基于 *K*-means-LSTM 模型的证券股价预测

肖田田

（安徽建筑大学数理学院，合肥 230601）

摘要： 鉴于股票数据具有非平稳、非线性等特征，传统的统计模型无法精准预测股票价格的未来趋势。针对这个问题，构建一种混合深度学习方法来提高股票预测性能。首先，通过将距离算法修改为 DTW（动态时间归整），令 *K*-means聚类算法拓展为更适用于时间序列数据的 *K*-means-DTW，聚类出价格趋势相似的证券；然后，通过聚类数据来训练 LSTM（长短时记忆网络）模型，以实现对单支股票价格的预测。实验结果表明，混合模型 *K*-means-LSTM 表现出更好的预测性能，其预测精度和稳定性均优于单一 LSTM 模型。

关键词： 股票价格预测；*K*-means；DTW（动态时间归整）；*K*-means-LSTM（*K* 均值-长短时记忆网络）混合模型

中图分类号： F830 **文献标志码：** A **文章编号：** 1671－1807(2024)03－0210－06

随着我国经济的蓬勃发展，越来越多的人积极融入金融投资领域，将其视为一种重要的理财手段。股票因其流动性强、投资收益高，一直是广大投资者青睐的投资方式。然而，股票数据表现出非平稳、非线性和高噪声等特点，同时受到多种因素综合影响。因此，建立一个科学合理且实用的股票价格预测模型变得尤为具有挑战性[1]。

传统的统计学模型，如自回归移动平均（auto regressive and moving average，ARMA）、自回归求和移动平均（autoregressive integrated moving average，ARIMA）已经被广泛应用于经济、社会等各个领域。吴玉霞和温欣[2]采用 ARIMA 模型对股票价格进行短期预测，取得了不错的效果。尽管传统统计模型在时间序列数据的预测中表现出一定的能力，但是也存在明显的局限性，这些模型本质上只能捕捉线性关系，无法捕捉非线性关系。这意味着它们在处理包含复杂非线性因素的数据时可能无法提供准确的预测。随着人工智能的快速发展，机器学习算法凭借其强大的特征提取和学习能力已广泛应用于自然语言处理、推荐系统、数据挖掘等领域[3]。邓烜堃等[4]采用 BP（back propagation）神经网络用于股票价格预测；邬春学和赖靖文[5]利用支持向量机预测股票价格。还有多种模型混合的方法，如崔文喆等[6]提出基于广义自回归条件异方差（generalized autoregressive conditional heteroskedasticity，GARCH）和 BP 神经网络模型的股票市场价格预测。尽管机器学习算法在处理非线性数据方面表现出显著的优势，但由于股票数据具有时序相关性，一般的机器学习算法在特征提取方面仍然存在一定的局限性。

随着大数据分析时代的来临，深度学习技术凭借其卓越的学习和泛化能力，在数据预测方面展现出超越传统机器学习的强大潜力[7]。循环神经网络（recurrent neural network，RNN）能够捕捉和记忆序列中的先前信息，适用于处理时间序列数据。然而，当处理长序列时，RNN 面临着梯度消失问题。为了解决这一挑战，长短时记忆网络（long short-term memory，LSTM）作为 RNN 的变种，通过引入门控机制，能够有效控制和更新信息流，更好地处理长期依赖关系。近年来，LSTM 模型及其相关的混合和改进版本在多个领域广泛应用，用于解决分类和预测问题。LSTM 模型为处理复杂、具有时序性的数据提供了更准确和可靠的解决方案。

以往的研究通常都是集中于单一股票的历史数据和相关特征，却忽略了不同股票之间的关联性。同一类股票之间往往具有相似的价格波动趋势，故提出了一种聚类数据的长短期记忆神经网络混合模型。首先，根据历史收盘价对 15 支证券股票进行聚类，聚类结果能够有效表达股票价格的历史相关性和趋势相关性；随后，利用聚类数据训练 LSTM 模型并进行预测；最后，将其结果与单一模

收稿日期： 2023-11-03

基金项目： 安徽省高校自然科学研究项目（KJ2020A0479）；安徽建筑大学博士启动基金（2020QDZ20）

作者简介： 肖田田（1996—），女，安徽安庆人，硕士研究生，研究方向为机器学习。

型 LSTM 进行对比。这一研究方法不仅考虑股票之间的相关性,还能捕捉相关股票之间的内在联系,有望提供更准确的股票价格预测,有助于投资者做出更明智的决策。

1 研究方法

1.1 欧式距离

K-means 算法是一种常用的聚类分析方法,其默认的距离计算公式为欧式距离。对两个时间序列 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$、$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)$,计算欧式距离为公式为

$$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

然而欧式距离在计算两个时间序列的距离时,存在一些限制。当两个时间序列的长度不等时,较长的一个时间序列会剩下无法匹配的点,此时欧式距离计算公式不再适用。此外,两个序列时间步对不齐,但总体的趋势是相似时,显然欧式距离按照点与点之间对应关系的计算方式无法满足研究需求。但动态时间归整(dynamic time warping,DTW)算法可以通过计算所有点之间的最小距离,实现一对多的匹配,解决时间轴上的失真[8]。

1.2 动态时间规划(DTW)

设 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$,$\boldsymbol{y}=(y_1,y_2,\cdots,y_m)$ 是长度分别为 n 和 m 的两个时间序列,为了对齐这两个序列,构建一个距离矩阵 $\boldsymbol{M}^{n\times m}$,矩阵元素 (i,j) 表明 $\boldsymbol{x}$ 中第 i 个点 x_i 与 $\boldsymbol{y}$ 中第 j 个点 y_j 的距离,x_i 和 y_j 两点间的距离定义为欧氏距离 $w_s=(x_i-y_j)^2$,在矩阵中找到一条通过若干格点的路径,路径通过的格点即为两个序列进行计算的对齐的点。DTW 算法就是找到一条最优路径,使得路径上所有匹配点对的距离和最小,该最小距离对应 DTW 距离为

$$\mathrm{DTW}(\boldsymbol{x},\boldsymbol{y})=\min\sqrt{\sum_{s=1}^{k}w_s} \tag{2}$$

矩阵 $\boldsymbol{M}$ 上的最优路径可以使用如下递归函数计算:

$$\gamma(i,j)=d(x_i,y_j)+\min\{\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1)\} \tag{3}$$

1.3 *K*-means-DTW 算法

将距离算法修改为 DTW 后,*K*-means 的目标函数 E 如式(4)所示。给定数据集 $D=\{x_1,x_2,\cdots,x_m\}$,将其聚为 k 个簇(k 根据需要设定);$C=\{c_1,c_2,\cdots,c_m\}$,C 为 k 个簇的中心。过程是先计算每个序列到当前对应簇的中心序列的 DTW 距离的累积。

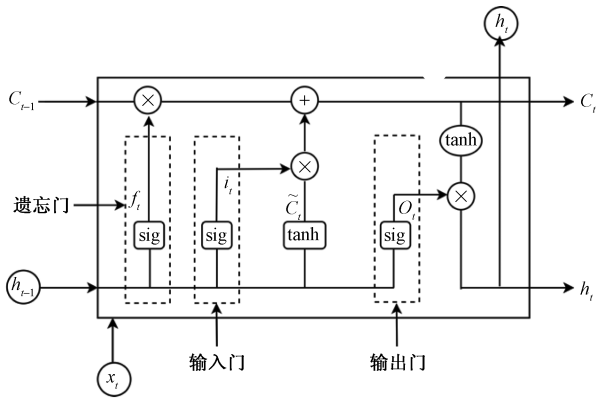
$$E=\sum_{i=1}^{k}\sum_{x\in c_i}\mathrm{DTW}(x,\mu_i) \tag{4}$$

式中:$\mu_i=\frac{1}{|c_i|}\sum_{x\in c_i}x$ 为 c_i 的中心。

具体步骤:①随机选择选择 k 个样本作为初始聚类中心 μ_i;②计算其余样本到聚类中心 μ_i 的 DTW,将样本划到距离最近的聚类中心所属的类别中;③重新计算每个簇的新中心 μ_i;④重复②、③步骤直到每个簇的中心不再变。

1.4 长短期记忆网络(LSTM)

LSTM 是由 Hochreiter 和 Schmidhuber[9] 首次提出的一种神经网络模型。LSTM 模型通过引入门控机制,确保梯度能长期维持传递,有效缓解了标准 RNN 中梯度消失、梯度爆炸问题,近年来,被广泛应用于时间序列数据的处理与分析。LSTM 模型存储单元由 3 部分组成,分别是遗忘门、输入门和输出门,如图 1 所示。



C_{t-1} 为上一个单元状态;C_t 为当前单元状态;$\widetilde{C}_t$ 为临时单元状态;x_t 为当前时刻的输入值;h_{t-1} 为上一时刻的输出值;h_t 为当前时刻的输出值;sig 为 sigmoid 函数;tanh 为 tanh 函数;f_t 为遗忘门的输出值;i_t 为输入门的输出值;O_t 为输出门的输出值;⊗为乘法运算符;⊕为加法运算符

图 1 LSTM 模型结构

LSTM 模型的计算过程如下。

将上一时刻的输出值和当前时刻的输入值输入到遗忘门中,经过计算得到遗忘门的输出值:

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f)\,f_t=\sigma \tag{5}$$

式中:f_t 的取值范围为(0,1);W_f 为遗忘门的权重;b_f 为遗忘门的偏置;x_t 为当前时刻的输入值;h_{t-1} 为上一时刻的单元输出;σ 为 sigmoid 函数。

将上一次的输出值和当前的输入值输入到输入门,经过计算得到输出值和输入门的候选细胞状态:

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i) \tag{6}$$

$$\widetilde{C}_t=\tanh(W_C[h_{t-1},x_t]+b_C) \tag{7}$$

式中：i_t 取值范围为(0,1)；W_i 为输入门的权重；b_i 为输入门的偏置；W_C 为候选输入门的权重；b_C 为候选输入的偏置门。

更新当前单元状态：

$$C_t = f_t C_{t-1} + i_t \widetilde{C}_t \tag{8}$$

式中：C_t 取值范围为(0,1)。

在 t 时刻接收输出值 h_{t-1} 和输入值 x_t 作为输出门的输入，得到输出门的输出 O_t：

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_O) \tag{9}$$

式中：O_t 取值范围为(0,1)；W_o 为输出门的权重；b_o 为输出门的偏置。

通过计算输出门的输出和细胞状态得到 LSTM 的输出值：

$$h_t = O_t \tanh C_t \tag{10}$$

2　数据描述

2.1　数据

以中国主要的上市证券公司为研究对象，涵盖了共计 15 支不同的证券公司。这些公司包括海通证券(HT)、东北证券(DB)、西南证券(XN)、太平洋证券(TPY)、长江证券(CJ)、国金证券(GJ)、国元证券(GY)、东吴证券(DW)、兴业证券(XY)、广发证券(GF)、光大证券(GD)、山西证券(SX)、吉林敖东证券(JLAD)、中信证券(ZX)和招商证券(ZS)。以天为单位收集信息，研究数据的时间跨度为 2011 年 12 月 12 日至 2021 年 12 月 17 日，总计涵盖了 2 346 个交易日的数据。实验的数据集均来自国泰安(CSMAR)数据库。

考虑不同股票的价格波动区间存在差异，所以在进行聚类和股票价格预测之前需要消除量纲不同带来的影响。常见的无量纲化处理方法有标准化和归一化，本文在聚类之前对数据进行标准化处理，公式为

$$X^* = \frac{X - \mu}{S} \tag{11}$$

式中：X 为样本数据；μ 和 S 分别为样本数据的均值和标准差；X^* 为标准化后的数据。

通过 LSTM 模型进行股票价格预测时，为了方便计算，对每个证券的收盘价进行归一化处理，使最终的输出值在 0 和 1 之间，计算公式为

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{12}$$

式中：X 为样本数据；$X_{\min}$ 为样本数据中的最小值；$X_{\max}$ 为样本数据中的最大值；X^* 为归一化后的数据。

2.2　相关性分析

计算 15 支证券的相关系数，如图 2 所示。太平洋证券和西南证券的相关系数最高，为 0.95；东北证券和中信证券的相关系数最低，为 −0.09。总体而言，不同证券公司之间的两两相关系数是不规则的，不存在一定的正负相关规律。其中，中信证券与绝大多数证券的相关性皆较低。

3　实验结果

实验环境基于 Python 3.8，并使用 Keras 神经网络库实现神经网络模型。LSTM 模型的层数为 2，每层的维度均为 64。引入 Relu 激活函数增强神经网络模型的非线性。使用均方误差作为损失函数。为了防止过拟合，并提高模型的泛化能力，使用 Dropout 正则化，其中概率设置在区间[0, 1)。在训练 LSTM 模型时，采用随机梯度下降法，批量大小设置为 32，迭代次数为 100，输入数据的时间步

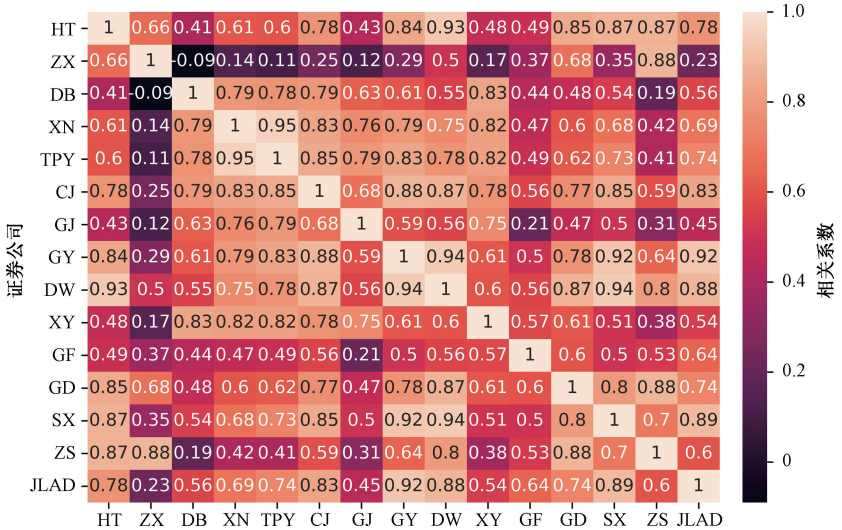


图 2　相关系数图

长为 10。优化器选择了 Adam，学习率被设置为 0.001。

3.1 聚类分析

对 15 支证券股票的收盘价进行标准化处理后，将 K-means 模型的距离算法修改为 DTW 后进行聚类，结果见表 1。其中，K 代表聚类簇的数量，在实验中分别设置了 2～5 个不同的聚类数目，以观察在不同簇数下的聚类情况。

表 1 显示，国元证券、东吴证券、山西证券和吉林敖东证券这 4 支的股票一直被聚在同一类别中，即国元证券、东吴证券、山西证券和吉林敖东证券之间的相似性较高，这表明它们的股价波动趋势存在显著的相似性。这种相似性意味着它们之间可能存在一些共同的市场因素或者投资者行为，导致它们的股价表现相近。这 4 支的股票价格走势也极为相似，如图 3 所示。此外，在 $K=3\sim5$ 时的聚类结果中，中信证券已经成为一个独立的类别，再次验证了中信证券与其他多数证券之间的低相关性。

选取来自不同集群的 5 支股票公司，分别是太平洋证券、广大证券、中信证券、吉林敖东证券和广发证券，以进一步分析。其股价走势如图 4 所示，与图 3 所示的趋势相比，这 5 支证券的趋势呈现出显著的差异。该结果证明了 K-means-DTW 聚类算法的有效性，进一步说明 DTW 方法在计算时间序列相似性方面的重要性。DTW 方法能够有效考虑时间序列上的畸变，有助于识别和分析不同证券之间的相似性和差异。

选择聚类后的 4 支证券进行股价预测，将数据集划分为训练集和测试集，使用前 80% 作为训练集，后 20% 作为测试集，训练集用于训练 LSTM 模型并调整模型参数，测试集用于模型性能测试。证券股价预测流程如图 5 所示。

实验选用 3 个评价指标衡量预测效果，分别为均方误差(MSE)、平均绝对误差(MAE)以及决定系

表 1 股票价格聚类结果

聚类簇数 K	证券	
2	0	海通证券、东北证券、西南证券、太平洋证券、长江证券、国金证券、国元证券、东吴证券、兴业证券、广发证券、光大证券、山西证券、吉林敖东证券
	1	中信证券、招商证券
3	0	东北证券、西南证券、太平洋证券、长江证券、国金证券、兴业证券
	1	海通证券、国元证券、东吴证券、广发证券、光大证券、山西证券、招商证券、吉林敖东证券
	2	中信证券
4	0	东北证券、西南证券、太平洋证券、长江证券、国金证券、兴业证券
	1	海通证券、广发证券、光大证券、招商证券
	2	中信证券
	3	国元证券、东吴证券、山西证券、吉林敖东证券
5	0	西南证券、太平洋证券、长江证券、国金证券
	1	海通证券、光大证券、招商证券
	2	中信证券
	3	国元证券、东吴证券、山西证券、吉林敖东证券
	4	东北证券、兴业证券、广发证券

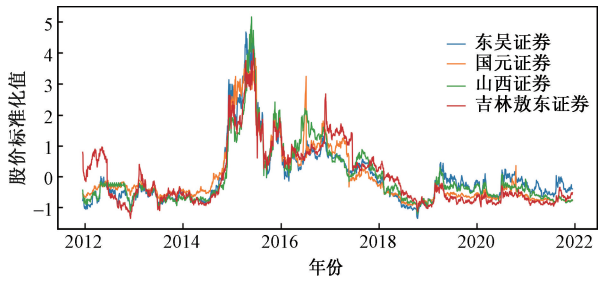


图 3 2012—2022 年同一簇的证券股价趋势

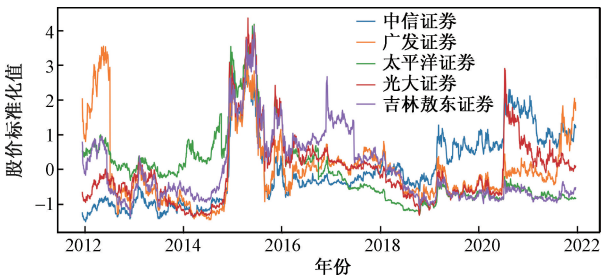


图 4 2012—2022 年不同簇的证券股价趋势

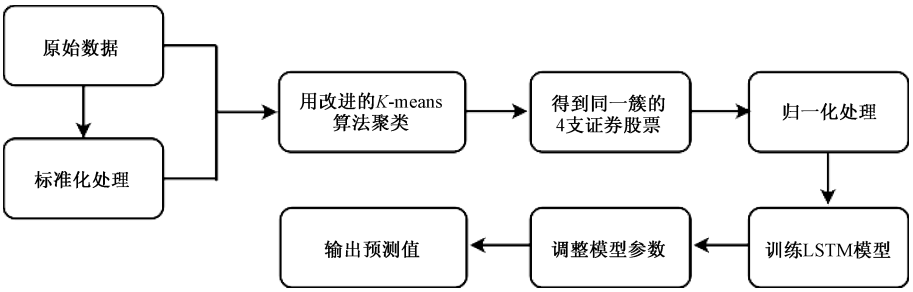


图 5 预测流程

数(R^2),具体计算公式为

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_1-y_i)^2 \tag{13}$$

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_1-y_i| \tag{14}$$

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_l)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_{l\ i})^2} \tag{15}$$

式中:N 为样本总数;y_i 为测试集的实际值;$\hat{y}_l$ 为测试集的预测值;$\bar{y}_l$ 为测试集的真实值的平均值。

3.2　预测分析和性能比较

通过单一模型和混合模型对股票价格进行预测。其中单一模型是指将每支股票的历史数据独立作为输入,然后使用 LSTM 模型进行预测。这意味着每支股票都有其独立的预测模型,不考虑其与其他股票之间的相关性。而混合模型即 K-means-LSTM,将经过聚类后的多支股票的数据一起输入到 LSTM 模型中进行预测。这意味着多支股票共享一个模型,模型能够有效考虑不同股票之间的相关性,从而更全面地捕捉市场的复杂性。最后,比较单一模型和混合模型的预测结果。

由表 2 可知,混合模型在不同证券的预测中展现出较为出色的综合表现。例如,在预测国元证券时,混合模型的 R^2 高于单一模型,MSE 和 MAE 低于单一模型。这意味着混合模型在一定程度上能更好地解释股价变动的方差,表现出更好的拟合效果,更接近实际股价趋势。混合模型在考虑不同股票之间相关性的情况下,具有更强的预测能力。

然而,混合模型在吉林敖东证券的预测中表现不如单一模型,这可能是因为吉林敖东证券受其他 3 支证券的影响较小,混合模型在这种情况下无法充分发挥其优势。这一观察结果强调了模型选择的重要性,应根据不同证券的特性来决定使用单一模型还是混合模型,以获得更准确的预测结果。

表 2　混合模型和单一模型评价指标

模型	证券	MSE	MAE	R^2
混合模型	东吴证券	0.057 22	0.155 88	0.934 34
	国元证券	0.047 44	0.134 39	0.958 75
	山西证券	0.023 34	0.096 77	0.953 79
	吉林敖东证券	0.079 64	0.187 19	0.892 52
单一模型	东吴证券	0.059 42	0.158 48	0.928 40
	国元证券	0.067 84	0.166 98	0.940 67
	山西证券	0.028 18	0.110 51	0.946 78
	吉林敖东证券	0.077 63	0.183 39	0.894 97

综合而言,混合模型在使用相似证券股价数据进行训练时通常表现优于在单一证券数据上进行训练的模型,同时也说明了聚类算法的有效性。混合模型的优势在于它能够更好地分析同类证券之间的相互影响,从而更准确地捕捉市场的复杂性。即通过考虑相似性高的证券,提高预测的准确性。

4　结论

为了提高股票预测的准确性,提出了一种混合深度学习模型 K-means-LSTM。该模型基于证券股价之间的相关性,运用 K-means-DTW 聚类算法,筛选出具有相似价格波动的股票,接着利用聚类数据训练 LSTM 模型并进行预测。选取 15 支证券的股票数据进行实验,结果表明混合模型 K-means-LSTM 表现出更好的预测性能,其预测精度和稳定性均优于单一模型 LSTM。混合模型综合考虑了具有相似股票价格趋势的公司之间的关联性,不仅弥补了仅使用历史数据的不足,还为股票价格预测提供了更为全面的数据基础,由于其更高的预测精度和良好的泛化能力,该模型在金融时间序列分析中表现出科学性和可行性,对于构建大数据和人工智能背景下的金融时间序列数据预测模型具有参考价值,有望为投资者提供更可靠的决策支持,促进投资组合模型的优化建立。

然而,K-means-LSTM 模型只考虑股票价格历史数据和相关企业的影响,但是影响股票价格的因素是多方面的,如新闻,市场情绪等。因此,后续的研究应该更全面地考虑外部因素,以更准确地分析和预测股票价格,以便投资者能够更好地规划其投资策略并降低风险。

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] 吴玉霞, 温欣. 基于 ARIMA 模型的短期股票价格预测[J]. 统计与决策, 2016, 32(23): 83-86.

[3] 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能, 2014, 27(4): 327-336.

[4] 邓烜堃, 万良, 黄娜娜. 基于 DAE-BP 神经网络的股票预测研究[J]. 计算机工程与应用, 2019, 55(3): 126-132.

[5] 邬春学, 赖靖文. 基于 SVM 及股价趋势的股票预测方法研究[J]. 软件导刊, 2018, 17(4): 42-44.

[6] 崔文喆, 李宝毅, 于德胜. 基于 GARCH 模型和 BP 神经网络模型的股票价格预测实证分析[J]. 天津师范大学学报(自然科学版), 2019, 39(5): 30-34.

[7] 胡越, 罗东阳, 花奎, 等. 关于深度学习的综述与讨论

[J]. 智能系统学报, 2019, 14(1): 1-19.

[8] BERNDT D J, CLIFFORD J. Using dynamic time warping to find patterns in time series: WS-94-03[R]. Palo Alto: AAAI, 1994: 359-370.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

Securities Stock Price Prediction Based on *K*-means-LSTM Model

XIAO Tiantian

(School of Mathematics and Science, Anhui Jianzhu University, Hefei 230601, China)

Abstract: In view of the non-stationary and non-linear characteristics of stock data, traditional statistical models cannot accurately predict the future trend of stock prices. To address this problem, a hybrid deep learning method is constructed to improve prediction performance. Firstly, the distance algorithm is modified to DTW (dynamic time warping) by expanding the K-means clustering algorithm to K-means-DTW, which is more suitable for time series data, to cluster securities with similar price trends. Then, the LSTM (long short-term memory)model is trained through clustering data to predict the price of a single stock. Experimental results show that the hybrid model K-means-LSTM shows better prediction performance and its prediction accuracy and stability are better than the single LSTM model.

Keywords: stock price prediction; K-means; DTW; K-means-LSTM(K-means-long short-term memory) hybrid model