基于知识图谱的小微企业贷款申请反欺诈方案

金磐石[1]，万光明[2]，沈丽忠[3]
1. 中国建设银行股份有限公司，北京 100033；2. 中国建设银行金融科技部，北京 100032；
3. 建信金融科技有限责任公司厦门事业群，福建 厦门 361008

摘要
近年来，在各大商业银行竞相开展小微企业信贷业务的同时，贷款欺诈风险也随之产生。针对小微企业信贷业务的特点，提出了一种基于全方位企业画像与企业关联图谱的贷前反欺诈模型。通过整合多源信息，形成完整的企业属性特征，并结合从图谱中提取的关系网络结构特征，把特征共同输入模型，以定量评估小微企业客户的欺诈风险。实验表明，使用隐含在关系图谱中的信息比单纯使用企业自身特征建模在测试集上的AUC提高了5%，有助于银行机构准确地对企业申贷欺诈行为进行评估。

关键词
反欺诈；企业画像；关联图谱

中图分类号：TP393　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2019035

Knowledge graph-based fraud detection for small and micro enterprise loans

JIN Panshi[1], WAN Guangming[2], SHEN Lizhong[3]
1. China Construction Bank, Beijing 100033, China
2. Mistry of Finance and Technology, China Construction Bank, Beijing 100032, China
3. China Construction Bank Fintech Co., Ltd., Xiamen 361008, China

Abstract
While major commercial banks have been providing various business loans, the risk of loan fraud has arisen at the same time. In order to overcome this challenge, an anti-fraud model was proposed based on the full-scale enterprise portrait and enterprise relation graph. By integrating multi-source information to form a concrete enterprise profile, and quantifying the interactions among enterprise entities, the fraud risk of new SMEs'loan applications could be quantitatively evaluated. Experiments show that compared with purely considering enterprise's attributes, the additional use of relational information contributes a 5% increase in the AUC of the test set, which is more helpful for banking institutions to accurately assess the corporate fraud risk.

Key words
anti-fraud, enterprise portrait, relation graph

1 引言

随着我国经济发展步入新常态，小微企业快速发展，并逐渐成为我国市场经济中较活跃的主体。据相关数据统计，目前我国约有近1亿户小微企业，这些小微企业广泛分布于实体经济的各行各业[1]。作为经济发展和社会稳定的重要支柱，小微企业在促进人才有序流动、维护市场活力、推动科技创新等方面发挥着不可或缺的作用。然而相对于大中型企业而言，小微企业在市场竞争中仍处于弱势地位，融资问题导致其很难保证经营的稳定性与持续性[2]。

银行贷款是企业融资的重要手段，为缓解小微企业的经营压力并满足强烈的融资需求，商业银行的小微贷款业务逐步扩张，各类信贷产品应运而生[3]。截至2018年第三季度，小微企业在银行机构的贷款总额达33.04万亿元，贷款授信户数超过1 800万户。但由于小微企业在提供自身资讯上的天然弱势，“不透明”“内部化”的非对称数据信息使得银行难以在业务申请阶段把控小微客户实质性的信贷风险，对小微企业信贷产品的管理比大型企业也困难得多[4]。如今，欺诈风险已经成为消费金融业务面临的主要风险之一，反欺诈也对银行的风险控制技术提出了更高的要求。科学合理地进行小微企业贷款欺诈风险管理已经成为商业银行亟须解决的问题。

小微企业贷款业务是近年来兴起的一种贷款手段，由于其“新”的特征，在金融行业内对小微信贷欺诈尚没有一个通用的定义。商业银行在信贷申请欺诈风险的防范工作方面已经积累了很多经验，但是当前的工作多为面向大中型企业或面向个人信贷的反欺诈方案。不论是较为原始的黑白名单甄别、实现自动决策的规则引擎，还是基于人工智能技术的风险量化与无监督的欺诈模式识别，反欺诈的分析和度量技术在小微企业领域仍在探索的过程中[5]。

为解决以上难题，本文针对小微企业的信贷业务特点提出了一种基于全方位企业画像与企业关联图谱的贷前反欺诈模型。该模型从企业自身、企业实控人、企业干系人以及企业网络关系图谱四大维度提取特征，然后在金融行业知识的辅助下，通过数据挖掘、特征筛选等方法定义多个欺诈场景，并利用LightGBM分类器对欺诈进行概率推断。相比传统的信贷欺诈分析模型，本文方案将关系型信息和企业节点自有属性信息有机地结合，能够有效地挖掘小微企业与欺诈目标的非线性关系，有助于全面、有效地在申请阶段评估小微企业的欺诈风险状况。相比于传统的企业风控反欺诈，本文方案的创新点如下。

- 针对小微企业信息不对称的特点，整合多源数据集，在提取企业自身特征的同时，利用其关联实控人及相关干系人的基本信息与征信状况生成全方位企业画像，更加完善地刻画小微企业可能存在的信贷风险。
- 结合图知识和金融行业知识，挖掘欺诈场景，通过分析小微企业命中欺诈场景的情况，输出欺诈概率。
- 挖掘并构建企业与企业间的关联信息，建立企业关联图谱，有利于识别出异常的欺诈行为。
- 归约掉图谱中的非企业节点，将复杂、庞大的异构关系网络折叠为仅保留小微企业实体的同构网络。这既解决了网络存储的开销问题和网络特征提取的计算复杂度高的问题，又能够隔离无关节点的干扰，直观地呈现企业之间的联系。

● 使用 LightGBM 分类框架，采用有监督学习模式在构建的关联图谱上进行基于节点自身属性特征与网络特征的欺诈概率预测，从关系和实体属性两个角度共同进行欺诈评估。

2 相关工作

2.1 企业传统信贷反欺诈

商业银行在贷款业务申请阶段的反欺诈手段主要集中在配置规则引擎和建立机器学习模型上。规则引擎起源于基于规则的专家系统，用来模拟人的行为，以实现计算机自动决策。它是一种建立在对欺诈行为的特点与模式充分认知的基础上，针对单一或组合欺诈行为设计的启动和触发机制[6]。规则引擎主要用来核实信息的真实性，如是否存在不良征信记录、是否命中风险名单等，但无法检测复杂的欺诈模式。

有监督学习是当下反欺诈检测中应用广泛的机器学习方法，能够基于历史数据（即已知的欺诈申请和正常申请）建立分类模型，输出欺诈概率，更好地量化欺诈风险。该方法收集客户申请信息并将其作为训练集，训练出的机器学习模型通过对用户特征的抽象理解，分析特征间的隐藏关系，填补并增强规则引擎无法覆盖的复杂欺诈行为。业界常用的模型算法包括逻辑回归[7]、决策树[8]、支持向量机[9-10]、XGBoost[11]和神经网络[12]等。

2.2 基于关联图谱的信贷反欺诈

在反欺诈场景中，除了考虑单一信息点的属性，企业间的隐藏关联往往包含更多未知的潜在信息。因此，企业信贷欺诈的识别问题也可以转化为企业关系图挖掘或社交网络分析问题。基于知识图谱的信贷反欺诈旨在将多源异构的数据整合成机器可以理解的知识，将“单点”的信息转换成“平面”的相互关联的图谱，进行异常风险检测[13]，从而实现欺诈的识别与防御。

基于构建好的知识图谱，可以使用半监督或无监督的方式进行异常子图挖掘、社群发现或标签染色[14]。除此之外，对网络特征（包括中心度、一度二度关联特征）的直接提取也可以供各类有监督风险评估模型使用。另外，随着深度学习算法的普及，知识图谱的关联分析引入了网络嵌入这种图表示学习的思想。在尽量保留网络信息的前提下，根据图谱中的拓扑关系，用低维向量表示每个节点与其周围节点的关联特性。作为一种表征学习的方法，图嵌入可以解决图数据的高维度、稀疏性等问题[15]。经过向量的表征后，再利用大量算法进行边的预测、分类、聚类。

2.3 小微企业信贷反欺诈的难点

尽管商业银行在反欺诈方面已经展开了较多尝试，但是针对小微企业贷前的欺诈识别仍处于探索阶段。其难点主要集中在以下方面。

（1）小微企业信息透明性差

小微企业与传统的银行贷款业务有很大差异，如果采用无差异化的风控模型和放贷标准，绝大部分小微企业无法通过审核。具体地，对于传统的中大型企业信贷业务，银行通常采用申请评分卡（application score card）的方式决定是否放贷以及贷款额度。这样的方式对企业申请贷款时提交信息的完整性和真实性有极高的要求，如果信息造假或者有缺值，

评分卡方式很难准确判断企业的信用情况。中大型企业通常采用标准的财务审计制度，经营信息也相对透明，是可以通过评分卡的方式授信放贷的。但是对于小微企业而言，其经营规范性差、信息透明度差，评分卡的方式在很大程度上并不适用于小微信贷业务。

（2）银行与企业之间信息不对称

银行与企业之间的信息不对称与欺诈标签的定义对有监督学习提出了挑战[1]。由于小微企业缺乏规范的财务报表和税务审计信息，经营透明度差，银行难以掌握小微企业的真实状况，无法全面刻画客户特征，从而进行风险评估。此外，由于银行针对小微企业推出的大多贷款业务开展时间不足一年，对于欺诈标签的定义尚不明确，真实欺诈标签的严重不足阻碍了模型对欺诈特征的学习与把控。

（3）图嵌入模型可解释性有限

基于知识图谱的方法能够捕捉到企业间的关联，利用这些关联，可以得到全新的反欺诈分析。金融反欺诈实践中常用的图谱均为异构图谱，即网络中的节点与边涵盖多种类型，这加大了图特征提取以及图嵌入方法选择的难度。以某金融机构信用卡申请关系图谱为例，个体类型包括申请人的IP地址、设备、账户联系人等，个体之间也存在不同的关系，如IP登录行为、设备登录行为等。尽管一些网络特征指标在实际反欺诈实践中效果非常显著，但由于反欺诈对实时决策的高需求，这些指标都需要迅速地计算提取。节点的中介中心性、集聚系数等指标存在较高的计算复杂度，如果不采取一些近似的算法并进行预计算，难以满足实施决策的业务需求。基于深度学习的图嵌入算法的选择同样受到异构网络的局限，大多网络嵌入方法（如Perozzi等人[16]提出的DeepWalk、Grover等人[17]提出的Node2vec、适用于大规模信息网络的LINE[18]）仅适用于同构网络的表征学习。而2017年由Dong等人[19]提出的Metapath2vec虽然对DeepWalk进行了扩展，使用基于元路径的漫游捕获不同类型节点之间的关系，进而解决异构网络的学习问题，但其网络嵌入可解释性有限，无法对任务决策做出解释。

（4）孤立企业信息难以利用

尽管现有的反欺诈方案可以分别从孤立的企业信息或整体的企业网络结构来甄别欺诈，但是还没有一种明确的方案同时从这两种维度着手。如何有效地处理并综合利用这两种信息，使二者相互辅助共同发挥作用是当前小微企业反欺诈着重关注的问题。

3 小微企业贷款申请反欺诈方案

在有监督机器学习预测分析中，信贷反欺诈的目标变量是既定的欺诈标签。在关系图谱的学习中，目标是在给定的网络中寻找高风险的欺诈模式。如果将二者结合，则需找出合适的方式在一组特征中概括网络，然后将这些特征与非网络（即节点属性）特征相结合进行预测，从而得到特定节点的欺诈概率。本文提出了一种解决方案，采用有监督的学习方法，从构建企业关联图谱入手，整合多个数据源，全面捕获企业的独立属性与企业间的隐藏联系。在构建图谱时，将多维度的企业特征作为节点属性，并将复杂的异构网络归约为仅保留申请贷款企业的同构网络，既使企业关系可以直观展现，又能降低图谱的存储空间与网络特征指标的计算成本。最终将节点属性与网络特征共同输入LightGBM分类模型进行欺诈预测。方案整体模型架构如**图1**所示。

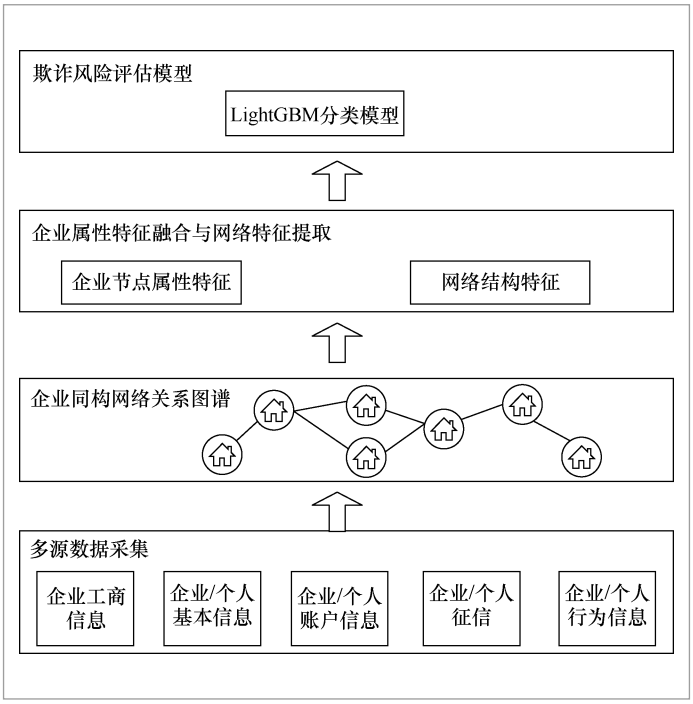


图1 模型架构

3.1 小微企业信贷欺诈的定义

与传统银行信贷不同，小微企业的信贷需求具有“短、小、频、急”的特点，其小额、短期、分散的特征更类似于零售贷款。由于小微企业对资金流动性的要求更高，目前银行机构对小微企业信贷业务通常采取定额循环贷款的方式，即小微企业在贷款审批通过后，可以在贷款周期内多次取出、还入贷款。这样的灵活性适应了小微企业的信贷需求，但同时也极大地限制了传统的信贷风控模型在小微企业信贷业务上的效用。

近年来，受政策利好的影响，我国小微企业贷款业务蓬勃发展。但是由于小微信贷的业务特性以及当前绝大多数小微企业信贷仍未到还款期，很难直接获取还款逾期名单，也很难采用一些传统信贷欺诈判别规则来判定小微信贷欺诈。对于什么是小微信贷欺诈这个问题，业界目前没有明确的定义，整个行业对这个问题的定义仍在不断探索。

但是，从业务角度出发，仍可以定义一些高置信度的小微企业信贷欺诈场景，例如：单日多次缴税、资金出入差异过大、与涉及P2P业务的公司频繁往来等。笔者基于企业画像和关联图谱，在金融行业知识的辅助下，挖掘出数十种小微企业信贷欺诈场景，经过对关键信息的筛选和合并，最终保留了8种高置信度信贷欺诈场景，并对筛选后的欺诈场景进行了相关性排序。通过对当前图谱中的节点进行分析，最终能够获得小微企业命中欺诈场景和小微信贷欺诈的概率分布。具体地，对于小微企业e以及欺诈场景集合S，如果该企业命中欺诈场景$s_1, s_2, \cdots, s_n \in S$，则可以给出该小微企业信贷欺诈概率$p(\hat{e}|s_1, s_2, \cdots, s_n)$，其中，$\hat{e}$代表该小微企业信贷欺诈。在实际应用中，通常设定概率阈值，若小微企业的信贷欺诈概率超过阈值，则认为该小微企业有欺诈风险。

3.2 建模粒度与标签定义

本方案以企业为粒度进行建模，后续网络的构建与特征工程都将围绕每一家待预测的小微企业展开。此外，由于反欺诈模型的构建仍采用有监督的学习模式，明确欺诈的定义是保证模型预测效果的前提。如第2.3节所述，由于历史数据中缺少对企业欺诈的准确描述，本方案将根据银行内部及相关部门披露的企业及企业干系人在2018年1月到2019年3月的严重违规记录建立欺诈标签，并将该标签作为目标变量。相关企业和个人严重违规数据包括银行内部欺诈系统中的名单以及行政违法记录黑名单。

3.3 图谱构建与特征提取

3.3.1 数据准备

为获取小微企业的全面信息，本方案用到的数据不仅来源于银行提供的企业数据、个人客户数据，还通过关联外部第三方数据形成具体的企业画像。数据的抽取可分为实体、属性的抽取以及关系的抽取。抽取范围确定在企业贷款申请时间处于2018年1月到2018年12月且有还款表现的企业中。

小微企业申请贷款的原始数据存在于多张数据表中，经过字段清洗、合并等预处理后，将数据存入SQL数据库中。本文方案的输入数据通过直接查询数据库获得。数据准备完成之后，按照预定义的键值处理方法对不同的数据类型进行规范化处理，例如：对于个人信息中的身份信息，使用“01”代指企业实控人，使用“02”代指股东、高管等其他企业干系人。

（1）实体和属性抽取

以企业为粒度进行建模，即实体是企业。对企业及其主控人、干系人的数据信息进行属性提取。

企业信息数据包括如下内容。

- 企业基础信息：企业技术编号、注册地址、行业分类、法人信息等。
- 企业工商类数据：企业注册日期、工商信息变更次数、法人信息变更次数等。
- 企业存款数据：贷款余额、存款余额、存款月积数等。
- 转账数据：半年内转出次数及金额、半年内转入次数及金额等。
- 征信数据：企业黑名单、企业第三方征信等。

实控人与干系人信息数据包括如下内容。

- 个人基础信息：年龄、学历、职务、婚姻状况、子女状况等。
- 个人存款数据：存款时点余额、时点个人资产管理规模、一年内申请贷款的次数等。
- 征信数据：个人黑名单、个人征信等。

（2）关系抽取与实体归约

构建知识图谱的前提是建立节点间的关联，而关联的定义需要依据业务需求，并且极为复杂。本文结合业务逻辑，将企业间风险较高的关系进行罗列并汇总为5类：企业与个人的对应关系（如控股关系）、企业干系人与企业干系人的对应关系（如亲属关系）、企业与相关属性的对应关系（如企业共同联系方式）、企业干系人与相关属性的对应关系（如干系人共同联系方式）、企业与企业的对应关系（如企业担保关系）。

具体关系见**表1**。**表1**中建立的原始企业关系涉及多个信息来源，不同的来源对节点和边处理差异很大，由此导致了数据异构、碎片化问题，关系中包含多个实体类型。这些非企业实体在关系的构建中仅作为连接企业节点的媒介，而非网络结构分析的主体。异构网络信息种类敏繁杂，很难从中提取到有用的信息。因此，保证整体企业关系网络同构，实体统一可使网络特征的提取更高效。因此，本文在构建出以上所有关系后，将关系两端的节点限定为企业，并将关系中与考察实体无关的其他类型节点（如干系人、联系方式等节点）进行折叠归约，减少无关节点的干扰，保证每条关系最终对应到企业本身。

3.3.2 同构网络关系图谱构建

基于实体和关系的提取，反欺诈知识

表1 关系的提取与构建

关系	表示	说明
企业与个人的对应关系	$C_1 \xrightarrow{r_1} (P) \xrightarrow{r_2} C_2$	C_1, C_2∈(企业), $C_1 \neq C_2$ P∈(个人) r_1, r_2∈(干系人关系)
企业干系人与企业干系人的对应关系	$C_1 \xrightarrow{r_1} (P_1) \overset{r_3}{\cdots} (P_2) \xrightarrow{r_2} C_2$	C_1, C_2∈(企业), $C_1 \neq C_2$ P_1, P_2∈(个人), $P_1 \neq P_2$ r_1, r_2∈(干系人关系) r_3∈(亲属关系)
企业与相关属性的对应关系	$C_1 \xrightarrow{r_4} (Attr\_c) \xrightarrow{r_4} C_2$	C_1, C_2∈(企业), $C_1 \neq C_2$ Attr_c∈(企业属性) r_4∈(企业属性拥有关系)
企业干系人与相关属性的对应关系	$C_1 \xrightarrow{r_1} (P_1) \overset{r_5}{\cdots} (Attr\_p) \overset{r_6}{\cdots} (P_2) \xrightarrow{r_2} C_2$	C_1, C_2∈(企业), $C_1 \neq C_2$ P_1, P_2∈(个人), $P_1 \neq P_2$ Attr_p∈(干系人属性) r_1, r_2∈(干系人关系) r_5,r_6∈(干系人拥有属性关系)
企业与企业的对应关系	$C_1 \xrightarrow{r_7} (C_3) \xrightarrow{r_7} C_2$	C_1, C_2, C_3∈(企业), $C_1 \neq C_2 \neq C_3$ r_7∈(担保关系)

图谱的具体构建过程可总结为以下3步。

- 以提交信贷申请的企业作为图谱的节点实体。
- 将企业之间的各种归约关系作为图谱的关系。
- 删除图谱中存在的孤立节点。

构建完成的图谱由一个个非连通的子图组成，这些子图可能描述的是某家族企业、某同一控制企业、存在相同类型潜在风险的企业。图2为放大后的子图。

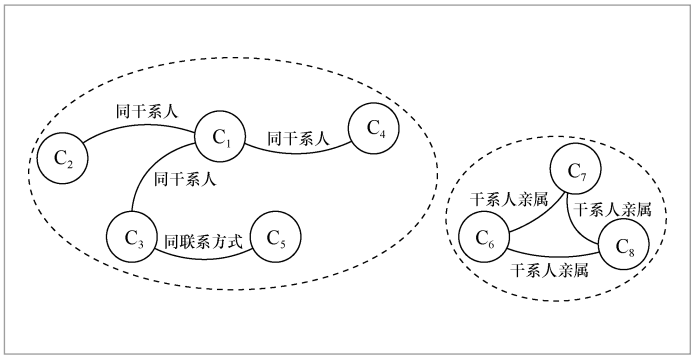


图 2 企业关系图谱子图

3.3.3 节点属性特征构建

企业节点的属性特征来自于抽取的企业信息与干系人信息。由于企业相关信息涉及的数据范围对企业风险的描述有限，为构造全面的企业画像，同时为每个企业匹配实控人及其他干系人的相关信息，生成基于企业的多维特征，增强总体数据的表征能力。每个企业具有唯一实控人和多个其他干系人，而实控人与企业的关联程度与其他干系人相比更为密切。本方案在处理这些特征时，单独将企业实控人的信息与企业信息拼接，将企业其他实控人的信息进行聚合后对企业特征进一步扩展。在对企业其他干系人进行聚合时，对不同变量选用的聚合函数有最大值、加和、中位数、均值、众数等。

最终处理过的实控人特征与干系人聚合特征被关联到企业上。通过异常值、缺失值、变量之间相关性的分析对特征进行

筛选，将剩余的60余个变量作为企业节点的属性特征。

3.3.4 网络特征构建

网络中存在的关系数据为提高分类模型的效果提供了帮助。如果两个实体相关，则一个实体的某些内容也可以帮助推断另一个对象[20]。具体到反欺诈的应用场景则可以解释为，如果一个企业运营中有违法操作，那么与该企业相关的企业同样存在违法风险。为充分利用关系信息，网络特征的提取规则为计算每个企业关联的一度邻居企业中欺诈标签的占比。

3.4 图数据的存储和维护

原始数据经过预处理和图关系的构建，最终获得的图包含约14万个节点、20万条边。每个节点代表一个小微企业，有60种节点属性。图数据被存储在Neo4j图数据库中，这样能够快速直观地对图数据进行可视化分析和展示。在应用阶段，利用基于Python的igraph或NetworkX等工具包加载图数据，能够方便、快捷地实现模型训练和服务部署。

与中大型企业不同，小微企业的工商信息、经营情况变化快且情况复杂，小微信贷业务也因其灵活性而迭代频繁。为了让模型拥有良好的时效性，对图数据进行定期更新和维护是很有必要的。在本文方案中，对图数据的维护分为两个模块。

- 图中节点属性信息的变更：每月进行小微企业工商信息的变更检查，从而对节点的属性进行更新。对由于节点属性变更而产生的边增加或减少情况进行更新。
- 新入节点的预处理与载入：每日对新产生的小微贷款业务的数据进行预处理、节点属性构建以及边的构建，并存入当前图数据库。

得益于本文方案采用的LightGBM算法较低的时空复杂度，每日新入节点更新完成后，能够实现当日快速模型迭代，从而保证模型的时效性，提高小微贷款业务审批流程的效率。

3.5 算法选择

本方案整体将贷款反欺诈问题定义为对欺诈的二分类问题。以企业为粒度，以融合干系人信息的企业多维变量为企业自身的属性特征，以基于关系图谱提取的邻居节点欺诈比例为企业的网络特征，预测企业最终的欺诈标签类型，同时输出欺诈概率。

大部分的机器学习模型，特别是金融行业普遍应用的逻辑回归模型需要较长的训练时间，难以应对变化多端的欺诈行为。除此之外，欺诈具有多样性和复杂性，即使一些征信类数据具有较强的指示作用，但其在总体样本上的高缺失率使得数据预处理的难度加大，最终能够提供的信息增益微乎其微。一些与存款相关的数据之间有较高的线性关系，这违背了传统回归模型不存在多重共线性的假设，使得模型评估失真。

真实的反欺诈业务场景更关注在保证泛化性、稳定性的前提下，模型可取得的最佳效果。为了能有效地利用这些大量类似的碎片化信息，本方案选用LightGBM算法进行建模。LightGBM于2016年被微软亚洲研究院团队在GitHub上开源，是一种基于梯度提升树的算法框架[21]。该算法采用基于直方图的决策树算法与带深度限制的叶子生长（leaf-wise）策略，在预测精度与XGBoost相似的情况下，可大幅提升训练速度，降低内存占用，更好地满足大规

模特征数据处理运算与银行对于小微企业贷款“秒申、秒审和秒贷”的业务需求。对比其他分类模型，LightGBM能够直接读取类别特征，无须进行独热编码处理，避免构建稀疏的特征空间。其对于缺省值的自动填充也极大地简化了数据前期预处理和后续的建模流程。

4 实验及结果

实验阶段将分别建立两种分类模型。模型1为基础模型，单纯使用节点属性，不考虑网络结构特征。模型2在使用节点属性的基础上，引入基于关联图谱提取的节点邻居欺诈比例这一额外特征共同训练。两种模型均使用LightGBM算法框架在相同训练集上进行模型训练，并在相同测试集上进行评估，旨在探究融合网络结构特征后的欺诈模型是否对单独的节点自身属性欺诈模型的预测效果有所提升。

4.1 实验设定

4.1.1 数据集划分

实验将所有存在申请贷款行为的140 885家企业样本作为总数据集，整体数据集按照4:1的比例划分为训练集和测试集，训练集用于模型训练，测试集用于模型验证。其中，正样本（欺诈用户）共计389个，负样本（正常用户）共计140 496个。正负样本比例约2.7‰，数据集存在不平衡现象。划分后的数据集中，正负样本的比例在训练集与测试集中保持一致，整体样本分布见**表2**。

4.1.2 网络特征处理

直接运用以上划分方式对数据集进行切割并单纯使用节点属性建模，是符合逻辑的，但当加入图谱网络的特征时会出现一些问题。其中与传统机器学习模型中经常做出的假设相违背的一点是数据特征的非独立同分布（independently identically distribution，IID）特性。在关系图谱中，节点之间的相关行为意味着一个节点的标签信息可能会影响其相关节点的标签信息。因此在对每个节点进行网络特征提取时，模型训练集和测试集的划分势必会因为整个网络的互联特性而受影响，难以实现完全独立的分割。即使本文假定所构建图谱中节点间的关系是全感知的，且在一定时间范围内不会发生变化，图谱中的每个节点的标签在真实应用场景中也不一定都是已知的。因此，本文在划分完训练集与测试集后对节点周围邻居欺诈比例这一特征进行计算时，如果某节点的邻居来自训练集，那么其标签被设为已知，直接加入特征计算过程中；如果来自测试集，那么标签被设为未知（NA），以此避免关系型特征带来的干扰。

4.2 模型训练与参数选择

为保证最优模型效果，两组实验在模型训练时均采用贝叶斯自动化调参[22]对模

表 2 数据集划分

数据集	样本数量/个	正样本/个	负样本/个	正负样本比
训练集	112 708	312	112 396	2.7‰
测试集	28 177	77	28 100	2.7‰

型参数进行调优。以优化受试者工作特征曲线下与坐标轴围成的面积（AUC）为目标函数，对训练集中的数据进行30轮训练，每轮采用5折交叉验证，并输出平均AUC作为当前模型的表现。最终输出AUC最大时选用的参数组合及训练出的模型。

4.3 评价指标

反欺诈的业务场景中数据的不平衡程度很高，因此应选择对数据倾斜敏感度较低的指标对模型表现进行客观评价。本文选择AUC以及KS值对模型效果进行评价。AUC与KS值均基于真正率（TPR）和假正率（FPR）计算得出。AUC值表示用不同阈值下TPR与FPR连成的ROC曲线下方的面积。AUC值越高，模型对于正负样本的区分能力越强，效果越好。KS值表示TPR和FPR的最大差值，反映的是模型的最优区分效果，KS值对应的阈值可作为区分好坏用户的最优阈值。

4.4 实验结果

第一组实验将企业节点的自身属性输入模型并进行训练，全部特征数目为60，涵盖了筛选过后的企业自身属性、企业唯一实控人属性以及企业其他干系人属性。第二组实验除了用到以上60个全部特征，还加入了当前企业关联周围邻居节点欺诈申请的比例，共计61维。比较两个模型在测试集上的表现，模型效果对比如**图3**所示，模型性能表现见**表3**。

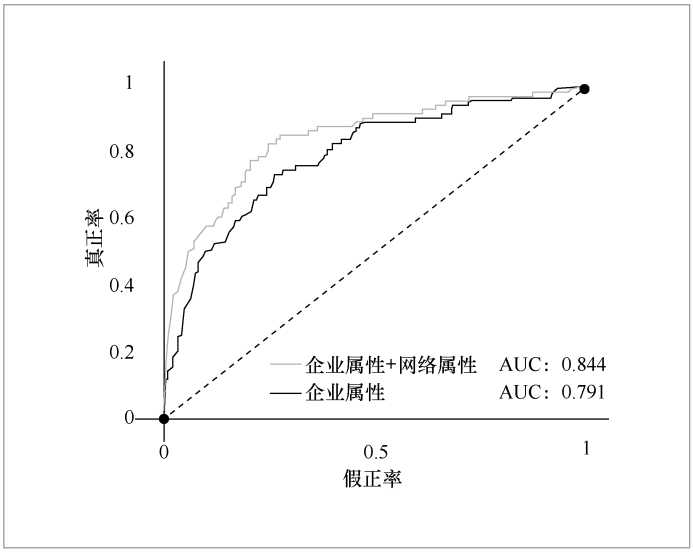


图3 模型效果对比

灰色和黑色两条曲线分别代表是否加入网络特征的模型的表现。其中引入网络特征的模型（AUC为0.844）对欺诈样本的区分能力明显好于仅使用节点属性的模型（AUC为0.791）。其在测试集上的AUC值提升了6.7%，KS值提升了24.7%。

除了模型表现的比较，模型对结果的可解释性同样确保了整个银行贷款业务的风险可控。由于LightGBM本质上采用基于决策树算法的梯度提升框架，在模型训练的过程中会根据特征的信息增益来分裂叶子节点。因此，可将其作为特征重要性信息，以评价不同特征对欺诈风险预测的贡献程度。**图4**列出了模型2中重要性排名靠前的特征。可以发现，企业邻居的欺诈比例对模型判断企业欺诈风险的影响最大，企业地理位置、企

表3 模型性能表现

模型	特征	测试集AUC值	测试集KS值
模型1	节点属性（60维）	0.791	0.466
模型2	节点属性+邻居特征（61维）	0.844	0.581

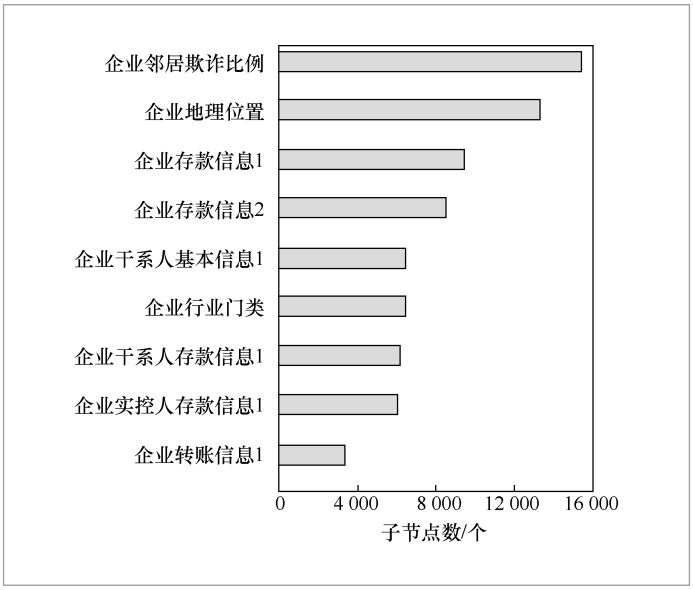


图 4 特征重要性排序

业存款信息位列其后。同时，在排名靠前的特征中出现了企业实控人与企业干系人的相关属性，证明了建立完整的企业画像在全面考量小微企业申贷风险时的重要性。

5 结束语

本文针对当前小微企业信贷反欺诈的痛点，提出了一种基于企业画像与企业关系图谱的有监督学习解决方案，实现从孤立的企业实体欺诈风险到全局网络风险的把控，为反欺诈实践提供了新思路。通过对比传统的信贷反欺诈风控方法，可以明显地看到本文方法为孤立的企业画像引入网络特征的优势，也能够更准确地刻画申贷小微企业与欺诈目标之间的非线性关系。在当前普惠金融的大市场环境下，该方法可在一定程度上解决小微企业贷前反欺诈的风控难题，帮助银行提供更加高效、可靠的金融服务。

参考文献:

[1] 张潇飞. 商业银行小微企业贷款信用风险研究[J]. 智富时代，2015(2).
ZHANG X F. Research on credit risk of small and micro enterprise loans in commercial banks[J]. The Fortune Times, 2015(2).

[2] 陈隆，闫真宇，邓舒仁. 对当前小微企业融资问题的若干思考[J]. 浙江金融，2018(1): 17-23.
CHEN L, YAN Z Y, DENG S R. Reflections on current financing problems of small and micro enterprises[J]. Zhejiang Finance, 2018(1): 17-23.

[3] 孙自通. 小微企业信贷业务流程与法律实务[M]. 北京：中华工商联合出版社，2017.
SUN Z T. Small and micro enterprise credit business process and legal practice[M]. Beijing: All-China Federation of Industry and Commerce, 2017.

[4] 于沛丰. 大数据金融破解小微企业融资难的分析[J]. 全国流通经济，2018，2189(29): 88-89.
YU P F. Analysis of big data finance cracking the financing difficulties of small and micro enterprises[J]. China Circulation Economy, 2018, 2189(29): 88-89.

[5] 陈平，王晓婷，黄一朕. 商业银行企业级反欺诈实践与趋势[J]. 中国银行业，2017(11): 23-25.
CHEN P, WANG X T, HUANG Y Z. Commercial bank's enterprise-level anti-fraud practice and trend[J]. China Banking, 2017(11): 23-25.

[6] 丁濛濛. 基于规则引擎的互联网金融反欺诈研究[J]. 电脑知识与技术，2018，14(1): 1-3.
DING M M. Internet finance anti-fraud research based on rule engine[J]. Computer Knowledge and Technology, 2018, 14(1): 1-3.

[7] 仵伟强，后其林. 基于机器学习模型的消费金融反欺诈模型与方法[J]. 现代管理科学，2018(10): 51-54.
WU W Q, HOU Q L. Consumer finance

anti-fraud model and method based on machine learning model[J]. Modern Management Science, 2018(10): 51-54.

[8] 何湘东, 魏吉勇. B2B平台的反欺诈问题研究[J]. 信息安全与技术, 2016, 7(11): 47-51.
HE X D, WEI J Y. Research on the B2B platform anti-fraud problem[J]. Information Security and Technology, 2016, 7(11): 47-51.

[9] 李苏, 周小惠, 宝哲. 基于支持向量机的商业银行对中小信贷企业选择方法的研究[J]. 数学的实践与认识, 2018(11): 299-305.
LI S, ZHOU X H, BAO Z. Research of loan enterprise selection for bank based on support vector machine[J]. Mathematics in Practice and Theory, 2018(11): 299-305.

[10] 张杰, 赵峰. 基于支持向量机的中小企业技术信贷违约预测[J]. 统计与决策, 2013(20): 66-69.
ZHANG J, ZHAO F. SME technology credit default forecast based on support vector machine[J]. Statistics and Decision, 2013(20): 66-69.

[11] 邱耀, 杨国为. 基于XGBoost算法的用户行为预测与风险分析[J]. 工业控制计算机, 2018, 31(9): 47-48.
QIU Y, YANG G W. User behavior prediction and risk analysis based on XGBoost algorithmv[J]. Industrial Control Computer, 2018, 31(9): 47-48.

[12] 陈俊清. 神经网络模型在互联网金融反欺诈领域的研究与实践[J]. 中国金融电脑, 2016(8): 42-46.
CHEN J Q. Research and practice of neural network model in the field of internet financial anti-fraud[J]. Financial Computer of China, 2016(8): 42-46.

[13] 胡鹏飞. 金融科技在互联网金融行业性风险防范领域的应用[J]. 大数据, 2018, 4(1): 117-123.
HU P F. Application of FinTech in internet financial industry risk prevention[J]. Big Data Research, 2018, 4(1): 117-123.

[14] 樊盛博. 金融社交网络在伪卡欺诈发现中的应用研究[J]. 中国金融电脑, 2017(3): 65-71.
FAN S B. Research on the application of financial social network in the detection of pseudo-card fraud[J]. Financial Computer of China, 2017(3): 65-71.

[15] BEUTEL A, AKOGLU L, FALOUTSOS C. Fraud detection through graph-based user behavior modeling[C]//The 22nd ACM SIGSAC Conference on Computer and Communications Security, October 12-16, 2015, Denver, USA. New York: ACM Press, 2015: 1696-1697.

[16] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2014, New York, USA. New York: ACM Press, 2014: 701-710.

[17] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[C]//The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016, San Francisco, USA. New York: ACM Press, 2016: 855-864.

[18] TANG J, QU M, WANG M, et al. LINE: large-scale information network embedding[C]//The 24th International Conference on World Wide Web, May 18-22, 2015, Florence, Italy. [S.l.:s.n.], 2015: 1067-1077.

[19] DONG Y, CHAWLA N V, SWAMI A, et al. Metapath2vec: scalable representation learning for heterogeneous networks[C]//The 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 13-17, 2017, Halifax, Canada. New York: ACM Press, 2017: 135-144.

[20] NEVILLE J. Iterative classification[M]. Heidelberg: Springer, 2000.

[21] KE G, MENG Q, FINLEY T, et al. Lightgbm: a highly efficient gradient boosting decision tree[C]//The 31st Conference on Neural Information Processing Systems, December 4-9,

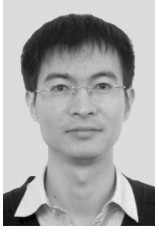
2017, Long Beach, USA. [S.l.:s.n.], 2017: 3146-3154.

[22] SNOEK J, LAROCHELLE H, ADAMS R P. Practical bayesian optimization of machine learning algorithms[C]// The 25th International Conference on Neural Information Processing Systems, December 3-6, 2012, Lake Tahoe, USA. Miami: Curran Associates Inc., 2012: 2951-2959.

作者简介

金磐石(1965-),男,中国建设银行股份有限公司信息总监,主要研究方向为数据处理领域技术研究与应用、人工智能领域技术研究与应用等。

万光明(1974-),男,中国建设银行金融科技部高级工程师,主要研究方向为应用架构管理、人工智能方向架构管理等。



沈丽忠(1978-),男,建信金融科技有限责任公司厦门事业群大数据平台架构师,主要研究方向为数据挖掘、分布式存储、分布式计算、流计算、数据分析等。

收稿日期:2019-05-02