

基于异构自回归模型的股市波动率研究

王 帅

河北工业大学, 天津

Email: 469151532@qq.com

收稿日期: 2021年3月22日; 录用日期: 2021年4月11日; 发布日期: 2021年4月27日

摘 要

本文以传统的异构自回归模型(HAR)为基础, 考虑到新闻信息文本对于股票市场波动的影响, 构建了新的情感词典, 将情绪指标(AI)与基础的HAR模型相结合, 并将深度学习中具有学习长短期依赖状态特点的门控循环单元(GRU)模型引入进来, 形成了GRU-AI模型。然后介绍了波动率预测模型中的评估方法——损失函数法以及模型信度集(MCS)检验法。最后进行实证分析表明将情绪指标加入到波动率模型中可以提高模型样本外的预测精度。

关键词

波动率, 异构自回归模型, 情绪一致指标, GRU模型

The Study on Stock Market Volatility Based on Heterogeneous Autoregressive Model

Shuai Wang

Hebei University of Technology, Tianjin

Email: 469151532@qq.com

Received: Mar. 22nd, 2021; accepted: Apr. 11th, 2021; published: Apr. 27th, 2021

Abstract

Based on the traditional heterogeneous autoregressive model (HAR), considering the impact of news information text on stock market fluctuations, this paper constructs a new emotion dictionary, combines emotion indicator (AI) with the basic HAR model, and introduces the Gated Recycle Unit (GRU) model, which has the characteristics of long and short term dependent state learning in deep learning, to form the GRU-AI model. And then the evaluation methods of volatility prediction model—loss function method and model reliability set (MCS) test method are intro-

文章引用: 王帅. 基于异构自回归模型的股市波动率研究[J]. 应用数学进展, 2021, 10(4): 1122-1131.

DOI: 10.12677/aam.2021.104122

duced. Finally, the empirical analysis shows that adding sentiment index to the volatility model can improve the prediction accuracy of the model out of sample.

Keywords

Volatility, Heterogeneous Autoregressive Models, Agreement Index, GRU Model

Copyright © 2021 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

波动率作为金融市场风险的指示标,广泛应用于金融市场各个领域,是金融市场资产价格的变化速度的体现,波动率与市场变化速度成正相关,故精准刻画波动率尤为关键。在对金融市场大量数据的实证研究中发现,波动率数据往往是不服从正态分布的,而是呈现尖峰厚尾、非对称性、聚集性、长期记忆性和持续性等特征[1]。

随着经济的发展,越来越多科学技术应用到金融领域,这为获取高频数据提供了极大方便,高频数据相比于早期使用的低频数据,包含了更为丰富与全面的日内市场信息,并且省去了复杂的参数估计过程,故交易者可利用高频交易数据对波动率进行更为精确的分析和预测[2],成为现阶段的研究热点。通过对已有金融数据的分析表明,金融市场有明显的异质性属性,Corsi [3]基于市场异质假说,认为每一个时间尺度的潜在波动率都会受相同时间尺度的过去实际波动率和更大时间尺度上的未来潜在波动率的影响。考虑不同时间尺度对实际波动率的影响,建立了异构自回归模型(HAR)。Andersan [4]通过对汇率、股票指数回报和债券收益率的实证分析,发现波动率跳跃分量不仅非常重要,而且明显比连续分量的持久性更差,将粗糙的跳跃移动与平稳的连续移动分开可以显著提高样本外波动率预测。考虑到跳跃的重要性,于是构建了 HAR-RV-J 模型。Patton 和 Sheppard [5]关注的是不对称效应(即杠杆效应),提出了 HAR-RV-RS 模型。

以上学者不断地对 HAR 模型进行完善,形成了相对完备的 HAR 框架。但随着人工智能的不断发展,一些机器学习方法也应用到波动率研究领域。2016年,Ryo Akita [6]等人在研究中加入了数据的文本类信息,通过深度学习中的 LSTM 算法对上市公司的股票进行了价格预测,得到不错的预测结果。2018年,王钧[7]利用了深度学习中的神经网络中的 LSTM、优化模型、RNN 及 Seq2Seq 对股票数据样本进行短期价格预测,其中 Seq2Seq 的预测效果最优。而深度学习中的 GRU 模型,由于对于非线性数据敏感,也受到了学者们的重视。黎镭等[8]利用神经网络对未来股票收盘价进行预测,精准度可以高达 98%。

本文受到文献[8]的启发,将深度学习中的 GRU 模型引入到股票市场波动率预测领域,并考虑新闻信息对于股票市场波动率的影响,构建了新的情感词典,将情绪指标与经典的 HAR 模型相结合,最后通过实证分析证明,GRU 模型预测效果最好,加入了情绪指标后,确实提高了模型的预测精度。

2. 模型构建

2.1. 情感词典的构建

在互联网的当今大数据飞速发展的环境下,各品牌网站的新闻信息大部分都是通过文本形式来展示。这是属于非结构化的数据形式,这种形式下通常会包含着大量的价值性高的信息,因此,需要对这些信

息进行量化处理。本篇讨论的重点是关于新闻类的文本数据在股市上关于波动性的影响，因此，关键点就是发现找出关于新闻的文本信息。一则新闻即可看作是一篇文本，而多个词也可以组合成一篇文本，而词又可以分为积极词和消极词，因此新闻中的情感表达可以通过文本中的积极词和消极词来展示。

在《中国金融情感词典》的编纂过程中，以《清华褒贬词典》、《大连理工大学情感词典》和《LM翻译词典》为基础词典。知网情感词典(HowNet)在金融领域涉及少之又少，所以没有收录。作为中国的主流媒体之一，新浪新闻实时报道着国际上的新闻，并与中国证券报、人民日报、CNKI等多家新闻网站和机构合作，新闻信息全面渗入整个金融领域。

所以，本文的多个主要板块优先选择新浪股票设置为新闻语料库。具体步骤如下：

1) 合并大连词典、清华词典等词典中的正反词，剔除重复项，删除单个字的单词。形成了《中国金融情感词典》的基础词典。

2) 选择2017年1月至2021年1月，新浪股票网站上的一些专业板块的新闻文本来形成新闻专业的语料库。同时使用python工具包，将全部的文档做出分段处理，然后选择最适用于文本分析工作的“精确模式”。

3) 对已经构建的基础词典中存在的正反单词出现在新闻语料中的频率进行统计分析。将频率未出现过的单词剔除出去，此时可得到初步的，版本为1.0。

4) 上一步得出的金融词典1.0版本中，全部的单词都是来自清华词典、大连词典等词典，以上所说的几部词典不存在“不足”和“崩溃”这两个典型的金融领域的消极性词汇，以及“双赢”和“转机”这两个经典的金融领域的积极性词汇，所以在本文统计的新闻语料库中，每个词汇出现的频率都会被监测到，然后根据人工筛选，筛选出频率的大小，最终选择出现频率最高的金融词语，以正反词各100个的形式，最终得到财经词典的2.0版本。

5) 使用word2vec将建立的新闻语料转变成词向量，然后依照相似率的高低，对金融词典2.0进行扩展，对扩展后情感极性分类有误的单词以人工的方式进行检查，得到金融词典3.0版本。

6) 合并金融词典1.0版本和金融词典3.0版本，消除重复的单词，最终形成本文所述的中文金融情感类大词典CFED。

最终通过以上的过程而建立的中文财经情感词典，里面共有5897个词语，其中正面3275个，负面2622个。如表1所示。

Table 1. Frequency comparison table of news words in various dictionaries

表 1. 各个词典新闻词语频率对照表

	TSINGHUA	DLUTSD	LM	CFED
负面词语	33.35%	14.21%	60.16%	100%
正面词语	44.01%	18.35%	75.56%	100%
总数	39.18%	16.34%	62.47%	100%

从表中可以看出来，新闻中出现的词语在中文财经情感词典中出现的比例是100，这说明后者是根据前者的范围来制定的，也说明我们所构建的词典是全面有效的。

2.2. 情绪指标的构建

为了一些能够体现股票市场舆论导向影响的一些投资者的情绪，我们建立了情绪指标，这些指标的建立主要是通过上文中的CFED对于每天的新闻来具体的定量分析。我们有如下定义：

为了一些能够体现股票市场舆论导向影响的一些投资者的情绪，我们建立了情绪指标，这些指标的建立主要是通过上文中的 CFED 对于每天的新闻来具体的定量分析。我们有如下定义：

看涨指数： $BI = (\text{poswords} - \text{negwords}) / \text{totalwords}$

情绪一致性指标： $AI = 1 - \sqrt{1 - BI^2}$

当 $BI = -1$ 时，表示一致看跌；

当 $BI = 1$ 时，表示一致看涨；

当 $BI = 0$ 时，表示看涨看跌持平。

可以得出 AI 取值为 $(0, 1)$ ， AI 接近 0，表示意见统一度低， AI 与 1 接近，表示意见统一度高。

本文将进行交易的那天的所以新闻都放在一起，并不是只拿当天的某一个，总共的新闻文档大概一天 120 份，然后再这个基础上计算情绪指标。这样的操作方式可以减少因为文本数目不够而造成情绪指标结果的不准确现象。如果看涨情绪变高，那么股票的收益率也会变高。呈现正相关，而情绪一致性于其的关系恰好和看涨情绪相反，呈现负相关。因为本文的研究目的是波动率变化，所以我们主要观察情绪一致性的变化。

2.3. 模型构建

2.3.1. 已实现波动率

Merton 等人提出，如果在足够大的采样频率前提下，高频平方变量之和能够作为独立同散布随机变量(在固定周期内)方差的一个足够准确的预估。因此，日已实现波动率的非参数估计可以通过将日内分时数据在一定频率下的对数收益率的平方根相加得到，即已实现波动率，具体为：

$$RV_t^{(d)} = \sqrt{\sum_{i=0}^{N-1} r_{t-i\Delta}^2} \quad (2.1)$$

其中： N ：日内观测频率， $\Delta = 1d/N$ ， $r_{t-i\Delta}^2 = X_{t-i\Delta} - X_{t-(i+1)\Delta}$ 。

2.3.2. HAR-RV 和 HAR-RV-AI 模型

Corsi 认为每一个时间尺度的潜在波动率都会受相同时间尺度的过去实际波动率和更大时间尺度上的未来潜在波动率的影响。考虑不同时间尺度对实际波动率的影响，构建了实际波动率预测的 HAR (Heterogeneous Autoregressive, 异构自回归)模型，即 HAR-RV 模型，刻画过去日、周、月(即 1 个、5 个、22 个交易日)实际波动率对未来日实际波动率的滞后影响，简称为(1, 5, 22)结构。模型形式如下：

$$RV_t = \beta_0 + \beta_d RV_{t-1} + \beta_w RV_{t-5,t-1} + \beta_m RV_{t-22,t-1} + \varepsilon_t \quad (2.2)$$

$$RV_{t-h,t-1} = (1/h)(RV_{t-h} + \dots + RV_{t-1}) \quad (2.3)$$

大量的研究表明，新闻媒体的情绪指标的确可以影响股票市场的波动，所以将情绪指标与预测模型相结合是有意义的，本文基于 HAR-RV 模型，建立了 HAR-RV-AI 模型，模型如下：

$$RV_t = \beta_0 + \beta_d RV_{t-1} + \beta_w RV_{t-5,t-1} + \beta_m RV_{t-22,t-1} + \alpha_d AI_{t-1}^d + \alpha_w AI_{t-1}^w + \alpha_m AI_{t-1}^m + \varepsilon_t \quad (2.4)$$

2.3.3. GRU 模型

门控循环单元(Gate Recurrent Unit)，以下简称 GRU，是循环神经网络(Recurrent Neural Network, RNN)的一种，是为了解决长期记忆和反向传播中的梯度问题提出的，应用于自然语言处理，图像识别以及其他领域。相比于一般的 RNN 模型，对于解决长期记忆性问题比较好，而与长短记忆循环网络(Long Short Term Memory network, LSTM)相比，GRU 参数比 LSTM 少，但是却可以高效的完成与 LSTM 相同的任务，既提高了计算效率，又节省了时间成本因此，近年来也被应用于预测股票收益。GRU 详细结构图如图 1 所示：

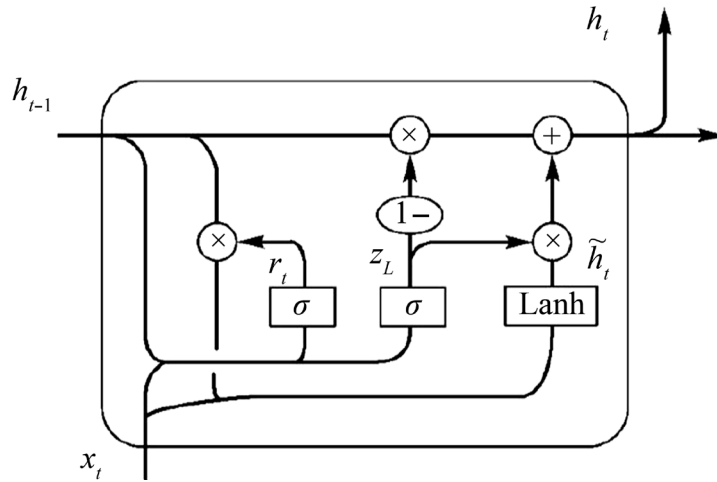


Figure 1. GRU structure diagram
图 1. GRU 结构示意图

这里的激发函数为 \tanh ，误差计算利用 MAE 损失函数。模型的“epoch”设定为 50，“batch”大小设定为 72。在本文中，我们只研究一个隐藏层，隐藏层节点的计算公式如下所示：

$$i = \sqrt{n+l+m}$$

i 表示隐藏层节点的个数， n 表示输入层节点个数， l 表示输出层节点个数， m 是一个 1~10 之间的常数。(5, 10, 22) 可以考虑为短期，中期，长期的波动，我们考虑这些参数根据公式得到相应的隐藏层的节点个数，然后训练神经网络模型得到相对误差的平均值，最后选择误差最小的隐藏层节点数，根据这种方法，对于 GRU 模型，输入层节点数为 22，隐藏层节点数为 5，输出层节点数为 1。对于 GRU-AI 模型，输入层节点数为 22，隐藏层节点数为 7，输出节点数为 1。

考虑到新闻媒体情绪对股票市场的影响，我们在 GRU 单变量模型中加入 AI 作为解释变量，建立 GRU-AI 模型来研究实现波动率。

2.4. 损失函数及 MCS 检验

2.4.1. 损失函数

损失函数有很多种表达方式，但是由于研究还不够深入，所以对于其选取还没有一个统一的说法。它主要是判断损失和误差的函数。

Hansen 和 Lunde [9] 认为损失函数的多少可以成为模型好坏的标准。由于不同的损失函数模型是从不同方面对于评价模型进行考虑的，本文采用下面介绍的六种损失函数：

$$MSE = \frac{1}{N} \sum_{t=T+1}^{T+N} (RV_t - RV_t^f)^2 \tag{2.5}$$

$$MAE = \frac{1}{N} \sum_{t=T+1}^{T+N} |RV_t - RV_t^f| \tag{2.6}$$

$$MSPE = \frac{1}{N} \sum_{t=T+1}^{T+N} \left(\frac{RV_t - RV_t^f}{RV_t} \right)^2 \tag{2.7}$$

$$MAPE = \frac{1}{N} \sum_{t=T+1}^{T+N} \left| \frac{RV_t - RV_t^f}{RV_t} \right| \tag{2.8}$$

$$\text{QLIKE} = \frac{1}{N} \sum_{t=T+1}^{T+N} \left(\ln RV_t^f + \left(\frac{RV_t}{RV_t^f} \right)^2 \right) \quad (2.9)$$

$$\text{R}^2\text{LOG} = \frac{1}{N} \sum_{t=T+1}^{T+N} \left(\ln \left(\frac{RV_t}{RV_t^f} \right)^2 \right)^2 \quad (2.10)$$

其中: RV_{t+i} : 真实值, RV_{t+i}^f : 预测值。

2.4.2. MCS 检验

损失函数受到外界的影响因素也有很多, 如果我们要讨论模型的猜测准确度的时候把损失函数作为参考物, 那么就必须进行很多次的样本猜测, 只通过一次猜测就得出结论的行为, 是不能被认同的, 因为没有做到科学严谨, 如果在其中出现了一些不正常的数值, 那么损失函数的结果就会快速的增大, 最终影响结论。

为了解决上述的问题, Hansen 和 Lunde [10]提出了一种新的检验方式, 这个方式可以先将模型聚集起来, 然后再对所有的进行检验。这种方法被称为, 模型信度集检验。

计算相对函数损失值的公式:

$$\begin{aligned} L_{i,u,t} &= L(RV_{u,t}^f, RV_t) \\ d_{i,uv,t} &= L_{i,u,t} - L_{i,v,t} \end{aligned} \quad (2.11)$$

MCS 检验过程:

- 1) 用 M^* 来表示所有预测模型的组合。
- 2) 在假设所有的模型的预测能力是一样的基础上, 给出 MCS 的最开始的假设:

$$H_0 : E(d_{i,uv,t}) = 0 \quad (2.12)$$

3) 在 α 置信水平确定的情况下, 用该方法把 M^* 中的模型两两分组, 然后再分别检验。由 Hansen 等提出, 定义如下:

$$\begin{aligned} T_R &= \max_{uv \in M^*} \frac{|\bar{d}_{i,uv}|}{\sqrt{\text{var}(\bar{d}_{i,uv})}} \\ T_{SQ} &= \max_{uv \in M^*} \frac{(\bar{d}_{i,uv})^2}{\text{var}(\bar{d}_{i,uv})} \\ \bar{d}_{i,uv} &= \frac{1}{N} \sum_{t=T+1}^{T+N} d_{i,uv,t} \end{aligned} \quad (2.13)$$

对应于统计量 T_R 和 T_{SQ} 的剔除准则 E_R 和 E_{SQ} 分别为:

$$\begin{aligned} E_R &= \arg \max_u \left\{ \sup_{v \in M^*} \frac{|\bar{d}_{i,uv}|}{\sqrt{\text{var}(\bar{d}_{i,uv})}} \right\} \\ E_{SQ} &= \arg \max_u \left\{ \sup_{v \in M^*} \frac{(\bar{d}_{i,uv})^2}{\text{var}(\bar{d}_{i,uv})} \right\} \end{aligned} \quad (2.14)$$

不断地将第三个步骤重复，一直到 M^* 中的模型都处于一定显著水平，即 p 值大于 α ，此时便不再剔除模型， p 值与预测精度成正相关。

3. 实证分析

3.1. 数据说明和描述性统计

本文采用的是 5 分钟高频沪深 300 指数交易数据，时间为 2017 年 1 月 3 日~2020 年 4 月 15 日，期间共 799 个交易日。将样本总量三分之二作为训练样本，三分之一作为测试样本。根据我国股市交易时间，每个交易日共有 48 个交易日数据。记为 $p_{t,j-1}, t=1,2,\dots,N; j=1,2,\dots,48$ ，样本总数为 38,252。

本文情绪指标的获取途径主要是我国的新浪新闻。我们首先获取了关于股票的网站中的各个新闻板块所发布的新闻的文字部分，而且在选取的时候要注意新闻发布的时间应该和股票进行买卖的时间是一样的，如果是在节假日发布了新闻，我们所采取的处理方式把这个行为放到下一个股票买卖的时间里面去，在确保我们组合的样本的时间长度和跨度适合股票指数的时间长度和跨度是一样的。从 2017 年 1 月 3 日至 2021 年的 4 月 15 日我们一共整理了，1174 个每日的新闻文本，共计 799 个股票买卖的时间，总共将近 12 万篇。我们再把每个股票买卖的时间的新闻文本分开词语，和 CFED 里面的正面词语和负面词语相比较，利用工具算出 AI，然后确定了我们所要利用的样本。

学者们进行了大量的研究发现，资产的波动率并不是正态分布。但是因为 HAR-RV 模型对于这些变化是有特殊要求的，要求它的扰动项必须符合正态分布，我们对比了对数和非对数的已经实现的变化率，发现前者的分布符合要求，各变量信息如表 2 所示：

Table 2. Descriptive statistical tables

表 2. 描述性统计表

	Rt	RVt	lnRVt	AI
Mean	0.011	2.476	-0.089	0.084
Min	-9.154	0.053	-2.947	0.003
Max	6.493	54.011	3.989	0.198
Std	1.654	5.875	1.278	0.034
Skew	-1.237	5.729	0.619	0.363
Kurt	6.849	37.9	0.116	-0.185
ADF	-21.778***	-5.856***	-4.087***	-7.406***
J-B	1735.132***	51455.110***	50.988***	18.649***
Q(5)	28.693***	1249.549***	2047.057***	782.464***
Q(10)	65.784***	1654.223***	3774.823***	1128.250***
Q(22)	97.242***	2081.473***	6930.548***	1382.891***
Hurst	0.489	0.569	0.619	0.591

由表我们可以得到以下结论：

1) 我们知道在正态分布中偏度是零，峰度也是零，由表 2 可知，每天的利润率 r_t 有着很明显超越正常额度的最高值，并且非正偏度十分的明显，由 J-B 检验知，每天的利润率和我们所假设的正态分布并不吻合；在置信水平 1% 的时候 RVt 的偏度为 5.729，峰度为 37.9，我们可以拒绝原假设；对数化的日实际波动率虽然也出现了类似情况，但是他的两个度值都很接近正态分布；这也是我们选择对数化已实现波动率为变量的原因。

2) 根据 ADF 检验的结果可以看出来, 每个序列均拒绝了原假设, 即每个序列存在 1% 置信水平下的单位根, 这种情况我们就说每个序列都是平稳的; 在检验和证明自相关性的时候, L-BQ 统计量显示, 在各个滞后阶的序列全部拒绝了原假设, 这种情况就说明每个序列都有自相关性。除此之外, 利用 Hurst 指数分析了序列长记忆性, 从表中显示的来看, 利润率的 Hurst 指数与 0.5 接近, 这表明了序列类似于随机游走过程。反观其他三项的指数都在(0.5, 1)区间, 对数化的已实现波动率更明显, 这也是将对数已实现波动率作为因变量建模的有力依据。

3.2. 样本外的预测结果

3.2.1. 基于损失函数下的模型预测精度评估

在公式(2.5)~(2.10)的基础上, 我们给出了文中四种波动率模型在不同损失函数下的样本外预测结果, 如表 3 所示:

Table 3. Results of each loss function

表 3. 各个损失函数结果

模型	MSE	MAE	MAPE	MSPE	QLIKE	R ² LOG
HAR-RV	1.281	0.386	0.466	0.459	2.604	1.302
HAR-RV-AI	1.243	0.387	0.461	0.434	2.596	1.293
GRU	0.852	0.296	0.398	0.321	2.47	1.035
GRU-AI	0.172	0.233	0.397	0.305	2.71	1.069

在表 3 中, 我们可以很明显的看出: 1) GRU 波动率模型(GRU 和 GRU-AI)比 HAR-RV 波动率模型(HAR-RV 和 HAR-RV-AI)具有更好的预测性能。这可能是因为 GRU 模型相对于的一般 RNN 模型, 可以解决长期依赖问题, 所以能有效地维持时间序列中每个时刻的状态, 体现已实现波动的非线性特点。另外, HAR-RV 模型本质上来讲是自回归线性模型, 它展现了不同时期波动率之间的线性关系。2) 将(MSE, MSPE、MAPE、QLIKE、R²LOG)五种损失函数为评估标准, 我们发现, HAR-RV-AI 模型的预测精度优于 HAR-RV 模型。我们还发现, 将 4 种损失函数(MAE、MSE、MAPE、MSPE)最为评价标准, GRU-AI 模型的预测准确性要优于 GRU 模型。这表明, 在某种程度上, 将 AI 情绪指数添加到的波动率预测模型中对原有模型的预测能力确实有所提高。

3.2.2. MCS 检验下的模型预测精度评估

为了使检验结果更有说服力, 我们给出了各个模型在 MCS 检验的结果, 如表 4 所示:

Table 4. MCS test results

表 4. MCS 检验结果

损失函数	统计量	HAR-RV	HAR-RV-AI	GRU	GRU-AI
MSE	TR	0.15	0.157	0.157	1
	TSQ	0.103	0.101	0.103	1
MAE	TR	0.078	0.078	0.18	1
	TSQ	0.03	0.023	0.171	1
MSPE	TR	0.007	0.001	1	0.502
	TSQ	0.213	0.398	1	0.505

Continued

MAPE	TR	0	0	0.997	1
	TSQ	0.015	0.015	0.998	1
QLIKE	TR	0.504	0.504	1	0.848
	TSQ	0.379	0.379	1	0.846
R ² LOG	TR	0.172	0.172	1	0.745
	TSQ	0.217	0.27	1	0.742

表 4 的结果表明, 所有模型都通过了基于 QLIKE 损失函数的测试。以三个损失函数(MSE、MAE、MAPE)作为评价标准, 我们发现 HAR 模型和 HAR-RV-AI 模型都没有通过检验, 只有 GRU-AI 模型通过检验, p 值为 1, 表明实现了基于神经网络的波动率预测模型具有较强的预测精度。同时, 通过对比 HAR-RV-AI 和 HAR-RV, 发现加入 AI 情绪指标的模型比原始模型具有更好的预测精度。因此, 加入 AI 情绪指标的模型可能提高预测模型的准确性。结果表明, GRU-AI 模型是最佳的预测模型。

4. 结论

本文主要研究股票市场波动的预测, 以便更准确地了解市场经济, 由此来帮助投资者把握股市波动走向。本文首先表明了金融股票市场确实可以受到新闻媒体的影响, 并证实了在金融领域挖掘新闻信息文本的重要性。在此基础上, 本文研究了文本大数据对股票波动性的影响, 我们利用机器学习的方法, 在股票波动预测方面提供了新的想法。

我们研究了两个方面: 考虑深度学习方法(GRU)预测波动率, 以及考虑新闻媒体对波动率的影响波动。因此, 首先构建一个面向金融领域的中文金融情感词典。在此基础上, 我们将新闻媒体中的情绪量化为要添加到模型中的情绪指数, 并建立了新的预测模型——HAR-RV-AI 和 GRU-AI。为了比较模型的预测能力, 我们以损失函数法和 MCS 检验为评价标准, 采用滚动窗口策略进行样本外预测。最后得出结论, 深度学习 GRU 模型在股票波动率预测领域拟合良好。GRU 神经网络模型对已实现波动率的预测能力明显高于 HAR-RV 模型。这可能是因为 GRU 模型相对于一般 RNN 模型, 可以解决长期依赖问题, 所以能有效地维持时间序列中每个时刻的状态, 体现已实现波动的非线性特点。另外, HAR-RV 模型本质上来讲是自回归线性模型, 它展现了不同时期波动率之间的线性关系。此外, 添加情感指数后, 预测模型的精度有所提高。这说明及时掌握新闻信息可以提高股票市场的风险控制能力。

参考文献

- [1] 王国华. 中国股票市场日内波动率研究[D]: [博士学位论文]. 武汉: 中南财经政法大学, 2017.
- [2] Engle, R.F. (2000) The Econometrics of Ultra-High Frequency Data. *Econometrica*, **68**, 1-22. <https://doi.org/10.1111/1468-0262.00091>
- [3] Corsi, F. (2009) A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*, **7**, 174-196. <https://doi.org/10.1093/jjfinec/nbp001>
- [4] Andersen, T.G., Bollerslev, T. and Diebold, F.X. (2007) Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility. *Review of Economics and Statistics*, **89**, 701-720. <https://doi.org/10.1162/rest.89.4.701>
- [5] Patton, A.J. and Sheppard, K. (2015) Good Volatility, Bad Volatility: Signed Jumps and the Persistence of Volatility. *Review of Economics and Statistics*, **97**, 683-697. https://doi.org/10.1162/REST_a_00503
- [6] Akita, R., Yoshihara, A., Matsubara, T., et al. (2016) Deep Learning for Stock Prediction Using Numerical and Textual Information. 2016 *IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Okayama, 26-29 June 2016, 7550882. <https://doi.org/10.1109/ICIS.2016.7550882>

-
- [7] 王钧, 张鹏, 袁帅. 基于股票预测的 Seq2Seq RNN 和 LSTM 模型比较[J]. 时代金融, 2018(35): 381-392.
 - [8] 黎镭, 陈嵩祥, 李伟书, 等. GRU 递归神经网络对股票收盘价的预测研究[J]. 计算机与现代化, 2018, 38(11): 103-108.
 - [9] Lunde, A. and Hansen, P.R. (2004) Realized Variance and IID Market Microstructure Noise. 2004 *North American Summer Meetings*, America: Econometric Society, 66-70.
 - [10] Hansen, P.R., Lunde, A. and Nason, J.M. (2011) The Model Confidence Set. *Econometrica*, **79**, 453-497.
<https://doi.org/10.3982/ECTA5771>