互联网金融背景下的大数据产品创设及机器人投顾模式的建立

中泰证券股份有限公司　股票雷达*

一、背景介绍

现有证券公司投顾业务主要是根据2010年10月中国证监会颁布的《证券投资顾问业务暂行规定》（以下简称《暂行规定》）开展。投资顾问在向客户提供服务的过程中，如果涉及向客户销售金融产品，须按照2012年12月实施的《证券公司投资者适当性制度指引》（以下简称《指引》）的规定对客户进行适当性管理。

互联网尤其是移动互联网时代，券商作为金融行业的重要参与者，势必也要转型迎接互联网金融浪潮。在上述背景下，大量用户从互联网渠道进入金融服务领域，券商客户结构本身正在发生剧烈变化，券商经纪业务呈现新的特征。

一是网上开户导致服务离散化。“服务离散化”是自网上开户实施以来，客户在全生命周期内与券商的社交黏性大为降低。客户在网上开户以后，所有产品买入和卖出操作均在网上直接完成，用户的整个投资过程甚至不需要与券商服务人员现场见面。但在移动互联网时代下，这些弱社交黏性、高频交易的小微客户，明显缺乏专业的投资经验。

二是互联网客户结构发生巨变。基于互联网平台开户的群体，呈现如下两个特征：第一，从新增的开户数来看，客户群体偏年轻化，“80后”、“90后”成为新增开户主力；第二，大批年轻客户涌入，让市场的长尾特征变得更为明显，新兴客户的理财行为也表现出新的特点。

这些特征使得券商现有的投顾服务模式很难满足客户的需求。在这样的背景之下，机器投顾的引入能够很好地弥补这一“需求缺口”。

---

* 小组成员：李肇嘉，何波，冯月，逯志军，李杜，石彦彦。原载于《中国证券》2016年第11期。

（1）机器人投顾的发展趋势——国外机器人投顾应用现状。随着大数据技术和移动互联网技术的发展，机器人投顾服务在美国已经迅速普及，有超过200家公司纷纷布局机器人顾问市场，这些公司包括新兴的机器人投顾公司Wealthfront、Betterment、Personal Capital等。花旗集团发布研究报告预测，在未来十年时间里，机器人投顾总额将达到5万亿美元。

我们从机器人投顾的用户风险偏好确定、服务模式、收费模式、投资标的选择以及目标用户定位5个维度，对14家国外机器人投顾公司的机器人投顾产品进行横向对比。

第一，用户风险偏好的确定。通过对国外机器人投顾模式研究发现，问卷成为确定用户风险偏好的主要方式。小部分公司如Vanguard、Rebalance IRA采用电话以及面谈的方式来了解用户的风险偏好。

第二，服务模式。国外机器人投顾从功能上提供两大类型服务，分别为“咨询建议型”和“资产管理型”，前者主要向客户提供一些证券分析和投资咨询建议，投资范围覆盖非标类资产、养老金规划和税收优化等场景；后者以代客理财形式，接受客户全权委托，基于投资者预期收益目标、风险承受能力等指标，提供相应的ETF基金投资组合。部分公司采用两者混合的服务模式。

第三，收费模式。从收费模式上看，多数公司都是以管理资产规模来确定管理费，部分公司设置免费管理资产规模上限。此外，现有公司均对投资咨询建议收取服务费。

第四，投资标的选择。由于客户结构、金融市场的开放程度以及投资品种丰富程度的差异，国外机器人投顾在投资标的的选择上范围相对更广。

第五，目标用户定位。各家公司推出的机器人投顾的客户资产要求在一定程度上能反映出平台对目标客户的定位。从机器人投顾公司来看，大多数公司对用户资产要求都比较低，仅少数公司资产门槛相对较高。

综上所述，国外机器人投顾在服务定位、客户管理和收费模式等方面已经相对比较成熟。国外机器人投顾的成熟经验，可被国内机器人投顾发展借鉴，但不能完全复制。结合国内应用现状，国内机器人投顾尚需做出大量改进，使得机器人投顾更能理解和适用国内投资者。

（2）机器人投顾的模式优势。较现有投顾业务模式，机器人投顾在客户定位、服务时长、服务手段、外部影响、成本费用和培育时间等方面具有较大优势（见表1）。

表1　　人工投顾和机器人投顾对比

对比维度	现有人工投顾	机器人投顾
客户定位	高净值人群	所有用户
服务时长	工作时间（5×8小时）	7×24小时
服务手段	电话、面对面访谈	移动互联网
外部影响	人为情绪因素强	公正、客观、透明
成本费用	边际成本高	边际成本几乎为零
培育时间	培育时间长，时效性低	学习时间短，深度学习能力强

资料来源：根据公开资料整理。

由此可见，相比于传统投资顾问，机器人投顾的优势显而易见。从短期来看，机器人投顾对现有投顾业务还未造成威胁，其应用还需要投资者根据自身情况进行平衡才能达到预期效果。但长期来看，客户的投资观念会逐渐发生变化，智能理财技术也会随着人工智能的进步而不断升级，这一替代趋势可能会不断被强化。

二、机器人投顾模式下的适当性管理

（一）机器人投顾模式介绍

1. 概念界定

机器人投顾模式是在大数据动态 KYC 的基础上，运用智能选配模型向投资者提供符合其自身风险承受能力及风险偏好的金融产品和服务的模式。

2. 特点介绍

机器人投顾模式主要特点是能够实现对客户风险承受能力的实时的动态调整。动态主要体现在：客户行为特征标签库的动态搜集与分类；客户画像系统日级别的调整；客户风险级别与金融产品的动态匹配调整；能够及时发现最大程度适合客户风险承受能力的金融产品。

（二）机器人投顾模式下的适当性管理——动态 KYC

1. 动态 KYC 介绍

（1）动态 KYC 概念。本文所指的动态 KYC 是对现有投资者适当性管理模式的创新，具体是基于公司大数据平台，通过对客户、产品数据的挖掘建立画像系统，从而实现动态了解客户、了解产品，及时识别客户的风险承受能力及风险偏好的变化，为智能选配系统的实现提供支持的系统。

（2）动态 KYC 与现行模式的差异。动态 KYC 是在现有投资者适当性模式基础上设计的，但是两者也存在明显差异。两者最重要的区别是：第一，从获取客户风险偏好信息来看，现行模式主要是通过问卷形式，在客户开户或者购买金融产品、享受金融服务前要填写风险调查问卷，而动态 KYC 获取客户风险偏好信息主要是通过算法模型抓取客户内外部行为特征，来实现客户画像；第二，从客户风险等级调整频率来看，当客户风险偏好发生变化时，现行模式无法及时调整，而动态 KYC 能够通过基于大数据的画像系统及时发现客户风险偏好的变化并及时调整其风险等级。

2. 机器人投顾模式与动态 KYC

本文所指的机器人投顾模式包括动态 KYC 的实现、智能选配模型构建两大模块。从模式设计思路来看，动态 KYC 的实现是机器人投顾模式的基础，同时也是智能选配的前提。

（三）动态 KYC 的设计思路

1. 整体设计思路

动态 KYC 整体设计思路是沿用现有模式：了解客户、了解产品。在设计动态 KYC 时，把问卷作为获取客户特征的途径之一。通过对客户行为数据的分析实现对客户的全方位了解。

在技术实现上，动态 KYC 是通过构建客户画像及产品画像系统来实现对客户及产品的

了解，通过建立大数据抓取系统，实时抓取客户行为特征数据，从而实现对客户风险承受能力及风险偏好的动态更新。

如图 1 所示，对客户的了解是通过调查问卷和客户画像系统来实现。调查问卷和客户画像系统分别对客户风险承受能力进行评级，前者的评级结果偏主观，后者偏客观。调查问卷主要是基于客户的主观判断划分风险等级，而客户画像系统是通过对客户真实行为数据的挖掘来对客户风险承受能力进行评级，同时获取客户的偏好信息。

当客户画像系统识别出的客户风险承受能力和风险偏好与调查问卷评级结果不一致时，系统会提示客户对原有风险评级结果进行调整。从而帮助客户了解其真实风险承受能力及风险偏好。当两种评级结果一致时，系统会自动根据客户的评级结果及偏好向客户推荐个性化的金融产品。

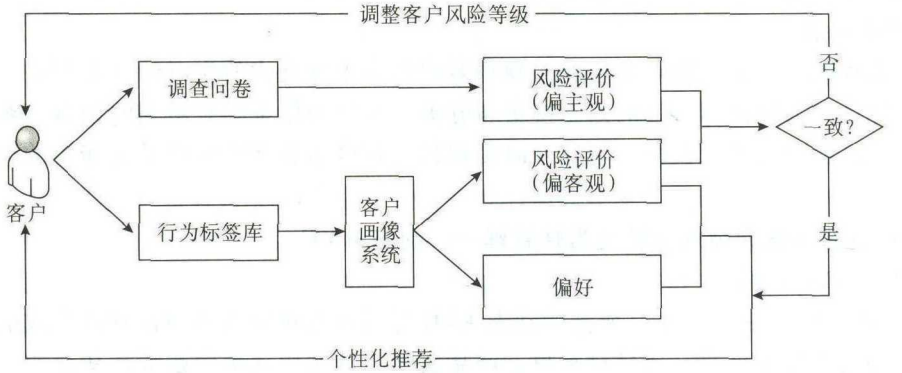

图 1　动态 KYC 逻辑图

2. 具体框架构建

（1）了解客户——客户画像系统（见图 2）。

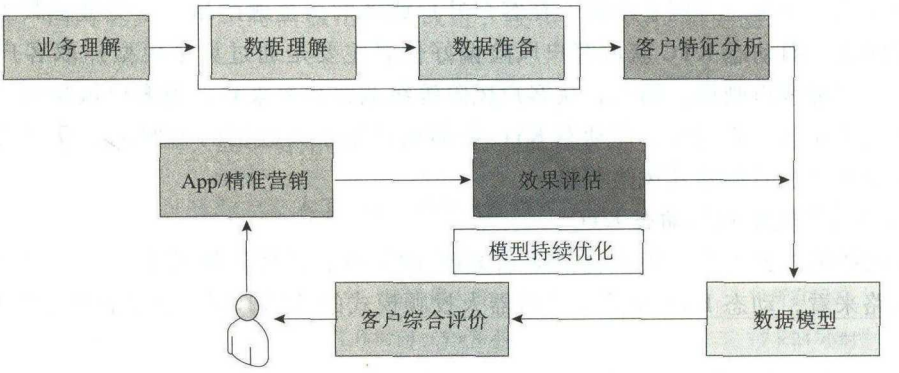

图 2　客户画像流程图

①客户行为特征的数据采集。动态确定数据源是整个 KYC 系统设计的关键，数据源在一定程度上决定了数据的质量。根据客户行为特征的属性不同，我们将客户行为特征数据标签划分为人口特征属性、交易属性、资产属性、互联网社交属性等维度。

②客户画像数据建模。客户画像的实现是一个数学建模的过程，客户行为数据标签库的

建立是为了便于计算机识别、运算。标签数据是整个画像实现的基础，标签数据的质量在某种程度上决定了客户画像系统的准确度。

③客户风险承受能力综合评估。在对客户画像的基础之上，赋予不同的行为标签以不同的评分和权重，根据标签分值及权重给客户风险承受能力综合评分。

④客户风险等级调整。当客户行为发生异常变化（超出临界范围），我们会根据影响标签的权重不同，适时调整该客户的风险等级，并通过会话式系统与客户确认。

（2）了解金融产品——产品画像系统。

①金融产品特征抽取。我们将金融产品特征划分为金融特征、销售情况和客户特征三个维度，并对每一个维度对应的特征数据进行提取，形成金融产品特征标签库。随着金融产品数据的不断丰富以及客户与金融产品交互数据的采集，金融产品标签库实现动态更新。

②产品画像及其应用。结合产品标签库，根据所筛选的评估指标与产品风险、收益的相关性构建产品的多维度评估表，从风险、收益等多个维度对金融产品画像，评估结果作为智能选配模型的输入因子。

③产品风险等级动态调整。金融产品画像并不是固定不变的，其风险和收益属性会随着市场环境的变化和行业发展阶段的变化而变化。产品画像系统动态监测产品标签和产品评估指标的有效性，对发生变化的立即进行调整。

（四）动态 KYC 的技术实现

客户和产品画像的精确度一是取决于标签数据库的完备性；二是取决于标签数据的有效性。可以说数据质量决定了整个画像系统的质量。

动态 KYC 是基于大数据用户和产品的画像系统，其实现从技术上来看，离不开大数据平台建设、数据抓取系统建设、算法平台的搭建；从业务上来看，离不开客户行为标签的抽取、标签权重划分、客户风险承受能力评估以及客户风险等级的调整等。

1. 公司客户/产品画像系统架构图

在公司大数据平台架构中，在数据集市[①]层根据业务模型和业务主题分为 12 个具体的业务模块，其中画像系统是属于业务模型中的模块之一。因为本项目除底层架构（数据平台层）通用之外，从业务应用场景来看，更多涉及客户、产品画像，所以我们绘制了基于公司大数据平台的客户/产品画像系统架构（见图 3）。

2. 客户画像系统的技术实现

（1）客户原始数据的采集。

①客户数据。客户数据是指客户的资料数据，包括客户的性别、年龄、家庭住址、学历、兴趣爱好，以及客户的持仓记录、风险等级等，这些客户资料为客户画像提供了基础数据。

②交易数据。交易数据包括客户的下单记录，客户的成交记录以及产品购买记录等。通过对客户基金以及理财产品的购买记录、客户信息以及交易记录进行关联，可以获取客户的行为和风险偏好。

① 数据集市就是一个为某个特殊的专业人员团体服务的数据的仓库，针对不同的部门，会创建不同主题的数据集市，包括基金、期货、股票、行业、公司等，以迎合各个不同部门对数据的具体需求。

图3 画像系统架构图

③外部金融数据。外部数据源包括行业数据、股票数据、公司数据、港股通数据。

④互联网金融数据。互联网金融数据是公司客户以及用户在移动App上、网页上留下的行为数据。这部分数据通过App埋点来采集。

（2）特征标签的属性划分。

从业务场景上来划分，将从各个渠道获取的客户数据划分为五大属性，分别为基础属性、交易属性、社交属性、金融属性以及互联网属性（见图4）。

图4 客户特征标签属性

（3）数据处理与客户标签提取。在完成数据收集过程后，我们要以此为据建立客户的行为标签，用以对客户风险等级和偏好进行评估。原始的大数据集中的数据质量难以保证，会有错误、冗余、缺失等情况发生，因此首先要进行数据清洗。然后，对数据进行同步。数据同步主要分为定期同步和实时同步。

（4）客户画像构建与风险评估。客户风险等级评估包括客户风险承受能力评估和风险偏好评估两方面，我们根据从原始数据中提取出的客户标签构建出精准的客户画像，并对其风险承受能力和风险偏好进行刻画。

（5）客户风险等级调整。客户在投资的过程中，随着市场表现、投资收益、投资经验、个人财务状况等各方面因素的变化，其风险承受能力与风险偏好可能会发生改变，因此我们的客户画像要随之进行调整，以满足推荐过程中的适当性要求。在客户风险等级调整方面，主要是将实时采集的客户数据代入模型进行计算，并以结果对客户标签画像进行定期的动态更新。

3. 产品画像系统的技术实现

产品画像和用户画像的实现逻辑大致相同，即将产品大数据通过算法操作实现产品特征的提取和分类，不同的是生成的标签库以及具体的产品画像体系的建立。

（1）产品画像系统标签库的建立。在产品画像特征提取部分，已经得到的是大数据平台里一系列未经处理的基础数据，其中既含有结构化的数据，也含有通过外部信息抽取得到的关于产品舆情热度的一些非结构化数据。本部分主要是介绍如何对这些数据进行处理，实现产品标签库的维度划分，并系统地对产品标签库的建立过程进行阐述。

①产品标签提取。产品标签提取时，关于金融产品基础特征的提取主要是根据各公司发布的文件对于产品属性的一些规定进行归类抽取；而一些抓取数据的标签化处理过程则是通过对内容的关键词提取，提取方法与用户标签提取过程类似。

②产品画像系统标签库。通过标签提取过程可以实现产品信息的标签化处理，标签化处理之后建立产品画像系统标签库，为后面对产品进行综合评价提供基础。经过处理后，产品标签库主要归类为产品金融特征、产品销售特征以及产品客户特征三个维度[①]（见图5）。

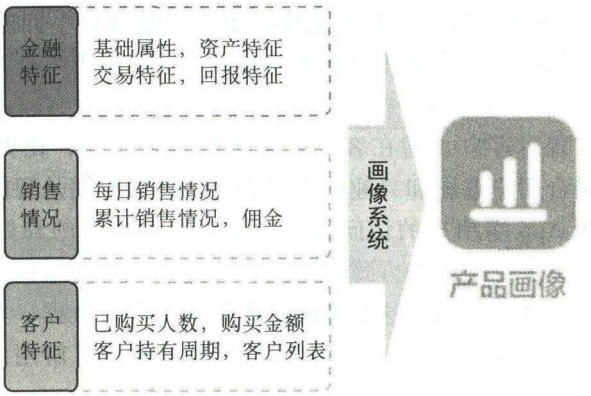

图5 产品画像标签属性

③产品标签库的动态更新。产品画像系统标签库建立以后，产品标签库时刻在更新。具体体现在，产品标签数据的动态搜集和产品标签有效性的动态评估与调整。

（2）多维度指标画像。金融产品作为客户的投资标的，其本身具有十分复杂的特征，如风险性、收益性和流动性等属性，且三种属性互相关联。流动性与收益性成反比，而风险与收益成正比，风险越大回报越高，客户通常期望其投资的金融产品流动性大、风险小而且收益回报高，但实际上这样的产品很难找到，客户选择产品只能根据自己的风险偏好和风险

① 产品特征的三个维度划分来自公司信息技术部门建设的产品画像标签库。

承受能力选择出符合自己预期的产品。

对产品进行画像，最大限度地挖掘出产品的各种属性特征，是了解产品的基础，也是保护投资者利益，对其投资需求进行最佳匹配的重要依据。

（3）画像系统的更新和优化。动态KYC的优化可以从以下几个方面进行：

①调整问卷内容。当用户表现出与现有画像不同的行为特征或偏好时，应该有针对性地对用户进行新的问卷调查。问卷内容也需要不断调整，以得到更加准确、全面的用户画像。

②丰富用户、产品特征。随着业务进行，我们需要不断向画像中增加新发现的有效特征，从外部抓取、购买新的特征数据。

③加快特征获取速度。提升特征获取速度可以在硬件上增加服务器数量，软件上考虑增加对特征数据的响应速度。

三、机器人投顾模式下智能选配模型的建立

（一）智能选配模型概述

1. 模型的提出

在证券市场上，如何在风险水平既定的条件下实现收益的最大化以及如何在目标收益率确定的条件下实现风险最小化？如何构造符合客户风险承受能力的投资组合并精准推荐给目标客户？这些问题其实都离不开投资组合的构造、筛选和匹配环节，据此我们设计了智能选配模型。

基于动态KYC设计的智能选配模型的优势在于：一是针对不同的客户风险承受能力匹配、推荐不同的金融产品；二是模型在保持风险收益比的前提下根据客户的需求能够快速生成个性化的金融产品。

智能选配模型可实现对客户（潜在客户）需求的信息挖掘，比如在客户浏览查询过程中，能自适应地提供贴合用户意图和兴趣的推荐内容。数据挖掘是智能选配模型的精髓，使用数据挖掘技术来分析App端用户的访问行为，准确识别用户的喜好，实现对用户行为数据的聚拢。数据挖掘突破了传统调查方式获取用户兴趣意向的局限性，并且能从收集的数据中准确判定一个用户的多维度特征；这些指向具体用户的即时数据，具有营销参考价值。用户在移动端的活动都能记录到后台数据库中，可以从中准确地了解用户潜在意图，从而找到受众目标。

2. 模型的技术组成

（1）输入/输出模块介绍。输入模块是获取客户行为数据的接口，分为“显性”和“隐性”两种获取方式。“显性”方式主要通过移动端、微信、PC端为客户设置的调查问卷来获取客户的目标收益率、初始投资金额、投资偏好等信息；“隐性”方式主要是通过读取移动端App埋点日志数据以及从公司交易系统、CRM系统读取客户的相关数据。

输出的形式主要为投资组合。具体输出包括了投资组合成分、各成分权重以及组合收益率水平。

（2）匹配模块介绍。模型的匹配模块主要用来实现模型生成的投资组合风险等级与客户风险承受能力匹配、投资组合收益率与客户目标收益率匹配。前者的匹配过程主要是根据动态KYC来实现，后者的匹配是通过ABL模型来实现。

（二）智能选配模型设计思路

该模型总体设计思路是在充分了解客户的前提下在众多产品中自动实现为客户挑选并匹配金融产品，目的是在符合投资者适当性管理的前提下帮助客户实现预期收益率。

智能选配模型核心涉及智能筛选、组合生成以及智能匹配、推荐的实现。主要通过智能化搜集市场股票、资讯、金融产品的交易、舆情信息来实现智能的择时、选股的量化模型。在智能选股的基础上，根据动态 KYC 把个股、组合匹配给适合的客户。在满足适当性的基础上，最大程度满足客户的预期收益率，算法强调的是收益率的拟合程度。

智能选配模型系统设计思路分为三大模块。具体包括用户潜在产品库的建立、推荐的实现以及基于推荐结果的自学习优化。

智能选配模型为每个客户建立一个潜在产品库。潜在产品库的建立是根据客户画像对客户风险承受能力、收益要求、风险偏好和产品画像描述各类可投资工具的风险收益。

匹配、推荐的实现通过择时模型选择合适的大盘环境、通过行业模型优选板块，通过推荐引擎对单个产品进行匹配分析，同时也可通过 ABL 模型自动生成符合客户预期收益率的产品组合。

模型的自学习优化通过对推荐结果和客户账户实际收益的跟踪，更新用户画像，优化选配模型（见图 6）。

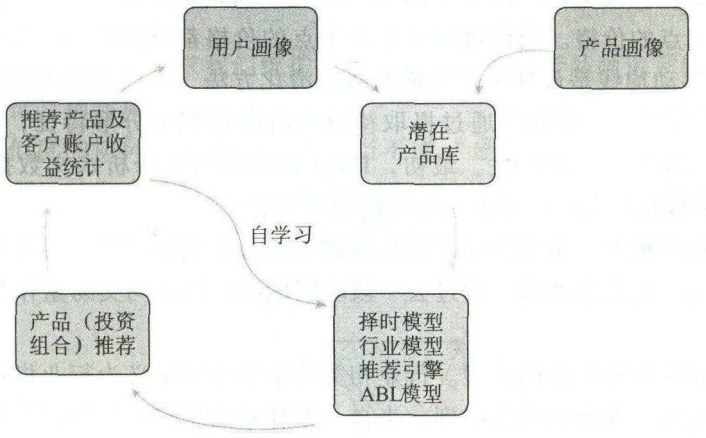

图 6　智能选配模型设计框架

（三）智能选配模型的实现

智能选配模型的实现主要包含筛选、构建和匹配三大环节，具体是指潜在产品库构建、组合的生成以及适当性匹配。模型考虑到众多中小投资者初始投资金额有限，并不具备购买组合的实力，所以整个模型设计时就将推荐标的分为单只产品和投资组合。

1. 单只产品智能选配的实现

（1）建立潜在产品库。根据客户画像对用户风险承受能力、收益要求、证券投资知识掌握程度的描述，匹配产品画像对各类可投资工具的风险收益描述，智能选配模型为每个客户建立一个潜在产品库。

潜在产品库是机器人投顾模式下适当性管理实现的第一步。客户画像多维度覆盖了调查问卷信息，潜在产品库方式是比仅通过调查问卷方式更准确更严格的适当性管理模式。

（2）行业（概念）模型。行业模型用于判断行业（概念）走势，思路如下：

①货币周期性模型。选取沪深300指数作为市场组合，并计算沪深300行业指数所对应行业的Beta值和均值方差，以此为基础将行业划分为周期性行业和非周期行业。

②行业动量与反转模型。动量效应，指的是行业表现具有持续性，强势的行业大概率会继续强势，低迷的行业大概率会继续低迷。

反转效应，指前期表现不好的行业跌破合理估值区间，在价值回归原理作用下，重回合理估值区间而产生的收益优于市场平均收益的表现。

③舆情模型。热点行业（概念）往往会被新闻媒体关注。利用抓取系统获取主流新闻网站上的新闻内容，对其进行分析，这样可以将新闻媒体关注的行业与其他行业区别开来。目前，舆情模型不能很好地判断新闻是正面或负面，需要配合其他行业模型，进一步将行业分为热点行业（大量正面新闻）、一般行业（缺少新闻关注）和夕阳行业（大量负面新闻）。

（3）择时模型。择时模型运用某种方法来判断大盘的走势情况，是上涨、下跌还是盘整。为了让智能选配模型更加有效，在向用户做出投资建议之前，我们通过择时模型对市场环境进行判断。

①基于价格的模型。

移动均线模型：该模型将一段时期内（一般是目标时间点前N天或月）的股票价格平均值代替目标时间点的价格。当价格曲线上每个点的价格都被代替后，形成新的价格曲线，称为N日均线。移动均线是最朴素的价格曲线光滑化方法，可以过滤大部分的价格噪声。

移动Hurst指数模型：该模型通过提取待分析时间序列的分形特征，计算出一个指数，判断该时间序列当前模式的持续性。最初，Hurst指数被用来分析水文数据，应用到证券市场上，移动Hurst指数可以帮助预警趋势的终结和反转。

②基于交易量的模型。市场的交易量是描述交易者乐观程度的最可靠数据，大的成交量代表乐观，小的成交量代表悲观。用过去一段时间区间内的平均交易量作为比较基准，可以定义交易量偏差。

③基于行业相关性集中度的模型。行业相关性集中度被定义为行业指数相关系数的均值与标准差的商。观察行业相关性集中度，来自一个朴素的思路：大盘趋势明显时，大部分行业表现出相同的趋势；大盘趋势不明显或趋势进入尾声时，各个行业的表现会呈现出较大的差异性。

（4）智能选配模型下的个性化推荐实现。在建立潜在产品库的基础上，由推荐引擎进行深度匹配，挑选出适合客户的股票或金融产品。推荐引擎算法考虑了如下方面：

①基于产品画像的推荐算法。当我们明确知道用户的偏好和待推荐产品的信息时，我们所要做的仅仅是找出与用户偏好匹配程度最高的产品（见图7）。

假设用户的偏好可以用 $\{a_i\}$ 来表示，产品的特征可以用 $\{b_j\}$ 表示，那么产品与用户偏好的匹配度可以用 $\frac{|\{a_i\}\cap\{b_j\}|}{|\{a_i\}|+|\{b_j\}|}$ 来计算。其中 $|\{a_i\}|$ 代表用户偏好特征的数目，$|\{b_j\}|$ 代表产品特征数目，$|\{a_i\}\cap\{b_j\}|$ 代表符合用户偏好的产品特征数目。

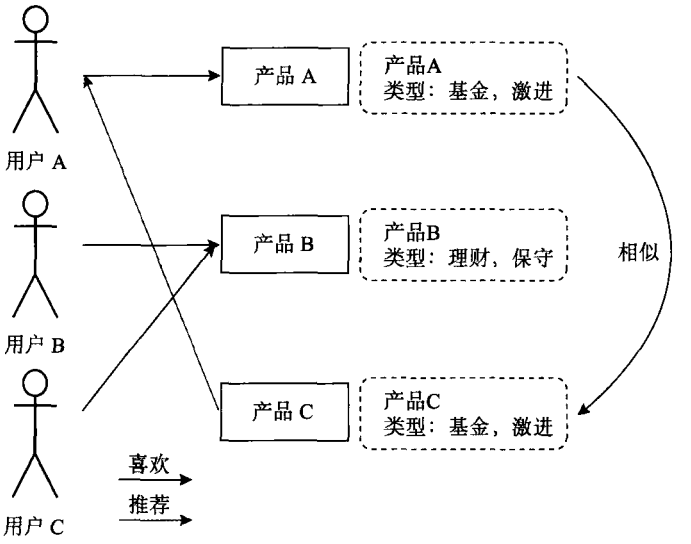

图 7　基于产品画像的推荐算法

在基于产品画像的推荐系统中，产品通过相关特征属性来定义，系统基于用户评价对象（例如客户购买过的、关注过的产品，阅读过的研报、新闻等）的特征，学习用户的兴趣，考察用户兴趣与待预测产品的相匹配程度。用户兴趣资料可以作为用户画像的一部分，随着用户的偏好改变而发生变化。

基于产品画像的推荐算法要求我们维护一个详细、有效的产品画像系统，并且需要用户充分的历史行为数据来描述用户兴趣。

②基于用户相似度的协同过滤推荐算法。当不能“显式”地得到客户偏好，例如，很难得到一位从未购买过基金产品的用户的基金偏好，可以通过相似用户的偏好推荐给用户。这个算法思路基于一个假设：相似的用户拥有相似的偏好（见图 8）。

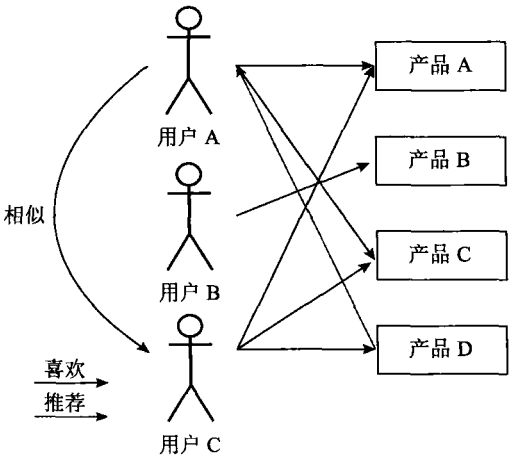

图 8　基于用户相似度的推荐算法

本模型中，我们利用算法自动动态生成一个用来综合衡量所有特征的距离函数。这个距离函数能给出在某个工作目标下用户之间的相似度。根据实际业务环境的不同，距离函数是

不同的，能更好地实现推荐精度。本模型的建立，参考了 k－means 聚类，梯度下降等算法。

③基于产品相似度的协同过滤推荐算法。类似地，通过产品相似度，也可以在不确知用户偏好的时候做出推荐。如果产品 A 与产品 C 相似，且用户 C 购买过产品 A，那么我们可以将产品 C 推荐给用户 C（见图 9）。

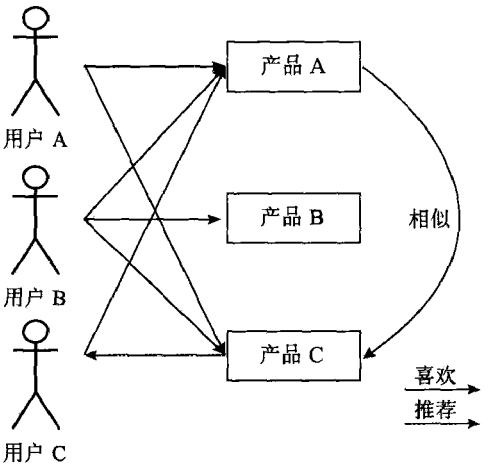

图 9　基于产品相似度的推荐算法

基于产品相似度的协同过滤推荐算法可以用类似基于用户相似度的算法实现。

④基于规则或概率的协同过滤推荐算法。一个规则通常有类似以下的形式："如果一个用户在股票账户中的现金资产超过 10 万元，且购买过 3 次以上 A 行业的股票，那么这个用户购买 X 基金的概率超过 50%"。

基于规则的协同过滤算法将找出这样的规则，并评估这些规则的有效性。有效性一般用支持度和可信度来描述。规则 $X \geq Y$ 的支持度和可信度的计算方式如下：

$$\text{支持度} = \frac{\text{包含 } X \cup Y \text{ 的样本量}}{\text{总样本量}}$$

$$\text{可信度} = \frac{\text{包含 } X \cup Y \text{ 的样本量}}{\text{包含 } X \text{ 的样本量}}$$

主要通过决策树模型和主成因分析模型进行规则挖掘，通过朴素贝叶斯模型进行概率方面的分析。

⑤会话式系统。当上述推荐算法都无法得出较好的推荐结果，需要通过会话式系统与用户进行交互，获得更多的客户偏好。会话式系统通过交互界面采集客户的语音或文字，提交给自然语言处理模型，提取出客户偏好信息，再将客户偏好信息补充给客户画像，直到客户画像中的信息足够推荐引擎给出较好的推荐结果（见图 10）。

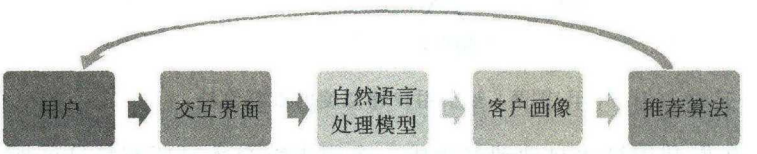

图 10　会话式系统流程

2. 投资组合智能选配模型的实现

以下主要介绍基于 Black - Litterman 模型（以下简称“B - L 模型”）的改进后的多因子模型（以下简称“ABL 模型”）的设计思路、投资组合产品生成逻辑以及具体组合生成模块的设计原理。

（1）B - L 模型核心思想。B - L 模型源于 Fischer Black 和 Robert Litterman 在 1992 年发表的一篇论文。在该文中，作者对传统的 Markowitz 理论和 CAPM 缺陷做出探讨和改进，一定程度上克服了结果对条件的敏感性问题和资本市场假设的误差问题，使得最终的结果更加分散、更加切合实际。B - L 模型使用贝叶斯分析方法将 CAPM 的先验市场均衡收益与投资者主观预期进行有效结合。

CAPM 的均衡收益率是基于其市场有效的假设得出的，B - L 模型将 CAPM 的均衡收益预期加入主观因素进行分析，根据投资者的预期和观点对基本投资组合进行调整，完善了 CAPM 的基本分析框架。模型采用观点对各资产的收益预期、对资产间关系的预期、观点的置信度等量化指标对观点进行描述，这些指标体现了观点的特征和有效程度。将量化指标加入基本的均衡组合配置方案后，可以对投资组合再次进行均值 - 方差优化。这种方法实际上是根据先验概率和观点试验调整产生最终的后验收益分布（见图 11）。

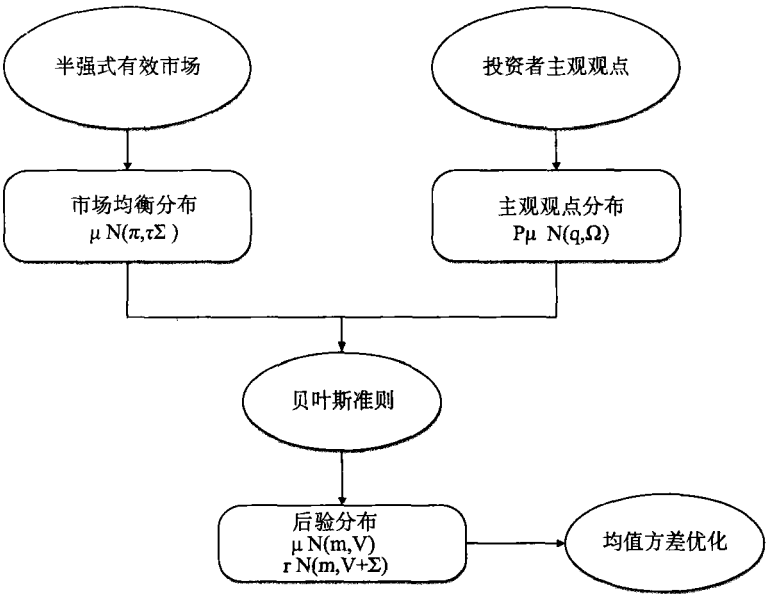

图 11 B - L 量化框架图

（2）ABL 模型。在 B - L 模型的基础之上，引入 GARCH 模型、美林投资时钟逻辑分析框架以及一致性预期来对传统 B - L 模型进行修正，形成 ABL 模型。

①ABL 模型设计思路。

整体思路：B - L 模型在 CAPM 模型的基础上根据贝叶斯统计的思想，将投资者对大类资产的观点作为试验条件，将 CAPM 模型中的市场均衡回报作为先验结果，通过二者结合最终产生后验的预期回报。该模型除采集市场产品基准数据外，还关注投资者对各产品的倾向性意见，并对各项意见进行量化分析，结合市场中各产品的基准数据和投资者的倾向性意见

给出新的资产配置策略。新的资产配置策略包含符合直觉的市场产品组合及可以理解的权重设置（见图12）。

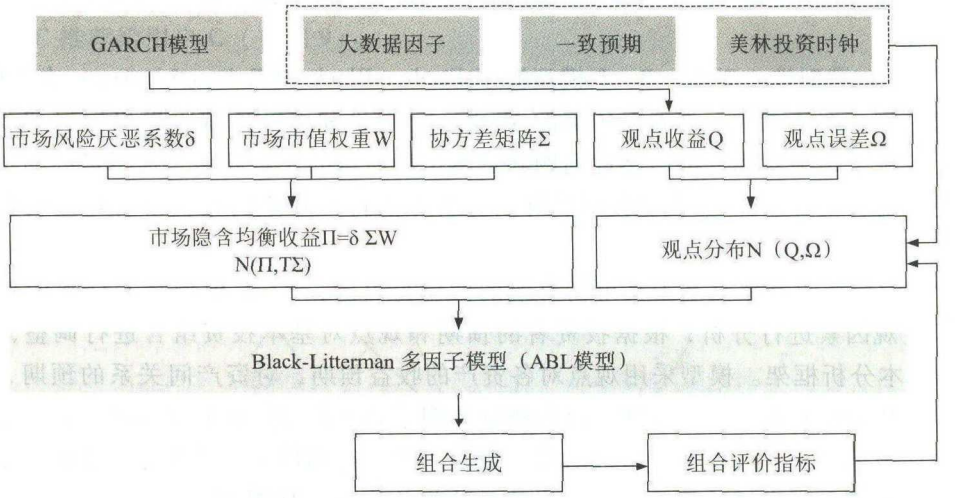

图12 智能选配模型原理

引入GARCH模型的考量：GARCH模型是研究时间序列数据相关关系的一类回归模型，该模型在传统模型对于数据本身进行建模的基础上，对数据的方差进行回归分析。由于方差体现了数据的波动性，因而GARCH模型除预测时间序列数据外，还对预测结果的波动性进行分析，能够给投资者的决策提供更多参考，从而优化投资者的决策，其意义很多时候超过了对数值本身的分析和预测。

引入美林投资时钟分析框架的考量：在投资标的选择上，采用美林时钟模型进行分析。该模型首先根据经济指标判断当前经济状况在经济周期中所处的时期，并选出该时期适合投资的大类资产构建基础资产池。美林时钟模型综合考量了资产状况、行业轮动、经济周期等方面的因素，能够较为全面地考量资产表现，在投资实践中具有较高的指导意义。借助美林时钟的分析框架，投资者可以有效识别经济周期的重要转折，从而实现资产配置的动态调整和优化。

引入产品库的考量：在模型中引入的产品库主要包括产品画像时得到的各类标签数据和此后的产品表现数据等，这些数据能够准确而充分地体现产品的收益、风险、期限等相关特性，为模型筛选出满足客户收益要求的产品并达到适当性管理的要求提供了基础支持。

引入大数据因子的考量：大数据因子的引入在加强模型的有效性、提升匹配结果的准确度方面能起到较好的效果。大数据因子主要包括产品历史大数据、舆情大数据、市场大数据、新闻资讯大数据、主题事件大数据、行业加权大数据、合作方社区用户大数据等。这些数据涉及产品、市场、投资者、舆情等多角度，从行为和心理两方面提供了全面的信息。这些数据将有助于我们对产品和客户构建精准画像，改进匹配结果（见表2）。

②构建投资组合生成模块。投资组合生成模块包括了输入、计算模型以及输出。输入模块包括用户标签、用户需求。用户标签是通过相关技术对用户基本特征、行为特征等数据标签作为模型的输入参数。用户需求数据是指用户的目标收益率预期、风险偏好等信息，这部

表 2　　大数据因子分析表

大数据类别	大数据描述	大数据来源	大数据分析方法
产品历史大数据	反映产品历史表现的海量数据	抓取产品的历史公开信息	数据挖掘技术与序列分析
舆情大数据	反映投资者观点舆情的海量数据	问卷调查、交易数据统计处理	合作方情绪量化分析方法
市场大数据	反映市场行情大势的海量数据	机构调研数据、问卷数据	数据挖掘技术
新闻资讯大数据	反映媒体新闻报道的海量数据	网络财经媒体新闻爬虫抓取	自然语言处理与数据挖掘技术
主题事件大数据	反映政策、个股重大变化的数据	研究数据、公开信息的收集	自然语言处理及数据挖掘技术
行业加权大数据	反映行业板块特征的海量数据	通过挖掘行业相关性获取	数据挖掘技术
合作方社区用户大数据	合作方平台的观点策略数据	通过合作方社区数据库获取	自然语言处理与数据挖掘技术

分数据通过问卷形式采集。输出模块中主要输出的是投资组合以及组合对应的收益率。计算模型作为整个智能选配系统的核心“黑箱”，是整个系统的“大脑”。

在具体构建策略时，我们采用 B－L 模型确定最终组合中各待选标的的权重。其中观点矩阵的确定，由于传统的主观设定方法存在缺陷，我们通过 GARCH 模型对标的历史数据进行拟合，并将模型对组合未来表现的预期作为观点指标，同时结合一致性预期、大数据因子分析技术构建市场观点矩阵（见图 13）。

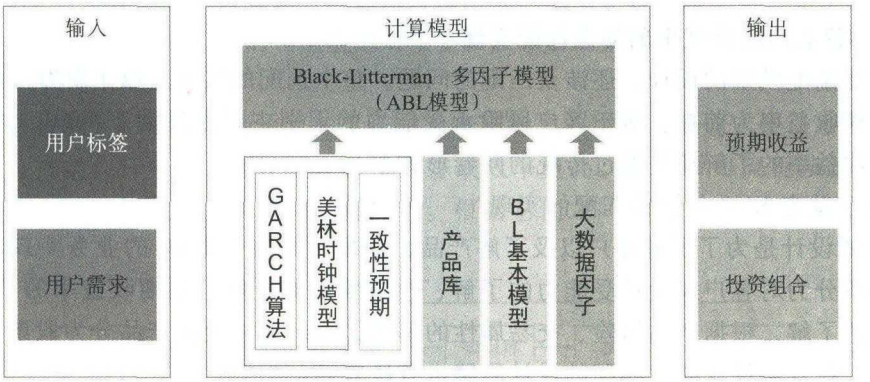

图 13　ABL 模型原理

③投资组合匹配。当客户输入具体的投资需求信息后，机器人根据客户输入以及采集到的信息，通过组合生成模块生成推荐的产品。如果组合标的是股票类产品，则直接推荐给客户。如果组合标的是理财类产品；根据适当性管理要求，对生成的理财组合产品的风险进行二次检验确认，确保推荐给用户的理财产品组合与其风险测评等级相一致。具体的组合推荐

过程如图14所示。

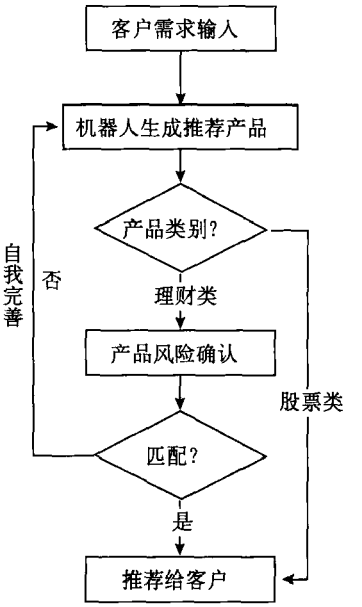

图14 产品匹配推荐流程图

通过动态KYC模块采集到客户的风险承受能力、风险偏好和预期收益率等信息之后，机器人会生成相对应的推荐产品。当输出的组合产品是理财组合时，需要对该组合的风险水平进行确认，以判断是否和客户的风险承受能力相匹配。如果匹配合适，则推荐给客户；如果两者不匹配，则调整参数进行第二次组合生成，直到找出符合客户风险承受能力的产品组合为止。

（四）智能选配模型下的动态匹配实现

对智能选配模型的设计，还涉及匹配的问题。整个选配的实现是以了解客户风险承受能力以及预期收益率为前提。对于客户风险承受能力的识别主要通过动态KYC来实现，对于客户预期收益率的了解还是通过传统的问卷形式。

1. 基于动态KYC的智能匹配的实现

KYC的设计是为了了解客户以及了解产品，匹配则是整个KYC的业务实现环节。对于客户的了解分为对客户风险承受能力的了解、投资偏好的了解、预期收益率的了解等方面。对于产品的了解，根据产品风险、交易属性的不同，我们先将金融产品分为股票类和理财类两个大类。在具体设计时，匹配的设计思路也有所差异。

（1）设计思路。基于动态KYC的匹配主要是根据客户的风险承受能力来匹配理财类金融产品。对于产品的选择，综合考虑产品风险等级与客户风险承受能力的匹配程度，而不以绝对收益率作为唯一考量标准。换句话说，我们推荐给客户的理财类产品不一定是收益率最高的，但是与客户风险承受能力是最匹配的。

（2）匹配的动态调整。一方面根据客户数据更新导致的标签变化对匹配结果进行调整。以风险标签为例，当一个用户历史数据证明其为高风险接受能力，通过后期该用户的交易行

为不断修正后系统发现其真实属性为低风险偏好时，一是对其发起二次调查问卷，对风险充分披露后要求其作正确评估；二是对其资产配置进行调整，用更多低风险属性的产品替换掉之前的高风险产品。

另一方面根据推荐产品组合的实际表现与预期表现之间的差异对产品组合进行更换或内部比例调整，保证策略的最终收益与预期相符。由于产品类别特性存在差异，需要对各类产品设定不同的衡量指标和阈值，当产品相关指标触发预先设定的调整阈值时便实行调仓。

2. 基于客户预期收益率动态再平衡的实现

机器人投顾中“再平衡”是确保收益长期稳定按照用户预期执行的关键功能。再平衡包括资产再平衡和策略再平衡。

（1）组合再平衡。当组合的实际收益率与客户预期收益率偏离达到某一设定的数值，如初始设定的止损线，ABL 模型将通过调整原有组合的资产标的或标的权重以达到最大程度地拟合客户预期收益率的目标。

（2）策略再平衡。基于 ABL 的策略再平衡设计的初衷是生成的组合的收益率最大程度地拟合客户的预期收益率。

当组合的预期收益率与客户预期收益偏离较大，且通过调整组合权重及组合标的仍难以实现对客户预期收益率的拟合，这时可以通过策略再平衡的方式重新调整已有策略，以保证策略一直处在最有效状态。

四、机器自学习完善机制

（一）自学习系统概述

1. 自学习系统的提出

本项目设计的动态 KYC、产品组合的生成以及推荐是由模型自动实现，由于模型的数据，如标签库数据、产品库数据、投资组合库数据以及与客户的交互数据都是动态更新的。模型能够实现同步数据的动态变化，同时自适应地对模型参数做出调整优化，以提升匹配精度。

2. 自学习系统介绍

自学习过程就是系统在不断重复的工作中对本身能力的增强或改进，使得系统在下一次执行同样任务或类似任务时，会比现在做得更好或效率更高。整个学习流程涉及六个单元：环境、选例、学习、知识库、执行和监督环节。根据它们之间的关系建立起如图 15 所示的学习模型。

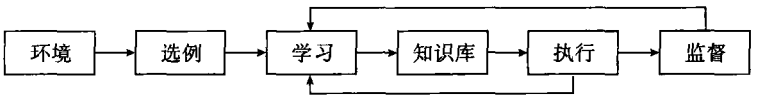

图 15　自学习流程图

如果整个流程没有自学习机制，那么基于先验知识的机器学习就类似于人们首先知道应该怎么解决这些问题，然后设定算法，规定机器怎么去处理问题。从这个意义上来说，机器相对于人工而言，仅是提高了效率，而计算的结果可能并非最优。

（二）自学习算法框架

大部分的机器学习算法都需要使用已有的历史数据去训练它。训练数据集的质量越高，训练后的算法越精确。部分算法，特别是推荐算法，非常缺乏历史数据。我们缺少客户对推荐结果的反馈以及推荐结果本身的效果，来确认该推荐是否有效。在这种情况下，需要一些模型来帮助缺少训练数据集的算法来实现自动修正和提升。

另一方面，我们拥有众多的择时模型和行业模型。在实际运用中，不同模型可能会给出截然相反的结论。这时，我们需要另一些模型来统筹分析所有模型的结论，进而得出一个信号，使判断的准确性超过任一单个模型。

从算法框架（见图16）来看，整个业务流程从用户和产品画像开始，到推荐产品及客户账户收益统计后，根据统计结果，通过增强学习算法对投资组合推荐模型进行优化，通过主动学习模型对推荐引擎进行优化，通过协同训练模型对择时模型和行业模型进行优化，实现算法学习的闭环。

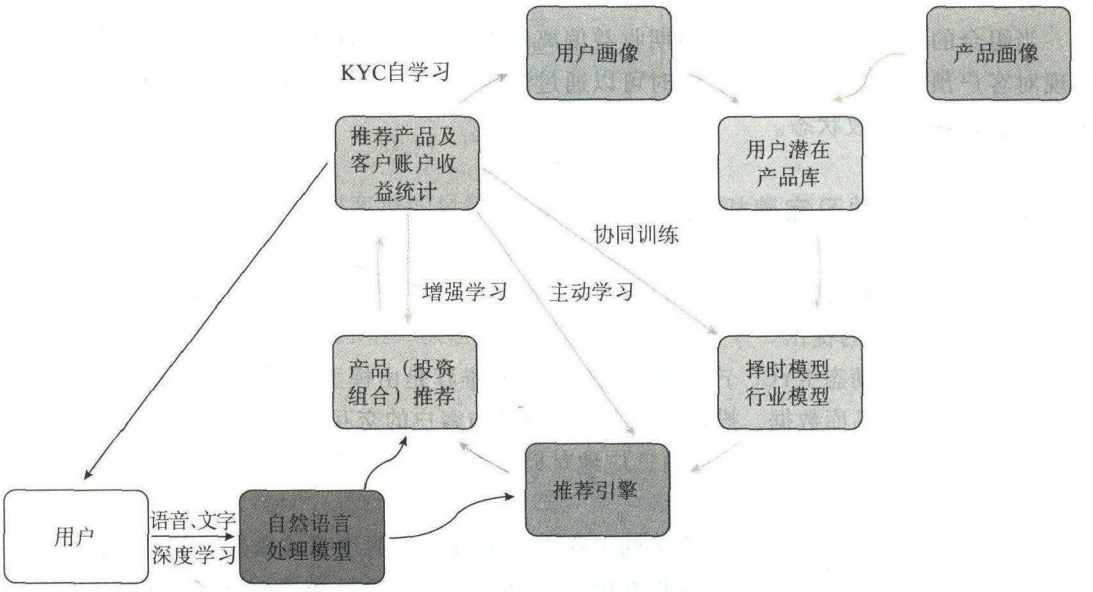

图16 自学习算法框架

（三）自学习在选配模型中的应用

模型的自学习功能在整个项目中主要应用于动态KYC的实现以及智能选配模型（择时、选股、推荐）。

1. 主动学习在推荐引擎的应用

在推荐算法中，缺少对推荐数据的检验结果。例如，市场上新推出一款基金产品，算法自动生成可能会对这款基金产品感兴趣的用户名单。但是，如果没有一个衡量算法结果好坏的标准，那么无法知道算法在哪些用户的匹配上效果不好，算法也无法进一步优化。为了解决这个问题，引入主动学习模型。

具体而言，当推荐引擎将推荐结果展示给客户，我们可以根据客户是否采纳推荐建议来判定该推荐是否有效，也可以根据推荐结果是否符合预期来判定该推荐是否有效。主动学习系统通过主动查询，将这些结果判定作为“神谕”，并以此标记示例，反馈到主动学习机制中以扩充训练集进行常规的监督学习。推荐引擎将结合主动学习机制的训练结果学习这些经验，从而优化推荐算法和模型，使得推荐结果更加精准（见图 17）。

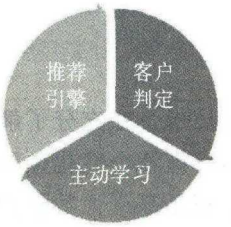

图 17　推荐引擎的主动学习机制

2. 协同训练在择时、行业模型的应用

协同训练是一种利用互补的学习机对未标识样本特征空间进行探索的半监督学习方法。它将样本特征集作为视图来构建学习机，利用学习机之间的相互训练来提高预测性能。可以弥补因单个学习机的预测偏差而对最终结果造成的影响。最终结果综合两个学习机预测的结果得到。

协同训练主要使用在择时模型和行业模型的训练中。择时模型和行业模型中不同模型结论产生自不同观察角度，天然满足条件独立假设。由于择时模型和行业模型可以使用历史行情数据，一般情况下，也满足视图充分冗余假设。通过协同训练，可以将一个模型的优势提供给其他模型进行学习，进而提升所有模型的效果。

3. 增强学习在产品组合推荐的应用

增强学习可以把动态决策的思路带入推荐引擎模型中。即每个推荐都是一连串后续决策的基础，也将承担后续一系列可能结果的风险。考虑到投资组合一般都要持有一段时间（短线操作更适合单个金融产品推荐），投资组合的建立和维护必须考虑一段时间内可能遇到的各种市场变化。本项目拟通过增强学习算法，学习各种组合在市场环境下的表现，使组合推荐引擎具有“未来最优”的思路，并具有动态组合调整的能力。

在配置产品组合过程中，由环境导入的数据源，通过增强学习，不断训练和优化模型，寻找最优的产品组合策略，使得产品组合的收益率、风险偏好更加符合投资者的投资需求和风险偏好。

4. 深度学习在自然语言处理中的应用

严格来说，深度学习是基于深层的神经网络的学习方法别称，并不是一种全新的机器学习方法。

2011 年 Seide 在深度学习研究中指出，在语音识别领域，自然语言处理模型可以成功地将识别错误率从 27.4% 降低到 18.5%，这使得深度学习在人工智能方面取得了突破性成果。

为了做到这些，深度学习模型需要学习大量的素材——数以亿计的文章、大量的语言信号。本项目中，自然语言模型中的深度学习算法将在上线后，采集与客户的交互数据，不断

训练优化自身。借助深度学习，自然语言处理模型可以通过语言或文字与用户进行互动，回答提问，采集用户需求并提交给推荐引擎，将推荐结果反馈给客户，实现动态的人机交互，真正做到“机器人投顾”。

五、机器人投顾产品化思路

（一）产品设计思路

1. 产品定位

在产品定位上，主要借助大数据技术，在充分了解用户风险承受能力、风险偏好以及需求的基础上，自动化、智能化地为客户推荐个性化金融产品。对于客户、产品的了解是基于画像系统实现；对于组合生成以及推荐是基于大数据相关算法来实现。

2. 目标用户

在客户定位上，产品主要服务于中小净值客户以及投资经验相对欠缺的客户。客户来源包括投资理财需求的公司内部客户和互联网外部用户。在权限设定上，考虑根据不同的客户级别开放不同的权限。

3. 投资标的

在投资标的选择上，结合国内用户的投资偏好以及投资品种大类，目前主要覆盖A股股票类产品和理财产品。其中，理财产品主要以基金为主，涉及权益类基金、（类）固定收益、黄金ETF、另类（非黄金）和现金管理类五大类产品，后续将逐步引入私募基金、资管类等产品，以达到全方位的资产配置效果。

4. 产品功能

根据大数据平台中的客户信息和产品信息，机器人投顾可以构建精准的客户画像和产品画像，结合择时模型、行业模型，为客户精准推荐符合其偏好和风险承受能力的产品；通过ABL模型，为客户提供个性化的资产配置建议。

5. 收费模式

国外机器人投顾主要以代客理财形式收取管理费用，而国内机器人投顾主要向投资者提供投资建议，所以基本以免费为主。本产品在结合国内行情的基础上，同样推出免费模式，且对投资者的投资金额无最低要求。

6. 账户体系

对于机器人推荐的理财产品组合，投资者可通过理财账户或交易账户实现购买，对开户无硬性要求，投资者只需注册为理财用户，绑定银行卡，即可实现购买；而对于股票产品组合，投资者只能通过交易账户进行买卖，用户通过开户实现购买流程。

（二）机器人投顾业务流程

我们设计的机器人投顾产品分为两个分支：一条分支是建立在投资者选择股票品种的基础上，经过一系列的流程，最后得到股票组合；另一条分支是建立在投资者选择理财产品的基础上，最后得到理财产品组合。具体业务流程如下（见图18）。

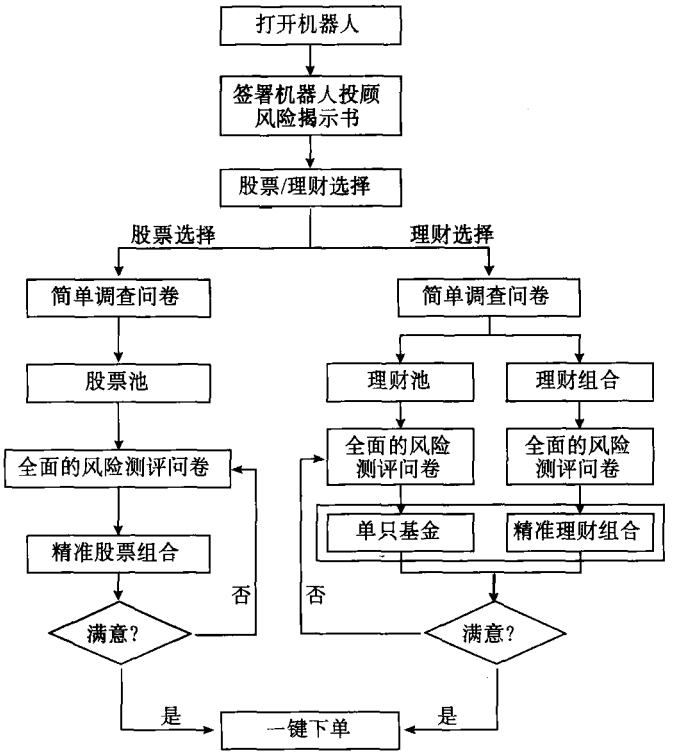

图 18　机器人投顾业务流程图

第一步：由用户根据投资需求，选取投资品种，目前主要以股票、理财为主；

第二步：填写简单调查问卷，主要从风险偏好、资金门槛、资金流动性三个维度来评估用户的投资偏好；

第三步：根据简单的调查问卷，系统首次推荐股票池、理财池[①]或粗略的理财组合；

第四步：当用户对推荐结果不满意时，可通过填写全面的风险测评问卷，得到精准的股票组合、理财组合或单只理财产品；

第五步：通过“一键下单”功能实现产品组合购买。在设计“一键下单”功能时，我们也考虑了客户持仓对推荐组合收益率的影响，通过组合再平衡策略进行调整。

（三）产品页面展示介绍

根据产品业务流程、业务功能以及用户行为习惯，我们设计了机器人投顾原型图。下文我们将对产品原型图页面逐一介绍。

用户打开机器人投顾功能，签署风险揭示书以后，会直接跳到产品首页（见图 19 左），首页根据业务流程分为两个入口。用户通过点击“股票”或“理财”按钮，可得到不同的产品配置建议。如果用户曾经保存过机器人推荐的产品组合，则可以通过“查看组合”按钮，直接查看原来的组合配置详情，以及组合产品的收益情况。

用户从首页入口进入，首先需要填写一份简单的调查问卷（见图 19 右），题目设置为 5

① 理财池是指从一系列理财产品中，选出适合投资者的一些理财产品，放入一个备选库，供投资者选择。

道题，用户只需根据自身状况进行选择。

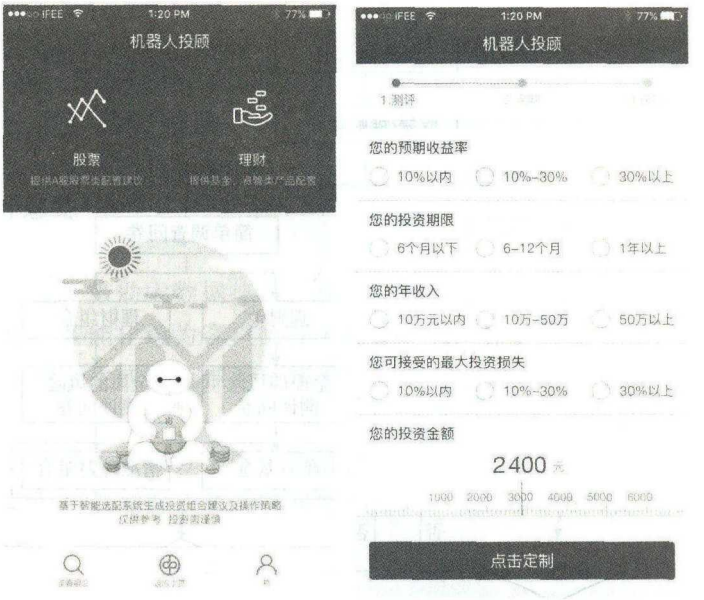

图19 产品首页

简单的调查问卷结果可以确保机器人投顾对用户有一个初步了解，在此基础上，智能选配系统会根据客户提供的信息为其定制个性化的推荐产品。通过“股票”入口进入的用户，系统将推荐一篮子的“股票池”（见图20左）；通过“理财”入口进入的用户，系统会对调查问卷中的“投资金额”做一个判断，当投资金额超过10 000元时，系统默认为其推荐风险相对分散的“理财组合”，反之，则将推荐“理财池”（见图20右）。

图20 系统推荐

如果用户对系统初步提供的“股票池”、“理财池”或“理财组合”不满意，则可通过重新填写一份包含有 15 道题目的风险测评问卷（见图 21 左），得到更加精准的产品配置结果。

从“股票”入口进入的用户，机器人投顾将根据全面的风险测评问卷结果，为其推荐股票组合（见图 21 右）。该组合有具体的仓位建议，尽可能满足用户的目标收益。

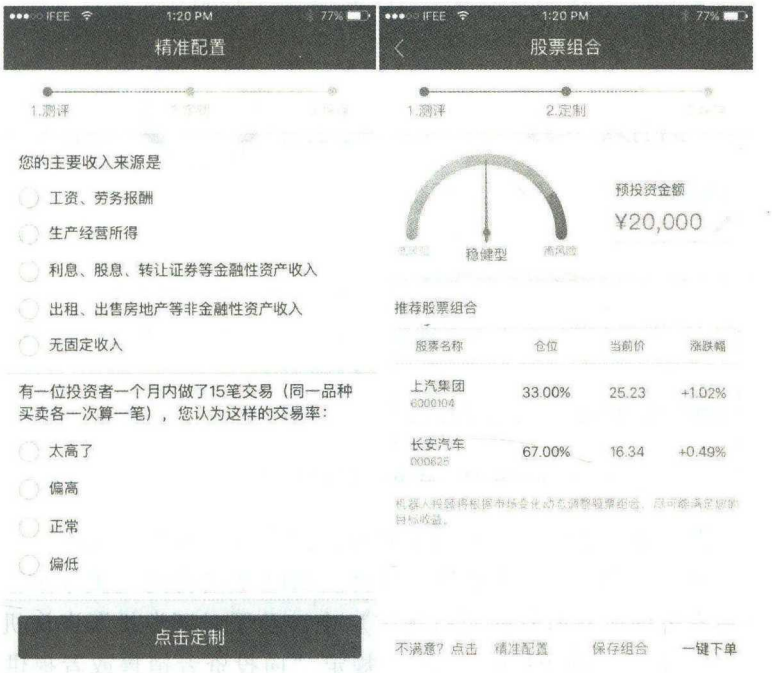

图 21　风险测评问卷

而从“理财”入口进入的用户，完成全面的风险测评问卷以后，系统则会匹配更加精确的理财产品：推荐投资标的为“基金池”的用户会将得到最优的单只基金（见图 22 左）；推荐投资标的为“理财组合”的用户，系统将提供更符合其投资需求的“理财组合”（见图 22 右）。另外，机器人投顾具有智能的个性化推荐功能。例如，对于投资金额小于 10 000 元的投资者，系统会向其推荐单只理财，但是根据动态 KYC 抓取的客户行为数据，如果投资者真实的风险承受能力和投资额度超过 10 000 元，系统则会同时提醒用户考虑理财组合建议，供投资者参考。

六、机器人投顾的挑战及发展前景

（一）机器人投顾发展面临的问题和挑战

机器人投顾在全球范围内尚属新兴事物，在其发展过程中存在许多不足之处，我们主要从政策监管、机器人投顾算法及产品本身等方面对其面临的一些问题进行简要分析。

1. 监管法规尚待完善

由于机器人投顾是近几年来兴起的一项业务，国内外均未形成相应的监管法规约束。

图22 系统匹配理财产品

国内在机器人投顾的法律定位方面尚不明确，仅在《证券投资顾问业务暂行规定》中涉及了“以软件工具、终端设备等为载体”字样。2013年出台的《关于加强对利用“荐股软件”从事证券投资咨询业务监管的暂行规定》中涉及了对证券投资咨询机构利用“荐股软件”从事证券投资咨询业务的规范，也只是规定“向投资者销售或者提供‘荐股软件’，并直接或者间接获取经济利益的，属于从事证券投资咨询业务，应当经中国证监会许可，取得证券投资咨询业务资格”。由此可见，关于机器人投顾如何进行监管，国内尚处于探索阶段，各种法律法规尚待完善。

2. 机器人投顾的算法风险

机器人投顾算法模型的改进需要大量的数据积累和深入研究，但目前机器人投顾应用尚处于起步阶段，积累的数据和学习经验不足，产品推荐精度以及生成的组合有效性也较低，而数据分析能力的提升和算法优化都是一个漫长的过程。

3. 产品规范化管理问题

国内外监管机构都对投资顾问业务的适当性分析环节高度重视，要求向合适的客户提供合适的产品，而判定机器人投顾的适当性分析是否有效则成为产品规范化管理中的重要问题。市场上一些企业仅是根据调查问卷推荐一些产品便打起机器人投顾的旗号，甚至掩盖资金去向，进行非法经营，适当性管理更是无从谈起。

（二）机器人投顾的发展前景

机器人投顾以智能化、科学化的资产管理方式向投资者提供客观、高效的服务，随着国民投资理财意识的觉醒，机器人投顾将拥有广阔的应用前景。

1. 在中小投资者中获得广泛应用

目前中国的证券市场中小投资者占大多数，但由于专业投顾人数较少，广大中小投资者无法享受到相关服务。在行业十分重视中小投资者保护的背景下，机器人投顾的出现将会使广大中小投资者获得便捷、低成本的服务，而迅速增长的中小投资者市场也为机器人投顾的发展和普及提供了机会。

2. 投资品种的多样化

就目前状况看，许多资产配置平台虽然选择了大量产品，但产品所属的资产大类只有一两种，并未真正分散风险。随着中国资本市场监管的逐步放开，投资者可选择的投资标的种类将趋于多样化。面对更加丰富的选择，机器人投顾也能够产生更多的资产配置方案，最大限度地降低非系统风险。

3. 建立人机合作机制

现有机器人投顾虽然多数只是运用大数据技术分析进行策略推荐，有些功能的实现可能还要依赖人工辅助，但机器人投顾作为未来投资理财的发展方向，随着人工智能技术的进步，其提供的策略将越来越精准，服务也将趋于个性化，投资者未来可以与投顾机器人进行需求交互，通过人机合作机制实现定制的个性化服务。