Advances in Applied Mathematics 应用数学进展, 2025, 14(8), 76-85
Published Online August 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.148372

Hans汉斯

基于LSTM模型的股票分类与预测研究

王秉基

中国海洋大学海德学院，山东 青岛

收稿日期：2025年7月5日；录用日期：2025年7月28日；发布日期：2025年8月7日

摘　要

本文以中国A股上市银行为研究对象，构建了一个融合PCA降维、DTW相似度分析与层次聚类技术的金融评价框架，并结合LSTM模型对股票价格走势进行趋势预测。首先，基于5年财务与市场数据，利用主成分分析提取银行关键特征，并应用动态时间规整算法量化银行间时间序列相似度。随后，采用加权DTW距离与层次聚类相结合的方法将银行划分为五类，并为每类构建加权指数。在此基础上，引入长短期记忆网络(LSTM)对各类指数进行预测建模。实证结果显示，该方法在训练集上取得了R^2超过0.93的拟合度，测试集多数类别的R^2也稳定在0.80以上，部分类别预测精度接近0.89，验证了分类结构与模型构建的有效性。该研究为金融时序建模与行业板块化评价提供了新的思路与实证依据。

关键词

主成分分析，加权DTW距离，层次聚类，LSTM模型

Research on Stock Classification and Prediction Based on LSTM Model

Bingji Wang

Haide College, Ocean University of China, Qingdao Shandong

Received: Jul. 5th, 2025; accepted: Jul. 28th, 2025; published: Aug. 7th, 2025

Abstract

Taking Chinese A-share listed banks as the research object, this paper constructs a financial evaluation framework integrating PCA dimensionality reduction, DTW similarity analysis and hierarchical clustering, and combines it with an LSTM model for trend prediction of stock price movements. First, based on 5-year financial and market data, key bank characteristics are extracted using principal component analysis, and the dynamic time regularization algorithm is applied to quantify the time series similarity among banks. Subsequently, a combination of weighted DTW distance and hierarchical clustering is used to classify banks into five categories, and a weighted index

文章引用: 王秉基. 基于 LSTM 模型的股票分类与预测研究[J]. 应用数学进展, 2025, 14(8): 76-85.
DOI: 10.12677/aam.2025.148372

is constructed for each category. On this basis, the Long Short-Term Memory (LSTM) network is introduced to predictively model the indices for each category. The empirical results show that the method achieves a goodness of fit of R^2 over 0.93 on the training set, and most categories of the test set are also stabilized above 0.80, and the prediction accuracy of some categories is close to 0.89, which verifies the validity of the categorization structure and the model construction. This study provides new ideas and empirical basis for financial time series modeling and industry segmentation evaluation.

Keywords

Principal Component Analysis, Weighted DTW Distance, Hierarchical Clustering, LSTM Model

1. 引言

股票市场的趋势预测一直是金融领域的重要研究方向，但由于股票市场受到多种因素的影响，股票价格走势具有高度的不确定性与波动性，单一的预测模型难以获得稳定准确的预测效果[1] [2]。因此，为了降低预测的不确定性，近年来越来越多的研究聚焦于股票市场的分类问题。通过将走势相似的股票归为一类，可使每一类股票的趋势预测更加稳定有效[3] [4]。随着人工智能技术的迅速发展，深度学习方法，尤其是长短期记忆网络(LSTM)在处理时间序列数据方面表现出了卓越的能力，逐渐成为股票预测领域的重要工具[5] [6]。本文在现有研究基础上，进一步融合主成分分析(PCA)与动态时间规整(DTW)算法，探索股票数据的潜在特征和动态相似性，并结合层次聚类方法，为不同类型股票分别构建精确的 LSTM 预测模型，以期提高股票趋势预测的准确性。

2. 方法阐述

2.1. 主成分分析

主成分分析(Principal Component Analysis, PCA)是一种常用的线性降维算法，旨在通过正交变换将原始高维特征映射到一组新的低维变量上[7]。这些新变量称为主成分，它们是原始变量的线性组合，彼此正交，并按方差大小依次排列，从而保留数据中最主要的变异信息。设原始样本数据为 X ($n\times d$ 维)，其中 n 为样本数，d 为特征维度。PCA 的基本步骤如图 1。

2.2. 动态时间规整算法

动态时间规整(Dynamic Time Warping, DTW)是一种常用于衡量两条时间序列之间相似性的经典算法，主要解决序列长度不一致或存在时间偏移的问题。该算法通过对两个序列进行非线性配对，找出最优对齐路径，从而衡量整体相似度[8]。

设有两条时间序列 $X=\{x_1,x_2,\cdots,x_n\}$ 和 $Y=\{y_1,y_2,\cdots,y_n\}$，构造距离矩阵 $D_{m\times n}$，其中每个元素 $d(a,b)$ 表示 x_a 与 y_b 的欧氏距离。定义累计距离矩阵 $G_{m\times n}$，其递推公式为：

$$g(a,b)=\min\begin{cases}g(a-1,b)+d(a,b),\\ g(a-1,b-1)+2d(a,b),\\ g(a,b-1)+d(a,b)\end{cases} \tag{1}$$

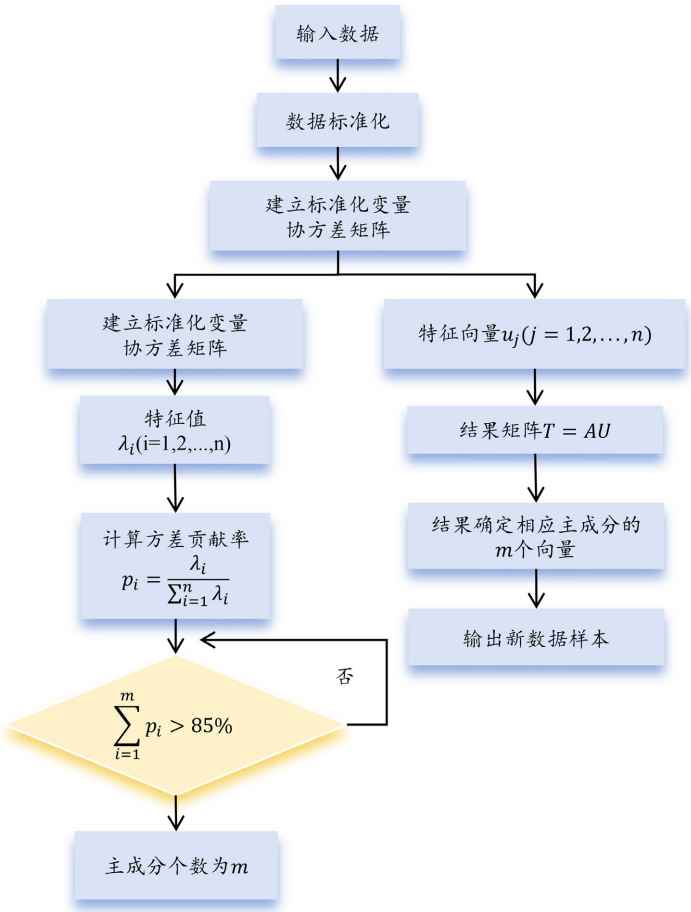


Figure 1. Flowchart of PCA algorithm
图 1. PCA 算法流程图

其中，$g(0,0)=g(0,b)=g(a,0)=0$，最终的 DTW 距离记作 $d_{\mathrm{DTW}}=g(m,n)$，表示从序列起点到终点的最短规整路径代价。DTW 的优点在于能捕捉序列之间局部变形或时间延迟带来的相似性，因此广泛应用于语音识别、金融数据分析等领域。在本文中，DTW 被用于识别不同个股之间历史价格走势的相似性，为后续模型引入参考序列提供依据。

2.3. 层次聚类算法

层次聚类(Hierarchical Clustering)是一类通过构建嵌套聚类结构实现数据划分的聚类方法，广泛应用于模式识别、文本分析及金融研究等领域[9]。该方法不需要预先设定簇的个数，而是通过一系列的合并或分裂操作，逐步建立样本间的层级关系，最终以树状图形式呈现聚类结构。设有样本集合 $X=\{x_1,x_2,\cdots,x_n\}$，初始时共有个 n 簇，记为 $C=\{C_1,C_2,\cdots,C_n\}$，聚类过程迭代执行以下步骤：

1) 计算簇之间的距离 $d(C_i,C_j)$；
2) 找出最小距离的簇对 (C_p,C_q)；
3) 合并 C_p 和 C_q 成新簇 C_{pq}；
4) 更新簇集合 C，重复步骤 1~3，直至满足终止条件。

最终聚类结构可表示为树状图，每一层表示不同的聚类粒度。可根据研究需求在树状图中选择截断高度，确定最终簇划分个数 k。

2.4. DTW-CH 算法

为了有效挖掘时间序列数据中潜在的相似结构与关联模式，本文引入动态时间规整与层次聚类相结合的融合方法，以构建基于时序相似性的样本分组机制，为后续分类预测任务提供结构化先验信息。

传统聚类方法在处理时间序列时通常依赖欧氏距离，无法应对序列间存在的时间错位、局部变形等非线性变化。而 DTW 能够通过弹性配对方式对齐两条时间序列，从而准确衡量其真实相似性，特别适用于金融、语音、生物序列等时序强相关数据场景。

具体方法如下：设样本集合为条长度可能不等的时间序列 $X=\left\{x^{(1)},x^{(2)},\cdots,x^{(n)}\right\}$，首先构建 DTW 距离矩阵 $D\in R^{n\times n}$，其中第 (i,j) 元素表示 $x^{(i)}$ 与 $x^{(j)}$ 间的 DTW 距离：

$$D_{ij}=\mathrm{DTW}\left(x^{(i)},x^{(j)}\right) \tag{2}$$

在获得全局距离矩阵后，基于该非线性距离量度，应用层次聚类算法，逐步构建时间序列间的层次聚类树状结构。簇间距离的计算可采用平均连接策略，即：

$$d\left(C_P,C_q\right)=\frac{1}{\left|C_P\right|\left|C_q\right|}\sum_{x\in C_P}\sum_{y\in C_q}D_{xy} \tag{3}$$

通过树状图(Dendrogram)可视化聚类结果，并根据预设的截断阈值(例如距离上限或类别数)完成聚类划分。该方法不仅能有效识别在整体走势或局部模式上相似的样本簇群，还能在不引入模型偏置的前提下提升预测模型的结构鲁棒性，算法流程如图 2。

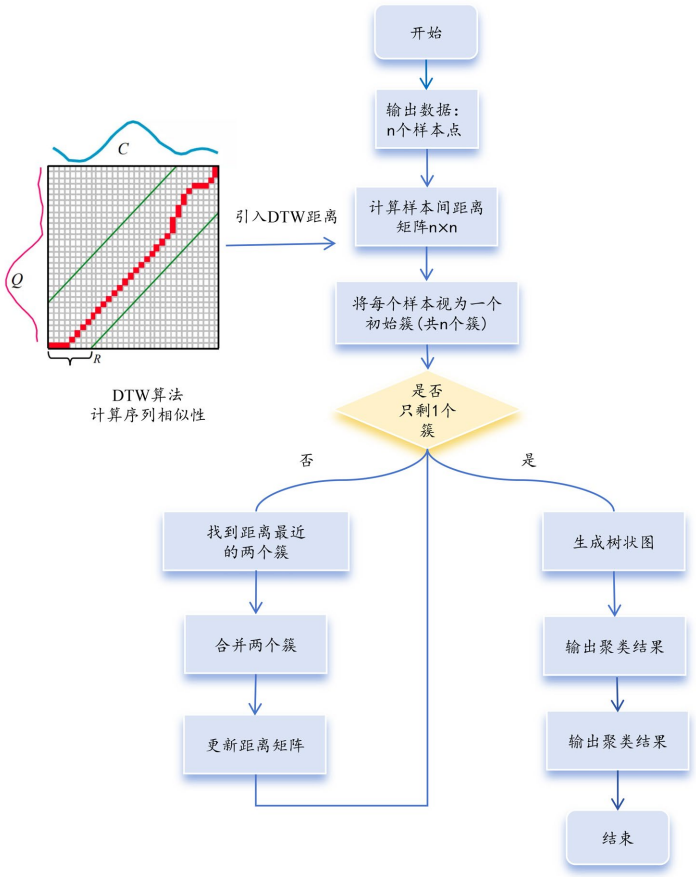


Figure 2. Flowchart of DTW-CH algorithm
图 2. DTW-CH 算法流程图

2.5. LSTM 算法

长短期记忆网络(Long Short-Term Memory，简称 LSTM)是一种专门为解决序列学习中信息长距离传递难题而设计的神经网络结构，其核心在于引入细胞状态 C_t——可视为贯穿网络的“记忆通道”，以及一系列门控机制，用于选择性地读取、写入或遗忘信息，从而实现对重要特征的持久保存与无关噪声的抑制[10]，LSTM 具体内容如下：

遗忘门决定保留上一步记忆中的哪些信息：

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{4}$$

其中，f_t 的每个元素在[0, 1]之间，数值越大表示越“保留”[10]。

输入门控制将多少新信息写入记忆，决定对即将生成的候选状态给予多大权重：

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{5}$$

候选细胞状态生成新的信息候选，这里的 tanh 保证候选信息在[−1, 1]范围内：

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{6}$$

细胞状态更新通过遗忘门和输入门的加权结合：

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{7}$$

输出门决定从更新后的记忆中提取哪些信息作为本步输出，控制最终输出的内容和强度：

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{8}$$

隐藏状态作为本时刻的最终输出，传递给下一时刻或后续网络层，同时也能作为额外的观测特征用于后续决策或预测：

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{9}$$

LSTM 的所有权重矩阵 W^* 和偏置向量 b^* 可以通过反向传播算法端到端学习，并常辅以梯度裁剪等技术来保证在较长序列上的训练稳定性[10]，LSTM 结构如图 3。

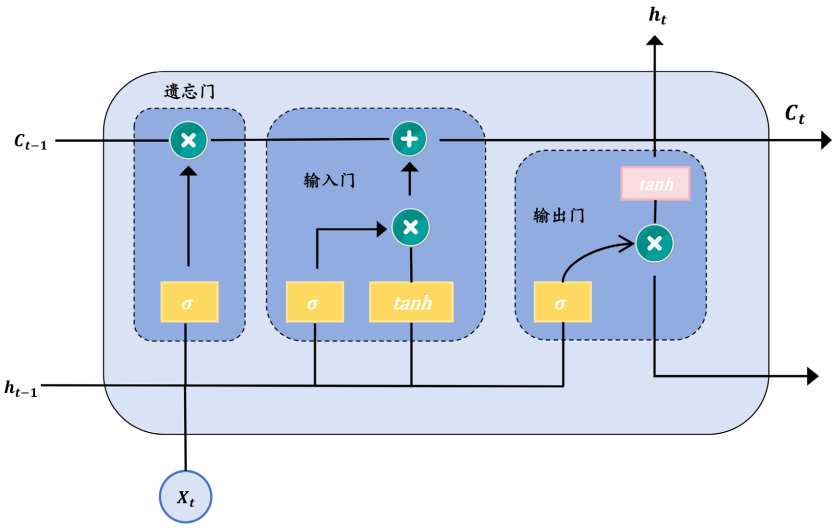


Figure 3. LSTM algorithm structure diagram
图 3. LSTM 算法结构图

3. 实证分析

3.1. 股票数据选取

为了具体展示上述方法在银行业金融评价模型中的应用，本文结合中国 A 股上市银行的数据进行了一个初步的实证分析。数据样本包括若干家 A 股上市银行自 2018 年 3 月到 2023 年 3 月的财务报表指标和股票市场表现。所采集数据包括各银行共计 5 年内的年报财务指标，涵盖资产规模、资本结构、盈利能力、风险控制与流动性等多个维度，变量总数数十项。

为提高数据完整性与分析的稳健性，本文首先对样本中存在缺失的股票价格数据进行预处理。针对个别缺失值，采用该变量在相应时间段内的样本均值进行填补，以减少由数据不完整带来的偏差。随后，为消除不同财务指标在量纲和尺度上的差异，统一变量的比较基础，本文对所有原始数据变量进行 Z-score 标准化处理，使得每个指标的样本均值为 0、标准差为 1，从而保证各变量在后续分析中具有可比性，避免高量纲变量对模型结果产生主导性影响。接着应用主成分分析(PCA)方法进行降维，保留累计解释方差达到 85%的前若干主成分，如图 4，最终将原始几十维指标压缩为 6 个代表性主成分。

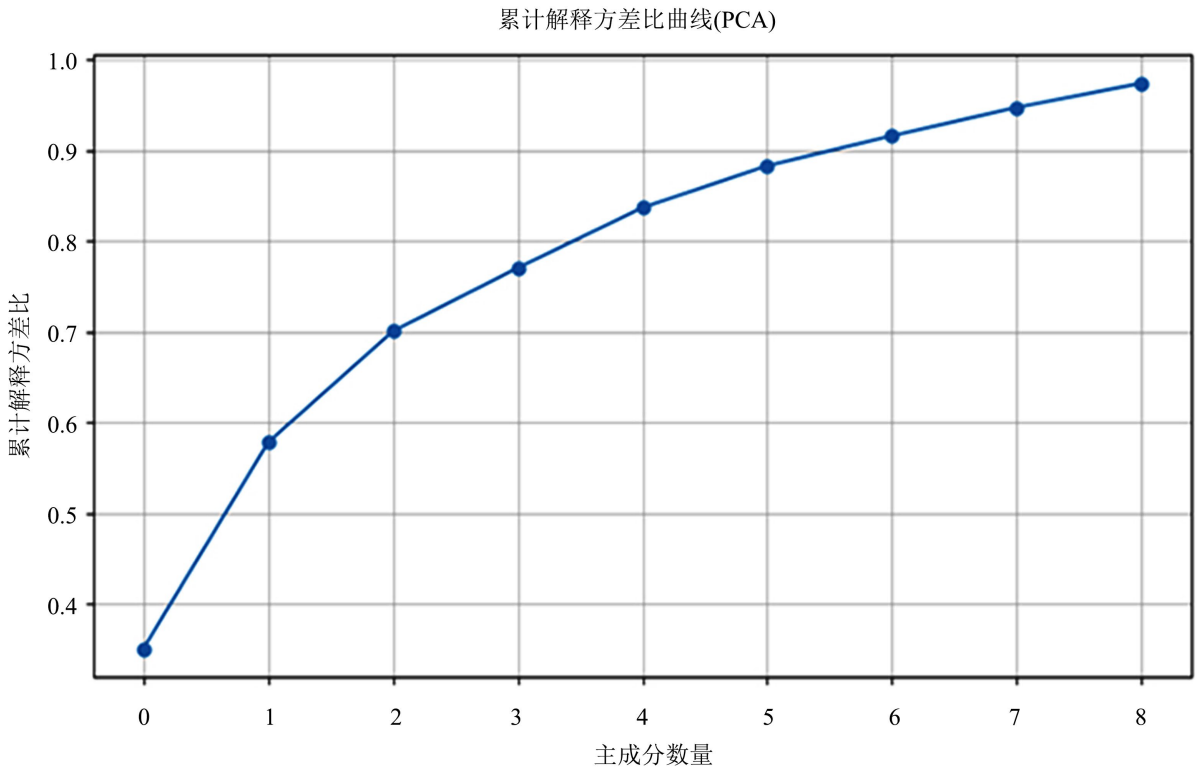


Figure 4. Cumulative variance curve
图 4. 累计方差曲线图

3.2. 相似度度量与聚类

基于降维后的主成分时间序列数据，我们计算了不同银行之间的相似度。在计算出任意两家银行之间的 DTW 距离后，我们进一步使用层次聚类方法对银行进行分类。在聚类过程中，不同类别的财务特征可以赋予不同权重，这里我们根据方差贡献度对上述主成分赋予了 0.35、0.23、0.12 等不同比重，从而构造综合加权距离。如表 1，结合 CH 准则，聚类结果将样本中的上市银行清晰地划分为五个类别。

Table 1. A-share stock clustering table
表 1. A 股股票聚类表

股票代码	种类	股票代码	种类	股票代码	种类
600000.SH	5	601187.SH	2	601838.SH	1
600015.SH	2	601229.SH	2	601860.SH	2
600016.SH	1	601288.SH	4	601916.SH	2
600036.SH	3	601328.SH	1	601939.SH	4
600919.SH	2	601398.SH	3	601963.SH	2
600926.SH	2	601577.SH	2	601988.SH	4
600928.SH	0	601665.SH	2	601997.SH	2
601166.SH	1	601818.SH	2	601998.SH	2
601169.SH	2	601825.SH	2	601128.SH	2
601077.SH	2	601009.SH	2		

每一类别代表了一组在财务特征演化上相似的银行(例如，一类银行可能以规模大、增长稳健为特征，另一类则可能资产规模较小但增长迅速)。如图 5 展示了聚类树状图，不同类别之间的距离较大，类内银行具有相近的发展轨迹。这种根据财务特征演变进行的分类为后续针对不同类型银行的分析奠定了基础。

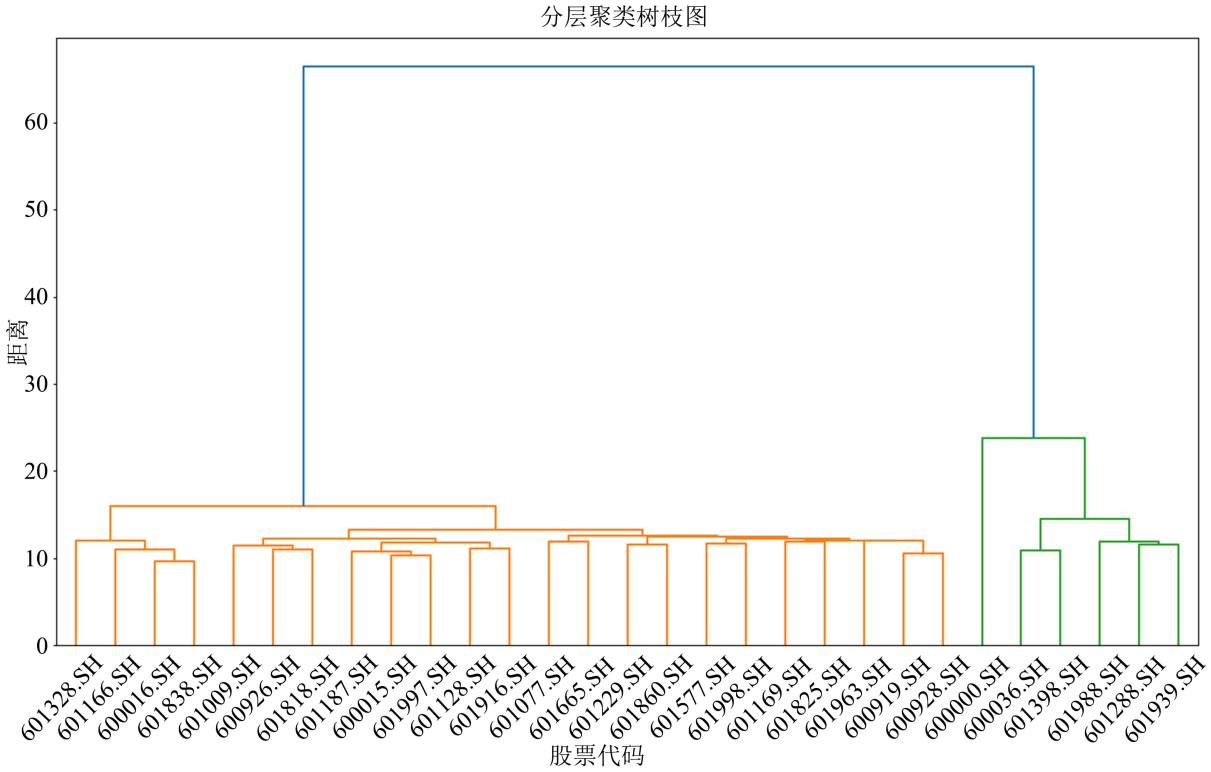
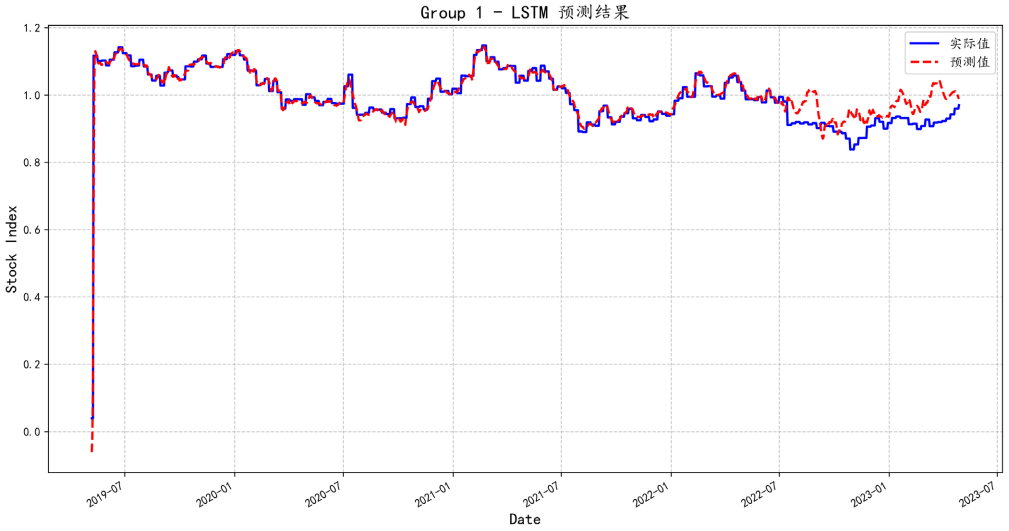


Figure 5. Stock clustering results
图 5. 股票聚类结果

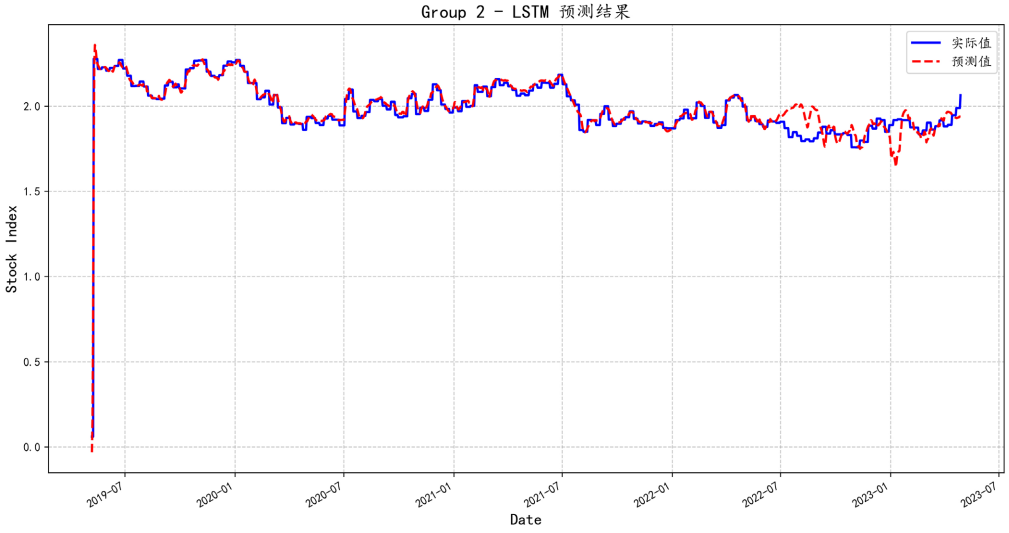
3.3. LSTM 分类预测结果与评价

完成聚类后，我们为了便于比较和预测，构建了每个类别的股票表现综合指数。考虑到不同银行市值规模差异较大，对指数的贡献应有所不同，我们采用流通市值为权重对类别内银行股价进行加权平均。我们将每只银行股每日收盘价乘以该银行的市值权重，再对同类别所有银行的加权收盘价求和，得到该类别在该日的综合指数值。这样得到的五条指数序列代表了五类银行板块的市场表现，既反映了类别内所有银行股价的整体走势，又考虑了市值大小使得龙头银行对指数的影响更大。

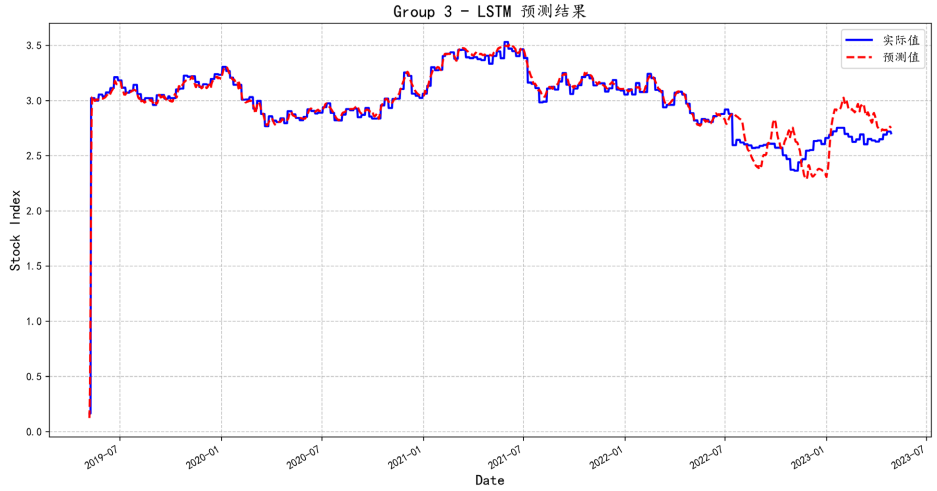
将这些数据序列作为 LSTM 模型的输入特征，用于预测各类银行板块的未来表现。LSTM 模型采用两层结构，每层 32 个神经元，并通过 Dropout 层(dropout rate = 0.3)有效防止过拟合。这里使用了 80 天历史数据预测未来一天的指数值。在模型训练中，数据被划分为训练集(80%)和测试集(20%)。为了提高模型的鲁棒性，输入数据经过 MinMaxScaler 标准化处理，并采用 Adam 优化器和均方误差损失函数进行模型优化。为评估模型效果，我们分别计算了均方误差(MSE)、平均绝对误差(MAE)和决定系数(R^2)等指标。分类预测结果如图 6。



(a)



(b)



(c)

(d)

(e)

Figure 6. Prediction results of LSTM algorithm
图 6. LSTM 算法预测结果

实验结果表明，LSTM 在训练阶段普遍表现出较高的拟合精度，训练集平均决定系数(R^2)大多超过0.93，说明模型能够有效学习时间序列中的非线性特征。在测试阶段，各组数据的预测效果存在一定差异，分别为 0.85、0.89、0.87、0.80、0.83。相较于已有的股票预测方法如基于单支股票的机器学习模型[11]，模型上升了 10.29%。

4. 总结

本文基于中国 A 股市场数据，提出了一种融合 PCA 降维、DTW 相似度计算与层次聚类分析的股票分类方法，并结合长短期记忆网络(LSTM)构建了面向各类股票的趋势预测模型。通过对高维财务与市场数据的压缩与结构化分类处理，模型在特征提取与样本划分层面有效提升了同质性，显著增强了 LSTM 在时间序列预测中的建模能力。实证结果表明，该方法在训练集上普遍获得较高的拟合度($R^2 > 0.93$)，测试集在五类股票板块中的预测 R^2 也稳定在 0.80 以上，部分组别甚至达到 0.89，也体现出该模型在趋势识别与短期预测方面的稳健性和有效性。

尽管如此，模型在部分类别中预测效果有限，表明仍需引入更丰富的数据(如宏观指标、舆情信息等)及更先进的机器学习模型，以提升模型泛化能力与解释力。综上，本文为银行业金融评价提供了一种融合机器学习的新思路，具有良好的发展前景和实践潜力。

参考文献

[1] Wang, J. and Wang, J. (2015) Forecasting Stock Market Indexes Using Principle Component Analysis and Stochastic Time Effective Neural Networks. *Neurocomputing*, **156**, 68-78. https://doi.org/10.1016/j.neucom.2014.12.084

[2] 郑磊, 李颖, 陈晓红. 基于深度学习的股票趋势预测研究综述[J]. 管理科学学报, 2017, 20(4): 87-104.

[3] Huang, W., Nakamori, Y. and Wang, S. (2005) Forecasting Stock Market Movement Direction with Support Vector Machine. *Computers & Operations Research*, **32**, 2513-2522. https://doi.org/10.1016/j.cor.2004.03.016

[4] 王铁山, 王敏, 张军辉. 基于聚类分析的股票市场分类预测模型研究[J]. 系统工程理论与实践, 2018, 38(7): 1790-1799.

[5] 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

[6] Chen, K., Zhou, Y. and Dai, F. (2015) A LSTM-Based Method for Stock Returns Prediction: A Case Study of China Stock Market. 2015 *IEEE International Conference on Big Data* (*Big Data*), Santa Clara, 29 October-1 November 2015, 2823-2824. https://doi.org/10.1109/bigdata.2015.7364089

[7] Jolliffe, I.T. and Cadima, J. (2016) Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A*: *Mathematical*, *Physical and Engineering Sciences*, **374**, Article ID: 20150202. https://doi.org/10.1098/rsta.2015.0202

[8] García-Escudero, L.A. and Gordaliza, A. (2019) Trimmed Fuzzy Clustering of Financial Time Series Based on Dynamic Time Warping. *Annals of Operations Research*, **299**, 1379-1395.

[9] Tang, G., Tian, R. and Wu, B. (2022) An Overview of Clustering Methods in the Financial World. *Proceedings of the* 2022 7*th International Conference on Financial Innovation and Economic Development* (*ICFIED* 2022), 14-16 January 2022, 524-529. https://doi.org/10.2991/aebmr.k.220307.084

[10] Das, N., Goswami, B. and Begum, R.N.A. (2023). Stock Prices Prediction Using Long Short Term Memory. 2023 4*th International Conference on Computing and Communication Systems* (*I3CS*), Shillong, 16-18 March 2023, 1-5. https://doi.org/10.1109/i3cs58314.2023.10127443

[11] Ding, Q., Wu, S., Sun, H., Guo, J. and Guo, J. (2020) Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Special *Track on AI in FinTech*, 4640-4646. https://doi.org/10.24963/ijcai.2020/640