

### 中信证券研究部

王兆宇

首席量化策略

分析师

S1010514080008

赵文荣

首席量化与配置

分析师

S1010512070002

张依文

量化策略分析师

S1010517080004

刘方

首席组合配置

分析师

S1010513080004

厉海强

首席金融产品

分析师

S1010512010001

### 核心观点

人工智能的浪潮已波及到投资领域，国内外各类资产管理机构正抓紧布局相关研究。本文探讨了在量化研究中使用机器学习算法的基本原则和注意事项，并整理了六大类常用机器学习模型在投资领域中应用的优势和特点，本文最后还总结了部分在投资领域中使用机器学习算法的场景与案例。

■ **量化型基金成对冲基金主流，人工智能/机器学习方法被广泛使用。**（1）2018年，全球最大的六家对冲基金均为量化型基金。为避免策略同质化，投资者开始探索和使用人工智能/机器学习方法。（2）巴克莱 2018 年的对冲基金调查显示已经有 56% 的投资经理使用 AI/ML 进行投资决策，而 2017 年仅 20%。

■ **各大金融机构在 AI/ML 领域争相布局。**（1）近年来，不少顶级人工智能专家转战证券投资领域，如：微软人工智能首席科学家和卡耐基梅隆大学机器学习系主任分别在 2017 年、2018 年加入了 Citadel 和摩根大通。（2）包括平安集团、华夏基金在内等多家国内资管机构也纷纷布局人工智能技术研究。

■ **近十年 AI 对冲基金表现不俗。**（1）国际对冲基金研究机构 Eurekahedge 针对使用 AI/ML 的基金编制的 AI 对冲基金指数在近十年时间能够实现年化 12.6% 的收益。而同期 Eurekahedge 的旗舰对冲基金指数年化收益为 4.9%。（2）近两年，市场中开始出现运用 AI/ML 进行投资的 ETF，且增长势头迅猛。2017 年有一只 ETF（AIEQ）成立，到 2018 年底就增长到 9 只 ETF。

■ **六种常用的机器学习模型梳理。**（1）神经模型：通过神经元的连接，能够拟合任意函数，实现复杂的映射。（2）图模型：擅长对因果关系和概率关系进行刻画，能对复杂场景建模并且解释性较强。（3）聚类/编码模型：一般用于对数据进行归类和可视化。（4）线性模型：结构简单，易于分析理解，使用度最广。（5）树模型：模拟树的拓扑结构对特征空间进行逐层划分，得到预测结果的模型，解释性强。（6）集成学习模型：综合多个的弱决策模型得到强决策模型。

■ **机器学习的应用逻辑。**相对传统方法而言，机器学习方法具备四大优势：非线性优势、数据化优势、速度优势和复杂度优势。而金融数据的高噪、低维输入、动态性的特点给应用带来了极大的挑战。为此，本文提出可沿两个方向尝试：（1）机器学习算法和专业知识相结合。专业知识的结合不仅可以帮助改进模型，还能从经济金融学的角度来理解模型的行为，得到相互印证。（2）跳出收益率预测的思维定式。收益率预测的问题难度较大，可选择信噪比高的对象来进行学习和预测。

■ **现有的机器学习方法在投资领域的应用。**本文总结了大量文献，按机器学习方法在投资中应用任务进行梳理，主要包括：预测、定价、交易、文本分析和资产组合。

## 目录

海外对冲基金加大人工智能/机器学习（AI/ML）投入 .....	1
量化型基金包揽对冲基金规模榜前六 .....	1
巴克莱对冲基金调查显示过半受访者使用 AI/ML 进行投资决策 .....	1
各大金融机构在 AI/ML 领域争相布局 .....	2
近十年 AI 对冲基金表现不俗 .....	4
机器学习算法概述 .....	6
机器学习算法的应用逻辑 .....	7
为什么要用机器学习算法？ .....	7
金融数据的特点：高噪、低维输入、动态 .....	8
机器学习算法和专业知识相结合 .....	9
跳出收益率预测的思维定式 .....	9
现有的机器学习方法在投资领域的应用场景 .....	10
文献中机器学习方法应用的汇总 .....	10
预测问题 .....	10
资产定价 .....	11
交易执行 .....	13
文本分析 .....	14
组合优化 .....	14
机器学习在量化投资中应用的经典文献整理 .....	15

## 插图目录

图 1：巴克莱 2017 和 2018 年调查：使用 AI/ML 的比例提升 .....	2
图 2：巴克莱 2018 调查中 AI/ML 的使用形式 .....	2
图 3：使用 AI/ML 的时间长度 .....	2
图 4：使用 AI/ML 管理的资产规模 .....	2
图 5：围棋 AI 击败人类顶尖棋手李世石 .....	3
图 6：德州扑克 AI 击败人类顶尖高手 .....	3
图 7：Man AHL 的模型里程碑 .....	4
图 8：Eurekahedge 的旗舰对冲基金指数与 AI 对冲基金指数表现 .....	5
图 9：机器学习算法导图 .....	6
图 10：神经元的连接 .....	7
图 11：肺病诊断的图模型表示 .....	7
图 12：机器学习算法在投资领域应用导图 .....	10
图 13：人工神经网络 .....	11
图 14：k 近邻 .....	11
图 15：支持向量机 .....	11

图 16: 两种状态的隐马尔可夫模型预测波动率 .....	11
图 17: 核方法的原空间 .....	12
图 18: 核方法的映射空间 .....	12
图 19: 自动编码器算法原理 .....	12
图 20: 决策树算法原理 .....	12
图 21: 强化学习原理示意图 .....	13
图 22: 循环神经网络 .....	13
图 23: 基于模式匹配方法的资产组合构建 .....	14

## 表格目录

表 1: 2004 和 2018 年全球对冲基金资管规模排名 .....	1
表 2: 金融机构布局机器学习相关消息梳理 .....	3
表 3: 机器学习顶级学术会议 NeurIPS 的投资相关赞助商 .....	3
表 4: 使用 AI/ML 投资的 ETF .....	5

## ■ 海外对冲基金加大人工智能/机器学习（AI/ML）投入

### 量化型基金包揽对冲基金规模榜前六

截至 2018 年，全球对冲基金资管规模排名显示，量化型基金强势包揽了前六。而在 2004 年，前 9 名都是主动性基金，仅桥水基金占据了第 10 名的位置。海外经验表明，量化型管理方式是一个行业趋势，经过十多年的发展，量化型基金已经成为对冲基金的主流。

对比国内现状，量化型基金还有相当大的开展空间。实际上，现在的量化基金不再局限于高频策略，而是所有的策略都有涉足，包括宏观策略（桥水基金）、股票基本面（AQR）、大众商品、债券等。

表 1：2004 和 2018 年全球对冲基金资管规模排名

2004			2018		
公司	资管规模 (亿美元)	公司分类	公司	资管规模 (亿美元)	公司分类
Caxton Associates	115	主观	Bridgewater Associates	1328	量化
GLG Partners	110	主观	AQR	837	量化
Citigroup Alternative Investments	99	主观	Man Group	591	量化+主观
Farallon Capital Management	99	主观	Renaissance Technologies	570	量化
Citadel Advisors	95	主观为主	Two Sigma	388	量化
Angelo, Gordon & Co.	90	主观	Millennium Mgmt	353	量化
Vega Asset Mgmt	85	主观	Elliott Management	350	主观
Andor Capital Mgmt	83	主观	Marshall Wace	348	量化
Soros Fund Mgmt	83	主观	Davidson Kempner Capital Mgmt	314	主观
Bridgewater Associates	81	量化	Baupost Group	310	主观

资料来源：Pensions&Investments，中信证券研究部

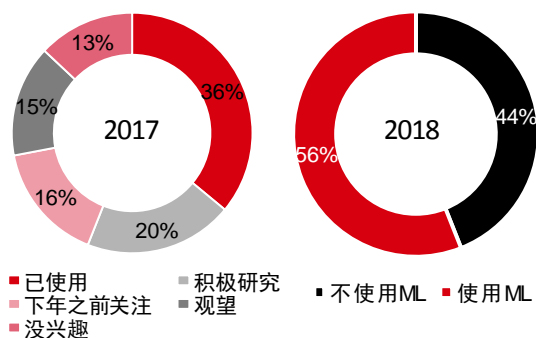
### 巴克莱对冲基金调查显示过半受访者使用 AI/ML 进行投资决策

降低策略的同质性是量化型基金必须要解决的一个重要问题，因此，使用新型方法开发量化策略是所有量化型基金关注的焦点。

2018 年 5 月，巴克莱对冲基金（BarclayHedge）对对冲基金专业人士做了一项关于 AI/ML 使用情况的调查，显示超过一半的对冲基金受访者（56%）使用人工智能进行投资决策。而在一年前的调查中仅 20%的受访者表示使用 AI/ML，一年时间增长一倍多。且一年前的调查显示已使用的比例与积极研究比例共计 36%，远低于 56%，这表明从业者自身都低估了 AI/ML 的扩张速度。

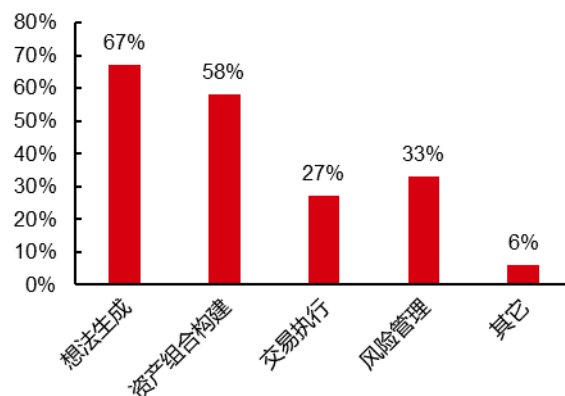
对于如何使用 AI/ML，三分之二的受访者表示主要用于生成交易想法和优化投资组合。另外，使用 AI/ML 进行交易执行和风险管理的受访者均接近三分之一。

图 1：巴克莱 2017 和 2018 年调查：使用 AI/ML 的比例提升



资料来源：巴克莱对冲基金，中信证券研究部

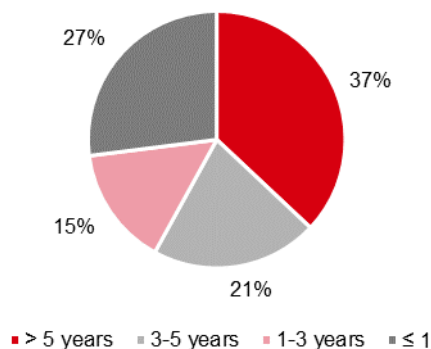
图 2：巴克莱 2018 调查中 AI/ML 的使用形式



资料来源：巴克莱对冲基金，中信证券研究部

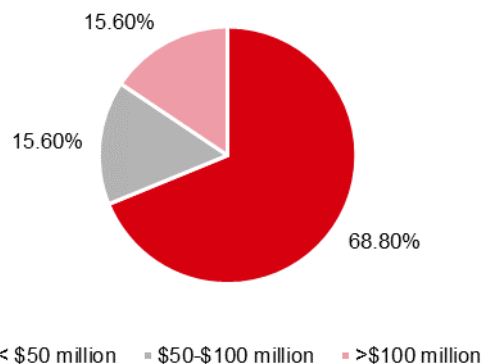
关于使用 AI/ML 的时长调查中，最高比例（37%）的人群已经使用 AI/ML 达五年及以上，超过三年的受访者占比达到 58%。另外，调查还显示使用 AI/ML 管理的资产规模大于 1 亿美金的占比为 15.6%，大部分小于 5,000 万美金，总体上使用 AI/ML 管理的资金规模仍较小。

图 3：使用 AI/ML 的时间长度



资料来源：巴克莱对冲基金，中信证券研究部

图 4：使用 AI/ML 管理的资产规模

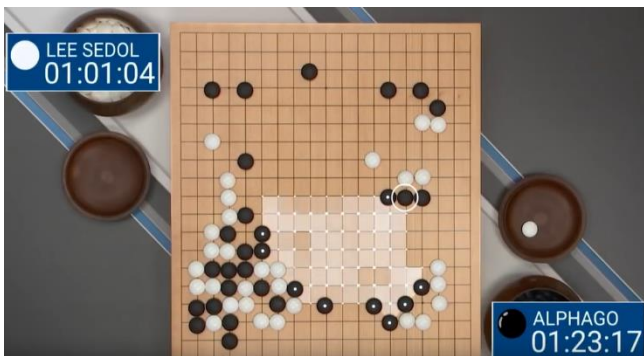


资料来源：巴克莱对冲基金，中信证券研究部

## 各大金融机构在 AI/ML 领域争相布局

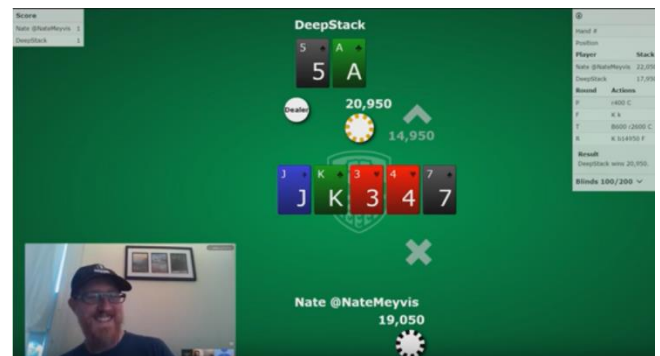
近年来 AI/ML 技术高速发展，技术迭代日新月异，目前 AI/ML 技术不仅在图像处理，自然语言处理，语音信号处理等多个计算机主流领域取得领先地位，还涉足围棋、赌博、游戏等博弈问题，同样取得了惊人的效果。这一成就也带给投资业充足的想象空间。

图 5：围棋 AI 击败人类顶尖棋手李世石



资料来源：youtube.com

图 6：德州扑克 AI 击败人类顶尖高手



资料来源：youtube.com

近年来不少顶级人工智能专家转战证券投资领域。2017 年 5 月，微软人工智能首席科学家、IEEE Fellow 邓力加入美国基金公司 Citadel 担任首席人工智能官。2018 年 5 月，摩根大通宣布聘请卡耐基梅隆大学机器学习系主任 Manuela Veloso 博士担任人工智能研究院负责人。三个月后，华盛顿大学计算机科学教授 Pedro Domingos 宣布加入对冲基金巨头 D. E. Shaw。

国内的公募基金和资管机构也纷纷布局。平安集团可能是行业内战略投入最大手笔的企业之一，一方面平安科技投入数百亿进行人工智能技术的研发；另一方面在 2017 年，平安资管也成立人工智能投资团队，采用智能算法进行投资。嘉实基金、天弘基金均成立了人工智能投资部门，华夏基金与科技巨头微软强强联手，研究智能投资。国寿资产和泰康资产分别成立智能投资部和上线智能投研平台。

表 2：金融机构布局机器学习相关消息梳理

海外	时间	机器学习专家	专家身份	加入公司
	2017 年 5 月	邓力	微软人工智能首席科学家、IEEE Fellow	Citadel
	2018 年 5 月	Manuela Veloso	CMU 机器学习系主任	JP Morgan
	2018 年 8 月	Pedro Domingos	华盛顿大学计算机科学教授	D. E. Shaw
国内	时间	布局		
	2016 年	嘉实基金成立了人工智能投资研究中心		
	2017 年	华夏基金与微软亚洲研究院开展战略合作研究		
	2017 年	国寿资产成立了智能投资部		
	2017 年	泰康资产上线智能投研深度学习分析平台		

资料来源：各公司官网，中信证券研究部

另外，机器学习最顶级的学术会议神经信息处理系统大会（NeurIPS）的赞助商中，越来越多出现了金融机构，特别是知名对冲基金的身影，如 Two sigma, D. E. Shaw, Citadel 等。从 2013 年到 2015 年该会议赞助商中金融机构的数目稳步增长，到 2016 迅速增长近一倍，并在后续几年保持稳定。

表 3：机器学习顶级学术会议 NeurIPS 的投资相关赞助商

年份	数量	对冲基金，资产管理相关公司赞助商
2013	4	Two Sigma, PDT Partners, D. E. Shaw group, DRW Trading

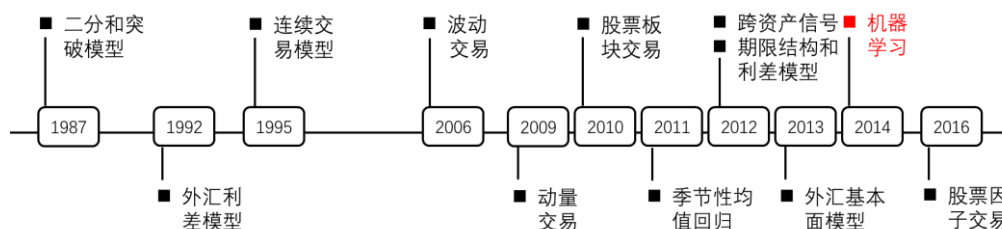


年份	数量	对冲基金，资产管理相关公司赞助商
2014	5	Two sigma, PDT Partners, D. E. Shaw group, Winton, Ketchum Trading
2015	6	Voleon, Man AHL, D. E. Shaw group, Winton, Cubist Systematic Strategies, Two sigma,
2016	11	Winton, Data Collective, Voleon, Two sigma, D. E. Shaw group, Cubist Systematic Strategies, Man AHL, Hutchin Hill, G-Research, Rosetta Analytics, RBC Research,
2017	10	Citadel, Voleon, Winton, D. E. Shaw group, Two sigma, PDT Partners, Rosetta Analytics, Man AHL, Cubist Systematic Strategies, Tudor
2018	11	Citadel, Voleon, Two sigma, Cubist Systematic Strategies, PDT Partners, D. E. Shaw group, Morgan Stanley, Squarepoint, HSBC, Edgestream, GSA Capital

资料来源：NeurIPS 官方网站，中信证券研究部

实际上，很多大型对冲基金早在多年前就将 AI/ML 方法应用到交易过程中，Man AHL 官网上显示其从 2014 年开始运用机器学习模型。Two sigma 的洞察报告（insight）也在持续关注优化，机器学习解释性等方面。

图 7：Man AHL 的模型里程碑



资料来源：Man AHL，中信证券研究部

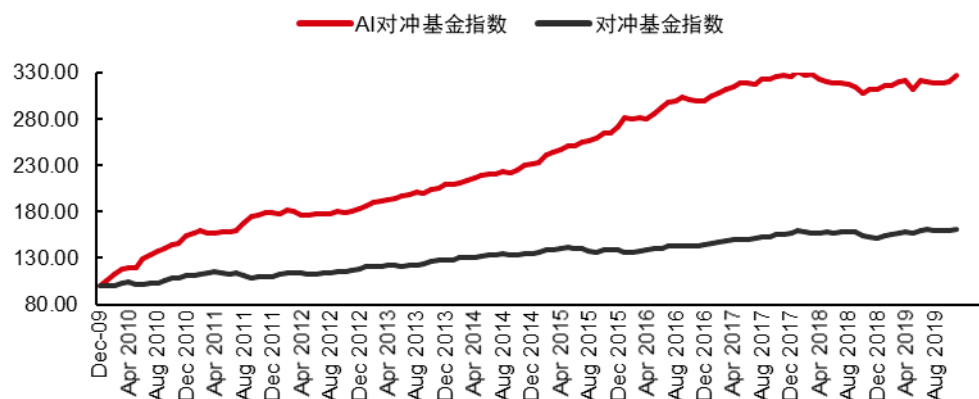
## 近十年 AI 对冲基金表现不俗

国际对冲基金研究机构 Eurekahedge 针对这类使用 AI/ML 的基金编制了一个 AI 对冲基金指数<sup>1</sup>。截止 2019 年 11 月，该指数在近十年时间获取了 227.4% 的收益，年化 12.6%。而同期 Eurekahedge 的旗舰对冲基金指数<sup>2</sup>收益为 61.6%，年化 4.9%。AI 对冲基金指数以稳定的收益大幅跑赢旗舰对冲基金指数。

<sup>1</sup> Bloomberg 代码：EHFI817

<sup>2</sup> Bloomberg 代码：EHFI251

图 8: EurekaHedge 的旗舰对冲基金指数与 AI 对冲基金指数表现



资料来源: EurekaHedge, 中信证券研究部, 截至 2019 年 11 月

近两年, 市场中开始出现运用 AI/ML 进行投资的 ETF, 且势头迅猛。2017 年有一只 ETF (AIEQ) 成立, 到 2018 年底就增长到 9 只 ETF。另外, 十只 ETF 中贝莱德一家就占了半壁江山, 表明其高度看好 AI/ML 技术在投资领域应用的前景。

表 4: 使用 AI/ML 投资的 ETF

基金代码	基金名称	基金发行者	基金规模	成立日期
AIEQ	AI Powered Equity ETF	ETF Managers Group	\$127.72M	2017/10/18
IETC	iShares Evolved U.S. Technology ETF	Blackrock	\$13.63M	2018/3/21
KOIN	Innovation Shares NextGen Protocol ETF	Exchange Traded Concepts	\$10.52M	2018/1/30
IEME	iShares Evolved U.S. Media and Entertainment ETF	Blackrock	\$7.17M	2018/3/21
MSUS	LHA Market State U.S. Tactical ETF	Little Harbor Advisors	\$6.08M	2018/4/4
IEDI	iShares Evolved U.S. Discretionary Spending ETF	Blackrock	\$6.02M	2018/3/21
IEFN	iShares Evolved U.S. Financials ETF	Blackrock	\$4.95M	2018/3/21
IECS	iShares Evolved U.S. Consumer Staples ETF	Blackrock	\$4.13M	2018/3/21
AIQ	AI Powered International Equity ETF	EquBot	\$3.82M	2018/6/5
QRFT	QRAFT AI-Enhanced U.S. Large Cap ETF	Exchange Traded Concepts	\$3.34M	2019/5/21

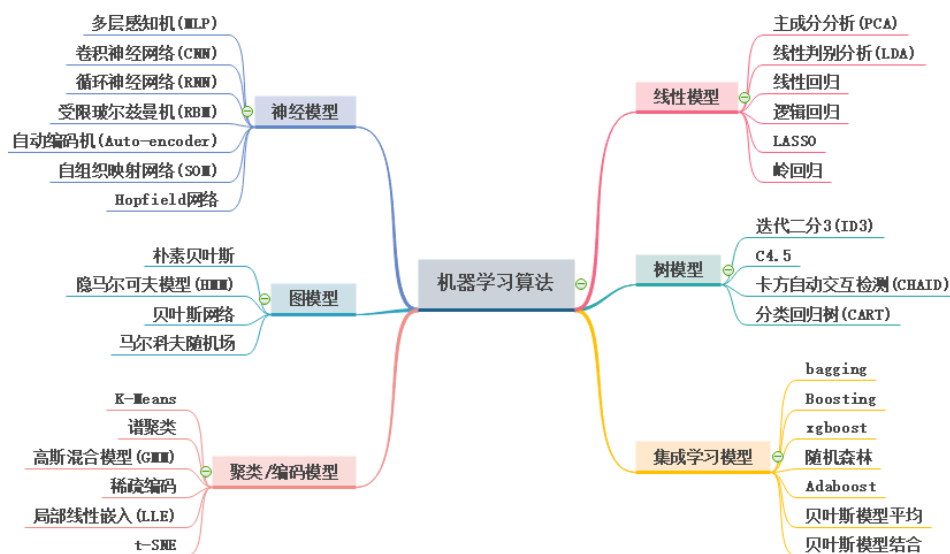
资料来源: www.etf.com, 中信证券研究部



## 机器学习算法概述

机器学习作为一门交叉学科，涉及统计学、数学、计算机科学等多个领域，内容庞杂，分支众多，归纳方式也有多种。比如一般按有无标签信息分为有监督学习和无监督学习；按学习思想又可归纳出对偶学习、迁移学习、对抗学习等。这里我们从模型类型的角度出发，对常用机器学习方法进行一个梳理。

图 9：机器学习算法导图



资料来源：中信证券研究部

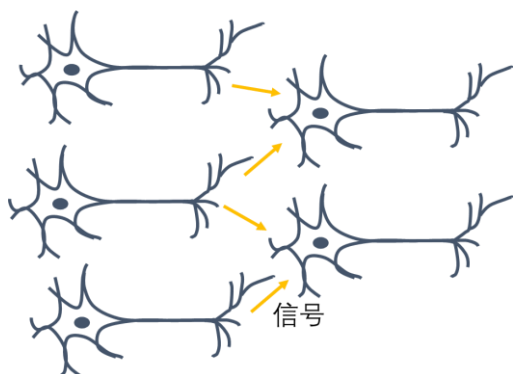
- 神经模型：神经模型起源于对生物神经元的模拟，通过神经元复杂的连接，实现复杂的映射。典型的如多层感知机、卷积神经网络和循环神经网络。后两种神经模型通过独特的连接方式能够更好地提取特定信号的特征，如卷积神经网络提取空间特征，循环神经网络多用于提取时序特征。

特点：神经模型作为典型的数据驱动模型，包含大量参数。优势是能够实现任意复杂的函数，能够对信号进行超出人类设计的表示。缺点是需要大量数据进行学习，且缺乏解释性。尤其是当想要学习的规律并不是一种稳定的模式时，很难学习出有价值的结果。

- 图模型：图模型擅长对因果关系和概率关系进行刻画，通过将某一领域的知识或逻辑转化为节点的概率关系，从而对复杂场景进行描述和推理。图 12 展示了一个简单的肺病诊断的图模型，当了解了患者是否吸烟和检查结果后我们就能计算出每种病的概率，从而得出诊断结论。

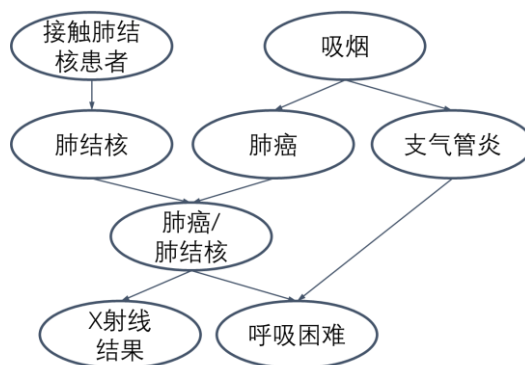
特点：图模型的优点是模型清晰，能够实现专业知识的刻画。缺点是要求训练者必须对该问题具备较深的专业知识，且当知识体系过于复杂时确定模型的拓扑结构的难度也会加大。

图 10：神经元的连接



资料来源：中信证券研究部

图 11：肺病诊断的图模型表示



资料来源：中信证券研究部

- 聚类/编码模型：聚类模型一般用于对数据进行归类和可视化，而编码模型能够将数据表示为典型类别的组合。

特点：工具集丰富，便于我们对信号进行观察和分析，而且不需要额外提供标签信息，处理成本低。缺点是模型的适用对象有限，可能失效。

- 线性模型：线性模型是机器学习中一类最简单、最常用的模型，但所包含的种类也极多，如主成分分析、线性判别分析、回归模型（线性回归，LASSO 和岭回归）和分类模型（逻辑回归）等。

特点：线性模型结构简单，易于分析理解，不易过拟合。缺点是假设较强，复杂的金融信号可能不满足线性假设。

- 树模型：一种应用树的拓扑结构对特征空间进行逐层划分，得到预测结果的模型，可以处理回归和分类任务。

特点：解释性强，非参数型。

- 集成学习模型：集成方法是由多个较弱的模型集成模型组，其中的模型可以单独进行训练，并且它们的预测能以某种方式结合起来去做出一个总体预测。

特点：通用性较强，通过集成学习一般都能超过单一模型的性能。且能够综合基于不同维度信息开发的模型。

通过梳理，我们看到机器学习的模型繁多，且优缺点各异。这意味着我们在尝试应用模型的时候要深入理解模型的内在逻辑，根据其特性设计数据形式和训练方法。

## ■ 机器学习算法的应用逻辑

### 为什么要用机器学习算法？

目前机器学习能受到如此追捧，是因为其充分借用了大数据和算力提升的优势，在很

多任务上实现了意想不到的效果，例如在人脸识别能力上超越人类，在围棋和德州扑克上战胜人类顶尖玩家等等。

但是在投资领域机器学习也能发挥作用吗？对此大家出现了分歧。有人认为机器学习在其它领域对传统技术的颠覆也会移植到投资领域，也许未来将不存在人类投资者了；而有的人认为投资的逻辑复杂，影响因素众多，所使用的数据大部分又体现为无结构化和高噪声的特点，人类都未必能做好对于机器还是太早了。关于机器投资是否会取代人类，我们不予评价，但是机器学习方法的很多特性确实能够弥补传统方法的不足，满足更精确估计和信息提取的需要。这些特性包括但不限于以下几个方面：

- **非线性优势。**在很多定价模型和多因子模型中传统的方法都假设因子与收益之间是线性关系，这点假设过强。机器学习方法能够提供非线性的模型。另外，深度的非线性模型可以提取数据中高层的概念模式，提升模型的描述能力。
- **数据化优势。**随着信息化变革的加深，全社会产生的数据在总量和维度上都已经今非昔比，传统数据已经不能满足专业投资者的需要。投资者开始从新闻，产品网站，卫星数据等另类数据中挖掘有价值的信息并进行数据化。机器学习在这些方面已经有成熟的技术工具可以使用，如自然语言处理技术和图像处理技术。
- **速度优势。**在投资领域处理信息的速度起着至关重要的作用，因为信息的价值会随着时间的衰减。机器学习技术可以快速对海量的数据进行处理来产生交易想法和信号，从而领先其他投资者。
- **复杂度优势。**在中高频的交易领域，传统交易信号指标多是基于量价数据，指标设计也往往根据个人经验，这样极易造成信号指标间的同质化，导致投资收益降低。机器学习技术可以批量的从基础数据中挖掘新的指标，且能保证差异化，最终保证了获利能力。

## 金融数据的特点：高噪、低维输入、动态

在投资领域应用机器学习算法仍然存在着不少问题和挑战。金融信号不同于图像，文本和语音信号，有着以下四点独特的性质。

- **极低的信噪比：**金融数据中含有大量噪声，特别是噪声不仅存在于输入信息，还存在于拟合的目标信号，这增大了对信息的提取难度。当输入信息有噪声，目标信号较纯净时，可以通过增加训练过程的稳健性来克服；然而，目标信号有大量噪声时，直接导致错误信号（error signal）的质量变差，影响训练结果。
- **少量样本和无结构化数据。**很多有价值的数据并不是大数据（big data），很难满足如深度学习等复杂的模型。复杂模型必须与较大的样本空间相匹配，只有少量数据时极易产生过拟合。另外很多有价值的数据是无结构的，如政策信息，新闻信息等，很难用统一的模型来处理。
- **低维的信息输入。**很多的量化策略为了使得收益来源更加清晰且便于开发，只会

使用单一或几个维度的信息。比如趋势策略可能只使用了交易信息，而不使用基本面信息，但市场的波动是交易信息，基本面信息，政策信息等综合的结果。对信息的简化降维可能带来巨大的弊端，即我们期待的是低维信息能够完全拟合潜在的客观规律，但所得结论可能只是严重过拟合的“伪结果”。

- **市场的动态性。**目前大部分使用机器学习算法的场景都假设数据在样本外和样本内有着同样的分布规律，但是实际的市场几乎是不可能的。而我们又需要坚信市场必然存在着一些稳定的规律，且一定存在某种方法能够捕捉这样的规律。

上述的四个问题是在投资领域中成功应用机器学习方法的挑战。但是这些问题也并非机器学习算法所独有，传统的量化投资与主动投资也同样面临。只不过在传统的认知与研究方法下这些问题常常被忽视，而当人类对复杂系统的认知技术有了突破后，这些问题才成为主要矛盾。

## 机器学习算法和专业知识相结合

机器学习最容易陷入的误区是对数据的过度挖掘。例如很多人尝试了决策树模型，发现效果一般，就开始转向支持向量机或者深度学习。然而科学的运用模型其实更像医生治病一样，需要对症下药，并不是越贵的药越好。

如何能够像医生一样明析症结所在，就需要专业知识的指导。很多情况下是由于想要学习的规律不明确，信噪比太低，导致所有的模型都半斤八两。此时，应该审视任务定义的逻辑，而不是换用所谓“更强”的模型。

专业知识的结合不仅可以帮助改进模型，还能从经济金融学的角度来理解模型的行为。如果能够得到相互印证，那么模型输出的可信度便可能大大提升。

## 跳出收益率预测的思维定式

机器学习算法固然潜能巨大，但目前的应用方式较为单一。大多数的应用是定义为涨跌幅预测的回归任务或者涨跌的分类任务，这面临的问题是金融信号的低信噪比。简单模型能滤掉噪声，同时也会滤掉信号，复杂模型则反之。那么，在对于收益率的预测上大多数人还是遵循奥卡姆剃刀原则<sup>3</sup>，即使用简单的线性模型。

但是收益率预测的问题难度较大，此时何不换一种思路，比如上述介绍的在公司基本面和风险方面进行预测就是一种尝试。不同预测对象的信噪比不同，风险和交易成本的预测就相对容易。

除了直接预测外，机器学习方法还可以用来改进传统投资的一些中间过程。比如在使用均值方差模型时需要估计资产收益率和协方差矩阵，但是传统的方法对协方差的估计效果较差，此时可以尝试使用模式匹配算法检索相似的历史时期，采用这些时期进行估计能够大大提升准确度。

---

<sup>3</sup> 奥卡姆剃刀原则：当有两个或多个处于竞争地位的理论能得出同样的结论时，那么简单或可证伪的那个更好，即“如无必要，勿增实体”。

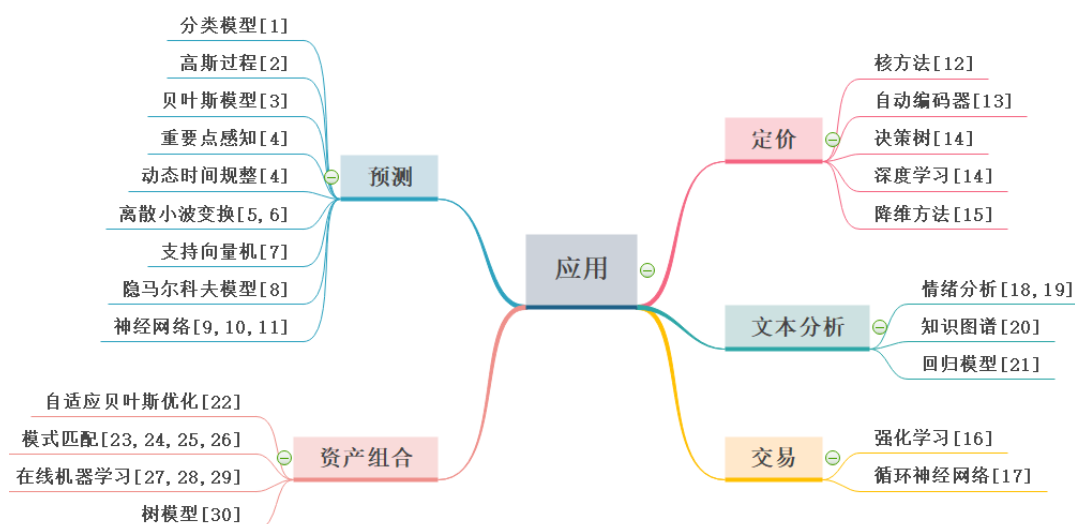
## ■ 现有的机器学习方法在投资领域的应用场景

### 文献中机器学习方法应用的汇总

现有相关研究文献中，机器学习方法在投资中主要被用来解决预测、定价、交易、文本分析和资产组合五个任务。

从下图可以看到，目前的应用采用的算法十分广泛。其中，最多的一类还是预测，除了分类模型外，还有很多其它的技术也可以用来做预测，如动态时间规整、隐马尔科夫模型等。另外，在一些投资过程中的传统任务如定价和交易中，也有利用机器学习方法进行优化，解决复杂映射的问题。随着自然语言处理技术的进步和大数据时代的来临，采用这类算法处理文本信息现在也广受关注。

图 12：机器学习算法在投资领域应用导图



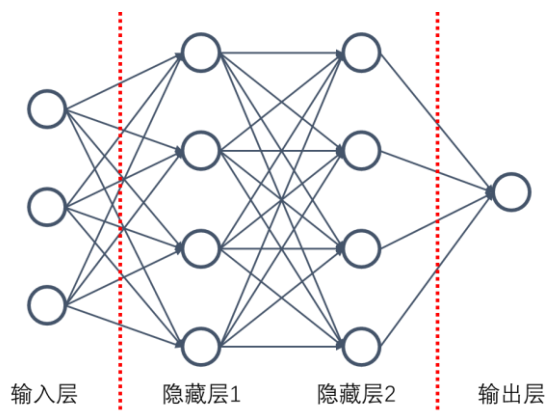
资料来源：中信证券研究部，文献见附录

### 预测问题

最基本的采用机器学习的方法即为对收益率进行预测，如文献[1]中采用了人工神经网络（artificial neural network, ANN）、k 近邻（k-nearest neighbor, kNN）和决策树（decision tree, DT）这三种方法来预测市场的涨跌。人工神经网络和决策树主要用于学习输入因子空间到预测目标的映射，前者通过深层的连接实现，后者通过逐层的因子划分实现。而 k 近邻基于相似的特征具有相似的预测值的假设，通过找到 k 个最相似历史点求取平均得到预测结果。

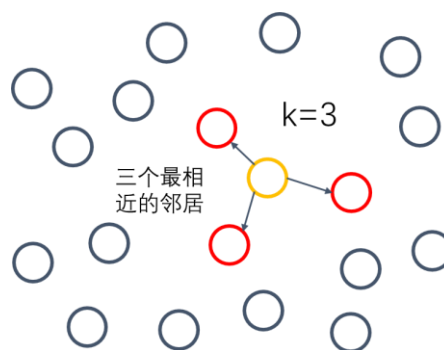


图 13: 人工神经网络



资料来源: 中信证券研究部

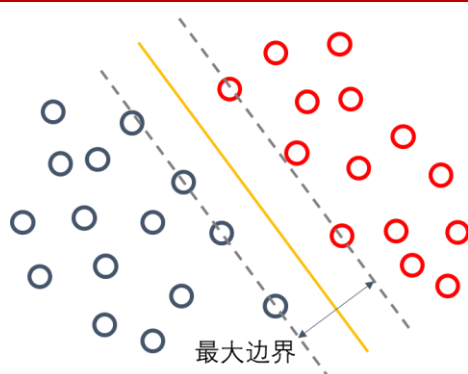
图 14: k 近邻



资料来源: 中信证券研究部

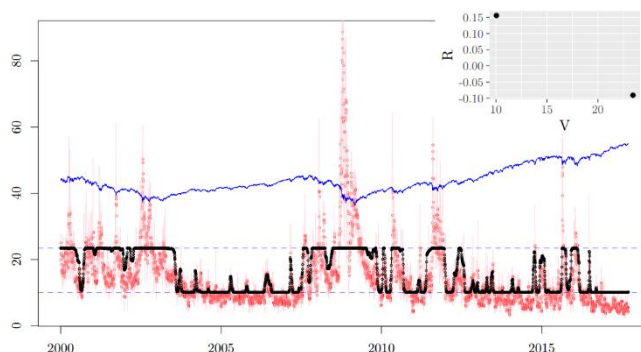
此外, 还有使用支持向量机, 隐马尔可夫模型和深度学习等模型预测资产收益率, 或采用贝叶斯方法预测市场拐点。如文献[11]回避信噪比较低的交易信息, 转而利用深度学习挖掘基本面因子间的关系, 预测未来基本面数据, 然后根据预测结果进行投资决策。

图 15: 支持向量机



资料来源: 中信证券研究部

图 16: 两种状态的隐马尔可夫模型预测波动率



资料来源: Stephen H.T. Lihn, 2017<sup>4</sup>, 中信证券研究部

## 资产定价

资产定价是除了预测问题外另一个应用最广泛的问题。例如, 因子模型 (APT) 和现金流折现模型 (DCF) 是两个最典型的定价模型, 但因子模型为线性模型, 而现金流折现模型为非线性模型, 两者不能完全融合。

根据因子模型, 股票的收益率是影响股票的多个因子的线性加和, 如以下公式所示。

<sup>4</sup> Lihn S H T. Hidden Markov Model for Financial Time Series and Its Application to S&P 500 Index[J]. Quantitative Finance, Forthcoming, 2017.



$$r_t = \sum_{i=1}^n w_i f_i + \varepsilon$$

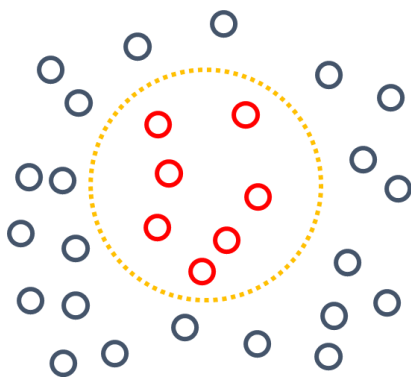
然而，根据股息贴现模型( $M_t$ 表示  $t$  时刻公司的市值， $Y_{t+\tau}$ 表示  $t + \tau$ 时刻的股权收益， $dB_{t+\tau}$ 为  $B_{t+\tau} - B_{t+\tau-1}$ ， $r$ 则为股东预期现金流的内部收益率，大约是长期预期股票收益率。)

$$M_t = \sum_{\tau=1}^{\infty} \frac{E(Y_{t+\tau} - dB_{t+\tau})}{(1+r)^\tau}$$

可以看到股票收益率与因子间是非线性关系，并且如果能够求解其解析解，会存在因子交叉项。文献[12]和[13]分别采用核方法和自动编码器（一种深度学习模型）来处理这一问题。

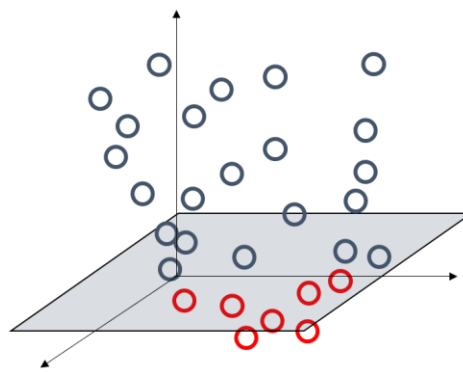
核方法的基本思路是通过核函数将原因子空间映射到高维空间，使得不适用线性模型的原空间经过映射后可以用线性模型处理。如图展示了一个分类任务，在二维空间( $x, y$ )中线性不可分，映射到三维空间( $x, y, x^2 + y^2$ )中，则可以用平面（三维空间的线性模型）分开。

图 17：核方法的原空间



资料来源：中信证券研究部

图 18：核方法的映射空间

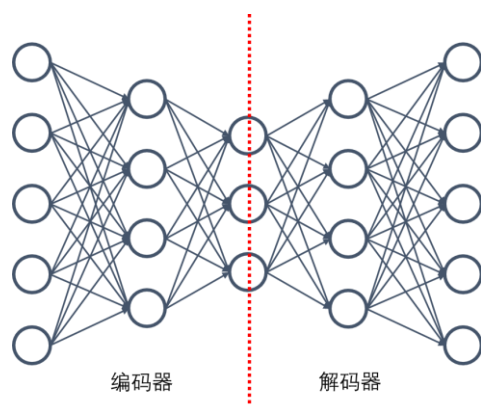


资料来源：中信证券研究部

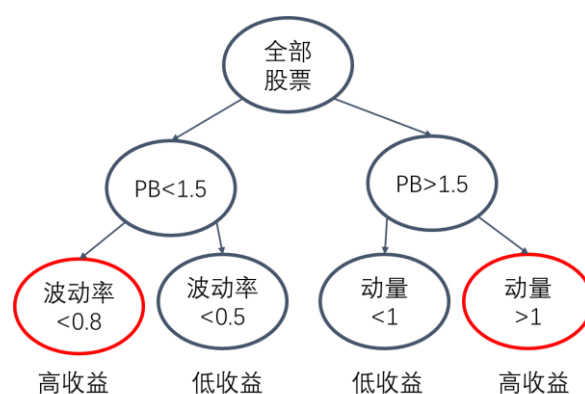
自动编码器是一种特殊的深度神经网络，呈对称的蝶形，编码阶段单元逐层减少最后形成瓶颈（bottleneck），然后经过解码器还原出输入信号，通过信息压缩的方式得到数据的另一种表示。在一定限制条件下，原空间不满足线性关系的可以在压缩空间满足。实际上自动编码器可以看成一种非线性的主成分分析（PCA），通过非线性映射实现信息压缩。

图 19：自动编码器算法原理

图 20：决策树算法原理



资料来源：中信证券研究部



资料来源：中信证券研究部

## 交易执行

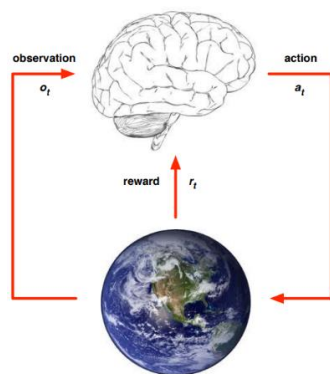
交易执行是投资活动中重要的一个环节，目标在于一定时间内买入（卖出）一定量的资产，使得成本（收益）更低（更高）。由于整个过程一般在较短的时间内完成，所以可以忽略公司基本面和宏观因素带来的影响。从而价值信息显得更加充分，缓解了我们在上文提到的低维输入的问题，有利于机器学习方法发挥作用。

文献[16]采用强化学习来优化交易执行，这是一种能够根据环境反馈直接学习决策的机器学习范式。通常的投资问题一般都忽略市场冲击或视为固定成本，而在交易执行的问题中恰恰是需要考虑市场反馈的，强化学习的特点正好符合这一问题的需求。

在一般的机器学习方法分类下，强化学习被认为是区别于监督学习和非监督学习的第三种类别，因为其没有监督学习中的标签信息，也不是和无监督学习一样完全没有指导信息，只是指导信息往往是延迟的，以一个全局的分数体现，比如玩游戏和下围棋。闻名世界的 DeepMind 围棋智能 AlphaGo 正是基于强化学习算法开发的。

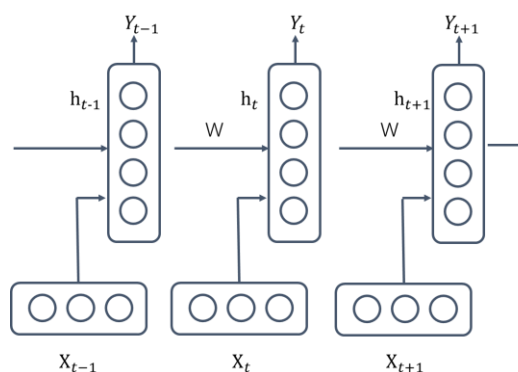
强化学习的常见模型是马尔科夫决策过程，如图 21 所示，在  $t$  时刻，智能体（agent）根据观察到的环境状态  $O_t$ ，选择了  $a_t$  的操作，然后环境给与了  $r_t$  的奖励，同时环境状态转变成  $O_{t+1}$ 。如此循环，强化学习的目标就是最大化累积奖励。

图 21：强化学习原理示意图



资料来源：icml.cc

图 22：循环神经网络



资料来源：中信证券研究部

在文献[16]中，状态表示为综合了市场交易信息及剩余时间和库存的向量，行为表示为现有库存的下单价格，而奖励表示为流进现金流（买入，卖出为流出现金流）。合理的表示了交易任务中的各种要素之后，就可以套用强化学习的方法对决策进行学习。

循环神经网络是神经网络的一个变种，适合对时间序列进行建模。文献[31]将盘口下单信息作为输入，来预测一定时间内的价格变化。相比于输入历史信息，盘口信息的时效性更好且更充分，有助于提高方法的有效性。

## 文本分析

随着信息化工具的发展，很多影响市场的信息最先都是以文本的形式进行传播，如公司公告、新闻资讯、论坛讨论等。因此，具有快速强大的文本信息提取能力可以极大地提高市场竞争力。而机器学习领域已经为文本处理准备了丰富的算法工具，我们面临的问题就是选择合适的工具并加以利用即可。

在这一方向上前人已经进行了不少尝试。文献[18]从文本的情绪分析入手挖掘超额收益。作者发现公司相关 Twitter 的情绪极性与股票未来几天的收益存在相关性，并根据情绪排名构造了市场中性组合，结果显示可以获得可观的收益。后者同样以 Twitter 作为研究对象，发现了相似的现象。

文献[32]认为单单对文本分类还不够，首先股票的热度不等价于观点，另外，不是每个人的情绪都是正确的，相反，大部分人可能不具备投资能力，其判断不具有参考价值。因此他们针对雅虎财经或 Raging Bull 这样的投资者交流平台，挖掘其中的投资专家。思路也很简单，首先对文本进行观点的识别，然后考察每一个用户的观点正确性。实证结果显示跟随挖掘出的专家确实能更准确的预测股价涨跌。

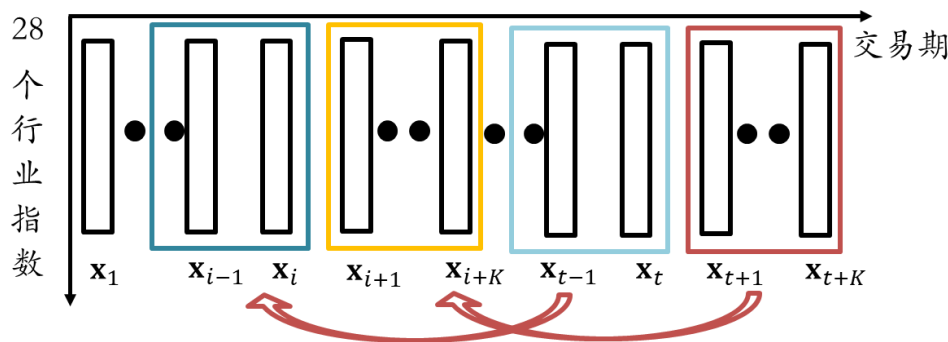
上述是对社交媒体数据的挖掘，也有直接对公司财报进行处理的工作。文献[21]采用词袋模型对财报进行建模，然后采用支持向量回归模型对股价未来的波动率进行回归，发现可以显著的优于历史信息估计的波动率。其基本原理就是提取出财报中的用词信息，从而带来了增量信息。

## 组合优化

资产组合是投资中的核心环节，其中面临的一大困难是市场的动态性。机器学习算法可以帮助跟踪市场的变化，如自适应贝叶斯方法和在线机器学习都是用来解决这一问题的。

另外，模式匹配技术（文献[26]）也被用来优化投资组合。该方法的基本假设为历史会重现，即相似的市场状态应该采用相似的资产组合。这一直觉性的假设实际上被研究人员不自觉地广泛应用，如判断市场未来趋势时去观察类似市场环境的历史，因此具有较强的解释性。

图 23：基于模式匹配方法的资产组合构建



资料来源：中信证券研究部

上图展示了模式匹配策略构建资产组合的基本思路，已知 $x_1$ 到 $x_{t-1}$ 所有历史信息，浅蓝色框表示的当前市场状态，红框是下一期的资产回报向量，需要为下一期各资产确定资产组合。深蓝色的框为历史中与当前时刻相似的市场状态，黄色的框是当时的资产回报向量。模式匹配策略的基本思路是搜集历史中相似的市场状态，则可以根据历史的资产回报优化下一期的资产组合。从参数估计的角度来看，采用相似时期进行估计要比现代资产组合采用近期数据估计的更加准确。

## ■ 机器学习在量化投资中应用的经典文献整理

- [1]. Qian B, Rasheed K. Stock market prediction with multiple classifiers[J]. Applied Intelligence, 2007, 26(1): 25-33.
- [2]. Rizvi S A A, Roberts S J, Osborne M A, et al. A Novel Approach to Forecasting Financial Volatility with Gaussian Process Envelopes[J]. arXiv preprint arXiv:1705.00891, 2017.
- [3]. Knoblauch J, Damoulas T. Spatio-temporal Bayesian on-line changepoint detection with model selection[J]. arXiv preprint arXiv:1805.05383, 2018.
- [4]. Tsinaslanidis P. Perceptually important points and dynamic time warping in time series prediction: Application to finance[J].
- [5]. Li S T, Kuo S C. Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks[J]. Expert Systems with applications, 2008, 34(2): 935-951.
- [6]. Kao L J, Chiu C C, Lu C J, et al. A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting[J]. Decision Support Systems, 2013, 54(3): 1228-1244.
- [7]. Kim K. Financial time series forecasting using support vector machines[J].

- Neurocomputing, 2003, 55(1-2): 307-319.
- [8]. Hassan M R, Nath B. Stock market forecasting using hidden Markov model: a new approach[C]//5th International Conference on Intelligent Systems Design and Applications (ISDA'05). IEEE, 2005: 192-196.
- [9]. Guresen E, Kayakutlu G, Daim T U. Using artificial neural network models in stock market index prediction[J]. Expert Systems with Applications, 2011, 38(8): 10389-10397.
- [10]. Chen K, Zhou Y, Dai F. A LSTM-based method for stock returns prediction: A case study of China stock market[C]//2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015: 2823-2824.
- [11]. Alberg J, Lipton Z C. Improving factor-based quantitative investing by forecasting company fundamentals[J]. arXiv preprint arXiv:1711.04837, 2017.
- [12]. Kozak S. Kernel Trick for the Cross Section[J]. Available at SSRN 3307895, 2019.
- [13]. Gu S, Kelly B T, Xiu D. Autoencoder asset pricing models[J]. Available at SSRN, 2019.
- [14]. Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning[R]. National Bureau of Economic Research, 2018.
- [15]. Giglio S, Xiu D. Asset pricing with omitted factors[J]. Chicago Booth Research Paper, 2018 (16-21).
- [16]. Nevmyvaka Y, Feng Y, Kearns M. Reinforcement learning for optimized trade execution[C]//Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 673-680.
- [17]. Dixon M. Sequence classification of the limit order book using recurrent neural networks[J]. Journal of computational science, 2018, 24: 277-286.
- [18]. Zhang W, Skiena S. Trading strategies to exploit blog and news sentiment[C]//Fourth international aAAI conference on weblogs and social media. 2010.
- [19]. Ranco G, Aleksovski D, Caldarelli G, et al. The effects of Twitter sentiment on stock price returns[J]. PloS one, 2015, 10(9): e0138441.
- [20]. Kamaruddin S S, Bakar A A, Hamdan A R, et al. Conceptual graph formalism


- for financial text representation[C]//2008 International Symposium on Information Technology. IEEE, 2008, 3: 1-6.
- [21]. Kogan S, Levin D, Routledge B R, et al. Predicting risk from financial reports with regression[C]//Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 272-280.
- [22]. Nyikosa F M, Osborne M A, Roberts S J. Adaptive Bayesian optimisation for online portfolio selection[C]//Workshop on Bayesian Optimization at NIPS. 2015, 2015.
- [23]. Györfi L, Lugosi G, Udina F. Nonparametric kernel - based sequential investment strategies[J]. Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics, 2006, 16(2): 337-357.
- [24]. Györfi L, Udina F, Walk H. Nonparametric nearest neighbor based empirical portfolio selection strategies[J]. Statistics & Decisions International mathematical journal for stochastic methods and models, 2008, 26(2): 145-157.
- [25]. Li B, Hoi S C H, Gopalkrishnan V. Corn: Correlation-driven nonparametric learning approach for portfolio selection[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 21.
- [26]. Wang Y, Wang D. Market Symmetry and Its Application to Pattern-Matching-Based Portfolio Selection[J]. The Journal of Financial Data Science, 2019, 1(2): 78-93.
- [27]. Li B, Zhao P, Hoi S C H, et al. PAMR: Passive aggressive mean reversion strategy for portfolio selection[J]. Machine learning, 2012, 87(2): 221-258.
- [28]. Li B, Hoi S C H, Zhao P, et al. Confidence weighted mean reversion strategy for online portfolio selection[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2013, 7(1): 4.
- [29]. Li B, Hoi S C H. On-line portfolio selection with moving average reversion[C]//Proceedings of the 29th International Conference on International Conference on Machine Learning. Omnipress, 2012: 563-570.
- [30]. Moritz B, Zimmermann T. Tree-based conditional portfolio sorts: The relation between past and future stock returns[J]. Available at SSRN 2740751, 2016.
- [31]. Dixon M F, Polson N G, Sokolov V O. Deep learning for spatio-temporal



modeling: Dynamic traffic flows and high frequency trading[J]. Applied Stochastic Models in Business and Industry, 2018.

- [32]. Bar-Haim R, Dinur E, Feldman R, et al. Identifying and following expert investors in stock microblogs[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1310-1319.

若中信证券以外的金融机构发送本报告，则由该金融机构为此发送行为承担全部责任。该机构的客户应联系该机构以交易本报告中提及的证券或要求获悉更详细信息。本报告不构成中信证券向发送本报告金融机构之客户提供的投资建议，中信证券以及中信证券的各个高级职员、董事和员工亦不为（前述金融机构之客户）因使用本报告或报告载明的内容产生的直接或间接损失承担任何责任。

 未经中信证券事先书面授权，任何人不得以任何目的复制、发送或销售本报告。  
中信证券 2020 版权所有。保留一切权利。

## 分析师声明

主要负责撰写本研究报告全部或部分内容的分析师在此声明：(i) 本研究报告所表述的任何观点均精准地反映了上述每位分析师个人对标的证券和发行人的看法；(ii) 该分析师所得报酬的任何组成部分无论是在过去、现在及将来均不会直接或间接地与研究报告所表述的具体建议或观点相联系。

## 评级说明

投资建议的评级标准		评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 到 12 个月内的相对市场表现，也即：以报告发布日后的 6 到 12 个月内的公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A 股市场以沪深 300 指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准；美国市场以纳斯达克综合指数或标普 500 指数为基准；韩国市场以科斯达克指数或韩国综合股价指数为基准。	股票评级	买入	相对同期相关证券市场代表性指数涨幅 20%以上
		增持	相对同期相关证券市场代表性指数涨幅介于 5%~20%之间
		持有	相对同期相关证券市场代表性指数涨幅介于-10%~5%之间
		卖出	相对同期相关证券市场代表性指数跌幅 10%以上
	行业评级	强于大市	相对同期相关证券市场代表性指数涨幅 10%以上
		中性	相对同期相关证券市场代表性指数涨幅介于-10%~10%之间
		弱于大市	相对同期相关证券市场代表性指数跌幅 10%以上

## 其他声明

本研究报告由中信证券股份有限公司或其附属机构制作。中信证券股份有限公司及其全球的附属机构、分支机构及联营机构（仅就本研究报告免责条款而言，不含 CLSA group of companies），统称为“中信证券”。

## 法律主体声明

本研究报告在中华人民共和国（香港、澳门、台湾除外）由中信证券股份有限公司（受中国证券监督管理委员会监管，经营证券业务许可证编号：Z20374000）分发。本研究报告由下列机构代表中信证券在相应地区分发：在中国香港由 CLSA Limited 分发；在中国台湾由 CL Securities Taiwan Co., Ltd. 分发；在澳大利亚由 CLSA Australia Pty Ltd. 分发；在美国由 CLSA group of companies（CLSA Americas, LLC（下称“CLSA Americas”）除外）分发；在新加坡由 CLSA Singapore Pte Ltd.（公司注册编号：198703750W）分发；在欧盟与英国由 CLSA Europe BV 或 CLSA（UK）分发；在印度由 CLSA India Private Limited 分发（地址：孟买（400021）Nariman Point 的 Dalalal House 8 层；电话号码：+91-22-66505050；传真号码：+91-22-22840271；公司识别号：U67120MH1994PLC083118；印度证券交易委员会注册编号：作为证券经纪商的 INZ000001735，作为商人银行的 INM000010619，作为研究分析商的 INH000001113）；在印度尼西亚由 PT CLSA Sekuritas Indonesia 分发；在日本由 CLSA Securities Japan Co., Ltd. 分发；在韩国由 CLSA Securities Korea Ltd. 分发；在马来西亚由 CLSA Securities Malaysia Sdn Bhd 分发；在菲律宾由 CLSA Philippines Inc.（菲律宾证券交易所及证券投资者保护基金会）分发；在泰国由 CLSA Securities (Thailand) Limited 分发。

## 针对不同司法管辖区的声明

**中国：**根据中国证券监督管理委员会核发的经营证券业务许可，中信证券股份有限公司的经营经营范围包括证券投资咨询业务。

**美国：**本研究报告由中信证券制作。本研究报告在美国由 CLSA group of companies（CLSA Americas 除外）仅向符合美国《1934 年证券交易法》下 15a-6 规则定义且 CLSA Americas 提供服务的“主要美国机构投资者”分发。对身在美国的任何人士发送本研究报告将不被视为对本报告中所评论的证券进行交易的建议或对本报告中所载任何观点的背书。任何从中信证券与 CLSA group of companies 获得本研究报告的接收者如果希望在美国交易本报告中提及的任何证券应当联系 CLSA Americas。

**新加坡：**本研究报告在新加坡由 CLSA Singapore Pte Ltd.（资本市场经营许可持有人及受豁免的财务顾问），仅向新加坡《证券及期货法》s.4A（1）定义下的“机构投资者、认可投资者及专业投资者”分发。根据新加坡《财务顾问法》下《财务顾问（修正）规例（2005）》中关于机构投资者、认可投资者、专业投资者及海外投资者的第 33、34 及 35 条的规定，《财务顾问法》第 25、27 及 36 条不适用于 CLSA Singapore Pte Ltd.。如对本报告存有疑问，还请联系 CLSA Singapore Pte Ltd.（电话：+65 6416 7888）。MCI (P) 086/12/2019。

**加拿大：**本研究报告由中信证券制作。对身在加拿大的任何人士发送本研究报告将不被视为对本报告中所评论的证券进行交易的建议或对本报告中所载任何观点的背书。

**欧盟与英国：**本研究报告在欧盟与英国归属于营销文件，其不是按照旨在提升研究报告独立性的法律要件而撰写，亦不受任何禁止在投资研究报告发布前进行交易的限制。本研究报告在欧盟与英国由 CLSA（UK）或 CLSA Europe BV 发布。CLSA（UK）由（英国）金融行为管理局授权并接受其管理，CLSA Europe BV 由荷兰金融市场管理局授权并接受其管理，本研究报告针对由相应本地监管规定所界定的在投资方面具有专业经验的人士，且涉及到的任何投资活动仅针对此类人士。若您不具备投资的专业经验，请勿依赖本研究报告。对于由英国分析员编纂的研究资料，其由 CLSA（UK）与 CLSA Europe BV 制作并发布。就英国的金融行业准则与欧洲其他辖区的《金融工具市场指令 II》，本研究报告被制作并意图作为实质性研究资料。

## 一般性声明

本研究报告对于收件人而言属高度机密，只有收件人才能使用。本研究报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。本研究报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。中信证券并不因收件人收到本报告而视其为中信证券的客户。本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。

本报告所载资料的来源被认为是可靠的，但中信证券不保证其准确性或完整性。中信证券并不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他损失承担任何责任。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

本报告所载的资料、观点及预测均反映了中信证券在最初发布该报告日期当日分析师的判断，可以在不发出通知的情况下做出更改，亦可因使用不同假设和标准、采用不同观点和分析方法而与中信证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。中信证券并不承担提示本报告的收件人注意该等材料的责任。中信证券通过信息隔离墙控制中信证券内部一个或多个领域的信息向中信证券其他领域、单位、集团及其他附属机构的流动。负责撰写本报告的分析师的薪酬由研究部门管理层和中信证券高级管理层全权决定。分析师的薪酬不是基于中信证券投资银行收入而定，但是，分析师的薪酬可能与投行整体收入有关，其中包括投资银行、销售与交易业务。

若中信证券以外的金融机构发送本报告，则由该金融机构为此发送行为承担全部责任。该机构的客户应联系该机构以交易本报告中提及的证券或要求获悉更详细信息。本报告不构成中信证券向发送本报告金融机构之客户提供的投资建议，中信证券以及中信证券的各个高级职员、董事和员工亦不为（前述金融机构之客户）因使用本报告或报告载明的内容产生的直接或间接损失承担任何责任。

未经中信证券事先书面授权，任何人不得以任何目的复制、发送或销售本报告。

中信证券 2020 版权所有。保留一切权利。