计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900127

机器学习在金融资产定价中的应用研究综述

许　杰[1]　祝玉坤[1]　邢春晓[2]
1 清华大学五道口金融学院　北京 100084
2 清华大学北京信息科学与技术国家研究中心　北京 100084
(xujie@pbcsf.tsinghua.edu.cn)

摘　要　金融资产配置的关键问题是资产的价格,资产定价是现代金融学的核心内容,揭示资产定价规律一直是金融研究热点之一。文中回顾了机器学习在资产定价领域使用的方法与研究进展,将机器学习资产定价的方法分类为基于特征处理的机器学习方法与端到端处理的深度学习方法;围绕当前机器学习资产定价遇到的主要问题,比较了不同算法在原理和应用场景方面的区别;指出了两类机器学习方法的适用性与局限性;讨论了机器学习资产定价未来可能的研究趋势。

关键词:机器学习;资产定价;投资组合;价格预测;深度学习
中图法分类号　TP181

Application of Machine Learning in Financial Asset Pricing:A Review

XU Jie[1],ZHU Yu-kun[1] and XING Chun-xiao[2]
1 PBC School of Finance,Tsinghua University,Beijing 100084,China
2 Beijing National Research Center for Information Science and Technology(BNRist),Tsinghua University,Beijing 100084,China

Abstract　The key problem of financial asset allocation is asset price. Asset pricing is the core content of modern finance,which indicates that asset pricing law has always been one of the hot topics of financial research. This paper reviews the methods used by machine learning in the field of asset pricing and research progresses,classifies machine learning asset pricing method into machine learning method based on the characteristics processing and deep learning method based on end-to-end processing,compares the differences between different algorithms in principle and application scenarios,points out the applicability and limitations of the two kinds of machine learning methods,prospects the research direction on machine learning asset pricing in the future.

Keywords　Machine learning,Asset pricing,Portfolio,Price forecasting,Deep learning

1　概述

资产定价指在不确定条件下对未来资产的价格或者价值进行重估。本文所指资产为金融工具或某类证券,而价格指反映了各种影响因素(如基本面、风险和情绪等),由市场需求与供给共同决定的价格。已有众多学者从不同角度研究该类资产定价的规律,如随机漫步理论、有效市场假说及行为金融学等。随机漫步理论指市场对随机事件的反应具有布朗运动随机性,该理论认为价格不具有可预测性,把预测股票走势认定为"傻瓜的游戏";有效市场假说将市场分为弱型有效市场、半强型有效市场和强型有效市场,该理论认为股票价格能完全反映所有关于该资产的有效信息,然而随着反转效应、动量效应以及市值效应被相继发现,有效市场假说理论的有效性较低[1];行为金融学则认为股价不仅受企业的内在价值影响,也受投资者个体行为、群体主体行为的影响。

金融市场是十分复杂且不断变化发展的系统,股票市场的运行规律一直受到高度关注[2],分析方法主要包括基本面分析和技术分析。众多方法分析了对股市有影响的因子,从而产生了因子动物园的说法[3]。这些方法包括简单的线性回归和非线性拟合、传统方法和机器学习等[2,4-8]。一方面,随着数据的爆发,金融市场包含大量的噪声及不确定性因素,当因子特征维度变大时,非线性的考虑使得预测函数形式的搜索复杂度急剧增加,因此传统计量方法和线性方法不适用于分析复杂、高维且具有噪声的金融市场数据序列[9];另一方面,机器学习在海量数据的处理与分析上取得了重要突破,已经被广泛应用于计算机、生物、医疗、传媒和金融等领域。其中,使用机器学习进行资产定价的相关研究具有算法效果好、适用性强、易于处理大数据的特点,带来了新的解决思路。

相比传统计量与统计模型资产定价分析方法,机器学习

到稿日期:2021-09-15　返修日期:2021-12-05
基金项目:科技部重点研发计划:现代服务可信交易理论与技术研究(2018YFB1402701)
This work was supported by the Key research and Development Plan of Ministry of Science and Technology:Research on the Theory and Technology of Modern Service Trusted Transaction(2018YFB1402701).
通信作者:邢春晓(xingcx@mail.tsinghua.edu.cn)

的优势主要体现在以下几个方面。1）机器学习利用端到端的处理方式，取消了复杂的经济金融学原理知识与模型设定，具备从历史数据中学习经验知识与特征的能力，并基于此预测资产的未来价格。2）机器学习算法具有天然的处理非结构化数据的能力，能抽取深层次的潜在特征，其取消了传统计量方法中假定数据特征关系与协方差矩阵等的计算，能更加全面地描述金融规律，是对结构化数据分析的重要补充。3）金融时序数据具有非线性、非平衡性、高维度特性和高噪声性质等特点，传统计量模型的研究范式难以取得进一步的突破，而机器学习并未对模型函数形式做出严格的假定，取消了金融市场的变量概率统计分布等假设，更为简便且更具优势。机器学习方法不仅能有效处理大量金融数据，更是一种新的思维研究模式。

从数学的角度来看，机器学习表现为一种变量空间的映射关系，其使学习到的函数能较好地表征原有数据规律，最大化逼近真实函数曲线。传统的计量资产定价研究主要关注市场规律研究，而机器学习主要关注数据处理与算法本身的改进、特征深层次提取和特征相互关系研究等，因此本文重点从技术角度综述资产定价，为从事资产定价的研究人员提供技术方面的借鉴。本文主要从基于特征处理的机器学习资产定价方法与基于端到端的深度学习资产定价方法两个方面进行了介绍，回顾了机器学习在资产定价领域已有的方法与研究进展，讨论了当前机器学习资产定价方法遇到的主要问题，指出了不同算法应用场景的区别，分析了不同算法的原理、优势与劣势，并对机器学习金融资产定价问题进行了展望，最后总结全文。

2　机器学习资产定价方法类别

本文所述的资产定价模型属于广义资产定价，包括风险-收益比（Risk-Return Trade-Off）模型与直接预测资产价格模型。资产定价数据包括结构化数据与非结构化数据，数据的质量决定了模型性能的上限。结构化数据包括股票交易数据、期货交易数据、外汇交易数据、宏观数据等，以及衍生出的市场因子、规模、价值、动量、盈利、换手率等。非结构化数据包括 Facebook，Twitter 和 Youtube 等社交媒体数据，以及上市公司季报内容、年报内容、金融中介分析报告、金融图片、企业知识图谱数据等。需要注意的是，结构化数据与非结构化数据并非完全对立，而是存在交叉的情况。根据对数据类型的处理方式，将机器学习模型分成两类：1）基于特征处理的机器学习方法，包括主成分分析、奇异值分解、独立成分分析、贝叶斯算法、马尔可夫算法、自动编码机、支持向量机（Support Vector Machine，SVM）、决策树算法、随机森林、粒子群算法、遗传算法、迁移算法和集成学习算法等；2）端到端处理深度学习方法，包括 CNN、LSTM、强化学习、文本分析、知识图谱和模型融合等。机器学习模型的基本架构如图 1 所示。

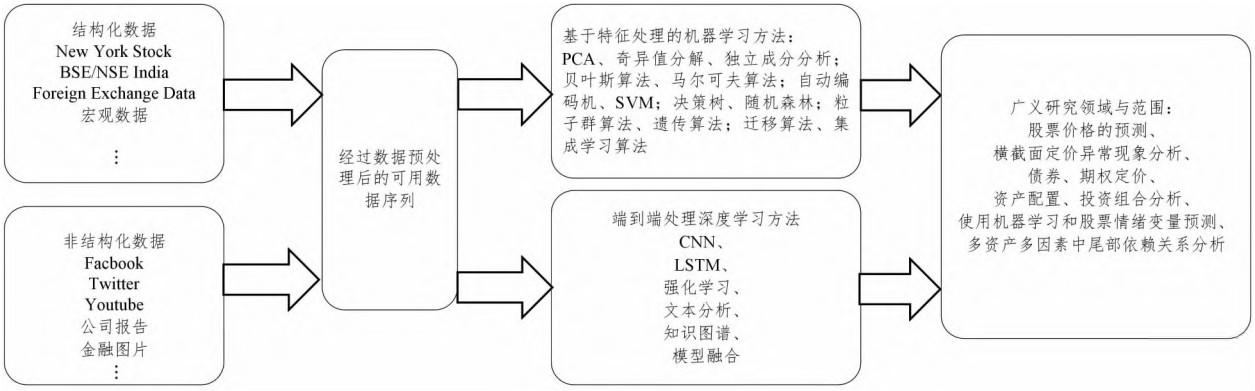

图 1　机器学习资产定价算法的流程

Fig. 1　Process of machine learning asset pricing algorithm

3　基于特征处理的机器学习方法

特征工程是提高预测准确率的关键步骤，其基于领域知识从原始数据中最大限度地提取输入特征，减少特征损失，以供后续模型使用。数据特征决定了机器学习的上限，特征工程包括数据清洗、去噪、异常样本处理、数据不均衡处理、数据归一化处理、数据离散化和缺失补齐等[10]。基于特征处理的机器学习方法又可细分为：1）基于高维数据维度约减算法，即主成分分析、奇异值分解和独立成分分析；2）基于数理统计算法，即贝叶斯算法和马尔可夫算法；3）基于高低维转换算法，即自动编码机和 SVM；4）基于传统机器学习分类理论算法，即随机森林、决策树算法；5）基于启发式算法，即粒子群算法、遗传算法；6）其他算法，如迁移算法、集成学习算法。

3.1　基于高维数据维度约减算法

（1）主成分分析

主成分分析（Principal Components Analysis，PCA）是一种统计分析与简化数据集矩阵的方法。传统金融数据分析需要考虑多个协同变量之间的相互作用，使分析和预测具有一定的难度，而 PCA 利用正交变换对变量观测值进行线性变换，产生一系列线性不相关变量，即主成分（Principal Components），从而达到变量降维的目的，其可用于消除金融时序数据多元回归方程多重共线性问题，是一种多变量大样本数据处理的有效方法，在经济金融应用中较为常见。文献[11]扩展了主成分分析，将主成分分析与套利定价相结合，解释了数据中同方向变化无套利因素，提出了一种针对高维度财务面板数据的预期超额收益模型，解释了产生预期回报和协方差结构的原因，考虑了预期收益中的定价误差，在 PCA 中加入

无套利惩罚项可以解决金融数据低信噪比的问题，并获取与内核定价相关的信息。文献[12]将 PCA 用于股票预测模型，使用指向性主成分分析，用可观察到的特征来引导那些不能观察到的动态因子，获得与风险补偿相应的收益/补偿关系。相反，如果此类因子不存在，则推断这些特征效应是无风险补偿，被认定为异象因子。文献[13]将高频数据与面板数据集相结合，使用 PCA 方法处理局部波动和跳跃协方差矩阵，非平均地捕捉跳跃性因子的时序变化。传统方法分析一个因素结构仅限于一组预先指定的因子，而文献[13]估计了面板数据上未知的连续跳跃因子的结构，该方法无须事先了解收益的结构类型。文献[14]提出了一种漂移参数渐近框架，当因子不能解释截面与时间相关的情况时，显式地将这些偏差与特殊项协方差结构联系起来，在只有中等偏置的情况下估计渐近参数，解释了主成分估计器失效的原因。MonteCarlo 实验证明了渐近公式和估计值在较小的可观测矩阵上也能取得较好的效果。文献[15]提出了一种在采样频率和协方差矩阵维数都增加的情况下，利用高频数据对高维数据公共因子进行估计的方法。股票投资组合协方差矩阵表示为低秩公共结构与稀疏残差矩阵，PCA 揭示的因子比可观察到的投资组合因子(如市场投资组合、Fama-French 投资组合以及 ETF 投资组合)解释的资产收益比例更大、样本外估计效果更好。文献[16]使用 PCA 方法研究基于期权未平仓权益，指出市场因子由隐含波动率回报加权日复利来决定，分析了 9 个决定性影响因子和数据冗余。

从上述文献可以看到，PCA 的优势在于其将随机信号中最有意义的子信号包含在方差逐渐减小的不相关随机变量中，善于处理高频高维金融数据。对大规模面板型数据进行矩阵分解，可提取潜在因子。此外，PCA 不需要预先指定可观察公共因子，降低了对研究人员的专业性要求。但是，PCA 未能利用已有条件变量来确定因素结构，增加了从数据中构建最有信息价值特征的难度 。基于 PCA 的统计因素分析在识别小方差因子方面效果一般，且对于大量冗余股票特征子集，其只能解释特定排序组合中的小部分变化，导致主成分向量方差减小，降低了算法的效果。

(2)奇异值分解

与 PCA 相似，奇异值分解(Singular Value Decomposition，SVD)也是一种有效处理维度约减的方法，在机器学习领域应用广泛，可用于特征分解降维、压缩去噪、推荐系统、自然语言处理等领域。奇异值分解并不要求被分解矩阵为方阵，通过 $M=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{*}$ 变换，其中 $\boldsymbol{U}$ 是 $m\times m$ 阶酉矩阵，$\boldsymbol{V}$ 是 $m\times n$ 阶实数对角矩阵，$\boldsymbol{\Sigma}$ 是奇异值矩阵，在低秩条件下，对矩形数据进行近似，可同时获得左奇异向量和右奇异向量矩阵。奇异值分解常被用于金融数据降维与压缩。文献[17]设计了基于奇异值分解的道琼斯工业指数成分股熵矩阵，计算每日数据和月数据的熵时变度量，熵越大，系统混乱情况越严重，此外还对时变熵指数和道琼斯工业平均指数进行格兰杰因果检验。文献[18]引入了多尺度奇异值分解熵概念。在两个非平稳、非高斯分布之间，两组序列滑动点积采用奇异值分解相关矩阵，在不同时间规模上估计两组时间序列的相关性，结果表明，在周期不足一个月时，奇异值分解熵对道琼斯工业平均指数具有较好的预测能力。文献[19]提出了一种基于可调奇异值分解算法，利用 SVD 提取股票收益因子模型中的财务因子，通过调整算法参数来减小模型对数据矩阵的误差敏感性，提高了模型的鲁棒性和有效性。

可以发现，当不同来源的数据的分量量纲不一致时，不能直接套用 PCA，如某国的 GDP(美元)、失业率(比率)、政府清廉指数(比率)等，对 PCA 分量进行直接运算没有实际意义。PCA 对噪声敏感，在低信噪比情况下，其算法效果较差。SVD 可以获得两个方向上的主成分，通常比直接使用 PCA 更稳定。

(3)独立成分分析

独立成分分析(Independent Component Analysis，ICA)是近年来迅猛发展的数学挖掘工具，其目的是通过观察随机信号 x 来估计混合信号矩阵 $\boldsymbol{A}$ 以及源信号 s，将数据或信号分离成独立的非高斯信号线性组合，并且在源信号非高斯的情况下，将独立成分分解成唯一线性组合，可携带更多信息。ICA 是 PCA 的一个扩展，其使用两步法来实现，即经 PCA 得到向量组 $\boldsymbol{y}$，再将 $\boldsymbol{y}$ 正交变换，输出独立信号。PCA 考虑了金融数据的二阶统计量，但对于更高阶的非高斯分布数据，ICA 更能发现数据的内在特征，能更有效地挖掘随机变量、测量数据和隐藏变量关系。对于股票金融市场数据，找到数据中最大的独立信号，即可最大限度地降低风险，获得最优投资组合。同样地，在进行 ICA 之前，需要对金融数据进行去噪等基础操作。文献[20]使用 ICA 处理多变量金融序列股票投资的组合应用，ICA 的核心思想是将观测到的多元时间序列线性映射到独立分量空间，具有导致股票价格发生重大变化、频率较高但对股票的总体水平贡献不大两类特征，噪声大小取决于幅度大小，与频率无关，ICA 为理解影响股票市场数据的机制提供了一个新的视角。

不管是 PCA 还是 ICA，均不需要假定源信号的具体分布。PCA 通过原始数据降维，可以成为 ICA 预处理的步骤。ICA 主要关注独立性，对信号的能量与方差不敏感。ICA 根据数据对证券价格的不同影响程度，将映射后的独立分量细化分类。混合信号经线性变换不会影响 ICA 的结果，但会对 PCA 的结果造成影响。如果观测到的信号为高斯分布，此时 PCA 和 ICA 等价。

3.2 基于数理统计算法

(1)贝叶斯分类算法

贝叶斯分类算法是一类以贝叶斯定理为基础的分类算法，由一个无环图 G 和 n 个条件概率组成的贝叶斯网络，又称为贝叶斯置信网络。基于网络拓扑结构的有向图描述，适用于推理不完全事件或概率性事件，在金融数据稀少的情况下可取得较好的分类或预测效果。金融领域在实际应用贝叶斯模型时，一般都会限定各事件之间的先后发生关系与内在因果关系，使用反向算法取得有向图中各状态之间的转移概率，然后由先验概率推导出后验概念。将先验知识纳入评估过程，不仅考虑了资产配置中所涉及的财务风险等因子，还考虑了资产定价模型中由不确定参数值带来的估计风险。投资者可以使用特定资产定价模型来做出投资决策[21]。文献[22]实现了一种贝叶斯学习方法，由于资产价格和基本因子

具有共同趋势，在正常时期，资产价格除以股息被假定遵循一个围绕长期均值波的动回归过程。在股市泡沫期间，资产价格剧烈波动，采用贝叶斯学习方法对所提模型的潜在状态和参数进行实时联合估计，在 S&P 500 上的实证分析表明，该模型能较好地发现股市是否存在泡沫。文献[23]研究了资产定价模型参数的先验信念对市场投资组合和动量投资组合之间的最优资金配置影响。投资者在决策时可获得的信息总量为下一期超额收益概率密度函数与收益相乘后再求总积分，使用蒙特卡罗方法求得解。贝叶斯方法检验动量效应强度，表明当投资者对 CAPM 定价缺乏信心时，先验知识越少，越依赖于动量翻转，持有更多头寸。文献[24]提出了一种显著随机波动特征与累积消费贝叶斯过程，通过捕获消费、股息增长和资产回报联合动态，使用高频数据，预测包含消费和红利的多个随机波动过程。文献[25]基于 CAPM 贝叶斯来预测未来 alpha 值，以推断最能反映投资者行为的一组先验信念，比较了贝叶斯估计共同基金业绩与频率测量方法获得的业绩，证实投资者相信基金经理可以获得超额回报。

贝叶斯方法使我们能够对参数以及潜在状态进行有限样本推断，实时监测市场动态，其对缺失数据的敏感度较低、训练速度较快、结果解释度较高。

(2)马尔可夫算法

与贝叶斯算法相同，马尔可夫算法也属于概率模型。马尔可夫过程为状态空间中从一个状态转换到另一个状态的随机过程，规定下一状态的概率分布只能由当前状态决定，与该状态之前的状态无关，该算法常用于时序数据处理。与贝叶斯网络相似，其也是一种概论图模型(Probabilistic Graphical Model，PGM)算法，但马尔可夫网络是一种有向图推导网络，两个状态之间具有不可逆向推导性，其结构简单、鲁棒性强且可解释性好，多应用于股票价格序列标注难度较大的情况。文献[26]中的模型是马尔可夫正态混合模型的推广，在混合成分数和马尔可夫过程的限制条件下，推导了时间序列自协方差函数和整数幂线性表示，使用模型先验预测分布来研究模型对资产回报收益函数的影响。文献[27]提出由不同利率调整引起的长期均衡背离马尔可夫误差修正模型，当股价长期偏离价值时，可以看成对偏离长期均衡关系的调整。为了捕捉这些数据特征，提出了自适调整率长期均衡二阶段误差修正模型，即马尔可夫纠错模型，将高价格本利比与长期均衡特征相融合。

3.3　基于高低维转换算法

由于金融市场数据维度较大，一种常用的思路是先将高维数据转换为低维数据，将低维数据进行分类或预测后，再将其还原为高维数据，以解决维数灾难与过学习问题。自动编码机与 SVM 便属于此类方法。区别于 PCA，自动编码机与 SVM 使用非线性的方式降维。

(1)自动编码机

由于对非线性模型线性逼近会导致预测股权溢价或回报幅度出现较大误差，因此可以采用自动编码机的方法来处理。自动编码机是机器学习中的维度约减模型，属于一种特殊的无监督神经网络模型。类似 PCA，自动编码器可用于特征降维，其由模型编码器和解码器组成，作为特征提取器，可以由神经网络提取特征，然后输入其他分类模型中。自动编码器适合处理高维数据，其将高维特征表达成低维特征，是一种压缩数据无监督算法，输入变量通过隐藏层中的少量神经元形成输入变量压缩表示，通过非线性激活函数 $g(\cdot)$ 进行变量转换，与 PCA 算法的思路相似，但自动编码器可通过神经网络进行非线性映射，因此其更灵活，应用范围更广。对于不平衡数据，自动编码器从样本中直接学习特征，对相似类的训练可以移植到新的样本中，属于有损特征抽取与信息压缩。传统自动编码方法没有考虑协变量对特征因子及风险-收益的影响，因此文献[28]使用自动编码器将收益压缩成低维因子集，为个股收益引入条件自动编码器模型，允许资产特征协变量对因子暴露的非线性影响，非线性特征通过协变量神经网络映射，以贝塔形式展现出来，构建具有经济指导意义的自动编码器，在贝塔限定下，通过嵌入神经网络增强传统自动编码器。文献[29]提出了一种灵活描述收益率的曲线模型，为评估政府债券市场价格提供了参考，3 个自动生成的因子分别代表了水平、曲率和斜率。文献[30]将自动编码器和多层感知器相结合，在预训练期间分别训练多个自动编码器，后两层 MLP 自动编码器和 sigmoid 层共享参数。这种贪婪分层过程比随机初始化深度网络能产生更好的局部极小值，实现更好的泛化。但在样本数较少的情况下，自动编码分类的效果差于其他分类方法。

对比 PCA、奇异值分解、ICA 和自动编码机这 4 种方法可知，通过增强、减少或改良原有的交易数据特征，减少外部干扰信息，可以有效地提高效果。上述 4 种方法既可以作为预测方法，也可以作为其他算法的“准备”工作，即经过处理后的数据作为其他算法的输入。

(2)支持向量机

支持向量机是在分类与回归分析中的一种监督式学习模型，早期主要应用于模式识别领域，是一种特殊的学习算法。由于金融数据不断变化，市场实效性很强，前期预定模型可能在后续的金融数据中发生变化，模型适用性降低。此外，小样本数据(如风险暴露数据、欺诈数据等)的样本规模不大，使用神经网络等方法存在欠拟合问题。SVM 建立在结构风险最小化理论的基础上，使用核函数 Kernels 连续变化，通过最小化泛化误差上界来估计函数，可以较好地处理高维数据，泛化性能较好。SVM 以训练误差为优化问题约束条件，将置信范围值最小化作为优化目标，将低维空间数据通过核函数转换为高维空间数据，从而实现高维空间分类。因此，SVM 在小样本、动态变化数据预测上能力更强。另外，SVM 的时间训练开销较小，算法运算速度更快。另一个关键特性是训练 SVM 等价于求解线性约束二次规划问题，因此 SVM 的解总是全局唯一最优，不容易陷入局部极小。文献[31]提出了一种使用支持向量回归(Support Vector Regression，SVR)及遗传演算法(Genetic Algorithm，GA)的选股方法，该方法使用 SVM 按照股票回报对各支股票进行排名，排名靠前的股票被选择形成资产组合，通过遗传演算法进行特征选择，在参数空间使用全局最优化方法寻找参数最优解，从而获得输入变量的最优子集。文献[32]采用 SVM 与混合特征选择方法对股票趋势进行预测，结合滤波方法和包装方法的优点，从原始

特征集中选择最优特征子集，在3个常用方面（信息增益、对称不确定性和基于相关性特征选择），将通配 t 检验与BPNN算法的性能进行比较。文献[33]使用SVM模拟市场每周的变化，在技术分析方法中，SVM的输入为RSI和MACD两个指标，输出为股票集合和市场波动的程度（看涨或看跌）。文献[34]提出了一种财务决策支持系统风险评估，该系统融合了统计智能技术（如多层感知器和支持向量机），训练了112个SI MLPs和SVM，并在不同的数据库上与基准算法进行了对比，实证结果显示，具有sigmoid AF函数以及线性核函数SVM训练能达到更好的预测结果。文献[35]使用SVM来预测东京证券交易所225只股票指数的周趋势，由于日本的利率几乎长期为零，国内市场的消费能力有限，最大的出口目标是美国，因此选择S&P 500作为模型的输入参数之一，并将其与线性判别分析、二次判别分析、Elman神经网络性能进行了比较。文献[36]用SVM与反向传播神经网络来预测亚洲六大市场的股票走势，分析了亚洲资本市场的不稳定性特点，即对好事或坏事反应过度，实证结果表明，与早期研究相比，这两种模型都具有优越性。

相比支持向量机，传统神经网络模型遵循经验风险最小化原则，而SVM遵循结构风险最小化原则。SVM的解可由凸优化得到，且具有全局唯一性，达到了较高的泛化性能。

可以看到，SVM的主要问题包括：选取与优化SVM惩罚因子、参数时间和空间的成本较高，SVM的训练速度受到训练集规模的影响。SVM属于浅层学习方法，其学习能力有限，在金融时间序列中受噪声干扰较多，但在解决小样本、高维数和非线性等问题方面性能较好。

以上方法的主要思路是使用其他方法提取特征集，然后输入SVM进行预测或分类，再使用传统的参数最优化方法求解SVM模型参数。

3.4 基于传统机器学习分类理论算法

（1）决策树与随机森林

决策树与随机森林是金融市场常用的方法之一，其优点在于对特征的重要性进行了排序，且对数据分布没有限定。决策树是一系列决策规则的集合，将特征空间划分成有限不相交子区域，叶子节点代表分类结果。随机森林是一种基于决策树的集成算法，适用于大量数据超高频交易等。在文献[37]中，当候选变量数量很大时，投资组合排序和Fama-MacBeth回归法很难判断出影响横截面变化的变量，因此基于随机森林方法，以数据驱动方式在每个投资组合中最优地选择阈值，产生基于树的条件投资组合排序，并以此来平滑多个决策树的边界参数，提升了预测效果。文献[38]调查了不同宏观经济变量的波动如何影响加纳股票市场的流动性，提出了利用宏观经济变量来预测股票市场的RF和RNN混合机器学习模型，消除宏观因子在股票市场预测中的多重共线性问题。通过改进交叉验证策略，选择关键宏观经济预测因子，将随机森林应用于数据集特征排序，减小非排列树和排列树的误差差异。文献[39]分析了深度神经网络、梯度增强树、随机森林以及这些方法在S&P 500数据集上统计套利的有效性，实证结果表明，随机森林的表现优于梯度增强树和深度神经网络，资本市场的异常现象大多集中在月度数据上，盈利模式优先从最近的收益和每日数据中发现。

决策树与随机森林的优势如下：1）金融数据可能存在过采样或欠采样的情况，需要重新调整不同数据的采样频率以及占比权重。对于不平衡数据，决策树与随机森林有着天然的优势，通常表现良好。2）用多种不同模型和不同数据子集，将效果较差的模型合并成最终的预测模型，有助于解决过拟合问题。

3.5 基于启发式算法

粒子群算法（Particle Swarm Optimization，PSO）属于群智算法，其模拟鸟类捕食的过程，在固定区域内通过信息共享找到最大的食物源（全局最优解），从随机解出发，通过迭代寻找最优解。该方法容易实现且收敛速度快。遗传算法是计算数学中用于解决全局最优化的搜索算法，属于运筹学方法。两种算法都在全局解空间中集中搜索高可能性部分。文献[40]综述了PSO在股票组合优化、股票价格趋势预测等方面的优势，以及粒子群算法的应用意义，对群体智能和进化算法进行了分类。与遗传算法类似，粒子群算法是一种自然启发的群体智能方法，可以解决局部最优问题，被应用于各种连续价值优化问题。传统金融理论算法未考虑投资中的一些实际限制因素，如卖空机制、交易成本、投资比例和金融摩擦等，如果采用约束条件下的线性或非线性方程来求解，效果较差，而启发式的粒子群和遗传算法则能较好地处理。一般而言，两种算法会结合其他模型（如深度学习算法）共同使用。启发式算法的主要作用在于寻找配套模型的最优化参数，缩短参数调试时间，其在一些研究工作中已取得不错的效果。文献[41]提出了由ARIMA，ESM和RNN组成的统一股票价格预测模型，包括混合线性和非线性方法，使用遗传算法确定各组模型对预测结果的影响权重。类似地，文献[42]提出了一种将模板匹配与遗传算法相结合的方法，在该算法交易系统中，模板匹配主要用于识别整体上升趋势，对于具体买卖时间确定、噪声消除、滑动窗口大小等参数，则使用遗传算法调优得到。

相比遗传算法，粒子群优化算法的实现流程更简单，该算法在位置变动上具有较好的导向性，空间最优解逼近能力较强，但容易陷入局部最优解。遗传算法不论是交叉操作还是变异操作，都缺乏明确的导向性，其对空间最优解的搜索能力较强，但最优解逼近能力较弱。遗传算法具有较强的鲁棒性，不依赖金融数据的具体特征，算法并行能较快速地完成全局搜索。

单独运用粒子群算法或遗传算法来分析金融数据的情况较为少见，比较常见的是将其与其他算法融合，转而求解其他模型的参数，如求解神经网络中的权重参数、选取SVM的最优参数，将问题转化为在全局解空间中搜索最优参数。

3.6 其他算法

（1）迁移学习

迁移学习（Transfer Learning，TL）是将已经学习过的源域（Source Domain）知识应用到新目标域（Target Domain）中进行辅助学习。在图像处理领域，迁移学习方法在解决小数据问题中取得了较好效果，包括基于样本的迁移、基于特征的迁移和基于模式的迁移。迁移学习包括减小在新的空间源域

和目标域的距离以及原有算法在源域和目标域上的性能差距，将源域特征样本集与目标域特征样本集经再生核 Hilbert 空间映射后，使迁移后两类特征样本集边的缘概率分布尽可能相同。在金融市场中，以下 3 种情况可考虑使用迁移算法：1）训练集标签数据过少；2）新数据集比原数据集大，但数据类型与数据规律不完全一致；3）训练新数据模型的成本太高。因此，在时间序列数据量较少、股票预测效果受到了限制或者股票市场中相似行业的公司的股票价格变动规律具有相关性和联动性的情况下，可以使用迁移学习的方法。文献[43]通过对样本应用不同权重，对不同幅度的价格波动进行处理，使用迁移学习来处理数据稀缺性，将股票价格预测问题表述为市场收益回归问题，讨论了每个数据点的不同权重，从相同股票分布中提取新闻、技术指标和价格波动之间关系，因此迁移学习应用于不同市场波动中表现出较好的性能。文献[44]提出了一种考虑了因果关系的金融新闻机器学习模型来预测股票的价格波动，该方法利用转移熵寻找因果关系，基于韩国市场数据集与样本的测试表明，即使目标公司没有财务新闻，该模型使用具有因果关系公司的财务新闻也能预测股价走势。由于一年内每日收集的数据样本有限，文献[45]提出了一种基于特征迁移的模型，首先选择与给定股票有关系的股票，相似定义由最高余弦相似度（Cs）、相似场（SF）和最高市值（HMC）确定，然后基于给定股票、指数（即 KOSPI 200 或 S&P 500）和相近股票的特征共同作为模型输入，精调模型从而提高其性能。新闻文章在不同股票中的分布为偏态分布，因此新闻少的股票训练样本少，文献[46]基于 3 个原则，即来源和目标股票的历史价格时间序列高度相关、来源和目标股票在相同工业分类部门和源数据具有最优预测性能，将源数据与目标财经新闻映射到同一个情感特征空间中，将源股票特征转移到目标股票，最后采用多数投票机制对生成的候选股票进行排名。

随着时间的推移，时效性要求很高的原有金融标签数据可能不可用，如利用上月份的训练样本学习得到的模型并不能较好预测本月的新样本。不同于传统机器学习，迁移学习方法不要求训练数据与测试数据作同分布假设，并且可避免对获得的金融数据重新标注标签而引起的人力和物力耗费。对于源数据，必须有足够可利用的训练样本才能学习得到一个好的分类模型。

（2）集成学习算法

集成学习算法并不是指某种具体算法，而是使用多种学习算法来获得比使用单一算法更好的性能。随着训练样本容量的增长，文献[47]对 bootstrap 样本训练的二进制预测器进行多数投票，以构造 bagging，实验结果表明，当训练样本规模较小且预测器不稳定时，bagging 是有效的。自适应增强（Adaptive Boosting，AdaBoost）也称增强学习或提升法，是一种重要的集成学习，当一个基本分类器被错误分类时，该样本权值增大，错误分类样本权值则减小，作为下一个次分类器训练的基础。基于综合少数过采样技术（Synthetic Minority Oversampling Technique，SMOTE）和时间加权支持向量机集成算法（Adaboost-SVM-tw），文献[48]提出了不平衡动态财务困境预测方法。该方法使用 SMOTE 处理类平衡，将 SMOTE 嵌入到 adaboost-SVM-tw 的迭代中，设计了一种样本加权机制，在每一轮迭代中，不仅通过时间加权对大多数财务样本进行重采样，而且合成少数样本，使训练数据集再平衡。

3.7　小结

对比以上基于特征处理的机器学习算法可以发现：

（1）金融时间序列一般是非平衡的，存在大量噪声。为了减少数据噪声和不确定性，通常手工提取金融特征，因此需要对金融市场有着较为深入的理解。从技术面、基本面和信息面提取的数据特征产生的结果不尽相同，限制了模型效果的提升。此外，外部的现实限制性因素（如交易成本、流动性、市场情绪、投资者心理等）可能并不能通过模型进行有效表示。

（2）采用多元分析方法来解决将宏观经济变量（如汇率、利率、货币供应量等）集中在一个模型中而导致的数据多重共线性问题，通过矩阵分解与维度约简进行预处理（PCA 和 ICA）的效果较好。

（3）多因子资产定价存在的问题。因子的过度重复使用导致因子失效。不同的金融数据选择的因子不同，导致不同资产组合对因子模型的选择也不同，进而影响算法结果。

（4）在传统的机器学习预测方法中，当市场出现牛熊市转换时，使用固定模型拟合数据的算法可能不再适用。

（5）早期的机器学习资产定价所使用的常见变量为开盘价、成交量等，随着研究的深入，许多“潜在”的金融组合特征被开发出来，提升了预测的效果。这种提取更深层次特征的方法也是深度学习的主要思路。

4　端到端处理深度学习方法

端到端的方法能够从大量原始数据中提取特征，无须专业人员深刻复杂的先验知识，其将复杂的深度特征分解为逐级嵌套简单特征表示，对不同来源的数据进行矩阵式表示，在将其向量融合后统一作为输入，增加了数据来源的多样性。深度学习具有强大的复杂特征提取能力和非线性函数拟合能力，部分方法取消了复杂的内部逻辑设计。随着金融大数据的发展，端到端的方法适应了金融大数据与高频金融数据分析的需要[49]。接下来将介绍的端到端机器学习方法包括 CNN、LSTM、强化学习、文本分析、知识图谱和模型融合。

4.1　CNN

卷积神经网络是一种前馈神经网络，其主要结构由输入层、卷积层、池化层、全连接层和输出层组成。卷积层使用一个卷积核，对输入的特征图进行卷积操作，此操作可以增强原输入中的某些特征。卷积神经网络应用包括两个步骤：1）将金融时间序列进行标准对齐等步骤，转化为可被 CNN 识别的向量或向量集数据；2）通过多层卷积操作深度抽取数据的潜在特征。其主要可用于时间序列价格预测模型、市场趋势分类和投资组合优化等问题。文献[50]提出了一种利用 CNN 来预测 ETF 价格预测股票价格变动方法，通过每天在受限窗口时间段内生成图像快照，提取常用的趋势指标、势头指标以及一些常用的基本面分析指标作为输入特征；在数据集方面，使用了多个 ETF 来增加数据集规模，减小了包含的信息方差。为了考察不同市场之间的相关性，文献[51]提出

了专用 CNN 框架，该框架可应用于多种来源的数据集合，主要有 S&P 500、纳斯达克、道琼斯 NYSE、道琼斯 DJI 和罗素指数，以此聚集成三维张量表，每个预测模型都可以使用三维张量表中所有的信息作为输入，预测某一具体市场的未来走势。文献[52]用相同数据的不同表示形式，生成股票时间序列和股票图表图像（包括价格 K 线图、最高价格和最低价格折线图、交易量数据的柱状图），融合 LSTM 与 CNN 模型以预测股票价格，并且创建了不同的模型表现形式来适应变化的数据。从以上文献可以发现，基于 CNN 算法的预测效果一般，但能多角度融合多种来源的数据，有效地稳定模型，提高算法的鲁棒性。但是，对于交易型金融时序数据，由于不存在类似图像像素之间的空间联系，因此 CNN 与循环神经网络（RNN）存在一定的差距。

4.2 LSTM

循环神经网络（RNN）已被有效地用于预测非平稳数据，LSTM 是解决消失梯度的一种特殊递归神经网络。LSTM 的本质上是一种特殊的循环神经网络，其中控制记忆单元包括忘记门（Forget Gate）、输入门（Input Gate）和输出门（Output Gate），每一个门单元由一个 Sigmoid 神经网络层和一个点乘法运算组成。LSTM 解决了 RNN 因逻辑单元距离增加而出现的长期性依赖问题，被广泛地应用于金融时间序列并取得了显著成果。文献[53]提出在线 LSTM 模型，用于处理非平稳高频股票市场数据，以大量的技术分析指标作为网络输入，按照每时段数据对模型贡献的重要性分别进行加权。在线 LSTM 模型包括两组 LSTM，第一层能够处理一般特征，第二层则针对特定特征。文献[54]提出利用 LSTM 模型来学习股票月收盘价，介绍了 5 种不同投资组合构建策略，包括对 smart-beta 策略的修改，将每支股票的月度历史数据（开、高、低、收盘价和成交量）作为模型输入，预测每月的收盘价，最终输出组合中股票的不同权重。

与 CNN 相比，LSTM 可以发现长、中、短不同周期的数据规律，快速捕捉新变化的市场规律，提升了模型效果。提取各类不同来源的金融交易数据指标与金融市场数据指标是一件复杂的事情，并且提取哪些指标始终未能取得共识。CNN 与 LSTM 则跳过了手工特征提取过程，其模型结果以一种端到端的方式呈现。CNN 与 LSTM 也存在一定的缺陷：1）暗箱操作与不可理解性；2）两种方法是基于过去的时间序列来表明历史发生的规律，即假定时间序列分布、数据分布模式是重复的，然而随着时间向前推移，模型的适用性程度不断下降，导致错误率上升，在实际交易过程中会产生新的风险。基于此类情况，基于强化学习的资产定价方法随之被提出。

4.3 强化学习

强化学习是机器学习中的子领域，其强调如何基于环境来行动，从而取得最大化预期利益，即如何在环境给予的奖励或惩罚的刺激下执行相应动作，产生最大化预期收益。强化学习是一种从状态到动作的映射，智能体在与环境的交互中学习策略，取得了最大回报收益。区别于监督和非监督式学习，强化学习由代理、观察、动作、激励和环境 5 部分组成，是一种广义的马尔可夫决策过程。深度强化学习通过计算长期马尔可夫决策的激励的折现值，使用深度神经网络前馈计算来解决状态空间过大的问题。深度强化学习融合了深度学习和强化学习，集成了深度学习在图形、语音等问题上的感知能力和强化学习的决策能力，扩大了策略函数的应用范围，能够直接从高维数据中学习特征并实时调整策略，将预测价格与投资动作相结合，直接以投资收益目标为优化目标，避免了因 CNN 与 LSTM 的高预测率而不能取得高收益的困境。文献[55]主要有两个目的：1）如果实际表现在某种程度上是异质的，那么这些认知心理参数对代理人的表现是否重要？2）探索个人心理特征对交易行为的影响。从实证结果来看，受试者的收入表现与他们使用限制订单交易之间存在着普遍的正相关关系。因此，通过估计强化学习模型，重点研究了测试者使用极限订单或市场订单方面的学习行为，结果显示了参考者之间的巨大异质性。该异质性反映了受试者人格特质的多样性，反过来又影响了受试者的收入表现。文献[56]讨论了股权溢价之谜和改善家庭财务结果的关系，以及个人 401(k) 储蓄率与 401(k)收益的高平均回报与低方差回报之间的关系。实证表明，在做储蓄决定时，个人投资者遵循一种朴素的强化学习法则，会过度从他们的个人回报经验中作推断。文献[57]研究了存在交易成本时，如何利用强化学习来推导衍生品最优对冲策略。当目标是最小化对冲成本和标准差时，使用两个不同的 Q 函数来获得不同状态与行为组合下的成本期望值；另外引入学习算法，扩大了可以使用的目标函数范围。文献[58]开发了深度神经网络增强递归模型，以估计交易的短期和长期影响，评价价格对股票收益的影响。在特定市场条件下采取交易行动后，临时价格影响为每个时间点预期影响的总和，证实了交易价格、交易量以及市场状况的关系。将股票交易时的跟单行为分为 8 类，预测每类行为发生的概率，描述了交易行为及其引起的行为之间的非线性关系。文献[59]将股票图像作为输入来构建数据矩阵，上下部分分别代表收盘价和成交量，使用当前股价值与前一日股票的变化比例来决定激励值大小，预测相近的股票价格在全球股票市场的变动。

强化学习可分为基于模型（Model Based）的强化学习与无模型（Model Free）的强化学习两类，与深度学习的结合使得强化学习变成一个数据驱动的自动动态决策问题。一般而言，深度强化学习算法的决策逻辑与人类决策逻辑相似，代表了自动化交易最有可能的发展方向，对基于深度强化学习资产定价的研究也越来越多。

4.4 文本分析

计算机技术的发展扩大了数据来源。多来源文本数据具有信息冗余高、信息密度低、数据量大和时频高等特点，可用于监测不同市场参与主体的情绪、舆情分析和投资者意见分歧等。文献[60]将文档表示为密集向量，使用事件嵌入神经张量网络进行训练，使用深度卷积神经网络对输入事件序列进行语义组合，从而预测股价变动。文献[61]解释了股票收益与市场误判之间的关系，这种关系由投资者的情绪导致。提取新浪微博上的宏观资讯来衡量投资者的宏观经济情绪，实证表明，不同类别情绪（包括愤怒、厌恶、恐惧、喜悦和悲伤）对上证综合指数影响显著；能否快速传递金融信息是影响金融市场稳定的重要因素之一，在某种程度上是资产交易定价

的核心竞争力。文献[62]表明，大众传媒可以有效缓解信息摩擦，即使传媒未能提供有价值的新闻，资产定价也能受其影响。通过研究媒体报道和股票预期收益之间的横截面关系可知，即使控制了已知的风险因素，没有媒体报道的股票也能比有媒体报道的股票获得更高的回报。

4.5　知识图谱

知识图谱是语义网络(Semantic Network)知识库，是包含多种类型节点和多种类型边的多关系图。其本质上是一种大型的语义网络，以实体概念为节点，以关系为边，从关系的角度抽象世界，更适合人类思维逻辑过程。知识图谱三元组由实体、属性和关系组成，知识图谱由知识抽取、知识融合和知识推理等部分组成。证券市场中同一板块中的股票数据有时会呈现相似走势，不同的市场之间存在某种关联关系，很多开源数据库(如 FreeBase 和 Wikidata)都提供了知识图谱数据，近年来各种知识图谱为资产定价研究提供了新的思路。传统的分析(如股权融资、公司合作、供应商供需关系等)忽略了公司之间的业务关系。文献[63]提出了一种利用企业知识图谱嵌入的方法来计算上述因子对股票的影响程度，不仅考虑了新闻中涉及的公司，也考虑了与之相关的公司，为每种股票构建了情绪向量，并在此基础上将焦点股票的新闻情绪向量和焦点股票量化特征结合起来，共同预测股票走势。文献[64]将目标公司的关联公司信息纳入其股价预测，根据真实市场的投资事实，构造公司投资知识图谱，并通过图节点嵌入方法学习每个公司的投资分布概率，通过卷积操作关联企业信息。整合相关企业信息的预测模型能够对股市做出更准确的预测。文献[65]以知识图谱为基础，通过多通道连接，将价格向量和事件嵌入作为预测模型的输入，展示事件之间的因果联系，指出由知识驱动的事件是价格反转的常见原因。

当前知识图谱的发展还不够成熟，主要用于描述企业股权关系、投资关系、合作与供应、客户关系等。其在资产定价应用领域以辅助决策为主，即形成各类金融知识库，然后对金融知识库使用知识关联、知识检索和知识推理等，提供了知识服务支撑。

4.6　模型融合

为了充分利用各模型的优势，常使用模型融合来研究资产定价问题。文献[66]融合统计方法和模式识别方法，将隐马尔可夫模型与决策树相结合，基于历史收盘价、股息和收益，利用决策树来预测孟买证券交易所敏感指数。文献[67]融合指数平滑模型(ESM)、自回归综合移动平均模型(ARIMA)和反向传播神经网络(BPNN)模型，不同模块之间的权重值由遗传算法确定。文献[68]采用玻尔兹曼约束机作为潜在特征提取器，将支持向量机作为分类器，并使用巴西证券市场 5 种资产的真实数据。文献[69]从中国社交网络新浪微博上选择内容，通过提取情感特征和潜在的 LDA 特征，使用 RNN 与 Adaboost 的混合预测模型来预测中国股票市场的波动率。文献[70]提出了一种基于不同时间跨度的双重特征提取方法，以获得更多市场数据特征；采用 PLR 和 CNN 从市场数据中提取长期时间特征和短期空间特征，使用基于双注意力机制的编码-解码框架来区别特征权重，预测股票价格趋势。

从以上研究可以看出，模型融合主要有以下几种形式。1)将前期抽取的数据特征向量与后期分类算法融合。2)融合多个不同模型以提高泛化能力，如基于 PCA 的支持向量回归(SVR)及遗传演算法(GAs)、LSTM 与 CNN 的融合等。3)启发式算法与分类算法相融合。启发式算法求解分类模型参数的最优解，如遗传算法优化 SVR 参数，或遗传算法先处理离散特征，然后确定神经网络连接权值。4)对来源不同的数据集进行融合，将每个不同来源的数据集构建为一个分类模型，最终合并成一个大的分类器。

4.7　小结

对比以上端到端处理深度学习方法可以发现，从已有文献总结来看，应用机器学习资产定价方法的研究人员主要是计算机领域的研究人员，这一点在深度学习资产定价领域更加明显。由于数据来源不一致，评价标准不同，公认性机器学习资产定价的文章较少。相对而言，对 PCA、SVM、贝叶斯和决策树的研究比较充分，对其他模型的研究则相对较少。端到端算法研究虽然较少，但呈现逐年增加的趋势。

5　未来展望

机器学习资产定价方法与现代信息技术融合具有较大的应用潜力，其未来可能的研究方向如下。

(1)拓展性。相比传统的分析研究方法，近年来在图像、声音和自然语言处理方面取得显著进展的 CNN 和 LSTM 等算法在资产定价方面使用得较少，这可能是这些深度学习模型本身的复杂性导致抽取了过多的深层次特征，这些深层次特征所具有的弱性因子暴露特性与研究发掘金融市场本质规律不一致，未来可进行进一步的研究。

(2)数据有限性。随着短样本和非结构化数据的增加，金融数据时间跨度相对较短。例如，社交媒体数据最多只有十余年的数据可供使用；比较完备的美国股市数据，即使从 1970 年算起，大概也只有 600 个月度数据。这些数据对于机器学习的大规模训练的作用十分有限，会影响模型的准确性；而非传统另类数据的作用则越来越明显，迁移算法就是一种可能性方案。现有的许多机器学习模型没有考虑现实交易的限定因素，如资产流动性、交易成本、交易摩擦和法律限制等，如何将研究成果应用到具体的商业环境中，以辅助或替代人力，是未来的重要研究方向。

(3)可解释性。机器学习方法面临准确率和模型复杂度之间的权衡，模型越复杂，金融规律解释难度越高。线性回归可解释性强，但无法处理复杂的数据关系，而深度神经网络可以处理变量之间的深层关系，但模型容易成为黑箱，只能使用准确率等来替代传统资产定价模型评估标准。投资者希望对投资逻辑有清晰的解释，以处理不断变化的市场状况。考虑加强事实认知推理的解释作用，采用知识计算方法，将全流程资产定价问题转换为多重函数组合而成的知识图谱计算问题，使得推理路径和推理逻辑可以为人类所掌握，这是提高资产定价模型效果的重要研究方向。

(4)跨学科融合性。对于衍生品类资产定价，如果采用传统的金融模型，由于生成模型经过多层嵌套，容易导致传统模型过于复杂，反而不易刻画变量与资产价格之间的关系；使用

机器学习则可以较为轻松地拟合，但同时增加了理解金融规律的难度。因此需要更多地从交易市场自身规律与交易者的博弈出发，行为金融和机器学习技术相结合的混合方法可能会被证明更有效。

(5)泛化性。机器学习资产定价需要解决如何有效地从数据中提取出需要的特征并形成样本外拟合能力(即模型的泛化能力)的问题，但市场动态性强，同一股票概率分布前后并不一致。此外，随着资讯越来越便捷，迅速处理交易者博弈，更快纠正价格，对金融领域的模型自适应能力有更高的要求。

结束语 本文通过梳理机器学习在资产定价领域的相关文献，总结了机器学习资产定价方法的应用现状、发展趋势和存在的问题，包括常用算法、常用框架以及不同算法的优势与缺点，较为全面地了解了该领域的发展现状及发展趋势，展望了未来可能的研究方向。总体而言，机器学习资产定价方法从最开始的手工提取特征，依赖假定模型建立并求解模型参数，逐步转向端到端的处理，增加了数据来源的多样性，特别是随着近些年深度强化学习的发展，模型可解释性得到了提高。利用模型融合优势，通过智能体直接与环境的交易，来展示推理路径与推理逻辑，成为了下一步研究的重点。

参 考 文 献

[1] LO A W, MACKINLAY A C. Stock market prices do not follow random walks: Evidence from a simple specification test[J]. The Review of Financial Studies, 1988, 1(1): 41-66.

[2] SHAH D, ISAH H, ZULKERNINE F. Stock market analysis: A review and taxonomy of prediction techniques[J/OL]. International Journal of Financial Studies, 2019, 7(2): 26. https://www.mdpi.com/2227-7072/7/2/26.

[3] HARVEY C R, LIU Y. A census of the factor zoo[J/OL]. SSRN, 2019, 3341728. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3341728.

[4] RUNDO F, TRENTA F, DI STALLO A L, et al. Machine learning for quantitative finance applications: A survey[J/OL]. Applied Sciences, 2019, 9(24): 5574. https://www.mdpi.com/2076-3417/9/24/5574.

[5] JURCZENKO, EMMANUEL E D. Machine Learning for Asset Management: New Developments and Financial Applications [M]. John Wiley & Sons, 2020.

[6] CHONG E, HAN C, PARK F C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies[J]. Expert Systems with Applications, 2017, 83: 187-205.

[7] NTI I K, ADEKOYA A F, WEYORI B A. A systematic review of fundamental and technical analysis of stock market predictions[J/OL]. Artificial Intelligence Review, 2019: 1-51. https://linkspringer.53yu.com/article/10.1007/s10462-019-09754-z.

[8] JIANG W. Applications of deep learning in stock market prediction: recent progress[J]. arXiv: 2003.01859, 2020.

[9] DIXON M F, HALPERIN I. The four horsemen of machine learning in finance[J/OL]. SSRN, 2019, 3453564. https://papers.ssrn.com/sol3/papers.cfm?.

[10] ZHENG A, CASARI A. Feature engineering for machine learning: principles and techniques for data scientists[M/OL]. O'Reilly Media, Inc., 2018.

[11] LETTAU M, PELGER M. Factors that fit the time series and cross-section of stock returns[J]. The Review of Financial Studies, 2020, 33(5): 2274-2325.

[12] KELLY B T, PRUITT S, SU Y. Characteristics are covariances: A unified model of risk and return[J]. Journal of Financial Economics, 2019, 134(3): 501-524.

[13] LETTAU M, PELGER M. Estimating latent asset-pricing factors[J]. Journal of Econometrics, 2020, 218(1): 1-31.

[14] ONATSKI A. Asymptotics of the principal components estimator of large factor models with weakly influential factors[J]. Journal of Econometrics, 2012, 168(2): 244-258.

[15] AIT-SAHALIA Y, XIU D. Using principal component analysis to estimate a high dimensional factor model with high-frequency data[J]. Journal of Econometrics, 2017, 201(2): 384-399.

[16] AVELLANEDA M, HEALY B, PAPANICOLAOU A, et al. PCA for Implied Volatility Surfaces[J]. The Journal of Financial Data Science, 2020, 2(2): 85-109.

[17] CARAIANI P. The predictive power of singular value decomposition entropy for stock market dynamics[J]. Physica A: Statistical Mechanics and its Applications, 2014, 393: 571-578.

[18] GU R, SHAO Y. How long the singular value decomposed entropy predicts the stock market?—Evidence from the Dow Jones Industrial Average Index[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 453: 150-161.

[19] WANG D. Adjustable robust singular value decomposition: Design, analysis and application to finance[J]. Data, 2017, 2(3): 29.

[20] BACK A D, WEIGEND A S. A first application of independent component analysis to extracting structure from stock returns [J]. International Journal of Neural Systems, 1997, 8(4): 473-484.

[21] BARILLAS F, SHANKEN J. Comparing asset pricing models [J]. The Journal of Finance, 2018, 73(2): 715-754.

[22] FULOP A, YU J. Bayesian analysis of bubbles in asset prices[J/OL]. Econometrics, 2017, 5(4): 47. https://www.mdpi.com/2225-1146/5/4/47.

[23] TURNER J A. Momentum Portfolios and the Capital Asset Pricing Model: A Bayesian Approach[J/OL]. Quarterly Journal of Finance and Accounting, 2010: 43-59. https://www.jstor.org/stable/23074629.

[24] SCHORFHEIDE F, SONG D, YARON A. Identifying long-run risks: A Bayesian mixed frequency approach[J]. Econometrica, 2018, 86(2): 617-654.

[25] BUSSE J A, IRVINE P J. Bayesian alphas and mutual fund persistence[J]. The Journal of Finance, 2006, 61(5): 2251-2288.

[26] GEWEKE J, AMISANO G. Hierarchical Markov normal mixture models with applications to financial asset returns[J]. Journal of Applied Econometrics, 2011, 26(1): 1-29.

[27] PSARADAKIS Z，SOLA M，SPAGNOLO F. On Markov error correction models，with an application to stock prices and dividends[J]. Journal of Applied Econometrics，2004，19(1)：69-88.

[28] GU S，KELLY B，XIU D. Autoencoder asset pricing models[J/OL]. Journal of Econometrics，2020. https://sciencedirect.53yu.com/science/article/pii/S0304407620301998.

[29] SUIMON Y，SAKAJI H，IZUMI K，et al. Autoencoder-Based Three-Factor Model for the Yield Curve of Japanese Government Bonds and a Trading Strategy[J]. Journal of Risk and Financial Management，2020，13(4)：82.

[30] LV S，HOU Y，ZHOU H. Financial Market Directional Forecasting With Stacked Denoising Autoencoder[J]. arXiv：1912.00712，2019.

[31] HUANG C F. A hybrid stock selection model using genetic algorithms and support vector regression[J]. Applied Soft Computing，2012，12(2)：807-818.

[32] LEE M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction[J]. Expert Systems with Applications，2009，36(8)：10896-10904.

[33] KARATHANASOPOULOS A，THEOFILATOS K A，SERMPINIS G，et al. Stock market predictionusing evolutionary support vector machines：an application to the ASE20 index[J]. The European Journal of Finance，2016，22(12)：1145-1163.

[34] ABEDIN M Z，GUOTAI C，MOULA F E，et al. Topological applications of multilayer perceptrons and support vector machines in financial decision support systems[J]. International Journal of Finance & Economics，2019，24(1)：474-507.

[35] HUANG W，NAKAMORI Y，WANG S Y. Forecasting stock market movement direction with support vector machine[J]. Computers & Operations Research，2005，32(10)：2513-2522.

[36] CHEN W H，SHIH J Y，WU S. Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets[J]. International Journal of Electronic Finance，2006，1(1)：49-67.

[37] MORITZ B，ZIMMERMANN T. Tree-based conditional portfolio sorts：The relation between past and future stock returns[J/OL]. SSRN，2016，2740751. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740751.

[38] NTI K O，ADEKOYA A，WEYORI B. Random forest based feature selection of macroeconomic variables for stock market prediction[J/OL]. American Journal of Applied Sciences，2019，16(7)：200-212.

[39] KRAUSS C，DO X A，HUCK N. Deep neural networks，gradient-boosted trees，random forests：Statistical arbitrage on the S&P 500[J]. European Journal of Operational Research，2017，259(2)：689-702.

[40] THAKKAR A，CHAUDHARI K. A comprehensive survey on portfolio optimization，stock price and trend prediction using particle swarm optimization[J/OL]. Archives of Computational Methods in Engineering，2020：1-32. https://linkspringer.53yu.com/.

[41] RATHER A M，AGARWAL A，SASTRY V N. Recurrent neural network and a hybrid model for prediction of stock returns[J]. Expert Systems with Applications，2015，42(6)：3234-3241.

[42] PARRACHO P，NEVES R，HORTA N. Trading in financial markets using pattern recognition optimized by genetic algorithms[C]// Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation. 2010：2105-2106.

[43] MERELLO S，RATTO A P，ONETO L，et al. Ensemble Application of Transfer Learning and Sample Weighting for Stock Market Prediction[C]// 2019 International Joint Conference on Neural Networks(IJCNN). 2019：1-8.

[44] NAM K H，SEONG N Y. Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market[J]. Decision Support Systems，2019，117：100-112.

[45] NGUYEN T T，YOON S. A novel approach to short-term stock price movement prediction using transfer learning[J]. Applied Sciences，2019，9(22)：4745.

[46] LI X，XIE H，LAU R Y K，et al. Stock prediction via sentimental transfer learning[J]. IEEE Access，2018，6：73110-73118.

[47] LEE T H，YANG Y. Bagging binary and quantile predictors for time series[J]. Journal ofEconometrics，2006，135(1/2)：465-497.

[48] SUN J，LI H，FUJITA H，et al. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting[J]. Information Fusion，2020，54：128-144.

[49] RUNDO F，TRENTA F，DI STALLO A L，et al. Machine learning for quantitative finance applications：A survey[J]. Applied Sciences，2019，9(24)：5574.

[50] GUDELEK M U，BOLUK S A，OZBAYOGLU A M. A deep learning based stock trading model with 2-D CNN trend detection[C]// 2017 IEEE Symposium Series on Computational Intelligence. 2017：1-8.

[51] HOSEINZADE E，HARATIZADEH S. CNNPred：CNN-based stock market prediction using several data sources[J]. arXiv：1810.08923，2018.

[52] KIM T，KIM H Y. Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data[J/OL]. PloS one，2019，14(2)：e0212320. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0212320.

[53] BOROVKOVA S，TSIAMAS I. An ensemble of LSTM neural networks for high-frequency stock market classification[J]. Journal of Forecasting，2019，38(6)：600-619.

[54] YILDIZ Z C，YILDIZ S B. A portfolio construction framework using LSTM-based stock markets forecasting[J/OL]. International Journal of Finance & Economics，2020. https://onlinelibrary.wiley.com/doi/abs/10.1002/ijfe.2277.

[55] CHEN S H，HSIEH Y L. Reinforcement learning in experimental asset markets[J]. Eastern Economic Journal，2011，37(1)：109-133.

[56] CHOI J J，LAIBSON D，MADRIAN B C，et al. Reinforcement

learning and savings behavior[J]. The Journal ofFinance, 2009, 64(6): 2515-2534.

[57] CAO J, CHEN J, HULL J, et al. Deep hedging of derivatives using reinforcement learning[J]. arXiv: 2013. 16409, 2021.

[58] CAO Y, ZHAI J. Estimating price impact via deep reinforcement learning[J/OL]. International Journal of Finance & Economics, 2020. https://onlinelibrary. wiley. com/doi/abs/10. 1002/ijfe. 2353.

[59] LEE J, KIM R, KOH Y, et al. Global stock market prediction based on stock chart images using deep Q-network[J]. IEEE Access, 2019, 7: 167260-167277.

[60] DING X, ZHANG Y, LIU T, et al. Deep learning for event-driven stock prediction[C]// Twenty-fourth International Joint Conference on Artificial Intelligence. 2015.

[61] XU Y, ZHAO J. Can sentiments on macroeconomic news explain stock returns? Evidence form social network data[J/OL]. International Journal of Finance & Economics, 2020. https://onlinelibrary. wiley. com/doi/abs/10. 1002/ijfe. 2260.

[62] FANG L, PERESS J. Media coverage and the cross-section of stock returns[J]. The Journal of Finance, 2009, 64(5): 2023-2052.

[63] LIU J, LU Z, DU W. Combining enterprise knowledge graph and news sentiment analysis for stock price prediction[C]// Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019.

[64] CHEN Y, WEI Z, HUANG X. Incorporating corporation relationship via graph convolutional neural networks for stock price prediction[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 1655-1658.

[65] DENG S, ZHANG N, ZHANG W, et al. Knowledge-driven stock trend prediction and explanation via temporal convolutional network[C]// Companion Proceedings of The 2019 World Wide Web Conference. 2019: 678-685.

[66] TIWARI S, PANDIT R, RICHHARIYA V. Predicting future trends in stock market by decision tree rough-set based hybrid system with HHMM[J/OL]. Iternational Journal of Electronics and Computer Science Engineering, 2010, 1(3). http://citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 261. 3105&rep=rep1&type=pdf.

[67] CREIGHTON J, ZULKERNINE F H. Towards building a hybrid model for predicting stock indexes[C]// 2017 IEEE International Conference on Big Data(Big Data). 2017: 4128-4133.

[68] ASSIS C A S, PEREIRA A C M, CARRANO E G, et al. Restricted Boltzmann machines for the prediction of trends in financial time series[C]// 2018 International Joint Conference on Neural Networks. 2018: 1-8.

[69] 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.

[70] CHEN Y, LIN W, WANG J Z. A dual-attention-based stock price trend prediction model with dual features[J]. IEEE Access, 2019, 7: 148047-148055.

XU Jie, born in 1986, Ph. D. His main research interests include machine learning, asset pricing and quantitative trading.

XING Chun-xiao, born in 1967, Ph.D supervisor. His main research interests include deep learning, big data and knowledge engineering, and fintech.

(责任编辑:柯颖)