THE FORELAND OF
TRADING TECHNOLOGY

内部资料 免费交流
《准印证》编号沪（K）0671

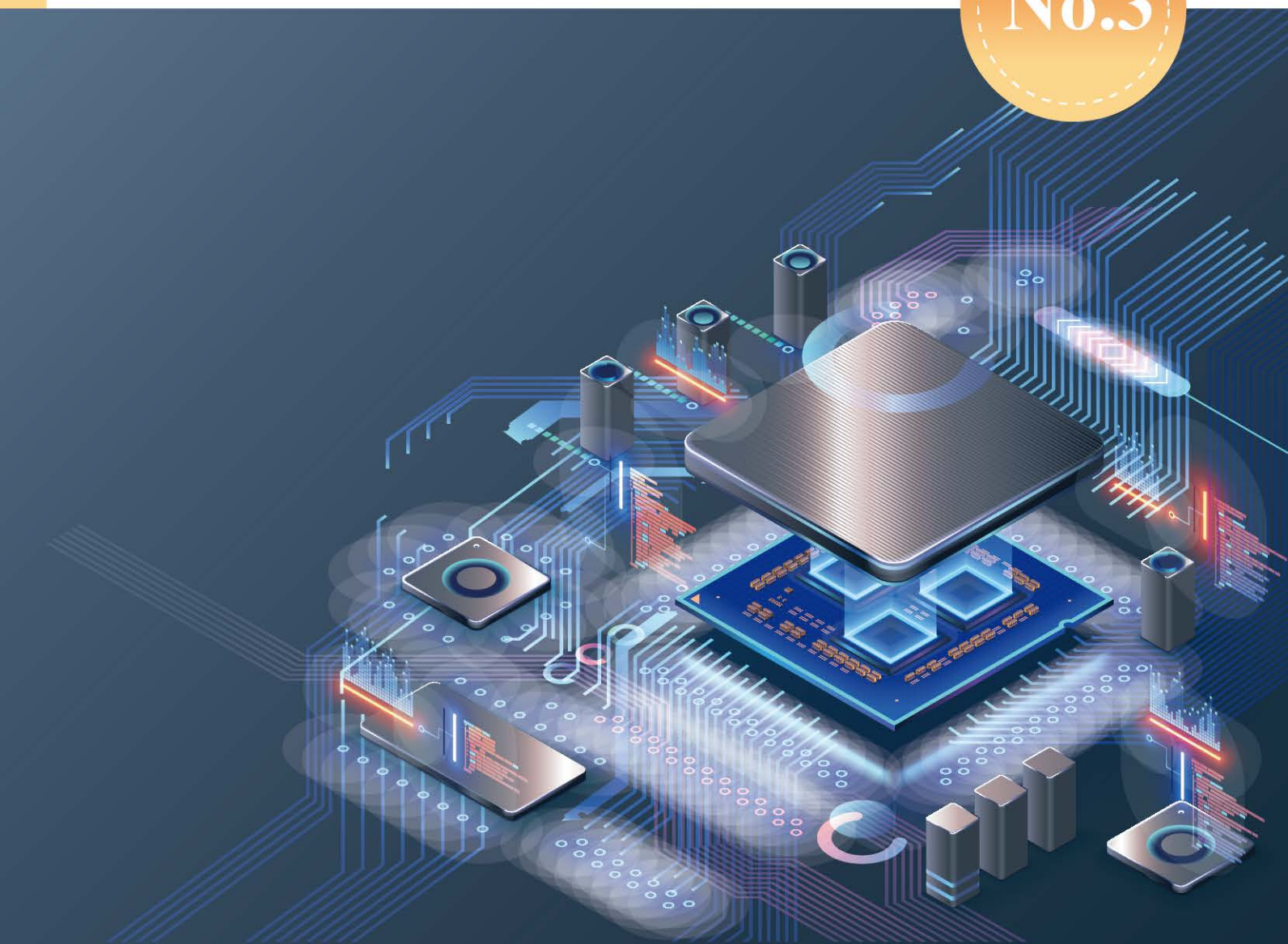
交易技术前沿

2021年 第三期 总第45期

本期主题

人工智能

No.3



上海证券交易所 SHANGHAI STOCK EXCHANGE
ITRDC 证券信息技术研究发展中心（上海）

内部资料 2021 年第三期（总第 45 期）
准印证号：沪（K）0671

NO.3

主管：上海证券交易所
主办：上交所技术有限责任公司
总编：黄红元
副总编：徐毅林
执行主编：王泊
责任编辑：黄俊杰、徐丹、郭望

上海市浦东南路 528 号
邮编：200120
电话：021-68607128, 021-68607131
传真：021-68813188
投稿邮箱：ftt.editor@sse.com.cn

扫码阅览历期杂志

篇首语

近年来，随着通讯、存储、计算能力的大幅度提高，30 年前学术领域的成熟理论成为现实，人工智能在计算机视觉、语言识别、自然语言处理、推荐系统等领域取得了瞩目的突破，并不断向各行业纵深发展。证券期货业乘着 AI 技术浪潮，在智能客服、智能投研、监管合规等业务场景创造新价值，并不断基于自动化方法重塑开发测试运维的传统流程。本期《交易技术前沿》以“人工智能”为主题，收录来自行业十三篇优秀文章，探讨行业技术前沿。

《基于机器学习的期货市场趋势预测方法的研究》立足业务思考，详述了基于历史数据特征和模型融合实现交易、持仓量预测的系统方案。《联邦学习在证券行业的应用初探》探索了联邦学习技术在基金推荐、休眠用户激活等场景的应用，展望了解决行业隐私数据共享的广阔前景。《海通证券智能外呼客服应用研究》分享了智能外呼场景下语义引擎应用实践所面临的痛点及解决方案。《基于私有云的智能灾备中心的实践》叙述了在数据中心异构环境下，对物理机、虚拟机采用温备的方式，在私有云架构中实现统一的灾备建设管理，并通过设立独立灾备演练区，实现传统灾备演练模式由人工向自动化转型的方案。《安信证券容器云平台落地实践分享》探索了云原生容器领域的最佳实践，包括平台架构设计、技术选型、安全防护、运维测试、上云推广和建设成果。《网络传输状态数学建模和实证检验》从分析数据成因入手，结合工程和物理设备对应的概念建立了传输状态模型，并根据实际采样数据进行了实证检验。

在可预见的未来，人工智能技术带来的模式挖掘与自动化能力将持续赋能证券行业，实现更强的信息整合、建模处理与算法分析能力，提升业务服务质量和风险防控能力，不断推动普惠金融、绿色金融发展，对行业产生深远的影响。

《交易技术前沿》编辑部

2021 年 6 月 30 日

目录 Contents

本期热点 Hotspot

实践探索 Exploration

行业观察 Observation

信息资讯采撷 Information

本期热点 Hotspot

1 基于机器学习的期货市场趋势预测方法的研究

2 联邦学习在证券行业的应用初探

3 基于列式存储和交互式数据分析的风险管理平台建设实践

4 海通证券智能外呼客服应用研究

基于机器学习的期货市场趋势预测方法的研究

宁晓冬、杨和国 / 郑州商品交易所
赵颖杰、宫朝辉 / 郑州易盛信息技术有限公司

长期以来，期货市场的稳定运行一直是交易所关注的重点，对于期货品种的功能发挥起到重要作用。合约交易持仓量是期货市场运行的重要指标，也是利用期货管理风险的基础指标。为增强对市场趋势的了解，提高运行预判能力，本文基于合约历史运行规律及风控措施参数，开展数据分析，提取历史数据及风控参数作为输入特征，建立基于多个机器学习算法的融合模型，利用网格搜索方式设置最优参数，进行期货合约未来五日交易量、持仓量的预测。实验结果显示，本文构建的算法模型预测交易量平均准确率接近70%、持仓量平均准确率达到 83%。同时，本文以案例分析的形式证实了融合模型和网格搜索技术对于预测准确率的提升存在显著效果。

一、项目背景

期货交易是现货市场的晴雨表，为商品远期定价提供基准，具有护航实体经济稳健运行的重要意义。期货交易价格由不同参与主体共同报价撮合而成。套期保值者利用市场锁定利润管理价格波动风险，投机者尝试判断行情获取利润。当市场交易过热时，期货价格会失真并偏离现货价格，可能给投资者和套保企业带来损失；反之，当缺乏流动性时，期货价格无法准确反映市场参与者的“共识”。因此，稳定的市场参与度是期货交易合理定价的重要基础。预判市场热度对于调节市场情绪，合理利用风控措施稳定行情至关重要。本文将核心问题定义为预测市场热度，即预测品种的交易持仓情况。通常来说，期货市场交易持仓趋势受到多方面因素的影响，如期货标的价格变化，突发舆情事件，政策影响等。多因素影响下，简单的规则算法难以有效预测交易持仓情况。基于此，本文尝试从多个维度提取有效特征，并利用三个独立的机器学习模型捕捉数据之间的不同关系。最后，利用网格搜索方法将三

个模型的结果进行加权融合输出最终的预测结果。本文组织架构如下：第二章针对历史运行数据开展分析，研究了风控参数、结算价等与交易持仓量的关系；第三章就特征提取及三个单一模型的构建进行了详细介绍；第四章描述了模型融合及权重网格搜索技术；第五章设计实验验证模型有效性并设计方法解释模型结果；第六章为模型可解释性研究；第七章为总结与展望。

二、历史运行分析

机器学习相关问题中，数据分析是整个数据建模的基础，决定了特征提取质量与模型最终效果。数据分析对于深入了解目标问题起到重要作用，并指导模型的迭代构建。本文数据分析涵盖了众多维度，下面挑选四个方面对交易所历史数据简要分析。

（一）主力合约的生命周期（双峰现象）

回溯历史数据，所有合约在挂牌摘牌整个周期中，都会经历交易持仓量逐渐放大随后下降的过程。其中，有较大比例的主力合约（接近40%）在挂牌摘牌的整个周期会呈现“双峰”现象。“双峰”现象即合约在成为主力之后，交易量与持仓量会经历两个峰值，其中交易量尤为明显。我们以图 1、2 分别展示玻璃期货 1705 与棉花期货 1801 两个合约的交易持仓走势。从图中可以看出，虽然玻璃、棉花分属于非农与农两个类别，但是交易量上都呈现出较为典型的双峰形态。该现象产生原因可能一是当合约成为主力合约之后，交易资金会快速流入，导致交易持仓量快速放大；二是前主力合约进入交割摘牌阶段，主力合约因此达到第二个高峰。双峰现象的周期性规律对于我们掌握品种运行规律及预测交易持仓起到指导性作用。

（二）品种交易持仓量与价格关系

为探索交易、持仓量的影响因素，本文着重分析价格波动与交易持仓量之间的关系。交易市场上存在一种“共识”，即认为价格的波动会引

图 1：玻璃 1705 合约交易持仓走势图

图 2：棉花 1801 合约交易持仓走势图

起交易持仓量的放大。因此，本文尝试计算价格波动与交易持仓量变化之间的皮尔森相关系数[1]，研究从长期运行维度来看价格波动是否会实质上引起交易量与持仓量的放大。我们定义如下三个指标：

$$P_{delta}=Abs(P_T-P_{T-N})/P_{T-N}$$

$$V_{delta}=(V_T-V_{T-N})/V_{T-N}$$

$$H_{delta}=(H_T-H_{T-N})/H_{T-N}$$

其中 T 表示当前日期，N 表示时间差，P_T 表示 T 日结算价格，V_T 表示 T 日交易量，H_T 表示 T 日持仓量；相应的，P_{T-N}、V_{T-N}、H_{T-N} 分别表示 T-N 日的对应数值；P_{delta} 表示以 T 日与 T-N 日之间价格波动比例的绝对值，V_{delta} 表示对应的交易量波动比例实际值，H_{delta} 表示持仓量波动比例实际值。我们分别计算当 N 设置为 [1-5] 日时，P_{delta} 与 V_{delta}、H_{delta} 之间的皮尔森系数。实验中，我们选取了 2016-2018 年郑商所已上市的所有品种，并对品种下的全部挂牌合约进行汇总。具体情况见下表。

表 1 中 PD1-HD1 表示当 N 值取 1 时，P_{delta} 与 H_{delta} 之间的关联系数；PD1-VD1 表示当 N 值取 1 时，P_{delta} 与 V_{delta} 之间的关联系数，以此类推。

表 1：价格波动与交易持仓变化相关系数表

品种	PD1_HD1	PD1_VD1	PD2_HD2	PD2_VD2	PD3_HD3	PD3_VD3	PD4_HD4	PD4_VD4	PD5_HD5	PD5_VD5
苹果	0.31	0.16	0.19	0.19	0.07	0.23	0.04	0.25	0.06	0.26
棉花	0.38	0.27	0.43	0.27	0.38	0.24	0.38	0.32	0.35	0.37
棉纱	0.48	0.25	0.57	0.42	0.58	0.52	0.71	0.68	0.64	0.71
玻璃	0.36	0.26	0.33	0.26	0.32	0.24	0.29	0.24	0.27	0.25
粳稻	0.04	0.08	0.06	0.06	0.04	0.01	0.06	0.11	0.06	0.1
晚籼稻	0.27	0.15	0.27	0.27	0.4	0.31	0.44	0.34	0.44	0.37
甲醇	0.23	0.2	0.29	0.14	0.29	0.11	0.27	0.12	0.28	0.14
菜油	0.36	0.31	0.44	0.32	0.43	0.31	0.42	0.32	0.43	0.32
普麦	0.44	0.28	0.47	0.24	0.5	0.06	0.48	0.09	0.49	0.08
早籼稻	0.24	0.18	0.17	0.1	0.16	0.09	0.15	0.15	0.13	0.14
菜粕	0.29	0.18	0.28	0.08	0.24	0.05	0.21	0.05	0.17	0.09
菜籽	0.43	0.15	0.49	0.12	0.48	0.1	0.49	0.02	0.49	0.08
硅铁	0.18	0.2	0.26	0.02	0.32	0.08	0.32	0.02	0.31	0.08
锰硅	0.29	0.12	0.31	0.1	0.33	0.13	0.35	0.16	0.36	0.19
白糖	0.28	0.28	0.16	0.33	0.11	0.3	0.07	0.27	0	0.24
PTA	0.3	0.28	0.26	0.2	0.18	0.24	0.18	0.23	0.12	0.29
强麦	0.2	0.2	0.21	0.17	0.18	0.11	0.19	0.14	0.2	0.11
动力煤	0.23	0.16	0.23	0.08	0.19	0.1	0.17	0.08	0.17	0.07

[1] 皮尔森相关系数通常用来衡量两个变量之间的线性相关性，取值范围为 [-1,1]。当为正时，表示两个变量之间为正相关，为负时表示为负相关。绝对值越接近于 1 表示相关性越强。

从表中易知，在不同 N 的取值下，所有品种的相关系数均为正。因此，交易持仓量的变化与价格波动的绝对值之间确实存在着正向关系。但是，学术界一般认为，当相关系数 |r|>0.8 时，两变量间存在高度相关性；当 0.6<|r|<0.8 时，可以认为两变量具备较强相关性；当 0.4<|r|<0.6 时，两变量具备中等相关性；当 0.2<|r|<0.4 时，可认为两变量相关性较弱。从表中可发现，除少数蓝色区域（大于 0.4），大多数品种交易持仓量的变化与价格波动幅度的关系均较弱，且间隔日期 N 的长短对于结论也无较明显影响。整体来看，价格波动对于交易量、持仓量趋势均有一定正向影响。在提取特征时，需要将价格波动相关数据引入模型，但需要设计模型结构捕捉非线性关系提高数据价值。

（三）品种交易持仓与风控措施参数的长期关系

除价格波动外，本文同时研究风控参数对于交易持仓量的长期影响。风控参数的设置拟在调节市场热度，平抑行情变化。考虑到保证金、手续费等参数与交易持仓量的变化量纲不同，在分析相关参数与交易持仓量变化波动相关性时，本文决定采用变异系数（Coefficient of Variation）来衡量不同风控参数下交易、持仓量的运行情况。具体计算方式如下。

$$变异系数 = \left(\frac{标准差}{平均值}\right) * 100\%$$

变异系数越大，交易持仓相比其平均值波动幅度越大。本文以 2016-2018 年各品种的相关数据为基础，分别计算不同品种运行的变异系数，并利用皮尔森系数计算风控参数与变异系数之间的关系。本文以保证金与平今仓手续费为代表进行重点分析。具体结果见表 2、3。考虑到相关系数的计算要求相关风控参数经历过多次调整，因此表 2、3 仅保留了所选区间内符合条件的品种进行分析。

表 2：保证金与品种长期波动的变异系数

品种	相关系数 r1（成交量变异系数&保证金）	相关系数 r2（持仓量变异系数&保证金）
PTA	-0.44	-0.52
强麦	0.21	0.1
早籼稻	-0.16	-0.07
晚籼稻	-0.16	-0.07
普麦	0.13	0.43
棉纱	0.24	-0.73
棉花	-0.34	-0.66
煤	-0.15	-0.56
玻璃	0.73	0.18
甲醇	0.86	-0.53
白糖	-0.52	-0.89
硅铁	-0.24	-0.18
粳稻	-0.16	0.87
苹果	-0.42	-0.28
菜油	0.29	-0.06
菜籽	0.26	0.52
菜粕	0.48	-0.11
锰硅	-0.22	-0.5

表 3. 平今仓手续费与品种长期波动的变异系数

品种	相关系数 r3（成交量变异系数&平今仓手续费）	相关系数 r4（持仓量变异系数&平今仓手续费）
PTA	-0.44	-0.52
强麦	0.21	0.1
早籼	-0.16	-0.07
晚籼	-0.16	-0.07
普麦	0.13	0.43
棉纱	0.24	-0.73
棉花	-0.34	-0.66

从表 2、3 可发现，整体来看，保证金及平今仓手续费数值大小与交易持仓波动变异系数的关系为负相关。当保证金或平今仓手续费增大，对应品种的交易持仓量波动比率相对更小，具体数值因品种差异而有较大的变化。观察表格，发现存在部分品种的相关系数为正的情况，可能是因为调整点聚集在单边行情或大波动行情下。基于上述分析，我们决定将风控措施参数引入特征序列，作为预测的基础。

（四）品种交易持仓行情与风控措施参数的短期关系

除长期维度外，本文以 2016-2018 年相关数据为基础，尝试探索风控参数的短期变化对市场运行的影响。经数据分析，从全市场角度来看，保证金及手续费变化对于交易持仓量的短期影响较为随机，没有显著规律。基于此，本文决定根据交易特征将客户分为四个群体，如长线客户、短线客户、大客户及小客户[2]。整体来看，保证

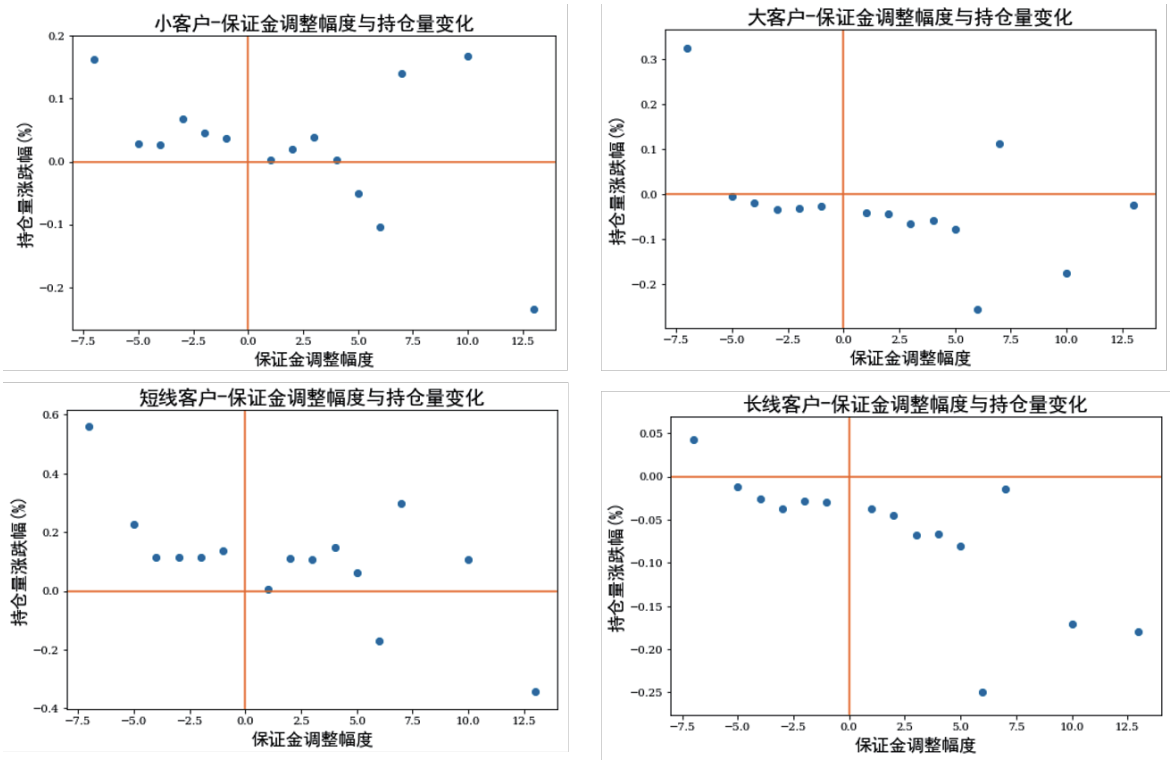


图 3：不同客户群体下保证金调整幅度与调整点前后五日持仓量均值变化关系

[2] 四类群体的划分主要基于客户的交易持仓周期及数量。

金对于持仓量影响较明显，而平今仓手续费对于交易量影响更显著。具体见图 3、4。图 3 中四个小图分别代表四个客户群体下保证金调整幅度与持仓量之间的关系。x 轴表示保证金调整前后数值变化的大小，y 轴表示保证金调整前后五日内持仓量均值的变化幅度。图中每一个点均代表历史上一次真实调整。图 4 中，x 轴代表平今仓手续费调整幅度，y 轴表示调整前后五日内交易量均值变化幅度。观察图 3 可知，当保证金上调时，大客户及长线客户持仓量呈现降低趋势；当保证金下调时，小客户及短线客户的持仓量呈现上涨趋势。观察图 4 可知，平今仓手续费上调对于短线客户交易量减少的影响较为明显，也符合普遍认知；相应的，下调手续费对于短线及小客户交易量促进有一定作用。

短期来看，风控措施参数对于不同客户群体的交易持仓有一定影响。因此，本文将风控参数措施的变化值也引入特征向量。

三、特征工程与模型构建

基于上述分析，交易持仓量受到多重因素的影响，不同因素影响程度不一而同。本文尝试利用多模型融合方式捕捉数据之间的不同关系，挖掘深层价值，并对未来交易持仓量进行预测。具体问题定义为：针对任一合约，第 T 日收盘后，根据当日现行风控措施参数及历史运行数据，预估未来 5 个交易日的交易量与持仓量。

数值预测相关问题中，特征选择是模型构建的重要基础，决定了模型效果。本文经过数据分析及实验迭代，最终决定选择包含结算参数、行情特征、客户特征及合约特征等四大类共 317 维特征。其中，结算参数包含如历史价格波动、合约间价差等多维度特征；行情特征包含了历史交易持仓相关数据特征；客户特征包含不同属性客户的特征数据及不同客户群体的交易特征；合约特征重点提取了合约运行特征及合约阶段，约束

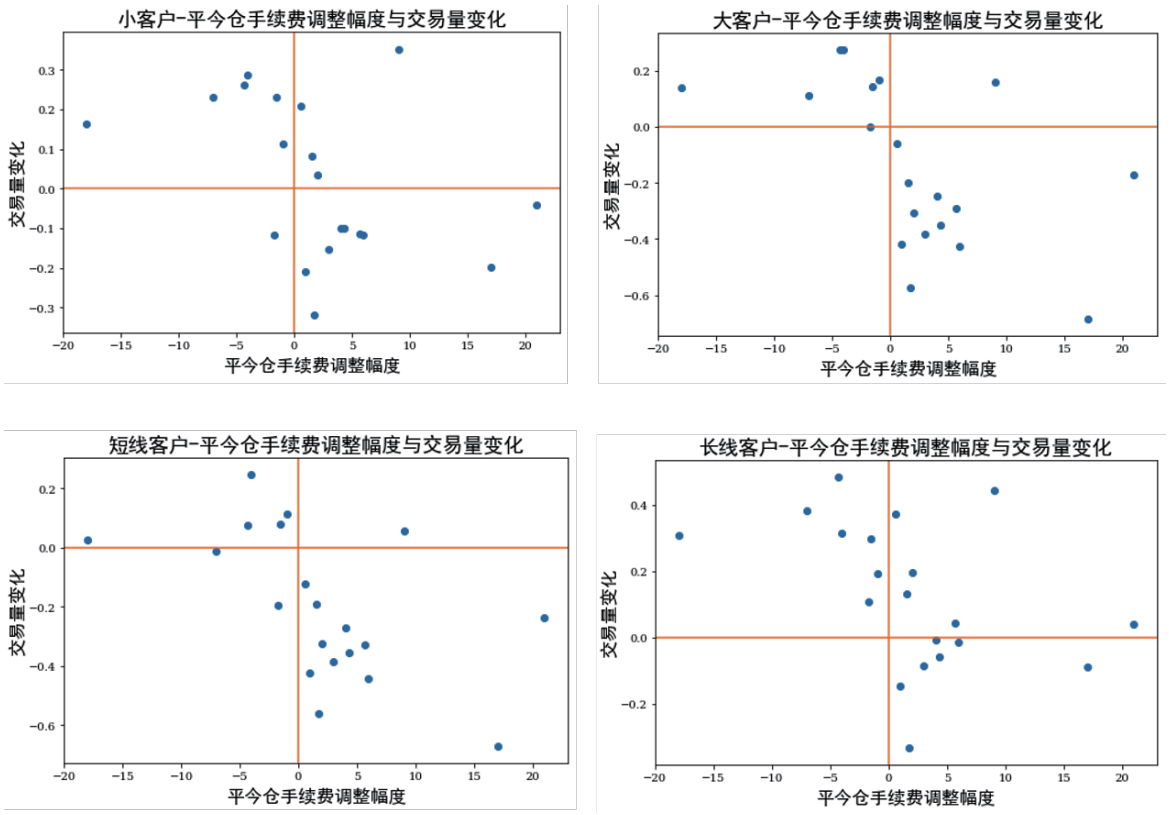


图 4：不同客户群体下平今仓手续费调整幅度与调整点前后五日交易量均值变化关系

预测结果。

下述 317 维特征中，7 维为合约约束性特征，310 维为历史交易相关的时序特征。完成特征提取后，本文开始构建三个机器学习模型。具体细节如下。

（一）整合移动平均自回归（ARIMA）

在统计与经济相关领域，ARIMA（Autoregressive integrated moving average）模型是一种常用的时间序列预测算法，该模型通常应用于平稳时间序列，或通过差分平稳过程消除均值方程的非平稳性的序列。其中，自回归（AR）是统计上处理时间序列的一种方法，衡量序列自身在不同时刻随机变量的相关性，利用变量以往时刻的取值来预测当期时刻的取值，并假设它们为线性关系。该方法被广泛的应用于金融序列相关的建模问题中。移动平均模型（MA）是另一种对单一变量进行时间序列建模的方法。因本文的问题较为契合 ARIMA 的常用场景，因此决定利用它捕捉交易持仓量序列的时序关系。

（二）基于支持向量机的回归模型（SVR）

ARIMA 模型是基于捕捉时序相关性直接预测未来交易量持仓量，同时我们希望利用更多的信息量以求获得更好的预测效果。我们决定利用支持向量机来预测未来交易量持仓量的涨跌幅度

表 4：数据特征表

特征类别	特征名称	维度
结算参数特征	历史最高价波动幅度	20
	历史结算价波动幅度	20
	历史开盘价波动幅度	20
	历史收盘价波动幅度	20
	历史现货价波动幅度	20
	历史主次合约间价差	10
	历史保证金波动幅度	10
	历史交易手续费相对于基准手续费的变化率	10
	历史平今仓手续费相对于基准手续费的变化率	10
行情特征	历史持仓量及波动幅度	20
	历史交易量及波动幅度	20
	历史持仓量与近 5 日最大值比值	10
	历史交易量与近 5 日最大值比值	10
	历史买、卖开仓占比	20
	历史法人买、卖持仓占比	20
	历史成交持仓比	10
客户特征	不同属性客户（N、L、S、A）的历史特征数据	20
	不同客户群体（短线、长线、大户、小户）历史交易持仓数据	40
合约特征	合约所处阶段	5
	合约类型	2

作为补充。支持向量机（Support Vector Machine）是一种广泛应用于分类与回归问题中的机器学习模型。该方法的核心是将低纬度不可分特征使用“核函数”有效的进行非线性处理，映射到高维特征空间。通过寻找高维空间中的超平面对数据进行分类或回归。

（三）序列到序列模型（Seq2Seq）

Seq2Seq 模型，全称为 Sequence to Sequence，是由谷歌大脑团队和 YoshuaBengio 团队提出的一种广泛运用在翻译、文本自动摘要及一些回归预测问题上的深度神经网络。在提出之初，Seq2Seq 主要被用来解决自然语言处理相关的问题。但因其强大的时序关系挖掘能力，近年来也被逐渐应用于数值型序列的预测问题中。如图所示，本文所用网络通过编码器（Encoder）与解码器（Decoder）两个过程将过往十日的行情特征作为输入序列，将其映射为未来五日的交易量或持仓量序列。编码器利用非线性函数将输入序列组合为隐藏层的隐藏向量，该向量具备表达输入序列信息及潜在关系的能力。解码器将传递来的隐藏向量进行解码，并结合输入的 T~T+4 日的市场行情特征，逐日预测未来 T+1~T+5 日的交易持仓情况。

四、多算法融合模型构建

期货市场行情瞬息万变，客户群体的交易持仓行为受到众多因素的影响，因此单一模型容易对历史数据产生过拟合现象，并且面对未知影响时模型鲁棒性与泛化能力较差。基于此，本文提出基于 ARIMA、SVR 及 Seq2Seq 三个单一模型的融合模型，并利用网格搜索技术进行参数权重配置。

（一）单一模型训练

构建融合模型之前，需要利用历史数据训练单一模型，设定最优参数。本文采用 2016-2018 年三年的数据作为数据集，并将数据集分为均等三份，前两份用做训练数据集，后一份用做验证数据集。实际训练中，利用训练数据拟合模型，根据验证数据效果优化参数，直到效果达到“瓶颈”即停止训练。

（二）模型融合

为增强模型的鲁棒性，本文决定构建基于单一模型的融合模型。为提高预测效果，我们需要为三个单一模型科学的分配权重，最优化模型效果。模型权重属于超参数，一般做法是根据实验结果不断进行人工判断调整。考虑到人工较难穷

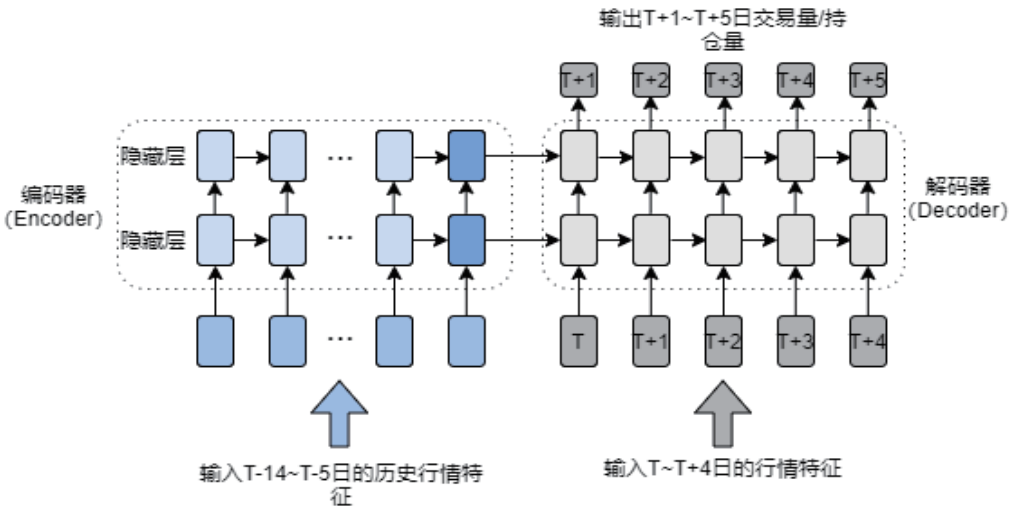


图 5：Seq2Seq 模型流程图

表 5：融合模型超参数权重配置表

	Seq2Seq	Arima	SVR
T+1 天交易量	0.25	0.25	0.5
T+2 天交易量	0.55	0.45	0
T+3 天交易量	0.5	0.3	0.2
T+4 天交易量	0.5	0.3	0.2
T+5 天交易量	0.5	0.3	0.2
T+1 天持仓量	0.25	0.25	0.5
T+2 天持仓量	0.5	0.4	0.1
T+3 天持仓量	0.5	0.3	0.2
T+4 天持仓量	0.5	0.3	0.2
T+5 天持仓量	0.5	0.3	0.2

尽各种组合，且比较难配置到最优参数。本文使用网格搜索（Grid Search[3]）进行自动化权重分配调优。

经过 Grid Search 的搜索，最终参数配置如表 5。

五、模型评估

（一）模型准确率及置信区间

为科学评估模型效果，本文定义如下参数评估模型预测效果，具体如下。

$$Acc = Relu(1 - 0.2 * \sum_{i=1}^{5} abs(\frac{P_i - T_i}{T_i}))$$

$$Relu(x)=Max(0,x)$$

假定当前日期为 T 日，Pi 表示第 T+i 日的交易量（或持仓量）预测值，Ti 表示第 T+i 日的交易量（或持仓量）的真实值，Acc 表示最终衡量实验结果的准确率，本文中我们使用 Relu 函数对最终结果进行平滑处理，减少极端值对于模型评估影响。从上述公式定义中易知，准确率结果处于 [0,1] 区间之内，数值越大表示预测效果越好。

为多角度评估模型效果，本文在准确率之外引入了置信区间概念。计算每个品种预测准确率置信度为 95% 的区间范围，具体计算方式如下：Avg 代表 N 个预测日的平均准确率，Delta 代表 N 个预测日的标准差，定义 95% 置信区间为 [Avg-delta*1.96, Avg+delta*1.96]。通常来说，置信区间的上下界代表了预测结果的可信度及波动性。上下界数值越高，差值越低表示预测效果更好。

本文选取 2020 年 12 月 21 日至 2021 年 1 月 22 日共 19 个交易日计算准确率与置信区间，结果见表 6。从表中可发现，融合模型对于持仓量的预测较好，平均值达到了 80% 以上，棉花、白糖甚至超越了 85%。交易量的波动相比于持仓量更大，因此预测结果也相对逊色一些，但是仍有 8 个品种超过了 70%。持仓量预测的置信区间上下界普遍较高，也和持仓量走势的稳定性较为一致；交易量预测的置信区间相对较宽，下界相比于持仓量更低，但整体比较稳定。

（二）融合模型与单一模型比较

为直观展示模型融合的优势，本文以 CF1805 整个挂牌周期为标的，比较融合模型与各单一模型的预测效果，结果见表 7。从表 7 中可知，融合模型的交易持仓量预测准确率均高于单一模型。将预测标的放宽至所有合约，计算统

[3]Grid Search 是一个自动化参数调优方法，通过人工设置参数可取的数值集合，该算法通过穷举搜索的方式找到最优参数。

表 6：各品种预测准确率及置信区间

品种名称	交易量 Acc	置信区间	持仓量 Acc	置信区间
棉花	0.77	[0.69,0.99]	0.86	[0.91,0.99]
白糖	0.75	[0.7,0.978]	0.86	[0.91,0.99]
甲醇	0.74	[0.75,0.99]	0.8	[0.9,0.98]
PTA	0.72	[0.69,0.99]	0.74	[0.9,0.98]
菜油	0.7	[0.61,0.99]	0.84	[0.88,0.98]
尿素	0.7	[0.48,0.99]	0.83	[0.86,0.97]
纯碱	0.71	[0.62,0.99]	0.86	[0.83,0.99]
菜粕	0.7	[0.64,0.97]	0.85	[0.9,1]
苹果	0.69	[0.66,0.97]	0.8	[0.86,0.97]
棉纱	0.69	[0.66,0.98]	0.85	[0.91,0.99]
玻璃	0.68	[0.67,0.98]	0.85	[0.89,0.99]
短纤	0.63	[0.52,0.99]	0.86	[0.92,0.99]
动力煤	0.57	[0.49,0.94]	0.82	[0.86,0.98]
硅铁	0.6	[0.53,0.93]	0.8	[0.84,0.99]
锰硅	0.65	[0.63,0.98]	0.76	[0.75,0.99]
红枣	0.51	[0.325,0.93]	0.83	[0.77,0.99]

计后，融合模型在 88% 的情况下效果优于单一模型，证实了融合方案的有效性。

同时，本文仍以 CF1805 为例，运用图形化模式展示各模型 T+1 日交易量预测值与真实值走势，具体见图 6、图 7、图 8 及图 9。从图中易知，融合模型拟合真实趋势的效果更好，Seq2Seq 与 Arima 模型在部分时间段与真实趋势背离较大。SVR 模型对于行情拟合较好，但缺点是极端行情下的偏离度太大。

（三）不同超参权重下的效果对比

融合模型中，单一模型的权重组合对于最终效果通常具有决定性因素。表 8 直观展示不同参数组合之间的差异性。表中黑体加粗一行为本文最终选定的超参权重，通过对比可发现，该组合准确率（0.92）高于表中列出的其他组合。

六、模型可解释性

在时序数据预测的业务问题中，预测结果的可解释性是一个重点关注领域。针对融合模型，如何高效直观的解释预测结果成为一个挑战。本文使用 LIME（Local Interpretable Model-agnostic Explanation）[4] 方法对预测结果进行解释，量化不同特征对于预测结果的贡献度。以 CF1805 在 2018 年 1 月 11 日的交易量预测为例，真实值为 46484 手，模型预测值为 49865 手。使用 LIME 方法解释该预测结果，按照倒序排列，挑选对于

表 7：主力合约交易持仓量预测效果对比表

预测结果	融合模型	Seq2Seq	Arima	SVR
交易量准确率	**0.74**	0.67	0.65	0.70
持仓量准确率	**0.92**	0.64	0.79	0.90

[4] LIME 是近年来机器学习界较为常见的模型解释工具，其核心原理是利用简单模型对复杂的黑盒模型进行拟合，进而完成结果解释。

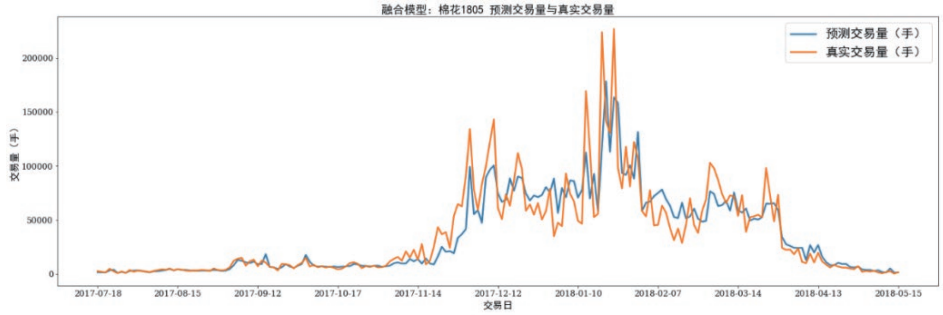


图 6：融合模型真实交易量与预测量走势图

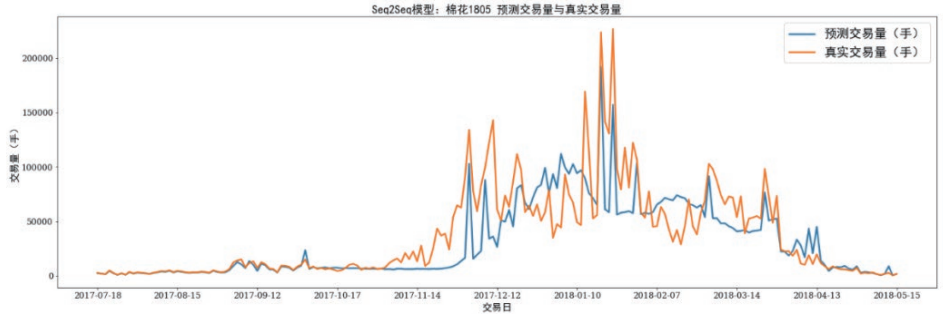


图 7：Seq2Seq 模型真实交易量与预测量走势图

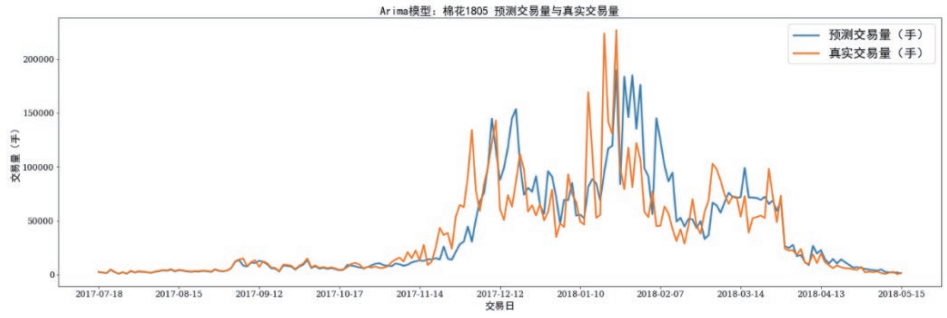


图 8：Arima 模型真实交易量与预测量走势图

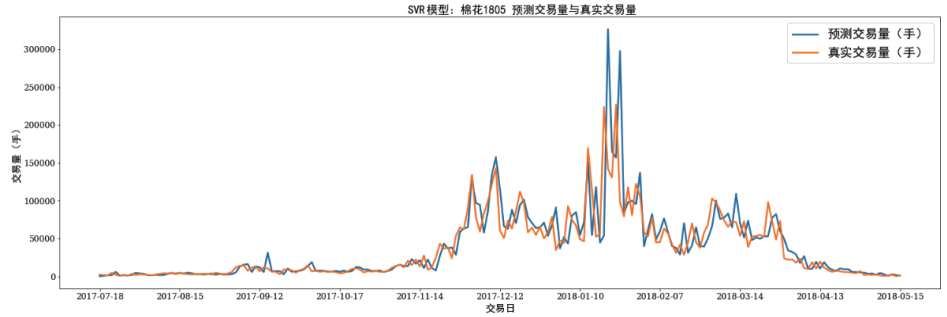


图 9：SVR 模型真实交易量与预测量走势图

预测结果影响最大的十个特征，具体见表 9。表中共四列，第一列为特征具体名称，第二列为根据 LIME 计算的特征分叉阈值，第三列为特征实际值，第四列为特征贡献度。其中，特征分叉阈值意味着当特征值超过阈值时，该特征将对于最终预测结果产生影响。一般来说，LIME 方法中的特征贡献度主要应用于评估分类问题中各个特征对于预测结果的影响大小。而本文的实际问题

表 8：不同超参数组合下融合模型预测准确率

Seq2Seq	Arima	SVR	交易量准确率	持仓量准确率
0	0.6	0.4	0.7	0.87
0	0.7	0.3	0.7	0.85
0.1	0.7	0.2	0.71	0.85
0.1	0.8	0.1	0.71	0.83
0.2	**0.2**	**0.6**	**0.7**	**0.92**
0.2	0.3	0.5	0.71	0.89
0.3	0.6	0.1	0.72	0.83
0.4	0.5	0.1	0.72	0.83
0.4	0.6	0	0.7	0.81
0.5	0	0.5	0.7	0.83
0.6	0	0.4	0.71	0.8
0.7	0.3	0	0.7	0.76
0.8	0	0.2	0.7	0.73
0.9	0	0.1	0.7	0.7
0.9	0.1	0	0.69	0.7

归属于数值回归，特征贡献度仍然可以表征该特征对于最终结果的正负影响度。举例来说，特征贡献度绝对值最大的是“第 T-3 日交易量振幅”这个特征，该特征分叉阈值为 0.36，而真实值为 1.11，远高于阈值。因此这个特征被 LIME 判定为对于最终结果贡献系数为 0.139，即该参数值对于模型最终结果贡献了 13.9% 的正向影响。同时，我们可观察到特征“第 T-3 交易日成交持仓比”贡献度为 -0.015，即该特征大小对最终预测结果影响为负 1.5%。

七、总结与展望

本文尝试基于历史数据特征的融合模型实现

表 9：特征贡献表

特征名称	特征分岔阈值	特征值	特征贡献度
第 T-3 日交易量振幅	X> 0.36	1.11	0.139
第 T-2 日交易量振幅	X<= -0.19	-0.2	0.112
第 T-2~T-1 日交易量振幅之和	X<= -0.11	-0.3	0.084
第 T-6 日交易量振幅	X<= -0.19	-0.56	0.054
第 T-3~T-1 日交易量振幅最大值	X> 0.71	1.11	0.037
第 T-3~T-1 日交易量振幅之和	X> 0.78	0.81	0.022
第 T-3 交易日成交持仓比	X> 0.45	0.61	-0.015
第 T-1 日交易量振幅	-0.17 <X <= 0.1	-0.1	-0.011
第 T-2 日成交持仓比	X> 0.46	0.49	-0.009
第 T-4 日成交持仓比	0.27 <X <= 0.45	0.3	-0.008

交易、持仓量的预测。整体来看，融合模型可以有效预测未来五日交易、持仓量。其中，持仓量准确率较高，交易量准确率与稳定性相对较差。作为探索性项目，本文仍存在以下不足。

一是交易量预测准确率与稳定性有待提升。融合模型预测持仓量的准确率及稳定性均显著高于交易量。原因在于期货交易作为T+0交易制度，行情剧烈变动、主力资金拉盘、舆情消息等对于日内交易量的影响较大，规律不明显且可预测性相对较差。

二是预测特征抽取不够丰富。虽然本文从历史数据中抽取逾300维特征，但是并未将宏观数据、品种基本面等信息纳入特征。同时，对于短期内对市场参与度影响极大的舆情消息，目前也暂未提出好的量化方法纳入预测模型。

三是预测结果存在一定滞后性。市场运行在外部因素影响下常出现大幅波动，模型较难及时捕捉到这种异动，因此预测趋势经常慢于实际波动一个交易日。该现象在时序数据的预测上较为常见，本文利用高维度特征与模型融合在一定程度上缓解了这种现象，但仍无法完全避免。

联邦学习在证券行业的应用初探

陈镇光、丁一 / 国信证券股份有限公司 邮箱：chenzguang@guosen.com.cn

一、概述

联邦学习的概念在 2016 年由谷歌率先提出，最初用于解决安卓手机更新本地模型的问题。而近年来随着各个国家对个人隐私和数据安全的重视逐步提高，联邦学习作为机器学习和隐私计算的结合体，为解决数据孤岛问题开辟了一片全新领域。在医疗、金融、零售等领域逐渐涌现出了一些很有价值的应用场景，有部分场景已进入具体的落地应用阶段；在业界也出现了支持联邦学习架构体系的工业级开源框架，如 FATE（Federated AI Technology Enabler）。

作为金融行业数字化转型的践行者，国信证券走在探索科技改变金融的道路前列，积极投入探索联邦学习在证券行业的应用，期望更好的支持未来证券业务的发展。联邦学习是否有价值，主要取决于它的关键应用场景。本文结合国信证券的研究和实践，重点介绍了联邦学习在证券行业可行应用场景的探索。

二、什么是联邦学习

当多个数据拥有方想要联合他们各自的数据训练机器学习模型时，传统做法是把数据整合到一方进行训练。然而该方案由于涉及隐私和数据安全等合规问题通常难以实施，于是联邦学习出现了。它是指在进行机器学习的过程中，各参与方可借助其他方数据进行联合建模，并且各方无需共享数据资源，即数据不出本地的情况下，进行数据联合训练，建立共享的机器学习模型。

因此，联邦学习总结起来具有如下 3 个方面的特点：（1）分布式计算，各参与方地位平等。（2）原始数据不出库，减少了数据泄露的风险。（3）模型共享，各方都可以从训练的结果中获益。

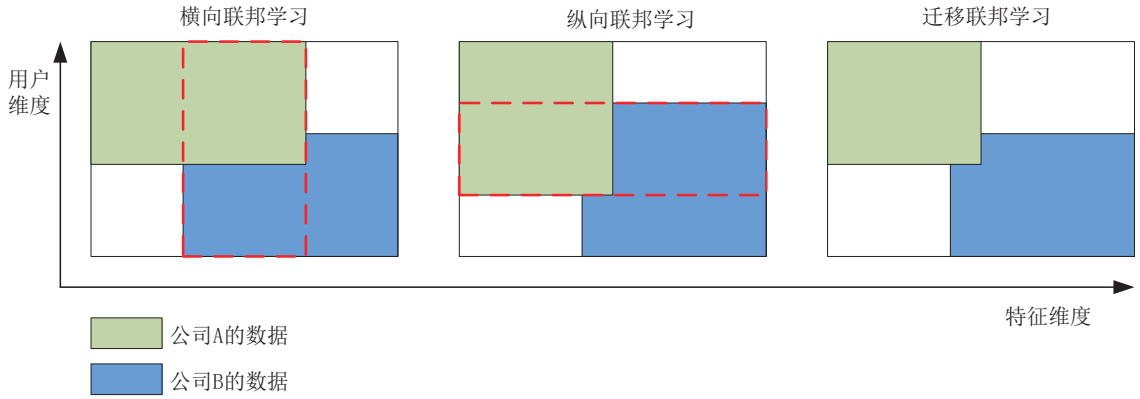


图 1：联邦学习的分类

从联邦学习的特点我们可以看到它的核心用途是通过分散的数据进行联合训练，以解决本地数据不足的问题。所以它特别适合如下的应用场景：

（1）某些行业或者公司本地数据量很少，但仍想应用先进的人工智能技术。现实中除了有限的几个行业外，更多领域存在数据有限且质量较差的问题，不足以支撑人工智能技术的实现或者实现的效果不佳，于是它们可以考虑借助外部的数据来达到目的。

（2）不同的机构为了获得更多的收益，有意愿推动数据做联合训练。例如在产品推荐服务中，产品销售方拥有产品的数据、用户购买商品的数据，而第三方支付公司有用户购买能力和支付习惯的数据，于是两家公司为了共同的利益可以进行数据的联合训练。

（3）数据源之间存在难以打破的壁垒。在大多数行业中，数据是以孤岛的形式存在的，由于行业竞争、隐私安全、法律法规、行政手续复杂等问题，即使是在同一个公司的不同部门之间，实现数据整合也面临着重重阻力。

三、联邦学习的应用方式

根据多方数据集的用户、特征重叠量的不同，可以将联邦学习分为横向联邦学习、纵向联邦学习和联邦迁移学习（如图 1 所示）。

横向联邦学习应用于用户交叉不多、特征重叠较多的场景。例如一家券商在不同地区的两个分支机构，或者两家不同区域的地方性券商，他

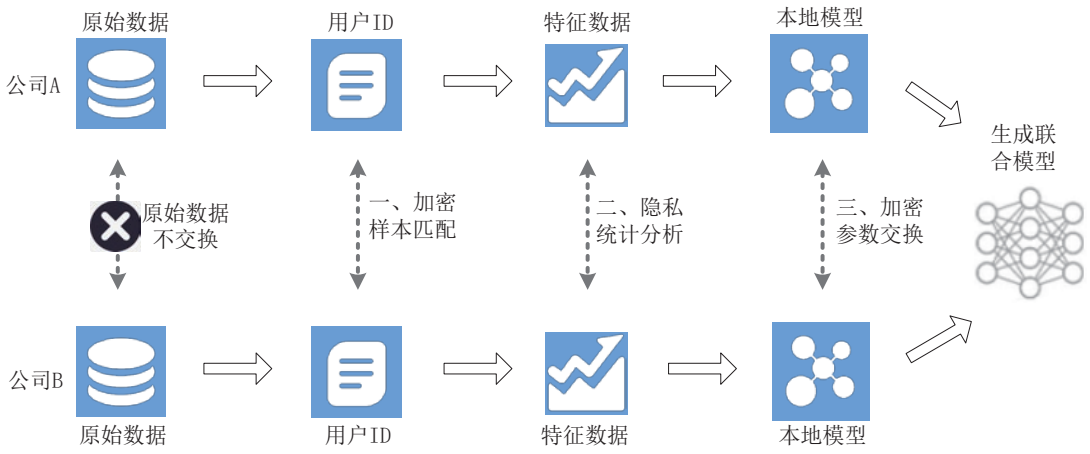


图 2：联邦学习应用步骤

们服务的是不同的客户群体，但客户的金融属性大致相似。纵向联邦学习则应用于特征重叠不多、用户交叉较多的场景。例如服务同个地区的券商和通信运营商，他们的客户必然会有很多的交叉，但由于业务领域不同他们拥有的客户标签属性也会有很大不同。而联邦迁移学习是应用于用户和特征的交叉都不多的场景，例如一家中国的银行和一家美国的互联网公司，各自的用户群和用户特征都不太相同。

3 种不同类型的联邦学习所采用的应用方式各不相同。目前生产中的应用主要还是以纵向联邦学习为主，因此下面以纵向联邦学习为例介绍它的应用方式，主要分为以下几个步骤（如图 2 所示）。

第一步隐私求交，就是双方在不暴露各自数据的前提下，通过加密标识对齐共有的客户，这些客户的数据就是后面训练的原始数据。

第二步隐私特征处理，利用同态加密等技术，在保证双方数据不泄漏的情况下，实现特征相关性分析、特征权重分析、特征预测能力分析，抽取出原始特征后再通过特征的组合、特征升维等衍生处理生成目标特征。

第三步就是最重要的联合建模，原始数据都在各自本地，每一轮迭代都要交换中间的梯度数据，同步更新各自模型参数。目前的实现方式中大都需要借助中间的控制结点完成一些控制指令的汇总、分发。

最后输出一个联合模型，根据这个模型，双方都可以应用在自己的业务场景中。

四、国信的试点应用

联邦学习在证券行业目前还没有先驱应用，但它显现出来的价值是非常明确的。国信证券正在尝试营销领域的试点应用，包括以下的场景。

（一）基金推荐

在财富管理领域，基金代销是其不可或缺的重要组成部分。国信的金太阳 APP 中就为客户提供了大量的代销基金产品。但面对海量的产品，客户往往无从下手。因此我们要从客户的角度出发，帮助客户挑选适合自己的理财产品。

传统的做法更多的是不管什么类型的客户都为其推荐收益率最靠前的基金，达到“吸引眼球”

图 3：联邦学习在基金推荐场景的应用

的效果。但这种方法往往不能带来高的转化率，因为没有考虑客户的风险偏好等因素，客户往往并不买账。

另一种逐渐流行的方法是引进 AI 技术，通过对券商自有数据进行分析形成客户画像和基金画像，再结合个性化推荐算法，能够为不同客户推荐更匹配自身投资需求的产品，达到了一定程度的“千人千面”的效果。但券商自有的数据毕竟有限，无法覆盖客户离开自家 APP 后的行为方向，也无法覆盖非自家客户的潜在客户，因此客户画像的精度和广度具有一定的局限性。

联邦学习为券商的 AI 画像技术打开了另一扇大门。

国信证券拥有客户的账户、交易，以及基金的资料等数据，能够对客户的购买能力、基金的特性等进行画像。而互联网公司拥有客户上网的行为数据，能够刻画用户的爱好和习惯。例如客户在搜索引擎中检索的关键字反映了客户某个时刻的关注点，通过收集客户的检索信息可以挖掘客户短期或长期的兴趣爱好；又比如通过收集浏览器数据，分析那些经常浏览金融、财经网页的人群，因为他们相对来说有更强烈的投资需求。

因此，通过结合这两类数据进行联合训练形成联合预测模型，能够达到任何单独一方无法取得的模型精度，从而能够为客户精准推荐更加合理的理财产品。

（二）休眠户激活

国信对休眠户的定义是指年手续费低于一定阈值的客户，这些客户虽然已经开了户但并不能为公司带来价值。根据统计，休眠户的数量随着整体客户数量的增加而呈每年上升的态势。如何充分挖掘这部分客户的潜在价值，是每家券商都需要面对的难题。

这个问题之所以难以解决，一方面是因为这部分客户的基数往往不少，远远超出了营销人员的覆盖能力，全部进行激活的成本太高；另一方面由于休眠客户的交易不活跃，公司内部缺乏有效的支撑数据去分析客户的意向。因此，一般的做法都是基于一些专家规则挑选出部分客户去做营销。但由于针对性不强，营销的效果往往并不理想。

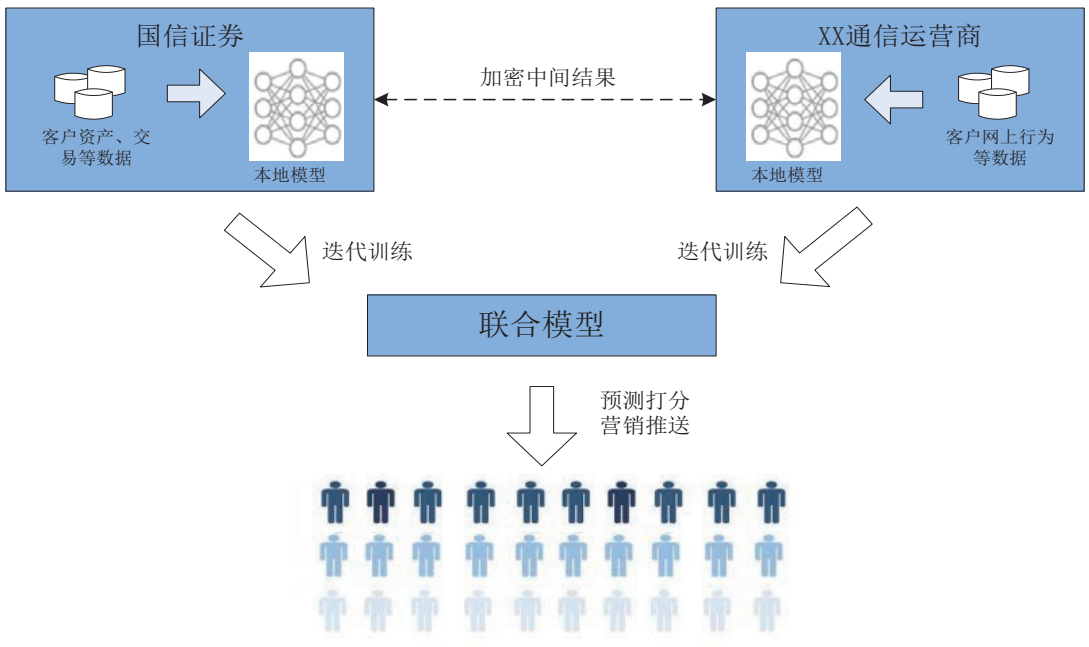


图 4：联邦学习在休眠户激活场景的应用

国信正在尝试的解决思路，是通过采用联邦学习的方法，借助第三方通信运营商的数据来补充本地数据的不足，从而对客户进行更精准的分析。

如果一个客户在国信内部看来处于休眠状态，但在国信外部却经常存在跟金融相关的网上行为，例如打开证券或银行 APP、浏览金融理财的网页等，说明他很可能是可以被激活的潜在客户，这时候我们就可以加紧对其做针对性营销。通过引入这种外部的相关数据进行联合建模，预测每个休眠户是否仍存在理财投资的意向，并对其进行打分排序，精准定位出潜在人群，从而提高营销的投入产出比。

五、存在的问题分析

联邦学习作为最近几年兴起的一项新技术，不可避免的存在一些仍待解决或优化的难题，例如：

（1）模型训练时间比较长。训练时由公司 A 训练出部分结果，B 训练出部分结果，再进行参数的交换，相比本地训练多了几次外网的交互。如果 A 的数据量较大，或者机器设备较差导致训练较慢，这时 B 需要等待;反过来也一样。因此，每一轮迭代时间延长了，而一个模型的训练需要成千上万轮迭代，这将导致训练时间大大延长，影响机器学习的效率。

（2）系统复杂度比较高。参与训练的相关方之间存在强依赖的关系，只要有一方出问题，整个训练流程就会被阻塞。如果涉及三方、甚至更多方的联合训练时，系统将更加复杂，因为它是一个网状的互联结构。这就涉及到网络的稳定性、异构系统的高可靠、分布式系统的故障排查等问题；参与方越多可能越增加系统的脆弱性。

（3）隐私问题。尽管联邦学习的目的是解决多方协同中的隐私问题，但目前的技术还难以做到百分之百规避隐私问题。例如在隐私求交时，双方通过加密 ID 的碰撞，能够知道对方的部分用户列表，尽管原始数据并没有暴露，但这也限制了某些敏感场景的应用。

六、总结和展望

联邦学习的用途是非常明确和有价值的，因为数据孤岛已经越来越成为制约人工智能全面爆发的瓶颈。因而，很多科技公司都纷纷布局并参与其中，逐步把联邦学习推向更加成熟的应用模式。国信证券率先在证券行业开展联邦学习应用的研究和实践，通过尝试新技术与现有业务的结合，用科技来改善业务效果，甚至期望在未来开创新的业务模式。可以预期的是，未来很多场景将需要行业共同参与和培育，才能发挥更大的价值，就像区块链的应用。

联邦学习的应用场景跟区块链中的联盟链有异曲同工之处。区块链是用去中心化、防篡改等技术实现数据的静态共享，而联邦学习则是用去中心化、隐私保护的技术实现数据的动态协同。随着这两项技术的快速发展，未来很可能走向融合，将数据共享的应用场景推向更加广阔的空间。

基于列式存储和交互式数据分析的风险管理平台建设实践

潘聪、杨光 / 光大理财有限责任公司 panc.ew@cebwm.com

2019 年 9 月光大理财有限责任公司作为首个股份制银行理财子公司获批开业。理财资管业务进入子公司经营阶段后，随着具有不同策略理财产品的推出，涉及到的交易市场、交易品种、交易体量不断增多，需要建设风险系统将公司的各项资产整合纳入其中，使得管理层能够从统一的视角进行风险控制；使得执行部门能够更加精确地评估投资绩效和控制投资风险，以达到“风险集中管理”的目标。搭建涵盖市场风险、信用风险、流动性风险和压力测试的一体化风险系统架构，系统能够通过灵活的风险指标配置，达到动态预警，同时依据绩效归因结果分析调整组合策略、优化组合持仓，减少不必要的风险敞口，获取更多的超额收益。

一、风险系统简介

依托光大理财企业架构和技术栈自主研发风险计量引擎，通过智能化的指标分组计算、自动调度次序，可以实现 VaR、久期、Greeks 等 1000 余个风险指标的计量，计量范围覆盖权益类、固收类等全部产品类别；可根据报表和前端功能自动配置所需数据、数据聚合和展现形式，实现应用快速响应。

风险数据集市作为企业级数据仓库的首次搭建与应用，接入了包括客户信息、交易信息、估值信息和资讯信息等 100 余张表、共 8000 多个字段，累积数据超过 500G。如发现仍有部分数据未纳入，可通过数据管理系统操作界面便捷地扩展补充。每个风险模型的知识和经验以卡片式积累，按照统一规范、命名规则通过 Jupyter Lab

存储，模型代码与文档实时保持一致，方便业务人员和技术通过可视化编辑器搭建和优化模型实现。

二、系统架构设计

2.1 功能架构

1）限额管理：通过限额监控模块，可随时灵活添加各种监控指标，包括合规性监控指标集和根据投资组合类型划分设置的监控指标子集。

2）组合管理：可维护投资组合的业绩比较基准，支持市场指数、市场利率和多样复合指数。可查询组合资产的配置情况，盈亏情况，交易数量和金额数据等。

3）组合风险：针对组合中不同资产类别，从市场风险、流动性风险、利率风险、信用风险等方面进行风险分析与控制，根据用户定义的不同维度对绝对风险 / 相对风险在任意时间段内进行计算和分解。主要包括 VaR 分析、波动率分析、Beta 分析、利率敏感性分析、流动性分析、情景分析等。

4）组合绩效：从收益分析、业绩归因、风险调整收益、业绩持续性以及多组合横向对比等多个方面来对组合的绩效做出量化分析。

5）压力测试：根据监管要求，提供银行理财压力测试报表计算，并保存测试结果。

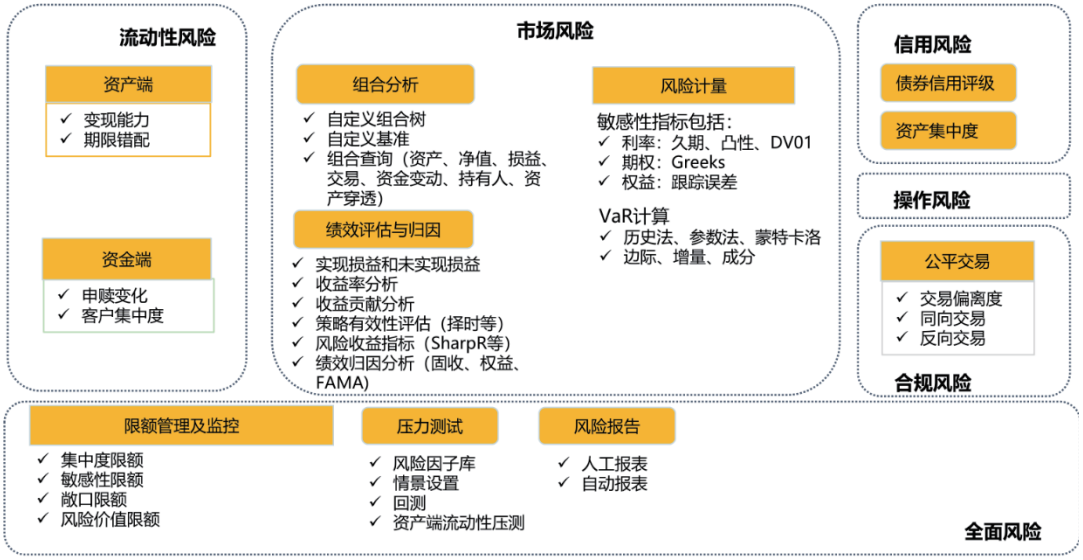
6）风控报告：可满足业务人员所需要的各类报告，支持人工自动配置报告栏目和内容模块，自动生成报表。

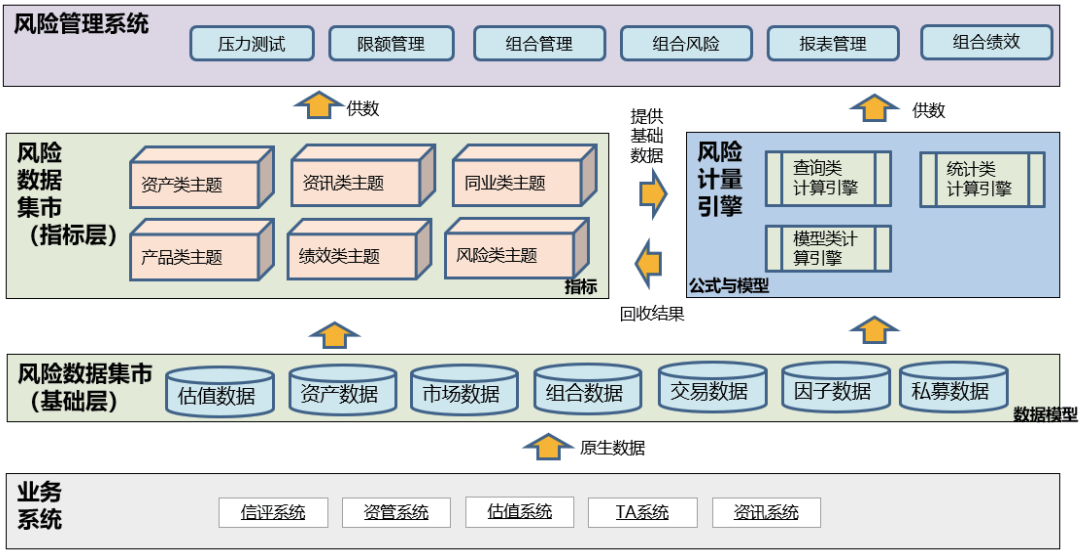
2.2 系统架构

业务系统为风险管理系统提供源数据，从估值系统获取持仓等组合和资产数据；从资管系统获取组合和产品等基本信息；从 TA 系统获取销售数据；从资讯系统获取市场行情数据、资讯数据等基本信息；从信评系统获取内部评级数据。

风险计量引擎实现各类风险的加总计量和限额计算，引用数据集市基础层数据，进行指标计量，并将计量结果返回至集市指标层和风险管理系统。

风险数据集市作为基础数据整合平台，统一多数据源，共用一套基础数据，支持跨风险计量。作为核心功能之一包括数据管理模块和指标管理模块，数据管理模块包括源数据确认（表与字段）、数据模型定义，数据校验、数据质量修正等；指标管理模块包括指标模型管理、指标分类体系、





指标间血缘关系追溯、指标计算逻辑与计算中间结果展示，指标逻辑与代码同步。

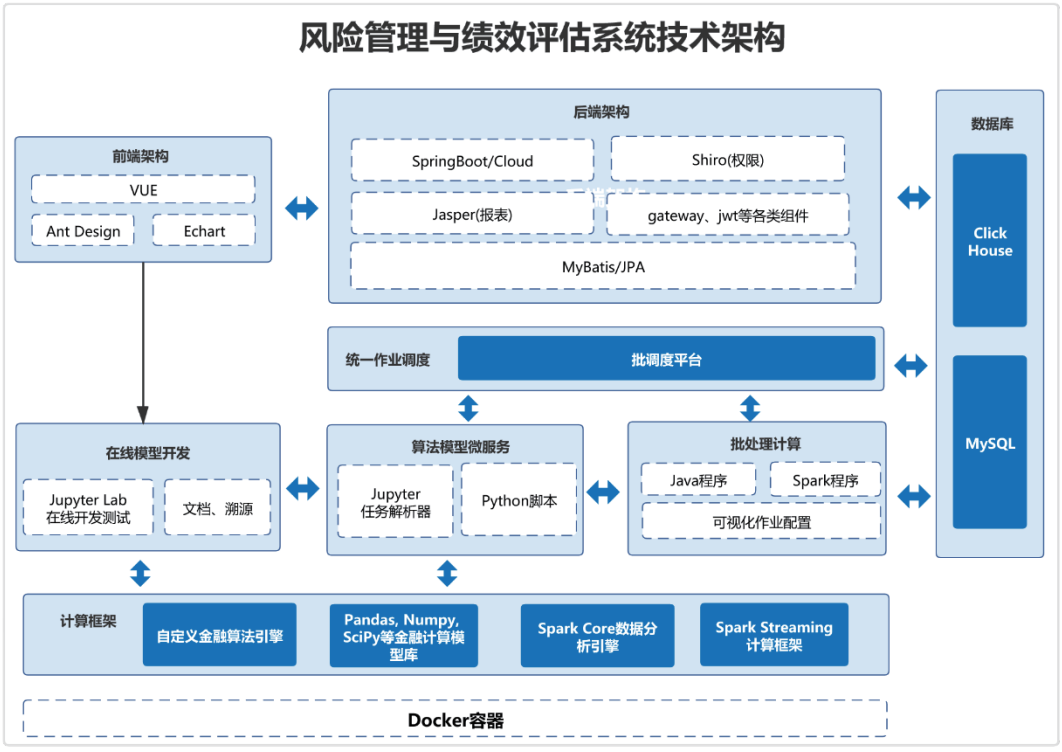
风险管理系统是基于风险计量引擎和风险数据集市的上层应用，目标是建立跨市场、跨品种、一体化的风险管理平台，实现公司一站式事后风险管理。

2.3 技术架构

在应用层面采用前后端分离技术，数据库设计了 ClickHouse - MySQL 读写分离方案。

前端架构包括 VUE 框架，Ant Design 界面组件库。

后端架构分为 Spring Cloud 微服务和 Python



金融微服务；前者提供基础应用功能，后者负责动态指标计算。

批处理任务组分为负责静态指标计算的 Python 金融指标计算引擎，负责限额计算的 Java Spring Batch 批处理任务和，负责外部数据导入 ETL。金融模型使用 Python 微服务及二次封装 Jupyter Lab，实现金融模型在线开发、指标血缘管理、指标算法卡片和金融模型在线发布。

2.4 数据架构

风险数据集市整体分成基础层和指标层两结构。基础层分为估值数据、资产数据、市场数据、组合数据、交易数据等 5 个主题数据，基础层模型基于证券期货行业数据模型进行设计，根据银行理财子公司的特点对模型进行了改造，适用于银行理财风险管理的需求，同时具备较强的可扩展性，目前涵盖 100 多张表，7000 多个字段。指标层基于基础层数据做金融指标计算，包括计算、统计、算法模型等加工形式，形成业务直接可用的指标数据，使用纯函数方式按需实时计算，一般情况下并不保存于数据库，除内外部的留痕要求外。

三、交互式数据分析

系统提供的实时交互式数据分析通过金融风险指标库和自助数据服务来实现。

3.1 金融风险指标库

3.1.1 指标库管理

金融指标库是金融指标的载体，作为金融指标的逻辑组合，与指标的物理存储结构无关。指标类型包括动态指标和静态指标，静态指标是指通过 ET 或批处理任务对指标进行计算，并写入数据集市供前台系统使用的指标。动态指标是指在系统或 API 服务对象发出请求后实时进行计算，并返回结果的指标。所有动态指标均由 Jupyter 实现计算逻辑，对外提供服务。金融指标库分成 6 大类，29 小类。

3.1.2 指标管理

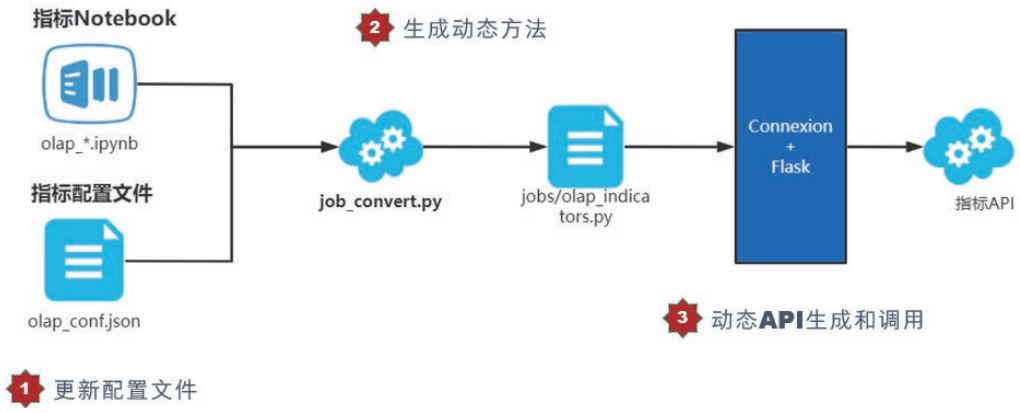
按照性质及生成方式的不同，金融指标分成事实维度类、查询计算类、统计计算类、模型计算类 4 种。

其中事实维度类为静态金融指标；查询计算类、统计计算类和模型计算类为动态金融指标，使用 Jupyter Lab 进行开发，将 Jupyter Lab 嵌入风险应用系统，通过算法编辑功能可直接跳转到 Jupyter Lab 中修改代码，所见即所得；通过审批后才可以发布使用。指标算法可以存档，并可使用版本控制回溯。使用 Jupyter Lab 的开发方式简要描述如右上图。

3.1.3 指标血缘分析

指标间互相依赖，存在复杂的血缘关系，关系的精准表达及存储是进行及时准确指标计算的基础，从而增加指标结果的可回溯性。无指标血

指标分类	实现方法	技术工具	指标举例
查询计算类	数据库查询统计为主，辅助基础数学运算	Python 库：Pandas，Numpy，Math 等	市值占资产净值比例、当日净现金流入
统计计算类	在基础数学运算的基础上需要进行回归、概率、相关性等统计计算	Python 库：SciPy，StatsModels，Pandas，Numpy，Math 等	Beta 系数、Sharpe 比例
模型计算类	金融模型计算指标	Campisi，Brinson，TM 等模型	归因相关指标



缘依赖关系则只能通过批量更新全部指标，无法支持细粒度及时调度和重计算。

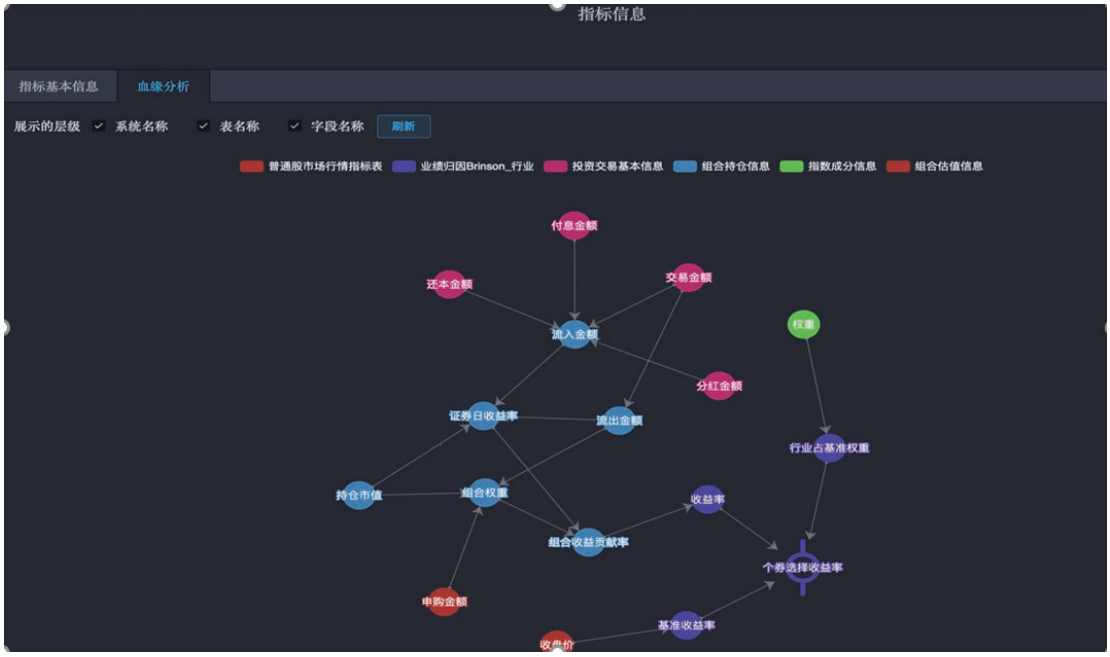
如下图所示:“个券选择收益率”由“收益率”、“行业占基准权重”和“基准收益率”三个业绩归因类指标计算测出。“组合收益贡献率”等 6 个组合持仓信息指标由“交易金额”等来自组合估值信息和投资交易基本信息计算得出。

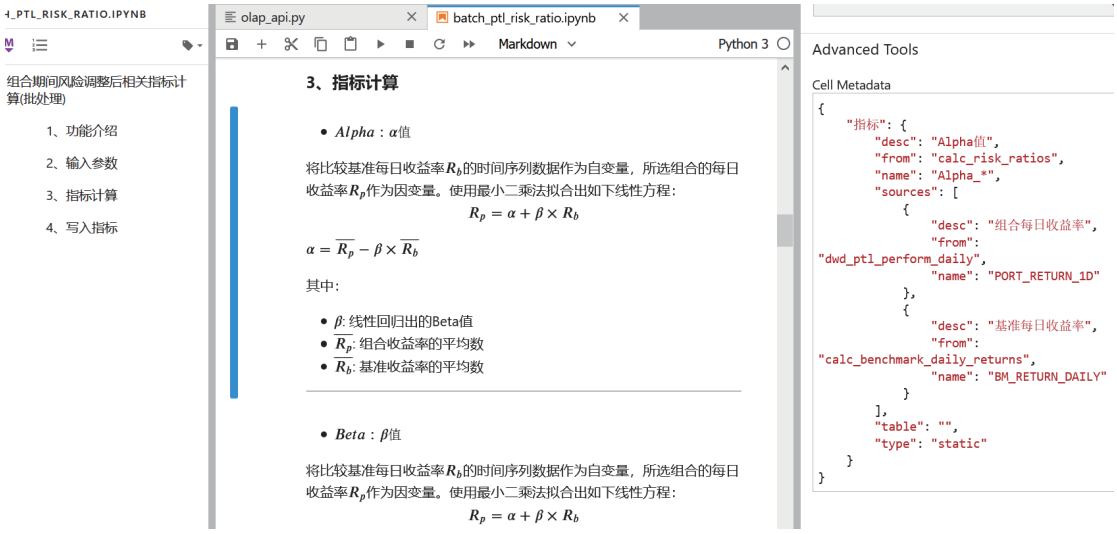
血缘分析分为两部分：1）ETL 把数据从外部数据源抽取导入到风险数据集市，通过解析 ETL SQL，获得 SQL 的数据逻辑，从而得到指标和基础层的表和字段的关系。2）金融模型计算指标通过 Python 开发模板（如图所示）上定义了每个指标的说明、算法、以及输入和输出。Jupyter 引擎解析之后实现生成文档、算法卡片以及血缘关系等。

3.1.4 指标计算引擎

由于 Jupyter Lab 具有交互性好，易于探索式开发和热部署等优点，从而方便业务人员、需求分析人员和开发人员协同进行指标研发、测试及后期维护。如下图所示，每个 Jupyter Lab 脚本由四块主要内容组成：

1. 指标介绍和描述

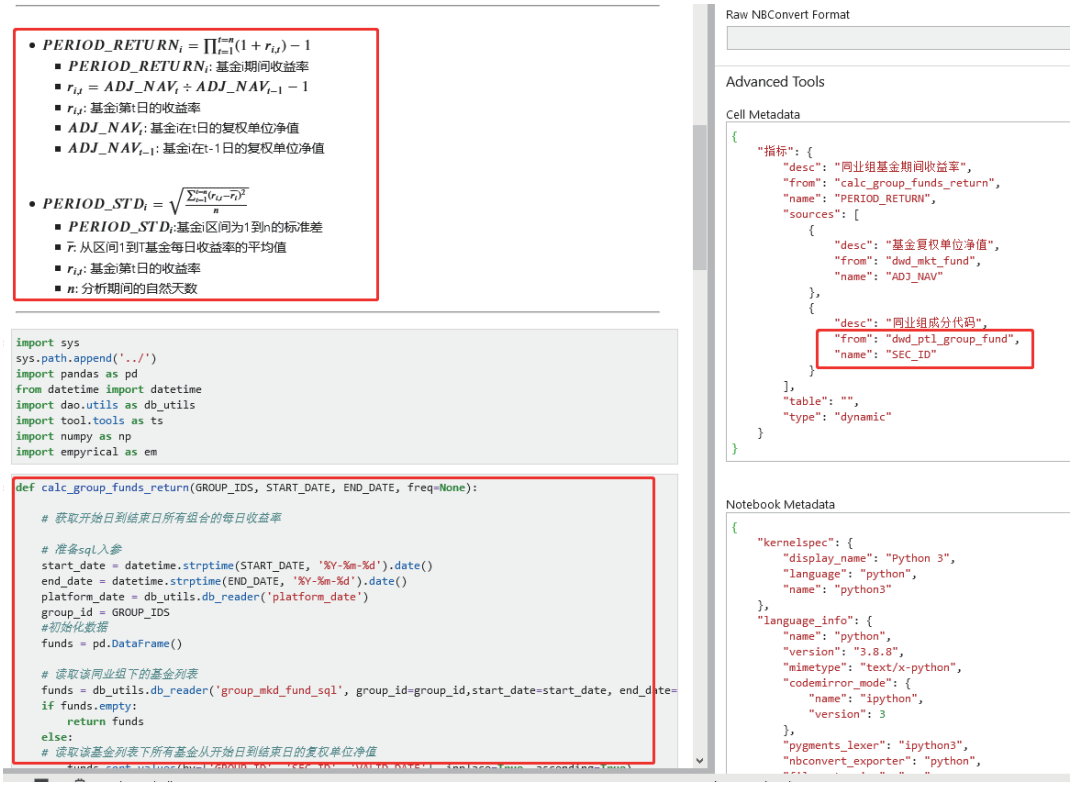




2. 指标的来源表 / 字段
3. 指标的算法描述
4. 指标的具体代码实现

3.2 自助数据服务

任何系统的功能和报表数量都是有限的，而风险前台应用的特点确要求对数据进行无限探索，固定的功能和报表不能满足风险前台应用对数据多样化的需求，在应用层面亟需支持用户自定义数据提取、加工、呈现的需求。数据服务支持用户可选择在指标库中的任何指标，指标可以来自一个或多个指标分类，多个指标分类之间需要定义各自的映射字段，中间结果集将以



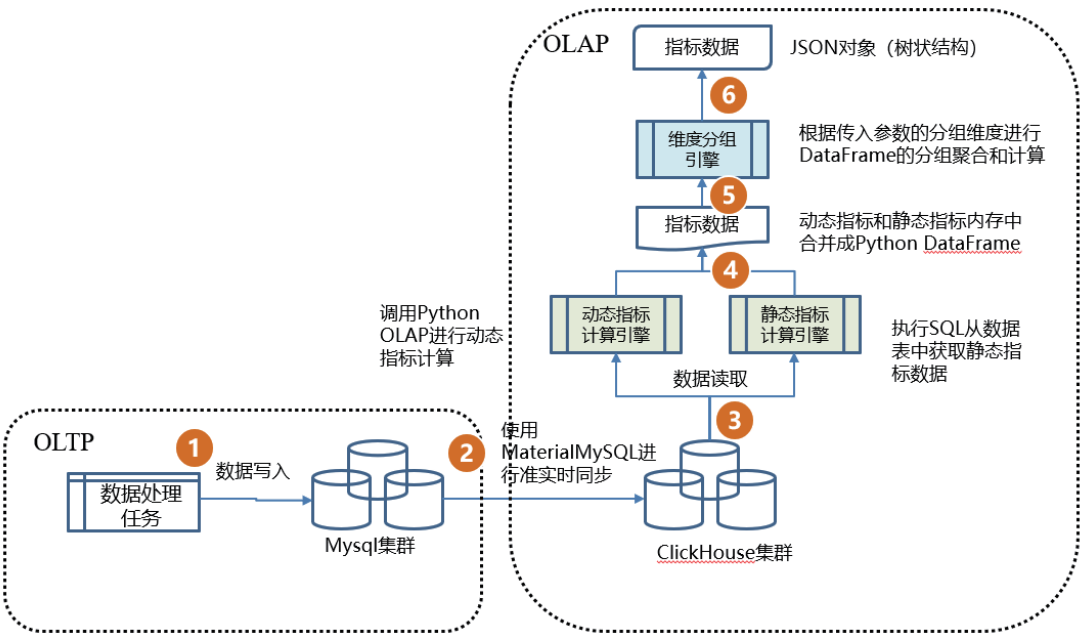
Dataframe 形式存在，在其基础上还可以定义排序规则，聚合和分组规则，以及筛选条件，最终结果以 Json 形式返回给前端应用或 API 调用端，使得数据定义和提取流程完全自助化。

四、数据处理和分析

整个数据处理和分析分为 OLTP 和 OLAP 两个流程：

风险管理系统的特点是需要从业务系统中快速读入大量数据，数据的读取次数远多于写入次数，支持用户进行任意维度的灵活探索，对数据做挖掘、分析，并生成报表，是一个对数据应用不断调整和持续优化的应用场景，是典型的 OLAP 应用。OLAP 类业务更关注写入吞吐，数据一旦导入完成，基本上不做更新和删除操作，对事务需求较少。ClickHouse 从 OLAP 场景需求出发，定制开发了一套全新的高效列式存储引擎；列式存储在分析场景下有着许多优良的特性，列式存储只需读取参与计算的列，IO cost 较低；提供更高的压缩比缩短磁盘中读取数据耗时，且对系统缓存使用的效果较好；在数据导入时全部是顺序写，充分利用了磁盘的吞吐能力，有着优异的写入性能。另外 ClickHouse 在计算层实现了单机多核并行、分布式计算、向量化执行与 SIMD 指令等工作，将硬件能力用到极致，极大地提升查询速度。而 OLTP 类业务对于写入延时要求高，因此采取 MYSQL 作为事务型数据库搭配 ClickHouse，实时从事务型数据库中进行数据同步的方案。

数据加载与处理任务通过任务调度系统集中管理，支持 ETL、Python 和 Java 等不同任务类型。通过定义任务组和任务的执行时间及优先级确定任务的调度次序。任务调度计划和执行结果以关系图的形式展现，任务执行失败将尝试预定义的重试策略并提供告警事件通知。通过 ETL 从上游系统通过文件形式加载数据，经过处理之后导入到数据集市；静态指标于 T-1 日终，Python 从 ClickHouse 读取数据，经过算法模型的计算出指标结果后，存入 MySQL 数据库。ClickHouse 挂载为 MySQL 的一个从库，借助于 ClickHouse 的 MaterializeMySQL 引擎，MySQL 的数据变动通过解析 binlog 的过程，将数据



同步至 ClickHouse。MySQL 和 ClickHouse 的数据同步是一个准实时的过程。在日间做报表查询分析过程中，只有读操作，或者有少量的数据修改操作，数据传输延迟性在毫秒级别。但是对于夜间的 ETL 抽数，每次有百万级别数据事务写入，这时可能会出现秒甚至分钟级别的延迟，因此在 ETL 结束后增加一个批处理任务监控 ClickHouse 是否已经完成了数据同步。

以下为数据写入 MySQL 的三个场景（见下表）。

前台应用所需指标的源数据从 ClickHouse 读取，ClickHouse 通过多线程执行 SQL，列存储顺序 IO 等特性提供高性能的复杂多表连接 SQL 执行效率。用户查询的指标由动态指标和静态指标构成。动态指标使用 Python 从 ClickHouse 获取源数据，实时计算出指标结果。该类型指标主要是通过区间查询开始日和查询结束日参数输入，实时计算时序指标结果集。静态指标通过 Python 从 ClickHouse 直接获取已存在的指标数据。静态指标一般分成两类：1）事实类指标：如证券外部评级为 AA，2）计算类指标：如该券前一天的市值，通过数量 * 价格，在 T-1 日的日终批处理已经过计算后存入数据库。该指标的结果和查询时间区间无关，只作用于某一天。Python 把静态指标和动态指标合并成一个 Data Frame 数据集，根据查询时传入的聚合维度和聚合算法参数，对数据集进行聚合处理，以树形或者二维数据集结构的形式返回给应用系统。

五、系统性能（见右表）

5.1 日终任务

1）ETL 任务组负责外部数据加载并处理，共 183 项任务，每天增量处理的数据量为 4 千万条，执行时间为 47 分钟。

2）Python 金融指标计算任务组负责计算静态指标，共 23 项任务，执行时间为 5 分钟。

3）Java 限额计算任务组负责监管、产品和内控限额计算，共 68 项任务，执行时间为 14 分钟。

5.2 系统功能

六、总结与展望

风险系统首期建设了风险分析、绩效归因、风险因子库、压力测试、情景分析、流动性风险分析等功能模块，提供直观集中的产品净值、持仓、风险和绩效情况的分析和展示。从长期来看，

功能	延迟性	方案
ETL	一次性事务百万数据，秒/分钟级别	增加监控批，待完成数据同步，再运行 Python 静态指标计算
Python 静态指标批处理	每次按照组合写入更新，毫秒级别	无需监控
日常配置	毫秒级别	无需监控

系统功能	时间区间	业务参数	平均用时
压力测试	1 天	组合数量：79 个； 情景：净值型情景	15.3 秒
VaR 分析	1 年	组合数量：86 个； 展望天数：1 天； VaR 方法：历史模拟法	1.81 秒
Brinson 归因	1 年	组合数量：5 个； 业绩基准：上证 50 指数 行业分类：申万一级	1.05 秒
Campisi 归因	1 年	组合数量：52 个；	13.62 秒

实现搭建具备量化风险因子库、建立各类风险分析模型，便捷地进行多种压力测试和虚拟组合试算分析，为投资提供前瞻性的分析结果；全面使用自动化系统报表替代手工风险报表，不断优化对理财产品的风险收益进行归因分析。

风险系统是典型的数据消费系统，业务数据源数据质量不高，会极大地影响风险数据计量的及时性和准确性。针对数据质量管理，期待建立例行化的长效运营体系，搭建数据质量管理端到端的闭环管理机制，做到事前治理，事中控制，事后预警相结合，全面地提升数据质量；同时建立多维度的数据监控体系，使例行化运营体系能持续运转。

底层数据关联关系复杂，指标数量多且粒度细；目前系统仅支持多维度查询，在风险数据可视化方面仍需进一步完善。

海通证券智能外呼应用研究

金鑫鑫、任荣、王东、王洪涛 / 海通证券股份有限公司
林金曙、齐海丰、梅锦 / 恒生电子股份有限公司

近年来，以大数据、自然语言处理为代表的新型信息技术迅猛发展，促进了人工智能的发展，并逐步进入商业领域应用。在金融行业中，智能客服发展较为成熟，已广泛应用在智能在线客服、智能语音导航、智能外呼等场景。

传统的人工客户服务，主要依赖对客服人员的培训和客服中心知识库的维护，实现对客户的主动和被动服务。根据实践，在这些服务场景下，大部分的客户咨询、业务问题是存在很大比例重合的，将人工智能应用于客户服务领域，实现大量简单、重复的工作交由智能客服来完成，不仅可以降低人工成本，也可以提升客户服务的标准化和效率，促进客户服务的智能化转型。

本文在简介智能化应用场景的基础上，具体阐述海通证券智能外呼场景下，语义引擎应用实践所面临的痛点及解决方案。

一、概述

结合客服中心的业务现状，目前主要有三种应用场景与智能客服结合比较紧密：1、智能在线客服、智能语音导航、智能外呼，场景简介如下：

1）智能在线客服：该在线客服主要依托 e 海通财 App、微信公众号展开。客户可以在“在线客服”服务中，唤起智能客服，建立与客服的对话。用户发起会话后，系统通过意图识别判断是否需要调用智能知识库，若是则进入智能回复模式，通过多轮对话尝试解决客户的业务问题。若遇到机器人无法回复的问题，也提供转人工服务的选项，用户选择后跳转至人工服务。

2）智能语音导航：客户通过拨打客服电话后，

可触发智能导航系统。该系统采用 ASR 语音识别及 TTS 语音合成技术，基于既定的业务场景应答模型，支持客户通过自然语言交互的方式直接在电话中与机器人进行沟通，机器人根据智能知识库关键字快速定位客户需求，直接进行业务回复，或将用户导航至特定功能节点。并对接传统 IVR，实现导航菜单的扁平化，缓解传统 IVR 导航菜单层层嵌套用户等待时间长的问题，并在智能导航时保留转人工服务的选项，提升电话接通率和客户满意度。

3）智能外呼：通过智能外呼服务，实现用机器人来模拟真人坐席，按模型设计回访话术、应对技巧，对投资者进行业务通知类回访，引导投资者获取通知及风险揭示，减轻人工回访工作压力，提高回访覆盖率。同时为了提升客户体验，可设置相关回访策略，以支持语音中断后转人工、发散性问题回复及主流程返回引导等功能。

关于智能外呼应用场景，在证券行业主要包括中签通知、状态过期通知、适当性回访、办理业务回访等业务类型。客户回访作为客户服务的重要组成部分，一方面是客户服务的重要手段和形式，另一方面也长期面临着服务资源匮乏和客户回访量大的两难问题。

随着证券业务的快速发展，新股发行常态化，及内外部管理要求的细化，分支机构一线人员及客服坐席的客户服务压力陡增。这些大量且逻辑简单的外呼工作，恰恰可以由整个智能外呼系统来完成，相比传统人工外呼，智能外呼机器人有着并发支持外呼量大、无需休息、工作状态稳定、训练成本低等优点。同时根据业务场景扩展，持续优化智能外呼机器人模型，保证智能外呼服务的持续稳定高效运行。

二、证券智能外呼现状

当今证券智能外呼系统大都已经结合深度学习、语音识别、自然语言处理等 AI 技术，基本解决了传统外呼单纯依靠人工带来的诸多问题。然而，伴随着使用人工智能外呼客服，业务人员对现有的语义智能引擎相关的服务也有了更高的要求，同时对相应的 AI 技术也有了更大的挑战。

语义相似度匹配是智能客服领域 [1] 普遍使用的技术手段，也是 NLP 领域研究的热点话题。语义相似度匹配最早经常被视为一个二分类问题进行解决，比如文章 [2] 采用孪生循环网络解决二分类问题；之后，有研究者使用三元组形成 triplet-loss 的做法进行训练，目的区分相似与非相似的句子，比如文章 [3] 采用卷积网络解决三元组问题。然而，这两种方法都有一个很严重的问题：负样本采样严重不足，导致效果提升非常慢。

同样在外呼场景，系统也需要计算客户的回答与知识库中已经配置好的待匹配回答之间的文本相似度，从而在知识库中识别出与客户回答最相近的意图，并给出对应回复话术。其中，主要难点在于为知识库中配置多个语义相同的扩展问，工作量繁重；其次，训练满足上线使用要求的相似度模型，准确率低、泛化性弱、容易过拟合。

因此，目前证券智能外呼面临的主要痛点为：

1）语义相似性模型不能在新的特定业务场景中发挥很好的作用，容易失效，准确率低。由于业务粒度及要求不同或业务人员配置策略不同，两个句子是否相似或是否属于同一个意图也是不一样的，见下表。

样例	场景A	场景B	
	①询问来电意图	②是本人并询问来电意图	③非本人并询问来电意图
找我干嘛	√	√	
找他干嘛	√		√
找我啥事情	√	√	
找他什么事	√		√

如表所示，“找我干嘛”、“找他干嘛“在场景 A 中属于一个意图，而在场景 B 中不属于一个意图，需要区分；因此不同的场景，句子相似

的粒度也会发生变化，需要采用不同的相似性模型解决。

2）语义相似性训练语料准备工作效率低，比较耗时耗力。外呼业务场景众多且更替频繁，随着场景的变化，相似性模型需要高效的语料准备工作以及模型的快速迁移。

3）扩展问配置效率低质量差。业务人员往往配置的扩展问数量有限、结构单一，不能满足客户回答的多样性，配置效率以及质量都无法满足上线要求，最终导致外呼回访的成功率明显下降。

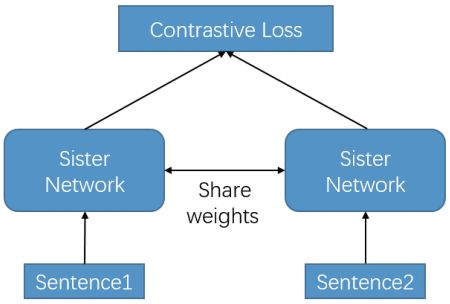
4）特定业务场景下的语料较少，模型极容易陷入过拟合的状态，丧失语义相似性的通用能力。

三、文本匹配算法及预训练模型介绍

（一）文本匹配算法

在智能外呼系统中，需要通过计算客户回答文本与知识库中的配置文本之间的语义相似度从而匹配到对应相似文本。这种计算文本之间的相似度的算法称为文本匹配算法。随着自然语言处理发展到深度学习阶段，文本匹配算法也步入到新的台阶，即“深度文本匹配”。深度文本匹配算法弥补了了统计学习模型在计算语义相似度时语义的缺失信息，进一步挖掘两个文本直接的语义相似度，因此效果更好。深度文本匹配方法主要分为表示型深度文本匹配模型和交互型深度文本匹配模型。表示型模型是先将两段文本转换成一个语义向量，然后计算两向量的相似度，其更侧重对语义向量表示层的构建，它的优势是结构简单、解释性强，且易于实现，是深度学习出现之后应用最广泛的深度文本匹配方法。基于交互式的匹配模型的思路是，假设文本之间的匹配度受到文本内局部特征匹配度的影响，因此在模型输入层就对文本编码表示做交互、对比，从而利用到了两个文本之间的局部特征结构信息。

1. 首先介绍表示型深度文本匹配模型，表示型深度文本匹配模型一般采用 Siamese 网络结构（即所有文本共享同一个模型网络参数），通过将文本经过深度网络模型编码成同一个尺度空间的向量，从而计算文本之间的相似度。典型的 Siamese 模型架构如下图所示：



根据 Sister Network 的不同，典型的表示型文本匹配模型主要有 DSSM、SiamLSTM、Multi-View、Sentence-BERT 等模型。

1.1 DSSM 文本匹配模型。DSSM 的全称是 Deep Structured Semantic Models，是匹配模型的鼻祖，由微软研究院在 2016 年提出。模型结构主要分为输入层、表示层、匹配层。输入层时将文本映射到一个向量空间中并输入到深度神经网络中。表示层是使用模型结构对输入层输入的向量做抽象抽取，得到一系列的特征向量来表示输入的文本。在匹配层，通过将表示层得到的向量做相似度匹配，进而利用极大似然估计构造损失函数。DSSM 模型的优点在于采用有监督的训练方法，准确度高，同时单个词或单个字处理不依赖切词的正确与否。缺点在于词向量的表示使用词袋模型，不考虑词语之间的位置信息，损失了很多语义信息。

1.2 SiamLSTM 模型在 Siamese 模型架构中采用 LSTM 作为 Encode 模型，通过共享权重的 LSTM 编码，将文本映射为同一空间中的向量，并采用 Manhattan 距离计算损失。SiamLSTM 模型通过使用 LSTM 网络，将文本中字词之间的结构信息加入到文本的表示向量中，使得文本向量表示语义更丰富。

1.3 百度在 EMNLP2016 中针对多轮对话的文本匹配问题，提出了 Multi-view 的 Q-A 匹配方式，输入的 query 是历史对话的拼接，分别编码了 word sequence view 和 utterance sequence view 两种表示。词级别的计算和 SiamLSTM 差不多，都是用 RNN 的最后一步输出做 Q-A 匹配，而句子级别的会对 RNN 每步输出做 max pooling 得到句子表示，然后再将句子表示输入到 GRU 中，取最后一步作为带上下文的表示与回答文本匹配计算。

1.4 Sentence-Bert 模型。Sentence-Bert 分别采用孪生网络双塔结构和三胞胎网络来更新模型参数（三个模型共享参数），通过将预训练语言模型 Bert 引入到模型训练中作为模型的 Encoder 部分，进一步丰富语义向量的表达，提高文本特征提取的能力。Sentence-Bert 的优点是将预训练语言模型和孪生网络架构相结合，进一步丰富文本的向量表达。

2. 下面介绍基于交互式的深度文本匹配模型。表示型的模型更侧重于对表示层的构建，缺点是分别从两个文本对象中单独提取特征，很难捕获文本之间的局部结构关联信息。交互型的深度语义匹配模型将文本的局部特征信息提前进行交互对比生成交互特征，从而提高匹配准确度。交互型匹配模型的主要算法有 ARC-II、MatchPyramid、ESIM 等。

2.1 在 ARC-II 模型中，将文本中的每个词表示为词向量后，每个句子构成一个矩阵，通过滑动窗口在矩阵中选择一个或多个词的向量组成词向量组。将两个文本中选择的向量组做卷积操作，通过两两组合的卷积操作构造成一个 2D 向量矩阵。以这个 2D 矩阵为基础在进行多次卷积和池化操作，最后得到一个表达两个文本之间关联度的向量并输入到 MLP 中来综合这个向量的每个维度得到的匹配值。ARC-II 模型考虑了句子中词的顺序和交互信息，但是缺乏对细微匹配关系的捕捉，匹配精准度上存在缺陷。

2.2 MatchPyramid 模型重新定义了两段文本的交互方式，即重新构造模型中的匹配矩阵。MatchPyramid 模型的核心思想是使用层次化思想构造匹配矩阵。类似于 CNN 在图像处理方面的原理，CNN 通过提取图像中像素、区域的相关性从而提取图像中的特征，MatchPyramid 模型将每个词语当作一个像素，对于两个单词数为 K,V 大小的句子，构造出一个大小为 K*V 的相似度矩阵。在匹配矩阵构造中，采用三种匹配矩阵构造方法：a. 在相似度矩阵中，对于两个序列中词相同的值置为 1，不相同置 0。b. 使用预训练的词向量将词转换为词向量并计算两个序列中每个词语之间两两的 cosine 相似度，填充到相似度矩阵中对应的位置。c. 与 b 类似，计算两个序列中两两词语向量之间的点积距离作为相似度矩阵中对应位置的值。对于得到的相似度矩阵，采用两层 CNN 对相似度矩阵进行特征提取，最后用两层全连接层对 CNN 得到的特征结果进行转换，进而使用 softmax 函数得到分类概率构造损失函数。MatchPyramid 模型的特点是采用多层的卷积，能够在单词或者句子级别自动捕获重要的特征。

2.3 ESIM 模型综合使用了 BiLSTM 和注意力机制，包含四个部分：输入层、交互层、聚合层和预测层。ESIM 的输入层采用与训练好的词向量或者初始化 embedding 层，然后接 BiLSTM 结构对词向量做特征提取。然后再用 attention 机制计算文本 a 中某一个单词和文本 b 中各个单词的相似度权重，再将权重赋给 b 中的各个单词的词向量，从而用加权后的 b 中各个单词的词向量来表征 a 中该单词，从而对 a 文本形成一个新的向量表征序列，文本 b 依然。简单理解是，假设 a 中有个单词“开户”，首先分析这个词语与文本 b 中各个词语直接的联系，计算得到的结果标准化之后作为权重，利用文本 b 中各个词语的词向量结合权重来去表征单词“开户”。对得到的新的向量序列做差异分析，判断两个句子之间的关系是否足够大，在模型上通过将得到的 a 的新向

量序列和 b 的新向量序列分别与原来 a 的向量序列和 b 的向量序列做差和积操作，并将得到的结果和原向量序列做拼接，形成新的向量序列。在聚合层对得到的新序列向量再用一次 BiLSTM 编码，将新序列的信息融合起来。模型的预测层是将聚合层的输出做 pooling 操作，并将 pooling 后的结果做拼接输入到分类器中做预测。

通过上面深度匹配算法模型的相关介绍可知，表示型的深度匹配模型更侧重于对表示层的构建，其特点是对将要匹配的两个句子分别进行编码与特征提取，最后进行相似度交互计算。缺点是分别从两个对象单独提取特征，很难捕获匹配中的结构信息。交互型的深度匹配模型在字词粒度对两个句子的特征做交互关联匹配，能够更精细的捕捉句子之间的关联关系。在工业生产环境应用中，多采用表示型深度匹配模型来作为文本匹配相似度的计算。受制于模型结构的影响，交互型的深度匹配模型无法满足系统实时性响应，因为交互型的深度匹配模型需要同时将待匹配的两个或者多个文本同时输入到模型中，而表示型深度匹配模型由于孪生网络结构共享参数的机制，可以提前将待匹配的所有文本的模型表征向量提前缓存到内存中，从而能够满足系统实时性的响应要求。本文提出的文本匹配改进算法也是属于表示型的深度文本匹配模型。

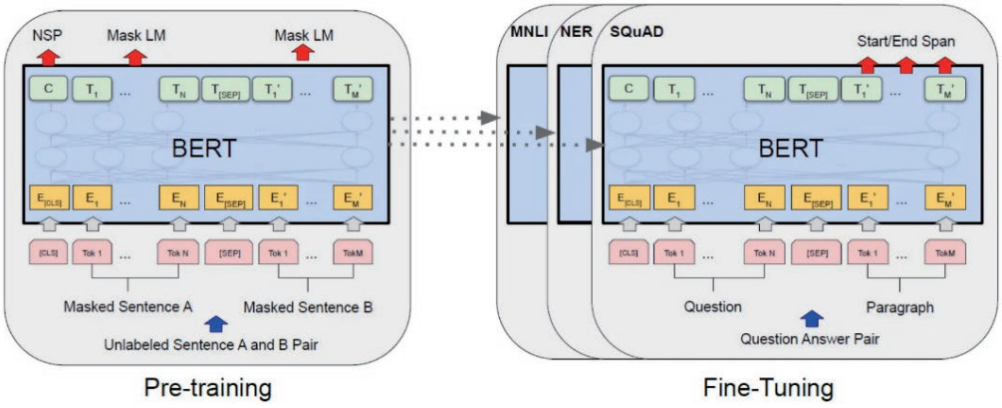
（二）预训练语言模型

在 NLP 研究领域中，随着计算机算力的增强，预训练语言模型不断的涌现出来，并形成一种新的 NLP 范式，即使用大规模文本语料库进行预训练得到预训练语言模型，然后再针对特定的下游任务对训练好的预训练语言模型做微调，从而避免了从零开始训练模型的步骤，降低了 NLP 任务的模型训练难度的同时也提高了 NLP 任务的模型性能。预训练模型的发展可以分为两个时代：1. 旨在学习词嵌入的预训练模型，这些模型往往采用浅层网络，如 word2vec、Glove 等。这些模型虽然可以捕捉词的语义，但是由于未基于上下午环境，不能捕捉到更深层次的概念。2. 第二代预训练模型是专注于学习上下文的词嵌入的预训练语言模型，如 ELMO、BERT、Albert 等。有别于第一代模型，第二代预训练模型参数更多，结构更复杂，能够捕捉语料中的上下文关系，文本中的位置结构关系等。同时训练好的模型可以直接拿来做下游任务的训练。本文所提出来的深度文本匹配算法就是在 Albert 预训练语言模型的基础上做的改造创新，下面介绍一下本文涉及到的预训练语言模型 BERT、Albert 模型。

1. BERT 预训练语言模型。BERT 模型是 Google 在 2018 年提出的一种 NLP 模型，成为最近几年 NLP 领域最具有突破性的一项技术。在 11 个 NLP 领域的任务上都刷新了以往的记录。下面从模型结构、训练方法及任务、优缺点等方面展开介绍。

1.1 BERT 模型结构介绍。BERT 模型是在 Vaswani 提出的多层双向 Tranformer 模型结构上发展而来，利用了 Transformer 模型的 Encoder 模型结构部分，但是比原始 Transformer 结构更深。Transformer 的 Encoder 模块包含 6 个 Encoder block，BERT-base 模型包含 12 个 Encoder block，BERT-large 包含 24 个 Encoder block。下图是 BERT 模型训练的结构图，左侧的图表示了预训练的过程，右边的图是对具体任务的微调过程。

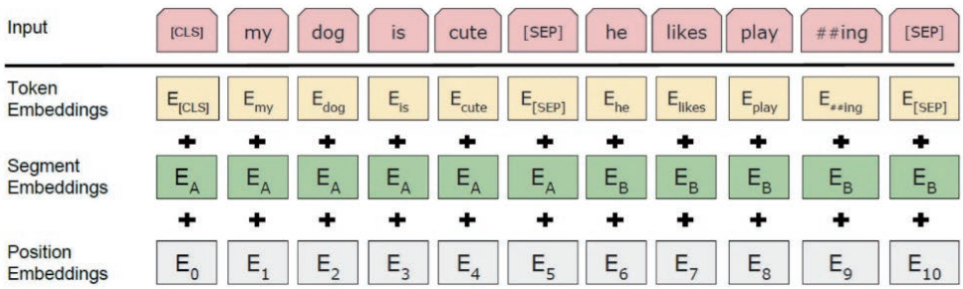
BERT 的输入可以表示一个单独的文本序列，也可以表示一对文本序列。另外，BERT 在训练时增加了一些有特殊作用的标志位：[CLS] 标志放置在第一个句子的首位，经过 BERT 得到的表征向量可以用于后续的分类任务；[SEP] 标志用于分开两个输入句子，例如输入句子 A 和 B，要在句子 A 和 B 后面增加 [SEP] 标志；[MASK] 标志用于遮盖句子中的一些单词，将单词用 [MASK] 遮盖后，再利用 BERT 输出的 [MASK] 向量预测单词是什么，属于 BERT 的一个预训练任务。例如给定两个句子 "my dog is



cute" 和 "he likes palying" 作为输入样本，BERT 会转为 "[CLS] my dog is cute [SEP] he likes play ##ing [SEP]"。BERT 里面用了 WordPiece 方法，会将单词拆成子词单元 (SubWord)，所以有的词会拆出词根，例如 "palying" 会变成 "paly" + "##ing"。BERT 得到要输入的句子后，要将句子的单词转成 Embedding，Embedding 用 E 表示。与 Transformer 不同，BERT 的输入 Embedding 由三个部分相加得到：Token Embedding，Segment Embedding，Position Embedding。如下图所示。

其中，Token Embedding 代表词的 Embedding 向量，通过模型学习而来。Segment Embedding 用于区分每个单词属于句子 A 还是句子 B，也是通过学习而来。Position Embedding 用来编码单词的位置信息，与 Transformer 使用固定的公式计算不同，BERT 的 Position Embedding 也是通过模型学习得到的。

1.2 BERT 模型预训练。在 BERT 模型的预训练阶段，BERT 采用两个无监督预训练任务：Masked LM 预测被 mask 的单词的训练任务和 NSP（Next Sentence Prediction）预测下一个句子的训练任务。Mask 操作是语言模型训练的常规操作，在 Word2Vec 中，CBOW 通过单词 i 的上下文信息来预测被 mask 的单词 i，不过这种训练方式不考虑语序，采用的是词袋模型。语言模型 ELMO 在训练时使用 BiLSTM，在预测被 mask 掉的单词 i 时，对于 i 后面的单词，前向 LSTM 会全部 mask 掉，只使用 i 前面的前向 LSTM 信息来预测单词 i。同时 i 前面的单词被后向 LSTM mask 掉，只使用后向 LSTM 的信息来预测单词 i。因此 ELMO 是将上下文信息分隔开进行预测的，而不是同时利用上下文信息进行预测。而 BERT 的作者认为在预测单词时，要同时利用单词的上文和下文信息才能更好的预测，因此提出在 Transformer Encoder 结构上训练出一种深度双向模型，直接使用上下文信息来预测被 mask 掉的单词。但是在后续使用预训练模型时，由于句子中并不会出现 [mask] 的单词，为了防止模



型性能的下降，在做训练时采用随机 mask 的策略。随机选择句子中 15% 的单词进行 Mask，在选择为 Mask 的单词中，有 80% 真的使用 [Mask] 进行替换，10% 不进行替换，剩下 10% 使用一个随机单词替换。以上就是 BERT 的第一个预训练任务 Masked LM。BERT 的第二个预训练任务是 Next Sentence Prediction (NSP)，即下一句预测，给定两个句子 A 和 B，要预测句子 B 是否是句子 A 的下一个句子。BERT 使用这一预训练任务的主要原因是，很多下游任务，例如问答系统 (QA)，自然语言推断 (NLI) 都需要模型能够理解两个句子之间的关系，但是通过训练语言模型达不到这个目的。BERT 在进行训练的时候，有 50% 的概率会选择相连的两个句子 A 和 B，有 50% 的概率会选择不相连得到两个句子 A 和 B，然后通过 [CLS] 标志位的输出来预测句子 A 的下一句是不是句子 B。

1.3 BERT 的优缺点。相较于以前的预训练模型，BERT 能够捕捉到真正意义上的上下文信息。缺点是做生成任务时因为和预训练过程不一致导致生成任务时表现不佳，而且由于输入噪声 [MASK] 造成预训练和精调阶段的不一致影响到模型性能表现。

2. Albert 模型简介。Bert 提出后，提高了很多 nlp 任务的 baseline，但是由于 bert 参数量巨大，导致在工业生产环境中做模型推断时耗时很大。为了解决上述问题，Albert 应运而生。Albert 模型结构与 Bert 无异，但是 Albert 使用一个新的自监督 loss 函数，提出的新的 SOP 预训练任务更能学习到句子间的内部特征。并且 Albert 提出两种模型参数机制缩小了预训练模型的参数量。

2.1 Albert 模型参数压缩方法。原始的 BERT 模型以及各种基于 Transformer 的预训练语言模型都有一个共同特点，及 Embedding 的维度 E 和隐藏层的维度 H 相同，一旦增加了 H，E 的维度也增大了，最终导致参数量呈平方级的增加。Albert 模型将 E 和 H 进行解绑，提出向量参数分解法，将一个非常大的词汇向量矩阵分解为两个小的矩阵。例如词汇量大小是 V，向量维度为 E，隐藏层向量为 H，则原始词汇向量参数大小为 V*H。Albert 不直接将 one-hot 向量映射到隐藏层，而是将 V*H 分解为两个矩阵，因此原来的参数量 V*H 变为了 V*E+E*H，E 是一个远远小于 H 的维度。这做法大幅降低了模型参数的参数量。在 BERT 模型中，模型的参数有 20% 在 word embedding 映射层，80% 的参数量都在 transformer 模块中。为了减少模型参数量，Albert 在提出另一种模型参数缩减的方法，即参数共享，由于 BERT 模型是由多层 Transformer 的 encode 部分堆叠而来，模型参数随着 Transformer 层数的增加而增加。Albert 提出在多层 transformer 单元共享参数，避免了模型随着深度的增加带来参数量的增大。

2.2 新的预训练任务 SOP。从上文可知，BERT 模型训练的 loss 由两个预训练任务组成，即 mask LM 和 NSP，maskLM 通过预测 mask 掉的词语来实现真正的双向 transformer，NSP 类似于语义匹配的任务，预测句子 A 和句子 B 是否匹配，是一个二分类的任务，其中正样本从原始语料获得，负样本随机负采样。NSP 任务可以提高下游任务的性能，比如句子对的关系预测。但是 NSP 任务太过简单，导致模型没有真正学习到句间的关联关系，Albert 提出一个新的预训练任务 SOP(sentence-order prediction)。SOP 关注句子间的连贯性，而非句子间的匹配性。SOP 正样本也是从原始语料中获得，负样本是原始语料的句子 A 和句子 B 交换顺序。 举个例子说明 NSP 和 SOP 的区别，原始语料句子 A 和 B，NSP 任务正样本是 AB，负样本是 AC；SOP 任务正样本是 AB，负样本是 BA。可以看出 SOP 任务更加难，学习到的东西更多了（句子内部排序）。

2.3 通过上面的介绍可知，Albert 克服了 BERT 模型参数量大、预训练任务简单等缺点。实验表明，在大幅缩减模型参数的情况下，

Albert 模型的性能相较于 BERT 仅仅下降 1-2 个点，但是模型推理速度却提高了一个数量级。

四、方案流程设计

针对目前智能外呼的痛点，本文在 Albert 的基础上设计了一个兼相似性检索与生成的语义相似性（预训练）模型框架，并采用 few-shot learning 的训练方法，通过弱监督信号对特定业务场景进行微调，既达到上线要求的准确率与泛化性，也满足模型快速迁移的需求。

方案流程见下图。

方案流程主要包括三个步骤。

步骤 1：利用预训练的生成能力，通过 beam-search 方法，结合人工勾选或修改，快速形成特定场景的知识库。

步骤 2：利用特定场景的知识库，在预训练模型的基础上快速微调，训练出适合该特定场景的相似性模型。

步骤 3：利用知识库以及微调模型，对用户提问进行相似性计算，准确识别意图。

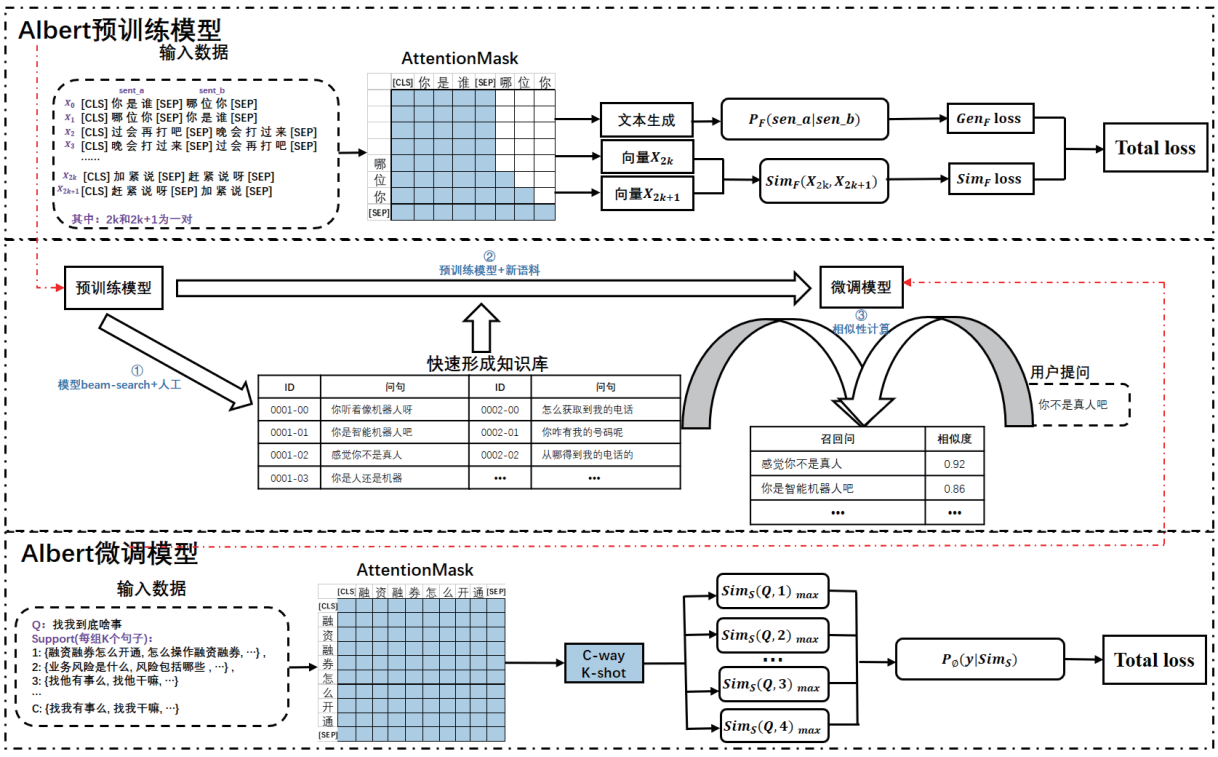
该方案流程的构思创新点：

1）预训练模型利用构造特殊的 Attention-Mask，在相似性任务的基础上，增加生成任务，充分利用文本相似与生成任务 [4] 之间的关联性，增加训练难度，可以明显提高相似性的准确率，也可以为业务人员提供大量多样性的扩展问。

2）微调模型结合 few-shot learning[5] 训练方法，采用弱监督信号，允许在特定业务场景语料少量的情况下，快速训练相似性模型，可以避免模型过拟合而失去相似性通用能力的问题。

3）在预训练模型中，每个 batch 数据有特别设计，按照一定比例加入了金融语料与通用语料，再次增加训练难度。恒生收集的 FAQ、外呼等金融场景的语料共 50 万组，通用语料共 712 万组。

4）整体方案流程节省大量计算资源以及人力成本，可根据特定业务场景快速迁移，灵活方



不同方案	是否重新标注	扩展问语料准备			模型训练		时间总成本
		时间消耗	数据量	数据质量	可扩展性	迭代次数	
传统分类方案	是	5d	5000	单一	需要足够的标注语料才能保证效果，足够的语料就需要大量的语料准备时间	3次以上	10d
传统相似性方案	否	2d	200	单一	无需训练，但是扩展性差，模型不适用于新场景		2d
本文方案	否	0.5d	200	多样性	扩展性强	1-2次	1d

便。假设一个业务场景 5 个问题，则从语料准备到训练出可上线模型所花费成本，与传统分类、传统相似性方案相比见上表。

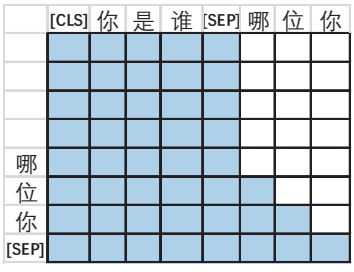
五、算法设计

（一）预训练模型

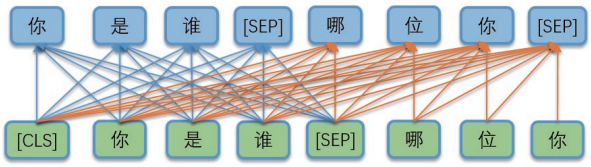
考虑到模型推理的性能问题，模型框架在 Albert[6] 基础上进行改造。

Albert 模型结构采用 BERT 作为 backbone，对嵌入字向量参数进行因式分解以及 encoder 跨层参数共享，参数量缩小为 1/18，并采用 SOP（Sentence-Order-Prediction）训练任务，用于解决原版 BERT 中 NSP（Next-Sentence-Prediction）任务损失低效的问题，因此 Albert 在不失精度的情况下可以保证快速的推理能力。

假如输入“你是谁”，相似句为“哪位你”，那将这两个句子拼成一块：“[CLS] 你是谁 [SEP] 哪位你 [SEP]”，则对应的 Attention-Mask 为：



其中“[CLS] 你是谁 [SEP]”这几个 token 之间是双向的 Attention，而“哪位你 [SEP]”这几个 token 则是单向 Attention，从而允许递归地预测“哪位你 [SEP]”这几个 token，所以它具备文本生成能力，生成过程见下图。



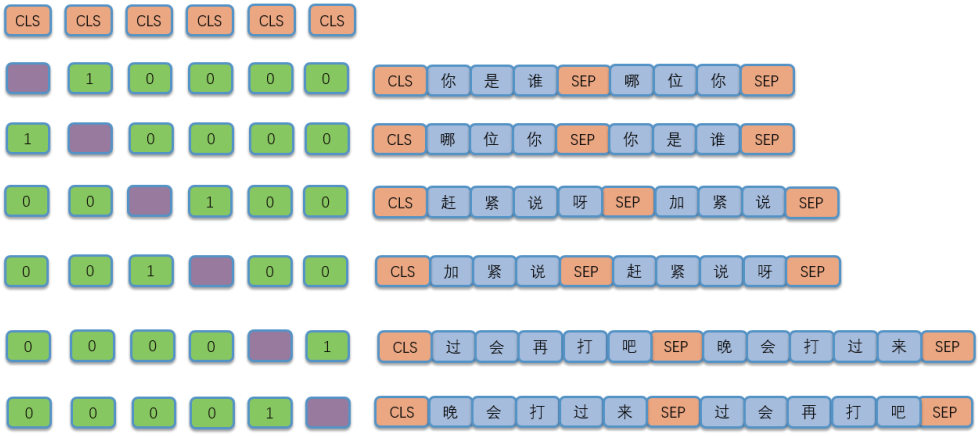
其中生成任务中的损失函数为：

$$Gen_F\ loss = \sum_t loss_t,$$

$$\text{loss}_t = 1 - P(y_t|y_0, y_1, \ldots, y_{t-1})$$

上述公式中，t 指预测第几个字，y_t 指预测的第 t 个字。

其次，“[CLS] 你是谁 [SEP]”这几个 token 只在它们之间相互 Attention，与“哪位你 [SEP]”没有关系，因此“[CLS]”对应的向量编码代表着“[CLS] 你是谁 [SEP]”的句向量。利用句向量，可以把整个 batch 内的“[CLS]”向量取出来，得到一个句向量矩阵 V，然后两两做内积，得到 b × b（b 指 batch_size）的相似度矩阵，并 mask 掉对角线部分，最后每一行进行 am-softmax，如下图：



上图一个 batch 中 6 个样本，1 代表相似、0 代表不相似，则将相似性问题转化为分类问题，当一个 batch 的数量较大时，即把 batch 内所有的非相似样本都作为负样本，这样就可以解决之前负样本采样不足的问题。

设 z 为模型最后一层的输出向量，W=(c_1,c_2,…,c_n) 为权重向量，则利用 am-softmax 得到相似度任务的损失函数为：

$$Sim_F\ \text{loss} = -\sum_n y_i \log(p_i),$$

$$p_i = \begin{cases} \dfrac{e^{s\times(\cos<z,c_t>-m)}}{e^{s\times(\cos<z,c_t>-m)} + \sum_{i\neq t} e^{s\times cos<z,c_i>}}, i = t \\ \dfrac{e^{\cos<z,c_i>}}{e^{s\times(\cos<z,c_t>-m)} + \sum_{i\neq t} e^{s\times cos<z,c_i>}}, i \neq t \end{cases}$$

其中，y_i 为第 i 个样本的真实标签，t 为目标标签，s 为比例缩放参数，默认值为 30，m 为一正数，默认值为 0.35。

最终模型的损失函数为：

Loss=Gen_F loss+Sim_F loss。

最终在近 760 万组相似性数据中，预训练模型训练近 4 周，模型在 10 万条测试数据中，top1、top2、top3 的准确率分别为 0.939、0.945、0.974。

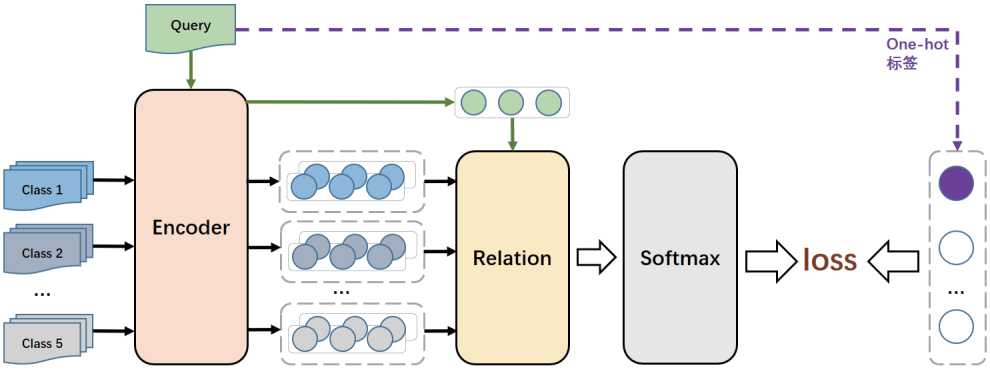
（二）微调模型

在外呼场景中，一个意图下的扩展问语料有时不具有语义层面的相似性，比如“忙”这个意图，扩展问有“开车呢”、“正在打比赛”、“单位忙着呢”等；如果在微调模型中，使用强监督信号训练语义相似性，势必会严重破坏预训练模型中的文本语义相似结构，导致微调模型效果不佳。

因此，本文微调模型采用 few-shot learning 的训练方式，通过弱监督信号，对特定场景的少量语料进行快速训练，旨在防止过拟合的前提下，有效学习不同意图之间的差异。

few-shot learning 算法研究主要集中在小样本图像识别的任务上，目标不是识别图像的类别，而是识别出不同图片之间的差异；近年来，在自然语言处理领域也开始出现 few-shot learning 的场景，比如问答领域的意图识别。few-shot 的训练集通过构建 support set 以及对应的 query 完成训练任务。这里 support set 包含 5 个类别（即 5 个意图），每个类别下含有 5 到 20 个样本，即训练时采用 5-way 20-shot 的方式，之后选取这 5 个类别中的一个样本作为 query（不在 support set 中），以这种方式，构造一系列的 episode 来训练网络，最后由 softmax 层获取训练误差，过程见下图。

图中 Encoder 指句子向量编码层，Class 指不同的意图类别，Relation 指各意图的扩展问与 Query 进行相似性计算，分别得到每个意图的最



大相似性，损失函数如下：

$$Loss = -\sum_{n} y_i \log(p_i),$$

$$p_i = \frac{core_i}{\sum_n core_i},$$

$$core_i = \max\{\text{Relation}(\text{Query}, \text{Class}_i)\},$$

其中 Class_i 指第 i 个意图，core_i 指 i 意图与 Query 的最大相似性，y_i 指是否属于该意图（属于为 1，不属于为 0）。

利用此微调模型，可以在不破坏相似性通用性的情况下，区分“找我干嘛”、找他干嘛以及“我同意”、“我不同意”等细分意图。

六、总结与展望

本文对证券智能外呼客服的现状以及痛点进行了系统的梳理与阐述，根据智能语义引擎中的应用实践，提出预训练 + 微调的相似性方案，赋予预训练模型相似句生成的能力，并对微调模型采用 few-shot lerning 的训练方式，通过弱监督信号在样本少量的情况下，快速训练出特定业务场景的相似性模型。

通过模型的持续训练结合业务场景挖掘，目前海通证券智能外呼服务已先后上线 6 个服务场景，最大并发达到 150 路，累积执行任务数超过 200 万次，剔除无人接听、忙线、错号等情况，平均成功率可达到 90%。在人工成本节省上，根据外呼电话时长核算，平均每个交易日可节省 25 个专职人员的工作量，有效的实现了释放分支机构一线人员和坐席资源到高级客户服务及客户拓展。

在人工智能技术不断发展完善的大背景下，基于外呼效果持续跟踪和用户需求不断迭代模型，优化知识训练和应答逻辑，提升智能外呼机器人的准确度和效率；持续扩展外呼场景，进一步提升回访成功率和业务覆盖率，实现客户与公司的双赢。