计算机科学
COMPUTER SCIENCE

基于动态选择预测器的深度强化学习投资组合模型

赵淼，谢良，林文静，徐海蛟

引用本文

赵淼，谢良，林文静，徐海蛟. 基于动态选择预测器的深度强化学习投资组合模型[J]. 计算机科学，2024，51(4)：344-352.

ZHAO Miao, XIE Liang, LIN Wenjing, XU Haijiao. Deep Reinforcement Learning Portfolio Model Based on Dynamic Selectors [J]. Computer Science, 2024, 51(4): 344-352.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230100048

基于动态选择预测器的深度强化学习投资组合模型

赵 淼[1] 谢 良[1] 林文静[1] 徐海蛟[2]
1 武汉理工大学理学院 武汉 430070
2 广东第二师范学院计算机学院 广州 510303
(2668770405@qq.com)

摘 要 近年来,投资组合管理问题在人工智能领域得到了广泛的研究,但现有的基于深度学习的量化交易方法还存在一些问题。首先,对股票的预测模式单一,通常一个模型只能训练出一个交易专家,交易决策也仅根据模型预测结果作出;其次,模型使用的数据源相对单一,只考虑了股票自身数据,忽略了整个市场风险对股票的影响。针对上述问题,提出了基于动态选择预测器的强化学习模型(DSDRL)。该模型分为3部分,首先提取股票数据的特征并传入多个预测器中,针对不同的投资策略训练多个预测模型,用动态选择器得到当前最优预测结果;其次,利用市场环境评价模块对当前市场风险进行量化,得到合适的投资金额比例;最后,在前两个模块的基础上建立了一种深度强化学习模型模拟真实的交易环境,基于预测的结果和投资金额比例得到实际投资组合策略。文中使用中证500和标普500的日k线数据进行测试验证,结果表明,此模型在夏普率等指标上均优于其他参照模型。

关键词:强化学习;LSTM;投资组合;股市预测;神经网络
中图分类号 TP391

Deep Reinforcement Learning Portfolio Model Based on Dynamic Selectors

ZHAO Miao[1],XIE Liang[1],LIN Wenjing[1] and XU Haijiao[2]
1 College of Science,Wuhan University of Technology,Wuhan 430070,China
2 School of Computer Science,Guangdong University of Education,Guangzhou 510303,China

Abstract In recent years,portfolio management problems have been extensively studied in the field of artificial intelligence,but there are some improvements in the existing quantitative trading methods based on deep learning. First of all,the prediction model of stocks is single,usually a model only trains a trading expert,and the decision of trading is only based on the prediction results of the model. Secondly,the data source used in the model is relatively single,only considering the stock's own data,ignoring the impact of the entire market risk on the stock. Aiming at the above problems,a reinforcement learning model based on dynamic selection predictor(DSDRL) is proposed. The model is divided into three parts. Firstly,the characteristics of stock data are extracted and introduced into multiple predictors. Multiple prediction models are trained for different investment strategies,and the current optimal prediction results are obtained by dynamic selector. Secondly,the market environment evaluation module is used to quantify the current market risk and obtain the appropriate proportion of investment amount. Finally,based on the first two modules,a deep reinforcement learning model is established to simulate the real trading environment,and the actual portfolio strategy is obtained based on the predicted results and the proportion of investment amount. In this paper,the daily k-line data of China Securities 500 and S & P 500 are used for test verification. The results show that the proposed model is superior to other reference models in Sharpe rate and other indicators.

Keywords Reinforcement learning,LSTM,Investment portfolio,Stock market forecast,Neural networks

到稿日期:2023-01-10 返修日期:2023-05-30
基金项目:广东省自然科学基金(2020A1515011208);广州市基础研究计划基础与应用基础研究项目(202102080353);广东省普通高校自然科学类特色创新项目(2019KTSCX117)
This work was supported by the Natural Science Foundation of Guangdong Province,China(2020A1515011208),Basic and Applied Basic Research Project of Guangzhou Basic Research Program(202102080353) and Characteristic Innovation Project of Natural Science in General Colleges and Universities in Guangdong Province(2019KTSCX117).
通信作者:谢良(whutxl@hotmail.com)

1　引言

随着人工智能的发展,投资组合管理问题的研究取得了很大进展。投资组合管理旨在使多重风险资产的预期回报最大化,其中,预测股票的未来趋势在股票投资中起着关键作用[1-2]。准确的股票预测可以一定程度上提高投资组合收益,而股票价格受到多种因素的影响,因此,要实现高精度的股票趋势预测是一个具有挑战性的难题。

近年来,深度强化学习已逐渐在金融领域展现出了巨大潜力[3-5],这类方法的核心思想是通过智能与环境的交互来实现模型的训练和优化,不仅节省了获取带标签样本的时间,还赋予了模型自主学习的能力。但是现有的方法仍存在许多亟待解决的问题[6],一方面大多数方法存在预测单一问题,未充分考虑股票市场的周期性特点,代理往往只学习了一种市场预测方法。在市场迅速变化的过程中,代理学习到的预测方法很可能不适合当下,当市场差异过大时,代理往往以遗忘旧的预测方法为代价来学习新的预测方法。因此,代理很难在剧烈的波动中学习到合适的策略。另一方面,部分投资策略没有考虑资金比例问题,也就是没有考虑投资总金额在不同股票中的分配比例。这种情况下,深度学习虽然可以相对有效地预测股票的未来趋势,并且根据市场的周期性变化进行合理的排名,但是仍然缺乏对资金投入比例方面的考虑。

考虑到真实世界中投资策略和经济周期的变化,本文设计了一种能够有效考虑不同交易模式的股票预测方法,可以有效增强模型对股票的交易决策能力[7]。代理基于不同时期的市场环境学习合适的股票交易环境,提高对不同资产的预测准确性,从而指导后续代理对股票的投资资金占比。一般来说,大多数方法通过历史数据判断得到股票的未来趋势。不同的投资专家对市场有不同的见解,因此对同一只股票的未来趋势的判断也有所不同。基于这种思想,本文构建资产权重模块,一方面让代理在市场中通过学习不同的市场解读方法,来训练得到不同模式下的预测器。另一方面让代理提取时间序列特征并结合不同预测器在强化学习环境中的回报,为预测样本在多个模式下的预测器中挑选最佳的预测结果。同时,部分个股的走势与市场的走势呈现一定的相关性,本文构建市场环境评价模块,让代理学习动态地调整做空做多的资金比例。本文的主要贡献如下:

1)针对预测单一问题,在强化学习网络中设计多种交易模式存在下的股票预测方法,代理将学习这些交易模式并选择当下最合适的交易模式,以解决预测模式单一导致的代理在跨时段中表现不好的问题。

2)为了解决资金比例问题,将投资组合问题分为两个部分,即投资权重问题与投资比例问题,并且设计了两个不同的模块处理对应的问题,即资产权重模块和市场环境评价模块。市场环境评价模块针对市场数据进行分析,从而对资金比例进行调控。

3)通过在真实市场数据上的实验验证了所提方法的收益能力,并且证明了基于动态选择器的深度强化学习投资组合模型能有效地平衡风险和收益。在训练过程中,该方法在训练集上的回报表现随着迭代次数的增加稳定上升。

2　相关工作

2.1　时间序列预测

时间序列预测是一种可以利用大量关于过去行为的时间序列数据来做出长期或短期的预测的方法,在很多领域中有着广泛的应用[8-9]。常用的两种预测方法分为统计方法和机器学习。统计方法包括自回归综合移动平均(ARIMA)[10]和指数平滑状态空间(ETS)模型等[11]。股票趋势预测中,股票价格会在极短的时间间隔内[12]急剧波动,因此,这些方法在金融市场的表现中往往会有所下降。

基于机器学习的预测方案被用于股票预测中,且被证明能够在异构数据中使用强大的非线性模型来捕获更复杂的模式[13]。其中,深度学习技术,如卷积神经网络(CNN)[14]、循环神经网络(RNN)[15-16]和长短期记忆(LSTM)[17-18]在预测中取得了不错的效果。预测时的数据一般分为3种:历史价格数据、技术分析指标和市场情绪。市场情绪,例如2020年的新冠疫情对全球股市造成了严重的影响。一些研究者使用主题模型对新闻文章进行分析,提取出多个主题,并选择与股票回报最相关的主题,以解释股票回报[19]。Azhikodan等将深度强化学习方法与情绪分析(来自新闻媒体的外部信息)相结合,并证明了所提方法可以学习股票交易的技巧[20-21]。历史价格数据,例如Lawrence等[3]利用由堆叠的限制玻尔兹曼机器组成的自动编码器来从个股价格的历史中提取特征,并在没有广泛的输入特性手工工程的情况下发现股票动量效应的增强版本。技术分析指标[22],例如Fu等[23]演示了如何应用机器学习算法来区分好股票和坏股票,利用不同的深度学习算法,来构建244个技术和基本特征用于描述每只股票的特征,并且根据股票收益率波动方面的排名来标记股票。这些深度算法的效果都优于传统算法。

这些算法在股票预测上取得了良好的效果,但是在投资组合策略方面,往往只是采取简单的预测排名,并且未将交易费率等交易因素与网络的更新进行结合,因此在实际的交易中取得的效果有限。

2.2　强化学习

强化学习通过代理与环境的不断交互找到控制行动的最优策略。目前,一些研究者将深度神经网络与强化学习相结合学习交易策略[3,6,21,24-25]。这类工作大致可以分为两类:1)结合不同的网络以及机器学习中的改进方法,来增强代理在金融问题中的学习效果。例如Neuneier[26]首先在外汇中使用RL算法,获得了比以往的监督学习算法更好的性能。Moody等[27]提出了一种递归强化学习(RRL)算法,通过优化目标函数来训练投资组合的交易系统,其中,使用递归神经网络(RNN)从交易环境中编码状态。Rundo[4]结合长短期记忆(LSTM)网络在RL算法中分层预测外汇市场的中短期趋势,并开发一种算法以最大化高频交易的投资回报率。Sezer等[5]将一维金融时间序列转换成二维图像的数据表示,以便能够利用深度卷积神经网络的力量来实现算法交易系统。2)利用不同的强化学习算法并结合其优势,根据实际金融问题的特性,在环境的设计或者代理的设计上,使用更丰富的数据,从多维度的方向提取更多的特征来优化投资策略。例如,

Deng 等[28]提出了一种任务感知的时间反向传播方法，通过RNN有效地训练代理，用RNN表示实时金融信号，可以从状态中提取更高层次的特征，以期在中国市场获得了可观的利润。

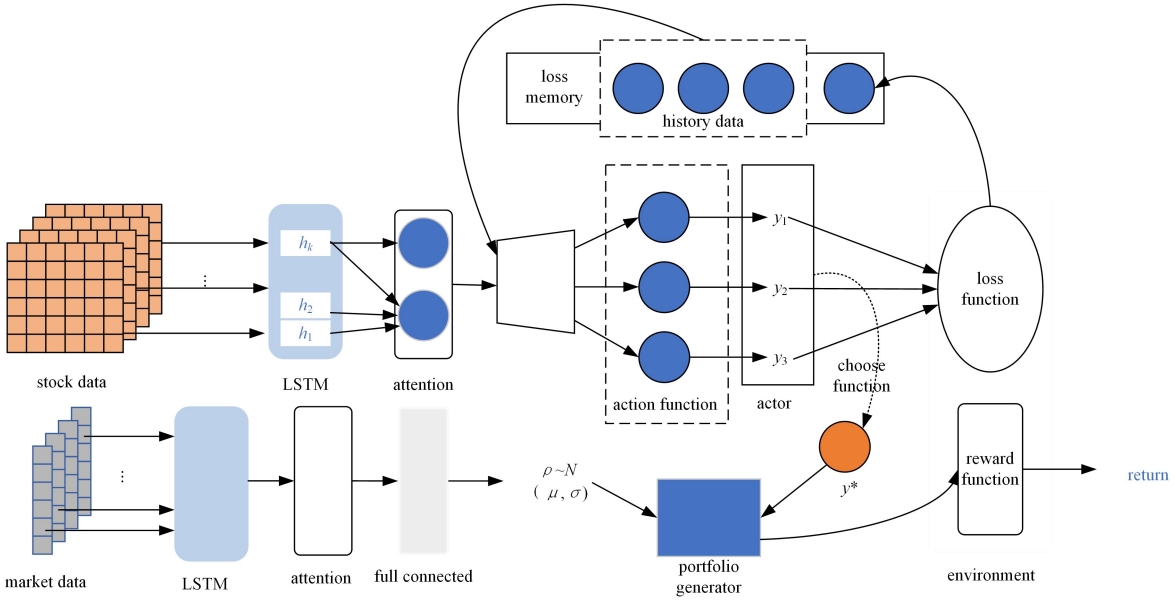

图1 DSDRL算法框架图

Fig. 1 Framework diagram of DSDRL algorithm

Yu 等[29]通过预测模块(IPM)、生成对抗数据增强模块(DAM)和行为克隆模块(BCM)的RL框架训练交易代理。然而，这些方法只训练了一种投资策略，难以在市场变化中捕捉到特点。针对此问题，本文设计了多种交易模式存在下的股票预测方法，代理将学习这些交易模式并且选择当下最合适的交易模式。

3 基于动态选择器的深度强化学习投资组合模型

本章提出了一个基于强化学习的投资组合模型，并且用夏普率作为算法的优化目标。如图1所示，该方法由多个部分组成，首先将资产 x^a 输入LSTM层并用注意力机制对隐藏层重新计算作为 k 个预测器的输入，最后得到未来股票的预测评分 y_k，同时将重新计算后的隐藏层结合预测损失记忆作为选择器的输入，并得到分类得分。选择器根据分类得分动态地选择最合适的预测评分，作为真实股票评分 v。同时，输入市场数据 x^b，通过全连接层将其转化为高斯分布的参数。最后通过投资组合管理，将股票评分 v 转化为做空比例 b 和做多比例 ω，并通过 ρ 选择做多资金比例。最后整个过程通过强化学习进行优化。

3.1 投资组合管理

3.1.1 马尔可夫过程

投资组合管理是根据当前市场状况将资源配置为一组金融资产的一个顺序决策过程，目的是在控制风险的同时实现回报最大化。这一过程符合马尔可夫过程 $M=\langle S,A,P,R\rangle$。

1) $S=((s_1,s_2,\cdots,s_t,\cdots)$ 为状态空间，描述股票当前状态的一组特性。一般来说，不同类型的信息，如历史价格变动、交易量、财务报表和情绪得分，可以作为当前状态。使用某一特定公司过去 T 天的收盘价序列和特征因子 $x^a\in R^{m\times k\times T}$ 与市场数据 $x^b\in R^{k\times T}$，即 $s_t=[x_t^a,x_t^b]$。其中 m 为股票数量，k 为特征因子数量。

2) $A=((a_1,a_2,\cdots,a_t,\cdots)$ 为动作空间，t 时刻执行动作 $a_t=[v,\rho]$，v 表示代理对不同股票的评分，并且根据评分选取适合做多的股票进行投资，计算出投资比例 $\omega_t\in R^{n\times T}$。其中 n 为选出的股票，ρ 为做多资金比例。

3)状态 s_t 根据过渡分布转化为 $s_{t+1}\sim P(s_{t+1}\mid s_t,a_t)=P(s_{t+1}\mid s_t)$，在投资组合问题中，代理对下一个状态的影响可以忽略不计。因此，状态 s_{t+1} 只受到 s_t 上一个状态的影响，这意味着，可以用真实的股票回报率去评价代理的行动。

4)代理获得的一个回报为 $r_t=R(s_t,a_t,s_{t+1})$，代理的目标是学习一个策略 π，最大化预期回报：

$$J=E_{p(\tau)}\left[\sum_{t=1}^{T}\gamma^{t-1}r_t\right] \tag{1}$$

其中，$\gamma\in[0,1]$为折扣因子，当 $\gamma=0$ 时，表示代理只关心一步的影响。

3.1.2 交易过程

本小节给出了多头投资和多空投资两种策略，分别对应SP500和CSI500两种市场。

1)多空投资

考虑一个同时包含多头和空头操作的交易策略来应对SP500市场。

(1)在 $t-1$ 持有期结束时，交易员将持有的股票卖出，得到当前的现金 $C_{t-1}^2=C_{t-1}^1+\omega_{t-1}\cdot\boldsymbol{p}_{t-1}$，其中，$\rho_{t-1}$ 为 $t-1$ 时期投资股票的比例，ω_{t-1} 为 $t-1$ 时期的做多投资组合比例，$\boldsymbol{p}_{t-1}$ 为 $t-1$ 时期的资产价格向量，C_{t-1}^1 为 $t-1$ 时期进行交易之后的现金。

(2)在做空的情况下，购买 $t-1$ 时期开始借入的股票并还给股票经纪公司，得到当前的现金 $C_{t-1}^3=C_{t-1}^2-b_{t-1}\cdot$

p_{t-1}，其中，b_{t-1}为 $t-1$ 时期的做空投资组合比例。

(3)根据交易员选择的做空比例 b_t 进行做空，得到当前的现金 $C_t^0=C_{t-1}^3+b_t\cdot p_t$。

(4)根据做多比例 ω_t 购买股票，得到当前的现金 $C_t^1=C_t^0-\omega_t\cdot p_t$。

2)多头投资

考虑一个只包含多头操作的的交易策略来应对 CSI500 市场。

(1)在 $t-1$ 持有期结束时，交易员将持有的股票卖出，得到当前的现金 $C_{t-1}^1=C_{t-1}^0+\omega_{t-1}\cdot \boldsymbol{p}_{t-1}$，其中，$\rho_{t-1}$为 $t-1$ 时期投资股票的比例，ω_{t-1}为 $t-1$ 时期的做多投资组合比例，$\boldsymbol{p}_{t-1}$为 $t-1$ 时期的资产价格向量，C_{t-1}^0为 $t-1$ 时期进行交易之后的现金。

(2)根据做多比例 ω_t 购买股票，得到当前的现金 $C_t^0=C_{t-1}^1-\omega_t\cdot p_t$。

3.1.3　优化目标

为了确保本文提出的策略平衡投资的风险和收益，使用夏普率作为优化目标。给定一个包含 T 个持有期的连续投资，其夏普比为：

$$SR=\frac{E(R_T)-R_f}{V_T} \tag{2}$$

其中，$E(R_T)$是每期的平均收益率，R_f 是无风险收益率，V_T 是用来衡量投资风险的波动率。

$$E(R_T)=\frac{1}{T}\sum_{i=1}^{T}(R_i-c_i) \tag{3}$$

$$R_T=\frac{C_T^1-C_{T-1}^1}{C_{T-1}^1} \tag{4}$$

其中，R_T 是第 T 时期的真实收益率，c_T 是第 T 时期的交易费率，c_T^2 是将第 T 时期未进行做空和做多的当前资金流。波动率 V_T 由式(5)计算得出：

$$V_T=\sqrt{\frac{\sum_{i=1}^{T}(R_t-\bar{R})^2}{T}} \tag{5}$$

其中，$\bar{R}=\sum_{i=1}^{T}R_i/T$ 是 R_T 的均值。

3.2　资产权重模块

资产权重模块是一个策略 $\pi_{\theta_1}^a(v|x^a;\theta_1)$，该模块的输入为资产数据 x^a，输出为资产评分 v。

首先使用长短期记忆(LSTM)层提取出资产的时间特征，再通过注意力机制在长时间跨度内有效地建模，将得到的信息传入选择预测器中对股票进行打分，选择预测器得到的结果中的一部分传入资产管理模块得到投资的股票以及相应权重，另一部分传入强化学习的环境中并将计算的结果存入损失记忆中。

3.2.1　资产打分

输入资产 $x^a=[x_0^a,x_1^a,\cdots,x_k^a]$用 LSTM 提取，计算公式如下：

$$h_{k-i+1}=LSTM(h_{k-i},x_k^a),i=0,1,2,\cdots,k-T+1 \tag{6}$$

其中，$h_{k-i+1}\in R^{C\times m}$表示 LSTM 在第 $k-i+1$ 时刻编码的隐藏状态，C 为隐藏层的维数。最后一个隐藏状态 h_k 可以看作是输入信号的全局表示。为了在长时间跨度内有效地建模，采用时间注意机制自适应建模非线性的非线性关系，注意权重计算式为：

$$e_k=V_e^{\mathrm{T}}\tanh(W[h_T;h_k]+Ux_k^m) \tag{7}$$

$$\alpha_k=\frac{\exp(e_k)}{\sum_{i=1}^{K}\exp(e_i)} \tag{8}$$

其中，$V\in R^C$，$W\in R^{C\times 2C}$，$U\in R^{C\times C}$都是需要学习的参数。最后隐藏状态重新计算为：

$$\hat{h}_k^a=\sum_{i=k-T+1}^{k}\alpha_i\cdot h_i \tag{9}$$

将得到的隐藏状态 $\hat{h}_k^a\in R^{C\times m}$和损失记忆 L_k 结合作为 actor 的输入，由式(10)给出：

$$v=actor(\hat{h}_k^a,L_k) \tag{10}$$

actor 层由两个模块组成：选择器和预测器。

3.2.2　选择器和预测器

预测器由 m 个网络组成，每个网络输出一个预测的结果，模拟 m 个专家的决策结果，由以下公式给出：

$$y_k^i=actor_i(\hat{h}_k^a),i=1,2,\cdots,m \tag{11}$$

$$y_k=((y_k^1,y_k^2,\cdots,y_k^m) \tag{12}$$

其中，$actor_i$ 采用全连接层。将第 k 时刻预测器的预测结果传入损失函数中，将计算的结果保存到损失记忆 L 中，计算式如下：

$$l_k^i=loss(y_k^i),i=1,2,3,\cdots,m \tag{13}$$

$$l_k=[l_k^1,l_k^2,\cdots,l_k^m] \tag{14}$$

$$L\leftarrow[L,l_k] \tag{15}$$

$loss$ 函数将计算所有动作 y_m 的回报与真实回报的差值，不参与网络的更新。其将网络的结果输出，并与强化学习的环境进行交互，计算出预测损失，存入预测损失记忆中。

选择器是一个动态选择器，它将历史的预测记忆损失和当前市场状态信息相结合，选择当前时间最合适的投资专家，其计算式如下：

$$p_1,p_2,\cdots,p_m=Routing(\hat{h}_k,L_k) \tag{16}$$

$$L_k\triangleq[l_{k-1},l_{k-2},\cdots,l_{k-t}] \tag{17}$$

其中，m 为分类器分类个数，$Routing$ 为全连接层，p_m 为分类得分，L_k 为 $k-T$ 时刻到 $k-1$ 时刻的记忆损失。最后根据 p_i 选择最合适的资产评分 v，计算式如下：

$$z=\arg\max(p_i),i=1,2,\cdots,m \tag{18}$$

$$v=y_k^z \tag{19}$$

上述过程可以描述为：每个 $actor_i$ 会得到一个动作，在不断与环境交互的过程中，会得到一个带有得分的轨迹 $\tau_i=(s_0^i,a_0^i,p_0^i,s_1^i,a_1^i,p_1^i,\cdots)$，$actor$ 在每步都会选择得分最大的动作，$a_j^*=\arg\max(R(a_j^i)),i=1,2,\cdots,m$。此外，每个动作可以在环境中计算出其真实回报，这些回报会存入损失记忆中，为之后选择器选取最佳动作提供输入。在训练过程中，每次迭代开始时，损失记忆需要清空，并且用 $m\times T$ 大小的零向量对其进行填充，当新的损失进入损失记忆时，最老的记忆损失将会被排出记忆损失。整个过程如图 2 所示。

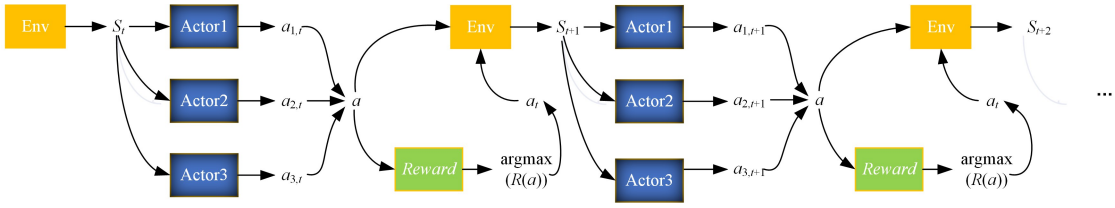

图 2 强化学习的轨迹图

Fig. 2 Track diagram of reinforcement learning

3.3 市场环境评价模块

在以往基于强化学习的投资模型中，最终的投资策略仅仅对股票进行分析，而忽略了市场的变化，在市场变坏的情况下，依然将所有的资金用于投资股票。因此，在进行投资时，不能有悖于市场的逻辑。在股市下跌时，应投资更多的钱进行做空或者持有现金；在股市上涨时，应投资更多的钱进行做多。通过这种方式来动态地调整资金。故为了平衡风险和收益，我们将市场状况作为输入，对资金做动态调整。

市场环境评价是一个策略函数 $\pi^b_{\theta_2}(\rho|x^b;\theta_2)$，输入为市场数据 x^b，输出为持有资金比例 ρ。

对于市场数据 x^b，用 LSTM 层提取出隐藏状态，再用时间注意力机制抓取更长时间的信息 $\hat{h}^b_k \in R^{c\times m}$，再用全连接层得 μ 和 σ，公式如下：

$$\mu,\sigma = U_m \cdot \hat{h}^b_k + b_m \tag{20}$$

3.4 投资组合策略

1)资金管理：因为市场数据具有不稳定性，在某一个状态采取某一个动作时，对每轮投资结束时的所有的回报进行统计，得到一个分布。如图 3 所示，不同的策略回报分布不同。因此，在训练时，取 $\rho\sim N(\mu,\sigma^2)$，通过取 $\mu\in[0.1,0.9]$，$\sigma\in[0,0.1]$，来保证 $\rho\in[0,1]$。在测试时，取 $\rho=\mu$，避免测试时动作的随机，并且这意味着，无论市场表现得多差，都进行持仓，无论市场表现得多好，都不满仓。

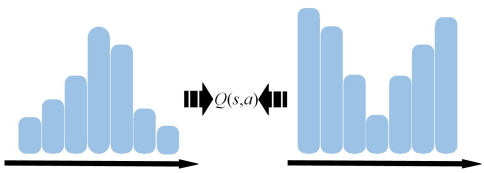

图 3 不同策略的分布

Fig. 3 Distribution of different strategies

2)投资组合：在计算资产得分 v 后，使用 softmax 函数将得分 v 转化为投资组合的权重。将资产得到的评分进行排名，选取排名靠前的股票进行做多，并归为集合 G；选取排名靠后的股票进行做空，并归为集合 H。做多和做空比例如下：

$$\omega_i = \begin{cases} \dfrac{\exp(v_i)}{\sum_{j\in G}\exp(v_j)}, & i\in G \\ 0, & i\notin G \end{cases} \tag{21}$$

$$b_i = \begin{cases} \dfrac{\exp(v_i)}{\sum_{j\in G}\exp(v_j)}, & i\in H \\ 0, & i\notin H \end{cases} \tag{22}$$

3.5 策略优化

对投资组合策略 $\pi((a\mid x^a,x^b;\theta)$ 进行优化，其中策略 π 分为两部分：

1)$\pi^a_{\theta_1}(v|x^a;\theta_1)$ 通过输入股票数据 x^a 得到做空做多的股票比例，持有期 t 的回报为 $r_t=\dfrac{p_t}{p_{t-1}}\cdot\pi^a_{\theta_1}$。对于强化学习的轨迹 τ，通过优化网络参数 θ_1 使得达到目标的最大值的计算式如下：

$$\theta^* = \arg\max_{\theta_1} J^a(\theta) = \arg\max_{\theta_1} \sum_{\tau}\prod_{t=1}^{T}\frac{p_t}{p_{t-1}}\cdot\pi^a_{\theta_1} \tag{23}$$

2)$\pi^b_{\theta_2}(\rho|x^b;\theta_2)$ 通过输入股票数据 x^b 得到做多资金的比例。对于强化学习的轨迹 τ，给定奖励 $R(\tau)$，优化网络参数 θ_2 的计算式如下：

$$\theta^* = \arg\max_{\theta_2} J^b(\theta) = \arg\max_{\theta_2} \sum_{\tau} R(\tau)\cdot\pi^b_{\theta_2}(\tau) \tag{24}$$

对式(23)和式(24)进行加权求和得到最终的目标函数：

$$J(\theta) = J^a(\theta_1) + \omega J^b(\theta_2) \tag{25}$$

算法 1 描述了基于动态选择预测器的强化学习算法的整个流程。

算法 1 基于动态选择预测器的强化学习算法

输入：x^a，x^b

输出：θ_1，，θ_2，

1. 初始化损失记忆 M，经验回放 D
2. 初始化行为网络 θ_1(股票评分)，θ_2(投资金额)
3. for i←1 to m do
   3.1. 初始化市场特征 x^a、历史趋势特征 x^b
   3.2. 预处理状态
   $h^a_t = \mathrm{LSTM}(X^a)$，$h^b_t = \mathrm{LSTM}(X^b)$，$\mu,\sigma = \mathrm{FC}(h^b_t)$
   3.3. for 每次环境更新 do
      3.3.1. 计算 $y_k = \mathrm{actor}(h^a_t)$，$l_k = \mathrm{loss}(y_k)$，并存入损失记忆 M=[M;$l_k$]
      3.3.2. 计算股票评分 $v\sim\pi_{\theta_1}(v|x^a_t)$
      3.3.3. 计算金额投资 $\rho\sim\pi_{\theta_2}(\rho|x^b_t)$
      3.3.4. 更新下一个状态
      $s_{t+1}\sim p(s_{t+1}|s_t,\rho,v)$，$s_t=[x^a_t,x^b_t]$
      3.3.5. 更新经验回放
      D←D∪{$(s_t,\rho,v,r(s_t,\rho,v),s_{t+1})$}
   3.4. end for
   3.5. for j←1 to n do
      3.5.1. 从经验回放中采集数据利用式(25)更新网络参数 $\theta\leftarrow\theta+\eta\nabla J(\theta)$
   3.6. end for
4. end for

4 实验结果与分析

本章首先介绍数据处理、基线和度量标准；然后在不同的市场上用基线来评本文的算法；最后通过消融实验，分析

了资产权重模块中的动态选择预测器和市场环境评价模块对奖励的影响。

4.1　数据集

实验中使用的股票数据来自中国中证 500 成份股和美国标普 500 的成分股，如表 1 所列。将中证 500 指数数据中数据有空缺的股票剔除，并且根据市值筛选出前五十只股票。具体来说，将开盘价、收盘价、高价、低价除以前一天的收盘价，以避免使用未来信息。对标普 500 的处理方式也是一样，并且，使用的 6 个特征因子是从 Qlib 平台提供的 alpha158 因子选出的[29]，分别是："KMID""KLEN""MA5""BETA5""RSV5""VMA5"。

4.2　训练集设置

对于模型中的一些超参数，设置持仓周期为 30 天，包含持仓 29 天，交易期为 1 天。强化学习的学习率为 1×10^{-5}，折扣因子 gamma 设置为 0.96。为了更加真实地模拟投资者的策略，预测分类器个数设为 $m=3$。并且对所有权值使用预处理初始化，并将 LSTM 输出空间设置为 128，每次输入数据批次为 128。使用 ReLU 激活函数和 Adam 优化器，设置最大迭代次数为 500 次。初始资金设置为十万元，交易费率为 0.1%。

为了更好地发现股票数据中的顺序模式，采用滚动训练、验证和测试的方法来学习。训练过程如图 4 所示，描述为：1）训练：在一段时间内（以交易日为单位）选择样本，在多个时期内进行训练；2）验证：选择训练窗口之后的样本进行验证，如果在多个时期内验证损失没有显著下降，则停止训练过程；3）测试：选择验证之后的样本进行测试；4）向前滚动窗口：当测试过程结束时，将验证样本放入训练数据集，并将窗口向前滚动；5）重复：重复过程 1）中的训练。

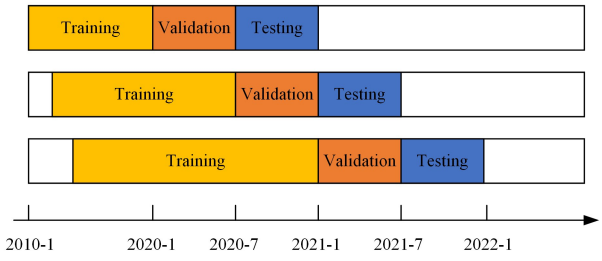

图 4　滑动训练的数据集分割

Fig. 4　Dataset segmentation for sliding training

在实验中使用网格搜索来调整训练、验证和测试周期的长度，调整的训练窗口长度为{1 200，1 500，1 800，2 100，2 400}，验证窗口长度为{120，150，180，210，240}，测试窗口长度在{120，150，180，210，240}。如表 1 所列，中证 500 指数的训练集为 2010-01－2019-02，验证集为 2019-02－2020-02，测试集为 2020-02－2022-03。

表 1　数据集的信息

Table 1　Dataset information

市场名称	股票数量	训练集	测试集	验证集
中证 500	50	2010-01－2019-02	2020-02－2022-03	2019-02－2020-02
标普 500	50	2014-01－2019-10	2020-06－2022-04	2019-10－2020-06

4.3　评价指标

实验中使用了 5 个指标，可分为 3 类：1）利润标准，包括年化回报率（ARR）；2）风险标准，包括年化波动率（AVol）和最大回撤（MDD）；3）风险利润标准，包括年化夏普比率（ASR）和风险比率（CR）。AVol 和 MDD 越低越好，其余指标越高越好。

4.4　对比方法

为了能够有效地评估本文提出的基于动态选择器的深度强化学习投资组合模型，实验对比了以下 5 种方法：

Market：从第一天开始买入并且持有。

FinRL[30]：一个用于量化金融自动交易的深度强化学习库，选取其中的 DDPG 和 PPO 方法，并用效果最好的方法。

TRA[7]：根据样本属于不同的时间段将样本分配给不同的预测器，得到对不同股票的动态评分来实现投资组合策略。

DT[31]：由市场评分单元、股票评分单元和投资组合管理器组成，是一种强化学习的投资组合方法。

DPRL[29]：由预测模块（IPM）、生成对抗数据增强模块（DAM）和行为克隆模块（BCM）的强化学习框架来训练交易代理，是一种基于模型的动态投资组合优化深度强化学习方法。

DSDRL：对股票评分时，用动态选择模块选择当前时间段最合适的评分策略，再结合市场环境评价模块，得到投资组合策略。

4.5　实验结果

本节在 CSI500 和 P500 上进行了许多实验来证明所提出的 DSDRL 算法的有效性。

4.5.1　与基线方法的比较

从表 2 中可以看到，基于深度强化学习（DRL）的方法中 DSDRL 方法的表现更好。通过比较其他评价指标可知，DSDRL 在两个数据集上的累积收益都高于其他方法，夏普比率有了很大的提高，这意味着投资者承受相同的每单位风险可以获得更多的回报。在风险评估指标（AVOL，MDD）上，DSDRL 的性能基本优于其他方法，证实了该策略的有效性和稳健性。

表 2　中国和美国股市的不同方法的回测表现

Table 2　Backtesting performance of all methods in Chinese and American stock markets

Algorithms	中证 500					标普 500				
	ARR/%	ASR	AVol	*MDD*%	CR	*ARR*/%	ASR	AVol	*MDD*%	CR
Market	9.72	0.44	0.220	22.00	0.813					
TRA	12.70	0.72	0.176	25.00	1.120	8.81	0.419	0.210	47.1	0.931
DDPG	23.50	0.98	0.239	30.60	1.450	20.60	0.789	0.261	23.6	2.910
PPO	31.10	1.17	0.265	22.90	1.780	39.10	2.010	**0.195**	22.0	3.310
DT	36.80	1.26	0.292	33.20	2.240	42.50	1.560	0.272	25.7	3.810
DPRL	7.51	0.45	**0.168**	21.30	0.728	45.20	2.080	0.217	21.3	4.100
DSDRL	**47.90**	**2.23**	0.221	**8.57**	**5.760**	**53.20**	**2.550**	0.208	**10.5**	**4.600**

图 5 给出了不同算法在 CSI500 的累积收益与交易日的关系。从市场收益变化中可以看到，虽然市场多次下跌，但是这一时期总体上呈现出不错的上升趋势。虽然 TRA 方法没有跑赢买入后持仓不动的方法，该方法只是训练多个预测器进行预测，但大多数策略的投资组合的价值持续攀升。DT 方法在初期的收益最高，但是在后续市场的表现略显不足。而 DSDRL 算法虽然不能保证在每一个时刻的收益达到最高，但是在总体表现上能做到稳步上升，达到最佳的收益效果。

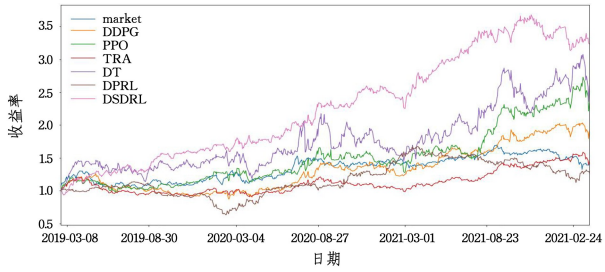

图 5 在 CSI500 上不同算法的表现

Fig. 5 Performance of different algorithms on CSI500

图 6 给出了不同算法在 SP500 的累计财富与交易日的关系。这段时期 SP500 市场优于 CSI500，并且结合做空机制，这些强化学习算法整体上都取得了很高的累积收益。在 2019 年底之前，这些强化学习算法取得的收益很接近，但是在 2020 年 3 月份之后，这些算法的收益开始出现分歧。而 DSDRL 算法在这段时期表现最好，通过市场环境评价模块可以有效规避风险时期，在短期内快速调整，并且完成超额的收益。如图 7 所示，市场环境评价通过控制资金比例来规避风险。

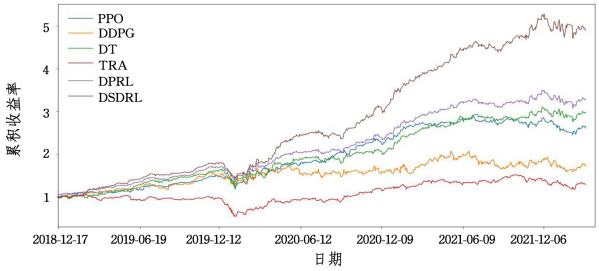

图 6 在 SP500 上不同算法的表现

Fig. 6 Performance of different algorithms on SP500

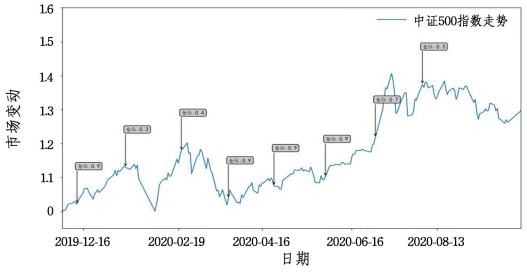

图 7 市场环境评价在市场崩溃下的分析

Fig. 7 Analysis of market environment evaluation under market collapse

4.5.2 算法性能表现

由于金融市场的特性，许多强化学习方法用于投资组合问题时都存在一个问题，即训练过程中的波动很大，很难完成收敛，特别是一些针对个股投资提出的强化学习算法。从图 8 中可以看到，DSDRL 在训练过程中总体效果稳步上升，并在 120 次迭代时达到一个峰值，后续的训练也没有产生剧烈的波动。

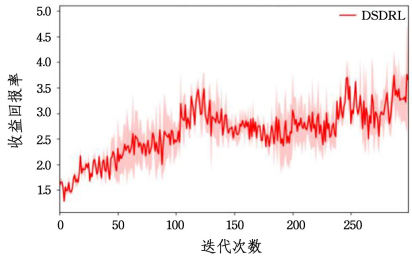

图 8 算法训练表现

Fig. 8 Algorithm training performance

4.5.3 消融实验

为充分解释动态选择预测器模块和市场环境评价模块的重要性，在同一数据集中进行了两部分的消融实验，以下为两部分消融实验的分析：

1)动态选择预测器的消融实验：本文训练了多种预测方法，并且由智能体选择最优的预测结果，根据这些预测结果进行投资组合。本实验去掉了动态选择预测器模块，这意味着模型只会训练一种预测方法，然后由智能体根据单个预测结果进行投资组合。我们用 DSDRL-1 表示该消融实验。

2)市场环境评价的消融实验：为了让智能体考虑市场环境的影响，让模型根据市场数据进行资金配置，本实验去掉了市场环境评价，这意味着智能体不会考虑资金配置，所有的资金将会用来做多。我们用 DSDRL-2 表示该消融实验。

消融实验的结果如表 3 所列，并且对结果保留 3 位有效数字。在去掉动态选择预测器模块和市场环境评价模块之后，DSDRL 方法取得的收益并不好，并且规避风险的能力较差。因此，动态选择预测器模块和市场环境评价的引入，能够有效地帮助代理做出更好的决策。

表 3 消融实验的回测表现

Table 3 Backtest performance of ablation test

指标	*ARR*/%	ASR	AVol	*MDD*/%	CR
DSDRL-1	27.4	1.07	0.255	21.20	1.29
DSDRL-2	34.6	1.43	0.242	13.30	2.59
DSDRL	47.9	2.23	0.221	8.57	5.76

4.5.4 市场环境评价表现分析

市场环境评价模块可以针对市场变化调整投入资金比例，在遇到极端市场的情况下，能有效减少投入资金，从而规避风险。因此，本小节在极端市场情况下对市场环境评价模块进行分析。

COVID-19 于 2019 年末被发现并于 2020 年初爆发，COVID-19 严重阻碍了全球的经济发展，全球股市也遭到了冲击，CSI500 也不例外。在 2020 年 2 月 3 日，中国春节法定假日之后，CSI500 遭遇了大跌，而 DSDRL 中的市场环境评价模块根据市场情况减少了投入资金的比例。2020 年 7 月 6 日之后，受注册制改革、大量资金的流入等积极因素的影响，中国市场各类指数大涨，CSI500 也不例外，其迎来了一波强大

的走势,DSDRL 的市场环境评价模块在此期间上调了投入资金的比例。图 8 反映了市场环境评价模块的作用,在交易间隔为一个月时,市场环境评价模块会判断市场情况,从而调整投入资金比例(图中简写为仓位),该比例的范围在[0.1,0.9],并且对该值进行了取整操作,以便在图中更加直观地显示结果。在 CSI500 大跌之前,市场环境评价模块将仓位(投入资金的比例)降低到 0.3,并在 2020 年 6 月底市场回暖时增加投资比例至 0.7。因此,市场环境评价模块在规避风险的同时,保证了一定的收益。

结束语　本文提出的 DSDRL 是一个基于动态选择预测器的强化学习模型,它结合了资产权重模块和市场环境评价模块。具体来说,资产权重模块主要由预测器和选择器两个板块构成。将股票数据的特征输入到多个预测器中得到不同的预测结果,再由选择器动态地选择出最优预测结果。市场环境评价则根据当前市场风险进行量化,对资金的比例进行配置来规避风险。将两个模块的结果传入投资组合管理模块中,得到最终的交易策略。在本文提出的损失函数的指导下,代理尽可能采取不同的投资行动,同时最大化自己的奖励。在股票数据上进行的实验表明,所提方法在夏普比率等指标上优于大多数现有的先进方法。

未来,我们可以在本文模型基础之上做进一步研究和改进,例如,动态选择预测器中可以加入更多不同的预测方法,模拟更多的交易策略使得模型能在更复杂的交易环境下做出更好的决策。

参 考 文 献

[1] QIN Y,SONG D,CHEN H,et al. A dual-stage attention-based recurrent neural network for time series prediction[J]. arXiv:1704.02971,2017.

[2] XIE C,RAJAN D,CHAI Q. An interpretable Neural Fuzzy Hammerstein-Wiener network for stock price prediction[J]. Information Sciences,2021,577:324-335.

[3] TAKEUCHI L,LEE Y Y. Applying deep learning to enhance momentum trading strategies in stocks[R]. Technical Report. Stanford,CA,USA:Stanford University,2013.

[4] RUNDO F. Deep LSTM with reinforcement learning layer for financial trend prediction in FX high frequency trading systems[J]. Applied Sciences,2019,9(20):4460.

[5] SEZER O B,OZBAYOGLU A M. Algorithmic Financial Trading with Deep Convolutional Neural Networks:Time Series to Image Conversion Approach[J]. Applied Soft Computing,2018,70:525-538.

[6] YANG H,LIU X Y,ZHONG S,et al. Deep reinforcement learning for automated stock trading:an ensemble strategy[C]// Proceedings of the first ACM International Conference on AI in Finance. 2020:1-8.

[7] LIN H,ZHOU D,LIU W,et al. Learning multiple stock trading patterns with temporal routing adaptor and optimal transport[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021:1017-1026.

[8] MILLER C,ARJUNAN P,KATHIRGAMANATHAN A,et al. The ASHRAE great energy predictor III competition:Overview and results[J]. Science and Technology for the Built Environment,2020,26(10):1427-1447.

[9] LIANG Y,KE S,ZHANG J,et al. Geoman:Multi-level attention networks for geo-sensory time series prediction[C]// IJCAI. 2018:3428-3434.

[10] GILBERT K. An ARIMA supply chain model[J]. Management Science,2005,51(2):305-310.

[11] DURBIN J,KOOPMAN S J. Time series analysis by state space methods[M]. OUP Oxford,2012.

[12] CAETANO M A L,YONEYAMA T. Characterizing abrupt changes in the stock prices using a wavelet decomposition method[J]. Physica A:Statistical Mechanics and its Applications,2007,383(2):519-526.

[13] DING Q,WU S,SUN H,et al. Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction[C]// IJCAI. 2020:4640-4646.

[14] KVAMME H,SELLEREITE N,AAS K,et al. Predicting mortgage default using convolutional neural networks[J]. Expert Systems with Applications,2018,102:207-217.

[15] CHEN W,YEO C K,LAU C T,et al. Leveraging social media news to predict stock index movement using RNN-boost[J]. Data & Knowledge Engineering,2018,118:14-24.

[16] PIAO G,BRESLIN J G. Financial aspect and sentiment predictions with deep neural networks:an ensemble approach[C]// Companion Proceedings of the The Web Conference 2018. 2018:1973-1977.

[17] SPILAK B. Deep neural networks for cryptocurrencies price prediction[D]. Berlin:Humboldt-Universität zu Berlin,2018.

[18] SI W,LI J,DING P,et al. A multi-objective deep reinforcement learning approach for stock index future's intraday trading[C]// 2017 10th International Symposium on Computational Intelligence and Design(ISCID). IEEE,2017:431-436.

[19] GLASSERMAN P,KRSTOVSKI K,LALIBERTE P,et al. Choosing news topics to explain stock market returns[C]// Proceedings of the First ACM International Conference on AI in Finance. 2020:1-8.

[20] BRAHIMI B,TOUAHRIA M,TARI A. Improving sentiment analysis in Arabic:A combined approach[J]. Journal of King Saud University — Computer and Information Sciences,2021,33(10):1242-1250.

[21] AZHIKODAN A R,BHAT A G K,JADHAV M V. Stock trading bot using deep reinforcement learning[C]// Innovations in Computer Science and Engineering:Proceedings of the Fifth ICICSE 2017. Singapore:Springer,2019:41-49.

[22] GRACE A. Can deep learning techniques improve the risk adjusted returns from enhanced indexing investment strategies[J/OL]. https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1129&context=scschcomdis.

[23] FU X Y,DU J H,GUO Y F,et al. A machine learning framework for stock selection[J]. arXiv:1806.01743,2018.

[24] CHEN L,GAO Q. Application of deep reinforcement learning on automated stock trading[C]// 2019 IEEE 10th International

Conference on Software Engineering and Service Science (ICSESS). IEEE, 2019: 29-33.
[25] PUN C S, WANG L, WONG H Y. Financial thought experiment: a gan-based ap-proach to vast robust portfolio selection [C]// Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'20). 2020: 451-458.
[26] NEUNEIER R. Optimal asset allocation using adaptive dynamic programming[J]. Advances in Neural Information Processing Systems, 1995, 8: 952-958.
[27] MOODY J, SAFFELL M. Reinforcement learning for trading [J]. Advances in Neural Information Processing Systems, 1998, 11: 917-923.
[28] DENG Y, BAO F, KONG Y, et al. Deep direct reinforcement learning for financial signal representation and trading [J]. IEEE transactions on Neural Networks and Learning Systems, 2016, 28(3): 653-664.
[29] YU P, LEE J S, KULYATIN I, et al. Model-based Deep Reinforcement Learning for Dynamic Portfolio Optimization[J]. arXiv: 1901.08740, 2019.
[30] LIU X Y, YANG H Y, CHEN Q, et al. Finrl: A deep reinforcement learning library forautomated stock trading in quantitative finance[J]. arXiv: 2011.09607, 2020.
[31] WANG Z C, HUANG B W, TU S, et al. Deeptrader: a deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 643-650.

ZHAO Miao, born in 1998, postgraduate. His main research interests include machine learning and quantitative trading.

XIE Liang, born in 1987, Ph.D, associate prefessor. His main research interests include multimedia retrieval and machine learning.

（责任编辑：何杨）

CCF 科学普及工作委员会召开 2024 年度工作会议

2024 年 3 月 17 日上午，CCF 科学普及工作委员会（以下简称科普工委）2024 年度工作会议在中国科学院计算所召开。科普工委 24 位新一届执行委员，通过线上和线下的形式参加了会议。会议由科普工委主任王元卓主持。

王元卓首先介绍了科普工委在 2023 年度的工作成果。2023 年，科普工委利用能力矩阵，以“群星计划”为核心，使科普图书、科普视频、走进中小学、科普教育、科普活动和科普教育基地各项工作产生联动，从静态的图书内容到动态的科普视频，从“走出去”进中小学做科普报告，到成体系的科普教育，再到“请进来”组织科普活动，直至科普活动常态化延伸出科普教育基地，取得了良好的效果。CCF 被科协评为 2023 年度全国学会科普工作优秀单位，且在 10－12 月连续三次位列科协系统科普新媒体传播榜前三，进入科普第一梯队。

随后，科普工委各工作组的召集人分别介绍了 2023 年度的工作成果和 2024 年度的工作规划。与会执委们就 2024 年工作规划进行了积极发言和热烈讨论。王元卓提出，2024 年科普工委将继续在群星计划、科普图书、科普视频、走进中小学、科普教育、科普活动、科普基地等方面深化，做大做强。2024 年，科普工委还规划了几个重点项目和活动，包括“群星计划四周年”嘉年华活动，筹建科普训练营，拓展影视渠道，与央视合作纪录片，尝试建立 CCF 产学合作科普基金等。他希望科普工委的委员们都能积极报名参与到各个工作组的工作中，在新的一年里共同努力，继续推动 CCF 科普事业不断向前发展。

据 CCF 微信公众号