量化投资新趋势（3）：驶向另类数据的信息蓝海

胡骥聪
SAC 执证编号：S0080521010007
SFC CE Ref：BRF083
jicong.hu@cicc.com.cn

郑文才
SAC 执证编号：S0080121120041
wencai3.zheng@cicc.com.cn

周萧潇
SAC 执证编号：S0080521010006
SFC CE Ref：BRA090
xiaoxiao.zhou@cicc.com.cn

从量化策略的发展历史来看，量化策略创新发展一直交织着明暗两条线索，明线是量化策略模型的发展，暗线则是金融数据的爆发。另类数据对比传统结构化金融数据蕴含了大量新增信息，但由于另类数据的非结构化性质，通常不能够直接作为传统量化模型的输入。近年来机器学习模型的发展使得处理另类数据并提取剥离其中的关键有效的信息成为可能。本文认为量化策略发展的重要趋势之一为量化策略对“另类数据”的挖掘和使用，如何发掘另类数据以及用正确的方法将其应用在合适的模型之中是量化策略未来发展的重要方向之一。

另类数据：数据增量，信息蓝海

1）另类数据的定义与分类：由于量化投资规模的总体上升，股票市场中由传统低频价量数据和基本面数据构造的量化策略池逐渐拥挤，另类数据受到越来越多投资者的关注。本文中，通过数据的原始生成目的是否服务于金融市场定价，以及数据被金融投资市场的利用程度这两个维度，我们对另类数据给出了一种有效的定义与分类方式。

2）另类数据种类越来越丰富：近年来随着科技发展，各种新型数据呈现指数级别增长趋势。根据IDC的数据显示，2020 年全球共有 64.2ZB的数据，而这个数量预计在 2025 年会增长到 175ZB。从新闻和各类报告的文本舆情信息，到卫星图片中的港口交运数据，大量可以反映金融市场情绪、企业运营状态以及经济发展情况的领先指标蕴含在这片信息蓝海中。

3）海内外积极布局：另类数据本身的历史很长，但应用在投资领域的时日较短。海内外投资机构对于另类数据的布局均迅速发展。我们估计国内量化研究投资对另类数据的应用阶段正处于早期探索期的尾部。

机器学习：“因地制宜”协助另类数据处理，填补传统模型不足

机器学习模型快速发展种类繁多，各有优劣。在不同的投资场景中需要根据不同需求，利用合适的模型与算法，这是量化策略应用机器学习的关键与难点之一。除在经典量化领域的应用之外，本文认为将非结构化的另类数据处理成量化模型能够理解的结构化形式将成为机器学习模型在量化领域的核心作用之一。

投资应用场景翻新，量化守正出奇

另类数据的不断出现和科技的发展，催化了各种另类模型的发展，而另类模型和另类数据的结合所发掘的信息增量，将产生更新型的量化投资场景。本文展示一些使用另类数据在量化投资中的应用场景，以供参考：

1）挖掘新的alpha信息：利用高频交易数据，或者上市公司关联的新闻数据等新数据源，从中挖掘具有新增量信息的alpha因子，或者配合已有数据改进传统因子，构造有超额收益的量化策略。

2）被动投资、主题构建：使用上市公司年报中的主营业务部分构造业务关键词，将市场关心的主题词组与公司业务词组计算文本相关性并排序，从而选取其中相关性较高的上市公司，编制相关主题指数。

3）ESG投资：挖掘上市公司的社会责任报告，通过文本向量化和语义相关性计算社会责任报告对于ESG各类话题的重视程度，构建ESG的多空组合策略，多数年份获得显著超额收益。

更多作者及其他信息请见文末披露页

目录

另类数据：金融投资的宝矿 4
数据种类丰富，历史源远流长：从工业时代到信息与数字时代 4
投资应用刚刚起步：从结构化数据到非结构化数据 7
量化投资寻求破局：从积极布局到更深入的理解 7
机器学习：填补传统模型不足 12
机器学习解决传统量化问题 12
NLP 技术助力量化策略数据开疆扩土 15
投资应用场景翻新，量化守正出奇 17
文本数据被动投资：主题构建 17
高频量化数据应用：市场微观结构观察与选股 alpha 挖掘 18
新闻舆情数据应用：文本数据中的 alpha 20
另类数据 ESG 策略 21
附录 23
A. 常见机器学习模型特点 23
B. 深度 NLP 模型特点一览 25
C. 高频微观流动性指标汇总表 26
参考文献 27

图表

图表 1：按照数据另类性质分类 5
图表 2：按照产生主体分类 5
图表 3：另类数据历史追溯时长 6
图表 4：另类数据和传统数据的优势对比 7
图表 5：使用另类数据的时长分布（2019 年） 8
图表 6：是否打算使用另类数据（2019 年） 8
图表 7：对另类数据的上升需求比例最高（2018 年） 8
图表 8：基金公司应用另类数据种类（2019 年） 8
图表 9：美国另类数据产业图 9
图表 10：滞后使用另类数据带来的风险 10
图表 11：另类数据应用周期 10
图表 12：基金经理于对冲基金使用另类数据的阻碍（2019 年） 11
图表 13：使用另类数据的风险 11
图表 14：机器学习模型分类 13
图表 15：机器学习模型与相关应用场景 13
图表 16：衍生品定价领域传统模型与机器学习模型对照表 14
图表 17：涉及金融领域的机器学习论文主题分布（2015-2019 年） 15
图表 18：涉及金融领域的机器学习模型主题分布（2015-2019 年） 15
图表 19：时间序列建模发展 15
图表 20：NLP 主要模型发展史 16
图表 21：上市公司年报章节内容有所调整 17
图表 22：山西焦化 2016 年主营业务词云图 18
图表 23：宁德时代 2018 年主营业务词云图 18
图表 24：主流宽基指数弹性均值 19
图表 25：主流宽基指数宽度均值 19
图表 26：主流宽基指数平均深度均值 19
图表 27：主流宽基指数有效深度均值 19
图表 28：盘口深度因子历史 IC 序列 19
图表 29：盘口深度因子分组收益表现 19
图表 30：数库新闻数据样本 20
图表 31：新闻动量因子 IC 序列测试 20
图表 32：单因子多空策略净值 21
图表 33：多空策略分年份表现 21
图表 34：利用企业 CSR 报告计算 ESG 重视程度指标 22
图表 35：公司 ESG 重视程度指标多空收益 22
图表 36：公司 ESG 重视程度指标多空收益曲线 22
图表 37：常见机器学习模型特点-上 23
图表 38：常见机器学习模型特点-下 24
图表 39：NLP 深度学习 25
图表 40：高频微观流动性指标汇总表 26

另类数据：金融投资的宝矿

站在漫长的金融市场的视角来看，量化投资兴起的时间相对较短，Alfred W. Jones 作为世界上第一家对冲基金创始于 1949 年，主要是使用多空股票策略获得收益。在纳斯达克交易所 1971 年上线并成为世界上第一个电子证券交易市场开始，股票的交易数据逐渐变得简单易得，量化策略在此基础上进一步发展。上个世纪 70-80 年代，基于股票价量数据的技术分析成为主流方法，道式理论得到发展并在现在仍作为多种技术分析手段的基础。1990 年代 Fama-French 和 Barra 的因子模型在 CAPM 模型的基础上利用公司基本面信息作为自变量解释股票的收益率。麦克刘易斯在《Flash Boys》中一书介绍了在 2009 年为了将交易数据的传输时间从 17 毫秒降低到 13 毫秒，从芝加哥到新泽西的一条上千公里造价不菲的光缆轨道建设的故事将高频交易的面纱揭开了一个小角。2010 年代机器学习技术在图像识别，自然语言处理领域崭露头角，也给量化交易公司带来了新的探索方向，各类资产管理公司通过不同方式希望将最新的机器学习技术应用到量化交易中。

不难发现，量化策略的发展一直交织着明暗两条线索，**明线是量化策略模型的发展，暗线则是金融数据的爆发**。从技术图形分析到因子理论再到高频交易，量化模型的发展与可用数据的发展紧密相关。近些年机器学习模型的井喷发展除了算力大幅提升外，更重要的是具有去学习的对象，也即输入数据，例如社交媒体上取之不尽的图像数据可供如 Google、Meta 这样的互联网巨头学习使用。

对于量化领域来说，现有的结构化数据在经历十几年发展后可供挖掘的信息逐渐匮乏，市场策略逐渐饱和且新策略开发速度难以跟上资金体量的增长。量化策略亟待新的发展方向，本文认为在机器学习发展的背后，**量化策略发展的重要趋势之一为量化策略对“另类数据”的挖掘和使用，如何发掘另类数据以及用正确的方法将其应用在合适的模型之中是量化策略未来发展的重要方向之一。**

数据种类丰富，历史源远流长：从工业时代到信息与数字时代

几乎每一次金融市场的变革都是由于科技的发展与进步。在工业时代，投资活动中使用的数据少量且零散。步入信息时代和数字时代后，可使用数据丰富程度大幅提升，数据量也呈指数级增长。

在纳斯达克交易所 1971 年上线并成为世界上第一个电子证券交易市场开始，股票的交易数据逐渐变得简单易得，量化策略在此基础上进一步发展。由于量化策略规模的总体上升，股票市场中由传统低频价量数据和基本面数据构造的量化策略池逐渐拥挤，策略有效性衰减速度也逐渐加快，策略创新成为量化策略中的重点工作之一。

创新主要可以分为两个方面，一方面为模型层面的创新，主要表现为策略构造方式逐渐复杂，非线性程度加深，但其本质上还是对现有信息的精加工；另一方面为数据创新，主要表现为主动挖掘和创造信息增量。

机器学习和另类数据的出现给以上两种方向分别提供了思路，另类数据对比传统结构化数据蕴含了大量新增信息，机器学习模型将关键有效的信息从中剥离，应用到相应的投资策略中，两者应有机结合才能取得更好的实际效果。

股票投资者利用另类数据进行投资的历史比应用结构化数据的历史更久。在交易所尚未步入电子化之前，投资者主要是通过阅读报纸观看新闻来了解对应投资标的的情况。这种投资行为其

实就是对新闻文本数据在投资方面的典型应用，实际上这种基于新闻信息的投资方式在现代仍是许多个人投资者的主要策略。一直以来由于另类数据的非结构化属性，导致传统的量化策略难以将其像价量数据一样作为输入变量加以利用。近年来随着 NLP 技术和机器学习模型的发展，许多非结构化数据例如新闻文本数据中可被模型挖掘的信息逐渐增多，量化策略对于此类数据的应用也逐渐加强。

► 另类数据的定义与分类

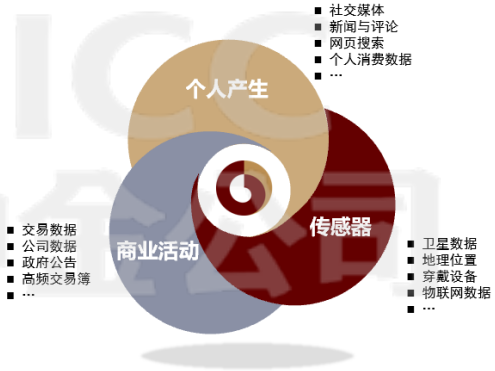
另类数据作为一个相对的概念，目前还没有一个统一的定义，而且另类数据的定义会随着时间变化：当某种另类数据逐渐被市场上大部分参与者所接纳和使用时，它也就不会再显得非常“另类”。目前来看，一个大多数人认可的定义为：另类数据是指投资研究中使用的非传统来源的新型数据。

结合另类数据的动态性和非传统性，我们对于另类数据做出以下定义：对量化策略来说，金融领域中应用较少的，或产生目的原本并非为了应用在投资和金融领域的数据即可被称之为另类数据。我们将常见的另类数据按照该标准划分类别，其中被标为深红色的数据类型其非结构化程度更深。也可按照产生主体划分[1]。

图表 1：按照数据另类性质分类

资料来源：中金公司研究部

图表 2：按照产生主体分类



资料来源：Eagle Alpha，中金公司研究部

► 另类数据的历史

新闻另类数据是各种另类数据中可追溯历史最长的数据之一。自活字印刷被发明以来，文本资料的传播和发行变得相对容易。16 世纪左右，现代新闻报社开始在欧洲兴起，自此新闻发行活动一直延续至今。

企业活动产生的另类数据历史也相当久远，企业和企业之间以及企业与个人之间的交易行为会产生大量数据。企业记录自身经营活动的行为逐渐标准化为如今的财务报表。这也是另类数据逐渐变为标准化数据的典型例子。此外仍有许多在企业经营生产活动中产生的未被规范化的数据，例如企业之间的产业链关系数据、公司公告、广告发布、岗位招聘以及企业的专利权数据等。

公共政策部门颁布的宏观统计数据和各类政策文件以及官员的公开讲话也蕴含着大量信息。举例来说，近期美国通货膨胀率达到了 40 年新高 7.5%，导致市场预期美联储将要在未来加息，

1 Eagle Alpha. 2019. “Alternative Data Use Cases Edition 6”

而加息的力度可以通过美联储主要官员在各类公开讲话的措辞中体现出来。挖掘美联储官员讲话内容以推测美联储加息力度就是对另类数据中文本数据在金融经济活动中的典型应用。

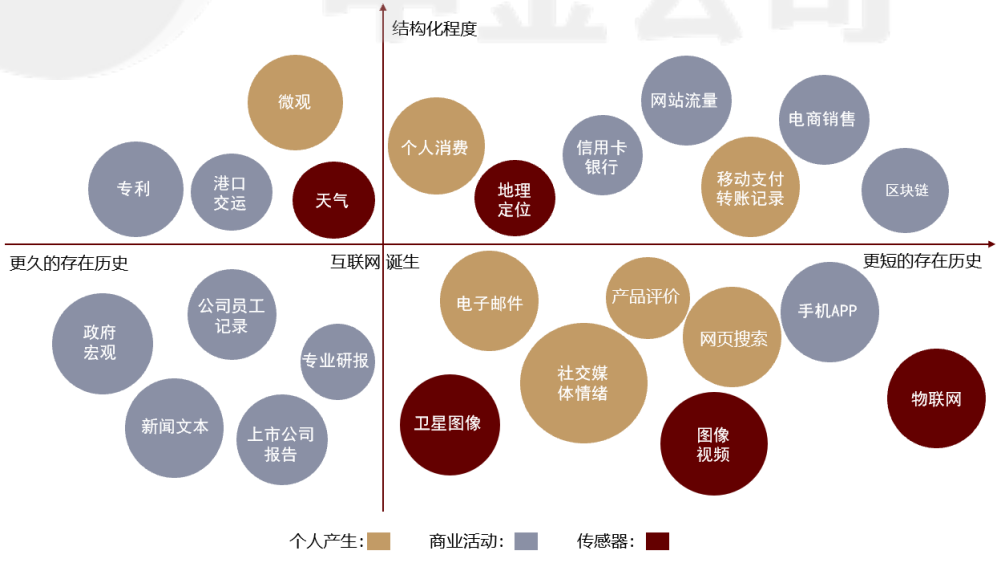
可系统性使用的另类数据主要产生在互联网诞生之后，蒂姆·伯纳斯于 1990 年底推出世界上第一个网页浏览器和第一个网页服务器，推动了万维网的产生，导致了互联网应用的迅速发展。尽管经历了 2000 年初期的互联网泡沫，随着互联网用户的增加，互联网在现代经济生活中正发挥着日益重要的作用。个人用户和企业在互联网上发布相关信息，产生了大量的数据交换。例如用户最常使用的谷歌百度线上信息检索功能蕴含了互联网热点趋势信息，同時各类企业也带来如百度地图、招聘网站等提供地图与卫星图、企业招聘数据等各种另类数据。

2010 年以后的移动互联网时代进一步催化了这一进程，智能手机让人随时随地使用互联网，进而创造更多种类的另类数据。用户会在 Instagram、B 站这样的虚拟社区中上传大量文本和视频，也会在通讯软件如 WhatsApp 和微信上进行日常交流，移动支付软件如支付宝、微信支付等数据隐含各类用户的消费习惯，美团、大众点评等软件上的评论信息则可以反应各类商铺的消费热度。

当移动互联网终端结合卫星 GPS 定位数据时，可以推断出各大商场的营业量的变化进而预测上市公司的财报数据。卫星数据属于感应器的一种，其中气象卫星的气象数据可以通过对天气的预测来判断农作物的收成状况进而预判相关商品期货的走势。目前应用最广的卫星数据类型仍为对地成像的图片类型数据，通过对特殊地点如特定港口和交通要塞的交通分析，对交通流量以及总体经济状况提供直观依据。

目前来说，互联网公司的中心化使得大量用户数据高度集中，互联网公司使用此类另类数据的方法较为单一，多为使用算法精准投放广告。随着监管的不断成熟，对于另类数据的使用也会更加规范。并且随着 web3.0 的不断发展，未来用户的数据有望会重新回到用户手中。

图表 3：另类数据历史追溯时长[2]



资料来源：Eagle Alpha，中金公司研究部

2 Niall Hurley. 2021. “B2021 Alternative Data Report: Year in Review”. Eagle Alpha

投资应用刚刚起步：从结构化数据到非结构化数据

虽然另类数据的历史源远流长，但量化策略对非结构化数据的应用历史却是刚刚启程，主要原因是在 NLP 等模型及高效的信息采集与传输技术出现之前，量化策略开发者们不能批量使用非结构化数据来满足量化交易策略体系化的特点。因此在 2010 年之前，量化策略主要使用的数据仍然是价量交易数据和上市公司披露的年报数据等，通过对此类数据的分析以及在各类模型中的应用，构造出可以形成超额收益的量化模型帮助做出交易决策。

由于价量及基本面数据的易得性，多种基于该类数据的量化策略迅速发展，模型的复杂性也随之升高，市场策略拥挤度一再上升，导致模型失效的周期也进一步缩短。一方面，近年来量化私募的快速发展导致了规模快速扩张，而单个量化策略本身的容量有限，同时很多量化策略难以随着规模的快速上升而迅速调整。另一方面，去年以来的量化基金收益率高波动和去年末的大幅回撤也使得市场出现对于量化策略高度同质性的一些质疑。

由于金融市场发展阶段的差异性，这种情况在海外发生的时间更早。参考美国私募市场发展情况，在 2000 年到 2008 年经历过一段快速增长期，后来由于市场策略的拥挤度逐渐升高，增长速度逐渐放缓。随着近 10 年来的科技发展，机器学习和 NLP 等技术使得文本数据等非结构化数据得以作为量化模型的输入数据，量化投资公司开始逐渐重视另类数据在量化策略中的应用，并总结出以下另类数据相对传统数据的优势。

图表 4：另类数据和传统数据的优势对比

	另类数据	传统数据	描述
数据体量和信息	高频更新 更短的历史 更宽的数据宽度	低频更新 更长的历史 更窄的数据宽度	大量数据被不断产生 高频更新使得基金经理研究更具前瞻性，加强组合构建优势
未被发掘的观点	更宽的数据宽度 对于资产类别和行业具有深刻指导价值	窄数据宽度 仅提供涉及少数资产的特定信息	新兴数据不仅基于财务数据产生，使趋势预测更全面准确 开发新信息可使得投资策略更分散
竞争优势	需要投资和数据处理能力	对所有人可用 研究开展更容易	需要人才技术才能从另类数据中挖掘信息，并帮助基金经理产生更高的收益
受托责任	利用所有可用的信息和数据	利用金融数据	需要人才技术才能从另类数据中挖掘信息，并帮助基金经理产生更高的收益
效率	高速和高效的研究 更广的覆盖度	高度人工研究 更窄的覆盖度	另类数据为多种资产提供研究见解 可替换特定现有研究流程 研究员可花费更多时间在建模和投资策略中

资料来源：Eagle Alpha，中金公司研究部

量化投资寻求破局：从积极布局到更深入的理解

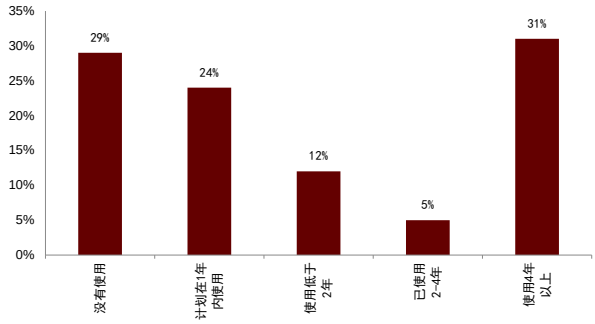
► 海外另类数据发展需求旺盛

2018 年 12 月至 2019 年 2 月，Greenwich Associates[3] 采访了全球投资管理公司的 42 位 CIO、投资组合经理和投资分析师。受访者回答了有关未来 5-10 年投资研究将如何变化的问题。

关于他们是否打算使用另类数据，近 50% 的投资经理表示他们正在使用另类数据，另有接近四分之一计划在未来 12 个月内使用该数据。而在回答有关管理者期望增加或减少他们对哪些信息来源时，另类数据也是受访者回答中希望增加使用比例最高的信息来源。

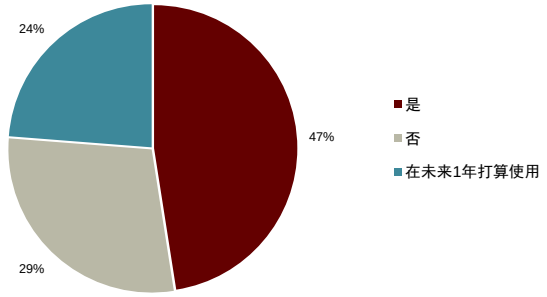
3 Brad Tingley. 2020. “FISD Alternative Data Council: A Guide to Alternative Data”. Greenwich Associates

图表 5：使用另类数据的时长分布（2019 年）



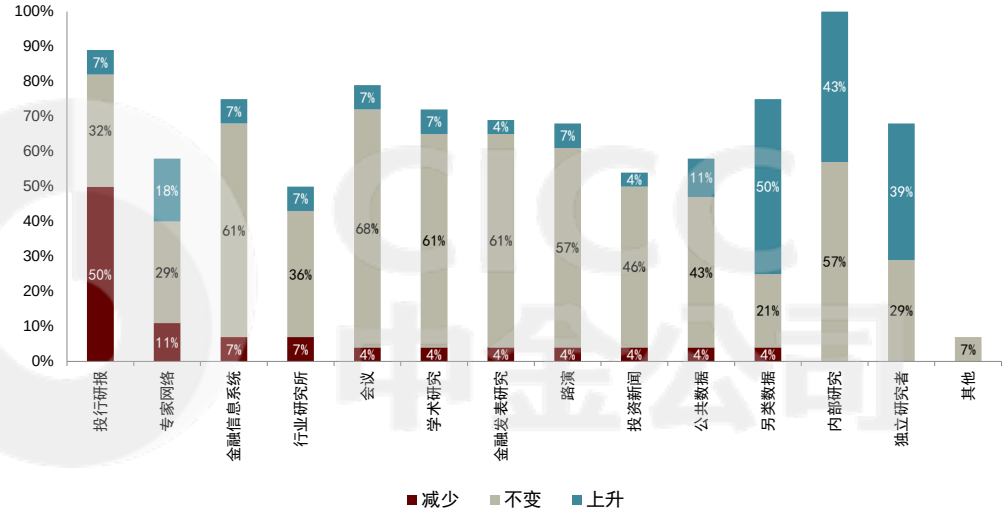
资料来源：格林尼治协会，中金公司研究部

图表 6：是否打算使用另类数据（2019 年）



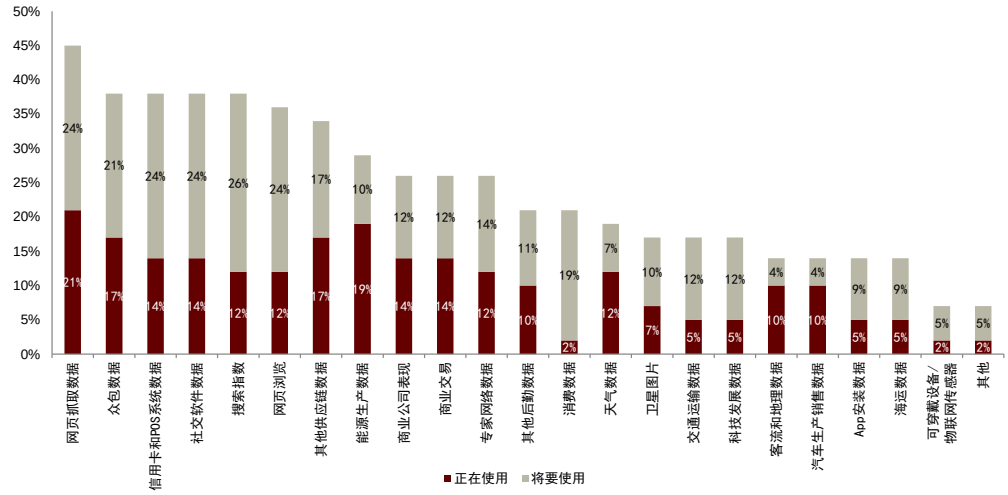
资料来源：格林尼治协会，中金公司研究部

图表 7：对另类数据的上升需求比例最高（2018 年）



资料来源：格林尼治协会，中金公司研究部

图表 8：基金公司应用另类数据种类（2019 年）



资料来源：格林尼治协会，中金公司研究部

► 另类数据市场发展较快

目前来看，另类数据仍在以较快的速度不断积累，另类数据公司数量迅速增加和另类数据市场空间巨大。从数据积累方面看，根据 IDC[4]（国际数据公司）的一份报告，2018 年全球有 33ZB 的数据，而这个数量预计在 2025 年会增长到 175ZB，这依赖于计算机算力的提升和存储设备技术的提高。

从另类数据公司数量上看，据 AlternativeData 的统计数据[5]，2018 年全球另类数据公司已增长到近 400 家，国内另类数据公司大约占 100 家，国内发展迅速。根据出售数据的处理程度，另类数据公司主要分为三类：1）原始数据提供者，这类供应商只收集最原始的另类数据，对于数据的处理程度最小；2）轻处理数据提供者，提供与金融资产相关的可视化数据；3）信号提供者，一般关注于某个特定行业，向资产管理公司提供打包好的量化投资信号。

从市场空间上看，AlternativeData 统计表明截止 2017 年全球已有约 800 支基金利用另类数据做投资决策，2017 年投资机构对另类数据的投入规模约为 4 亿美金，行业正处于快速发展期，到 2020 年另类数据市场已增长到 17 亿美元。Grandview research 预计 2021 年至 2028 年 CAGR 复合增长率将达 58.5%。

图表 9：美国另类数据产业图

资料来源：格林尼治协会，中金公司研究部

► 滞后使用以及另类数据自身风险

贝莱德 SAE（Systematic Active Equity）团队曾表示[6]，为了产生持续的阿尔法，投资者应该接受获取、分析和理解快速增长的数据领域。那些无法做到这一点的投资者将面临在快速变化的投资环境中落后的风险。在德勤发布的一篇报告中显示，滞后使用另类数据将面临以下三类风险[7]：

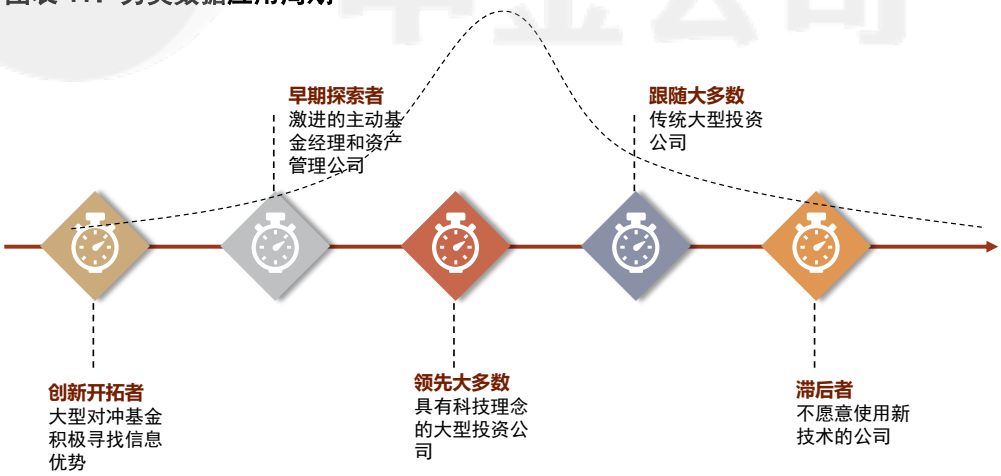
[4] David Reinsel. 2018. The Digitization of the World: From Edge to Core. IDC
[5] Alternativedata. 2022. https://alternativedata.org/alternative-data/
[6] Blackrock. 2015. 'The Evolution of Active Investing. Finding Big Alpha in Big Data'
[7] Deloitte. 2017. Warming up to collective intelligence investing.

图表 10：滞后使用另类数据带来的风险

资料来源：德勤，中金公司研究部

对于新型科学技术或新型数据的研究与应用，市场上的参与群体大多会遵循以下周期：在技术出现的前期，有一小部分创新开拓者由于体量较大，有更多的试错成本，积极利用新兴技术寻找竞争优势，随着市场的发展，更多市场参与者意识到该新兴技术或数据中蕴含的价值并及时跟进。随着早期探索者开始因此获益，越来越多的中小公司也开始跟进，直到该技术市场被参与者完全挤占，最后无法再获得新的竞争优势。我们认为，另类数据的使用周期符合这一模型的设定，并且按我们估计，国内量化研究与投资正处于早期探索期的尾部。

图表 11：另类数据应用周期

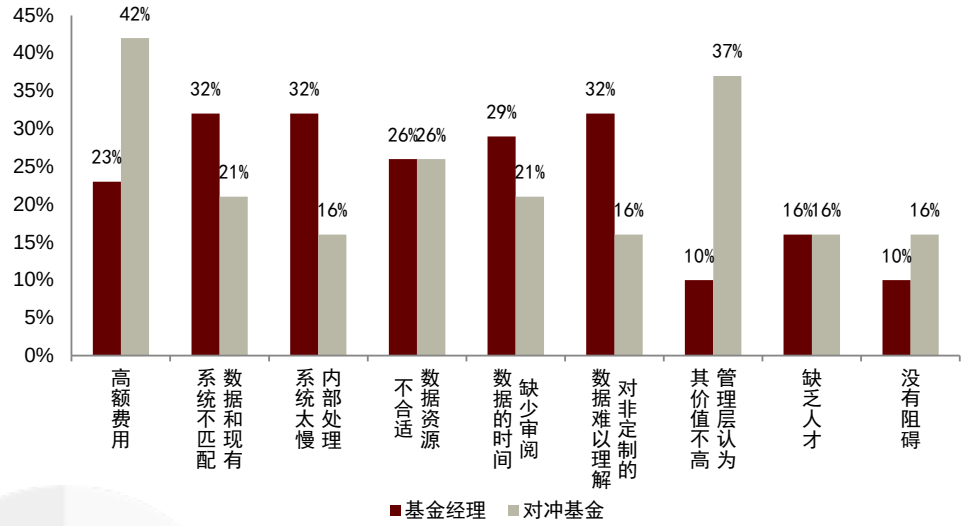


资料来源：德勤，中金公司研究部

不能及时跟进新型技术或新型数据固然会面临压力，但如何正确使用另类数据这样的新型数据同样值得思考。使用另类数据也会面临一定的阻碍和风险，主要包括另类数据缺乏处理另类数据的工具和人才等。并不是所有另类数据都可以帮助基金公司获取 Alpha，因此另类数据公司在收集、清洗数据之前，需要基金经理来评判数据是否有价值，同时高效的处理和应用将是降低成本和提升效率的关键，因此另类数据公司应当具备机器学习等技术开发能力和高效的产品策略。目前另类数据正在全面崛起，国内外出现了一批创新型的另类数据公司，为传统金融机

构提供新的创新动能。未来，另类数据在金融服务中的角色将越来越重要。目前使用另类数据设计的风险有以下几种：1.无效应用和错误指导；2.早期投入高成本；3.隐私问题；4.监管风险；5.另类数据信息在多家跟进后信息衰减[8]。

图表 12：基金经理于对冲基金使用另类数据的阻碍（2019 年）



资料来源：格林尼治协会，中金公司研究部

图表 13：使用另类数据的风险

资料来源：德勤，中金公司研究部

[8] Deloitte US. 2019. “Alternative Data – Perspectives and Insights”

机器学习：填补传统模型不足

另类数据最主要的特点之一是非结构化。传统量化模型在处理新型数据时有心无力，主要的限制在于另类数据的格式和大小与传统量化领域使用的结构化数据可以存在较大差异。例如新闻类的文本数据多以 pdf、txt 形式储存，卫星数据可能含有大量的图片格式数据。为提取和处理此类信息，各种新型另类模型高速发展，我们认为目前对于量化策略来说其中最重要的两个领域为机器学习模型和自然语言处理模型，他们是将非结构化数据转化为量化策略可以使用的结构化数据的核心。

机器学习是一个混合领域，结合了两个领域的概念和研究：统计学（数学）和计算机科学。与其起源相关，机器学习大致可以分为三类：经典机器学习、深度学习与强化学习。有监督和无监督学习的经典机器学习方法是统计学的自然延伸。量化分析策略的大多数应用都属于经典机器学习的范畴。深度学习方法对于处理卫星图像、自然语言处理和其他非结构化数据的分析中应用更广。例如在实践中，分析师可以构建卷积神经网络框架来直接从卫星图像文件中计算汽车数量变化的时间序列；强化学习方法可能构造出比人工交易更快速且胜率更高的自动交易策略。

机器学习解决传统量化问题

在机器学习领域，一个基本的共识就是“没有免费的午餐”。换言之，就是没有一类算法能完美地解决所有问题。机器学习不是一种一劳永逸的解决方法，越好的工具更需要知道正确使用方法才能完全发挥其优势，将机器学习模型生搬硬套到现有的量化策略中可能会过犹不及。传统量化模型使用的大多是 20 年以内的日度或月度数据，相对于适用于大量图片等大型数据集的深度学习等机器学习模型来说数据体量相对较小，强行使用会出现过拟合等问题。

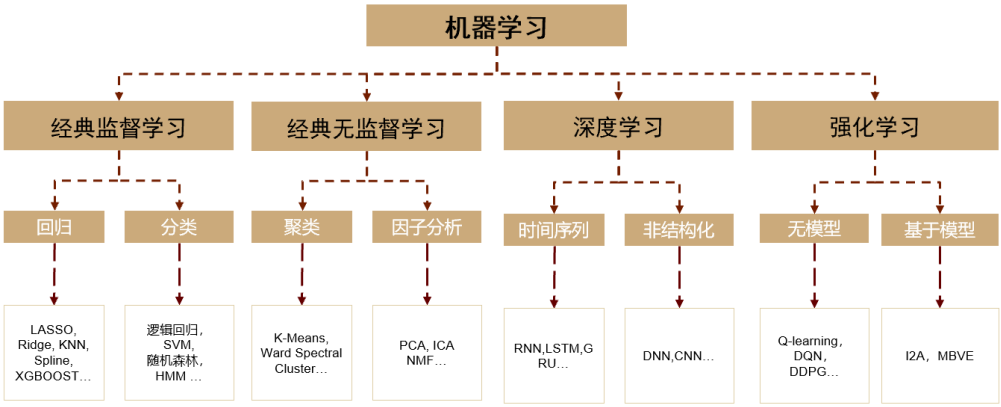
不同机器学习模型由于其特性的差异导致其适用的数据特征、适用场景以及期望目标各有不同，根据特定场景差异化地使用相应的机器学习模型能够事半功倍，发挥出机器学习真正的能力。举例来说，不能去说神经网络任何情况下都能比决策树更有优势，反之亦然。它们要受很多因素的影响，比如你的数据集的规模或结构。其结果是，在用给定的测试集来评估性能并挑选算法时，我们应当根据具体的问题来采用不同的算法。当然，所选的算法必须要适用于相应的问题，这就要求选择正确的机器学习任务。

- 机器学习模型宜“因地制宜”

机器学习模型（Machine Learning）是近年来量化领域尝试应用的重点模型。但目前大多数机构对于机器学习模型使用方法仍是使用传统价量数据和财务数据等高度结构化数据对模型进行训练，本质上仍是通过将数据集非线性化处理达到对存量信息的精细化提取。这种使用方法一方面仍然只使用了存量数据的信息，另一方面，传统数据不仅从维度还是从数据量来说都不太适合直接盲目应用在现有的机器学习模型中，尤其是具有复杂结构的深度学习模型，数据量的不足会直接导致过拟合等一系列问题，使得模型在样本内表现优秀，而在样本外很难应对市场风格（Regime）的调整。因此对于不同的问题，不同的数据应该根据相关模型的特点“因地制宜”。

机器学习发展经历了从感知机到支持向量机再到神经网络强化学习的几个阶段，随着计算能力的提升，机器学习模型结构也逐渐复杂。机器学习模型具体可以分为以下几类：

图表 14：机器学习模型分类[9]



资料来源：Iqbal H，中金公司研究部

我们整理了一些主流机器学习模型及其所常用的经典应用场景。更多关于不同机器学习模型的原理、优缺点及其合适的应用场景细节可参考文末附录 A。

图表 15：机器学习模型与相关应用场景

模型	应用场景
线性回归	因子模型，ARIMA，GARCH
Logistic Regression	信用评估，预测收益，风险事件预测，经济预测
朴素贝叶斯	文本分类，文本情感分析
KNN	文本分类，模式识别，聚类分析，多分类问题
SVM	涨跌分类，违约分类
决策树	投资决策，公司信用判断
随机森林	信用分类问题，收益率预测
Adaboosting	市场模式识别，收益率预测
GBDT	小特征空间，收益率预测
XGBoost	小特征空间，收益率预测
LightGBM	大特征空间，信用基差预测，收益率预测
神经网络	信用违约，自然语言处理，衍生品定价
CNN	图像处理检测分类，目标检测
RNN	自然语言处理，语音识别
LSTM	大数据量时间序列
K-Means	无标记的文本分类
PCA	收益率分解
EM最大期望算法	低维特征空间，小规模数据聚类
密度聚类	文本数据分类
层次聚类	产业链聚类
GRU	小数据量时间序列
RBM	降维，特征学习，波动率预测
GAN	自动提取特征，自动判断和优化
自编码器	异常监测，数据去噪，数据降维
卡尔曼滤波	数据去噪，提取交易信号
强化学习	组合构建，高频交易，策略优化

资料来源：Iqbal H，中金公司研究部

[9] Iqbal H. 2021. “Machine Learning: Algorithms, Real-World Applications and Research Directions”. SN Computer Science

► 机器学习应用场景广泛：资产配置、选股、衍生品定价等

传统量化模型多在价量、基本面以及宏观数据上开发择时、因子选股以及资产配置策略例如 ARIMA、Fama-French 因子模型与 Black-Litterman 模型等，机器学习也可以解决一些传统的股票量化模型的问题，例如卡尔曼滤波提取交易信号，LightGBM 预测资产收益率等。时间序列方面，在经历 ARIMA，GARCH 模型的传统线性模型的发展后，机器学习中的 RNN，LSTM 等模型又对时间序列的预测提供了非线性的思路。

衍生品量化模型多借助常微分方程等数学模型计算衍生品的理论价格将其作为定价和交易的基础如 Black-Scholes 模型，而机器学习也可以帮助计算如何给某些奇异期权定价或帮助其对冲风险。

图表 16：衍生品定价领域传统模型与机器学习模型对照表

	模型基础	技术方法
量化金融模型	随机过程，偏微分方程， 组合优化，因子模型，...	有限差分法，蒙特卡洛模拟， 有限制优化，线性模型， ...
深度学习模型	深度学习/机器学习，反应式学习， 无模型数据驱动， ...	神经网络，有监督/无监督学习，强化学习， 深度生产模型，信号向量化，非线性模型，...

资料来源：CQF，中金公司研究部

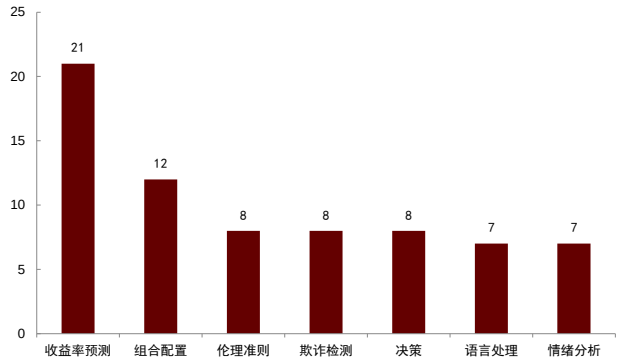
► 机器学习模型在量化研究领域广受欢迎

从 Markowitz 投资组合优化到最近的 CAPM、EMH 和因子模型，量化投资者已经表明他们愿意接受新技术和策略。将机器学习技术应用于金融投资问题主要有两种，其一是通过机器学习模型捕获数据和预测结果的非线性关系，第二是将非结构化的数据转化为量化模型可以理解的结构化数据或直接转化为交易信号。机器学习已在量化投资的许多领域中应用并取得了积极成果，包括投资组合优化、因子投资、债券风险可预测性、衍生品定价、对冲和交易策略等。

近年来机器学习模型在量化领域应用越发广泛，通过对 2015 年至 2019 年 118 篇机器学习在金融方向的研究论文后，Sophie[10]得出量化领域研究涉及话题最高频的两个主题是收益率预测和组合配置。其中涉及的模型最多的是多层神经网络模型、支持向量机模型以及 LSTM。

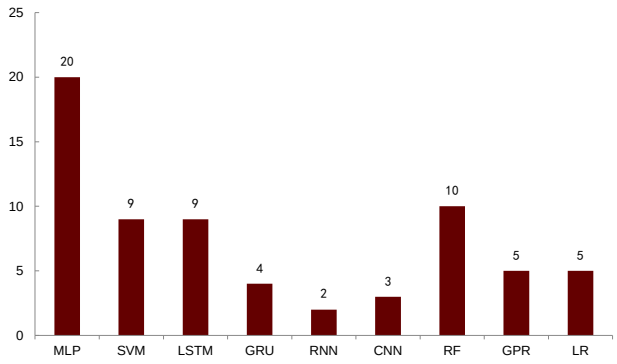
[10] Sophie Emerson. 2019. “Trends and Applications of Machine Learning in Quantitative Finance”

图表 17：涉及金融领域的机器学习论文主题分布（2015-2019 年）



资料来源：Sophie Emerson，中金公司研究部

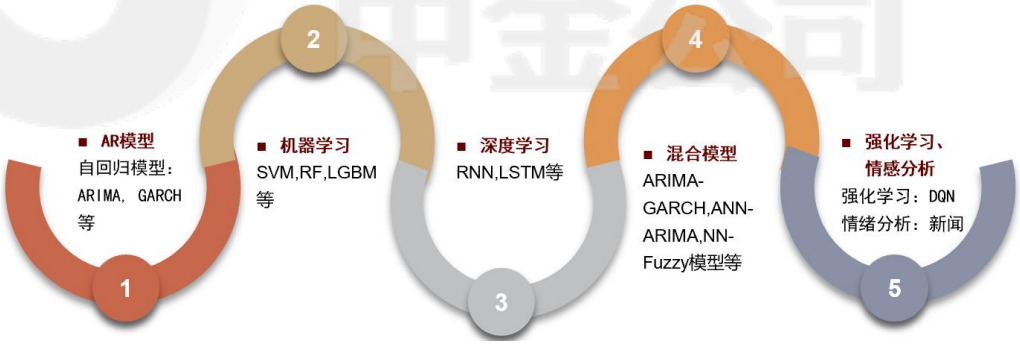
图表 18：涉及金融领域的机器学习模型主题分布（2015-2019 年）



资料来源：Sophie Emerson，中金公司研究部

机器学习提供了比以前更复杂的金融分析的能力。从这些文献中可以看出，量化投资者已经接受了新出现的工具和技术。越来越多的论文将机器学习技术应用于投资问题。多种机器学习方法已应用于量化金融领域，最流行的方法是神经网络模型，其次是 SVM 和 LSTM。机器学习已应用于回报预测、投资组合构建和风险建模等领域的问题。这些机器学习方法利用传统的财务数据，以及利用新型另类数据。大数据提供了需要分析的新数据集，而机器学习技术能够对复杂（非线性）关系进行建模并分析新型数据。在金融投资领域最重要模型之一的时间序列模型方面，机器学习应用也得到了长足的发展。

图表 19：时间序列建模发展



资料来源：CQF，中金公司研究部

NLP 技术助力量化策略数据开疆扩土

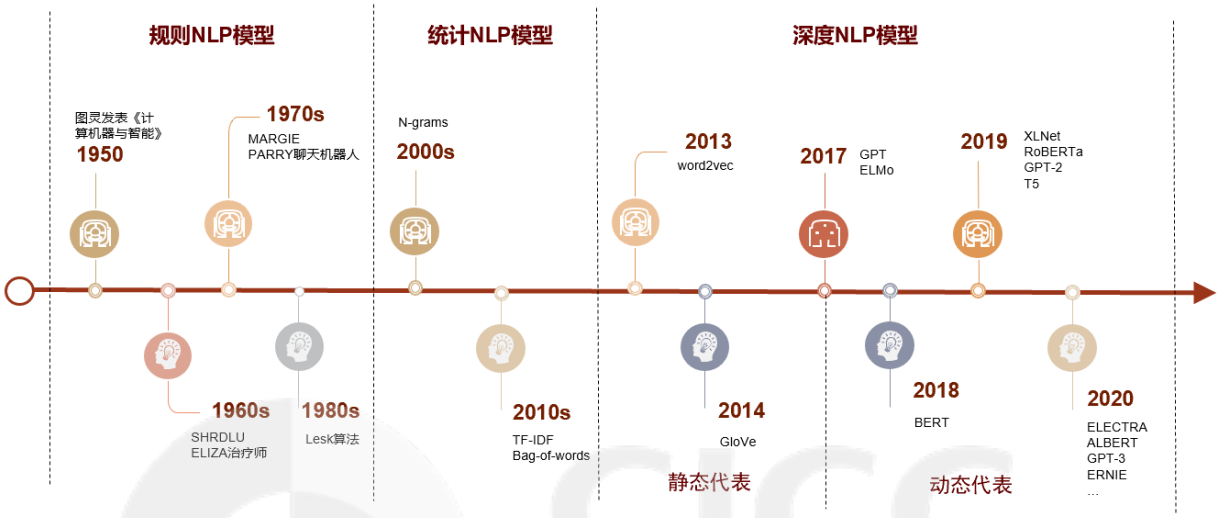
另类数据中占比较大的一部分为文本数据，例如社交软件、新闻和政府文件数据。为使量化模型取得文本数据中蕴含的信息，首先需要 NLP 模型来讲非结构化数据处理成量化模型可以理解的结构化数据，例如将新闻文本转化成新闻舆情指数。NLP 全称为自然语言处理（Natural Language Processing）,它的目的是尝试从传统模型难以理解的人类语言文本中取得其中包含的信息。

► NLP 结合深度学习模型发展迅速

20 世纪 50 年代到 70 年代自然语言处理主要采用基于规则的方法。70 年代以后随着互联网的

高速发展，自然语言处理思潮由理性主义向经验主义过渡，基于统计的方法逐渐代替了基于规则的方法。从 2008 年到现在，在图像识别和语音识别领域的成果激励下，人们也逐渐开始引入深度学习来做自然语言处理研究。由最初的词向量模型到 2013 年 word2vec 再到 2018 年的 BERT，将深度学习与自然语言处理的结合推向了高潮，并在机器翻译、问答系统、阅读理解等领域取得了一定成功。关于 NLP 模型的特点细节请参考附录 B。

图表 20：NLP 主要模型发展史



资料来源：NLP Institutes，中金公司研究部

- NLP 在量化策略里的直接应用

文本相关性

计算文本相关性是 NLP 中的一类基本任务类型。主要是通过语义识别，计算两文本之间在语义上的相关关系，其中又主要包括两句话意思的相似性，和两句话在上下文中的关联度。LAUREN COHEN[11]通过计算上市公司年报文本在时间序列上的相似性，构造出文本相似性因子和有效的交易策略。

文本舆情

通过对新闻、社交媒体、政府工作文件、基金公告和公司年报数据打分来系统性地判断大量文本的方向。体现出量化策略的优势，传统分析师没有覆盖到的标的，没有精力去逐个分析的文本文件，隐藏着宏观和微观的大量信息，如何将其中的信息有效提取是量化分析师面临的主要问题。在下一章节我们展示了 NLP 技术在另类投资策略中的一些应用场景。

[11] LAUREN COHEN. 2020. “Lazy Prices”. Journal of Finance

投资应用场景翻新，量化守正出奇

另类数据的不断出现和科技的发展，催生了各种类别的另类模型的发展，而另类模型和另类数据的结合所发掘的信息增量，将在量化策略中大有可为。

文本数据被动投资：主题构建

使用文本分析的主题投资更加直观，补充了从文本主题到投资标的中间缺失的重要一环。传统的主题投资主要是使用人工对上市公司股票打相应的标签，并不定时手动更新。但使用机器学习模型和上市公司的年报数据，我们可以将上市公司的主营业务的关键词全部有效提取出来，并根据其重要性排序，进而得到一个上市公司标的对于相关主题的重要性。

在相关热点新闻出现后，如“东数西算”、“元宇宙”概念出现，即可使用主题词组与所有上市公司的主营业务高频词汇计算相关系数，构造出主营业务最贴近相关主题的股票投资组合。

由于年报相关披露章节不固定，需要根据年报披露年份抓取不同部分文本。

图表 21：上市公司年报章节内容有所调整

	2012年修订	2014年修订	2015年修订	2016年修订	2017年修订	2021年修订
施行时间	2013年1月1日	2014年5月28日	2015年11月9日	2016年12月9日	2017年12月26日	2021年6月28日
对应财报公布年份	2013年	2015年	2016年	2017年	2018年	2022年
第一节	重要提示、目录和释义	重要提示、目录和释义	重要提示、目录和释义	重要提示、目录和释义	重要提示、目录和释义	重要提示、目录和释义
第二节	公司简介	公司简介	公司简介和主要财务指标	公司简介和主要财务指标	公司简介和主要财务指标	公司简介和主要财务指标
第三节	会计数据和财务指标摘要	会计数据和财务指标摘要	公司业务概要	公司业务概要	公司业务概要	管理层讨论与分析
第四节	董事会报告	董事会报告	管理层讨论与分析	经营情况讨论与分析	经营情况讨论与分析	公司治理
第五节	重要事项	重要事项	重要事项	重要事项	重要事项	环境和社会责任
第六节	股份变动及股东情况	股份变动及股东情况	股份变动及股东情况	股份变动及股东情况	股份变动及股东情况	重要事项
第七节	董事、监事、高级管理人员和员工情况	优先股相关情况	优先股相关情况	优先股相关情况	优先股相关情况	股份变动及股东情况
第八节	公司治理	董事、监事、高级管理人员和员工情况	董事、监事、高级管理人员和员工情况	董事、监事、高级管理人员和员工情况	董事、监事、高级管理人员和员工情况	优先股相关情况
第九节	内部控制	公司治理	公司治理	公司治理	公司治理	债券相关情况
第十节	财务报告	内部控制	财务报告	公司债券相关情况	公司债券相关情况	财务报告
第十一节	备查文件目录	财务报告	备查文件目录	财务报告	财务报告	
第十二节		备查文件目录		备查文件目录	备查文件目录	

资料来源：中国证券监督管理委员会，中金公司研究部

以山西焦化和宁德时代年报内容中抓取出的主营业务关键词组成词云实例：

图表 22：山西焦化 2016 年主营业务词云图

图表 23：宁德时代 2018 年主营业务词云图

资料来源：山西焦化 2016 年公司年报，宁德时代 2018 年公司年报，中金公司研究部

计算主营业务关键词与市场热词词组之间的相关性，并将所有词组与上市公司相关性加总，得到该市场热点与上市公司的相关性，并可以据此计算哪些上市公司与市场热点主题最相关，并根据最相关上市公司标的组成市场主题投资组合。

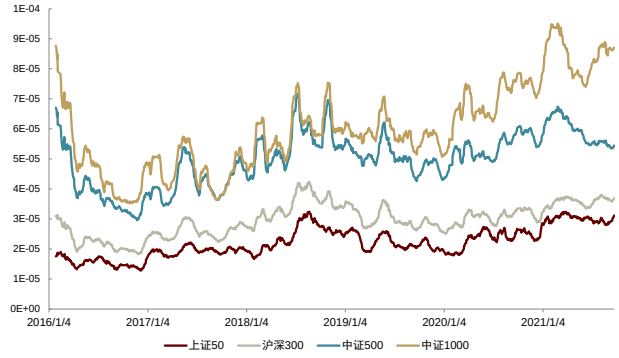
高频量化数据应用：市场微观结构观察与选股 alpha 挖掘

高频因子由 tick 级别的交易数据衍生出来，由于这类数据每秒钟都可能出现大量买单或卖单，所以数据总体体量会比常规数据要高几十甚至几百倍，这其中也包含了更多信息可供挖掘。在之前的微观指标构建的报告《市场微观结构系列（2）： 高频视角下的微观流动性与波动性》中，我们定义了一些微观流动性指标和波动性指标用以对市场微观特征进行跟踪观测（具体微观特征构建详见文末附录 C）。

我们以 2016 年 1 月至 2021 年 9 月的 A 股股票 tick 数据计算所有股票的微观特征指数，并结合主流指数的成分股权重数据，构造主流宽基指数及行业指数的微观流动性特征。通过观察主流宽基指数在各个不同维度的微观流动性特征指标时间序列变化，可以得出结论：小盘股指数流动性总体低于大盘股，而 2021 年开始，中证 500 高频流动性持续改善。而从具体的细分指标来看：

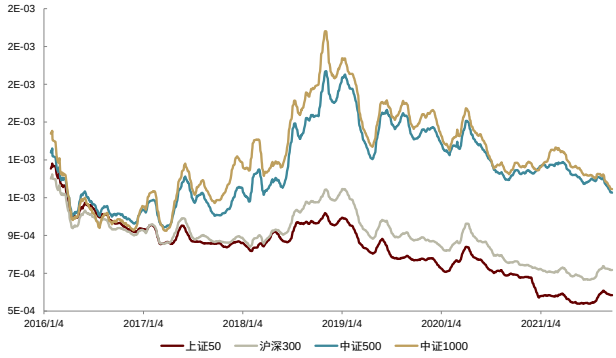
- 高频弹性指标在中证 500 和中证 1000 在 2020 年初前走势基本相同，但 2021 年后开始分化，中证 1000 指数在弹性维度的流动性特征有较明显的恶化。
- 高频宽度指标方面，中证 500 与中证 1000 指数从 2017 年 4 月开始在宽度维度流动性上明显削弱，并与沪深 300 和上证 50 发生明显分化，这种分化程度在 2018 年 10 月份达到最大值。此后慢慢降低到 2019 年 5 月进入局部低值，并进入到流动性不断改善的宽度长期下行阶段。而上证 50 和沪深 300 则是呈现长期缓慢下行的形态。
- 不同深度指标之间走势基本一致：2016 到 2018 年上证 50 与沪深 300 和中证 500 与中证 1000 之间的差距最大，在此期间前者展现出稳定高于后者的流动性特征，在 2018 年之后两者逐渐趋于一致直到 2021 年年初，此后中证 500 的市场深度特征逐渐超过上证 50 和沪深 300 的市场深度，表现出流动性方面的显著改善。

图表 24：主流宽基指数弹性均值



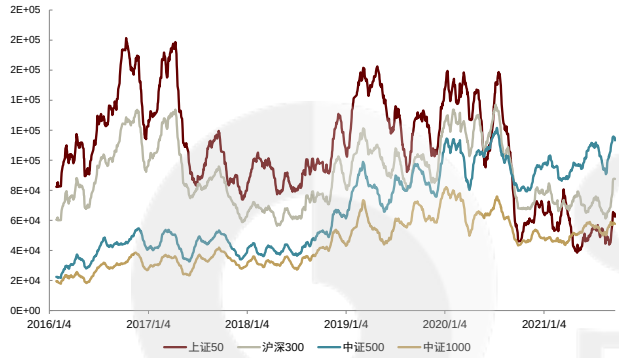
资料来源：万得资讯，中金公司研究部

图表 25：主流宽基指数宽度均值



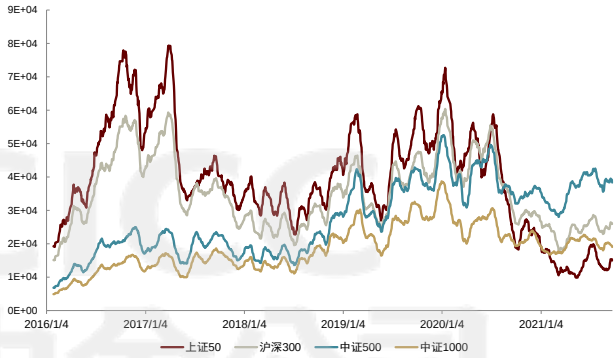
资料来源：万得资讯，中金公司研究部

图表 26：主流宽基指数平均深度均值



资料来源：万得资讯，中金公司研究部

图表 27：主流宽基指数有效深度均值



资料来源：万得资讯，中金公司研究部

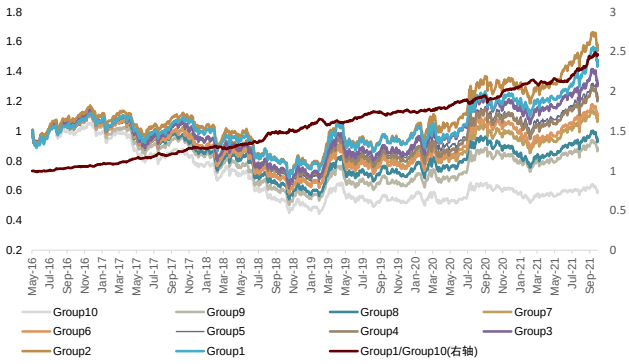
更多细节请参考《市场微观结构系列（2）： 高频视角下的微观流动性与波动性》。

除了利用高频量化用以刻画市场特征，也可以从中挖掘 alpha 信息。通过 tick 数据中的 level2 数据，构建**盘口深度因子。**在 2016 年以来盘口深度因子 IC 序列较为稳定，在中证 1000、中证 500 与沪深 300 股票池内，ICIR 分别达 0.76，0.26 与 0.36。

图表 28：盘口深度因子历史 IC 序列

资料来源：万得资讯，中金公司研究部（统计期：2016-02-01 至 2021-09-30）

图表 29：盘口深度因子分组收益表现



资料来源：万得资讯，中金公司研究部（统计期：2016-05-01 至 2021-09-30）

更多细节请参考《市场微观结构系列（3）：微观特征因子及其策略应用》。

新闻舆情数据应用：文本数据中的 alpha

新闻舆情指数是从新闻数据中抓取的文本中使用 NLP 技术构建出的情绪指数。从与上市公司相关的新闻中可以分析出该新闻对于上市公司的情绪，将其作为对于股价影响的参考。

本文使用原始数据为数库从大型金融新闻网站获取所有新闻并按照正向、负向和中性对每条新闻进行情绪分类，并将其映射到相关上市公司。本次因子测试主要使用加权到单日单个上市公司的新闻舆情指数。数据示例如下表：

图表 30：数库新闻数据样本

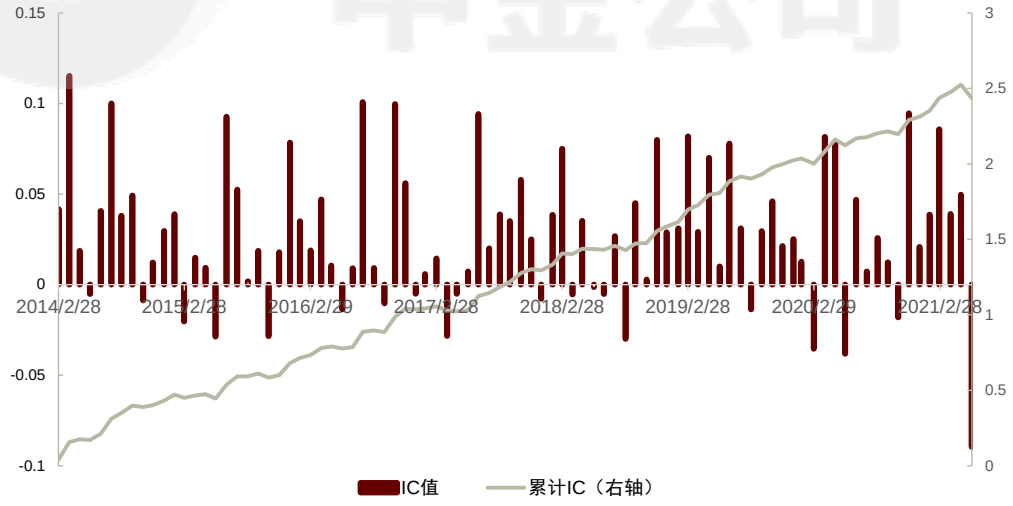
stockCode	companyId	chineseName	englishName	newsId	newsTs	relevance	emotionIndicator	emotionWeight	emotionDetail
002707_SZ_EQ	CSF0000195868	众信旅游	UTOUR TRAVEL	2452395	2014-01-01T21:08:44+0800	0.8333	1	0.97	{0=0.0309, 1=0.9668, 2=0.0023}
600369_SH_EQ	CSF0000001229	西南证券	SOUTHWEST SECURITIES	2452400	2014-01-01T21:16:46+0800	0.0162	0	0.7	{0=0.696, 1=0.2698, 2=0.0342}
000950_SZ_EQ	CSF0000002206	重药控股	C.Q. Pharmaceutical	2452400	2014-01-01T21:16:46+0800	0.75	2	0.9	{0=0.0974, 1=0.0034, 2=0.8992}
002261_SZ_EQ	CSF0000000761	拓维信息	Talkweb Information	2452423	2014-01-01T21:30:34+0800	1	2	0.9	{0=0.0927, 1=0.0071, 2=0.9002}
300362_SZ_EQ	CSF0000003035	天翔环境	Techcent Environment	2452431	2014-01-01T21:27:52+0800	0.6875	1	0.65	{0=0.3422, 1=0.6541, 2=0.0037}
601179_SH_EQ	CSF0000195699	中国西电	XD Electric	2452453	2014-01-01T21:36:46+0800	0.75	1	0.88	{0=0.0989, 1=0.8835, 2=0.0177}

资料来源：数库，中金公司研究部

获取公司舆情指数需要 NLP 的技术手段，但如何使用上市公司的舆情数据更为关键。以下为新闻动量因子构造方法：将上市公司过去一个月内有新闻的交易日的收益率综合计算作为当月新闻动量因子，并将其作行业与市值中性化处理。其原理是投资者是有限理性的，他们的对于信息的认知能力有限，而导致市场反应不足。并且金融分析师也会根据新闻信息调整他们对于个股的收益预测，但这个过程通常会存在几天的时差，这就会导致这些信息所导致的股价变动也存在一定程度上的时间滞后，成为新闻动量效应的驱动力之一。

经过市值和行业中性化之后的新闻动量的因子测试序列如下：

图表 31：新闻动量因子 IC 序列测试

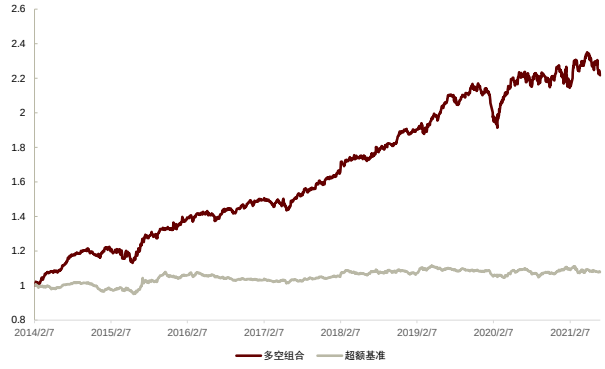


资料来源：数库，中金公司研究部

除 2021 年 6 月份 IC 值负值较大外，大部分月份 IC 值都稳定为正，累计 IC 值也呈现一条稳定上升的折线。上图 IC 均值为 2.99%，波动率为 4.12%，ICIR 为 0.73，胜率为 73.86%。

通过构建多空策略，买入因子得分高的一组，卖出因子得分低的一组，得到最近 2014 年以来的收益曲线如图。最近八年多空策略的年化收益为 11.17%，Sharpe 比例为 1.4，最大回撤为 11.81%，胜率达 67.05%。

图表 32：单因子多空策略净值



图表 33：多空策略分年份表现

	年化收益	年化波动率	夏普比率	最大回撤	月度胜率
2014	19.8%	0.06	3.12	4.4%	80.0%
2015	12.9%	0.10	1.26	7.6%	50.0%
2016	8.9%	0.06	1.54	4.1%	83.3%
2017	8.8%	0.05	1.64	4.7%	75.0%
2018	14.4%	0.06	2.49	1.9%	75.0%
2019	13.2%	0.06	2.21	3.2%	75.0%
2020	4.3%	0.10	0.41	10.7%	50.0%
2021	0.1%	0.12	0.01	5.7%	33.3%
综合	11.2%	0.08	1.40	11.8%	67.0%

资料来源：数库，中金公司研究部

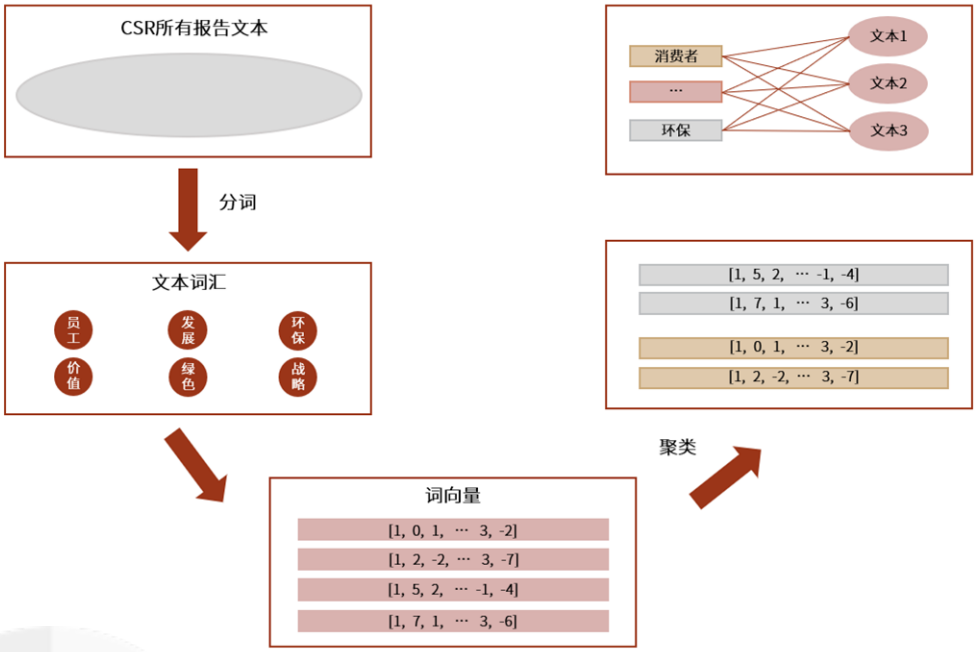
另类数据 ESG 策略

ESG（Environment、Social、Governance）可持续社会责任投资概念是使用非结构化数据的另一典型应用场景。对于上市公司的环保情况、社会责任和公司治理目前来说都不存在类似财务报表的标准化披露指标，所以对于上市公司在 ESG 方面的评价很大程度上是依赖于如污染排放量、岗位提供数量、员工离职率等非结构化的数据进行评估。而不同地区和不同行业的相关指标可比性有较大差异，例如重工业公司的污染排放和金融企业的污染排放量显然是不可比的，如何构建出一套一定程度上标准化的测评框架是当期 ESG 工作的难点之一。

在上述非结构化数据之外，企业本身也会像年报一样披露企业社会责任报告，挖掘企业社会责任报告中的有效信息也能在一定程度上帮助对于一家企业 ESG 计划的评估。

为了获得上市公司在 CSR 报告中对 ESG 的重视程度指标，我们采用 NLP（自然语言处理）模型对 CSR 报告的文本进行量化分析，NLP 建模分析步骤如下图所示：

图表 34：利用企业 CSR 报告计算 ESG 重视程度指标



资料来源：中金公司研究部

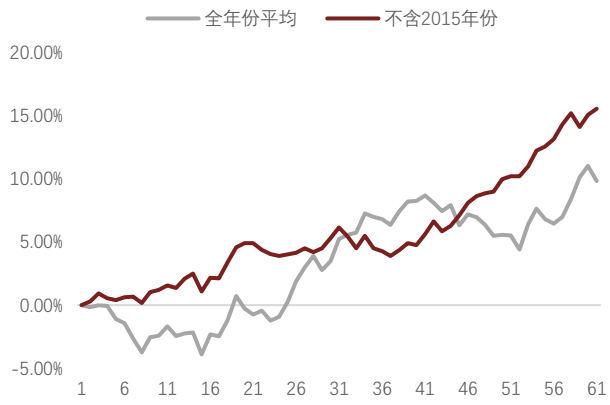
CSR 报告的 ESG 重视程度指标存在短期超额收益。我们使用 CSR 报告的 ESG 重视程度得分进行排序，以得分排名前 20%和后 20%构建出多空组合进行分析。总体来看 ESG 重视程度更高的组合大多具有更高的短期收益，CSR 报告 ESG 重视程度与短期股价存在正相关。从 CSR 报告发布 60 个交易日内的收益表现来看，ESG 高分组相比 ESG 低分组的收益优势较明显。这种收益优势在除 2015 年之外年份均稳定存在，且在 2016 年和 2017 年最为明显。不仅如此，我们发现 CSR 报告发布的事件对股价的影响力有逐渐提高的趋势，说明投资者对于 CSR 报告的关注度在提升。

图表 35：公司 ESG 重视程度指标多空收益

	5 日	10 日	30 日	60 日
2011	0.83%	1.34%	2.04%	3.40%
2012	-0.65%	-1.49%	-0.28%	1.77%
2013	-0.27%	0.91%	0.51%	0.38%
2014	0.56%	0.58%	0.29%	1.05%
2015	-2.07%	-3.23%	-0.90%	-5.72%
2016	0.29%	0.99%	1.84%	5.32%
2017	-0.13%	-0.77%	1.74%	3.62%
2018	-0.09%	1.44%	2.62%	3.24%
2019	0.25%	2.49%	2.17%	2.44%
2020	-0.57%	-0.44%	0.39%	-0.63%

资料来源：万得资讯，中金公司研究部

图表 36：公司 ESG 重视程度指标多空收益曲线



资料来源：万得资讯，中金公司研究部

更多细节请关注报告《ESG 投资系列（2）： ESG 与企业经营、企业价值》。

附录

A. 常见机器学习模型特点

图表 37：常见机器学习模型特点-上

	基础原理	优点	缺点	应用场景
线性回归	最小二乘法	思想简单，实现容易。建模迅速，对于小数据量、简单的关系更有效 线性回归模型十分容易理解，结果具有很好的可解释性，有利于决策分析	对于非线性数据或者数据特征间具有相关性多项式回归难以建模 难以很好地表达高度复杂的数据	因子模型 ARIMA GARCH
Logistic Regression	线性回归+Sigmoid 函数（非线形）映射	实现简单 分类时计算量非常小，速度很快，存储资源低 便利的观测样本概率分数 结合正则化允许多重共线性 计算代价不高，易于理解和实现	当特征空间很大时，逻辑回归的性能一般 容易欠拟合，一般准确度不太高 不能很好地处理大量多类特征或变量 只能处理两分类问题，且必须线性可分 对于非线性特征，需要进行转换	信用评估 预测收益 风险事件预测 经济预测
朴素贝叶斯	结合先验概率和后验概率，假设目标值互相独立	数学基础强，分类效率稳定 对超大规模的训练集速度较快 对小规模的数据表现很好，能个处理多分类任务，适合增量式训练 对缺失数据不太敏感，算法较简单 朴素贝叶斯对结果解释较容易	需要计算先验概率 对输入数据的形式敏感 样本属性有关联时效果不好	文本分类 文本情感分析
KNN	采用向量空间模型分类，输入包含特征空间（Feature Space）中的k个最接近的训练样本	理论成熟，思想简单，既可以用来做分类也可以用来做回归 可用于非线性分类 对数据没有假设，准确度高，对奇异值不敏感 新数据可以直接加入数据集而不必进行重新训练	样本不平衡问题效果差； 对于样本容量大的数据集计算量比较大，需要大量内存 样本不平衡时，预测偏差比较大 KNN每一次分类都会重新进行一次全局运算 k值大小的选择没有理论选择最优	文本分类 模式识别 聚类分析 多分类问题
SVM	对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面（maximum-margin hyperplane）	可以解决高维问题 解决小样本下机器学习问题 能够处理非线性特征的相互作用 无局部极小值问题 无需依赖整个数据集 泛化能力较强	当观测样本很多时，效率不高 对非线性问题没有通用解决方案，找到一个合适的核函数较难 对于核函数的高维映射解释力不强，尤其是径向基函数 常规SVM只支持二分类 对缺失数据敏感	涨跌分类 违约分类
决策树	一颗由多个判断节点组成的树。 在树的每个节点做参数判断，进而在树的最末枝(叶结点)能够对所关心变量的取值作出最佳判断	决策树易于理解和解释，可以可视化分析，容易提取出规则 可以同时处理标称型和数值型数据 比较适合处理有缺失属性的样本 能够处理不相关的特征 测试数据集时，运行速度比较快 在相对短的时间内能够对大型数据源做出可行且效果良好的结果	容易发生过拟合 容易忽略数据集中属性的相互关联 对于那些各类别样本数量不一致的数据，在决策树中，进行属性划分时，不同的判定准则会带来不同的属性选择倾向 ID3算法计算信息增益时结果偏向数值比较多的特征	投资决策 公司信用判断
随机森林	随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个体树输出的类别的众数而定	随机森林具有防止过拟合能力，精度比大多数单个算法要好 随机森林分类器可以处理缺失值 在训练过程中，能够检测到特征间的互相影响，计算特征的重要性 每棵树可以独立、同时生成，容易并行化 具有一定的特征选择能力	随机森林已经被证明在某些噪音较大的分类或回归问题上会过拟合 对于有不同取值的属性的数据，取值划分较多的属性会对随机森林产生更大的影响，所以随机森林在这种数据上产出的属性权值不可信	信用分类问题 收益率预测
Adaboosting	Adaboost是模型为加法模型，学习算法为前向分步学习算法，损失函数为指数函数的算法	高精度的分类器 可以使用各种方法构建子分类器，Adaboost算法提供的是框架 当使用简单分类器时，计算出的结果是可以理解的 弱分类器的构造简单 不用做特征筛选 不易发生过拟合	对outlier比较敏感 AdaBoost迭代次数不好设定，可以使用交叉验证来进行确定 数据不平衡导致分类精度下降 训练比较耗时，每次重新选择当前分类器最好切分点	市场模式识别 收益率预测
GBDT	每一次建立模型是在之前建立模型损失函数的梯度下降方向 弱学习器限定只能使用CART回归树模型	GBDT属于强分类器，一般情况下比逻辑回归和决策树预测精度要高 GBDT可以自己选择损失函数，当损失函数为指数函数时，GBDT变为Adaboost算法 GBDT可以做特征组合，往往在此基础上和其他分类器进行配合	弱学习器之间存在依赖关系，难以并行训练数据 和其他树模型一样，不适合高维稀疏特征	小特征空间 收益率预测
XGBoost	对梯度提升算法的改进，求解损失函数极值时使用了牛顿法，将损失函数泰勒展开到二阶，损失函数中加入了正则化项	对特征值进行提前排序，实现一定程度的并行运算，提升运算速度	只适合处理结构化数据，不适合处理超高维特征数据	小特征空间 收益率预测
LightGBM	基于GBDT算法的框架，支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率、支持分布式可以快速处理海量数据	更快的训练速度和更高的效率 更低的内存占用 由于在训练时间上的缩减，具有处理大数据的能力 支持并行学习	可能会长出比较深的决策树，产生过拟合 由于LightGBM是基于偏差的算法，所以会对噪点较为敏感 在寻找最优解时，依据的是最优切分变量，没有将最优解是全部特征的综合这一可能考虑进去	大特征空间 信用基差预测 收益率预测
神经网络	模仿了生物神经元信号相互传递的方式。 人工神经网络(ANN)由节点层组成，包含一个输入层、一个或多个隐藏层和一个输出层	分类的准确度高 并行分布处理能力强,分布存储及学习能力强 对噪声神经有较强的鲁棒性和容错能力 具备联想记忆的功能，能充分逼近复杂的非线性关系	神经网络需要大量的参数，如网络拓扑结构、权值和阈值的初始值 黑盒过程，不能观察之间的学习过程，输出结果难以解释，影响到结果的可信度 学习时间过长，容易过拟合而且易陷入局部最优	信用违约 自然语言处理 衍生品定价

资料来源：Iqbal H，中金公司研究部

图表 38：常见机器学习模型特点-下

	基础原理	优点	缺点	应用场景
CNN	由一个或多个卷积层和顶端的全连通层（对应经典的神经网络）组成，同时也包括关联权重和池化层（pooling layer）。这一结构使得卷积神经网络能够利用输入数据的二维结构	重共享策略减少了需要训练的参数，相同的权值可以让滤波器不受信号位置的影响来检测信号的特性，使得训练出来的模型的泛华能力更强 池化运算可以降低网络的空间分辨率，从而消除信号的微小偏移和扭曲，从而对输入数据的平移不变性要求不高	容易出现梯度消散	图像处理检测分类 目标检测
RNN	基于神经网络，其中节点之间的连接形成一个有向图沿着序列，允许展示时间序列的时间动态行为	模型是时间维度上的深度模型，可以对序列内容建模	需要训练的参数多，容易造成梯度消散或梯度爆炸问题 不具有特征学习能力 长时依赖问题（Long-Term Dependencies）	自然语言处理 语音识别
LSTM	基于RNN,通过遗忘门和输出门忘记部分信息来解决梯度消失的问题	通过遗忘门和输出门忘记部分信息来解决梯度消失的问题	信息在过远的距离传播中损失很大 无法很好地并行	大数据量时间序列
K-Means	基于欧式距离的聚类算法，其认为两个目标的距离越近，相似度越大	算法简单，容易实现，速度很快 对处理大数据集，该算法是相对可伸缩的和高效率的，因为它的复杂度大约是O(nkt)，其中n是所有对象的数目，k是簇的数目,t是迭代的次数。通常k<<n。这个算法通常局部收敛 当簇是密集的、球状或团状的，且簇与簇之间区别明显时，聚类效果较好	对数据类型要求较高，适合数值型数据 可能收敛到局部最小值，在大规模数据上收敛较慢 分组的数目k是一个输入参数，不合适的k可能返回较差的结果 对初值的簇心值敏感，对于不同的初始值，可能会导致不同的聚类结果 不适合于发现非凸面形状的簇，或者大小差别很大的簇 对于"噪声"和孤立点数据敏感，少量的该类数据能够对平均值产生极大影响	无标记的文本分类
EM最大期望算法	通过迭代进行极大似然估计（Maximum Likelihood Estimation, MLE）的优化算法，通常作为牛顿迭代法（Newton-Raphson method）的替代用于对包含隐变量或缺失数据的概率模型进行参数估计	比K-means算法计算结果稳定、准确	比K-means算法计算复杂，收敛也较慢 不适于大规模数据集和高维数据	低维特征空间 小规模数据聚类
密度聚类	假设聚类结构能通过样本分布的紧密程度确定。通常情形下，密度聚类算法从样本密度的角度来考察样本之间的可连接性，并基于可连接样本不断扩展聚类簇以获得最终的聚类结果	可以对任意形状的稠密数据集进行聚类，相对的K均值之类的聚类算法一般只适用于凸数据集 可以在聚类的同时发现异常点，对数据集中的异常点不敏感 初始值对聚类结果影响不大	如果样本集的密度不均匀、聚类间距差相差很大时，聚类质量较差 如果样本集较大时，聚类收敛时间较长，此时可以对搜索最近邻时建立的KD树或者球树进行规模限制来改进 调参相对于传统的K-Means之类的聚类算法稍复杂，不同的参数组合对最后的聚类效果有较大影响	文本数据分类
层次聚类	通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点	距离和规则的相似度容易定义，限制少 不需要预先制定聚类数 可以发现类的层次关系 可以聚类成其它形状	计算复杂度太高 对奇异值敏感 算法很可能聚类成链状	产业链聚类
GRU	基于RNN,取消了LSTM中的cell state，只使用了hidden state,并且使用更新门来替换LSTM中的输入们和遗忘门，取消了LSTM中的输出门，新增了重置门	GRU的参数量少，减少过拟合的风险	学习效率不高	小数据量时间序列
RBM	一种无向图，属于生成式随机双层神经网络，该网络由可见单元和隐藏单元构成，可见变量和隐藏变量都是二元变量	能够处理缺失/不规则数据 不需要数据的标签信息	计算时间长 对抽样噪音敏感	降维 分类 特征学习 主题建模
GAN	一种生成模型，模型包括生成器（Generator）和判别器（Discriminator）两个网络，训练的过程用到了二人零和博弈的思想	能更好建模数据分布 理论上，GANs 能训练任何一种生成器网络。其他的框架需要生成器网络有一些特定的函数形式，比如输出层是高斯的 无需利用马尔科夫链反复采样，无需在学习过程中进行推断	难训练，不稳定。生成器和判别器之间需要很好的同步，但是在实际训练中很容易D收敛，G发散 GANs的学习过程可能出现模式缺失，生成器开始退化，重复生成同样的样本点，无法继续学习	自动提取特征 自动判断和优化
自编码器	一种利用反向传播算法使得输出值等于输入值的神经网络，它先将输入压缩成潜在空间表征，然后通过这种表征来重构输出	泛化性强 无监督不需要数据标注 训练速度快	需要大量清洁数据 信息损失较大	异常监测 数据去噪 数据降维
强化学习	描述和解决智能体（agent）在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题	不需要带标签的输入输出对 无需对非最优解的精确地纠正 擅长实现长期目标	缺乏可扩展性 需要大量数据 计算资源需求大	组合构建 高频交易 策略优化

资料来源：Iqbal H，中金公司研究部

B. 深度 NLP 模型特点一览

图表 39：NLP 深度学习

模型类别	模型名称	应用
基础嵌入模型	NNLM	词向量转换
	word2vec	
	Glove	
	BPE	
	n-gram	
	COVE	
	ELME	
	GPT	
	BERT	
	skip thought	句向量转换
	quick thought	
	doc2vec	文档向量转换
	FastText	文本分类
基于CNN模型	TextCNN	文本分类
	DCNN	语义分割
	VDCNN	文本分类、加深网络
	GCNN	文本分类、引进门机制
	TCN	语言模型
基于RNN模型	RNN	语言模型
	LSTM/GRU	加GATE机制的RNN
	Bi-LSTM	词性标注、命名实体识别、情感分析
	RCNN	基于CNN的RNN
	RNN/LSTM seq2seq	机器翻译
	树模型LSTM	情感分析
	栈LSTM	句法分析
	基于RNN的VAE	文本生成
	基于LSTM-CNN的GAN	CNN作为分类器取分真是样本与生产样本
	DMN	词性标注、动态储存网络
	Bi-LSTM-CRF	词性标注、基于LSTM建模
	Seq2Seq+注意力机制	机器翻译、语音识别
	ATAAE-LSTM	情感分析
注意力机制模型	ABCNN	文本分类
	Transformer	语音识别
	GPT	文本生成
基于Transformer模型	BERT	基于Transformer的预训练模型
	XLM	基于BERT的跨语言模型,机器翻译

资料来源：NLP Institutes，中金公司研究部

C. 高频微观流动性指标汇总表

图表 40：高频微观流动性指标汇总表

指标类型	指标名称	计算公式	指标逻辑
宽度	盘口价差 spread	$\frac{2*(a_1-b_1)}{(a_1+b_1)}$	使用日内tick级别数据计算买一价与卖一价的价差再除以买一价和卖一价的算数平均值。其基本逻辑是买一卖一的价差越大，其市场宽度越大，流动性越差，和流动性为负相关； 将该指标的日内高频值取标准差以体现该指标在当日均匀程度，和流动性为负相关；
弹性	弹性 resiliency	$\frac{high-low}{turnover}$	使用日内tick级别数据计算每个tick最高价与最低价的差再除以换手率的值，反映了当日价格弹性。其基本逻辑是当最高价与最低价价差越大时，股价整体弹性越大，流动性越弱。当换手率越高时，整体弹性越小，流动性越强。以日内所有tick的价格弹性序列为样本，计算价格弹性均值体现整体流动性水平，计算价格弹性标准差体现当日流动性在日内的均匀程度，与流动性为负相关； 将该指标的日内高频值取标准差以体现该指标在当日均匀程度，和流动性为负相关；
深度	卖盘深度 ask_depth	$\sum\lvert\frac{av_i*mid_{price}}{a_i-last_{mid}+\epsilon}\rvert$	使用日内tick级别数据，使用五档买单与当下时刻中间价格的价差的倒数对买单的数量进行加权，分别计算买单的市场深度指标。当单边所有买单档位为空时，按照定义规定该边市场深度为0。该指标反映单边市场深度。其基本逻辑是五档买单报价中每一档当离中间价格越远时，它的交易量在它所代表的市场深度权重应该越小。将该指标的日内高频值简单平均，和流动性为正相关； 将该指标的日内高频值取标准差以体现该指标在当日均匀程度，和流动性为负相关；
	买盘深度 bid_depth	$\sum\lvert\frac{bv_i*mid_{price}}{b_i-last_{mid}+\epsilon}\rvert$	使用日内tick级别数据，使用五档卖单与当下时刻中间价格的价差的倒数对卖单的数量进行加权，分别计算卖单的市场深度指标。当所有卖单档位为空时，按照定义规定该边市场深度为0。该指标反映单边市场深度。其基本逻辑是五档卖单报价中每一档当离中间价格越远时，它的交易量在它所代表的市场深度权重应该越小。将该指标的日内高频值简单平均，取绝对值后和流动性为正相关； 将该指标的日内高频值取标准差以体现该指标在当日均匀程度，和流动性为负相关；
	盘口深度 avg_depth	$\frac{(av_1+bv_1)}{2}$	使用日内tick级别数据，计算最优买单与卖单的挂单量的简单平均值，当单边最优挂单量为0时，整体深度为0。该指标反映整体市场深度。其基本逻辑是双边最优价单的挂单量的平均值越高，市场总体深度越大。将该指标的日内高频值简单平均，和流动性为正相关； 将该指标的日内高频值取标准差以体现该指标在当日均匀程度，和流动性为负相关；
	有效深度 effective_depth	$min(av_1,bv_1)$	使用日内tick级别数据，计算最优买单与卖单的挂单量中的较小值，反映市场实际上的有效深度。其基本逻辑是双边最优价单的挂单量较小值越大，市场实际有效深度越大。将该指标的日内高频值简单平均，和流动性为正相关； 将该指标的日内高频值取标准差以体现该指标在当日均匀程度，和流动性为负相关；
交易集中度	开盘集合竞价成交量占比 first_call_ratio	$\frac{\sum_{09:15}^{09:25}(Volume)}{\sum_{09:15}^{15:00}(Volume)}$	使用日内tick级别数据计算上午开盘9：25之前的集合竞价的总交易量占全天交易量的比例，如果该比例越大,说明当日大量成交集中在开盘竞价阶段，流动性分布越不均匀，和流动性负相关；
	收盘集合竞价成交量占比 last_call_ratio	$\frac{\sum_{14:57}^{15:00}(Volume)}{\sum_{09:15}^{15:00}(Volume)}$	使用日内tick级别数据计算下午收盘前14：57-15：00的集合竞价的总交易量占全天交易量的比例，如果该比例越大,说明当日大量成交集中在收盘竞价阶段，流动性分布越不均匀，和流动性负相关；
	卖盘深度前15分钟占比 ask_depth_cct	$\frac{ask\_depth_{15min}}{ask\_depth_{allday}}$	使用日内tick级别数据，计算开盘一段时间（如15分钟）内市场卖盘深度和占交易日所有市场交易深度之和的比例，计算以上四个深度的集中度。其基本逻辑为当市场深度越集中于开盘后15分钟，说明当天流动性分布越不均匀，和流动性负相关；
	买盘深度前15分钟占比 bid_depth_cct	$\frac{bid\_depth_{15min}}{bid\_depth_{allday}}$	使用日内tick级别数据，计算开盘一段时间（如15分钟）内市场买盘深度和占交易日所有市场交易深度之和的比例，计算以上四个深度的集中度。其基本逻辑为当市场深度越集中于开盘后15分钟，说明当天流动性分布越不均匀，和流动性负相关；
	盘口深度前15分钟占比 avg_depth_cct	$\frac{avg\_depth_{15min}}{avg\_depth_{allday}}$	使用日内tick级别数据，计算开盘一段时间（如15分钟）内市场平均盘口深度和占交易日所有市场交易深度之和的比例，计算以上四个深度的集中度。其基本逻辑为当市场深度越集中于开盘后15分钟，说明当天流动性分布越不均匀，和流动性负相关；
	有效深度前15分钟占比 effective_depth_cct	$\frac{effective\_depth_{15min}}{effective\_depth_{allday}}$	使用日内tick级别数据，计算开盘一段时间（如15分钟）内市场有效盘口深度和占交易日所有市场交易深度之和的比例，计算以上四个深度的集中度。其基本逻辑为当市场深度越集中于开盘后15分钟，说明当天流动性分布越不均匀，和流动性负相关；

资料来源：中金公司研究部

参考文献

[1] Eagle Alpha. 2019. “Alternative Data Use Cases Edition 6”

[2] Niall Hurley. 2021. “B2021 Alternative Data Report: Year in Review”. Eagle Alpha

[3] Brad Tingley. 2020. “FISD Alternative Data Council: A Guide to Alternative Data”. Greenwich Associates

[4] David Reinsel. 2018. The Digitization of the World: From Edge to Core. IDC

[5] Alternativedata. 2022. https://alternativedata.org/alternative-data/

[6] Deloitte. 2017. Warming up to collective intelligence investing.

[7] Blackrock. 2015. “The Evolution of Active Investing. Finding Big Alpha in Big Data”

[8] Deloitte US. 2019. “Alternative Data – Perspectives and Insights”

[9] Iqbal H. 2021. “Machine Learning: Algorithms, Real-World Applications and Research Directions”. SN Computer Science

[10] Sophie Emerson. 2019. “Trends and Applications of Machine Learning in Quantitative Finance”. 8th International Conference on Economics and Finance Research

[11] LAUREN COHEN. 2020. “Lazy Prices”. Journal of Finance

作者信息

胡骥聪
SAC 执证编号：S0080521010007
SFC CE Ref：BRF083
jicong.hu@cicc.com.cn

郑文才
SAC 执证编号：S0080121120041
wencai3.zheng@cicc.com.cn

周萧潇
SAC 执证编号：S0080521010006
SFC CE Ref：BRA090
xiaoxiao.zhou@cicc.com.cn

刘均伟
SAC 执证编号：S0080520120002
SFC CE Ref：BQR365
junwei.liu@cicc.com.cn

王汉锋
SAC 执证编号：S0080513080002
SFC CE Ref：AND454
hanfeng.wang@cicc.com.cn

法律声明

一般声明

本报告由中国国际金融股份有限公司（已具备中国证监会批复的证券投资咨询业务资格）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但中国国际金融股份有限公司及其关联机构（以下统称“中金公司”）对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供投资者参考之用，不构成对买卖任何证券或其他金融工具的出价或征价或提供任何投资决策建议的服务。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐或投资操作性建议。投资者应当对本报告中的信息和意见进行独立评估，自主审慎做出决策并自行承担风险。投资者在依据本报告涉及的内容进行任何决策前，应同时考量各自的投资目的、财务状况和特定需求，并就相关决策咨询专业顾问的意见对依据或者使用本报告所造成的一切后果，中金公司及/或其关联人员均不承担任何责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断，相关证券或金融工具的价格、价值及收益亦可能会波动。该等意见、评估及预测无需通知即可随时更改。在不同时期，中金公司可能会发出与本报告所载意见、评估及预测不一致的研究报告。

本报告署名分析师可能会不时与中金公司的客户、销售交易人员、其他业务人员或在本报告中针对可能对本报告所涉及的标的证券或其他金融工具的市场价格产生短期影响的催化剂或事件进行交易策略的讨论。这种短期影响的分析可能与分析师已发布的关于相关证券或其他金融工具的目标价、评级、估值、预测等观点相反或不一致，相关的交易策略不同于且也不影响分析师关于其所研究标的证券或其他金融工具的基本面评级或评分。

中金公司的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。中金公司没有将此意见及建议向报告所有接收者进行更新的义务。中金公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见不一致的投资决策。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现。过往的业绩表现亦不应作为日后回报的预示。我们不承诺也不保证，任何所预示的回报会得以实现。分析中所做的预测可能是基于相应的假设。任何假设的变化可能会显著地影响所预测的回报。

本报告提供给某接收人是基于该接收人被认为有能力独立评估投资风险并就投资决策能行使独立判断。投资的独立判断是指，投资决策是投资者自身基于对潜在投资的目标、需求、机会、风险、市场因素及其他投资考虑而独立做出的。

本报告由受香港证券和期货委员会监管的中国国际金融香港证券有限公司（“中金香港”）于香港提供。香港的投资者若有任何关于中金公司研究报告的问题请直接联系中金香港的销售交易代表。本报告作者所持香港证监会牌照的牌照编号已披露在报告首页的作者姓名旁。

本报告由受新加坡金融管理局监管的中国国际金融（新加坡）有限公司（“中金新加坡”）于新加坡向符合新加坡《证券期货法》定义下的认可投资者及/或机构投资者提供。提供本报告于此类投资者，有关财务顾问将无需根据新加坡之《财务顾问法》第 36 条就任何利益及/或其代表就任何证券利益进行披露。有关本报告之任何查询，在新加坡获得本报告的人员可联系中金新加坡销售交易代表。

本报告由受金融服务监管局监管的中国国际金融（英国）有限公司（“中金英国”）于英国提供。本报告有关的投资和服务仅向符合《2000 年金融服务和市场法 2005 年（金融推介）令》第 19（5）条、38 条、47 条以及 49 条规定的人士提供。本报告并未打算提供给零售客户使用。在其他欧洲经济区国家，本报告向被其本国认定为专业投资者（或相当性质）的人士提供。

本报告由中国国际金融日本株式会社（“中金日本”）于日本提供，中金日本是在日本关东财务局（日本关东财务局长（金商）第 3235 号）注册并受日本法律监管的金融机构。本报告有关的投资和服务仅向符合日本《金融商品交易法》第 2 条 31 项所规定的专业投资者提供。本报告并未打算提供给日本非专业投资者使用。

本报告将依据其他国家或地区的法律法规和监管要求于该国家或地区提供。

特别声明

在法律许可的情况下，中金公司可能与本报告中提及公司正在建立或争取建立业务关系或服务关系。因此，投资者应当考虑到中金公司及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。

与本报告所含具体公司相关的披露信息请访 https://research.cicc.com/footer/disclosures，亦可参见近期已发布的关于该等公司的具体研究报告。

中金研究基本评级体系说明：

分析师采用相对评级体系，股票评级分为跑赢行业、中性、跑输行业（定义见下文）。

除了股票评级外，中金公司对覆盖行业的未来市场表现提供行业评级观点，行业评级分为超配、标配、低配（定义见下文）。

我们在此提醒您，中金公司对研究覆盖的股票不提供买入、卖出评级。跑赢行业、跑输行业不等同于买入、卖出。投资者应仔细阅读中金公司研究报告中的所有评级定义。请投资者仔细阅读研究报告全文，以获取比较完整的观点与信息，不应仅仅依靠评级来推断结论。在任何情形下，评级（或研究观点）都不应被视为或作为投资建议。投资者买卖证券或其他金融产品的决定应基于自身实际具体情况（比如当前的持仓结构）及其他需要考虑的因素。

股票评级定义：

- 跑赢行业（OUTPERFORM）：未来 6~12 个月，分析师预计个股表现超过同期其所属的中金行业指数；
- 中性（NEUTRAL）：未来 6~12 个月，分析师预计个股表现与同期其所属的中金行业指数相比持平；
- 跑输行业（UNDERPERFORM）：未来 6~12 个月，分析师预计个股表现不及同期其所属的中金行业指数。

行业评级定义：

- 超配（OVERWEIGHT）：未来 6~12 个月，分析师预计某行业会跑赢大盘 10%以上；
- 标配（EQUAL-WEIGHT）：未来 6~12 个月，分析师预计某行业表现与大盘的关系在-10%与 10%之间；
- 低配（UNDERWEIGHT）：未来 6~12 个月，分析师预计某行业会跑输大盘 10%以上。

研究报告评级分布可从https://research.cicc.com/footer/disclosures 获悉。

V190624
编辑：樊荣

北京
中国国际金融股份有限公司
中国北京建国门外大街 1 号
国贸写字楼 2 座 28 层
邮编：100004
电话：(86-10) 6505 1166
传真：(86-10) 6505 1156

上海
中国国际金融股份有限公司上海分公司
上海市浦东新区陆家嘴环路 1233 号
汇亚大厦 32 层
邮编：200120
电话：(86-21) 5879-6226
传真：(86-21) 5888-8976

深圳
中国国际金融股份有限公司深圳分公司
深圳市福田区益田路 5033 号
平安金融中心 72 层
邮编：518048
电话：(86-755) 8319-5000
传真：(86-755) 8319-9229

香港
中国国际金融（香港）有限公司
香港中环港景街 1 号
国际金融中心第一期 29 楼
电话：(852) 2872-2000
传真：(852) 2872-2100

东京
中国国际金融日本株式会社
〒100-0005 東京都千代田区丸の内３丁目２番３号 丸の内二重橋ビル２ 1 階
Tel: (+813) 3201 6388
Fax: (+813) 3201 6389

旧金山
CICC US Securities, Inc. San Francisco Branch Office
One Embarcadero Center, Suite 2350,
San Francisco, CA 94111, USA
Tel: (+1) 415 493 4120
Fax: (+1) 628 203 8514

纽约
CICC US Securities, Inc
32nd Floor, 280 Park Avenue
New York, NY 10017, USA
Tel: (+1-646) 7948 800
Fax: (+1-646) 7948 801

新加坡
China International Capital Corporation (Singapore) Pte. Limited
6 Battery Road, #33-01
Singapore 049909
Tel: (+65) 6572 1999
Fax: (+65) 6327 1278

伦敦
China International Capital Corporation (UK) Limited
25th Floor, 125 Old Broad Street
London EC2N 1AR, United Kingdom
Tel: (+44-20) 7367 5718
Fax: (+44-20) 7367 5719

法兰克福
China International Capital Corporation (Europe) GmbH
Neue Mainzer Straße 52-58, 60311
Frankfurt a.M, Germany
Tel: (+49-69) 24437 3560