Computer Science and Application 计算机科学与应用, 2025, 15(6), 45-55
Published Online June 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.156156

Hans汉斯

层次聚类和长短期记忆网络(LSTM)混合机器学习模型的数据资产估值模型

彭守斌[1,2]，杨学军[2]

[1]上海大学智慧城市研究院，上海
[2]上海荟宸信息科技有限公司研究中心，上海

收稿日期：2025年5月11日；录用日期：2025年6月9日；发布日期：2025年6月17日

摘　要

研究数据资产估值在宏观层面能为数字经济发展、资源配置优化、行业规范和国家竞争力提升等方面提供支撑，对推动社会经济数字化转型意义重大。在数据资产估值过程中，会遇到诸多复杂问题，如数据异质性问题、时间序列特征挖掘问题、数据维度高和复杂性问题、缺乏通用估值标准问题和突发外部事件影响问题。层次聚类和长短期记忆网络(LSTM)混合机器学习模型将层次聚类划分异质数据成簇，LSTM挖掘各簇时间序列特征，应对高维复杂数据，结合制定估值标准，快速适应突发变化；可有效应对以上诸多复杂问题。

关键词

层次聚类，长短期记忆网络(LSTM)，AgglomerativeClustering (层次聚类算法)，门控机制(LSTM)，梯度消失/爆炸(RNN缺陷)，泛化能力，经典估值模型(成本法/市场法/收益法)，单一估值模型(K-Means/MLP/RNN)

A Data Asset Valuation Model of a Hybrid Machine Learning Model of Hierarchical Clustering and Long- and Short-Term Memory Networks

Shoubin Peng[1,2], Xuejun Yang[2]

[1]Intellectual City Research Institute of Shanghai University, Shanghai
[2]Research Center of Shanghai Hui Chen Information Technology Co., Ltd., Shanghai

Received: May 11th, 2025; accepted: Jun. 9th, 2025; published: Jun. 17th, 2025

文章引用: 彭守斌, 杨学军. 层次聚类和长短期记忆网络(LSTM)混合机器学习模型的数据资产估值模型[J]. 计算机科学与应用, 2025, 15(6): 45-55. DOI: 10.12677/csa.2025.156156

Abstract

Research on data asset valuation can support the development of the digital economy, optimize resource allocation, standardize industries, and enhance national competitiveness at a macro level, playing a significant role in promoting the digital transformation of society and the economy. In the process of data asset valuation, numerous complex issues arise, such as data heterogeneity, time series feature mining, high-dimensional and complex data, lack of universal valuation standards, and the impact of sudden external events. A hybrid machine learning model combining hierarchical clustering and long short-term memory (LSTM) clusters heterogeneous data into groups through hierarchical clustering, and LSTM mines temporal features within each cluster to handle high-dimensional and complex data. By formulating valuation standards, it quickly adapts to sudden changes, effectively addressing these various complex issues.

Keywords

Hierarchical Clustering, Long and Short Term Memory Network, AgglomerativeClustering (Hierarchical Clustering Algorithm), Gated Mechanism (LSTM), Gradient Disappearance/Explosion (RNN Defect), Generalization Ability, Classical Valuation Model (Cost Method/Market Method/Benefit Method), Single Valuation Model (K-Means/MLP/RNN)

1. 引言

数据资产估值是数字经济时代资源配置的核心环节，传统估值方法(如成本法、市场法、收益法)在应对多模态、高维度、动态变化的数据资产时，面临数据异质性处理不足、时间序列特征挖掘有限、估值标准不统一等挑战。随着机器学习技术的发展，融合层次聚类与长短期记忆网络(LSTM)的混合模型成为突破方向。层次聚类通过数据相似性分组降低复杂度，LSTM 则针对各簇时间序列特性建模，二者协同可有效解决单一模型在数据结构分析和动态预测中的缺陷。本文聚焦该混合模型的构建逻辑、协同机制及实际应用，旨在为数据资产估值提供更精准、自适应的解决方案，推动数字经济下数据要素的高效流通与价值释放。

2. 现有数据资产估值方法的回顾和比较

国内外常用的方法和模型主要围绕成本、市场、收益等维度构建，各有其适用场景与优劣。目前主流方法包括经典估值方法与新兴技术辅助估值模型。经典估值方法主要有成本法、市场法、收益法和综合法[1]；新兴技术辅助估值模型主要有基于机器学习的估值模型和区块链赋能的估值模型。

在应对多模态、大数据、跨领域跨行业、市场波动敏感等一系列应用场景，经典估值方法存在明显的不足，如成本法易忽略数据的潜在市场价值和未来收益、难以准确核算数据在不同阶段的成本；收益法难以合理预测未来收益和选择合适的折现率；市场法难以找到完全匹配的可比案例、市场波动会影响估值；综合法权重的选择往往缺乏客观标准[2]。在机器学习、大模型及区块链等新型技术持续发展的驱动下，新兴技术辅助估值模型越来越收到关注。其中，区块链赋能的估值模型，利用区块链的不可篡改、可追溯和智能合约特性，记录数据资产的产生、交易和使用过程，确保数据的真实性和完整性，为估值

提供可靠的数据基础[3]。在数据资产估值过程中，会遇到诸多复杂问题，如数据异质性问题、时间序列特征挖掘问题、数据维度高和复杂性问题、缺乏通用估值标准问题和突发外部事件影响问题[4]。

基于机器学习的数据估值模型，主要运用机器学习算法，对大量历史数据和市场信息进行学习和分析，挖掘数据特征与价值之间的复杂关系，从而实现对数据资产的估值[5]。可处理多源异构数据，能捕捉传统方法难以发现的价值驱动因素，提高估值的精准度和适应性。常见的主流模式有回归分析模型、神经网络模型和聚类分析模型。

回归分析模型是一种通过建立数学模型来描述一个或多个自变量与因变量之间关系的统计方法。它基于观测数据，尝试找出变量之间的规律和趋势，从而实现对因变量的预测或者对变量间关系的解释。主要包括：线性回归、多项式回归和逻辑回归。线性回归模型通过建立数据资产特征与价值之间的线性关系进行估值。该模型简单直观，易于理解和解释，计算复杂度低，在数据特征与价值呈现近似线性关系时，能快速得出估值结果。但它假设数据特征与价值之间是线性关系，实际情况中这种关系可能较为复杂，线性回归模型无法准确反映数据资产的真实价值。多项式回归模型允许数据资产特征与价值之间存在非线性关系。通过引入数据特征的多项式项，能够更好地拟合复杂的数据分布。但是，多项式次数过高可能导致过拟合，使模型在训练数据上表现良好，但在新数据上泛化能力较差，且模型的解释性会随着多项式次数增加而变弱。逻辑回归模型：虽名为“回归”，实则用于分类问题，但在数据资产估值中，可用于判断数据资产是否达到某个价值阈值，或者将数据资产价值划分为不同等级。它主要适用于分类任务，对于连续型的价值预测能力有限。

神经网络模型，又称人工神经网络(Artificial Neural Network, ANN)，是一种模拟生物神经网络结构和功能的计算模型，旨在通过对大量数据的学习来实现复杂的模式识别、预测和决策任务。主要包括多层感知机(MLP)、循环神经网络(RNN)和长短期记忆网络(LSTM)等。多层感知机(MLP)：是一种前馈神经网络，由输入层、隐藏层和输出层组成。隐藏层中的神经元通过权重连接，能够自动学习数据资产的复杂特征表示。在数据资产估值中，MLP 可以处理多个输入特征，通过对大量数据的训练，挖掘数据之间的非线性关系，从而实现精准的估值预测。它具有强大的非线性映射能力，能处理高度复杂的数据关系，对复杂数据的拟合效果好。但 MLP 需要大量的训练数据和计算资源，训练时间较长，且容易出现过拟合现象，模型的解释性较差。循环神经网络(RNN)特别适用于处理具有序列特性的数据，如时间序列数据。但是，传统 RNN 存在梯度消失或梯度爆炸问题，导致训练困难，难以学习长序列中的长期依赖关系。长短期记忆网络(LSTM)是 RNN 的改进版本，通过引入新的门控机制，有效解决了传统 RNN 的梯度消失和梯度爆炸问题[6]。在数据资产估值中，对于需要长期跟踪和分析的数据资产，如企业的历史交易数据，LSTM 可以更好地利用历史信息进行价值评估和预测。但它结构相对复杂，训练参数较多，计算成本较高，模型的训练和调优难度较大。

聚类分析模型是数据挖掘和统计学领域中，用于将物理或抽象对象集合分组为相似对象类别的方法。其核心是基于数据对象间的相似性度量，将数据集划分为多个簇，使同一簇内对象相似度高，不同簇间对象相似度低，旨在发现数据的内在结构和分布模式。常见的聚类模型包括 K-Means 聚类模型、层次聚类模型。K-Means 聚类模型将数据资产按照相似性划分为 K 个簇，通过计算数据点之间的距离(如欧氏距离)来确定簇的划分。在估值中，可先对数据资产进行聚类，将相似的数据资产归为一类，然后对每一类数据资产进行统一估值。K-Means 算法简单高效，计算速度快，对大规模数据的处理能力较强。但 K 值需要事先确定，选择不当会影响聚类效果，对初始聚类中心敏感，不同的初始值可能导致不同的聚类结果。层次聚类模型通过构建树形的聚类结构，逐步合并或分裂数据点，形成不同层次的聚类[7]。在数据资产估值中，可以根据数据资产的特征相似度，从单个数据资产开始，逐步合并相似的数据资产，直到形成一个大的聚类。

3. 单一数据资产估值模型面临的问题

在采用单一数据资产估值模型对数据资产估值过程中，会遇到诸多复杂问题，主要有：

数据异质性问题，数据资产往往具有高度的异质性，不同类型的数据资产在特征、价值驱动因素和变化规律上差异显著。例如，金融行业的数据资产，如股票交易数据和信贷违约数据，前者具有高频波动、受市场宏观因素影响大的特点，后者则更多与客户信用特征、经济环境等因素相关。传统单一模型难以同时捕捉这些不同类型数据的特征，导致估值不准确。

时间序列特征挖掘问题，数据资产的价值通常随时间变化，具有明显的时间序列特征。例如，电商平台的用户流量数据、社交媒体的用户活跃度数据等，都会随着时间呈现出周期性、趋势性等变化。传统的估值方法往往难以充分挖掘这些时间序列中的信息，导致对数据资产未来价值的预测能力不足。

数据维度高和复杂性问题，随着信息技术的发展，数据资产的维度越来越高，包含了大量的特征信息。高维度的数据增加了数据的复杂性，使得传统模型在处理时容易出现“维度诅咒”问题[7]，导致模型性能下降。

缺乏通用估值标准问题，数据资产的估值目前还缺乏统一的通用标准，不同类型的数据资产价值评估方法差异较大。而且数据资产的价值受到多种因素的影响，包括市场需求、数据质量、应用场景等，这些因素相互交织，使得估值变得更加复杂。

突发外部事件影响问题，数据资产的价值可能会受到突发外部事件的影响，如政策法规的变化、市场突发事件、技术创新等[8]。这些事件具有不可预测性，会导致数据资产的价值在短时间内发生剧烈变化。传统的估值模型通常难以快速适应这些变化，导致估值结果与实际价值偏差较大。

4. 层次聚类和长短期记忆网络(LSTM)混合机器学习模型及特点

为有效解决采用单一数据资产估值模型对数据资产估值过程中遇到诸多复杂问题，拟采用层次聚类和长短期记忆网络(LSTM)混合机器学习模型(本文称为“混合机器学习模型”)。混合机器学习估值模型结合了两者的优势，在数据资产估值等领域展现出显著优点和价值。

混合机器学习模型是一种将层次聚类的结构发现能力与 LSTM 的时间序列处理能力相结合，用于处理复杂数据任务的模型。该模型先通过层次聚类算法依据数据的相似性对数据进行分组，挖掘数据的内在结构，然后利用 LSTM 对每个聚类中的具有时间序列特征的数据进行建模，以捕捉数据在时间序列上的长期和短期依赖关系，从而更全面、准确地分析和预测数据[9]。

在数据资产估值过程中，层次聚类模型可以根据数据的特征相似度将数据资产划分为不同的簇，使得同一簇内的数据具有较高的相似性。这样就可以将异质的数据进行有效的分类，为后续的分析提供更有针对性的分组。例如，将金融数据资产分为股票、债券、信贷等不同的簇。而 LSTM 模型可以针对每个聚类单独进行训练，能够更好地学习每个聚类内数据的时间序列特征和价值变化规律，从而提高估值的准确性。LSTM 模型是专门处理时间序列数据的强大工具，它具有记忆单元和门控机制，能够有效地捕捉时间序列中的长期依赖关系和复杂模式。通过对历史数据的学习，LSTM 可以预测数据资产未来的价值变化趋势。结合层次聚类，先将数据按特征聚类，再对每个聚类使用 LSTM 进行时间序列分析，能够更精准地挖掘不同类型数据资产的时间序列特征，提高估值的及时性和准确性。层次聚类可以在高维空间中对数据进行降维和分类，通过将相似的数据点聚集在一起，减少数据的复杂度[10]。同时，将高维数据划分为不同的聚类后，每个聚类的数据维度相对降低，更便于 LSTM 模型进行处理。LSTM 可以专注于每个聚类内数据的时间序列特征，避免了在高维空间中直接建模的困难，提高了模型的训练效率和性能。层次聚类模型和 LSTM 混合模型可以根据数据的实际特征和变化规律进行个性化的估值。通过层次聚类将数据资产分类，再为每个聚类训练 LSTM 模型，能够更好地适应不同类型数据资产的特点，为

不同的聚类制定更合适的估值策略，从而解决缺乏通用估值标准的问题。虽然层次聚类和 LSTM 混合模型不能完全预测突发外部事件，但 LSTM 模型具有一定的动态学习能力。在事件发生后，通过持续更新训练数据，LSTM 可以快速调整模型参数，学习到新的价值变化模式，从而及时反映数据资产价值的变化。层次聚类则可以将受同一类外部事件影响的数据资产归为一类，便于模型集中处理和分析这些数据的变化规律。

混合机器学习模型融合了层次聚类在数据结构分析和 LSTM 在时间序列处理方面的专长，为数据资产估值带来系统性优势。

数据处理层面：层次聚类模型能依据数据资产的特征，将其划分成具有相似特性的群组。以电商企业的数据资产为例，可按照用户购买行为、地域分布等特征，把用户数据聚类。这一操作能有效降低数据的复杂性，使后续分析聚焦于每个群组的独特模式，避免不同特性数据相互干扰，提升对数据资产价值评估的精准度。而 LSTM 作为处理时间序列数据的强大工具，可充分挖掘数据资产价值随时间的变化规律。像金融数据资产，其价值受市场动态影响，LSTM 能捕捉到如每日股价波动、交易量变化等时间序列特征，对金融数据资产价值的动态变化进行有效建模。

模型性能层面：混合模型针对不同聚类训练专属的 LSTM 模型，能够更精准地适配各类数据的独特模式。不同类别的数据资产，其价值驱动因素和变化规律存在差异，统一模型难以全面捕捉[11]。同时，层次聚类提供的初始结构，有助于 LSTM 模型更高效地学习。它减少了模型训练时的噪声干扰，让 LSTM 专注于每个聚类内数据的时间序列特征，加速模型收敛，提高训练效率，减少训练所需的时间和计算资源[12]。

泛化能力层面：该混合模型具有更强的泛化能力，能更好地适应不同的数据分布和特征变化。在实际应用中，数据资产的特征和分布复杂多变，单一模型难以适应各种情况。混合模型通过层次聚类对不同特性数据分别处理，再利用 LSTM 的动态学习能力，使其在面对新数据时，能更灵活地调整预测，有效应对数据的多样性和不确定性，提升模型在不同场景下的适用性。

可解释性层面：层次聚类的结果具备直观的可解释性，它展示了数据资产之间的层次关系和相似性。通过聚类结果，分析师能清晰了解不同数据资产的分类依据，如哪些数据资产因相似特征被归为一类，这在一定程度上弥补了 LSTM 模型解释性不足的缺陷，帮助使用者更好地理解模型的决策过程和依据。

5. 混合机器学习模型设计及实验验证

5.1. 混合机器学习模型设计

混合机器学习模型在完成需求理解与数据收集、数据预处理的基础上为每个聚类训练一个 LSTM (长短期记忆网络)模型，利用层次聚类将数据分组，再针对每组数据的特点分别训练 LSTM 模型，从而提升模型对不同模式数据的适应性和预测准确性。

1) 需求理解与数据收集

需求理解：明确数据估值的具体目标和场景，例如对金融数据、电商用户数据等进行估值。确定估值的指标，如数据的市场价值、潜在价值等。

数据收集：收集与数据资产相关的多维度数据，包括但不限于数据的基本特征(如数据量、数据质量等)、历史价值数据(如果有)、时间序列信息等。确保数据的准确性和完整性。

2) 数据预处理

数据清洗：处理缺失值、异常值和重复数据。对于缺失值，可以采用删除、填充(如均值填充、中位数填充)等方法；对于异常值，可以通过统计方法(如 Z-score)进行识别和处理；删除重复的数据记录。

特征选择与提取：根据业务需求和领域知识，选择与数据估值相关的特征。可以使用特征选择算法(如相关性分析、递归特征消除等)来筛选重要特征。同时，也可以提取一些新的特征，如时间序列的趋势特征、季节性特征等[13]。

数据标准化：对数据进行标准化处理，将不同特征的取值范围统一到相同的尺度上，例如使用 Min-Max Scaler 或 Standard Scaler。这有助于提高模型的训练效率和稳定性。

3) 层次聚类

选择聚类算法：选择合适的层次聚类算法，如 AgglomerativeClustering。该算法是一种自底向上的聚类方法，一开始将每个样本看作一个单独的簇，然后不断合并相近的簇，直到达到预设的簇数量或者满足其他停止条件。以下从原理、sklearn 库中实现步骤和代码示例进行详细介绍：

算法原理如下：(1) 计算距离：计算所有样本对之间的距离，可以使用不同的距离度量，如欧氏距离、曼哈顿距离等。(2) 合并簇：每次迭代中，找到距离最近的两个簇，并将它们合并成一个新的簇。(3) 更新距离：合并簇后，更新剩余簇之间的距离。距离的更新方式有多种，如单链接(两个簇中最近样本的距离)、全链接(两个簇中最远样本的距离)、平均链接(两个簇中所有样本对距离的平均值)等。重复步骤(2)和(3)：直到达到预设的簇数量或者满足其他停止条件。

在 sklearn 库中的实现步骤：(1) 导入必要的库：需要导入 sklearn.cluster 中的 AgglomerativeClustering 类。(2) 准备数据：将数据整理成适合输入模型的格式，通常是一个二维数组，每行代表一个样本，每列代表一个特征。(3) 建模型实例：设置模型的参数，如簇的数量、距离度量、链接方式等。(4) 拟合模型：使用准备好的数据对模型进行训练。(5) 获取聚类结果：通过模型的 labels_属性获取每个样本的聚类标签。(6) 确定聚类参数：确定聚类的关键参数，如簇的数量、距离度量(如欧氏距离、曼哈顿距离等)和链接方式(如单链接、全链接、平均链接等)。可以通过肘部法则、轮廓系数等方法来选择合适的簇数量[14]。

4) LSTM 模型构建与训练

数据准备：对于每个聚类，将数据转换为适合 LSTM 模型输入的时间序列格式。通常，需要将数据整理成三维张量【样本数，时间步长，特征数】。

模型构建：为每个聚类构建一个 LSTM 模型。模型通常包含 LSTM 层、全连接层等。可以根据具体情况调整模型的结构和参数。

模型训练：使用每个聚类的数据对对应的 LSTM 模型进行训练。可以使用验证集来监控模型的性能，避免过拟合。

5) 模型评估与优化

模型评估：使用评估指标(如均方误差(MSE)、平均绝对误差(MAE)等)对每个 LSTM 模型进行评估，以衡量模型的性能[15]。

模型优化：如果模型性能不理想，可以尝试以下优化方法[16]：调整 LSTM 模型的结构和参数，如增加 LSTM 层的单元数、添加 Dropout 层等。尝试不同的层次聚类参数，以获得更合适的聚类结果。增加训练数据量或进行数据增强。

6) 预测与估值

新数据预处理：对新的数据进行预处理，包括清洗、特征选择、标准化等操作，使其与训练数据具有相同的格式和尺度[17]。聚类预测：使用训练好的层次聚类模型预测新数据所属的聚类。

估值预测：使用对应聚类的 LSTM 模型对新数据进行估值预测。

7) 模型部署与监控

模型部署：将训练好的模型部署到生产环境中，使其能够实时处理新的数据并进行估值预测。模型监控：定期监控模型的性能，收集新的数据并对模型进行更新和优化，以确保模型的准确性和稳定性。

5.2. 实验验证

1. 实验设计与数据准备

1) 实验数据

数据集：某金融机构的客户信用数据(含 2018~2023 年季度性指标)，包含以下特征：

静态特征：客户年龄、收入水平、资产规模(高/中/低)

动态特征：月度还款记录(时间序列长度 ＝24)、信用评分变化

标签：数据资产价值(基于收益法计算的未来 3 年预期收益现值)

数据划分：训练集：3000 条(含 2000 条正常数据 ＋1000 条高波动数据)、验证集：800 条、测试集：1200 条

2) 对比方法

单一模型：LSTM (直接处理全量数据)、层次聚类 ＋ 线性回归(HC + LR)、K-Means + LSTM (KM + LSTM)

经典方法：收益法(DCF)、市场法(对标同类数据交易价格)

3) 评价指标

均方误差(MSE)、平均绝对误差(MAE)、决定系数(R^2)

2. 混合模型 vs 单一模型/经典方法的对比实验

1) 模型配置

层次聚类参数：算法为 AgglomerativeClustering (距离度量 ＝ 欧氏距离，链接方式 ＝ ward，簇数量 = 3)

LSTM 参数：层数为 2 层、隐藏单元为 100 units (第 1 层)，50 units (第 2 层)、时间步长为 12、优化器采用 Adam (学习率 ＝1e − 3)

2) 实验结果，见表 1：对比实验结果

Table 1. Results of the comparative experiment
表 1. 对比实验结果

方法	MSE	MAE	R^2	训练时间(秒)
混合模型(HC + LSTM)	**0.028**	**0.019**	**0.92**	185
单-LSTM	0.045	0.032	0.85	152
HC + LR	0.061	0.048	0.78	45
KM + LSTM	0.035	0.026	0.89	168
收益法	0.123	0.091	0.62	-
市场法	0.105	0.083	0.68	-

3) 结果分析

精度优势：混合模型 MSE 比单一 LSTM 降低 37.8%，比市场法降低 73.3%，表明层次聚类的分组显著提升了 LSTM 对异质数据的建模能力。

效率权衡：混合模型训练时间比单一 LSTM 增加 21.7%，但精度提升明显，说明分组带来的收益超过计算成本。

经典方法局限性：收益法和市场法因无法捕捉动态时间特征，误差显著高于机器学习模型。

3. 层次聚类参数对模型性能的影响实验

1) 参数变量：簇数量(K = 2, 3, 4, 5)、距离度量(欧氏距离、曼哈顿距离、余弦距离)、链接方式(ward、average、complete)

2) 结果矩阵(测试集 MSE)，见表 2：结果矩阵

Table 2. Result matrix
表 2. 结果矩阵

簇数量	距离度量	链接方式	MSE	R^2
2	欧氏距离	ward	0.034	0.88
3	欧氏距离	ward	**0.028**	**0.92**
4	欧氏距离	ward	0.031	0.90
5	柗瑪明氏距离	ward	0.035	0.87
3	曼哈顿距离	ward	0.033	0.89
3	余称厕弦距离	ward	0.030	0.91
3	欧氏距离	average	0.032	0.90
3	欧氏距离	complete	0.036	0.86

3) 关键发现

簇数量优化：K = 3 时 MSE 最低，表明数据存在 3 种主要模式(如高信用、中信用、低信用客户簇)，过多簇会导致单簇样本量不足，过少则无法分离异质模式。

距离度量影响：欧氏距离优于曼哈顿距离和余弦距离，因数据特征为连续型数值，欧氏距离更能反映绝对差异。

链接方式：ward 链接(最小化簇内方差)效果最佳，符合混合模型“簇内同质性最大化”的需求。

4. LSTM 参数调优实验

1) 参数变量

隐藏层单元数(50, 100, 150)、时间步长(6, 12, 18, 24)、正则化系数(L2 = 0, 0.001, 0.01)

2) 结果矩阵(验证集 MSE)，见表 3：参数调优结果矩阵

Table 3. Parameter tuning result matrix
表 3. 参数调优结果矩阵

隐藏单元	时间步长	L2 正则	MSE	过拟合指数(训练/验证 MSE 比)
50	12	0	0.035	1.12
100	12	0	**0.028**	1.05
150	12	0	0.031	1.20 (过拟合)
100	6	0	0.042	1.03
100	18	0	0.034	1.07
100	24	0	0.038	1.09
100	12	0.001	0.030	1.04
100	12	0.01	0.033	1.02

3) 关键发现

隐藏单元数：100 units 时验证集 MSE 最低，增加至 150 units 导致过拟合(训练/验证 MSE 比 ＞1.2)。

时间步长：12 步(对应 1 年数据)最优，过短(6 步)丢失长期依赖，过长(24 步)引入噪声。

正则化：L2 = 0.001 时轻微抑制过拟合，但会牺牲少量精度，默认不使用正则化更平衡。

5. 动态适应性测试(突发外部事件模拟)

1) 实验设计

在测试集中注入“政策收紧”事件(第 18 个月)，人为降低高风险客户的信用评分(模拟政策对数据价值的冲击)。

对比混合模型与单一 LSTM 在事件前后的预测误差变化。

2) 结果对比，见表 4：动态适应性测试对比结果

Table 4. Dynamic adaptability test comparison results
表 4. 动态适应性测试对比结果

模型	事件前 MSE	事件后 MSE	误差增幅
混合模型	0.027	0.031	14.8%
单-LSTM	0.043	0.068	58.1%

3) 分析

混合模型因层次聚类将高风险客户单独分组，LSTM 可快速捕捉该簇的异常变化，误差增幅仅为单一 LSTM 的 25.5%，表明分组机制提升了模型对局部异常的响应能力。

6. 层次聚类与 LSTM 的协同与模型调整策略

6.1. 层次聚类与 LSTM 的协同

层次聚类与 LSTM 在混合模型中形成互补协同关系，通过数据结构发现与时间序列建模的有机结合，实现对复杂数据资产的精准估值。层次聚类与 LSTM 的协同关系包括数据结构驱动的建模分工、噪声过滤与特征增强和泛化能力的协同提升。

通过层次聚类的预处理作用和 LSTM 的动态特征挖掘构建数据结构驱动的建模分工的协同。通过自底向上的合并策略(如 AgglomerativeClustering)，依据欧氏距离、曼哈顿距离等度量标准，将高维异质数据划分为具有相似特征的簇。这一过程降低数据复杂度，避免不同模式数据混合导致的特征干扰，为 LSTM 提供同质性更强的输入子集。对每个簇内数据，LSTM 通过门控机制(遗忘门、输入门、输出门)捕捉时间序列中的长期依赖(如季度性消费趋势)和短期波动(如促销活动带来的流量激增)。

通过层次聚类的噪声隔离与 LSTM 的特征抽象能力来构建噪声过滤与特征增强的协同。聚类将差异较大的数据点分配至不同簇，减少同一训练集中无关特征的干扰。针对每个簇的时间序列数据，LSTM 的隐藏层单元可自动提取高阶特征。例如，在用户活跃度数据簇中，LSTM 可从原始访问时长、页面跳转次数等特征中，抽象出“用户粘性指数”等非线性组合特征，提升估值模型的表达能力。

通过层次聚类的场景适配与 LSTM 的动态适应来实现泛化能力的协同提升。通过调整簇数量(如肘部法则确定最优聚类数)，模型可灵活适应不同数据分布。例如，当数据资产包含“常规业务数据”与“创新业务数据”两类时，分为 2 个簇可避免单一模型对创新业务的异常模式过拟合。每个簇的 LSTM 模型独立训练，可针对该簇数据的时间特性调整参数(如时间步长、隐藏层维度)。

6.2. 基于聚类结果的 LSTM 模型调整策略

调整策略主要包括：模型结构调整、关键参数调整和训练策略优化。模型结构调整主要通过输入维

度适配和网络深度调整来实现。若某簇数据的特征维度较低(如仅包含时间戳与交易额)，可减少 LSTM 层的输入特征数；若簇数据包含多模态特征(如文本评论 + 数值型交易数据)，可引入嵌入层(Embedding Layer)对文本进行向量化，再输入 LSTM。

若簇内数据呈现明显线性趋势(如稳定增长的用户基数)，可采用单层 LSTM (如 50 个隐藏单元)，避免过拟合。复杂模式簇：若簇内数据存在多重周期或非线性波动(如受节假日影响的零售数据)，可堆叠多层 LSTM (如 2 层，每层 100 个隐藏单元)，增强特征提取能力。

调整关键参数：时间步长(seq_length)、学习率与优化器和正则化参数。时间步长主要根据簇内数据的时间相关性确定。通过遍历不同时间步长(如 5、10、20)，选择在验证集上 MSE 最小的参数，对高频数据簇(如股票分钟级交易数据)，设置较小时间步长(如 20 分钟)；对低频数据簇(如年度经济指标)，设置较大时间步长(如 5 年)。

学习率与优化器的调整策略：对样本量较小的簇(如新兴业务数据)，采用较小学习率(如 1e-4)和 Adam 优化器，避免梯度震荡；对样本量充足的簇(如成熟业务的历史数据)，可采用较大学习率(如 1e-3)和 RMSprop 优化器，加速收敛。采用正则化参数，对噪声较大的簇(如含异常值的传感器数据)，在 LSTM 层后添加 Dropout 层(如 dropout = 0.2)，或在损失函数中加入 L2 正则项，抑制过拟合。

通过数据增强与动态权重分配来实现训练策略优化。对样本量少的簇(如罕见事件数据)，采用时间序列平移、缩放等数据增强方法，生成虚拟样本。在多簇模型集成时，根据簇的估值误差动态调整权重。例如，对误差较小的簇赋予更高权重(如加权平均预测时权重为 0.6)，对误差较大的簇降低权重(如 0.4)。

7. 混合机器学习模型的潜在不足及解决方案

7.1. 潜在不足

虽然混合机器学习估值模型虽结合了两者优势，但仍存在潜在的不足，需要继续优化模型及场景。既有层次聚类、LSTM 模型自身不足，也有两者之间协同、应用场景的不足。

层次聚类的潜在不足主要两类：其一，聚类结果对数据顺序敏感，稳定性不足；主要原因：层次聚类(如 AgglomerativeClustering)的合并顺序依赖初始数据排列，随机顺序可能导致簇划分差异，影响后续 LSTM 训练数据的一致性。其二，聚类数量确定缺乏客观标准。主要原因：肘部法则、轮廓系数等传统方法在高维数据中可能失效，导致簇数量过拟合或欠拟合。

LSTM 模型的潜在不足主要有：训练效率低，计算资源消耗大；主要原因：LSTM 结构复杂(含遗忘门、输入门、输出门)，参数数量多，处理大规模数据时迭代速度慢。对突发外部事件的响应滞后；主要原因：LSTM 依赖历史数据学习模式，突发事件(如政策突变、市场崩盘)的非结构化信息未被显式建模。

混合模型协同性不足主要有：1) 聚类与 LSTM 的特征交互不足；原因：层次聚类仅提供静态分组，未与 LSTM 的动态特征提取形成反馈机制；2) 高维数据下聚类效果下降；原因：层次聚类在高维空间中面临“维度诅咒”，欧氏距离等度量失效，导致簇内相似性计算不准确。

最后，混合模型需基于历史数据训练，新类型数据(如新兴业务数据)缺乏足够样本时，聚类和 LSTM 均难以有效建模[18]，对场景应用具有一定的适应性困难。

7.2. 解决方案

层次聚类的潜在不足，可以通过：1) 多重随机初始化 + 结果集成、引入约束条件与使用概率聚类算法，提高稳定性；2) 基于贝叶斯信息准则(BIC)的自动选择和动态聚类框架的方案，降低簇数量过拟合或欠拟合。

LSTM 模型的潜在不足，可以：1) 用模型轻量化，如采用门控循环单元(GRU)替代 LSTM 和引入瓶

颈结构压缩特征维度、分布式训练和早停机制(Early Stopping)的方案应对训练效率低，计算资源消耗大的问题；2) 用事件嵌入(Event Embedding)、动态调整学习率和混合模型集成的方案应对突发外部事件的响应滞后的问题。

混合模型协同性的潜在不足，可以 1) 用端到端联合训练和注意力机制引入来应对聚类与 LSTM 的特征交互不足；2) 用特征降维预处理：通过主成分分析(PCA)、t-SNE 或自编码器(AE)将高维数据映射至低维空间，再进行聚类，提升距离度量的有效性；度量学习：采用马氏距离替代欧氏距离，来应对高维数据下聚类效果下降的问题。

模型依赖历史数据，对无历史数据的新数据资产估值困难的应用场景局限的问题，可以引入：小样本学习(Few-Shot Learning)：引入元学习(Meta-Learning)框架(如 MAML)，利用先验簇知识快速适应新簇，仅需少量样本即可初始化 LSTM 参数。迁移学习(Transfer Learning)：将相似领域(如电商用户数据迁移至新社交平台数据)的预训练聚类结果和 LSTM 参数作为初始化，通过微调适应新数据。

参考文献

[1] International Valuation Standards Council (IVSC) (2020) Valuation of Data and Intangible Assets. https://www.markables.net/international-valuation-standard-for-intangible-assets-ivs-210

[2] European Commission (2020) Guidance on the Regulation on a Framework for the Free Flow of Non-Personal Data in the European Union. https://ec.europa.eu/commission/presscorner/detail/en/ip_19_2749

[3] (2019) ISO/IEC 20546:2019. Information Technology-Big Data-Overview and Vocabulary. https://www.en-standard.eu/bs-iso-iec-20546-2019-information-technology-big-data-overview-and-vocabulary

[4] OECD (2019) Guidelines for Measuring the Value of Data as an Asset. https://www.oecd.org/en/publications/measuring-the-digital-transformation_9789264311992-en.html

[5] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[6] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Murtagh, F. and Legendre, P. (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, **31**, 274-295. https://doi.org/10.1007/s00357-014-9161-z

[8] Taleb, N.N. (2007) The Black Swan: The Impact of the Highly Improbable. Random House.

[9] Gers, F.A., Schmidhuber, J. and Cummins, F. (2000) Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, **12**, 2451-2471. https://doi.org/10.1162/089976600300015015

[10] Zhang, T., Ramakrishnan, R. and Livny, M. (1996) BIRCH: An Efficient Data Clustering Method for Very Large Databases. *ACM SIGMOD Record*, **25**, 103-114. https://doi.org/10.1145/235968.233324

[11] Zhang, G.P. (2003) Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing*, **50**, 159-175. https://doi.org/10.1016/s0925-2312(01)00702-0

[12] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[13] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning. 2nd Edition, Springer.

[14] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[15] Raschka, S. and Mirjalili, V. (2019) Python Machine Learning. 3rd Edition, Packt Publishing.

[16] Chollet, F. (2021) Deep Learning with Python. 2nd Edition, Manning Publications.

[17] Han, J., Kamber, M. and Pei, J. (2011) Data Mining: Concepts and Techniques. 3rd Edition, Morgan Kaufmann.

[18] Shiller, R.J. (2015) Irrational Exuberance. 3rd Edition, Princeton University Press.