基于股票预测模型LSTM的降维比较

马致远

(成都理工大学工程技术学院, 四川 乐山 614007)

摘要: 量化模型是投资者挑战股票动态预测的核心之一,原始LSTM(长短期记忆网络)股票预测模型由于输入的数据存在噪声,干扰了预测效果。针对影响股价的因子有259项指标,先用降维方法对输入数据进行降维,保留了关键信息,再输入LSTM,组成改进的预测模型,即PCA-LSTM(主成分分析-长短期记忆网络)模型、ISOMAP-LSTM(等距映射-长短期记忆网络)模型与PCA-ISOMAP-LSTM模型。通过实证对比,相较于原始LSTM预测模型和注意力机制模型(MHA-LSTM),PCA-LSTM模型与ISOMAP-LSTM模型减少了训练时间,预测误差评估指标中平均绝对误差(MAE),平均相对误差(MAPE),均方根误差(RMSE)都有显著降低,平均涨跌准确率(ARRF)有显著提高,但PCA-ISOMAP-LSTM模型误差率有所增长,准确率有一定降低。Diebold-Mariano检验也表明,PCA-LS TM模型、ISOMAP-LSTM模型股票预测能力都强于原始LSTM模型和MHA-LSTM,而PCA-ISOMAP-LSTM模型和MHA-LSTM模型均比原始LSTM模型预测能力弱,PCA-LSTM与ISOMAP-LSTM两种模型预测精度差异不显著,都可作为股票量化投资的一种新的技术支持。

关键词: 降维; 主成分分析; ISOMAP(等距映射); LSTM(长短期记忆网络); 股票预测; 注意力机制

中图分类号: TP301 **文献标志码:** A **文章编号:** 1671-1807(2025)11-0008-09

股票市场的复杂性源于其动态性和不确定性,受众多因素如宏观经济环境、公司业绩、政策变动等影响,使得动态预测股价极具挑战性[1]。随着人工智能和大数据技术的发展,投资者普遍希望找到准确率高的量化模型帮助投资,因此量化模型是投资者进行成功投资的核心之一。深度学习技术,尤其是长短期记忆网络(long short-term memory, LSTM)因其强大的时间序列建模能力,适合处理和预测时间序列中间隔和延迟相对较长的情况,这一技术特征与股票预测问题有着很高的契合度[2]。随着股票市场信息量的不断增加,股票行情越来越难预测,维度冗余可能导致预测模型效率低下和泛化能力下降,研究者趋向于用数据降维算法与LSTM模型的组合研究[3]。王东等[4]先利用主成分分析法(principal components analysis, PCA)对股票行情基础数据进行降维,再结合一些股票技术指标作为LSTM的输入变量,减少了LSTM模型的输入维度,实验结果表明降低训练时间的同时提高了预测精度;李辉等[5]通过随机森林(RF)选择最优特征集,降低数据维度和训练复杂度,再利用LSTM模型对股票进行预测,实验结果表明组合模型预测的误差有明显的降低;肖田田[6]构建了一种混合深度模型 K-means-LSTM来提高股票预测性能;范辉等[7]使用了多头注意力机制提取了股票特征,再由全连接层预测股票,实验表明注意力机制有一定的效果;朱瑞琪等[8]在LSTM中引入自注意力机制(self-attention)解决了其在处理长序列时的全局建模和多维向量间复杂交互关系不足的问题;曹帅等[9]在电力数据调度模型中引入Attention机制,提升网络的特征提取能力。周志轩等[10]采用压缩感知(CS)与LSTM相结合的模型,预测农产品价格趋势,预测精度较高。研究表明,LSTM神经网络对股票进行预测时,数据降维、注意力机制与压缩感知(CS)都是一种有效的数据预处理手段,可以降低模型复杂度的同时提升模型效果,去除噪声,同时保留关键信息,这对于提升模型的预测精度和稳定性具有重要意义。本文深入分析主成分分析法(PCA)与等距映射(ISOMAP)两种降维方法提取数据特征,与注意力机制提取数据特征对比,并通过实证研究来验证这些方法在股票预测中的实际效果,体现降维技术在优化LSTM模型在股票预测中的作用。

收稿日期: 2024-12-18

基金项目: 乐山市科技局项目(22GZD031)

作者简介: 马致远(1971—),男,四川宜宾人,硕士,讲师,研究方向为人工神经网络与数值计算。

在股票预测中,降维可以帮助缓解过拟合,提高模型的泛化能力,同时降低计算成本,使得LSTM 模型在大规模数据集上更加高效[11]。尽管降维在股票预测领域的潜力已经得到了初步探索,但大部分探索都是单一降维方法与 LSTM 模型的结合,对多种降维方法与 LSTM 模型相结合,探索在各种方法中研究其效果差异,从而选择最合理的降维方法方面研究还较少。本文选择线性降维方法中的主成分分析法,非线性降维方法中的ISOMAP,通过降维提取股票数据中的特征信息,用这些特征信息结合 LSTM 模型预测股票价格。

1 基于降维算法的 LSTM 模型

1.1 获取股票训练数据

通过 Python 环境下 tushare 财经数据接口包,获取了所选股票的 261 项行情数据:股票代码(ts_code)、交易日期(trade_date)、开盘价(open)、前复权开盘价(open_hfq)、后复权开盘价(open_qfq)、后复权最高价(high_hfq)、收盘价(close)、最高价(high)、最低价(low)、换手率(turnover_rate)、涨跌幅度 pct_chg、涨跌额 change、成交量 vol 以及成交金额 amount 等,见表 1。

LSTM 模型需要大量数据训练,模型预测才会更加精确。本文选择浦发银行 600000.SH 从 2020 年 5 月 6 日到 2023 年 5 月 4 日,每个开盘日共 700 个行情数据作为 LSTM 模型的建模训练数据集,选择其中的 259 项指标因子,见表 2。

所有指标分为 3 类:行情因素指标、基本面指标和技术指标,包括开盘价、收盘价、最高价、最低价、换手率、涨跌幅度、涨跌额、成交量、成交金额、市盈率、流通股本、多周期振动升降、真实波动、多空指标(BBI)、乖离率(BIAS)、布林线指标(BOLL)、情绪指标(BRAR)、顺势指标(CCI)、价格动量指标(CR)、平行线差指标、动向指标、连跌天数、连涨天数、区间震荡线、多周期指数移动平均、简单移动平均、随机指标(KDJ)、肯特纳交易通道、异同移动平均线指标(MACD)、梅斯线、股票技术指标(RSI)、动量指标、能量潮指标、投资者对股市涨跌产生心理波动的情绪指标、变动率指标、唐安奇通道(海龟)交易指标、三重指数平滑平均线、容量比率(VR)、威廉指标(W&R)、薛斯通道等,其中有些指标行当前指标,前复权指标,后复权指标,如收盘价包括当日实际收盘价,还有前复权收盘价,后复权收盘价,还有的指标有多个周期,如简单移动平均有 5 日、10 日、20 日、30 日、60 日、90 日、250 日多周期指标值,用这些指标的 T 日数据来预测 $T+1$ 日的收盘价。在经过训练的降维 LSTM 模型中,选取浦发银行 600000.SH 从 2023 年 5 月 6 日到 2024 年 4 月 6 日,共 220 日行情数据为测试数据。

1.2 利用主成分分析法将训练数据降维

主成分分析法(PCA)[11]通过正交变换将可能相关的变量转换为一组数值较少的线性不相关变量,即主成分,降维的步骤如下。

(1)将原始数据标准化,对股票数据 x_{ij} 进行标准化处理。

$$y_{ij}=\frac{x_{ij}-\overline{x}_j}{s_j},\ i=1,2,\cdots,n;\ j=1,2,\cdots,m \quad (1)$$

式中:$\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$;$s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2}$;$n$ 为股票行情数据总天数;m 每个行情数据的指标总数;y_{ij} 为标准化后的值。

(2)计算相关系数矩阵 $\boldsymbol{R}$。

$$\boldsymbol{R}=\frac{1}{n-1}\sum_{k=1}^{n}y_{ki}y_{kj},$$
$$i=1,2,\cdots,n;\ j=1,2,\cdots,m, \quad (2)$$

通过求解特征方程 $|\lambda\boldsymbol{E}-\boldsymbol{R}|=0$,$\boldsymbol{E}$ 为单位矩阵,得到特征值 $\lambda_i(i=1,2,\cdots,m)$,特征值 λ_i 为各个主成分的方差,$\lambda_i\Big/\sum_{k=1}^{m}\lambda_k$ 为第 i 个主成分的贡献率,$\sum_{k=1}^{i}\lambda_k\Big/\sum_{k=1}^{m}\lambda_k$ 为前 i 个主成分的累计贡献率,根据特征值都大于 1 且累计贡献率到一个较高的百分数以上(85%以上),可以保证主成分中包含原始数据的大部分信息。假设选取最大的前 k 个特征值

表 1 股票基础数据

序号	ts_code	trade_date	open	open_hfq	open_qfq	high	high_hfq	…
0	600000.SH	2023-05-04	7.48	111.184 9	6.903 71	7.75	115.198 33	…
1	600000.SH	2023-04-28	7.47	111.036 3	6.894 48	7.65	113.711 90	…
2	600000.SH	2023-04-27	7.48	111.184 9	6.903 71	7.51	111.630 89	…
3	600000.SH	2023-04-26	7.56	112.374 1	6.977 55	7.56	112.374 11	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
700	600000.SH	2020-05-06	10.44	132.729 9	8.241 49	10.49	133.366 10	…

表2　股票多维度因子

指标维度		指标名称	指标代号
行情因素指标		开盘价	open
		最高价	high
		最低价	low
		收盘价	close
		涨跌额	change
		涨跌幅	pct_chg
		成交量/手	vol
		成交额/千元	amount
		换手率/%	turnover_rate
		换手率(自由流通股)	turnover_rate_f
		量比	volume_ratio
基本面指标		市盈率(总市值/净利润，亏损的PE为空)	pe
		市盈率(TTM,亏损的PE为空)	pe_ttm
		市净率(总市值/净资产)	pb
		市销率	ps
		市销率(TTM)	ps_ttm
		股息率 /%	dv_ratio
		股息率(TTM)/%	dv_ttm
		总股本/万股	total_share
		流通股本/万股	float_share
		自由流通股本/万元	free_share
		总市值/万元	total_mv
		流通市值/万元	circ_mv
技术面指标	振动指标	振动升降指标	asi
		真实波动平均值	atr
	波动指标	简易波动指标	emv
		梅斯线	mass,ma_mass
		平行线差指标	dfma_difma
	趋向指标	连跌天数	downdays
		连涨天数	updays
		指数移动平均	ema_5,ema_10,ema_20,ema_30,ema_60,ema_90,ema_250
		EMA指数平均数指标	expma_12,expma_50
		三重指数平滑平均线	trix,trma
	反趋向指标	MACD指标	macd, macd_dea, macd_dif
		区间震荡线	dpo,madpo
		BIAS乖离率	bias1,bias2,bias3
		动量指标	mtm,mtmma
		RSI指标	rsi_6,rsi_12,rsi_24
		KDJ指标	kdj,kdj_d,kdj_k
	情绪指标	BRAR情绪指标	brar_ar,brar_br
		投资者对股市涨跌产生心理波动的情绪指标	psy,psyma
	压力支撑指标	肯特纳交易通道	ktn, ktn_mid, ktn_upper
		薛斯通道II	xsii_td1, xsii_td2, xsii_td3
	成交量指标	能量潮指标	obv
		VR容量比率	vr
		MFI指标	mfi

$\lambda_1,\lambda_2,\cdots,\lambda_k$，并求出其对应的特征向量为 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k$，其中 $\boldsymbol{\mu}_j=(\mu_{1j},\mu_{2j},\cdots,\mu_{mj})^{\mathrm{T}}$。

(3)计算主成分，并求得每个样本投影后作为新的训练数据。对每个样本 $\boldsymbol{X}^i$，原来的特征 $(x_1^i,x_2^i,\cdots,x_m^i)^{\mathrm{T}}$，投影之后的新特征为 $(y_1^i,y_2^i,\cdots,y_k^i)$，计算公式如下：

$$\begin{cases} y_1^i=\boldsymbol{\mu}_1^{\mathrm{T}}(x_1^i,x_2^i,\cdots,x_m^i)^{\mathrm{T}}= \\ \quad \mu_{11}x_1^i+\mu_{12}x_2^i+\cdots+\mu_{1m}x_m^i \\ y_2^i=\boldsymbol{\mu}_2^{\mathrm{T}}(x_1^i,x_2^i,\cdots,x_m^i)^{\mathrm{T}}= \\ \quad \mu_{21}x_1^i+\mu_{22}x_2^i+\cdots+\mu_{2m}x_m^i \\ \vdots \\ y_k^i=\boldsymbol{\mu}_k^{\mathrm{T}}(x_1^i,x_2^i,\cdots x_m^i)^{\mathrm{T}}= \\ \quad \mu_{k1}x_1^i+\mu_{k2}x_2^i+\cdots+\mu_{km}x_m^i, \\ i=1,2,\cdots,n \end{cases} \tag{3}$$

1.3　利用ISOMAP算法将训练数据降维

在选择的指标变量中，除了有些指标存在较强的线性相关性以外，还有一些指标存在某些未知的非线性关系，这些非线性关系表现为数据中的非线性结构和局部关系，通过非线性降维将数据映射到低维空间，能够更好地保留这些非线性结构和局部关系。等距映射Isomap是一种考虑数据几何结构的降维工具，通过保持数据点间的测地距离，有助于揭示股票数据中的潜在空间结构[12]。尽管计算成本较高，但对于理解股票价格的非线性关系和市场动态具有独特价值，降维步骤如下。

(1)对原始数据构建邻域图. 取高维空间 $\boldsymbol{R}^n$ 中的样本点 x_i 和 x_j，其两点之间的欧氏距离为 $d_x(i,j)$。若 x_i 和 x_j 是相邻点，则说明领域图有边，边长为 $d_x(i,j)$。

(2)计算最短路径矩阵 $\boldsymbol{D}_{\mathrm{G}}$，当 x_i 和 x_j 之间的邻域图有边时，$d_{\mathrm{G}}(i,j)=d_x(i,j)$；当 x_i 和 x_j 之间的邻域图没有边时，$d_{\mathrm{G}}(i,j)=\infty$；当 x_i 和 x_j 存在 N 个邻近点时，取 $d_{\mathrm{G}}(i,j)=\min\{d_{\mathrm{G}}(i,j),d_{\mathrm{G}}(i,k)+d_{\mathrm{G}}(k,j)\}$。

(3)构建低维映射，即找到原始高维数据集 $X=\{x_1,x_2,\cdots,x_N\}$ 相对应的 d 维数据集 $Y=\{y_1,y_2,\cdots,y_d\}(d\leqslant N)$，把最短路径矩阵 $\boldsymbol{D}_{\mathrm{G}}$ 作为MDS(multidimensional scaling，多维尺度)算法的输入，利用ISOMAP算法将训练数据降维返回的结果的集合即为 X 在低维空间的映射。

1.4　综合利用PCA与ISOMAP算法将训练数据降维

为了同时发现股票行情数据的决策指标之间

较强的线性相关性与非线性关系,采用二次降维法,结合两类降维方法,先分别用线性降维中的主成分分析法与非线性降维中的等距映射法,得到降维后的数据集,然后从其中各自取部分数据组成新的训练数据集,这样训练数据集就综合了原数据的两类特征信息。

1.5 利用 LSTM 模型进行预测

LSTM 网络通过对长期信息的学习,成功解决了普遍循环神经网络的缺陷,在时间序列预测等领域中能达到较好的应用效果[13]。LSTM 单元结构如图 1 所示。

LSTM 中有 3 个门控结构,分别是遗忘门、输入门和输出门,LSTM 通过这 3 个门控结构的组合来控制细胞状态,进而实现长期记忆保留的功能,解决梯度爆炸或消失问题。

遗忘门:决定了当前时刻单元状态 c_t 从上一刻状态舍弃多少信息,其数学表达式为

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{4}$$

输入门:更新当前时刻状态 c_t,即决定哪些输入信息 x_t 可以加入 c_t,其数学表达式为

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{5}$$

$$\tilde{c}_t = \tanh(W_c h_{t-1} + U_c x_t + b_c) \tag{6}$$

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

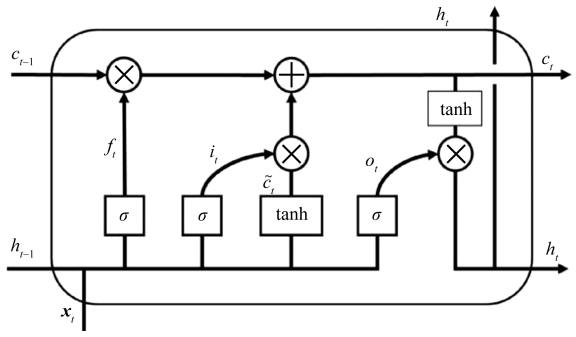
输出门 o_t:通过 c_t 函数激活前时刻单元状态,决定了最终的输出,

LSTM 在 t 时刻的输出 h_t 为

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o h_{t-1} + \boldsymbol{U}_o x_t + \boldsymbol{b}_o) \tag{8}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \tanh(c_{t-1}) \tag{9}$$

式中:$\boldsymbol{W}$、$\boldsymbol{U}$ 为权重矩阵;$\boldsymbol{b}$ 为偏置向量。在 LSTM 模型中,模型通过遗忘门忘记哪些内容,通过输入门选择保留哪些内容,以便模型分析那些与任务最相关的数据。LSTM 模型还可以学习数据的更加抽象表示,以便模型学习数据更多的特征[11]。这些特性使得 LSTM 模型应用于股票中时,可以更有效地分析股票走势。



$\boldsymbol{x}_t$ 为当前时间步的输入向量;h_{t-1} 为上一时间步的隐含层输出;c_{t-1} 为上一时间步的长期记忆;f_t 为遗忘门;i_t 为输入门;$\tilde{c}_t$ 为当前输入的新信息;c_t 为当前时间步的细胞状态;o_t 为输出门;h_t 为当前时间步的隐藏状态;σ 为 Sigmoid 激活函数,将值压缩到 [0,1]

图 1 LSTM 网络结构图

1.6 利用多头注意力机制提取股票数据特征

注意力机制应用于机器翻译领域,通过权重分配对关键信息的聚焦,从而提高模型的预测能力。在股票输入数据中,多头注意力机制能够评估输入特征对股票收盘价的影响力,并为各特征赋予相应权重,从而突出对股票价格影响显著的信息,提高模型的准确性。该机制通过不同的线性变换分别映射到查询(Query)、键(Key)和值(Value)空间,形成 3 个参数矩阵 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$,将每个矩阵拆分为 h 个头,得到 $\boldsymbol{Q}_i$、$\boldsymbol{K}_i$、$\boldsymbol{V}_i$,其计算公式为

$$\boldsymbol{A}_i = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_i \tag{10}$$

$$\boldsymbol{A}(x) = \mathrm{Concat}(\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_h)W^0 \tag{11}$$

式中:$i = 1, 2, \cdots, h$;$\boldsymbol{A}_i$ 为第 i 个头的注意力输出;d_k 为键向量的维度;softmax 用于产生注意力权重;$\boldsymbol{A}$ 为 h 个头的输出注意力相加;$\boldsymbol{W}^0$ 为权重向量。

2 实例分析

2.1 选取样本与指标数据

在原始 LSTM 模型中,取建模训练数据集 700 个样本,每个样本有 259 个决策指标,即构成700×259 决策参数矩阵,作为 LSTM 模型的输入数据,以第 2 天的收盘价为目标数据,组成 LSTM 模型。

在选择的指标变量中,有些存在较强的线性相关性,对模型的预测泛化能力有干扰影响,用主成分分析法来提取几个独立的主成分,用较少的主成分来作为 LSTM 模型的输入来训练模型,既保留了大部分有用信息,又降低了模型的复杂度与训练难度。由式(2)先计算出原始数据的相关系数矩阵 $\boldsymbol{R}$ 再求 $\boldsymbol{R}$ 的特征值与方差贡献率,见表 3,根据方差贡献率从大到小排序,根据主成分选择规则[4],前 17 个主成分的特征值均大于 1,且累计方差贡献率达到了 96%,故取前 17 个主成分。

主成分方差累积贡献率如图 2 所示,从图 2 上可以看出,在主成分数达到 7 以后,方差累积贡献率增加速度变缓。

经过计算得到前 17 个主成分的载荷矩阵,见表 4。

表 3　特征值与贡献率

成分	初始特征值			提取载荷平方和		
	合计	方差百分比/%	累积百分率/%	合计	方差百分比/%	累积百分率/%
1	113.591	44.028	44.028	113.591	44.028	44.028
2	67.223	26.055	70.083	67.223	26.055	70.083
3	21.069	8.166	78.249	21.069	8.166	78.249
4	10.267	3.980	82.229	10.267	3.980	82.229
5	7.324	2.839	85.068	7.324	2.839	85.068
6	5.608	2.174	87.241	5.608	2.174	87.241
7	4.013	1.555	88.797	4.013	1.555	88.797
8	3.312	1.284	90.080	3.312	1.284	90.080
9	3.206	1.243	91.323	3.206	1.243	91.323
10	2.803	1.087	92.409	2.803	1.087	92.409
11	2.162	0.838	93.248	2.162	0.838	93.248
12	1.565	0.607	93.854	1.565	0.607	93.854
13	1.347	0.522	94.376	1.347	0.522	94.376
14	1.241	0.481	94.858	1.241	0.481	94.858
15	1.090	0.422	95.280	1.090	0.422	95.280
16	1.074	0.416	95.696	1.074	0.416	95.696
17	1.035	0.401	96.097	1.035	0.401	96.097
18	0.875	0.339	96.437	—	—	—
19	0.779	0.302	96.739	—	—	—
20	0.753	0.292	97.030	—	—	—
⋮	⋮	⋮	⋮	⋮	⋮	⋮

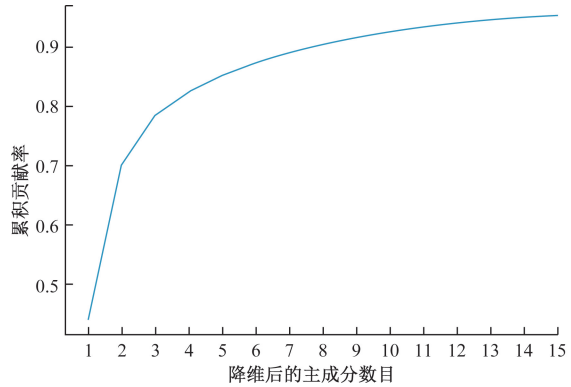


图 2　主成分方差累积贡献率

表 4　主成分载荷

指标	主成分 1	主成分 2	主成分 3	…	主成分 17
open	0.093	0.013	−0.005	…	−0.013
open_hfq	0.093	0.015	0.000	…	0.004
open_qfq	0.093	0.015	0.000	…	0.004
high	0.093	0.015	−0.008	…	−0.014
high_hfq	0.092	0.018	−0.004	…	0.002
high_qfq	0.092	0.018	−0.004	…	0.002
low	0.093	0.014	−0.009	…	−0.013
low_hfq	0.092	0.016	−0.005	…	0.004
low_qfq	0.092	0.016	−0.005	…	0.004
close_hfq	0.092	0.018	−0.009	…	0.007
close_qfq	0.092	0.018	−0.009	…	0.007
pre_close	0.093	0.013	−0.006	…	−0.015
change	−0.005	0.030	−0.081	…	0.051
pct_chg	−0.004	0.031	−0.083	…	0.047
⋮	⋮	⋮	⋮	⋮	⋮

再利用式(3)得到 17 个主成分的得分值，作为 LSTM 模型的训练数据，组成 PCA-LSTM 模型。

利用等距映射 Isomap 对数据进行非线性降维。为了对比试验，取与主成分分析一致的降维维度，即 17 维。取建模训练数据集 700 个样本，每个样本有 259 个决策指标，即构成 700×259 决策参数矩阵，经过等距映射 ISOMAP，得到降维后的数据，形成 700×17 决策参数矩阵，作为 LSTM 模型的输入数据，以第 2 天的收盘价为目标数据，组成 ISOMAP-LSTM 模型。

经过主成分分析降维得到 700×17 决策参数矩阵，作为新的样本数据再经过等距映射 ISOMAP 降维，得到新的 700×4 决策参数矩阵，作为 LSTM 模型的输入数据，以第 2 天的收盘价为目标数据，组成 PCA-ISOMAP-LSTM 模型。

利用多头注意力机制(MAH)，由式(11)把建模训练数据集 700×259 矩阵产生 700×16 决策参数矩阵，作为 LSTM 模型的输入数据，以第 2 天的收盘价为目标数据，组成 MHA-LSTM 模型。

2.2　确定 LSTM 模型参数

原始 LSTM 模型的输入值为表 1 中每日 259 个指标数据，第 2 日的收盘价为输出值，模型参数设定为：输入层维数为 259，输出层维数为 1，由隐含层神经元个数的经验公式，经过多次测试，隐含层神经元个数取为 20。

$$h = \sqrt{m+N}+K \tag{12}$$

式中：m 为输入层神经元个数；N 为输出层神经元个数；K 为 1～10 的常数。学习率为 0.000 6，步长取为 28。

将训练集的每个交易日的相关行情数据作为主成分分析数据，得到载荷矩阵，由载荷矩阵计算出每日的每个主成分得分，以这 700 个降维后的数据作为 LSTM 的输入，第 2 日的收盘价为输出值，通过配对数据来训练模型。LSTM 模型参数设定为：输入层维度为 17，输出层维度为 1，隐含层层数为 2，经过反复测试，确定隐含层神经元个数为 10 效果较好。

PCA+ISOMAP+LSTM 模型中，输入层维度为 17，输出层维度为 1，隐含层层数为 2，确定隐含层神经元个数为 20 效果较好，700 组训练数据前 400 组数据为主成分分析降维数据，后 300 组数据为 Isomap 降维数据。

利用网格搜索法与试验，确定上述 4 种模型的训练参数设置，见表 5。

表 5　4 种模型的训练参数设置

模型	参数设置
LSTM	optimizer=adam，dropout=0.1，epochs=200，batch_size=60 units=10，time_step=12，num_layers=2
PCA-LSTM	optimizer=adam，dropout=0.05，epochs=150，batch_size=80，units=80，time_step=12，num_layers=2
ISOMAP-LSTM	optimizer=adam，dropout=0.08，epochs=150，batch_size=80，units=80，time_step=12，num_layers=2
PCA-ISOMAP-LSTM	optimizer=adam，dropout=0.05，epochs=400，batch_size=16 units=20，time_step=12，num_layers=2

2.3　特征提取效果评估指标

以 220 日行情数据为测试数据，通过比较模型预测与实际值差异，评估模型特征提取效果。所选误差评估指标包括平均绝对误差(MAE)，平均相对误差(MAPE)，均方根误差(RMSE)，涨跌准确评估指标用平均涨跌准确率(ARRF)，如果预测收盘价差与实际收盘价差符号相同，表明预测方向与实际方向相同，即预测方向准确，否则方向相反，预测方向错误。其中 MAE、RMSE 越小，ARRF 越大，表示预测效果越好，评估指标的公式见表 6。

表 6　评估指标公式对照

评估指标	评估指标公式
平均绝对误差	$\mathrm{MAE}=\frac{1}{l}\sum_{j=1}^{l}\lvert y_j-\hat{y}_j\rvert$
平均相对误差	$\mathrm{MAPE}=\frac{1}{l}\sum_{j=1}^{l}\left\lvert\frac{y_j-\hat{y}_j}{y_j}\right\rvert$
均方根误差	$\mathrm{RMSE}=\sqrt{\frac{1}{l}\sum_{j=1}^{l}(y_j-\hat{y}_j)^2}$
平均涨跌准确率	$\mathrm{ARRF}=\frac{1}{m}\sum_{i=1}^{m}a_i$，其中 $a_i=\begin{cases}1, & (\hat{y}_{i+1}-\hat{y}_i)(y_{i+1}-y_i)\geqslant 0\\ 0, & \text{其他}\end{cases}$

注：y_j 为第 j 个指标的实际值；$\hat{y}_j$ 为第 j 个指标的预测值；l 为测试集的数据组数。

此外，采用 Diebold-Mariano(DM)检验平均绝对偏差(MAE)，该检验的原假设和备择假设如下。

H0：$D=0$，即两种模型的预测效果是等价的；

H1：$D\neq 0$，即两种模型的预测效果是存在差异的。

DM 检验假定股票时间序列为 y_t，$t=1,2,\cdots,T$，两个模型的预测值分别为 $\hat{y}_{1t}$ 和 $\hat{y}_{2t}$，则两个模型的预测误差分别为 $\varepsilon_{1t}=y_t-\hat{y}_{1t}$ 和 $\varepsilon_{2t}=y_t-\hat{y}_{2t}$，对应的损失函数为 $l_1=g(\varepsilon_{1t})$ 和 $l_2=g(\varepsilon_{2t})$，两个模型的损失函数之差：

$$d_t=g(\varepsilon_{1t})-g(\varepsilon_{2t}) \tag{13}$$

DM 统计量：

$$d_{\mathrm{mean}}=\frac{1}{T}\sum_{t=1}^{T}\lvert g(\varepsilon_{1t})-g(\varepsilon_{2t})\rvert \tag{14}$$

$$d_{\mathrm{std}}=\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\lvert d_t-d_{\mathrm{mean}}\rvert} \tag{15}$$

$$\mathrm{DM}=\frac{d_{\mathrm{mean}}}{d_{\mathrm{std}}} \tag{16}$$

在 P 小于显著性水平的情况下，若 DM<0，认为模型 1 的预测效果好于模型 2；若 DM>0，则认为模型 2 的预测效果优于模型 1。

2.4　对比实验与分析

将原始 LSTM、基于主成分的 PCA-LSTM 模型、基于等距映射的 ISOMAP-LSTM 模型、综合主成分与等距映射的 PCA-ISOMAP-LSTM 模型分别训练、基于多头注意力机制的 MHA-LSTM，为防止过拟合，LSTM 模型的 keep_prob 取 0.8，记录每个模型的模型训练时间，用测试集数据，计算模型预测的平均绝对误差、平均相对误差、均方根误差以及涨跌准确率，模型的性能见表 7。

从表 7 可以看出，在保持隐含层神经元不变的情况下，将 700×259 决策参数矩阵降维为 700×17 决策参数矩阵，相对原始 LSTM 模型，改进的模型改变了网络结构，输入维数由 259 变为 17，大幅降低了输入变量的维数，减少了预测误差与网络训练时间，提高了预测精度。PCA-LSTM、ISOMAP-LSTM 与 PCA-ISOMAP-LSTM 模型比 LSTM 模型

表 7　5 种 LSTM 模型的性能

模型	网络结构	训练时间/s	MAE	MAPE	RMSE	ARRF
LSTM	258-20-1	209.00	0.131 0	0.018 2	0.170 6	0.637 2
MHA-LSTM	16-64-1	69.00	0.143 1	0.020 5	0.183 0	0.720 9
PCA-LSTM	17-20-1	87.38	0.081 3	0.011 0	0.098 6	0.748 8
ISOMAP-LSTM	17-20-1	90.00	0.128 4	0.018 4	0.161 5	0.730 2
PCA-ISOMAP-LSTM	17-20-1	119.00	0.151 6	0.021 6	0.191 3	0.683 7

网络训练时间，由原来模型的 209 s 分别减少到 88、90、119 s，表明降维后模型更加简洁，可以实现更快地训练。除了模型训练时间的大幅减少外，PCA-LSTM 与 ISOMAP-LSTM 模型的预测误差显著降低，平均涨跌准确率 ARRFPCA-LSTM 模型提升了 6%，ISOMAP-LSTM 模型提升了 9%，表明降维后改进的这两个模型预测涨跌方向效果明显，这是因为 PCA-LSTM 去除了一些指标数据的相关性再输入 LSTM 模型，相对原始 LSTM 模型减少了数据的噪声干扰，并使低维数据最大程度保持原始高维数据的方差信息；而 ISOMAP-LSTM 模型通过保持高维数据的测地距离与低维数据的欧式距离的不变性来找到低维特征表示，低维特征输入使 LSTM 模型更容易学习股票数据的特征，从而提高模型的预测精度。虽然模型 PCA-ISOMAP-LSTM 的 ARRF 提升了 5%，但误差评估指标也显著增加了，表明该模型预测效果不佳，这是因为虽然 PCA 去除了数据的相关性，Isomap 通常在处理非线性数据集时表现良好，所以预测时大的趋势准确性高，但是经过两次降维去掉了绝大部分噪声，减少了随机性，因此预测误差也随着增大。MHA-LSTM 模型效果不如原始 LSTM 模型，说明多头注意力机制在提取股票数据特征效果不理想，可能是多头注意力机制提取的信息过少，输入 LSTM 模型的股票影响信息不完整，造成预测效果较差。

图 3～图 7 为 5 种模型的预测曲线，预测曲线与真实值曲线基本拟合，能够预测股票较长时间的走势。对比图 3～图 7 可以看出，在长期的底部阶段或顶部阶段，由于方向不明，可能存在一些人为因素，改进的模型预测效果不及原始模型；但当确定方向后，PCA-LSTM 与 ISOMAP-LSTM 曲线拟合程度很高，两者的相似性很强。

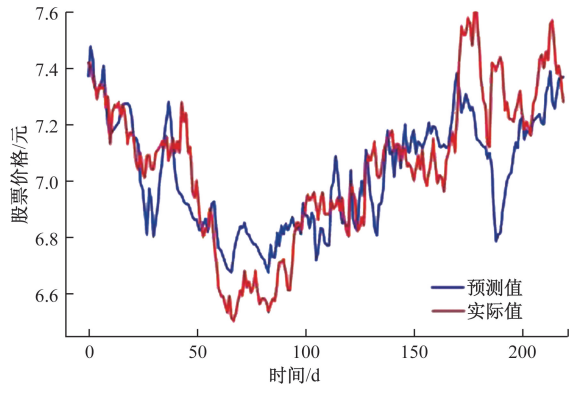


图 3　原始 LSTM 模型预测曲线

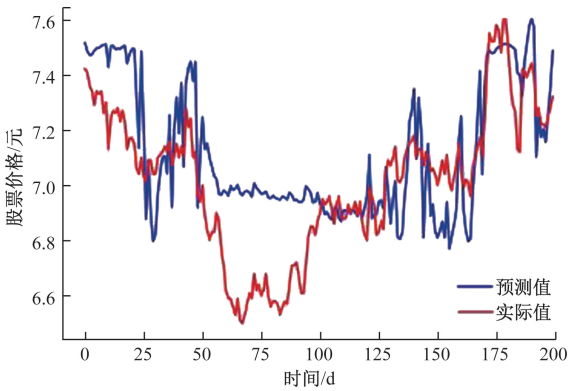


图 4　MAH-LSTM 模型预测曲线

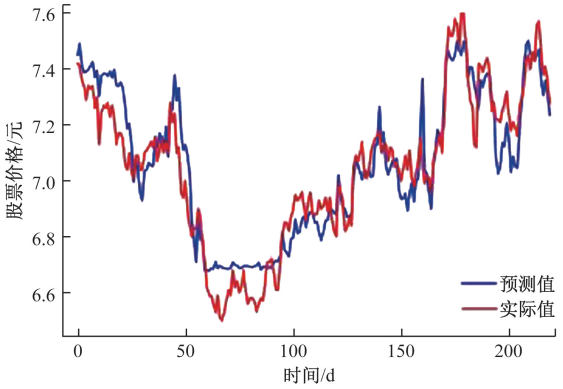


图 5　PCA-LSTM 模型预测曲线图

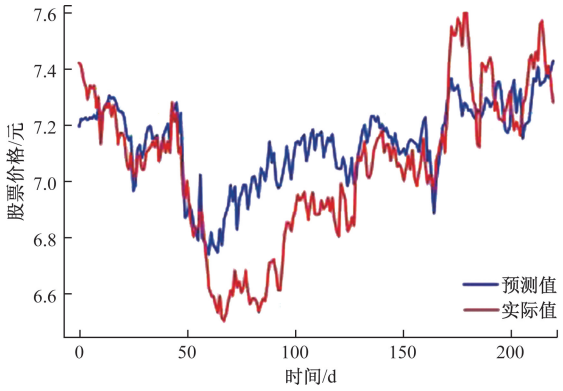


图 6　ISOMAP-LSTM 模型预测曲线

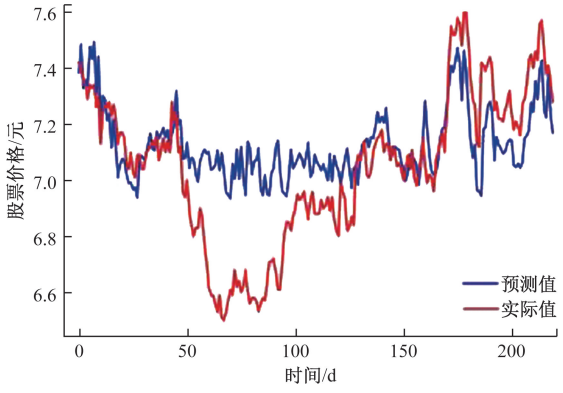


图 7　PCA-ISOMAP-LSTM 模型预测曲线

在模型预测的基础上,DM 检验比较以上几种预测模型的效果优劣。选取 4 个预测模型两两比较,共 12 中情形,对平均绝对误差 MAE 进行检验,由式(13)~式(16)计算,结果见表 8。

表 8　MAE 的 DM 检验结果

模型 1	模型 2	DM 统计量	P
原始 LSTM	PCA-LSTM	5.174 15	9×10^{-7}
原始 LSTM	ISOMAP-LSTM	5.883 27	4×10^{-6}
原始 LSTM	PCA-ISOMAP-LSTM	−4.378 00	3×10^{-5}
原始 LSTM	MAH-LSTM	−6.861 00	4×10^{-10}
PCA-LSTM	ISOMAP-LSTM	1.561 70	0.121
PCA-LSTM	PCA-ISOMAP-LSTM	−7.313 00	3×10^{-11}
PCA-LSTM	MAH-LSTM	−9.254 20	1×10^{-13}
ISOMAP-LSTM	MAH-LSTM	−10.917 00	2×10^{-13}
ISOMAP-LSTM	PCA-ISOMAP-LSTM	−8.800 00	1.3×10^{-13}

DM 检验的原假设是两个模型的预测效果相同,如果 P 小于 0.05 的显著性水平,则拒绝原假设,说明不同的模型预测效果显著不同,再结合 DM 统计量取值的正负,可判断哪个模型的预测效果更优。在表 8 中,以原始 LSTM 模型为基准,P 都小于 0.05,DM 统计量为正的有 PCA-LSTM 和 ISOMAP-LSTM,DM 统计量为负的是 PCA-ISOMAP-LSTM 和 MAH-LSTM,表明 PCA-LSTM 和 ISOMAP-LSTM 模型预测性能比原始 LSTM 模型更优,即经过降维提取到了影响股价的关键信息,而 PCA-ISOMAP-LSTM 模型比原始 LSTM 模型更差,表明两种降维方法融合效果不理想,因为 PCA 提取线性关系信息,而 ISOMAP 提取空间位置信息,两种信息提取时造成信息损失,训练时造成网络参数无法收敛。PCA-LSTM 与 ISOMAP-LSTM 比较,P 大于 0.05,两种模型预测性能无法比较优劣,表明两种降维方法都能提取数据的关键信息,预测效果相差不大;PCA-LSTM、ISOMAP-LSTM 与 MAH-LSTM 比较,DM 统计量都为负,表明 MAH-LSTM 模型预测性能比前两者都差,多头注意力机制提取特征信息不如主成分分析法与等距映射法。同理,PCA-ISOMAP-LSTM 也比前两者性能较差。为了进一步通过 DM 检验比较模型的预测效果,可以增加之前多年的训练数据样本和检验样本,或者在实际数据加入噪声增加一定的随机性,未来将在这方面进行研究。

3　结论

在大量指标数据的基础上,研究了降维后的 LSTM 模型预测股票价格问题。首先收集大量的行情数据,利用主成分分析法提取出决策数据的主成分,计算样本的得分值为新的样本数据集,提取数据内部的线性关系;等距映射法把高维空间数据映射到低维空间后,用低维空间的映射作为样本数据集,提取数据内部的非线性关系与结构。通过降维处理,消除了输入特征的线性关系和一些噪声干扰,也简化了网络模型的结构。实验中分别构建线性降维模型、非线性降维模型、线性与非线性综合降维模型、注意力机制提取特征模型,通过仿真与 DM 检验表明,线性降维的 PCA-LSTM 模型和非线性降维的 ISOMAP-LSTM 模型预测精度都有所显著提高,但线性与非线性综合降维模型 PCA-ISOMAP-LSTM 与 MAH-LSTM 预测性能显著降低,表明综合两种降维方法和注意力机制模型,预测效果不佳。由此可见,PCA-LSTM 与 ISOMAP-LSTM 这两种模型为量化投资提供了一种新的技术支持和实践经验。

参考文献

[1] GANDHMAL D P, KUMAR K. Systematic analysis and review of stock market prediction techniques[J]. Computer Science Review, 2019, 34: 100190.

[2] 包振山, 郭俊南, 谢源, 等. 基于 LSTM-GA 的股票价格涨跌预测模型[J]. 计算机科学, 2020, 47(S1): 467-473.

[3] BAEK Y J, KIM H Y. Modaugnet: a new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module[J]. Expert Systems with Applications, 2018, 113: 457-480.

[4] 王东, 王霄鹏, 杨川东. 一种基于主成分 LSTM 模型在股票预测中的研究[J]. 重庆理工大学学报(自然科学), 2021, 35(2): 282-288.

[5] 李辉, 化金金, 邹波蓉. 基于 RF-LSTM 组合模型的股票价格预测[J]. 河南理工大学学报(自然科学版), 2022, 41(1): 136-142.

[6] 肖田田. 基于 K-means-LSTM 模型的证券股价预测[J]. 科技和产业, 2024, 24(3): 210-215.

[7] 范辉, 朱勇丞, 李晋江. 基于注意力机制和特征融合的股票预测方法[J]. 山东工商学院学报, 2024, 38(1): 57-68, 76.

[8] 朱瑞琪, 陆佳敏, 陆佳艳, 等. 一种基于自注意力机制的 CNN-BiLSTM 非侵入式负荷分解方法研究[J]. 机电信息, 2023(16): 77-81.

[9] 曹帅, 李晓君, 贺成铭, 等. 基于改进 LSTM 的电力调度数据预测模型设计与仿真[J]. 电子设计工程, 2024, 32(19): 173-177.

[10] 周志轩, 陈仲民, 邓君丽. 基于压缩感知和深度学习的农产品价格预测[J]. 武汉大学学报(工学版), 2024, 57(9): 1327-1334.

[11] HTUN H H, BIEHL M, PETKOV N. Survey of feature selection and extraction techniques for stock market prediction[J]. Financial Innovation, 2023, 9(1): 26.
[12] TENENBAUM J B, SILVA V D, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290: 2319-2323.
[13] MALHOTRA P, VIG L, SHROFF G, et al. Long short term memory networks for anomaly detection in time series[J]. Computational Intelligence and Machine Learning, 2015, 24(3): 210-215.

Dimensionality Reduction Comparison Based on Stock Prediction Model LSTM

MA Zhiyuan
(The Engineering & Technical College of Chengdu University of Technology, Leshan 614007, Sichuan, China)

Abstract: Quantitative models are one of the core challenges for investors in stock dynamic prediction. The original LSTM(long short-term memory) stock prediction model was affected by noise in the input data, which interfered with the prediction effect. In this paper, there are 259 indicators that affect stock prices. Firstly, the input data was reduced in dimensionality using dimensionality reduction methods to preserve key information, and then input into LSTM to form an improved prediction model, namely PCA-LSTM model, ISOMAP-LSTM model, and PCA-ISOMAP-LSTM model. Through empirical comparison, compared with the original LSTM prediction model and the attention mechanism model MHA-LSTM, the PCA-LSTM model and ISOMAP-LSTM model reduce training time. The average absolute error (MAE), average relative error (MAPE), and root mean square error (RMSE) in the prediction error evaluation indicators are significantly reduced, and the average rise and fall accuracy (ARRF) is significantly improved. However, the PCA-ISOMAP-LSTM model has an increase in error rate and a certain decrease in accuracy. The Diebold Mariano test also showed that the PCA-LSTM model and ISOMAP-LSTM model have stronger stock prediction abilities than the original LSTM model and MHA-LSTM model, while the PCA-ISOMAP-LSTM model and MHA-LSTM model have weaker prediction abilities than the original LSTM model. The difference in prediction accuracy between the PCA-LSTM and ISOMAP-LSTM models is not significant, and both can be used as a new technical support for quantitative stock investment.

Keywords: dimensionality reduction; principal component analysis; ISOMAP(isometric mapping); LSTM(long short-term memory); stock prediction; attention mechanism