doi:10.19920/j.cnki.jmsc.2026.01.008

基于工商注册信息的中小微企业信用评价研究①

李 铁[1], 寇 纲[2], 彭 怡[1*]
(1. 电子科技大学经济与管理学院, 成都 611731; 2. 西南财经大学工商管理学院, 成都 611130)

摘要: 企业信用的量化评价是我国建立高效市场经济监管制度的重要基石. 由于缺乏公开的经营和财务数据,针对中小微企业的信用评价一直是征信领域的一个难题. 鉴于互联网上存在大量公开易得的中小微企业工商注册信息,本研究提出通过挖掘其中的企业名称、企业间投资关系、经营范围等非常规、非结构化的数据,提升对中小微企业信用评价的效果. 为了对非结构化、异质的工商信息进行结构化表征和融合,本研究针对工商文本提出了一种能够提取文本中的序列信息并将文本呈现为数值向量的门控循环神经网络表征学习方法;针对企业间的投资关系提出了一种能够编码图结构信息并将图嵌入到向量空间的图注意力神经网络表征学习方法. 另外,由于征信领域对评价结果可解释性的高要求,本研究提出了文本信用评价的可解释方法和信用风险传递路径的可解释方法. 基于某市约6.8万家企业工商信息和对应参考信用评分的实验结果表明,企业名称和企业间投资关系均蕴含了传统的数值型工商字段所缺乏的信用信息,能显著提高中小微企业的信用分类准确率. 本研究的结果表明,工商注册信息可以作为当前中小企业信用评价体系的有效补充,对改善当前中小微企业信用信息稀缺的难题有重要意义.

关键词: 企业信用评价; 工商注册信息; 文本数据挖掘; 图数据挖掘; 异质数据融合
中图分类号: F832 **文献标识码**: A **文章编号**: 1007-9807(2026)01-0130-13

0 引 言

随着我国社会主义市场经济制度改革向纵深发展,企业信用的量化评价成为政府制定产业政策和金融机构融资管理的重要依据[1]. 我国现有的企业信用评价体系主要是针对大型集团企业或上市公司制定的,而亟需政策与资金扶持的中小微企业,由于公开的业务数据少,面临着量化信用评价的难题[2]. 与此同时,互联网上可以公开获得的海量文本、图片、关联关系等非结构化信用数据,仍未得到充分利用. 如何从公开的非结构化数据中进一步提炼企业的信用信息,是征信领域当前亟需解决的难题之一[3].

目前,主流的企业信用评价方法主要建立在数据挖掘的基础之上. 根据有无信用标签,可分为监督式评价和无监督式评价[4]. 监督式评价依赖信用标签,通常使用分类器或者回归方法. 例如,肖斌卿等[5]提出了基于模糊神经网络的信用评分方法; Sebastián等[6]提出了基于代价敏感支持向量机的特征选择方法; Verbraken等[7]提出了以利润作为分类标签的信用评估方法; 黄宝凤结合企业的动态交易数据和数值型工商注册数据,并以天眼查评分为参考标签,建立了小微企业信用评价体系[8]. 无监督式评价主要是建立在数据特征的统计基础之上. 例如,Deo等[9]提出了基于拟合最大似然估计和高斯过程的企业信用评价模

① 收稿日期: 2021-02-02; 修订日期: 2023-02-17.
基金项目: 国家社会科学基金资助项目(23BTJ040); 国家自然科学基金资助项目(72471047; 71910107002).
通讯作者: 彭 怡(1975—), 女, 四川成都人, 博士, 教授, 博士生导师. Email: pengyi@uestc.edu.cn

型；Alexander 等[10]提出了基于参数估计协方差矩阵目标选择的信用评分方法；Fernandes 等[11]提出了基于空间相关性的信用风险评估方法；Hong 等[12]使用蒙特卡洛方法评估了信用风险相关因子的敏感性. 可以看出，以上评价体系主要是针对数值化、结构化的特征而展开，无法应用于企业间的关联关系和文本等非结构化数据. 而近期研究表明，企业之间的关联关系[13]和新闻财经报道[14, 15]等信息也是监测风险传递和评价企业实力的重要切入点.

针对企业之间的关联关系，信用体系通常从两个方面分别进行评估，一方面是对企业影响力的评估（正向指标），另一方面是对企业信用风险的评估（负向指标）[17]. 企业影响力方面，网络节点的“度”经常被用来衡量节点企业的重要性[18]；Freeman[19]提出的介数指标衡量了节点的繁忙程度，可以用于评估企业在风险传播中的关键性；而网页重要性评价的一些算法，如 Wei 等[18]也可以用来评价企业节点的影响力. 企业信用风险方面，Battiston 等[20]从复杂网络的角度研究了金融风险的检测理论方法，并提出了金融规则制定的原则；周开国等[21]基于市场间的关联关系建立了风险传递网络；陈道富和刘新海[21]从网络图角度研究了担保圈的金融风险问题，并提出相应的风险管控策略；Van 等[23]提出了一种通过社会网络检测企业欺诈性破产倒闭以及逃税的风险.

随着深度神经网络在文本领域的突破性进展，文本中序列化语义信息的向量化编码成为可能. 这为文本数据在信用评价中发挥作用提供了契机. 例如，姚加权等[14]提出了基于情绪词典的金融风险分析方法；俞红海等[15]针对企业首次公开募股文本披露信息提出了基于主题模型的分析方法；洪亮等[16]提出了基于金融股权知识大图的风险关联发现算法. 总体上来说，相对数值型和关联关系（网络）型数据，基于文本的信用评价研究仍处于起步阶段.

通过以上文献梳理，可以发现当前针对中小微企业的信用评价有如下瓶颈：现有的企业信用评价体系，多以结构化的财务报表和市场交易等定量数据为基础，而征信机构对中小微企业进行信用评估时，很难收集到这些传统的财务定量指标数据. 在使用了工商注册信息的研究中，也仅使用了易于数值化的字段（如注册资本、成立时长等）[8]，工商注册文本、投资关系等异质的、非结构化的数据中所蕴含的潜在信用价值仍未被充分发掘.

本研究的目的是探讨如何提炼工商注册中的文本型和关系型这两种非常规金融数据中的信用信息，并用于中小微企业信用评价. 具体贡献包括：1）提出了工商注册信息中文本、投资关系等非结构化数据的结构化表征学习方法；2）建立了文本类信息的信用评价结果、信用风险传递路径的解释方案. 总体研究框架如图 1 所示.

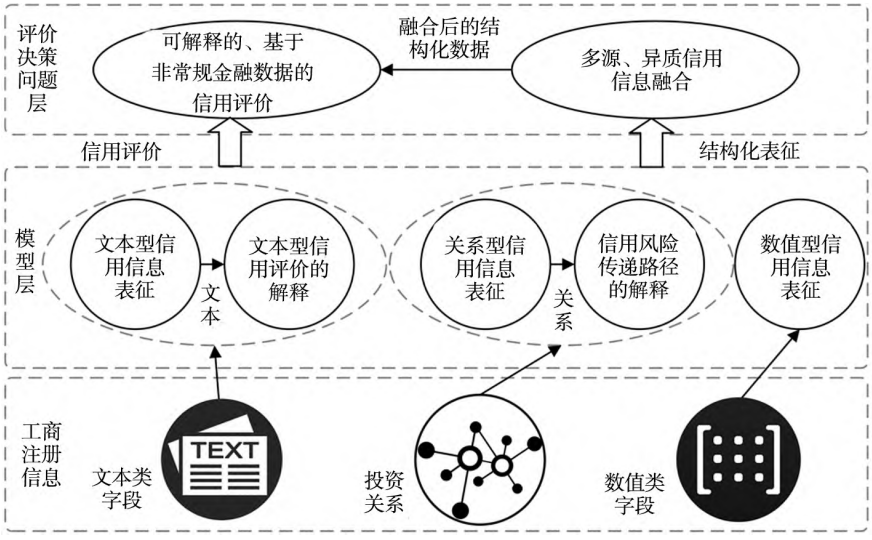


图 1　研究框架

Fig. 1 Research framework

1 数据、特征及标签

本研究共收集了某市高新区约 13 万家企业的工商注册信息,并以该区政府所提供的企业信用评分——“钛金分②”作为标签划分依据. 下面对拟使用的主要字段信息加以介绍.

1.1 工商注册文本类信息

工商注册信息中的文本,如企业名称、注册地址和经营范围,在某种程度上反映了公司的行业、规模和地理区域,可能与企业的信用状况具有相关性. 本研究将分别针对企业名称、注册地址、经营范围进行结构化表征学习并用于信用评价.

1.2 投资关联关系网络

工商注册信息中包含有股东投资情况,而企业之间的投资关系和共有股东关系体现了企业之间的资本关联性. 本研究将企业视为节点、企业之间的直接投资或共有股东关系视为边,然后以这些节点和边构建投资关系网络图. 这种关系型数据既能体现企业的资本潜力,也能体现企业之间潜在的信用风险传递,因此是重要的信用信息源.

1.3 数值型工商信息

工商注册信息中除文本和投资关系外,还有其它如公司成立时间、注册资本等数值型数据可用于信用评价. 这些也是传统信用评价方法常使用的工商内容[8],具体如表 1 所示.

表 1 数值型工商字段

Table 1 Numerical fields of business registration

字段	说明
start_date	成立时间,转换为了成立到当前的天数(续存时间);
econ_type	企业类型,如“有限责任公司”、“有限合伙企业”,共有 75 种,将进行独热(One-Hot)编码;
regi_capital	注册资本,以万为单位;
check_date	上一次年检时间,转换为了年检到当前时间的天数;
capital_type	注资币种,如人民币、美元等,共 8 种,将进行独热编码;
change_times	近三年内的工商变更次数.

1.4 参考信用评分及分类标签划分

前述的“钛金分”是某市高新区管委会综合了企业的工商简项、纳税欠税、违法涉诉、行政处罚、动产质押、经营异常、法院公告、股权出质、知识产权、失信执行等信息的一种综合性、试用性评分方案,主要为该区高新技术企业孵化和高新产业政策制定提供参考依据. 为了使信用的区分度更好,本研究取信用评分最低的约 25%企业作为“低信用”样本,取信用评分最高的约 25%企业作为“高信用”样本,共约 6.8 万条,如表 2 所示.

表 2 信用分及信用标签划分依据

Table 2 The credit scores and the credit labeling method

“钛金分”分值区间	分位段	分类标签	数据量	总数据量
[350, 664)	Q1	0	33 481	136 510
[664, 678)	Q2		34 127	
[678, 692)	Q3		33 879	
[692, 850]	Q4	1	35 023	

本研究使用该评分的目的一方面是检验工商注册信息和现有评分的关联关系,另一方面是反推当前信用评价体系存在的不足和可以改进的方向.

1.5 中小微企业界定

由于我国当前对中小微企业的划分标准主要是基于年营业收入,而本研究所使用的工商注册信息无法体现年营收情况,因此界定了以下标准进行大中小微企业划分(见表 3).

表 3 大中小微企业划分依据

Table 3 Large, medium and micro enterprises division method

划分标准	大型	中型	小型	微型
注册资本 Y/(万元)	$Y\geqslant$ 10 000	1 000$\leqslant Y<$10 000	100$\leqslant Y<$1 000	$Y<100$

基于以上划分标准,表 4 统计了不同信用类别的数量和比例.

表 4 不同信用类别在各企业类型中的分布数量及比例

Table 4 Distribution quantities and proportion of different credit labels in each enterprise type

企业类型	数量/比例	高信用数量/比例	低信用数量/比例
大型	1 238/1.81%	629/1.80%	609/1.82%
中型	14 542/21.23%	9 404/26.85%	5 138/15.35%
小型	36 215/52.87%	19 913/56.86%	16 302/48.69%
微型	16 509/24.10%	5 077/14.50%	11 432/34.14%
总计	68 504/100%	35 023/100%	33 481/100%

② 脱敏后的虚拟名称,非真实评分体系名称.

如表 4 所示，注册资本在 1 亿元以下的中小微企业占比 98.19%，显示了中小微企业在本样本中的重要性. 另外，高和低两种信用类别在不同企业类型中的分布具有显著的结构性差异. 这种差异会在数据的分布模式上形成众多的局部子模式，使不同标签类别的数据在总体上线性不可分[17].

2　基于工商信息的信用评价及结果解释

信用评价领域建模常用的思路是：1）简单特征工程 + 复杂分类器；2）复杂特征工程 + 简单分类器[28]. 本研究采用第二种思路：使用较复杂的模型将原始线性不可分的特征呈现为一种较简单的数据分布形态，即表征学习，然后使用简单的信用评价模型，如逻辑斯蒂回归（Logistic Regression，LR）或线性支持向量机（Linear Support Vector Machine，L-SVM），即可对变换后的特征进行分类或评分. 另外，由于经济与管理模型必须是对客户、业务人员和管理人员可解释的[30]，可解释性也将是本研究的重点内容.

2.1　文本型工商信息的表征与可解释信用评价

基于门控循环单元（Gate Recurrent Unit，GRU）提出文本类工商注册信息的表征与可解释信用评价方案. GRU 是一种长短记忆神经网络（Long Short-Term Memory，LSTM）的优化版本，具备原 LSTM 相近的性能，但计算量大幅降低[31].

1）基于 GRU 的文本型工商信息表征

GRU 的定义如下[31]

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \hat{h}_t ,$$
$$z_t = \sigma(\boldsymbol{W}_z x_t + \boldsymbol{U}_z h_{t-1} + b_z),$$
$$\hat{h}_t = \tanh(\boldsymbol{W}_h x_t + \boldsymbol{U}_h (r_t \odot h_{t-1}) + b_h),$$
$$r_t = \sigma(\boldsymbol{W}_r x_t + \boldsymbol{U}_r h_{t-1} + b_r)$$

其中 $\sigma(\cdot)$ 为 Logistic 函数，x_t 为第 t 个词汇的向量输入（可用 Word2Vec、GloVe 或 BERT 向量），$z_t \in [0,1]^d$ 被称为更新门，d 为原始词汇向量维度，$\hat{h}_t$ 表示时刻 t 的候选状态，$r_t \in [0,1]^d$ 被称为重置门，用来控制候选状态 $\hat{h}_t$ 对上一时刻状态 h_{t-1} 的依赖程度，$\boldsymbol{W}$ 与 $\boldsymbol{U}$ 均为线性变换矩阵，b 为偏置. 可以看出，相较一般的神经网络，GRU 的核心特点就是通过 z_t、r_t 等门控机制学习各个时刻隐态之间的依赖关系. 由于本研究使用 GRU 进行信用分类，因此还在最后一个词汇的 GRU 输出基础上加了一层线性分类层，即

$$\hat{y}_i = \underset{\text{index}}{\operatorname{argmax}}(\boldsymbol{W} h_{i\text{-length}} + b)$$

其中 $\boldsymbol{W} \in \mathbb{R}^{2\times d}$，$b \in \mathbb{R}^2$（二维偏置项向量，因为本研究的信用有两种类别，因此输出层有两个神经元，每个神经元都需要一个偏置值），i-length 为第 i 个文本字段句子所包含的词汇长度，$\hat{y}_i$ 为句子的预测标签. 由于该神经网络的最后一层为线性分类器，在训练过程中，会最大程度地将 $h_{i\text{-length}}$ 的值逐渐呈现为一种线性可分的理想状态，即实现文本语句的表征学习. 本研究使用该值作为第 i 个文本字段句子的向量表征，即

$$\boldsymbol{O}^{\text{Text}}_{i\text{-Target}} = \boldsymbol{h}_{i\text{-length}}$$

2）基于 GRU 的文本型信用评价解释

基于词汇在 GRU 信用分类过程中的作用，提出一种新的基于文本对信用评价结果进行解释的方法. 由于前述 GRU 文本向量化表征模型的最后一层是线性分类器，企业名称等工商注册文本经过神经网络的各层到 GRU 的输出层后已被映射为简单的信用线性可分状态，可以被认为是一种理想的分布态. 由于 GRU 是一种循环神经网络，词汇向量的输出不仅取决于当前词汇的向量编码，还取决于前一个词汇的隐态编码. 企业名称等完整工商文本的词汇序列所对应的编码由语汇向量和隐态编码构成，句子中的词汇犹如一个个小的动量，将句子向量逐步拉动向用于信用分类任务的理想空间区域. 词汇序列中每一个词在 GRU 层的输出可以被理解为一条优化路径，而这条路径上每增加一个词汇所行走的方向和距离，可以被认为是该词汇对最后分类任务的边际贡献. 当信用分类结果正确时，本研究把这种影响力解读为词汇对信用评价的边际影响，如图 2 所示.

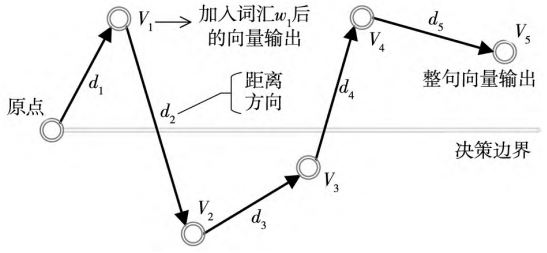


图 2　基于向量差分的词汇贡献解释示意图

Fig. 2 Word contribution interpretation based on vector differencing

根据以上思路,本研究对边际词汇向量的行走方向、距离、贡献、影响力作了如下量化定义

$$\cos\theta_{it} = \frac{\boldsymbol{V}_i^{\mathrm{T}}(\boldsymbol{V}_{it} - \boldsymbol{V}_{i(t-1)})}{\| \boldsymbol{V}_i \| \cdot \| \boldsymbol{V}_{it} - \boldsymbol{V}_{i(t-1)} \|},$$

$$d_{it} = \| \boldsymbol{V}_{it} - \boldsymbol{V}_{i(t-1)} \|,$$

$$Contribution_{it} = \cos\theta_{it} \times d_{it},$$

$$Influence_{it} = \frac{contribution_{it}}{\sum_{t=1}^{m} contribution_{it}},\ \text{if}\ \hat{y}_i = y_i$$

其中 $\boldsymbol{V}_{it}$ 为加入第 t 个词汇后的目标句子向量输出,$\boldsymbol{V}_i$ 为整个句子 i 的向量输出,$\cos\theta_{it}$ 为第 t 个词汇的边际行走向量与整个句子向量的夹角余弦,d_{it} 为第 t 个词汇所行走的边际距离,$Contribution_{it}$ 为第 t 个词汇的边际贡献. 可以看出,增加词汇 t 后的行走方向与句子向量越一致、行走距离越远,边际贡献越大. $Influence_{it}$ 为第 t 个词汇在句子 i 中的边际影响力,是边际贡献归一化后的结果.

按照以上机制,基于 GRU 对工商注册文本字段的分类结果,可以根据词汇所对应的边际影响力来解读每个词汇对当前分类结果所发挥的作用. 另外,在对每个工商文本句子中的高影响力词语进行累加统计后,可以从总体上解释工商注册信息中词汇与信用评价结果的相关关系.

2.2 投资关系型工商信息的表征、信用评价及信用风险传递路径解释

针对企业之间由直接投资关系或者共有股东关系所形成的投资网络图,本研究使用图注意力神经网络(Graph Attention Network, GAT)[32]进行图数据的结构化表征. 相对于传统的图卷积神经网络(Graph Convolutional Network,GCN)等模型,GAT 最大的优点是能够表征企业节点间长距离的依赖关系.

1)基于 GAT 的关系型工商信息表征

GAT 由堆叠的图注意力层实现. 图计算力层针对“节点对 (i,j)”计算注意力系数 a_{ij} 的方式为

$$a_{ij} = \frac{\exp(\sigma(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i \oplus \boldsymbol{W}\boldsymbol{h}_j]))}{\sum_{k\in N_i} \exp(\sigma(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i \oplus \boldsymbol{W}\boldsymbol{h}_k]))}$$

其中 N_i 表示节点 i 的邻节点集合,$\boldsymbol{h}_i$ 为节点的原始输入特征向量,$\boldsymbol{W}$ 为每一个节点上应用的线性变换权重矩阵,$\boldsymbol{a}$ 为权重向量.

为了稳定学习过程,GAT 一般使用多头注意力机制,即同时学习 M 个独立的注意力机制,然后取 M 个注意力变换的输出向量的均值或拼接向量然后进行非线性激活作为表征输出.

$$\boldsymbol{h}_i' = \sigma\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{j\in N_i} a_{ij}^{(m)}\, \boldsymbol{W}^{(m)}\, \boldsymbol{h}_j\right)$$

其中 $a_{ij}^{(m)}$ 和 $\boldsymbol{W}^{(m)}$ 分别表示第 m 个注意头中的注意力系数和线性变换矩阵,σ 为非线性激活函数. 从上式可以看出,GAT 仍需要使用企业节点本身的原始特征 $\boldsymbol{h}_j$ 作为输入. 但如果使用企业的数值或者文本表征特征作为输入,在学习结果中将难以区分网络图结构特征和固有特征的各自贡献. 为解决这种问题,本部分直接使用独热方式对企业节点进行唯一性编码,作为 GAT 的输入.

同 2.1 节类似,为了实现信用分类任务,本部分也在 GAT 多头注意力输出的基础上加入一个线性分类层. 由于最后一层为线性分类器,在训练过程中,神经网络会最大程度将 GAT 多头注意力的输出逐渐呈现为一种线性可分状态. 本部分使用 GAT 的输出作为关系型信息的向量表征,即

$$\boldsymbol{O}_{i\text{-Target}}^{Graph} = \boldsymbol{h}_i'$$

2)基于 GAT 的信用风险传递路径解释

GAT 中的注意力系数体现了信用评价过程中两个企业节点之间关联关系的重要程度. 基于此,本部分提出了一种使用注意力系数解释信用风险传递路径的方案,如算法 1 所示.

算法 1 信用风险传递路径解释

输入 V_N – 所有企业节点;信用评价模型 $cls(\cdot)$

输出 RiskPath – 信用风险传递路径

算法

```
for v_i in V_N
if cls(O^Graph_i-Target) == 'Q_1' and cls(O^Other_i-Target) != 'Q_1'
    for v_j in V_N
        ā_ij = (1/M) Σ_{m=1}^{M} a_ij^(m)  #多头注意力值取平均
        if ā_ij ≥ δ : # δ 为注意力阈值,默认值 1E-3
            add (v_i, v_j) to RiskPath
        end if
    end for
end if
end for
```

算法 1 的核心思想是筛选出从企业的数值和文本特征角度看为信用正常(高信用,Q4),而从

投资关系角度看为信用异常(低信用,Q1)的数据. 本研究认为导致这种现象的原因是:其它异常企业通过投资关系将风险传递给了原本信用应正常的企业. 通过捕捉企业间投资关系的重要程度,可在众多投资关系中筛选出主要的风险关系,建立易于解释的企业信用风险传递路径.

2.3　数值型工商注册信息的表征与信用评价

由于 1.3 节所介绍的数值型工商信息可以直接向量化,本部分将原始线性不可分的数值向量变换为分布较为简单的线性可分状态,以期提升线性分类器的信用评价质量. 本部分使用多层感知器(Multi-Layer Perception, MLP)对原数据进行非线性变换. MLP 中的每一层为一种"仿射变换" + "非线性激活",数据输入多层堆叠的 MLP 时,相当于进行了非线性变换. MLP 的定义如下

$$z_l = \boldsymbol{W}_l a_{l-1} + b_l \ , \ a_0 = X_b \ , \ a_l = \sigma_l(z_l)$$

其中 a_{l-1} 为第 $l-1$ 层的输出,也为第 l 层的输入,X_b 为原始数据输入,z_l 为第 l 层的仿射变换结果,$\boldsymbol{W}_l$ 为第 l 层的线性变换矩阵,也代表第 l 层的连接权重,b_l 为第 l 层的偏置,σ_l 为第 l 层的非线性激活函数.

本研究使用 MLP 倒数第二层的输出作为数值型工商信息的向量表征,即

$$O_{\text{Target}}^{Numeric} = a_{\text{total}-1} = \sigma_{\text{total}-1}(z_{\text{total}-1})$$

其中 total 为 MLP 的总层数. 由于 MLP 的最后一层是一个线性分类器,在神经网络的训练过程中,会最大程度地将倒数第二层的输出 $O_{\text{target}}^{\text{T}}$ 逐渐呈现为一种线性可分状态,解决 1.5 节所述的结构性差异所导致的线性不可分问题,完成表征学习.

另外,数值型工商信息最终用于信用评价的特征是原数值字段的非线性重组,已有大量此方面的可解释性研究[26,27],本处不再赘述.

2.4　异质工商信息的融合

前述各部分已将文本和关系型数据进行了向量化,现需拼接各部分的输出向量. 由于各部分的向量表征都是线性可分的,拼接后的整体向量仍然线性可分. 集成的拼接向量为

$$\boldsymbol{X} = O_{\text{Target}}^{\text{Text1}} \oplus O_{\text{Target}}^{\text{Text2}} \oplus O_{\text{Target}}^{\text{Text3}} \oplus O_{\text{Target}}^{\text{Graph}} \oplus O_{\text{Target}}^{Numeric}$$

在以上拼接的向量基础上,训练 LR 等线性分类器,即可进行综合信用评价或者评分.

3　信用评价实验及结果讨论

本研究使用 Spark 对工商注册数据进行了预处理,使用 PyTorch 实现了第 2 章中所提出的 GRU、GAT、MLP 等工商信息表征模型. 所有的实验均在 1 台 10 核 I9 英特尔处理器、64G 内存的联想服务器上展开.

3.1　基于文本类工商信息的信用评价实验

本部分使用企业名称、注册地址、经营范围三种工商文本类字段训练了前文所述的 GRU 模型(基于"人民日报"语料的 GloVe 预训练词向量③作为输入,词向量维度为 300),然后使用 GRU 最后一层的输出作为文本的向量化表征,其中 GRU 设置了 2 层,线性仿射变换层设置了 1 层,隐层单元数均为 64,三种文本向量表征的维度均为 64.

1)文本类信息表征学习的性能评估

在以上模型输出的向量表征和已有信用标签的基础上,本部分使用三种在信用评价领域常用的"广义线性模型":LR、L-SVM 和线性判别分析(Linear Discriminant Analysis, LDA)对结果进行评估. 表 5 列出了三种信用评价模型在工商注册文本表征的结果上进行训练后的三折交叉检验结果.

从表 5 可以看出以下较为显著的现象.

a)基于企业名称的最高信用分类准确率高达 94.63%. 这一方面说明企业名称和"钛金分"信用评分数据具有很高的相关性(企业名称中所体现的企业类型和行业等信息,可以在很大程度上解释本研究所使用的"钛金分"的评分依据),另一方面也说明本研究提出的文本表征方法是一种有效的信用信息呈现方法.

b)基于注册地址和经营范围的最高分类准确率分别为 65.04%和 76.49%. 说明它们并不是完全的随机变量,和"钛金分"有一定的相关性. 注册地址分类的准确性不高,说明其并不是很强的决定性因素. 经营范围的准确率中等,说明其相对较为重要.

c)企业名称较高的 AUC 得分和较低的交叉

③　https://github.com/Embedding/Chinese-Word-Vectors.

检验折间标准差也反映了使用其进行信用评价时 鲁棒性要优于使用注册地址和经营范围.

表 5 三种文本类工商注册信息的单独信用评价表现

Table 5 Respective performances of the credit evaluation based on three types of textual business registration information

信用评价模型	文本字段	TPR/%	Accuracy/%	F-Measure/%	AUC/%	MCC/%
LR	企业名称	94.11 ±0.17	94.62 ±0.11	94.48 ±0.12	99.07 ±0.04	89.24 ±0.22
	注册地址	56.68 ±0.51	65.04 ±0.10	61.31 ±0.26	71.21 ±0.16	30.15 ±0.19
	经营范围	67.33 ±0.26	76.49 ±0.04	73.68 ±0.05	83.41 ±0.16	53.56 ±0.12
L-SVM	企业名称	93.91 ±0.15	94.63 ±0.10	94.47 ±0.11	94.61 ±0.10	89.25 ±0.20
	注册地址	56.67 ±0.48	64.98 ±0.08	61.27 ±0.26	64.80 ±0.09	30.03 ±0.15
	经营范围	67.60 ±0.27	76.48 ±0.04	73.75 ±0.06	76.29 ±0.04	53.49 ±0.14
LDA	企业名称	94.11 ±0.13	94.56 ±0.11	94.24 ±0.11	98.91 ±0.05	89.11 ±0.22
	注册地址	57.17 ±0.69	65.01 ±0.24	61.49 ±0.42	71.14 ±0.14	30.05 ±0.46
	经营范围	67.71 ±0.26	76.45 ±0.03	73.76 ±0.06	83.33 ±0.16	53.41 ±0.10

注:“±”符号后为交叉检验折间标准差.

2)文本类信用评价结果的解释

前述实验结果表明,企业名称字段在中小微企业信用分类中具有重要作用. 据此可以认为企业名称经表征学习后处于一种“理想分布态”. 而 2.1 节中所提出的解释方案,可以对企业名称中每个词汇对促进这种“理想分布态”的边际贡献和影响力进行解释. 例如,一家被正确识别为“高信用”的名称为“艾斯玛特仪器贸易有限公司”的企业,其名称中每个词汇对导致识别为“高信用”的边际影响力为

艾斯玛特:0.18,仪器:0.25,贸易:0.21,有限:0.23,公司:0.13

最后,通过对每一个企业名称中的高边际影响力词汇进行累加统计,可获得文本词汇中与总体信用评价强相关的高频词汇,如表 6 所示.

表 6 与信用强相关的高频词汇

Table 6 High-frequency words strongly associated with credit evaluation

“高信用”强相关高频词汇(排名前 100)	“低信用”强相关高频词汇(排名前 100)
科技、工程、技术、设备、贸易、建设、电子、商务、家具、生物、国际、机电、医疗、软件、智能、投资、医药、材料、智慧、连锁、器械、品牌、通信、酒业、服饰、电气、自动化、世纪、机械、互动、基金、股份、家居、未来、信息、精密、时代、食品、发展、鑫业、联创、知识、蓝图、石油、环保、制造、科创、互联、优品、仪器、电梯、系统、化工、天成、光电、天地、股权、电器、用品、控制、药业、产权、展览、高科、宏图、数码、开发、珠宝、集团、立方、产业、工业、智创、电科、基业、互娱、创新、照明、安全、农业、化妆品、配件、众联、旅游、智联、在线、众合、孵化器、天下、制冷、联盟、生活、测控、轨道、物联、空间、魔方、智控、环球、伟业	经营部、服务、文化、合伙、建筑、传媒、餐饮、网络、中心、＊＊市、装饰、管理、物业、建材、分店、汽车、商贸、服务部、营销、专业、餐饮店、诊所、咨询、传播、合作社、体育、美容、五金、酒店、养殖、家政、商业、教育、健身、租赁、保洁、家庭、学校、广告、饮店、户外、供应链、影业、农场、清洁、服装、房产、分社、第一、绿化、人力、健康、美发、事务所、产品、俱乐部、天宇、中和、旅行社、联合、图文、盛景、财产、眼镜店、创富、会计师、汇通、运动、星辰、大道、劳务、房屋、网吧、大成、培训、大地、中恒、建工、信达、运输、实业、门诊部、市场、饮品、装修、暖通、家园、税务师、财富、岩土、共享、门诊部、先生、招标、保利、汇金、嘉业、生鲜、鞋业、华鑫

注:“＊＊市”指的是近期有部分企业并入该高新区的某邻近地级市.

从表 6 可以看出,“高信用”强相关的企业名称常用词汇主要集中在科技、贸易或其它投入较多的行业;而“低信用”强相关的企业名称常用词汇主要集中于中小型的传统产业.

3.2 投资关系类信息表征学习的性能评估

下面把企业作为节点、企业间的投资关系作为边,然后取出表 2 中的高、低信用企业,构建了企业的投资关系网络图.

1)投资关系网络特征分析

在本研究的实验数据中投资关系数据为 9 298(并不是所有企业都和其它企业有投资关系,投资关系数据通常仅涉及全部企业中的一部分),其中高、低信用企业数分别为5 429和 3 869. 投资关系网络特征统计如表 7 所示.

表 7　网络图特征统计

Table 7 Statistical properties of the network graph

企业节点数	边数量	平均度	网络直径	平均路径长度	平均聚类系数	模块度	社团数	弱连接组件数
9 298(Q4:5 429,Q1:3 869)	12 042	2.59	17	6.22	0.77	0.96	3 176	3 161

图 3 展示了使用 Gephi 的 OpenOrd 布局对上述投资关系网络进行可视化的结果.

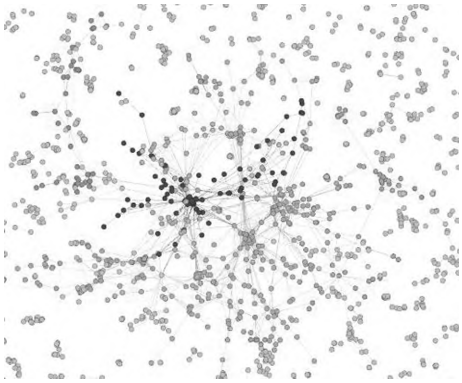


图 3　投资关系网络图中的节点及社团分布图

Fig. 3 Distribution of the nodes and communities in the investment relationship graph

从图 3 中的节点分布及表 7 中的模块度、社团数等统计指标可以看出,投资关系图中分布着大量的社团,即部分企业群之间存在着紧密的投资关联关系.

2)投资关系信息的表征学习结果

使用 2.2 节所述的 GAT 模型将上述投资关系网络进行了向量表征. GAT 的注意力头数量和向量表征维度分别设置为了 32 和 8. 为了检测 GAT 的相对性能,本部分还引入 GCN(2 层, 64 神经元/层)与 GraphSage 图表征方法进行结果对比. 表 8 列出了信用评价的实验结果.

表 8　使用 GAT 及对比方法将关系型工商信息向量表征后的信用评价结果

Table 8 The performances of the credit evaluation based on relational business registration information vectorized by GAT and its comparative methods

信用评价模型	表征学习方法	TPR/%	Accuracy/%	F-Measure/%	AUC/%	MCC/%
LR	GAT	**70.54 ±1.26**	**84.86 ±0.61**	**79.49 ±0.94**	**78.25 ±0.47**	**69.18 ±1.24**
	GCN	69.32 ±0.69	83.15 ±0.13	77.39 ±0.24	77.47 ±0.23	65.37 ±0.30
	GraphSage	62.08 ±0.94	83.55 ±0.83	71.21 ±0.93	71.58 ±0.52	55.32 ±1.89
L-SVM	GAT	**70.46 ±1.26**	**84.86 ±0.62**	**79.47 ±0.95**	**82.79 ±0.70**	**69.20 ±1.26**
	GCN	65.39 ±0.60	82.30 ±0.25	75.46 ±0.11	79.87 ±0.13	63.96 ±0.71
	GraphSage	63.21 ±1.30	82.37 ±4.36	70.29 ±1.26	69.37 ±1.35	56.72 ±5.96
LDA	GAT	**70.46 ±1.26**	**84.87 ±0.58**	**79.48 ±0.91**	**78.21 ±0.47**	**69.22 ±1.17**
	GCN	62.01 ±0.24	81.66 ±0.10	73.78 ±0.10	77.55 ±0.20	63.10 ±0.29
	GraphSage	62.36 ±0.86	83.17 ±0.59	66.85 ±0.91	65.98 ±0.47	55.61 ±1.03

注:在每种信用评价模型上表现最佳的表征结果被标粗;“±”符号后为交叉检验折间标准差.

从表 8 可以看出,GAT 在所有信用评价模型上表现均为最佳,基于投资关系的最高信用分类准确率为 84.87%. 实验结果一方面说明企业间的投资关系和本研究所使用的“钛金分”有较高的相关性,是重要的信用评价信息源;另一方面也说明所提出的表征方法能够较好地呈现投资关系中的信用信息.

3)信用风险传递路径的解释

基于 2.2 节中的信用风险传递路径可解释算法 1,在此共记录了风险传递关系 641 对,涉及 738 家企业. 路径的分布情况如图 4 所示.

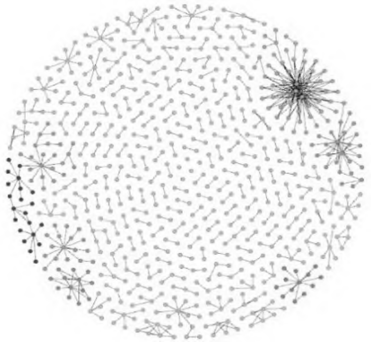


图 4　风险传递路径(Fruchterman 布局)

Fig. 4 Risk transferring path (Fruchterman layout)

从图 4 可以看出,大部分信用风险在 2 家 ~3 家

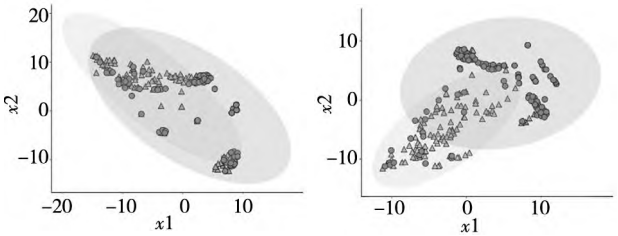
企业之间传递,有 7 个紧密风险关联的企业群形成了大型风险社团,这种社团内部一般都有较严重的投资异常情况. 这些包含较多企业的风险社团是金融监管机构监控系统性金融风险的重要切入点.

3.3 数值型工商信息及数据融合后的性能评估

使用 2.3 节所介绍的 MLP 方法对表 1 中所述的数值型工商字段进行了表征(隐层及输出维度均为 100). 为了显示 MLP 变换的效果,本文使用 t – 分布领域嵌入算法分别将原始向量和 MLP 变换后的向量降维到了 2 维,可视化效果如图 5 所示.

从图 5 可以看出,原始数据中用不同形状表示的两种信用类别的样本点重叠在一起,而经过 MLP 的变换后,原始的线性不可分状态变为了基本线性可分状态. 表 9 中每个分类器的前两行记录了原始数值型字段及经 MLP 变换后的数据的信用评价表现.

另外,为了验证融合文本型及关系型信息后的信用评价效果,表 9 中记录了逐步融合了文本型(企业名称 GRU 表征)和关系型数据(投资关系 GAT 表征)后的性能变动情况. 最后,由于数值和文本数据对应的企业中仅有部分被包含在了投资关系网络中,全集与子集的性能指标值没有可比性,因此对投资关系所涉及的企业子集数据做了单独验证.



(a)原数据 (b)基于 MLP 的变换

图 5 原始数值型字段及 MLP 变换后的分布图

Fig. 5 Distribution of the original numerical features and the ones transformed by MLP

表 9 数值型工商注册信息及融入文本与关系型数据后的信用评价结果对比

Table 9 Comparisonof credit evaluation results after integrating numerical business registration information with textual and relational data

信用评价模型	工商信息类型		TPR/%	Accuracy/%	F-Measure/%	AUC/%	MCC/%
LR	全部企业	原始数值型	57.09 ±0.32	66.11 ±0.30	62.22 ±0.30	74.46 ±0.13	32.37 ±0.63
		数值型(MLP 变换后)	76.52 ±0.20	82.96 ±0.24	81.44 ±0.23	89.09 ±0.14	66.28 ±0.52
		数值型 + 文本型	96.01 ±0.10	96.64 ±0.12	96.54 ±0.12	99.57 ±0.03	93.27 ±0.24
	投资关系涉及的子集	原始数值型	15.69 ±0.10	62.83 ±0.50	26.00 ±0.36	70.09 ±1.41	21.34 ±1.98
		数值型(MLP 变换后)	67.54 ±0.60	81.50 ±0.65	75.24 ±0.81	87.26 ±0.30	61.80 ±1.41
		数值型 + 关系型	82.27 ±1.09	88.72 ±0.35	85.85 ±0.42	92.82 ±0.33	76.72 ±0.76
		数值型 + 文本型	91.65 ±0.25	93.62 ±0.27	92.28 ±0.32	98.05 ±0.19	86.85 ±0.55
		数值型 + 文本型 + 关系型	93.56 ±0.22	95.08 ±0.14	94.06 ±0.16	98.44 ±0.10	89.88 ±0.30
L-SVM	全部企业	原始数值型	56.61 ±10.96	67.24 ±8.76	68.85 ±24.16	67.36 ±8.32	36.56 ±15.88
		数值型(MLP 变换后)	75.96 ±0.13	82.56 ±0.21	80.98 ±0.20	82.41 ±0.21	65.48 ±0.46
		数值型 + 文本型	96.01 ±0.08	96.70 ±0.15	96.61 ±0.15	96.69 ±0.15	93.41 ±0.31
	投资关系涉及的子集	原始数值型	46.05 ±31.16	58.71 ±10.8	43.64 ±15.52	56.89 ±6.97	18.78 ±18.14
		数值型(MLP 变换后)	65.57 ±0.99	81.74 ±0.19	74.92 ±0.46	79.42 ±0.29	62.59 ±0.36
		数值型 + 关系型	82.37 ±0.97	89.25 ±0.20	86.44 ±0.35	88.26 ±0.30	77.83 ±0.40
		数值型 + 文本型	92.30 ±0.47	94.79 ±0.08	93.65 ±0.12	94.43 ±0.14	89.27 ±0.17
		数值型 + 文本型 + 关系型	94.80 ±0.39	96.43 ±0.13	95.67 ±0.16	96.20 ±0.13	92.65 ±0.27
LDA	全部企业	原始数值型	52.71 ±0.84	65.33 ±0.39	59.77 ±0.60	73.58 ±0.13	31.11 ±0.80
		数值型(MLP 变换后)	75.41 ±0.16	81.92 ±0.19	80.30 ±0.18	88.48 ±0.10	64.19 ±0.40
		数值型 + 文本型	95.29 ±0.17	95.83 ±0.17	95.72 ±0.18	99.34 ±0.05	91.66 ±0.34
	投资关系涉及的子集	原始数值型	14.71 ±0.75	62.76 ±0.56	24.73 ±1.23	69.89 ±1.53	21.41 ±2.06
		数值型(MLP 变换后)	65.75 ±0.70	81.58 ±0.38	74.81 ±0.59	87.37 ±0.15	62.17 ±0.79
		数值型 + 关系型	77.38 ±1.05	87.63 ±0.34	83.89 ±0.53	93.17 ±0.29	74.67 ±0.69
		数值型 + 文本型	91.76 ±0.29	94.51 ±0.19	93.30 ±0.20	98.74 ±0.11	88.69 ±0.39
		数值型 + 文本型 + 关系型	92.84 ±0.52	95.43 ±0.08	94.41 ±0.13	99.12 ±0.07	90.59 ±0.17

注:“ ±”符号后为交叉检验折间标准差.

从表 9 中的指标可以看出,经过 MLP 变换后,线性信用评价模型的准确率和 AUC 大幅提

升，准确率的折间标准差降低，说明原数据中存在的较复杂的线性不可分模式，经过多层 MLP 变换后已被解耦，数据被呈现为了较简单的分布状态，评价鲁棒性也得以提升. 表 9 中 MLP 变换后最高的准确率为 82.96%，说明“钛金分”与注册资本、企业类型、续存时间、年检时间、币种及工商变换次数等信息也有较高的相关性.

另外，在数值型工商信息经 MLP 变换的基础上融入文本类信息（企业名称、注册地址、经营范围）后，综合信用评价准确率提升到了 96.7%，高于数值或文本类型单独进行信用评价的结果，显示异质数据的融合对信用评价具有重要价值.

最后，在投资关系所涉及的企业节点上，如果仅使用数值型的 MLP 表征进行信用评价，最高准确率为 81.74%；在数值型基础上融入投资关系信息后，最高准确率提升为了 89.25%，说明投资关系中含有部分数值型没有包含的重要信用信息；在数值基础上融入文本字段后，最高准确率提升到了 94.79%，说明文本对数值型字段具有重要的信息补充作用；继续在数值和文本基础上融入投资关系信息后，最高准确率为 96.43%，仅提升了 1.64%，说明投资关系、文本与数值信息之有较强的信息重叠. 但本研究认为这 1.64%的提升仍是重要的，因为其中包含了由文本或数值字段无法识别的、由企业投资关系转移的纯风险传递关系，这对研究风险传递路径具有重要作用.

3.4　实验结果讨论

下面对实验结果进行讨论和总结.

1）企业名称为何与企业信用高度关联

用企业名称可以较准确地预测“钛金分”信用类别，主要原因如下：a）企业名称是对企业性质、所属行业、经营范围等工商注册信息的高度提炼，工商注册信息中的大部分有效企业信用信息已被精炼地包含在了企业名称中；b）本研究所使用的“钛金分”信用标签，来源于某市高新区对有关企业进行孵化时所做的信用评估，其所关注的企业多为初创公司，积累的经营风险信息不足，并且这种评估场景较重视企业的发展潜力，而名称中的行业信息非常适合这种使用场景；c）从文本分类技术角度看，企业名称中均是精准的关键词，几乎不存在噪声问题，而经营范围和注册地址中包含了太多噪声文本信息. 当前的语句表征技术较难处理噪声词汇问题，这也在技术上导致使用企业名称时信用分类准确率较高.

2）投资关系网络对企业信用评价的意义

由 GAT 模型所表征的企业间投资关系与“钛金分”信用标签之间有较强的相关性，分类准确率为 84.87%，高于传统的数值型工商信息，但低于文本型工商信息. 虽然投资关系用于预测企业个体的信用时不是最准确的，但其独特优势在于其反映了由投资关系而引发的信用风险. 投资关系解释了企业间的风险传递路径，对防范系统性金融风险具有重要意义[21, 23]，也是金融监管机构应重点关注的风险源[34].

3）工商注册信息的优缺点分析

企业的信用体系建设应该从正反两方面展开，既要看到企业的发展潜力和影响力一面，也要看到企业的经营风险一面[17, 33]. 工商注册数据在某些特定类型的信用评估场景（如高新技术企业孵化）中具有重要意义，主要体现在评估企业的发展潜力这个正向企业信用评分方面. 同时，也应注意到工商注册信息具有实时性不强等缺点，即无法及时捕捉动态出现的企业信用风险. 在应用中，如果信用评价的使用场景为监控企业异常经营行为或者预测企业破产，企业名称等静态工商注册文本信息并不是反映此类问题的有效特征. 信用信息价值的高低，取决于具体的应用领域. 鉴于中小微企业信用数据高度稀缺的现状，工商注册数据仍可以作为中小微企业信用评价体系的有效补充[34]，但使用时要区分特定的场景和用途.

4）政策建议

基于上述实验结果所反映的问题，本研究提出如下政策建议：a）建立更完善的企业信用信息数据开放和共享机制. 任何算法和模型都无法进行无米之炊，当前很多信用评价的难题都源自信用信息的缺失. 除了工商注册信息，可靠的企业信用评价依赖于高质量的法律、财务、税务、银行流水等方面的数据源；b）建立面向异质、非常规金融数据的信用风险评价制度. 在信用体系建设中，应更加重视从多源的、非常规的数据中对企业的信用违约和风险等方面问题进行分析，并建立面向非常规金融数据的企业信用风险量化评价体系；c）建立企业信用评价体系的动态调整机制. 现代社会的经济、金融市场信息瞬息万变，再完善

的信用评价制度也难以应对持续变化的市场环境. 因此,必须建立一套全面的市场信息监督机制及与之相对应的动态信用评价调整机制. 最后,本研究认为信用评价仍是一个见仁见智的开放性、非监督或半监督学习问题,本研究所提出的政策建议仍需未来的实践检验.

4 结束语

本研究提出了面向工商注册信息中传统上未被有效利用的、非常规、非结构化数据——文本型与投资关系型数据的表征学习、信息融合,及信用评价结果的可解释方法. 基于某市高新区约 6.8 万家企业的工商注册信息和对应的"钛金分"信用评分的实验结果表明,工商注册信息中企业名称与投资关系蕴含了传统的数值型字段所没有的独特信用信息,是信用评价的重要信息源,具有较高的信用分类准确率. 信用评价的可解释性实验表明,企业名称中包含的特定词汇能通过其所反映的行业、科技、规模等属性关联到特定的信用类别,可以实现基于词汇对文本信用评价的结果进行解释;而通过推导企业投资关系中与"低信用"强相关的关系(高注意力系数),可以得出由纯投资关系所传导的信用风险,并解释风险在企业中的传递路径,为系统性信用风险的管控提供依据.

工商注册信息是企业信用评价中的重要信息源,但在信用评价的实时性等方面还存在许多局限. 近年来,随着大数据知识图谱技术的兴起,为企业间多样性的异质关系建模提供了一种全新的视角[25],笔者未来的工作将关注基于异质知识图谱的信用信息挖掘;另外,鉴于工商注册信息的局限性,笔者下一步将考虑融入资金流水、金融评论、财务报表等数据,完善数据层面的多样性,实现全景情境化下基于多模态数据的信用决策分析.

参 考 文 献:

[1]宋凌峰, 阳 浪. 经济下行、信用风险反馈和政府隐性救助[J]. 管理科学学报, 2016, 19(11): 103 – 113.
Song Lingfeng, Yang Lang. Economic downturn, credit risk feedback and government implicit bailout[J]. Journal of Management Sciences in China, 2016, 19(11): 103 – 113. (in Chinese)

[2]迟国泰, 李鸿禧, 潘明道. 基于违约鉴别能力组合赋权的小企业信用评级——基于小型工业企业样本数据的实证分析[J]. 管理科学学报, 2018, 165(3): 110 – 131.
Chi Guotai, Li Hongxi, Pan Mingdao. Small enterprises credit rating based on default identification ability of combination weighting: Based on an empirical analysis of small industrial enterprises[J]. Journal of Management Sciences in China, 2018, 165(3): 110 – 131. (in Chinese)

[3]马长峰, 陈志娟, 张顺明. 基于文本大数据分析的会计和金融研究综述[J]. 管理科学学报, 2020, 23(9): 19 – 30.
Ma Changfeng, Chen Zhijuan, Zhang Shunming. A survey on accounting and finance research based on textual big data analysis[J]. Journal of Management Sciences in China, 2020, 23(9): 19 – 30. (in Chinese)

[4]Wang Z, Jiang C Q, Zhao H M. Know where to invest: Platform risk evaluation in online lending[J]. Information Systems Research, 2022, 33(3): 765 – 783.

[5]肖斌卿, 杨 旸, 李心丹, 等. 基于模糊神经网络的小微企业信用评级研究[J]. 管理科学学报, 2016, 19(11): 114 – 126.
Xiao Binqing, Yang Yang, Li Xindan, et al. Research on the credit rating of small and micro enterprises based on fuzzy neural network[J]. Journal of Management Sciences in China, 2016, 19(11): 114 – 126. (in Chinese)

[6]Sebastián M, Juan P, Cristián B. Cost-based feature selection for support vector machines: An application in credit scoring[J]. European Journal of Operational Research, 2017, (261): 656 – 665.

[7]Verbraken T, Bravo C, Weber R, et al. Development and application of consumer credit scoring models using profit-based classification measures[J]. European Journal of Operational Research, 2014, (238): 505 – 513.

[8]黄宝凤, 李金玲, 徐 剑. 基于交易大数据的小微企业信用动态评价[J]. 中国统计, 2020, (4): 32 – 35.
Huang Baofeng, Li Jinling, Xu Jian. Dynamic credit evaluation of small and micro enterprises based on transaction big data[J]. China Statistics, 2020, (4): 32 – 35. (in Chinese)

[9]Deo A, Juneja S. Credit risk: Simple closed-form approximate maximum likelihood estimator[J]. Operations Research, 2021, 69(2): 361 – 379.

[10]Aduenko A A, Motrenko A P, Strijov V V. Object selection in credit scoring using covariance matrix of parameters estima-

tions[J]. Annals of Operations Research, 2018, (260): 3 – 21.
[11]Fernandes G B, Artes R. Spatial dependence in credit risk and its improvement in credit scoring[J]. European Journal of Operational Research, 2016, (249): 517 – 524.
[12]Hong L J, Juneja S, Luo J. Estimating sensitivities of portfolio credit risk using Monte Carlo[J]. Informs Journal on Computing, 2014, 26(4): 848 – 865.
[13]刘海飞, 柏　巍, 李冬昕, 等. 沪港通交易制度能提升中国股票市场稳定性吗？——基于复杂网络的视角[J]. 管理科学学报, 2018, 21(1): 97 – 110.
Liu Haifei, Bai Wei, Li Dongxin, et al. Does Shanghai-Hongkong stock connect trading mechanism improve the stability of Chinese stock market?: A complex network perspective[J]. Journal of Management Sciences in China, 2018, 21(1): 97 – 110. (in Chinese)
[14]姚加权, 冯　绪, 王赞钧, 等. 语调、情绪及市场影响：基于金融情绪词典[J]. 管理科学学报, 2021, 24(5): 26 – 46.
Yao Jiaquan, Feng Xu, Wang Zanjun, et al. Tone, sentiment and market impacts: The construction of Chinese sentiment dictionary in finance[J]. Journal of Management Sciences in China, 2021, 24(5): 26 – 46. (in Chinese)
[15]俞红海, 范思妤, 吴良钰, 等. 科创板注册制下的审核问询与 IPO 信息披露——基于 LDA 主题模型的文本分析[J]. 管理科学学报, 2022, 25(8): 45 – 62.
Yu Honghai, Fan Siyu, Wu Liangyu, et al. Registration system review inquiry and IPO information disclosure on star market: Textual analysis based on IDA topic model[J]. Journal of Management Sciences in China, 2022, 25(8): 45 – 62. (in Chinese)
[16]洪　亮, 欧阳晓凤. 金融股权知识大图的知识关联发现与风险分析[J]. 管理科学学报, 2022, 25(4): 44 – 66.
Hong Liang, Ouyang Xiaofeng. Knowledge association discovery and risk analysis of financial equity knowledge graph[J]. Journal of Management Sciences in China, 2022, 25(4): 44 – 66. (in Chinese)
[17]Gang K, Yong X, Yi P, et al. Bankruptcy prediction for SMEs using transactional data and two-stage multi-objective feature selection[J]. Decision Support Systems, 2021, (140): 113429.
[18]Wei Y, Yildirim P, Van den Bulte C, et al. Credit scoring with social network data[J]. Marketing Science, 2016, 35(2): 234 – 258.
[19]Freeman L C. Centrality in social networks conceptual clarification[J]. Social Networks, 2002, 1(3): 215 – 239.
[20]Battiston S, Farmer J D, Flache A, et al. Complexity theory and financial regulation[J]. Science, 2016, 35(6725): 818 – 819.
[21]周开国, 季苏楠, 杨海生. 系统性金融风险跨市场传染机制研究——基于金融协调监管视角[J]. 管理科学学报, 2021, 24(7): 1 – 20.
Zhou Kaiguo, Ji Sunan, Yang Haisheng. Cross-market contagion mechanism of systemic risk from the perspective of coordinated supervision[J]. Journal of Management Sciences in China, 2021, 24(7): 1 – 20. (in Chinese)
[22]陈道富, 刘新海. 我国担保圈大数据分析与风险管控[J]. 中国农村金融, 2016, (12): 25 – 27.
Chen Daofu, Liu Xinhai. Big data analysis and risk control in China's guarantee circle[J]. China Rural Finance, 2016, (12): 25 – 27. (in Chinese)
[23]Van Vlasselaer V, EliassiRad T, Akoglu L, et al. GOTCHA! Network-based fraud detection for social security fraud[J]. Management Science, 2016, 63(9): 3090 – 3110.
[24]Liu Y Y, Slotine J J, Barabási A L. Control lability of complex networks[J]. Nature, 2011, 473(7346): 167 – 173.
[25]Chuan S, Philip S Y. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, (29): 17 – 37.
[26]董路安, 叶　鑫. 基于改进教学式方法的可解释信用风险评价模型构建[J]. 中国管理科学, 2020, 28(9): 45 – 53.
Dong Luan, Ye Xin. Interpretable credit risk assessment modeling based on improved pedagogical method[J]. Chinese Journal of Management Science, 2020, 28(9): 45 – 53. (in Chinese)
[27]Kelley S, Ovchinnikov A, Hardoon D R, et al. Anti-discrimination laws, artificial intelligence, and gender bias: A case study in nonmortgage fintech lending[J]. Manufacturing & Service Operations Management, 2022, 24(6): 3039 – 3059.
[28]梅子行, 毛鑫宇. 智能风控：Python 金融风险管理与评分卡建模[M]. 北京：机械工业出版社, 2020.
Mei Zixing, Mao Xinyu. Intelligent Risk Control: Python Financial Risk Management and Scorecard Modeling[M]. Beijing: China Machine Press, 2020. (in Chinese)

[29]Senoner J, Netland T H, Feuerriegel S. Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing[J]. Management Science, 2022, 68(8): 5704 – 5723.

[30]Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]. //Proceeding of the Twenty-Eight Annual Conference on Neural Information Processing Systems. Canada Montréal, 2014.

[31]Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[C]. //Proceeding of the Sixth International Conference on Learning Representation. Canada Vancouver, 2018.

[32]徐 驰, 汪冬华, 庆 楠. 风险相关性下的银行非预期操作风险集成度量——基于动力学模型视角[J]. 管理科学学报, 2018, 21(5): 58 – 69.
Xu Chi, Wang Donghua, Qing Nan. Unexpected operational risk aggregation considering risk correlation: A dynamical modeling perspective[J]. Journal of Management Sciences in China, 2018, 21(5): 58 – 69. (in Chinese)

[33]吴 刚, 刘晓惠, 冉淑青. 西部地区市场主体成长力评价及路径优化研究——基于工商登记注册信息数据的分析[J]. 西北工业大学学报(社会科学版), 2018, (1): 98 – 104.
Wu Gang, Liu Xiaohui, Ran Shuqing. The research on the growth capacity evaluation and path optimization of the market subject in the western region: Analysis of information data based on industrial and commercial registration[J]. Journal of Northwestern Polytechnical University (Social Sciences), 2018, (1): 98 – 104.

Credit evaluation of medium, small, and micro enterprises based on business registration information

LI Tie[1], *KOU Gang*[2], *PENG Yi*[1]*

1. School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, China

Abstract: The quantitative credit evaluation of enterprises is a cornerstone of establishing a more efficient market economy supervision system in China. However, the credit evaluation of medium, small, and micro enterprises has been challenging due to the lack of public business and financial data. Based on the observation that there are a large number of open and unstructured business registration data of medium, small, and micro enterprises on the Internet, this paper proposes to mine unconventional and unstructured data such as enterprise names, investment relationships between enterprises, and business scope to improve the credit evaluation results. Specifically, this paper proposes two data representation methods. The first uses a Gated Recurrent Neural Network to extract sequential information from the business registration text and transform the text into numerical data. The second uses a Graph Attention Network to encode the graph structure formed from the investment relationships into a numerical space. As a result, the heterogeneous information can be easily fused by merging the numerical vectors. Since the interpretability of credit evaluation models is crucial in financial applications, this paper further proposes interpretable solutions for the text-mining-based credit evaluation model and for identifying the credit risk transmission path. The experimental results based on 68 504 enterprises revealed that both enterprise names and investment relationships contain credit information that cannot be identified in traditional numerical data. The results showed that business registration information can be used as a useful supplement to the current enterprise credit evaluation system, which is valuable in dealing with the scarcity of credit information for medium, small, and micro enterprises.

Key words: enterprise credit evaluation; business registration information; textual data mining; graph data mining; heterogeneous data fusion