计算机科学
COMPUTER SCIENCE

基于网络表示学习的车险欺诈溯因分析研究

李炜卓, 卢冰洁, 杨骏铭, 那崇宁

引用本文

李炜卓, 卢冰洁, 杨骏铭, 那崇宁.基于网络表示学习的车险欺诈溯因分析研究[J]. 计算机科学, 2023, 50(2): 300-309.

LI Weizhuo, LU Bingjie, YANG Junming, NA Chongning. Study on Abductive Analysis of Auto Insurance Fraud Based on Network Representation Learning [J]. Computer Science, 2023, 50(2): 300-309.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800169

基于网络表示学习的车险欺诈溯因分析研究

李炜卓[1,3,4]　卢冰洁[2]　杨骏铭[1]　那崇宁[2]
1 南京邮电大学现代邮政学院　南京 210003
2 之江实验室金融科技研究中心　杭州 311100
3 南京大学计算机软件新技术国家重点实验室　南京 210093
4 东南大学计算机网络和信息集成教育部重点实验室　南京 211189
(liweizhuo@amss.ac.cn)

摘　要　车险欺诈检测对促进汽车保险业的良性健康发展有着重要意义。由于欺诈的判断涉及公民权利等核心内容,需要车险专家对案件进行核查,提供欺诈原因。尽管基于机器学习的方法泛化能力强、精确度高,但缺少可解释性,而基于专家系统的规则方法尽管有较好的可解释性,但受限于规则复杂的触发条件。为了解决未触发专家系统欺诈规则而被机器学习方法检测为“欺诈”的案件无法被解释的问题,文中提出了基于网络表示学习的车险欺诈溯因分析方法。该方法首先定义了车险欺诈溯因分析任务,然后采用网络表示学习对已触发专家系统中欺诈规则的案件进行案件-规则因子网络的建模,学习欺诈规则中因子的分布式向量表示。为了更好地度量“欺诈”案件与专家系统中因子未全部触发规则之间的相似度,该方法基于溯因缺省原理,设计了一种规则因子的加权拼接策略来缓解训练数据不足的问题。实验结果表明,所提方法相较于已有方法在车险欺诈溯因预测任务的3项指标中均能取得更好的效果。

关键词:汽车保险欺诈;网络表示学习;溯因推理;专家系统;可解释性
中图法分类号　TP391

Study on Abductive Analysis of Auto Insurance Fraud Based on Network Representation Learning

LI Weizhuo[1,3,4],LU Bingjie[2],YANG Junming[1] and NA Chongning[2]
1 School of Modern Posts,Nanjing University of Posts and Telecommunications,Nanjing 210003,China
2 Fintech Research Center,Zhejiang Lab,Hangzhou 311100,China
3 State Key Laboratory for Novel Software Technology,Nanjing University,Nanjing 210093,China
4 Key Laboratory of Computer Network and Information Integration(Southeast University),Ministry of Education,Nanjing 211189,China

Abstract　Auto insurance fraud detection plays an important role in promoting the healthy development of auto insurance. As the judgment of fraud involves the core content of civil rights,it is necessary for auto insurance experts to check the case and provide the reasons for fraud. Although the method based on machine learning have strongscalability and high accuracy,it lacks interpretability,while the rule method based on expert system has good interpretability,but it is limited by the trigger conditions of complex rules. To address the unexplainable problem of cases detected as“fraud” by machine learning methods without triggering the expert system fraud rules,this paper puts forward an analysis method of auto insurance fraud traceability based on network representation lear-ning. It first defines the abductive analysis task of auto insurance fraud. That is,for cases that are identified as “fraud ”ones by machine learning methods without triggering the expert system,it returns the ranking of the most likely fraud rules to auto insurance experts. Then,the method models the case-rule factor network based on the network representation learning according to the fraud cases that have triggered the rules of the expert system,and learns the vector representation of these factors in fraud rules. To better measure the similarity between fraud cases and rules with incomplete triggering factors in the expert system,a weighted splicing strategy of factors in fraud rules is designed based on the principle of abductive reasoning,which can alleviate the problem of insufficient training data to some extent. Experimental results show that the proposed method can ob-

到稿日期:2022-08-16　返修日期:2022-11-06
基金项目:国家自然科学基金(62006125);江苏省双创博士项目(JSSCBS20210532);南京邮电大学引进人才科研启动基金(NY220171);之江实验室科研攻关项目(2020NF0AC01,2022NF0AC01)
This work was supported by the National Natural Science Foundation of China(62006125),Foundation of Jiangsu Provincial Double-Innovation Doctor Program(JSSCBS20210532),NUPTSF(NY220171) and Key Research Project of Zhejiang Lab(2020NF0AC01,2022NF0AC01).
通信作者:卢冰洁(lubj@zhejianglab.com)

tain better performances than existing methods in terms of three metrics.

Keywords Auto insurance fraud, Network representation learning, Abductive reasoning, Expert system, Interpretability

1 引言

保险作为四大金融支柱之一，具有分摊风险、补偿损失、融通资金以及管理社会等功能，有利于保障人民的基本生活水平、维护社会稳定、推动国家 GDP 稳定增长[1]。自改革开放以来，我国保险事业取得了长足的发展。根据银保监会 2022 年布的保险业发展报告，截至 2021 年底，我国保险机构数量达 206 家，总资产达 24.9 万亿，成为了全球第二大保险市场。汽车保险作为财产险中的第一大险种，占比高达 69%，对保险行业的稳定发展有着至关重要的作用。然而，近年来，车险欺诈案件数量逐年上升，这使得保险公司的赔付成本也不断攀升。据学者统计，我国车险欺诈渗透率至少为 20%[2]，2021 年我国保险公司在车险欺诈方面的损失则高达 900 亿元以上。这不仅增加了汽车保险公司的运营成本，侵害了保险公司的合法权益，而且对整个社会构成了危害。因此，如何有效检测车险欺诈并对其案件进行分析，对促进汽车保险的良性健康发展有着重要意义。

相比其他保险欺诈，汽车保险欺诈存在犯罪手段隐蔽、手法多样等特点，近年来逐渐呈现团伙化的作案方式[3]，受到了各国监管部门、保险公司、研究机构的广泛关注。目前，车险欺诈问题可以通过两类方法来进行检测分析。一类是以规则推理为主的车险欺诈检测专家系统。它将车险专家的经验形式化为推理规则来进行决策，通过激活判定规则的条件因子来达到欺诈检测的目的[4]。另一类是以机器学习为主的分类技术，它通过对被保险人、保险标的、出险情况等各方面属性取值进行收集提炼，采用相应的标注数据对机器学习模型进行训练，继而针对新案件所对应的属性取值来进行车险欺诈的判别[5]。

尽管如此，上述方法在车险欺诈检测的实际场景中均存在一定的局限性。专家系统的优势在于自身基于规则的因果推理策略使它具备了良好的可解释性。然而，由于专家系统中的每条车险欺诈规则都关联了较多数量的因子，一旦规则中的因子无法全部触发，通过专家系统的规则推理技术来进行车险欺诈检测的实用性就会大打折扣[6]。倘若欺诈方对系统规则有所了解，就能够在犯罪中绕过这些规则的触发条件，从而规避专家系统的检测。相对地，尽管基于机器学习的方法在车险欺诈检测场景中具备泛化能力强与预测准确性高等优点，但它并非像专家系统一样具备良好的可解释性[7]，难以为案件提供欺诈因果检验的调查方向。由于欺诈的判断涉及公民权利等核心内容，因此车险专家往往需要耗费大量的时间和精力来查证此类案件，提供欺诈原因。

综上，本文将机器学习关联推理与专家系统因果推理进行优势互补，提出了一种基于网络表示学习的车险欺诈溯因分析方法。该方法首先定义了车险欺诈溯因分析任务，即针对未触发专家系统规则但被机器学习方法检测为“欺诈”的案件，为这些“欺诈”案件提供最可能的欺诈规则排序，然后基于网络表示学习对已触发专家系统规则的欺诈案件进行案件-规则因子网络的建模，通过提取训练数据中所有车险欺诈案件的属性取值，将它们与专家系统中欺诈规则的触发因子建立映射索引，最终形成案件-规则因子网络，再利用网络表示学习模型进行训练，来学习欺诈规则中因子的分布式向量表示。此外，该方法基于溯因缺省原理，设计了一种规则因子的加权拼接策略，该策略不仅可以更好地度量“欺诈”案件与专家系统中因子未全部触发规则之间的相似度，还可以在一定程度上缓解训练数据不足的问题。在真实的车险数据集上进行实验评估，结果表明所提方法在$Rank_{\min}$，MAP 以及 $p@10$ 这 3 项评估指标上的整体效果均好于已有方法。

2 相关研究

2.1 基于机器学习的车险研究

国内外均有不少学者将机器学习模型应用在车险欺诈检测技术上，并取得了较好的研究成果。譬如，国外学者 Viaene 等[8]、Payam 等[9]、Kašćelan 等[10]、Li 等[11]分别研究了贝叶斯模型、聚类模型、数据挖掘、随机森林等技术在车险欺诈检测领域的效果；He 等[12]、Guo 等[13]、Wang 等[14]则进一步探索了深度学习模型在该任务上的应用价值；Subudhi 等[15]、Majhi 等[16]则从混合模型的角度切入，提出了一种有效的建模方法。在国内，学者 Guo 等[17]、Liu 等[18]最早从车险理论的角度出发，对车险欺诈检测技术进行了探究；Zhao 等[19]、Tang 等[20]、Wang[21]根据国内的车险欺诈的实际情况，应用传统机器学习模型对其进行建模。近年来，Yan 等[22]、Yu 等[23]、Xu 等[24]从深度学习网络和混合模型的角度出发，在车险欺诈检测任务的精度上取得了较大的进展。

此外，针对车险欺诈数据的特征空间庞大且特征之间有着复杂的依赖关系等一系列问题，Panigrahi 等[25]采用了 3 种特征选择算法，提取车险欺诈数据中的重要特征，并利用机器学习算法进行检测，挑选出不同机器学习模型的最佳特征选择方法。另一方面，Hassan 等[26]、Padhi 等[27]分别使用了欠采样、过采样等策略来缓解车险欺诈任务所存在的数据不均衡挑战。

2.2 基于溯因推理的车险研究

溯因推理是从已知结论出发，反向寻找某事例成立的原因，以此对结果进行解释，属于因果推理的分支[28]。人工智能中的溯因推理方法可分成集合覆盖溯因方法、基于逻辑的推理方法以及知识层次的溯因方法 3 种不同形式[29]。在车险相关的研究中，Sun 等[30]对缺省逻辑表示下的溯因框架进行了研究，提出了一种基于规则的命题逻辑缺省溯因诊断的求解方法；Lin 等[31]针对我国商业车险奖惩的业务场景，设计了一类线性约束下的动态转移规则与方法，继而计算出最小化索赔频率真实值与预测值之间的均方误差，最终该方法通过溯因方法求得最优奖惩系数；Huang 等[32]提出了基于粗糙集的续保规则溯因推理模型，该模型可以有效挖掘出汽车保险信息系统在不同简化层次上满足置信度要求的规则；Fan 等[33]基于溯因分析提出了可变精度的粗糙集模型来对车险

高利润客户进行挖掘；Zhu[34]基于溯因推理建立了欺诈预警与奖惩系统挖掘的车险评价指标体系，该体系可以提供更全面、准确的特征来识别保险领域中的欺诈异常行为。

2.3 研究现状分析

上述两类方法仅从机器学习和溯因推理各自的角度进行车险欺诈研究，没有将机器学习模型的强泛化能力、高准确率与溯因推理良好的可解释性进行有效融合。本文则尝试将两类推理技术进行优势互补，提出了一种基于网络表示学习的车险欺诈溯因分析方法，用于车险溯因分析的场景。本文方法通过已触发专家系统规则的欺诈案件数据与专家系统中的欺诈规则因子构建了案件-规则因子网络，采用网络表示学习模型学习得到欺诈规则中因子的分布式向量表示。针对机器学习方法检测为“欺诈”而专家系统中车险欺诈规则无法触发的案件，本文方法可以为车险专家返回这些“欺诈”案件最可能的欺诈规则排序，继而提供可供查证的欺诈原因。

3 基于网络表示学习的车险欺诈溯因方法

该节重点介绍了本文提出的基于网络表示学习的车险欺诈溯因分析方法。首先定义车险欺诈溯因分析任务；然后给出方法的总体技术路径；最后对框架中的各个模块进行了叙述。

3.1 车险欺诈的溯因分析

车险欺诈判定与法律罪名判定类似，都需要对判定的结果给出合理解释。然而，当机器学习应用在车险欺诈检测的过程中时，由于分类器与特征模块主要是基于提供的训练语料自动学习得到的，因此将机器学习应用在车险欺诈检测技术上会存在可解释性缺失的问题。依据 Cong 等对可解释需求的分类[7]，车险欺诈判定涉及当事人实体权利或程序性权利等核心内容。因此，车险专家需对“判断结果”进行核查，并提供详实的欺诈原因。

相对地，专家系统中大量的规则可以较好地为车险专家提供合理的欺诈线索，这在一定程度满足了上述需求。为此，本文借助专家系统中的规则来对机器学习方法识别为“欺诈”且无法完全触发专家系统中欺诈规则因子的案件进行溯因推理，即从案件“欺诈”的结论出发，寻找最可能的欺诈规则，为车险专家提供因果解释[6]。接下来，给出车险欺诈溯因分析任务的定义。

定义 1（车险欺诈溯因分析） 给定专家系统的规则集合 $\mathcal{R}$，车险欺诈溯因分析在于学习到一个目标函数 $f(c,r)$，使得对于每一个车险欺诈案件 c，都能推理计算出 c 与规则库中任意规则 $r\in\mathcal{R}$ 的相似度，并返回相似度最高的前 k 条规则。

可以看出，该任务属于机器推理中的因果推理，即通过数据驱动的方式，找到车险欺诈案件最有可能的规则排序。尽管如此，对于无法完全触发专家系统中欺诈规则因子的案件，该任务只能为车险专家提供最可能欺诈规则的排序，无法进行精准决策。因此，仍需要车险专家根据所提供的规则中的触发因子与规则的类型进行因果检验。

3.2 总体技术路线

图 1 为本文针对车险欺诈溯因分析任务提出的技术流程图。该方法主要由案件-规则因子网络构建模块与车险欺诈案件溯因预测模块组成。案件-规则因子构建模块通过将触发专家系统规则的每个案件的属性取值与专家系统规则库中因子的集合建立映射索引，继而建立每个案件与规则因子的关联，形成案件-规则因子网络。通过网络表示学习模型对构建的网络进行训练，得到网络中案件与规则因子的分布式向量表示。车险欺诈案件溯因预测模块则是将机器学习检测为“欺诈”但无法触发专家系统欺诈规则的案件根据建立的映射索引进行匹配，得到该类案件在每一条欺诈规则中的已触发的因子，再根据训练得到的规则因子向量表示，结合规则的向量表示策略来对“欺诈”案件与每条规则的语义关联进行计算，最终得到“欺诈”案件中最有可能的欺诈规则排序，从而为车险专家提供案件核查的思路。

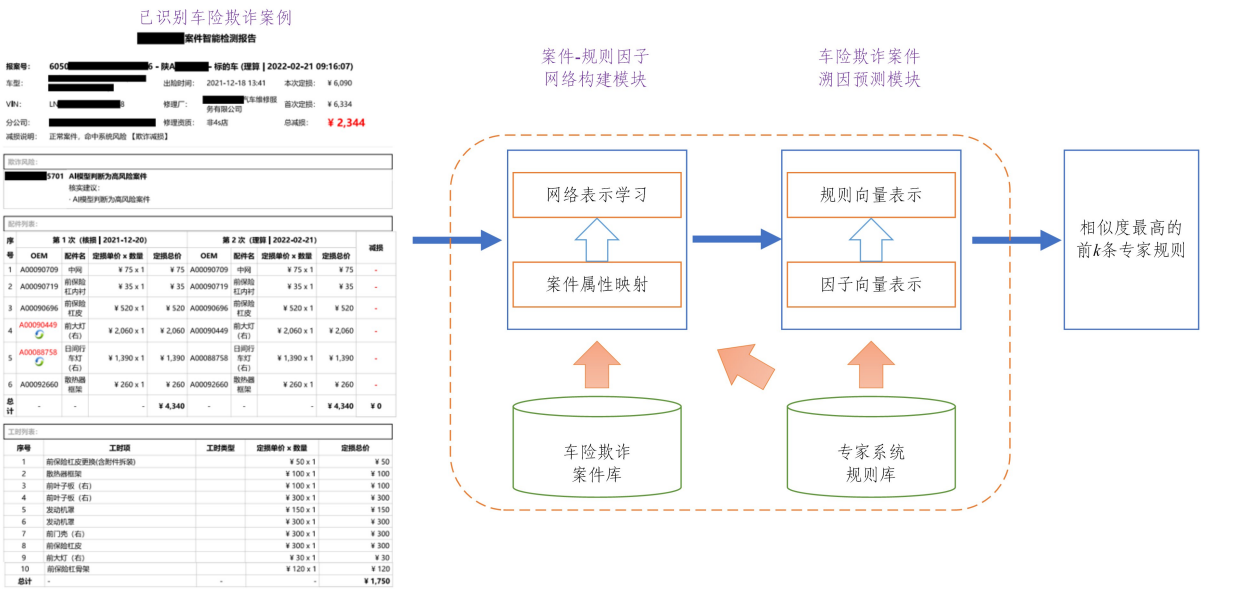


图 1 车险欺诈溯因分析的技术流程图

Fig. 1 Technical flow chart of auto insurance fraud of abductive analysis

值得注意的是，在案件-规则因子构建模块中建立的案件-规则因子网络是由车险欺诈案件与触发的规则因子构建而成，并非由原始案件与其自身的属性值构建。因此，案件-规则因子网络可以通过案件关联的欺诈规则的类型与触发的因子为案件提供因果解释，达到溯因推理的效果，而后者无法直接做到。

3.3 案件-规则因子网络的构建

网络的表示方式可以有效反映数据之间的关联，并成为衔接新任务数据的桥梁。

定义 2（网络） 一个网络记为 $G=(V,E)$，其中 V 是网络节点集合，E 是网络中边的集合，边 $e=(v_i,v_j)\in E$ 表示节点 v_i 到节点 v_j 的一条边。

通常，一个网络的关系可以用邻接矩阵 $\boldsymbol{A}\in\boldsymbol{R}^{|V|\times|V|}$ 来表示。然而，由于邻接矩阵 $\boldsymbol{A}$ 需要占用 $|V|\times|V|$ 的存储空间，因此当 $|V|$ 增长到 10 万甚至 100 万级时，计算的代价通常是不可接受的。另一方面，由于矩阵 $\boldsymbol{A}$ 中大部分节点之间并无关联，因此邻接矩阵的稀疏性使得传统的机器学习方法对其进行快速有效的应对变得十分困难。为此，研究者们提出了基于网络表示学习的方法[35-37]，该方法尝试将网络中的节点编码为低维稠密的实值向量 $\boldsymbol{v}\in\boldsymbol{R}^d$，其中 $d\ll|V|$，在模型编码的过程中，通过拟定的目标函数将网络的结构信息与实体之间的语义关联进行有效的保留。

受到网络表示学习的启发，本文提取了所有车险欺诈案件每个属性的取值，并将它们与专家系统中规则的触发因子建立映射索引。由于欺诈案件中每个属性取值可能是数值或类别，因此在建立属性与规则因子的映射索引时，需要对属性进行归一化处理，以便将其映射至规则因子中进行二值判断。为了清晰地体现欺诈案件与规则因子之间的关联，本文将构建的案件-规则因子网络中的节点分为上下两层，上层节点用车险欺诈案件的 ID 来表示其唯一性，下层节点为欺诈规则所对应的因子。若欺诈案件所关联的属性值与触发规则的因子之间存在映射关系（这里的映射关系只考虑案件属性的取值是否会触发某条规则因子的情况，即将属性取值代入规则因子后，输出为“真”或“假”），则在案件节点与规则因子节点之间添加一条边。构建案件-规则因子网络样例如图 2 所示，对于案件“酒驾换驾案件 99”中数值型属性“出险时间”的取值（如：2017/5/11 23:08:00），映射为网络中某条规则因子中“出险时间为 21 点至 5 点”。由于满足映射关系，需在案件与规则因子之间添加一条边。对于类别型属性“车辆使用性质”的取值（如：家庭自用汽车），映射后则为网络中某条规则因子中“标的不是运营车辆”。类似地，由于属性取值与规则因子之间满足映射关系，因此在案件与规则因子之间添加一条边。最终，通过建立所有欺诈案件与规则之间的关联形成案件-规则因子网络。通过网络表示学习模型可以学习到图 2 网络中所有节点（如：非人伤）的分布式向量表示（如：$\boldsymbol{v}_{非人伤}=(0.64,0.36,\cdots,0.67)\in\boldsymbol{R}^d$），其中维度 d 的取值远小于网络中的节点数量。接下来，本文以网络表示学习模型 LINE[38] 为例进行网络建模介绍。

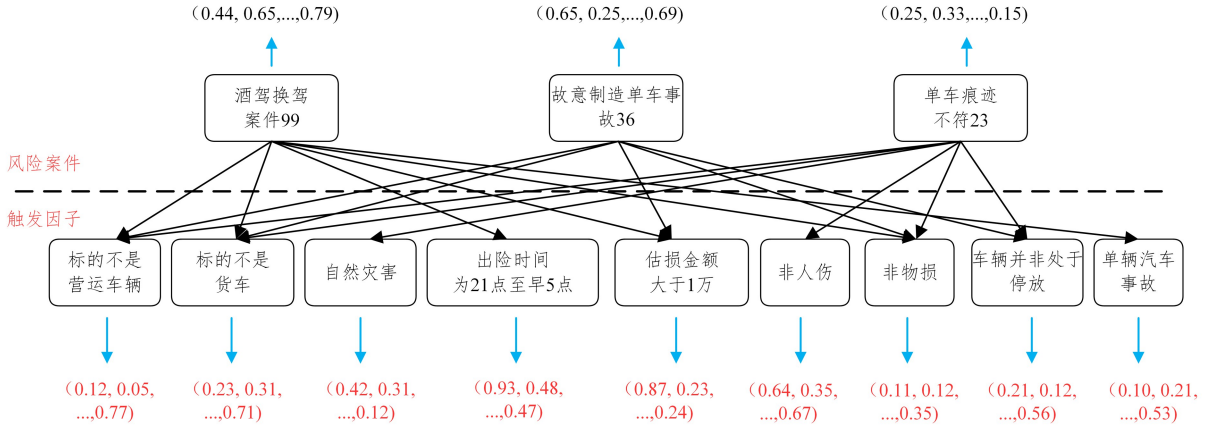


图 2 案件-规则因子网络样例

Fig. 2 Example of case-rule factor network

由于实际业务中，车险欺诈案件之间的邻接关系未知，因此本文方法直接采用了 LINE 的二阶邻接关系来建模。在二阶邻接关系中，LINE 模型假设两个节点（车险欺诈案件）共享连接的规则因子越多，那么这两个节点彼此越相似。这些共享的节点从溯因推理的原理可以被视为特定的“上下文”（“上下文”指以该节点为起点，经过两条边所形成的关联节点集合），并且模型进一步假设在“上下文”中具有相似分布的节点通常也是相似的。为了建模二阶邻接关系，模型 LINE 要求每一个低维向量表示的节点所指定的上下文的条件分布 $p_2(\cdot|v_i)$ 接近经验分布 $\hat{p}_2(\cdot|v_i)$。相应的目标函数定义如下：

$$O_2=\sum_{i\in V}\lambda_i d(\hat{p}_2(\cdot|v_i),p_2(\cdot|v_i)) \tag{1}$$

其中，O_2 表示 LINE 二阶邻接关系的损失函数，v_i 表示索引为 i 的节点，V 表示网络所有的节点集合。$\lambda_i=\sum_{k\in N(i)}w_{ik}$ 为前节点的声誉值，其中 w_{ik} 为节点 v_i 出度的权重，$p_2(\cdot|v_i)$ 表示节点的条件分布，$\hat{p}_2(\cdot|v_i)$ 表示节点的经验分布，$d(\cdot,\cdot)$ 表示两个分布的距离，即 KL 散度。LINE 通过最大化这个联合概率来更新节点的向量表示。模型计算序列中每个节点的条件概率，即该节点出现的情况下序列中其他节点出现的概率的 log 值，约简一些无关常数，如式(2)所示：

$$O_2=\sum_{(i,j)\in E}w_{ij}\log p_2(v_j|v_i)) \tag{2}$$

由于整体优化式(2)的代价太高,为此,本文方法采用了LINE提出的负采样策略,根据每条边$(u_i,u_j)\in E$的噪声分布进行负采样。具体的目标函数如式(3)所示:

$$\log\sigma(\boldsymbol{u}_j^{\mathrm{T}}\cdot\boldsymbol{u}_i)+\sum_{i=1}^{K}E_{v_n}\sim P_n(v)[\log\sigma(-\boldsymbol{u}_n^{\mathrm{T}}\cdot\boldsymbol{u}_i)] \tag{3}$$

其中,$\sigma=1/(1+\exp(-x))$为sigmoid函数,$\boldsymbol{u}_i$,$\boldsymbol{u}_j$为边(u_i,u_j)中节点的向量表示,$P_n(v)\propto d_v^{3/4}$为模型定义的负采样分布。简单来说,对于每条边$(u_i,u_j)\in E$,在节点集合V中按照负采样分布寻找其他的节点进行替换,生成模型负例(u_i,u_j')或者(u_i',u_j)进行学习。

此外,考虑到欺诈案件中属性取值类型上存在的差异性,如:数值型与类别型,在网络表示学习的负采样过程中,当某种类型属性所映射的规则因子作为节点时,本文方法会尽可能排除该类型节点作为负例进行替换的情况。

3.4 车险欺诈案件溯因预测

车险欺诈案件溯因预测的核心挑战在于对未触发专家系统规则的"欺诈"案件,如何基于学习得到规则因子向量,将这些"欺诈"案件与专家系统中存在的欺诈规则进行关联计算,得到最可能的欺诈规则排序。因此,对欺诈规则与"欺诈"案件的向量表示进行建模十分重要。

目前,欺诈规则向量表示的常用做法是将每条规则所有的因子向量进行算术平均来获得该规则的语义向量表示,相关的定义如下:

$$r_i=\frac{1}{m}\sum_{k=1}^{m}f_k \tag{4}$$

其中,$r_i\in\mathcal{R}$为专家系统中的第i条规则,m为每个规则所关联的因子个数,$\boldsymbol{f}_k\in\mathbb{R}^d$为基于网络表示学习模型训练后学习得到的因子向量表示,d为因子向量的维度。相对地,对于"欺诈"案件的向量表示,它仍可用式(4)进行建模计算,只是需要将案件中未触发的规则因子f_k设置为零向量,即$\boldsymbol{f}_k=(0.0,0.0,\cdots,0.0)$。

尽管算术平均的策略可以有效解决案件中规则因子未全部触发时不同规则与"欺诈"案件关联计算时向量维度不统一的问题,但该策略仍存在一定的局限性。一方面,在车险欺诈案件触发规则的真实场景中,由于案件触发的规则较少而规则对应的因子数量较多,导致构建的网络较为稀疏。如果仅用训练得到的因子向量进行算术平均来获得欺诈规则的向量表示,则可能无法对规则进行有效的区分。另一方面,由于部分欺诈规则所对应的欺诈案件类型(如:酒驾、故意制造交通事故)可能存在训练数据不均衡的情况,对于真实场景中未触发专家系统规则的"欺诈"案件,直接将"欺诈"案件未触发的规则因子$\boldsymbol{f}_k$设置为零向量来进行案件的向量表示,这样则会导致算术平均之后的"欺诈"案件向量表示过于平滑,将它们与欺诈规则的向量进行语义关联计算时则无法突显出欺诈案件类型上的差异。

为了解决上述问题,本文设计了一种规则因子加权拼接策略。首先,对于未触发专家系统规则的"欺诈"案件与专家系统所有的欺诈规则,均采用因子拼接策略来替代算术平均策略,这样可以有效地缓解规则中未触发因子过多所导致算术平均后的规则无法区分的问题。其次,考虑到训练数据中部分欺诈规则所对应的欺诈案件类型较少,策略进一步采用因子的TF-IDF数值对"欺诈"案件已触发的规则因子向量进行加权。规则向量$\boldsymbol{r}_i$的表达式如式(5)所示:

$$\boldsymbol{r}_i=\boldsymbol{f}_1\oplus\boldsymbol{f}_k\oplus\cdots\oplus\boldsymbol{f}_m,k\in1,\cdots,m \tag{5}$$

其中,$\boldsymbol{r}_i\in\mathcal{R}$为专家系统的第$i$条规则,$\oplus$为拼接操作$\mathbb{R}^{a\times d}\oplus\mathbb{R}^{b\times d}\rightarrow\mathbb{R}^{(a+b)\times d}$,用来连接规则因子的向量表示,$\boldsymbol{f}_1,\boldsymbol{f}_k,\boldsymbol{f}_m\in\mathbb{R}^d$表示维度为$d$的向量,$m$为每个规则所关联的因子个数。

类似地,假设车险欺诈案件c中对应的触发因子集合与规则r_i的因子交集为$\mathcal{F}'$,那么该案件c相对于规则$\boldsymbol{r}_i$的向量表示$\boldsymbol{r}_c'$如式(6)所示:

$$\boldsymbol{r}_c'=\omega_1\boldsymbol{f}_{c1}\oplus\omega_k\boldsymbol{f}_{ck}\oplus\cdots\oplus\omega_m\boldsymbol{f}_{cm},k\in1,\cdots,m \tag{6}$$

其中,$\oplus$为拼接操作,ω_k为因子在所有规则中的TF-IDF权重,$\boldsymbol{f}_{c1},\boldsymbol{f}_{ck},\boldsymbol{f}_{cm}\in\mathbb{R}^d$为因子交集$\mathcal{F}'$中维度为$d$的因子向量表示。由于规则$\boldsymbol{r}_i$中部分因子未被案件的属性值所激活,因此仍需要用维度为d的零向量$\mathbf{0}=(0.0,0.0,\cdots,0.0)$进行替换。

具体的基于网络表示学习的车险欺诈溯因算法如算法1所示。给定当前的车险欺诈案件c,专家系统的欺诈规则集合$\mathcal{R}$,所有规则的因子集合$\mathcal{F}$以及基于网络表示学习模型$\mathcal{M}$训练后的规则因子向量表示$\mathbb{F}$。首先,通过步骤1、步骤2对案件进行预处理,获得案件c的属性值集合$\{v_1,v_2,\cdots,v_n\}$,将车险欺诈案件c的所有属性值与规则因子集合$\mathcal{F}$建立映射索引,得到案件c对应的因子集合$\mathcal{F}_c=\{f_{c1},f_{c2},\cdots,f_{cl}\}$。接下来,步骤3—步骤13计算案件$c$与规则集合$\mathcal{R}$中的每一条规则的相似度。具体来说,对规则集合$\mathcal{R}$中的每一条规则$\boldsymbol{r}_i$,得到其规则因子集合与这些因子相应的向量表示,再通过拼接策略获得规则的向量表示$\boldsymbol{r}_i$(见步骤4—步骤5)。步骤6—步骤12将案件的因子集合与规则的因子集合进行交集运算,从而得到案件相对于规则的向量表示$\boldsymbol{r}_c$,其中$\boldsymbol{r}_c$的向量表示维度与规则向量表示维度保持一致。若案件对应的因子在规则中被激活,则因子$\boldsymbol{f}_{ck}$为规则的原始向量表示$\boldsymbol{f}_k$。反之,则用零向量$\mathbf{0}$进行替代。类似地,欺诈案件的向量表示也采用因子向量拼接的策略获得,两者的区别在于融入了案件因子在所有规则中的TF-IDF权重ω_k。步骤13通过cosine余弦计算每条规则向量$\boldsymbol{r}_i$与案件向量$\boldsymbol{r}_c$的关联相似度。最终,对所有相似度排序后,返回相似度最高的k条欺诈规则,即完成了车险欺诈案件的溯因分析任务。

算法1 基于网络表示学习的车险欺诈溯因算法

输入:车险欺诈案件c,专家系统的欺诈规则集合$\mathcal{R}$,所有规则的因子集合$\mathcal{F}=\{f_1,f_2,\cdots,f_n\}$,基于网络表示模型$\mathcal{M}$训练后的所有规则因子向量表示集合$\{\mathbf{f}_1,\mathbf{f}_2,\cdots,\mathbf{f}_n\}$

输出:前k条与案件关联的规则

1. 获得案件c的属性值集合$\{v_1,v_2,\cdots,v_n\}$/*案件c属性值预处理*/
2. $\mathcal{F}_c=\{f_{c1},f_{c2},\cdots,f_{cl}\}\xleftarrow{\mathcal{F}}\{v_1,v_2,\cdots,v_n\}$ /*将案件c的属性值与规则因子库$\mathcal{F}$建立映射索引,得到案件对应的因子集合*/
3. for each $r_i\in\mathcal{R}$
4. $\mathcal{F}_r\leftarrow\{f_1,f_2,\cdots,f_m\}$/*取得规则的因子集合*/
5. $\mathbf{r}_i\leftarrow\mathbf{f}_1\oplus\mathbf{f}_k\oplus\cdots\oplus\mathbf{f}_m$/*形成规则的向量表示*/
6. $\mathcal{F}'\leftarrow\mathcal{F}_r\cap\mathcal{F}_c$/*取得规则因子与案件因子的交集*/
7. for k←1 to m/*从每条规则的因子角度,判断案件是否激活了该因子*/

8. if $f_k \in \mathcal{F}'$ /*若规则的因子被激活*/
9. $\mathbf{f}_{ck}=\mathbf{f}_k$/*案件因子的向量表示则为原规则因子的向量表示*/
10. else
11. $\mathbf{f}_{ck}=0$/*因子未被激活则用零向量替代*/
12. $\mathbf{r}_c=\omega_1\mathbf{f}_{c1}\oplus\omega_k\mathbf{f}_{ck}\oplus\cdots\oplus\omega_m\mathbf{f}_{cm}$/*形成案件的向量*/
13. $\mathrm{Sim}(\mathbf{r}_i,\mathbf{r}_d)\leftarrow\cos(\mathbf{r}_i,\mathbf{r}_c)$
14. Return 规则集合 $\mathcal{R}$ 中前 k 条与案件 c 相似度最高的欺诈规则。

可以注意到，算法中学习得到所有因子的分布式向量表示也可以用其他表示学习模型（如图表示学习[39]、知识图谱表示学习[40]）进行建模，该思路将作为今后研究的探索方向。

4 实验分析

4.1 数据集与评估标准

本文的训练数据集共选取真实业务场景下某车险公司的采样数据[5]，时间跨度为2014年3月－2019年8月。其中训练数据集为触发了专家系统欺诈规则的5 706个案例，包括酒驾、痕迹不符、故意制造交通事故等欺诈类型。相应触发规则为183条，具体的统计如表1所列。

表1 训练数据集的统计情况

Table 1 Statistics of training datasets

	酒驾	痕迹不符	故意制造交通事故	其他类型
车险欺诈案件数量	116	574	54	4962
专家库规则数	5	84	38	56

基于上述训练数据集，本文提取了所有车险欺诈案件中每个属性的取值，并将它们与专家系统中欺诈规则的因子建立映射索引。在构建的案件-规则因子网络中，将欺诈案件的ID与规则库中所有的因子作为节点。如果欺诈案件所关联的属性值与触发规则的因子存在映射关系，则在案件节点与规则因子节点之间添加一条边。最终，构建网络中节点的数量为5977、边的数量为458733。

为了验证本文方法的有效性，本文共收集了酒驾(DUI)、痕迹不符(TD)、故意制造交通事故(ITA)3种不同欺诈类型的数据集进行测试。该些欺诈案件虽然被车险专家判定为欺诈，但无法触发专家系统中的欺诈规则，满足案件溯因预测设定的测试条件，具体测试集的统计情况如表2所列。

表2 测试数据集的统计情况

Table 2 Statistics of test datasets

数据集	数量	平均属性个数	平均触发规则的因子数
酒驾数据集(DUI)	10	6.20	3.66
痕迹不符数据集(TD)	246	10.25	5.61
故意制造交通事故数据集(ITA)	16	8.77	4.85

考虑到评估过程中，需要先将每个测试模型输出的欺诈规则相似度结果进行降序排序后，才能评估模型在溯因任务中效果，因此实验中采用了搜索排序算法的评估指标来进行度量，分别为$Rank_{\min}$，MAP以及$p@10$这3项指标，相应的指标描述如下。

(1)$Rank_{\min}$：记录规则相似度排序结果中首次出现相关规则的位置，并对所有数据的评估结果求平均，对应的计算公式为：

$$Rank_{\min}=\frac{1}{n}\sum_{i}^{n}\arg\min_{i} rank_{i,j}$$

其中，n为该类欺诈案件测试的总数目，$rank_{i,j}$表示对于第i条欺诈案件，模型溯因预测后该类欺诈案件标签对应规则首次出现的位置为j。$Rank_{\min}$越小，表明模型能更快找到案件原因，因此模型表现的效果越好。

(2)MAP：即平均精度均值。它是搜索排序中常用的评估指标，同时考虑了预测精准度和相对顺序，可以直接衡量多条车险欺诈案件的溯因质量。

$$MAP=\frac{\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\frac{j}{position_{i,j}}}{n}$$

其中，n为该类欺诈案件测试的总数目，m表示规则的总个数，$position_{i,j}$表示模型为第i条欺诈案件溯因预测后，规则依照相似度从高到低排序后，第j条与案件关联规则在排序中的位置。MAP指标越大，表明模型能找到关联规则的位置整体越靠前，因此模型表现的效果越好。

(3)$p@10$：计算规则相似度排序结果中前10条规则中包含的相关规则数量，并对所有数据的评估结果求平均，对应的计算式为：

$$p@10=\frac{\sum_{i=1}^{n}\frac{1}{10}\sum_{i=1}^{10}k_{i,j}}{n}$$

其中，n为该类欺诈案件测试的总数目，$k_{i,j}$表示对于第i条欺诈案件，模型预测出相似度最高的第j条规则与数据d的标签是否关联。若关联，则$k_{i,j}=1$；否则$k_{i,j}=0$。$p@10$越大表明模型能找到关联原因的覆盖面越大，模型表现的效果越好。

4.2 模型测试与结果分析

实验分别选取基于规则因子等价方法（记作FactorEqual）与TF-IDF作为基线测试方法，本文中的网络表示学习模型一共选用DeepWalk[41]，LINE[38]，Node2Vec[42]，GraRep[43]，M-NMF[44]与近期提出的ProNE[45]进行评估测试（本文根据网络表示学习最新综述文献[36-37]收集了所有已开源的模型，发现最新开源的网络表示学习模型大部分只适配于自身构建的数据集，仅有M-NMF[42]，ProNE[43]能够较好地适配于车险欺诈的溯因预测任务）。其中，欺诈案件与规则的向量表示策略默认为因子拼接策略。本文方法主要是在LINE模型的基础上，采用加权拼接策略进行欺诈规则的向量表示，并对其负采样方式进行了改进（记作Our Method）。所有方法均在Linux操作系统上进行评估，编程语言为Python 3.7，测试模型的训练与测试均采用个人工作站来完成，其中CPU为16核的Inter Xeon 2.99 GHz，内存为64 GB。

表3列出了不同测试方法在车险欺诈案件中的对比结果，通过分析可以总结出以下结论。

(1)整体来看，基于网络表示学习的模型DeepWalk，LINE与Node2Vec发挥较为稳定，它们相比基线测试方法FactorEqual和TF-IDF在酒驾(DUI)与故意制造交通事故(ITA)两个测试数据集上均有较大程度的提升。相对而言，GraRep，M-NMF与ProNE在痕迹不符(TD)数据集中有较好的表现，它们在$p@10$上获得了最佳的效果。分析认为，

主要是网络表示模型通过网络建模的方式更容易学到规则因子与欺诈案件之间的潜在语义关联。可以看出，模型DeepWalk，LINE与Node2Vec在酒驾(DUI)、故意制造交通事故(ITA)这种少量样本数据集上就可以达到较好的效果，而GraRep，M-NMF与ProNE则需要更多的训练数据才能使训练模型实现收敛。

(2)TF-IDF方法在痕迹不符(TD)中取得了不错的效果。分析认为，主要是由于该类欺诈案例所对应的规则与因子数量较多，因此TF-IDF方法能够有效区分所有规则中涉及痕迹不符规则因子的重要性。尽管网络表示学习方法整体上在痕迹不符(TD)上与TF-IDF有轻微的差距，然而，模型在$Rank_{\min}$的评估数值均处于2～3之间，因此车险专家根据溯因规则进行核查时，痕迹不符的相关规则仍能够快速出现在车险专家面前。

(3)本文提出的方法在酒驾(DUI)与故意制造交通事故(ITA)两个数据集中，$Rank_{\min}$，MAP以及$p@10$这3项评估指标均取得了最佳的效果，这表明本文采用的网络表示学习模型结合所提出的规则因子加权拼接策略能够进一步对溯因预测后的欺诈规则进行有效的区分，并能获得更好的预测结果。

表3 不同方法在车险欺诈溯因分析任务中的对比结果

Table 3 Result comparison of different methods in abductive analysis task of auto insurance fraud

对比算法	酒驾(DUI)			痕迹不符(TD)			故意制造交通事故(ITA)		
	$Rank_{\min}$	MAP	$p@10$	$Rank_{\min}$	MAP	$p@10$	$Rank_{\min}$	MAP	$p@10$
FactorEqual	29.00	0.079	0.083	2.89	0.491	0.589	3.25	0.409	0.463
TF-IDF	16.50	0.115	0.067	2.23	**0.506**	0.563	1.88	0.374	0.450
DeepWalk	13.00	0.132	0.067	2.99	0.489	0.551	1.88	0.456	0.487
LINE	17.17	0.130	0.083	2.97	0.473	0.495	1.63	0.458	0.600
Node2Vec	13.50	0.120	0.067	2.91	0.489	0.593	1.50	0.457	0.463
GraRep	27.00	0.081	0.083	2.74	0.492	0.554	2.75	0.406	0.425
M-NMF	27.67	0.096	0.083	2.73	0.493	0.597	3.00	0.421	0.475
ProNE	26.83	0.104	0.083	2.74	0.492	**0.610**	2.38	0.434	0.512
Our Method	**6.67**	**0.168**	**0.116**	**2.16**	0.504	0.510	**1.38**	**0.483**	**0.650**

表4列出了不同规则向量表示策略在6种网络表示学习方法上的对比结果，通过分析可以总结出以下结论。

表4 不同规则向量表示策略的对比结果

Table 4 Result comparison of different rule vector representation strategies

规则因子表示策略	对比算法	酒驾(DUI)			痕迹不符(TD)			故意制造交通事故(ITA)		
		$Rank_{\min}$	MAP	$p@10$	$Rank_{\min}$	MAP	$p@10$	$Rank_{\min}$	MAP	$p@10$
算术平均策略	DeepWalk	39.33	0.078	0.033	2.71	0.491	0.524	2.63	0.494	0.513
	LINE	48.67	0.055	0.033	4.03	0.453	0.380	2.75	0.499	0.512
	Node2Vec	38.83	0.072	0.050	2.33	0.506	0.489	3.38	0.432	0.375
	GraRep	52.50	0.039	0.017	2.84	0.478	0.520	2.63	0.428	0.475
	N-NMF	72.67	0.036	0.017	3.25	0.468	0.524	1.75	0.502	0.525
	ProNE	63.33	0.034	0.017	2.77	0.473	0.549	2.87	0.479	0.500
加权的算术平均策略	DeepWalk	18.00↑	0.098↑	0.083↑	2.33↑	0.497↑	0.532↑	1.88↑	0.507↑	0.550↑
	LINE	22.67↑	0.069↑	0.067↑	3.41↑	0.469↑	0.402↑	2.00↑	**0.512**↑	0.600↑
	Node2Vec	16.17↑	0.099↑	0.083↑	2.28↑	**0.509**↑	0.515↑	2.38↑	0.436↑	0.475↑
	GraRep	32.50↑	0.044↑	0.033↑	2.31↑	0.485↑	0.537↑	1.88↑	0.504↑	0.563↑
	N-NMF	45.50↑	0.027↓	0.033↑	2.63↑	0.471↑	0.544↑	1.63↑	0.515↑	0.550↑
	ProNE	38.33↑	0.037↑	0.067↑	2.51↑	0.485↑	0.598↑	1.75↑	0.483↑	0.537↑
因子拼接策略	DeepWalk	13.00	0.132	0.067	2.99	0.489	0.551	1.88	0.456	0.487
	LINE	17.17	0.130	0.083	2.97	0.473	0.495	1.63	0.458	0.600
	Node2Vec	13.50	0.120	0.067	2.91	0.489	0.593	1.50	0.457	0.463
	GraRep	27.00	0.081	0.083	2.74	0.492	0.554	2.75	0.406	0.425
	N-NMF	27.67	0.096	0.083	2.73	0.493	0.597	3.00	0.421	0.475
	ProNE	26.83	0.104	0.083	2.74	0.492	**0.610**	2.38	0.434	0.512
加权的因子拼接策略	DeepWalk	8.33↑	0.158↑	0.100↑	2.61↑	0.501↑	0.505↓	1.63↑	0.471↑	0.585↑
	LINE	**6.67**↑	**0.168**↑	**0.116**↑	**2.16**↑	0.504↑	0.510↑	**1.38**↑	0.483↑	**0.650**↑
	Node2Vec	6.83↑	0.165↑	0.100↑	2.68↑	0.503↑	0.508↓	1.38↑	0.466↑	0.613↑
	GraRep	16.5↑	0.115↑	0.067↓	2.23↑	0.506↑	0.563↑	1.88↑	0.374↓	0.450↑
	N-NMF	16.5↑	0.115↑	0.067↓	2.23↑	0.506↑	0.563↓	1.88↑	0.374↓	0.450↓
	ProNE	16.5↑	0.115↑	0.067↓	2.23↑	0.506↑	0.563↓	1.88↑	0.374↓	0.450↓

(1)本文设计的基于规则因子的拼接策略相比算术平均策略在度量指标$Rank_{\min}$，MAP以及$p@10$的整体性能上更好。在酒驾(DUI)、痕迹不符(TD)、故意制造交通事故(ITA)3个数据集共9项的指标统计中，使用(加权)规则因子拼接策略的方法能取得7项最佳的效果，仅在痕迹不符(TD)与故意制造交通事故(ITA)的MAP指标上存在轻微的差距。分析发现，由于评估指标MAP考虑了所有规则的排序，当欺诈案件类型对应的规则数量已到达一定规模时，采用算术平均的规则向量表示策略更加契合该类指标的计算。

(2)无论是算术平均策略还是基于因子拼接策略，通过本文设计的加权方式，网络表示学习模型在整体性能上均有不小的提升。在算术平均策略的基础上使用因子加权之后，

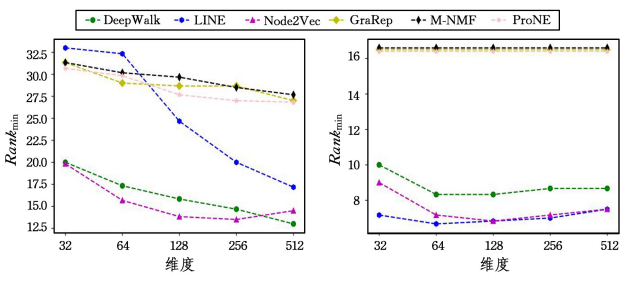
酒驾(DUI)数据集上的算法性能提升最为明显，$Rank_{\min}$平均提升了20个名次，*MAP*与$p@10$整体提升了20%。然而，在少数情况下，模型使用该加权策略后，*MAP*与$p@10$会出现轻微下降的情况。

(3)本文方法由于采用了加权拼接策略并改进了负采样方法，整体相比原始LINE模型在3个数据集上均有大幅度的提升。其中，指标$Rank_{\min}$提升了0.25～10.5个名次，*MAP*提升了5.5%～29.2%，$p@10$提升了3.0%～39.8%。

本文进一步分析了网络表示学习模型中规则因子不同维度设置(32维、64维、128维、256维、512维)对溯因预测结果的影响。图3－图5分别给出了网络表示学习模型基于拼接策略与加权拼接策略分别在酒驾(DUI)、痕迹不符(TD)、故意制造交通事故(ITA)数据集上不同维度下$Rank_{\min}$上的结果(由于节点的向量在网络表示学习模型中采用随机初始化，因此模型在最终预测结果上会存在一定浮动)。其中左侧标记为(a)的子图采用原始的规则因子拼接策略，右侧标记为(b)的子图采用的是加权后的规则因子拼接策略。通过对比曲线趋势，可以总结出以下结论。

(1)对于采用原始拼接策略的模型而言，规则因子向量维度设置为512维时，整体效果比其他维度更好。分析认为，在车险欺诈训练数据不足的情况下，向量维度越大具备刻画的语义信息能力越强。因此，规则的向量表示经因子向量拼接之后能够使得规则之间有更好的区分效果。尽管如此，当模型维度设置过大时，仍需要更多的训练数据形成相对“稠密”的网络才能使训练模型的损失函数收敛。这样，训练后的因子向量表示在溯因预测任务中才会有更好的性能表现。

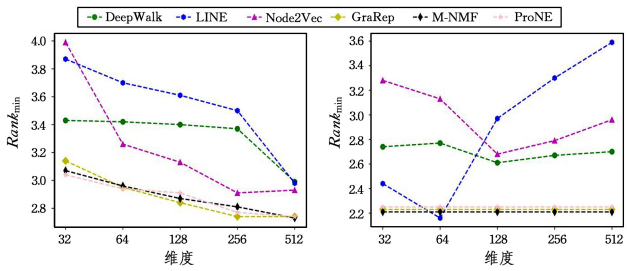
(2)本文提出的规则因子加权拼接策略可在一定程度上缓解训练样本不足的问题。可以发现，将网络表示学习模型的规则因子维度设置为32，64，128这些较低维度时，它们在$Rank_{\min}$的整体表现效果比原始的拼接策略更好。分析认为，在训练样本有限的情况下，较低维度的规则因子向量表示能够与因子的权重形成优势互补，通过规则因子的TF-IDF数值对欺诈案件中的触发因子进行加权，可以进一步区分案件与规则之间的差异性，从而使得模型在溯因预测任务中获得更好的效果。



(a)拼接策略 (b)加权拼接策略

图3 网络表示学习模型基于拼接策略与加权拼接策略分别在酒驾(DUI)数据集上不同维度下的$Rank_{\min}$结果

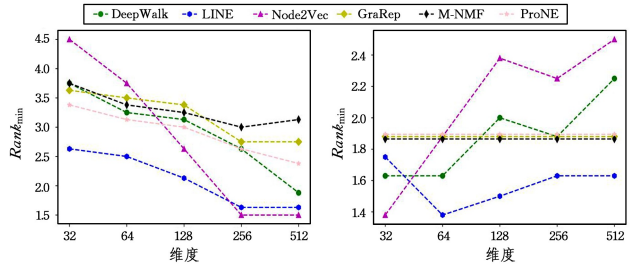
Fig. 3 $Rank_{\min}$ results of network embedding models with splicing strategy and weighted ones in different dimensions on DUI dataset



(a)拼接策略 (b)加权拼接策略

图4 网络表示学习模型基于拼接策略与加权拼接策略分别在痕迹不符(TD)数据集上不同维度下的$Rank_{\min}$结果

Fig. 4 $Rank_{\min}$ results of network embedding models with splicing strategy and weighted ones in different dimensions on TD dataset



(a)拼接策略 (b)加权拼接策略

图5 网络表示学习模型基于拼接策略与加权拼接策略分别在故意制造交通事故(ITA)数据集上不同维度下的$Rank_{\min}$结果

Fig. 5 $Rank_{\min}$ results of network embedding models with splicing strategy and weighted ones in different dimensions on ITA dataset

结束语 本文从车险欺诈案件核查的可解释性需求出发，针对现有方法存在的局限性，提出了一种基于网络表示学习的车险欺诈溯因分析方法。该方法首先给出了车险欺诈溯因分析任务的定义，然后基于已触发专家系统规则的欺诈案件数据与专家系统中的欺诈规则因子构建了案件-规则因子网络，采用网络表示学习模型学习得到了欺诈规则中因子的分布式向量表示。本文方法进一步基于溯因缺省原理设计了一种规则因子加权拼接策略，以此来更好地度量未触发专家系统规则的“欺诈”案件与专家系统中因子未全部触发规则之间的相似度。在酒驾、痕迹不符、故意制造交通事故等多个车险欺诈数据集上的实验结果表明，本文方法在$Rank_{\min}$，*MAP*以及$p@10$这3项评估指标上，整体的车险欺诈溯因任务效果均好于已有方法。

在未来工作中，将尝试在更多类型的车险欺诈案件中进行评估测试，并探究如何与知识图谱推理技术[46]进行融合，进一步提升模型效果。

参 考 文 献

[1] DIAO L, WANG N. Research on Premium Income Forecast Based on X12-GLSTM Model [J]. Computer Science, 2020, 47(S1): 512-516.

[2] YU W, FENG G F, ZHANG W J. A Research on Fraud Detec-

tion System and Gang Identification of Vehicle Insurance [J]. Insurance Studies, 2017(2): 63-73.

[3] ZHANG B Y, XIAO Y G, ZENG Y Z. A Comparative Study on Measuring Variable Importance in Auto Insurance Pricing—Based on Ensemble Learning and Generalized Linear Regression [J]. Insurance Studies, 2019(10): 73-83.

[4] ŠUBELJ L, FURLAN Š, BAJEC M. An expert system for detecting automobile insurance fraud using social network analysis [J]. Expert Systems with Applications, 2011, 38(1): 1039-1052.

[5] LU B, LI W, NA C, NIU Z, et al. Auto Insurance Fraud Detection with Machine Learning Models: A Survey [J]. Computer Engineering and Applications, 2022, 58(5): 34-49.

[6] DING M, LAN X, PENG R, et al. Progress and Prospect of Machine Reasoning [J]. Pattern Recognition and Artificial Intelligence, 2021, 34(1): 1-13.

[7] CONG Y, WANG Z, ZHU J, et al. Insights into Dataset and Algorithm Related Problems in Artificial Intelligence for Law [J]. Computer Science, 2022, 49(4): 74-79.

[8] VIAENE S, DEDENE G, DERRIG R A. Auto claim fraud detection using Bayesian learning neuralnetworks[J]. Expert Systems with Applications, 2005, 29(3): 653-666.

[9] PAYAM H, NEDA R P. A Data Mining Model for Risk Assessment and Customer Segmentation in the Insurance Industry[J]. International Journal of Strategic Decision Sciences(IJSDS), 2013, 4(1): 52-78.

[10] KAŠĆELAN L, KAŠĆELAN V, NOVOVI-BURIĆ M. A Data Mining Approach for Risk Assessment in Car Insurance: Evidence from Montenegro [J]. International Journal of Business Intelligence Research(IJBIR), 2014, 5(3): 11-28.

[11] LI Y, YAN C, LIU W, et al. A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification [J]. Applied Soft Computing, 2018, 70: 1000-1009.

[12] HE X, CHUA T S. Neural factorization machines for sparse predictive analytics [C] // Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 355-364.

[13] GUO J, LIU G, ZUO Y, et al. Learning sequential behavior representations for fraud detection [C] // Proceedings of the 2018 IEEE International Conference on Data Mining(ICDM). IEEE, 2018: 127-136.

[14] WANG R, FU B, FU G, et al. Deep & cross network for ad click predictions[C] // Proceedings of the ADKDD'17. 2017: 1-7.

[15] SUBUDHI S, PANIGRAHI S. Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection [J]. Journal of King Saud University-Computer and Information Sciences, 2020, 32(5): 568-575.

[16] MAJHI S K. Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection[J]. Evolutionary intelligence, 2021, 14(1): 35-46.

[17] GUO G Z, DUAN J X. Game analysis of insurance fraud [J]. Journal of Capital University of Economics and Business, 1999 (3): 51-54.

[18] LIU X H, JIN J L. The Insurance Fraud Game and Insurance Contract Based on Optimal Game Strategies [J]. System Engineering Theory and Practice, 2004(2): 19-24.

[19] ZHAO G Q, WU H. Is There Moral Hazard in Chinese Automobile Insurance Market? —— Evidence from Dynamic Renewal Data [J]. Journal of Financial Research, 2010(6): 175-188.

[20] TANG J, MO Y W. Construction of auto insurance anti-fraud system based on data mining technology [J]. Shanghai Insurance, 2013(11): 39-42, 63.

[21] WANG H W. A Research on Chinese Insurers' Moral Hazard Screening in Operation: From the Big Data Hadoop Clustering Analysis Technology Perspective [J]. Insurance Studies, 2016 (2): 59-67.

[22] YAN C, LI Y Q, SUN H T. A Research on Automobile Insurance Fraud Identification Based on Random Forest Model and Ant Colony Optimization Algorithm [J]. Insurance Studies, 2017(6): 114-127.

[23] YU W, FENG G F, ZHANG W J. A Research on Fraud Detection System and Gang Identification of Vehicle Insurance [J]. Insurance Studies, 2017(2): 63-73.

[24] XU X, WANG Z X, WANG M Q. The model and empirical study of motor vehicle insurance fraud identification based on deep learning technology [J]. Shanghai Insurance, 2019(8): 53-58.

[25] PANIGRAHI S, PALKAR B. Comparative analysis on classification algorithms of auto-insurance fraud detection based on feature selection algorithms[J]. International Journal of Computational Science and Engineering, 2018, 6(9): 72-77.

[26] HASSAN A K I, ABRAHAM A. Modeling insurance fraud detection using imbalanced data classification[C] // Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing. Cham: Springer, 2016: 117-127.

[27] PADHI S, PANIGRAHI S. Decision Templates based Ensemble Classifiers for Automobile Insurance Fraud Detection[C] // Proceedings of the 2019 Global Conference for Advancement in Technology(GCAT). IEEE, 2019: 1-5.

[28] KUANG K, LI L, GENG Z, et al. Causal Inference [J]. Engineering, 2020, 6(3): 107-130.

[29] CHEN R, JIANG Y F, LIN L. Study of Abductive Reasoning: State of the Art and Problems[J]. Computer Science, 2003(5): 25-27, 38.

[30] SUN J G, LIU R S, CHEN R. Abductive Diagnosis from Propositional Default Theories[J]. Journal of Jilin University(Science Edition), 1998(4): 34-38.

[31] LIN Y. The risk of motor vehicle and the principle of the study of the proxy commission in China[J]. Journal of Central South University(Social Sciences) 2006, 12(3): 274-278.

[32] HUANG P, LI J. Mining Model for the Insurance Retainment Rules Based on Rough Sets[J]. Journal of Shanghai Jiaotong University, 2004(4): 641-645.

[33] FAN X J. The Extraction of High Profitable Custom's Characteristics Based on Variable Precision Rough Set[J]. Journal of

Donghua University(Natural Science)，2004(3)：43-47.

[34] ZHU J Z. Research on data notification and protection mechanism of UBI auto insurance under networking[J]. Financial Regulation Research，2020(8)：102-114.

[35] TU C C，YANG C，LIU Z Y，et al. Network representation learning：an overview[J]. Scientia Sinica(Informationis)，2017，47(8)：980-996.

[36] CUI P，WANG X，PEI J，et al. A Survey on Network Embedding [J]. IEEE Transactions on Knowledge and Data Engineering，2019，31(5)：833-852.

[37] LIU X Y，TANG J. Network representation learning：A macro and micro view[J]. AI Open，2021(2)：43-64.

[38] TANG J，QU M，WANG M Z，et al. LINE：Large-scale Information Network Embedding[C]// Proceedings of the 24th International Conference on World Wide Web. ACM，Italy，2015：1067-1077.

[39] CHEN F，WANG Y C，WANG B，et al. Graph representation learning：a survey[J]. Transactions on Signal and Information Processing，2020，9(1)：e15.

[40] WANG Q，MAO Z，WANG B，et al. Knowledge graph embedding：A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering，2017，29(12)：2724-2743.

[41] PEROZZI B，AL-RFOU R，SKIENA S. Deepwalk：Online learning of social representations[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，USA. 2014：701-710.

[42] GROVER A，LESKOVEC J. node2vec：Scalable Feature Learning for Networks[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，USA. 2016：855-864.

[43] CAO S，LU W，XU Q. GraRep：Learning Graph Representations with Global Structural Information [C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management，Australia. 2015：891-900.

[44] XIAO W，CUI P，WANG J，et al. Community Preserving Network Embedding[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. USA，2017：203-209.

[45] JIE Z，DONG X Y，WANG Y，et al. ProNE：Fast and Scalable Network Representation Learning[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence，China. 2019：4278-4284.

[46] MA R X，LI Z Y，CHEN Z K，et al. Review of Reasoning on Knowledge Graph[J]. Computer Science，2022，49(S1)：74-85.

LI Weizhuo，born in 1989，Ph.D，associate professor，is a member of China Computer Federation. His main research interests include knowledge graph and insurance fraud an-alysis.

LU Bingjie，born in 1996，postgraduate. Her main research interests include deep learning and insurance fraud analysis.

（责任编辑：喻藜）