基于 TF-IDF_VaR_GRU_Attention 模型的股票预测研究

蔡伟斌　赵婧　强嘉豪　王安奇
西京学院　陕西 西安　710123

摘　要：针对股票的特征选取和风险量化问题，提出了一种基于 TF-IDF 的特征评分、VaR 风险评估和 Attention 机制的股票预测模型。通过消融实验，对比了 5 种基于 LSTM 神经网络的模型和 3 种基准模型，发现本文提出的 TF_IDF_GRU_VaR_Attention 模型在 RMSE、MSE 和 MAE 评价指标均优于对比模型。可知基于 TF-IDF_GRU_VaR_Attention 的预测模型对比传统基准模型和当下流行的模型，能更为准确地预测股票的涨跌趋势。

关键词：股票预测；TF-IDF；GRU 模型；VaR 风险评估；注意力机制

引言

近年来，全球资产排名前 4 以及第 6 位的五家资管机构，均依靠计算机技术来开展投资决策。从国内情况来看，对量化投资的探索起步于 2004 年，当时光大保德量化核心基金和上投摩根阿尔法基金是最先尝试量化投资的机构。到 2009 年，由于美国遭遇金融危机，部分华尔街精英选择归国就业，同时带来了量化投资的经验 [1]。可见量化投资的概念不算新颖，但目前国内真正的量化投资基金还较为少见。

数学模型是量化分析股票趋势的重要工具，从股票预测的发展趋势看，传统预测模型如 BP 神经网络和 SVM 等技术被用于股票预测。但浅层的机器学习算法泛化效果一般，故引入深度学习框架，包括 LSTM、BiLSTM 和 GRU 等循环神经网络。再针对股票指标的选取，文献 [2] 人为地选取了常见指标，文献 [3] 中客观选定了股票因子，但未对股票的风险因子进行研究，由此本文引入 TF-IDF 算法 [4] 针对券商研报客观提取指标。再基于 VaR 算法量化股票的风险因子，最后采用 Attention 机制，对每个时间序列分配权重，反映该时间序列对输出的重要性，并引入最终的输出。

本文主要基于股票的历史数据建立循环神经网络模型，模拟股票的变化趋势，实现了股票的指标选取、风险量化分析和多类算法对比，训练并预测了各支股票未来 30 个交易日的涨跌趋势，并进行了对比验证，为制定股票量化投资的策略提供一定的参考。

1 模型结构

1.1 GRU 模型结构

本文研究的股票行情预测属于典型的时序问题，故选用 GRU 模型进行研究。GRU 为长短时记忆神经网络（LSTM）的简化版本，具有能够学习长期依赖信息的能力。将 LSTM 中的三个门（遗忘门、输入门和输出门）减少到两个（重置门和更新门），因此 GRU 在捕捉和学习时间序列数据中的长期相关性方面中，表现出了更强的熟练度，同时还降低了模型训练的复杂度和计算成本，提升训练的效率，具体结构如下：

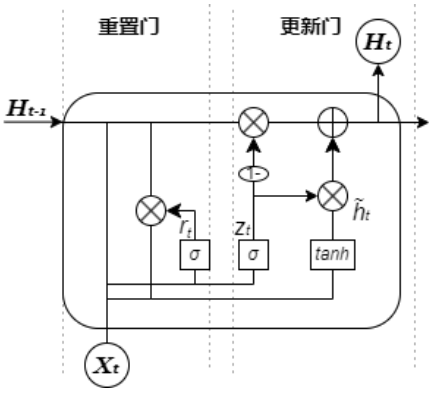

图 1 GRU 模型结构

（1）重置门

$$G_r = \sigma(W_r \cdot [H_{t-1}, X_t] + b_r) \quad (1)$$

$$\tilde{h}_t = \tanh(W_h \cdot [H_{t-1} \cdot G_r, X_t] + b_r) \quad (2)$$

其中，W_r 和 b_r 表示权重和偏置常数，$\tanh$ 和 σ 表示激活函数，H_{t-1} 表示上层输出。

（2）更新门

$$G_z = \sigma(W_z \cdot [H_{t-1}, X_t] + b_z) \quad (3)$$

$$H_t = H_{t-1} \cdot (1 - G_z) + \widetilde{h}_t \cdot G_z \quad (4)$$

其中，$\widetilde{h}_t$ 表示候选值向量，H_t 表示该层的输出。

1.2 TF-IDF 模型

TF-IDF（term frequency-inverse document frequency）是一种衡量词语权重的方法。它通过分析每个文本中词语的词频和在整个语料库中的逆文档频率来计算词语的重要性，可知：

$$TF\text{-}IDF = \frac{D_{ij}}{\sum_{k=1}^{K} D_{kj}} \cdot \log\left(\frac{D}{|\{j : i \in j\}| + 1}\right) \quad (5)$$

其中，D_{ij} 表示单词 i 在文本 j 中出现的次数，D 表示文本样本的总数，$\{j : i \in j\}$ 表示包含单词 i 的文本数量。

1.3 VaR 风险评估模型

VaR 方法的解释是风险价值方法，是用来评估金融风险的工具。早在 20 世纪 90 年代，G30 成员发表了一份关于金融衍生工具的报告，首次建议使用“风险价值系统来评估金融风险”，其一出现就受到了广大金融机构的青睐并被采用，可知：

$$VaR_l = a \cdot std\left(\ln \frac{E_{lo}}{E_{l(o-1)}}\right) \cdot h \quad (6)$$

其中，E_{lo}、R_{lo} 分别表示第 l 支股票在第 o 个交易日的收盘价与收益率，a 表示 R_{lo} 在服从正态分布下对应置信区间的分位点，std 表示 R_{lo} 的标准差，h 表示持有期。

1.4 环境配置及模型参数

本文研究环境基于 Python3.8 语言环境，以 Tensorflow2.13 为深度学习框架。其中，为了更好地评估各时间窗口中隐藏状态的影响，本文引入时间注意力机制，用于度量不同状态对收盘价预测的重要性，故网络架构设定为一层 GRU 和一层 Attention 注意力机制层，并加入 Dropout 层防止模型过拟合。同时选取 Adam 为优化器，学习率为 0.001，损失函数为 MSE 函数，时间步长为 30 天，设置训练集与测试集数据比例为 8：2，同时设定训练 epoch 为 50。

2 实验与结果分析

2.1 数据来源

本文通过第三方模块 Tushare，选取了华发股份、丽珠集团、亿纬锂能、白云山、瀚蓝环境、顺络电子、风华高科、分众传媒、粤水电和华侨城 A 等 10 支股票的相关数据，且时间跨度为 2011 年 11 月 7 日至 2021 年 12 月 16 日。

2.2 数据预处理

（1）研报分词

券商研究报告是研究人员对证券及相关产品价值，以及影响其市场价值的因素进行分析所作出的报告，本文基于“慧博智能”终端平台，获取 10 支股票近 1 年的券商研报，共计 63 篇。利用 Stanford CoreNLP 分词工具，对研报内容进行关键字提取处理，再基于 TF-IDF 算法，得到研报关键指标的评分和累加贡献度，设定 90% 累加贡献度的指标作为股票指标，并将其代入待训练的股票预测模型中，各指标关键字的评分指数，如表 1 所示：

表 1 股票指标及评分指数

指标名称	评分指数（归一化）
涨跌幅	1.000
市盈率	0.921
周转率	0.875
日振幅	0.437
日利润增长率	0.313
净资产收益率	0.126
……	……

可知，90% 贡献度以内的股票指标包括：涨跌幅、市盈率、周转率、日利润增长率和净资产收益率。

（2）风险因子

对各支股票的收盘价采取 VaR 风险评估法，得到“单日 VaR 值”作为股票的风险指标加入股票预测模型中，作为各支股票的风险因子。

2.3 模型性能比较

本文以收盘价为目标变量，其余指标以及风险因子为特征变量，建立 TF-IDF_GRU_VaR_Attention 模型，与 LSTM、BiLSTM、GRU 等模型进行对比实验。

表 2 是各股票模型在 10 支股票数据集的平均指标对比，表 3 是各模型在 10 支股票数据集下的评价指标对比。可知，本文提出的 TF_IDF_GRU_VaR_Attention 模型综合评价均优于其他模型，表明了该模型预测效果的准确性，由表 3 看出，单支股票对应的各预测模型对比效果，发现本文提出的模型在大部分股票的对比效果中位于前列。

同时，本文针对 TF-IDF、VaR 和 Attention 模块进行了消融实验，将各模型的评价效果逐步进行对比。

表 2：各股票模型综合评分对比

模型	MAE	MSE	RMSE
LSTM	0.6207	0.7793	0.7342
BiLSTM	0.5938	0.7119	0.7148
GRU	0.6132	0.7852	0.7464
TF_IDF_GRU	0.3610	0.3110	0.4563
GRU_VaR	0.5758	0.7095	0.7216
TF_IDF_GRU_VaR	0.3056	0.2685	0.4038
BiLSTM_Attention	0.4209	0.4414	0.5473
TF_IDF_GRU_VaR_Attention	0.2232	0.1519	0.2942

可知，本文提出的 TF-IDF_GRU_VaR_Attention 模型的 RMSE、MSE 和 MAE 均低于文献中的 Spatial_Temporal_BiLSTM 模型和未引入 Attention 的 TF-IDF_GRU_VaR 模型，分别降低了 35.68%、60.40% 和 34.36%，以及 12.82%、34.90% 和 9.59%，说明本文模型较与当前流行模型，以及未引入 Attention 的模型，在效果上有一定的改进提升，充分说明了本文模型的有效性。

本文选择了 2 支具有代表性的股票预测效果图，分别为 2020 年 8 月 17 日至 2021 年 6 月 12 日的“华发股份”和“华侨城 A”2 支股票，具体预测效果如图 2 和图 3：

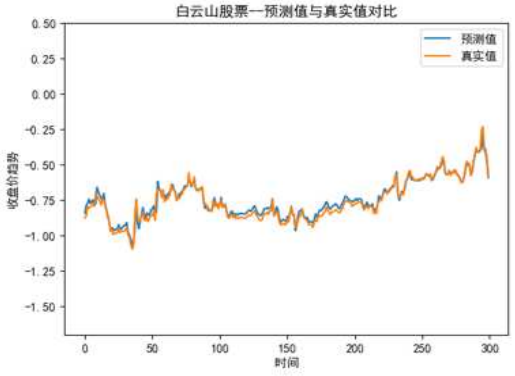

图 2 白云山股票预测效果图

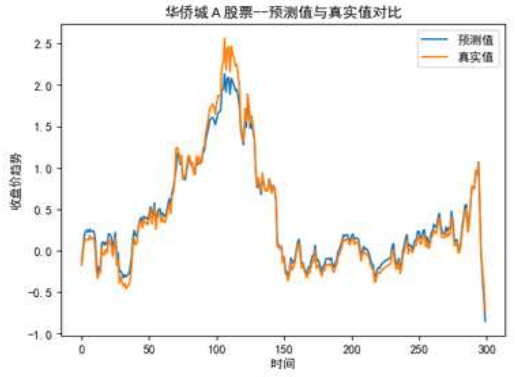

图 3 华侨城 A 股票预测效果图

可知，两支股票的收盘价预测效果整体较好，预测值与真实值基本重合。

3 结论

相较于传统的量化投资，本文讨论了股票指标的客观提取、投资风险的客观因素以及预测算法的对比，得出如下结论：

（1）基于 TF-IDF 对各公司近 1 年的券商研报进行分析，选取了综合贡献率达到 90% 的指标，对比主观地选择股票指标，预测效果更佳，说明客观、正确地选取股票指标有助于提升模型的准确度。

（2）基于 VaR 风险评估法计算各支股票的“每日 VaR 值”，对比加入风险指标前后的预测效果，发现大部分模型的预测均得到了提升，证明了该风险指标的有效性。

（3）基于 TF-IDF_GRU_VaR_Attention 神经网络模型构建股票行情预测模型，对各支股票未来 30 天的收盘价和收盘价增长率进行了预测，预测结果与实际结果基本相符，并通过消融实验进行对比，得出本文的模型综合评分最优，说明该模型能较准确地反映各支股票的行情趋势。

参考文献：

[1] 张兆瑞 , 穿透量化投资的迷雾 , in 新金融观察 2021. 20.

[2] 彭燕 , 刘宇红与张荣芬 , 基于 LSTM 的股票价格预测建模与分析 . 计算机工程与应用 , 2019. 55(11): 209-212.

[3] 傅廷君 , 基于券商研究报告的股票价格趋势预测 , 2017, 暨南大学 .

[4] 杨蓦 , 王静 . 基于时空注意力机制的双向长短期记忆神经网络的股指预测研究 [J]. 运筹与管理 ,2023,32(08):174-180.

基金项目：

“陕西省 2021 年自然科学基础研究计划”项目突发公共卫生事件网络异常评论关键节点定位及干预机制研究（项目编号：2021JQ-878）

作者简介：

蔡伟斌（2001.01-），男，汉族，江西上饶人，硕士研究生，研究方向：谣言检测；

赵婧（1983.05-）女，陕西西安人，副教授，博士，研究方向：模式识别。