DOI:10.3969/j.issn.1003-5060.2025.03.015

融入股票论坛 UGC 时序特征的上市公司财务困境预测方法

张　玉[1]，　蒋翠清[1,2]

(1.合肥工业大学 管理学院,安徽 合肥　230009; 2.过程优化与智能决策教育部重点实验室,安徽 合肥　230009)

摘　要:股票论坛用户生成内容(user generated content,UGC)能反映上市公司利益相关者对公司经营业绩和相关事件的关注和观点,具有及时性和动态性,是对财务信息的有效补充。为有效提取动态变化 UGC,文章提出一种融入股票论坛 UGC 时序特征的上市公司财务困境预测方法。首先,针对用户评论和用户阅读中的时间序列信息,考虑情感特征时序性和互动信息时序性,采用门控循环网络(gated recurrent unit,GRU)模型,挖掘时间序列中的动态信息;其次,不同时间段下发生的事件对财务困境预测的影响程度不同,采用注意力机制聚合重大事件对财务困境预测的影响;最后,基于 UGC 时序特征,并结合财务特征对上市公司财务困境进行预测。研究表明,所提方法能够有效地提取并聚合时序特征,提高财务困境预测效果。

关键词:股票论坛;时序特征;门控循环网络;注意力机制;财务困境预测

中图分类号:F832.5　　**文献标志码**:A　　**文章编号**:1003-5060(2025)03-0387-08

Research on financial distress prediction of listed companies integrating UGC time series characteristics of stock forum

ZHANG Yu[1],　JIANG Cuiqing[1,2]

(1. School of Management, Hefei University of Technology, Hefei 230009, China; 2. Key Laboratory of Process Optimization and Intelligent Decision-making of Ministry of Education, Hefei 230009, China)

Abstract: The user generated content(UGC) of the stock forum can reflect the concerns and opinions of the stakeholders of the listed company on the company's operating performance and related events. It is timely and dynamic, and is an effective supplement to financial information. In order to effectively extract the dynamic UGC, this paper proposes a method for predicting the financial distress of listed companies that integrates the UGC time series characteristics of the stock forum. Firstly, for the time series information in user comments and user reading, considering the time series of emotional features and interactive information, the gated recurrent unit(GRU) deep recurrent network model is used to mine dynamic information in time series. Secondly, events in different time periods have different effects on financial distress prediction, and attention mechanism is used to aggregate the effects of major events on financial distress prediction. Finally, the financial distress of listed companies is predicted based on the UGC time series characteristics extracted and combined with the financial features. The research shows that the method proposed in this paper can effectively extract and aggregate time series characteristics, thus improving the prediction effect of financial distress.

Key words: stock forum; time series characteristics; gated recurrent unit(GRU); attention mechanism; financial distress prediction

收稿日期:2022-07-26;**修回日期**:2022-10-20

基金项目:国家自然科学基金重点资助项目(71731005)

作者简介:张　玉(1997—),女,安徽巢湖人,合肥工业大学硕士生;
蒋翠清(1965—),男,安徽无为人,博士,合肥工业大学教授,博士生导师.

上市公司在促进经济发展、稳定就业等方面发挥着重要作用。然而,上市公司发展的内外部环境正发生复杂深刻变化,财务风险不断增加。资料显示,2016 年中国上市公司陷入财务困境的一共有 56 家。其中:2017 年有 61 家;2020 年达到 104 家[1]。陷入财务困境的上市公司不断增加,给社会稳定和国家经济发展带来重大影响。科学有效的财务困境预测方法能够提前预判公司是否会陷入财务危机,为投资者和债权人提供早期预警,帮助利益相关者及时规避风险或降低损失。

经典财务困境预测主要使用财务报表信息。财务报表能够综合反映过去一段时间内生产经营成果和财务状况,已被广泛应用于财务困境预测[2-3],但财务报表所反映的信息具有静态性和滞后性,且存在财务舞弊可能[4],不能完全真实反映公司运营状态。文献[5]指出,财务信息的价值相关性已经大幅下降,因此财务信息对财务困境预测效果受限。近年来,学者开始探究非财务信息对财务困境的影响,如年报文本[6]、投资者情绪[7]、管理层讨论与分析[8]等。文献[6]使用词向量和深度学习方法从上市公司年报和审计报告中构造文本情绪特征。文献[9]采用情感词典对管理层讨论与分析部分统计正面和负面情感词语词频,并构建管理层语调特征,验证其对财务困境预测的效用。以上提取的非财务特征能在一定程度上提升财务困境预测模型的效果,然而年报、管理层讨论与分析等都是公司自身公开发布的信息,只能反映公司自身对运营状况的总结,不能反映其他投资人等利益相关者对公司运营发展的看法,并且含有的时间信息量不足,无法动态地表征公司运营情况的变化。因此,迫切需要新的、及时有效的外部补充信息来提高财务困境预测性能。

股票论坛是一个可以给投资者、专家、股民等提供在线交流和共享信息的重要平台,汇聚了从各个途径收集的相关信息产生了大量的用户生成内容(user generated content,UGC),更新数量最多的公司每年有超过 20 万条评论。UGC 以秒为单位不停更新,具有覆盖范围广、信息量大、更新频率快的特点。相比从公司自身披露的报告获取信息,UGC 具有更高的时效性和专业性,同时大大降低了信息的不对称性。在不同事件的冲击下,股票论坛能在第一时间汇聚大量的专家和投资者等,他们主动进行交流,并发表对这些事件的看法,指出公司可能存在的风险[10]。在分析和讨论中所表达的观点被证明包含价值相关的信息,已被用于预测未来的股票收益[11]。文献[12]研究发现客户情绪可以作为评估公司业绩的线索和信号。文献[13]研究表明股票论坛情感特征能有效提升财务困境预测效果,但是并没有考虑情感的时间序列特征,且忽视了每一条 UGC 的用户阅读量以及其他用户对 UGC 的评论数量这些互动信息,阅读量和评论数量也具有时间序列特征。有关这些互动信息的时间序列特征对财务困境预测评价性能的影响研究并不多。

本文研究股票论坛 UGC 的情感时间序列特征以及每个 UGC 阅读量和评论数量的时间序列特征对财务困境预测性能的提升效果。本文首次考虑了 UGC 情感特征和 UGC 互动信息的时序性对财务困境预测的影响,采用门控循环网络(gated recurrent unit,GRU)挖掘 UGC 情感特征、用户评论和用户阅读中的动态信息,并利用注意力机制区分不同时间段下信息的重要程度,以此来聚合更多与财务困境相关的时序信息,提升财务困境预测模型性能。

1 相关理论

1.1 基于互信息法的特征筛选

财务指标是进行财务困境预测的基本指标,也是本文研究的基准模型,但收集的财务指标一般有 20 多个[3],这些指标之间可能存在共线性。过多的指标在训练过程中会降低模型训练速度,因此需要对财务特征进行特征筛选。现有特征筛选方法有过滤法、包裹法和嵌入法。本文采用互信息法作为特征筛选方法。互信息法是过滤法的一种,它不需要复杂的评价指标或者模型来处理,最终只需要根据阈值或者排序得分结果将其他变量剔除,以此来完成特征筛选工作。

从信息增益的角度来看,互信息表示由于 X 的引入而使 Y 的不确定性减少的量。信息增益越大,意味着特征 X 包含的有助于将 Y 分类的信息越多,即 Y 的不确定性越小。

如果随机变量服从分布 $X\sim p(x)$、$Y\sim p(y)$、$(X,Y)\sim p(x,y)$,那么 X、Y 之间的互信息计算公式为:

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\lg\frac{p(x,y)}{p(x)p(y)} \quad (1)$$

1.2 GRU 网络原理

循环神经网络(recurrent neural network,RNN)被用于序列建模,在处理时间序列中取得

了显著的效果[13]，然而 RNN 网络在反向传播过程中存在梯度消失问题，因此在处理较长的时间序列问题中效果明显下降。GRU 作为循环神经网络的一种，是为了解决长期记忆和反向传播中的梯度消失等问题而提出来的，相比 RNN 计算时只使用上一时间步的结果，GRU 网络中含有更新门和重置门，用于控制上一时刻传入的记忆信息。同时 GRU 参数调节更为简单，在参数较少的情况下，模型的计算复杂度大大降低。本文选用 GRU 网络作为时序特征提取模型。

GRU 的网络结构主要由重置门 $\boldsymbol{r}_t$ 和更新门 $\boldsymbol{z}_t$ 组成，内部结构如图 1 所示。

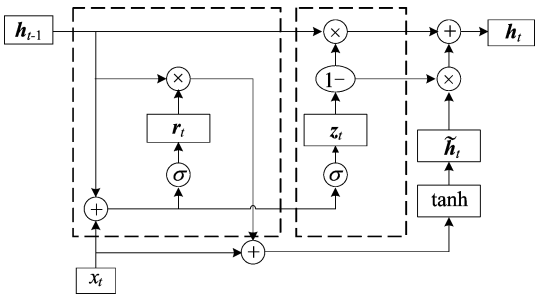

图 1　GRU 网络内部结构

1）重置门 $\boldsymbol{r}_t$ 计算公式为：

$$\boldsymbol{r}_t = \text{sigmoid}(\boldsymbol{W}_r \boldsymbol{x}_t + \boldsymbol{W}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r) \tag{2}$$

其中：$\boldsymbol{W}_r$ 为重置门权重矩阵；$\boldsymbol{b}_r$ 为偏置项；sigmoid 函数可以将数据变换为 0～1 之间，充当门控信号。

2）更新门 $\boldsymbol{z}_t$ 用来控制需要从上一时刻状态 $\boldsymbol{h}_{t-1}$ 保留的信息量。

$$\boldsymbol{z}_t = \text{sigmoid}(\boldsymbol{W}_z \boldsymbol{x}_t + \boldsymbol{W}_z \boldsymbol{h}_{t-1} + \boldsymbol{b}_z) \tag{3}$$

其中：$\boldsymbol{W}_z$ 为更新门的权重矩阵；$\boldsymbol{b}_z$ 为偏置项。

3）候选隐含记忆单元 $\tilde{\boldsymbol{h}}_t$ 的计算公式为：

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_h \boldsymbol{x}_t + \boldsymbol{W}_h (\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h) \tag{4}$$

其中：⊙为点乘；$\boldsymbol{W}_h$ 和 $\boldsymbol{b}_h$ 分别为整个单元的权重矩阵和偏置项。

4）最后计算得到下一时刻的隐藏层状态 $\boldsymbol{h}_t$，其公式为：

$$\boldsymbol{h}_t = \boldsymbol{z}_t \odot \boldsymbol{h}_t + (1 - \boldsymbol{z}_t) \odot \tilde{\boldsymbol{h}}_t \tag{5}$$

1.3　注意力机制

经过 GRU 网络对时间序列信息的学习，可获得输出向量，然而 GRU 网络对隐藏层向量采用统一的权重分配，模型的预测性能可能达不到预期效果[14]。注意力机制能选择对模型预测更重要的信息予以学习，作为一种权重分配机制，能有效聚合重要信息的影响[15]。

在实际应用中，注意力可以分为软注意力和硬注意力。软注意力考虑了所有信息并采用加权聚合计算，硬注意力则选择输入序列某一位置上的信息。本文选择软注意力作为时序特征提取方法，考虑了每一向量所含的信息，并将重要信息分配更多的权重。计算步骤如下：

首先，根据注意力机制计算要求，获得键值向量（$\boldsymbol{K}$）和查询向量（$\boldsymbol{Q}$）：

$$\boldsymbol{K} = \boldsymbol{W}_k \boldsymbol{H}_i \tag{6}$$

$$\boldsymbol{Q} = \boldsymbol{W}_q \boldsymbol{h}_t \tag{7}$$

其中：$\boldsymbol{H}_i$ 为 GRU 网络隐藏层输出向量；$\boldsymbol{h}_t$ 为 GRU 网络最后一个时刻隐藏层输出向量；$\boldsymbol{W}_k$、$\boldsymbol{W}_q$ 为随机初始化向量矩阵。

然后，计算 $\boldsymbol{K}$ 与 $\boldsymbol{Q}$ 之间的相似度，得到注意力得分 $\boldsymbol{S}_i$：

$$\boldsymbol{S}_i = \boldsymbol{K}^{\mathrm{T}} \boldsymbol{Q} \tag{8}$$

最后，经过 softmax 层得到归一化后的注意力权重 $\boldsymbol{p}_i$：

$$\boldsymbol{p}_i = \text{softmax}(\boldsymbol{S}_i) \tag{9}$$

经过注意力机制后，可得输出向量 $\boldsymbol{a}_i$：

$$\boldsymbol{a}_i = \boldsymbol{p}_i \boldsymbol{H}_i \tag{10}$$

2　融入时序特征财务困境预测方法

为了从股票论坛中抽取时序特征并检验其对财务困境预测的有效性，本文构建了融入时序特征的上市公司财务困境预测研究框架，如图 2 所示。其中：$(E_1, E_2, \cdots, E_i)$ 表示公司样本；$(C_1, C_2, \cdots, C_j)$ 表示每一家公司对应的财务数据；$(t_1, t_2, \cdots, t_n)$ 表示不同时间区间；$(G_1, G_2, \cdots, G_m)$ 表示每一家公司对应的股票 UGC 数据。本框架采用 Attention_GRU 模型对股票论坛中 UGC 时间序列信息进行抽取。同时为了验证提取的 UGC 时序特征的增量作用，将 UGC 时序特征与财务特征进行结合，比较不同财务困境预测模型的效果。模型包括以下 4 个部分。

1）财务特征提取。首先对财务指标中的缺失值采用均值法进行填补；其次为消除各指标间量纲差异，采用归一化方法对数据进行标准化处理；为了降低指标间冗余和存在的共线性问题，采用互信息法进行筛选得到最终的财务特征。

2）时序特征构建。分为时间步长划分、GRU 网络参数调整和注意力机制权重整合 3 个关键步骤。将股吧 UGC 时间序列数据输入经过参数调整的 GRU 网络得到 GRU 隐层输出向量，再经过注意力权重分配后获得的输出向量作为时序特征。

3) 特征融合层。特征融合是将提取的财务特征与时序特征融合,构建新的融入时序特征的财务困境预测体系。

4) 预测模型。选择业界通用的逻辑回归(logistic regression,LR)、随机森林(random forest,RF)、支持向量机(support vector machine,SVM)、极端梯度增强算法(extreme gradient boosting,XGBoost)、自适应提升算法(adaptive boosting,AdaBoost)作为财务困境预测模型[16-18]对提取特征的有效性进行验证。

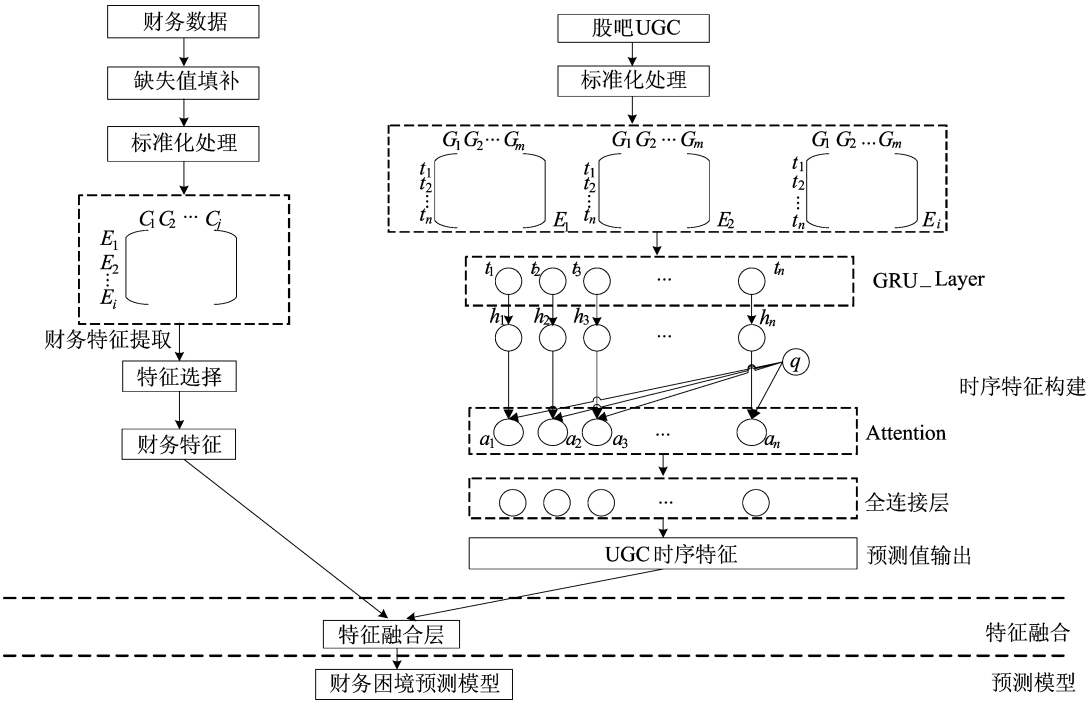

图 2 财务困境预测方法框架图

3 实验与结果分析

3.1 实验数据

本实验以中国上市公司为研究对象,公司被特别处理(special treat,ST)定义为陷入财务困境[19]。上市公司被 ST 指连续两年出现负净利润,或因一年实质性亏损,其每股净资本低于每股面值的公司[18]。上市公司在$(t+1)$年是否被 ST 是参照 t 年的财务数据,因此选取样本时需要至少间隔一年。根据以上要求,本文选取 2018 年未被 ST 的上市公司作为研究对象,选用 2020 年公布的 ST 作为是否陷入财务困境的预测标签,构建二元变量财务困境预测指标,将 ST 公司设为 1,将正常公司设为 0。财务指标从中国经济金融研究数据库中获得,初步得到 3 554 家公司样本,其中初始财务指标 103 个。

从东方财富网获取上市公司股票论坛 UGC 时序数据。为了保持样本区间的一致性,选择了 2018 年全年数据作为股票 UGC 时间序列数据。由于部分公司股票论坛信息较少,时序标签也相对少,影响整个模型的训练效果,为此删除了 2018 年全年中超过 90 d 没有 UGC 更新的公司,最终确定用于建模的样本共 2 953 家。

3.2 财务特征筛选

财务指标分为五大类,分别是发展能力、经营能力、盈利能力、偿债能力和现金流。参考现有文献中的财务指标框架[20-21],初步确定了 27 个初始财务特征。为了降低财务变量间的冗余,进一步提高财务困境预测模型的准确度,采用互信息法对财务特征进行筛选,财务特征筛选结果如图 3 所示。

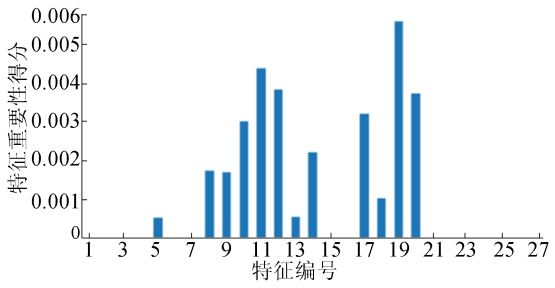

图 3 财务特征筛选结果

筛选后特征及其对应的含义见表 1 所列，财务特征描述性统计结果见表 2 所列。

表 1　财务指标

类别	指标	指标名称
发展能力	C_1	总资产增长率
	C_2	净利润增长率
经营能力	C_3	存货周转率
	C_4	总资产周转率
	C_5	流动比率
偿债能力	C_6	保守速动比率
	C_7	利息保障倍数
	C_8	资产负债率
	C_9	长期负债权益比率
盈利能力	C_{10}	总资产净利润率
现金流	C_{11}	净利润现金净含量
	C_{12}	营业收入现金净含量

表 2　财务指标描述性统计值

指标	平均值	标准差	最小值	最大值
C_1	0.210 8	0.495 4	−0.506 3	12.415 6
C_2	−1.661 2	74.513 2	−3 851.530 0	298.419 6
C_3	102.001 0	2 177.95	0	85 125.400 0
C_4	0.387 0	0.306 7	0	5.655 4
C_5	2.560 7	2.791 0	0	56.780 2
C_6	1.462 2	1.688 0	0	35.478 0
C_7	86.854 4	1 102.870 0	−494.402 0	41 213.010 0
C_8	0.406 7	0.200 4	0.010 5	1.330 4
C_9	0.199 1	0.521 6	−2.840 4	17.619 4
C_{10}	0.027 1	0.049 5	−1.404 0	0.293 8
C_{11}	−1.265 5	30.490 0	−590.670 0	1 224.940 0
C_{12}	−0.038 7	1.457 3	−53.970 0	21.750 0

3.3　基于 Attention_GRU 的时序特征抽取方法

3.3.1　股票 UGC 输入数据

1）数据获取。使用第三方采集器从东方财富网收集数据。根据研究内容，收集每家公司股票论坛中每月 UGC 阅读总量、评论数总量、标题链接下对应的具体内容和 UGC 发布时间。对每条 UGC 的具体内容进行情感分析，得到 UGC 的情感特征，归类为正面、负面、中性，再按月为时间戳统计 UGC 情感数量特征：正面情感 UGC 数量（G_1）、负面情感 UGC 数量（G_2）和中性情感 UGC 数量（G_3），另外还有互动信息特征：阅读数总量（G_4）、评论数总量（G_5）。

2）数据标准化处理。为了消除各指标间的差异，采用 min-max 标准化，将原始数据（G_1，G_2，G_3，G_4，G_5）进行线性变化，并映射到区间［0，1］。

3）时间步长划分。为了充分学习时间变化过程中非财务特征的变化情况，需要选择合适的时间步长将数据划分后再放入模型中。本文按照月份统计各个指标的基本信息，时间步长划分为 12 个月。输入数据按照每一批次处理数量（Batch_size）、时间步长（Seq_len）、UGC 特征数维度（hidden_size）输入到 GRU 网络层中。

3.3.2　GRU 网络参数设置

根据一般隐藏层选择的范围[22]，本文在模型中尝试了 1 层～5 层的隐藏层数设置，同时输出维度1、2、3、4进行实验，最终根据实验效果，确定最终模的参数为：GRU 输入维度为 5，隐藏层数为 2，学习率为 0.01，Batch_size 为 64。优化器选择 Adam 时，模型的训练效果达到最好。

不同隐藏层层数对应的训练误差见表 3 所列。

表 3　不同隐藏层模型训练误差

隐藏层数	1	2	3	4	5
训练误差	0.019 7	0.018 5	0.021 3	0.221 0	0.022 6

从表 3 可以看出，当隐藏层层数为 2 时，模型的训练误差最低，即模型预测效果能达到最高精度。

训练集的训练误差值与训练次数（epoch）的对应关系如图 4 所示。从图 4 可以看出，随着训练次数的增大，训练误差不断减小，训练次数为 6 时训练误差基本保持稳定。本实验中选择训练次数为 10。

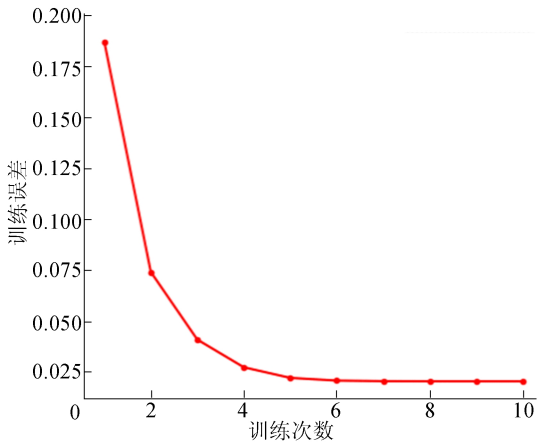

图 4　不同训练次数对应的训练误差值

3.4　评价指标

选用接收者操作特征曲线（receiver operating characteristic curve，ROC）下面积 S_{ROC} 作为评价指标之一。S_{ROC} 指标反映模型对陷入财务困境

公司与正常公司的区分能力。

选用 K-S(Kolmogorov-Smirnov)统计量 K_S 作为另一评价指标,是衡量正、负样本累积分布差异的指标,体现了对正常公司和陷入财务困境公司这两种状态的区分程度。K_S 计算公式为:

$$K_S = \max(T_{PR} - F_{PR}) \tag{11}$$

其中:T_{PR} 表示实际为 1 且预测也为 1;F_{PR} 表示实际为 0 且预测为 1。

S_{ROC} 和 K_S 值越大,模型的预测性能越好。

3.5 对比实验设置

为了验证本文提出的方法在财务困境预测模型中的有效性,选择仅有财务特征(C)作为基准实验。从股票论坛提取的特征作为增量信息,分为 4 种特征提取方法:① 不含时序信息,直接加入股票论坛 UGC 特征全年平均值(G_{avg});② 采用 RNN 提取 UGC 时序特征 G_{RNN};③ 采用 GRU 提取 UGC 时序特征 G_{GRU};④ 采用 Attention_GRU 提取 UGC 时序特征(G_{Att_GRU})。将从股票论坛提取的 4 种特征分别与财务特征融合放入财务困境预测模型中,检验其预测效果。

1) 仅使用财务特征(C)。将经过特征筛选后得到的财务特征作为财务困境预测模型输入,得到财务特征预测效果。

2) 使用财务特征与股票论坛 UGC 特征全年平均值($C+G_{avg}$)。使用财务特征,同时加入股票论坛特征全年平均值 G_{avg} 放入财务困境预测模型作为的对比实验,观察加入股票特征全年平均值在财务困境预测模型中的效果。

3) 使用财务特征与采用 RNN 提取股票论坛 UGC 时序特征($C+G_{GRU}$)。采用 RNN 对序列化数据的特征提取,挖掘数据中的时序信息,作为时序特征提取对比实验,验证加入 RNN 提取股票论坛 UGC 时序特征的预测效果。

4) 使用财务特征与采用 GRU 提取股票论坛 UGC 时序特征($C+G_{GRU}$)。使用 GRU 提取股票 UGC 时序特征与财务特征融合,观察财务困境预测模型效果。

5) 使用财务特征与采用 Attention_GRU 提取股票论坛 UGC 时序特征($C+G_{Att\_GRU}$)。使用 Attention 注意力机制区分不同月度下信息对财务困境预测的重要程度,进一步提高模型预测效果。

3.6 实验结果分析

为了更加准确地度量提取的股票 UGC 时序特征在财务困境预测模型中的效果,选用常用的机器学习模型和一种 AdaBoost 集成学习模型作为财务困境预测模型,并且对比了采用不同方法提取的特征在不同预测模型下的效果,通过 S_{ROC} 和 K_S 的大小衡量所提取特征的预测效果。S_{ROC} 和 K_S 的实验结果分别见表 4 和表 5 所列。

表 4 不同特征在不同预测模型下的 S_{ROC} 值

预测模型	C	$C+G_{avg}$	$C+G_{RNN}$	$C+G_{GRU}$	$C+G_{Att\_GRU}$
LR	0.688 3	0.695 4	0.710 4	0.701 7	0.726 9
RF	0.760 7	0.787 8	0.786 1	0.796 9	0.810 9
SVM	0.650 8	0.655 6	0.676 8	0.686 7	0.701 9
XGBoost	0.772 9	0.775 5	0.809 7	0.812 7	0.820 1
AdaBoost	0.798 7	0.810 3	0.823 1	0.836 3	0.841 8

从表 4 可以得到如下结论:

1) 通过对比 C 和 $C+G_{avg}$ 两组实验,发现直接加入股票 UGC 特征全年平均值的财务困境预测能力优于仅使用财务特征,但是 5 个预测模型有 4 个预测模型的 S_{ROC} 提升都是在千分位上变动,提升效果不明显。

2) 比较 $C+G_{avg}$ 和 $C+G_{RNN}$ 实验结果,采用 RNN 提取股票 UGC 时序特征加入财务困境预测模型中,除了 RF 模型预测的 S_{ROC} 下降0.17%,其他预测模型预测的 S_{ROC} 均上升,且上升幅度均高于 2%;相较于仅使用财务特征,预测效果提升均超过 2%,从股票 UGC 提取的时序特征能显著提升财务困境预测模型效果。

3) 比较 $C+G_{RNN}$ 和 $C+G_{GRU}$ 实验结果,发现采用 GRU 提取股票 UGC 时序特征后,除 LR 模型预测的 S_{ROC} 下降 0.87%,其他预测模型的 S_{ROC} 在采用 RNN 提取时序特征的基础上又有了少量提升。

4) 比较 $C+G_{GRU}$ 和 $C+G_{Att\_GRU}$ 实验结果,采用 Attention_GRU 提取股票论坛 UGC 时序特征加入财务困境预测模型中,预测模型的 S_{ROC} 最好,相较于仅使用财务特征,LR、RF、SVM、XGBoost、AdaBoost 预测效果提升 3.80%、6.83%、5.11%、4.72%、4.30%。

综合以上对比实验，采用 Attention_GRU 提取的 UGC 时序特征在财务困境中取得了最好的效果。因此，本文从股票论坛中提取的 UGC 时序特征能为财务困境预测提供增量信息，并且将时序特征加入到财务困境预测模型中能极大地提升模型的预测精度。

表 5　不同特征在不同预测模型下的 K_S 值

预测模型	C	$C+G_{avg}$	$C+G_{RNN}$	$C+G_{GRU}$	$C+G_{Att\_GRU}$
LR	0.369 3	0.370 6	0.377 5	0.369 3	0.414 3
RF	0.500 6	0.572 8	0.523 4	0.553 5	0.561 6
SVM	0.324 7	0.326 5	0.355 0	0.354 1	0.357 1
XGBoost	0.575 3	0.580 3	0.561 4	0.581 7	0.604 0
AdaBoost	0.716 9	0.710 9	0.697 6	0.711 0	0.737 8

从表 5 可以看出，将使用 Attention_GRU 提取的 UGC 时序特征加入到财务困境预测模型，相比仅使用财务数据，LR 预测模型的 K_S 提升了 4.51%，RF 预测模型的 K_S 提升了 6.10%，SVM 预测模型的 K_S 提升了 3.24%，XGB 预测模型的 K_S 提升了 2.87%，AdaBoost 预测模型的 K_S 提升 2.05%。K_S 指标的提升再次验证了本文提取的股票 UGC 时序特征能有效地为财务困境预测提供增量信息。

4　结　　论

针对现有上市公司财务困境预测研究大多利用财务信息或公司自身发布的年报和管理层讨论分析等文本信息，未考虑投资人和利益相关者的意见信息，尤其是动态时序信息的不足，本文考虑了股票论坛 UGC 的情感时间序列特征以及每个 UGC 阅读量和评论数量的时间序列特征，并采用 Attention_GRU 提取股票论坛中 UGC 的时序特征，结合财务特征预测公司财务困境。在实验中，比较了直接使用股票 UGC 特征全年平均值和 RNN 提取 UGC 时序特征的财务困境模型预测效果，研究发现：

1）股票论坛中含有的 UGC 确实可以为财务困境预测提供增量信息，直接使用股票 UGC 特征的平均值加入财务困境预测模型中时预测模型效果提升不明显，加入 UGC 时序特征后预测模型效果提升显著。

2）采用 GRU 提取股票 UGC 时序特征优于 RNN，采用 RNN 提取股票 UGC 时序特征融入财务困境预测模型效果虽然会有提升，但是相比只使用财务数据预测，并不是每个模型的预测效果都会有明显的提升，而采用 GRU 后，财务困境预测模型的效果会进一步提升。

3）使用 Attention_GRU 提取 UGC 时序特征，财务困境预测模型的预测效果相比只使用财务特征或者财务特征和其他网络模型提取时序特征都要显著，进一步验证了本文所提方法的有效性。

从本文研究中可以得出：① 可以从投资者和利益相关者收集和发布的信息作为财务困境预测的增量信息，补充财务信息的不足，更加全面地反映公司的运营状况；② 充分利用时序信息在上市公司财务困境预测中所起作用，监管机构可以重点关注公司运营过程中时序信息的变化，警惕公司运营可能存在的潜在风险。

在后续研究中，可针对股票论坛中还含有作者信息、粉丝数、回帖信息等，结合作者之间发布内容的关联性、回帖信息的互动性对股票 UGC 做深入处理，进一步研究时序信息对财务困境预测的效用。

[参　考　文　献]

[1] 吴敬琏，张卓远，李京文，等. 中国经济金融研究数据库[DB/OL]. (2021-05-30)[2022-07-28]. https://www.gtarsc.com/.

[2] BATCHELOR T. Corporate bankruptcy: testing the efficacy of the Altman Z-score[J]. International Research Journal of Applied Finance, 2018, 9(9): 404-414.

[3] SUN J, FUJITA H, ZHENG Y J, et al. Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods[J]. Information Sciences, 2021, 559: 153-170.

[4] 刘云菁，伍彬，张敏. 上市公司财务舞弊识别模型设计及其应用研究：基于新兴机器学习算法[J]. 数量经济技术经济研究，2022，39(7)：152-175.

[5] BARUCH L, FENG G. The end of accounting and the path forward for investors and managers[M]. [S. l.]: John Wiley & Sons, Inc., 2016: 1812-1816.

[6] HUANG B, YAO X, LUO Y Q, et al. Improving financial distress prediction using textual sentiment of annual reports

[J]. Annals of Operations Research,2023,330:457-484.
[7] 徐维军,彭子衿,张卫国,等. 基于文本信息考虑投资者情绪的均值回归策略设计:以东方财富股吧发帖文本和 A 股市场为例[J]. 运筹与管理,2022,31(3):193-198.
[8] 陈云,杨晓雪. 基于新闻文本的上市公司财务困境组合预测模型[J]. 计算机应用研究,2017,34(6):1663-1667.
[9] 陈艺云. 基于信息披露文本的上市公司财务困境预测:以中文年报管理层讨论与分析为样本的研究[J]. 中国管理科学,2019,27(7):23-34.
[10] DONG W,LIAO S Y,ZHANG Z J. Leveraging financial social media data for corporate fraud detection[J]. Journal of Management Information Systems, 2018, 35 (2): 461-487.
[11] WANG X J,XIANG Z Q,XU W K,et al. The causal relationship between social media sentiment and stock return: Experimental evidence from an online message forum[J]. Economics Letters,2022,216:110598.
[12] ROSARIO A B,SOTGIU F,DE V K,et al. The effect of electronic word of mouth on sales:a meta-analytic review of platform, product, and metric factors[J]. Journal of Marketing Research,2016,53(3):297-318.
[13] ZHAO S P,XU K,WANG Z,et al. Financial distress prediction by combining sentiment tone features[J]. Economic Modelling,2022,106:105709.
[14] 李洁,林永峰. 基于多时间尺度 RNN 的时序数据预测[J]. 计算机应用与软件,2018,35(7):33-37,62.
[15] VASHISHTH S,UPADHYAY S,TOMAR G S,et al. Attention interpretability across nlp tasks[EB/OL]. [2022-07-06]. http://arXiv/pdf/1909.11218.pdf.
[16] CHEN M Y. Predicting corporate financial distress based on integration of decision tree classification and logistic regression[J]. Expert Systems with Applications, 2011, 38(9):11261-11272.
[17] HUANG Y P,YEN M F. A new perspective of performance comparison among machine learning algorithms for financial distress prediction[J]. Applied Soft Computing, 2019,83:105663.
[18] SUN J,JIA M Y,LI H. AdaBoost ensemble for financial distress prediction:an empirical comparison with data from Chinese listed companies[J]. Expert Systems with Applications,2011,38(8): 9305-9312.
[19] 吕喜梅,蒋翠清,丁勇,等. 融合临时报告软信息的新三板企业财务困境预测研究[J]. 中国管理科学,2023,31(11):140-150.
[20] 肖毅,熊凯伦,张希. 基于 TEI@I 方法论的企业财务风险预警模型研究[J]. 管理评论,2020,32(7):226-235.
[21] 梁墨,李鸿翔,张顺明. 基于 ST 预测的财务困境测度与股票横截面收益[J]. 中国管理科学,2023,31(2):138-149.
[22] STATHAKIS D. How many hidden layers and nodes? [J]. International Journal of Remote Sensing, 2009, 30 (8): 2133-2147.

(责任编辑 李 凯)

(上接第 375 页)

[18] ISTA. International rules for seed testing[M]. Bassersdorf:[s. n.],1985:1-152.
[19] 陆开形,汤飞峰,丁沃娜. 镉胁迫对不同基因型小白菜种子萌发的影响[J]. 宁波大学学报(理工版),2012,25(2):6-16.
[20] SCHABENBERGER O,THARP B E,KELLS J J,et al. Statistical tests for hormesis and effective dosages in herbicide dose response[J]. Agronomy Journal,1999,91(4):713-721.
[21] SEEFELDT S S,JENSEN J E,FUERST E P. Log-logistic analysis of herbicide dose-response relationships[J]. Weed Technology,1995,9(2):218-227.
[22] 杨肖松,刘月仙,解小凡,等. 基于物种敏感性分布法预测茚对白菜毒害的生态风险阈值[J]. 农业环境科学学报,2018,37(10):2127-2134.
[23] LI U S,LIU M,WANG X,et al. Absorption and accumulation of cadmium in different Chinese cabbage cultivars [J]. Advanced Materials Research,2013,26:21-27.
[24] SHAO Q. Estimation for hazardous concentrations based on noec toxicity data: an alternative approach[J]. Environmentrics,2000,11(5):583-595.
[25] 杨彬,李婷,张野,等. 不同价态外源硒对芥菜生长及富硒量的影响[J]. 辣椒杂志,2019(3):5-10.
[26] 胡婷,李文芳,向昌国,等. 硒对常见蔬菜种子萌发的影响及在植株中的分布[J]. 食品科学,2015,36(7):45-49.
[27] 谢文文. 芸薹属主要蔬菜作物富硒比较研究[D]. 重庆:西南大学,2018.
[28] 毛晖,王朝辉. 硒的价态与浓度水平对 6 种植物种子发芽和根际生长的影响[J]. 农业环境科学学报,2011,30(10):1958-1965.

(责任编辑 吴 亮)