arXiv:2512.14042v1 [cs.CE] 16 Dec 2025

基于多源金融数据和投资者知识表征的股票市场指数预测动态堆叠集成学习

Ruize Gao[1], Mei Yang[2], Yu Wang[3*], ShaozeCui[41] 数字经济实验室，北京应用数学研究院，中国北京。[2] 管理学院，浙江大学，中国杭州。[3*] 经济与工商管理学院，重庆大学，中国重庆。[4] 管理学院，北京理工大学，中国北京。

*通讯作者。电子邮箱：yuwang@cqu.edu.cn；贡献作者：gaoruize@bimsa.cn；yangm152@zju.edu.cn；shaoze-cui@foxmail.com；

摘要

不同金融数据源的模式差异显著，因此投资者在信息处理过程中表现出异质性认知行为。为捕捉不同模式，我们提出一种基于投资者知识特征表示的新方法，称为两阶段动态堆叠集成模型，旨在有效提取和整合多源金融数据中的特征。第一阶段，我们从投资者视角出发，识别全球股票市场指数、行业指数和财经新闻中不同的金融数据属性。随后，我们针对这些属性设计合适的神经网络架构，以生成有效的特征表示。基于学习到的特征表示，我们设计了多个元分类器，并针对每个时间窗口动态选择最优分类器，使模型能够有效捕捉并学习不同时间周期中呈现的不同模式。为评估所提出模型的性能，我们将其用于预测中国股市中上证综合指数、深证成份指数和创业板指数的每日走势。实验结果证明，我们的模型能够有效提升预测性能。在准确率指标方面，我们的方法在上证综合指数、SZEC 和 GEI 指数上的表现分别比最佳竞争模型高出 1.42%、7.94% 和

7.73%。此外，我们基于所提出的模型设计了一种交易策略。经济效益结果表明，与其他竞争性交易策略相比，我们的策略在累积收益率和夏普比率方面表现更优。

关键词： 股票市场指数、多源金融数据、特征表示、深度学习、动态堆叠集成

1 引言

预测股票市场指数（SMI）走势对于投资者和监管机构而言都是一项重要任务。对于投资者来说，准确的预测对于制定有效的投资策略至关重要，有助于降低风险并提高收益率。对于监管机构而言，精准的SMI预测能够支持制定有据可依的政策，以维持市场稳定。然而，预测SMI走势是一项困难且具有挑战性的任务，因为股票市场本质上具有非线性、动态性和复杂性，并且受到高度噪声的影响（Zhong and Enke 2017；Gao 等人，2022a）。

影响目标SMI的数据源多种多样。首先，目标市场的历史市场数据是一个重要的数据源，例如开盘价、收盘价、最高价、最低价、成交量、交易额和收益率（Yun 等人，2021）。其次，相关市场的历史数据也发挥着重要作用。例如，Hoseinzade 和 Haratizadeh （ 2019 ）纳入了多个股票市场指数、宏观经济指标、原油市场数据和贵金属价格，以预测目标SMI的走势。他们的研究结果表明，整合不同市场的数值数据能够显著提高预测性能。此外，行业层面数据也具有相关性，因为不同行业的繁荣或衰退会对整体SMI走势产生重大影响。除数值数据外，文本信息也日益得到利用，主要包括财经新闻和社交媒体内容。财经新闻通常由篇幅较长、专业性较强的文本构成，而社交媒体评论通常是简短且无结构的内容。由于财经新闻文本具有较强的专业性，其往往比社交媒体评论更能吸引投资者关注度。

为了准确预测SMI走势，现有研究日益采用多源数据视角，提出先进的方法，将不同数据源整合起来，以提升目标SMI的预测性能（Zhong and Enke 2017；Ye et al. 2024）。在这些方法中，最常用的可分为三大类：基于主成分分析（PCA）的方法（Zhong andEnke 2017；Weng et al. 2018）、基于张量的方法（Liet al. 2017；Zhang et al. 2018；Li et al. 2021；Wang et al. 2023）以及基于深度学习的方法（Wang etal. 2018；Xuetal. 2020；Ma et al. 2022；Yeet al.2024）。第一类方法采用主成分分析（PCA）、模糊鲁棒主成分分析（FRPCA）和基于核的主成分分析（KPCA）等技术进行降维。随后，运用机器学习方法学习输入变量与股票市场价格或指数之间的关系。第二类方法使用张量表示从多模态数据中提取的信息

。例如，Li et al. (2021) 将张量的三个维度分别表示为公司特定特征、专业观点特征和情感特征。基于张量表示，通过张量分解与重构来提取信息。张量分解类似于高阶主成分分析，而张量重构能够揭示不同数据源之间的潜在信息关系。第三类方法着重设计针对各类输入的专用特征提取层，以便从多模态数据中有效提取相关特征。随后，利用集成层合并从不同数据模态提取的特征，从而实现更准确的股票价格预测。例如，Ma et al. (2022) 采用一种策略：利用线性层提取目标股票的新闻情感特征，并利用图卷积层捕捉相关股票的新闻情感特征。这种方法不仅凸显了复杂神经网络架构在处理复杂数据输入方面的效用，还增强了模型辨别并整合多方面市场信息的能力。

尽管以往研究采用多源数据视角和方法融合异构信息、提升预测性能，但仍存在若干重要局限。首先，这些方法主要侧重于特征级表示学习和静态融合。因此，基于主成分分析的技术在降维过程中往往会丢失重要的非线性结构，而基于张量的模型则依赖相对稳定的模态间关系，这与真实金融市场随时间变化的动态特征并不相符。深度学习方法虽然具有更强的特征提取能力，但通常采用固定的融合架构，无法适应不断变化的市场状况。更重要的是，大多数现有模型在输入与输出之间建立直接的非线性映射，却未考虑投资者如何在认知层面处理多源信息。实际上，多源金融数据对SMI走势的影响是通过投资者的行为传递的，即投资者如何识别各数据源中的模式、形成预期并作出决策。近期研究还表明，视觉信息或专家经验（两者均蕴含隐性的投资者认知）能够显著提升预测性能（Jiang 等（2023）；Cao 等（2024））。因此，有必要提出一种多源金融数据融合框架，纳入投资者认知视角，从不同数据源中提取深层特征表示，并在不同时间窗口内动态调整预测机制。

因此，我们提出一种基于投资者知识表征的两阶段动态堆叠集成（TDSE）模型，旨在有效提取和整合多源金融数据中的特征。在第一阶段，我们通过考察区域内股市指数之间的趋势相似性、行业轮动效应以及不同新闻提供商的影响差异，考虑不同数据源的内在模式。为捕捉全球股票市场指数、行业指数数据和财经新闻数据中的多样化模式并提取其内在特征，我们采用了量身定制的特征提取方法，即多分支卷积神经网络、基于谱聚类的多分支卷积神经网络，以及基于证据规则的循环神经网络。每种模型都旨在有效提取深层

特征表示，从每种金融数据源中提取特征。在第二阶段，我们采用多个元分类器，从提取的特征中有效学习。鉴于提取特征与SMI走势之间的关系具有内在的时间动态特性，我们会自适应地选择最优元分类器，以有效捕捉每个时间段内不断演变的数据模式。为验证所提出模型的有效性，我们将其应用于中国股市SMI走势预测。实验结果表明，我们的模型在预测表现和经济绩效方面均优于竞争方法。

本文其余部分的组织结构如下。第2节全面回顾相关文献。第3节深入介绍本研究提出的方法论。第4节报告实验研究结果。第5节讨论理论和实践意义。最后，我们在第6节总结全文并展望未来工作。

2 相关工作

2.1 金融预测中使用的多源金融数据

2.1.1 金融预测中使用的历史市场数据

在金融预测中，市场数据发挥着至关重要的作用。它包括各种组成部分，如开盘价、收盘价、最高价、最低价、成交量、交易额和收益率。这些市场数据反映了投资者交易行为。以往研究广泛探讨了历史市场数据中的模式，并利用这些模式预测SMI走势（Constantinou etal. 2006；Orimoloye et al. 2020；Yunet al. 2021）。实验结果表明，历史市场数据对SMI走势具有显著的预测能力。

随着经济全球化的发展，人们对理解不同股票市场之间的相互作用产生了日益浓厚的兴趣。研究人员已致力于探索多个股票市场指数之间的信息流及其关系。Marschinski 和 Kantz (2002) 运用转移熵方法分析国际股票市场指数之间的相互依赖关系，揭示了市场间传导过程中的重要见解。这项研究为通过考量全球股票市场指数来预测目标SMI提供了重要的理论基础。在近期研究中，Kia 等人 (2018年) 研究了全球股票市场指数、黄金价格和原油数据对预测目标 SMI 走势的能力。研究结果强调，与仅依赖目标股票市场历史数据的方法相比，纳入其他市场的历史数据能够显著提升预测性能。同样，Hoseinzade 和 Haratizadeh (2019) 将多个股票市场指数、经济数据、原油市场数据和贵金属价格纳入模型，以预测目标 SMI 走势。研究结果表明，整合不同市场的数值数据能够提升预测性能。

除了目标市场的历史数据和相关数据外，行业数据对目标SMI走势的影响也不容忽视。在股票市场中，行业由提供相似产品或服务的上市公司群体构成。行业指数根据上市公司的股票价格计算，

反映各行业的整体发展状况。值得注意的是，行业指数数据呈现出行业轮动效应，即不同行业交替经历上涨和下跌时期（Su2017）。以往研究深入探讨了这种行业轮动效应对目标SMI的影响。例如，He (2001)发现，在稳定时期，中国股市中行业轮动效应出现的频率很高，并且会影响目标SMI。此外，Pei 和 Zhu(2019)建立了行业轮动效应与经济周期之间的联系，并指出这种效应在不同时间段各有差异。根据统计分析，Su (2017)发现，2013年至2014年，房地产行业领先钢铁行业和家电行业1至2个时间段。然而，2016年这一时间间隔延长至1至3个时间段。

2.1.2 金融预测中使用的财经新闻数据

在信息快速扩张、在线数据浩如烟海的时代，文本数据的可用性呈指数级增长。在金融领域，文本数据主要包括财经新闻和社交媒体数据。财经新闻文本由具备专业金融知识的作者撰写，并在发布前经过编辑的严格审查。因此，凭借其固有的专业性，与相对不受限制的社交媒体数据相比，财经新闻文本更能吸引投资者关注。由于财经新闻具有公认的可靠性并提供专家见解，投资者更倾向于关注财经新闻。

鉴于财经新闻属于非结构化数据，在用作预测模型的输入之前，必须将其转换为结构化数据。一种方法是利用从财经新闻中提取的词向量来预测股市趋势。例如，Schumaker 和 Chen (2009)采用词袋、名词短语和命名实体等技术来表示财经新闻，并表明将这些表示与市场数据相结合可以提升预测性能。Shynkevich 等人(2016年)研究了从财经新闻中提取的词向量对股票走势的影响。实验结果表明，整合多个行业的词向量有助于提升预测性能。Nam 和 Seong (2019)利用涉及相关公司的财经新闻词向量来预测目标股票走势，并表明考虑个股之间的相关性可以增强财经新闻的预测能力。

虽然词向量可用作预测 SMI 走势的输入，但其处理方式会削弱财经新闻的可解释性，使投资者难以仅凭词向量做出明智决策。因此，研究人员转而采用情感分析技术，以增强财经新闻的可解释性并助力投资者进行决策。Li 等人(2014) 强调，金融新闻情感在连接新闻与股市走势的链条中发挥着至关重要的作用。（Li 等人 2014）中的实验结果表明，在股票、行业和股票指数层面，利用新闻情感进行预测的效果优于使用新闻向量。

2.1.3 金融预测中的多源数据融合

多源数据融合的相关研究见表 1。从表 1 可以看出，数值数据类型主要包括市场数据、技术指标、基本面数据、宏观经济数据和汇率数据。其中，市场数据是最常用的数值数据，体现了其在金融预测中的重要性。另一方面，在文本数据表示方面，情感分析是应用最广泛的文本分析技术。此外，以往研究通常依靠词嵌入方法，有效表示来自财经新闻和社交媒体的数据。值得注意的是，研究人员通常会在研究中纳入多个数据源，数量从三个到五个不等。这表明研究人员认识到，利用多模态数据有助于提升预测性能。

表1 多源金融数据整合相关研究

作者	金融数据来源	融合方法
Zhong and Enke (2017年)	市场数据，宏观经济数据，汇率数据	特征降维：主成分分析模型
Li 等人（2017年）	市场数据、财经新闻：名词项向量与情感；社交媒体数据：情感	张量分解与重构—构建
Weng 等人(2018年)	市场数据、技术指标、Google趋势，维基百科点击量、财经新闻：情感	特征降维：主成分分析模型
Zhang 等人 (2018年)	市场数据，财经新闻：事件；社交媒体数据：情感	耦合矩阵 and 张量分解
Wang 等人(2018年)	技术指标、社交媒体数据：情感	深度随机集成学习
Xu 等人(2020)	市场数据、社交媒体数据：词语嵌入	Deep 基于学习的特征提取
Li 等人(2021)	市场数据、基本面数据、金融新闻：情感	张量分解与重构
Ma 等(2022)	市场数据、技术指标、金融新闻：基于词嵌入的	Deep 学习特征提取
Wang 等人 (2023)	基本面数据、公司经营状况，财经新闻：情感；	特征降维：张量稳健性PCA
Li 等人(2023)	成交量和价格、技术指标及宏观因素data	特征融合残差长短期记忆网络
Gao 等人 (2024)	市场数据、社交媒体数据：投资者情感	OLS
Dong 等人 (2024)	市场数据，在线评论：词语嵌入	BERT-BiLSTM模型
Zhao 等(2025)	市场数据、宏观经济数据、其他市场数据类型（债券、外汇、商品）	随机森林
沈等人（2026）	市场数据、宏观数据、社交媒体	深度学习

关于多源金融数据的整合过程，大致可分为两种类型：特征降维方法和特征提取方法。在特征降维方法中，PCA方法能够有效降低多维向量数据的维度。另一方面，基于张量的方法使用张量表示输入数据，并利用分解和重构技术捕捉多源数据之间的相互作用。至于特征提取方法，深度学习方法是一种有效整合

多源数据的方法。基于端到端机制，深度学习方法逐步从不同数据源提取特征，并利用这些特征预测目标 SMI 走势。

2.2 使用多源金融数据的预测技术

基于前述综述，我们发现，基于数值数据或文本数据的各种金融预测方法通常采用支持向量机模型、人工神经网络模型及其他单一技术，这些方法适用于处理单模态数据。然而，在处理多模态金融数据的复杂性时，混合模型已成为金融预测中广泛采用的方法。这些混合模型由多种技术组成，例如基于主成分分析的机器学习模型、基于张量的机器学习模型以及先进的深度学习模型。通过结合这些不同的方法，混合模型能够有效捕捉不同数据模态之间的交互作用，从而提升预测性能。

第一类模型称为基于主成分分析的机器学习模型，它将主成分分析方法与支持向量机、人工神经网络模型等机器学习技术相结合，用于预测SMI走势。主成分分析方法是一种特征降维方法，本质上是通过构建新的坐标系，将多模态数据的原始特征空间映射到新的特征空间。这种转换能够生成新的特征，有效减少冗余信息。例如，Zhong and Enke (2017年)采用主成分分析、核主成分分析和FRPCA模型，对市场数据、宏观经济数据和汇率数据进行降维。实验结果表明，与其他模型相比，将主成分分析类方法与人工神经网络模型相结合，在预测标普500指数走势方面具有更优的预测性能。同样，Weng 等人 (2018年)利用主成分分析模型从市场数据、技术指标、财经新闻、Google 趋势和维基百科点击量等多种来源中提取特征。随后，他们采用Boosting回归树（BRT）模型预测多只股票的走势。结果显示，所提出的模型表现出良好的效果，在预测准确率方面优于其他竞争方法。这些研究探讨了基于主成分分析的方法从多模态金融数据中提取有效特征的潜力。

PCA方法主要处理基于向量的数据表示。相比之下，第二类模型称为基于张量的机器学习模型，它将张量用作基本数据结构（Lee 等人2025）。张量通过纳入高阶维度扩展了向量和矩阵的概念。基于张量的方法能够捕捉不同模态之间复杂的相互作用，因此在多模态金融数据分析领域日益受到关注。例如，Li 等人(2017年)采用基于张量的模型，整合市场数据、财经新闻和讨论论坛。研究结果表明，所提出的基于张量的模型优于Tucker模型和迭代优化模型，证明了张量在捕捉不同模态之间复杂关系方面的有效性。Zhang 等人(2018年)采用基于张量的分解与重构方法，将股票量化数据、社交媒体数据和网络新闻数据结合起来。研究结果表明，所提出的模型优于传统方法，例如

支持向量机和PCA+支持向量机模型在准确率方面的表现。Wang 等人(2018年)设计了一种新方法，利用基于张量的表示来表征基本面数据、情绪特征和市场信息。他们进一步提出了一种基于注意力机制的长短期记忆模型，用于预测股票走势。实验结果表明，他们提出的模型在准确率和马修斯相关系数方面优于竞争方法。此外，Li 等人(2021)采用张量变换方法来表示市场数据和新闻数据，随后使用长短期记忆模型预测股票走势。实验结果表明，他们提出的模型在分类性能和模拟投资收益方面优于其他竞争模型。Sn´aˇsel 等人(2024)提出了一种针对选股问题的新框架，该框架基于多源数据融合和决策级融合。框架采用基于向量的数据表示，并运用动态时间规整（DTW）和定制损失函数，以提高股票时间序列预测的准确率。总体而言，这些研究探讨了基于张量的模型在有效整合多模态金融数据，以及提升金融领域不同应用场景预测准确率方面的潜力。

第三类模型，即先进的深度学习模型，模拟了人类行为的认知过程。这类模型首先从原始数据中提取低层次、简单的特征，然后通过增加层数，逐步提取更高层次、更抽象的特征。深度学习模型已广泛应用于相关研究中的多模态金融数据整合。例如，Xu 等人（2020）提出了一种基于注意力机制的深度学习方法，用于整合市场数据和社交媒体数据。实验结果表明，所提出的模型优于竞争方法，展现出更优的预测性能。同样，Wang 等人（2018年）提出了一种深度随机子空间集成方法，用于融合技术指标和社交媒体数据。结果验证了所提出的模型在有效结合这两类数据方面的效果。在另一项研究中，Ma 等人（2022）采用综合方法，提取目标股票的量化指标、目标股票的新闻特征以及重要相关股票的新闻特征。随后，他们使用双向长短期记忆网络（BiLSTM）模型预测股价走势。实验结果表明，所提出的模型在分类性能和金融评价指标方面均优于其他竞争方法。Li 等人（2023）提出了一种基于特征融合残差长短期记忆网络（FFRL）的模型，通过结合成交量和价格、技术数据及宏观数据等多源、多频信息来预测中国股市。实验结果表明，与传统深度学习模型相比，FFRL的性能有显著提升。此外，Dong 等人（2024）开发了一种基于多源数据融合和词嵌入技术的BERT-双向长短期记忆网络预测模型，将在线社区中的用户评论转换为特征向量。该模型在上证50指数成分股的价格预测中表现良好，显著优于传统方法。基于上述研究，我们发现，深度学习技术为整合多模态数据和提升预测性能提供了稳健的框架。

2.3 我们的贡献

本研究从以下几个方面为该领域作出贡献。

(1) 我们采用以投资者为中心的视角，全面考量各种影响因素背后的内在模式，并开发了一种创新的特征提取框架来表示异构数据源。学习得到的特征表示使模型能够发现并利用各数据源中独特的模式，从而提取信息量更丰富且与预测相关的特征，以预测目标 SMI 走势。

(2) 我们提出了一种新颖的动态堆叠集成模型，可有效整合从多源金融数据中提取的特征。通过考虑提取特征与目标 SMI 走势之间关系所固有的时间动态，我们的模型能够针对每个时间段动态选择最优元分类器。

(3) 我们开发了一种分阶段优化方法，用于优化所提出的模型并增强其泛化能力。分阶段优化过程使我们能够以系统的方式微调模型超参数。分阶段优化方法有助于降低解空间的复杂度。

3 方法论

3.1 问题陈述

有多种因素会对SMI走势产生显著影响。在这些数据源中，有三种尤为重要：全球SMI数据$\boldsymbol{X}_{1,\ t} = \{x_{11,t}, x_{12,t}, \ldots, x_{1n_1,t}\}$行业指数数据 $\boldsymbol{X}_{2,\ t} = \{x_{21,t}, x_{22,t}, \ldots, x_{2n_2,t}\}$以及财经新闻数据 $\boldsymbol{X}_{3,\ t} = \{x_{31,t}, x_{32,t}, \ldots, x_{3n_3,t}\}$。值得注意的是，这些数据源具有不同的模态：全球SMI数据和行业指数数据属于数值型数据，而财经新闻数据属于文本型数据。投资者根据这些数据做出决策时，会分析特定数据类别中的整体模式，以评估未来的市场表现。因此，识别并利用各数据源中的内在模式，对于提升目标SMI走势的预测表现至关重要。然而，从投资者视角发现整体模式、捕捉这些独特特征，并在预测模型中整合多模态数据，都是极具挑战的任务。克服这些挑战对于提高预测模型的准确率和有效性至关重要。

为应对上述挑战，我们提出了一种由特征提取模型和SMI预测模型组成的两阶段动态堆叠集成模型。具体而言，我们观察到，全球股票市场指数数据内部存在较强的区域内相关性，行业指数数据中存在行业轮动效应，不同新闻提供商的影响也存在差异。有效整合这些数据源固有的内在模式，并提取其特征以准确预测目标SMI走势，已成为一项关键挑战。为此，我们采用了三种不同的特征提取过程，如下列公式所示：

$fe_{1,t} = FE_1(\boldsymbol{X}_{1,t})$、$fe_{2,t} = FE_2(\boldsymbol{X}_{2,t})$、$fe_{3,t} = FE_3(\boldsymbol{X}_{3,t})$。 FE_1是多分支卷积神经网络（MBCNN）模型，其中每个分支内的卷积单元能够有效地从区域内股市指数中提取特征。基于参数共享的卷积运算使这些单元能够将区域股市指数相似性整合到特征提取过程中。 FE_2 是基于谱聚类的MBCNN（SC-MBCNN）模型，其中首先使用谱聚类算法，根据行业指数数据形成多个行业聚类。在这些已建立的行业聚类基础上，随后利用MBCNN模型提取综合特征，从而捕捉行业轮动效应。 FE_3 是基于证据推理规则的循环神经网络（RNN-ER），其中首先使用RNN模型为每个新闻提供商构建基分类器。该方法能够有效捕捉不同新闻提供商在预测能力上的差异。随后，使用证据推理规则整合来自多个新闻提供商的提取特征，增强整体特征表示。这些特征提取过程是捕捉各数据源潜在模式的关键步骤，有助于后续预测实现准确预报。

随后，利用预测模型有效地组合提取特征，如下式所示：

$$y_{t+1} = f(fe_{1,t}, fe_{2,t}, fe_{3,t}), \tag{1}$$

其中， f 是动态堆叠集成函数，能够捕捉提取特征与目标 SMI 走势之间关系所固有的时间动态。 y_{t+1} 是目标 SMI 走势，其计算公式如下：

$$y_{t+1} = \frac{Close_{t+1} - Close_t}{Close_t}, \tag{2}$$

其中， $Close_t$ 表示第 $t\text{-}th$个交易日的收盘价。

3.2 模型框架

如前所述，由于各数据源具有独特的内在模式，利用多模态财经数据预测目标 SMI 走势是一项重大挑战。为了考虑各数据源固有的内在模式，并从不同数据源中有效提取信息丰富的特征，我们提出了 TDSE 模型，用于整合三个关键数据源：全球SMI数据、行业指数数据和财经新闻文本。所提出的模型框架如图 1 所示，由两个阶段组成：投资者知识驱动的特征提取（阶段1）和自适应特征融合（阶段2）。

在阶段1中，我们首先从投资者的视角分析不同金融数据源的特性，然后采用三种合适的模型从中提取特征。具体而言，MBCNN模型用于捕捉与区域内股市指数相似性相关的特征，从而揭示全球金融格局中的相关性。SC-MBCNN模型着重提取行业轮动效应的特征，为了解特定板块的周期性提供参考。

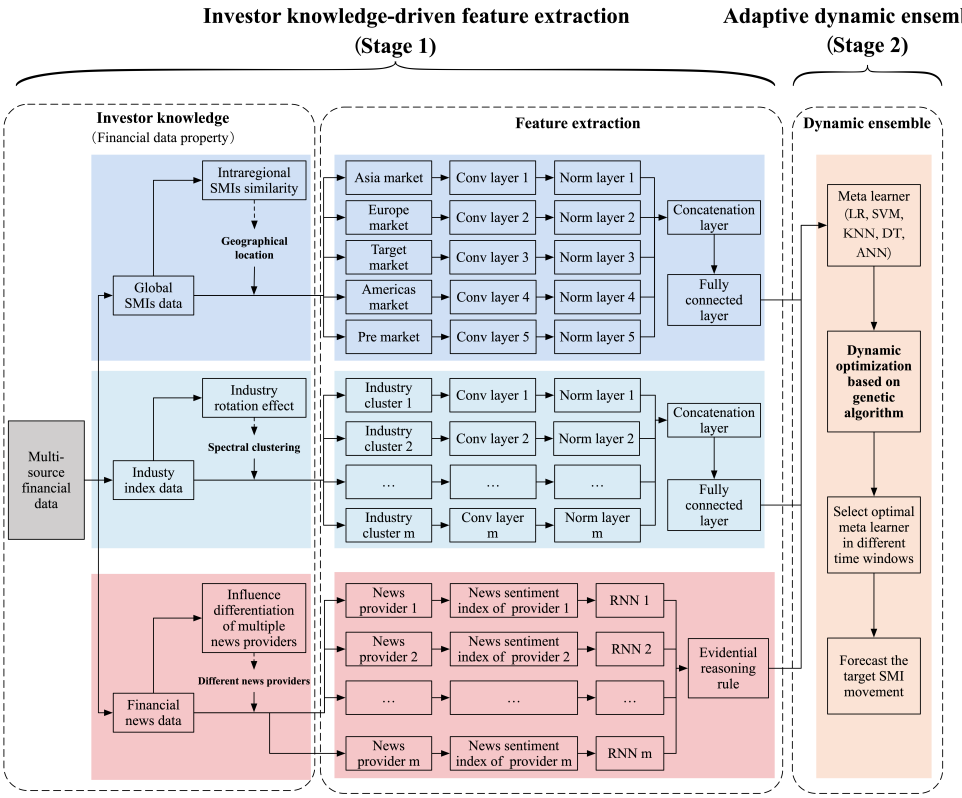

图1 所提出模型的框架

RNN-ER模型用于提取与不同新闻提供商之间影响差异相关的特征，从而全面了解不同新闻来源的影响。该系统化方法确保特征提取过程运用投资者知识，揭示各数据源的内在模式，为准确预测目标SMI走势提供有价值的参考。

在阶段2中，阶段1提取的特征会在动态堆叠集成模型中得到自适应利用，以进行特征融合和目标 SMI 走势预测。具体而言，我们采用了五种不同的元分类器：逻辑回归、支持向量机、K近邻、决策树和人工神经网络。之所以选择多种元分类器，是因为它们各有优势，且适用于不同的时间段。由于每个时间段都需要选择最合适的元分类器，因此元分类器的选择对预测性能有重要影响。基于所提出的模型，我们有效聚合了提取特征，确保这些特征协同发挥作用，从而提升整体预测性能。

3.3 阶段1：投资者知识驱动的特征提取

本节介绍投资者知识驱动的特征提取（阶段1）过程，该过程主要涉及全球股票市场指数、行业数据和财经新闻的特征提取。

3.3.1 全球股票市场指数的特征提取

本研究使用的第一个数据源是全球SMI数据。在这些数据中，我们观察到区域内股市指数之间存在显著的相似性，即同一区域的股市指数往往呈现相似的走势。考虑这一数据源的独特属性不仅至关重要，而且有助于揭示关键信息，从而显著提升预测性能。因此，为了有效提取区域内股市指数的相似性特征，本研究采用多分支卷积神经网络，具体如下所述。

根据地理位置，我们将市场划分为五个分支进行特征提取：亚洲市场、欧洲市场、美洲市场、Tar- get市场、盘前市场。对于各区域市场中的股票市场指数，我们使用若干 $R_{O\text{-}C}$、$R_{C\text{-}C}$和 $R_{C\text{-}O}$ 作为模型的输入。 $R_{O\text{-}C}$、 $R_{C\text{-}C}$ 和 $R_{C\text{-}O}$ 分别表示 $Open\text{-}to\text{-}Close$、$Close\text{-}to\text{-}Close$和 $Close\text{-}to\text{-}Open$的收益率。本研究所用股票市场指数的详细说明见附录A。

对于 $i\text{-}th$ 分支，输入向量表示为 $\boldsymbol{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}\}$，其中 $x_{n_i}^{(i)}$ 是 $i\text{-}th$ 分支中的 $n_i\text{-}th$ SMI。给定卷积核 $\boldsymbol{w}^{(i)} = [w_1^{(i)}, w_2^{(i)}, \ldots, w_F^{(i)}]$，其中 F表示卷积核的长度，使用卷积运算提取区域内股市指数的相似度特征。选择卷积运算的原因在于其局部相关性和权重共享的特性，这有助于捕捉区域内股市指数的相似度。卷积运算的计算过程如下式所示：

$$V_j^{(i,l)} = \theta(\sum_{k=1}^{F} w_k^{(i)} x_{j+k-1}^{(i)}), \tag{3}$$

其中， $V_j^{(i,l)}$ 是 $i\text{-}th$ 分支中 $l\text{-}th$ 层的 $j\text{-}th$ 列的值， $w_k^{(i)}$ 是卷积核 $k\text{-}th$ 列的权重， θ 是修正线性单元（ReLU）激活函数 $\theta(x) = \max(x, 0)$。

为提高模型训练效率，同时确保模型收敛至最佳性能，采用归一化层对特征提取层的分布进行有效归一化。给定所获得的 $l\text{-}th$ 层 $\boldsymbol{V}_{(i,\ l)} = \{V_{(i,\ l)1}, V_{(i,\ l)2}, \ldots, V_{(i,\ l)k}, \ldots, V_{(i,\ l)E}\}$，通过以下公式对每个维度的值进行归一化：

$$\hat{V}_k^{(i,l+1)} = \frac{V_k^{(i,l)} - \mu_V}{\sigma_V}, \tag{4}$$

其中， μ_V 是均值， σ_V 是标准差，$\hat{V}_k^{(i,l+1)}$ 是 $l+1$ 层中$k\text{-}th$ 维的值。由于该

层的特征表达的含义可能会在归一化过程中发生变化，为确保特征变换实现恒等变换，采用以下公式解决该问题：

$$Z_k^{(i,l+1)} = \gamma_k^{(i)} \hat{V}_k^{(i,l+1)} + \beta_k^{(i)}, \tag{5}$$

其中，$Z_k^{(i,l+1)}$ 表示 $l+1$ 层中 $k\text{-}th$ 维度的归一化值，$\gamma_k^{(i)}$ 和 $\beta_k^{(i)}$ 是待训练的模型参数。

基于卷积层和归一化层的特征提取，$i\text{-}th$ 分支中 $l+1$ 层的输出为 $\boldsymbol{Z}^{(i,l+1)} = [Z_1^{(i,l+1)}, Z_2^{(i,l+1)}, \ldots, Z_k^{(i,l+1)}, \ldots, Z_E^{(i,l+1)}]$。给定 m（$m=5$）个分支，$l+1$ 层中会产生新的特征 $\boldsymbol{Z}^{(1,l+1)}$，$\boldsymbol{Z}^{(2,l+1)}$、……、$\boldsymbol{Z}^{(m,l+1)}$。基于多个分支生成的新特征，我们采用拼接层融合这些特征，从而形成 $l+2$ 层的输入特征 $\boldsymbol{V}^{(l+2)} = [\boldsymbol{Z}^{(1,l+1)}, \boldsymbol{Z}^{(2,l+1)}, \ldots, \boldsymbol{Z}^{(m,l+1)}] = [V^{(l+2)\,1}，\ V_{(l+2)\,2}$、……、$V_{(l+2)n}\ ]$。全连接层与拼接层相连。此处采用全连接层的原因在于，它能够从多个分支中提取全局特征。给定全连接层（$l+3$ 层）的输入向量 $V^{(l+2)}$ 和ReLU激活函数 θ，$l+3$ 层中 $i\text{-}th$ 神经元的输出由以下公式计算：

$$V_i^{l+3} = \theta(\sum_k w_{k,i}^{l+2} V_k^{l+2}), \tag{6}$$

其中，$w_{k\,i}^{l+2}$，是 $l+2$ 层的 $k\text{-}th$ 神经元与 $l+3$ 层的 $i\text{-}th$ 神经元之间的连接权重。最后，通过采用softmax激活函数的输出层获得提取特征。共有 n（$n=2$）种类型，其中 $i\text{-}th$ 类型的预测概率由以下公式计算：

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \tag{7}$$

其中，z_j 表示输出层中 $j\text{-}th$ 神经元的输出值。

基于从globalSMIs数据中进行特征提取的过程，MBCNN模型确定了预测概率结果 $\boldsymbol{p}_G$，该结果可在阶段2中进一步用于提升预测性能。

3.3.2 行业数据的特征提取

在本研究中，行业指数数据是第二种有价值的数据源。根据以往研究（Su 2017年; He 2001），行业指数数据存在行业轮动效应，即股票市场中不同行业在一段时间内轮流上涨或下跌的模式。行业轮动效应可为目标SMI预测提供重要信息。针对行业指数数据中的行业轮动效应，我们采用SC-MBCNN模型提取行业轮动效应特征。

为有效提取行业指数数据的特征，我们为特征提取过程设计了两个模块：基于谱聚类算法的行业聚类划分（模块1）和行业轮动效应特征提取（模块2）。特征提取过程的步骤如下。首先，根据行业指数和目标SMI数据，将数据集划分为多个时间段，从而为每个时间段构建行业指数矩阵，即 $\boldsymbol{II}_{n\times x}$ 其中，n 表示行业指数的数量，l 表示所选时间段的长度。随后，我们采用谱聚类（SC）算法，根据每个时间窗口内的行业指数数据形成行业聚类。SC算法由Ng 等人（2002）提出，该算法将样本视为空间中的点，并通过边连接样本点。距离较远的样本点之间被赋予较低的边权重，而距离较近的样本点之间被赋予较高的边权重。随后，通过图划分方法将样本点划分为子图，每个子图对应一个簇。最终，不同子图内的边权重之和应尽可能高，而子图之间的边权重之和应尽可能低。由于SC算法具有对数据分布适应性强、计算量小等优点，因此已成功应用于金融时间序列数据（Li 和 Yang 2022）。

谱聚类的行业聚类划分过程描述如下。给定行业指数矩阵 $\boldsymbol{II} = \{F_1, F_2, \ldots, F_n\}$，其中 F_i 是 $i\text{-}th$ 行业的属性，即 $F_i = \{\text{Value}_{i,\ 1},\ \text{Value}_{i,\ 2},\ \ldots,\ \text{Value}_{i,\ l}\}$，其中 l 是时间段的长度。将行业指数特征连接起来，构成无向加权图 $G(\boldsymbol{II},\ A)$，其中 A_{ij} 表示 $i\text{-}th$ 行业指数特征与 $j\text{-}th$ 行业指数特征之间的相似度，该相似度通过以下公式计算：

$$A_{ij} = \exp(-\frac{d\left(F_i, F_j\right)}{2\sigma^2}), \tag{8}$$

其中 $exp()$是以自然常数为底的指数函数，σ 决定邻域的宽度，$d(F_i,\ F_j)$是两个行业指数特征之间的欧几里得距离，计算公式如下：

$$d(F_i, F_j) = \sqrt{\sum_{k=1}^{l}(\ \text{Value}\ _{i,k} -\ \text{Value}\ _{j,k})^2}. \tag{9}$$

根据公式（8）和（9），当尺度参数 σ 保持不变时，两个行业指数之间的距离越近，连接权重越大。基于所构建的无向加权图，采用图划分方法，使每个子图内的权重总和最大化，并使子图之间的权重总和最小化。SC算法的详细步骤如算法 1所示。

输入： 行业指数数据集 $\boldsymbol{F}=\{F_1,F_2,\ldots,F_n\}$；聚类数量$k$。**输出：** k 个行业聚类。1 根据行业指数的相似度构建邻接矩阵；2 计算拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$，其中 $\boldsymbol{D}=\text{diag}(d_1,\ d_2,\ \ldots\ldots,\ d_n)$，$d_i=\sum_{j=1}^{n}A_{ij},\ i=1,\ 2,\ \ldots\ldots,\ n$；3 计算归一化拉普拉斯矩阵 $\tilde{L}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{1/2}$；计算归一化拉普拉斯矩阵中前 k 个最小特征值对应的特征向量 $z_1,\ z_2,\ \ldots\ldots,\ z_k$；4 构建由 k 个维特征向量组成的矩阵 $\boldsymbol{Z}_{n\times k}$；5 使用$k$-均值算法将矩阵 $\boldsymbol{Z}_{n\times k}$ 聚为$C_1,\ C_2,\ \ldots\ldots,\ C_k$类。

算法 1： 用于行业指数矩阵的谱聚类算法

在 SC 算法中，聚类数量 k 是预先确定的。本研究采用肘部法确定聚类数量。基于 SC 算法，将行业指数矩阵 $II=\{F_1,F_2,\ldots,F_n\}$ 聚类为 k 个行业聚类$C_1,\ C_2,\ \ldots\ldots,\ C_i,\ \ldots\ldots,\ C_k$其中 $C_i=\{F_i^1,F_i^2,\ldots,F_i^{n_i}\}$，其中 n_i 是第 n_i-th 个簇的第 i-th 个行业指数。对于第 i-th 个簇，输入向量为 $\boldsymbol{x}_i=[x_i^1,x_i^2,\ldots,x_i^{n_i}]$，其中 n_i-th 表示第 i-th 个行业聚类的行业指数收益率。

基于形成的行业聚类，每个行业聚类代表一段时间内高度相关的行业指数。谨慎选择最具预测力的行业聚类，可以提供关键信息，显著提升预测性能。重要行业聚类的选择过程可视为一种动态行业轮动效应，我们会策略性地识别并优先考虑更有可能提供有价值见解的行业聚类。选择过程使模型能够捕捉并利用行业指数数据中固有的模式。因此，为了从相应的聚类中提取行业轮动效应的特征，我们采用多分支卷积核。特征提取过程的细节见公式（3）至公式（7）。基于上述步骤，预测概率结果 $\boldsymbol{p}_I$ 由SC-MBCNN模型获得。

3.3.3 财经新闻的特征提取

所提出的模型中的第三个数据源是财经新闻文本。在现实中，不同新闻提供商对目标 SMI 走势的权威性和预测能力各不相同。为了捕捉多个新闻提供商之间影响力的差异，我们采用 RNN-ER模型提取财经新闻文本中蕴含的复杂特征，从而揭示不同新闻提供商的影响和预测能力差异。特征提取过程的步骤如下所述。

假设有 m 个新闻提供商，对应于 m 个数据集，其中每个数据集都包含市场数据以及各新闻提供商的情绪指数。市场数据包括开盘价、收盘价、最高价、最低价、成交量和收益率。新闻提供商的情绪指数包括正面情绪指数、负面情绪指数和中性情绪

指数。基于准备好的数据，我们首先采用基于情感词典的方法来计算一篇财经新闻文本的情感极性（Pang 和 Lee（2008））。基于词典的情感分类方法包含以下三个步骤：1）进行文本预处理，主要包括分词、去除标点符号和停用词；2）利用词向量表示财经新闻，其中词向量中的每个值都是一个词的加权值，该值通过词频-逆文档频率（TF-IDF）方法计算得出；3）基于金融情感词典，将词向量转换为情感向量。随后，我们采用先前研究中的情感指数法（Gao 等人，2022a）来获取不同新闻提供商的情绪指数。因此，结合历史市场数据，$i\text{-}th$ 新闻提供商的输入为 $\boldsymbol{x}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,s}\}$。

基于输入 $\boldsymbol{x}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,s}\}$，我们构建双层循环神经网络模型作为基分类器。采用双层循环神经网络模型的原因是，新闻提供商的情绪指数和历史市场数据均为序列数据，而循环神经网络模型能够很好地处理这类数据。下面简要介绍从多个新闻提供商进行特征提取的过程。

给定时间步 t_1 $\boldsymbol{x}_i^{t_1} = \{x_{i,1}^{t_1}, x_{i,2}^{t_1}, \ldots, x_{i,s}^{t_1}\}$ 的输入，第一循环层和第二循环层中第$j\text{-}th$ 个神经元的输出如下式所示：

$$h_{i,j}^{t_1(1)} = \theta(\boldsymbol{U} \cdot \boldsymbol{x}_i^{t_1} + \boldsymbol{W}^{(1)} \cdot \boldsymbol{h}_i^{t_1-1,(1)} + b_{i,j}^{(1)}), \tag{10}$$

$$h_{i,j}^{t_{1,1}(2)} = \theta(\boldsymbol{V} \cdot \boldsymbol{h}_i^{t_1(1)} + \boldsymbol{W}^{(2)} \cdot \boldsymbol{h}_i^{t_1-1,(2)} + b_{i,j}^{(2)}), \tag{11}$$

其中，θ 为激活函数，$b_i^{(1)}{}_{,j}$ 和 $b_i^{(2)}{}_{,j}$ 分别为偏置，$\boldsymbol{U}$，$\boldsymbol{V}$，$\boldsymbol{W}^{(1)}$ $\boldsymbol{W}^{(2)}$ 分别为输入层与第一循环层、第一循环层与第二循环层、时间步 $t_1 - 1$ 的第一循环层与时间步 t_1 的第一循环层、时间步 $t_1 - 1$ 的第二循环层与时间步 t_1 的第二循环层之间的连接权重。随后，我们采用全连接层，根据 $i\text{-}th$新闻提供商，得到输出值 $\boldsymbol{r}_i$。

基于 m多个新闻提供商的预测结果，我们采用基于证据推理的规则来整合上述结果。在基于证据推理的集成规则中，每个新闻提供商都被视为一项证据。使用多项证据的目的是提高预测性能。我们采用基于证据推理的规则，是因为它不仅考虑每项证据的权重，也考虑每项证据的可靠性。在目标SMI预测任务中，考虑证据的可靠性至关重要，因为并非所有证据来源都同样可靠。此前的研究探讨了多个新闻提供商的优势及其计算过程（Gao 等人，2022a）。基于RNN-ER模型，我们获得了 $\boldsymbol{p}_M$ 财经新闻提供的预测概率结果。

3.4 阶段2：动态堆叠集成模型

利用全球SMI数据、行业指数数据和金融新闻文本数据，采用MBCNN模型、SC-MBCNN模型和RNN-ER模型生成预测概率，分别记为 $\boldsymbol{p}_G$、 $\boldsymbol{p}_I$、 $\boldsymbol{p}_M$。这三组预测概率共同构成提取特征 $\boldsymbol{x} = \{x_1, x_2, \ldots, x_6\}$，其中 x_1和 x_2 分别表示MBCNN模型的上涨和下跌预测概率， x_3和 x_4 分别表示SC-MBCNN模型的上涨和下跌预测概率，x_5 和 x_6 分别表示RNN-ER模型的上涨和下跌预测概率。

为增强预测能力，必须整合从异构分类器中获得的提取特征。常见的集成技术包括多数投票、相对多数投票、加权投票和堆叠方法等。其中，堆叠方法被认为特别适合整合异构分类器（Wolpert 1992年；Papouskova 和 Hajek 2019；Cui 等 2021）。在基于堆叠的集成框架中，基分类器在第一阶段生成。随后，元分类器用于学习这些基分类器产生的结果。

对于时间序列数据，提取特征与目标 SMI 走势之间的交互表现出动态特征，即提取特征与目标 SMI 走势之间存在明显的非线性模式。因此，为有效捕捉这些时间动态，利用多个元分类器进行特征整合是一种有效策略。为此，我们设计了一个动态堆叠集成模型来整合这些提取特征。具体而言，我们首先在集成模型中使用多个元分类器。随后，针对每个时间窗口动态选择最优元分类器。这些元分类器包括 LR、K 近邻、SVM 类模型（径向基函数和多项式函数）、极端随机树（ET）模型、随机森林（RF）模型和人工神经网络模型。在附录 B 中，我们全面说明了多个元分类器所采用的计算过程。下面，我们重点介绍最优元分类器的选择过程。

采用（Gevaand Zahavi 2014）中的滑动时间窗口划分方法，我们将数据集划分为十个时间窗口，分别记为 D_1， D_2，……， D_{10}。在每个时间窗口内，首先将数据集划分为训练集和验证集。随后，再将训练集划分为两个部分：第一部分用于从不同的数据源中提取特征，第二部分用于训练元分类器。需要注意的是，各元分类器的预测能力可能因时间窗口而异。考虑到这一特点，我们在每个时间窗口内动态选择预测性能最优的元分类器，并以准确率指标作为评价指标。

基于生成的元分类器集合 $M\text{-}C\text{-}L$ $=\{M\text{-}C_1$， $M\text{-}C_2$、……、 $M\text{-}C_7\}$，我们使用以下公式为每个时间窗口选择最优元分类器：

$$M\text{-}C_i = \arg\max(\text{Accuracy}(M\text{-}C_{i,j})), j = 1, 2, \ldots, 7. \tag{12}$$

随后，所选元分类器会整合从全球SMI数据、行业指数数据和财经新闻数据中生成的特征，以预测目标SMI走势。

3.5 分阶段优化

所提出的TDSE模型包括两个阶段：阶段1着重于通过整合投资者知识和合适的网络结构，从不同数据源中提取特征；阶段2采用动态堆叠集成模型进行特征融合和目标SMI走势预测，该模型结合了逻辑回归模型、K近邻模型、径向基函数支持向量机模型、多项式支持向量机模型、随机森林模型、极端随机树模型和人工神经网络模型。

由于两个阶段都涉及许多重要的超参数，因此优化这些超参数对于提升泛化能力至关重要。为此，我们采用优化算法对阶段1和阶段2中的重要超参数进行微调。通过优化过程，我们能够有效地确定使所提出的模型预测性能最大化的最优超参数设置。

在所提出模型的阶段1中，共有42个超参数需要优化。具体而言，MBCNN模型包含8个超参数，RNN-ER模型包含26个超参数，SC-MBCNN模型包含8个超参数。在阶段2中，共有11个超参数需要优化，包括逻辑回归模型的惩罚值、径向基函数支持向量机模型的惩罚值和伽马值、多项式支持向量机模型的惩罚值和多项式值、随机森林模型的决策树数量、人工神经网络模型第一层、第二层和第三层的神经元数量、K近邻模型的邻域样本数量，以及极端随机树模型的决策树数量。

同时优化全部 53 个超参数会导致解空间过大，降低搜索效率，并增加陷入局部最优的风险。为解决这一问题，我们采用分阶段优化方法来确定超参数。具体而言，我们首先优化阶段1中的重要超参数，包括微调超参数，以使阶段1的每个特征提取过程都达到最佳性能。确定阶段1的超参数后，我们继续优化阶段2中的重要超参数。所提出的分阶段优化方法可以提高搜索效率并降低陷入局部最优的风险。

在阶段1和阶段2的优化过程中，共有53个超参数需要优化。由于这项任务涉及优化大量参数，因此选择合适的优化算法至关重要。常见的优化算法包括遗传算法（GA）、粒子群优化（PSO）、灰狼优化器（GWO）、萤火虫算法（FA）和和声搜索（HS）算法。其中，遗传算法（GA）尤为值得关注，因为它稳健性强，且能有效搜索大型复杂空间。GA具有诸多优势：（1）通过基于种群的方法保持解的多样性；（2）能够同时探索解空间中的多个区域；（3）防止过早收敛

陷入局部最优。此外，其强大的全局搜索能力降低了停留在次优解上的风险。因此，我们采用遗传算法优化上述超参数。所提出的TDSE-GA模型的详细信息见算法2。

输入： 种群规模 n；最大代数 H；最大停滞代数；全球SMI数据 D_G；行业指数数据 D_I；财经新闻数据 D_M；元分类器 $M\text{-}C$；适应度函数 $fit()$。**输出：** 阶段1中的最优分类器：$MBCNN_{L,\ b}$, $SC\text{-}MBCNN_{L,\ b}$, $RNN\ ER_{L\ b}$ $2\ \ M\ C\ L_b$ -, ，阶段中的最优分类器：--。1 使用时间滑动窗口方法将数据集 $\boldsymbol{D}$ 划分为十个子数据集 ; 2 **对于** $j = \{1, 2, \ldots, 10\}$ **执行**3 并行优化阶段1中的特征提取模型：$\boldsymbol{p}_{G,\ j} = MBCNN(D_{G,\ j})$, $\boldsymbol{p}_{I,j} = SC\text{-}MBCNN(D_{I,j})$, $\boldsymbol{p}_{M,j} = RNN\text{-}ER(D_{M,j})$; 4 end5 $h \leftarrow 0$, $sh \leftarrow 0$; 6 在阶段 2 中初始化种群：$P^{(h)} \leftarrow \{M\text{-}C\text{-}L_1^{(h)}, M\text{-}C\text{-}L_2^{(h)}, \ldots, M\text{-}C\text{-}L_1^{(h)}, \ldots, M\text{-}C\text{-}L_n^{(h)}\}$; 7 for $i = \{1, 2, \ldots, n\}$ do 8 for $j = \{1, 2, \ldots, 10\}$ do 9 for $m = \{1, 2, \ldots, 7\}$ do 10 计算元分类器的准确率：$Accuracy_{j,m}(M\text{-}C_{j\ m}^{(h)}, (\boldsymbol{p}_{G,j}, \boldsymbol{p}_{I,j}, \boldsymbol{p}_{M,j}))$; 11 end 12 选择最优元分类器：$M\text{-}C\text{-}L_{i\ ,j}^{(h)} \leftarrow \arg\max(Accuracy(M\text{-}C_i^{(h)},_{j,m}))$ $m = 1$,2, ... ,7; 13 end 14 end 15 评估初始化种群中的每个个体：$fit_i^{(h)} \leftarrow fit(M\text{-}C\text{-}L_i^{(h)})$; 16 选择初始化种群中适应度值最高的个体：$M\text{-}C\text{-}L_b^{(h)}$;17 $M\text{-}C\text{-}L_b \leftarrow M\text{-}C\text{-}L_b^{(h)}$, $fit_b \leftarrow fit_b^{(h)}$; 18 while $h < H$ do 19 从种群 $P^{(h)}$中选择精英个体； 20 基于交叉操作和变异操作生成子代个体； 21 基于精英个体和子代个体生成新种群 $P^{(h+1)}$； 22 评估 $P^{(h+1)}$中的每个个体：$fit_i^{(h+1)} \leftarrow fit(M\text{-}C\text{-}L_i^{(h+1)})$; 23 选择 $P^{(h+1)}$中适应度值最高的个体：$M\text{-}C\text{-}L_b^{(h+1)}$; 24 if $fit_b^{(h+1)} > fit_b$ then 25 $M\text{-}C\text{-}L_b \leftarrow M\text{-}C\text{-}L_b^{(h+1)}$, $M\text{-}C\text{-}L_b \leftarrow M\text{-}C\text{-}L_b^{(h+1)}$, $sh \leftarrow 0$; 26 else27 $sh \leftarrow sh + 1$; 28 if $sh > SH$ then break; 29 end30 $h \leftarrow h + 1$; 31 end 算法2：TDSE-GA模型

给定全球SMI数据、行业指数数据和财经新闻数据，算法2求得最优解。

步骤1：并行　　优化阶段1中的特征提取模型　　在阶段1
首先，将数据集划分为十个子数据集（第1行）。随后，针对每个子数据集，并行优化MBCNN、SC-MBCNN和RNN-ER模型（第2—4行）。因此，在 $j\text{-}th$ 子数据集中提取的特征分别为$\boldsymbol{p}_{G、j}$、 $\boldsymbol{p}_{I、j}$和$\boldsymbol{p}_{M、j}$。

步骤2：优化阶段2中的超参数
在第5行中，将阶段2的迭代代数和停滞代数初始化为0。同时，初始化的种群为 $P^{(H)}$。随后，在每个时间段内计算每个个体的元分类器准确率，以帮助评估每个元分类器的性能（第10行）。为适应不同时间段的数据模式，我们根据准确率为每个特定时间段选择最优元分类器（第12行）。然后，按照以下公式计算每个个体的适应度值：

$$fit_i^{(h)} = \frac{1}{K}\sum_{k=1}^{K} Accuracy_{i,k}, \tag{13}$$

其中， $K = 9$ 表示我们使用前九个子数据集来判断某个个体是否为优选个体。选取适应度值最高的个体作为当前最佳个体（第 16 行），同时暂时将当前最佳个体选为最佳个体（第 17 行）。

第 18 至 32 行描述了遗传算法的操作，即选择、交叉和变异操作。首先，从当前种群中选出适应度值较高的精英个体。以这些精英个体为基础，通过交叉和变异操作生成子代个体。随后，将这些个体与子代个体合并，形成新的种群（第 21 行）。然后，根据公式（13）评估新种群中每个个体的适应度值。

该过程（第 24 至 31 行）根据适应度值判断是否用 $P_{(h+1)}$($M\text{-}C\text{-}L_b^{(h+1)}$) 中的最佳个体替换临时最佳个体。如果该个体的适应度值高于临时最佳个体的适应度值，则替换临时最佳个体，并将停滞代数重置为 0。否则，临时最佳个体保持不变，停滞代数增加 1。

迭代过程（第18—31行）会持续进行，直到满足以下任一终止条件：（1）TDSE-GA模型达到最大代数；（2）TDSE-GA模型达到最大停滞代数。当所提出的模型终止时，将输出阶段1和阶段2中的最优个体，即阶段1的最优分类器：$MBCNN_{L,\ b}$，$SC\text{-}MBCNN_{L,\ b}$,$RNN\text{-}ER_{L,\ b}$；阶段2的最优分类器： $M\text{-}C\text{-}L_b$。

4 实验研究

为验证所提出模型的有效性，我们将其用于预测中国股市中上海证券综合指数、深证成份指数和创业板指数的每日走势，并将其表现与其他方法进行比较。基于所提出的模型，我们设计了一种交易策略，并从累积收益率和夏普比率方面将其与其他交易策略进行比较。实验结果如下。

4.1 数据描述

目标SMI包括上海证券综合指数（SSEC）、深圳证券市场综合指数（SZEC）和创业板指数（GEI）。这些目标SMI的数据来自中国股票市场与会计研究数据库（CSMAR）。我们从 investing.com 网站获取全球SMI数据，其中包括来自亚洲、欧洲和美洲的30个主要SMI。行业指数数据来自万得数据库，包括97个行业指数。我们从 Tushare 平台获取财经新闻文本数据，该平台是中国最大的金融大数据开放社区之一。我们从五个财经新闻来源收集了共751,186条新闻文本：新浪财经、华尔街见闻、同花顺、东方财富和云财经。收集的数据时间跨度为2018年10月11日至2021年4月28日。所收集财经新闻文本的概况见表2。

表 2 收集的财经新闻文本简要信息

No.	新闻提供商	新闻文本数量	百分比（%）
1	新浪财经	255,225	33.98
2	华尔街见闻	67,962	9.05
3	同花顺	117,206	15.60
4	东方财富	140,462	18.70
5	云财经	170,331	22.67
	总计	751,186	100.00

对于从各种金融新闻来源收集的文本数据，我们会去除每篇财经新闻中的标点符号、数字和停用词，并对剩余词语进行分词。预处理完成后，我们采用（Gao 等人，2022a）中的情绪指数计算方法，确定每个新闻提供商的情绪指数。在计算过程中，我们使用了中文情感词典（Jiang 等人，2021），该词典包含5,890个积极词语和3,338个消极词语。不同金融新闻提供商的情绪指数和新闻数量统计结果见附录中的表 C.3。

4.2 评估指标

我们采用（Geva and Zahavi2014）提出的时间滑动窗口方法，将数据集划分为训练集和测试集。具体而言，我们将滑动窗口数量设为10。鉴于数据集涵盖31个月，在每个时间窗口中，前11个月的数据用作训练集，随后3个月的数据用作测试集。我们的目标是预测目标 SMI 走势，这属于分类任务。因此，我们采用分类任务中常用的评估指标，包括准确率、精确率、召回率、F-度量值和AUC值。分类结果汇总于表3所示的混淆矩阵中。

表3 分类任务的混淆矩阵

实际结果	预测结果	
	Up	Down
Up	TU	FD
Down	FU	TD

在表3中，TU（真上涨）表示预测结果和实际结果均为“上涨”。TD（真下跌）表示预测结果和实际结果均为“下跌”。如果预测结果为“下跌”，而实际结果为“上涨”，则标记为FD（假下跌）。如果预测结果为“上涨”，而实际结果为“下跌”，则标记为FU（假上涨）。准确率、精确率、召回率和F度量指标按以下公式计算：

$$Accuracy = \frac{TU + TD}{TU + TD + FU + FD}, \tag{14}$$

$$Precision = \frac{TU}{TU + FU}, \tag{15}$$

$$Recall = \frac{TU}{TU + FD}, \tag{16}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{17}$$

曲线下面积（AUC）是ROC曲线下的面积，其中假阳性率（FPR）为横轴，真阳性率（TPR）为纵轴。FPR和TPR通过以下公式计算：

$$FPR = \frac{FU}{FU + TD}, \tag{18}$$

$$TPR = \frac{TU}{TU + FD}. \tag{19}$$

AUC值衡量分类器在所有可能分类阈值下的整体性能，主要用于评估分类器的泛化能力。AUC值越高，分类性能越好；数值越接近1，表明分类器区分正样本和负样本的能力越强。

4.3 实验设置

为验证所提出模型的有效性，本研究设计了四项实验。第一项实验旨在比较不同数据源对目标SMI走势的预测能力。在这项实验中，我们采用以下模型作为基线：人工神经网络（ANN）、支持向量机（SVM）、随机森林（RF）、卷积神经网络（CNN）、循环神经网络（RNN）、长短期记忆网络（LSTM）、门控循环单元（GRU）和二维卷积神经网络（2D-CNN）。上述模型的输入包括市场数据、全球SMI数据、行业指数数据和财经新闻文本。

第二项实验旨在验证所提出的TDSE-GA模型的有效性，该模型考虑了不同数据源的金融数据模式。对比模型包括随机预测模型、标签t-1模型、MBCNN-GA模型（Gao等，2022b）、SC-MBCNN-GA模型、RNN-ER-GA模型（Gao 等人，2022a）、ANN+PCA模型（Ma等，2022）、MKL模型（Shynkevich等，2016）、FCN模型（Wang等，2017）和Transformer模型（Vaswani等，2017）。对于MBCNN-GA、SC-MBCNN-GA和RNN-ER-GA模型，其输入分别为全球SMI数据、行业指数数据和财经新闻文本。对于MKL、ANN+PCA、FCN、Transformer和所提出的TDSE-GA模型，输入包括全球SMI数据、行业指数数据和财经新闻文本。全球SMI数据和财经新闻文本的滞后期设为1，行业指数数据的滞后期设为5。该设置与先前的研究一致（Gao 等人，2022a，b）。

在第三项实验中，我们基于不同的交易策略评估经济结果。为考察所提出的TDSE-GA 模型能否带来超额收益，我们设计了一种基于该模型的交易策略，并将其与另外两类交易策略进行比较：简单交易策略，以及基于相关研究中所用模型的交易策略，即 MKL、ANN+PCA、FCN 和 Transformer 模型。

第四项实验聚焦于元分类器的动态选择过程。由于所提出的模型通过在不同时间段逐步选择最优分类器来提升预测性能，因此本实验旨在考察模型的稳健性以及元分类器的动态选择过程。

4.4 实验结果

4.4.1 不同数据源的预测结果

在本实验中，我们比较不同数据源的预测结果，具体包括市场数据、全球SMI数据、行业指数数据和财经新闻文本。

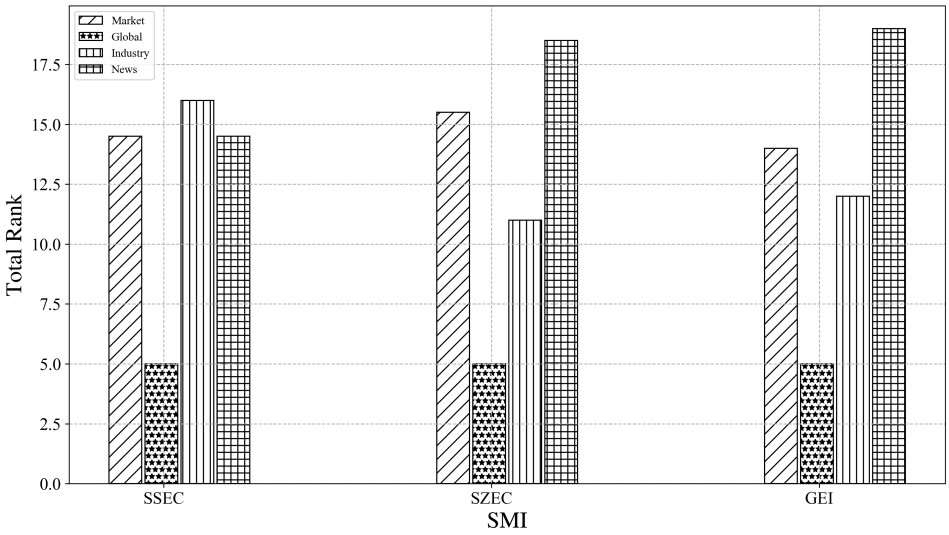

图 2 上证综合指数、SZEC 和 GEI 数据集上不同数据源的总排名结果

本实验旨在探究在历史市场数据的基础上加入其他数据源，是否能够提升预测模型的预测能力。为全面展示不同数据源在 SSEC、SZEC 和 GEI 数据集上的预测能力，我们引入了总排名指标。该指标定义为各数据源“排名”的平均值。“排名”指标表示：针对每项评价指标（准确率、召回率、精确率、*F*-度量值和曲线下面积），根据不同数据源在各种机器学习方法上的平均预测表现进行排序。总排名值是预测性能的关键指标，数值越低，表示该数据源的表现越好。图 2展示了不同数据源在 SSEC、SZEC 和 GEI 数据集上的总排名结果。详细的计算过程和结果见附录 C.2 节。

从图 2所示结果中，我们首先观察到，将全球SMI数据与历史市场数据结合，显著优于其他数据源，在 SSEC、SZEC 和 GEI 数据集上的总排名值均为最佳（5）。此外，我们发现，在 SZEC 和 GEI 数据集中加入行业指数数据后，预测性能有所提升。相反，在 SZEC 和 GEI 数据集中加入新闻数据会对预测性能产生不利影响。最后，我们发现，在 SSEC 数据集中，无论加入行业指数数据还是新闻数据，都未能明显提升预测性能。

4.4.2 不同预测模型的预测结果比较

第 4.4.1 节所示的实验结果揭示了全球股票市场指数数据、行业指数数据和财经新闻文本各自不同的预测能力。具体而言，将历史市场数据与全球股票市场指数数据结合，能够明显提升预测性能。相比之下，加入行业指数数据或财经新闻数据并不能显著提升预测能力。因此，为了充分发挥上述数据源的预测潜力并提升预测性能，我们提出了 TDSE-GA 模型，专门用于整合多模态数据，并考量不同数据源中的金融数据模式。为验证所提出的模型的有效性，我们将其与其他竞争方法进行了比较。所提出的模型及竞争方法的详情见表 4。为与相关研究（Gao 等人，2022a）中的参数设置保持一致，我们确定了所提出模型的超参数，详见附录中的表 C.3。

表 4 所提出的模型和竞争方法的描述

Type	预测模型	描述
简单模型	随机预测	预测值由随机选择确定
	标签 t-1	预测值由以下标签确定：周期 t-1
基于单一数据源的预测模型	MBCNN-GA	模型输入为全球SMI数据
	SC-MBCNN-GA	模型输入为行业指数数据
	RNN-ER-GA	模型输入为财经新闻文本
多源数据融合模型	MKL	模型输入包括全球SMI数据，行业指数数据和财经新闻文本
	人工神经网络+主成分分析	
	FCN	
	Transformer	
所提出的模型	TDSE-GA	模型输入包括全球SMI数据、行业指数数据和财经新闻文本

表5展示了所提出的模型和竞争方法在上证综合指数、SZEC 和 GEI 数据集上的预测结果。结果取 10 个时间窗口的平均值，以进行全面评估。在表5中，每列以粗体标出的数值代表特定 SMI 和评估指标的最高值。改进率是指所提出的 TDSE-GA 模型相对于对比模型最佳结果的提升幅度。根据表8中的预测结果，我们可以得出以下主要观察结论。(1) 在准确率（上证综合指数、SZEC 和 GEI）、召回率（上证综合指数）、精确率（上证综合指数、SZEC 和 GEI）、F度量（上证综合指数、SZEC 和 GEI）以及曲线下面积（上证综合指数、SZEC 和 GEI）方面，所提出的 TDSE-GA 模型优于其他竞争方法。在准确率这一股票市场指数预测任务中最重要的指标方面，所提出的 TDSE-GA 模型在上证综合指数、 SZEC 和 GEI 指数上的表现分别比最佳对比模型提升了 11.42%、7.94% 和 7.73%。(2) 标签

t-1模型在上证综合指数、SZEC 和 GEI 数据集上的预测结果优于随机预测方法。(3) 在采用不同数据源的模型中，MBCNN-GA 模型的准确率优于 SC-MBCNN-GA 模型和 RNN-ER-GA 模型。(4) 对于融合多模态数据的模型，Transformer 模型在上证综合指数和 GEI 数据集上的表现更好，而 ANN+PCA 模型在 SZEC 数据集上的表现优于其他模型。(5) 考虑单个数据源金融模式的模型，通常比以简单直接的方式整合多源数据的模型能提供更好的预测结果。例如，MBCNN-GA 模型在上证综合指数、SZEC 和 GEI 数据集上的预测结果均优于 Transformer 模型。

为了检验所提出的模型与竞争方法的预测结果之间差异的显著性水平，我们进行了配对t检验。表t 列出了所提出的模型与竞争方法在 SSEC、SZEC 和 GEI 数据集上的配对检验结果。根据表6所示结果，我们发现，在 SSEC 数据集上，TDSE-GA模型与其他竞争方法在准确率、召回率、精确率、F值和曲线下面积指标方面存在显著差异（与Transformer模型比较时的精确率指标除外）。准确率和曲线下面积指标的显著性水平低至 0.05，而F值指标的显著性水平为 0.01。这些发现表明，所提出的 TDSE-GA模型通过考虑多模态数据的金融数据模式，有效提升了目标 SMI走势的预测性能。

4.4.3 基于不同交易策略的经济结果

为评估我们提出的 TDSE-GA模型的经济价值，我们根据其预测信号设计了一种交易策略。随后，我们将这一策略与其他方法进行比较：朴素交易策略，以及基于相关研究预测信号的交易策略。在本实验中，我们不考虑交易成本和税费。以下介绍这三种类型交易策略的具体内容。

(1) 基于 TDSE-GA模型预测信号的做多交易策略（TDSE-GA）：如果 TDSE-GA模型预测价格将上涨，交易系统便买入或持有目标投资产品；如果预测信号显示价格将下跌，交易系统则卖出目标投资产品的所有持仓并持有现金。

(2) 朴素交易策略包括买入并持有交易策略和基于随机预测的做多交易策略。买入并持有交易策略是在第一个交易日买入目标投资产品，并持有至最后一个交易日。基于随机预测的做多交易策略遵循与 TDSE-GA交易策略相同的原则，但使用随机预测生成信号。

（3）基于相关研究中模型的预测信号的做多交易策略。该交易策略的原则与 TDSE-GA交易策略相同，但用于生成预测信号的模型不同。预测信号由相关研究中的模型生成，包括 MKL模型（Shynkevich等，2016）、

表5 不同方法的预测结果

方法类型	模型	准确率	召回率	精确率	F值	AUC
			SSEC			
简单模型	随机预测	0.4657	0.4890	0.4488	0.4619	0.4687
	标签 *t*-1	0.5255	0.4905	0.4919	0.4911	0.5123
单一数据源模型	MBCNN-GA	0.5790	0.6112	0.5537	0.5808	0.5800
	SC-MBCNN-GA	0.5573	0.6087	0.5367	0.5732	0.5565
	RNN-ER-GA	0.5561	0.5058	0.5649	0.4929	0.5468
多源数据融合模型	人工神经网络+主成分分析	0.5112	0.5438	0.5029	0.5171	0.5021
	MKL	0.4885	0.4369	0.2852	0.3382	0.4220
	FCN	0.5134	0.5844	0.5054	0.5343	0.5204
	Transformer	0.5511	0.5659	0.5549	0.5439	0.5558
	TDSE-GA	**0.6451**	**0.7236**	**0.6442**	**0.6521**	**0.6453**
	改进率(%)	11.42	18.39	14.04	12.28	11.26
			SZEC			
简单模型	随机预测	0.4587	0.4915	0.4836	0.4781	0.4619
	标签 *t*-1	0.5399	0.5484	0.5456	0.5468	0.5282
单一数据源模型	MBCNN-GA	0.5767	0.6268	0.5868	0.6068	0.5776
	SC-MBCNN-GA	0.5703	0.6104	0.5800	0.5865	0.5668
	RNN-ER-GA	0.5696	**0.8172**	0.5614	0.6194	0.5398
多源数据融合模型	人工神经网络+主成分分析	0.5415	0.5519	0.5682	0.5518	0.5536
	MKL	0.5145	0.4411	0.3193	0.3621	0.4584
	FCN	0.5219	0.5786	0.5414	0.5506	0.5209
	Transformer	0.5223	0.5401	0.5551	0.5345	0.5383
	TDSE-GA	**0.6225**	0.7708	**0.6312**	**0.6624**	**0.6192**
	改进率(%)	7.94	-5.68	7.57	6.94	7.20
			GEI			
简单模型	随机预测	0.4618	0.4986	0.5217	0.4992	0.4668
	标签 *t*-1	0.5035	0.5373	0.5373	0.5372	0.4888
单一数据源模型	MBCNN-GA	0.5809	0.7035	0.6088	0.6418	0.5718
	SC-MBCNN-GA	0.5706	0.6436	0.6170	0.6162	0.5745
	RNN-ER-GA	0.5807	**0.7533**	0.5781	0.6122	0.5374
多源数据融合模型	人工神经网络+主成分分析	0.5546	0.5746	0.6078	0.5847	0.5259
	MKL	0.5048	0.6525	0.4381	0.5154	0.4124
	FCN	0.5512	0.5817	0.6029	0.5851	0.5343
	Transformer	0.5579	0.5987	0.6114	0.5938	0.5958
	TDSE-GA	**0.6258**	0.7419	**0.6823**	**0.6663**	**0.6188**
	改进率（%）	7.73	-1.51	10.58	3.82	3.86

人工神经网络+主成分分析模型（Zhongand Enke 2017）、FCN模型（Wang等，2017）和Transformer模型（Vaswani等，2017）。

这些交易策略的时间段为2019年9月1日至2021年4月30日。不同策略在上证综合指数、SZEC 和 GEI 数据集上的累计收益曲线分别如图3、图4和图5所示。根据分析结果，我们得出以下观察结论。首先，其他交易策略的累计收益曲线都围绕买入并持有策略的曲线波动。这一现象是因为收益来源于正确的

表6 所提出的模型与竞争方法之间的配对 t 检验

方法类型	模型	准确率	召回率	精确率	F值	AUC
			SSEC			
简单模型	随机预测	0.001***	0.003***	0.003***	0.001***	0.001***
	标签 t-1	0.010***	0.010***	0.002***	0.006***	0.004***
单一数据源模型	MBCNN-GA	0.032**	0.032**	0.028**	0.004***	0.009***
	SC-MBCNN-GA	0.000***	0.024**	0.001***	0.002***	0.000***
	RNN-ER-GA	0.004***	0.046**	0.042**	0.006***	0.000***
多源数据融合模型	人工神经网络+主成分分析	0.001***	0.007***	0.004***	0.000***	0.012**
	MKL	0.004***	0.033**	0.001***	0.004***	0.001***
	FCN	0.001***	0.058*	0.004***	0.004***	0.009***
	Transformer	0.043**	0.008***	0.122	0.006***	0.049**
			SZEC			
简单模型	随机预测	0.000***	0.002***	0.000***	0.001***	0.000***
	标签 t-1	0.005***	0.000***	0.010***	0.001***	0.003***
单一数据源模型	MBCNN-GA	0.043**	0.001***	0.079*	0.003***	0.049**
	SC-MBCNN-GA	0.029**	0.007***	0.101	0.015**	0.050**
	RNN-ER-GA	0.106	0.613	0.082*	0.348	0.006***
多源数据融合模型	人工神经网络+主成分分析	0.064*	0.002***	0.126	0.017**	0.161
	MKL	0.002***	0.023**	0.002***	0.006***	0.001***
	FCN	0.006***	0.010***	0.027**	0.009***	0.003***
	Transformer	0.000***	0.001***	0.026**	0.001***	0.008***
			GEI			
简单模型	随机预测	0.001***	0.009***	0.003***	0.001***	0.003***
	标签 t-1	0.000***	0.005***	0.001***	0.001***	0.000***
单一数据源模型	MBCNN-GA	0.099*	0.553	0.028**	0.420	0.109
	SC-MBCNN-GA	0.073*	0.101	0.046**	0.080*	0.115
	RNN-ER-GA	0.209	0.882	0.006***	0.366	0.013**
多源数据融合模型	人工神经网络+主成分分析	0.047**	0.023**	0.017**	0.055*	0.031**
	MKL	0.002***	0.555	0.010***	0.141	0.000***
	FCN	0.111	0.037**	0.107	0.057*	0.071*
	Transformer	0.115	0.054*	0.109	0.056*	0.610

* 在 0.1 显著性水平上具有统计显著性。** 在 0.05 显著性水平上具有统计显著性。***在 0.01 显著性水平上具有统计显著性。

预测信号，而该信号与目标SMI直接相关。因此，买入并持有交易策略恰当地反映了目标SMI的收益表现。其次，基于所提出的模型的交易策略在上证综合指数、SZEC 和 GEI 数据集上始终比其他交易策略产生更高的累计收益。这一结果凸显了所提出的模型在提升预测性能方面具有显著的经济价值。第三，在相互竞争的交易策略中，基于随机预测的策略产生的累计收益最低。相比之下，基于 Transformer模型的交易策略以及

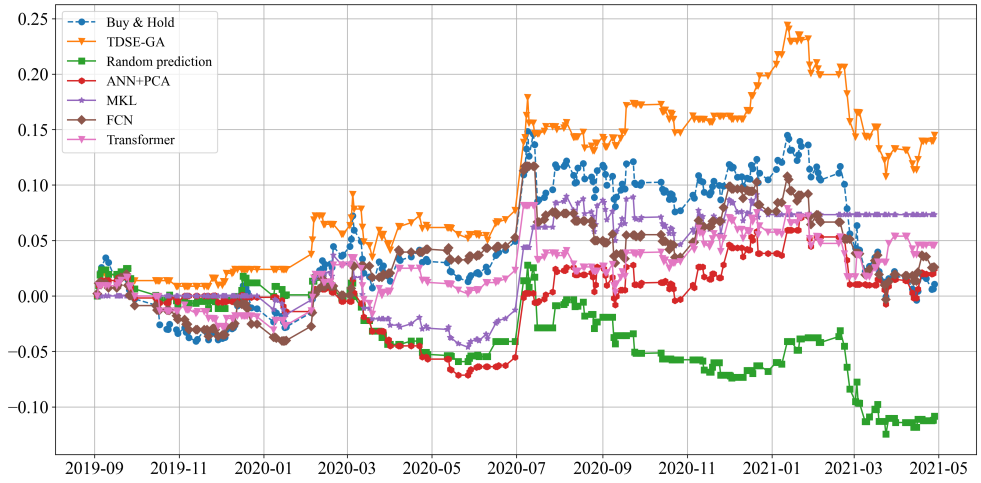

图 3 上证综合指数上不同策略的累计收益曲线

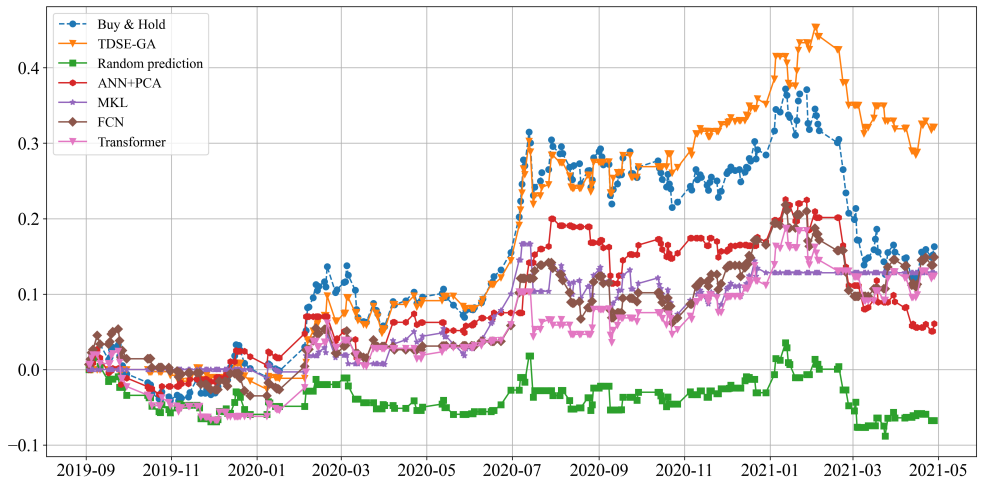

图 4 SZEC 上不同策略的累计收益曲线

买入并持有策略表现优于其他策略，在上证综合指数、SZEC 和 GEI 数据集上取得了更高的累计收益。

为全面评估各种交易策略的表现，我们计算了每种策略的累计收益、月收益率、夏普比率、日收益率和最大回撤率。表7 展示了不同交易策略在 SSEC、SZEC 和 GEI 数据集上的比较结果，其中每列的粗体数字表示对应 SMI 的最佳结果。改进率是指

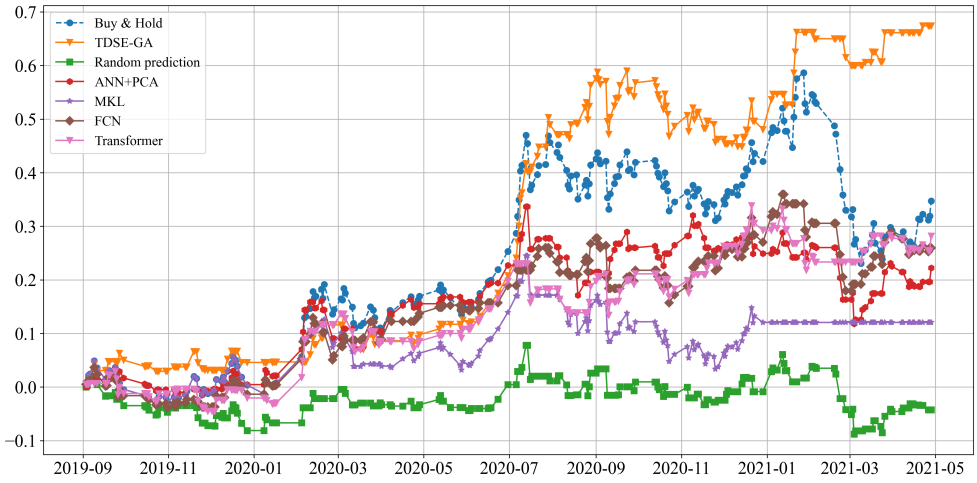

图 5 GEI上不同策略的累计收益曲线

基于所提出的模型的交易策略，在特定评价指标下相较于表现最佳的竞争策略所取得的性能提升。表7 有一个有趣的现象：基于人工神经网络+主成分分析模型、FCN模型和Transformer模型的交易策略，其月收益率与日收益率的变化方向并不一致。这种差异源于计算方法不同：月收益率考量的是一个月内首个与最后一个交易日之间的差异，而日收益率则表示累计收益的平均值。

根据表 7中的结果，我们首先发现，基于TDSE-GA模型的交易策略在SSEC（14.49%）、SZEC（32.13%）和GEI（67.40%）数据集上取得了最高的累计收益率。这些数值分别是买入并持有交易策略的13.6倍、1.97倍和1.94倍。就累计收益率指标而言，基于所提出的模型的交易策略相比表现最佳的竞争交易策略，改进幅度超过90%。具体而言，SSEC、SZEC和GEI的改进率分别为97.68%、97.25%和93.99%。其次，与其他竞争交易策略相比，基于TDSE-GA模型的交易策略在SSEC（0.4059）、SZEC（0.5543）和GEI（0.6507）数据集上均取得了更高的夏普比率。这表明，所提出的模型能够在承担单位风险的同时实现更高收益。最后，就最大回撤率而言，我们观察到不同股票市场指数对应的最优交易策略各不相同。具体来说，在SSEC数据集上表现最佳的是基于Transformer模型的交易策略（7.26%），在SZEC数据集上表现最佳的是基于MKL模型的交易策略（8.52%），而在GEI数据集上表现最佳的是基于TDSE-GA模型的交易策略（8.90%）。

本研究采用遗传算法作为优化方法。不过，也可以使用其他优化技术，例如粒子群优化和萤火虫算法。为验证所选算法的有效性，我们将遗传算法与其他优化方法进行比较。比较标准包括各模型的预测性能和计算时间。详细结果见附录 C.4 节。

表7 上证综合指数、SZEC 和 GEI 上不同交易策略的结果比较

SMI	交易策略	累计 收益率	月 收益率	夏普 比率	日 收益率	最大 回撤 rate
SSEC	买入并持有	0.01063	0.00390	0.17326	0.00004	0.13246
	随机预测	-0.10826	-0.00599	-0.34054	-0.00044	0.14822
	人工神经网络+主成分分析	0.02447	-0.00261	-0.18592	0.00010	0.08644
	MKL	0.07331	0.00237	0.10780	0.00030	0.07994
	FCN	0.02601	-0.00076	-0.03984	0.00011	0.10765
	Transformer	0.04584	-0.00299	-0.14398	0.00019	**0.07263**
	TDSE-GA	**0.14492**	**0.00489**	**0.40594**	**0.00059**	0.10970
	改进率(%)	+97.68	+25.38	+134.30	+96.67	-51.04
SZEC	买入并持有	0.16288	0.01573	0.38773	0.00066	0.18378
	随机预测	-0.06758	-0.00348	-0.15676	-0.00027	0.11936
	人工神经网络+主成分分析	0.06075	0.01275	0.36238	0.00025	0.14239
	MKL	0.12826	0.00802	0.34007	0.00052	**0.08520**
	FCN	0.14902	0.00267	0.17118	0.00060	0.09965
	Transformer	0.12400	0.00011	0.00425	0.00050	0.08917
	TDSE-GA	**0.32128**	**0.01850**	**0.55426**	**0.00130**	0.11661
	改进率(%)	+97.25	+17.61	+42.95	+96.97	-36.87
GEI	买入并持有	0.34745	0.02486	0.47677	0.00141	0.22434
	随机预测	-0.04239	0.00038	0.01382	-0.00017	0.15402
	人工神经网络+主成分分析	0.22210	0.01435	0.55245	0.00090	0.16357
	MKL	0.12049	0.00572	0.16746	0.00049	0.16998
	FCN	0.25968	0.01387	0.58775	0.00105	0.13228
	Transformer	0.28170	0.00569	0.13305	0.00114	0.09032
	TDSE-GA	**0.67401**	**0.03820**	**0.65072**	**0.00273**	**0.08900**
	改进率(%)	+93.99	+53.66	+10.71	+93.62	+1.46

4.4.4 元分类器的动态选择过程

所提出的 TDSE-GA模型通过逐步确定不同时间段的最优分类器，提升了预测性能。因此，需要围绕 TDSE-GA模型探讨两个关键方面：鲁棒性和动态选择过程。首先，需要考察模型的预测结果是否逐步收敛。其次，需要确定哪些分类器在不同时间段对模型的预测性能有显著影响。为研究 TDSE-GA模型的收敛过程以及元分类器的动态选择过程，我们在 SSEC数据集上开展了以下实验。首先，提取每一代最优个体的适应度函数值及所选元分类器。随后，使用折线图和热力图呈现这些数据，如图 6 所示。

在图6中，元分类器采用缩写表示，对应以下模型：L（逻辑回归模型）、K（K近邻模型）、R（径向基函数支持向量机模型）、P（多项式支持向量机模型）、RF（随机森林模型）、E（极端随机树模型）和A（人工神经网络模型）。根据图10中的结果，我们发现以下几点。首先，在子图(a)中，我们注意到，随着迭代代数的增加，适应度函数值持续上升，最终达到稳定并收敛的状态。这表明优化过程有效地收敛到了令人满意的解。其次，从子图(b)到子图(k)，我们观察到不同时间窗口对应的最优元分类器各不相同，这为动态选择设计的成功实施提供了证据。该自适应方法确保为每个特定时间段选择最合适的元分类器，从而提升

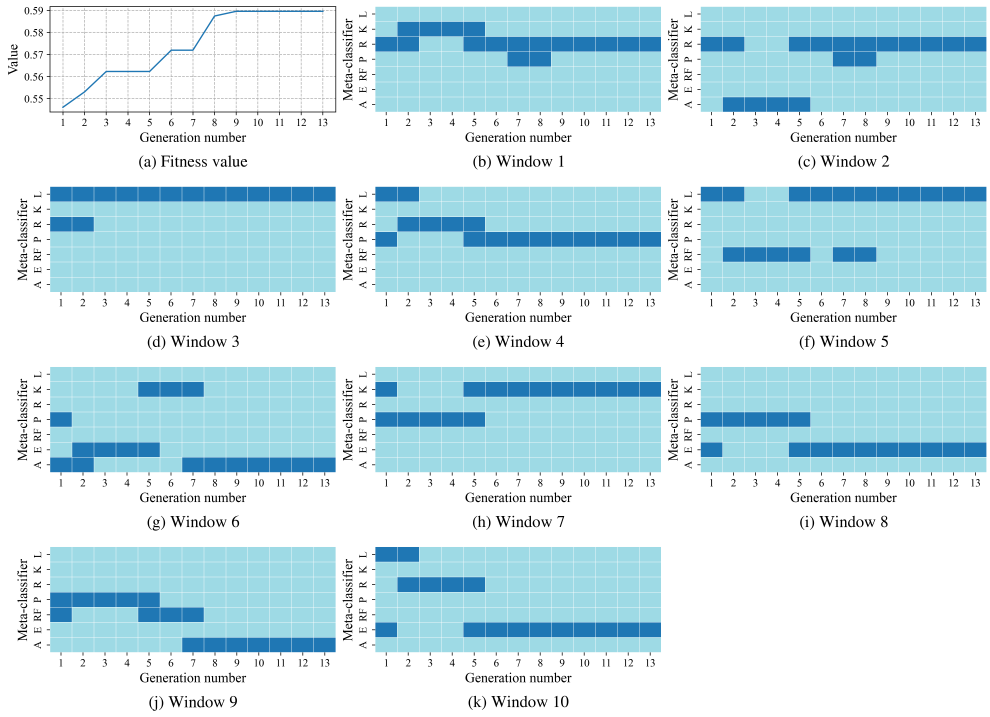

图 6 不同代次中最优个体的适应度及所选元分类器

预测性能。最后，我们发现逻辑回归模型、径向基函数支持向量机模型、极端随机树模型和人工神经网络模型最常被选为元分类器（各被选中两次）。这一现象表明，它们能够有效整合来自多个数据源的数据。动态选择验证了这些模型有助于提升TDSE-GA模型整体预测能力的有效性。

5 讨论

5.1 理论意义

本文的理论意义包括：深化对多模态数据在金融预测中的应用的理解，推动预测建模理论的发展，并为行为金融理论带来启示。在多模态数据建模方面，先前研究强调了信息融合和多模态学习在提升预测性能方面的优势（Thakkar 和 Chaudhari, 2021；Bao 等，2025）。在此基础上，我们考虑了不同的数据源，即区域内股市指数之间的趋势相似性、行业指数数据中的行业轮动效应，以及不同新闻提供商带来的影响差异。与现有的多模态数据整合研究相比（Weng 等，2018；Ma等，2022），我们的研究基于投资者知识，纳入了多模态数据的不同模式。考虑到

这些特征，我们采用 MBCNN、SC-MBCNN 和 RNN-ER 模型，从这些数据源中提取内在特征。通过利用每种数据类型的特定属性，我们的模型不仅提高了预测准确率，还能更深入地洞察市场动态，从而有助于在金融预测中做出更明智的决策。

在预测建模理论方面，我们提出了一种基于深度学习技术的新型两阶段动态堆叠集成模型。该模型采用多个元分类器，动态整合从各数据模态中提取的特征。通过考虑提取特征与目标 SMI 走势之间的时间动态，我们的模型能够针对每个时间段自适应地选择最优元分类器。这一自适应框架充分发挥不同分类器的优势，利用各时间窗口中识别出的特定模式来提升预测性能。与以往研究（Weng 等人 2018年；Wang 等人 2018年）相比，该方法不仅展示了动态集成方法在金融预测中的实用价值，也为探讨集成模型捕捉复杂市场动态的效能贡献了理论见解。此外，我们采用分阶段优化方法来增强所提出模型的泛化能力。首先，我们优化基分类器，确保其能够有效捕捉和处理特定模态的特征。随后，我们优化集成组件——元分类器，使其能够有效整合来自不同数据源的提取特征。这种分阶段优化方法显著缩小了解空间的复杂度，降低了过拟合风险，并提升了模型整体的泛化能力。我们为复杂预测模型的优化提供了宝贵的理论见解，这些见解可广泛应用于机器学习和数据科学的各个领域。

关于对行为金融理论的启示，本研究深入探讨了投资者经验和知识如何影响市场走势，为行为金融理论提供了重要关联。通过将这些行为因素系统地整合到预测模型中，我们不仅提高了模型的准确率，也加深了对驱动金融市场的心理和认知因素的理解。这种方法提供了一套结构化方案，用于量化投资者情绪、对新闻的反应以及不确定性下的决策模式等行为指标。此外，我们的研究表明这些行为因素在市场动态中发挥着重要作用，或可启发新的理论发展。

5.2 实践意义

本研究的实践意义包括：利用多种数据源提升市场分析工具的性能，从而实现更准确的预测；通过先进的多模态特征提取提高预测表现；以及优化交易策略，以提升经济价值。

首先，我们发现不同数据源的预测能力各不相同，其中全球SMI数据的表现优于行业指数

数据和财经新闻文本。这一重要发现为投资者带来了战略优势，使其能够识别最可靠的市场预测数据源。因此，这有助于开发更精细的市场分析工具，使其能够根据数据源经过实证验证的预测能力进行优先排序。这些工具可以改进投资策略框架中的决策流程，促进更明智的财务决策。

其次，我们提出了一种先进模型，旨在从多模态金融数据中提取内在特征，从而提升预测性能。这一增强能力为专业投资者带来了关键优势，使他们能够改进分析模型。整合该模型有助于他们实现更准确的预测，并改善风险管理实践。因此，在波动性市场环境中，这将促使决策更加具有战略性，并有望提升财务稳定性和盈利能力。

最后，基于所提出的模型预测结果制定的交易策略优于其他策略，展现了其理论稳健性和实际经济价值。投资组合经理和算法交易者可以采用该模型，并将其整合到交易系统中。通过实施该模型，交易者有望获得显著更高的收益率，并实现更好的投资组合多元化。

6 结论与未来工作

本研究针对基于多源金融数据中重要因素的SMI走势预测展开研究。每种金融数据源都呈现出不同的模式，这些模式从原始金融数据中可能并不明显，但经验丰富的投资者能够发现它们。识别并纳入这些数据模式对于提升预测性能至关重要。因此，我们提出了TDSE模型，该模型能够有效提取多源金融数据中的特征，并提升预测性能。

基于实验结果，我们得出以下结论。首先，不同数据源展现出不同的预测能力，其中全球SMI数据的表现优于行业指数数据和财经新闻文本。其次，所提出的模型具备出色的多模态金融数据特征提取能力，其预测性能优于竞争方法。第三，在PSO、GWO、FA和HS等常用优化算法中，遗传算法在预测性能和计算效率方面均表现出更好的结果。此外，基于所提出模型的交易策略优于其他交易策略，进一步证实了我们方法的经济价值。最后，所提出的模型能够快速且稳定地收敛，同时有效识别最优元分类器，证明了其稳健性和有效性。通过考虑每个金融数据源的独特模式，所提出的模型为提升目标SMI走势的预测性能提供了有前景的解决方案。

尽管本研究取得了改进，但也必须承认其局限性。以下方面有待进一步关注。（1）本研究主要聚焦于日频，忽略了分钟频率、小时频率或周频等其他频率。为

了深入理解多尺度数据在预测目标SMI走势中的作用，我们将探究不同交易频率的重要性。（2）本研究以SMI层面为中心。然而，探索其他层面也具有潜在价值，包括个股和板块。通过拓展研究范围，我们可以全面了解预测性能。在这些研究的基础上，我们可以提出一个投资组合构建研究框架，综合考虑个股、板块和SMI层面的预测结果。

缩略语。
二维卷积神经网络 二维卷积神经网络 曲线下面积 ROC曲线下面积 人工神经网络 人工神经网络 双向编码器表示（来自变换器） 双向编码器表示（来自变换器） 双向长短期记忆网络 双向长短期记忆网络 卷积神经网络 卷积神经网络 萤火虫算法 萤火虫算法 全卷积网络 全卷积网络 FRPCA 模糊鲁棒主成分分析 遗传算法 遗传算法 GEI 创业板指数 GWO 灰狼优化器 和声搜索 和声搜索算法 核主成分分析 核主成分分析 长短期记忆网络 长短期记忆网络 MBCNN 多分支卷积神经网络 多核学习 多核学习 主成分分析 主成分分析 粒子群优化 粒子群优化算法 随机森林 随机森林 循环神经网络 循环神经网络 RNN-ER 具有事件表示的循环神经网络 谱聚类 谱聚类 SC-MBCNN 基于谱聚类的多分支卷积神经网络 SMI 股票市场指数 上证综合指数 上海证券综合指数 支持向量机 支持向量机 SZEC 深圳成分指数 TDSE-GA 基于遗传算法的两阶段动态堆叠集成

致谢。 本研究得到国家自然科学基金（项目编号：72401029）、中国博士后科学基金（项目编号：2023M740237 和 2024M750254）以及中央高校基本科研业务费专项资金（项目编号：2024CDJSKPT14）的资助。

附录A 不同数据源的描述

表 A.1：全球股票市场指数代码说明

No.	SMI代码	说明
1	SSEC	上证综合指数，包括在上海证券交易所上市的所有股票
2	AXJO	标普/澳大利亚证券交易所200指数，由在 澳大利亚证券交易所
3	N225	日经225指数，包括在东京证券交易所上市的225家公司
4	KS11	韩国综合指数，涵盖在韩国上市的780多家公司 证券交易所
5	TWII	台湾加权指数，涵盖在台湾证券交易所上市的所有公司 证券交易所（特别股和全额交割股除外）
6	HSI	恒生指数，涵盖香港证券交易所的蓝筹股 交易所
7	BSESN	由孟买证券交易所上市的30家公司组成的BSE Sensex指数 交易所
8	JKSE	包括印度尼西亚 证券交易所
9	PSI	菲律宾马尼拉指数，包括市值最大的30家公司 在菲律宾证券交易所上市
10	SPX	标普500指数 ，其中包括在纽约证券 交易所和纳斯达克证券交易所上市的500家公司
11	GSPTSE	标普/多伦多证券交易所指数，包含在多伦多证券 交易所
12	BVSP	博维斯帕指数，包含在圣保罗证券 巴西交易所上市的公司
13	COLCAP	哥伦比亚COLCAP指数，包含在哥伦比亚证券交易所上市的20家公司 证券交易所
14	MXX	墨西哥BOLSA指数，涵盖了在墨西哥证券 交易所
15	SPBLPGPT	秘鲁利马指数，涵盖了在利马证券交易所上市的20多家 证券交易所
16	BEL 20	比利时BEL 20指数由在 布鲁塞尔证券交易所
17	OMXC20	OMX哥本哈根20指数包含在 哥本哈根证券交易所
18	OMXH25	OMX赫尔辛基25指数，包含在 赫尔辛基证券交易所
19	FCHI	CAC 40指数，包含规模最大的40家公司以及交易最活跃的 在泛欧交易所巴黎证券交易所上市的公司
20	GDAXI	DAX 30指数，包含规模最大且交易最活跃的30家公司 在法兰克福证券交易所上市
21	ATG	雅典综合指数，由在雅典证券 交易所
22	FTMIB	富时MIB指数，包含在 意大利证券交易所
23	AEX	阿姆斯特丹交易所指数，由规模最大且交易最活跃的25家 在泛欧交易所阿姆斯特丹证券交易所上市的公司
24	OBX	奥斯陆证券交易所指数，包括在奥斯陆证券交易所上市的交易最活跃且流动性最强的25只股票 在奥斯陆证券交易所
25	WIG30	华沙证券交易所指数，由规模最大、流动性最强的 30 家 家在华沙证券交易所上市的公司组成
26	PSI20	葡萄牙股票指数，包含规模最大、流动性最强的 20 家公司 在泛欧交易所里斯本上市
27	IMOEX	莫斯科交易所俄罗斯指数，涵盖在 莫斯科交易所上市的
28	IBEX 35	伊比利亚指数，由35只流动性最高且交易最活跃的股票组成 在马德里证券交易所上市
29	BIST 100	包含流动性最高且交易最活跃的100只 在伊斯坦布尔证券交易所上市的股票
30	FTSE	由在伦敦证券交易所上市的100家最大公司组成的金融时报证券交易所100指数 在伦敦证券交易所上市

表A.2：行业指数数据说明

行业名称		No.	子行业名称
农业， 渔业	林业，畜牧业 and	1-1	农业
		1-2	林业
		1-3	畜牧业
		1-4	渔业
		1-5	农业，林业，畜牧业和渔业 服务业
采矿业		2-1	煤炭开采和洗选业
		2-2	石油和天然气开采业
		2-3	黑色金属矿采选业
		2-4	有色金属矿采选业
		2-5	非金属矿采选业
		2-6	采矿辅助活动
制造业		3-1	农副食品加工业
		3-2	食品制造业
		3-3	酒、饮料和精制茶制造业
		3-4	纺织业
		3-5	纺织服装、服饰业
		3-6	皮革、毛皮、羽毛及其制品和制鞋业- try
		3-7	木材加工和木材、竹、藤、棕榈、草 制品业
		3-8	家具制造业

行业名称	No.	子行业名称
	3-9	造纸及纸制品业
	3-10	印刷和记录媒介复制业
	3-11	教育、体育及娱乐用品制造业
	3-12	石油加工、炼焦及核燃料加工行业
	3-13	化学原料和化学制品制造业
	3-14	医药制造业
	3-15	化学纤维制造业
	3-16	橡胶和塑料制品业
	3-17	非金属矿物制品业
	3-18	黑色金属冶炼和压延加工业
	3-19	有色金属冶炼和压延加工业-try
	3-20	金属制品业
	3-21	通用设备制造业
	3-22	专用设备制造业
	3-23	汽车制造业
	3-24	铁路、船舶、航空航天及其他运输设备制造业
	3-25	电气 机械 and 设备 制造业行业
	3-26	计算机、通信及其他电子设备制造业
	3-27	仪器仪表制造业
	3-28	其他制造业
	3-29	废弃资源综合利用业
电力、热力、燃气及水生产和供应业	4-1	电力、热力生产和供应业
	4-2	燃气生产和供应业
	4-3	水生产和供应业
建筑业	5-1	房屋建筑业
	5-2	土木工程建筑业
	5-3	建筑安装业
	5-4	建筑装饰及其他建筑业
批发和零售业	6-1	批发业
	6-2	零售业
交通运输、仓储和邮政业	7-1	铁路运输
	7-2	道路运输
	7-3	水上运输
	7-4	航空运输
	7-5	装卸搬运和运输代理业
	7-6	仓储业
	7-7	邮政业
住宿和餐饮业	8-1	住宿业
	8-2	餐饮业
信息传输 , 软件和信息-技术服务	9-1	电信业 , 广播、电视和卫星传输服务
	9-2	互联网及相关服务
	9-3	软件和信息技术服务
金融	10-1	货币和金融服务
	10-2	资本市场服务
	10-3	保险
	10-4	其他金融服务
租赁和商务服务	11-1	租赁
	11-2	商务服务
科学研究和技术服务	12-1	研究与试验发展
	12-2	专业和技术服务
	12-3	科学技术推广和应用服-务
水利、环境和公共设施管-理	13-1	生态保护和环境治理
	13-2	公共设施管理
卫生和社会工作	14-1	卫生
	14-2	社会工作
文化、体育和娱乐	15-1	新闻与出版
	15-2	广播、电视、电影和视频录制制作
	15-3	文化艺术
	15-4	体育
房地产业		房地产
教育		教育
住宅服务、维修和		机动车、电子产品和日用消费品
一般工业		一般工业

表 A.3 财经新闻文本描述

新闻提供商	输入描述
新浪财经	新浪财经的正面指数、负面指数、中性指数、新闻数量；SMI 的历史数据（开盘价、最高价、最低价、收盘价、成交量，日收益率）；滞后期为 $t-1, t-2, .., t-L$。
华尔街见闻	正面指数、负面指数、中性指数、华尔街见闻的新闻数量；SMI 的历史数据（开盘价、最高价、最低价、收盘价、成交量、日收益率）；滞后期为 $t-1, t-2, ..., t-L$。
同花顺	同花顺的正面指数、负面指数、中性指数、新闻数量；SMI 历史数据（开盘价、最高价、最低价、收盘价、成交量、日收益率）；滞后期为 $t-1$，$t-2$、.....、$t-L$。
东方财富	东方财富的正面指数、负面指数、中性指数、新闻数量；其——SMI 历史数据（开盘价、最高价、最低价、收盘价、成交量，日收益率）；滞后期为 $t-1$、$t-2$、.....、$t-L$。
云财经	云财经的正面指数、负面指数、中性指数、新闻数量；其 SMI 的历史数据（开盘价、最高价、最低价、收盘价、成交量、日收益率）；滞后期为 t-1、$t-1$、$t-2$、.....、$t-L$。

附录 B 预测模型

B.1 逻辑回归模型

给定输入向量 $\boldsymbol{x} = (x_1, x_2, ..., x_6)$，线性模型如下所示：

$$z = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b = \boldsymbol{w}^T \boldsymbol{x} + b, \tag{B.1}$$

其中，$\boldsymbol{w} = (w_1, w_2, ..., w_d)$为权重向量，$d = 6$ 且 b 为偏置。本研究的问题是一个二分类问题，所提出的模型的输出为目标 SMI 走势，记作 $y \in \{0, 1\}$。基于逻辑函数，y 可由以下方程表示：

$$y = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\boldsymbol{w}^T \boldsymbol{x} + b)}}. \tag{B.2}$$

基于对数变化，式（B.2）转换为以下方程：

$$\ln \frac{y}{1 - y} = \boldsymbol{w}^T \boldsymbol{x} + b. \tag{B.3}$$

若将 y 视为后验概率估计 $p(y = 1 \mid \boldsymbol{x})$，则式（B.3）转换为以下方程：

$$\ln \frac{p(y = 1 \mid \boldsymbol{x})}{p(y = 0 \mid \boldsymbol{x})} = \boldsymbol{w}^T \boldsymbol{x} + b. \tag{B.4}$$

基于式（B.4），参数 $\boldsymbol{w}$ 和 b 通过最大似然法确定。最后，根据参数 $\boldsymbol{w}$ 和 b，我们可以得到逻辑回归模型的预测结果。

B.2 K近邻模型

给定训练集 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2)....., (\boldsymbol{x}_n, y_n)\}$，其中 $\boldsymbol{x}_i$ 是 $i\text{-}th$ 样本的概率输入向量，y_i 是 $i\text{-}th$样本的类别。首先，K近邻模型寻找

k 个与训练数据集中的点 $\boldsymbol{x}_i$ 最近的邻居。邻域 $N_k(\boldsymbol{x}_i)$由上述 k 个最近邻组成，其中距离通过欧几里得距离计算：

$$L(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{l=1}^{d} (\boldsymbol{x}_i^l - \boldsymbol{x}_j^l)^2}. \tag{B.5}$$

根据邻域 $N_k(\boldsymbol{x}_i)$ 中的多数规则，点 $\boldsymbol{x}_i$ 的未知类别由其 k 个最近邻中的多数类别决定：

$$y_i = \arg\max_{c_j} \sum_{x_a \in N_k(x_i)} I(y_a = c_j), \tag{B.6}$$

其中，I 是指示函数。如果 $y_a = c_j$，则 $I = 1$；否则 $I = 0$。

B.3 支持向量机模型

给定训练集 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \ldots, (\boldsymbol{x}_n, y_n)\}$，假设原始特征空间中不存在能够正确划分两个类别样本的超平面。然后，假设 $\phi(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的映射特征向量，则特征空间的超平面如下式所示：

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b, \tag{B.7}$$

其中，$\boldsymbol{w}$ 和 b 分别是权重向量和偏置。为确定具有最大间隔的分离超平面，式（B.7）需满足以下方程：

$$\begin{aligned} \min_{\boldsymbol{w}, b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + b) \geq 1, \quad i = 1, 2, \ldots, n \end{aligned} \tag{B.8}$$

其中，支持向量是这些样本（$y_i(\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + b) = 1$）。基于拉格朗日乘子法和软间隔，式（B.9）的对偶问题如下式所示：

$$\begin{aligned} \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) \\ s.t. \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, \mathrm{n} \end{aligned} \tag{B.9}$$

其中，α_i 是拉格朗日乘子，C 是正则化常数。由于特征空间的维度可能很高，计算难度也随之增加，因此支持向量机模型采用核函数来计算 $\phi（\boldsymbol{x}_i）^T \phi（\boldsymbol{x}_j）$：

$$k(x_i, x_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) \tag{B.10}$$

核函数通常包括线性核函数、多项式核函数、径向基核函数、Sigmoid核函数和拉普拉斯核函数。如果方程（B.9）的解满足Karush-Kuhn-Tucker条件，则可以得到原问题和对偶问题的解。根据训练得到的参数 $\boldsymbol{w}$ 和 b，我们可以得到支持向量机模型的预测结果。

B.4 随机森林和极端随机树模型

随机森林（RF）模型和极端随机树（ET）模型都基于决策树模型，并包含多个决策树模型。上述模型的预测结果由多个决策树输出类型的众数决定。决策树主要包括一个根节点、多个内部节点和叶节点，其中叶节点表示决策结果，其他类型的节点为属性测试。决策树的属性被递归划分，以构建从根节点到叶节点的测试序列。由于决策树模型使用所有样本和所有特征来训练模型，因此过拟合风险会增加。为降低过拟合风险，随机森林模型和极端随机树模型会随机生成多个决策树。随机森林模型和极端随机树模型的具体细节如下所述。

对于随机森林模型，给定一个包含 n 个样本的数据集，其中每个样本具有 K个特征，每棵决策树都会从该数据集中有放回地随机选择 m 个训练样本（$m < n$），并随机使用 k 个特征（$k < K$）。基于所选样本和特征，可以构建一棵决策树。随后，重复这一随机过程以构建多个决策树。通过样本随机性和特征随机性来保证所生成决策树的多样性。

ET模型与RF模型的区别包括以下两点：(1) RF模型采用有放回的方式选择样本，而ET模型使用全部样本。(2) 在所选特征的基础上，RF模型递归地选择最佳划分特征来构建决策树，而ET模型随机选择特征来构建决策树。

B.5 人工神经网络模型

给定训练数据集 {($\boldsymbol{x}_1$，y_1)、($\boldsymbol{x}_2$，y_2).....、($\boldsymbol{x}_n$，y_n)}，人工神经网络模型的第一层为输入层，其中包含六个神经元 $\boldsymbol{x}$ =(x_1，x_2，.....，x_6)。所设计的人工神经网络模型包含三个隐藏层。基于输入层，第一个隐藏层的 j-th 神经元的输出通过以下公式计算：

$$o_j^{(1)} = \theta(\sum_{i=1}^{d} \boldsymbol{w}_{i,j}^{(1)} \boldsymbol{x}_i + b_j) \tag{B.11}$$

其中，θ 为激活函数，$\boldsymbol{w}_{i,\ j}$ 为第一隐藏层中 i-th 神经元与第二隐藏层中 j-th 神经元之间的连接权重，b_j 为偏置。第二隐藏层和第三隐藏层中神经元的输出也通过上述公式计算。最后一层为输出层，其中包含两个神经元。在最后一层中，采用Softmax函数作为激活函数，

并通过以下公式计算 $i\text{-}th$ 类型的概率：

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{B.12}$$

其中， n 是类型的数量， z_j 是 $j\text{-}th$ 神经元的输出值。根据输出层，将支持概率最高的类型作为预测结果，即预测的SMI走势。

附录C 详细实验结果

C.1. 数据描述

表 C.1 展示了不同财经新闻提供商的情绪指数和新闻数量的统计结果。

表 C.1 情绪指数和新闻数量的统计结果

媒体指标	Mean	Std	25%	中位数	75%	Mean	Std	25%	中位数	75%
	新闻提供方1: 新浪财经					新闻提供方2: 华尔街见闻				
正面指数	0.502	0.070	0.445	0.505	0.551	0.519	0.060	0.481	0.518	0.561
负面指数	0.220	0.049	0.185	0.216	0.248	0.209	0.040	0.180	0.205	0.230
中性指数	0.279	0.041	0.250	0.281	0.307	0.273	0.041	0.244	0.271	0.299
新闻数量	546.6	330.5	220.5	556.0	901.5	351.3	213.8	203.0	322.0	418.5
	新闻提供方3: 同花顺					新闻提供方4: 东方财富				
正面指数	0.534	0.064	0.497	0.538	0.573	0.540	0.060	0.499	0.541	0.585
负面指数	0.203	0.052	0.165	0.198	0.232	0.222	0.048	0.187	0.217	0.253
中性指数	0.263	0.041	0.241	0.263	0.288	0.238	0.045	0.206	0.232	0.266
新闻数量	294.9	136.9	164.0	326.0	405.5	331.4	103.8	297.5	349.0	385.0
	新闻提供商5: 云财经					市场新闻				
正面指数	0.567	0.060	0.533	0.571	0.601	0.525	0.055	0.489	0.524	0.562
负面指数	0.175	0.041	0.147	0.169	0.197	0.208	0.040	0.180	0.205	0.230
中性指数	0.258	0.039	0.238	0.259	0.280	0.267	0.035	0.244	0.268	0.291
新闻数量	348.8	135.5	285.0	370.0	440.5	1404	588.1	977.5	1481.0	1835

根据表C.1的结果，我们发现，五家新闻提供商的平均正值均高于负面和中性指数，这表明这些中国新闻提供商倾向于以热情的语气发布新闻文章。相比之下，以往研究表明，英语新闻媒体倾向于发布更多悲观的新闻[1]。此外，我们对表C.1 的分析显示，与其他新闻提供商相比，新浪财经每天发布的新闻文章数量更多。

C.2. 不同数据源的预测结果

具体而言，我们将上述数据源作为输入，输入一系列预测模型，包括人工神经网络、支持向量机、随机森林、卷积神经网络、循环神经网络、长短期记忆网络、门控循环单元和二维卷积神经网络。不同数据源对上证综合指数、SZEC 和 GEI 的预测结果如表 C.2 所示，其中“市场”表示模型的输入为历史市场数据，“全球”表示模型的输入为全球股票市场指数数据，“行业”

[1]Gao, R., Cui, S., Xiao, H., Fan, W., Zhang,H., &Wang,Y.(2022).整合多家新闻提供商的情绪以预测股票市场指数走势：一种基于证据推理规则的深度学习方法。Information Sciences, 615, 529–556.

表示以行业指数数据作为输入，“新闻”表示以不同新闻提供商的情绪指数作为输入。在表 C.2中，每行带下划线的数字表示该行的最佳值。“均值”指标表示每行中所有预测模型的平均值。“均值”列中的粗体数字表示针对某一评价指标，不同数据源所得到的最大预测结果。此外，“排名”指标表示不同数据源对应的“均值”指标的排序结果。

表 C.2：不同数据源对 SSEC、SZEC 和 GEI 的预测结果

指标	输入	ANN	SVM	RF	CNN	RNN	LSTM	GRU	二维卷积神经网络	Mean	Rank
目标SMI：上证综合指数											
准确率	市场	0.496	0.525	0.511	0.513	0.527	0.498	0.521	0.486	0.510	3
	全球	0.525	0.568	0.521	0.527	0.502	0.536	0.519	0.519	**0.527**	1
	行业	0.498	0.458	0.491	0.501	0.494	0.504	0.497	0.483	0.491	4
	News	0.497	0.531	0.526	0.512	0.519	0.494	0.501	0.511	0.511	2
召回率	市场	0.451	0.518	0.426	0.578	0.460	0.537	0.449	0.581	0.500	3
	全球	0.546	0.605	0.619	0.605	0.514	0.581	0.583	0.607	**0.582**	1
	行业	0.554	0.469	0.556	0.543	0.517	0.589	0.592	0.539	0.545	2
	News	0.433	0.446	0.407	0.520	0.457	0.463	0.413	0.531	0.459	4
精确率	市场	0.486	0.528	0.416	0.498	0.476	0.516	0.562	0.454	0.492	3
	全球	0.508	0.554	0.521	0.520	0.494	0.525	0.515	0.518	**0.519**	1
	行业	0.496	0.454	0.489	0.499	0.496	0.500	0.493	0.480	0.488	4
	News	0.498	0.544	0.549	0.505	0.518	0.489	0.484	0.544	0.517	2
F-度量	市场	0.449	0.496	0.399	0.513	0.426	0.459	0.411	0.468	0.452	3.5
	全球	0.521	0.570	0.551	0.543	0.494	0.540	0.533	0.543	**0.537**	1
	行业	0.511	0.457	0.507	0.512	0.499	0.533	0.532	0.496	0.506	2
	News	0.433	0.463	0.420	0.481	0.458	0.451	0.420	0.491	0.452	3.5
AUC	市场	0.518	0.509	0.541	0.544	0.559	0.577	0.567	0.569	0.548	2
	全球	0.491	0.597	0.542	0.546	0.540	0.561	0.545	0.567	**0.549**	1
	行业	0.516	0.459	0.519	0.507	0.473	0.509	0.485	0.471	0.492	4
	News	0.495	0.505	0.551	0.506	0.483	0.507	0.466	0.527	0.505	3
目标SMI：SZEC											
准确率	市场	0.480	0.521	0.493	0.488	0.488	0.495	0.512	0.487	0.496	3.5
	全球	0.529	0.562	0.557	0.531	0.533	0.542	0.551	0.532	**0.542**	1
	行业	0.483	0.489	0.496	0.516	0.504	0.491	0.510	0.511	0.500	2
	News	0.501	0.481	0.495	0.500	0.501	0.480	0.490	0.519	0.496	3.5
召回率	市场	0.382	0.399	0.300	0.462	0.436	0.482	0.475	0.454	0.424	3
	全球	0.596	0.608	0.611	0.605	0.541	0.575	0.581	0.564	**0.585**	1
	行业	0.469	0.479	0.499	0.558	0.506	0.517	0.534	0.533	0.512	2
	News	0.416	0.313	0.342	0.435	0.472	0.404	0.455	0.528	0.421	4
精确率	市场	0.472	0.456	0.499	0.525	0.511	0.472	0.565	0.595	0.512	3
	全球	0.544	0.578	0.571	0.534	0.550	0.550	0.565	0.545	**0.555**	1
	行业	0.512	0.516	0.515	0.532	0.525	0.518	0.538	0.538	0.524	2
	News	0.510	0.460	0.502	0.483	0.519	0.493	0.490	0.521	0.497	4
F-衡量	市场	0.391	0.405	0.334	0.418	0.399	0.431	0.430	0.396	0.400	4
	全球	0.563	0.581	0.581	0.557	0.540	0.555	0.564	0.543	**0.561**	1
	行业	0.483	0.490	0.501	0.540	0.508	0.507	0.525	0.527	0.510	2
	News	0.434	0.355	0.358	0.407	0.476	0.417	0.445	0.499	0.424	3
AUC	市场	0.487	0.533	0.516	0.518	0.548	0.507	0.529	0.509	0.518	2
	全球	0.546	0.590	0.586	0.541	0.556	0.552	0.558	0.556	**0.561**	1
	行业	0.498	0.487	0.487	0.481	0.492	0.463	0.486	0.499	0.487	3
	News	0.494	0.502	0.489	0.436	0.435	0.466	0.452	0.478	0.469	4
目标SMI：GEI											
准确率	市场	0.544	0.493	0.561	0.533	0.561	0.558	0.583	0.507	0.542	2
	全球	0.550	0.578	0.561	0.572	0.517	0.570	0.559	0.541	**0.556**	1
	行业	0.524	0.519	0.509	0.522	0.553	0.539	0.525	0.559	0.531	3
	News	0.517	0.496	0.523	0.497	0.515	0.513	0.511	0.509	0.510	4
召回率	市场	0.521	0.336	0.480	0.620	0.613	0.725	0.650	0.560	0.563	3.5
	全球	0.540	0.646	0.679	0.662	0.562	0.659	0.624	0.685	**0.632**	1
	行业	0.564	0.558	0.622	0.597	0.547	0.581	0.590	0.574	0.579	2
	News	0.534	0.491	0.468	0.599	0.591	0.611	0.597	0.611	0.563	3.5
精确率	市场	0.615	0.477	0.583	0.600	0.613	0.603	0.652	0.502	0.581	3
	全球	0.601	0.620	0.596	0.611	0.573	0.611	0.608	0.577	**0.600**	1
	行业	0.580	0.581	0.560	0.569	0.618	0.599	0.573	0.610	0.586	2
	News	0.525	0.541	0.687	0.526	0.551	0.562	0.562	0.566	0.565	4
F-测量	市场	0.517	0.376	0.491	0.548	0.567	0.615	0.591	0.505	0.526	3.5
	全球	0.563	0.620	0.622	0.625	0.558	0.623	0.606	0.616	**0.604**	1
	行业	0.561	0.558	0.575	0.572	0.571	0.575	0.572	0.585	0.571	2
	News	0.509	0.500	0.460	0.544	0.552	0.560	0.547	0.535	0.526	3.5
AUC	市场	0.529	0.492	0.537	0.545	0.591	0.582	0.596	0.534	0.551	2
	全球	0.569	0.618	0.589	0.556	0.535	0.592	0.585	0.522	**0.571**	1
	行业	0.498	0.529	0.514	0.529	0.562	0.571	0.548	0.551	0.538	3
	News	0.510	0.483	0.514	0.486	0.505	0.502	0.480	0.494	0.497	4

根据表C.2中的结果，我们的分析得出了两项重要见解。首先，我们研究了在历史市场数据之外整合补充数据源对预测性能的影响。我们发现，这些不同数据

源对预测能力的贡献差异显著。具体而言，对于全球SMI数据，将其纳入后，三个数据集（SSEC、SZEC和GEI）上的预测模型预测能力均显著增强。这种增强在以下五项指标中尤为明显：准确率、召回率、精确率、*F*值和曲线下面积。对于行业指数数据，将其整合后，SZEC和GEI数据集上的整体预测性能有所提升，但在SSEC数据集中整合这类数据则会对预测性能产生负面影响。对于新闻数据，我们发现，在SSEC、SZEC和GEI数据集中，将其纳入预测模型并未带来明显改善。第二项观察涉及不同模型预测性能的比较。表C.2 中的结果显示，支持向量机、随机森林、长短期记忆网络和二维卷积神经网络模型在SSEC、SZEC和GEI数据集上表现出色。具体而言，在SSEC数据集中，支持向量机模型取得最优结果的频率最高，其次是随机森林模型和长短期记忆网络模型。在SZEC数据集中，二维卷积神经网络模型最常表现出色，其次是支持向量机模型和卷积神经网络模型。同样，在GEI数据集中，长短期记忆网络模型取得优异结果的频率最高，二维卷积神经网络模型和随机森林模型则紧随其后。

C.3. 不同预测模型的预测结果比较

表 C.3 TDSE-GA模型的超参数设置

Type	模型	参数	值
优化模型	GA	种群规模	50
		最大迭代代数	20
		最大停滞代数	5
		选择率	0.3
		交叉率	0.8
特征表示	MBCNN	优化器	Adam
		卷积层的激活函数和全连接层	RELU
		输出层的激活函数	Softmax
	SC-MBCNN	确定聚类数量的方法	肘部法
		优化器	Adam
		卷积层的激活函数及全连接层	RELU
		输出层的激活函数	Softmax
	RNN-ER	优化器	Adam
		输出层的激活函数	Softmax
元学习器	LR	C 值	{0.1、1、10、100、 1000}
	径向基函数支持向量机	C 值	{0.1、1、10、100、 1000}
		伽马值	{0.1、0.5、1、1.5、2、2.5}
	多项式支持向量机	C 值	{0.1、1、10、100、 1000}
		多项式值	{2, 3, . . . , 10}
	RF	决策树数量	{5, 10, 15, . . . , 50}
	ET	决策树数量	{5, 10, 15, . . . , 50}
	ANN	各层的神经元数量	{2, 3, . . ., 50}
	KNN	邻居样本的数量	{2, 3, . . ., 20}

C.4. 不同优化算法的比较

在本研究中，我们采用GA作为优化算法。为验证所选算法的有效性，我们将GA与其他优化算法进行比较，具体包括粒子群优化（PSO）算法、灰狼

优化器（GWO）、萤火虫算法（FA）和和声搜索（HS）算法。比较标准包括每个模型的预测分类性能和计算时间。这些优化方法在SSEC、SZEC和GEI数据集上的预测结果如表C.4 所示，其中每列的粗体数字代表相应SMI的最高值。根据表C.4中的结果，我们得出以下两点结论。(1) 遗传算法优于其他优化方法，在准确率（SSEC和SZEC）、召回率（SSEC和GEI）、精确率（SSEC和GEI）、F值（SSEC、SZEC和GEI）以及曲线下面积（SSEC和SZEC）方面取得了最高结果。(2) 萤火虫算法在SSEC上的准确率、召回率、精确率和F值指标方面取得了次优结果。类似地，粒子群优化算法在SZEC上的准确率、精确率和AUC值方面表现次优。此外，灰狼优化算法在GEI数据集上的准确率、召回率、精确率、F值和AUC值方面取得了次优结果。

表 C.4 不同优化方法在上证综合指数、SZEC 和 GEI 上的预测结果

SMI	模型	准确率	召回率	精确率	F-度量	AUC
SSEC	TDSE-PSO	0.6059	0.6343	0.5885	0.5925	0.6042
	TDSE-GWO	0.6029	0.6134	0.5867	0.5822	0.5956
	TDSE-FA	0.6307	0.7021	0.6136	0.6265	0.6191
	TDSE-HS	0.5883	0.6372	0.5744	0.5852	0.5802
	TDSE-GA	**0.6451**	**0.7236**	**0.6442**	**0.6521**	**0.6453**
SZEC	TDSE-PSO	0.6075	0.7291	**0.6460**	0.6290	0.6148
	TDSE-GWO	0.5917	**0.7851**	0.6171	0.6474	0.5932
	TDSE-FA	0.6035	0.7167	0.6164	0.6337	0.6009
	TDSE-HS	0.5987	0.7403	0.6358	0.6342	0.6051
	TDSE-GA	**0.6225**	0.7708	0.6312	**0.6624**	**0.6192**
GEI	TDSE-PSO	0.5902	0.6236	0.6349	0.6188	0.5823
	TDSE-GWO	**0.6273**	0.7037	0.6719	0.6526	**0.6263**
	TDSE-FA	0.6123	0.6612	0.6606	0.6363	0.6070
	TDSE-HS	0.6072	0.6819	0.6638	0.6402	0.6037
	TDSE-GA	0.6258	**0.7419**	**0.6823**	**0.6663**	0.6188

除了比较预测结果外，我们还通过比较不同优化算法的运行时间来评估计算效率。不同优化方法在SSEC、SZEC和GEI数据集上的运行时间如图所示。C.1。根据图C.1，我们首先发现，与其他优化算法相比，遗传算法在SSEC、SZEC和GEI数据集上的运行时间更短，分别为257.15秒、221.47秒和220.68秒。其次，萤火虫算法在这些数据集上的运行时间明显更长，分别为3110.25秒、6760.76秒和2792.34秒。萤火虫算法运行时间较长，可能是因为该算法在每一代中都需要通过个体之间的两两比较来确定最优解。

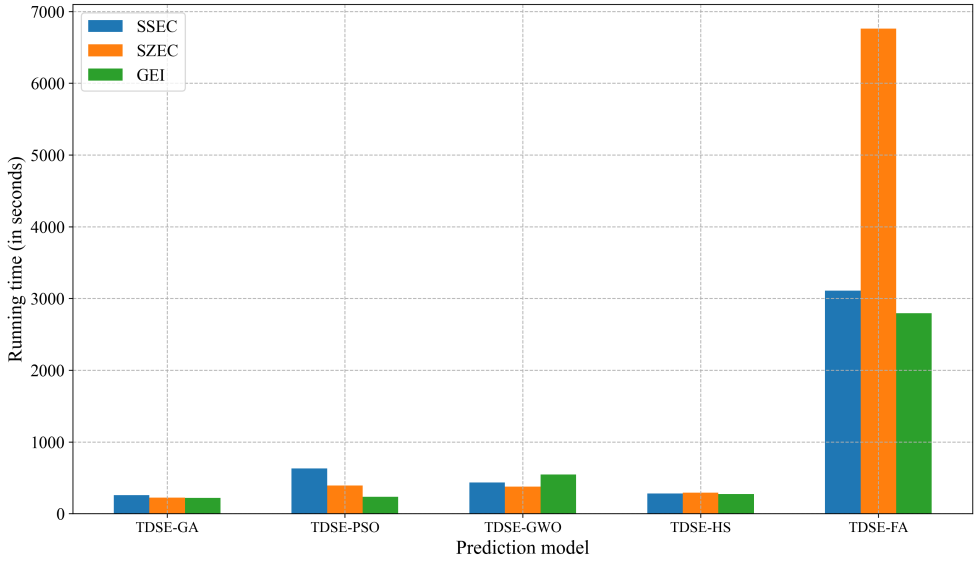

图 C.1 不同优化方法在上证综合指数、SZEC 和 GEI 上的运行时间

根据表C.4 和图C.1所示的结果可以看出，遗传算法在预测性能和计算效率方面均优于粒子群优化算法、灰狼优化算法、和声搜索算法和萤火虫算法。与其他优化方法相比，遗传算法在SSEC、SZEC和GEI数据集上的准确率、召回率、精确率、F值和AUC值始终更高。此外，遗传算法的运行时间明显更短，因此是优化TDSE模型第二阶段的更高效选择。这些实验结果凸显了遗传算法作为所提出的模型中一种优化方法的优越性。

参考文献

Bao W、Cao Y、Yang Y 等（2025）基于神经网络的数据驱动的股票预测模型：综述。InformationFusion 113:102616。https://doi.org/10.1016/j.inffus.2024.102616

Cao S、Jiang W、Wang J 等（2024）从人类与机器到人类 + 机器：股票分析的艺术与人工智能。Journal of Financial Economics 160:103910。https://doi.org/10.1016/j.jfineco.2024.103910Constantinou E、Georgiades R、Kazandjian A 等（2006）体制转换与人工神经网络预测在塞浦路斯证券交易所日收益率中的应用。International Journal of Finance & Economics 11(4):371–383。https://doi.org/10.1002/ijfe.305

Cui S、Qiu H、Wang S 等（2021）用于汽油辛烷值损失预测的两阶段堆叠异质集成学习方法。应用软计算 113:107989。https://doi.org/10.1016/j.asoc.2021.107989Dong L、Zheng H、Li L 等（2024）在在线社区中提炼群体智慧：利用评论发布者构建新型预测市场。决策支持系统 180:114190。https://doi.org/10.1016/j.dss.2024.114190Gao R、Cui S、Xiao H 等（2022a）整合多个新闻提供商的情绪进行股票市场指数走势预测：一种基于证据推理规则的深度学习方法。Information Sciences 615:529–556。https://doi.org/10.1016/j.ins.2022.10.029Gao R、Zhang X、Zhang H 等（2022b）结合全球市场指数预测股票市场指数隔夜收益方向：一种多分支深度学习方法。专家系统与应用 194:116506。https://doi.org/10.1016/j.eswa.2022.116506Gao Y、Zhao C、Wang Y（2024）投资者情绪与股票收益：中国碳中和股票市场基于多源数据的新证据。Interna-tional Review of Economics & Finance 92:438–450。https://doi.org/10.1016/j.iref.2024.02.049Geva T、Zahavi J（2014）结合市场数据与文本新闻的自动化日内股票推荐系统的实证评估。决策支持系统 57:212–223。https://doi.org/10.1016/j.dss.2013.09.013He C（2001）中国股市“板块现象”分析。经济研究 12:82–87 Hoseinzade E、Haratizadeh S（2019）Cnnpred：使用多样化变量集的基于卷积神经网络的股市预测。专家系统与应用 129:273–285。https://doi.org/10.1016/j.eswa.2019.03.029Jiang F、Meng L、Tang G（2021）媒体文本情绪与中国股票收益可预测性。中国经济季刊 21(04):1323–1344。https://doi.org/10.13821/j.cnki.ceq.2021.04.10Jiang J、Kelly B、Xiu D（2023）（重新）想象价格趋势。The Journal of Finance 78(6):3193–3249。https://doi.org/10.1111/jofi.13268Kia AN、Haratizadeh S、Shouraki SB（2018）一种用于预测全球股票市场和大宗商品价格次日走势的混合监督与半监督图模型。专家系统与应用 105:159–173。https://doi.org/10.1016/j.eswa.2018.03.037Lee K、Kim YJ、Kim M 等（2025）用于股市预测的分层多模态融合架构搜索。应用软计算 183。https://doi.org/10.1016/j.asoc.2025.113581

Li B, Yang Y (2022) 中国股市的无向与有向网络分析。Co
mputationalEconomics 60(3):1155–1173.
https://doi.org/10.1007/s10614-021-10183-wLi Q, Wang T, Li P, 等 (2014) 新闻和公众情绪对股票走势的影响。Information Sciences 278:826–840.
https://doi.org/10.1016/j.ins.2014.03.096Li Q, Wang J, Wang F, 等 (2017) 社交情绪在股票市场中的作用：从多种信息源的联合效应看。Multimedia Tools and Applications 76(10):12315–12345. https://doi.org/10.1007/s11042-016-3643-4Li Q, Tan J, Wang J, 等 (2021) 基于在线新闻进行股票走势预测的多模态事件驱动长短期记忆网络模型。IEEE Transactions on Knowledge and Data Engineering 33(10):3323–3337. https://doi.org/10.1109/TKDE.2020.2968894Li S, Tian Z, Li Y (2023) 融合多源多频信息的残差长短期记忆网络：在中国股市中的应用。Information Sciences 622:133–147.https://doi.org/10.1016/j.ins.2022.11.136Ma Y, Mao R, Lin Q, 等 (2022) 用于股价走势预测的多源聚合分类。Information Fusion 91:515–528. https://doi.org/10.1016/j.inffus.2022.10.025

Marschinski R, Kantz H (2002) 分析金融时间序列之间的信息流。《欧洲物理学杂志 B：凝聚态与复杂系统》30(2):275–281。https://doi.org/10.1140/epjb/e2002-00379-2

Nam K, Seong N（2019）基于韩国股市影响因果关系分析的财经新闻股票走势预测。决策支持系统 117：100–112。 https://doi.org/10.1016/j.dss.2018.11.004Ng A, Jordan M, Weiss Y（2002）论谱聚类：分析与算法。载于：神经信息处理系统进展, 第14卷。MIT出版社 Orimoloye LO, Sung MC, Ma T, 等（2020）比较深度前馈神经网络与浅层架构在预测股票价格指数方面的有效性。专家系统与应用 139：112828。https://doi.org/10.1016/j.eswa.2019.112828Pang B, Lee L（2008）观点挖掘与情感分析。基础与趋势® 信息检索 2（1–2）：1–135。
https://doi.org/10.1561/1500000011Papouskova M, Hajek P（2019）采用异质集成学习的两阶段消费信贷风险建模。决策支持系统 118：33–45。
https://doi.org/10.1016/j.dss.2019.01.002Pei Q, Zhu（2019）中国股市的金融传染及其渠道——基于行业数据的实证研究。管理科学学报 22（03）：90–112

Schumaker RP, Chen H（2009）利用突发财经新闻进行股票市场预测的文本分析：azfin文本系统。ACM信息系统学报 27（2）：1–19。https://doi.org/10.1145/1462198.1462204Shen F, Huang S, Zhao W, 等（2026）旅游股票指数动态预测：融合新兴媒体数据的多尺度深度学习框架。专家系统与应用 298：129566。 https://doi.org/10.1016/j.eswa.2025.129566Shynkevich Y, McGinnity T, Coleman SA, 等（2016）利用多核学习，根据不同类别的新闻文章预测医疗保健股票价格走势。决策支持系统 85：74–83。https://doi.org/10.1016/j.dss.2016.03.001Sn´a˘selV, Vel´asquez JD, Pant M, 等（2024）对基于多源融合的股票选择框架进行推广。Information Fusion 102：102018。 https://doi.org/10.1016/j.inffus.2023.102018

Su M（2017）中国A股市场行业轮动研究。《南方金融》02：40–48

Thakkar A、Chaudhari K（2021）股票市场预测中的融合：关于其必要性、最新进展及未来潜在方向的十年综述。《Information Fusion》65：95–107。https://doi.org/10.1016/j.inffus.2020.08.019

Vaswani A、Shazeer N、Parmar N 等（2017）注意力机制是你所需要的一切。载于：Guyon I、Luxburg UV、Bengio S 等（编），《神经信息处理系统进展》第30卷。神经信息处理系统，拉霍亚

Wang J, Hu Y, Jiang TX, 等（2023）用于多模态信息驱动股票走势预测的本质张量学习。《知识型系统》262：110262。https://doi.org/10.1016/j.knosys.2023.110262Wang Q, Xu W, Zheng H（2018）利用深度随机子空间集成，结合群体智慧与技术分析进行金融市场预测。神经计算 299：51–61。https://doi.org/10.1016/j.neucom.2018.02.095Wang Z, Yan W, Oates T（2017）从零开始使用深度神经网络进行时间序列分类：一个强基线。载于：2017年国际神经网络联合会议（IJCNN），第1578–1585页，https://doi.org/10.1109/IJCNN.2017.7966039Weng B, Lu L, Wang X, 等（2018）利用集成方法和在线数据源预测短期股票价格。《专家系统与应用》112：258–273。https://doi.org/10.1016/j.eswa.2018.06.016Wolpert DH（1992年）堆叠泛化。《神经网络》5（2）：241–259。https://doi.org/10.1016/S0893-6080(05)80023-1Xu H, Chai L, Luo Z, 等（2020）基于推文和历史价格、采用融合注意力机制的股票走势预测网络。神经计算418：326–339。https://doi.org/10.1016/j.neucom.2020.07.108

Ye W, Yang J, Chen P（2024）利用高频限价订单簿数据成像进行短期股价趋势预测。《国际预测杂志》40（3）：1189–1205。https://doi.org/10.1016/j.ijforecast.2023.10.008Yun KK, Yoon SW, Won D（2021）采用三阶段特征工程流程的混合遗传算法-xgboost算法进行股票价格方向预测。《专家系统与应用》186：115716。https://doi.org/10.1016/j.eswa.2021.115716Zhang X, Zhang Y, Wang S, 等（2018）通过异构信息融合改进股票市场预测。《知识型系统》143：236–247。https://doi.org/10.1016/j.knosys.2017.12.025Zhao Y, Gong X, Zhang W, 等（2025）基于类别因子代理变量的股票收益率预测。《金融创新》11（1）。https://doi.org/10.1186/s40854-025-00779-8Zhong X, Enke D（2017）利用降维预测每日股票市场收益率。《专家系统与应用》67：126–139。https://doi.org/10.1016/j.eswa.2016.09.027