武汉大学 经济与管理学院
Economics and Management School of Wuhan University

时间序列方法+机器学习在股指预测中的应用

——基于LSTM-RF Regression方法

汇 报 人 ： 刘郅哲（第11组）
专 业 ： 工商管理类
学 号 ： 2020301052098
时 间 ： 2022-5-31

目录

1 研究背景及目的

沪深300指数

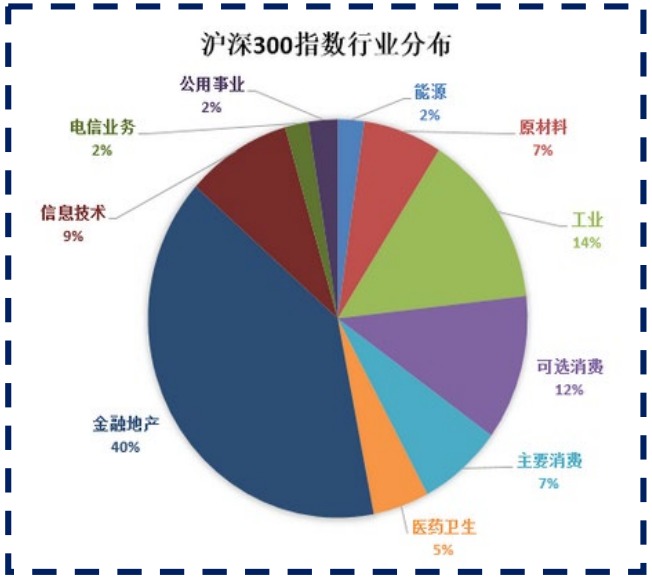
沪深300指数

000300.SH，399300.SZ

是中证指数有限公司编制的

用以反映沪深两市价格变动

总览的**跨市场指数**。



指数样本选自沪深两个证券市场，覆盖了大部分流通市值。成份股为市场中**市场代表性好**，**流动性高**，**交易活跃**的主流投资股票，能够反映市场主流投资的收益情况。

金融时间序列预测

时间序列方法

移动平均自回归模型

广义自回归条件异方差模型

指数平滑模型

深度学习方法

多层感知机

卷积神经网络

循环神经网络

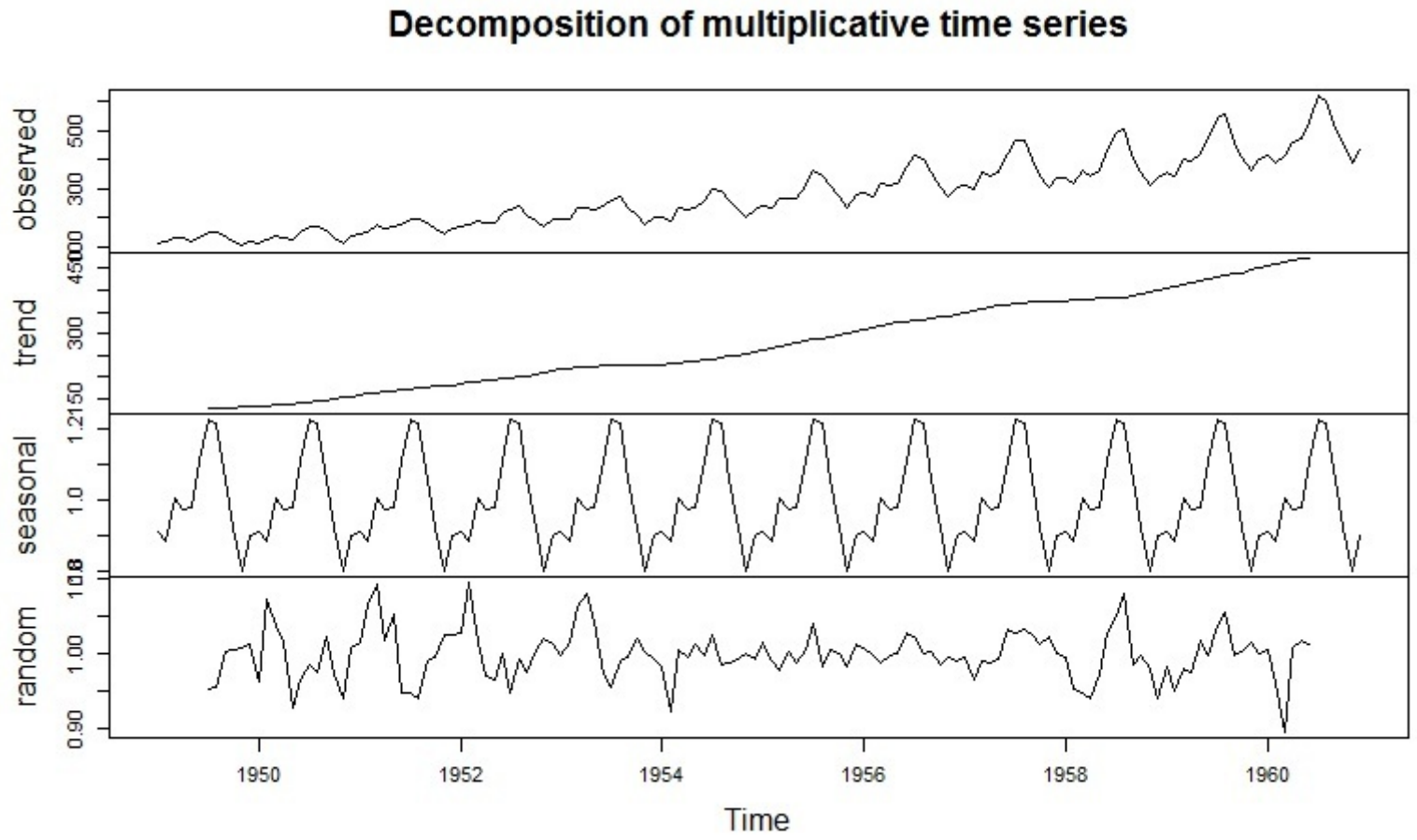
注意力机制

组合方法

ARIMA+SVR

GARCH+SVR

经济理论基础及研究目的



时序分解

- 神经网络处理时序趋势信息
- 机器学习处理其他市场信息

研究应用

- 量化交易 – 择时
- 风险管理 – 异常值

2 数据收集与预处理

表 1：变量表

特征	变量名称
开盘价	*open*
收盘价	*close*
日内最高价	*high*
日内最低价	*low*
当日涨跌幅	*ret*
换手率	*turnrate*
成交量	*volume*
动态买卖气指标	*ADTM*
均幅指标	*ATR*
顺势指标	*CCI*
异同移动平均	*MACD*
动量指标	*MTM*
变动率指标	*ROC*
盈潮-S 指标	*SOBV*
标准差（26 周期）	*STD_26*
标准差（5 周期）	*STD_5*

数据来源

数据采集于Wind数据库与Choice数据库，选取了常见量价指标与衍生技术指标、市场情绪指标的16个特征

样本数量

自2010-01-04至2022-5-17共3003个交易日的样本

数据预处理

归一化

采用极值归一化方法，将全部数据映射入$(-1,1)$

张量化

将数据转化为张量形式，方便传进神经网络处理

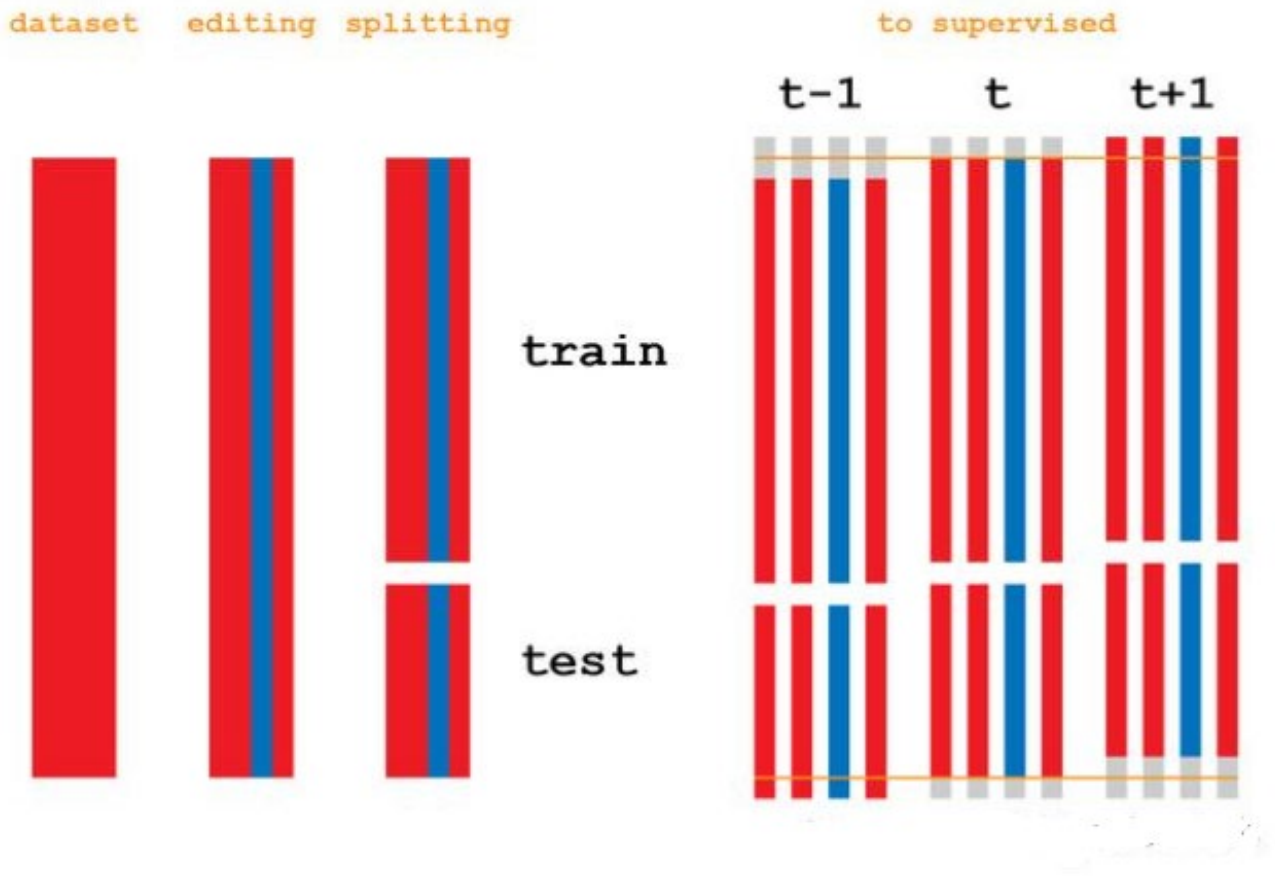
窗口化

设置滑动输入窗口为30，输出窗口为1，符合神经网络的输入规范

批次化

对数据标记传入批次，标记在新的张量维度

序列分割



数据集划分

在长序列上进行7：3的数据划分，得到：

训练集为[2371, 2, 30]的张量

测试集为[570, 2, 30]的张量

有监督学习

序列分割后转化为有监督学习问题

3 数据描述性统计

表 2：变量描述性统计表

Variables	Min	Max	Mean	Median
open	*-0.2386*	*1.0000*	*0.3540*	*0.4282*
close	*-0.2242*	*1.0000*	*0.3957*	*0.4708*
high	*-0.2303*	*1.0000*	*0.3631*	*0.4351*
low	*-0.2053*	*1.0000*	*0.4072*	*0.4851*
ret	*-0.8879*	*0.8646*	*0.1318*	*0.1380*
volume	*-0.7785*	*0.1854*	*-0.5455*	*-0.5771*
turnrate	*-0.9049*	*-0.0976*	*-0.7101*	*-0.7360*
ADTM	*-0.9768*	*0.9559*	*0.2235*	*0.2848*
ATR	*-0.9517*	*0.5744*	*-0.6919*	*-0.7400*
CCI	*-1.0000*	*0.8080*	*0.0725*	*0.0916*
MACD	*-0.4227*	*0.9582*	*0.1225*	*0.1163*
MTM	*-0.3076*	*0.9834*	*0.2463*	*0.2534*
ROC	*-0.6364*	*0.6266*	*-0.0410*	*-0.04074*
SOBV	*0.6398*	*1.0000*	*0.8497*	*0.8391*
STD_26	*-0.85681*	*-0.0237*	*-0.5389*	*-0.6154*
STD_5	*-0.9627*	*0.3265*	*-0.6695*	*-0.7192*

偏度

中位数-均值间距离

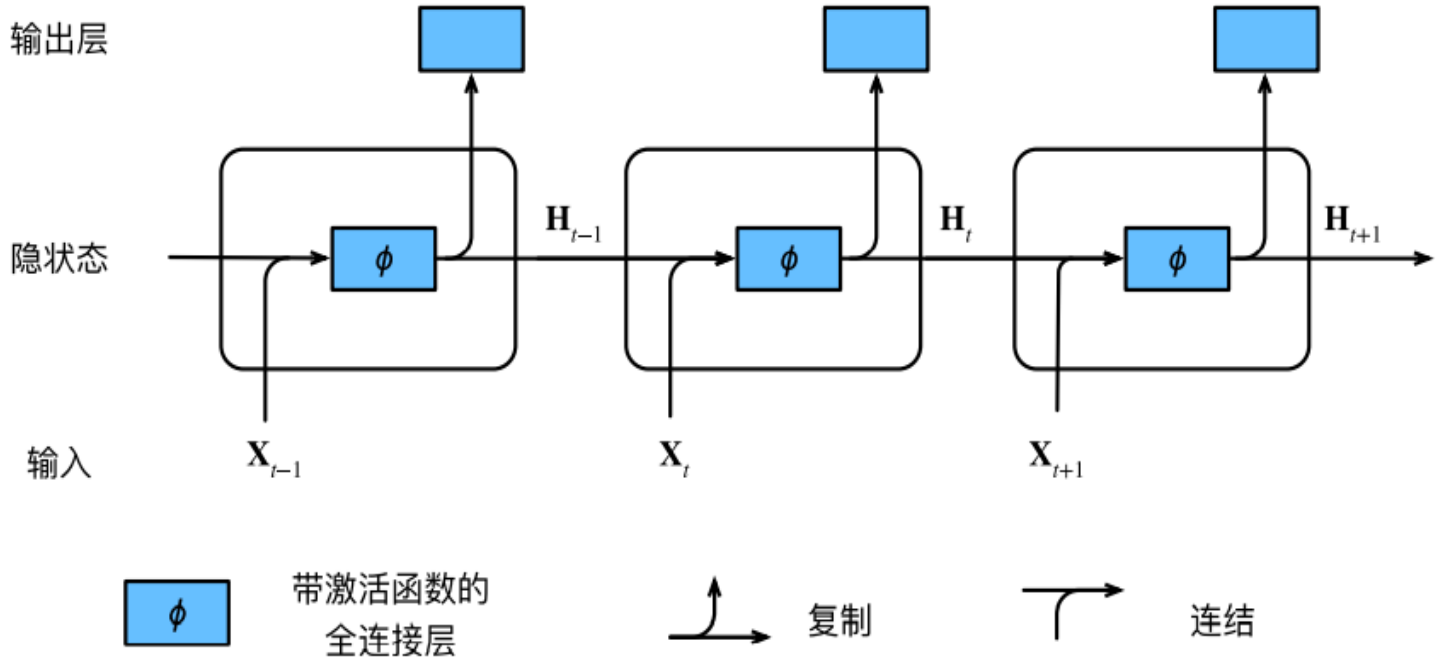
趋势

上升趋势

近期表现

近期跌幅较大

循环神经网络（RNN）



循环神经网络

处理序列数据的神经网络

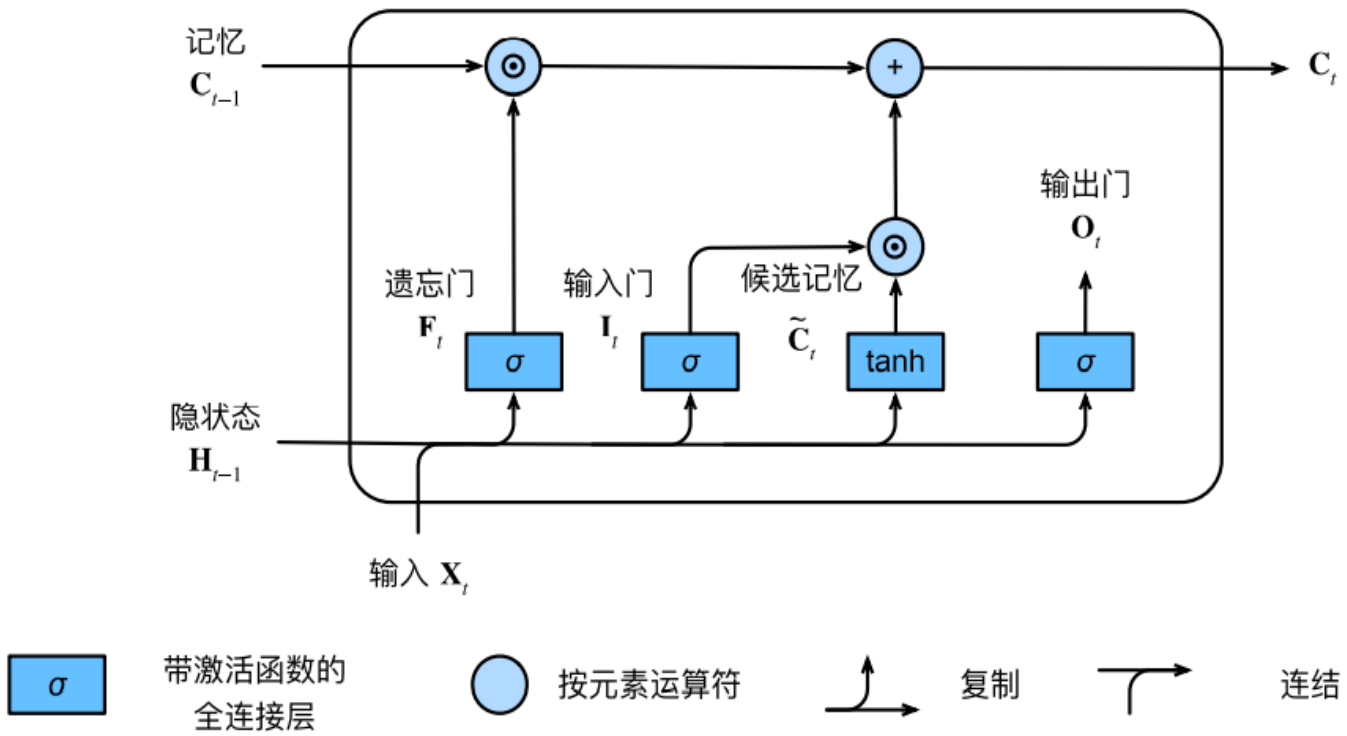
隐状态

含有过去时间信息的状态

循环层

循环计算隐状态的网络层

长短期记忆神经网络（LSTM）



长短期记忆神经网络

处理长序列训练中的梯度消失、梯度爆炸问题

门控结构

输入门、输出门、遗忘门

记忆元

对过去记忆元信息进行处理

4 模型构建


```
class LSTM(nn.Module):
    def __init__(self, input_dim = 30, hidden_layer_dim = 100, output_dim = 1):
        super().__init__()

        self.hidden_layer_dim = hidden_layer_dim

        self.lstm = nn.LSTM(input_dim, hidden_layer_dim).cuda()

        self.linear = nn.Linear(hidden_layer_dim, output_dim).cuda()

        self.hidden_cell = self.init_hidden()

    def init_hidden(self):
        return(torch.zeros(1, 1, self.hidden_layer_dim).cuda(),
               torch.zeros(1, 1, self.hidden_layer_dim).cuda())

    def forward(self, input_seq):
        lstm_out, self.hidden_cell = self.lstm(input_seq.view(len(input_seq), 1, -1),
                                               self.hidden_cell)

        predictions = self.linear(lstm_out.view(len(input_seq), -1))

        return predictions[-1]
```

类的继承

pytorch库中，nn.lstm作为衍生结构继承了nn.rnn特征

参数定义

输入长度、输出长度与设置窗口长度相等，分别为30与1

GPU训练

网络模型训练

梯度裁剪

采用裁剪方法，防止在反向传播过程中，随层数增加而产生的梯度爆炸问题

动态学习率

在学习过程中，随训练步数的增加不断降低学习率，令训练初期模型收敛较快，后期趋于稳定

定义超参

该模型训练过程需要定义的超参为初始学习率lr，学习率衰减率gamma，总步数epochs

模型结果

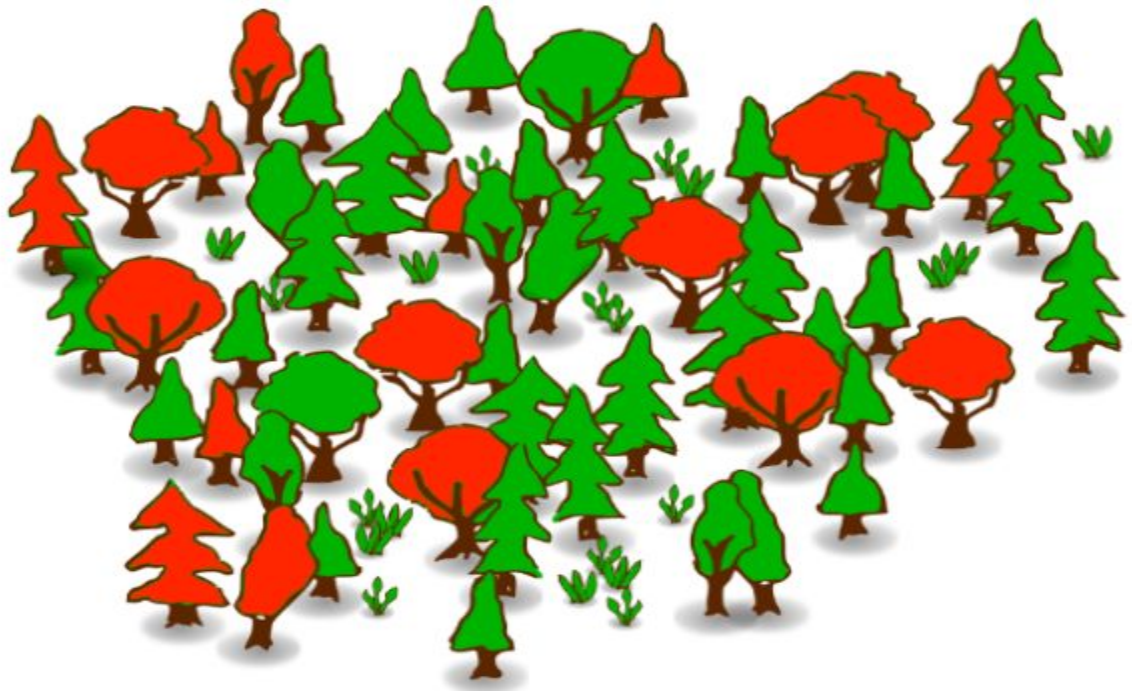
预测问题

预测延后、不够精确

单一LSTM网络不能精确预测股指序列

引入随机森林

随机森林回归（Random Forest Regression）



随机

随机取样本，随机取特征

森林

生成多棵回归树，由投票加权得出最终预测

4 模型构建


```
forestmodel <- randomForest(diff ~ ret + turnrate + ADTM + ATR + CCI + MACD + MTM + ROC
+ SOBV+ STD_26 + STD_5,
                            data_train,
                            ntree = 1000,
                            mtry = 3,
                            nodesize = 25)
for(i in 1:168){
  predictForest[i] <- predict(forestmodel, type = "response", newdata = data_test[i,])
  data_train <- bind_rows(data_train, data_test[i,])
  forestmodel <- randomForest(diff ~ ret + turnrate + ADTM + ATR + CCI + MACD + MTM +
ROC + SOBV+ STD_26 + STD_5,
                            data_train,
                            ntree = 1000,
                            mtry = 3,
                            nodesize = 25)
}
```

预测目标

LSTM模型未能解释的残差

超参定义

ntree = 1000,

nodesize = 25, mtry = 3

滚动建模

预测后将实际值纳入回训练集，迭代训练、预测

模型结果

表 3：不同模型的均方误差表

Model	$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$
LSTM	0.00664
RF	0.16556
LSTM-RF	0.00129

随机森林独立预测

对趋势的解释效果较差

LSTM-RF

对残差序列拟合，将随机森林产生的拟合值与LSTM的预测值加总，得到新的预测

均方误差

LSTM-RF模型表现最佳

5 研究结论

- **时间序列方法预测趋势+机器学习拟合特征**的思路是正确的，模型效果也是有效的；
- 尽可能**克服了 LSTM 的时滞弱点**，在随机森林回归模型的加入下，整体鲁棒性较高；
- 较为精确的预测能够给指数产品的**量化投资**提供方向指引与研判；
- 在**风险管理**领域，能够识别较大程度的跌幅，有利于机构与个人避险。

6 优化展望

- 参数调优仍然需要探索，LSTM模型细节仍待优化，考虑后续引入同源的**GRU模型**或引入**注意力机制（Transformer）**进行预测对比；
- 鉴于神经网络的高度可调节性，可以添加多种**非同质信息**作为神经网络的输入，附加**小波分解或主成分分析**等数据预处理技术进行模型优化。

谢谢观看

自强，弘毅，求是，拓新

汇 报 人 ： 刘郅哲
时 间 ： 2022-4-11