软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(3):845−864 [doi: 10.13328/j.cnki.jos.005689]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于强化学习的金融交易系统研究与发展*

梁天新, 杨小平, 王 良, 韩镇远

(中国人民大学 信息学院,北京 100872)

通讯作者: 王良, E-mail: wangliang@ruc.edu.cn

摘 要: 近年来,强化学习在电子游戏、棋类、决策控制等领域取得了巨大进展,也带动着金融交易系统的迅速发展.金融交易问题已经成为强化学习领域的研究热点,特别是股票、外汇和期货等方面具有广泛的应用需求和学术研究意义.以金融领域常用的强化学习模型的发展为脉络,对交易系统、自适应算法、交易策略等方面的诸多研究成果进行了综述.最后讨论了强化学习在金融领域应用中存在的困难和挑战,并对今后强化学习交易系统发展趋势进行展望.

关键词: 强化学习;深度学习;金融交易系统;自适应算法;交易策略

中图法分类号: TP18

中文引用格式: 梁天新,杨小平,王良,韩镇远.基于强化学习的金融交易系统研究与发展.软件学报,2019,30(3):845−864. http://www.jos.org.cn/1000-9825/5689.htm

英文引用格式: Liang TX, Yang XP, Wang L, Han ZY. Review on financial trading system based on reinforcement learning. Ruan Jian Xue Bao/Journal of Software, 2019,30(3):845−864 (in Chinese). http://www.jos.org.cn/1000-9825/5689.htm

Review on Financial Trading System Based on Reinforcement Learning

LIANG Tian-Xin, YANG Xiao-Ping, WANG Liang, HAN Zhen-Yuan

(School of Information, Renmin Universityof China, Beijing 100872, China)

Abstract: In recent years, reinforcement learning has made great progress in the fields of electronic games, chess, and decision-making control. It has also driven the rapid development of financial transaction systems. The issue of financial transactions has become a hot topic in the field of reinforcement learning. Especially, it has wide application demand and academic research significance in the fields of stock, foreign exchange, and futures. This paper summarizes the research achievements of transaction systems, adaptive algorithms, and transaction strategies based on the progress of reinforcement learning models, which are commonly used in the financial field. Finally, the difficulties and challenges of reinforcement learning in financial trading system are discussed, and the future development trend is prospected.

Key words: reinforcement learning; deep learning; financial trading system; adaptive algorithm; trading strategy

自从Fama提出有效性市场假说(efficient markets hypothesis,简称EMH)[1,2]以来,EMH就被奉为经典金融理论,并走过了接近50年的历程.到20世纪80年代,许多研究者发现并记录了几个与有效市场假说相互背离的金融现象,由此形成了关注人类交易心理和行为的行为金融学.在经过长期的检验之后,研究者又发现市场也不像行为金融学解释得那样持续无效,相反,很多金融现象在相关论文公开后出现了减少或消失的迹象.这两大学派的争论促进了金融学的发展,也说明金融市场的复杂性可以包容不同学派的存在.金融学家 Andrew Lo 结合进化论和有限理性的概念提出了适应性市场假说(adaptive markets hypothesis,简称 AMH)[3,4],主要观点包含:

* 基金项目: 国家自然科学基金(71531012)
Foundation item: National Natural Science Foundation of China (71531012)
本文由智能数据管理与分析技术专刊特约编辑樊文飞教授、王国仁教授、王朝坤副教授推荐.
收稿时间: 2018-07-19; 修改时间: 2018-09-20; 采用时间: 2018-11-01

(1) 市场中的个体基于自身利益做出决策;(2) 市场中的个体会犯错;(3) 市场个体会学习和适应;(4) 竞争导致个体适应和更新;(5) 自然选择塑造市场生态,进化决定市场动态.

根据 Lo 的理论,金融市场可以被看成一个进化的环境[4].在这个环境中,包含着不同的参与者,如对冲基金、做市商、退休基金和零售投资商等.这些参与者的理性表现并不是即时的,他们对金融产品价格的影响作用也不全是直接发生的,这就促进了金融市场上积极的流动性,流动性则意味着存在套利的机会,这些机会随时会被参与者吃掉,同时,新的机会又会再次出现.这种在进化压力下的流动性同时改变着交易环境和商业环境.这就意味着,一个有效的金融交易系统要能够随时根据交易市场的变化进行自我调整,在感知市场变化的同时,采取相应的行动,如做多(long)、做空(short)、空仓(观望).市场会在行动的基础上给予一定的反馈,如收益、亏损.金融交易系统(financial trading system,简称 FTS)的有效与否不在于执行单次交易的回报,而在于一段时间内交易的总回报,比如年化回报、季度回报等,总回报往往具有延迟性.基于以上原因,Lo[3]提出了如下理论:第一,回报和收益之间关系不太可能一直稳定;第二,相对于经典的 EMH,AMH 认为套利机会一直存在;第三,投资策略在特定环境表现良好,而在其他环境表现较差,既有繁荣也有衰败.针对以上问题,强化学习可以提供很好的解决方案.强化学习技术的基本原理是[5]:如果智能体(agent)的某个动作导致环境正向奖励,则智能体随后产生这个动作的趋势便会加强;反之,智能体产生这个动作的趋势就会减弱.强化学习的目标是学习一个行为策略,使智能体选择的动作能够获得环境最大的奖赏.在一个标准的强化学习框架结构中,它主要有 4 个要素,即策略(policy)、奖惩反馈(reward)、值函数(value function)和环境模型(model of environment).在这 4 个要素中,首先要解决的就是实时环境的数学模型.强化学习可以有效提升金融交易模型的适应性:首先,强化学习擅长解决具有延迟回报的非线性问题;其次,强化学习可以定义灵活的目标函数,在训练中促进模型向最优的目标函数逼近,实践中可以将平衡回报和收益的技术指标作为目标函数;最后,随着 EMH 有效性的提高,固定参数的交易模型难以保证统计套利获得最大利润,而强化学习具有传统探索和利用(exploration and exploitation)机制[6],即通过探索尝试新的参数,利用已有的信息获得最佳回报.

本文综述了强化学习交易系统的各类算法、交易策略、系统构成等方面.第 1 节介绍强化学习在金融交易中应用的关键技术.第 2 节介绍自适应交易系统的应用与发展.第 3 节主要介绍策略轮动模型.第 4 节重点讲解基于值函数的强化学习交易系统和多智能体的发展.第 5 节着重阐述基于策略梯度的交易系统.第 6 节重点介绍深强化学习的应用历史和现状,随后分析了强化学习金融交易系统的研究趋势和应用前景.最后做出总结.

1 金融交易领域的强化学习

1.1 RRL在金融交易系统中的应用

Moody 等人将循环强化学习算法模型(recurrent reinforcement learning,简称 RRL)应用在单一股票和资产投资组合等领域[7],测试了日内外汇市场(USD/GBP)、标准普尔 500(S&P 500 Index)、美国短期国债等金融资产.以收益率为输入,微分夏普比率(Sharp ratio)为目标函数,在交易成本为 5‰的情况下进行实验.RRL 策略获得的回报超过 Q 学习(Q-learning)策略和买入持有策略,并在交易次数上明显小于 Q 学习策略[7].

1999 年,Moody 和 Wu 详尽地解释了 RRL 的理论依据和组织构成,此外,还比较了信息比率(information ratio)与斯特林比率(Sterling ratio)作为目标函数时的收益情况.在标准普尔 500 指数和部分美股测试中,采用斯特林比率作为目标函数的强化学习模型收益最高[8].2001 年,Moody 等人[9]在 RRL 的基础上加入空仓观望动作 $F_t\in\{-1,0,1\}$,$F_t=0$ 表示某段时间内暂停交易,降低风险;此外,还使用下降偏差比率(downside deviation ratio)代替夏普比率作为目标函数,测试市场下行时模型的收益状况.这是将 RRL 首次应用在英镑兑美元的外汇高频交易中.RRL 与 Q 学习的比较结果看,RRL 在多方面优于 Q 学习策略,也证明了 RRL 更适合用在高频交易中.

2003 年,Gold[10]提出在 RRL 模型中用多层神经网络替代单层神经网络.Gold 在 25 个不同的高频外汇交易市场上进行了测试,测试结果表明:单层 RRL 和多层 RRL 都能够实现盈利,且多层 RRL 表现差于单层.无独有偶,2011 年,Gorse 也做过类似的实验,尝试使用多层神经网络代替单层网络.实验结果表明,多层 RRL 的性能相比单层没有明显提升[11].从文献[10,11]的实验中可以看出:当时这些学者采用的仅仅是多层的神经网络,并没有采用

Hinton 等人提出的深度神经网络(DNN)[12],缺少预训练、正则化、Dropout 等深度学习的训练方法,因此,Gorse 的实验出现过拟合现象不足为奇.

Gold[10]的实验结果表明了 RRL 适用于高频金融交易,如外汇交易、指数交易等.同时,有少数文献提到低频交易中的 RRL 应用,例如 Moody[8]和 Gorse[11]等人一直致力于在股票指数的日均时间序列上测试收益效果.本文需要强调:真正的股票交易市场中,有些股票无法像股票指数一样做多或做空.股票指数与个股不同,指数可能存在价格自相关性,受市场基本面影响更多.2013 年,Zhang 等人发现,RRL 在个股日收益率的数据上收益并不理想[13].他们引入了遗传算法(genetic algorithm,简称 GA)来改进 RRL 模型在单只股票上的表现,通过在模型中加入股票的传统指标,提高了 RRL 在个股低频交易上的效果,这种方法称为 RRL-GA.Zhang 等人引入 8 类股票指标,如阳性波动指数(IPVI)、阴性波动指数(INVI)、相对强弱指数(IRSI)、条件波动率(CVOL)等,将其加入到 RRL 的输入序列中.在训练中,利用 GA 找寻 8 类指标的布尔数字组合,实现收益的最大化.实验证明:引入某些指标后,RRL-GA 的收益高于 RRL.2016 年,Zhang 等人不满足于此成果,精选了 10 类单只股票的上述指标作为输入部分加入到 RRL-GA 之中,为了减少 RRL 输入端的噪声,最终仅加入了可能会提高表现的指标.最后,在 180 支美国股票数据的实验中,Zhang 证实了 RRL-GA 比 RRL 有更高的收益[14].因此可以得出这样的结论:RRL 模型在交易单只股票时,交易系统的设计者需要参考来自基本面的分析数据和各类量化交易指标.这样做的好处是利用传统交易手法上积累的经验,规避风险,提升利润.

在交易领域中,最终利润或者基于风险的收益,代表交易模型的回报.通过专家标签和分析一定长度金融时间序列做出交易决策,这种监督方式交易系统存在以下弊端:首先,金融交易获得的回报不是即时的,而是交易终止时的总回报,这导致每一步决策的回报不明确,这正是强化学习中的临时信用分配和结构信用难题,即“系统获得的奖赏如何分配到每个操作上[5]”;其次,标签数据是基于已知的金融时间序列,忽略了不断变化的市场风格对输入变量有效性的影响,导致交易系统不能及时调整策略;最后,随着交易价格的变化,交易成本也在不断变化,无法实时调整交易成本的模型,即便是预测准确,依然会由于交易成本失控导致交易亏损.实践证明,监督学习方式在金融自动交易系统中应用效果并不理想.

相比监督式的交易系统,Moody 等人提出的 RRL 算法是一种在线模式,可以找到随机动态规划问题的近似解[7].RRL 不需要标记信息,通过行动得到的环境反馈来调整内部参数,增加未来回报的期望值.基于 RRL 建立的交易系统通过循环算法解决优化时间信用分配问题和结构信用分配问题[5],RRL 获得的交易利润是路径依赖决策的结果,既包含基于时间的反向传播算法(back propagation through time,简称 BPTT),也包含在线自适应算法(adaptive algorithm).RRL 既可以应用在单一金融资产交易领域,也可以应用到投资组合管理领域.它在金融领域的应用如图 1 所示.

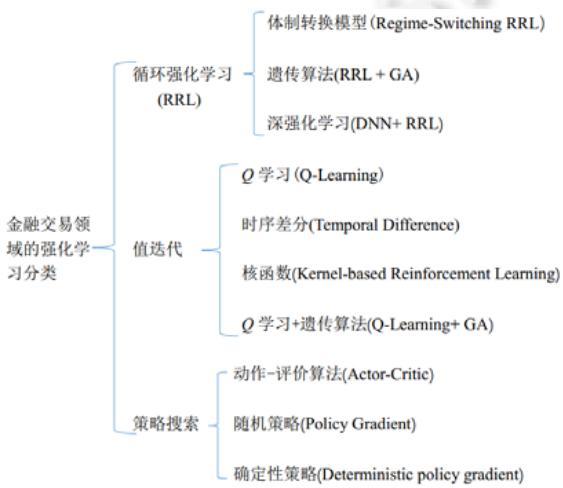

Fig.1 Classification of reinforcement learning application in field of financial

图 1 金融交易领域的强化学习算法分类

1.2 RRL模型

Moody 等人[7]提出的 RRL 模型把金融时间序列作为输入,以最大化微分夏普比率为目标函数,设计两类金融领域常见的操作:做空、做多.RRL 将动作定义为 $F_t\in\{-1,1\}$,代表着在 t 时刻的操作(空/多),RRL 单层神经网络的预测模型如公式(1)所示:

$$F_t=\tanh\left(\sum_{i=0}^{M}w_i r_{t-i}+w_{M+1}F_{t-1}+v\right) \tag{1}$$

向量 $\vec{w}$ 和变量 v 是神经网络权重和阈值;r_t 代表收益率,有如下两种表示方法:

$$r_t=\frac{p_t-p_{t-1}}{p_{t-1}} \tag{2}$$

$$r_t=\ln(p_t)-\ln(p_{t-1}) \tag{3}$$

研究中常采用对数收益率,对数收益率比价格差值更容易体现价格的变动,也更容易计算夏普比率、最大回撤率(max drawdown)等风险度量指标.当价格变化幅度小时,公式(2)和公式(3)中的 r_t 近似相等,但使用对数处理数据更平滑,克服数据本身的异方差,具有对称性.选用 tanh 作为激活函数也正好符合 F_t 的值选择范围. RRL 的基本结构如图 2 所示.

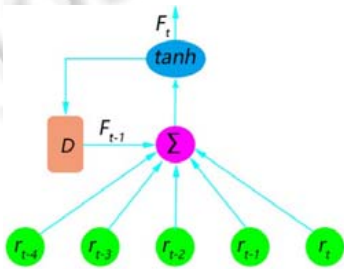

Fig.2 RRL model

图 2 RRL 模型

RRL 算法以最大化利润为目标,但通常不使用最高累计利润作为模型表现衡量指标,最高累计利润 p_T 见公式(4):

$$p_T=\sum_{t=1}^{T}R_t \tag{4}$$

$$R_t=\mu(F_{t-1}r_t-\delta|F_t-F_{t-1}|) \tag{5}$$

其中,μ代表交易额度;δ代表交易成本在每次交易中的比率,包含税率和券商收取费用;R_t 代表单笔交易利润.最大化的累计收益 p_T并不适合作为目标函数,因为 p_T不能体现交易中存在的回撤.在实际金融交易市场中,投资者的本金数额有限,当回撤非常大时,交易的头寸会被损失所吞没,无法继续投资.而夏普比率作为衡量风险和收益的指标更能表现投资是否稳健,较高的夏普比率代表着较高的收益和较低的风险.夏普比率 S 见公式(6):

$$S=\frac{Average(R_t)}{Standard\ Deviation(R_t)} \tag{6}$$

S 作为目标函数时,模型的时间复杂度为 $O(T^2)$.为降低时间复杂度,通常用微分夏普比率(differential Sharpe ratio)替代它.微分夏普比可以看成是一个滑动平均式夏普比率,其推导见公式(7)~公式(10).

$$\hat{S}(T)=\frac{A_t}{B_t} \tag{7}$$

$$A_t=A_{t-1}+\eta(R_t-A_{t-1})=A_{t-1}+\eta\vartriangle A_t \tag{8}$$

$$B_t=B_{t-1}+\eta(R_t^2-B_{t-1})=B_{t-1}+\eta\vartriangle B_t \tag{9}$$

A_t 和 B_t 是代表收益率 R_t 的一阶矩和二阶矩,$\vartriangle A_t$ 和$\vartriangle B_t$ 代表参数的增量.微分夏普比率将移动平均值扩展到

自适应参数η的一阶展开,并使用η的一阶导数作为衡量夏普比率的瞬时性能指标.微分夏普比率见公式(10).

$$D_t=\left.\frac{d\hat{S}_t}{d\eta}\right|_{\eta=0}=\frac{B_{t-1}\triangle A_t-\frac{1}{2}A_{t-1}\triangle B_t}{(B_{t-1}-A_{t-1}^2)^{3/2}} \tag{10}$$

RRL 是一种在策略(on policy)学习方式,微分夏普比率有利于在训练的过程中直接优化 RRL 参数,加速训练的收敛过程,为强化学习提供了一个便捷的评估方法.

2003 年,Gold[10]提出了使用多层神经网络替代公式(1)介绍的单层神经网络,即增加一个隐含层,如公式(11)和公式(12):

$$F_t=\tanh\left(\sum_{j=0}^{N}w'_jx_j+v'\right) \tag{11}$$

$$x_j=\tanh\left(\sum_{i=0}^{M}w_{i,j}r_{t-i}+w_{M+1}F_{t-1}+v_j\right) \tag{12}$$

w'_j 和 v'是新增隐含层的权重和阈值,x_j 代表第 1 层神经网络输出值.

1.3 RRL优化方式

RRL 的目标是通过梯度上升的方式在一个循环神经网络中优化权重 w_t,见公式(13).

$$w_t=w_{t-1}+\rho\frac{dU_t}{dw_t}=w_{t-1}+\triangle w \tag{13}$$

w_t 代表 t 时刻循环网络中的权重,U_t 代表交易体系的某种指标或目标函数,ρ代表学习率.由公式(1)可知:RRL 是一个路径依赖算法,权重更新需要依靠目标时间的梯度传导.循环模型权重更新的梯度值依赖于前段时间整个序列的总导数,这类似于 BPTT 决策序列中的时间依赖性通过参数梯度的递归更新方程来解释.$\triangle w$ 在时刻 t 的值见公式(14).

$$\triangle w_t=\rho\sum_{t'=1}^{t}\frac{dU_{t'}}{dR_{t'}}\left\{\frac{dR_{t'}}{dF_{t'}}\frac{dF_{t'}}{dw_{t'}}+\frac{dR_{t'}}{dF_{t'-1}}\frac{dF_{t'-1}}{dw_{t'-1}}\right\} \tag{14}$$

不同于监督学习在获得最终值时才回传误差和梯度,RRL 模型在前向传播时就不断调整参数,使得目标函数值最大化.如果更新参数仅考虑最近的操作所产生的回报,则公式(14)可简化为公式(15):

$$\triangle w_t=\rho\frac{dU_t}{dw_t}\approx\rho\frac{dU_t}{dR_t}\left\{\frac{dR_t}{dF_t}\frac{dF_t}{dw_t}+\frac{dR_t}{dF_{t-1}}\frac{dF_{t-1}}{dw_{t-1}}\right\} \tag{15}$$

RRL 以最大化微分夏普比率为目标函数时,公式(15)中的参数意义如下:通过公式(4)和公式(5)可知交易动作和回报关系,回报函数对交易动作的导数表示为公式(16)和公式(17):

$$\frac{dR_t}{dF_t}=-\mu\delta\tanh(F_t-F_{t-1}) \tag{16}$$

$$\frac{dR_t}{dF_{t-1}}=r_t+\mu\delta\tanh(F_t-F_{t-1}) \tag{17}$$

鉴于整个模型是循环型结构,导数 $\frac{dF_t}{dw_t}$ 的计算类似于 BPTT 的方式,如公式(18):

$$\frac{dF_t}{dw_t}\approx\frac{\partial F_t}{\partial w_t}+\frac{\partial F_t}{\partial F_{t-1}}\frac{dF_{t-1}}{dw_{t-1}} \tag{18}$$

由公式(18)可知,RRL 的权重 w_t 可以通过梯度求导的方式进行更新.

RRL 是强化学习在交易领域的基础算法,它的价格自适应性和目标函数多样性得到研究者的青睐,许多 FTS 都以 RRL 为核心来搭建,下面的章节中有详细的论述.

2 基于 RRL 自适应交易系统

2.1 金融交易自适应问题

高收益的金融量化模型系统必须具有良好的自适应性,这样才能应对市场频繁的变化.自适应动态规划(adaptive dynamic programing,简称 ADP)由 Werbos 于 20 世纪 70 年代提出[15],在 Bertsekas[16],Lewis[17],Liu[18],Zhang[19]等学者的努力下日臻成熟.ADP 是一种针对连续状态空间的最优控制方法.

基于金融资产时间序列交易是一个复杂问题,它的状态空间和动作空间往往是连续的,规模较大.由于维度爆炸的缘故,不能采用传统的查表法来得到性能函数,需要使用函数逼近器,例如线性函数逼近器和神经网络逼近器等来逼近性能函数.

市场有效性和行为金融学在市场中交替发挥作用,这对交易系统有如下影响:第一,当市场有效性逐渐提高,某些策略的获利机会逐渐消失,传统的静态常数难以保证获利最大,需要对交易参数进行优化,而且还要动态、自适应地调整优化值;第二,常规交易模型的参数往往采用静态常数,由于金融资产时间序列有明显的异方差性,限制了模型使用.对于传统模型的缺陷,一些参数调整方案已经取得了一定效果,但是始终受到新的条件约束.

2.2 RRL自适应交易系统

2003 年,Atiya 等人提出了基于 Q 学习的自适应模拟退火算法,该算法在测试表现中强于传统的 Q 学习算法,证明了良好的自适应性是交易算法的必备特性[20].2006 年,Jangmin 等人提出了基于 RRL 的自适应投资组合策略,它能够有效利用来自特定股票和基金的时间序列信息进行训练,并在投资组合中合理配置高风险资产和无风险资产的份额.Jangmin 将这种资产配置策略应用于韩国股市,它的表现比一些经典的资产配置策略更好[21].

基于 RRL 的、完善的交易系统出现在 2006 年,Dempster 等人创建了三层结构的自动金融交易系统,其模型如图 3 所示.

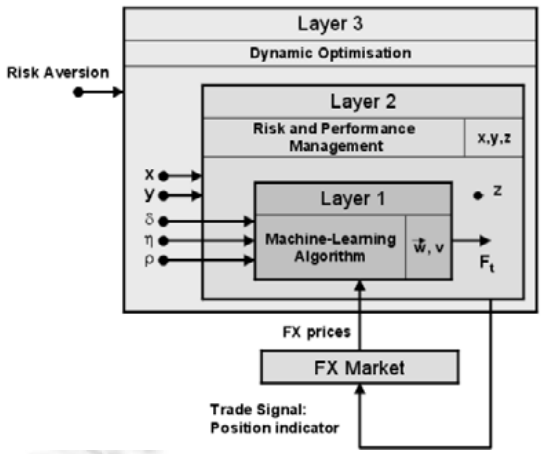

Fig.3 Automated trading system architecture

图 3 自动交易系统架构

结构的 3 个层分别是 RRL 机器学习层(layer 1)、风险管理层(layer 2)和动态优化层(layer 3)[22].风险管理层的作用是在最终决定交易之前使机器学习层的输出决策受到一定的风险限制,z 代表终止交易的被激活值,Risk Aversion 代表控制风险的系统外参数.动态优化层的作用是通过模型的自适应性为模型寻找最佳参数,其中,x 是止损度,y 是交易阈值.RRL 机器学习层的参数 δ 代表交易成本,η 代表自适应参数,ρ 代表学习率.在两年期欧元

兑美元分钟级数据测试中,该系统利润明显高于单独的 RRL 模型.同时,Dempster 等人还引入 14 个常用技术指标作为系统的一部分输入.然而除了少数指标外,大部分技术指标并没有明显增加交易利润.

2007 年,Bertoluzzo 等人在 Moody 的三动作模型 $F_t\in\{-1,0,1\}$的基础上[9],加入风险管理策略来对 RRL 模型进行止损[23].在金融交易中,交易员通过观察不同资产的回报分布不对称性来判断投资的下行风险,其中一个重要的指标是加权对称性(weighted-symmetric).该模型以加权对称指数(weighted direction symmetry index)作为目标函数,而不是微分夏普比率.这样做的目的在于:当市场出现下行风险时,可以更好地控制决策.该系统在 9 个世界主要股市指数上测试的结果令人鼓舞,至少有 8 个指数有盈利表现.

2011 年,Gorse 等人提出一种控制交易成本的自适应金融交易系统[11],该系统将公式(1)中的固定阈值 v_j 替换为如公式(19)中的带有权重的可变阈值 w_{M+2}:

$$F_t=\tanh\left(\sum_{i=0}^{M}w_i r_{t-i}+w_{M+1}F_{t-1}+w_{M+2}\right) \tag{19}$$

从公式(1)和 Moody[8]中可知:阈值与交易成本息息相关,可以通过不断调整阈值应对交易成本的变化.但是,这并不意味着交易成本上升就可以通过提升阈值来应对.若阈值设置不合理,同样会造成交易损失.Gorse 设置这种自适应方式来代替手动调整阈值,在训练中实现阈值的自动调整,以达到收益最大.该模型也尝试使用多层神经网络代替单层网络,然而实验中发现,多层的 RRL 并未提升模型性能.同样是在 2011 年,Tan 等人提出一种非套利型的高频交易系统[24],在 RRL 中加入自适应网络模糊推理构成一种混合模型(adaptive network fuzzy inference system,简称 ANFIS).ANFIS 的优势在于可以通过模糊推理的方式进行模式转换,使 RRL 系统适应不同的股票市场周期.例如,股票上行趋势会持续几天或几周,股市的大波动率后往往有大波动伴随,小波动率后往往有小波动伴随,这种周期规律已被市场经验所验证,敏锐地适应这样的周期会产生可观的利润.ANFIS 根据这种趋势规律实现了股市拐点的预测.使用 5 只美国股票的 13 年时间序列数据测试 ANFIS,均取得了稳定的利润.Almahdi 等人在 2017 提出了自适应能力的 RRL 交易系统[25],他们研究发现:在资产投资组合交易中,使用动态止损(stop loss)策略同时配合不同的目标函数(objective function)使用,得到的收益远高于单一目标函数的策略.例如:使用斯特林比率作为目标函数的 RRL 模型可以抵消市场长期下行风险,而市场平稳上行时,使用夏普比率的模型收益更高.卡玛比率(Calmar ratio)对损失的大小很敏感,当交易成本逐渐上升,并且期望最大回撤(expected maximum drawdown,简称 EMDD)很大时,使用卡玛比率的投资组合收益始终优于基于夏普比率和斯特林比率.交易系统整体流程如图 4 所示.

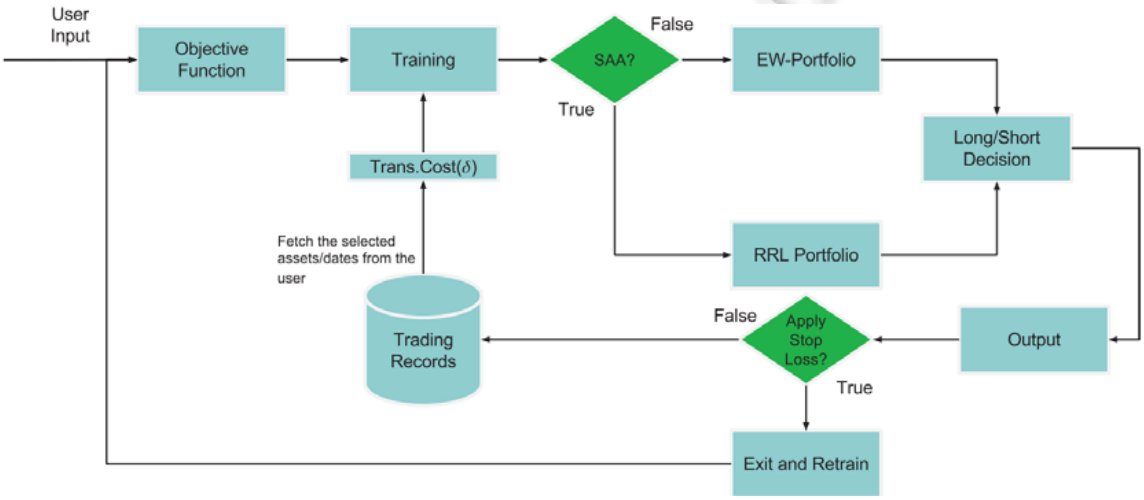

Fig.4 RRL based trading decision system

图 4 基于 RRL 交易决策系统

Almahdi 等人使用上述 3 种不同的目标函数开发出新型 RRL 自适应交易系统.文献[25]从雅虎财经收集金融资产数据,选择 5 个常见的 ETF 的投资组合进行实验.此系统中,交易者先选择一个最有利的目标函数, RRL 系统将使用雅虎财经历史数据来学习和训练参数;然后,允许用户选择两种资产组合方式:加权投资组合(EW-portfolio)和 RRL 投资组合(EW-portfolio).RRL 系统将输出对每个资产的多/空决策(long/short decision)以及投资组合方式.系统还会询问投资者是否愿意使用动态止损退出策略,这将停止交易并重新训练系统.如果不想止损,那么输出将被存储(trading records)以供系统继续从给定的产出中学习.假定系统训练时的预定交易成本为每股 0.1%,在训练阶段没有止损.在真实的交易系统中,投资者可以根据自己的交易成本估算他们过去的交易记录.由于交易成本在不同时期会发生变化,系统会提醒投资者改变目标函数重新训练参数,以适应这些变化.当交易成本超过每股 0.15%时,系统会建议用户设定卡玛比率作为目标函数,这将有助于系统承受交易成本上升的影响.此外,如果投资者担心出现大幅回撤,那么改用卡玛比率训练系统以应对预期最大回撤将是非常明智的.

综上,通过研究我们发现,成功的自适应交易系统有如下 3 个特征.

(1) 正确选择用于交易的自适应算法和模型目标函数;

(2) 使用明确的规则定义进场和出场时机;

(3) 良好的风险控制方法,根据市场情况及时转换交易策略.

3 具有策略轮动的 RRL 金融交易系统

通过第 2 节可得知交易策略转换关系到系统能否成功.根据适应性市场假说理论,单一策略不可能长期有效,总会有一段时间策略 A 效果特别好,而过一段时间策略 B 效果更好.交易系统不仅要在适当的时候持有合适的股票、基金、债券,还要重仓合适的策略模型.业界将一段金融资产时间序列的不同时期定义为不同的状态,择优选择策略,这就是策略轮动,本文称为体制转换模型.

最简单的情况下,金融资产时间序列状态的转换可以用一阶马尔可夫链描述,称为马尔可夫体制转换模型,体制转换模型属于变参数模型.Hamilton 将体制转换模型与自回归模型(GRACH)相互结合,用 GARCH 模型计算动态价差标准差[26].GARCH 模型的参数变化是一个离散状态马尔可夫过程,可以描述变量的趋势转变.Hamilton,Susmel[27]和 Gray [28]将体制转换模型与 ARCH 模型结合,描述了波动率在不同大小的波动状态之间的转换.体制转换模型不是一个独立的模型,需要结合其他模型一起来判定趋势.

RRL 交易系统不能完全应对金融交易市场的复杂情况,Gold 的实验已经证明:在金融数据包含噪音的环境下,多层神经网络非常容易出现过拟合现象,神经网络的黑盒式方法也难以总结关系之间的联系[10].因此,Maringer 等人提出的体制转换模型(regime-switching recurrent reinforcement learning,简称 RS-RRL)更适合于模拟非线性的变化情况[29].该模型让 RRL 模型在不同的波动率下选择不同的权重,以应对市场风格连续发生变化的情况.2010 年,Maringer 和 Ramtohul 首次提出阈值自回归模型(threshold RRL,简称 TRRL)[30],此模型设置一个转换阈值控制两个模式的转换,如图 5 所示.

图 5 中,变量描述如公式(20)~公式(22)所示:

$$G_t=I[q_t>c] \text{ for TRRL} \tag{20}$$

$$F_t=y_{t,1}G_t+y_{t,2}(1-G_t) \tag{21}$$

$$y_{t,j}=\tanh\left(\sum_{i=0}^{m}w_i r_{t-i}+w_{m+1,j}F_{t-1}+w_{m+2,j}v\right) \text{ for } j=\{1,2\} \tag{22}$$

其中,$y_{t,1}$ 和 $y_{t,2}$ 代表两个不同的 RRL 网络,q_t 代表指示变量,c 代表阈值,G_t 代表权重.TRRL 可以被看成两个 RRL 网络,每个网络对应一种交易风格,系统总的输出 F_t 是单个网络 $y_{t,1}$ 和 $y_{t,2}$ 的加权和,权重受到 q_t 的直接作用.

在金融市场中,波动率是描述金融市场风格的重要标志之一.初始阶段,$y_{t,1}$ 和 $y_{t,2}$ 有同样的权重;训练期间,该模型进行选择性的学习,每个网络有一组独特的权重,阈值是一组门控制器,在不同的时间序列阶段步骤选择不同的网络.实际上,指示变量 q_t 的作用是让模型能够在高波动率和低波动率之间转换,适应不同的市场风格,

公式(20)~公式(22)共同组成 TRRL.在使用 4 只欧洲股票的测试上,TRRL 均有超出 RRL 的表现[30].2012 年,Maringer 等人对 TRRL 模型进行改进,新模型称为平滑转换自回归模型(STRRL)[29].TRRL 模型的阈值是二元数,只能在[0,1]间进行转换而不能平滑过渡,STRRL 的模型设计中则包含平滑过渡的方式,见公式(23).

$$G_t=I[1+\exp(-\gamma[q_t-c])]^{-1} \text{ for STRRL} \tag{23}$$

STRRL 的网络结构如图 6 所示.TRRL 的每个网络学习一个独特的映射对应一个特定的区域,在指标变量 q_t 转换的过程中,TRRL 作为一个开关在每个时间步选择合适的网络.STRRL 则允许两个模型有一定量的重叠,重叠的程度由 γ 来规定.STRRL 的 G_t 可以取[0,1]之间的任何值,参数 γ 决定了转换的平滑性.当 γ 趋近于无穷大时,STRRL 趋近于 RRL.从图 6 中可以看出,STRRL 相比 TRRL 在体制转换上有更好的平滑性,更容易及时应对金融市场风格变化.

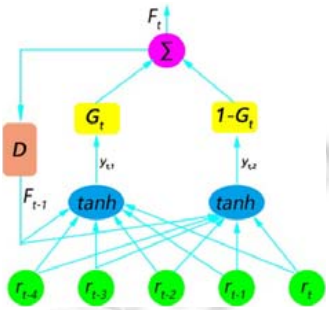

Fig.5 TRRL model

图 5 TRRL 模型

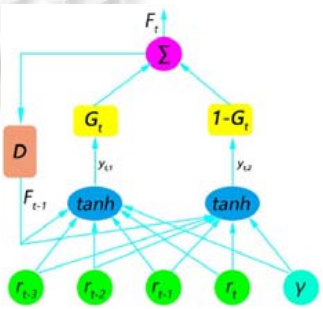

Fig.6 STRRL model

图 6 STRRL 模型

为了验证 STRRL 的有效性,文献[29]在人工生成数据和 12 只美股数据上进行了测试,结果显示,在人工生成数据上,微分夏普比率没有太大的区别;但是在 12 只美股的测试上,STRRL 比 TRRL 和 RRL 获得更高的微分夏普比率.可见 STRRL 能够适应真正包含趋势信息的金融资产时间序列,而不是人工生成的随机序列.

Maringer 等人[31]探讨过指示变量 q_t 对于市场风格变化的映射关系应该由哪些参数确定,不仅 GARTH 模型生成的波动率可以体现市场风格的变化,成交量(trading volume)、日内信息到达率(daily rate of information arrival)都可以作为 q_t 的衡量指标.当市场条件发生剧烈变化时,单独的衡量指标不足以描述市场变化,多指标更有把握.交易量与价格变化绝对值之间存在正相关性,交易减少通常伴随着价格下跌,交易量增大通常意味着价格上涨.例如,新的股票公告或者新闻稿也会直接导致价格的波动.因此,日内信息到达率通常也影响着市场风格的变化.通过在 15 只美股数据上的测试,基于成交量和波动率的 RS-RRL 模型比基础的 RS- RRL 模型有更好的表现,这足以证明引入更多的指标信号会对交易有积极的作用.

本文将 Maringer 的模型称为 RS-RRL1.0.在此系统中,无论是 TRRL 还是 STRRL,如果没有人工干预 G_t 中的参数,系统无法实现自动模式转换,旧的转换模型未必适用于当前的金融交易环境.基于可能性推理的转换函数在交易方面会弱化 RRL 的自适应性.以上的缺点让 Maringer 和 Zhang 在 2014 年提出 RS- RRL2.0,用以提高 RS-RRL 交易系统的表现[32].新模型用一个 *sigmoid* 函数的求和公式代替权重 G_t、指示变量 q_t、阈值 c,见公式(24):

$$G_t = sigmoid(w_{t-1}^G \times r_{t-1}) = \frac{1}{1+\exp(-w_{t-1}^G \times r_{t-1})} \tag{24}$$

w_{t-1}^G 和 r_{t-1} 与公式(1)中的 w_i 和 r_i 相同,代表输入金融时间序列和权重.RS-RRL2.0 的结构如图 7 所示.由图 7 可知,权重 G_t 由单层神经网络所替代,可以通过训练的方式实现权重的自动转换.RS-RRL2.0 模式的转换功能是动态系统利用最大化夏普比率驱动,不再由基于高低波动率之间的转换来驱动.利用单层神经网络可以提供对市场风格更敏锐的洞察,在 20 只瑞士股票数据上的实验结果表明,RS-RRL2.0 交易系统的夏普比率在统计上要比 RS-RRL1.0 交易系统高出 5%以上.由此可见,动态的、自适应性的交易策略能更好地服务于投资管理,

特别是在市场风格变化频繁时,能够更好地适应节奏,转换到合适的策略模式下.

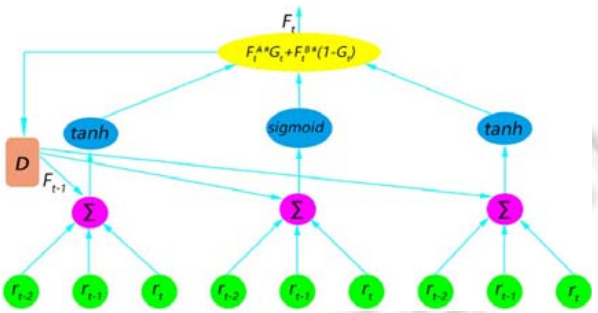

Fig.7 RS-RRL2.0 model

图 7 RS-RRL2.0 模型

4 值函数与 Q 学习的金融交易系统

马尔可夫决策过程(Markov decision process,简称 MDP)是强化学习建模的经典算法,其主要思想是在 MDP 上进行动态规划,寻找最大化累计回报.假设一个策略能够在一个状态上就预测到未来的累计回报,那么意味着存在动态规划的最优解,这种求解方法被称为值函数(value function)方法.

最近 20 年,基于值函数方法,如时间差分学习(TD-learning)和 Q 学习,一直是该领域的主要研究课题[33,34].Q 学习是最早最重要的在线强化学习算法,由 Watkins 在其博士论文中提出[35].该算法的主要思想为:将在线观测到的数据带入到更新公式中对 Q 函数进行迭代学习,得到精确解.

Q 学习是一种离策略(off policy)的学习算法,使用合理的策略来产生动作,根据该动作与环境交互所得到的下一个状态以及奖赏来学习得到另一个最优的 Q 函数.Q 学习只能在有一定限制条件并且理论上能够收敛的情况下才能得到最优控制策略[36−38].当 Q 学习中离散状态很多时,行动选择过多会陷入贝尔曼维度诅咒[5].并且用 Q 学习做函数逼近时,某些情况下马尔可夫决策过程不收敛.在 Q 学习算法下,微小的噪音往往也会导致无法选择最优策略[39−41].

理论上,值函数有值函数(25)和状态-动作对值函数(26)两种:

$$V^{\pi}(x)=\sum_{a}\pi(x,a)\sum_{y}p_{xy}(a)\{D(x,y,a)+\gamma V^{\pi}(y)\} \tag{25}$$

$$Q^{\pi}(x,a)=\sum_{y}p_{xy}(a)\{D(x,y,a)+\gamma\max_{b}Q^{\pi}(y,b)\} \tag{26}$$

其中,$\pi(x,a)$是在状态 x 下采取行动 a 的概率;$p_{xy}(a)$是在动作 a 下从状态 x 到状态 y 的转移概率;$D(x,y,a)$是即时回报,在金融交易里面可以是最大的微分夏普比率、最大利润或其他指标;γ是折扣率,取值范围是[0,1],越远的动作回报率越低.

公式(25)和公式(26)都是通过获得最优值函数来获得最大化累计回报.如果当前的策略在值函数下获得值超过之前的其他策略,则称为最优策略.通过对公式(25)的迭代,可以实现值函数的最终收敛.公式(25)满足贝尔曼方程(Bellman equation),通过迭代优化得到公式(27):

$$V^{*}(x)=\max_{\pi}V^{\pi}(x)=\max_{a}\sum_{y}p_{xy}(a)\{D(x,y,a)+\gamma V^{*}(y)\} \tag{27}$$

这也意味着公式(26)和公式(27)两个函数存在以下关系:

$$V^{*}(x)=\max Q^{*}(x,a) \tag{28}$$

相应的最佳动作就可以表示为公式(29):

$$a^{*}=\arg\max_{a}\{Q(x,a)\} \tag{29}$$

Q 学习依据上面的公式不断迭代,寻找更高的回报,近似函数的更新规则可以通过梯度的方差进行迭代,最

优动作决定最大回报,最优动作的选择策略由完全贪心策略(ε-greedy)决定,ε-greedy 会在一定的概率限制下进行探索,而不是完全使用贪心算法.2001 年,Moody 等人将 Q 学习算法应用在资产组合配置和金融交易中[9],他们定义了 3 个动作 $F_t\in\{-1,0,1\}$,分别在人工生成数据、外汇交易数据和 S&P500 指数上测试,结果显示,RRL 胜过 Q 学习算法.可见,当时在交易中 RRL 自适应方式优于 Q 学习.但 Q 学习更好的灵活性和扩展性,在之后的研究中逐渐显现出来.

相比于 RRL 的简单动作 $F_t\in\{-1,0,1\}$,Q 学习的动作 a 定义方法非常多.2003 年,Lee 等人提出了基于 Q 学习的多智能体自动交易系统[42],它考虑交易过程中交易限价单的情况,对不同的价格状态做出判断并执行相关动作.该系统首先通过买信号智能体(buy signal agent)判断是否有必要买入,待确定后,再唤醒买单智能体(buy order agent)下单.买单智能体根据交易数据的涨跌判断是否到达抛售点,比如涨 30%、跌 20%.在达到或接近抛售点时,唤醒卖入信号智能体(sell signal agent).每个智能体都有自己确定的动作和回报设定,如:买方智能体只有不买和买入两个动作,而回报需要卖出后才能得到,卖方订单完成后,有对买的回报,不买回报始终是 0.而卖出信号智能体在完成交易并扣除交易成本后才能得到回报.文献[42]在韩国综合股票指数(KOSPI200)上测试时,得到了远超过买入持有策略的回报.2007 年,Lee 等人再次完善多智能体 Q 学习自动交易系统,命名为 MQ-Trader[43].它定义多个 Q 学习智能体,有效地克服了之前在复杂环境中股票交易存在的问题.

基于值函数的强化学习经典理论是通过策略π求得最大回报 $V^*(S)$,其回报公式为

$$V^*(s)=E\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k}\mid s_t=s,a_t=a,\pi\right\} \tag{30}$$

金融投资领域中某些人是风险厌恶型投资者,比如母基金(fund of funds,简称 FOF)或者养老保险基金等.相比于高利润,这些投资者在保证基本利润的同时更关心风险系数的大小.2006 年,Jian Li 等人提出通过调整回报的方式规避 Q 学习中存在的交易风险问题[44].Li 将交易回报同 GARCH 模型得到的风险标准差合并,从而得到回报调整强化学习模型(reward adjustment reinforcement learning,简称 RARL),回报值改为公式(31):

$$r_t^a=r_t-\alpha\sigma(g_t) \tag{31}$$

r_t 代表模型定义的基本回报值,α代表厌恶风险的情绪值,$\sigma(\cdot)$代表方差公式,g_t 代表从 GARCH 模型获得的风险值.经过公式(31)的变化,RARL 的回报从公式(30)升级为公式(32):

$$V_{new}^{\pi}(s)=E\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k}^{\alpha}\mid s_t=s,\pi\right\} \tag{32}$$

在香港股票数据的测试上,此方法泛化性能明显优于传统的 Q 学习金融模型.2012 年,Bertoluzzo 在文献[23]工作的基础上继续完善 FTS 系统,使用值函数的方式构建 FTS 替换之前的 RRL 模型.Bertoluzzo 又测试了基于时序差分模型(temporal difference,简称 TD)和核函数的强化学习模型(kernel-based reinforcement learning,简称 KbRL)作为 FTS 系统的主模型,动作设置为 $a\in\{-1,0,1\}$,采用经典夏普比率而非微分夏普比率作为目标函数[45,46].文献[46]提出构建 FTS 系统时不采用动态规划或蒙特卡洛方式,原因如下.

- 首先,动态规划需要一个模型来计算一个状态到另外一个状态的实际转移概率,在金融交易中,这样的模型通常是未知的;
- 其次,为了改进策略需要等到全部交易结束之后才能进行估算,而 FTS 交易是无限次数的.

Q 学习的方式比较符合 FTS,Q 学习源于无模型强化学习的 TD 学习,Q 学习不需要等到交易结束就可以让模型在近似状态下收敛.

鉴于 Q 学习的关键问题是定义环境、状态、动作、回报这四者之间的关系,Bertoluzzo 等人在 2014 年又对 FTS 系统做了进一步完善,重新定义了金融市场状态变量[47],如公式(33):

$$s_t=(e_{t-4},e_{t-3},e_{t-2},e_{t-1},e_t) \tag{33}$$

最后 5 个交易日结束时的对数收益率 $e_\tau=\ln(p_\tau/p_{\tau-1})$为系统状态变量,$\tau$代表间隔时间,$p_\tau$代表价格;同时引入多种目标函数,如夏普比率、净值对数回报、净值对数收益之和比率等,经过在意大利股票指数数据上的测试,结果总体令人满意.

在 Q 学习的交易算法中,不仅是状态 S_t 的定义具有灵活性,交易动作 a 也可以做必要的扩展.在交易中如果持有金融资产,那么每一个时间步骤中无论价格是上涨还是下跌,都需要设置相应的动作(买入或卖出).Du 等人设置了 4 种组合操作来应对这种情况[48],如公式(34)所示:

$$\begin{pmatrix} rise/buy & rise/sell \\ fall/buy & fall/sell \end{pmatrix}=\begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{34}$$

Du 详细对比了 RRL 和 Q 学习的交易方式后指出:应用于包含大量噪音数据集合时,在正确的目标函数下,RRL 在稳定性和计算收敛性上优于 Q 学习,但是 Q 学习的操作选择更加灵活多样.

综上所述,在价格自适应上,RRL 一定程度优于 Q 学习.但是 Q 学习的动作设置上可以多种多样,不仅应用于买卖,还可以用来触发各类交易信号.同时,Q 学习还可以将很多金融资产的各种状态定义到 Q 学习的状态 S_t 中,这比 RRL 有更大的优势.此外,Q 学习还可以定义多智能体的应用方式,在买入、卖出等交易环节处应用,比传统的 Q 学习有更高的灵活性.在 FTS 应用中,常有多种策略同时使用,比如配对交易、股票中性等,这些策略往往同时操作多种金融资产,单纯的买入和卖出不能满足系统的操作需求,因此,基于 Q 学习的多种算法值得深入研究.

5 基于策略梯度的金融交易系统

理论上,值函数方法在离散状态空间中可以收敛到最优策略,但收敛速度可能极慢.值函数的一个微小变动都可能导致动作选择的错误,这种变化会影响算法的收敛性.同时,值函数的方法有两个局限性[41].

(1) 值函数算法最终得到的是一个确定性策略,而最优策略有可能是随机的;

(2) 值函数存在策略退化问题,即使值函数估计得很准确,通过值函数获得的策略仍然不是最优策略.

为解决寻找最优策略问题,Sutton 等人提出了策略梯度算法(policy gradient)[41],该类算法不会出现策略退化现象[49,50].策略梯度是一种直接逼近的优化策略,直接在策略空间进行求解得到策略.

基于值函数的方法,通过迭代计算每一轮(state-action-reward)的交互,选择回报最大的动作 a,这是一种间接做法.直接的做法是通过神经网络直接求得下一次的状态或动作.2014 年,Eilers 等人提出用策略梯度将交易决策与回报紧密联系起来[49],描述见公式(35):

$$State\ s_t \in S \quad s_t \xrightarrow{a_t} s_{t+1} \tag{35}$$

$$s_{t+1}=\sigma(s_t,a_t) \tag{36}$$

$$r_t=r(s_t,a_t) \tag{37}$$

S_t 代表交易前状态,S_{t+1} 代表交易后状态,a 代表交易动作,σ代表状态转换函数.Eilers 等人使用三层神经网络的 RRL 作为σ函数,奖励直接从函数 r 中获得.r_t 代表回报值,用正负表示积极或消极,最终依据策略梯度的方式收敛.Eilers 分析并介绍了不同月份以及不同季度对股票期权市场的影响和交易风格的变化,使用人工神经网络结合 RRL 的方式,借助 RRL 的自适应性,让交易系统在不同的月份之间自动转换交易风格.实验显示,Eilers 等人的交易系统没有将目光局限在获得最大收益或高夏普比率值,而是最大限度地提高每次交易的即时回报,将最好的交易动作分配给最合适的情况.

基于策略梯度的方法在求解上相比值函数更加方便,但也更容易陷入到局部最优解.因为策略梯度过分关注获得最大期望回报,而不是最优解.Actor-Critic 在这两种方法中找到了平衡[50].行动网络(actor)的目的是通过一个函数——输入状态 S_t,输出动作 a,在这个过程中尽量获得最高的回报,使用策略梯度来更新参数.为了训练 Actor,创建一个值函数评价网络(critic)评估 Actor 的表现.用值函数替代采样的回报(reward),提高样本利用率,降低策略梯度求解时的梯度(估计)方差.这个两个网络通常使用人工神经网络来近似模拟.Actor-Critic 的示意如图 8 所示.

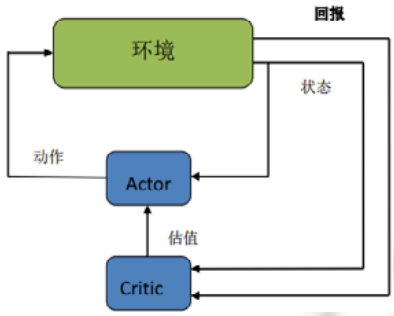

Fig.8 Actor-Critic schematic diagram

图 8 Actor-Critic 原理图

2007 年,Haili 等人提出了基于 Actor-Critic 算法的结合强化学习和监督学习共同预测金融数据短期走势的模型[51],分别是 Actor 结合监督学习模型(actor-supervised llearning)和动作-评价模型(actor-critic).前者结合了两种模型的优点,监督学习具有快速收敛的优点,洞察短时间的市场惯性并给出最佳预测插值;Actor 使用的是基于 RRL 的模型,缩小搜索空间.这表明在没有正确标记训练样本时,可以利用 RRL 的自适应性,通过强化学习对监督学习的缺陷进行微调.此外,RRL 缩小了搜索领域,也避免过早收敛陷入局部最优解.后者模型则通过两个 MLP 网络拟合各自的函数,实现预测:首先,实验将数据定义为一些离散状态 S_t,然后将状态 S_t 和 Actor 网络的输出 F_t 作为 Critic 网络的输入,将下一个时刻的价格状态 S_{t+1} 定义为输出,训练 Critic 网络.得到训练的 Critic 可以更好地评价 Actor,实现 Actor-Critic 模型的共同训练.这里,Actor 与 Critic 在参数更新上是异步的.实验中, Hailin 使用时间跨度达 20 年的数据,分别将两个模型用在 3 种股票价格预测上,如 S&P 500 指数、纳斯达克综合指数、IBM 股票.结果显示:S&P 500 指数预测效果良好,纳斯达克综合指数和 IBM 也在盈利的范围内.在其他个股的金融时间序列预测中,单独的 Actor-Critic 模型通常表现更好.

2010 年,Stelios 等人也使用 Actor-Critic 构建了 FTS 系统[52],并提出一个自适应模糊强化学习模型.该模型能够准确迅速地识别市场方向.模糊推理最初应用在控制领域,它提供了一种用不精确数据来表示不确定的方法.这意味着它可以成为不确定条件下智能体选择决策的极好工具.模糊推理直接将数字术语翻译成为语言术语,通过 IF-THEN 表示模糊推理规则指定语句将模糊输入与模糊输出相互关联,例如:出现条件“西红柿非常红”,立即可得出结论“西红柿非常熟”.模糊推理将有效的经验法则与非结构化知识结合起来,这种方法同上文中的 RS-RRL 模型非常类似,可以通过经验把握金融市场风格的变化.

金融时序价格一般都具有两个属性:预期收益(expected return)和条件波动(conditional volatility).Stelios 利用这两点定义了 8 个不同的状态空间.通过模糊推理系统提供的输出作为 Actor 的输入,利用波动性的指标实现强化学习的可预测性,最后使用选定的参数为 Critic 构建基于交易的决策.在纳斯达克综合指数、英国富时 100 指数、Nikkei 亚洲 300 可投资指数等指数的实验中,Stelios 提出的 FTS 系统收益高于循环神经网络、马尔可夫模型和买入持有策略.

综上所述,在状态和交易动作选择上,Actor-Critic 算法远多于 Q 学习算法和 RRL 算法.RRL 算法在自适应性有上一定的优势,易于根据当前金融信息自动转换交易风格.Actor-Critic 算法和 Q 学习则可以通过引入系统外变量,如模糊推理和市场风格变换信号实现自适应性.此外,当深度学习模型替代 Actor-Critic 中的策略函数后,该算法在围棋等领域取得了突破.未来,融合深度学习的 Actor-Critic 模型也必将在 FTS 上得到推广和应用,在下文中将介绍深度学习对强化学习的促进作用.

Actor-Critic 模型多智能体算法近年发展迅速,Lee 等人提出过 Q 学习多个智能体算法,它主要用多个智能体执行不同环节的任务.这类算法存在两个问题.

(1) 这种方法无法解决强化学习运算效率低的难题,无法利用多核 CPU 和分布式计算;

(2) 无法解决金融数据序列时间上强相关的问题.

神经网络训练不稳定的最主要原因是数据存在着较强的时间相关性,且不满足独立同分布的条件.DQN[53,54]和 DDPG[55]方法都利用经验回放的技巧打破数据之间的相关性,然而,在经验回放之外的一种方法是异步方法.

A3C[56]便是主流的异步方法,全称为异步优势动作评价算法(asynchronous advantage actor-critic),源于Actor-Critic 算法,训练时利用多个线程而非单线程.每个线程相当于一个智能体在随机探索,多个智能体共同探索,并行计算策略梯度,维持一个总的更新量.相比于经验回放,这种方式同样能让数据实现独立同分布,并且可以利用 CPU 多核实现分布计算,提升训练的速度.因此,基于 A3C 异步策略算法构建的 FTS 将具有广阔的应用前景.

6 基于深强化学习的交易系统

2006 年,Hinton 等人在 Science 期刊上提出了基于深度信任网(deep belief network,简称 DBN)的非监督训练算法,实现了深度学习(deep learning,简称 DL)的重大突破[12].目前,已经在图像分析[57,58]、语音识别[59,60]、自然语言处理[61,62]、视频分类[63]等领域取得了令人瞩目的成就.DL 的基本思想是:通过多层网络结构和非线性变化组合低层特征,形成抽象的、易于区分的高层表示,以发现数据的分布式特征表示[64].因此,DL 方法侧重于对事物的深层特征提取,而强化学习侧重于提出解决问题的策略.随着社会的飞速发展,在复杂问题中,利用 DL 自动学习大规模输入数据的抽象特征,并以此表征进行自我训练的 RL,已成为解决问题的策略.Deep Mind 团队创新性地将具有感知能力的 DL 和具有决策能力的 RL 相结合,形成了深度强化学习(deep reinforcement learning,简称 DRL).

由第 1.1 节可知,早在 2003 年,Gold 等人就尝试使用多层神经网络替代经典 RRL 中的单层神经网络,但是多层网络容易发生过拟合现象,效果提升有限[10].在后来的 FTS 系统构建中,学者们也尝试使用多层神经网络替代单层神经网络.Bertoluzzo 等人提出的用多层感知机替代单层神经网络的 FTS,但是并未证明多层神经网络优于单层神经网络[23].此后,学者们并没有放弃将深度神经网络应用在 RRL 之中.

上述问题的难点在于,金融交易中始终存在两个问题.

(1) 财务数据包含大量的噪音,这种不确定性导致时间序列高度不稳定.因此,能否从数据中直接获得特征一直是研究的目标;

(2) 动态交易的执行问题.强化学习是通过连续性操作获得回报,即使有一套稳定的策略,也会因为频繁交易带来巨大的交易成本,这反而对实际利润没有贡献.

基于以上原因,需要把当前的市场条件同先前的交易动作相结合,用前一个时刻的多空操作和持仓数量来决定当前的操作.虽然 RRL 也具备这样的能力,但要想在更长的时间段中运行,交易模型需要具有一定的记忆能力.与此同时,在模型融入记忆力的设置中不能增加额外的复杂性,避免忘记过去的训练成果.2017 年,Deng 等人提出了一种结合模糊学习(fuzzy learning)、DNN、RRL 的 FTS 系统,称为 FRDNN[65].与文献[24,51]类似,文献[65]首先使用模糊学习减少数据的不确定性;其次,使用 DNN 对数据进行降噪和特征提取,通过范数正则化、数据增强、Dropout、自编码器预训练及 CNN 权值共享等方式解决多层神经网络过拟合问题,将处理过的数据交给 RRL;最后,进行交易行为选择.从前面的文献中可以了解到:RRL 强于自适应性,弱于特征提取.因此,通过加入模糊学习和深度神经网络可以整体提升原始模型的能力,公式(1)变化为如下公式(38)和公式(39):

$$F_t = \tanh\left(\sum_{i=0}^{M} w_i k_{t-i} + w_{M+1} F_{t-1} + v\right) \tag{38}$$

$$k_j = g_d(u(r_j)) \tag{39}$$

$u(\cdot)$代表模糊学习函数解决金融数据的不确定性问题,$g_d(\cdot)$代表 DNN 用来解决特征提取不充分和缺少记忆能力的问题,通过 DNN 将 $u(r_j)$映射为更深层次的向量,结构如图 9 所示.

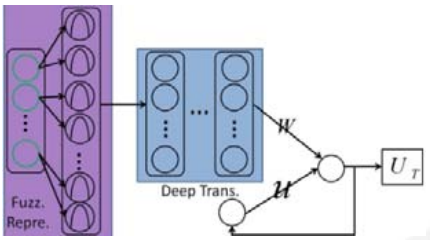

Fig.9 FRDNN framework

图 9 FRDNN 结构

在图 9 中,Fuzz Repre 代表模糊学习模块;而 Deep Trans 则代表使 DNN 的特征提取模块;W 代表 $\sum_{i=0}^{M} w_i k_{t-i}$;u 代表 $w_{M+1}F_{t-1}$;U_T 同于公式(4)的 p_T,代表时间 T 内的累计期望,即最大收益.Deng 分别在沪深 300 的期货交易数据和白银、白糖的商品期货分钟级别的高频数据上进行测试.实验结果表明,FRDNN 的收益极高,RRL 模型在某些交易上亏损非常严重.FRDNN 还与预测型 DNN 做了对比实验,分别使用 CNN,RNN,LSTM 在无交易成本时,DNN 模型的收益同 FRDNN 不相上下,一旦交易成本上升,DNN 模型的盈利能力迅速下降.可见:不能只注重模型预测能力,忽略交易成本,频繁交易的获利会被巨大的交易成本所吞没.这也进一步证明了 FRDNN 模型的合理性.同时,Deng 的实验中还对比了最高累计总利润和最高夏普比率分别作为目标函数时的收益情况.显而易见,最高夏普比率的模型收益明显要高,特别是在市场进入下行轨道时.

同样在 2017 年,Lu 等人发现,在文献[65]中使用 DNN 作为特征提取时常出现梯度消散问题,因此采用 LSTM 替换上述 DNN[66],并加入了 Dropout 技术来调试 LSTM 避免过拟合.Lu 在美元兑英镑的外汇交易数据上测试:首先,作者观察到公式(1)中的阈值 v 对交易频率和策略的影响,当 v 逐渐增大时,交易频率下降;之后,使用 LSTM 进行特征提取,并加入市场下行信号;最后,尝试使用下降偏差比率代替夏普比率作为损失函数.这些操作的结果都证明:在市场下行时,通过精确的做空,依然可以取得较高的交易利润.

文献[65,66]中可以看到:深度强化学习的算法应用在特征提取上,可以依靠确定性策略直接从采样特征中找寻下一次操作[67].无模型的策略搜索可以分为随机策略搜索方法和确定性策略搜索方法.2014 年以前,学者们都在发展随机策略搜索方法,直到 2014 年,Silver 提出了确定性策略理论[67].确定性策略意味着在应用策略函数 π_θ 时,在状态 s_t 下,下一步的动作 a 是确定的,即 $a=\pi_\theta(s_t)$.随机策略中,即使在相同的状态,每次采用的动作也很可能不一样.当然,当采用高斯策略的时候,相同的策略在同一个状态处,采样动作差别不大.确定性策略不需要像随机策略一样在空间进行大量采样.通常来说,确定性策略方法的效率比随机策略方法高 10 倍,这也是确定性策略方法最主要的优点.

2017 年,Jiang 等人将深度学和确定性策略应用在加密货币的投资组合中,通过将资金不断分配到不同的加密货币,获得更大累计收益[68].该系统包括独立评估集合(ensemble of identical independent evaluators,简称 EIIE)、投资组合内存(portfolio-vector memory,简称 PVM)、在线随机批量学习(online stochastic batch learning,简称 OSBL)和针对即时奖励的奖励函数.

Jiang 等人重新设计了 Actor-Critic 方法的状态、回报和动作,Actor 使用确定性策略梯度实现,Actor 的交易动作定义为下一个时间段 t 下各类资产分配的权重数值,用矢量 $w_t=\{x_1,\ldots,x_i\}$ 表示,x_i 的和为 1,见公式(40).

$$a_t=w_t \tag{40}$$

状态 s_t 则由当前时刻的价格张量 X_t(由最高价、最低价、收盘价组成)和前一时刻的资产分配权重 w_{t-1} 组成,见公式(41).

$$s_t=(X_t,w_{t-1}) \tag{41}$$

回报则用收益率的对数回报率表示.Jiang 采用深度神经网络作为确定性策略梯度函数 π_θ,并测试了 CNN,RNN,LSTM 这 3 个模型.例如,用 CNN 模型对输入特征(X_t,w_{t-1})进行采样,直接用 *softmax* 层的输出作为权重分

配值 w_{t-1},而在通常的分类任务中,常取 *softmax* 的最大值作为分类答案.同时,在训练过程中,依靠投资组合内存(portfolio-vector memory,简称 PVM)和小批量训练这两种机制进行训练.PVM 与强化学习的 DQN 经验回放机制非常相似:首先,通过引入外部存储机制,存储数据不断加入到训练数据中,使得训练数据尽量满足均衡分布,避免过拟合;然后,用小批量数据训练,每个批次内的数据必须是完整时间序列.对神经网络训练而言,即使它们具有显着重叠的间隔,不同时期的数据依然被认为是独特而有效的.这个系统依托在线随机批量学习方式,可以直接应用到在线上项目.在模型对比中,CNN,RNN 和 LSTM 占据了前三名,在比特币的虚拟交易中,即便在佣金率高达 0.25%的情况下,该系统仍然能够在 50 天内使收益增长为原来的 4 倍.

综上所述,深度强化学习在金融交易系统中的应用已经越来越多.随着深度强化学习在 2014 年后的强势兴起,带动了新一轮研究热潮.从模型结构上看,深度学习与强化学习的结合方式多种多样,在不同的应用领域各有优势:在单资产投资中,借助深度学习提取特征的 RRL 学习方法有效性依然很高,依托不同的目标函数应对不同的市场风格变化;而在资产组合交易中,基于策略搜索的深度强化学习方法显得更加灵活,状态和动作设计也不受模型局限.

7 结 论

本文综述了强化学习在金融交易领域的应用进展情况,包括 RRL、*Q* 学习、Actor-Critic、A3C 算法和结合深度神经网络的各类强化学习算法;以及依托强化学习构建的各类金融交易系统,在股票、指数、期货、投资组合、虚拟货币等交易领域的应用,基于强化学习的各类金融交易系统在风险控制、交易进出场时机、资金管理等方面都取得了突破.

基于强化学习将促进自动交易系统的进一步发展,可预见的趋势至少有两个方面.

(1) 经典的 RRL 模型将继续发展,但是 RRL 基于循环的自适应框架将会得到保留.在目标函数的选择上将变得更加灵活多样,在金融资产序列的特征提取上将更多地采用深度学习模型;

(2) 随着 A3C 算法的进一步发展,产业界与学术界将目光投向多智能体并行处理的方式,A3C 是在策略(on policy)算法,效果、时间和资源消耗上都优于 DQN 和 DDPG,它的应用有望部分解决强化学习策略受到的限制.

本文认为,上述研究中仍然存在着亟待解决的问题.

(1) 金融市场具有不稳定性,趋势实时变化.从历史的训练数据中学到的知识可能不会在后续测试数据中有良好的效果,这对强化学习模型的适应性提出了更高的要求,不同市场条件下如何选择合适的强化学习模型和深度学习模型仍然是一个悬而未决的问题;

(2) 构建基于强化学习的交易软件或系统.通常,一种算法不能解决全部问题,针对不同的市场情况,需要设置不同的配置模块.风险层、策略轮动层、自适应层等层次结构的设计至今没有统一解决方案,业界仍然在探索中;

(3) 大部分强化学习模型系统都是专攻某一类金融交易,单纯地做多、做空或空仓观望等,投资组合方式也仅是对各类金融资产的权重进行重新分配.但是,如股票中性、期货中性等策略需要对多种资产同时进行复杂的多空对冲操作时,仍缺少充分的研究;

(4) 强化学习领域最近提出了确定性策略和蒙特卡罗树搜索结合的算法,并应用于围棋领域[69],获得了突破.如何将蒙特卡罗树搜索策略应用在交易系统中,值得深入研究.

最后还要强调,深入研究强化学习理论、完善金融交易系统的组成结构、在提高交易的利润的同时降低交易风险,这是基于强化学习的金融交易系统研究的核心问题.

References:

[1] Fama Eugene F. Random walks in stock market prices. Financial Analysts Journal, 1965,21(5):55−59.

[2] Farmer JD. Market force, ecology and evolution. Computing in Economics & Finance, 1998,11(5):895−953(59). [doi: 10.1093/icc/11.5.895]

[3] Lo AW. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. Social Science Electronic Publishing, 2004. [doi: 10.3905/jpm.2004.442611]

[4] Lo AW. Reconciling efficient markets with behavioral finance: The adaptive markets hypothesis. Journal of Investment Consulting, 2005. http://ssrn.com/abstract=728864

[5] Sutton RS, Barto AG. Introduction to Reinforcement Learning. Vol.135. Cambridge: MIT Press, 1998. http://legacydirs.umiacs.umd.edu/~hal/courses/2016F_RL/RL9.pdf

[6] Kuleshov V, Precup D. Algorithms for the multi-armed bandit problem. Journal of Machine Learning Research, 2000,1:1−48. http://cn.arxiv.org/pdf/1402.6028

[7] Moody J, Saffell M. Reinforcement learning for trading. In: Proc. of the Conf. on Advances in Neural Information Processing Systems II. MIT Press, 1999. 917−923.

[8] Moody J, Wu L, Liao Y, Saffell M. Performance functions and reinforcement learning for trading systems and portfolios. Journal of Forecasting, 1998,17(5-6):441−470. [doi: 10.1002/(sici)1099-131x(1998090)17:5/6<441::aid-for707>3.3.co;2-r]

[9] Moody J, Saffell M. Learning to trade via direct reinforcement. IEEE Trans. on Neural Networks, 2001,12(4):875−889. [doi: 10.1109/72.935097]

[10] Gold C. FX trading via recurrent reinforcement learning. In: Proc. of the IEEE Int'l Conf. on Computational Intelligence for Financial Engineering. IEEE, 2003. 363−370. [doi: 10.1109/cifer.2003.1196283]

[11] Gorse D. Application of stochastic recurrent reinforcement learning to index trading. In: Proc. of the Esann 2011, European Symp. on Artificial Neural Networks. Bruges: DBLP, 2011. http://pdfs.semanticscholar.org/e7aa/08a404bb879cae6fcb751394a29465078e56.pdf

[12] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006,313(5786):504−507. [doi: 10.1126/science.1127647]

[13] Zhang J, Maringer D. Indicator selection for daily equity trading with recurrent reinforcement learning. In: Proc. of the Conf. Companion on Genetic and Evolutionary Computation. ACM Press, 2013. 1757−1758. [doi: 10.1145/2464576.2480773]

[14] Zhang J, Maringer D. Using a genetic algorithm to improve recurrent reinforcement learning for equity trading. Computational Economics, 2016,47(4):551−567. [doi: 10.1007/s10614-015-9490-y]

[15] Werbos PJ. Advanced forecasting methods for global crisis warning and models of intelligence. General Systems Yearbook, 1977, 22(6):25−38.

[16] Bertsekas DP, Tsitsiklis JN. Neuro-dynamic programming: An overview. In: Proc. of the IEEE Conf. on Decision and Control. IEEE, 1995. 560−564. [doi: 10.1109/cdc.1995.478953]

[17] Lewis FL, Vrabie D. Reinforcement learning and adaptive dynamic programming for feedback control. IEEE Circuits and Systems Magazine, 2009,9(3):32−50. [doi: 10.1109/MCAS.2009.933854]

[18] Liu D, Wei Q. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. IEEE Trans. on Neural Networks and Learning Systems, 2014,25(3):621−634. [doi: 10.1109/tnnls.2013.2281663]

[19] Zhao H, Wang B, Liao J, Wang H, Tan G. Adaptive dynamic programming for control: algorithms and stability. Communications & Control Engineering, 2013,54(45):6019−6022.

[20] Atiya AF, Parlos AG, Ingber L. A reinforcement learning method based on adaptive simulated annealing. In: Proc. of the 2003 IEEE Midwest Symp. on Circuits and Systems. IEEE, 2003. 121−124. [doi: 10.1109/mwscas.2003.1562233]

[21] Jangmin O, Lee J, Lee JW, Zhang BT. Adaptive stock trading with dynamic asset allocation using reinforcement learning. Information Sciences, 2006,176(15):2121−2147. [doi: 10.1016/j.ins.2005.10.009]

[22] Dempster MAH, Leemans V. An automated FX trading system using adaptive reinforcement learning. Expert Systems with Applications, 2006,30(3):543−552. [doi: 10.1016/j.eswa.2005.10.012]

[23] Bertoluzzo F, Corazza M. Making financial trading by recurrent reinforcement learning. In: Proc. of the Int'l Conf. on Knowledge-based and Intelligent Information and Engineering Systems. Berlin, Heidelberg: Springer-Verlag, 2007. 619−626. [doi: 10.1007/978-3-540-74827-4_78].

[24] Tan Z, Quek C, Cheng PYK. Stock trading with cycles: A financial application of ANFIS and reinforcement learning. Expert Systems with Applications, 2011,38(5):4741−4755. [doi: 10.1016/j.eswa.2010.09.001]

[25] Almahdi S, Yang SY. An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. Expert Systems with Applications, 2017,87:267−279. [doi: 10.1016/j.eswa.2017.06.023]

[26] Hamilton JD. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 1989, 57(2):357−384. [doi: 10.2307/1912559]

[27] Hamilton JD, Susmel R. Autoregressive conditional heteroskedasticity and changes in regime. Journal of Econometrics, 1994, 64(1-2):307−333. [doi: 10.1016/0304-4076(94)90067-1]

[28] Gray SF. Modeling the conditional distribution of interest rates as a regime-switching process. Journal of Financial Economics, 1996,42(1):27−62. [doi: 10.1016/0304-405x(96)00875-6]

[29] Maringer D, Ramtohul T. Regime-switching recurrent reinforcement learning for investment decision making. Computational Management Science, 2012,9(1):89−107. [doi: 10.1007/s10287-011-0131-1]

[30] Maringer D, Ramtohul T. Threshold recurrent reinforcement learning model for automated trading. In: Proc. of the Applications of Evolutionary Computation, Evoapplications 2010: Evocomnet, Evoenvironment, Evofin, Evomusart, and Evotranslog. Istanbul: DBLP, 2010. 212−221. [doi: 10.1007/978-3-642-12242-2_22]

[31] Maringer D, Ramtohul T. Regime-switching recurrent reinforcement learning in automated trading. In: Proc. of the Natural Computing in Computational Finance. Berlin, Heidelberg: Springer-Verlag, 2011. 93−121. [doi: 10.1007/978-3-642-23336-4_6]

[32] Maringer D, Zhang J. Transition variable selection for regime switching recurrent reinforcement learning. In: Proc. of the Computational Intelligence for Financial Engineering & Economics. IEEE, 2014. 407−413. [doi: 10.1109/cifer.2014.6924102]

[33] Wierstra D, Förster A, Peters J, Schmidhuber J. Recurrent policy gradients. Logic Journal of Igpl, 2010,18(2010):620−634. [doi: 10.1093/jigpal/jzp049]

[34] Baird L, Moore A. Gradient descent for general reinforcement learning. In: Proc. of the Conf. on Advances in Neural Information Processing Systems II. MIT Press, 1999. 968−974.

[35] Watkins CJCH. Learning from delayed rewards. Robotics & Autonomous Systems, 1989,15(4):233−235.

[36] Jaakkola T, Jordan MI, Singh SP. On the convergence of stochastic iterative dynamic programming algorithms. Neural Computation, 1993,6(6):1185−1201. [doi: 10.21236/ada276517]

[37] Tsitsiklis JN. Asynchronous stochastic approximation and *Q*-learning. Machine Learning, 1994,16(3):185−202. [doi: 10.1007/bf00993306]

[38] Watkins CJCH, Dayan P. Technical note: *Q*-learning. Machine Learning, 1992,8(3-4):279−292. [doi: 10.1007/978-1-4615-3618-5_4]

[39] Moore AW, Atkeson CG. Prioritized sweeping: Reinforcement learning with less data and less time. Machine Learning, 1993,13(1): 103−130. [doi: 10.1007/bf00993104]

[40] Mahadevan S, Maggioni M. Proto-value functions: A laplacian framework for learning representation and control in markov decision processes. Journal of Machine Learning Research, 2007,8:2169−2231. [doi: 10.1145/1102351.1102421]

[41] Sutton RS. Policy gradient methods for reinforcement learning with function approximation. Submitted to Advances in Neural Information Processing Systems, 1999,12:1057−1063.

[42] Lee JW, Jangmin O. A multi-agent *Q*-learning framework for optimizing stock trading systems. In: Proc. of the Int'l Conf. on Database and Expert Systems Applications. Springer-Verlag, 2002. 153−162. [doi: 10.1007/3-540-46146-9_16]

[43] Lee JW, Park J, Jangmin O, Lee J, Hong E. A multiagent approach to Q-learning for daily stock trading. IEEE Trans. on Systems Man & Cybernetics—Part A: Systems & Humans, 2007,37(6):864−877. [doi: 10.1109/tsmca.2007.904825]

[44] Li J, Chan L. Reward adjustment reinforcement learning for risk-averse asset allocation. In: Proc. of the IEEE Int'l Joint Conf. on Neural Network. 2006. 534−541. [doi: 10.1109/ijcnn.2006.246728]

[45] Bertoluzzo F, Corazza M. Reinforcement learning for automatic financial trading: Introduction and some applications. Working Papers, 2012. [doi: 10.2139/ssrn.2192034]

[46] Bertoluzzo F, Corazza M. Testing different reinforcement learning configurations for financial trading: Introduction and applications. Procedia Economics & Finance, 2012,3(338):68−77. [doi: 10.1016/s2212-5671(12)00122-0]

[47] Corazza M, Bertoluzzo F. *Q*-learning-based financial trading systems with applications. Social Science Electronic Publishing, 2014. [doi: 10.2139/ssrn.2507826]

[48] Du X, Zhai JJ, Lv KP. Algorithm trading using *q*-learning and recurrent reinforcement learning. 2016. http://cs229.stanford.edu/proj2009/LvDuZhai.pdf

[49] Eilers D, Dunis CL, von Mettenheim HJ, Breitner MH. Intelligent trading of seasonal effects: A decision support algorithm based on reinforcement learning. Decision Support Systems, 2014,64:100−108. [doi: 10.1016/j.dss.2014.04.011]

[50] Konda V. Actor-critic algorithms. Siam Journal on Control & Optimization, 1999,42(4):1143−1166. http://papers.nips.cc/paper/1786-actor-critic-algorithms.pdf

[51] Li H, Dagli CH, Enke D. Short-term stock market timing prediction under reinforcement learning schemes. In: Proc. of the IEEE Int'l Symp. on Approximate Dynamic Programming and Reinforcement Learning. IEEE, 2007. 233−240. [doi: 10.1109/adprl.2007.368193]

[52] Bekiros SD. Heterogeneous trading strategies with adaptive fuzzy actor—Critic reinforcement learning: A behavioral approach. Journal of Economic Dynamics & Control, 2010,34(6):1153−1170. [doi: 10.1016/j.jedc.2010.01.015]

[53] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D. Playing atari with deep reinforcement learning. Computer Science, 2013.

[54] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540):529. [doi: 10.1038/nature14236]

[55] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa, Y. Continuous control with deep reinforcement learning. Computer Science, 2015,8(6):A187.

[56] Mnih V, Badia AP, Mirza M, Graves A, Lillicrap TP, Harley T. Asynchronous methods for deep reinforcement learning. 2016.

[57] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. of the 26th Annual Conf. on Neural Information Processing Systems. Nevada, 2012. 1097−1105. [doi: 10.1145/3065386]

[58] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S. Image net large scale visual recognition challenge. Int'l Journal of Computer Vision, 2015,115(3):211−252. [doi: 10.1007/s11263-015-0816-y]

[59] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In: Proc. of the IEEE Conf. on Acoustics, Speech and NAL Processing. Vancouver, 2013. 6645−6649. [doi: 10.1109/icassp.2013.6638947]

[60] Li YX, Zhang JQ, Pan D, Hu D. A study of speech recognition based on RNN-RBM language model. Journal of Computer Research a Development, 2014,51(9):1936−1944 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2014.20140211]

[61] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Doha, 2014. 1724−1734. [doi: 10.3115/v1/d14-1179]

[62] Yang Z, Tao DP, Zhang SY, Jin LW. Similar handwritten Chinese character recognition based on deep neural networks with big data. Journal on Communications, 2014,35(9):184−189 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-436x.2014.09.019]

[63] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F. Large-scale video classification with convolutional neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Columbus, 2014. 1725−1732. [doi: 10.1109/cvpr.2014.223]

[64] Sun ZJ, Xue L, Xu YM, Wang Z. Overview of deep learning. Application Research of Computers, 2012,29(8):2806−2810 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-3695.2012.08.002]

[65] Deng Y, Bao F, Kong Y, Ren Z, Dai Q. Deep direct reinforcement learning for financial signal representation and trading. IEEE Trans. on Neural Networks and Learning Systems, 2017,28(3):653−664. [doi: 10.1109/tnnls.2016.2522401]

[66] Lu DW. Agent inspired trading using recurrent reinforcement learning and LSTM neural networks. Papers, 2017. https://arxiv.org/pdf/1707.07338.pdf

[67] Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proc. of the Int'l Conf. on Machine Learning. 2014. 387−395.

[68] Jiang ZY, Xu DX, Liang JJ. A deep reinforcement learning framework for the financial portfolio management problem. arXiv preprint arXiv:1706.10059, 2017. https://arxiv.org/abs/1706.10059

[69] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A. Mastering the game of Go without human knowledge. Nature, 2017,550(7676):354−359. [doi: 10.1038/nature24270]

附中文参考文献:

[60] 黎亚雄,张坚强,潘登,等.基于 RNN-RBM 语言模型的语音识别研究计算机研究与发展, 2014,51(9):1936−1944.

[62] 杨钊,陶大鹏,张树业,等.大数据下的基于深度神经网的相似汉字识别.通信学报,2014,35(9):184−189.

[64] 孙志军,薛磊,许阳明,等.深度学习研究综述.计算机应用研究,2012,29(8):2806−2810.

梁天新(1984－),男,黑龙江齐齐哈尔人,博士生,CCF 学生会员,主要研究领域为自然语言处理,深度学习,机器学习,强化学习.

王良(1963－),男,博士,副教授,CCF 高级会员,主要研究领域为智能科学,数据库管理系统,数据库系统评价和性能优化.

杨小平(1956－),男,博士,教授,博士生导师,主要研究领域为信息系统工程,电子政务,网络安全技术.

韩镇远(1993－),男,硕士生,主要研究领域为深度学习,自然语言处理.