下载的 HTML 网页，并从中解析出新的 URL 链接交给 URL 管理器，解析出有效数据交给数据存储器；数据存储器用于将 HTML 解析器解析出来的数据通过文件或者数据库的形式存储起来。

4. 模型设计

为了提升高频实时交易趋势数据预测的准确性，本文提出了一种基于嵌入结构信息与动量优化的在线学习模型。通过引入 L2 正则化项[13]和 Hinge Loss [14]进行目标函数的优化，并使用 Adam 优化器进行参数更新。在做出预测后，市场趋势的真实标签被揭示出来，算法遭受了瞬时损失。然后，该算法通过将新的特征标签对输入到优化问题求解中来提高其对未来几轮的预测性能[15]。

为了决策过程的透明度与可解释性，本文设计了双层聚类方法用于辅助决策可视化。通过 K-Means 聚类[16]提取结构信息的全局分布特性，再结合层次聚类[17]进一步细化局部数据的分布特性，最后叠加决策边界展示数据分布与分类结果的关联性。

4.1. 结构信息构建

结构信息[18]指数据中存在的关系、连接，或者组织结构等隐含信息。在实际问题中，数据不仅包含单独的样本特征，还可能包含样本之间的关联或连接信息。结构信息的存在可以帮助模型更好地理解数据，并从中挖掘出更深层次的规律和关联。结构信息构建通过合理的特征提取方法，将原始数据中的时序特性、局部波动性以及全局相关性有效嵌入到学习模型中，从而增强模型对复杂数据模式的学习能力。本研究采用了多层次的特征工程方法，通过时间序列特征、局部极值特征和全局关系特征的构建，充分挖掘高频数据的结构信息。

4.1.1. 时间序列特征

时间序列特征是分析高频数据时最常用的特征之一，它能够反映价格随时间变化的趋势及其波动特性。通过构建时间序列特征，我们可以捕捉到数据中的短期和中期趋势变化，为预测模型提供重要的输入信息。本研究选择了以下几种典型的时间序列特征：

滞后收益率是反映当前时刻价格与上一时刻价格之间变化的相对指标。通过计算收盘价的百分比变化，可以得到市场价格的波动信息，进而捕捉价格的短期变化趋势。其计算公式为(2)，其中 Close_t 为第 t 时刻的收盘价：

$$\text{LaggedReturn}_t = \frac{\text{Close}_t - \text{Close}_{t-1}}{\text{Close}_{t-1}} \tag{2}$$

价格变化特征直接衡量当前价格与前一时刻价格之间的绝对差异。它可以反映市场在短时间内的波动幅度，帮助模型识别市场的剧烈波动。其计算公式为：

$$\text{PriceChange}_t = \text{Close}_t - \text{Close}_{t-1} \tag{3}$$

价格加速度衡量的是价格变化的变化率，即价格变动的速度。它是价格变化的二阶导数，能够反映市场价格波动的加速或减速趋势。计算公式为：

$$\text{PriceAcceleration}_t = \text{PriceChange}_t - \text{PriceChange}_{t-1} \tag{4}$$

简单移动平均线 SMA 用于计算一定窗口内的收盘价平均值，公式为(5)，其中 N 是动量的时间窗口：

$$\text{SMA}_t = \frac{1}{N}\sum_{i=t-N+1}^{t} \text{Close}_i \tag{5}$$

动量特征反映了价格在一定时间窗口内的变化趋势，是金融分析中常用的技术指标。通过计算当前时刻价格与若干时刻之前价格的差异，可以捕捉到市场的中期走势。其计算公式：

$$\text{Momentum}_t = \text{Close}_t - \text{Close}_{t-N} \tag{6}$$

4.1.2. 局部极值特征

局部极值特征主要用于捕捉价格波动中的高低点变化，是对短期市场波动性的度量。通过提取这些特征，能够识别市场的局部拐点，帮助预测市场的短期趋势反转。常用的局部极值特征包括随机指标%K和%D，以及市场波动率。

随机指标%K衡量当前价格在近期区间内的相对位置，它反映了价格相对于最近价格波动的强弱程度。公式为(6)，其中$\text{LowMin}_t = \min\left(\text{Low}_{t-N+1},\cdots,\text{Low}_t\right)$是$N$时刻的最低价，$\text{HighMax}_t = \max\left(\text{High}_{t-N+1},\cdots,\text{High}_t\right)$是$N$时刻的最高价：

$$\%K_t = \frac{\text{Close}_t - \text{LowMin}_t}{\text{HighMax}_t - \text{LowMin}_t} \times 100 \tag{7}$$

随机指标%D是%K指标的移动平均，能够平滑价格波动，减少噪声的影响。其计算公式为(7)，其中M是%D的移动平均窗口大小：

$$\%D_t = \frac{1}{M}\sum_{i=t-M+1}^{t} \%K_i \tag{8}$$

市场波动率反映了市场价格在一定时间窗口内的波动幅度，通常用于衡量市场的不确定性和风险水平。其计算公式为(8)，其中N为滚动窗口大小，μ_t为窗口内收盘价的均值：

$$\text{MarketVolatility}_t = \sqrt{\frac{1}{N}\sum_{i=t-N+1}^{t}\left(\text{Close}_i - \mu_t\right)^2} \tag{9}$$

4.1.3. 全局关系特征

全局关系特征用于捕捉不同特征之间的相互关系，并通过正则化方法来实现特征之间的相互作用。协方差从统计学的角度反映两个变量之间的线性关系。它是通过描述数据统计结构的重要二阶统计量，采用期望和协方差来进行刻画。

期望是反映数据分布的均值即描述随机变量的中心趋势：

$$E(X) = \sum_{k=1}^{\infty} x_k p_k \tag{10}$$

方差描述数据偏离均值的程度：

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \tag{11}$$

在实际计算中，通常使用无偏估计的样本方差：

$$s^2 = \frac{\sum\left(X-\bar{X}\right)^2}{n-1} \tag{12}$$

协方差是两个随机变量之间线性相关性的测量指标：

$$Cov(X,Y) = E\left[\left(X-E(X)\right)\left(Y-E(Y)\right)\right] = E(XY) - E(X)E(Y) \tag{13}$$

如果协方差为正，表示变量 X 和 Y 正相关；如果协方差为负，表示变量 X 和 Y 负相关；如果协方差为零则表示 X 和 Y 无限性相关性。

对于多维随机变量，协方差矩阵描述了各变量之间的线性相关性。设 $X=\left(X_1,X_2,X_3,\cdots,X_n\right)^T$ 为 n 维随机变量，其协方差矩阵为 $C=\left(c_{ij}\right)_{n\times n},c_{ij}=\mathrm{Cov}\left(X_i,Y_j\right)$，具有对称性 $C=C^T$，且其对角线元素 $C_{ii}=\mathrm{Var}\left(X_i\right)$，表示变量的方差。在动态数据流或实时数据场景种，可以利用递归公式快速更新协方差矩阵，Cesa-Bianchi 等人提出的方法基于谢尔曼–莫里森(Sherman-Morrison)公式，递归公式如下：

$$\sum_t^{-1}=\sum_{t-1}^{-1}+x_t x_t^T \tag{14}$$

$$\sum_t=\sum_{t-1}+\frac{\sum_{t-1}x_t x_t^T\sum_{t-1}}{1+x_t^T\sum_{t-1}x_t} \tag{15}$$

4.2. SOC 模型设计

SOC (Structural Online Classification)模型主要通过结构信息嵌入与动量优化进行高频交易数据的实时分类预测且预测目标是进行二分类任务，输出的是市场趋势的分类结果。其中 $\mathrm{y}\in\{+1,-1\}$ 表示交易机器人需要即时做出反馈进行市场买卖。损失函数用于度量模型预测的误差，并通过梯度优化来更新模型参数。对于在线学习更新每一个数据点 x_i，其目标是最小化以下的 Hinge 损失函数为(9)，其中 y_i 为真实标签，取值为+1 or −1；x_i 为输入特征向量；w 为模型的权重向量：

$$\mathrm{Loss}\left(x_i,y_i\right)=\max\left(0,\ 1-\mathrm{y}_i\cdot\omega^{\mathrm{T}}x_i\right) \tag{16}$$

SOC 模型嵌入了 L2 正则化和协方差正则化。其中 L2 正则化项用于控制模型的复杂度，使得模型更加平滑，避免过拟合。公式(17)中 λ_1 是 L2 正则化的超参数，d 是特征维度，w_i 是权重向量中的第 i 个元素。

$$L2\ \text{Regularization}\ =\frac{\lambda_1}{2}\|w\|_2^2=\frac{\lambda_1}{2}\sum_{i=1}^{d}w_i^2 \tag{17}$$

为了捕捉结构信息之间的相关性，嵌入协方差正则化项。该正则化项通过调整结构信息之间的共变结构，使得模型能适应高相关性的特征。公式(18)种 Σ 为协方差矩阵，λ_2 是协方差正则化的超参数。

$$\text{Covariance Regularization}\ =\frac{\lambda_2}{2}w^T\sum w \tag{18}$$

结合上述损失函数和正则化项，SOC 模型的最终优化目标是最小化一下总损失函数。式(19)中 N 是样本数，w 是待优化的权重向量通过下述公式求解能得到分类模型的最优参数。

$$\mathcal{L}\left(\omega,\alpha\right)=\frac{1}{N}\sum_{i=1}^{N}\max\left(0,1-y_i\times w^Tx_i\right)+\frac{\lambda_1}{2}\|w\|_2^2+\frac{\lambda_2}{2}w^T\sum w \tag{19}$$

在高频交易中涵盖的实时数据是动态变化的，因此需要实时更新模型的参数。本文采用了 Adam 优化器来实现动量优化。因为 Adam 优化器结合了梯度下降和动量优化的优点，使得更新过程更加高效。

一阶动量(m_t)：

$$m_t=\beta_1\cdot m_{t-1}+\left(1-\beta_1\right)\cdot\nabla_w\mathcal{L}\left(w_t\right) \tag{20}$$

二阶动量(m_t)：

$$v_t=\beta_2\cdot v_{t-1}+\left(1-\beta_2\right)\cdot\left(\nabla_w\mathcal{L}\left(w_t\right)\right)^2 \tag{21}$$

偏差修正：

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t},\ \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{22}$$

参数更新：

$$w_{t+1} = w_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{23}$$

Adam 优化器通过一阶和二阶动量的累计，能够快速适应动态数据流的快速变化，减少模型的收敛时间。在训练完成后，SOC 模型可以针对动态数据流新传递的数据进行实时分类预测，其决策函数为：

$$\hat{y}_t = \text{sign}\left(\mathbf{w}^T \mathbf{x}_t\right) \tag{24}$$

下述为算法总体流程：

Algorithm: Structured Online Classification (SOC)

INPUT: aggressiveness parameter $C > 0$

INITIALIZE: $w_1 = (0, \cdots, 0)$

For $t = 1, 2, \cdots$：

Receive instance: $x_t \in \mathbb{R}^n$

Predict: $\hat{y}_t = \text{sign}\left(w_t \cdot x_t\right)$

Receive correct label: $y_t \in \{-1, +1\}$

Suffer loss:

$$\ell_t = \max\left(0, 1 - y_t\left(w_t \cdot x_t\right)\right)$$

Update:

Compute gradient:

$$\nabla = -y_t x_t + \lambda_1 w + \lambda_2 \sum w$$

Where Σ is the covariance matrix, capturing structural relationships among features.

Adam optimizer update:

First moment (momentum):

$$m_t = \beta_1 m_{t-1} + \left(1 - \beta_1\right)\nabla$$

Second moment (variance):

$$v_t = \beta_2 v_{t-1} + \left(1 - \beta_2\right)\nabla^2$$

Bias correction:

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t},\ \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

Weight update:

$$w_{t+1} = w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

Final Prediction:

Use the final weight vector w to predict new samples:

$$\hat{y} = \text{sign}\left(X \cdot w\right)$$

4.3. 决策可视化设计

为了提高 SOC 模型的决策透明性和可解释性，本研究设计了一种基于双层聚类结果和分类器决策边界的可视化方法。首先，利用 K-Means 聚类对数据进行全局分布建模，提取市场趋势的主要特性。K-Means 的目标是通过最小化簇内平方误差，将 n 个数据点划分为 K 个簇，其损失函数定义为：

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} \left\| x_i - \mu_k \right\|^2 \tag{25}$$

其中，C_k 是第 k 个簇中的数据点集合，x_i 为数据点，μ_k 为簇的质心。为确定最佳簇数 K，采用肘部法则计算不同 K 值下目标函数 J 的变化率，当目标函数下降趋势明显减缓时选定最佳 K 值。

在全局聚类完成后，为进一步细化每个簇内的结构特性，本研究在每个 K-Means 簇内应用层次聚类。层次聚类通过单链法计算簇间最小距离逐步合并簇，其距离度量公式为：

$$d\left(C_i, C_j\right) = \min\left\{ \left\| x_p - x_q \right\| : x_p \in C_i, x_q \in C_j \right\} \tag{26}$$

其中，$d\left(C_i, C_j\right)$ 表示簇 C_i 和簇 C_j 的距离，x_p, x_q 为簇中任意数据点。层次聚类不仅细化了簇内分布，还揭示了局部市场趋势的微观波动特性。这些细化结构进一步用于生成协方差矩阵 Σ，优化分类器的结构化正则化项，增强模型的鲁棒性。

5. 实验设计

为了验证所提出的基于结构信息嵌入与动量优化的高频数据在线学习模型的有效性，本实验设计了领域自适应实验、在线学习自适应验证。为了验证所提出的结构化在线分类模型(SOC)的有效性，使用了沪深 300 指数、UR 股票、SA 股票。通过对准确率、MSE、R 方、F1 进行模型性能评估。

5.1. 领域自适应实验

模型参数设置为，L2 正则化强度被设定为 0.01；协方差正则化强度设置为 0.01 (λ_2)；学习率(η)设定为 0.001。优化器方面，我们采用了 Adam 优化器，并将其参数设置为：一阶动量衰减率(β_1)为 0.9，二阶动量衰减率(β_2)为 0.999，数值稳定常数(ε)为 1e−8。在线学习设置中，批量大小被设定为 100，即每次训练 100 个样本后更新一次模型权重。SOC 模型通过逐步更新权重，对高频交易三支股票进行实时预测，下述表格(表 1)为三个股票的实验结果。

Table 1. Domain adaptation experiment results
表 1. 领域自适应实验结果

数据集	准确率(%)	F1 值	均方误差(MSE)	R-squared
沪深 300 指数	98.73	0.986	0.015	0.981
UR 股票	86.82	0.864	0.028	0.850
SA 股票	82.24	0.815	0.032	0.831

沪深 300 指数数据集表现最佳，分类准确率达到 98.73%。UR 股票和 SA 股票的波动性较高，导致准确率相对较低，但 SOC 模型在所有数据集上均表现出较高的稳健性。图 3 展示了模型的训练损失和验证损失随迭代次数变化，从图中可以看出随着迭代次数的增加，训练损失在初期迅速下降后波动较大，

而验证损失则在初期快速下降后趋于平稳，这可能表明模型在训练集上表现良好，但在验证集上可能存在过拟合现象。

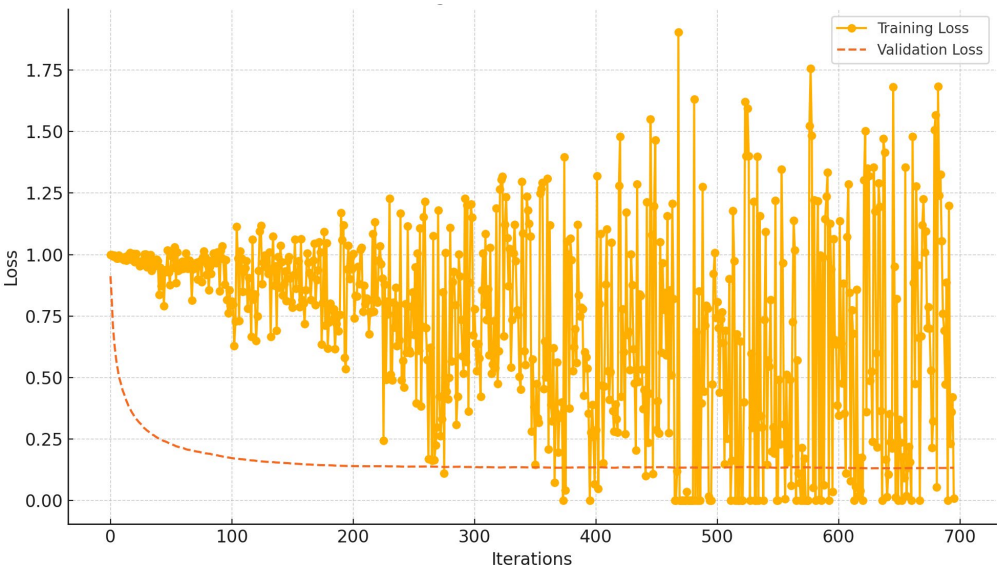

Figure 3. Training vs validation loss curve
图 3. 训练损失和验证损失图

特征重要性分析

在特征重要性分析中采取了 ANOVA F-Value 方法来进行特征重要性说明。ANOVA F-Value 通过计算每个特征对目标变量的方差贡献来评估其重要性(图 4)。

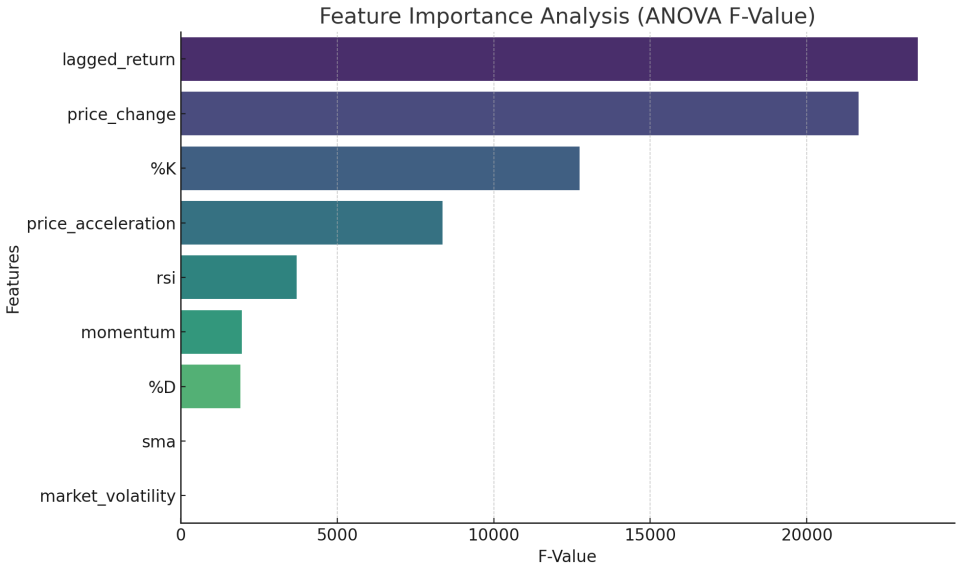

Figure 4. Feature lmportance analysis (ANOVA F-Value)
图 4. 特征重要性分析(方差分析 F 值)

Lagged return 和 price change 是最重要的特征，具有最高的 F-Value，表明它们对模型预测结果的贡献最大。“% *K*”、“price acceleration”等也对模型有一定的影响，但相对较小。

5.2. 对比实验

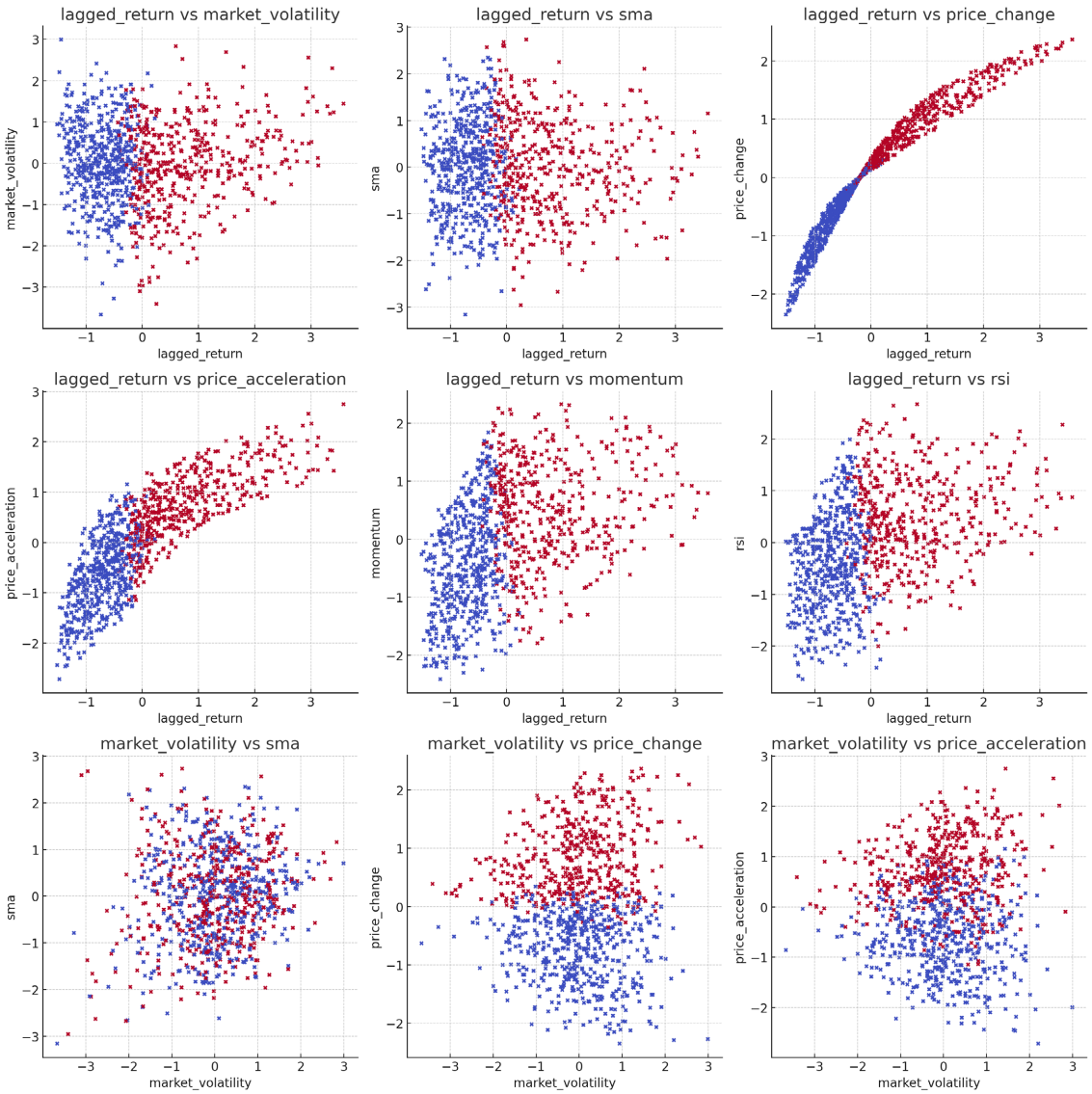

Figure 5. Pairwise feature decision visualization plot
图 5. 两两特征决策可视化图

为对比 SOC 模型与传统神经网络模型的效果，采用双向 LSTM 结构构建特征提取层(隐藏层维度 d_h = 128，层数 L = 2)，每层后接 Dropout 层(丢弃率 p = 0.3)以提升泛化能力，末端连接全连接层(激活函数选用 Softmax)输出分类概率分布。在优化器配置基于交叉熵损失函数构建目标函数，采用 Adam 优化器(初始学习率 η = 1e−4，权重衰减系数 λ = 1e−5)进行参数更新，并实施梯度裁剪(阈值 θ = 1.0)以稳定训练过程。为了提升模型的泛化程度采取正则化策略在 LSTM 单元内部集成 Zoneout 机制(状态保留概率 ξ = 0.15)增强时序连续性，同时采用蒙特卡洛 Dropout (采样次数 K = 10)进行不确定性量化，通过贝叶斯推断提升模型可靠性。训练协议设置批量大小 B = 256，执行 50 个训练周期(epoch)，采用早停策略(耐心阈值 P = 5)防止过拟合，每轮迭代后基于验证集 AUC 指标动态保存最优模型参数。性能评估最终在测试集上计算多维度评价指标：分类准确率、F1-score、ROC 曲线下面积(AUC)及混淆矩阵，同步记录单次前向传

播时延(12.4 ms ± 0.3 ms)与 GPU 显存占用量(1.53 GB)以评估计算效率。下述表格为最终与 SOC 模型的效果对比如表 2。

Table 2. Comparison of experimental results
表 2. 对比实验结果

指标	基线模型	SOC 模型	改进幅度
Accuracy	47.36%	98.73%	+108.4%
F1 Score	31.20%	98.73%	+216.5%
MSE	2.106	0.0507	−97.6%

通过上述对比实验结果得到 SOC 模型在准确率(Accuracy)、F1 分数(F1 Score)和均方误差(MSE)等指标上都显著优于基线模型，分别实现了 108.4%、216.5%和 97.6%的改进。

5.3. 决策可视化

由于本模型为二元分类器，在进行决策可视化的过程中因为是多指标数据导致在绘制决策过程的时候只能进行两两特征对应展示决策可视化如图 5 所示。

针对上述情况无法进行全局展示决策过程可视化，因此采用了双层聚类决策，下述图 6 为 K-Means 聚类结果。

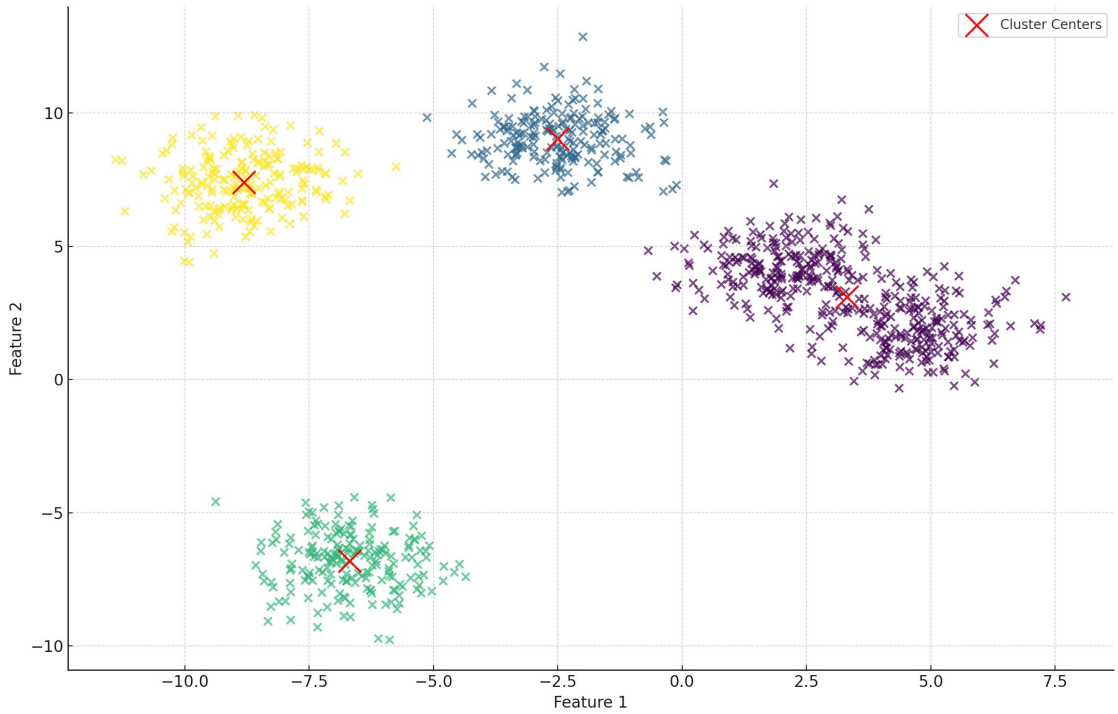

Figure 6. K-Means clustering visualization (4 clusters)
图 6. 聚类可视化情况图

结合聚类与分类器的决策边界，图 7 展示了分类效果。红色和蓝色区域分别对应上涨趋势和下跌趋势。

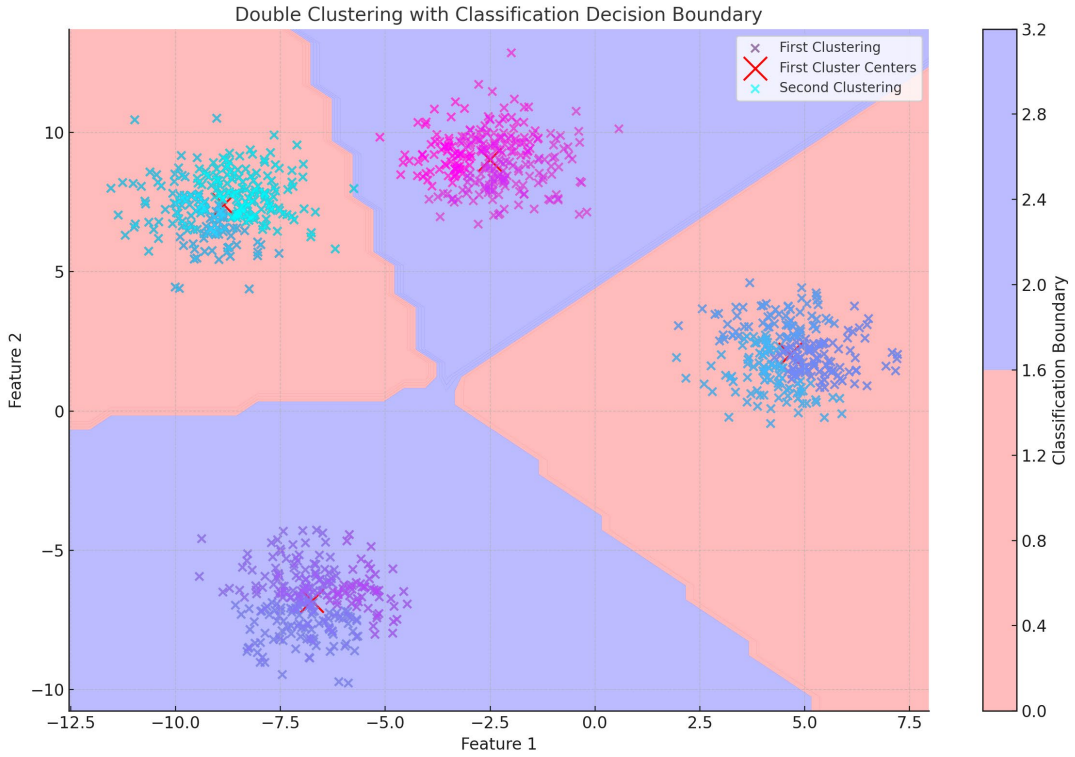

Figure 7. Double clustering with classification decision boundary
图 7. 双聚类与分类决策边界图

5.4. 在线学习自适应验证实验

通过分析在线学习技术累计的损失边界可以更加直观地展示模型的在线学习自适应性。与其他在线学习分类器不同的是本模型应用在高频交易预测涨幅中，涵盖多种特征提供有用的边界十分困难，为了评估 SOC 模型的自适应性能力，定义了一组动态数据。在每个阶段 t，正类和负类数据分别从两个多元正态分布中产生。其中正类数据为：$X_{\text{pos}} \sim N(\mu_t, \Sigma_t)$，负类数据为 $X_{neg} \sim N(-\mu_t, \sum_t)$。其中 $\mu_t \in \mathbb{R}^d$ 是阶段 t 的均值向量，其值逐步变化；$\sum_t \in \mathbb{R}^{d\times d}$ 是阶段 t 的协方差矩阵，数值可增大用以模拟数据的扩散。$\sum_t = \begin{bmatrix} \sigma_1^2 + t\cdot k & 0 \\ 0 & \sigma_2^2 + t\cdot k \end{bmatrix}$，$k > 0$ 控制协方差增长的速率。接着生成目标标签，其中 y_i 标签由分布的类别决定：$y_i = \begin{cases} +1, & if\ X_i \in X_{\text{pos}} \\ -1, & if\ X_i \in X_{\text{neg}} \end{cases}$，接着将生成的标签随机翻转达到添加噪声的效果 $y_i = -y_i$ for randomly selected i, with probability p_{noise}。每阶段生成 N_{phase} 个样本，阶段间分布通过 μ_t 和 $\sum_t$ 动态调整：其中阶段切换均值调整为 $\mu_{t+1} = \mu_t + \Delta\mu$，阶段切换协方差调整为 $\sum_{t+1} = \sum_t + \Delta\Sigma$。最终数据组合成：$X = \bigcup_{t=1}^{T} \{X_{\text{pos,t}} \cup X_{\text{neg,t}}\}$, $y = \bigcup_{t=1}^{T} \{y_{\text{pos,t}} \cup y_{\text{neg,t}}\}$。模型包含 lambda1、lambda2 和 eta 三个关键参数，lambda1 和 lambda2 分别为 L2 和协方差正则化强度，用以防止过拟合和约束特征相关性；eta 是学习率，控制权重更新步长，代码中分别设为 0.01、0.01 和 0.001。此外，Adam 优化器采用 0.9 的一阶动量衰减率与 0.999 的二阶动量衰减率，确保模型训练稳定并自适应调整参数更新步长。

为了分析 SOC 模型的在线自适应性，采取了每一百个样本生成一个带有决策边界的图，图 8 为节选出的决策边界可视化，从左到右依次进行变化。

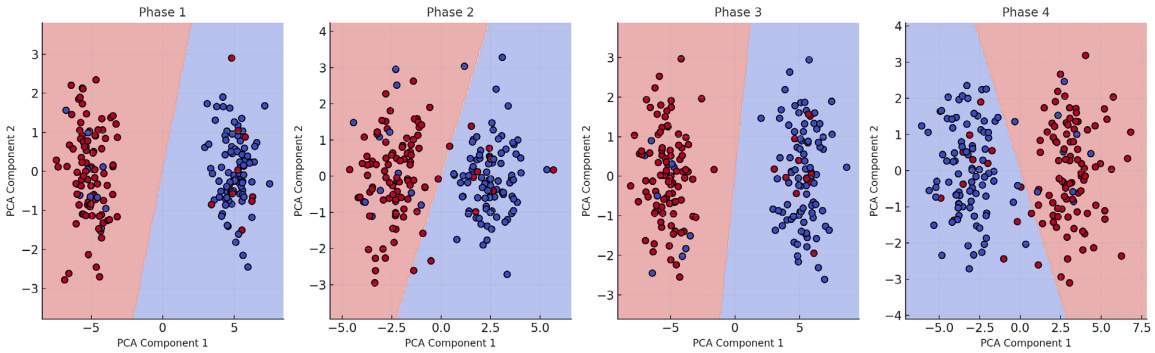

Figure 8. Time series visualization of features
图 8. 在线学习自适应验证决策过程可视化

虽然这些样本变化的数量较少，但是能够直观展示 SOC 模型对动态数据类别分布的变化。

6. 结论

SOC 模型提出结合结构化特征嵌入与动量优化技术，显著增强了模型的鲁棒性与实时学习能力，有效降低了因市场波动带来的预测误差。通过引入双层聚类方法，对高维特征进行降维分析，优化了指标的全局与局部特性提取，并通过决策边界的可视化显著提升了高频交易中分类决策的透明度与可解释性。实验结果表明，通过结构信息嵌入和动量优化，SOC 模型能够有效地处理高频数据并做出精确的市场趋势预测，相比于传统的 LSTM 神经网络模型，SOC 模型的效果和泛化能力提升巨大。

基金项目

2023 年江苏省大学生创新创业训练计划项目(202310305112Y)：嵌入结构信息的高频实时数据量化分析与预测方法；盐城工学院校企合作课题(G20240909007)；2024 年江苏省高校“人工智能通识教育教学改革研究”专项课题(2024AIGE54)。

参考文献

[1] 孙达昌, 毕秀春. 基于深度学习算法的高频交易策略及其盈利能力[J]. 中国科学技术大学学报, 2018, 48(11): 923-932.

[2] Chen, J. (2024) High-Frequency Trading (HFT): What It Is, How It Works, and Example. Investopedia. https://www.investopedia.com/terms/h/high-frequency-trading.asp

[3] 李志杰, 李元香, 王峰, 等. 面向大数据分析的在线学习算法综述[J]. 计算机研究与发展, 2015, 52(8): 1707-1721.

[4] Li, Y. and Long, P.M. (2000) The Relaxed Online Maximum Margin Algorithm. *NIPS Conference*, Denver, 29 November-4 December 1999, 498-504.

[5] Huang, G.-B., Liang, N.-Y., Rong, H.-J., Saratchandran, P. and Sundararajan, N. (2005) On-Line Sequential Extreme Learning Machine. *IASTED International Conference on Computational Intelligence*, Calgary, 4-6 July 2005, 232-237.

[6] Zhao, P., Wang, J., Wu, P., Jin, R. and Hoi, S. C. (2013) Fast Bounded Online Gradient Descent Algorithms for Scalable Kernel-Based Online Learning. *Pattern Recognition*, **45**, 495-499.

[7] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. and Singer, Y. (2006) Online Passive-Aggressive Algorithms. *Journal of Machine Learning Research*, **7**, 551-585.

[8] Gentile, C. (2001) A New Approximate Maximal Margin Classification Algorithm. *Journal of Machine Learning Research*, **2**, 213-242.

[9] Dredze, M., Crammer, K. and Pereira, F. (2008) Confidence-Weighted Linear Classification. *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, 5-9 July 2008, 264-271. https://doi.org/10.1145/1390156.1390190

[10] Crammer, K., Kulesza, A. and Dredze, M. (2009) Adaptive Regularization of Weight Vectors. *Advances in Neural*

Information Processing Systems, Vancouver, 7-10 December 2009, 414-422.

[11] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/bf00994018

[12] Tushare 团队. Tushare Pro: Python 金融数据接口[EB/OL]. http://tushare.pro/, 2024-12-26.

[13] 魏永合, 陈懿翀, 谷晓娇. 基于 SincNet 网络结合 L2 正则化的故障诊断[J]. 组合机床与自动化加工技术, 2024(8): 158-162.

[14] 徐龙飞, 郁进明. 基于 ML loss 的 SVM 分类算法[J]. 计算机应用研究, 2021, 38(2): 435-439.

[15] Malik, A.S., Boyko, O., Aktar, N. and Young, W.F. (2001) A Comparative Study of MR Imaging Profile of Titanium Pedicle Screws. *Acta Radiologica*, **42**, 291-293. https://doi.org/10.1080/028418501127346846

[16] 安琪, 梁宇飞, 王耀强, 等. 基于 K-Means 聚类与 PSO 特征优选 KNN 的分级负荷识别方法[J]. 河北科技大学学报, 2022, 43(3): 249-258.

[17] 陈斌, 谢文波, 付勋, 等. 基于改进局部密度的可扩展层次聚类算法[J]. 南京大学学报(自然科学), 2024, 60(3): 370-382.

[18] 蔡启航, 徐彬, 董晓迪. 利用语义增强提示和结构信息的知识图谱补全模型[J/OL]. 计算机科学, 2025: 1-17. http://kns.cnki.net/kcms/detail/50.1075.TP.20241028.1439.034.html, 2025-02-10.