基于模型融合思想的程序化交易投资者识别研究

袁钰坤, 徐刚, 吴畏, 徐力

引用本文

袁钰坤, 徐刚, 吴畏, 徐力. 基于模型融合思想的程序化交易投资者识别研究[J]. 计算机科学, 2023, 50(11A): 230300131-6.

YUAN Yukun, XU Gang, WU Wei, XU Li. Study on Programmatic Trading Investors Recognition Based on Model Fusion [J]. Computer Science, 2023, 50(11A): 230300131-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300131

基于模型融合思想的程序化交易投资者识别研究

袁钰坤[1]　徐　刚[1]　吴　畏[1]　徐　力[2]
1 中证数据有限责任公司　北京 100032
2 中国科学院计算技术研究所网络数据科学与技术重点实验室　北京 100190
(yuanstanly123@163.com)

摘　要　近年来,随着信息化、电子化技术在金融市场中快速发展,程序化交易成为了越来越多金融机构选择的交易方式,对证券期货市场的影响力也逐渐增强,已受到监管层及广大投资者的关注。文中基于模型融合的思想,构建了程序化交易投资者的识别模型,将专家规则与机器学习算法进行叠加融合,并在中国A股市场投资者交易数据上验证了模型的有效性。研究表明,模型能以超过90%的准确率和召回率识别出程序化交易投资者账户,超过了当下的前沿效果,相关研究成果可以为证券期货行业程序化交易识别相关的科技监管工作提供支持。

关键词:程序化交易,模型融合,机器学习,识别模型,科技监管
中图法分类号　TP391

Study on Programmatic Trading Investors Recognition Based on Model Fusion

YUAN Yukun[1],XU Gang[1],WU Wei[1] and XU Li[2]
1 China Securities Data CO,. LTD,Beijing 100032,China
2 Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190,China

Abstract　Programmatic trading has recently gained popularity among financial institutions due to the advancements of information and electronic technology in the financial market. It makes a significant impact on futures markets and draws the attention of regulators and investors. This paper develops recognition models based on the idea of model fusion for programmatic trading investors,combing the rule-based models and machine learning models,and proves the validity of the model on investor data in China's A-share market. The proposed model achieves over 90% accuracy and recall on recognizing programmatic trading accounts, which is better than the state of the art. Our experiments show that the proposed model is able to support the technical regulation on programmatic trading.

Keywords　Programmatic trading,Model fusion,Machine learning,Recognition model,Technical regulation

1　引言

近几年,尤其是2020年以来,国内量化基金发展迅速,不断涌现出百亿量级私募,甚至一度出现千亿级私募。同时,规模居前的头部量化私募越来越多地采用超级计算机等先进算力工具和机器学习等智能算法进行程序化的投资决策,这相对于中小散户投资者具有较大的技术优势。证监会主席易会满在第60届世界交易所联合会(WFE)上针对程序化交易等新型交易方式的监管问题,强调:"量化交易、高频交易在增强市场流动性、提升定价效率的同时,也容易引发交易趋同、波动加剧、有违市场公平等问题"[1]。这体现出了监管机构对于程序化交易的关注与重视。

近年来,我国关于国际金融监管机构对程序化交易的监管策略研究文献逐步增多,但针对具体监管实践的研究则并不多见,尤其是对基于人工智能等前沿技术的应用研究,这与国际的监管实践存在一定差距[2]。同时,受期货市场T+0的交易机制的影响,高频交易在期货市场中的规模逐步提高,而国内在相关领域上的监管手段研究却较为缺乏。当前,程序化交易特别是高频交易主要在ETF套利、跨市场交易、期货交易中进行。境内的现货市场程序化交易相对衍生品市场而言程度更低,但随着衍生品市场在近些年的发展,大多数的交易开始基于复杂的算法或者自动化的程序。这些算法的设计策略是以特定的方式进行交易,对市场价格、行业趋势、经济数据以及跨越交易市场不同板块等因素的变化做出了相应的反应。因此,当存在跨市场、跨品种的监管套利机会时,自动化程序交易会尽可能抓住任何获利的机会,同时规避监管。

基于以上原因,为了助力程序化交易的科技监管工作,本文结合专家规则和机器学习算法,针对投资者行为特征进行建模分析,并构建了一套用于自动识别程序化交易投资者的模型。本文的创新之处在于:1)基于模型融合思想,在A股市场的程序化交易投资者识别领域上,首次融合了专家规则、聚类算法和分类算法,构建了一套模型,其效果超过了已有研究;2)巧将聚类算法应用于分类场景,并与其他种类的算法有机结合,拓宽了聚类算法的应用思路;3)构建了一套将机器

基金项目:国家自然科学基金(61902380);北京市科技新星计划(Z201100006820061)
This work was supported by the National Natural Science Foundation of China(61902380) and Beijing Nova Program(Z201100006820061).
通信作者:徐刚(xug@csdata.cn)

学习算法赋能程序化交易投资者监管工作的思路，为监管实践工作提供了应用方向。

2 相关研究

2.1 程序化交易的发展

国内外关于高频交易为代表的程序化交易的研究较为丰富，但关于程序化交易的具体定义在学术界和工业界均未达成统一。如 Xiong 等[3]指出，我国的程序化交易涉及许多高频交易和算法交易特征。Lu[4]认为，基于特定的交易数据、算法模型以及交易策略，通过计算机进行自动或半自动化的决策，择时执行交易指令的方法均可称为程序化交易。Xiang 等[5]指出，程序化交易是基于计算机技术，按给定的程序进行大规模、高速的自动化交易。Zhang[6]认为，程序化交易是通过程序进行市场分析、决策选择、时机判断和报单指令传达。在国外，基于美国纽约证券交易所官方网站 2013 年出台的相关规定，任何一笔同时买卖十五支及以上股票的集中性交易均可被认定为程序化交易[3]。

作为程序化交易的一个很大占比的子集，高频交易已经发展成为重要的趋势，然而学术界对高频交易对股市的影响和作用存在一定程度上的分歧。有部分学者认为高频交易增强了市场流动性[7-8]，降低了交易的成本[9]，强化了市场稳定性[10]，如 Kirilenko[11]发现高频交易不会对股市的多空力量产生影响，Conrad 等[12]发现高频交易可以让股价更为稳定。然而，也有学者[13-14]认为高频交易会在一定程度上加重信息不对称问题，认为是新的市场不稳定因素，同时普通交易者会减少交易，损害了证券市场的公平性[1,14]，这会降低市场质量，进而导致高频交易在实质上并未提供流动性[16-18]，在某些情况下会损害市场的流动性[19,21]。

随着人工智能及大数据技术在许多领域被应用，这些技术在资本市场监管中也逐渐被应用和落地[22]，例如应用于资本市场情绪分析[23-24]、市场的风险分析[25-26]、股市的走势分析[27]等。因此，探索将人工智能算法应用于程序化交易投资者识别的监管场景具有重要的意义，可以更好地将前沿技术与资本市场科技业务场景相结合，实现以科技赋能监管工作。

2.2 程序化交易投资者的识别研究

程序化交易发展引起的市场风波日益受到各界关注，全球各主要市场的监管部门逐步加大对程序化交易的监管力度。近些年，国内学者开始研究程序化交易风控策略[28-29]，同时纷纷呼吁监管部门应当未雨绸缪，在国内建立相应的程序化交易风险防范机制[24-31]，认为全面认识程序化交易的功能和作用尤为关键[32]。

2015 年，中国证监会就《证券期货市场程序化交易管理办法(征求意见稿)》公开征求意见。同年，沪深交易所纷纷编制了程序化交易相关管理实施细则并公开进行意见征求，可见监管方对程序化交易管理工作的重视。为更加高效地开展对程序化交易的分析和识别，近年国内外对相关内容的研究逐渐增加[19,33-34]，以下是对程序化交易相关识别的两类研究，一类是基于专家规则的研究，另一类是基于机器学习的研究。

2.2.1 基于专家规则的识别研究

目前，基于专家规则的程序化交易投资者识别的研究相对较多，主要集中在以下两类方式。

1)基于特征指标的定量分析和识别

基于特征指标的定量分析和识别主要依赖于业务专家的经验，通过人工合成特征指标对程序化交易投资者的行为进行量化分析和判别。目前，这类方法在欧美证券期货市场使用较多，主要有以下方法。

第一种是基于交易信息量指标的识别方法。芝加哥商品交易所和美国洲际交易所都采用了加权交易信息量比例 WVR(Weighed Volume Ratio)[35]，如式(1)所示，以识别高频程序化交易并予以限制。

$$WVR=\frac{\sum(\text{发至交易系统的信息量}\times\text{加权比例})}{\sum\text{总成交量}} \tag{1}$$

第二种是基于订单成交比机制的识别方法。欧洲期货交易所和欧洲洲际交易所均采用了订单进行监测分析，以限制撤单比率。OTR 机制的两档衡量标准分别是基于订单数与基于成交量。通过计算 OTR 模型指标进行监测，其计算频率一般为日度。

$$\text{订单数 } OTR=\left(\frac{\text{订单总数}}{\text{成交的订单数量}}\right)-1 \tag{2}$$

$$\text{成交量 } OTR=\left(\frac{\text{所有订单的手数}}{\text{成交量}}\right)-1 \tag{3}$$

除以上主要的两种特征指标外，还有撤单率(Cancellation Rates)、日内回转率(Daily Turnover)、订单速率(Message Profiling)[36]等多种基于专家经验构建的指标被用于识别高频类程序化交易投资者。

2)基于多规则判断的定性分析和识别

基于多规则判断的分析主要是通过构建一套判定标准，基于所有标准对高频类程序化交易投资者进行漏斗式地筛查，最终找到识别目标。这种通过一系列标准进行定性分析的方法目前主要在美国得到了广泛使用。

美国商品交易委员会的技术咨询委员会在 2012 年 10 月的草案中明确提出了高频交易属于程序化交易的范畴，并根据高频交易行为特征，形成了基于多个判定规则的高频交易识别流程模式，以便进行回顾分析。其中涉及的行为特征包括以下 4 个方面：(1)基于算法进行交易，在进行交易决策、执行指令生成等各环节中必有 1 个节点没有人为参与；(2)通过低延时技术，将响应时间最小化，涉及主机托管以及邻近地方法；(3)将指令迅速接入到场内；(4)维持在较高水平的指令信息速率，采用撤单率、市场参与者信息比率、市场参与者交易量比率 3 种衡量指标。高频交易的识别流程如图 1 所示。

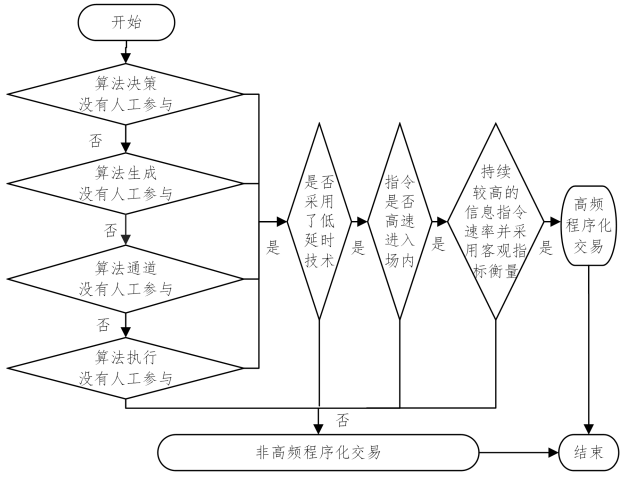

图 1 基于交易技术判断高频交易的流程

Fig. 1 Judgement process of high-frequency trading based on trading technique

2.2.2 基于机器学习的识别研究

近年来，随着大数据和人工智能的发展，逐渐出现将机器学习算法应用于不同证券期货业投资者识别场景的研究。如Niu等[37]基于复旦大学研发的IGE(Interaction Graph Embedding)[19-20]特征提取算法，并以随机森林算法作为分类器进行了投资者类型的自动识别；Xu等[34]基于循环神经网络RNN(Recurrent Neural Network)，以准确率约70%的效果实现对A股程序化交易的智能识别。

综上，目前主要的程序化交易投资者的识别方法是基于专家规则的方法，其识别效果主要依赖于规则设定的好坏，且存在无法自动更新的弊端，识别方法的泛化能力较为局限，难以应对不断变化的程序化交易模式和方法。另一方面，传统的程序化交易识别方法中存在多个定性识别，该方法的主观性较强，难以用定量的方式进行衡量。此外，纵观已公开的研究中，基于机器学习的程序化交易投资者识别效果目前停留在约70%的准确率水准上，还有一定的提升空间。

对此，本文希望基于机器学习方法，结合多种模型的优点，构建一套具备自我更新能力、能够定量分析、识别效果相对于现有研究更优的模型，旨在更好地为我国证券期货业科技监管工作服务。本文以Xu等[34]的模型为本文的基准模型(Baseline Model)，用于比较，该模型是基于市场交易数据和持仓等数据所构建的，且在A股市场程序化交易投资者识别场景下具有当下前沿的模型效果(State of the Art)，即准确率约为70%，召回率为60%。

3 模型构建

3.1 模型构建的总体流程

本文构建基于机器学习的程序化交易投资者识别模型，主要包括数据预处理、特征工程、构建专家规则、聚类分析、分类分析这5个步骤，如图2所示。

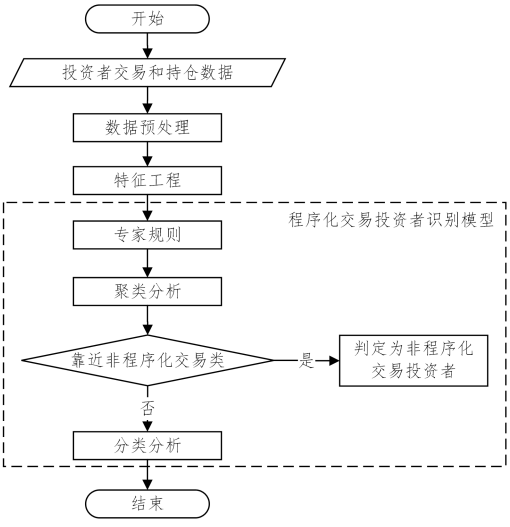

图2 基于机器学习的程序化交易识别流程

Fig. 2 Constructing process of recognition for programmatic trading based on machine learning

1)数据预处理：对缺失交易信息的投资者数据进行清洗和降噪，将持仓市值过低的中小散户剔除等。

2)特征工程：从投资金额、持仓周期、投资标的和交易频次等四大维度，基于业务经验构建投资者特征指标。

3)构建专家规则：基于业务经验构建多维度的筛选规则，识别出显著不符合程序化交易特征的投资账户。

4)聚类分析：基于投资者特征指标进行聚类，结合程序化交易投资者名单判定非程序化交易簇，并将其他疑似程序化交易投资者群体用于后续的投资者分类分析。

5)分类分析：结合程序化交易投资者名单，将聚类分析中判定的疑似程序化交易投资者进行分类分析。最后与专家规则、聚类分析步骤中判定的非程序化交易投资者结果进行统一的识别效果分析。

3.2 模型特征工程

基于证券期货行业投资者的交易行为研究经验，本文从投资者的交易频率、交易时间、交易标的和交易金额4个维度构建了9个投资者交易特征指标(指标和对应计算式详见表1)，用于刻画投资者的证券交易行为画像。其中，frequency为交易频次，period为交易周期，trading_day为交易日数量，houlding_day为持仓日数量，stockvolume为投资标的数量，stock为投资者标的对象，amount为交易金额。

表1 特征指标

Table 1 Feature indicators

类型	指标	公式
频次	总频次	$\sum_{i=1}^{n=period} frequency_i$
	日度频次	$\frac{\sum_{i=1}^{n=period} frequency_i}{period}$
	最大频次	$MAX\{\times_{i=1}^{n} frequency_i\}$
周期	总交易周期	$\sum trading\_day_i$
	平均持有周期	$\frac{\sum_{i=1}^{n=stockvolume} holding\_day_i}{stockvolume}$
标的	总标的量	$\sum_{i=1}^{n=period} stock_i$
	日度标的量	$\frac{\sum_{i=1}^{n=period} stock_i}{period}$
金额	总交易额	$\sum_{i=1}^{n=period} amount_i$
	日度交易额	$\frac{\sum_{i=1}^{n=period} amount_i}{period}$

3.3 识别模型主体构建

基于上文构建的9大投资者交易特征指标，本节基于模型融合的思想，采用将专家规则、聚类算法和分类算法叠加融合的方式，构建了一套用于识别程序化交易投资者的模型。模型主体逻辑示意图如图3所示。

首先，模型的第一层为专家规则。基于对程序化交易投资者的研究经验，本文对9个特征指标设置了不同的阈值，用于识别并剔除显著的非程序化交易投资者，剩下的疑似程序交易投资者将进入下一层进行识别。

然后，模型的第二层为聚类算法。该层中涵盖了基于质心、基于图、基于密度和基于连通性等4种类型的聚类算法，本文结合程序化交易投资者名单，对上述4类聚类算法进行效果分析。分析逻辑如下：首先，将基于各类聚类算法对投资者特征指标数据进行聚类，以给定的程序化交易投资者名单为基准，将不同聚类结果中与名单有部分重合的类均作为聚类算法筛选出的疑似程序化交易投资者群体，并基于该筛选结果进行准确率和召回率计算，以此达到将聚类算法用于分类场景的效果。而重合度为0的类将判定作为非程序化交易

类，更靠近这类的投资者将被聚类算法判定为非程序化交易投资者。最后，疑似程序化交易的投资者将进入下一层进行最后的识别。

最后，模型的第三层为分类算法。在该层中，我们分别将数据输入到两大类算法中，即可解释性较强的算法（如 Logistic Regression，Decision Tree）和可解释性弱但可能有更好识别效果的算法（如 SVM，FCneural network）。其中，可解释性较强的算法主要用于辅助寻找用于判定程序化交易投资者的主要影响因素，其他算法主要用于追求最更好的识别效果。经过分类算法分析后的结果将作为最终的识别结果。通过上述模型融合的方法，能够有效兼顾基于规则的模型和机器学习模型的优点，同时基于规则的模型筛选剔除掉的非程序化交易投资者后，能节约后续机器学习模型的训练时间和并缩小识别范围，进而提高识别效果。

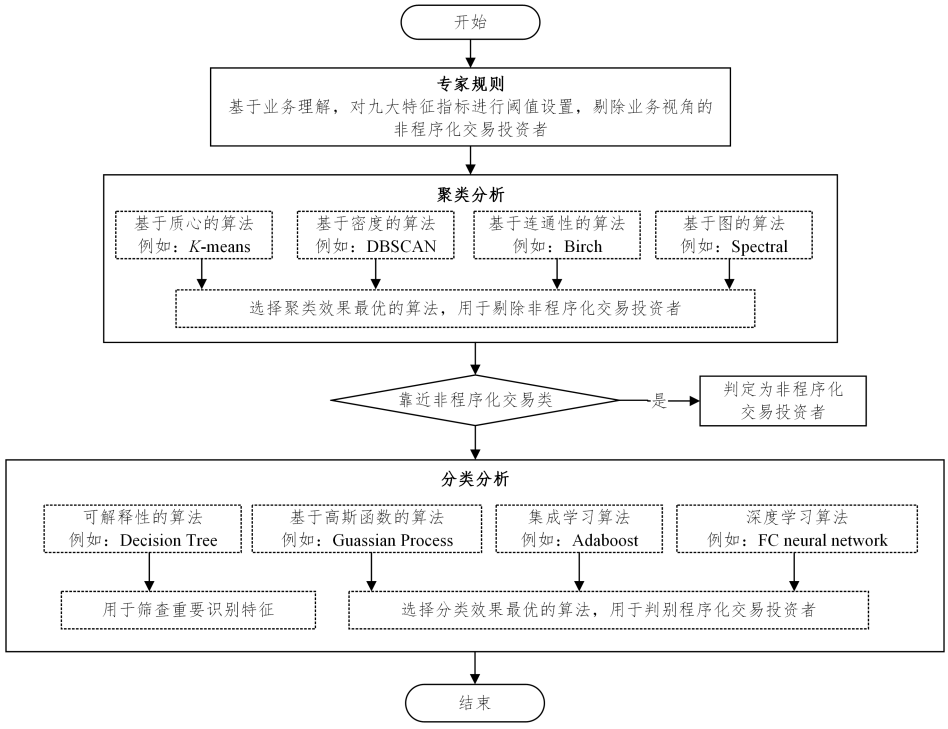

图 3 基于模型融合思想的程序化交易识别模型示意图

Fig. 3 Schematic diagram of recognition model for programmatic trading based on idea of model fusion

4 实验与结果分析

4.1 实验数据和环境

本文数据选取的是 2021 年上半年（118 个交易日）中，A 股市场全量投资者的交易和持仓明细数据，包括交易时间、交易标的、交易金额、交易方向、交易笔数等信息。分析的投资者对象包括了散户投资者、国内机构投资者和外资等在 A 股市场有进行过交易记录的投资群体。研究目标时间段为日内 9:30－15:00 的盘中竞价阶段。同时，数据集中还包含经过交易所标注和基金业协会备案的数万条程序化交易投资者数据。

本文实验环境中的硬件设备采用 Teradata 的数据仓库一体机作为数据仓库服务器，其中数据实验建模环境采用 6 节点 Teradata 2800，建模的机器 CPU 为 2 * 14 核，内存为 512GB。实验使用的软件环境中，操作系统是 Ubuntu16.04.4LTS，数据预处理和特征工程采用 SQL 脚本进行 ETL，机器学习模型训练和测试的开发语言为 Python3.7。

4.2 数据准备和预处理

4.2.1 数据样本和标签准备

考虑到资金量较小的散户投资者一般不会采用程序化交易的方式进行投资，且该类投资者个体对市场行情的影响可以忽略不记。因此，为了在不影响模型构建目标的前提下，我们剔除了持仓市值较小的投资者，将剩下超百万数量级的投资者信息作为数据挖掘样本集合（考虑到数据的敏感性，本文不作具体描述）。

为了衡量模型的效果，本文以沪深交易所曾认定的程序化交易账户以及基金业协会备案信息为程序化基金的私募基金账户作为模型对比衡量的数据，一共涉及数十万余个证券账户。本文在进行数据特征工程和模型训练前，为这些账户标注上了程序化交易投资者的标签，用于识别模型的训练和优化。

4.2.2 数据清洗和指标加工

针对待识别的证券交易投资者，实验首先筛选出交易时间与既定程序化交易投资者统计时间重叠的投资账户，即 2021 年上半年的窗口期（118 个交易日）范围内有过交易行为的投资账户。

然后，将样本数据中部分维度数据存在缺失或重复的样本账户进行数据清洗，保证样本账户在给定时间范围内存在交易，且各个研究维度的数据均不存在缺失或重复的情况。

最后，基于上文特征工程中的投资者交易特征指标计算公式，实验进行了特征指标的计算，为每一位待识别的投资者形成唯一的特征向量，并将数据按照大约 80％的训练集和 20％的验证集来划分。对所有投资者账户的每个特征采用 Z-Score 方法进行数据标准化，以减小量纲对模型训练的影响，避免单一特征过大地干扰模型效果。经过上述加工后的投资者交易特征向量将被输入到识别模型中。

4.3 模型中途识别效果分析

为了解基于模型融合的模型效果,本文进行了模型中途识别效果的分析(考虑到模型的敏感性,本文不作具体核心代码的描述),具体如下。

首先,对基于专家规则的模型进行了识别分析,模型效果如表 2 所列。可以明显看出,基于专家规则的模型的准确率不高,低于当下前沿水平。但模型的召回率 Recall 达到了 100%,即真正的程序化交易投资者均在模型识别出的名单中,同时还能筛选掉部分判定为非程序化交易的投资者,这为后续的聚类和分类分析奠定了良好的分析基础。

表 2 专家规则模型效果

Table 2 Performance of model based on expert rules

(单位:%)

Algorithm	*Precision*	*Recall*	*F*1
专家规则	62.2	100	76.6

然后,将经过专家规则模型判定的疑似程序化交易投资者输入到聚类算法中,通过 Grid Search 方法,本文在常用的参数范围中找到了各聚类算法表现最优的参数组合。同时,基于专家规则识别结果,叠加聚类算法模型的识别效果如表 3 所列。

表 3 聚类算法效果

Table 3 Performance of clustering algorithms

(单位:%)

Algorithm	*Precision*	*Recall*	*F*1
K-means	61.6	100	81.3
DBSCAN	81.2	100	89.6
BRICH	78.2	100	87.7
Spectral	67.2	100	80.3

由表 3 可以看出,基于密度的 DBSCAN 聚类算法综合表现最佳,在保证 100%的召回率下有超过 80%的准确率,故程序化交易投资者识别模型将以 DBSCAN 算法为第二层的聚类算法。基于该算法的聚类结果,靠近程序化交易类的投资者将被输入到下一层的分类算法中,其他的靠近非程序化交易类的投资者将被判为非程序化交易投资者并被筛选出来。

4.4 模型整体效果分析

经过专家规则模型和聚类算法的叠加判定,疑似程序化交易投资者将输入到分类算法中。通过 Grid Search 方法,本文在常用的参数范围中找到了各种分类算法最优的参数组合。在专家规则和聚类算法识别结果的基础上,分类算法模型的识别效果如表 4 所列。

表 4 基于各分类算法的模型效果

Table 4 Performance of classification algorithms

(单位:%)

Algorithm	Precision	Recall	F1
Logistic Regression	82.6	99.2	90.1
Decison Tree	86.1	99.3	92.2
Guassian Process	99.3	66.7	79.7
Random Forest	90.3	77.5	83.4
Adaboost	92.1	72.1	80.8
Gradient Boosting	93.9	65.6	77.2
FC Neural Network	90.1	97.7	93.7

从各算法的整体结果可以看出,以 Gaussian Process 和集成学习为代表的几类算法在程序化交易特征账户的识别上相对准确率较高,均超过 90%,但召回率相对较低。以 Logistic Regression 和 Decision Tree 这类可解释性较强的算法在程序化交易特征账户的识别上更为全面,其召回率均超过 98%,但是其准确率相对较低。作为近几年来较为流行的深度学习神经网络在准确率和召回率上综合表现最佳,在测试集上均达到了 90%以上,超过了当下的前沿效果。故程序化交易投资者识别模型将基于专家规则、DBSCAN 和全连接神经网络叠加构建。同时,考虑到决策树和逻辑回归的识别效果 F1 均超过了 90%,因此可将其作为模型辅助筛选重要投资者交易特征的工具。

结束语 程序化交易在我国证券期货业的迅速发展,引起了监管方和投资者的关注,相关的识别模型也逐渐出现。本文基于模型融合的思想,充分利用专家规则、聚类算法和分类算法的优势,有层次地将其融合,构建了一套程序化交易投资者的智能识别模型。所提模型以 90.1%的准确率和 97.7%的召回率水平,超过了当下的前沿效果(准确率 72%,召回率 60%)。同时,还加入了决策树和逻辑回归,增强了模型的可解释性,便于监管业务人员理解。

值得注意的是,本文中的实验具有一定的局限性。一方面识别模型所用到的标签数据主要来自沪深交易所和基金业协会已经识别出的程序化交易账户,模型在其他潜在程序化交易投资者识别上的通用性和普适性有待进一步研究。另一方面,本文所选数据的时间范围有限,模型的泛化能力和有效周期有待进一步研究。

参 考 文 献

[1] WU L H. Yi Huiman:Research and launch of relevant measures to further expand opening up[N]. Economic Information Daily, 2021-09-07(1).

[2] BA S S, WANG Y C. The impact of high-frequency trading on securities market:aA review[J]. Securities Market Herald, 2019(7):42-51.

[3] XIONG X, YUAN H L, ZHANG W, et al. Program trading and its risk analysis [J]. Journal of UESTC(Social Sciences Edition), 2011, 13(3):32-39.

[4] LU Y. Development status, problems and suggestions of programmatic trading in China's capital market[J]. Chinese and Foreign Entrepreneur, 2015(1):106.

[5] XIANG Y T, SUN Y H, ZHENG J H, et al. International trend and enlightenment of high-frequency trading regulation[J]. Review of Financial Development, 2015(9):51-56.

[6] ZHANG M X. Frequent Cancellation and Market Manipulation Determination of High-frequency Trading:From the Perspective of False Statement Manipulation Case of US Treasury Futures [J]. Securities Market Herald, 2016(5):73-78.

[7] BOEHMER E, FONG K, WU J. International evidence on algorithmic trading[C]// AFA 2013 San Diego Meetings Paper. 2012:35-82.

[8] HENDERSHOTT T, JONES C M, MENKVELD A J. Does algorithmic trading improve liquidity? [J]. The Journal of Finance, 2011, 66(1):1-33.

[9] JENNIFE R, CONRA D, SUNI L, et al. High-frequency quoting, trading, and the efficiency of prices[J]. Journal of Financial Economics, 2015, 116(2):271-291.

[10] YU Y R. Research on the relationship between treasury bond

futures and spot prices based on high-frequency data[D]. Harbin: Harbin Institute of Technology, 2017.
[11] KIRILENKO A, KYLE A S, SAMADI M, et al. The flash crash: High frequency trading in an electronic market[J]. The Journal of Finance, 2017, 72(3): 967-998.
[12] CONRAD J, WAHAL S, XIANG J. High-frequency quoting, trading, and the efficiency of prices[J]. Journal of Financial Economics, 2015, 116(2): 271-291.
[13] HENDERSHOTT T, MOULTON P C. Automation, speed, and stock market quality: The NYSE's hybrid[J]. Journal of Financial Markets, 2011, 14(4): 568-604.
[14] BARON M, HAGSTRMER B, KIRILENKO A. Risk and Return in High-Frequency Trading[J]. GRU Working Paper Series, 2017, 53(3).
[15] CLARK C. How to keep markets safe in the era of high-speed trading? [J]. Chicago Fed Letter, 2012(303): 1.
[16] BROGAARD J, CARRION A, MOYAERT T, et al. High frequency trading and extreme price movements[J]. Journal of Financial Economics, 2018, 128(2): 253-265.
[17] BUDISH E, CRAMTON P, SHIM J. The high-frequency trading arms race: Frequent batch auctions as a market design response [J]. The Quarterly Journal of Economics, 2015, 130(4): 1547-1621.
[18] MENKVELD A J. High-Frequency Traders and Market Structure[J]. Financial Review, 2014, 49(2): 333-344.
[19] MENG X J, MENG X L, HU Y. Research on investor sentiment index based on text mining and Baidu index[J]. Macroeconomic Research, 2016(1): 144-153.
[20] ZHANG Y, XIONG Y, KONG X N, et al. IGE+: A Framework for Learning Node Embeddings in Interaction Graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(3): 1032-1044.
[21] WU X L, LI J G, WANG Z M. The impact of high-frequency trading on the market[J]. Tsinghua Financial Review, 2016(2): 16-24.
[22] YUAN Y K. Application of artificial intelligence in securities market supervision[J]. Information Technology and Standardization, 2019(5): 15-19.
[23] HUANG D L, WEN F H, YANG X G. Investor sentiment index and empirical evidence of China's stock market[J]. Systems Science and Mathematics, 2009, 29(1): 1-13.
[24] MENG H. International experience of high-frequency trading supervision[J]. Journal of Financial Market Research, 2014(1): 98-102.
[25] FU Q. Research on risk early warning of China's stock market in the post-stock reform era based on support vector machine [D]. Chengdu: Chengdu University of Technology, 2014.
[26] ZOU J J, ZHANG Z Y, QIN Z. Application of GARCH model in calculating the value at risk of China's stock market[J]. Systems Engineering-Theory & Practice, 2003(5): 20-25, 135.
[27] YUAN Y K, LI G, ZHAO Z X, et al. Research on influencing factors of stock market inflection point based on machine learning[J]. Computer Science, 2021, 48(S1): 165-168, 177.
[28] YE W. Risk control strategy of programmatic trading in China's capital market[J]. Securities Market Herald, 2014(8): 46-52.
[29] YONG C. Research on high-frequency trading supervision in securities and futures market[J]. Finance and Economics, 2019 (2): 83-87.
[30] JIANG Z. Research on potential risks and regulatory system of programmatic trading[J]. Research on Financial Regulation, 2017(6): 78-94.
[31] LUO Z L. Research on international experience in programmatic trading regulation and legislation[C]// China Futures Association. China Futures Association, 2016.
[32] ZHANG P P. High-frequency trading regulation from the perspective of Everbright oolong finger incident[J]. Journal of Financial Development, 2013(10): 57-60.
[33] LU S. On the identification of high-frequency trading manipulation market and its legal regulation[J]. Social Science Dynamics, 2017(9): 68-76.
[34] XU G B, ZHANG W. DeepEye: A programmatic transaction recognition and classification method based on deep learning[J]. Big Data, 2018, 4(5): 94-102.
[35] GUO C G, CUI E T, XIONG X P. The Enlightenment of Overseas High-frequency Trading Regulatory System on China's Futures Market Supervision-Based on the Perspective of Economic and Technical Means[J]. Journal of Financial Development Research, 2021(2): 65-72.
[36] ZHU W H, ZHANG W. Research on the regulatory dynamics and regulatory indicators of programmatic trading in overseas markets[J]. The Foreland of Trading Technology, 2014(19): 28-35.
[37] NIU Z, XIONG Y, ZHANG Y. Rresearch on automatic identification method of investor type based on artificial intelligence [J]. The Foreland of Trading Technology, 2017(28): 57-60.

YUAN Yukun, born in 1994, postgraduate. His main research interests include machine learning and natural language processing.

XU Gang, born in 1969, postgraduate. His main research interests include data mining and quantitative finance.