



复杂担保网络中传染路径的风险评估

程大伟^{1,2}, 牛志彬^{3,2}, 刘新海⁴, 张丽清^{1,2*}

1. 上海交通大学计算机科学与工程系, 上海 200240
2. 上海交通大学人工智能教育部重点实验室, 上海 200240
3. 天津大学智能与计算学部, 天津 300354
4. 北京大学金融智能研究中心, 北京 100871

* 通信作者. E-mail: zhang-lq@cs.sjtu.edu.cn

收稿日期: 2020-02-16; 修回日期: 2020-04-30; 接受日期: 2020-06-15; 网络出版日期: 2021-06-23

国家重点研发计划 (批准号: 2018AAA0100704) 和中国博士后科学基金 (批准号: 2019M651499) 资助项目

摘要 中小企业贷款在促进技术创新、推动经济发展、改善民生和增加就业等方面有着重要的作用. 为了满足商业银行的贷款评估标准, 很多中小企业选择互相提供担保以获得授信, 形成了结构复杂的担保网络. 当借款方的贷款违约时, 风险则沿着担保方向在网络中层层传播, 由此造成的潜在系统性风险给国家的金融安全和监管带来了严峻的挑战. 因此, 迫切需要发展相应的方法从系统角度对复杂金融担保网络中的传染路径进行风险评估和预测. 本文提出了一种基于深度学习的风险评估模型, 该方法应用图神经网络和注意力机制直接从网络化的贷款行为数据中学习风险特征, 无需依赖于金融领域专业知识的人工特征工程. 实验结果表明, 本文设计的方法在多数评价指标上均优于现有的 7 个对比的基准模型. 在传染路径风险评估任务中, 比基准方法在精确率和召回率的调和平均数 (F1-score) 方面平均提升了 2%~15%. 在新路径风险评估任务中, 比最好的基准方法平均提升了 3.5%. 结果表明了本文设计方法在传染路径风险评估中的有效性, 可为监管部门和金融机构对担保网络进行系统性风险评估提供方法理论基础.

关键词 风险评估, 传染路径, 担保网络, 图神经网络, 注意力机制

1 前言

在我国, 担保贷款是一种非常普遍的经济现象. 由于目前商业银行信贷评估标准主要是针对中大型股份制企业设计的, 中小企业很难满足银行的授信准入要求^[1]. 面对“融资难”、“融资贵”的资金获取难题, 众多民营中小企业选择互相担保来获取银行贷款. 担保是一种法律行为, 担保方对借款方的贷款提供代偿责任, 即如果借款方无力偿还贷款, 则其担保方具有法律责任进行偿还^[2]. 所以在担保

引用格式: 程大伟, 牛志彬, 刘新海, 等. 复杂担保网络中传染路径的风险评估. 中国科学: 信息科学, 2021, 51: 1068–1083, doi: 10.1360/SSI-2020-0028
Cheng D W, Niu Z B, Liu X H, et al. Risk assessment for contagion path in complex loan network (in Chinese). Sci Sin Inform, 2021, 51: 1068–1083, doi: 10.1360/SSI-2020-0028

贷款中,个体企业的违约会在网络中层层扩散传播,进而导致系统性的金融风险.相比于现有针对个体企业的风险评估,从传染路径角度进行分析可以识别高风险担保群体,从系统性角度鉴别个体风险在传染路径中的重要性,预警关联风险的集中爆发,可以为行业性、区域性等担保传染关联风险的监控提供依据.

随着中小企业的发展,商业银行发放的担保贷款也同步逐年递增.中国人民银行和前瞻产业研究院的研究数据表明^[3],截至 2018 年,担保贷款在保余额达到了 3.6 万亿元,年增长速度为 6.8%.与此同时,担保贷款的代偿率(指借款人违约,由担保方进行代偿的贷款金额比例)也持续走高,从 2011 年的 0.50% 增长至 2017 年的 2.78%,年复合增长率达到了 26%,显著高于贷款发放金额的年均增长率和商业银行的平均信贷不良率,由此风险传染所引发的系统性风险也逐渐开始显现^[4].但是,商业银行目前针对担保贷款的评估方法仍然主要依赖于借款方的财务信息和信用历史,使用经典的机器学习模型评估客户偿还贷款的可能性^[5].其主要存在 3 个方面的缺点:(1) 这些方法主要针对个体企业的自身风险进行评估,GAT^[6],HGAR^[7]等图神经网络方法虽然可以学习网络的全局信息,但是其学习任务的目标仍然是针对个体风险的评估,忽略了违约在企业间传播可能导致的系统性风险.因此无法有效对复杂担保网络中的传染路径风险进行精确评估.(2) 现有方法需要大量的人工特征工程,这些手动特征的抽取依赖于深度的金融领域专业知识.而且其中不少特征是从历史数据中聚合而来的,因此计算也比较耗时.(3) 如果风险传染路径是由新的信贷客户担保贷款生成的,其并没有相应的信贷历史记录,传统方法很难进行有效的评估.

因此,为了克服上述挑战,本文提出了一种基于深度学习的担保网络传染路径风险评估方法(CR-DAN).首先,针对原始信贷记录,我们高效地构建出对应的担保网络和风险传播路径.然后基于设计的图神经网络特征学习层直接学习网络化的贷款行为数据,生成高阶隐含表征.因此无需进行复杂的特征工程.随后,我们开发注意力机制模型对网络中相邻传播路径的影响进行动态学习,即学习网络中其他路径对目标对象风险评估的重要性,进而可以缓解评估对象没有信贷历史记录情况下的风险评估难题.最后,通过合作金融机构的真实数据集实验验证了本文设计方法的有效性和可靠性.综上,本文的主要贡献表述如下.

(1) 在当前复杂的金融环境下,针对系统性风险监控的紧迫性要求和网络化担保贷款风险评估的难题,我们提出了从风险传播角度对信贷担保网络的风险进行探索评估,设计了一种新型的基于深度学习的担保网络传染路径风险评估模型.

(2) 根据担保网络特性,设计了基于图神经网络的输入特征表示学习层来避免复杂的特征工程,同时开发了风险注意力机制模型,用来学习网络中的其他传播路径对评估目标影响的重要性.并且通过实验验证了该模块在整体评估模型中的有效性和可解释性.

(3) 设计了上述模块的总体网络模型,该模型融合了担保网络的结构特征和传染过程动态特性进行联合优化学习.实验结果表明本文设计的方法在传染路径风险评估中具有更高的准确度和实用性.

本文剩余部分内容安排如下:第 2 节介绍担保网络风险评估的相关研究工作,包括金融网络风险评估和网络特征学习方法.第 3 节详细介绍本文提出的担保网络传染路径风险评估方法及各步骤细节.第 4 节介绍实验设计、实验结果及分析、实验讨论等.最后总结全文并展望下一步研究工作.

2 相关工作

本文提出了基于图神经网络的深度学习方法对担保网络传染路径进行信贷风险评估,接下来简要回顾担保网络特征学习和金融网络风险评估的相关研究进展.

2.1 担保网络特征学习

通常来说, 对担保网络的风险评估过程可以看成在网络中进行特征学习并应用于后续分类或者回归任务的过程. 如前文所述, 由于传统的忽略网络信息仅进行个体风险评估的方法存在明显缺陷, 因此本小节重点对担保网络的特征学习进行回顾.

担保网络作为一种典型的网络结构数据, 传统方法中使用网络的邻接向量作为节点特征, 但是邻接向量存在着高维稀疏的缺点, 因此近些年网络表示学习方法的提出可以将网络信息转化为低维稠密的实数向量, 并将其用作机器学习算法的输入^[8]. 这些方法包括: DeepWalk^[9], LINE^[10], Node2vec^[11], SDNE^[12]等, 但是上述方法往往仅能表示网络的结构特征, 因此需要结合外部信息的网络表示学习, 如 STNE^[13]则引入节点属性信息学习网络的低维表征; Wang等^[14]基于随机游走的策略提出了异构网络表示学习的方法; 这些方法有效地从多个角度对网络信息进行特征表示. 在端到端的网络特征学习方面, 研究者们提出了图神经网络来处理图数据^[15]. 例如: 图注意力网络^[6]可以用注意机制确定节点邻域的权重; 图时空网络可以同时捕捉时空图的时空相关性, 来预测未来时空图中的节点值或标签, 在预测不同区域的网约车用车量^[16]、网络节点类别预测^[17]等方面得到了很好的应用. 在担保网络中, BHONEM^[18]基于担保网络的节点角色的二元性和节点距离的高阶性提出了二元高阶担保网络表示学习模型用于单个节点的风险特征提取.

可以看到目前针对担保网络特征学习的方法主要集中在个体节点的分析 and 预测, 很少从风险传播角度和系统性风险角度对网络化的贷款进行建模. 而后者往往对国民经济和金融机构带来更大的挑战和不确定性. 本文尝试从风险传染的角度建立违约传播路径建模方法, 可以为监管和金融机构的隔离机制提供基础, 通过对高风险的传染路径进行评估和预测、阻断关键风险传播者进而控制由于链式传播导致的系统性风险.

2.2 金融网络和风险评估

对系统性风险的评估及其导致的金融危机的预警一直是金融机构和政府监管部门的重大关切, 目前已经得到了广泛的研究^[19]. 网络(图)作为实体和关系的一种直接的表示方式, 在金融系统的分析中被广泛采用, 例如银行间的资金流向网络、企业间股权关系网络和本文中的信贷担保网络. 研究表明网络结构特征在系统性风险监控和预警方面起着重要的作用^[20]. 2008年金融危机之后, 金融系统中的复杂网络分析和学习得到了更多的关注^[21]. 由雷曼兄弟银行倒闭所引发的与其相连的金融网络的系统性危机与2003年爆发的非典型肺炎(SARS)的传播过程显示出一定的相似性, 都是由少数个体节点损害并且击中了广泛相连的网络系统, 进而导致严重性的后果^[22]. 随后《自然物理》刊登了专刊汇报使用复杂网络理论来理解金融系统中的基础性问题的相关研究^[23]. 例如, 银行间隔夜拆借的动态网络可以对流动性风险作出有效警示. 在经济危机中当金融机构体量太大而不能倒闭时(too big to fail), 则必须动用国家力量进行救援, 如美国联邦储备系统(The Federal Reserve System)注资其最大的保险公司美国国际集团(American International Group, AIG). 如今在网络化的金融体系中, 参与实体在网络中的拓扑结构信息已经成为了其规模大小之外的新的重要指标^[20].

尽管现有工作已经开始探索使用网络科学的手段来理解金融系统中的风险问题, 但是对担保网络中的由于风险传染导致系统性风险进行有效评估尚未得到充分研究. Meng等^[24]论证了企业所处节点的Kshell散度和企业在担保网络中的违约率存在相关性. Cheng等^[7]进一步探索了使用网络表示学习方法在个体企业风险评估中的作用, 表明了网络特征的重要性. 现有针对担保网络的信贷风险评估主要针对网络中的个体节点进行预测, 本文从风险传播的传染路径的角度入手, 探索对担保网络的

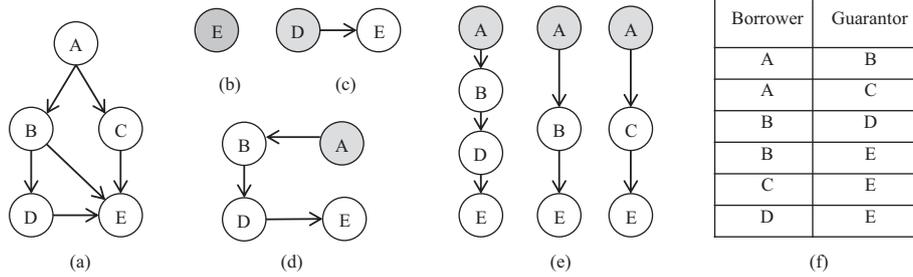


图 1 典型的担保网络及其传播过程说明. (a) 一个典型的担保网络; (b) 个体节点; (c) 传播风险; (d) 一条传染路径; (e) 所有传染路径; (f) 担保关系表

Figure 1 Illustration of a typical guarantee network and the diffusion process. (a) A typical guarantee network; (b) individual risk; (c) contagious risk; (d) a typical contagion path; (e) all contagion paths; (f) guarantee relationship table

系统性风险进行评估和分析, 为风险监控提供全局的数据和方法支撑.

3 本文方法

3.1 问题定义

定义1 (担保网络) 担保网络是一个有向图 $G = (V, E)$, 其中 G 代表网络, V 是其中的节点集合, 每个节点 $v \in V$ 代表一个中小企业. E 是边的集合, 每个边 $e \in E$ 代表一个担保关系, 从借款方指向担保方.

定义2 (自身风险) 给定担保网络 $G = (V, E)$, 任一顶点 v_i 的自身风险是指由于自身经营情况和财务状况导致其贷款违约的概率.

定义3 (传播风险) 在担保网络 $G = (V, E)$ 中, 如果顶点 $v_i \in V$ 给其他顶点 v_j 提供担保, 则当企业 v_j 贷款违约时, v_i 具有连带偿还责任, 我们将由于 v_j 违约的连带责任导致 v_i 违约的概率称为传播风险. 一个典型的担保网络及其风险传播过程如图 1 所示. 图 1(a) 展示了由 5 个顶点构成的担保网络. 图 1(b) 展示了传统的风险评估针对单个节点 E 时, 只能对其自身风险进行衡量, 无法全局计算出其他节点对 E 的影响. 图 1(c) 展示了如果 D 违约, 则其风险会沿着担保关系 (边) 的方向传染至节点 E 的过程.

定义4 (传染路径) 在担保网络 $G = (V, E)$ 中, 如果顶点 $v_i \in V$ 和顶点 v_j 是可达的, 则 v_i 和 v_j 之间的所有可达路径, 我们称之为传染路径 $c_{i,j} = \{(v, e) | v \in \text{Path}(v_i, v_j) \cup e \in \text{Path}(v_i, v_j)\}$.

如图 1(d) 展示了一个典型的传染路径. E 的违约风险依赖于节点 D 的概率, D 又依赖于 B, 依次层层依赖, 形成了 ABDE 的一条风险传染路径. 该担保网络的所有传染路径如图 1(e) 展示, 我们以只有出度的节点为起始节点, 以只有入度的节点为终止节点, 利用图论中的所有可达路径方法^[25] 构建出该担保网络中的所有传播路径: 分别为 ABDE, ABE, ACE. 图 1(f) 展示了每条边的风险传播对应的传播起始和终止节点情况. 针对大规模网络的情况, 我们采用分布式图计算框架 GraphX¹⁾ 来实现担保网络传染路径的快速构建.

定义5 (相邻路径) 如果两个传染路径拥有一个或者多个共同顶点, 我们定义其为相邻路径.

1) <https://spark.apache.org/graphx/>.

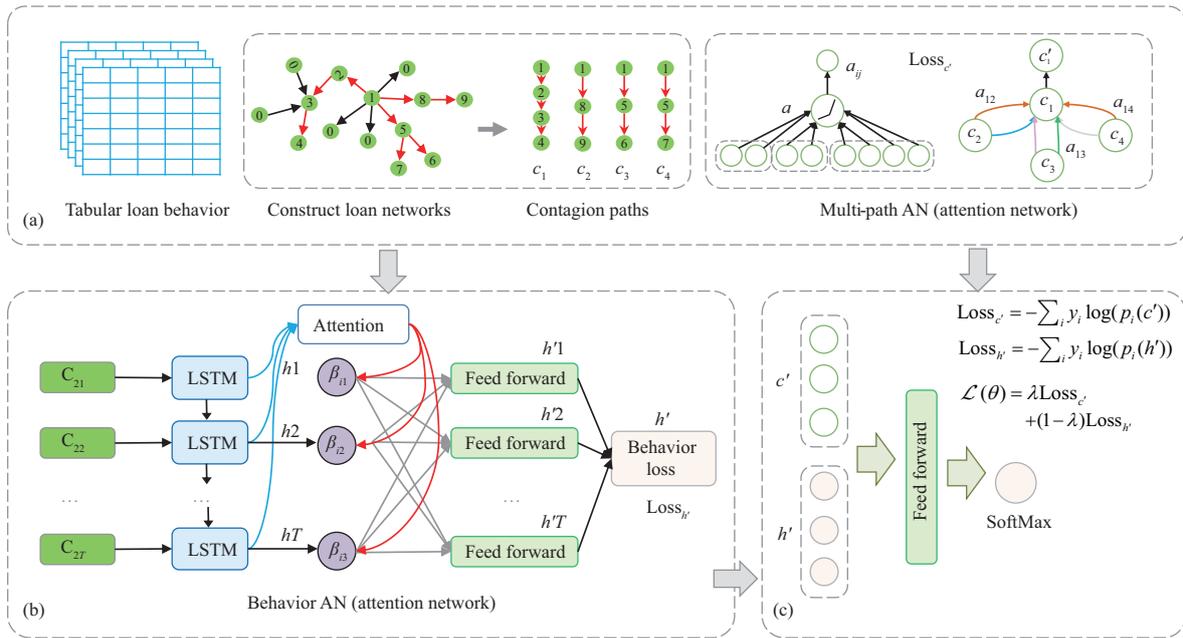


图 2 (网络版彩图) 本文设计的担保网络传染路径风险评估方法 (CRDAN) 模型框架图. (a) 图神经网络学习层; (b) 行为注意力网络; (c) 风险评估网络

Figure 2 (Color online) The model framework of this paper proposed contagion path risk rating model based on deep attention neural network (CRDAN). (a) Multi-path attention network; (b) behavior attention network; (c) risk rating network

定义6 (问题定义) 综上, 本文研究的问题定义: 输入: 担保网络 $G = (V, E)$, 以及网络中节点和边的属性; 输出: 网络中每一个传染路径 $C = \{c_1, c_2, \dots, c_{|C|}\}$ 的风险评级, $|C|$ 代表传染路径的数量; 目标: 风险评估方法具有较高的精度和计算效率.

3.2 模型框架

图 2 展示了本文设计的担保网络传染路径风险评估方法 (CRDAN) 模型框架图. 总体来说, 模型分为 3 个主要部分: (1) 图神经网络特征学习层, 该模块的输入是原始信贷表结构的记录数据, 输出为嵌入网络结构信息的高阶的路径表征. 详细来说, 首先将原始的表结构的数据记录构建为网络化的担保图, 网络中的属性由原始记录提供. 随后, 在构建好的担保网络中抽取传染路径和相邻路径, 抽取方法如定义 4 和 5 所述. 由于相邻路径之间存在公共节点, 所以在违约传播的过程中, 风险会跨路径传染. 即路径的风险特性不仅由其内部的节点决定, 也取决于其相邻路径的风险特性, 因此, 随后在图神经网络中我们使用注意力机制动态学习不同的相邻路径对目标传染路径的重要性. 路径的特征为其内部节点和边的属性的直接聚合. (2) 行为注意力网络, 其输入为构建的传染路径的时序特征, 随后应用递归神经网络和时序注意力机制学习其在时序上的特征表示. (3) 风险评估网络, 将图神经网络特征学习层和行为注意力网络的输出作为本模块的输入, 通过一个浅层神经网络估计该路径在预测的时间窗口中产生风险的概率. 在本文设计模型的 3 个子模块中, 输入来源于原始信贷记录, 因此不需要进行复杂的特征工程, 直接通过图神经网络学习结构特征和循环神经网络学习历史特征, 体现了深度神经网络在特征自动学习方面的优势.

3.3 图神经网络特征学习层

图神经网络特征学习层的输入来源于原始的信贷表结构的数据记录,我们首先由企业基本信息和担保关系记录构建初始担保网络 $G = (V, E)$. 其中节点代表企业 $V = \{v_1, v_2, \dots, v_{|V|}\}$, 边代表担保关系 $E = \{e_1, e_2, \dots, e_{|E|}\}$, $|V|$ 和 $|E|$ 分别指企业和担保关系的个数. 随后, 为节点和边嵌入属性, 节点属性包括企业注册资本、员工数量、企业规模、财务状态等基本信息. 边的属性包括担保开始时间、担保结束时间、担保金额、贷款金额、贷款利息等. 由于每条边都有各自的生效时间, 即从开始时间到结束时间, 因此我们在训练和预测阶段对边进行动态生成和更新. 最后对本文研究的对象传染路径进行构建. 路径 $C = \{c_1, c_2, \dots, c_{|C|}\}$ 随着边的更新同步变化. 由于原始网络是基于历史数据离线构建的, 所以动态更新边和路径时, 仅仅对有变化的子图进行更新, 时效性得到了很大的提升.

如定义 5 所述, 当两个传染路径存在一个或者多个共同节点时, 我们将其视为相邻路径. 由于传染路径的风险不仅由其内部节点决定, 也同样依赖于网络中的相邻路径的状态. 因此我们设计了多路径注意力网络机制, 来学习每个路径在担保网络中的高阶结构表征. 设给定传染路径 $C = \{c_1, c_2, \dots, c_{|C|}\}$, 其特征由其内部节点和边的属性聚合得到, $c_i \in \mathbb{R}^F$, 其中 F 是路径特征的维数, $|C|$ 是路径的总数. 图注意力网络机制学习网络中所有路径的高阶表征, 表示为 $C' = \{c'_1, c'_2, \dots, c'_{|C|}\}$, 其中 $c'_i \in \mathbb{R}^{F'}$. 详细来说, 对于任一传染路径 c_i , 我们首先抽取出其所有的相邻路径 $c_j \in \mathcal{N}_i$, \mathcal{N}_i 代表 c_i 所有相邻路径的集合. 则 c_i 和 c_j 之间的注意力权重表示为

$$\alpha_{i,j} = \frac{\exp(\sigma(a_c^T [W_c c_i \parallel W_c c_j]))}{\sum_{k \in \mathcal{N}_i} \exp(\sigma(a_c^T [W_c c_i \parallel W_c c_k]))}, \quad (1)$$

其中 $W_c \in \mathbb{R}^{F' \times F}$ 是注意力网络中全连接层的权重矩阵. \parallel 代表向量拼接, a_c 是注意力网络中需要学习的参数之一, \cdot^T 代表转置操作, σ 代表 sigmoid 函数, $\alpha_{i,j}$ 代表 c_i 和 c_j 之间的注意力权重.

为了注意力机制训练过程的稳定性, 我们使用多头注意力机制^[26](multi-head attention) 来独立学习 K 个如公式 (1) 所述的注意力实例, 然后对学习出的结果特征进行算术平均, 公式表示为

$$c'_i = \sigma \left(\frac{1}{K} \sum_{k=1}^K \sum_{j \in \mathcal{N}_i} \alpha_{i,j}^k (W_c^k c_j) \right), \quad (2)$$

其中 K 代表多头注意力机制的多头数量, $\alpha_{i,j}^k$ 和 W_c^k 为第 k 个实例的注意力权重和网络全连接层中的权重矩阵. 最后学习出的高阶表征 $C' = \{c'_1, c'_2, \dots, c'_{|C|}\}$ 作为后续风险评估网络的输入.

3.4 行为注意力网络

行为注意力网络的输入为在传染路径构建阶段构建好的路径序列 $C = \{c_1, c_2, \dots, c_{|C|}\}$, 同时输入的还有每一个路径的历史行为特征, 表示为 $c_i = \{c_{i1}, c_{i2}, \dots, c_{iT}\}$, 其中 T 为时间序列的长度. 行为注意力网络基于路径的基本属性学习其历史特征, 可以避免复杂的基于历史聚合操作的特征工程过程. 由于传染路径的属性分为点的属性和边的属性, 其历史特征的分布和变化频率均不同, 因此我们引入一种改进的循环神经网络(门控长短时记忆网络)^[27] 来对路径中点 v 和边 e 的历史特征进行分别学习. 为了公式表述的简便, 我们对 c_i 中的第 t 时刻点和边的属性特征 v_{it} 和 e_{it} 简写为 v_t 和 e_t , 得到循环神经网络的公式为

$$i_t = \sigma(W_{ei}e_t + W_{vi}v_t + W_{hi}h_{t-1} + b_i), \quad (3)$$

$$f_t = \sigma(W_{ef}e_t + W_{vf}v_t + W_{hf}h_{t-1} + b_f), \quad (4)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_{gc}g_t + W_{hc}f_{t-1} + b_c), \quad (5)$$

$$o_t = \sigma(W_{eo}e_t + W_{vo}v_t + W_{ho}h_{t-1} + b_o), \quad (6)$$

$$h_t = o_t \cdot \tanh(c_t), \quad (7)$$

其中 i_t , f_t 和 o_t 分别代表在 t 时刻的输入门、遗忘门和输出门. 为了保持和长短时记忆网络的公式的一致性, 式 (5) 和 (7) 中的 c_t 特指门控长短时记忆网络中的中间隐含激活单元.

为了进一步获得不同时序中的传染路径特征的重要性, 我们在此基础上设计了时序自注意力层来动态学习不同时刻特征的重要性. 时序自注意力层的输入是门控长短时记忆网络的输出, 即 $h = \{h_1, h_2, \dots, h_T\}$, 输出为该路径在时序上的隐含表征, 表示为 $h' = \{h'_1, h'_2, \dots, h'_T\}$, 时序自注意力的公式如下:

$$h'_t = \text{NN}_t \left(\left\|_{j=1}^T \beta_{t,j} W_t h_j \right\| \right), \quad \beta_{t,j} = \frac{\exp(\sigma(W_n h_t, W_n h_j))}{\sum_{i=1}^T \exp(\sigma(W_n h_t, W_n h_i))}, \quad (8)$$

其中 $\beta_{t,j}$ 是时序自注意力机制学习出的在时刻 t 和 j 之间的权重. NN_t 表示为一个浅层前馈神经网络, 由两个非线性激活单元层 (ReLU) 构成. W_n 和 W_t 为该浅层前馈神经网络的权重矩阵.

3.5 风险评估网络和模型优化

上述图神经网络特征学习层和行为注意力网络得到的对于担保网络风险传染路径在网络结构和时序上的高阶表征是风险评估网络的输入, 该网络输出路径的不同的风险等级 $r = \{r_1, r_2, \dots, r_l\}$, 其中 l 代表风险的等级数. 因此, 我们可以把传染路径风险评估视为一个多分类任务, r_1 代表风险等级最低, r_l 代表风险等级最高. 风险评估网络的代价函数为

$$p_i(x) = \frac{\exp(\text{NN}_r(x_i : W_r))}{\sum_{j=1}^l \exp(\text{NN}_r(x_j : W_r))}, \quad x \in [c', h'],$$

$$\text{Loss}_{c'} = - \sum_i y_i \log(p_i(c')), \quad i = \{1, \dots, N\}, \quad (9)$$

$$\text{Loss}_{h'} = - \sum_i y_i \log(p_i(h')), \quad i = \{1, \dots, N\},$$

其中 x 是传染路径的输入特征, 由 c' 或者 h' 构成. y 代表路径的实际观测样本值, N 为样本的数量. $\text{NN}_r(x_i; W_r)$ 是风险评估中的预测网络, 由一个非线性激活单元层 (ReLU) 和一个 Sigmoid 网络层构成, 将输入 x_r 映射到一个 $[0, 1]$ 的实数数值中, 代表该路径在 i 个风险类别上的概率值. W_r 是需要学习的权重.

综上, 总体风险评估网络的完整代价函数表示为

$$\mathcal{L}(\theta) = \lambda \text{Loss}_{c'} + (1 - \lambda) \text{Loss}_{h'}, \quad (10)$$

其中 θ 代表所有待学习的参数, $\lambda \in (0, 1)$ 是惩罚系数的超参数, 用来调整整个网络中图神经网络特征学习层和行为注意力层在损失函数中的重要性.

本文设计的端到端模型 (CRDAN) 可以直接使用标准的梯度下降法进行优化学习. 在实现时, 我们使用 Tensorflow 和 Keras 进行模型开发, 使用 Adam^[28] 作为模型的优化求解器. 其初始的学习率设置为 0.001, 学习的块大小 (batch size) 设置为 256.

Customer profile	Loan account info	Repayment status	Guarantee profile
Date Loan card ID Customer ID Sector Capital registered	Date Loan card ID Customer ID Loan contract ID Guarantee type	Date Loan card ID Customer ID Repayment amount Repayment interest	Guarantee contract ID Guarantee ID Amount Key: guarantee ID
Key: customer ID	Key: loan card ID	Key: loan card ID	Customer credit
Guarantee contract	Loan contract	Default status	Customer ID Rating Key: customer ID
Date Loan contract ID Guarantee contract ID Guarantee ID Start date End date	Date Customer ID Loan contract ID Guarantee type Start date End date	Date Loan card ID Customer ID Default amount Default interest	Guarantee relationship
Guarantee contract ID	Key: loan card ID	Key: loan Card ID	Start time End time Guarantee contract ID Guarantee contract ID

图 3 传染路径风险评估中所使用的担保网络数据集描述

Figure 3 The description of networked-guarantee loan datasets for contagion path risk rating

4 实验与分析

本节首先介绍实验所使用的担保网络数据集、对比方法、参数设置,接着介绍算法性能衡量指标,最后展示实验结果并对结果进行分析和讨论.

4.1 实验设置

4.1.1 数据集描述

本文实验使用的数据集来源于项目合作金融机构的从 2013 年 1 月 1 日至 2016 年 12 月 31 日的对公客户信贷记录,详细原始数据表结构如图 3 所示,其中包含了 112872 个顶点(中小企业),以及它们之间形成的 124957 个边(担保关系).在数据预处理阶段,我们筛选出内部节点数量大于 6 的子网络,一共包括 28360 个顶点和 32961 个边,其中顶点的属性特征 11 个,边的属性特征 6 个.经过预处理后,我们得到了实验中所使用的 33473 个担保网络传染路径.在实验过程中,我们将第一年的数据作为离线构建网络和模型训练的基础数据,随后在接下来的三年中,按照时间关系,依次预测新产生的记录对应的风险等级.经过标注后的新记录被用来更新模型,在本文实验中,我们的模型每天离线训练更新,用来预测接下来一天的传染路径风险.数据集的标注采用我国金融业对贷款质量评价的标准五级分类制度,即正常(normal)、关注(low)、次级(medium)、可疑(high)和损失(loss).基于传染路径每天真实违约信息,计算违约金额占路径中全部贷款金额的比例,根据商业银行对贷款系统性风险的容忍度,设置每个类别的违约比例为:正常(0%)、关注(0%~5%)、次级(5%~15%)、可疑(15%~50%)、损失(>50%),其中比例越高,说明该路径的风险等级越高.商业银行根据传染路径的风险状态,可以结合自身经营指标来判断行内担保贷款的系统性风险程度.

4.1.2 对比方法和参数设置

为了验证本文提出方法的有效性,我们选取了当前在网络行为特征学习和风险评估上比较优秀的

7 个基准方法进行对比实验, 它们分别是: 逻辑斯蒂回归 (Logistic regression, LR)、梯度提升决策树 (gradient boosting decision tree, GBDT)、深度神经网络 (deep neural network, DNN) [29]、长短时记忆神经网络 (long short term memory, LSTM)、快速图卷积神经网络 (simplifying graph convolutional networks, SGC) [30]、图注意力机制模型 (graph attention network, GAT) [6] 和高阶图注意力网络 (high-order graph attention representation, HGAR) [7]. 其中对于个体风险评估的方法, 我们使用算术平均聚合将其扩展至路径的风险评估任务中. 在逻辑斯蒂回归中, 使用 one-vs-the-rest 进行多分类扩展, 使用 L2 范数 $\lambda = 1$ 的参数惩罚项, 最大迭代次数为 1000. 针对梯度提升决策树设置树的最大深度为 3, 学习率为 0.1, 最大提升学习次数 (boosting stages) 为 500. 深度神经网络一共包含 5 层, 隐含层分别设置为 ReLU(256), ReLU(256), ReLU(128), 输出层为 softmax, 使用 Adam [28] 作为网络学习的优化器, 块大小设置为 128, 最大学习 epoch 设置为 1000. 长短时记忆循环神经网络的隐含层神经元设置为 128, 学习率为 0.001, 块大小设置为 128. 上述 4 种经典方法我们人工构建了约 500 个特征, 充分使用了网络的节点和边的信息及其属性特征. 对于 SGC, GAT 和 HGAR 3 种图神经网络的基准算法, 我们直接将网络及其属性作为输入, 参数选择方面采用对应文献中的推荐设置. 本文方法有两个子模块的基准 CRDAN-noPA 和 CRDAN-noBA, 分别代表在模型中将多路径注意力模块和行为注意力模块去除, 用来验证每个子模块的有效性. CRDAN-all 代表完整的本文提出的方法. 我们使用 Tensorflow 和 Keras 进行模型实现, 设置学习率为 0.001, 块大小为 256, 使用 Adam [28] 作为模型学习的优化器, 超参数 λ 由交叉验证决定, 在实验中分别被设置为 0.6.

4.1.3 衡量指标

本文实验中, 我们采用 Precision@k, Micro-F1 和 Macro-F1 作为传染路径风险评估的衡量指标. 详细来说, 我们将预测正确的正样本的数量记为正阳性 (TP), 预测错误的正样本数量记为假阳性 (FP), 预测正确的负样本记为正阴性 (TN) 和预测错误的负样本记为假阴性 (FN). 则相应的计算方法为

$$\begin{aligned} \text{Precision}_i &= \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \quad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}, \quad \text{F1}_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}, \\ \text{Precision} &= \frac{\sum_{i=1}^l \text{TP}_i}{\sum_{i=1}^l (\text{TP}_i + \text{FP}_i)}, \quad \text{Recall} = \frac{\sum_{i=1}^l \text{TP}_i}{\sum_{i=1}^l (\text{TP}_i + \text{FN}_i)}, \end{aligned} \quad (11)$$

其中 l 表示为传染路径风险评估分级的类别数, 本文分为五级风险. 因此我们得出衡量指标 Micro-F1 和 Macro-F1 的计算方法为

$$\text{Micro-F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Macro-F1} = \frac{1}{l} \sum_{i=1}^l \text{F1}_i. \quad (12)$$

Precision@k 用来衡量新路径风险评估问题, 即传染路径是由新的贷款记录构建出来的情况下, 没有针对该路径的信贷评估历史. 在实验中我们衡量模型预测出的置信度最高的 k 个样本的精确度, 其计算方法为

$$\text{Precision@k} = \frac{|\{i | i \in C_p \cap C_o\}|}{|C_p|}, \quad (13)$$

其中 C_p 代表模型预测出的 k 个置信度最高的传染路径, C_o 代表这些路径的实际观测值, $|\cdot|$ 代表集合中样本的数量.

4.2 传染路径风险评估实验

本小节分别验证了本文设计的方法和其他对比方法在传染路径风险评估上的精度表现. 传染路径风险评估任务是本文的主要目标任务, 表 1 中汇报了它们在五级风险分类上的 F1-score, 以及综合评

表 1 不同方法在传染路径风险评估的精度对比

Table 1 Comparison results of different methods in contagion path risk rating

Methods	Normal	Low	Medium	High	Loss	Macro-F1	Micro-F1	N features
LR	0.7561	0.6099	0.4812	0.9605	0.7537	0.7354	0.7048	~500
GBDT	0.7722	0.6422	0.5412	0.9635	0.8086	0.7626	0.7270	~500
DNN	0.7820	0.6637	0.5598	0.9659	0.8081	0.7707	0.7394	~500
LSTM	0.7817	0.6684	0.5704	0.9626	0.8046	0.7711	0.7412	~500
SGC	0.7971	0.6916	0.6087	0.9679	0.8241	0.7895	0.7599	17
GAT	0.8076	0.7164	0.6419	0.9673	0.8179	0.7952	0.7726	17
HGAR	0.8129	0.7216	0.6522	0.9688	0.8393**	0.8067	0.7805	17
CRDAN-noPA	0.8125	0.7224	0.6488	0.9690	0.8359	0.8050	0.7798	17
CRDAN-noBA	0.8205	0.7323	0.6565	0.9673	0.8391	0.8083	0.7859	17
CRDAN-all	0.8344**	0.7539**	0.6953**	0.9697**	0.8298	0.8189**	0.8023**	17

估指标 Macro-F1 和 Micro-F1 的数值. 同时, 针对 LR, GBDT, DNN 和 LSTM 4 个没有图神经网络层的机器学习模型, 我们基于银行业金融领域知识中常用的 Recency-Frequency-Monetary (RFM)^[31] 特征工程方法, 手动构建约 500 个人工特征, 其他方法直接使用原始的属性信息作为模型的输入. 本实验中, 我们对每一个方法测试 10 次, 表 1 中展示了结果的均值. 同时将最优方法与其余次优方法进行配对 t 检验, ** 表明精度的提升经过了 $p < 0.01$ 的统计显著性测试.

从表 1 中可以看出, 在风险评估领域最常用的方法 LR 的精度并不能够达到理想效果. GBDT 和 DNN 的精度略高于 LR, 表明增加模型的容量有助于提升传染路径风险预测的精度. LSTM 的效果略高于 DNN 和 GBDT, 说明运用深度学习历史行为的模式在风险预测中的有效性. 引入图神经网络的 SGC, GAT 和 HGAR 的方法精度明显优于上述 4 种模型, 进一步体现了学习网络特征的重要性. 高阶图注意力机制的 HGAR 方法是所有基准方法中效果最好的, 证明了注意力机制在传染路径风险预测中的领先效果. 最后本文方法的两个变体 CRDAN-noPA 和 CRDAN-noBA 分别和基准方法 GAT 以及 HGAR 效果相当, CRDAN-all 完整模型显著优于上述方法, CRDAN 只使用 17 个原始特征所获得精度比进行了复杂的特征工程后 LR 方法在次级 (medium) 类别上 F1-score 提升了 43%, 对最好的基准方法 HGAR 提升 5% 以上. 在全部 7 个衡量指标中, CRDAN 有 6 个是所有对比方法中最优的, 其中综合评估指标上, CRDAN 比最好的基准方法 HGAR 和 GAT 提升了大约 2%, 进一步表明了针对传染路径风险进行方法开发的必要性和本文方法在该任务中的有效性.

4.3 新路径风险评估实验

当传染路径是由没有历史记录的新用户构成时, 大多数依赖历史特征进行学习的传统方法在此类任务中难以进行有效评估. 在本实验中, 我们将 CRDAN 和各个对比方法在新构建的路径上的性能进行详细评估. 首先, 我们按照时间顺序依次抽取新出现的传染路径给不同的方法进行预测和评估, 将模型预测的前 k 个置信度最高的样本的准确度 (Precision@k) 汇报在图 4 中. 图 4 中 x 轴代表预测的前 k 个置信样本的 k 值, y 轴代表对应的准确度 (precision). 图 4 的 4 个子图分代表有风险的 4 个等级上的结果, 即关注 (low)、次级 (medium)、可疑 (high) 和损失 (loss). 从中可以看出, 带有图

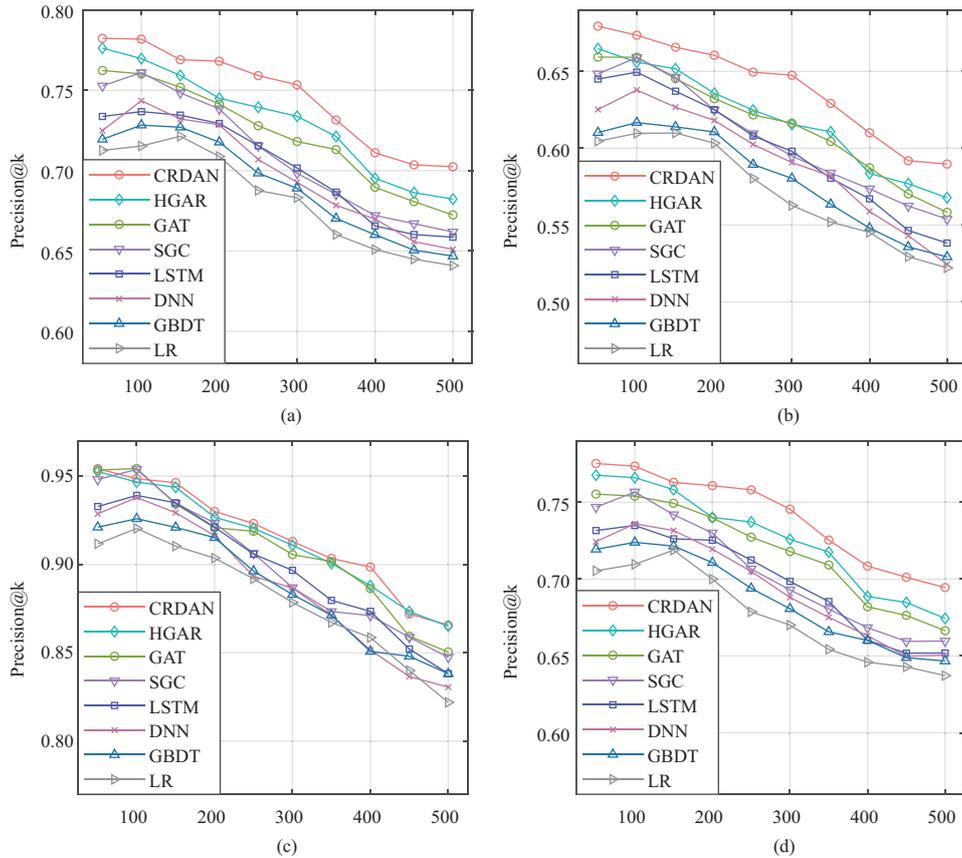


图 4 (网络版彩图) 新传染路径评估任务中每个风险类别 ((a) 关注, (b) 次级, (c) 可疑, (d) 损失) 的 Precision@k 精度值对比图

Figure 4 (Color online) The Precision@k of each risk category ((a) low, (b) medium, (c) high and (d) loss) in new constructed contagion paths

神经输入层的方法要显著高于其他的 4 种方法 (LR, GBDT, DNN 和 LSTM), 其中带有注意力机制的 GAT 和 HGAR 是所有基准方法中精度最高的, 表明了使用注意力机制在新路径评估任务中的重要作用, 而传统机器学习方法依赖于比较充足的历史特征才能够训练和学习出有效的风险判别能力, 因此这些方法在新路径评估任务中的精度显著低于 GAT, HGAR 以及 CRDAN 等由图注意力机制网络构成的神经模型. 本文提出的方法 CRDAN 由于同时使用了结构注意力机制和时序注意力机制对图神经网络的输出层以及时序神经网络进行联合训练和学习, 所以在所有的基准方法中以及所有的风险评估类别中都是效果最好的, 甚至在关注和次级的预测任务中, CRDAN top-500 所获得的准确率和 LR top-50 的值相当, 平均的精度提升达到了 12%, 在所有风险级别上的精度提升超过 8%. 结果充分表明了本文设计的方法在新传染路径评估任务中的优秀的性能, 能够有效缓解和克服前文提出的没有历史样本记录进行风险评估的挑战.

4.4 注意力机制的可解释性探索

对可解释性的要求是金融行业区别于其他领域的数据科学研究的主要特点. 为了验证本文设计的方法在精准评估传染路径风险方面能够带来的业务指导意义, 本小节将注意力机制学习到的权重分别

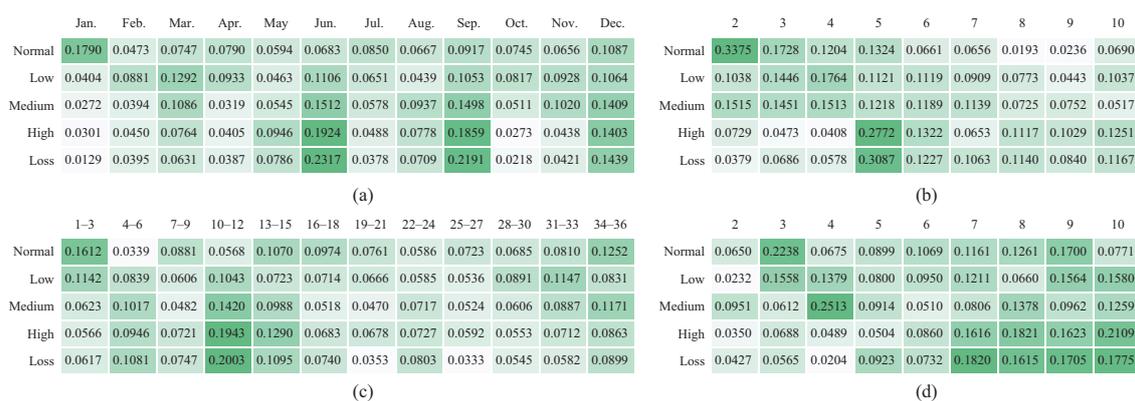


图 5 (网络版彩图) 在 4 个不同视图下的注意力权重的可视化展现

Figure 5 (Color online) Visualization of attention weights in four different views. (a) Attention weights on each month; (b) the number of inner-chain nodes; (c) the number of chain duration months; (d) the number of neighbor chains

在不同的风险等级以及不同的视图上进行平均聚合, 随后使用矩阵热力图进行呈现和探索分析. 我们分别抽取了时序注意力网络中的发生月视图和持续时间视图, 以及多路径注意力网络中的路径内节点视图和相邻路径视图进行展现, 结果如图 5 所示.

- 发生月视图: 我们在构建阶段中, 将传染路径标注上其初始发生的月份 (1~12 月), 对应的注意力机制矩形热力图如图 5(a) 所示. 可以看出, 1 月份对于正常的贡献度较大, 即为多数构建于 1 月份的路径在鉴别是否为正常上比较重要. 低风险类别中没有明显重要性的月份特征; 高风险和损失类别中, 6 月、9 月和 12 月的显著性都比较高, 经过金融业务专家的分析得出可能是因为大多数中小企业在 2~4 季度末的时候面临还款的主要压力, 因此在银行业务的风险规则制定时以季度为单位进行重点风险排查.

- 持续时间视图: 图 5(c) 中展示了传染路径持续的时间对于不同风险等级的注意力权重. 可以看出 1~3 个月的路径对于正常等级评估重要性较高, 持续了 10~12 个月的特征在高风险和损失类别的路径评估中最为显著. 因此, 在业务实践中, 需要重点关注持续了 10~12 个月的路径.

- 路径内节点视图: 在路径内节点视图中, 我们把内部节点的数量进行了可视化, 如图 5(b) 所示. 路径内节点数量较低对于正常类别和低风险的评价的重要性较高, 而含有 5 个左右的内部节点的路径对于高风险的识别重要性最为显著. 该发现与我们在前期工作中的结果基本吻合^[7].

- 相邻路径视图: 图 5(d) 中展示了相邻路径个数对于目标传染路径风险评估的重要性, 可以看出 3~4 个相邻路径在正常和低风险的分类中比较重要, 拥有 7 个以上的邻居路径在所有风险类别中都比较重要, 尤其在高风险和损失类别中重要性更加显著. 因此, 在风控实践中, 需要对该类规模的传染路径进行重点监控.

4.5 模型部署和案例研究

本小节介绍 CRDAN 在金融系统中的部署细节、性能表现以及实证研究的内容. 与大多数深度网络模型一样, 在部署中, CRDAN 分为离线训练和在线评估两个流程. 其中离线训练包括传染路径构建和模型训练部分, 以定时任务的形式每天更新模型. 由于离线训练过程中, 耗时较长的主要是需要对历史数据进行聚合的特征工程, 本文方法无需进行上述操作. 同时仅对每天变化的部分进行更新, 所以时效性得到了很大的提升. 在实际场景中, 使用 2 台双 Telsa P100 GPU 的服务器, 仅需 26 min 即完成了数据集中全量对公客户每日模型离线训练的批处理任务, 而现有基于评分卡的方法仅在特征工

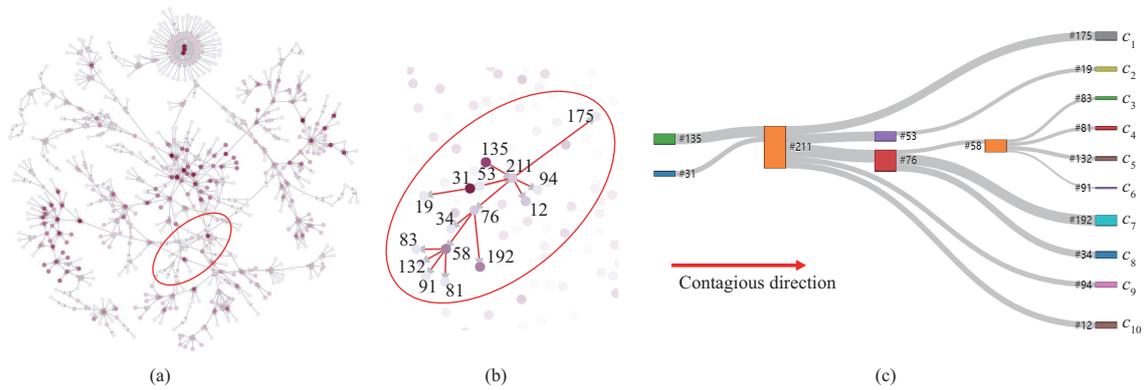


图 6 (网络版彩图) 真实担保网络中检测出的高风险传染路径案例研究

Figure 6 (Color online) The visualization of the real-world loan network and the predicted high-risk contagion paths. (a) A real-world complex loan network; (b) contagion paths; (c) the Sankey diragram view of contagion paths

程的过程中就花费了 1.5 h. 在线评估阶段, 本文方法单线程能够在 20 ms 内完成对单路径的评估, 同时保留了并发的可扩展性, 完全满足商业银行每日新增评估的需求. 由于在参数量相近的情况下, 深度网络预测任务耗时基本相近, 所以直接针对传染路径进行评估的本文方法, 比需要先对每个个体预测后再得出路径评估的 HGAR 和 GAT 等方法性能提升了 3~6 倍不等 (取决于网络中的平均传播路径的长度). 而传统方法由于在线预测时需要进行特征工程的加工, 性能方面远远低于本文方法.

在部署后的观察期中, 我们对本文方法的预测结果进行实证研究. 图 6(a) 中展示了代号为 GN173 的真实担保网络, 图 6(b) 对使用本文方法 CRDAN 进行传染路径风险评估检测出的高风险路径群组进行了放大展示, 图中节点颜色的深浅代表个体违约的概率, 颜色越深代表个体违约的概率越高. 传染路径在图 6(b) 中用红色描边的有向边来表示. 可以看出, 检测出的高风险路径群一共包含了 10 条, 涉及 16 家中小企业, 编号为 211, 76 和 58 的节点是该路径群中的关键节点, 占据着传染过程的关键位置. 编号为 135, 31, 192 和 58 的节点自身违约概率相对较高.

从图 6(a) 中, 我们可以看出, 检测出的高风险路径群的区域, 其平均个体违约风险和全网络基本相同. 为了深究该区域被本文方法识别为高风险区域的根本原因, 我们对该传染路径群进行了实证研究. 图 6(c) 展示了路径群的桑基图, 图中传染边的宽度代表企业间的担保金额. 我们发现节点 211 是关键因素, 该节点为编号为 135 和 31 的企业提供担保, 这两个企业都有着较高的自身违约风险, 因此即使 211 号节点自身风险较低, 其仍然是最容易被传染导致违约的企业. 更重要的是, 该路径群中, 众多的下游其他企业为 211 号公司提供担保, 一旦节点 211 发生违约, 则该路径极易发生系统性的信贷危机, 进而导致整个路径群中的企业全部发生违约. 在实证研究中, 该路径群构成的子网络涉及贷款金额超过 6000 万元人民币. 本文提出的方法预测出来的高风险群体中, 即暴露在风险敞口中的资金额度超过 50% 的贷款企业构成的传播路径, 使用金融机构传统的风险评估方法, 该传染路径仅被评级为关注类别, 意味着传统方法认为只有低于 5% 的资金处于风险中. 随后, 我们同合作金融机构的风控经理对其中的企业进行了实证研究, 对涉及的贷款进行风险摸底和排查. 实证研究结果证明了本文方法的有效性.

5 结论和展望

本文提出了一种基于深度图神经网络的担保贷款传染路径风险评估方法. 该方法联合学习中小企

业贷款的行为特征及其在担保网络中的结构特征,进而显著提升了违约风险评估的精确度.其中,图网络输入层可以直接学习担保贷款的原始行为记录,无需进行复杂的特征工程.多路径注意力机制可以为没有历史记录的人群动态学习其相邻路径的隐含特性,进而做出针对该路径的可信的风险评估.在真实场景的实验中,我们验证了本文提出的方法 CRDAN 的有效性,在风险评级和新路径评估任务中,本文方法能够比基准方法平均性能提升 2% 至 15%.同时通过实证研究探索了本方法中注意力机制在实际业务问题中的可解释性,表明了本文方法在传染路径系统性风险监控方面的实际价值和应用前景.

在未来的工作中,我们计划通过深度神经网络和复杂网络的最新理论研究进展,尝试探索担保网络传染路径风险评估的可解释性机理和更加高效的动态评估方法,进而为全方位的信贷风险评估和系统性的金融信贷风险监控提供理论和方法支撑.

参考文献

- 1 Jian M, Xu M. Determinants of the guarantee circles: the case of Chinese listed firms. *Pac-Basin Finance J*, 2012, 20: 78–100
- 2 Meng X F, Tong Y H, Liu X H, et al. Netrating: credit risk evaluation for loan guarantee chain in China. In: *Proceedings of Pacific-Asia Workshop on Intelligence and Security Informatics, Korea, 2017*. 99–108
- 3 Wu F N. An overview of guaranteed-loan industry in China. Qianzhan Industry Research Institute. 2019. <https://www.qianzhan.com/analyst/detail/220/190312-07aee546.html>
- 4 Li W, Ben S, Hommel U, et al. Default contagion and systemic risk in loan guarantee networks. *Account Finance*, 2019, 59: 1923–1946
- 5 Khandani A E, Kim A J, Lo A W. Consumer credit-risk models via machine-learning algorithms. *J Banking Finance*, 2010, 34: 2767–2787
- 6 Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks. In: *Proceedings of the 5th International Conference on Learning Representations, Vancouver, 2018*. 1–12
- 7 Cheng D W, Tu Y, Ma Z W, et al. Risk assessment for networked-guarantee loans using high-order graph attention representation. In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, 2019*. 5822–5828
- 8 Cui P, Wang X, Pei J, et al. A survey on network embedding. *IEEE Trans Knowl Data Eng*, 2019, 31: 833–852
- 9 Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2014*. 701–710
- 10 Tang J, Qu M, Wang M Z, et al. Line: large-scale information network embedding. In: *Proceedings of the 24th International Conference on World Wide Web, Florence, 2015*. 1067–1077
- 11 Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016*. 855–864
- 12 Wang D X, Cui P, Zhu W W. Structural deep network embedding. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016*. 1225–1234
- 13 Liu J, He Z C, Wei L, et al. Content to node: self-translation network embedding. *IEEE Trans Knowl Data Eng*, 2019, 1: 1–14
- 14 Wang X, Zhang Y D, Shi C. Hyperbolic heterogeneous information network embedding. In: *Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, 2019*. 5337–5344
- 15 Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: *Proceedings of Advances in Neural Information Processing Systems, Barcelona, 2016*. 3844–3852
- 16 Geng X, Li Y G, Wang L Y, et al. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In: *Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, 2019*. 3656–3663
- 17 Cheng D W, Xiang S, Shang C C, et al. Spatio-temporal attention-based neural network for credit card fraud detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence, New York, 2020*. 362–369
- 18 Cheng D W, Tu Y, Ma Z W, et al. BHONEM: binary high-order network embedding methods for networked-guarantee

- loans. *J Comput Sci Technol*, 2019, 34: 657–669
- 19 Detering N, Meyer-Brandis T, Panagiotou K, et al. Managing default contagion in inhomogeneous financial networks. *SIAM J Finan Math*, 2019, 10: 578–614
 - 20 Battiston S, Puliga M, Kaushik R, et al. Debtrank: too central to fail? Financial networks, the FED and systemic risk. *Sci Rep*, 2012, 2: 541
 - 21 Acemoglu D, Ozdaglar A, Tahbaz-Salehi A. Systemic risk and stability in financial networks. *Am Economic Rev*, 2015, 105: 564–608
 - 22 Bougheas S, Kirman A. Complex financial networks and systemic risk: a review. In: *Complexity and Geographical Economics*. Cham: Springer, 2015. 19: 115–139
 - 23 Catanzaro M, Buchanan M. Network opportunity. *Nat Phys*, 2013, 9: 121–123
 - 24 Meng X, Liu X H. Credit risk evaluation for loan guarantee chain in China. In: *Proceedings of the 13th Credit Scoring and Credit Control Conference*, Edinburgh, 2015. 1–42
 - 25 Brohée S, Faust K, Lima-Mendez G, et al. Network analysis tools: from biological networks to clusters and pathways. *Nat Protoc*, 2008, 3: 1616–1629
 - 26 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Proceedings of Advances in Neural Information Processing Systems*, Long Beach, 2017. 5998–6008
 - 27 Chung J, Gulcehre C, Cho K, et al. Gated feedback recurrent neural networks. In: *Proceedings of International Conference on Machine Learning*, Lille, 2015. 2067–2075
 - 28 Chen X Y, Liu S J, Sun R Y, et al. On the convergence of a class of Adam-type algorithms for non-convex optimization. In: *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, 2019. 1–30
 - 29 Tan F, Hou X R, Zhang J, et al. A deep learning approach to competing risks representation in peer-to-peer lending. *IEEE Trans Neural Netw Learning Syst*, 2019, 30: 1565–1574
 - 30 Wu F L, Zhang T Y, Souza A H, et al. Simplifying graph convolutional networks. In: *Proceedings of International Conference on Machine Learning*, Long Beach, 2019. 6861–6871
 - 31 Colombo R, Jiang W. A stochastic RFM model. *J Interactive Marketing*, 1999, 13: 2–12

Risk assessment for contagion path in complex loan network

Dawei CHENG^{1,2}, Zhibin NIU^{3,2}, Xinhai LIU⁴ & Liqing ZHANG^{1,2*}

1. *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;*

2. *MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China;*

3. *School of Intelligence and Computing, Tianjin University, Tianjin 300354, China;*

4. *Center of Finance Intelligence Research, Peking University, Beijing 100871, China*

* Corresponding author. E-mail: zhang-lq@cs.sjtu.edu.cn

Abstract Small and medium-sized enterprises (SMEs) loans play an essential role in many aspects: including technological innovation, economic development, employment, and people’s livelihood, etc. In order to meet the loan evaluation criteria of commercial banks, many SMEs choose to guarantee each other to obtain loans, thus forming a complex guarantee network. If the borrower defaults on the loan, the risk will be diffused to its guarantors along with the contagion path, which may lead to systemic risk across the loan networks. This has brought severe challenges to the nation’s financial security and regulation. Thus, accurately rating the contagion path is an urgent task for systematic risk management in the loan network. Therefore, we present a deep learning-based approach to the risk rating of contagion paths in the bank industry. We leverage the graph neural network and attention mechanism on graph-structured loan behavior to learn high-order representations, which do not require handcraft feature engineering. We demonstrate that our approach outperforms the existing baselines with 2%~15% improvements in risk rating and 3.5% in the newly constructed path rating problem. The result demonstrates the effectiveness of our proposed approach, which provides an effective method and theory basis for regulatory commissions and financial institutions to monitor systematic risks in networked-guarantee loans.

Keywords risk assessment, contagion path, loan network, graph neural network, attention mechanism



Dawei CHENG was born in 1987. He received his Ph.D. degree from Department of Computer Science and Engineering, Shanghai Jiao Tong University in 2018. He is currently a Postdoc researcher with MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University. His research interests include machine learning, data mining, and pattern recognition, with a focus on complex network analysis.



Zhibin NIU was born in 1984. He is a recipient of Marie Curie Fellow and received his Ph.D. degree in computer sciences from Cardiff University, UK, in 2016. He is currently a lecturer with School of Intelligence and Computing at Tianjin University, China. His research interests include reverse engineering, data mining, and visual analytics for the financial data.



Xinhai LIU was born in 1976. He received his Ph.D. degree from Katholieke Universiteit Leuven (K.U. Leuven), Belgium, in 2011. He is currently a researcher in China Mergers & Acquisitions Association, and a researcher with Center of Finance Intelligence Research, Peking University. His research interests include financial network analysis, alternative data based credit scoring.



Liqing ZHANG was born in 1961. He received the Ph.D. degree from Sun Yet-sen University, Guangzhou, China, in 1988. He is now a full professor with Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include visual cognitive computing, uncertainty reasoning, and statistical machine learning.