

基于 LSTM 神经网络模型的量化投资研究

摘 要

本文针对股票各项指标对量化投资的影响问题，对股票的各项影响因素与“数字经济”板块的关系进行了分析研究，建立均值插补模型、皮尔逊相关系数模型、基于 LSTM 神经网络的股票预测模型等多种数学模型，分析出与“数字经济”板块有关的主要指标，从而对“数字经济”板块指数进行预测，给出预测结果及总收益率等参数值。

针对问题一，首先对各项指标信息进行数据预处理，删除不在分析时间范围内的数据，通过建立均值插补模型，完善缺失数据。以日作为研究单位，分别求出“数字经济”板块各项每日的均值作为当日该类的平均水平。建立评价指标规范化模型，将所有数据标准化、无量纲化处理。最后，绘制并观察各指标图像的变化规律，建立皮尔逊相关系数模型，计算各项指标与上述六项信息的相关性数值并排序，分别得到各指标对于开盘价、收盘价、最高价、最低价、成交量和成交额各项指数的影响程度。再计算出各指标对于上述六项信息的相关性系数的均值，仅保留均值大于 0.6 的强相关性和极强相关性指标，从而得出与“数字经济”板块有关的 9 项主要指标为：EXPMA、MA、BBI、BOV、创业板指数、VMA、深证成分指数、BOLL、互联网电商。

针对问题二，参考问题一所得出的对于成交量的 12 项主要指标，建立基于 LSTM 神经网络的股票预测模型。由于无法在股票交易当天就获知上述 12 项指标的全部当日数据。因此，本问选用 2021 年 7 月 13 日至 2021 年 12 月 30 日的指标数据为输入层特征值，2021 年 7 月 14 日至 2021 年 12 月 31 日的每日每五分钟成交量为输出值，对 LSTM 神经网络模型进行训练，从而得出 12 项指标与成交额之间的非线性映射关系。最后，将 2022 年 1 月 3 日至 2022 年 1 月 27 日的指标数据作为输入，测试该神经网络，预测出 2022 年 1 月 4 日至 2022 年 1 月 28 日每日每五分钟的股票成交量。

针对问题三，本问直接参考问题二所建立的 LSTM 神经网络模型和训练集数据，对 2022 年 1 月 4 日至 2022 年 1 月 28 日每五分钟的收盘价进行预测。

针对问题四，基于问题三所得的 2022 年 1 月 4 日至 2022 年 1 月 28 日，共 852 个“数字经济”板块收盘价数据，以 100 万元为初始资金，先对总收益率进行分析。建立基于数据降维法的收益最大化模型，同时考虑交易佣金和投资者获利需求，给出交易者最合理的投资策略，从而得到最大的投资总收益率。接着，根据题目所给公式计算出信息比率和最大回撤率，分别反映出一段时间内的收益和该股票的投资风险。最后，利用题目所给的 2022 年 1 月 4 日至 2022 年 1 月 28 日的实际收盘价计算相应的总收益率、信息比率和最大回撤率，与使用预测值计算的结果相比较，从而得出预测值的准确性。

综上所述，本文通过建立多个数学模型，依据题目所给数据对股票的各项影响指标与“数字经济”板块数据的关系进行了分析研究，给出了对板块指数的预测结果及总收益率等参数值。最后，对模型进行了优缺点的评价及推广，使其更具现实意义。

关键词：均值插补模型；皮尔逊相关系数；LSTM 神经网络；数据降维

目录

一、 问题重述.....	3
1.1 问题背景.....	3
1.2 问题提出.....	3
二、 问题分析.....	3
2.1 问题一的分析.....	3
2.2 问题二的分析.....	3
2.3 问题三的分析.....	4
2.4 问题四的分析.....	4
三、 基本假设.....	4
四、 符号说明.....	4
五、 模型的建立与求解.....	5
5.1 问题一模型的建立与求解.....	5
5.1.1 问题分析	5
5.1.2 均值插补模型的建立	6
5.1.3 评价指标规范化模型的建立	6
5.1.4 皮尔逊相关系数模型的建立	7
5.1.5 模型的求解	8
5.2 问题二模型的建立与求解.....	10
5.2.1 问题分析	10
5.2.2 基于 LSTM 神经网络的股票预测模型的建立.....	11
5.2.3 模型的求解	14
5.3 问题三模型的建立与求解.....	15
5.3.1 问题分析	15
5.3.2 基于问题二 LSTM 神经网络模型的补充.....	15
5.3.3 模型的求解	16
5.4 问题四模型的建立与求解.....	18
5.4.1 问题分析	18
5.4.2 总收益率、信息比率、最大回撤率计算方法	19
5.4.3 基于数据降维法的收益最大化模型的建立	19
5.4.4 模型的求解	20
六、 模型的评价.....	21
6.1 模型的优点.....	21
6.2 模型的缺点.....	21
七、 模型的推广.....	22
八、 参考文献.....	22

一、问题重述

1.1 问题背景

随着时代的发展，量化投资在全球金融交易市场的地位愈加重要。据统计，近年来全世界越来越多的投资者开始加入量化投资交易市场。

投资者可以通过网络渠道获取已知交易数据，使用量化的方式对投资项目的数据进行客观分析从而得出投资策略。但是由于受到市场数据繁杂、产品价格波动、交易规则变化等诸多不确定因素的影响，投资者要根据海量信息进行客观分析，从而规避风险，得出自己的投资策略，是一个具有挑战性的工作。

1.2 问题提出

根据以上背景，及题目所给的数据信息，需要解决以下四个问题：

问题一，根据附件所提供的指标信息，分析得出与“数字经济”板块有关的主要影响指标。

问题二，根据问题一得到的主要指标，以 2021 年 7 月 14 日至 2021 年 12 月 31 日的“数字经济”板块指数为训练集，对“数字经济”板块指数每 5 分钟的成交量进行预测。并以 2022 年 1 月 4 日至 2022 年 1 月 28 日的实际“数字经济”板块指数为测试集进行分析验证。

问题三，根据上述两问题建立模型，以 2021 年 7 月 14 日至 2021 年 12 月 31 日的“数字经济”板块指数的数据为训练集，对每 5 分钟的收盘价进行预测。并以 2022 年 1 月 4 日至 2022 年 1 月 28 日的“数字经济”板块指数为测试集进行分析验证。

问题四，根据问题三所得结果，以 100 万元为初始资金，0.3%为交易佣金，对“数字经济”板块指数进行交易。从而得出 2022 年 1 月 4 日至 2022 年 1 月 28 日期间交易的总收益率、信息比率和最大回撤率。

二、问题分析

2.1 问题一的分析

问题一要求根据附件所提供的指标信息，分析得出与“数字经济”板块有关的主要影响指标。首先，对附件所提供的各项指标数据进行数据清洗和筛选。采用均值插补法，将缺失数据完善。以日作为单位进行研究，将“数字经济”板块的六项信息分别看作独立的一类，以各类每日的均值作为当日该类的平均水平。接着，建立评价指标规范化模型，将各项指标的数据进行标准化、无量纲化处理。最后，绘制出各指标数据的变化规律图像。观察曲线的变化规律，建立皮尔逊相关系数模型，确定各项指标与上述六项信息的相关系数值。将六项信息与各指标的相关系数分别排序，得到各指标对于开盘价、收盘价、最高价、最低价、成交量和成交额各项的影响程度。再计算出各指标对于上述六项信息的相关性系数的均值，舍去其中均值小于 0.6 的指标，仅保留强相关性和极强相关性的指标，从而得出与“数字经济”板块有关的主要指标。

2.2 问题二的分析

问题二要求根据问题一所提取的各项指标，对“数字经济”板块指数每 5 分钟成交量进行预测。使用 2021 年 7 月 14 日至 2021 年 12 月 31 日的“数字经

济”板块指数为训练集，2022 年 1 月 4 日至 2022 年 1 月 28 日数据为测试集。首先，根据问题一的分析结果，提取出了与“数字经济”板块相关性最强的 12 项指标。因此，本问的预测将基于上述 12 项指标，建立基于 LSTM 神经网络的股票预测模型，利用问题一中 2021 年 7 月 13 日至 2021 年 12 月 30 日规范化处理后的数据，对所建立的 LSTM 神经网络模型进行学习训练。神经网络通过学习可以得到 12 项指标与成交额之间的非线性映射关系，从而建立起该神经网络的输入输出关系。最后，将 2022 年 1 月 4 日至 2022 年 1 月 28 日的数据作为输入，测试该神经网络，得出该时间段内每日每五分钟的股票成交量。

2.3 问题三的分析

问题三需要在问题一和问题二所建立的数学模型的基础上，来预测每 5 分钟的“数字经济”板块指数的收盘价。本文可直接参考问题二所建立的 LSTM 神经网络模型，对未来一段时间的收盘价进行预测。

2.4 问题四的分析

问题四要求根据问题三所得结果，以 100 万元为初始资金，0.3%为交易佣金，对“数字经济”板块指数进行交易。从而得出 2022 年 1 月 4 日至 2022 年 1 月 28 日期间交易的总收益率、信息比率和最大回撤率。首先，利用问题三所得的 2022 年 1 月 4 日至 2022 年 1 月 28 日，共 852 个“数字经济”板块收盘价数据，以 100 万元为初始资金，先对总收益率进行分析。建立基于数据降维法的收益最大化模型，同时考虑交易佣金和投资者获利需求，给出交易者最合理的投资策略，从而得到最大的投资总收益率。接着，根据题目所给公式计算出信息比率和最大回撤率，分别反映出一段时间内的收益和该股票的投资风险。最后，利用题目所给的 2022 年 1 月 4 日至 2022 年 1 月 28 日的实际收盘价计算相应的总收益率、信息比率和最大回撤率，与使用预测值计算的结果相比较，从而得出预测值的准确性。

三、基本假设

1. 假设附件所给数据全部准确，无异常。
2. 假设国内股票市场指标、国际股票市场指标等各指标间相互无影响。
3. 假设消费者可以随意买卖任意数额的股票，不受限制。

四、符号说明

表 1 符号说明

符号	说明
Q_i	指标 Q 第 i 天数值的插补
z	极小型指标
M	指标 z 可能取值的最大值
m	指标 z 可能取值的最小值
o_i	指标 o 第 i 天的数值 ($i = 1,2,3 \dots n$)
K	“数字经济”板块中六项数据的其中一项
$E(S)$	数学期望
$D(S)$	方差

Cov	协方差
$Corr(S, K)$	二者的相关系数
x_1	输入特征值
p	偏执值
y	输出值预测值
a_1	x_1 运算时对应的权重值
Z	隐藏节点
$f(z)$	激活函数
E_t	总误差
δ	学习率
a_n^{new}	新的权重值
G	隐藏层神经元个数
f_t	表示遗忘门
c_t	t 时间下神经元的状态
$\{X_{nt}\}$	时间序列
z_i	第 <i>i</i> 天的当日收盘总资产市值
P_i	当日收益率
P_t	总收益率
q_i	第 <i>i</i> 日超额收益率
r_i	第 <i>i</i> 日中证 500 指数收益率
\bar{q}_i	T 日内超额收益率的均值
s	超额收益率标准差
q_i'	日均超额收益率
m	信息比率
D_i	第 <i>i</i> 天的产品净值

五、模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 问题分析

问题一要求根据附件所提供的指标信息，分析得出与“数字经济”板块有关的主要影响指标。首先，由于分析是基于 2021 年 7 月 14 日至 2022 年 1 月 28 日的数字经济板块的数据，要先对附件所提供的各项指标进行数据清洗和筛选。对于各项指标中不在上述日期范围内的数据，本文分析暂时不作考虑，可直接删除；对于在分析时间范围内的数据，观察发现有部分指标的数据缺失。通过建立均值插补模型，将缺失数据完善。

由于“数字经济”板块信息是以每五分钟为一周期的统计数据，而各项指标数据大多以每日为一个周期；因此，本问将以日作为单位进行研究。（对于宏观市场指标 1 和宏观市场指标 2，由于其变化周期分别为每月和每季度，数据过少且参考价值不高，本文不对其进行分析）。将“数字经济”板块的六项信息分别看作独立的一类，以各类每日的均值作为当日该类的平均水平。接着，为防止各项指标因数量级大小不同对分析造成影响，建立评价指标规范化模型，将各项指标的数据进行标准化、无量纲化处理，得到标准数据。

最后，绘制出各指标数据的变化规律图像。通过观察各指标与“数字经济”板块的六项信息的曲线变化规律，建立皮尔逊相关系数模型，确定各项指标与上述六项信息的相关性数值。将六项信息与各指标的相关系数排序，分别得到各指标对于开盘价、收盘价、最高价、最低价、成交量和成交额各项的影响程度。再计算出各指标对于上述六项信息的相关性系数的均值，舍去其中均值小于 0.6 的指标，仅保留强相关性和极强相关性的指标，从而得出与“数字经济”板块有关的主要指标。

具体思维流程图如下：

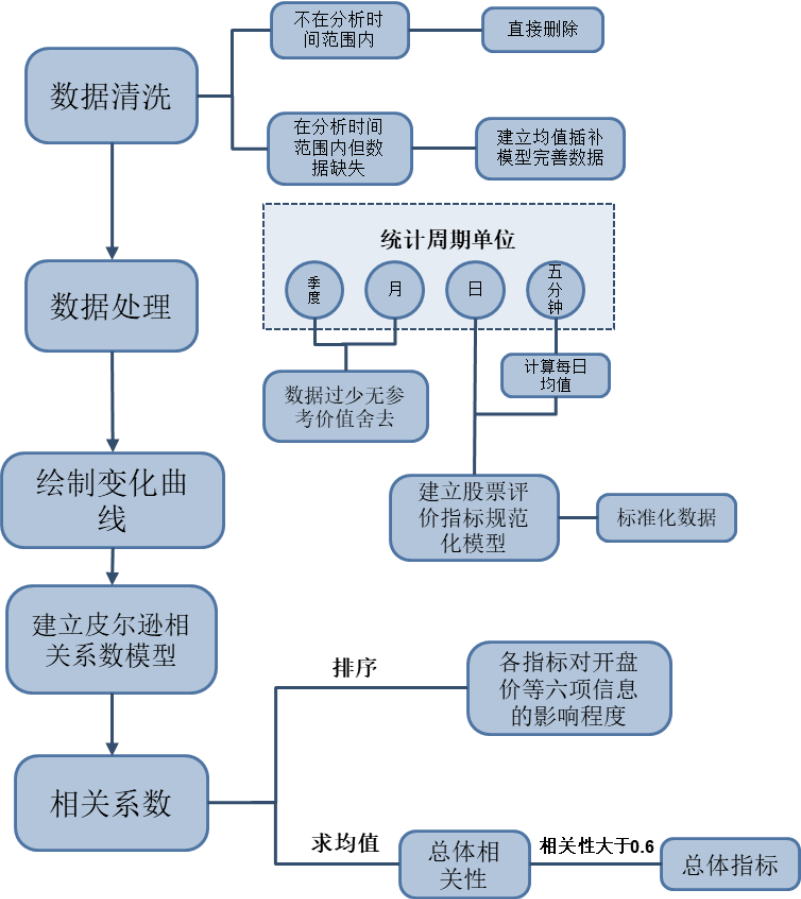


图 1 问题一思维流程图

5.1.2 均值插补模型的建立

为保证分析结果的准确性，对于附件中各指标数据的缺失项进行补全。与热卡填充法和回归替换法相比，均值插补法更为简便且可行，因此，此处使用均值插补法对缺失数据进行补全。

设缺失数据为指标 Q 第 i 天的数值，为确保插补数值的准确性，取缺失数据前后两天的数值的均值进行插补，则指标 Q 第 i 天数值的插补结果 Q_i 的计算公式如下：

$$Q_i = \frac{Q_{i+1} + Q_{i+2} + Q_{i-1} + Q_{i-2}}{4} \tag{1}$$

从而将表格中的缺失数据补齐。

5.1.3 评价指标规范化模型的建立

为防止各项指标因数量级大小不同对分析造成影响，此处对各项指标的数据进行一致化、无量纲化处理。

(一) 评价指标一般可以分为以下四种类型:

极大型指标: 总是期望指标的取值越多越好;

极小型指标: 总是期望指标的取值越小越好;

中间型指标: 总是期望指标的取值既不要太大, 也不要太小, 取适当的中间值最好;

区间型指标: 总是期望指标的取值落在某个确定的区间内最好。

通过模型对各种类型的指标进行一致化处理如下^[1]:

极小型指标: 设某极小型指标为 z , 则对其极大化处理公式如下:

$$z' = \frac{1}{z} (z > 0) \quad (2)$$

中间型指标: 设某中间型指标为 z , 则对其极大化处理公式如下:

$$z' = \begin{cases} \frac{2(z-m)}{M-m}, & m \leq z \leq \frac{1}{2}(M+m) \\ \frac{2(M-z)}{M-m}, & \frac{1}{2}(M+m) \leq z \leq M \end{cases} \quad (3)$$

其中, M 和 m 分别为指标 z 可能取值的最大值和最小值。

区间型指标: 设某区间型指标为 z , 则对其极大化处理公式如下:

$$z' = \begin{cases} 1 - \frac{a-z}{c}, & z < a \\ 1, & a \leq z \leq b \\ 1 - \frac{z-b}{c}, & z > b \end{cases} \quad (4)$$

其中, $[a, b]$ 为指标 z 的最佳稳定区间, $c = \max\{a-m, M-b\}$, M 和 m 分别为指标 z 可能取值的最大值和最小值。

(二) 对一致化处理后的数据进行标准化处理, 从而得到无量纲数据, 方法如下:

设指标 Q_i 为指标 Q 第 i 天的数值, 指标 o_i 为指标 o 第 i 天的数值 ($i = 1, 2, 3 \dots n$), 则将其进行标准化处理公式如下:

$$Q_i = \frac{Q_i - Q_{\min}}{Q_{\max} - Q_{\min}} \quad (5)$$

$$o_i = \frac{o_i - o_{\min}}{o_{\max} - o_{\min}} \quad (6)$$

5.1.4 皮尔逊相关系数模型的建立

判断两个事物的相关程度, 可以通过计算两事物数据间的相关系数的大小进行判定, 相关系数反映了两变量的变化趋势的方向以及程度, 其取值范围为-1到+1, 0表示两个变量不相关, 正值表示正相关, 负值表示负相关, 绝对值越大则表示相关性越强^[2]。常用的相关性系数分为三种:

(1) 斯皮尔曼相关性系数。它是衡量两个变量的依赖性的非参数指标, 利用单调方程评价两个统计变量的相关性。

(2) 肯德尔相关性系数。它也是一种秩相关系数, 不过, 它的目标对象是有序类别变量。

(3) 皮尔逊相关系数。它广泛用于度量两个变量之间的相关程度, 定义为两个变量之间的协方差和标准差的商。

考虑到我们所要分析的所有变量都是连续数据。且两个变量的观测值是成对

的，每对观测值之间相互独立。因此，本文采用皮尔逊相关系数进行分析计算。

设 S 为某指标数据， K 为“数字经济”板块中六项数据的其中一项，且 $E(S), E(K)$ 分别为二者的数学期望， $D(S), D(K)$ 分别为二者的方差，其协方差 Cov 计算公式如下：

$$Cov(S, K) = E((K - E(K))(S - E(S))) \quad (7)$$

则二者的相关系数 $Corr(S, K)$ 的计算公式为：

$$Corr(S, K) = \frac{Cov(S, K)}{\sqrt{D(S) \times D(K)}} \quad (8)$$

其中，系数的绝对值所反应的相关性情况如下：

- 1.0-0.8 极强相关
- 0.8-0.6 强相关
- 0.6-0.4 中等程度相关
- 0.4-0.2 弱相关
- 0.2-0.0 极弱相关或无相关

5.1.5 模型的求解

（一）将“数字经济”板块的六项信息分别看作单独的一类，计算各类每日的均值作为当日该类的平均水平进行分析，计算部分结果如下表：

表 2 “数字经济”板块各项每日均值

时间	开盘价	收盘价	最高价	最低价	成交量	成交额
2022/1/28	1383.21	1382.61	1384.85	1380.96	11954095	644624958
2022/1/27	1404.69	1403.70	1405.96	1402.50	13425853	709336791
2022/1/26	1433.52	1433.66	1435.15	1432.25	12001096	651272125
...
2021/7/16	1638.34	1637.98	1640.02	1636.61	74763547	1638032541
2021/7/15	1629.32	1629.50	1631.01	1627.48	67199635	1508407208
2021/7/14	1646.11	1645.87	1647.63	1644.22	82722858	1706918062

根据评价指标规范化模型，将所有指标数据进行一致化和标准化处理，处理后部分结果如下表所示（由于数值过于繁杂，此处仅展示部分数据，详细数据见附件）：

表 3 规范化处理后的数据表

时间	最高价	成交额	上证综合指数	VMA	标准普尔 500 指数	美元兑人民币	快手概念
2022/1/28	0.0000	0.0418	0.0401	0.0513	0.3240	0.2129	0.4772
2022/1/27	0.0724	0.0710	0.0140	0.0490	0.1271	0.2102	0.4294
2022/1/26	0.1725	0.0448	0.0000	0.0466	0.1709	0.2482	0.5909
...
2021/7/16	0.8750	0.4900	0.2318	0.9967	0.1283	0.8319	0.2545
2021/7/15	0.8441	0.4315	0.2247	0.9391	0.1898	0.7345	0.2610
2021/7/14	0.9010	0.5211	0.2476	0.9511	0.2164	0.7808	0.2975

（二）根据上述规范化处理后的数据，绘制各指标在分析时间范围内的变化曲线及“数字经济”板块内的六项数据的变化曲线进行比较。以收盘价，EXPMA

和深市：股票：流通市值为例，绘制变化折线图如下图所示：

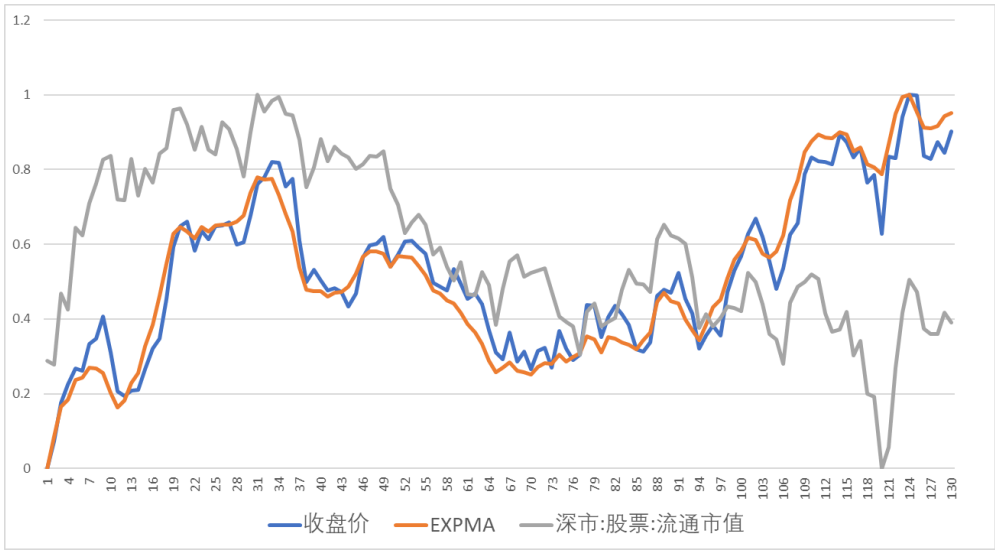


图 2 收盘价、EXPMA、深市：流通市值数据变化折线图

通过观察曲线图走势，发现各项指标与“数字经济”板块存在一定的相关性，利用皮尔逊相关性分析模型，确定各项指标与“数字经济”板块六项信息的相关性数值的绝对值，部分相关性系数如下表所示：

表 4 部分相关性系数表

名称	开盘价	收盘价	最高价	最低价	成交量	成交额
成交量:上证综合指数	0.0171	0.0178	0.0163	0.0189	0.3569	0.3378
成交金额:上证综合指数	0.2421	0.2417	0.2437	0.2397	0.6139	0.6572
VMA	0.5324	0.5329	0.5343	0.5306	0.8929	0.8428
VMACD	0.3817	0.3823	0.3840	0.3797	0.7887	0.7350
东京日经 225 指数	0.1592	0.1581	0.1611	0.1561	0.0393	0.1404
伦敦金融时报 100 指数	0.4362	0.4365	0.4372	0.4348	0.6709	0.5997
...

将所得的相关性系数排序，分别得到各指标对于开盘价、收盘价、最高价、最低价、成交量和成交额各项的影响程度由大到小如下表所示：

表 5 各指标影响程度大小排序表

开盘价	收盘价	最高价	最低价	成交量	成交额
EXPMA	EXPMA	EXPMA	EXPMA	VMA	VMA
MA	MA	MA	MA	VMACD	VMACD
BBI	BBI	BBI	BBI	数字孪生	欧元兑美元
OBV	OBV	OBV	深证成分指数	快手概念	美国证交所
深证成分指数	深证成分指数	深证成分指数	OBV	欧元兑美元	成交金额：上
创业板指数	创业板指数	创业板指数	创业板指数	恒生指数	沪市：流通
MACD	MACD	MACD	MACD	伦敦金融 100	恒生指数
DMA	DMA	DMA	DMA	美元兑人民币	美元兑人民币
BOLL	沪深 300 指数	BOLL	沪深 300 指数	美国证交所	互联网电商
沪深 300 指数	BOLL	沪深 300 指数	BOLL	互联网电商	伦敦金融 100
互联网电商	互联网电商	互联网电商	互联网电商	成交金额：上	快手概念

美国证交所 数字媒体 VMA 互联网 KDJ	美国证交所 数字媒体 VMA 互联网 KDJ	美国证交所 数字媒体 VMA 互联网 KDJ	美国证交所 数字媒体 VMA 互联网 KDJ	法国巴黎指数 沪市：流通 深市：流通 意大利 MIB 标准普尔 500 创业板指数 EXPMA MA OBV 深市：流通 法国巴黎指数 创业板指数	BBI 数字孪生 BOLL EXPMA 意大利 MIB MA OBV 深市：流通 法国巴黎指数 创业板指数
深证综合指数 上证 50 指数 伦敦金融 100 BIAS RSI	深证综合指数 上证 50 指数 伦敦金融 100 BIAS RSI	深证综合指数 上证 50 指数 伦敦金融 100 BIAS RSI	深证综合指数 上证 50 指数 伦敦金融 100 BIAS RSI		

其中，深蓝色代表二者极强相关，蓝色代表二者强相关，浅蓝色代表二者中等程度相关。弱相关和极弱相关此处不再展示。

（三）根据上述模型，计算各指标对于上述六项信息的相关性系数的均值，舍去其中均值小于 0.6 的相关性不强的指标，从而得出与“数字经济”板块总体有关的主要指标，如下表所示：

表 6 与“数字经济”板块总体有关的主要指标

序号	相关性系数均值	指标
1	0.8085	EXPMA
2	0.7850	MA
3	0.7736	BBI
4	0.6847	OBV
5	0.6736	创业板指数
6	0.6443	VMA
7	0.6331	深证成分指数
8	0.6237	BOLL
9	0.6050	互联网电商

5.2 问题二模型的建立与求解

5.2.1 问题分析

问题二要求根据问题一所提取的各项指标，使用 2021 年 7 月 14 日至 2021 年 12 月 31 日的“数字经济”板块指数为训练集，对 2022 年 1 月 4 日至 2022 年 1 月 28 日“数字经济”板块指数每 5 分钟成交量进行预测。首先，根据问题一的分析，提取出了 VMA、VMACD、欧元兑美元等 12 项与“数字经济”板块相关性最强指标。因此，本问的预测将基于上述 12 项指标。

根据预测数据的特点，本问选择建立基于 LSTM 神经网络的股票预测模型进行求解。通过查找相关数据可知，无法在股票交易当天就获知上述 12 项指标的全部当日数据。因此，对于神经网络的训练集，本文选用交易前一日的 12 项指标数据和交易当日的成交量作为一组训练数据。即以问题一中 2021 年 7 月 13 日至 2021 年 12 月 30 日规范化处理后的指标数据为输入层特征值，以问题一中 2021 年 7 月 14 日至 2021 年 12 月 31 日规范化处理后的每日每五分钟的成交量为输出值，对所建立的 LSTM 神经网络模型进行学习训练。

神经网络通过学习得出 12 项指标与成交额之间的非线性映射关系，从而建立起该神经网络的输入输出关系。最后，将 2022 年 1 月 3 日至 2022 年 1 月 27

日的指标数据作为输入，测试该神经网络，得出 2022 年 1 月 4 日至 2022 年 1 月 28 日每日每五分钟的股票成交量。

5.2.2 基于 LSTM 神经网络的股票预测模型的建立

（一）神经网络基本结构

神经网络是在现代神经科学的基础上提出和发展起来的，旨在反映人脑结构及功能的一种抽象数学模型^[3]。神经网络由许多小的单元相互连接组成，这些单元就叫做神经元。神经元在神经网络中位置不同则作用不同，相同作用的神经元可以分为同一层：输入神经元组成输入层、隐藏神经元组成隐藏层、输出神经元组成输出层。神经网络的基本结构如下图所示：

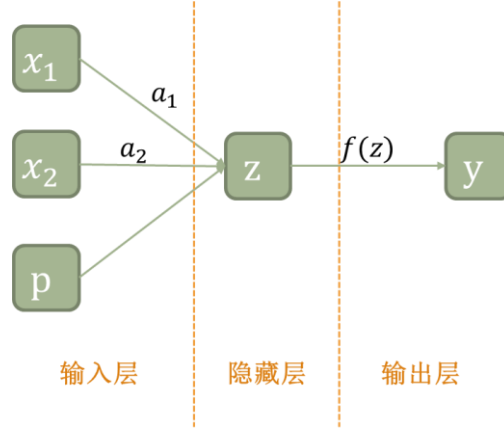


图 3 神经网络基本结构图

其中， x_1, x_2 为输入特征值， p 为偏执值， y 为输出值。 a_1, a_2 为 x_1, x_2 运算时对应的权重值。 z 为隐藏节点， $f(z)$ 为激活函数。

（二）前向传播和后向传播

首先，要训练神经网络得到合适的权重值 a ，则要先进行训练数据的前向传播，前向传播的过程如下^[4]：

$$z = x_1 a_1 + x_2 a_2 + p \quad (9)$$

$$y = f(z) \quad (10)$$

常用的激活函数为 sigmoid，公式如下：

$$\theta(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

在前向传播之后，还需对神经网络进行反向传播，即从后往前计算每层的误差，从而不断更新权重值 a ，使该神经网络对数据有更好的拟合度，从而使预测更加准确。反向传播基本结构原理如下图所示：

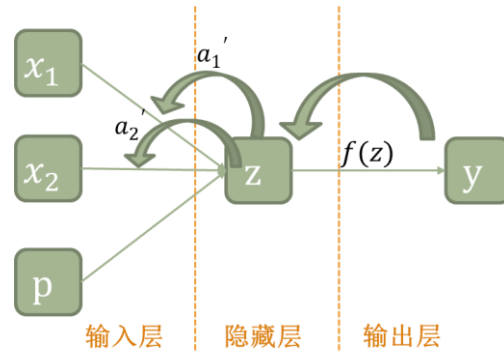


图 4 神经网络反向传播基本结构

将运算所得的预测值和实际值比较产生误差,此处采用均方误差法计算总误差 E_t , 公式如下:

$$E_t = \sum \frac{1}{2} (y_s - y_o)^2 \quad (12)$$

其中, y_s 为实际值, y_o 为输出的预测值。

接着,用总误差对每个权重值求偏导,从而得到对应误差的大小,公式如下:

$$\frac{\partial E_t}{\partial a_n} = \frac{\partial E_t}{\partial z} \cdot \frac{\partial z}{\partial a_n} \quad (13)$$

最后,引入学习率 δ ,通过上述所得误差值对权重进行适当调整,使模型与数据拟合度更高,计算公式如下,其中 a_n^{new} 为新的权重值:

$$a_n^{\text{new}} = a_n - \delta * \frac{\partial E_t}{\partial a_n} \quad (14)$$

(三) BP 神经网络基本原理

以上述神经网络为基础, BP 神经网络通常以一个输入层,一个输出层和 n 个隐藏层组成,基本结构如下图所示:

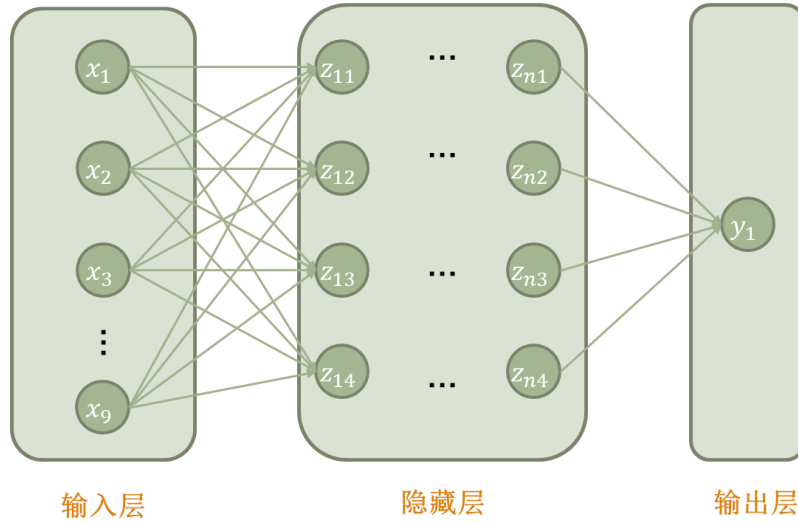


图 5 BP 神经网络基本结构

每层内的神经元与相邻层的所有神经元都互相连接。其中,对于本题而言,输入层的神经元个数为 12,分别对应 12 个指标的数值;输出层的神经元数为 1,对应成交量的数值;隐藏层的神经元个数并不固定,若个数过多,则会导致网络过于敏感,而个数过少,又会使学习能力降低。此处使用经验公式确定隐藏层神经元个数 G , 公式如下:

$$G = \log_2(G_1 + G_2)/2 \quad (15)$$

其中, G_1 为输入神经元个数, G_2 为输出神经元个数。

(四) LSTM 神经网络基本原理

由相关资料显示, BP 神经网络具有容易陷入局部极值、收敛速度慢、预测能力和训练能力存在矛盾等诸多缺点。因此,此处先引入更加优化的 RNN 神经网络算法。

RNN 神经网络在 BP 神经网络的基础上,其隐藏层不仅可以接受当前的输入,还可以记忆前一个时间段的输出,从而影响之后的权重值。其基本结构如下图所示:

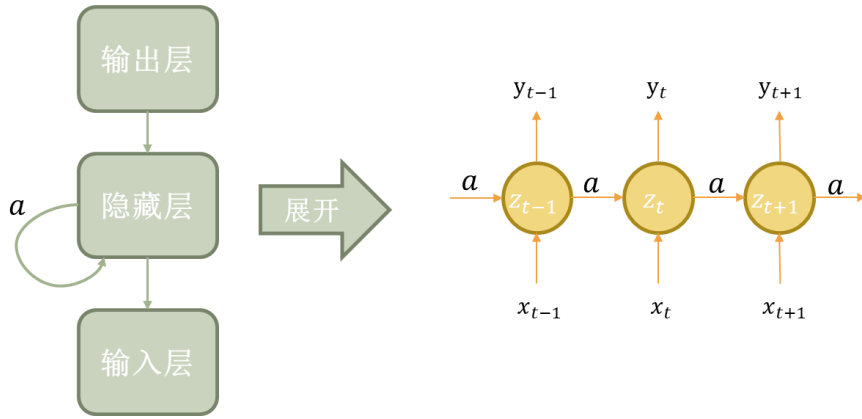


图 6 RNN 神经网络基本结构

由于 RNN 神经网络无法进行长期记忆，还有可能发生梯度消失现象。因此引入 LSTM 神经网络模型。LSTM 模型与 RNN 模型结构相似，但在每个神经元内部加入了输入门、输出门和遗忘门三道控制门。三道门可以对内部信息进行选择性的记忆。其结构如下图所示：

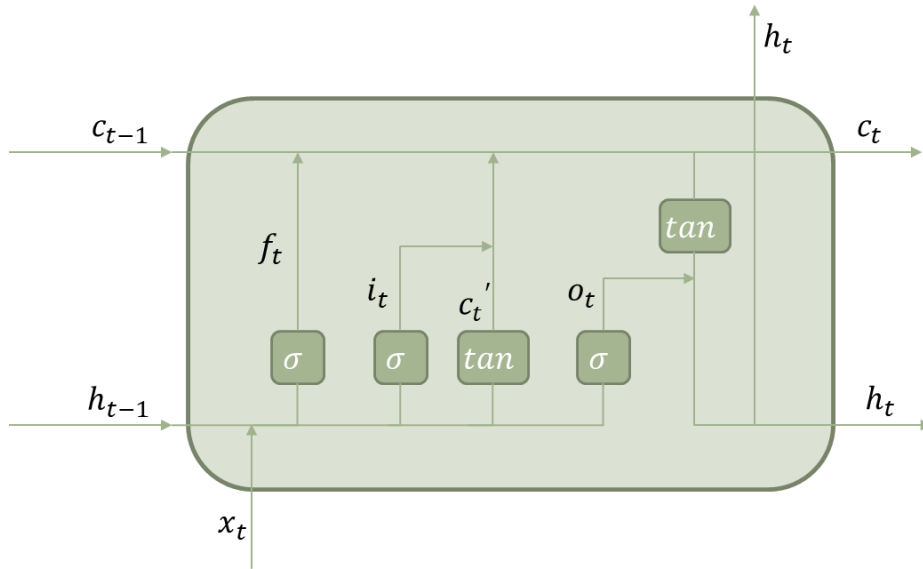


图 7 LSTM 神经网络基本结构

σ 层可以选择是否让相关信息通过，其输出值介于 0~1 之间，表示信息是否可通过的权重。1 表示全部信息都能通过，0 表示信息都不能通过。

LSTM 模型原理如下：

f_t 表示遗忘门，可以决定丢弃哪些信息，其公式如下：

$$f_t = \sigma(a_f[h_{t-1}, x_t] + b_f) \quad (16)$$

其中， σ 为 sigmoid 函数， a_f 为遗忘门的权重， $[h_{t-1}, x_t]$ 表示上一时刻的输入值和该时刻的输出值信息， b_f 为该门的偏执值。

i_t 表示输入门，可以决定要新增哪些信息，其公式如下：

$$i_t = \sigma(a_i[h_{t-1}, x_t] + b_i) \quad (17)$$

其中， a_i 为输入门的权重， $[h_{t-1}, x_t]$ 表示上一时刻的输入值和该时刻的输出值信息， b_i 为该门的偏执值。

根据上述两门的运算结果，更新其状态，公式如下：

$$c'_t = \tanh(a_c[h_{t-1}, x_t] + b_c) \quad (18)$$

$$c_t = f_t c_{t-1} + i_t c'_t \quad (19)$$

其中， c_t 表示 t 时间下神经元的状态， a_c 为神经元状态的权重， b_c 为该状态的偏执值， \tanh 为处理函数,由该函数可以生成备选的更新内容。

o_t 表示输出门，可以决定信息的输出，其公式如下：

$$o_t = \sigma(a_o[h_{t-1}, x_t] + b_o) \quad (20)$$

$$h_t = o_t \tanh(c_t) \quad (21)$$

其中， a_o 为输出门的权重， b_o 为该门的偏执值， h_t 表示 t 时间下神经元的输出值。

5.2.3 模型的求解

通过构建上述 LSTM 神经网络模型，并将训练集输入，令该神经网络进行机器学习，该过程如下图所示：

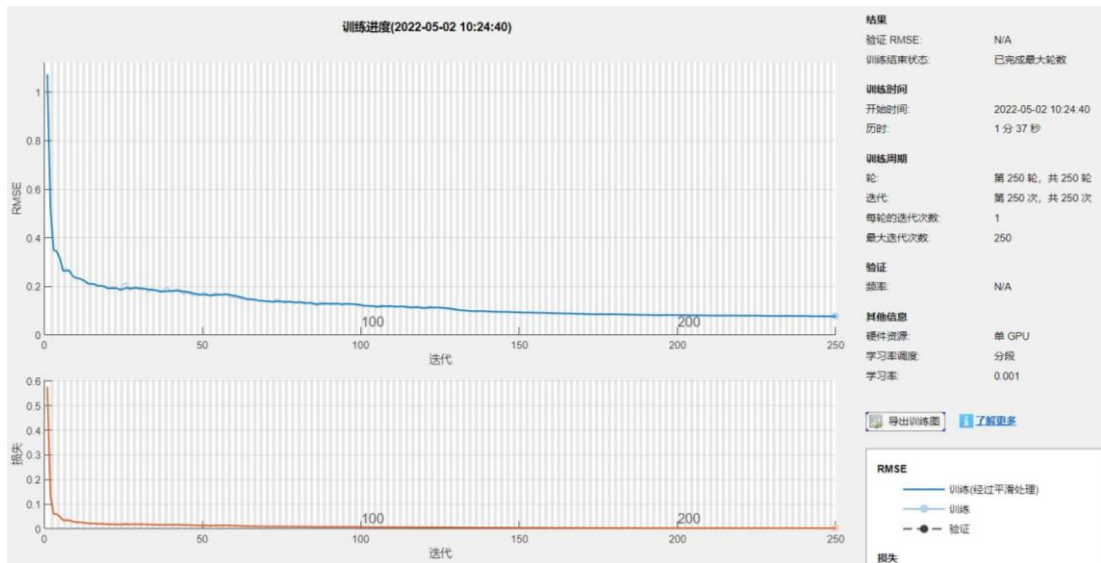


图 8 神经网络训练过程图

使用 2021 年 7 月 13 日至 2021 年 12 月 30 日的“数字经济”板块指数对基于 LSTM 神经网络的股票预测模型进行训练，从而得出 2022 年 1 月 4 日至 2022 年 1 月 28 日“数字经济”板块指数每 5 分钟成交量的部分预测值如下表所示：

表 7 部分成交量预测值与真实值统计表

时间	成交量预测值	成交量实际值
2022/1/28 15:00	14,877,705	15,251,620
2022/1/28 14:55	20,957,029	22,079,060
2022/1/28 14:50	14,992,859	16,963,480
...
2022/1/28 9:55	17,589,218	15,118,169
2022/1/28 9:50	19,013,101	19,917,294
2022/1/28 9:45	17,300,874	19,301,382
...
2022/1/17 9:55	23,300,083	22,433,560
2022/1/17 9:50	23,622,747	29,784,291
2022/1/17 9:45	28,025,037	27,987,554
...
2022/1/4 9:55	20297455	21194010

2022/1/4 9:50	23382252	24464949
2022/1/4 9:45	19479007	24013795

绘制成交量预测结果与真实值曲线对比图和训练数据测试结果残差图如下图所示：

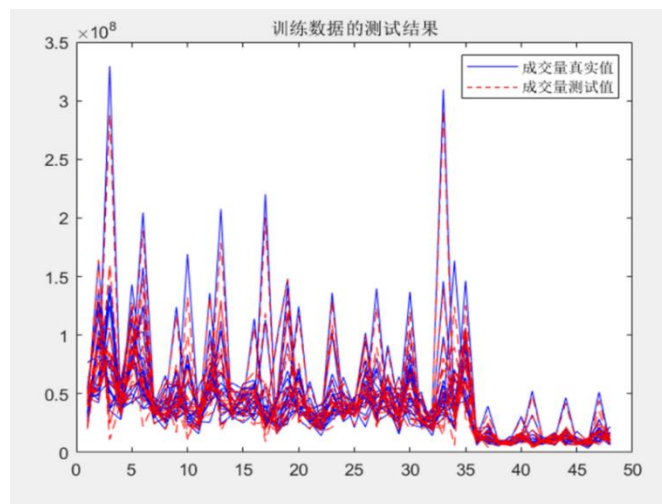


图 9 成交量预测结果与真实值曲线对比图

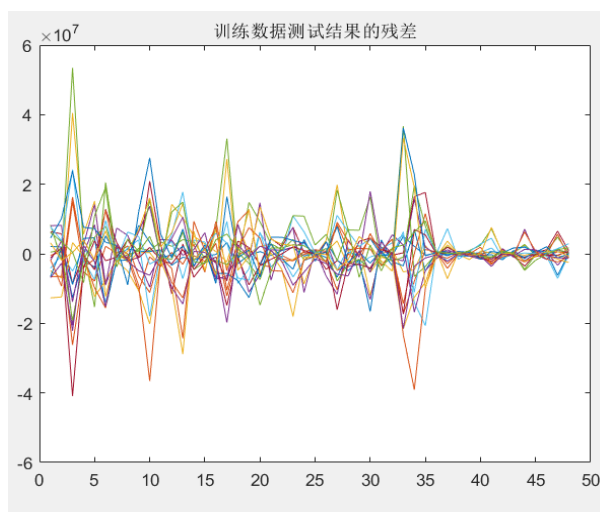


图 10 训练数据测试结果残差图

残差图中横坐标表示每天的 48 个五分钟，纵坐标表示测试结果的残差，不同颜色的折线则表示预测的 19 天数据。由上面两图可以看出，成交量的预测值与实际值十分接近，证明预测结果较为准确。

5.3 问题三模型的建立与求解

5.3.1 问题分析

问题三需要在问题一和问题二所建立的数学模型的基础上，来预测每 5 分钟的“数字经济”板块指数的收盘价。本文可直接参考问题二所建立的 LSTM 神经网络模型，对未来一段时间的收盘价进行预测。

5.3.2 基于问题二 LSTM 神经网络模型的补充

LSTM 神经网络模型的原理框图如下：

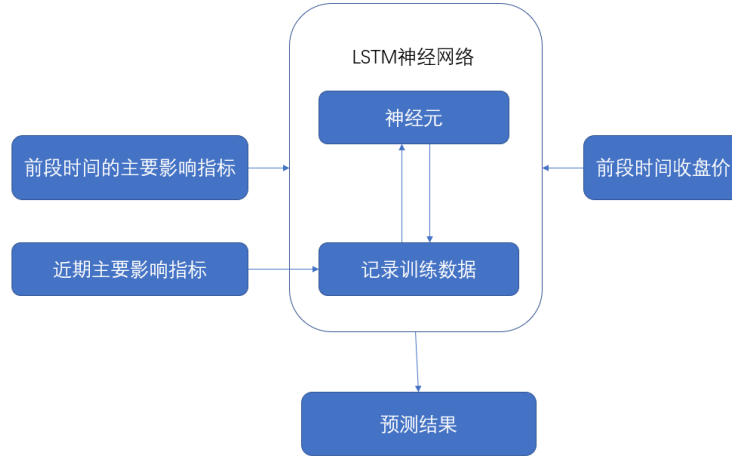


图 11 LSTM 神经网络模型原理框图

影响收盘价主要指标和前段时间收盘价是一个强非线性关系，可以用公式表示为：

$$y = e^{0.001x_1 \cdot x_2 \cdot x_3 \dots x_{n-1} \cdot x_n} \quad (22)$$

y 表示收盘价格， x_1 、 x_2 、 $x_3 \dots x_{n-1}$ 、 x_n 表示影响收盘价格的主要指标。

影响收盘价的主要指标、收盘价、时间三者是一个非线性时间序列关系，人为很难精准推测出这三者之间的关系。此时，可以用问题二的 LSTM 神经网络来解决。

由问题二可知，LSTM 神经网络对时间序列有很强的预测能力，它在训练与学习的过程中会把时间序列状态保留和传递。设时间序列为 $\{X_{nt}\}$ (n 表示第几种影响指标， t 表示时刻)，目标序列为 $\{Y_t\}$ 。在 LSTM 中可以表示为：

$$\begin{aligned} \{X_{n1}, X_{n2}, X_{n3}, \dots, X_{nm}\} &\rightarrow \{Y_1\}, \\ \{X_{n2}, X_{n3}, X_{n4}, \dots, X_{n(m+1)}\} &\rightarrow \{Y_2\}, \\ \{X_{n3}, X_{n4}, X_{n5}, \dots, X_{n(m+2)}\} &\rightarrow \{Y_3\}, \\ &\dots\dots \end{aligned} \quad (23)$$

通过成千上万这样循环训练学习，在 LSTM 神经网络系统中会保留时间序列 $\{X_{nt}\}$ 与目标序列 $\{Y_t\}$ 的一个函数。当输入时间序列时，LSTM 神经系统会输出对应的一个目标序列。

5.3.3 模型的求解

(一) 根据问题要求以 2021 年 7 月 14 日至 2021 年 12 月 31 日的每 5 分钟“数字经济”板块指数为训练集，以 2022 年 1 月 4 日至 2022 年 1 月 28 日的每 5 分钟“数字经济”板块指数为测试集。训练集的天数与测试集的天数共 130 天，有 10 个主要指标影响着收盘价。将上述时间序列做成一个 130×10 的一个矩阵。将该矩阵按训练时间与测试时间划分，前 111 天作为训练集，训练完成后来预测后 19 天的收盘价。在 matlab 中建立一个 LSTM 神经训练网络如下图所示：

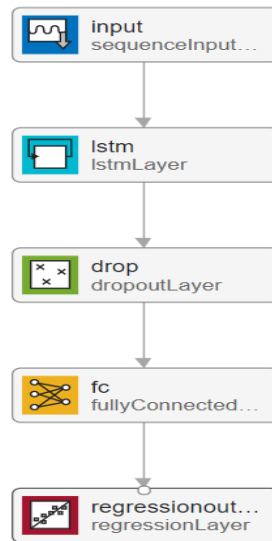


图 12 在 matlab 中建立 LSTM 神经网络流程图

训练网络搭建好后，将训练次数、学习率、单个 LSTM 中的神经元个数等训练参数设计好。将训练集与测试集归一化后输入到这个训练网络中，进行训练和预测，得到后 19 天每天每 5 分钟的收盘价，如下表所示：

表 8 分成交易预测值与真实值统计表

时间	收盘价真实值	收盘价预测值
2022/1/28 15:00	1,365.1880	1,368.9808
2022/1/28 14:55	1,366.1420	1,368.9858
2022/1/28 14:50	1,368.8356	1,373.3757
2022/1/28 14:45	1,367.4127	1,375.3030
...
2022/1/10 10:55	1464.8661	1460.782812
2022/1/10 10:50	1464.3232	1460.932777
2022/1/10 10:45	1463.0012	1460.793358
2022/1/10 10:40	1459.0932	1460.232351
...
2022/1/4 9:50	1566.6184	1551.809509
2022/1/4 9:45	1570.8784	1545.257744
2022/1/4 9:40	1572.6151	1577.348379

绘制成交量预测结果与真实值曲线对比图如下图所示：

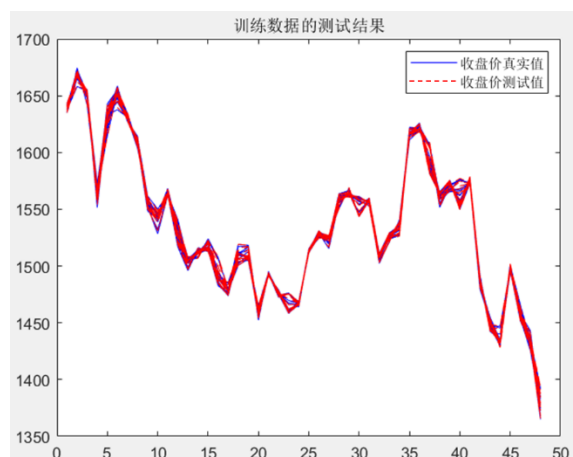


图 13 成交量预测结果与真实值曲线对比图

（二）误差分析

对上表中的预测数据与实际数据进行对比分析，预测误差率不超过 0.8%。说明预测值与真实值非常接近，该预测模型预测的值非常合理科学。预测模型具有一定的实用价值。其中，误差率计算公式如下：

$$\text{误差率} = \frac{\text{预测值} - \text{实际值}}{\text{实际值}} \quad (24)$$

绘制训练数据测试结果残差图如下图所示：

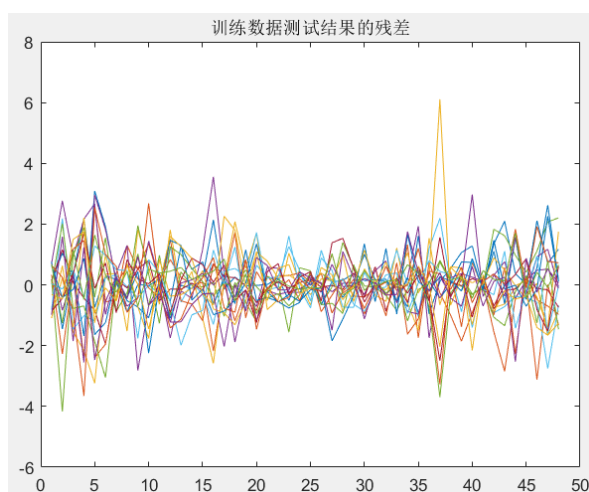


图 14 训练数据测试结果残差图

5.4 问题四模型的建立与求解

5.4.1 问题分析

问题四要求根据问题三所得结果，以 100 万元为初始资金，0.3%为交易佣金，对“数字经济”板块指数进行交易。从而得出 2022 年 1 月 4 日至 2022 年 1 月 28 日期间交易的总收益率、信息比率和最大回撤率。首先，利用问题三所得的 2022 年 1 月 4 日至 2022 年 1 月 28 日，共 852 个“数字经济”板块收盘价数据，以 100 万元为初始资金，先对总收益率进行分析。由于每笔交易需支出 0.3%的佣金，同时考虑到一般投资者想尽可能多获利的心理。本文基于数据降维法，建立收益最大化模型，给出交易者最合理的投资策略，从而得到最大的投资总收益率。接着，根据题目所给公式计算出信息比率和最大回撤率，分别反映出一段时

间的收益和该股票的投资风险。

最后，利用题目所给的 2022 年 1 月 4 日至 2022 年 1 月 28 日的实际收盘价计算相应的总收益率、信息比率和最大回撤率，与使用预测值计算的结果相比较，从而得出预测值的准确性。

5.4.2 总收益率、信息比率、最大回撤率计算方法

(一) 总收益率

设第 i 天的当日收盘总资产市值为 z_i ，则当日收益率 P_i 的计算公式如下：

$$P_i = (z_i - z_{i-1}) / z_{i-1} \quad (25)$$

那么，总收益率 P_t 的计算公式如下：

$$P_t = ((1 + P_1)(1 + P_2)(1 + P_3) \dots (1 + P_n) - 1) \times 100\% \quad (26)$$

其中， n 表示第 n 天；结合上述两公式，总收益率计算公式可化简表示为：

$$P_t = \left(\frac{z_n}{z_0} - 1 \right) \times 100\% \quad (27)$$

(二) 信息比率

设第 i 日中证 500 指数收益率为 r_i ，则第 i 日超额收益率 q_i 的计算公式如下：

$$q_i = P_i - r_i \times 90\% \quad (28)$$

设共交易 T 日，则 T 日内超额收益率的均值 \bar{q}_i 为：

$$\bar{q}_i = \frac{q_1 + q_2 + q_3 + \dots + q_T}{T} \quad (29)$$

则超额收益率标准差 s 计算公式如下：

$$s = \sqrt{\frac{(q_1 - \bar{q}_i)^2 + (q_2 - \bar{q}_i)^2 + \dots + (q_T - \bar{q}_i)^2}{T}} \quad (30)$$

日均超额收益率 q_i' 计算公式如下：

$$q_i' = \frac{q_i}{T} \quad (31)$$

信息比率 m 计算公式如下：

$$m = \frac{q_i'}{s} \quad (32)$$

(三) 最大回撤率

设 D_i 为第 i 天的产品净值， D_j 是 D_i 后面第 j 天的净值，则最大回撤率计算公式如下：

$$Drawdown = \max \frac{D_i - D_j}{D_i} \quad (33)$$

5.4.3 基于数据降维法的收益最大化模型的建立

根据题目所给计算方法，参考上述公式分析可知，要想得到最大总收益率，仅需求出 z_n 的最大值即可，即求出第 n 天收盘总资产市值的最大值。

首先列出由第三问预测所得的每五分钟收盘价，分析每相邻五分钟收盘价的涨跌情况，对其进行类别划分，方法如下：

- 针对上述分类结果，对类别一和类别二数据进行二次归类。分别计算出类别一和类别二每段数据起始和终止时刻的收盘价差，判断其差值是否

大于 0.3%的交易佣金。其中，将上涨幅度大于 0.3%的类别标记为 1；将下跌幅度大于 0.3%的类别标记为 2；上涨及下跌幅度低于 0.3%的类别标记为 0。

- 再对类别三进行二次归类。计算类别三中每段数据的最大收盘价和最小收盘价的差值，判断其差值是否大于 0.3%的交易佣金。同样，将上涨幅度大于 0.3%的类别标记为 1，下跌幅度大于 0.3%的类别标记为 2，上涨及下跌幅度低于 0.3%的类别标记为 0。

模型基本结构如下图所示：

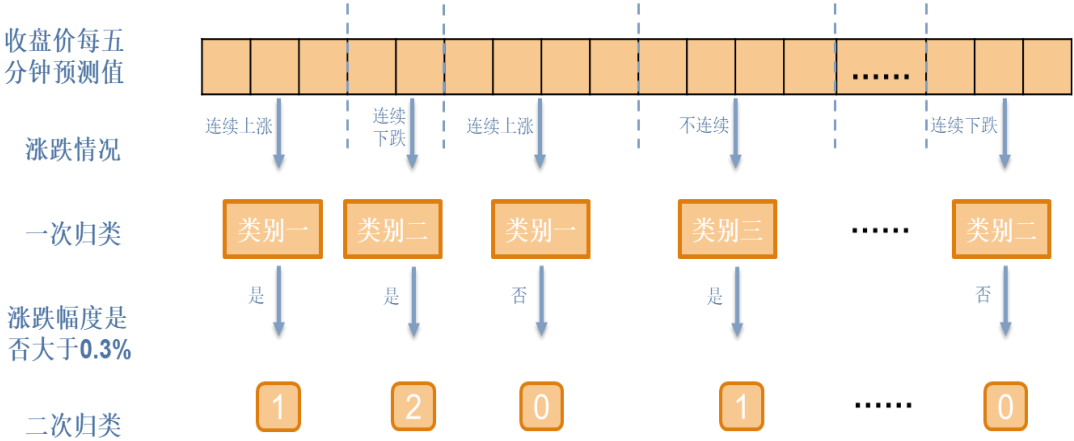


图 15 基于数字降维法的收益最大化模型基本结构图

针对上述结果，采用高抛低吸法。以每一类别为一个阶段。若未来阶段收盘价上涨超过 0.3%则买入，下跌超过 0.3%则卖出，变化幅度低于 0.3%则持有。即 1 表示买入，2 表示卖出，0 表示持有。

5.4.4 模型的求解

（一）首先，根据上述模型，利用问题三所得数据对总收益率和最大回撤率进行计算，计算部分结果如下表所示：

表 9 总收益率和最大回撤率	
总收益率	最大回撤率
0.0518	0.066200706

信息比率部分计算结果如下表所示：

表 10 信息比率				
日期	中证 500 指数 收益率	日均超额收 益率	信息比率	超额收益率 标准差
2022-01-28	-0.0038	8.76E-05	0.0067	0.0130
2022-01-27	-0.0260	1.20E-03	0.0918	
2022-01-26	0.0057	-2.65E-04	-0.0203	
2022-01-25	-0.0337	1.38E-03	0.1059	
...	
2022-01-07	-0.0075	2.29E-04	0.0176	
2022-01-06	0.0010	3.65E-04	0.0280	
2022-01-05	-0.0179	8.47E-04	0.0650	
2022-01-04	-0.0007	2.42E-03	0.1853	

(二) 根据真实值数据对总收益率和最大回撤率进行计算, 计算部分结果如下表所示:

表 11 总收益率和最大回撤率	
总收益率	最大回撤率
0.0942	0.0861

信息比率部分计算结果如下表所示:

表 12 信息比率表				
日期	中证 500 指数收益率	日均超额收益率	信息比率	超额收益率标准差
2022-01-28	-0.0038	0.0008	0.1332	0.00637
2022-01-27	-0.0260	0.0005	0.0940	
2022-01-26	0.0057	0.0008	0.1382	
2022-01-25	-0.0337	0.0011	0.1799	
...	
2022-01-07	-0.0075	0.0004	0.0746	0.00637
2022-01-06	0.0010	0.0002	0.0391	
2022-01-05	-0.0179	0.0007	0.1245	
2022-01-04	-0.0006	0.0001	0.0270	

综上所述, 通过预测值计算所得的指标数据与用真实值计算的指标数据相差不大, 且总收益率相对较高, 最大回撤率较低。由此可反映出问题二、三中建立的预测模型和本文中建立的组合交易模型的优良性。

六、模型的评价

6.1 模型的优点

1. 在求解问题时, 对题目所给数据做了充分的预处理, 以保证对数据的更好利用和所得结果的准确性。
2. 问题二的求解中, 使用了 LSTM 神经网络, 相较于 BP 神经网络更具优势。
3. 在问题四的求解中, 利用模型求出了最大总收益率, 考虑到了大多数投资者对于股票获利的需求。同时, 上述模型可推广使用于其他各项投资交易情景之中。
4. 综合运用了 matlab、Excel、spss 等多个数据处理软件, 结合各软件优势, 提高效率, 使分析结果更准确。

6.2 模型的缺点

1. 本文求解前做了适当的假设, 难免造成一定的误差。
2. 本文使用了 LSTM 的神经网络, 但若训练集数据过少, 则训练效果不佳, 预测不准确。
3. 问题一使用了皮尔逊相关系数, 该系数对异常值比较敏感, 可能对分析造成一定影响。

七、模型的推广

此模型适用于股票、基金、期货市场中任意板块，可以对其进行板块发展趋势预测以及制定较准确的投资方案。同时，对于大多数已知影响因素的投资项目或指标，可直接套用该模型进行分析和预测。

八、参考文献

- [1]杨丹.数据拟合方法探讨[J].大众标准化,2020(10):144-145.
- [2]陶洋,祝小钧,杨柳.基于皮尔逊相关系数和信息熵的多传感器数据融合[J/OL].小型微型计算机系统:1-7[2022-05-01].
- [3]杨波,梁伟.深度神经网络模型构建及优化策略[J].计算机时代,2022.
- [4]林升. 基于 LSTM 的股票预测研究[D].广州大学,2019.
- [5]Ghosh Pushpendu,Neufeld Ariel,Sahoo Jajati Keshari. Forecasting directional movements of stock prices for intraday trading using LSTM and random forests[J]. Finance Research Letters,2022,46(PA).
- [6]谭浩强主编. C 语言程序设计[M].清华大学出版社,1999.12
- [7]刘卫国主编. MATLAB 程序设计与应用（第 2 版）[M]. 北京高等教育出版社,2006

附录：
支撑材料目录

1	程序源码
2	模型结果结论
3	作品 Word 文档

附录 1:
问题一模型解决方案代码
<pre> load 'A'; %导入板块指标数据 load 'A1'; load 'A2'; load 'A3'; load 'A4'; load 'A5'; a=0; k=1; % 求数字经济板块每天的数值总和 for i=1:6 for j=1:6240 a=A(j,i)+a; if(mod(j,48)==0) B(k,i)=a; k=k+1; a=0; end end k=1; end % 数字经济板块求最大值 D=max(B); D1=min(B); % 求最小值 %无量钢化，使指标达到标准一致 for i=1:6 for j=1:130 C(j,i)=(B(j,i)-D1(1,i))/(D(1,i)-D1(1,i)); end end % 国际市场指标求最大数值 D2=max(A1); D3=min(A1); % 求最小值 for i=1:12 %无量钢化 for j=1:130 C1(j,i)=(A1(j,i)-D3(1,i))/(D2(1,i)-D3(1,i)); end end end </pre>

```

% 技术指标求最大数值
D4=max(A2);
D5=min(A2); % 求最小值
for i=1:14 %无量纲化
    for j=1:130
        C2(j,i)=(A2(j,i)-D5(1,i))/(D4(1,i)-D5(1,i));
    end
end
% 国际市场指标求最大数值
D6=max(A3);
D7=min(A3); % 求最小值
for i=1:11 %无量纲化
    for j=1:130
        C3(j,i)=(A3(j,i)-D7(1,i))/(D6(1,i)-D7(1,i));
    end
end
% 汇率指标求最大值
D8=max(A4);
D9=min(A4); % 求最小值
for i=1:2 %无量纲化
    for j=1:130
        C4(j,i)=(A4(j,i)-D9(1,i))/(D8(1,i)-D9(1,i));
    end
end
% 其他板块指标求最大值
D10=max(A5);
D11=min(A5); % 求最小值
for i=1:5 %无量纲化
    for j=1:130
        C5(j,i)=(A5(j,i)-D11(1,i))/(D10(1,i)-D11(1,i));
    end
end

load 'F'; %无量纲化后总的数字经济板块指数
load 'M'; %无量纲化后的指标总集
for j=1:6 %求各个相关性
    for i=1:44
        P(j,i)=corr(F(:,j),M(:,i),'type','pearson');
    end
end
for i=1:6 %相关性进行转置
    for j=1:44
        P1(j,i)=P(i,j);
    end
end

```



```

end
[B2 index2]=sort(P1); %正常从小到大排序
P2=abs(P1); %求绝对值
[B1 index1]=sort(P2); %绝对值从小到大排序
for i=1:44 %计算平均值相关性并经过排序
    a=0;
    for j=1:6
        a=a+P2(i,j);
    end
    P3(i,1)=a/6;
end
[B3 index3]=sort(P3);

```

附录 2:**问题二模型解决方案代码**

```
k=1; %排序收盘价训练集
for i=1:130 %导入各个指标
    for j=1:48
        D1(k,1)=X1(i,1); D2(k,1)=X2(i,1);
        D3(k,1)=X3(i,1); D4(k,1)=X4(i,1);
        D5(k,1)=X5(i,1); D6(k,1)=X6(i,1);
        D7(k,1)=X7(i,1); D8(k,1)=X8(i,1);
        D9(k,1)=X9(i,1); D10(k,1)=X10(i,1);
        k=k+1;
    end
end
k=1; %排序成交量训练集
for i=1:130
    for j=1:48
        W1(k,1)=Y1(i,1); W2(k,1)=Y2(i,1);
        W3(k,1)=Y3(i,1); W4(k,1)=Y4(i,1);
        W5(k,1)=Y5(i,1); W6(k,1)=Y6(i,1);
        W7(k,1)=Y7(i,1); W8(k,1)=Y8(i,1);
        W9(k,1)=Y9(i,1); W10(k,1)=Y10(i,1);
        W11(k,1)=Y11(i,1); W12(k,1)=Y12(i,1);
        k=k+1;
    end
end
j=1; %求上下反转各个训练集
for i=6240:-1:1
    L1(j,1)=D1(i,1); L2(j,1)=D2(i,1);
    L3(j,1)=D3(i,1); L4(j,1)=D4(i,1);
    L5(j,1)=D5(i,1); L6(j,1)=D6(i,1);
    L7(j,1)=D7(i,1); L8(j,1)=D8(i,1);
    L9(j,1)=D9(i,1); L10(j,1)=D10(i,1);
    M1(j,1)=W1(i,1); M2(j,1)=W2(i,1);
    M3(j,1)=W3(i,1); M4(j,1)=W4(i,1);
    M5(j,1)=W5(i,1); M6(j,1)=W6(i,1);
    M7(j,1)=W7(i,1); M8(j,1)=W8(i,1);
    M9(j,1)=W9(i,1); M10(j,1)=W10(i,1);
    M11(j,1)=W11(i,1); M12(j,1)=W12(i,1);
    C(j,1)=CG(i,1); S(j,1)=SP(i,1); j=j+1;
end

%预测代码
load 'T1';%时间序列
load 'T';%目标序列
```

```

A=T; B=T1; x=A'; y=B';
[d,mse1,test_ty]=shengjingwangluo(x,y)
yuce=test_ty;
D=d; M=mse1;
disp('预测值') yuce
disp('均方误差')
function [d,mse1,test_ty]=shengjingwangluo(x,y)
trainx =x; trainy =y;
[ww,mm]=size(trainx); testx = x;
%% 创建神经网络
% 包含 15 个神经元，训练函数为 traingdx
net=elmannet(1:2,15,'traingdx');
% 设置显示级别
net.trainParam.show=1;
% 最大迭代次数为 10000 次
net.trainParam.epochs=1000;
% 误差容限，达到此误差就可以停止训练
net.trainParam.goal=0.00001;
% 最多验证失败次数
net.trainParam.max_fail=5;
% 对网络进行初始化
net=init(net);
%% 网络训练
%训练数据归一化
[trainx1, st1] = mapminmax(trainx);
[trainy1, st2] = mapminmax(trainy);
% 测试数据做与训练数据相同的归一化操作
testx1 = mapminmax('apply',testx,st1);
% 输入训练样本进行训练
[net,per] = train(net,trainx1,trainy1);
%% 测试。输入归一化后的数据，再对实际输出进行反归一化
% 将训练数据输入网络进行测试
train_ty1 = sim(net, trainx1);
train_ty = mapminmax('reverse', train_ty1, st2);
% 将测试数据输入网络进行测试
test_ty1 = sim(net, testx1);
test_ty = mapminmax('reverse', test_ty1, st2);

%% 显示结果
% 1.显示训练数据的测试结果
figure(1)
x=1:length(train_ty);
% 显示真实值
plot(x,trainy,'b-');

```

```

hold on
% 显示神经网络的输出值
plot(x,train_ty,'r--')
legend('收盘价真实值','收盘价测试值')
title('训练数据的测试结果');
% 显示残差
figure(2)
plot(x, train_ty - trainy)
title('训练数据测试结果的残差')
% 显示均方误差
%mse1 = mse(train_ty - trainy);
%fprintf('mse = \n %f\n', mse1)
% 显示相对误差
disp('相对误差: ')
d=(train_ty - trainy)./trainy;
fprintf('%f ', (train_ty - trainy)./trainy );
fprintf('\n')
% 显示预测值
disp('预测值: ')
fprintf('%f ', test_ty );
fprintf('\n')
mse1 = mse(train_ty - trainy);
fprintf('mse = \n %f\n', mse1)

```

附录 3:

问题三模型解决方案代码

```
load 'F1'; load 'F';
A=F; B=F1; x=A' y=B'
%输出目标数据 8 天对应的 48 个成交量 %m 为检测点的个数
[d,mse1,test_ty]=shengjingwangluo(x,y)
yuce=test_ty;
%yucezhi
D=d;
M=mse1;
disp('预测值') yuce
disp('均方误差')
function [d,mse1,test_ty]=shengjingwangluo(x,y)
trainx =x; trainy =y;
[ww,mm]=size(trainx); testx = x;
%% 创建神经网络
% 包含 15 个神经元, 训练函数为 traingdx
net=elmannet(1:2,15,'traingdx');
% 设置显示级别
net.trainParam.show=1;
% 最大迭代次数为 10000 次
net.trainParam.epochs=5000;
% 误差容限, 达到此误差就可以停止训练
net.trainParam.goal=0.00001;
% 最多验证失败次数
net.trainParam.max_fail=5;
% 对网络进行初始化
net=init(net);
%% 网络训练
%训练数据归一化
[trainx1, st1] = mapminmax(trainx);
[trainy1, st2] = mapminmax(trainy);
% 测试数据做与训练数据相同的归一化操作
testx1 = mapminmax('apply',testx,st1);
% 输入训练样本进行训练
[net,per] = train(net,trainx1,trainy1);
%% 测试。输入归一化后的数据, 再对实际输出进行反归一化
% 将训练数据输入网络进行测试
train_ty1 = sim(net, trainx1);
train_ty = mapminmax('reverse', train_ty1, st2);
% 将测试数据输入网络进行测试
test_ty1 = sim(net, testx1);
test_ty = mapminmax('reverse', test_ty1, st2);
% 1.显示训练数据的测试结果
```

```

figure(1)
x=1:length(train_ty);
plot(x,trainy,'b-'); hold on % 显示真实值
plot(x,train_ty,'r--') % 显示神经网络的输出值
legend('成交量真实值','成交量测试值')
title('训练数据的测试结果'); hold on
figure(2) % 显示残差
plot(x, train_ty - trainy)
title('训练数据测试结果的残差')
mse1 = mse(train_ty - trainy); % 显示均方误差
fprintf('mse = \n %f\n', mse1)
disp(' 相对误差: ') % 显示相对误差
d=(train_ty - trainy)./trainy;
fprintf('%f ', (train_ty - trainy)./trainy );
fprintf('\n')
disp(' 预测值: ') % 显示预测值
fprintf('%f ', test_ty );
fprintf('\n')

```

附录 4:**问题四模型解决方案代码**

```
%求各个经济板块指数增长率
for i=912:-1:2
    L(i-1,1)=(K(i-1,1)-K(i,1))/K(i,1);
end
%将数据按时间从小到大，进行上下反转
j=1;
for i=911:-1:1
    L1(j,1)=L(i,1); j=j+1;
end
% 第一次排序，将大于 0 的作为一个参考集合
e=0;
for i=1:911
    if(L1(i,1)>0)
        e=e+L1(i,1);
    end
    if(L1(i,1)<0)
        L1(i,2)=e; e=0;
    end
end
%再次排序，将小于 0 的作为一个参考集合
e=0;
for i=1:911
    if(L1(i,1)<0)
        e=e+L1(i,1);
    end
    if(L1(i,1)>=0)
        L1(i,2)=e; e=0;
    end
end
%分组排序一次完毕
for i=1:911
    if(U1(i,1)>0.003)
        U1(i,2)=1;
    end
    if(U1(i,1)<-0.003)
        U1(i,2)=2;
    end
    if(U1(i,1)>=-0.003&&U1(i,1)<=0.003)
        U1(i,2)=0;
    end
end
%求中间总和
```

```

t=0;
for i=1:911
    if(U1(i,2)==0)
        t=t+U1(i,1);
    end
    if(U1(i,2)~=0)
        U1(i-1,3)=t; t=0;
        U1(i,3)=U1(i,1);
    end
end
%第二次排序，按上面的方法编号
for i=1:911
    if(X(i,1)>0.003)
        X(i,2)=1;
    end
    if(X(i,1)<-0.003)
        X(i,2)=2;
    end
    if(X(i,1)>=-0.003&&X(i,1)<=0.003)
        X(i,2)=0;
    end
end
%求进行多次分层后，求最后分组情况
for i=1:911
    if(X(i,2)==1)
        X(i,3)=X(i,1);
    end
    if(X(i,2)==0)
        X(i,3)=X(i,1);
    end
end
for i=1:911    %求和排序
    X(i,5)=X(i,3)+X(i,4);
end
j=1;    %求下上反转
for i=912:-1:1
    G(i,1)=P(j,1);
    j=j+1;
end
for i=1:912    %分类，1 为买，2 为卖
    if(G(i,1)>0.003)
        G(i,2)=1;
    end
    if(G(i,1)<-0.003)

```



```

        G(i,2)=2;
    end
    if(G(i,1)>=-0.003&&G(i,1)<=0.003)
        G(i,2)=0;
    end
end
%初始金额为 1000000 元，1 月 4 日~1 月 28 日每五分钟交易开盘总市值
G(912,3)=1000000;
for i=911:-1:1
    if(G(i,2)==1&&G(i+1,2)==0)    %根据分层，与预测值求每 5 分钟的收益
总值
        G(i,4)=G(i+1,3)*(1-0.003)*(1+G(i,1));
    end
    if(G(i,2)==0&&G(i+1,2)==0)
        G(i,3)=G(i+1,3)*(1+G(i,1));
    end
    if(G(i,2)==2&&G(i+1,2)==0)
        G(i,3)=G(i+1,3)*(1-0.003);
    end
    if(G(i,2)==2&&G(i+1,2)==1)
        G(i,3)=G(i+1,3)*(1-0.003);
    end
    if(G(i,2)==1&&G(i+1,2)==2)
        G(i,3)=G(i+1,3)*(1-0.003)*(1+G(i,1));
    end
    if(G(i,2)==0&&G(i+1,2)==1)
        G(i,3)=G(i+1,3)*(1+G(i,1));
    end
    if(G(i,2)==0&&G(i+1,2)==2)
        G(i,3)=G(i+1,3)*(1+G(i,1));
    end
end
%求最大总收益率，收盘价日期按从大到小排序，H（为总收益 20 个，也需
日期按从大到小排序）
load 'H'; %预期的收盘价
for i=19:-1:1
    L5(i,1)=(H(i,1)-H(i+1,1))/H(i+1,1);    %当日收益率
end
sum=1;
for i=19:-1:1    %计算总收益率
    sum=sum*(1+L5(i,1));
end
sum=sum-1;

```

```

%求当日中证 500 指数收益率
load 'Y'; %导入中证 500 指数
for i=19:-1:1
    L6(i,1)=(Y(i,1)-Y(i+1,1))/Y(i+1,1);
end
%求超额收益率
for i=1:19
    L7(i,1)= L5(i,1)-L6(i,1)*0.9;
end
%求日均超额收益率
for i=1:19
    L8(i,1)=L7(i,1)/19; %19 为交易日天数
end
D=std(L7); %求超额收益率标准差
for i=1:19 %求信息比率
    L9(i,1)=L8(i,1)/D;
end
MaxDD = maxdrawdown(H(:,1)); % 计算最大回撤
risk(1) = MaxDD % 将最大回撤赋值给 risk，作为股票的风险评估

```